METHODS FOR ACCURATE AND EFFICIENT BAYESIAN ANALYSIS OF TIME SERIES

Agnieszka Borowska

ISBN 978 90 361 0561 3

Cover design: Crasborn Graphic Designers bno, Valkenburg a.d. Geul

This book is no. 742 of the Tinbergen Institute Research Series, established through cooperation between Rozenberg Publishers and the Tinbergen Institute. A list of books which already appeared in the series can be found in the back.

VRIJE UNIVERSITEIT

METHODS FOR ACCURATE AND EFFICIENT BAYESIAN ANALYSIS OF TIME SERIES

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor of Philosophy
aan de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. V. Subramaniam,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de School of Business and Economics
op woensdag 3 juli 2019 om 13.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door
Agnieszka Borowska
geboren te Ciechanowiec, Polen

promotor: prof.dr. S.J. Koopman
copromotor: dr. L.F. Hoogerheide

To my parents

Acknowledgements

First and foremost I would like to express my sincere gratitude to my thesis supervisors Siem Jan Koopman and Lennart Hoogerheide for their great support and guidance during my PhD study. I was very lucky to work with two so acknowledged and excellent promotors. I have always found Siem Jan's encouragement and confidence in me very motivating. His insightful comments allowed me to learn lot not only about time series but also research in general. Lennart's expertise in Bayesian methods, his energy, enthusiasm and patience helped me greatly in completing my PhD thesis. I am also grateful I could learn from him about teaching and supervision, he will always be a role model of an academic teacher to me.

I am also indebted to the members of my thesis committee, James Mitchell, Peter Schotman, Patrick Groenen, Jacques Commandeur and Anne Opschoor, for accepting the invitation, reading my thesis and providing valuable feedback.

I am deeply grateful to Ruth King for hosting me at the School of Mathematics and Statistics at the University of Edinburgh as a visiting PhD student and for her support since then. I have benefited a lot from our collaboration, which has been a highly enjoyable research experience to me.

I also owe a great deal of gratitude to Herman K. van Dijk for providing me with guidance and support, even prior my PhD. Our numerous meetings in Rotterdam were always very inspiring for me and helped me to deepen my understanding of various methodological problems.

I would like to extend my sincere thanks to Nalan Baştürk, Stefano Grassi and István Barra. I benefited a lot from our collaboration and many discussions, which helped me to structure my thoughts on a lot of research problems.

I would like to express my great appreciation to my current postdoc supervisor Dirk Husmeier, for his endless patience and understanding for me finishing this thesis. Thank you for turning a blind eye to my thesis being "almost done" for much loner than expected.

I would like to thank the Tinbergen Institute for their unique and versatile support, including academic, financial and administrative aid. I am especially thankful to Ester van den Bragt, Arianne de Jong, Judith van Kronenburg, Christina Månsson and Carolien Stolting for their positive attitude and helping me with various administrative problems, back during my MPhil at the Tinbergen Institute and during finishing my PhD.

Also, a big thank you to Charles Bos for motivating me to use the computing facilities at the Department and for helping me with these.

I also wish to thank my friends and colleges at the VU, Andries, Dieter, Ilka, Marc, Mengheng, Paolo, Shihao. You made my life much more enjoyable.

I will always be grateful to my parents, Mariola and Radosław, and my sister Joanna, who always have given me unconditional love and support throughout my entire life. I am forever indebted to you for giving me the conditions and opportunities that have allowed me to realise my goals.

Most of all, I would like to thank Lukasz for his love, support, encouragement and patience. Thank you that we have managed to overcome the long distance between us.

Glasgow, April 2019

Contents

Chapter 1

Introduction

This thesis investigates Bayesian inference over time series models with the emphasis put on applications in economics and finance. We note, however, that the methods developed are general and can be employed in various fields. We adopt simulation-based techniques which are necessary in any nontrival problem in this setting. The main motivation behind the presented research is to increase the efficiency and accuracy of these computationally intensive methods in several different contexts. One of the main topics addressed is efficient and precise risk estimation, or rare event analysis. Another problem studied below is the efficiency of various sampling algorithms, in particular importance sampling (IS) and Markov chain Monte Carlo (MCMC) algorithms. Finally, we address the issue of forecasting, from a single model as well as from a combination of models.

A Bayesian approach provides a flexible, coherent and convenient framework for the analysis of time series for a number of reasons, see Robert (2007, Ch. 11) for "a defence of the Bayesian choice". Of these, one of the most relevant in practice is capturing of parameter uncertainty. Treating parameters as random variables and inference based on posterior distributions allows us to easily deal with the task of uncertainty quantification. This is of particular importance in the context of risk analysis, where the objective is precise estimation of rare events and where even a small degree of incorrectness might have tremendous and serious consequences (Chapters 2 and 3). Another advantage of a Bayesian approach is that it highly facilitates dealing with complex data and combining information stemming from different sources (Chapters 4 and 5). A related aspect is that standard Bayesian methods can be easily and naturally extended to develop hybrid approaches or schemes exceeding typical inference problems to allow e.g. for built-in optimisation elements (Chapters 3, 4 and 5).

1.1 Time series

The problem of identification of properties of processes evolving in time to characterise the observed patterns and to make predictions about their future realisations is ubiquitous in science. In economics and finance time series analysis has a well-established position and it comes naturally given the studied phenomena such as business cycle or stock market trends. There have been several approaches to investigate time-series, depending on the research area, e.g. signal processing approach (including spectral analysis) in physics and engineering; function mapping approach (including Gaussian processes and neural networks) in machine learning; statistical and econometric approach prevailing in economics and finance. We adopt on the latter methodology as it provides a useful explanation of the nature of the modelled processes and allows for a structural interpretation of the estimation results.

As far as the statistical analysis of time series is concerned, we focus on time-varying parameter models, which following Cox (1981) can be grouped into two classes of models with distinct advantages and disadvantages: observation-driven models and parameter-driven models. The former specify model parameters as deterministic functions of observations allowing for a perfect one-step-ahead predictability of the parameters given the current information set. The latter allow for idiosyncratic innovations governing the parameters dynamics and can be represented as state space models, see Durbin and Koopman (2012) for an extensive exposition of the state space methodology. In consequence, the likelihood of an observation-driven model is available in closed form while the likelihood of a parameter-driven model is typically analytically intractable. This makes observation-driven models easier to work with and faster to estimate but also gives extra flexibility and an intuitive structure to parameter-driven models. Well known examples of observation-driven models in econometrics include the generalized autoregressive conditional heteroskedasticity (GARCH) model of Engle (1982) and Bollerslev (1986) as well as the more recent generalized autoregressive score (GAS) models of Creal et al. (2013); a quintessential parameter-driven model in econometrics is the stochastic volatility (SV) model of Taylor (1994), with other important instances being the dynamic factor model of Geweke (1977), see also Stock and Watson (2002), and the factor-augmented vector autoregression (FAVAR) model of Bernanke et al. (2005) and Stock and Watson (2005). Given distinct merits of both classes we do not see it necessary to limit our attention to one certain class in advance. Depending on the main idea and the goal of each chapter we consider a specification which is more natural and convenient in the given context.

1.2 Bayesian inference

As mentioned above, we adopt a Bayesian approach to statistical inference by which we understand estimation and prediction from a given model (as well as model comparison and selection). We refer to Robert (2007) for an in-depth exposition of the Bayesian principles including philosophical foundations of this inference paradigm and to Gelman et al. (2013) for a more practical treatment adhering to common-sense merits of Bayesian thinking. It is worth mentioning that Gelman et al. (2013) understand Bayesian methods more broadly, as consisting of three steps: (1) model building, (2) inference conditional on the model, (3) model checking. To our view their points (2) and (3) both belong to the inference problem, while step (1) does not necessarily need to be related to "Bayesian" statistics, in the sense that Bayesian reasoning can be applied to any model (even to a non-statistical one). The strength of the Bayesian paradigm is that its basic principles and rules are universal, regardless of the exact model specification, in particular the choice between observation-driven and parameter-driven models in our case. In this work we are less concerned with the modelling stage of data analysis but rather with developing accurate and efficient techniques to enhance Bayesian inference over existing models.

The key principle of Bayesian statistics is conditioning on the observed data $\mathbf{y}$, which is formalised by the *likelihood principle*[1]. Hence, the observed data are seen as fixed, which stays in contrast to classical statistics treating the data as random realisations of a sampling process. On the other hand, Bayesian statistics sees the model parameters (and all unobserved quantities) $\boldsymbol{\theta}$ as random variables and makes statements about them in terms of probability distributions – differently than frequentist reasoning which assumes that parameters are fixed. Further, Bayesian analysis allows for incorporating prior beliefs about these unknown quantities, e.g. expert knowledge or information stemming from complimentary sources – but also lack of any knowledge – into the inference process via the *prior distribution* $p(\boldsymbol{\theta})$. After recording the data, these initial beliefs are updated using the *likelihood* $p(\mathbf{y}|\boldsymbol{\theta})$, or the data distribution, to form the *posterior distribution* $p(\boldsymbol{\theta}|\mathbf{y})$. This step is formalised by applying the Bayes' theorem

[1]The likelihood principle states that for inference on an unknown parameter all of the evidence from any observation is entirely contained in the likelihood function of this observation, see Robert (2007, p. 15–16) and Gelman et al. (2013, p. 6).

to obtain the relationship between these three distributions as follows

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{1.2.1}$$

$$\propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{1.2.2}$$

The denominator of (1.2.1) is called the *marginal likelihood* and is a normalising constant (as it does not depend on $\boldsymbol{\theta}$). It is often analytically intractable and even hard to estimate, hence one typically needs to use the unnormalised posterior distribution (1.2.2), also known as the *kernel* of the posterior distribution. The posterior distribution is then used to obtain estimators of $\boldsymbol{\theta}$, which in the Bayesian setting are formally represented as integrals (see Robert and Casella, 2004, Ch. 1.2). This implies the associated computational problem of integration, often in high dimensions, which turns out not analytically solvable in many practical applications. For this reason most of Bayesian analysis is concerned with simulation-based computations, which frequently are computationally intensive.

1.3 Simulation methods and numerical efficiency

The Bayesian paradigm is conceptually clear and intuitive however it was not until the "*Markov chain Monte Carlo* revolution" in the 1990s that it has gained broader popularity, see Robert (2007) and Robert and Casella (2011). Prior to that Bayesian analysis was mostly limited to the use of conjugate[2] prior distributions (Green et al., 2015). It is crucial to point out that even though the popularisation of the MCMC methods in statistics and econometrics has triggered the growing acceptance and usage of Bayesian methods in these fields, current Bayesian computations are not necessarily limited to MCMC algorithms. Alternatives to MCMC include IS (Hammersley and Handscomb, 1964), sequential Monte Carlo (SMC, Doucet et al., 2001), approximate Bayesian computation (ABC, Marin et al., 2012) or variational Bayes (VB, Blei et al., 2017). All these methods but the last one are sampling-based, predominantly adopting Monte Carlo (MC) methods. A detailed treatment of MC techniques is beyond the scope of this chapter and we refer to Robert and Casella (2004) for a comprehensive examination of this broad area as well as to Green et al. (2015) for a review of the history and the current state of Bayesian computations.

[2]A family of prior distributions is said to be conjugate for the likelihood function if the resulting posterior distribution belongs to this family. Conjugacy allows for obtaining the posterior distribution simply by updating of the hyperparameters of the prior distribution.

Amongst the above mentioned methods for Bayesian computations IS deserves a special mention, as we argue below. IS is a relatively old technique dating back to Kahn and Marshal (1953) and Marshall (1956), with a long tradition in econometrics originating from Kloek and van Dijk (1978). Not only serves it as a building block for the SMC methods and provides intuition for MCMC algorithms, but also it can be used as a variance reduction method. Variance reduction is of key interest in risk analysis and Chapters 2 and 3 are concerned with this topic.

Even though performance capabilities of modern computers have been continuously increasing and the limitations in computations faced in the past are less of an issue nowadays, numerical accuracy and efficiency are still of major interest in statistics and econometrics. One of the reasons for this is that currently decisions need to be made quicker than ever, often in real time, whether it is for central bankers, stock market analysts or financial institution managers. Algorithms such as IS, MCMC or SMC are theoretically sound and deliver *exact* estimates when the number of simulations diverges to infinity. However, they all require a distribution to sample from, known as an importance, proposal or candidate distribution. The quality of this distribution is crucial for the ultimate performance of the algorithm. Hence different variants of the same algorithm, only differing with respect to the choice of this sampling distribution, are likely to result in dissimilar outcomes in terms of the uncertainty of the associated estimator. Therefore, methods which are reliable only "in the limit" (after excessively long simulation runs) are not suited for the purpose of real time decision making, with fast and precise methods being naturally preferred. We address the issue of constructing efficient sampling based algorithms – in different contexts – in all the chapters of this thesis, in particular in Chapter 2 and 4.

Furthermore, typical modelling and inference methods are designed to explain *average* scenarios. Since "all models are wrong", as famously stated by George Box, we cannot expect that our inference *conditional* on a model will be fully accurate. Nevertheless, because "some models are useful" though, we can aim at finding valuable aspects of a model or models at hand, while hedging against its or their potential shortcomings. For instance, suppose our ultimate goal is inference over a particular region of the posterior distribution, such as its tail. We then suggest to still use standard models due to their well-documented ability to capture stylised facts of the data, but simply in a problem-adjusted manner, with a tail-focused estimation. Chapter 3 is dedicated to this problem. Alternatively, suppose that our aim is to construct a profit-maximising portfolio while still caring about the associated investment risk. We cannot expect that there is a universally dominating (e.g. over time) single model and a single invest-

ment strategy but we can construct the portfolio based on an appropriately specified combination of models and strategies. We discuss and illustrate this idea in Chapter 5.

1.4 Thesis outline

This thesis consists of four self-contained chapters all related to Bayesian inference in time-series models. Below we present their overview.

Chapter 2 is titled "Bayesian Risk Evaluation for Long Horizons" and is based on joint work with Lennart Hoogerheide and Siem Jan Koopman. We present an accurate and efficient method for Bayesian estimation of two financial risk measures, Value-at-Risk and Expected Shortfall, for a given volatility model. We obtain precise forecasts of the tail of the distribution of returns not only for the 10-days-ahead horizon required by the Basel Committee but even for long horizons, like one-month or one-year ahead. The latter has recently attracted considerable attention due to the different properties of short term risk and long run risk. Precise forecasts of the tail of the distribution can also be useful for option pricing. The key insight behind our proposed IS based approach is the sequential construction of marginal and conditional importance densities for consecutive periods. For robustness, these importance densities are efficiently constructed as mixtures of Student's t densities. By oversampling the extremely negative scenarios and giving them lower importance weights, we achieve a much higher precision in characterising the properties of the left tail. We report substantial accuracy gains for all the considered horizons in empirical studies on two datasets of daily financial returns, including a highly volatile period of the recent financial crisis. We analyse two workhorse models used by financial practitioners, GARCH(1,1)-t and GAS(1,1)-t. To illustrate the flexibility of the proposed construction method, we present how it can be adjusted to the frequentist case, for which we provide counterparts of both Bayesian applications.

Chapter 3 is titled "Partially Censored Posterior for Robust and Efficient Risk Evaluation" and is based on joint work with Lennart Hoogerheide, Siem Jan Koopman and Herman K. van Dijk. We introduce a novel approach to inference for a specific region of the predictive distribution. An important domain of application is accurate prediction of financial risk measures, where the area of interest is the left tail of the predictive density of logreturns. Our proposed approach originates from the Bayesian approach to parameter estimation and time series forecasting, however it is robust in

the sense that it provides a more accurate estimation of the predictive density in the region of interest in case of misspecification. The first main contribution of this chapter is the novel concept of the Partially Censored Posterior (PCP), where the set of model parameters is partitioned into two subsets: for the first subset of parameters we consider the standard marginal posterior, for the second subset of parameters (that are particularly related to the region of interest) we consider the conditional censored posterior. The censoring means that observations outside the region of interest are censored: for those observations only the probability of being outside the region of interest matters. This approach yields more precise parameter estimation than a fully censored posterior for all parameters, and has more focus on the region of interest than a standard Bayesian approach. The second main contribution is that we introduce two novel methods for computationally efficient simulation: Conditional MitISEM, an MCMC method to simulate model parameters from the Partially Censored Posterior, and PCP-QERMit, an IS method that is introduced to further decrease the numerical standard errors of the Value-at-Risk and Expected Shortfall estimators. The third main contribution is that we consider the effect of using a time-varying boundary of the region of interest, which may provide more information about the left tail of the distribution of the standardized innovations. Extensive simulation and empirical studies show the ability of the introduced method to outperform standard approaches.

Chapter 4 is titled "Semi-Complete Data Augmentation for Efficient State Space Model Fitting" and is based on joint work with Ruth King. We propose a novel efficient model-fitting algorithm for state space models. State space models are an intuitive and flexible class of models, frequently used in practice. This flexibility, however, often comes at the price of substantially more complicated fi

tting of such models to data due to the associated likelihood being analytically intractable. For the general case a Bayesian data augmentation approach is often employed, where the true unknown states are treated as auxiliary variables and imputed within the MCMC algorithm. However, standard "vanilla" MCMC algorithms may perform very poorly due to high correlation between the imputed states and/or parameters, leading to the need for specialist algorithms. The proposed method circumvents the inefficiencies of traditional approaches by combining data augmentation with numerical integration in a Bayesian hybrid approach. This approach permits the use of standard "vanilla" updating algorithms that perform considerably better than the traditional approach in terms of considerably improved mixing and hence lower autocorrelation. We use the proposed Semi-Complete Data Augmentation algorithm in different application areas and associated types of models, leading to distinct imple-

mentation schemes and demonstrating efficiency gains in empirical studies.

Chapter 5 is titled "Forecast Density Combinations of Dynamic Models and Data Driven Portfolio Strategies" and is based on Baştürk, Borowska, Grassi, Hoogerheide, and van Dijk (2018). We propose a novel dynamic asset allocation approach in which model-based forecasts are directly combined with a set of data driven portfolio strategies, without the necessity to define a utility or other scoring function. The specification of the underlying models is motivated by findings of a scrupulous analysis of typical stylized facts of the time series of monthly returns of ten US industries over the period 1926M7–2015M6. The portfolio strategies are based on the practice in financial investing to take advantage of a positive or negative momentum in industry returns. In probabilistic terms, the resulting dynamic asset-allocation model is specified as a combination of return distributions stemming from multiple pairs of models and strategies. The combination weights are defined through feedback mechanisms that enable learning, to allow for cross-correlation and correlation over time. We base our Bayesian inference over the proposed model on its representation as a nonlinear non-Gaussian state space model. To increase the efficiency and robustness of the simulations we introduce a new nonlinear filter based on mixtures of Student's t distributions. Diagnostic analysis of posterior residuals gives insight into the model and strategy incompleteness or misspecification. An extensive empirical application reveals that a combination of a smaller set of flexible models outperforms a larger combination of basic model structures in terms of expected return and risk. We believe that dynamic patterns in combination weights and diagnostic learning provide useful signals from a risk-management perspective and can help enhancing modelling and policy.

Chapter 6 summarises the main findings and concludes the thesis.

Chapter 2

Bayesian Risk Evaluation for Long Horizons

The global financial crisis stressed the importance of appropriate risk management which requires the accurate prediction of the market risk related to fluctuations of stock or index prices. It also emphasised the necessity of precise prediction of the long-term financial risk: as noted by The Volatility Laboratory (2012)[1], the turbulent events of 2008 moved the focus of risk management from solely short term horizons to the longer ones. This is because most portfolios consist of assets that are held longer than just a few days, so that e.g. excess leverage is likely to pose a much higher risk in the long run than in the short run (Engle, 2009). Hence, increased attention has been recently devoted to risk measurement for one-month-ahead or even one-year-ahead horizons, and not only the standard, 1-day-ahead or 10-days-ahead measures required by the Basel Committee on Banking Supervision (1995).

One of the potential reasons why the main focus was previously on short-run measures is the difficulty of obtaining precise evaluations of risk for long horizons. As noted by McNeil et al. (2015) and Embrechts et al. (2005), an obvious approach to long-term risk evaluation where the so-called scaling rule is applied to short term risk measures might be inappropriate[2]. Furthermore, Christoffersen et al. (1998) state that generally

[1] As it describes itself, the Volatility Laboratory (V-Lab) of the Volatility Institute provides real time measurement, modelling and forecasting of financial volatility, correlations and risk for a wide spectrum of assets and it produces volatility forecasts up to a year in advance. The Volatility Institute was created at New York University Stern School of Business in 2009 under the direction of Robert Engle.

[2] The performance of the scaling rule crucially depends on the data generating process, in particular its "closeness" to a normal random walk model, where indeed a quantile of H-days-ahead distribution is given by the quantile of the 1-day-ahead distribution multiplied by $\sqrt{H}$ (see Daníelsson and Zigrand,

conventional parametric models are ill-suited for extreme events analysis because they focus on "average" scenarios in order to obtain a high goodness of fit. This misperformance may be even more severe when the horizon of analysis increases.

McNeil and Frey (2000) distinguish three main approaches for computing tail related measures: non-parametric historical simulations (HS); parametric methods based on an econometric model where the volatility dynamics are explicitly specified; methods based on extreme value theory (EVT). They argue that a parametric model of volatility is essential in order to capture the volatility dynamics exhibited by financial returns, which allows for prediction of risk based on the current volatility background. Moreover, parametric time series models provide a framework to extrapolate the analysis beyond the observed data – as opposed to the HS methods. For these reasons it is a natural starting point for our analysis to build upon parametric methods from the second group. As the main drawback of these models McNeil and Frey (2000) indicate their common conditional normality assumption, which seems to be invalid for most financial series. Hence, they apply EVT to estimate extreme quantiles of the distribution of the standardised residuals from a normal generalized autoregressive conditional heteroskedasticity (GARCH) model. The EVT approach for capturing the properties of extreme tails was also suggested by Christoffersen et al. (1998).

In this chapter we decide to proceed differently: in order to address the issue of precise long-run risk evaluation we build upon the approach of Hoogerheide and van Dijk (2010). These authors suggest evaluation of the probability distribution of extreme events via importance sampling (IS) based on a specially designed importance density focusing on the left tail, for a given volatility model. To cope with heavy tails of conditional return distributions we consider volatility models with Student's t distributed error terms. We propose an accurate and efficient approach to forecasting two standard measures of market risk, Value-at-Risk (VaR) and Expected Shortfall (ES), in a situation when the prediction horizon is long, e.g. 40, 100 or 250 days ahead. The latter is a noticeable contribution compared to Hoogerheide and van Dijk (2010), who proposed a method suited for standard short-run analysis[3]. To this end we first redesign the original approach of Hoogerheide and van Dijk (2010) using a more flexible approximation algorithm. Second, we suggest a novel sequential construction of the importance density which allows for "guiding" of the subsequent simulated returns over

2006; Diebold et al., 1997).

[3]This limitation was pointed out by the cited authors themselves, as they note that the relative performance of their method may decrease with the prediction horizon length, due to the so-called "curse of dimensionality of importance sampling", and is likely to vanish for very long horizons, such as 100-days-ahead.

time so that the cumulative return falls in the "high-loss" region, so that analysis of long horizons becomes feasible. In our approach the properties of the subsequent conditional importance densities depend on the previous simulated returns in the sense that at each step we take into consideration the cumulative return up to that time point. This allows us to assess how much the situation still needs to deteriorate in order to qualify for being a "high-loss" scenario. We focus on the 99% quantile of the profit-loss distribution, as required by the Basel Committee on Banking Supervision (1995); such an extreme tail is also more challenging to precisely predict than e.g. the 95% quantile, which is also commonly analysed.

It is important to stress that our method is universal, i.e. it can be applied for any chosen parametric volatility model. Hence, we abstract from the issue of model selection, but aim at a precise and efficient evaluation of risk implied by the given model. Nevertheless, our method is still highly advantageous in the context of model selection because by reducing the uncertainty related to the simulation noise the comparison between models is more likely to be based on their "true" quality.

As a variance reduction technique, IS has been already applied in the context of market risk evaluation. Importantly, Glasserman et al. (1999), Glasserman et al. (2000) and Glasserman et al. (2002) combine IS with stratified sampling to obtain precise estimates of VaR. They, however, do not consider time series models and carry out barely a "numerical example", not an empirical study with real data. Furthermore, they restrict their attention to a 10-days-ahead horizon and analyse portfolio loss probabilities from the frequentist perspective. However, risk forecasting, and in particular for long horizons, is subject to a considerable parameter uncertainty. That is why the Bayesian approach seems to be particularly suited for long-run risk analysis. In addition, not only it naturally captures parameter uncertainty but also provides a convenient starting point for considering model uncertainty via Bayesian model averaging. Therefore we follow Hoogerheide and van Dijk (2010) and focus primarily on the analysis from the Bayesian perspective. However, to illustrate the merits and the flexibility of the proposed method, we demonstrate how the method can be adjusted to the frequentist case, for which we provide the counterparts of the Bayesian applications.

The outline of the chapter is as follows. In Section 2.1 we first recall the approach of Hoogerheide and van Dijk (2010) to show how IS can be applied in the context of risk evaluation; second, we present how our proposed method allows to mitigate the "curse of dimensionality", inherent to IS, to allow for more accurate and efficient long run VaR and ES forecasts. We illustrate the performance of our novel method in Section 2.2 with two workhorse models, commonly used by practitioners, i.e. GARCH(1,1)-t and

GAS(1,1)-t. In Section 2.3 we consider the alternative, frequentist method for long run prediction of VaR and ES: we discuss the necessary methodology modifications and provide the frequentist counterparts of the Bayesian applications from Section 2.2. Section 2.4 concludes and presents an outline for the further research.

2.1 Bayesian risk evaluation using importance sampling

Let $\{y_t\}_{t\in\mathbb{Z}}$ be a time series of daily logreturns $y_t = 100(\log p_t - \log p_{t-1})$ on a financial asset with price p_t at the end of day t, with $\mathbf{y}_{1:T} := \{y_1, \ldots, y_T\}$ denoting the observed data. We assume that $\{y_t\}_{t\in\mathbb{Z}}$ is subject to a dynamic stationary process parametrised by $\boldsymbol{\theta}$, on which we put a prior $p(\boldsymbol{\theta})$. Let $\mathbf{y}^*_{1:H} = \{y_{T+1}, \ldots, y_{T+H}\}$ denote the vector of H future returns and consider the posterior predictive distribution of profit/loss $PL(\mathbf{y}^*_{1:H}) = 100\left[\exp\left(\sum_{t=T+1}^{T+H} y_t/100\right) - 1\right]$ (converting the sum of the logreturns to the percentage return) defined as

$$p\left(PL(\mathbf{y}^*_{1:H})|\mathbf{y}_{1:T}\right) = \int p\left(PL(\mathbf{y}^*_{1:H})|\mathbf{y}_{1:T}, \boldsymbol{\theta}\right) p(\boldsymbol{\theta}|\mathbf{y}_{1:T}) d\boldsymbol{\theta}, \tag{2.1.1}$$

obtained by marginalisation over the parameter with respect to the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$. We are interested in Bayesian estimation of the $100\alpha\%$ VaR, i.e. the $100(1-\alpha)\%$ quantile of the posterior predictive distribution of profit/loss within a horizon of the next H trading days, i.e.

$$100\alpha\%\,\text{VaR} = \inf\left\{x \in \mathbb{R} :\, \mathbb{P}(PL(\mathbf{y}^*_{1:H}) \geq x|y_{1:T}) \geq \alpha\right\}.$$

We also consider ES as an alternative risk measure, due to its advantageous properties compared to VaR, mainly sub-additivity (which makes ES a coherent risk measure in the sense of Artzner et al., 1999). Given $100\alpha\%$ VaR, the conditional ES is defined as

$$100\alpha\%\,\text{ES} = \mathbb{E}\big[PL(\mathbf{y}^*_{1:H})|PL(\mathbf{y}^*_{1:H}) < 100\alpha\%\,\text{VaR}\big].$$

Since (2.1.1) is usually analytically intractable, simulation based methods need to be applied in order to estimate VaR and ES. Following Hoogerheide and van Dijk (2010) we distinguish two approaches to that. The first one, which we will refer to as the *direct approach*, is straightforward:

1. draw a sample of model parameter $\boldsymbol{\theta}^{(i)}$, $i = 1, \ldots, M$, from the posterior distribution (using e.g. the Metropolis-Hastings algorithm);
2. generate the corresponding paths of H future log-returns $\mathbf{y}^{*(i)} = \{y_{T+1}^{(i)}, \ldots, y_{T+H}^{(i)}\}$;
3. compute the resulting profits/losses $PL(\mathbf{y}^{*(i)})$;
4. sort in ascending order the values of $PL(\mathbf{y}^{*(i)})$ to obtain the permutation $PL^{(j)}$, $j = 1, \ldots, M$;
5. obtain the $100\alpha\%$ VaR and ES as

$$\widehat{VaR}_{DA} = PL^{((1-\alpha)N)}, \tag{2.1.2}$$

$$\widehat{ES}_{DA} = \frac{1}{(1-\alpha)N} \sum_{j=1}^{(1-\alpha)N} PL^{(j)}. \tag{2.1.3}$$

The Volatility Laboratory (2012) uses this direct approach for the non-Bayesian evaluation of long-run VaR, where step 1 is replaced by frequentist estimation. The drawback of the direct approach is that it is subject to an inherent problem of rare events simulations, i.e. that most of the generated scenarios are not the ones of the ultimate interest, the extremely negative ones. This makes direct estimators very inefficient and the only way to increase their precision is to consider many more draws. Obviously, the latter is costly, in terms of both computing time and computing resources (e.g. the available memory).

To illustrate the problem, let us introduce a toy example of white noise returns[4]

$$\begin{aligned} y_t &\sim \sqrt{\sigma^2}\varepsilon_t \qquad \varepsilon_t \sim \mathcal{N}(0,1), \\ \sigma^2 &\sim p(\sigma^2), \end{aligned}$$

where p is a conjugate prior distribution. Then, the future profits/losses follow $PL(\mathbf{y}^*_{1:H}) \sim \mathcal{N}(0, H\sigma^2)$. If we treat σ^2 as known and equal to 1, i.e. under the assumption that the data were generated from a standard normal distribution, the value for the 10-days-ahead 99% VaR is given by $\Phi^{-1}(0.01)\sqrt{10} = -7.3566$, for 100-days-ahead it is equal to -23.2635, while for 250-days-ahead to -36.7828. Figure 2.1.1 presents the outcome of the direct approach for the shortest horizon of 10-days-ahead. One can see that – as

[4]In this example we consider for simplicity the cumulative logreturn over $H = 10$ days (instead of the percentage return), so that the profit/loss is just the sum of the H logreturns, i.e. $PL(\mathbf{y}^*_{1:H}) := \sum_{h=1}^{H} y^*_h$.

discussed above – only a very small fraction of roughly 1/100 of the generated paths corresponds to the high losses that we are interested in, which indeed leads to a low efficiency.

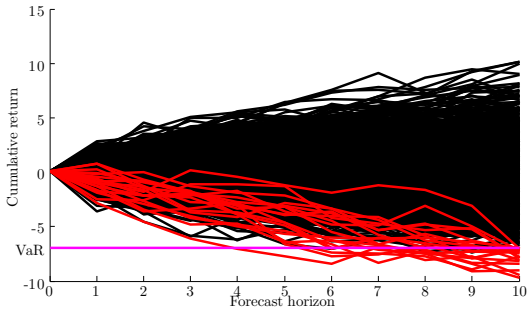


Figure 2.1.1: Direct simulation results in very few paths (the red ones) below the 99% VaR value (the violet horizontal line). White noise returns, 10-days-ahead horizon, simulated 10,000 paths.

2.1.1 Tail focused importance density

To overcome the inefficiency of the direct approach, Hoogerheide and van Dijk (2010) suggest importance sampling (IS), a well known variance reduction technique. Its main merit is the potential focus on the *important* subspace by adopting an appropriate sampling density, which in the context of VaR and ES should be tail-focused. Hoogerheide and van Dijk (2010) propose the Quick Evaluation of Risk using Mixture of t approximations (QERMit) algorithm, where the key idea is to oversample the high-loss scenarios and to give them lower importance weights. The theoretical insight for their method comes from the properties of the optimal importance density for the Bayesian estimation of $\bar{f} \equiv \mathbb{E}[f(X)]$ for a variable X with density kernel $p(x)$, outlined by Geweke (1989)[5], which is given by

$$q_{opt}(x) \propto |f(x) - \bar{f}|p(x), \tag{2.1.4}$$

provided that $\mathbb{E}[|f(X) - \bar{f}|] < \infty$. For the case of $f(x) = \mathbb{I}_S(x)$, i.e. the indicator function of the set S, we have

$$\mathbb{E}[f(X)] = \mathbb{P}[X \in S] =: \bar{p}$$

[5]Here, the optimality refers to minimisation, given the specified number of draws, of the numerical standard error of the IS estimator of $\bar{f} \equiv \mathbb{E}[f(X)]$, where f is the function of interest of the random variable X, which has the density $\tilde{p}(x)$ with the kernel $p(x)$.

and the optimal importance density is given by

$$q_{opt}(x) \propto \begin{cases} (1-\bar{p})p(x), & \text{for } x \in S \\ \bar{p}p(x), & \text{for } x \notin S \end{cases}, \quad \text{or} \quad q_{opt}(x) = \begin{cases} c(1-\bar{p})\tilde{p}(x), & \text{for } x \in S \\ c\bar{p}\tilde{p}(x), & \text{for } x \notin S \end{cases},$$

where c is a constant, which results in[6]

$$\int_{x \in S} q_{opt}(x)dx = \int_{x \notin S} q_{opt}(x)dx = \frac{1}{2}. \tag{2.1.5}$$

Condition (2.1.5) implies that half of the total probability mass of the importance distribution shall be located in the region of interest S, and the remaining half outside that region. Such a split is the consequence of using only the kernel of the target distribution and not its proper density, which makes it necessary to adequately normalise the importance weights via sampling from the whole domain instead of merely sampling high loss scenarios, which is the optimal method in the frequentist approach that we consider in the sequel of this chapter.

Hoogerheide and van Dijk (2010) apply the above result in the context of VaR and ES estimation. Then, S is interpreted as the "high loss region", i.e. the subspace of the profits/losses space with the $100(1-\alpha)\%$ lowest values, while the optimal importance density prescribes that 50% of draws shall represent high losses while the other 50% the remaining profit/loss realisations. Figure 2.1.2 illustrates the construction of the optimal importance density for the VaR estimation.

Notice that in the case of Bayesian estimation of VaR and ES we have a joint density $p(\boldsymbol{\theta}, \mathbf{y}^*_{1:H}|\mathbf{y}_{1:T})$ of the parameters $\boldsymbol{\theta}$ and future returns $\mathbf{y}^*_{1:H}$ of which we have kernel

$$p(\boldsymbol{\theta}, \mathbf{y}^*_{1:H}|\mathbf{y}_{1:T}) \propto p(\boldsymbol{\theta})p(\mathbf{y}_{1:T}|\boldsymbol{\theta})p(\mathbf{y}^*_{1:H}|\boldsymbol{\theta}, \mathbf{y}_{1:T}),$$

the product of the posterior density kernel and the future returns' density. The IS estimator $\widehat{VaR}_{IS}$ of the $100(1-\alpha)\%$ VaR is obtained by solving x in

$$\mathbb{P}[\widehat{PL(\mathbf{y}^*_{1:H})} \leq x]_{IS} = 1 - \alpha,$$

[6]This is obtained by noting that

$$\int_{x \in S} q_{opt}(x)dx = c(1-\bar{p}) \int_{x \in S} \tilde{p}(x)dx = c\bar{p}(1-\bar{p}) = c\bar{p} \int_{x \notin S} \tilde{p}(x)dx = \int_{x \notin S} q_{opt}(x)dx,$$

while $\int_{x \in S} q_{opt}(x)dx + \int_{x \notin S} q_{opt}(x)dx = 1$.

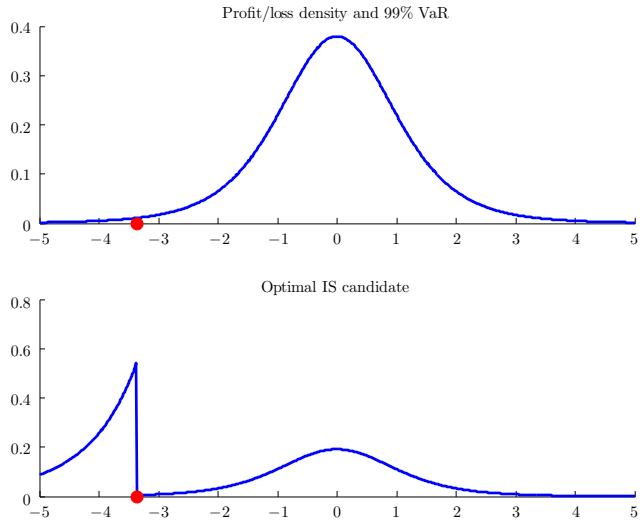


Figure 2.1.2: Construction of the optimal importance density. Exemplary density function (Student's t with 5 degrees of freedom) of profit/loss and the implied 99% VaR (top). The optimal importance density for the VaR estimation (bottom).

which in practice can be done via the following procedure:

1. draw a sample of parameter vectors $\boldsymbol{\theta}^{(i)}$ and corresponding future returns $\mathbf{y}_{1:H}^{*(i)}$, $i = 1, \ldots, M$, from their joint importance density $q(\boldsymbol{\theta}^{(i)}, \mathbf{y}_{1:H}^{*(i)} | y_{1:T})$;
2. compute the corresponding importance weights $w^{(i)} = \frac{p(\boldsymbol{\theta}^{(i)}, \mathbf{y}_{1:H}^{*(i)} | \mathbf{y}_{1:T})}{q(\boldsymbol{\theta}^{(i)}, \mathbf{y}_{1:H}^{*(i)} | \mathbf{y}_{1:T})}$, $i = 1, \ldots, M$;
3. compute the resulting profits/losses $PL(\mathbf{y}_{1:H}^{*(i)})$;
4. sort in ascending order the values of $PL(\mathbf{y}_{1:H}^{*(i)})$ to obtain the permutation $PL^{(j)}$, $j = 1, \ldots, M$, with the corresponding weights $w^{(j)}$;
5. set $\widehat{VaR}_{IS}$ as $PL^{(k)}$ for which

$$\sum_{j=1}^{k} w^{(j)} \leq 1 - \alpha \quad \text{and} \quad \sum_{j=1}^{k+1} w^{(j)} > 1 - \alpha,$$

and given $\widehat{VaR}_{IS}$

$$\widehat{ES}_{IS} = \sum_{j=1}^{k} w^{(j)} PL^{(j)} \Bigg/ \sum_{j=1}^{k} w^{(j)}.$$

2.1.2 Approximations by mixtures of Student's t distributions

The choice of the importance density is crucial for the performance of the IS estimation. Clearly, as pointed out by Geweke (1989), the importance density should resemble the

target density and at the same time remain easy to sample from. Moreover, the tails of the importance density need to be thicker than those of the target density, in order to minimise the risk of omitting subsets of the target's support. Finding an appropriate importance density becomes particularly cumbersome when the shape of the target density is non-elliptical. As illustrated by Figure 2.1.2, the optimal importance density for Bayesian VaR estimation is generally bimodal.

A standard approach to overcome this problem is to approximate the target density with a mixture of basis densities[7], for which Student's t densities are often chosen. Several methods to construct the approximating mixture of Student's t have been developed, see Peel and McLachlan (2000), Svensén and Bishop (2005), Hoogerheide et al. (2007) and Hoogerheide et al. (2012). We employ the latter algorithm, Mixture of t by Importance Sampling weighted Expectation-Maximization (MitISEM). This is a noticeable distinction compared to Hoogerheide and van Dijk (2010), whose original QERMit algorithm relies on another approximation algorithm, Adaptive Mixture of t (AdMit) of Hoogerheide et al. (2007). To explain our motivation behind this change of the employed method, below we provide a brief discussion of the differences between both techniques.

First, the objective function in AdMit is the coefficient of variation of the importance weights (i.e., the standard deviation divided by the mean), which is directly minimised via numerical optimisation. In contrast, MitISEM aims at minimising the Kullback-Leibler divergence, which is an indirect way to minimise the variance of the IS estimator. This makes the latter method quicker and more reliable, as it allows the optimization of the importance density to be performed with an EM algorithm, so that no Newton-Raphson based algorithm (such as the BFGS method) is needed. Second, MitISEM is a "fully adaptive" algorithm, as each time a new component is added to the old mixture, the parameters of all the components in the new mixture are jointly optimised, whereas in AdMit only the parameters of the new component are optimised, with those of the old mixture not being adjusted any more. Third, the only inputs to MitISEM are draws from the importance density and their importance weights, while in AdMit one needs to use the kernel of the joint target density. Thus, the latter method cannot be applied to conditional or marginal densities, which makes it useless in our Bayesian analysis which is based on the factorisation of the joint target density of the parameters and future returns.

[7]Zeevi and Meir (1997) show that such mixtures can provide an arbitrarily close approximation to any strictly positive density over a compact domain.

2.1.3 Sequential construction of marginal/conditional importance densities

If the horizon of the future returns increases, then it becomes more difficult to obtain an appropriate importance density for the parameters and future returns. Hence, we want to construct an approximation "sequentially", in each future time period conditioning the properties of the current conditional importance density of the return on the simulated parameters and returns in the previous periods. Intuitively, the idea is to "guide" the draws to fall into the high-loss region: if so far certain losses have been recorded, we know by how much the situation must additionally deteriorate to end up in the tail. Such a sequential and conditional construction of the importance densities can be easily carried out using the Partial MitISEM (PMitISEM) method of Hoogerheide et al. (2012). This algorithm aims at approximating the joint target density indirectly, by approximating the product of marginal and conditional target densities of subsets of model parameters – and in our case future returns.

To explain how the "guiding" process is carried out, below we discuss the details of PMitISEM. We express the joint target density $p(\boldsymbol{\theta})$ as a product of a marginal density and conditional densities:

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}_S|\boldsymbol{\theta}_{S-1}, \ldots, \boldsymbol{\theta}_2, \boldsymbol{\theta}_1) \ldots p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)p(\boldsymbol{\theta}_1),$$

where $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_S)$ is a partition of a k-dimensional vector $\boldsymbol{\theta}$ into S subsets with respective dimensions k_s, $s = 1, \ldots, S$, where naturally $\sum_{s=1}^{S} k_s = k$. Then it may be desirable to iteratively approximate each of the marginal and conditional densities due to the implied dimensionality reduction for each of the sub-problems. In general, the basic MitISEM could be applied to each of them to optimise the *modes*, scale matrices, degrees of freedom and weights independently for each subset. However, this would naturally result in a very poor joint importance density (unless the subsets $\boldsymbol{\theta}_s$ are independent) as the conditional structure would be neglected. In order to capture the interdependence between the subsets, in the PMitISEM algorithm the modes of the components in the subsequent conditional subsets are based on fitted values in the regression of the current subset parameters on (a function of) the parameters from the previous subsets (and potentially other "global" variables, e.g. functions of the data). PMitISEM optimises the *regression coefficients* for the conditional importance densities (corresponding to the subsets $\boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_S$), instead of optimising their *modes*. Below we discuss the details of the regression.

The underlying idea comes from the basic result in multivariate regression theory. For

the sake of simplicity of the exposition we restrict ourselves to the case $S = 2$; the extension to more subsets is straightforward. Consider the (asymptotically valid) approximating normal distribution $\mathcal{N}(\mu, \Sigma)$ for $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$, where $\mu = \arg\max_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$ and $\Sigma = -\mathcal{H}(\log f(\boldsymbol{\theta}))^{-1}|_{\boldsymbol{\theta}=\mu}$, where $f(\boldsymbol{\theta})$ is the posterior density kernel. Let

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Then

$$\begin{aligned} \boldsymbol{\theta}_1 &\sim \mathcal{N}(\mu_1, \Sigma_{11}), \\ \boldsymbol{\theta}_2|\boldsymbol{\theta}_1 &\sim \mathcal{N}(\underbrace{\mu_2 + \Sigma_{22}^{-1}\Sigma_{21}(\boldsymbol{\theta}_1 - \mu_1)}_{\beta X}, \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}). \end{aligned}$$

The PMitISEM algorithm replaces both the marginal and conditional normal distributions with mixtures of Student's t distributions. The mixture for the marginal distribution for $\boldsymbol{\theta}_1$ is constructed with the basic MitISEM algorithm. The mixture for the conditional density for $\boldsymbol{\theta}_2$ given $\boldsymbol{\theta}_1$ is constructed with a modified version of the algorithm, based on a regression of the parameters $\boldsymbol{\theta}_2$ on a constant term and some functions of parameters from the subsequent subset $\boldsymbol{\theta}_1$ (and potentially the data), all kept in the matrix X. Then, the above mentioned modification pertains to the optimisation of the coefficients of regression β instead of the modes.

In the **basic MitISEM** algorithm the maximisation step for the modes and the covariance matrices of the c-th mixture component is given by

$$\begin{aligned} \mu_c^{(L)} &= \left[\sum_{i=1}^{N} W^i \widetilde{z/w}_c^i\right]^{-1} \left[\sum_{i=1}^{N} W^i \widetilde{z/w}_c^i \boldsymbol{\theta}^i\right], \\ \hat{\Sigma}_c^{(L)} &= \frac{\sum_{i=1}^{N} W^i \widetilde{z/w}_c^i (\boldsymbol{\theta} - \mu_c^{(L)})(\boldsymbol{\theta}^i - \mu_c^{(L)})^T}{\sum_{i=1}^{N} W^i \widetilde{z}_c^i}, \end{aligned}$$

where W^i are the importance weights, and where $\widetilde{z/w}_c^i$ and $\widetilde{z}_c^i$, $i = 1, \ldots, N$, are computed in the expectation step of the algorithm. The exact formulae for their computation, together with other details of the basic MitISEM algorithm are provided in Appendix 2.A. In the **partial MitISEM** algorithm, the maximisation step for the regression coefficients β and the covariance matrices (for the conditional densities) of

the c-th mixture component becomes as follows

$$(\beta_c^{(L)})^T = \left[\sum_{i=1}^N W^i \widetilde{z/w}_c^i X_s^i (X_s^i)^T\right]^{-1} \left[\sum_{i=1}^N W^i \widetilde{z/w}_c^i X_s^i (\boldsymbol{\theta}^i)^T\right],$$
$$\hat{\Sigma}_c^{(L)} = \frac{\sum_{i=1}^N W^i \widetilde{z/w}_c^i (\boldsymbol{\theta}^i - \beta_c^{(L)} X_s^i)(\boldsymbol{\theta}^i - \beta_c^{(L)} X_s^i)^T}{\sum_{i=1}^N W^i \widetilde{z}_c^i}.$$

Notice that in the current partial setting each draw $\boldsymbol{\theta}_s^i$ (of length k_s) from the subset s ($s = 2, \ldots, S$) has a different conditional mean $\mu_c^i = \beta_c X_s^i$, where X_s^i is an $r \times 1$ vector (with elements being a constant and some $r-1$ functions of y and $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{s-1}$) and β_c is a $k_s \times r$ matrix. Intuitively, for each subset (and each component h in the conditional importance density for this subset) β_c characterises the common dependence of the s-th subset of parameters $\boldsymbol{\theta}_s$ (for component h) on the previous $s-1$ subsets of parameters (and on the data). The details of the procedure are presented in Appendix 2.B.

Let us return to the introductory toy example of white noise returns, with only one model parameter σ^2 and H other 'parameters' corresponding to the future disturbances $\varepsilon_1, \ldots, \varepsilon_H$. The sampling scheme is then as follows

$$\begin{aligned}
(\sigma^2, \varepsilon_1) &\sim q_1, \\
\varepsilon_2 | \sigma^2, \varepsilon_1 &\sim q_2, \\
\varepsilon_3 | \sigma^2, \varepsilon_1, \varepsilon_2 &\sim q_3, \\
&\vdots \\
\varepsilon_H | \sigma^2, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{H-1} &\sim q_H.
\end{aligned}$$

To construct the conditional mixture importance densities q_h with the PMitISEM algorithm we put for $h = 2, \ldots, H$

$$X_h = \left[\mathbf{1}, \sum_{t=1}^{h-1} y_t^*\right],$$

i.e. a column of ones and the cumulative returns in the previous periods. The latter choice is motivated by our aim to keep track of the evolution of the returns, i.e. how bad the situation has become up to now. In order to construct the marginal and conditional importance densities in the PMitISEM approach we need a preliminary set of parameter draws and corresponding high loss paths of future returns. For this purpose we use the high loss paths (and corresponding parameter draws) of a preliminary run of the direct

approach (illustrated in Figure 2.1.1), which also yields a preliminary VaR estimate. Given the preliminary VaR, this can allow us to assess how much "down" we still need to go in order to get to the high loss region. In Figure 2.1.3 this aim can be seen as ending up below the violet line.

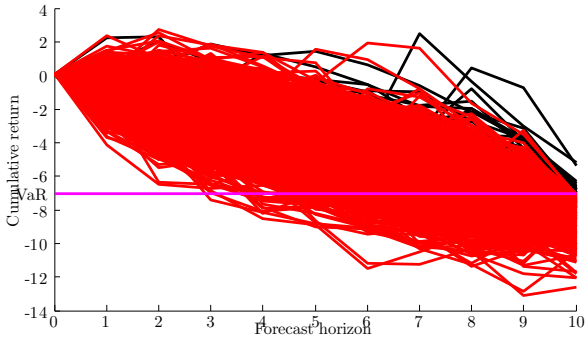


Figure 2.1.3: Simulations using PMitISEM result in almost all paths falling into the high-loss region (the red ones) below the 99% VaR value (the violet horizontal line). White noise returns, 10-days-ahead horizon, $10,000$ simulated paths.

2.2 Bayesian applications

In this section we discuss our key results for the 99% VaR and ES evaluation from the Bayesian perspective. We analyse two benchmark models of volatility, commonly employed by practitioners, the Generalized Autoregressive Conditional Heteroscedasticity model (GARCH, Engle, 1982; Bollerslev, 1986) and the Generalised Autoregressive Score model (GAS, Creal et al., 2013), both with Student's t innovations.

The main purpose of our applications is to illustrate the proposed IS-based evaluation method, i.e. how it is implemented and what remarkable efficiency gains it can yield. Keeping this in mind we apply each model to a different dataset, one used in the original paper of Hoogerheide and van Dijk (2010) and another one consisting of more recent data. Importantly, the former is a "calm" series, collected shortly before the financial crisis of 2008, while the latter contains the "wild" period of that financial distress, which makes the analysis much harder. Nevertheless, we record considerable efficiency gains for all the considered horizons also for that difficult dataset.

2.2.1 GARCH(1,1)-t

As our first illustration we consider the most advanced application from Hoogerheide and van Dijk (2010), where the authors apply the GARCH(1,1)-t model to the daily

logreturns of the S&P 500, from January 2, 1998 to December 31, 2007 (2513 observations, Figure 2.2.1) to evaluate the 10-days-ahead 99% VaR and ES. This is a natural starting point for our analysis, as with the AdMit algorithm, employed in the original paper, it was already difficult to obtain 10-days-ahead forecasts, while with the MitISEM algorithm "shorter" horizons, such as 10-day-ahead or 20-day-ahead, are easily reachable. Moreover, adopting the Partial MitISEM algorithm allows us to extend the original analysis much further, to record time–precision gains even for the one-year-ahead horizon.

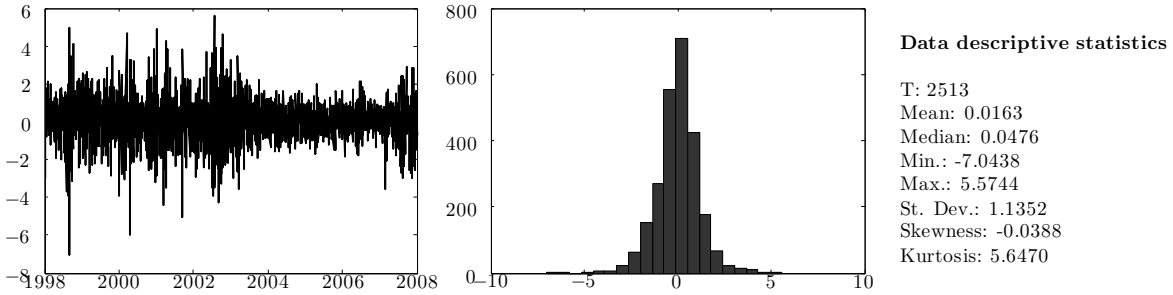


Figure 2.2.1: The data from the original Hoogerheide and van Dijk (2010) paper: the daily logreturns of the S&P 500, from January 2, 1998 to December 31, 2007.

The model is specified as follows:

$$
\begin{aligned}
y_t &= \mu + \sqrt{\rho h_t}\varepsilon_t, \\
\varepsilon_t &\sim t(\nu), \\
\rho &:= \frac{\nu - 2}{\nu}, \\
h_t &= \omega + \alpha y_{t-1}^2 + \beta h_{t-1},
\end{aligned}
$$

and we stack the model parameters into the vector $\boldsymbol{\theta} = (\omega, \alpha, \beta, \mu, \nu)$. We put flat priors on $\omega > 0$, $\alpha \in (0, 1)$, $\beta \in (0, 1)$ with $\alpha + \beta < 1$, to enforce that the conditional variance is positive and to ensure covariance stationarity, while for the degrees of freedom we set an uninformative yet proper prior: $\nu - 2 \sim \text{Exp}(0.01)$.

Table 2.2.1 presents the simulation results for the two direct approaches that we consider. In the naive-direct approach the candidate density is based on a single Student's t distribution with the mode equal to the MLE, the scale matrix equal to minus the inverse of the Hessian of the loglikelihood function evaluated at the mode, and the number of degrees of freedom set to 3 to allow for fat tails (as suggested by Geweke, 1989). To obtain the candidate with the adapted-direct approach we employ the MitISEM algorithm (Hoogerheide et al., 2012) to approximate the posterior of the model

	ML		MH (naive candidate)			MH (adapted candidate)		
Parameter	MLE	SD	Mean	SD	IF	Mean	SD	IF
ω	0.0082	0.0036	0.0091	0.0035	5.8174	0.0092	0.0034	5.4216
α	0.0726	0.0121	0.0702	0.0110	5.7170	0.0707	0.0109	4.7439
β	0.9238	0.0123	0.9241	0.0118	5.7776	0.9236	0.0117	4.8040
μ	0.0481	0.0169	0.0486	0.0171	5.5711	0.0489	0.0169	4.0058
ν	9.9964	1.9873	10.2582	1.9389	5.9288	10.2512	1.8897	4.2826
		AR		0.4376			0.6802	
	Time construction			0.93 s			60.89 s	
	Time sampling			10.86 s			13.72 s	
	No. of draws			10,000			10,000	

Table 2.2.1: Estimation results in the **GARCH(1,1)-t** model for Maximum Likelihood (ML) method (reported for comparison) and the Bayesian direct approach with naive (Student's t) and adapted (MitISEM mixture of Student's t) candidate distributions in the independence chain Metropolis-Hastings (MH) method: estimated posterior mean and standard deviation (SD), inefficiency factor (IF), acceptance rate (AR) in the MH method, and computing times for construction of the candidate distribution and for performing the direct approach.

parameters with the resulting candidate being a two-component mixture of Student's t distributions. Here, and in the subsequent applications, computation times refer to computations performed on an Intel(R) Core(TM) i5–3470 processor with 3.20 GHz. The "adaptation" of the candidate takes around one minute but allows for much closer approximation to the posterior distribution. The acceptance rate (AR) in the independence Metropolis-Hastings (MH) with the adapted candidate is almost 70%, which is much higher than when the naive candidate is adopted, in which case the AR is roughly 44%. Similarly, the adapted candidate results in less autocorrelated draws as measured by the inefficiency factors (IF)[8].

For the 99% VaR and ES evaluation we consider, next to both direct methods, two QERMit (i.e. IS-based) approaches. In these methods we apply different methods to approximate the "high-loss" density. The first one uses the basic MitISEM algorithm and targets the posterior predictive density as a whole. For this reason, it usually becomes infeasible to use for prediction horizons longer than 20, because then the covariance matrices of the Student's t components are hard to work with. The second approximation algorithm is PMitISEM, based on the sequential construction of the

[8]The inefficiency factor is defined as the variance of the parameter estimate divided by the variance in case the sampling scheme would generate independent posterior draws and it is the inverse of the relative numerical efficiency (see Pitt et al., 2012). For a sample of draws of a parameter ζ we compute IF as $\text{IF}(\zeta) = 1 + 2\sum_{\tau=1}^{\max\{L,1000\}} \rho_\tau(\zeta)$, where $\rho_\tau(\zeta)$ is the τ-th order autocorrelation in the sequence of draws of parameter ζ and L is the lowest order τ for which ρ_τ is not significant.

Subset	Parameters	No. of components	Weighted* μ or β*
1	$\{(\theta, \varepsilon_1)\}$	4	[0.0089 0.0698 0.9250 0.0473 9.8126 -1.0284]
2	$\{\varepsilon_2\}$	5	[-1.0863 -0.0948]
3	$\{\varepsilon_3\}$	5	[-1.1816 -0.1083]
4	$\{\varepsilon_4\}$	5	[-1.2589 -0.1070]
5	$\{\varepsilon_5\}$	5	[-1.4608 -0.1390]
6	$\{\varepsilon_6\}$	5	[-1.6167 -0.1471]
7	$\{\varepsilon_7\}$	5	[-1.8912 -0.1697]
8	$\{\varepsilon_8\}$	5	[-2.4583 -0.2202]
9	$\{\varepsilon_9\}$	4	[-2.8833 -0.2568]
10	$\{\varepsilon_{10}\}$	5	[-5.0261 -0.4842]

*Weighted with the mixture weights.

****The mode μ (for subset 1) or the regression coefficients β (for the other subsets).

Table 2.2.2: Properties of the marginal and conditional importance densities from the PMitISEM method for $H = 10$ in the **GARCH(1,1)-t** model.

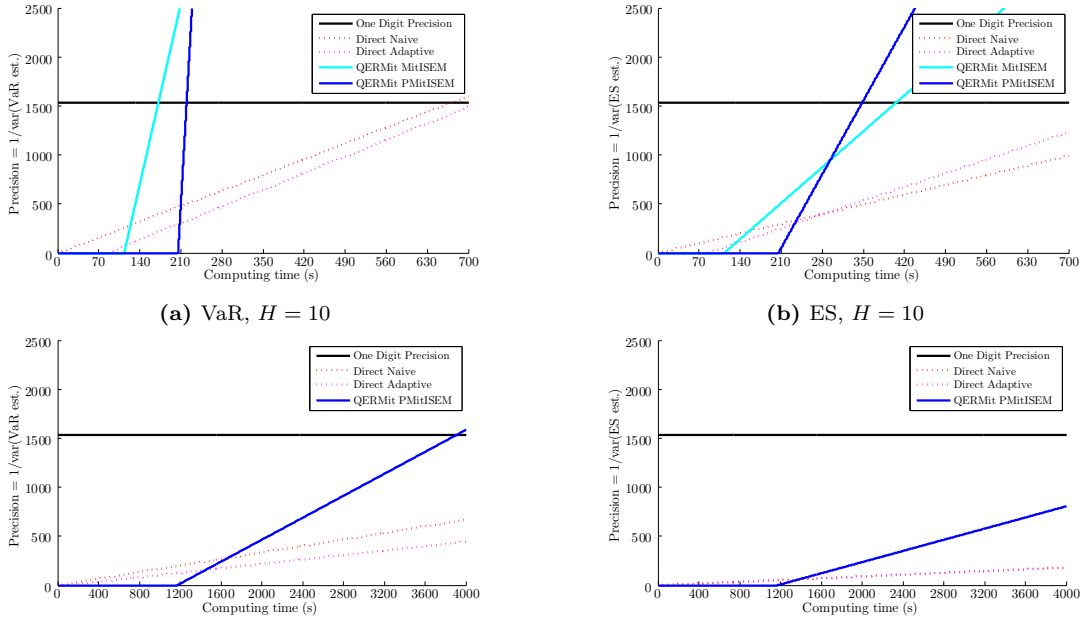
marginal and conditional importance densities as discussed in Section 2.1.3, which allows to extend the analysis way further than the basic QERMit of Hoogerheide and van Dijk (2010). We refer to these two methods by subscripts *mit* and *pmit*, respectively. Table 2.2.2 presents the properties of the partial mixture generated by PMitISEM for the 10-days-ahead case. Similarly as in the "toy" example of white noise returns we regress the draws from the current conditional importance density s on

$$X_s = \left[\mathbf{1}, \sum_{t=1}^{s-1} y_t^* \right],$$

to update the mode of the current conditional density. The last column contains the weighted mode of the marginal importance density, i.e. for $s = 1$, and weighted coefficients of regression for the conditional importance densities, i.e. for $s = 2, \ldots, 10$. The latter show how PMitISEM "guides" the subsequent draws into the "high-loss region". As expected, the later the period, the more negative the regression coefficient (at the cumulative return up to period $s-1$), with a noticeable jump in the last period to guarantee that the whole scenario becomes a high-loss one.

Table 2.2.3 compares the results for the 99% VaR and ES evaluation for different horizons, for which a visualisation is provided in Appendix 2.C.1. For each method, the results are based on $10,000$ draws, while to obtain the NSEs and interquantile ranges (IQR) we performed 20 Monte Carlo replications of the evaluation experiment. Here, and in the next applications, we consider five horizon lengths, $H \in \{10, 20, 40, 100, 250\}$. This selection ranges from the standard, intermediate horizon of two weeks, required by Basel Committee on Banking Supervision (1995), through the one month horizon, up to the long run, one-year-ahead horizon. The QERMit based methods clearly outperform the direct approaches, with both RNEs and IQRs being roughly 6 times higher for $H = 10$ VaR and almost 3 times for $H = 10$ ES. For the longest horizon of $H = 250$ QERMit delivers two to three times more accurate results than its direct competitors, both for VaR and ES evaluations. As expected, for long horizons, with $H = 40$ or more, the basic MitISEM becomes infeasible, due to the too high dimensionality of the scale matrices of the mixture components it would need to tackle. Fortunately, owing to the partial candidate construction, PMitISEM is still able to deliver satisfactory results even for these long horizons. Notice that PMitISEM outperforms the basic MitISEM already for the shorter horizons ($H = 10$ and $H = 20$), where its VaR forecasts are over twice more accurate than those obtained with basic MitISEM; for the ES the relative advantage of PMitISEM over basic MitISEM is smaller, yet still existing (the results from the latter algorithm are almost 50% less accurate than these from the former one). Interestingly, a better approximation to the posterior does not need to lead to a better performance in the tail: in some cases the adapted direct approach yields worse results than its naive counterpart, in particular when one considers just the IQR and not the NSE (cf. the NSE and the IQR for the VaR at $H = 250$ or just the IQR for the VaR at $H = 10$ or the ES at $H = 20$). This confirms the remark of Christoffersen et al. (1998) that standard, goodness-of-fit-focused methods are not bound to succeed in the tail estimation problems.

Naturally, for any method it holds that the longer the horizon, the lower the prediction accuracy. Also the advantage of the QERMit method over the direct approach diminishes when the horizon gets extended. The crucial question is then whether there is still a gain, in terms of the time-precision trade-off, of adopting a more accurate but also a more complex and time consuming method. To quantify that trade-off we consider the gain in precision (defined as the inverse of the variance) for one unit of computing time. We refer to it as the *slope*, as it characterises the steepness of a function determining the dependence between precision and computing time. A method with a higher slope will eventually require less computing time to achieve a certain (high) precision, even



(a) VaR, $H = 10$ **(b)** ES, $H = 10$

(c) VaR, $H = 250$ **(d)** ES, $H = 250$

Figure 2.2.2: Precision ($1/var$) of the estimated VaR (left) and ES (right), as a function of the amount of computing time for different approaches, for the **GARCH(1,1)-t** model, for the shortest and the longest horizon. The horizontal line corresponds to a precision of 1 digit ($1.96NSE \leq 0.05$). A missing line for the MitISEM-based importance density corresponds to a situation when it was not possible to construct such an importance density.

after accounting for an inevitable fixed "investment cost" of time needed to construct a reliable importance density. The results of the investigation on this issue are presented in Table 2.2.4, and the plots corresponding to the shortest and longest horizon are presented in Figure 2.2.2 (Appendix 2.D.1 provides plots for all the horizons, with additional details on the plots construction). The QERMit based methods turn out to be not only more accurate but also more efficient than the direct approaches, in a sense that they require less computing time and fewer draws to achieve the same accuracy as the direct methods, or, stated differently, they yield higher precision in the same time and using the same number of draws. Importantly, the conditioning of partial MitISEM allows us to increase efficiency for all horizons, including the longest horizon of $H = 250$, for both, VaR and ES evaluations.

Finally, following Hoogerheide and van Dijk (2010) we also consider the benchmark of 1 digit precision with 95% confidence. It is defined as $1.96\text{NSE} \leq 0.05$, which corresponds to the required precision level of 1536. Then, the *time required* and *draws required* refer to the computing time and the number of draws necessary to achieve this precision level. Notice, that this benchmark is set somewhat arbitrarily and considering a higher confidence would mean a much higher required precision. For instance, changing of

H		VaR_{naive}	VaR_{adapt}	VaR_{mit}	VaR_{pmit}	ES_{naive}	ES_{adapt}	ES_{mit}	ES_{pmit}
10		-8.1484	-8.1257	-8.2091	-8.1808	-9.9134	-9.7853	-9.9209	-9.8759
	NSE	(0.1836)	(0.1748)	0.0531	(0.0267)	(0.2329)	(0.1922)	(0.1192)	(0.0838)
	IQR	[0.2066]	[0.2478]	[0.0840]	[0.0367]	[0.4104]	[0.2563]	[0.1543]	[0.1384]
	RNE	1.02	1.01	5.8198	12.37	1.59	1.60	11.28	44.66
20		-11.2028	-11.2846	-11.3024	-11.2265	-13.5991	-13.7225	-13.6589	-13.5866
	NSE	(0.2907)	(0.2151)	0.1454	(0.0626)	(0.3923)	(0.3436)	(0.1683)	(0.1141)
	IQR	[0.3382]	[0.3125]	[0.2157]	[0.0958]	[0.5424]	[0.6844]	[0.2118]	[0.1536]
	RNE	1.01	1.03	2.6315	8.75	1.63	1.65	2.23	12.05
40		-15.2151	-15.2188	–	-15.3329	-18.5758	-18.6593	–	-18.7022
	NSE	(0.3520)	(0.3094)	(–)	(0.1020)	(0.5806)	(0.5470)	(–)	(0.1991)
	IQR	[0.3605]	[0.3839]	[–]	[0.1213]	[0.9029]	[0.5279]	[–]	[0.2513]
	RNE	1.03	1.00	–	7.79	1.77	1.67	–	25.85
100		-22.6319	-22.6711	–	-22.6115	-28.4722	-28.3719	–	-28.6178
	NSE	(0.6497)	(0.4005)	(–)	(0.2433)	(0.8134)	(0.7701)	(–)	(0.3119)
	IQR	[0.8049]	[0.5865]	[–]	[0.4399]	[1.0842]	[1.2843]	[–]	[0.4846]
	RNE	1.03	1.05	–	5.43	1.64	1.65	–	9.08
250		-32.0179	-32.0471	–	-32.1617	-41.3818	-41.8261	–	-41.3818
	NSE	(0.6737)	(0.7966)	(–)	(0.3266)	(1.2958)	(1.2476)	(–)	(0.4583)
	IQR	[0.7134]	[0.9548]	[–]	[0.4905]	[2.2109]	[1.6169]	[–]	[0.5894]
	RNE	1.02	1.03	–	3.73	1.65	1.65	–	10.11

Missing value (–): it was not possible to generate the particular result with the corresponding algorithm.

Table 2.2.3: Results for the 99% VaR and ES, in the **GARCH(1,1)-t** model, based on $N = 10,000$ draws and 20 replications to obtain the numerical standard error (NSE) and the interquartile range (IQR). The RNE is the relative numerical efficiency, the inverse of the inefficiency factor. The results are obtained using the direct approach (with naive and adapted candidate distribution in the Metropolis-Hastings algorithm), and the QERMit method (with the basic MitISEM and PMitISEM methods), respectively.

the confidence to 99% would raise it to 2654. Table 2.2.4 shows that even for the longest considered horizon of $H = 250$ the QERMit method is almost 2.5 times faster in estimating the 99% VaR with such a reasonable precision and requires over 4 times fewer draws to achieve that than the direct approach. For the ES the relative gain is even higher as QERMit turns out to be more than 5 times faster and nearly 8 times less draw-requiring than the naive direct approach. Notice that demanding a higher confidence on the precision would make QERMit even more advantageous relative to the direct approach.

	Direct		QERMit		Direct		QERMit	
H	Naive	Adapted	MitISEM	PMitISEM	Naive	Adapted	MitISEM	PMitISEM
	Total time							
10	13.89 s	98.65 s	127.00 s	218.56 s				
20	13.78 s	98.57 s	270.51 s	150.68 s				
40	13.96 s	98.61 s	–	328.77 s				
100	13.95 s	98.93 s	–	544.84 s				
250	14.09 s	99.20 s	–	1193.52 s				
	Construction time				Sampling time			
10	0.88 s	85.14 s	113.56 s	205.31 s	13.01 s	13.52 s	13.44 s	13.26 s
20	0.88 s	85.01 s	257.03 s	136.29 s	12.91 s	13.56 s	13.48 s	14.39 s
40	0.91 s	85.01 s	–	314.87 s	13.05 s	13.60 s	–	13.90 s
100	0.87 s	85.16 s	–	530.03 s	13.08 s	13.77 s	–	14.81 s
250	0.87 s	85.30 s	–	1176.81 s	13.22 s	13.90 s	–	16.72 s
	VaR slope*				**ES slope***			
10	**2.28**	**2.42**	**26.34**	**105.78**	**1.42**	**2.00**	**5.23**	**10.74**
20	**0.92**	**1.59**	**3.51**	**17.75**	**0.50**	**0.62**	**2.62**	**5.34**
40	**0.62**	**0.77**	–	**6.92**	**0.23**	**0.25**	–	**1.81**
100	**0.18**	**0.45**	–	**1.14**	**0.12**	**0.12**	–	**0.69**
250	**0.17**	**0.11**	–	**0.56**	**0.05**	**0.05**	–	**0.28**
	VaR time required**				ES time required**			
10	674.42 s	719.98 s	171.89 s	219.84 s	1,085.54 s	852.25 s	407.13 s	348.36 s
20	1,677.00 s	1,049.22 s	694.79 s	222.84 s	3,053.12 s	2,544.50 s	843.71 s	423.86 s
40	2,485.72 s	2,085.40 s	–	536.97 s	6,762.26 s	6,338.28 s	–	1,161.81 s
100	8,486.05 s	3,480.60 s	–	1,877.18 s	13,299.21 s	12,637.37 s	–	2,743.92 s
250	9,220.75 s	13,640.17 s	–	3,917.09 s	34,108.73 s	33,336.37 s	–	6,573.53 s
	VaR draws required**				ES draws required**			
10	517,761	469,580	43,392	10,959	833,801	567,412	218,386	107,921
20	1,298,426	711,093	324,786	60,162	2,364,446	1,813,836	435,278	199,891
40	1,903,737	1,470,764	–	159,768	5,180,194	4,597,643	–	609,227
100	6,486,508	2,465,138	–	909,605	10,165,937	9,113,098	–	1,494,828
250	6,975,069	9,750,193	–	1,639,192	25,803,439	23,917,916	–	3,228,229

Missing value (–): it was not possible to generate the particular result with the corresponding algorithm.

*Slope = increase in precision per unit of computing time.

**Required for % estimate with 1 digit of precision (with 95% confidence).

Table 2.2.4: Trade-off of precision versus computing time for the 99% VaR and ES in **GARCH(1,1)-t** model for different horizons.

2.2.2 GAS(1,1)-t

Having referred to the benchmark GARCH application of Hoogerheide and van Dijk (2010), in our second illustration we consider a more recently developed model for more recent data. Creal et al. (2013) propose an alternative approach to modelling volatility based on the updating of the time-varying parameter with the scaled score of the observation's contribution to the likelihood function. We employ their Generalised Autoregressive Score (GAS) model to the daily logreturns of the S&P 500, from January 3, 2005 to June 30, 2016 (2893 observations, Figure 2.2.3) to evaluate the 99% VaR and ES at the same horizons as in the previous section[9]. The data span over the 2008 financial crisis resulting in very high sample kurtosis, so that one would expect potential difficulties in obtaining precise risk forecasts.

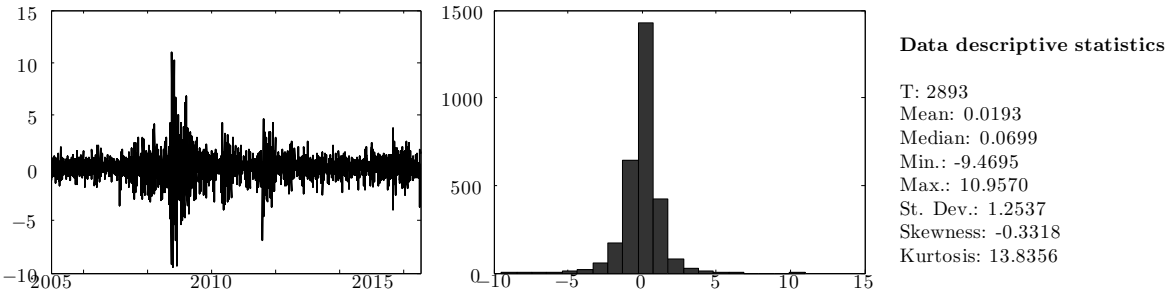


Figure 2.2.3: The series including the 2008 Financial Crisis: the daily logreturns of the S&P 500, from January 3, 2005 to June 30, 2016.

We adopt the following basic specification of the GAS model, referred to as GAS(1,1)-t,

$$
\begin{aligned}
y_t &= \mu + \sqrt{\rho h_t}\varepsilon_t, \\
\varepsilon_t &\sim t(\nu), \\
\rho &:= \frac{\nu-2}{\nu}, \\
h_t &= \omega + A\frac{\nu+3}{\nu}\Big(C_{t-1}(y_t-\mu)^2 - h_{t-1}\Big) + Bh_{t-1}, \\
C_t &= \frac{\nu+1}{\nu-2}\left(1+\frac{(y_{t-1}-\mu)^2}{(\nu-2)h_{t-1}}\right)^{-1},
\end{aligned}
$$

where we stack the model parameters into vector $\boldsymbol{\theta} = (\mu, \omega, A, B, \nu)^T$. Finally, we

[9]We also considered "complimentary" applications, i.e. employing the GAS model to the "old" dataset, as well as running the GARCH model on the "crisis" series. The former application performed better than the originally analysed model, yielding even more noticeable efficiency gains than those reported in Section 2.2.1. Regarding the latter, the GAS model as expected, provided a much better framework for modelling extreme returns present in the crisis data compared to the GARCH model, which is a result also reported by Jelsma and Lasak (2016).

put flat priors on μ, ω, A and B, with $\omega > 0$ and $B \in (0, 1)$ to guarantee that the conditional variance is positive and to ensure covariance stationarity, and uninformative exponential prior on ν, $\nu - 2 \sim \text{Exp}(0.01)$.

Table 2.2.5: Estimation results in the **GAS(1,1)-t** model for Maximum Likelihood (ML) method (reported for comparison) and the Bayesian direct approach with naive (Student's t) and adapted (MitISEM mixture of Student's t) candidate distributions in the independence chain Metropolis-Hastings (MH) method: estimated posterior mean and standard deviation (SD), inefficiency factor (IF), acceptance rate (AR) in the MH method, and computing times for construction of the candidate distribution and for performing the direct approach.

	ML		MH (naive candidate)			MH (adapted candidate)		
Parameter	MLE	SD	Mean	SD	IF	Mean	SD	IF
μ	0.0702	0.0141	0.0738	0.0140	4.8736	0.0739	0.0140	3.6611
ω	0.0219	0.0050	0.0222	0.0048	4.9919	0.0221	0.0048	3.6370
A	0.0996	0.0111	0.1026	0.0111	4.7300	0.1022	0.0110	3.6902
B	0.9817	0.0061	0.9818	0.0059	4.6926	0.9819	0.0059	3.7480
ν	6.8979	1.0376	7.0853	1.0256	5.0386	7.0762	1.0163	3.7607
		AR		0.5547			0.7776	
		Time construction		0.98 s			108.83 s	
		Time sampling		17.24 s			17.80 s	
		No. of draws		10,000			10,000	

Table 2.2.5 presents the simulation results for the two direct approaches. This time, due to a bit longer series and a more complex volatility update formula, the adaptation of the direct candidate takes slightly more than 1.5 minutes. However, the resulting AR is much higher than in the previous application, reaching nearly 78%; it also exceeds the one obtained with the naive candidate, which somewhat exceeds 55%. The superiority of the adapted candidate is also reflected in lower IF values for all the parameters. Notice that the degrees of freedom for the observation disturbances ν are estimated at a lower level than in the previous application (around 7 compared to roughly 10 before), which corresponds to a much more volatile nature of the current dataset.

Table 2.2.6 presents the properties of the partial mixture generated by PMitISEM for the 10-day-ahead case. Given an uneasy character of the current time series it is interesting to notice that with the GAS model a lower number of mixture components was required by the PMitISEM algorithm to approximate the tails, compared to the previous application. Now two or three components are sufficient while with the GARCH model as many as four to five components were necessary – and this was the case for much more regular data. Again, the last column presents decreasing values of the regression coefficient (at the cumulative return up to period $s - 1$) used to determine

Subset	Parameters	No. of components	Weighted* μ or β^*
1	$\{(\theta, \varepsilon_1)\}$	1	[0.0731 0.0225 0.1045 0.9823 7.0176 -1.0027]
2	$\{\varepsilon_2\}$	2	[-1.1023 -0.0975]
3	$\{\varepsilon_3\}$	2	[-1.1874 -0.0887]
4	$\{\varepsilon_4\}$	2	[-1.2748 -0.0966]
5	$\{\varepsilon_5\}$	2	[-1.4993 -0.1150]
6	$\{\varepsilon_6\}$	1	[-1.6575 -0.1231]
7	$\{\varepsilon_7\}$	2	[-1.9947 -0.1557]
8	$\{\varepsilon_8\}$	3	[-2.3280 -0.1755]
9	$\{\varepsilon_9\}$	3	[-2.9323 -0.2234]
10	$\{\varepsilon_{10}\}$	3	[-4.8901 -0.3954]

*Weighted with the mixture weights.

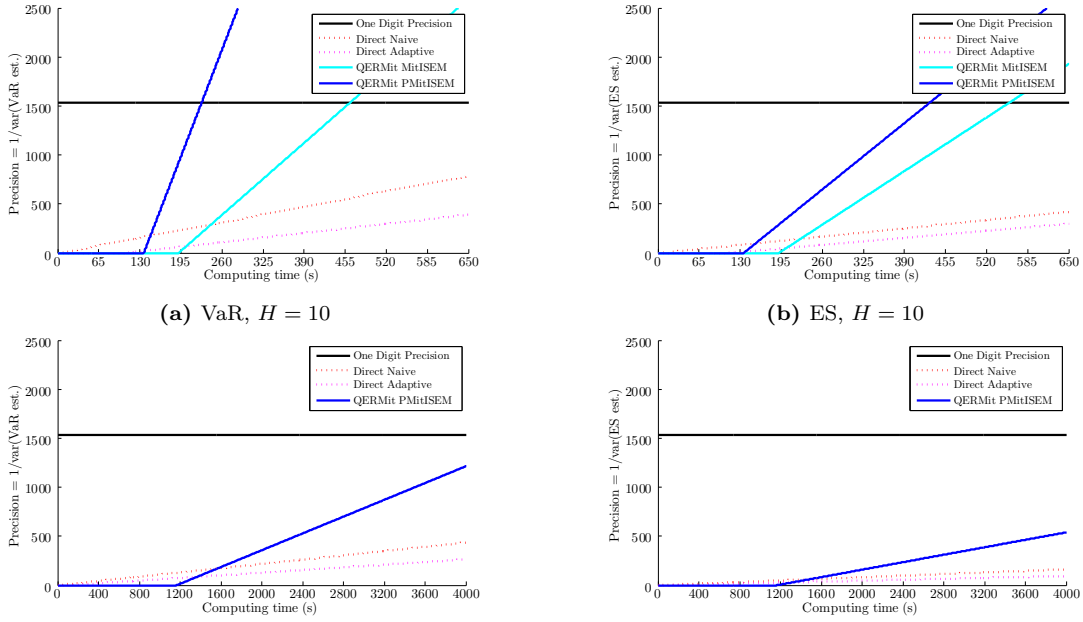
**The mode μ (for subset 1) or the regression coefficients β (for the other subsets).

Table 2.2.6: Properties of the marginal and conditional importance densities from the PMitISEM method for $H = 10$ in the **GAS(1,1)-t** model.

the modes of subsequent conditional mixtures, exhibiting the process of "guiding" of the draws to the tail by PMitISEM.

Table 2.2.7 reveals that also this time we observe substantial accuracy gains for our proposed methods for all horizons, for both the 99% VaR and ES (for the corresponding visualisation we refer to Appendix 2.C.1). For the VaR evaluations at $H = 10, 20, 40$ the NSE is around four times smaller, while for $H = 100$ and $H = 250$ it is roughly 2.5 times smaller. Again, the ES turns out to be somewhat harder to precisely estimate than the VaR, yet also in this case we report considerable gains. For $H \leq 40$ the computed NSEs are around three times lower with the PMitISEM based QERMit than with the direct approaches, while for the two longest horizons they diminish more than twice. Broadly speaking, a similar pattern pertains to the computed IQRs.

Finally, the most important results on time-precision trade-off are provided in Table 2.2.8 with the plots corresponding to the shortest and longest horizon presented in Figure 2.2.4 (Appendix 2.D.1 provides plots for all the horizons). For all horizons, for both the VaR and the ES, the slopes obtained with the PMitISEM algorithm are much higher than in the case of the direct approach, often by more than one order of



(a) VaR, $H = 10$

(b) ES, $H = 10$

(c) VaR, $H = 250$

(d) ES, $H = 250$

Figure 2.2.4: Precision ($1/var$) of the estimated VaR (left) and ES (right), as a function of the amount of computing time for different approaches, for the **GAS(1,1)-t** model, for the shortest and the longest horizon. The horizontal line corresponds to a precision of 1 digit ($1.96NSE \leq 0.05$). A missing line for the MitISEM-based importance density corresponds to a situation when it was not possible to construct such an importance density.

magnitude. Also basic MitISEM outperforms the direct approaches, but it is clearly inferior to PMitISEM. Eventually PMitISEM requires less time (and fewer draws) to achieve the same precision as the direct approaches. For instance, when the 1 digit precision with 95% confidence is considered, to accurately evaluate the 99% VaR and ES, the PMitISEM based QERMit needs, respectively, almost 3 and over 4 times less time than the naive direct approach (which outperforms the adaptive direct method).

H		VaR_{naive}	VaR_{adapt}	VaR_{mit}	VaR_{pmit}	ES_{naive}	ES_{adapt}	ES_{mit}	ES_{pmit}
10		-9.4284	-9.4076	-9.4290	-9.4358	-11.5862	-11.4901	-11.6038	-11.5870
	NSE	(0.2183)	(0.2793)	0.1040	(0.0601)	(0.2988)	(0.3205)	(0.1205)	(0.1078)
	IQR	[0.2697]	[0.3666]	[0.1865]	[0.0891]	[0.4290]	[0.5576]	[0.1183]	[0.1505]
	RNE	1.02	1.03	5.0467	9.92	1.67	1.59	8.82	26.39
20		-12.5332	-12.6962	-12.6807	-12.6483	-15.5819	-15.6293	-15.7741	-15.6556
	NSE	(0.3039)	(0.3145)	0.1569	(0.0686)	(0.4837)	(0.4070)	(0.3280)	(0.1310)
	IQR	[0.5002]	[0.3988]	[0.2253]	[0.1264]	[0.6433]	[0.5856]	[0.3339]	[0.1832]
	RNE	1.01	1.03	2.4818	8.43	1.65	1.60	4.62	24.17
40		-16.4218	-16.4804	–	-16.4626	-20.7435	-20.8218	–	-20.8775
	NSE	(0.3907)	(0.3582)	(–)	(0.0907)	(0.7497)	(0.5630)	(–)	(0.2182)
	IQR	[0.3910]	[0.5375]	[–]	[0.1363]	[0.7104]	[0.8472]	[–]	[0.2692]
	RNE	1.00	1.01	–	7.67	1.76	1.65	–	30.58
100		-21.7043	-21.7532	–	-21.6031	-28.3618	-28.6295	–	-28.4508
	NSE	(0.4918)	(0.5725)	(–)	(0.2135)	(1.1395)	(0.7797)	(–)	(0.4737)
	IQR	[0.6407]	[0.8066]	[–]	[0.3593]	[2.0389]	[1.1382]	[–]	[0.3251]
	RNE	1.04	1.02	–	5.73	1.71	1.69	–	14.57
250		-25.2962	-25.4476	–	-25.1630	-34.9541	-34.4421	–	-34.3317
	NSE	(0.7228)	(0.9014)	(–)	(0.3332)	(1.1825)	(1.5043)	(–)	(0.4997)
	IQR	[1.0707]	[1.2386]	[–]	[0.5608]	[1.8279]	[1.4069]	[–]	[0.5731]
	RNE	1.02	1.01	–	4.41	1.71	1.71	–	3.10

Missing value (–): it was not possible to generate the particular result with the corresponding algorithm.

Table 2.2.7: Results for the 99% VaR and ES, in the **GAS(1,1)-t** model, based on $N = 10000$ draws and 20 replications to obtain the numerical standard error (NSE) and the interquartile range (IQR). The RNE is the relative numerical efficiency, the inverse of the inefficiency factor. The results are obtained using the direct approach (with naive and adapted candidate distribution in the Metropolis-Hastings algorithm), and the QERMit method (with the basic MitISEM and PMitISEM methods), respectively.

	Direct		QERMit		Direct		QERMit	
H	Naive	Adapted	MitISEM	PMitISEM	Naive	Adapted	MitISEM	PMitISEM
	Total time							
10	21.64 s	126.74 s	208.47 s	152.90 s				
20	21.52 s	126.74 s	189.78 s	164.66 s				
40	21.57 s	126.79 s	–	186.77 s				
100	21.65 s	126.86 s	–	297.01 s				
250	21.87 s	127.13 s	–	1191.88 s				
	Construction time				Sampling time			
10	4.41 s	108.94 s	192.11 s	136.38 s	17.23 s	17.81 s	16.36 s	16.52 s
20	4.28 s	108.92 s	171.95 s	147.82 s	17.24 s	17.82 s	17.83 s	16.83 s
40	4.29 s	108.94 s	–	169.85 s	17.28 s	17.85 s	–	16.92 s
100	4.29 s	108.91 s	–	279.05 s	17.36 s	17.94 s	–	17.96 s
250	4.30 s	108.96 s	–	1170.85 s	17.57 s	18.17 s	–	21.02 s
	VaR slope*				**ES slope***			
10	**1.22**	**0.72**	**5.66**	**16.76**	**0.65**	**0.55**	**4.21**	**5.21**
20	**0.63**	**0.57**	**2.28**	**12.61**	**0.25**	**0.34**	**0.52**	**3.46**
40	**0.38**	**0.44**	–	**7.18**	**0.10**	**0.18**	–	**1.24**
100	**0.24**	**0.17**	–	**1.22**	**0.04**	**0.09**	–	**0.25**
250	**0.11**	**0.07**	–	**0.43**	**0.04**	**0.02**	–	**0.19**
	VaR time required**				ES time required**			
10	1,266.27 s	2,243.21 s	463.78 s	228.07 s	2,368.24 s	2,919.50 s	557.28 s	431.26 s
20	2,450.42 s	2,817.30 s	846.93 s	269.72 s	6,201.54 s	4,643.55 s	3,120.15 s	591.50 s
40	4,057.15 s	3,628.94 s	–	383.75 s	14,926.56 s	8,803.87 s	–	1,408.28 s
100	6,455.40 s	9,146.02 s	–	1,537.72 s	34,635.95 s	16,871.56 s	–	6,473.64 s
250	14,112.09 s	22,790.17 s	–	4,756.85 s	37,756.56 s	63,277.38 s	–	9,238.90 s
	VaR draws required**				ES draws required**			
10	732,326	1,198,603	166,065	55,495	1,371,858	1,578,403	223,218	178,472
20	1,419,270	1,519,981	378,493	72,408	3,595,706	2,544,903	1,653,200	263,554
40	2,345,620	1,971,873	–	126,410	8,636,375	4,870,819	–	731,891
100	3,716,418	5,036,326	–	700,689	19,950,925	9,341,719	–	3,448,462
250	8,028,977	12,485,292	–	1,705,594	21,485,424	34,772,225	–	3,837,372

Missing value (–): it was not possible to generate the particular result with the corresponding algorithm.

*Slope = increase in precision per unit of computing time.

**Required for % estimate with 1 digit of precision (with 95% confidence).

Table 2.2.8: Trade-off of precision versus computing time for the 99% VaR and ES in **GARCH(1,1)-t** model for different horizons.

2.3 Frequentist QERMit

The theoretical reason for the 50%-50% formula (2.1.5) was that in Bayesian analysis usually only a kernel of the target density is available, i.e. the normalising constant for the posterior density is unknown, so that one needs to normalise the importance weights by their sum. If the target was known there would be no need to normalise the importance weights and the optimal sampling density would put all the probability mass into the region of interest. This is typically the case in frequentist inference, where we only need to simulate the future returns (or future innovations) of which we know the exact density (including the scaling constant) given the parameter vector θ. Let $p(\boldsymbol{\varepsilon}^*)$ denote the target density of future disturbances, $\boldsymbol{\varepsilon}^* = (\varepsilon_1^*, \ldots, \varepsilon_H^*)$, and suppose that the vector of model parameters θ is fixed (this can be seen as either the "true" model parameters being known or the MLE being available). Then the optimal importance density is a function only of ε^*, given θ, and it is constructed solely over the tail.

In general, the optimal candidate density for estimation of $\mathbb{E}_p[g(X)]$ is given by

$$q_{opt}(x) = C|g(x)|p(x)$$

with the normalising constant $C = 1/\mathbb{E}_p[|g(X)|]$ (see Kahn and Marshall, 1953). In the case of estimating probability $\bar{p}$ of an event S we have $g(x) = \mathbb{1}_S(x)$, hence

$$q_{opt}(x) = \mathbb{1}_S(x)p(x)/\bar{p},$$

so it is a density proportional to the target over the set S. Then

$$\begin{aligned}\mathbb{E}_p[\mathbb{1}_S(X)] &= \int_S p(x)dx \\ &= \int_S \frac{p(x)}{q_{opt}(x)} q_{opt}(x)dx \\ &= \mathbb{E}_{q_{opt}}[\mathbb{1}_S(X)w(X)],\end{aligned}$$

where $w(x) = p(x)/q_{opt}(x)$, and its *unbiased* and consistent MC estimator is then given by

$$\widehat{\mathbb{E}_p[\mathbb{1}_S(X)]} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_S(x^{(i)})w(x^{(i)}),$$

where $x^{(1)}, \ldots, x^{(N)}$ are i.i.d. draws from q_{opt}. Notice that using q_{opt} results in *zero-variance IS*, as $\frac{p(x)}{q_{opt}(x)}\mathbb{1}_S(x)$ is constant (equal to $\mathbb{E}_p[g(X)]$). Hence, in the case when the target density is known, there is *no limit* on the potential relative gain in precision from using IS rather than direct simulation[10], which means that the RNE can be unbounded. In practice, however, the problem is that to implement sampling based on q_{opt}, one needs to know $\bar{p}$, which obviously is infeasible as the evaluation of $\bar{p}$ is the goal of the undertaken analysis in the first place. In the context of risk evaluation, using the previously introduced notation, we would need to know the $100\alpha\%$ VaR. Hence, similarly as in the Bayesian case, we can approximate q_{opt} based on some preliminary value VaR_{prelim} obtained with the direct approach. To this end we again use a mixture of Student's t distributions delivered by MitISEM.

As already noted in the introduction, the advantages of IS as a variance reduction technique in the frequentist case have already been noticed in the literature. Glasserman et al. (1999) and Glasserman et al. (2000) combine IS with stratified sampling to obtain precise estimates of VaR, while Glasserman et al. (2002) extend their analysis to also include ES. They specify an importance density based on a quadratic "delta-gamma" approximation to the change in portfolio value. They, however, do not consider time series models and do not carry out an empirical study on the real data, which are of key interest to us. Hence, we do not consider their approach in our research, although some insights from those studies might be useful in further research.

2.3.1 GARCH(1,1)-t

Below we discuss the frequentist counterpart of the Bayesian analysis of the GARCH(1,1)-t model from Section 2.2.1. We fix the model parameters to their MLE values (reported in Table 2.2.1), compute the corresponding volatility for the last in-sample time period and simulate only the i.i.d. future disturbances ε_h, $h = 1, \ldots, H$. Since there is no posterior density to approximate in this case, now we have only one direct approach, where we simulate ε_h directly from the target, which in this case is the standard Student's t density with roughly 10 degrees of freedom. The QERMit approaches are based on the approximations to the tail of the target, i.e. the tail of the predictive density.

Table 2.3.1 shows that also in the frequentist case we achieve noticeable improvements in the accuracy of the VaR and ES evaluations for all horizons (the corresponding plots

[10] See Hoogerheide and van Dijk (2010), who derive the limit of the potential relative gain in precision from using IS rather than direct simulation for VaR evaluation in the Bayesian context, which is equal to $(4(1-\bar{p})\bar{p})^{-1}$. This is 25.25 for $\bar{p} = 0.01$, the case of the 99% VaR. Note that the relative precision gain may be higher for the ES.

are provided in Appendix 2.C.2). This time, the NSEs for the VaR are four to eight times lower when computed using the PMitISEM based QERMit than in the case of direct sampling. For the ES PMitISEM outperforms the direct approach by more than three times. Notice that the RNEs for the QERMit based methods are astonishingly high, for the VaR ranging from over 3000 for $H = 10$, to 30 at $H = 250$, and for the ES from 250 to 10, respectively. This clearly demonstrates that in the frequentist case using IS rather than the direct simulation faces no limits on the relative precision gain and is "barely" constrained by our ability to construct an accurate candidate density.

Regarding the crucial time-precision trade-off, Table 2.3.2 shows that once again the slopes for the QERMit-based methods are higher than these of the direct approach (Appendix 2.D.2 provides the corresponding plots for all the horizons), usually two to three times. For some horizons, however, the increase in the slope due to adopting our IS-based method is much higher (for $H = 100$ it is over 7 for the VaR and over 13 for the ES). Interestingly, for $H = 10$ the superior algorithm turns out to be basic MitISEM and not PMitISEM, which delivers a slightly lower slope for the VaR than the direct approach.

An important remark must be made on the differences in sampling times between the Bayesian and the frequentist applications. *Any* frequentist method is extremely fast compared to any Bayesian method. Considering Table 2.2.4 for the Bayesian case and Table 2.3.2 for the frequentist case reveals that in the latter the sampling is usually faster by two orders of magnitude than in the former. The obvious reason for this speed of the frequentist sampling is that each logreturn draw $\mathbf{y}^{*(i)}$, $i = 1, \ldots, M$, is based on the common value of the parameter $\boldsymbol{\theta}$, fixed at the MLE. Therefore, not only no time is spent on drawing parameters from the posterior, but also on calculating the implied time T volatilities $h_T^{(i)}$, necessary for prediction of the future volatilities. In the frequentist case the direct sampling time consists therefore barely of drawing i.i.d. variates from the Student's t target (i.e. $\varepsilon_1^{(i)}, \ldots, \varepsilon_H^{(i)}$) and running the H-step-ahead recursion implied by the model to obtain the final $PL(\mathbf{y}^{*(i)})$ value. When the QERMit methods are adopted, $\varepsilon_h^{(i)}$, $h = 1, \ldots, H$ are no longer independently drawn from a univariate target, but from more complex densities with an inner dependence structure, which makes the sampling more time consuming.

The fact that the direct approach is so fast in the frequentist case results in PMitISEM-based QERMit methods requiring relatively more time to reach the benchmark 1 digit precision (with 95% confidence), even though they are characterised by higher slopes. Fortunately, for QERMit based on basic MitISEM (when it is feasible) our method requires less time than the direct approach to achieve this benchmark precision level.

Interestingly, however, both QERMit methods require far fewer draws than the direct approach to estimate 99% VaR and ES with the above specified precision, which again needs to be related to the differences in sampling time. Finally, recall once again that if more precise evaluations are required or a higher confidence for the precision is considered, the *time required* would of course change in favour of the QERMit-based methods, due to their higher slopes.

2.3.2 GAS(1,1)-t

Finally, we turn to the frequentist analysis of the GAS(1,1)-t model applied to the highly volatile "crisis" data from Section 2.2.2. As in the previous frequentist application we fix the model parameters at their MLE values (reported in Table 2.2.5). Hence, now the future observation disturbances are drawn from the Student's t distribution with roughly 7 degrees of freedom.

Table 2.3.3 presents the results for the VaR and ES evaluation (see Appendix 2.C.2 for the corresponding plots). One can see that also this time the QERMit-based methods generate much more accurate forecasts. For shorter horizons the NSE for the VaR is 5 to 6 times lower when evaluated with PMitISEM based QERMit than when computed directly, while for the ES the improvement ranges from 3 to 6 times. For both the VaR and the ES, the accuracy gain for long horizons is slightly lower, but still above 3 times. The IQR follows a similar pattern to the NSE, with the relative advantage of the QERMit-based methods being greater for the VaR than for the ES, and gradually slightly diminishing with the length of the forecast horizon.

Regarding the time-precision trade-off, Table 2.3.4 shows that for all the measure-horizon combinations we obtain considerable efficiency gains from adopting tail-focused densities (the visualisation of the results can be found in Appendix 2.D.2). This translates to fewer draws being required to achieve the 1 digit precision by the QERMit methods. Due to the specific nature of the frequentist sampling time, the PMitISEM-based QERMit in some cases requires more time for that purpose, yet with MitISEM we obtain gains also in this regard. A more demanding precision requirement would make both QERMit methods more competitive compared to the direct sampling both in terms of time and draws required.

H		VaR_{naive}	VaR_{mit}	VaR_{pmit}	ES_{naive}	ES_{mit}	ES_{pmit}
10		-7.9417	-7.8999	-7.8988	-9.5686	-9.5494	-9.5218
	NSE	(0.1496)	0.0256	(0.0179)	(0.2284)	(0.0900)	(0.0633)
	IQR	[0.1829]	[0.0235]	[0.0284]	[0.3452]	[0.1167]	[0.0812]
	RNE	1.00	1530.93	3129.10	1.00	123.57	249.28
20		-10.7175	-10.7775	-10.7904	-13.0425	-13.1050	-13.1076
	NSE	(0.2484)	0.0786	(0.0421)	(0.3270)	(0.0844)	(0.0969)
	IQR	[0.3229]	[0.0771]	[0.0404]	[0.4686]	[0.0834]	[0.0761]
	RNE	1.00	161.73	565.28	1.00	140.24	106.51
40		-14.5069	-14.4811	-14.5548	-17.7166	-17.8923	-17.8710
	NSE	(0.2999)	0.1981	(0.0630)	(0.5337)	(0.3440)	(0.0913)
	IQR	[0.3746]	[0.3215]	[0.0565]	[0.8246]	[0.2560]	[0.0760]
	RNE	1.00	25.49	251.59	1.00	8.45	120.01
100		-20.8270	–	-20.7822	-26.2797	–	-26.1151
	NSE	(0.7637)	(–)	(0.0882)	(1.1799)	(–)	(0.0956)
	IQR	[0.7992]	[–]	[0.0881]	[1.2810]	[–]	[0.1464]
	RNE	1.00	–	128.59	1.00	–	109.42
250		-27.6190	–	-27.3962	-35.6952	–	-35.4605
	NSE	(0.7819)	(–)	(0.1804)	(1.3967)	(–)	(0.3207)
	IQR	[1.0447]	[–]	[0.2453]	[2.1390]	[–]	[0.3624]
	RNE	1.00	–	30.73	1.00	–	9.73

Missing value (–): it was not possible to generate the particular result with the corresponding algorithm.

Table 2.3.1: Results for the 99% VaR and ES, in the **GARCH(1,1)-t** model, based on $N = 10000$ draws and 20 replications to obtain the numerical standard error (NSE) and the interquartile range (IQR). The RNE is the relative numerical efficiency, the inverse of the inefficiency factor. The results are obtained using the direct approach (with naive and adapted candidate distribution in the Metropolis-Hastings algorithm), and the QERMit method (with the basic MitISEM and PMitISEM methods), respectively.

	Direct	QERMit		Direct	QERMit	
H	Naive	MitISEM	PMitISEM	Naive	MitISEM	PMitISEM
	Total time					
10	0.90 s	1.30 s	2.15 s			
20	0.92 s	2.08 s	6.85 s			
40	0.96 s	1.17 s	23.98 s			
100	1.04 s	–	146.35 s			
250	1.27 s	–	1015.70 s			
	Construction time			Sampling time		
10	0.88 s	1.26 s	2.01 s	0.02 s	0.04 s	0.15 s
20	0.89 s	2.04 s	6.53 s	0.03 s	0.05 s	0.33 s
40	0.90 s	1.09 s	23.35 s	0.06 s	0.08 s	0.62 s
100	0.89 s	–	144.71 s	0.15 s	–	1.63 s
250	0.90 s	–	1010.97 s	0.37 s	–	4.74 s
	VaR slope*			**ES slope***		
10	**2,464.45**	**42,196.32**	**21,071.42**	**1,056.97**	**3,405.84**	**1,678.66**
20	**490.24**	**3452.54**	**1,719.31**	**282.90**	**2,993.64**	**323.94**
40	**177.31**	**300.83**	**405.28**	**55.98**	**99.70**	**193.33**
100	**11.44**	–	**78.70**	**4.79**	–	**66.96**
250	**4.42**	–	**6.49**	**1.39**	–	**2.05**
	VaR time required**			ES time required**		
10	1.51 s	1.30 s	2.08 s	2.34 s	1.71 s	2.92 s
20	4.02 s	2.48 s	7.42 s	6.32 s	2.55 s	11.27 s
40	9.56 s	6.20 s	27.15 s	28.35 s	16.50 s	31.30 s
100	135.22 s	–	164.24 s	321.53 s	–	167.66 s
250	348.18 s	–	1,247.76 s	1,108.99 s	–	1,759.19 s
	VaR draws required**			ES draws required**		
10	343,912	10,037	4,911	801,869	124,356	61,643
20	948,367	95,010	27,184	1,643,393	109,575	144,275
40	1,381,752	602,821	61,078	4,376,720	1,818,906	128,039
100	8,962,809	–	119,498	21,394,002	–	140,435
250	9,395,216	–	500,025	29,977,657	–	1,580,009

Missing value (–): it was not possible to generate the particular result with the corresponding algorithm.

*Slope = increase in precision per unit of computing time.

**Required for % estimate with 1 digit of precision (with 95% confidence).

Table 2.3.2: Trade-off of precision versus computing time for the 99% VaR and ES in **GARCH(1,1)-t** model for different horizons.

H		VaR_{naive}	VaR_{mit}	VaR_{pmit}	ES_{naive}	ES_{mit}	ES_{pmit}
10		-9.3886	-9.3681	-9.3562	-11.4654	-11.4724	-11.4604
	NSE	(0.2346)	0.0588	(0.0346)	(0.2467)	(0.1043)	(0.0711)
	IQR	[0.2717]	[0.0677]	[0.0430]	[0.3244]	[0.1365]	[0.0797]
	RNE	1.00	288.88	835.34	1.00	91.88	198.07
20		-12.5140	-12.4591	-12.5526	-15.3418	-15.4223	-15.4870
	NSE	(0.3144)	0.1222	(0.0673)	(0.4080)	(0.2497)	(0.0658)
	IQR	[0.4695]	[0.1413]	[0.0562]	[0.5067]	[0.1851]	[0.0760]
	RNE	1.00	66.96	220.75	1.00	16.04	231.05
40		-16.2119	–	-16.3093	-20.4478	–	-20.4527
	NSE	(0.2933)	–	(0.0769)	(0.5824)	(–)	(0.0941)
	IQR	[0.4169]	[–]	[0.0991]	[0.9540]	[–]	[0.1494]
	RNE	1.00	–	169.30	1.00	–	112.83
100		-21.4720	–	-21.2891	-27.5570	–	-27.3327
	NSE	(0.6093)	–	(0.1695)	(0.8992)	(–)	(0.1591)
	IQR	[0.7680]	[–]	[0.1984]	[1.2750]	[–]	[0.2958]
	RNE	1.00	–	34.81	1.00	–	39.50
250		-24.3314	–	-24.2340	-32.3997	–	-32.2739
	NSE	(0.7701)	–	(0.2084)	(1.4013)	(–)	(0.2369)
	IQR	[0.9156]	[–]	[0.2901]	[1.5927]	[–]	[0.3614]
	RNE	1.00	–	23.02	1.00	–	17.81

Missing value (–): it was not possible to generate the particular result with the corresponding algorithm.

Table 2.3.3: Results for the 99% VaR and ES, in the **GAS(1,1)-t** model, based on $N = 10000$ draws and 20 replications to obtain the numerical standard error (NSE) and the interquartile range (IQR). The RNE is the relative numerical efficiency, the inverse of the inefficiency factor. The results are obtained using the direct approach (with naive and adapted candidate distribution in the Metropolis-Hastings algorithm), and the QERMit method (with the basic MitISEM and PMitISEM methods), respectively.

	Direct	QERMit		Direct	QERMit	
H	Naive	MitISEM	PMitISEM	Naive	MitISEM	PMitISEM
	Total time					
10	4.35 s	0.67 s	3.12 s			
20	4.35 s	4.78 s	9.46 s			
40	4.38 s	–	23.71 s			
100	4.48 s	–	145.31 s			
250	4.85 s	–	988.06 s			
	Construction time			Sampling time		
10	4.33 s	0.64 s	2.98 s	0.02 s	0.03 s	0.14 s
20	4.31 s	4.73 s	9.16 s	0.03 s	0.05 s	0.29 s
40	4.31 s	–	23.10 s	0.06 s	–	0.61 s
100	4.32 s	–	143.79 s	0.15 s	–	1.52 s
250	4.47 s	–	983.40 s	0.38 s	–	4.66 s
	VaR slope*			**ES slope***		
10	**974.26**	**9,752.17**	**5789.43**	**880.80**	**3,101.69**	**1,372.74**
20	**293.21**	**1,389.32**	**749.19**	**174.13**	**332.71**	**784.13**
40	**182.63**	–	**276.93**	**46.30**	–	**184.56**
100	**17.45**	–	**22.90**	**8.01**	–	**25.98**
250	**4.48**	–	**4.94**	**1.35**	–	**3.82**
	VaR time required*			ES time required*		
10	5.91 s	0.80 s	3.24 s	6.08 s	1.14 s	4.09 s
20	9.55 s	5.83 s	11.22 s	13.14 s	9.35 s	11.12 s
40	12.73 s	–	28.65 s	37.50 s	–	31.43 s
100	92.40 s	–	210.89 s	196.11 s	–	202.93 s
250	347.47 s	–	1,294.69 s	1,140.28 s	–	1,385.65 s
	VaR draws required**			ES draws required**		
10	845,752	53,194	18,395	935,498	167248	77,581
20	1,519,296	229,489	69,609	2,558,257	958,281	66,507
40	1,321,478	–	90,766	5,212,265	–	136,195
100	5,705,631	–	441,394	12,423,807	–	388,979
250	9,112,650	–	667,648	30,176,110	–	862,735

Missing value (–): it was not possible to generate the particular result with the corresponding algorithm.

**Slope = increase in precision per unit of computing time.

**Required for % estimate with 1 digit of precision (with 95% confidence).

Table 2.3.4: Trade-off of precision versus computing time for the 99% VaR and ES in **GAS(1,1)-t** model for different horizons.

2.4 Conclusions

We have proposed an efficient importance sampling based method for the Bayesian risk evaluation, given a chosen model of volatility. We focus on two standard risk measures, Value-at-Risk and Expected Shortfall. The proposed method enables an accurate analysis even for long horizons, such as one-month or one-year-ahead. We have carried out two empirical studies for daily S&P 500 returns in different time periods, a calm period and a highly volatile crisis period. Both applications confirm that our method not only yields more accurate results than the direct sampling approach, commonly used in practice (see The Volatility Laboratory, 2012), but also achieves this in a time efficient way, resulting in a considerable gain in terms of time-precision trade-off. This substantial extension of the applicability of importance sampling to the simulation of returns for long horizons is to be attributed to the sequential construction of the marginal and conditional importance densities, which are flexible mixtures of Student's t distributions.

The proposed method succeeds also for the frequentist applications, in terms of yielding a higher precision gain for a unit of computing time. However, due to generally very fast computations in the frequentist case, the advantage of the QERMit method relative to the direct approach depends on the required precision level or on the chosen confidence for the precision. We do stress that in the context of long run risk evaluation, Bayesian analysis provides a more natural framework due to accounting for parameter uncertainty.

An interesting and important topic for further research is the application of our method to a multidimensional case. We intend to investigate portfolios of several assets using e.g. a copula based on a GAS model for volatility. Nevertheless, already the current study might be useful in a multidimensional context, since we consider the ES and not merely the VaR. Because the ES is a subadditive measure, a sum of the ES estimates for single assets provides a conservative risk measure for a portfolio consisting of these assets.

Another possible line for further research would be to build on the insights from QERMit in the context of credit risk evaluation. The phenomena of default dependence can be elegantly captured by Bayesian inference (see McNeil and Wendin, 2007) while importance sampling is an advantageous variance reduction technique also in this context (see Glasserman and Li, 2005). Analysis of portfolio defaults would naturally require different modelling techniques, yet the key element of QERMit, i.e. the focus on the tail, is expected to yield considerable efficiency gains also in this area.

Appendix 2.A MitISEM Algorithm

2.A.1 Approximation by minimisation of Kullback-Leibler divergence

We want to approximate the target density $\tilde{p}(\boldsymbol{\theta})$ of which only the kernel $p(\boldsymbol{\theta})$ is required with the candidate density $q_\zeta(\boldsymbol{\theta})$, parametrised by vector ζ, such that the Kullback-Leibler divergence (Kullback and Leibler, 1951)

$$\int p(\boldsymbol{\theta}) \log p(\boldsymbol{\theta}) d\boldsymbol{\theta} - \int p(\boldsymbol{\theta}) \log q_\zeta(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{2.A.1}$$

is minimised. The target density p will usually be the posterior density given the data y, but we omit the conditioning on y for the notational convenience. Moreover, we will take as the candidate q_ζ the mixture of Student's t distributions, so that the minimisation will be carried out with respect to the mixture parameters ζ, consisting of the mixture weights and the modes, scale matrices and degrees of freedom of each component as well as the number of mixture components H. Since the first term in (2.A.1) does not depend on ζ, the minimisation of (2.A.1) amounts to the maximisation of

$$\begin{aligned}
\int \log q_\zeta(\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} &= \int \log q_\zeta(\boldsymbol{\theta}) \frac{p(\boldsymbol{\theta})}{q_\zeta(\boldsymbol{\theta})} q_\zeta(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
&= \mathbb{E}_{q_\zeta}\left[\log g(\boldsymbol{\theta}) \frac{p(\boldsymbol{\theta})}{q_\zeta(\boldsymbol{\theta})}\right], \\
&\approx \frac{1}{N} \sum_{i=1}^{N} \log q_\zeta(\boldsymbol{\theta}^{(i)}) \frac{p(\boldsymbol{\theta}^{(i)})}{q_\zeta(\boldsymbol{\theta}^{(i)})} \\
&= \frac{1}{N} \sum_{i=1}^{N} \log q_\zeta(\boldsymbol{\theta}^{(i)}) w(\boldsymbol{\theta}^{(i)}),
\end{aligned}$$

where $\boldsymbol{\theta}^{(i)} \overset{i.i.d.}{\sim} q_{\zeta_{old}}(\boldsymbol{\theta})$ were drawn from the previous candidate, and

$$w(\boldsymbol{\theta}^{(i)}) = \frac{p(\boldsymbol{\theta}^{(i)})}{q_\zeta(\boldsymbol{\theta}^{(i)})}. \tag{2.A.2}$$

Importantly, the draws $\boldsymbol{\theta}^{(i)}$, $i = 1, \ldots, N$, and their weights $w(\boldsymbol{\theta})^{(i)}$ are fixed during the optimization and they do not depend on ζ.

2.A.2 EM step in MitISEM

Consider a mixture of C Student-t densities

$$q_\zeta(\boldsymbol{\theta}) = \sum_{c=1}^{C} \eta_h t(\boldsymbol{\theta}|\mu_c, \Sigma_c, \nu_c), \tag{2.A.3}$$

where $t(\boldsymbol{\theta}|\mu, \Sigma, \nu)$ denotes the d-dimensional Student-t density

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

and $\zeta = \{\mu_c, \Sigma_c, \nu_c, \eta_c\}_{h=1}^{H}$ is the set of the mixture parameters: modes, scale matrices, degrees of freedom and mixing probabilities. The aim is to maximise the weighted log-density

$$\frac{1}{N}\sum_{i=1}^{N} w^{(i)} \log q_\zeta(\boldsymbol{\theta}^{(i)}), \tag{2.A.4}$$

with respect to ζ, where $w^{(i)} = w(\boldsymbol{\theta}^{(i)}) = \frac{p(\boldsymbol{\theta}^{(i)})}{q_\zeta(\boldsymbol{\theta}^{(i)})}$ is the importance weight of the draw $\boldsymbol{\theta}^{(i)}$. Using the fact the a Student's t distribution can be represented as a mixture of normal distributions with the covariance matrices scaled by the random variables following an Inverse-Gamma distribution, one can equivalently represent the draws $\boldsymbol{\theta}^{(i)}$ from the mixture (2.A.3) in (2.A.4) as

$$\boldsymbol{\theta}^{(i)} \sim \mathcal{N}(\mu_c, \kappa_c^{(i)}\Sigma_c), \qquad \text{if } z_c^{(i)} = 1,$$

where $z^{(i)} \in \mathbb{R}^H$ is a latent vector from the standard base with one on the place corresponding to the component h which the draw $\boldsymbol{\theta}^{(i)}$ has been drawn from. The probability $\mathbb{P}[z^{(i)} = e_c]$ of belonging to the component h is given by η_c. The scaling factor $\kappa_c^{(i)}$ follows the Inverse-Gamma distribution

$$\kappa_c^{(i)} \sim \mathcal{IG}(\nu_c/2, \nu_c/2).$$

Such a representation introduces the latent data $\tilde{\boldsymbol{\theta}} = \{z_c, \kappa_c\}_{h=1}^{C}$ into the logdensity $\log p(\boldsymbol{\theta})$, so that the standard numerical maximisation of the data-augmented $\log p(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}|\zeta)$ density is infeasible. To find the optimal mixture parameters ζ one can resort to the expectation-maximisation (EM) algorithm of Dempster et al. (1977), which allows for the maximum likelihood estimation for the incomplete data problems. The

core of the procedure is to iterate between two steps, the Expectation step and the Maximisation step. In the former, one calculates the conditional expectation of the loglikelihood function with respect to the latent variables $\tilde{\boldsymbol{\theta}}$, given the parameter values from the previous iteration, ζ. In the latter, the expected loglikelihood is maximised with respect to the parameters.

Expectation step The conditional expectations in the Expectation step are given by

$$\begin{aligned}
\tilde{z}_c^{(i)} &\equiv \mathbb{E}\left[z_c^{(i)} \middle| \boldsymbol{\theta}^{(i)}, \zeta \right] = \frac{\eta_c t(\boldsymbol{\theta}^{(i)} | \mu_c, \Sigma_c, \nu_c)}{\sum_{l=1}^{H} \eta_l t(\boldsymbol{\theta}^{(i)} | \mu_l, \Sigma_l, \nu_l)}, \\
\widetilde{z/\kappa}_c^{(i)} &\equiv \mathbb{E}\left[\frac{z_c^{(i)}}{\kappa_c^{(i)}} \middle| \boldsymbol{\theta}^{(i)}, \zeta \right] = \tilde{z}_c^{(i)} \frac{d + \nu_c}{\rho_c^{(i)} + \nu_c}, \\
\tilde{\xi}_c^{(i)} &\equiv \mathbb{E}\left[\log \kappa_c^{(i)} \middle| \boldsymbol{\theta}^{(i)}, \zeta \right] \\
&= \left[\log\left(\frac{\rho_c^{(i)} + \nu_c}{2} \right) - \psi\left(\frac{d + \nu_c}{2} \right) \right] \tilde{z}_c^{(i)} + \left[\log\left(\frac{\nu_c}{2} \right) - \psi\left(\frac{\nu_c}{2} \right) \right] (1 - \tilde{z}_c^{(i)}), \\
\tilde{\delta}_c^{(i)} &\equiv \mathbb{E}\left[\frac{1}{\kappa_c^{(i)}} \middle| \boldsymbol{\theta}^{(i)}, \zeta \right] = \widetilde{z/\kappa}_c^{(i)} + (1 - \tilde{z}_c^{(i)}),
\end{aligned}$$

where $\rho_c^{(i)} = (\boldsymbol{\theta}^{(i)} - \mu_c)^T \Sigma_c^{-1} (\boldsymbol{\theta}^{(i)} - \mu_c)$ and ψ denotes the digamma function.

Maximisation step The updates at the iteration L of the Maximisation step are as follows

$$\begin{aligned}
\mu_c^{(L)} &= \left[\sum_{i=1}^{N} w^{(i)} \widetilde{z/\kappa}_c^{(i)} \right]^{-1} \left[\sum_{i=1}^{N} w^{(i)} \widetilde{z/\kappa}_c^{(i)} \boldsymbol{\theta}^{(i)} \right], \\
\Sigma_c^{(L)} &= \frac{\sum_{i=1}^{N} \kappa^{(i)} \widetilde{z/\kappa}_c^{(i)} (\boldsymbol{\theta}^{(i)} - \mu_c^{(L)})(\boldsymbol{\theta}^{(i)} - \mu_c^{(L)})^T}{\sum_{i=1}^{N} w^{(i)} \tilde{z}_c^{(i)}}, \\
\eta_c^{(L)} &= \frac{\sum_{i=1}^{N} w^{(i)} \tilde{z}_c^{(i)}}{\sum_{i=1}^{N} w^{(i)}},
\end{aligned}$$

while the updates for the degrees of freedom $\nu_c^{(L)}$ parameters come from solving of the first-order conditions with respect to ν_c

$$-\psi(\nu_c/2) + \log(\nu_c/2) + 1 - \frac{\sum_{i=1}^{N} w^{(i)} \xi_c^{(i)}}{\sum_{i=1}^{N} w^{(i)}} - \frac{\sum_{i=1}^{N} w^{(i)} \delta_c^{(i)}}{\sum_{i=1}^{N} w^{(i)}} = 0.$$

A more detailed discussion of the MitISEM algorithm can be found in Hoogerheide et al. (2012).

Appendix 2.B PMitISEM Algorithm

Below we present the details of the Partial MitISEM algorithm of Hoogerheide et al. (2012).

Step 0: Initialisation

$\boldsymbol{\theta}^{(i)} \sim g_{\text{naive}}$, $i = 1, \ldots, N$

$(\mu_{\text{naive}} = \hat{\boldsymbol{\theta}} \equiv \arg\max_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$, $\Sigma_{\text{naive}} = -\mathcal{H}^{-1} \left(\log f(\boldsymbol{\theta})\right)|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$)

Step 1: Adaptation

Use $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^{N}$ to IS-estimate the mean and the covariance matrix of f by μ_{adapt} and Σ_{adapt}.

Use μ_{adapt} and Σ_{adapt} to construct g_{adapt}.

Set $g_0 = g_{\text{adapt}}$.

$\boldsymbol{\theta}^{(i)} \sim g_0$, $i = 1, \ldots, N$.

$w_0^{(i)} = \frac{f(\boldsymbol{\theta}^{(i)})}{g_0(\boldsymbol{\theta}^{(i)})}$, $i = 1, \ldots, N$.

Step 2: Construction

for $s := 1$ **to** S **do**

Step 2a: ISEM

Run ISEM with $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^{N}$ and $\{w_0^{(i)}\}_{i=1}^{N}$ to optimise C_s components of g_s.

$(g_s(\boldsymbol{\theta}) = g(\boldsymbol{\theta}_s|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{s-1})$ for $s = 2, \ldots, S$, and $g_s(\boldsymbol{\theta}) = g(\boldsymbol{\theta}_1)$ for $s = 1)$

Calculate the *current* weights[1] of the draws $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^{N}$ from g_0 using the optimised candidate with C_S components with formula:

$$w_{\text{curr}}^{(i)} = \frac{f(\boldsymbol{\theta}^{(i)})}{\prod_{k=1}^{S} g_k(\boldsymbol{\theta}^{(i)})}. \tag{2.B.1}$$

Compute CoV_s for g_s using $\{w_{\text{curr}}^{(i)}\}_{i=1}^{N}$.

Step 2b: Iterate

while CoV_s not converged **do**

Find $\tilde{\boldsymbol{\theta}}^{j}$, $j \in \mathcal{M}$, where $\mathcal{M}$ is the set of indices of $x\%$ of the draws $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^{N}$ which correspond to the highest weights $\{w_{\text{curr}}^{(i)}\}_{i=1}^{N}$.

Use $\{\tilde{\boldsymbol{\theta}}^{(j)}\}_{j\in\mathcal{M}}$ and $\{w_{\text{curr}}^{(j)}\}_{j\in\mathcal{M}}$ to IS-construct the mode/coefficients and the covariance matrix of the C_s+1-th component of g_s as $\mu^s_{C_s+1}/\beta^s_{C_s+1}$ and $\Sigma^s_{C_s+1}$.
Update the current mixture g_s[2].
Run ISEM with $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^N$ and $\{w_0^{(i)}\}_{i=1}^N$ to optimise C_s+1 components of the updated g_s.
Calculate the new current weights $\{w_{\text{curr}}^{(i)}\}_{i=1}^N$ of the draws $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^N$ from g_0 using the latest candidate with formula (2.B.1).
(note that now g_s is an updated mixture of C_s+1 components)
Compute CoV_s for the latest candidate.

end while

end for

Step 3: Resimulation and convergence check

$\boldsymbol{\theta}^i \sim \prod_{s=1}^S g(s)(\boldsymbol{\theta})$, $i=1,\ldots,N$.
$w^{(i)} = \frac{f(\boldsymbol{\theta}^{(i)})}{\prod_{s=1}^S g(s)(\boldsymbol{\theta}^{(i)})}$, $i=1,\ldots,N$.
Update g to $\tilde{g}$:

for $s := 1$ **to** S **do**

Run ISEM with $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^N$ and $\{w^{(i)}\}_{i=1}^N$ to optimise components of g_s to obtain $\tilde{g}_s$.

end for

Calculate the new current weights $\{w_{\text{curr}}^{(i)}\}_{i=1}^N$ of the draws $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^N$ from g using the optimised candidate $\tilde{g}$ with formula (2.B.1).
If CoV has not converged set $g_0 := \tilde{g}$ and $w_0^{(i)} := w_{\text{curr}}^{(i)}$; else STOP.

---

[1] "Current" because these are not the "real" importance weights as the draws are fixed and coming from g_0, not from the updated candidate.

[2] Updating is done in the "standard" way: μ^s_h/β^s_h, Σ^s_h and ν^s_h for the old components $h=1,\ldots,C_s$, remain unchanged; $\mu^s_{C_s+1}/\beta^s_{C_s+1}$ and $\Sigma^s_{C_s+1}$ for the new components is are set to the current estimates based on $\{\tilde{\boldsymbol{\theta}}^{(j)}\}_{j\in\mathcal{M}}$; $\nu^s_{C_s+1}$ is set to some chosen initial value; $\eta^s_h := 0.9\eta^s_h$, $h=1,\ldots,C_s$ and $\eta_{C_s+1} := 0.1$.

Appendix 2.C Accuracy plots

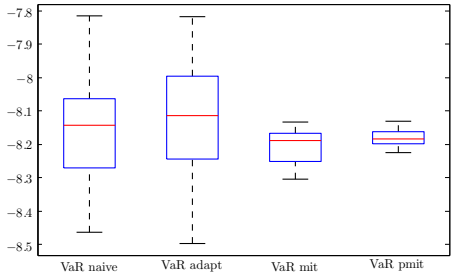
The plots in this appendix present the accuracy of 99% VaR and ES evaluations obtained with different algorithms. We consider two types of plots, standard box plots and error bar plots. The motivation behind this particular choice is that the former plot type is a popular and commonly used in dispersion visualisation, while the latter is more suited to illustrate the results based on QERMit. The underlying objective of QERMit is minimisation of the NSE, however this measure is not illustrated in a box plot, which focuses on the IQR instead. In a sense both plot types provide complimentary information regarding the accuracy of a certain evaluation method. Notice possible outliers as these might be of crucial importance in the context of risk evaluation.

2.C.1 Bayesian applications

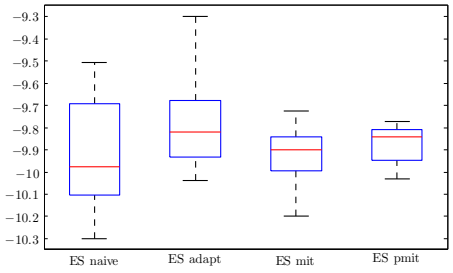
Figure 2.C.1 presents the results for the GARCH(1,1)-t model while Figure 2.C.2 for the GAS(1,1)-t model. Clearly, the precision of the QERMit-based methods greatly exceeds the one from both direct approaches. Moreover, the latter often generate outliers, which may have serious practical consequences.

2.C.2 Frequentist applications

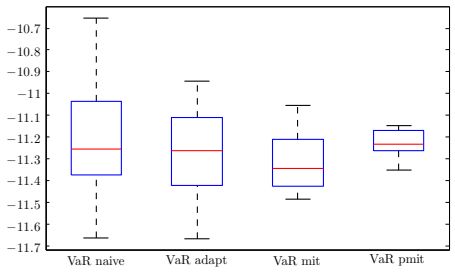
Figure 2.C.3 and Figures 2.C.4 are the frequentist counterparts of those presented in 2.C.1. We can see that the outcomes are similar to the Bayesian ones, with a much higher accuracy achieved with QERMit than with the direct approach.



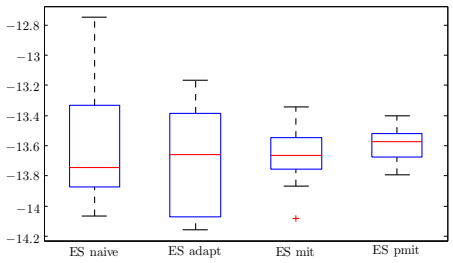
(a) VaR, $H = 10$



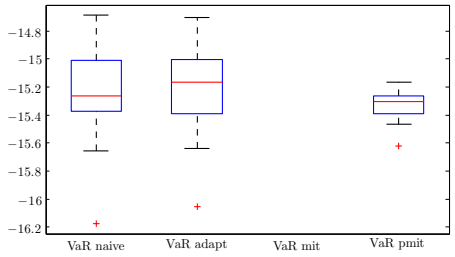
(b) ES, $H = 10$



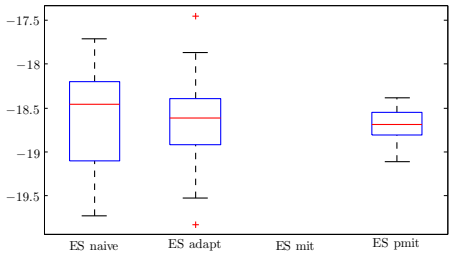
(c) VaR, $H = 20$



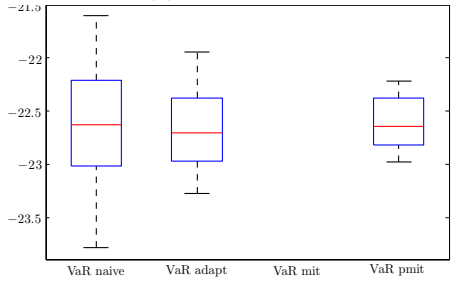
(d) ES, $H = 20$



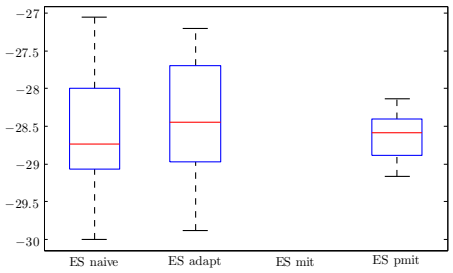
(e) VaR, $H = 40$



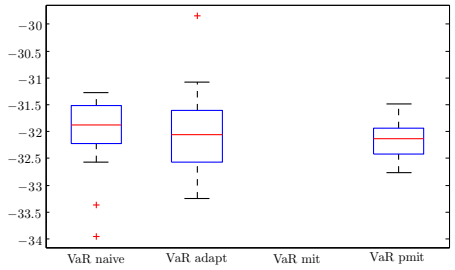
(f) ES, $H = 40$



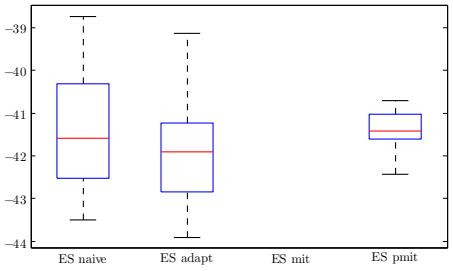
(g) VaR, $H = 100$



(h) ES, $H = 100$

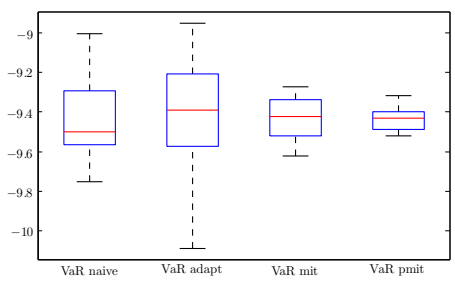


(i) VaR, $H = 250$

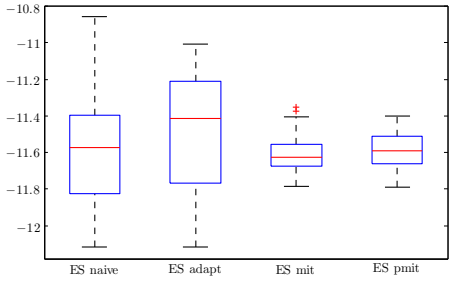


(j) ES, $H = 250$

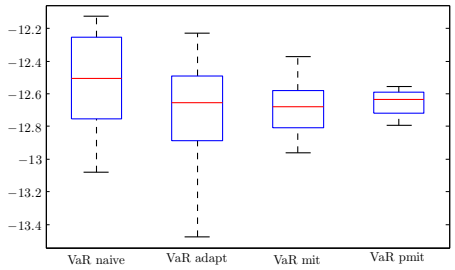
Figure 2.C.1: Accuracy of 99% VaR (left) and ES (right) results for the **GARCH(1,1)-t** model for different horizons, based on 20 MC replications. Two left boxes correspond to the direct approach (based on the naive and adapted candidate, respectively), two right ones – to the QERMit approach (with the candidate constructed with MitISEM and PMitISEM, respectively). A missing box for the MitISEM-based candidate corresponds to a situation when it was not possible to construct such a candidate.



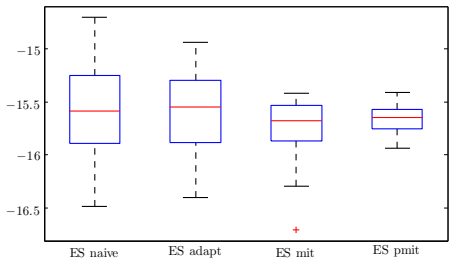
(a) VaR, $H = 10$



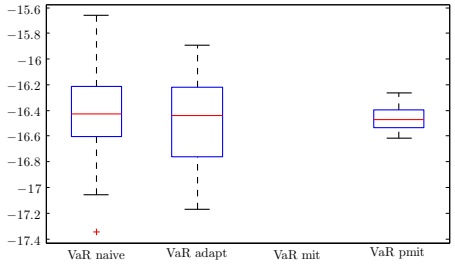
(b) ES, $H = 10$



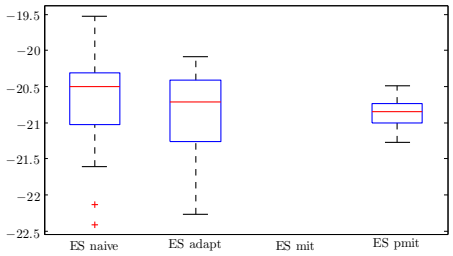
(c) VaR, $H = 20$



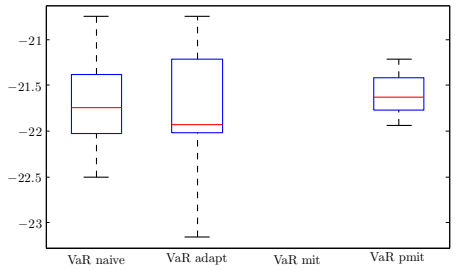
(d) ES, $H = 20$



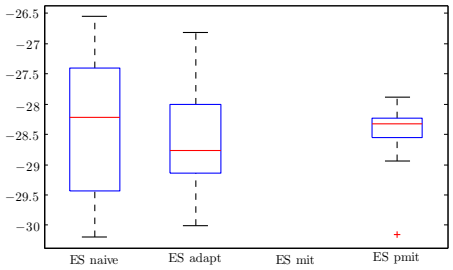
(e) VaR, $H = 40$



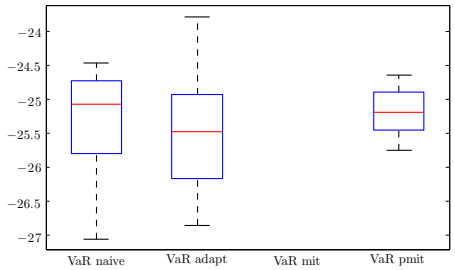
(f) ES, $H = 40$



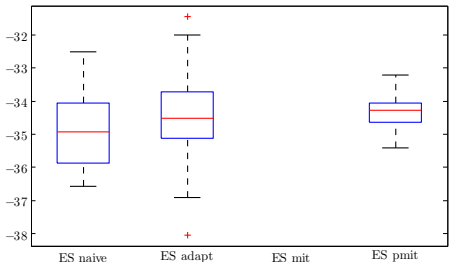
(g) VaR, $H = 100$



(h) ES, $H = 100$

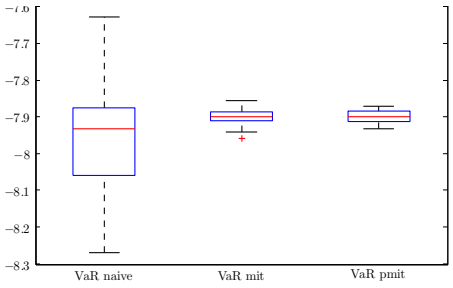


(i) VaR, $H = 250$

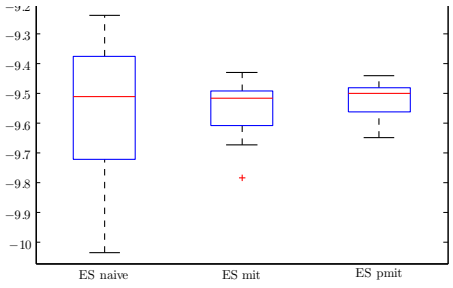


(j) ES, $H = 250$

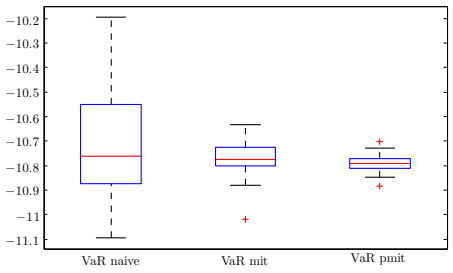
Figure 2.C.2: Accuracy of 99% VaR (left) and ES (right) results for the **GAS(1,1)-t** model for different horizons, based on 20 MC replications. Two left boxes correspond to the direct approach (based on the naive and adapted candidate, respectively), two right ones – to the QERMit approach (with the candidate constructed with MitISEM and PMitISEM, respectively). A missing box for the MitISEM-based candidate corresponds to a situation when it was not possible to construct such a candidate.



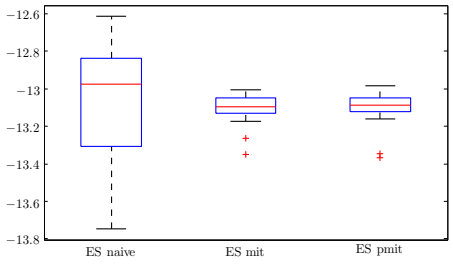
(a) VaR, $H = 10$



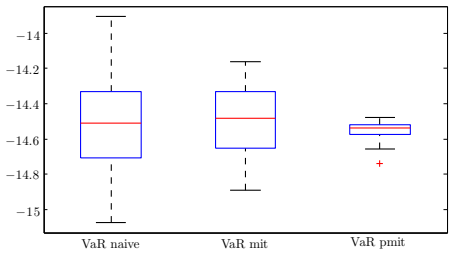
(b) ES, $H = 10$



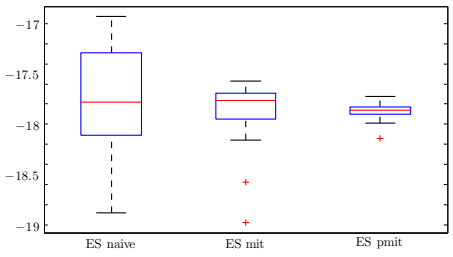
(c) VaR, $H = 20$



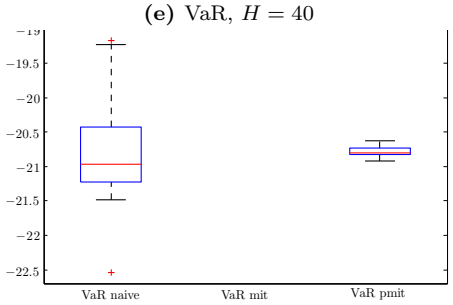
(d) ES, $H = 20$



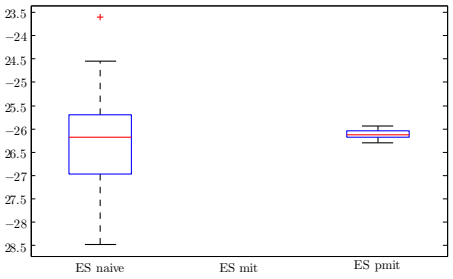
(e) VaR, $H = 40$



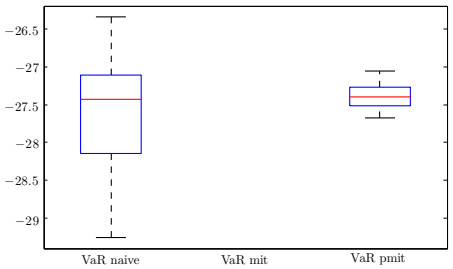
(f) ES, $H = 40$



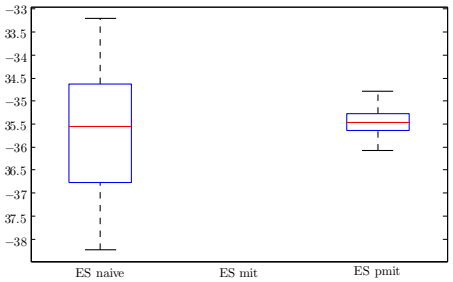
(g) VaR, $H = 100$



(h) ES, $H = 100$

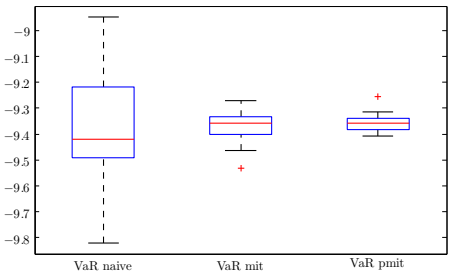


(i) VaR, $H = 250$

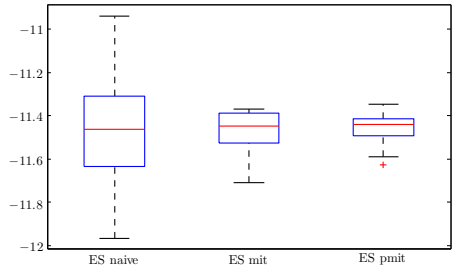


(j) ES, $H = 250$

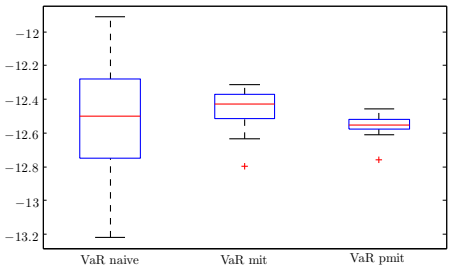
Figure 2.C.3: Accuracy of 99% VaR (left) and ES (right) results for the frequentist **GARCH(1,1)-t** model for different horizons, based on 20 MC replications. The left boxes correspond to the direct approach (based on the naive candidate), the middle and the right one – to the QERMit approach (with the candidate constructed with MitISEM and PMitISEM, respectively). A missing box for the MitISEM-based candidate corresponds to a situation when it was not possible to construct such a candidate.



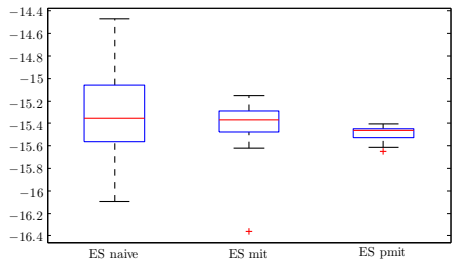
(a) VaR, $H = 10$



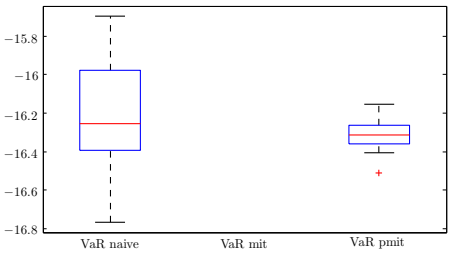
(b) ES, $H = 10$



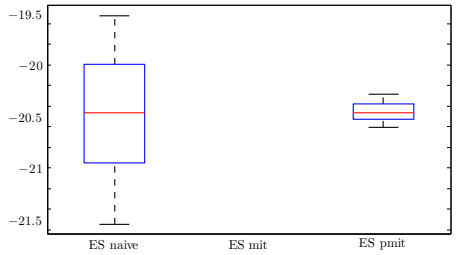
(c) VaR, $H = 20$



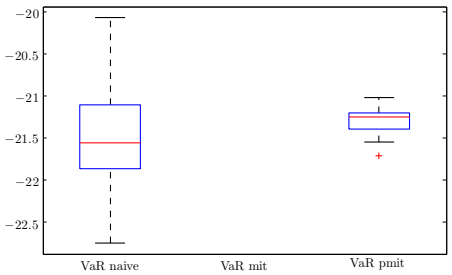
(d) ES, $H = 20$



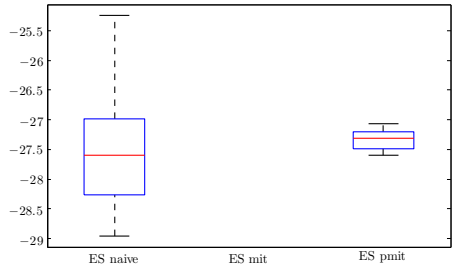
(e) VaR, $H = 40$



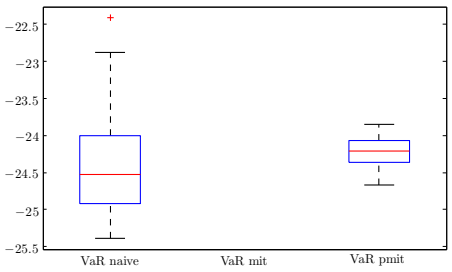
(f) ES, $H = 40$



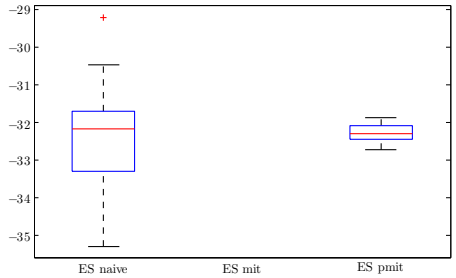
(g) VaR, $H = 100$



(h) ES, $H = 100$



(i) VaR, $H = 250$



(j) ES, $H = 250$

Figure 2.C.4: Accuracy of 99% VaR (left) and ES (right) results for the frequentist **GAS(1,1)-t** model for different horizons, based on 20 MC replications. The left boxes correspond to the direct approach (based on the naive candidate), the middle and the right one – to the QERMit approach (with the candidate constructed with MitISEM and PMitISEM, respectively). A missing box for the MitISEM-based candidate corresponds to a situation when it was not possible to construct such a candidate.

Appendix 2.D Time-precision plots

The plots in this appendix illustrate the time-precision trade-off for 99% VaR and ES evaluation. Precision is defined as the inverse of the variance of the results obtained in a Monte Carlo study, where we carried out 20 computations of VaR and ES. Computing time includes the "fixed cost" for the candidate construction, for which the lines are flat (negligible yet non-zero for the direct Bayesian methods, and more noticeable for the QERMit methods). The "variable cost" of the computing time refers to the time needed to perform a single VaR and ES evaluation, based on $N = 10,000$ parameter draws (for both, the direct and the QERMit approach). Note that the scales for the time axis differ among horizons. Then the slope of the non-flat part of each line is specified as the ratio of precision and sampling time. Following Hoogerheide and van Dijk (2010) we also consider the benchmark line of 1 digit precision with 95% confidence. It is defined as $1.96\text{NSE} \leq 0.05$, which corresponds to the required precision level of 1536, and is depicted in the plots as a black horizontal line.

2.D.1 Bayesian applications

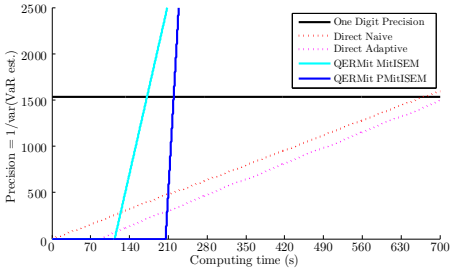
Figure 2.D.1 presents the results for the GARCH(1,1)-t model and Figure 2.D.2 for the GAS(1,1)-t. Importantly, for longer horizons ($H = 40$ and longer) there are no lines for QERMit based on the MitISEM algorithm, as it was not possible to apply it in such multidimensional cases. For both models the steepness of the QERMit methods is higher the for the direct approaches for both VaR and ES evaluations (see Tables 2.2.4 and 2.2.8 for the quantitative results), which means that if a high precision is required, then the proposed QERMit based methods will need less computing time to achieve this.

2.D.2 Frequentist applications

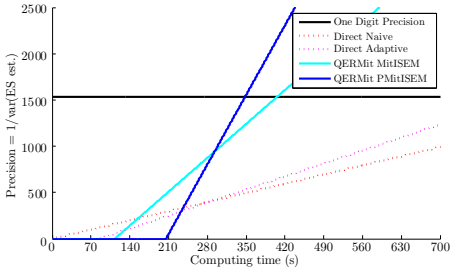
Figures 2.D.3 and 2.D.4 are the frequentist counterparts of the time-precision trade-off plots from Appendix 2.D.1. The main difference between the current plots and the previous ones is that now there are at most three lines in each plot, as we do not consider the adaptive direct method for the frequentist applications. Moreover, the "fixed cost" for the direct approach is exactly zero because is based on sampling of i.i.d. variates from a univariate standard Student's t distribution (with the number of degrees of freedom set equal to its MLE value) and not on a mixture of multivariate

Student's t distributions which needs to be constructed. Again, for longer horizons ($H = 100$ and over for GARCH and for $H = 40$ and over for GAS) there are no lines for QERMit based on MitISEM due to its infeasibility in high dimensions.

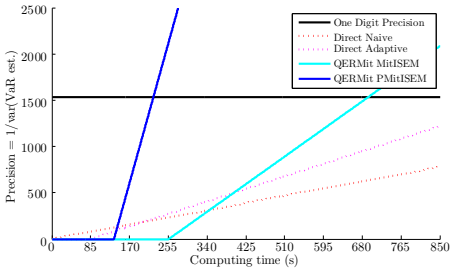
Even though for the longest horizons, $H \geq 100$ for GARCH and $H \geq 40$ for GAS, the direct approach is faster then QERMit in crossing the benchmark 1 digit precision line, higher slopes of the latter (see Tables 2.3.2 and 2.3.4) imply that eventually it is more efficient than the former. Notice, that the 1 digit precision line (with 95% confidence) was set somewhat arbitrarily and considering a higher confidence would mean a much higher line. For instance changing of the confidence to 99% would raise it from $1,536$ to $2,654$ so that in more cases less computing time would be needed to reach the required precision level with the QERMit approaches than with the direct one. This would be seen as more "crossings" of the lines for the direct and QERMit-based methods occurring below the required precision line.



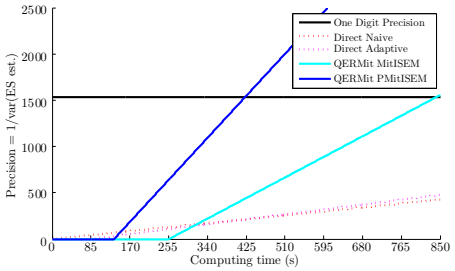
(a) VaR, $H = 10$



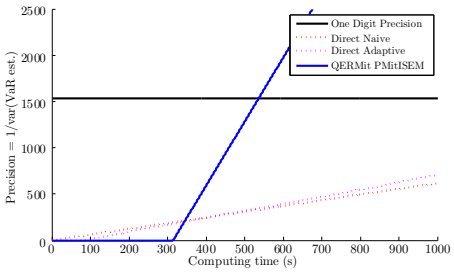
(b) ES, $H = 10$



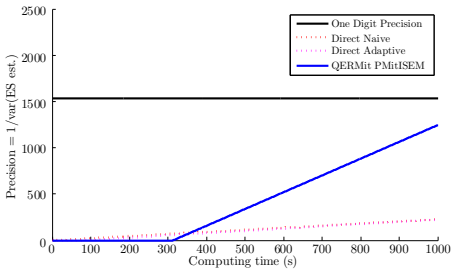
(c) VaR, $H = 20$



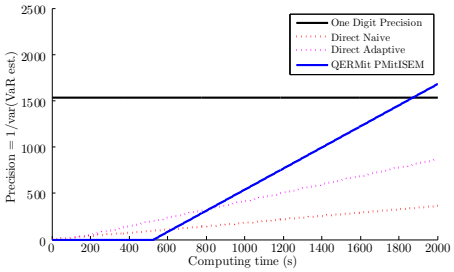
(d) ES, $H = 20$



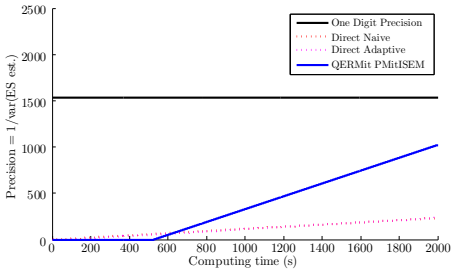
(e) VaR, $H = 40$



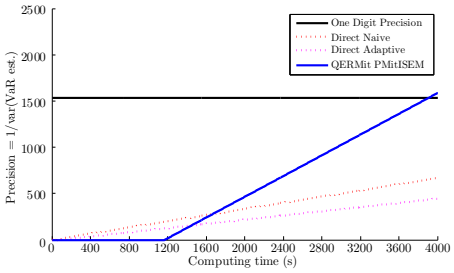
(f) ES, $H = 40$



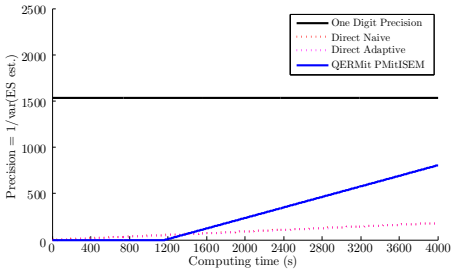
(g) VaR, $H = 100$



(h) ES, $H = 100$

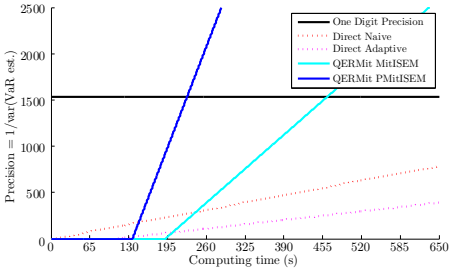


(i) VaR, $H = 250$

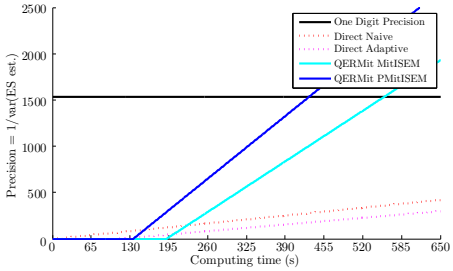


(j) ES, $H = 250$

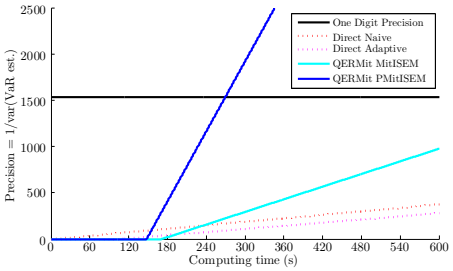
Figure 2.D.1: Precision ($1/var$) of the estimated VaR (left) and ES (right), as a function of the amount of computing time for different approaches, for the **GARCH(1,1)-t** model, for different horizons. The horizontal line corresponds to a precision of 1 digit ($1.96NSE \leq 0.05$). A missing line for the MitISEM-based candidate corresponds to a situation when it was not possible to construct such a candidate.



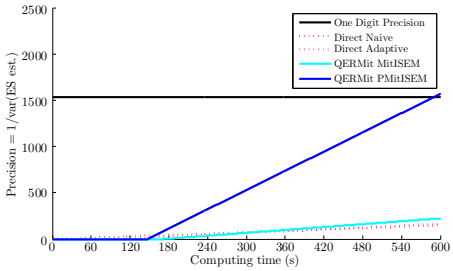
(a) VaR, $H = 10$



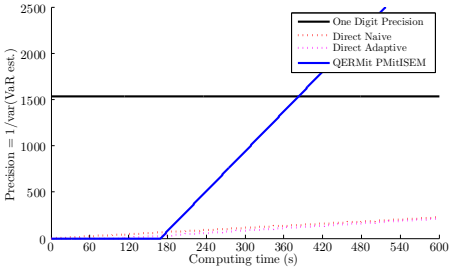
(b) ES, $H = 10$



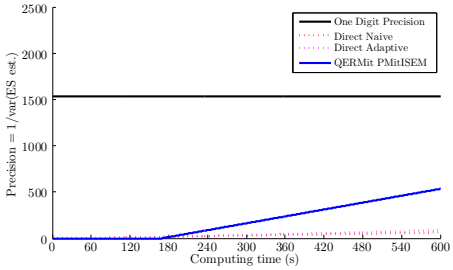
(c) VaR, $H = 20$



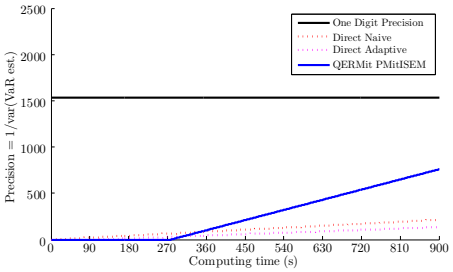
(d) ES, $H = 20$



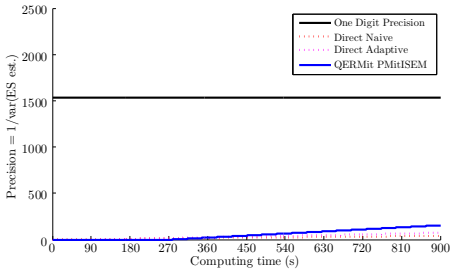
(e) VaR, $H = 40$



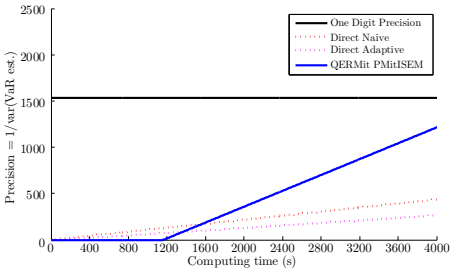
(f) ES, $H = 40$



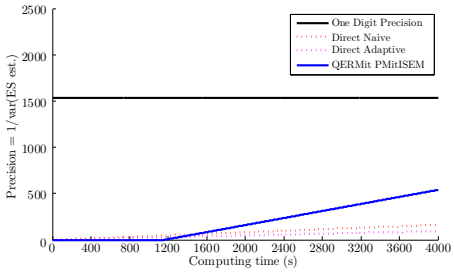
(g) VaR, $H = 100$



(h) ES, $H = 100$

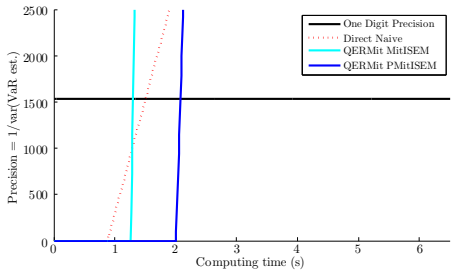


(i) VaR, $H = 250$

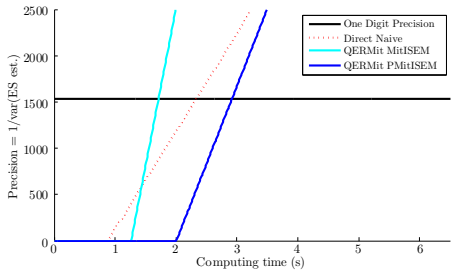


(j) ES, $H = 250$

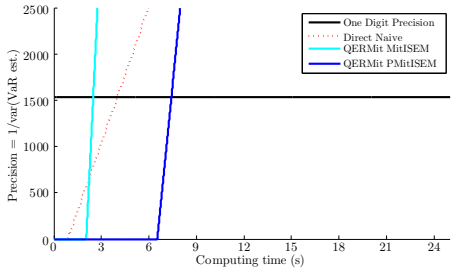
Figure 2.D.2: Precision ($1/var$) of the estimated VaR (left) and ES (right), as a function of the amount of computing time for different approaches, for the **GAS(1,1)-t** model, for different horizons. The horizontal line corresponds to a precision of 1 digit ($1.96NSE \leq 0.05$). A missing line for the MitISEM-based candidate corresponds to a situation when it was not possible to construct such a candidate.



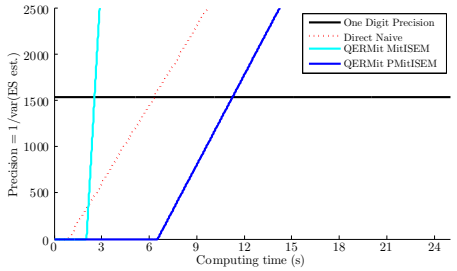
(a) VaR, $H = 10$



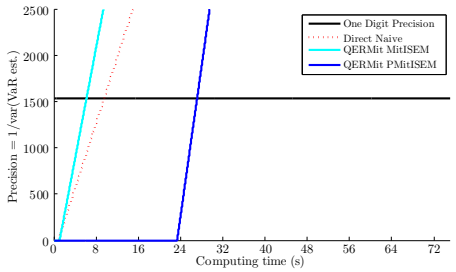
(b) ES, $H = 10$



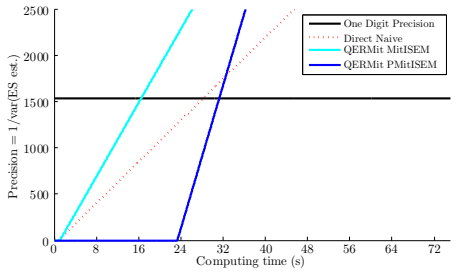
(c) VaR, $H = 20$



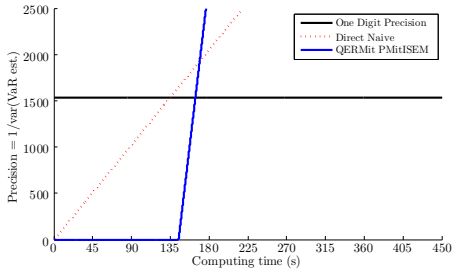
(d) ES, $H = 20$



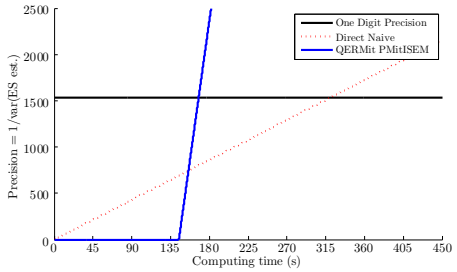
(e) VaR, $H = 40$



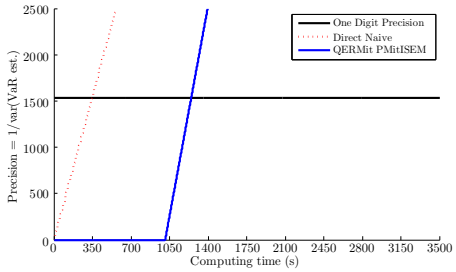
(f) ES, $H = 40$



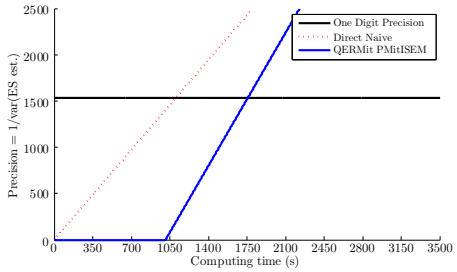
(g) VaR, $H = 100$



(h) ES, $H = 100$

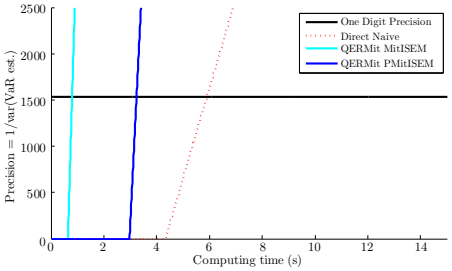


(i) VaR, $H = 250$

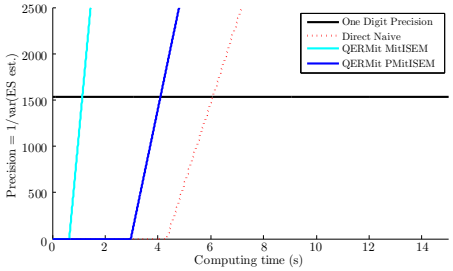


(j) ES, $H = 250$

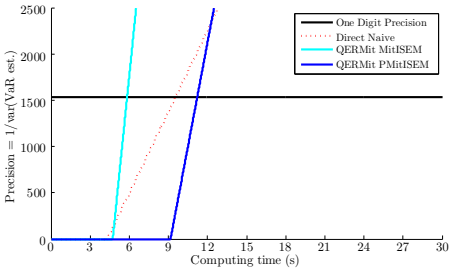
Figure 2.D.3: Precision ($1/var$) of the estimated VaR (left) and ES (right), as a function of the amount of computing time for different approaches, for the frequentist **GARCH(1,1)-t** model, for different horizons. The horizontal line corresponds to a precision of 1 digit ($1.96NSE \leq 0.05$). A missing line for the MitISEM-based candidate corresponds to a situation when it was not possible to construct such a candidate.



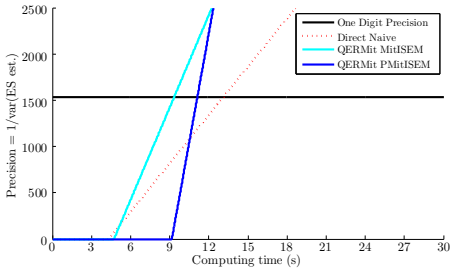
(a) VaR, $H = 10$



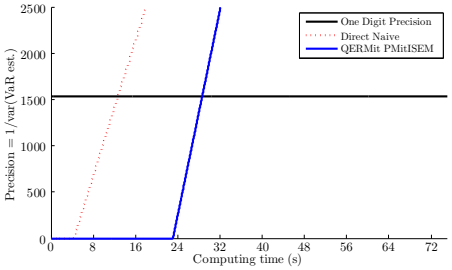
(b) ES, $H = 10$



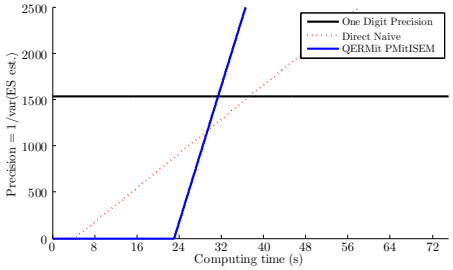
(c) VaR, $H = 20$



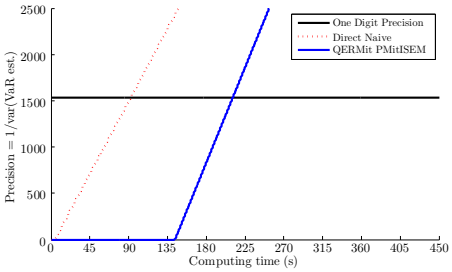
(d) ES, $H = 20$



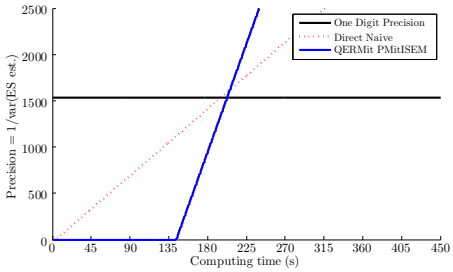
(e) VaR, $H = 40$



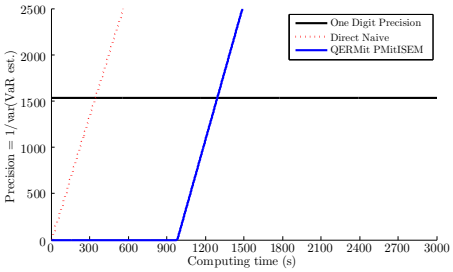
(f) ES, $H = 40$



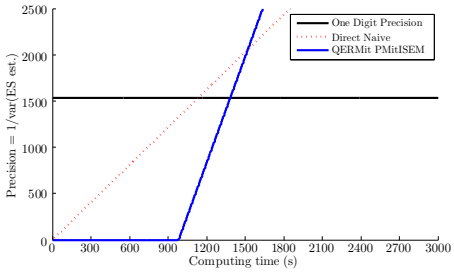
(g) VaR, $H = 100$



(h) ES, $H = 10$



(i) VaR, $H = 25s0$



(j) ES, $H = 250$

Figure 2.D.4: Precision ($1/var$) of the estimated VaR (left) and ES (right), as a function of the amount of computing time for different approaches, for the frequentist **GAS(1,1)-*t*** model, for different horizons. The horizontal line corresponds to a precision of 1 digit ($1.96NSE \leq 0.05$). A missing line for the MitISEM-based candidate corresponds to a situation when it was not possible to construct such a candidate.

Chapter 3

Partially Censored Posterior for Robust and Efficient Risk Evaluation

The issue of accurate estimation of the left tail of the predictive distribution of returns is crucial from the risk management perspective and is thus commonly investigated by both academics and practitioners. One of the main reasons for its importance is that it is used to obtain measures of downside risk for investments such as Value-at-Risk (VaR) and Expected Shortfall (ES), see McNeil and Frey (2000) and McNeil et al. (2015). The task of tail prediction is a special case of density forecasting where the focus is on a specific subset of the domain of the predictive distribution. Density forecasting in general has been rapidly growing in econometrics, finance and macroeconomics due to increased understanding of the limited informativeness of point forecasts, see Diks et al. (2011). In contrast to these, density forecasts provide a full insight into the forecast uncertainty. For a survey of the evolution of density forecasting in economics, see Aastveit et al. (2018b).

A natural framework, therefore, for analysing density forecasts is the Bayesian framework, as it treats all unobserved quantities as parameters to be estimated; see e.g. Geweke and Amisano (2010a) for a comparison and evaluation of Bayesian predictive distributions. This includes the predictions for the observation process. Importantly, the Bayesian approach incorporates the parameter uncertainty into analysis and facilitates dealing with model uncertainty, usually via Bayesian model averaging. However, the issue of Bayesian model misspecification still seems to be an open question.[1] A

[1] At the time of writing there is an active, ongoing debate in the Bayesian community about the issue

formal approach to this problem is provided by Kleijn and van der Vaart (2006), who show (under stringent conditions) that given an incorrectly specified model, the posterior concentrates *'close'* to the points in the support of the prior that minimise the Kullback-Leibler divergence with respect to the true data generating process (DGP). This result can be seen as the Bayesian counterpart of the MLE being consistent for the pseudo-true values in frequentist statistics. Nevertheless, differently than the asymptotic distribution of the MLE, the estimated posterior variance is incorrect in case of misspecification (Kleijn and van der Vaart, 2006). Müller (2013) shows that one can rescale the posterior so that credible sets have the correct coverage. As a practical solution to the problem, Geweke and Amisano (2012) apply the so-called model pooling, which relaxes the key assumption behind model averaging that the true model is in the set of models under consideration.

In the context of tail forecasting, the crucial question is: *what if "close" is not close enough*? From the perspective of accurate tail prediction obtaining estimates being just *"close"* to their real values is likely to lead to incorrect risk measures and hence to poor managerial decisions in cases where the misspecification is severe. To improve inference on a particular region of the predictive density, Gatarek et al. (2014) introduce the Censored Posterior (CP) for estimation and the censored predictive likelihood for model combination using Model Averaging. A concept underlying their approach is the censored likelihood scoring function of Diks et al. (2011), an adaptation (with specific focus on the left tail) of the popular logarithmic scoring rule, see Hall and Mitchell (2007) and Amisano and Giacomini (2007). Diks et al. (2011) use the censored likelihood scoring function only for comparing density forecasts in tails, not for estimation. The censoring means that observations outside the region of interest are censored: for those observations only the probability of being outside the region of interest matters. However, as we discuss in the later part of this chapter, for densely parametrised models applied in practice the Censored Posterior approach is likely to lose too much information.

To overcome these shortcomings the first main contribution of this chapter is the novel concept of the Partially Censored Posterior (PCP), where the set of model parameters is partitioned into two subsets: the first, for which we consider the standard marginal posterior, and the second, for which we consider a conditional censored posterior. In the second subset we choose parameters that are expected to especially benefit from censoring (due to their particular relationship with the tail of the predictive

of Bayesian model misspecification. Interestingly, it seems that there is no common ground on it (yet)! See Robert (2017) and Cross Validated (2017).

distribution). This approach leads to more precise parameter estimation than a fully censored posterior for all parameters, and has more focus on the region of interest than the standard Bayesian approach (that is, with no censoring).

The second main contribution is that we introduce two novel simulation methods. The first method is a Markov chain Monte Carlo (MCMC) method to simulate model parameters from the Partially Censored Posterior. Here we extend the *Mixture of t by Importance Sampling weighted Expectation-Maximization* (MitISEM) algorithm of Hoogerheide et al. (2012) to propose the *Conditional MitISEM* approach, where we approximate the joint censored posterior with a mixture of Student's t distributions and use the resulting conditional mixture of Student's t distributions as a candidate distribution for the conditional censored posterior. The high quality of the (conditional) candidate distributions leads to a computationally efficient MCMC method. The second method is an importance sampling method that is introduced to further decrease the numerical standard errors of the VaR and ES estimators. Here we adapt the *Quick Evaluation of Risk using Mixture of t approximations* (QERMit) algorithm of Hoogerheide and van Dijk (2010) to propose the *PCP-QERMit* method, where an adaptation is required since we do not have a closed-form formula for the partially censored posterior density kernel.

The third main contribution is that we consider the effect of using a time-varying boundary of the region of interest. To the best of our knowledge, the literature on the censored likelihood scoring rule, the censored likelihood and the censored posterior has been limited to a time-constant threshold defining the left tail. However, a constant threshold might be suboptimal when we focus on the left tail of the conditional distribution (given past observations). Even if the interest is in the unconditional left tail, then the time-varying threshold may be still more advantageous than the time-constant one. This is simply because the time-varying threshold allows us to obtain more information about the left tail of the distribution of the standardized innovations compared to the time-constant one.

The outline of this chapter is as follows. In Section 3.1 we consider the risk measure concepts, discuss the censored posterior and present a simple toy example to illustrate potential benefits and disadvantages of the censored posterior. Moreover, we introduce our novel concept of the Partially Censored Posterior and the novel simulation methods of Conditional MitISEM and PCP-QERMit. As an other extension of the existing literature on censored likelihood based methods, in Section 3.3 we introduce a time-varying threshold for censoring. In Section 3.4 we provide an empirical application using a GARCH model with Student's t innovations for daily IBM logreturns. Section

3.5 concludes.

3.1 Censored likelihood and censored posterior

Let $\{y_t\}_{t\in\mathbb{Z}}$ be a time series of daily logreturns on a financial asset price, with $y_{1:T} = \{y_1, \ldots, y_T\}$ denoting the (in-sample) observed data. We denote $\mathbf{y}_{s:r} = \{y_s, y_{s+1}, \ldots, y_{r-1}, y_r\}$ for $s \leq r$. We assume that $\{y_t\}_{t\in\mathbb{Z}}$ is subject to a dynamic stationary process parametrised by $\boldsymbol{\theta}$, on which we put a prior $p(\boldsymbol{\theta})$. We are interested in the conditional predictive density of $\mathbf{y}_{T+1:T+H}$, given the observed series $\mathbf{y}_{1:T}$. In particular, we are interested in the standard risk measure given by the $100(1-\alpha)\%$ VaR (in the sense of McNeil and Frey, 2000), the $100(1-\alpha)\%$ quantile of the predictive distribution of $\sum_{t=T+1}^{T+H} y_t$ given $y_{1:T}$.

$$100(1-\alpha)\%\,\mathrm{VaR} = \sup\left\{x \in \mathbb{R} : p(x|\mathbf{y}_{1:T}) \leq \alpha\right\}.$$

We also consider the ES as an alternative risk measure, due to its advantageous properties compared to the VaR, mainly sub-additivity (which makes ES a coherent risk measure in the sense of Artzner et al., 1999):

$$100(1-\alpha)\%\,\mathrm{ES} = \mathbb{E}\left[\sum_{t=T+1}^{T+H} y_t \,\middle|\, \sum_{t=T+1}^{T+H} y_t < 100(1-\alpha)\%\,\mathrm{VaR}\right].$$

The regular (uncensored) likelihood is given by the standard formula

$$p(\mathbf{y}_{1:T}|\boldsymbol{\theta}) = \prod_{t=1}^{T} p(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \boldsymbol{\theta})$$

and the posterior predictive density is

$$p(\mathbf{y}_{T+1:T+H}|\mathbf{y}_{1:T}) = \int p(\mathbf{y}_{T+1:T+H}|\mathbf{y}_{1:T}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|y_{1:T}) d\boldsymbol{\theta}.$$

We recall the details of Bayesian forecasting over an out-of-sample period of length H in Appendix 3.A. Given the data $y_{1:T}$ and a set of parameter draws $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^{M}$ from the posterior, the posterior predictive density can be estimated as:

$$p(\mathbf{y}_{T+1:T+H}|y_{1:T}) \approx \frac{1}{M}\sum_{i=1}^{M} p(\mathbf{y}_{T+1:T+H}|\mathbf{y}_{1:T}, \boldsymbol{\theta}^{(i)}). \tag{3.1.1}$$

As mentioned above, we are interested in a particular region of the predictive distribution, i.e. the left tail. For generality let us denote the region of interest by $A = \{A_1, \ldots, A_T\}$, where $A_t = \{y_t | y_t < C_t\}$ with threshold C_t potentially time-varying. For assessing the performance of forecast methods, i.e. comparing accuracy of density forecasts for such a region, Diks et al. (2011) introduce the censored likelihood scoring (CLS) function, which Gatarek et al. (2014) employ to define the censored likelihood (CL), where the CL is obtained by taking the exponential transformation of the CSL. The CL is given by

$$p^{cl}(\mathbf{y}_{1:T}|\boldsymbol{\theta}) = \prod_{t=1}^{T} p^{cl}(y_t|\mathbf{y}_{1:t-1}, \boldsymbol{\theta}), \tag{3.1.2}$$

where $p^{cl}(y_t|\mathbf{y}_{1:t-1}, \boldsymbol{\theta})$ is the conditional density of the mixed continuous-discrete distribution for the censored variable $\tilde{y}_t$

$$\tilde{y}_t = \begin{cases} y_t, & \text{if } y_t \in A_t, \\ R_t, & \text{if } y_t \in A_t^C. \end{cases} \tag{3.1.3}$$

Definition (3.1.3) means that the censored variable $\tilde{y}_t$ is equal to the original one in the region of interest, while everywhere outside it it is equal to the value $R_t \in A_t^C$. In consequence, the distribution of $\tilde{y}_t$ is mixed: continuous (in A_t) and discrete (in R_t). We have

$$\begin{aligned} p^{cl}(y_t|\mathbf{y}_{1:t-1}, \boldsymbol{\theta}) &= [p(y_t|\mathbf{y}_{1:t-1}, \boldsymbol{\theta})]^{I\{y_t \in A_t\}} \times \left[\mathbb{P}(y_t \in A_t^C|y_{1:t-1}, \boldsymbol{\theta})\right]^{I\{y_t \in A_t^C\}} \\ &= [p(y_t|\mathbf{y}_{1:t-1}, \boldsymbol{\theta})]^{I\{y_t \in A_t\}} \times \left[\int_{A_t^C} p(x|\mathbf{y}_{1:t-1}, \boldsymbol{\theta}) dx\right]^{I\{y_t \in A_t^C\}}. \end{aligned} \tag{3.1.4}$$

Differently than with a likelihood of a *censored dataset* where all $y_t \in A_t^C$ are censored and their exact values are completely ignored, with the censored likelihood the exact value of $y_t \in A_t^C$ still plays a role in conditioning in subsequent periods, in the sense that we condition on the *uncensored* past observations $y_{t-1}, y_{t-2}, \ldots$. Only in the case of i.i.d. observations when $p(y_t|\mathbf{y}_{1:t-1}, \boldsymbol{\theta}) = p(y_t|\boldsymbol{\theta})$ both approaches would be equivalent. We do this for two reasons. First, the purpose is to improve the left-tail prediction based on the actually observed past observations. By censoring the past observations $y_{t-1}, y_{t-2}, \ldots$ we would lose valuable information. Second, it would typically be much more difficult to compute the likelihood for censored data (where one would also condition on censored past observations). Therefore, the (Partially)

Censored Posterior is a *quasi*-Bayesian concept.

Gatarek et al. (2014) use the CL to define the censored posterior (CP) density as

$$p^{cp}(\boldsymbol{\theta}|\mathbf{y}_{1:T}) \propto p^{cl}(\mathbf{y}_{1:T}|\boldsymbol{\theta})p(\boldsymbol{\theta}), \tag{3.1.5}$$

where $p(\boldsymbol{\theta})$ is the prior density kernel on the model parameters. Typically, the censored posterior density $p^{cp}(\boldsymbol{\theta}|\mathbf{y}_{1:T})$ is a proper density in the same cases (i.e., under the same choices of the prior $p(\boldsymbol{\theta})$) where the regular posterior $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$ is a proper density (i.e., with finite integral $\int p^{cl}(\mathbf{y}_{1:T}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} < \infty$), as long as there are enough observations $y_t \in A_t$ that are not censored.

3.1.1 Scoring rules in density forecasting

Below we briefly recall the basic concepts related to scoring rules used for comparing accuracy of density forecasts. A scoring rule S is a loss function depending on the forecast density and the data, typically used to perform Diebold and Mariano (1995, DM) tests of equal predictive accuracy. The DM test is based on the average differential $\bar{d}_{T,H}$ of two log scores for a sequence of H one-step-ahead forecasts

$$\begin{aligned} \bar{d}_{T,H} &= \frac{1}{H}\sum_{h=1}^{H} d_h, \\ d_h &= S(p(y_{t+h}|\mathbf{y}_{1:T}, \mathcal{M}_1)) - S(p(y_{t+h}|\mathbf{y}_{1:T}, \mathcal{M}_2)). \end{aligned} \tag{3.1.6}$$

Notice that in (3.1.6) $p(y_{t+h}|\mathbf{y}_{1:T}, \mathcal{M}_i)$ is not a predictive distribution but its evaluation on the realised value of y_{t+h} (as opposed to y_h^*), where the predictive distribution originates from a method $\mathcal{M}_i$[2] The test statistic has the form

$$t_{T,H} = \frac{\bar{d}_{T,H}}{\sqrt{\hat{\sigma}^2_{T,H}/H}}, \tag{3.1.7}$$

where $\hat{\sigma}^2_{T,H}$ is a heteroskedasticity and autocorrelation-consistent variance estimator of the variance $\sigma^2_{T,H}$.

To begin with, we discuss presumably the most popular scoring rule, the log score, see Hall and Mitchell (2007) and Amisano and Giacomini (2007). Next, we present the censored likelihood scoring rule of Diks et al. (2011), which is the basis for the family

[2]In this chapter by a "method" we understand an an estimation method, but we note that often it is understood as an econometric model.

of censored posterior methods and which we use in our empirical study in Section 3.4 to compare the forecasting performance of competing estimation methods.

The popular log score rule is based on the Kullback-Leibler Information Criterion (KLIC), which is an information theoretic goodness-of-fit measure. Formally, for the density forecast $p(y_{t+h}|\mathbf{y}_{1:T}, \mathcal{M})$ obtained by a method $\mathcal{M}$, the KLIC defined as

$$\begin{aligned}\text{KLIC}\left[p(y_{t+h}|\mathbf{y}_{1:T}, \mathcal{M})\right] &= \mathbb{E}_t\left[\log \tilde{p}(y_{t+h}) - \log p(y_{t+h}|\mathbf{y}_{1:T}, \mathcal{M})\right] \\ &= \int \tilde{p}(y_{t+h}) \log\left(\frac{\tilde{p}(y_{t+h})}{p(y_{t+h}|\mathbf{y}_{1:T}, \mathcal{M})}\right) dy_{t+h},\end{aligned}$$

where $\tilde{p}_{t+h}$ denotes the true conditional density. Diks et al. (2011) show that for two competing density forecasts from methods $\mathcal{M}_1$ and $\mathcal{M}_2$, their relative KLIC values correspond exactly to the difference of their logarithmic scoring rules, with S in (3.1.6) specified as

$$S(p(y_{t+h}|\mathbf{y}_{1:T}, \mathcal{M})) = \log p(y_{t+h}|\mathbf{y}_{1:T}, \mathcal{M}).$$

The logarithmic scoring rule assigns a high score to a density forecast if the observation y_{t+h} falls within a region with high predictive density $p(y_{t+h}|\mathbf{y}_{1:T}, \mathcal{M})$, and a low score if the observation y_{t+h} is unlikely given the constructed forecast.

In terms of tail forecasting, where the events are rare by definition, taking into account the total probability of the region of interest is crucial from forecasting perspective, as noted by Diks et al. (2011). Yet some region-focused score functions, such as the weighted logarithmic scoring rule of Amisano and Giacomini (2007), simply ignore the observations outside the region of interest. In consequence the information on how frequently the region of interests occurs is lost and the scoring rule is *improper*[3]. As a possible solution to this problem Diks et al. (2011) specify the censored likelihood score function as

$$\begin{aligned}S^{cls}(p(y_{T+h}|\mathbf{y}_{1:T}, \mathcal{M})) = &\mathbb{I}(y_{T+h} \in A_{T+h}) \log p(y_{T+h}|\mathbf{y}_{1:T}, \mathcal{M}) \\ &+ \mathbb{I}(y_{T+h} \in A^c_{T+h}) \log\left(\int_{A^c_{T+h}} p(s|\mathbf{y}_{1:T}, \mathcal{M}) ds\right).\end{aligned} \tag{3.1.8}$$

This scoring rule does not neglect observations falling outside the region of interest

[3]An improper scoring rule may favour (i.e. assign a higher average score) an incorrect density forecast over the true conditional density. This can happen if the incorrect density forecast simply puts more probability mass in the region of interest.

but only as far as their total mass is concerned; the shape of the predictive density over A^c_{T+h} is ignored. Given a set of parameter draws $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^M$ the first term on the right hand side of (3.1.8), i.e. the part of corresponding to the region of interest A_{T+h}, can be approximated using (3.1.1). Suppose that the region of interest is chosen to be the tail of the conditional predictive distribution $A_{T+h} = \{y_{T+h} : y_{T+h} \leq C_{T+h}\}$, for a chosen value of C_{T+h}. Then the second term on the right hand side of (3.1.8), i.e. the "normalising" part corresponding to the complement of the region of interest A^c_{T+h}, is given by

$$\begin{aligned}\int_{A^c_{T+h}} p(s|y_{1:T}, \mathcal{M})ds &= \int_{C_{T+h}}^{\infty} p(s|y_{1:T}, \mathcal{M})ds \\ &= 1 - \int_{-\infty}^{C_{T+h}} p(s|y_{1:T}, \mathcal{M})ds \\ &= P(C_{T+h}|y_{1:T}, \mathcal{M}),\end{aligned}$$

where $P(\cdot|y_{1:T}, \mathcal{M})$ is the predictive cumulative distribution function. For a set of parameter draws $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^M$ the latter can be approximated by

$$\begin{aligned}P(C_{T+h}|\mathbf{y}_{1:T}, \mathcal{M}) &\approx \frac{1}{M}\sum_{i=1}^{M} P(C_{T+h}|\mathbf{y}_{1:T}, \boldsymbol{\theta}^{(i)}, \mathcal{M}), \\ P(C_{T+h}|\mathbf{y}_{1:T}, \boldsymbol{\theta}^{(i)}, \mathcal{M}) &= \int_{-\infty}^{C_{T+h}} p(s|\mathbf{y}_{1:T}, \boldsymbol{\theta}^{(i)})p(\boldsymbol{\theta}^{(i)}|\mathbf{y}_{1:T})ds.\end{aligned}$$

3.1.2 Advantages and disadvantages of CP: toy application

To illustrate the advantages and disadvantages of estimation based on the censored posterior, we start with a toy simulation study in which we consider as the data generating process (DGP) for y_t an i.i.d. split normal $\mathcal{SN}(\mu, \sigma_1^2, \sigma_2^2)$ model. The split normal density, analysed by e.g. Geweke (1989) and De Roon and Karehnke (2016), is given by

$$p(y_t) = \begin{cases} \phi(y_t; \mu, \sigma_1^2), & y_t > \mu, \\ \phi(y_t; \mu, \sigma_2^2), & y_t \leq \mu, \end{cases}$$

where $\phi(x; m, s)$ denotes the Gaussian density with mean m and variance s evaluated at x. The mean of a random variable distributed according to $\mathcal{SN}(0, \sigma_1^2, \sigma_2^2)$, i.e. with a split at zero, is equal to $-\frac{\sigma_2 - \sigma_1}{\sqrt{2\pi}}$, which is non-zero for any asymmetric case. Hence, shifting of the split point accordingly to the chosen variances allows us to consider a

zero-mean random variable: $y_t \sim \mathcal{SN}(\mu, \sigma_1^2, \sigma_2^2)$ with $\mu := \frac{\sigma_2 - \sigma_1}{\sqrt{2\pi}}$ results in $\mathbb{E}[y_t] = 0$. Such a specification is then equivalent to $y_t = \mu + \varepsilon_t$ with $\varepsilon_t \sim \mathcal{SN}(0, \sigma_1^2, \sigma_2^2)$.

We consider two cases of the true parameters of the DGP: a symmetric case with $\sigma_1 = 1$ and $\sigma_2 = 1$; and an asymmetric case with $\sigma_1 = 1$ and $\sigma_2 = 2$. In that latter case we set $\mu = \frac{1}{\sqrt{2\pi}}$ to impose $\mathbb{E}[y_t] = 0$. For both cases we generate $T = 100$, $T = 1000$ and $T = 10000$ observations from the true model. We are interested in evaluating the 95% and 99% VaR, i.e. in the estimation of the 5% and 1% quantiles of the distribution of y_t. For the symmetric case the true values for these quantities are -1.6449 and -2.3263, while for the asymmetric case -2.8908 and -4.2538.

For each case we estimate an i.i.d. normal model with unknown mean μ and variance σ^2. We specify the usual non-informative prior $p(\mu, \sigma) \propto \frac{1}{\sigma}$ (for $\sigma > 0$). We perform an estimation based on the uncensored posterior and two specifications for the censored posterior. In each the threshold value C is constant over time, $A_t = \{y_t : y_t \leq C\}$, and we consider two different values for the threshold C: one equal to the 10% quantile of the generated sample and another one equal to zero, where in both cases all the uncensored observations stem from the left half of the distribution. All the simulations are carried out with $M = 10000$ posterior draws after a burn-in of 1000 using an independence chain Metropolis-Hastings (IC-MH) algorithm with target density kernel (3.1.5) the candidate density being a single Student's t distribution.

Tables 3.1.1a and 3.1.1b report simulation results for 100 Monte Carlo (MC) experiments for the symmetric and asymmetric case, respectively. Figure 3.1.1 presents kernel density estimates of the asymmetric case for a single simulation for $T = 1000$; we refer to Appendix 3.C.1 for the plots for the remaining cases. In the misspecified case the regular posterior provides incorrect estimates from the left tail perspective, because the estimated model aims to approximate the distribution over the whole domain. The CP provides parameter estimates with a much better location (regarding the left tail of the predictive distribution) by focusing on the relevant region. The cost of a better location is, however, a larger variance of the estimates since censoring leads to an analysis based on effectively a smaller sample. Obviously, the precision of the estimates from the CP depends on the degree of censoring: the more censoring, the less information, the lower the precision. In the symmetric case we can see that, as expected, the only cost of censoring is a higher variance, but the locations of the regular posterior and the CP are similar. We observe that for the larger datasets ($T = 1000$ and $T = 10000$) the VaR from the regular posterior is only slightly better (in the sense of a slightly smaller MSE) in the case of no misspecification (with a normal DGP), whereas in the case of misspecification (with a split normal DGP) the censored posterior leads to much

more accurate VaR estimates. However, the VaR is substantially better for the regular posterior than for the censored posterior in case of a small dataset ($T = 100$) where the loss in precision due to censoring has a severe effect. We introduce the Partially Censored Posterior (PCP) in the next subsection, exactly for the reason of limiting this harmful effect of loss of information due to censoring.

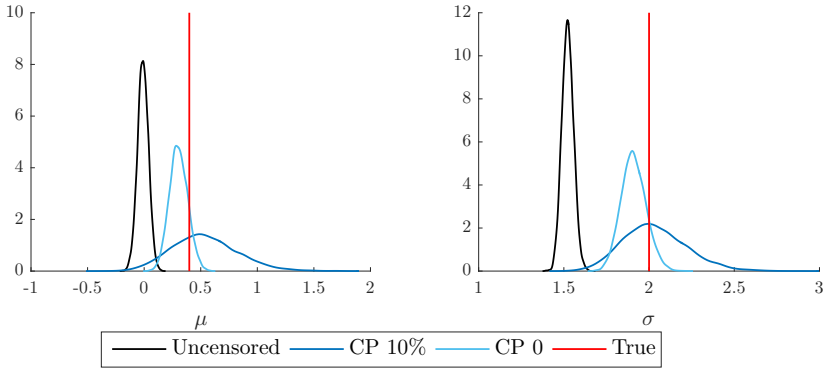


Figure 3.1.1: Estimation results in i.i.d. normal $N(\mu, \sigma^2)$ model for $T = 1000$ observations from DGP of i.i.d. split normal ($\sigma_1 = 1$, $\sigma_2 = 2$). Kernel density estimates of regular posterior and censored posterior (CP) with two different thresholds, at 0 (CP0) and at the 10% data percentile (CP10%) together with the true parameter values (corresponding to left tail).

3.2 Partially Censored Posterior

The previous subsection illustrated the advantages and disadvantages of the CP with respect to obtaining accurate evaluations of lower quantiles of the predictive posterior distribution: the CP has clearly a better location in case of misspecification, but this comes at the price of a lower precision of the estimates. Moreover, the estimated i.i.d. normal model had only 2 parameters whereas obviously most models have many more parameters, so that obtaining precise estimates becomes even harder. However, not all of the parameters are typically expected to particularly relate to the region of interest of the predictive distribution. For this reason we propose the *Partially Censored Posterior*, where only a selected subset of parameters is estimated with the conditional CP, while for the remaining parameters we consider the regular posterior.

3.2.1 Definition and MCMC algorithm *Conditional MitISEM*

Below we formally define the Partially Censored Posterior (PCP) and devise an MCMC algorithm to simulate from it. The PCP is a novel concept based on combining the standard posterior for the "common" parameters and the Censored Posterior of Gatarek

Value	True	Posterior	CP10%	CP0
		$T = 100$		
μ	0.3989	-0.0147	0.9452	0.5321
		(0.1658)	(1.0768)	(0.3171)
σ	2.0000	1.6414	2.5853	2.2724
		(0.1157)	(0.7684)	(0.2935)
99% VaR	-4.2538	-3.6551	-4.5968	-4.4697
		[0.5082]	[0.6438]	[**0.3506**]
95% VaR	-2.8908	-2.5675	-2.8886	-2.9773
		[0.1984]	[0.2612]	[**0.1402**]
		$T = 1000$		
μ	0.3989	-0.0103	0.5348	0.3043
		(0.0481)	(0.2895)	(0.0811)
σ	2.0000	1.5229	2.0338	1.9095
		(0.0343)	(0.1872)	(0.0732)
99% VaR	-4.2538	-3.5549	-4.2739	-4.2701
		[0.5063]	[0.0527]	[**0.0293**]
95% VaR	-2.8908	-2.5101	-2.8895	-2.8882
		[0.1540]	[0.0158]	[**0.0145**]
		$T = 10000$		
μ	0.3989	0.0334	0.4279	0.4290
		(0.0152)	(0.0901)	(0.0273)
σ	2.0000	1.5125	1.9825	1.9778
		(0.0106)	(0.0568)	(0.0250)
99% VaR	-4.2538	-3.5654	-4.2610	-4.2583
		[0.4787]	[0.0098]	[**0.0091**]
95% VaR	-2.8908	-2.5226	-2.8919	-2.8917
		[0.1369]	[0.0031]	[**0.0029**]

(b) Asymmetric (misspecified) case: $\sigma_2 = 2$.

Value	True	Posterior	CP10%	CP0
		$T = 100$		
μ	0.0000	0.0926	1.5715	0.1727
		(0.1014)	(1.5261)	(0.1529)
σ	1.0000	1.0226	1.8779	1.1289
		(0.0741)	(0.9051)	(0.1418)
99% VaR	-2.3263	-2.1245	-2.2322	-2.1519
		[**0.5763**]	[0.6812]	[0.6041]
95% VaR	-1.6449	-1.4899	-1.4668	-1.4951
		[**0.2922**]	[0.3123]	[0.2986]
		$T = 1000$		
μ	0.0000	0.0071	0.0196	0.0230
		(0.0304)	(0.1473)	(0.0387)
σ	1.0000	0.9604	0.9446	0.9725
		(0.0215)	(0.0921)	(0.0349)
99% VaR	-2.3263	-2.0998	-2.1020	-2.1074
		[**0.5464**]	[0.5500]	[0.5476]
95% VaR	-1.6449	-1.4816	-1.4823	-1.4858
		[**0.2725**]	[0.2734]	[0.2735]
		$T = 10000$		
μ	0.0000	0.0031	0.0300	-0.0049
		(0.0100)	(0.0433)	(0.0123)
σ	1.0000	0.9960	1.0053	0.9865
		(0.0071)	(0.0281)	(0.0111)
99% VaR	-2.3263	-2.0965	-2.0876	-2.0972
		[**0.5427**]	[0.5432]	[0.5428]
95%VaR	-1.6449	-1.4802	-1.4767	-1.4815
		[**0.2712**]	[0.2713]	[0.2713]

(a) Symmetric (correctly specified) case: $\sigma_2 = 1$.

Table 3.1.1: Estimation results in i.i.d. normal $\mathcal{N}(\mu, \sigma^2)$ model for data from DGP of i.i.d. normal $\mathcal{N}(\mu = 0, \sigma = 1)$ and i.i.d. split normal $\mathcal{SN}(\mu, \sigma_1 = 1, \sigma_2 = 2)$. Simulation results for the regular posterior and for the censored posterior with two different thresholds, at 0 (CP0) and at the 10% data percentile (CP10%). Standard deviations in parentheses. MSEs in brackets, with the best MSE in bold.

et al. (2014) for the parameters that particularly affect the properties of the region of interest. Consider a vector of model parameters $\boldsymbol{\theta}$ and suppose that some subset of parameters, call it $\boldsymbol{\theta}_2$, is particularly related to the (left) tail of the distribution so that it may benefit from censoring, while the other parameters, in the subset $\boldsymbol{\theta}_1$, would not benefit from censoring, or could even be adversely affected by censoring. In other words, we consider a partitioning $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$. How this partitioning is done depends on the model under consideration. We propose that a sensible way is to collect in $\boldsymbol{\theta}_2$ the parameters determining the shape of the conditional distribution of y_t (e.g., the degrees of freedom parameter of a Student's t distribution, the shape parameter of a generalized error distribution), but also parameters for the (unconditional) mean and variance. Next, we propose to collect in $\boldsymbol{\theta}_1$ the other parameters, such as the coefficients determining the dynamic behaviour of the conditional mean/variance in ARMA/GARCH models.

Definition and algorithm We define the PCP as

$$p^{pcp}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y}) = p(\boldsymbol{\theta}_1|\mathbf{y}) p^{cp}(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \mathbf{y}),$$

where $\mathbf{y}$ denotes the observed data, $p(\boldsymbol{\theta}_1|\mathbf{y})$ is the standard marginal posterior of $\boldsymbol{\theta}_1$ and $p^{cp}(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \mathbf{y})$ is the *conditional* censored posterior of $\boldsymbol{\theta}_2$ given $\boldsymbol{\theta}_1$. For a given value of $\boldsymbol{\theta}_1$, a kernel of the *conditional* censored posterior density of $\boldsymbol{\theta}_2$ given $\boldsymbol{\theta}_1$ is given by:

$$p^{cp}(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \mathbf{y}) = \frac{p^{cp}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y})}{p^{cp}(\boldsymbol{\theta}_1|\mathbf{y})} \propto p^{cp}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y}) \propto p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) p^{cl}(\mathbf{y}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2),$$

with prior density kernel $p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ and censored likelihood $p^{cl}(\mathbf{y}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ in (3.1.2). We propose the following MCMC procedure to simulate from the PCP, the *Conditional MitISEM* method.

1. Simulate $(\boldsymbol{\theta}_1^{(i)}, \boldsymbol{\theta}_2^{(i)})$, $i = 1, \ldots, M$, from posterior $p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y})$ using the IC-MH algorithm, using as a candidate density a mixture of Student's t densities obtained by applying the *Mixture of t by Importance Sampling weighted Expectation-Maximization* (MitISEM) algorithm of Hoogerheide et al. (2012) to the posterior density kernel $p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y})$.

2. Keep $\boldsymbol{\theta}_1^{(i)}$ and ignore $\boldsymbol{\theta}_2^{(i)}$, $i = 1, \ldots, M$.

3. For each $\boldsymbol{\theta}_1^{(i)}$ simulate $\boldsymbol{\theta}_2^{(i,j)}$, $j = 1, \ldots, N$, from the conditional censored posterior $p^{cp}(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1^{(i)}, \mathbf{y})$.

(a) Construct joint candidate density $q_{mit}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, a mixture of Student's t densities obtained by applying the MitISEM algorithm to the censored posterior density kernel $p^{cp}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y})$.

(b) Use conditional candidate density $q_{cmit}(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^{(i)})$, the mixture of Student's t densities implied by the joint candidate density $q_{mit}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, as a candidate density to simulate $\boldsymbol{\theta}_2^{(i,j)}$ from $p^{cp}(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1^{(i)}, \mathbf{y})$ in a run of the independence chain MH algorithm.

The use of MitISEM in step 3a implies that this step is efficiently performed with a relatively high acceptance rate in the IC-MH algorithm. To perform the conditional sampling in step 3b we use the fact that the conditional distribution of a joint mixture of Student's t distributions is itself a mixture of Student's t distributions and we provide its details in Appendix 3.B.

This implies that if we have obtained $q_{mit}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, a mixture of Student's t densities that approximates the joint censored posterior $p^{cp}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y})$, then we can use the M implied conditional mixtures of Student's t densities $q_{cmit}(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^{(i)})$, $(i = 1, \ldots, M)$, as candidate densities for $p^{cp}(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1^{(i)}, \mathbf{y})$ $(i = 1, \ldots, M)$. Hence, we only need one MitISEM approximation to obtain all the conditional candidate densities. In step 3b we do need a separate run of the IC-MH algorithm to simulate $\boldsymbol{\theta}_2^{(i,j)}$ for each given $\boldsymbol{\theta}_1^{(i)}$ $(i = 1, \ldots, M)$. However, given the typically high quality of the conditional MitISEM candidate density, a small burn-in will typically suffice, after which we can choose to use $N = 1$ draw $\boldsymbol{\theta}_2^{(i,j)}$. Note that step 3b can be performed in a parallel fashion. As an alternative, to further speed up the simulation method with only a small loss of precision, we can also choose to use $N \geq 2$ draws $\boldsymbol{\theta}_2^{(i,j)}$ $(j = 1, \ldots, N)$ from each run, for example $N = 10$, combined with a thinning approach for $\boldsymbol{\theta}_1^{(i)}$, where only every Nth draw of $\boldsymbol{\theta}_1^{(i)}$ is used.

3.2.2 Variance reduction with *PCP-QERMit*

Putting much effort in obtaining more accurate estimates of risk measures such as VaR and ES, using the specific left-tail focus of the PCP, might be wasteful if counteracted by large simulation noise affecting these estimates (i.e. high numerical standard errors). Hence, we aim to increase numerical efficiency of the proposed PCP method. For this purpose, we adapt the *Quick Evaluation of Risk using Mixture of t approximations* (QERMit) algorithm of Hoogerheide and van Dijk (2010) for efficient VaR and ES estimation.

QERMit is an importance sampling (IS) based method in which an increase in efficiency is obtained by oversampling "high-loss" scenarios and assigning them lower importance weights. The theoretical result of Geweke (1989) prescribes that the optimal importance density (in the sense of minimising the numerical standard error for a given number of draws) for Bayesian estimation of a probability of a given set (here, the left tail of the predictive distribution) should be composed of two equally weighted components, one for the high-loss scenarios (corresponding to the tail) and one for remaining realisations of returns. I.e. there is a 50%-50% division between "high-loss" draws and other draws. Such an approach allows for a substantial increase in efficiency compared to the so-called *direct approach* for VaR evaluation, in which predictions are obtained by simply sampling future innovations from the model and combining these with the posterior draws of model parameters to generate future paths of returns. One then simply computes the VaR estimate as the required percentile of the sorted (in ascending order) simulated returns. The QERMit method of Hoogerheide and van Dijk (2010) works for the regular (uncensored) Bayesian approach, i.e. based on the regular posterior and the regular predictive distribution. This method does require a closed-form formula for the target density, which is used as the numerator of the IS weights in the final step where the draws from the importance density are used to estimate the VaR. In case of the PCP we do not have a closed-form formula for the target density $p^{pcp}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y}) = p(\boldsymbol{\theta}_1|\mathbf{y})p^{cp}(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \mathbf{y})$, since we do not have closed-form formulas for the density kernels $p(\boldsymbol{\theta}_1|\mathbf{y})$ and $p^{cp}(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \mathbf{y})$.

New IS algorithm To overcome this problem, we propose a new IS-based method to reduce the variance of the H-step-ahead VaR estimator obtained with the PCP. Given the draws of $(\boldsymbol{\theta}_1^{(i)}, \boldsymbol{\theta}_2^{(i)})$, $i = 1, \ldots, M$, from the PCP, we aim to sample the future innovations in the model $\boldsymbol{\varepsilon}_{T+1:T+H}$ *conditionally* on $(\boldsymbol{\theta}_1^{(i)}, \boldsymbol{\theta}_2^{(i)})$ such that the resulting joint draws $(\boldsymbol{\theta}_1^{(i)}, \boldsymbol{\theta}_2^{(i)}, \boldsymbol{\varepsilon}_{T+1:T+H})$ will lead to "high losses". This relates to the idea of oversampling the negative scenarios underlying the QERMit approach of Hoogerheide and van Dijk (2010), however we do not require to evaluate the target density kernel of the PCP. The proposed *PCP-QERMit* algorithm proceeds as follows.

1. **Preliminary steps**

   (a) Obtain a set of draws from the PCP, $(\boldsymbol{\theta}_1^{(i)}, \boldsymbol{\theta}_2^{(i)})$, $i = 1, \ldots, M$, using the *Conditional MitISEM* algorithm of the previous subsection.

   (b) Simulate future innovations $\boldsymbol{\varepsilon}_{T+1:T+H}^{(i)}$ from their model distribution.

   (c) Calculate the corresponding predicted returns $\mathbf{y}_{T+1:T+H}^{(i)}$.

(d) Consider those joint draws $(\boldsymbol{\theta}_1^{(i)}, \boldsymbol{\theta}_2^{(i)}, \boldsymbol{\varepsilon}_{T+1:T+H}^{(i)})$ that have led to e.g. the 10% lowest returns $\sum_{t=T+1}^{T+H} y_t^{(i)}$ (the "high loss draws").

2. **High loss draws**

(a) Use the "high loss draws" from step 1d to approximate the joint PCP "high-loss" density of $\boldsymbol{\theta}$ and $\boldsymbol{\varepsilon}_{T+1:T+H}$ with a mixture of Student's t densities $q_{mit}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\varepsilon}_{T+1:T+H})$ by applying the MitISEM algorithm to the draws $(\boldsymbol{\theta}_1^{(i)}, \boldsymbol{\theta}_2^{(i)}, \boldsymbol{\varepsilon}_{T+1:T+H}^{(i)})$.

(b) Sample $\tilde{\boldsymbol{\varepsilon}}_{T+1:T+H}^{(i)} | \boldsymbol{\theta}_1^{(i)}, \boldsymbol{\theta}_2^{(i)}$, $i = 1, \ldots, M$, from its conditional importance density (aimed at high losses) $q_{cmit}(\boldsymbol{\varepsilon}_{T+1:T+H} | \boldsymbol{\theta}_1^{(i)}, \boldsymbol{\theta}_2^{(i)})$, the conditional mixture of Student's t distributions implied by $q_{mit}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\varepsilon}_{T+1:T+H})$ (see Appendix 3.B).

3. **IS estimation of the VaR (or ES)**

(a) Compute the importance weights of the draws $(\boldsymbol{\theta}_1^{(i)}, \boldsymbol{\theta}_2^{(i)}, \tilde{\boldsymbol{\varepsilon}}_{T+1:T+H}^{(i)})$, $i = 1, \ldots, M$, as

$$w^{(i)} = \frac{p(\tilde{\boldsymbol{\varepsilon}}_{T+1:T+H}^{(i)} | \boldsymbol{\theta}_1^{(i)}, \boldsymbol{\theta}_2^{(i)})}{q(\tilde{\boldsymbol{\varepsilon}}_{T+1:T+H}^{(i)} | \boldsymbol{\theta}_1^{(i)}, \boldsymbol{\theta}_2^{(i)})},$$

where the numerator $p(\boldsymbol{\varepsilon}_{T+1:T+H}^{(i)} | \boldsymbol{\theta}_1^{(i)}, \boldsymbol{\theta}_2^{(i)})$ is simply the density of the innovations in the model (and where the kernel of the partially censored posterior density $p^{pcp}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{y}) = p(\boldsymbol{\theta}_1 | y) p^{cp}(\boldsymbol{\theta}_2 | \boldsymbol{\theta}_1, \mathbf{y})$ drops out of the importance weight, as it appears in both numerator and denominator).

(b) Compute the future returns $\mathbf{y}_{T+1:T+H}^{(i)}$ corresponding to the joint draws $(\boldsymbol{\theta}_1^{(i)}, \boldsymbol{\theta}_2^{(i)}, \tilde{\boldsymbol{\varepsilon}}_{T+1:T+H}^{(i)})$, $i = 1, \ldots, M$, and the resulting total return over H periods $\sum_{t=T+1}^{T+H} y_t$.

(c) Estimate the $100(1-\alpha)$%VaR as the value C such that

$$\widehat{\mathbb{P}}\left(\sum_{t=T+1}^{T+H} y_t < C \right) = \alpha,$$

with

$$\widehat{\mathbb{P}}\left(\sum_{t=T+1}^{T+H} y_t < C \right) = \frac{1}{M} \sum_{i=1}^{M} w^{(i)} \mathbb{I}\left(\sum_{t=T+1}^{T+H} y_t^{(i)} < C \right),$$

where $\mathbb{I}(\cdot)$ denotes the indicator function.

For the ES the method continues in a similar fashion. Step 2b is crucial in the above algorithm, as it allows us to "guide" the future disturbances to the "high-loss" region without the necessity of evaluating the kernel of the partially censored posterior density $p^{pcp}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|y) = p(\boldsymbol{\theta}_1|y)p^{cp}(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, y)$. Note that we do not need to use the 50%-50% division between "high-loss" draws and other draws, which was the case in the regular QERMit method for Bayesian VaR/ES prediction, but we can fully focus on the high losses. Such a concentration of all the mass of the importance density in the "high-loss region" is valid since we do not use the *self-normalised* IS weights $w^{(i)}/\sum_{j=1}^{M} w^{(j)}$. Normalising of the IS weights is typically necessary in Bayesian IS estimation as often only the posterior kernel is available. Since we have the exact target and candidate densities of the innovations $\varepsilon_{T+1:T+H}$, we use the *unscaled* IS weights $w^{(i)}$ that only matter for "high-loss" draws with indicator $\mathbb{I}\left(\sum_{t=T+1}^{T+H} y_t^{(i)} < C\right) = 1$.

Illustration To illustrate the benefits of the PCP-QERMit method we consider a simple example involving the AR(1) model. Building upon the toy example from Subsection 3.1.2, we consider the true DGP of the form

$$y_t = \mu + \rho y_{t-1} + \varepsilon_t,$$

with split normally distributed innovations $\varepsilon_t \sim \mathcal{SN}(\delta, \sigma_1^2, \sigma_2^2)$ with $\delta = \frac{\sigma_2 - \sigma_1}{\sqrt{2\pi}}$ so that $E(\varepsilon_t) = 0$. We simulate $T = 1000$ observations from the model with $\mu = 0$, $\sigma_1 = 1$, $\sigma_2 = 2$ and $\rho = 0.8$.

We estimate the AR(1) model with normally distributed innovations $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$. The priors for μ and σ are the same as in the i.i.d. case, while for ρ we adopt a uniform prior over the stationarity region (i.e. $|\rho| < 1$).

We estimate the 1-step-ahead 99.5%, 99% and 95% VaR and ES (and compute the numerical standard error from 50 MC replications) using the PCP where $\boldsymbol{\theta}_1 = \{\rho\}$ stems from the regular marginal posterior, whereas $\boldsymbol{\theta}_2 = \{\mu, \sigma\}$ stems from the conditional censored posterior. Both the PCP direct approach (Conditional MitISEM) and the PCP-QERMit method make use of $10,000$ draws. (The PCP has a time-constant threshold C_t given by the 10% quantile of the in-sample data.) Table 3.2.1 shows the results, where the smaller numerical standard errors stress the usefulness of the PCP-QERMit method for obtaining more accurate estimates of both VaR and ES.

Risk measure	PCP direct approach	PCP-QERMit
99.5% VaR	-4.3557	-4.3379
	[0.1050]	[0.0500]
99.5% ES	-4.9877	-4.9786
	[0.1328]	[0.0830]
99% VaR	-3.8461	-3.8308
	[0.0813]	[0.0340]
99% ES	-4.5311	-4.5183
	[0.1003]	[0.0587]
95% VaR	-2.4682	-2.4675
	[0.0429]	[0.0100]
95% ES	-3.3130	-3.3055
	[0.0524]	[0.0228]

Table 3.2.1: Estimated 1-step-ahead 99.5%, 99% and 95% VaR and ES (and numerical standard error from 50 MC replications within brackets) for estimated AR(1) model with normally distributed innovations, for $T = 1000$ observations from DGP of AR(1) model with split normally distributed innovations ($\sigma_1 = 1, \sigma_2 = 2$). The PCP direct approach (Conditional MitISEM) and PCP-QERMit method make use of 10000 draws. (The PCP has a time-constant threshold C_t given by the 10% quantile of the in-sample data.)

3.2.3 Simulation study: AR(1) model

Below, we compare the quality of the left-tail density forecasts from the PCP with the regular posterior and the full CP. We consider the same estimated model and the same DGP as in the previous subsection: an estimated AR(1) model with normally distributed innovations for data from an AR(1) model with split normally distributed innovations.

We keep $\mu = 0$, $\rho = 0.8$ and $\sigma_1 = 1$ in the DGP. We do vary the level of misspecification by considering the correctly specified case of $\sigma_2 = 1$ and the misspecified cases of $\sigma_2 = 1.5$ and $\sigma_2 = 2$. Further, we analyse the effect of the sample size T by considering estimation windows of size $T = 100$, 200, 500 and 1000.

For each DGP we consider 1000 out-of-sample observations for 20 simulated datasets, where for each observation we compute the (one-step-ahead) censored likelihood score function (3.1.8) of Diks et al. (2011) (with time-constant threshold $C_t = C$ given by the 5% quantile of the returns). For each simulated dataset we compute the Diebold-Mariano test statistic (with Newey-West standard error; see Diebold and Mariano, 1995), where the loss differential is the difference in the censored likelihood score function. We use the average of the 20 Diebold-Mariano test statistics to test the null hypothesis of equal left-tail density prediction, where the critical values in a two-sided test at 5% significance are simply given by $\pm\frac{1.96}{\sqrt{20}} \approx \pm 0.44$ (as the 20 simulated datasets are independent, and the test statistics have approximately the standard normal dis-

tribution under the null).

Tables 3.2.2a–3.2.2c show the results. We observe the following findings. First, as expected, in the case without misspecification ($\sigma_2 = 1$), the regular posterior performs better than the PCP or CP. In this case it is obviously optimal to use all (uncensored) observations. Moreover, in this case, the PCP performs better than the CP, as "the less censoring, the better". Second, in the cases of misspecification and a large estimation window ($T = 500$ or $T = 1000$) the PCP and CP outperform the regular posterior. The more severe the misspecification, the smaller the sample size T for which censoring becomes beneficial. Third, in the case of misspecification and a small estimation window ($T = 100$ or $T = 200$) the regular posterior outperforms the CP and the PCP, caused by the loss of information due to censoring. Fourth, the PCP is never significantly outperformed by the CP. In the case of misspecification and a large estimation window, we do not reject the equality of their performance. In the cases of no misspecification and/or a small estimation window the PCP significantly outperforms the CP.

3.3 Time-varying threshold

Notice that the region of interest A_t used to define the censored variable in (3.1.3) is potentially time-varying. However, to the best of our knowledge, the literature on the censored likelihood scoring function, the censored likelihood and the censored posterior has been limited to a time-constant threshold. Gatarek et al. (2014) set the "censoring boundary" to the 20% or 30% percentile of the estimation window, leaving the topic of a time-varying threshold for further research. Opschoor et al. (2016) focus on the 15% percentile of a two-piece Normal distribution or a certain percentile (15% or 25%) of the empirical distribution of the data. Diks et al. (2011) investigate the impact of a time-varying threshold, which, however, is understood slightly differently. These authors evaluate the forecasting methods using a rolling window scheme and set the time-varying constant equal to the empirical quantile of the observations in the relevant estimation window. Obviously, a time-constant threshold implied by a certain empirical percentile differs between different data windows.

However, a constant threshold might be suboptimal when we focus on the left tail of the conditional distribution (given past observations). Even if the interest is in the unconditional left tail, so only in the most negative returns, then the time-varying threshold might be still more advantageous than the time-constant one. This is simply because the time-varying threshold provides more information about the left tail of the distribution of the standardized innovations compared to the time-constant one.

T	$\sigma_2 = 1$	$\sigma_2 = 1.5$	$\sigma_2 = 2$
100	7.379***	5.868***	2.137***
200	4.315***	1.097***	**-0.872*** **
500	5.261***	-0.367	**-1.221*** **
1000	2.026***	**-0.959*** **	**-1.648*** **

(a) Posterior vs PCP.

T	$\sigma_2 = 1$	$\sigma_2 = 1.5$	$\sigma_2 = 2$
100	4.471***	3.957***	1.894***
200	2.987***	1.458***	-0.739***
500	1.923***	0.065	-1.370***
1000	1.084***	-0.778***	-1.810***

(b) Posterior vs CP.

T	$\sigma_2 = 1$	$\sigma_2 = 1.5$	$\sigma_2 = 2$
100	**-1.561*** **	**-2.157*** **	**-2.312*** **
200	**-2.041*** **	**-0.924*** **	-0.419*
500	**-1.410*** **	-0.135	0.320
1000	**-0.857*** **	0.031	-0.157

(c) CP vs PCP.

Table 3.2.2: Left-tail density forecast comparison based on the censored likelihood score function (3.1.8) (with time-constant threshold $C_t = C$ given by the 5% quantile of the returns) between three estimation methods, the regular posterior, the full CP and the PCP. The tables show the average of 20 Diebold-Mariano test statistics (with Newey-West standard errors) for 20 simulated data sets. The loss differential (computed for 1000 out-of-sample observations for each simulated dataset) is the difference in the censored likelihood score function (3.1.8). Positive values indicate superior left-tail forecast performance of the CP; negative values indicate superior left-tail forecast performance of the PCP. The significance (in a two-sided test) is indicated by * for $p \leq 0.1$, ** for $p \leq 0.05$ and *** for $p \leq 0.01$. Bold numbers indicate a significantly better performance of our proposed PCP approach (at 5% significance level).

Therefore, we consider the time-varying threshold C_t given by a certain percentile of the estimated conditional distribution of y_t (given the past) that is implied by the Maximum Likelihood Estimate (MLE) $\hat{\boldsymbol{\theta}}_{ML}$. Note that the threshold C_t must be equal for all draws $\boldsymbol{\theta}^{(i)}$ $(i = 1, \ldots, M)$ from the (partially) censored posterior, as the threshold C_t affects the (partially) censored posterior. Making C_t depend on draws $\boldsymbol{\theta}^{(i)}$ $(i = 1, \ldots, M)$ from the (partially) censored posterior would lead to a circular reasoning. Hence, the MLE $\hat{\boldsymbol{\theta}}_{ML}$ provides a usable solution. As an alternative, one could use the regular posterior mean of $\boldsymbol{\theta}$.

The above discussion relates to *estimation* based on a censored posterior. However, note that the choice of a threshold C_{T+1} can also be important *for the assessment of the quality of the left-tail prediction*. Indeed, (3.1.8) can be computed with time-varying C_{T+1}. In our empirical study in Section 3.4 we consider, next to the standard

time-constant threshold for the CSL rule (the 10% percentile of the in-sample data), a time-varying threshold given by the 10% percentile of the MLE-implied conditional distribution.

Simulation study: GARCH(1,1) and AGARCH(1,1) models Our aim in the remaining part of this section is threefold. First, we investigate the role of the exact model specification on the usefulness of the PCP. There exists an immense amount of models of volatility, including an extensive family of GARCH-type models, see Bollerslev (2008) but also recently introduced Generalized Autoregressive Score models (GAS) of Creal et al. (2013). Not all model specifications may be expected to equally benefit from censoring. In other words, we consider the robustness of our results for different model specifications, where for the practical usefulness of the PCP its use should not only be beneficial in certain "convenient" models. That is, preferably one does not need to particularly adapt the model specification in order to make the PCP useful.

Second, we check what gains can be obtained from censoring with small and large estimation windows. Intuitively, partial censoring should be particularly useful for smaller datasets as then there is not enough information over the region of interest to accurately estimate the fully censored parameter vector.

Third, we analyse how extreme the tails need to be for censoring-based methods to be beneficial. For instance, the Basel requirement involves the 99% VaR, so the 1% percentile. However, for more conservative risk managers the 99.5% VaR may be of interest, while more "risk-seeking" approaches may consider the 95% VaR. One may expect that the focus on the left tail during the estimation is particularly useful when one is interested in the deep tail[4].

For illustration, consider the DGP of the following GARCH(1,1) model with split normal errors:

$$\begin{aligned} y_t &= \mu + \sqrt{(\kappa^{-1}h_t)}\varepsilon_t, \\ h_t &= \omega(1-\alpha-\beta) + \alpha y_{t-1}^2 + \beta h_{t-1}, \\ \varepsilon_t &\sim \mathcal{SN}(\delta, \sigma_1^2, \sigma_2^2), \end{aligned}$$

where $\delta = \frac{\sigma_2-\sigma_1}{\sqrt{2\pi}}$ so that $E(\varepsilon_t) = 0$, and where $\kappa = \frac{1}{2}\left((\sigma_1^2+\sigma_2^2) - \frac{(\sigma_2-\sigma_1)^2}{\pi}\right)$ is the

[4]On the other hand, if one would be interested in the median of the predictive distribution, then one should obviously not focus on the tail during estimation.

variance of ε_t[5]. We set $\mu = 0$, $\omega = 1$, $\alpha = 0.1$, $\beta = 0.8$, and for ε_t we again choose $\sigma_1 = 1$ and $\sigma_2 = 2$. Then $\kappa \approx 2.34$, which effectively implies the standard deviation of the right and the left tail of around 0.65 and 1.3, respectively. In a single experiment we simulate $T = 1000$ and $T = 2500$ observations from the DGP, and we carry out 50 MC repetitions of such an experiment.

To answer the question about the role of a "convenient" model specification, we estimate two (misspecified) models: the standard GARCH(1,1) model with normally distributed innovations and the Asymmetric GARCH(1,1) model (AGARCH(1,1)) of Engle and Ng (1993). The latter is characterised by two mean parameters: an actual mean μ_1 and a parameter μ_2 for defining the squared "demeaned" lagged return $(y_{t-1} - \mu_2)^2$ in the GARCH equation:

$$\begin{aligned} y_t &= \mu_1 + \sqrt{h_t}\varepsilon_t, \\ h_t &= \omega(1 - \alpha - \beta) + \alpha(y_{t-1} - \mu_2)^2 + \beta h_{t-1}, \\ \varepsilon_t &\sim \mathcal{N}(0, 1). \end{aligned}$$

The GARCH(1,1) model results from the AGARCH(1,1) model by setting $\mu = \mu_1 = \mu_2$. In the PCP for the GARCH(1,1) model we choose the tail related parameters $\boldsymbol{\theta}_2 = \{\mu, \omega\}$ and other (dynamics related) parameters $\boldsymbol{\theta}_1 = \{\alpha, \beta\}$, whereas for the AGARCH(1,1) model we choose $\boldsymbol{\theta}_2 = \{\mu_1, \omega\}$ and $\boldsymbol{\theta}_1 = \{\mu_2, \alpha, \beta\}$. Note that the AGARCH(1,1) model may seem a "convenient" counterpart of the GARCH(1,1) model for the PCP, as it separates the "direct" effect of μ on the conditional distribution of y_t (as the mean) and the effect of μ on the dynamics of the GARCH process in two different parameters μ_1 and μ_2.

For both models we take flat priors over the standard domains to impose stationarity and positivity of the volatility process, i.e. $\omega > 0$, $\alpha \in (0, 1)$, $\beta \in (0, 1)$, $\alpha + \beta < 1$. We analyse 99.5%, 99% and 95% one-step-ahead VaR and ES forecasting over a horizon of 100 days and we carry out 50 independent MC replications. For the estimation we consider, next to the regular posterior, two types of thresholds (for both the CP and PCP): one time-constant threshold (at the 10% data percentile) and the time-varying MLE-based threshold (at the 10% percentile of the estimated conditional distribution).

Tables 3.3.1 and 3.3.2 show the mean MSEs (i.e., the average of 50 MSEs for the 50 simulated datasets) for the 100 one-step-ahead VaR and ES predictions from the GARCH(1,1) and AGARCH(1,1) models, respectively. We observe the following find-

[5]Differently than in the i.i.d. or AR(1) model examples, here we need to impose the unit variance restriction on the innovations ε_t, so that the actual volatility of y_t is determined by h_t.

ings. First, the PCP and CP outperform the regular posterior for the 99.5% and 99% VaR and ES. On the other hand, the regular posterior outperforms the PCP and CP for the 95% VaR and ES, where the 5% quantile is apparently not "deep enough" in the tail to make the left-tail focus of censoring beneficial. Second, the PCP outperforms the CP for the small estimation window of $T = 1000$ observations (where it is apparently crucial to limit the loss of information due to censoring), whereas the performance of PCP and CP is similar for the large estimation window of $T = 2500$ observations. Notice that it obviously depends on the model whether an estimation window is "small", where a GARCH(1,1) model requires many more observations than an AR(1) model for accurate estimation, and an AGARCH(1,1) model requires somewhat more observations than a GARCH(1,1) model. Third, typically the PCP (and CP) with time-varying thresholds perform slightly better than their counterparts with time-constant threshold. Fourth, the use of the PCP is equally or less beneficial in the AGARCH(1,1) model than in the GARCH(1,1) model. Hence, we can say that the PCP approach can perform well even when applied to standard models, so that no specific models need to be used to make the PCP beneficial.

3.4 Empirical application

In this section we compare the left-tail forecasting performance for the regular posterior, the censored posterior and the partially censored posterior using empirical data. We consider daily logreturns of the IBM stock, from the 4th January 2007 to the 22nd December 2016 (2512 observations, Figure 3.4.1).

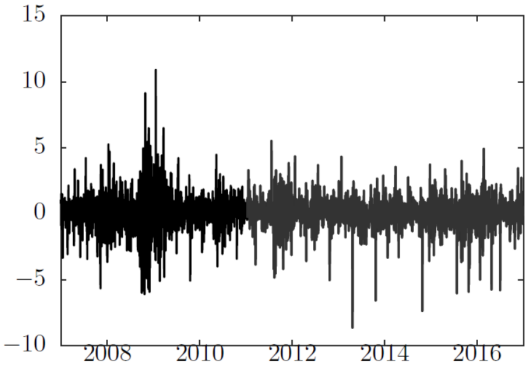


Figure 3.4.1: Daily logreturns of the IBM stock from the 4th January 2007 to the 22nd December 2016. Black: the in-sample period, gray: the out-of-sample period.

In our empirical study we analyse a benchmark model of volatility, commonly employed by practitioners, the generalized autoregressive conditional heteroscedasticity model

Risk measure	Posterior	CP (const. C)	PCP (const. C)	CP (var. C_t)	PCP (var. C_t)
			$T = 1000$		
99.5% VaR	0.1556	0.1188	0.1061	0.1156	**0.1040**
99.5% ES	0.2385	0.1531	0.1431	0.1470	**0.1385**
99% VaR	0.1056	0.0953	0.0835	0.0930	**0.0820**
99% ES	0.1784	0.1259	0.1146	0.1219	**0.1118**
95% VaR	**0.0225**	0.0565	0.0491	0.0554	0.0467
95% ES	**0.0645**	0.0770	0.0674	0.0753	0.0655
			$T = 2500$		
99.5% VaR	0.1572	0.0696	**0.0804**	0.0712	0.0807
99.5% ES	0.2447	0.0831	**0.1005**	0.0840	0.1006
99% VaR	0.1056	0.0589	0.0669	0.0602	**0.0659**
99% ES	0.1815	0.0711	0.0843	0.0724	**0.0841**
95% VaR	**0.0194**	0.0411	0.0424	0.0417	0.0420
95% ES	**0.0627**	0.0495	0.0548	0.0508	0.0546

Table 3.3.1: Simulation results in estimated (misspecified) GARCH(1,1) model with normally distributed innovations for data from DGP of GARCH(1,1) model with split normally distributed innovations (with $\sigma_1 = 1$ and $\sigma_2 = 2$). The table reports the averages of MSEs (over 50 MC replications) for one-step-ahead VaR and ES prediction over an out-of-sample window of $H = 100$ for standard posterior, censored posterior (CP) and partially censored posterior (PCP) – the latter two with time-constant threshold (const. C) and time-varying threshold (var. C_t), at the 10% percentile of the empirical distribution and the 10% percentile of the MLE-implied conditional distribution, respectively. Bold numbers indicate the lowest average MSE.

Risk measure	Posterior	CP (const. C)	PCP (const. C)	CP (var. C_t)	PCP (var. C_t)
			$T = 1000$		
99.5% VaR	0.1369	0.1374	0.1238	0.1372	**0.1201**
99.5% ES	0.2110	0.1802	0.1738	0.1793	**0.1654**
99% VaR	0.0924	0.1079	0.0952	0.1078	**0.0923**
99% ES	0.1572	0.1465	0.1361	0.1458	**0.1305**
95% VaR	**0.0188**	0.0593	0.0546	0.0583	0.0505
95% ES	**0.0551**	0.0857	0.0765	0.0853	0.0732
			$T = 2500$		
99.5% VaR	0.1627	0.0772	0.0901	**0.0764**	0.0893
99.5% ES	0.2509	0.0944	0.1160	**0.0937**	0.1140
99% VaR	0.1101	**0.0627**	0.0731	0.0632	0.0728
99% ES	0.1871	0.0791	0.0957	**0.0788**	0.0944
95% VaR	**0.0212**	0.0377	0.0437	0.0388	0.0434
95% ES	0.0660	**0.0503**	0.0592	0.0512	0.0591

Table 3.3.2: Simulation results in estimated (misspecified) AGARCH(1,1) model with normally distributed innovations for data from DGP of GARCH(1,1) model with split normally distributed innovations (with $\sigma_1 = 1$ and $\sigma_2 = 2$). The table reports the averages of MSEs (over 50 MC replications) for one-step-ahead VaR and ES prediction over an out-of-sample window of $H = 100$ for standard posterior, censored posterior (CP) and partially censored posterior (PCP) – the latter two with time-constant threshold (const. C) and time-varying threshold (var. C_t), at the 10% percentile of the empirical distribution and the 10% percentile of the MLE-implied conditional distribution, respectively. Bold numbers indicate the lowest average MSE.

(GARCH, Engle, 1982; Bollerslev, 1986) with Student's t innovations. We adopt the following specification

$$\begin{aligned} y_t &= \mu + \sqrt{h_t}\varepsilon_t, \\ \varepsilon_t &\sim t(0, 1, \nu), \\ h_t &= \omega(1 - \alpha - \beta) + \alpha(y_{t-1} - \mu)^2 + \beta h_{t-1} \end{aligned}$$

and we put flat priors and impose the standard variance positivity and stationarity restrictions (i.e. $\omega > 0$, $\alpha \in (0, 1)$, $\beta \in (0, 1)$ with $\alpha + \beta < 1$), except for the degrees of freedom, where we use an uninformative yet proper exponential prior (with prior mean 100) for $\nu - 2$.

As a benchmark and the starting point for the PCP approach, we first carry out the standard posterior analysis; second, we perform the estimation based on the CP. Each time we run $M = 10000$ iterations (after a burn-in of 1000) of the IC-MH using as a candidate the mixture of Student's t distributions obtained with the MitISEM algorithm of Hoogerheide et al. (2012) For the PCP, given the posterior draws of $\boldsymbol{\theta}_1 = \{\alpha, \beta\}$ of the dynamics parameters, we conditionally sample $\boldsymbol{\theta}_2 = \{\nu, \mu, \omega\}$ from the conditional censored posterior. For both the CP and PCP we consider two thresholds, a time-constant threshold at the 10% quantile of the in-sample data and a time-varying threshold, the 10% quantile of the MLE-implied conditional distribution.

Table 3.4.1 reports the estimation results and Figure 3.4.2 presents the corresponding kernel density estimates. For the CP leaving α and β to be estimated based on effectively few observations leads to much higher variances. Interestingly, for ν the CP and PCP lead to very different estimation results than the regular posterior. The latter implies a very fat-tailed distribution with a low degrees of freedom parameter, while both the CP and the PCP suggest an almost normal shape of the left tail of the distribution of the innovations (which comes with a high value of ω, suggesting that the left tail may be more like a normal distribution with a higher variance than like a Student's t distribution with a smaller variance). This huge discrepancy between the results from the regular posterior and the (P)CP can be interpreted as evidence of model misspecification.

In our forecasting study we consider $H = 1500$ out-of-sample density forecasts, where we have an in-sample period of $T = 1012$ observations. As our primary interest is accurate left-tail density prediction, we compare the density forecasts based on the censored likelihood scoring rule (3.1.8) of Diks et al. (2011). A novelty of this chapter is that we also allow the threshold *for the assessment of the quality of the left-tail*

Parameter	Posterior	CP (const. C)	PCP (const. C)	CP (var. C_t)	PCP (var. C_t)
ν	7.0943	45.6427	38.1495	43.2466	33.7796
	(1.4748)	(25.3582)	(26.3139)	(25.3678)	(24.8307)
μ	0.0905	0.6684	0.6287	0.5821	0.4492
	(0.0362)	(0.2725)	(0.3825)	(0.2310)	(0.3131)
ω	16.4548	260.5379	47.3642	288.6082	42.7302
	(18.9873)	(369.5474)	(60.8834)	(423.7474)	(59.0946)
α	0.1260	0.1605	0.1264	0.1683	0.1264
	(0.0271)	(0.0515)	(0.0273)	(0.0562)	(0.0273)
β	0.8652	0.8317	0.8650	0.8234	0.8650
	(0.0280)	(0.0530)	(0.0281)	(0.0572)	(0.0281)

Table 3.4.1: Empirical application to daily IBM logreturns: estimation results (means and standard deviations) for the GARCH(1,1)-t model estimated with the regular posterior, the censored posterior (CP) and the partially censored posterior (PCP) – the latter two with time-constant threshold (const. C) and time-varying threshold (var. C_t), at the 10% percentile of the empirical distribution and the 10% percentile of the MLE-implied conditional distribution, respectively.

prediction to be time-varying, which we set to the 0.5%, 1% and 5% percentile of the MLE-implied conditional distribution. We also consider a time-constant threshold for evaluation, as in the previous literature, which we set at the 10% percentile of the in-sample data.

Tables 3.4.2 and 3.4.3 present the results of the Diebold-Mariano test based on the censored scoring rule with time-constant and time-varying threshold, respectively. A positive number indicates that the corresponding row method provides better left-tail density forecasts than the corresponding column method. The plots of the loss differentials used in the Diebold-Mariano tests, provided in Figure 3.D.2 in Appendix 3.D, show that the PCP provides substantially better left-tail density predictions than the CP and the regular posterior on multiple days, whereas it is never (or hardly ever) substantially outperformed.

We can see that the censored likelihood scoring rule with time-constant threshold for evaluation clearly prefers the PCP over the CP: for all the quantile levels considered the PCP significantly outperforms the fully censored approach (at 1% significance level). Moreover, the PCP performs significantly better for the extreme left tail than the regular posterior (at 5% significance level). In this application full censoring is only harmful compared to the regular posterior, which stresses the merit of the introduced *partial* censoring.

Also the conclusions from the evaluations based on the time-varying threshold are supportive for the PCP approach. For the extreme left tail the regular posterior is

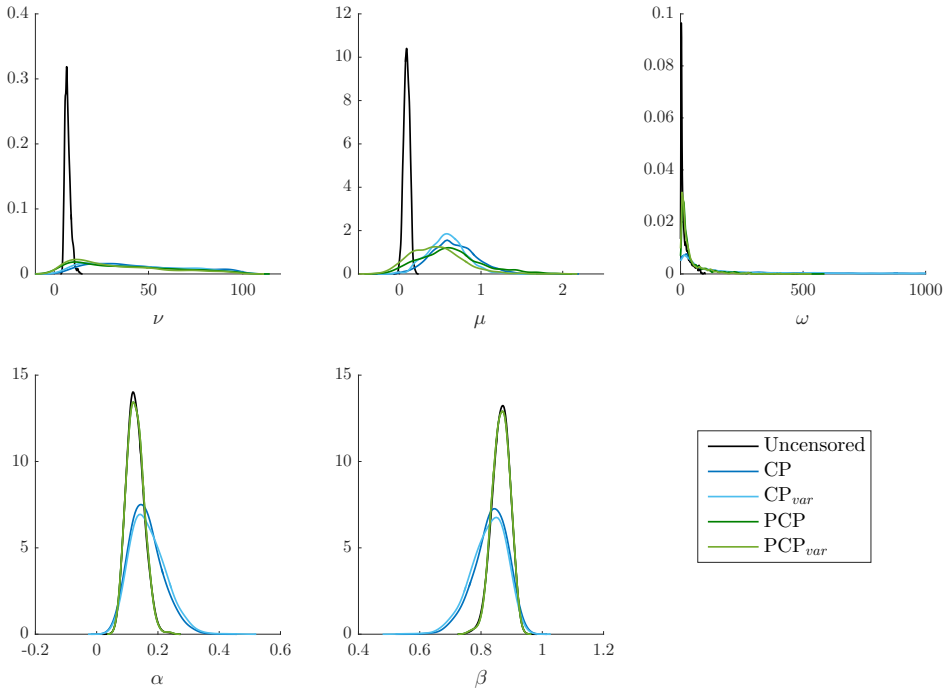


Figure 3.4.2: Empirical application to daily IBM logreturns: kernel density estimates for the regular posterior, censored posterior (CP) and partially censored posterior (PCP) – the latter two with time-constant threshold (const. C) and time-varying threshold (var. C_t), at the 10% percentile of the empirical distribution and the 10% percentile of the MLE-implied conditional distribution, respectively.

significantly outperformed by PCP (at 5% significance level).

We can conclude that for more complex models, usually applied in empirical practice, the role of *partial* censoring becomes crucial. With multiple parameters to be estimated based on effectively few observations, it might be hard for the fully censored posterior to provide accurate left-tail density forecasts, so that it may be more beneficial to use the regular posterior. On the other hand, with an "appropriately" chosen subset of parameters to apply censoring, we can achieve better left-tail density forecasts than with the standard posterior.

3.5 Conclusions

We have proposed a novel approach to inference for a specific region of interest of the predictive distribution. Our Partially Censored Posterior method falls outside the framework of regular Bayesian statistics as we do not work with the regular likelihood but with the censored likelihood based on the censored likelihood scoring rule of Diks et al. (2011). This allows us to keep the merits of the regular Bayesian analysis, e.g. taking into account parameter uncertainty, and at the same time to allow for robust

	Posterior	CP (const. C)	PCP (const. C)	CP (var. C_t)
Threshold for censored likelihood scoring rule = 0.5% percentile				
Posterior	–	–	–	–
CP (const. C)	-2.5013**	–	–	–
PCP (const. C)	2.0464**	2.5867***	–	–
CP (var. C_t)	-2.9143***	-2.0998**	-2.5866***	–
PCP (var. C_t)	2.0369**	2.6460***	-1.8986*	2.6533***
Threshold for censored likelihood scoring rule = 1% percentile				
Posterior	–	–	–	–
CP (const. C)	-3.8343***	–	–	–
PCP (const. C)	1.3922	2.5763***	–	–
CP (var. C_t)	-4.0150***	-1.7450*	-2.5415**	–
PCP (var. C_t)	1.3609	2.7439***	-1.3752	2.7008***
Threshold for censored likelihood scoring rule = 5% percentile				
Posterior	–	–	–	–
CP (const. C)	-4.9013***	–	–	–
PCP (const. C)	-1.8209*	3.5258***	–	–
CP (var. C_t)	-5.0946***	0.0735	-3.2159***	–
PCP (var. C_t)	-2.2713**	3.8532***	-0.7073	3.5323***

Table 3.4.2: Empirical application to daily IBM logreturns: Diebold-Mariano test statistics for pairwise method comparison of forecasting performance in the left tail based on the censored likelihood scoring rule with *time-constant* threshold for evaluation (i.e., for computing the censored likelihood scoring rule), for $H = 1500$ out-of-sample observations, between the regular posterior, censored posterior (CP) and partially censored posterior (PCP) – the latter two with time-constant threshold (const. C) and time-varying threshold (var. C_t), at the 10% percentile of the empirical distribution and the 10% percentile of the MLE-implied conditional distribution, respectively. *, **, *** indicate significance at 10%, 5%, 1% level, respectively.

	Posterior	CP (const. C)	PCP (const. C)	CP (var. C_t)
Threshold for censored likelihood scoring rule = 0.5% percentile				
Posterior	–	–	–	–
CP (const. C)	0.9642	–	–	–
PCP (const. C)	2.1526**	2.2572**	–	–
CP (var. C_t)	0.9848	-0.2301	-2.1132**	–
PCP (var. C_t)	2.3944**	2.4226**	-0.3568	2.3016**
Threshold for censored likelihood scoring rule = 1% percentile				
Posterior	–	–	–	–
CP (const. C)	-0.1137	–	–	–
PCP (const. C)	1.3287	2.8183***	–	–
CP (var. C_t)	-0.0889	0.4013	-2.5591**	–
PCP (var. C_t)	1.5510	3.2282***	0.3566	2.9846***
Threshold for censored likelihood scoring rule = 5% percentile				
Posterior	–	–	–	–
CP (const. C)	0.6637	–	–	–
PCP (const. C)	2.2503**	3.5778***	–	–
CP (var. C_t)	0.5552	-2.2570**	-3.6174***	–
PCP (var. C_t)	2.1086**	3.1398***	-2.5594**	3.2996***

Table 3.4.3: Empirical application to daily IBM logreturns: Diebold-Mariano test statistics for pairwise method comparison of forecasting performance in the left tail based on the censored likelihood scoring rule with *time-varying* threshold for evaluation (i.e., for computing the censored likelihood scoring rule), for $H = 1500$ out-of-sample observations, between the regular posterior, censored posterior (CP) and partially censored posterior (PCP) – the latter two with time-constant threshold (const. C) and time-varying threshold (var. C_t), at the 10% percentile of the empirical distribution and the 10% percentile of the MLE-implied conditional distribution, respectively. *, **, *** indicate significance at 10%, 5%, 1% level, respectively.

inference focused on the left tail in cases of potential model misspecification. The latter is vital for risk management, where the shape of the left tail of the conditional distribution is of crucial importance.

Partitioning of the parameter set into two subsets, one of which is likely to benefit from censoring, increases the precision of the parameter estimates compared to the fully censored posterior of Gatarek et al. (2014) and allows us to obtain better left-tail density forecasts. Further, we have introduced two novel simulation methods, the MCMC method of Conditional MitISEM and the importance sampling method of PCP-QERMit. Finally, we have considered novel ways of time-varying censoring, which allow us for an even better focus on the left tail of the distribution of the standardized innovations. We have demonstrated the usefulness of our methods in extensive simulation and empirical studies.

To further exploit the power of our quasi-Bayesian framework, in future research we intend to employ the PCP in the context of forecast combination via Model Averaging using partially censored predictive likelihoods. Also extensions of the classical approach of Opschoor et al. (2016) based on so-called pooling are relevant in this regard. The Bayesian approach of Aastveit et al. (2018a) can be used in this context. Another interesting extension will be to investigate the impact of using the smoothly-censored likelihood of Diks et al. (2011) in our PCP setting, to make the PCP approach even more robust w.r.t. the choice of the threshold C_t. An important domain of application of the proposed PCP methodology would be portfolio optimization and portfolio risk management, where the evaluation of the probability of y_t lying outside the region of interest ($\mathbb{P}(y_t \in A_t^C | \mathbf{y}_{1:t-1}, \boldsymbol{\theta})$) may require an efficient simulation method. Finally, an interesting extension would be the analysis of credit risk and defaults.

Appendix 3.A Bayesian out-of-sample forecasting

Generally, the H-step-ahead predictive posterior distribution can be specified via the decomposition

$$
\begin{aligned}
p(\mathbf{y}^*_{1:H}|\mathbf{y}_{1:T}) &= p(y^*_1|\mathbf{y}_{1:T}) \prod_{h=2}^{H} p(y^*_h|\mathbf{y}^*_{1:h-1}, \mathbf{y}_{1:T}) \qquad (3.A.1)\\
&= \int p(y^*_1|\mathbf{y}_{1:T}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}_{1:T}) d\boldsymbol{\theta} \prod_{h=2}^{H} \int p(y^*_h|\mathbf{y}^*_{1:h-1}, \mathbf{y}_{1:T}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}_{1:T}) d\boldsymbol{\theta},
\end{aligned}
$$

which is a forecast over all H out-of-sample periods where the posterior density is obtained using only the in-sample data $\mathbf{y}_{1:T}$. A common approach in practice is, however, to consider H one-step-ahead forecasts $\prod_{h=1}^{H} p(y_h^*|\mathbf{y}_{1:(T+h-1)})$ over the out-of-sample period with the new incoming data used to formulate one-step-predictions. The difference with respect to (3.A.1) is that the factors in the product on the right-hand-side in (3.A.1) are replaced by $p(y_h^*|\mathbf{y}_{1:(T+h-1)})$ to deliver

$$\begin{aligned}\prod_{h=1}^{H} p(y_h^*|\mathbf{y}_{1:(T+h-1)}) &= p(y_1^*|\mathbf{y}_{1:T}) \prod_{h=2}^{H} p(y_h^*|\mathbf{y}_{1:(T+h-1)}) \\ &= \int p(y_1^*|\mathbf{y}_{1:T}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}_{1:T}) d\boldsymbol{\theta} \prod_{h=2}^{H} \int p(y_h^*|\mathbf{y}_{1:(T+h-1)}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}_{1:(T+h-1)}) d\boldsymbol{\theta}.\end{aligned} \tag{3.A.2}$$

Such a procedure would in principle require to sequentially update the posterior $p(\boldsymbol{\theta}|\mathbf{y}_{1:(T+h)})$ for each incoming observation, which for large in-sample and out-of-sample periods might be computationally prohibitive. Hence, the pragmatic solution commonly adopted in practice is based on the approximation $p(\boldsymbol{\theta}|\mathbf{y}_{1:(T+h)}) \approx p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$ which for T sufficiently large should not be too crude. The resulting approximation to (3.A.2) is

$$\prod_{h=1}^{H} p(y_h^*|\mathbf{y}_{1:(T+h-1)}) \approx \int p(y_1^*|\mathbf{y}_{1:T}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|y_{1:T}) d\boldsymbol{\theta} \prod_{h=2}^{H} \int p(y_h^*|\mathbf{y}_{1:h-1}^*, \mathbf{y}_{1:T}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}_{1:T}) d\boldsymbol{\theta}.$$

Appendix 3.B Conditional density of (mixture of) multivariate Student's t distributions

Student's t distribution Let $\mathbf{x} \in \mathbb{R}^d$ follow the Student's t distribution with mode μ, scale matrix Σ and ν degrees of freedom, denoted $t(\mathbf{x}; \mu, \Sigma, \nu)$, where we assume $\nu > 2$ so that $\text{Var}(\mathbf{x}) = \frac{\nu}{\nu-2}\Sigma$. Then, the probability density function (pdf) of $\mathbf{x}$ is given by (see Zellner, 1996; Roth, 2013)

$$p(\mathbf{x}) = \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{d}{2}\right)(\pi\nu)^{\frac{d}{2}}}|\Sigma|^{-\frac{1}{2}}\left(1 + \frac{(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)}{\nu}\right)^{-\frac{d+\nu}{2}}.$$

Next, consider a partitioning of $\mathbf{x}$ into $\mathbf{x} = (\mathbf{x}_1', \mathbf{x}_2')'$ with $\mathbf{x}_1$ and $\mathbf{x}_2$ of dimensions d_1 and d_2, respectively. The corresponding parameter partitionings are then

$$\mu = \begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix}, \qquad \Sigma = \begin{pmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}.$$

Then, the conditional density of x_2 given x_1 is also a Student's t density, which is given by

$$p(\mathbf{x}_2|\mathbf{x}_1) = \frac{p(\mathbf{x}_1, \mathbf{x}_2)}{p(\mathbf{x}_1)} = t(\mathbf{x}_2; \mu_{2|1}, \Sigma_{2|1}, \nu_{2|1}),$$

with

$$\begin{aligned}
\mu_{2|1} &= \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{x}_1 - \mu_1), \\
\Sigma_{2|1} &= \frac{\nu + (\mathbf{x}_1 - \mu_1)'\Sigma_{11}^{-1}(\mathbf{x}_1 - \mu_1)}{\nu + d_1}\left(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right), \\
\nu_{2|1} &= \nu + d_1.
\end{aligned}$$

Mixture of Student's t distributions The above result extends to mixtures of Student's t distributions. Now let x follow an H component mixture of Student's t distributions $t(\mathbf{x}; \mu_h, \Sigma_h, \nu_h)$, with component probabilities η_h, $h = 1, \ldots, H$, so that its pdf is given by

$$p(\mathbf{x}) = \sum_{h=1}^{H} \eta_h t(\mathbf{x}; \mu_h, \Sigma_h, \nu_h).$$

Let $\mathbf{z}$ denote a (latent) H-dimensional vector indicating from which component the observation x stems: if $\mathbf{x}$ stems from the hth component then $\mathbf{z} = \mathbf{e}_h$, the hth vector of the standard basis of $\mathbb{R}^H$, i.e. $z_h = 1$ and $z_l = 0$ for $l \neq h$. Obviously, unconditionally

$\mathbb{P}(z = e_h) = \eta_h$. The conditional probability of $\mathbf{x}$ stemming from the hth component is

$$\begin{aligned}\mathbb{P}[\mathbf{z} = \mathbf{e}_h|\mathbf{x}] &= \frac{p(\mathbf{z} = \mathbf{e}_h, \mathbf{x})}{p(\mathbf{x})} \\ &= \frac{\mathbb{P}[\mathbf{z} = \mathbf{e}_h]p(\mathbf{x}|\mathbf{z} = \mathbf{e}_h)}{\sum_{m=1}^{H} \mathbb{P}[\mathbf{z} = \mathbf{e}_m]p(\mathbf{x}|\mathbf{z} = \mathbf{e}_m)} \\ &= \frac{\eta_h t(\mathbf{x}; \mu_h, \Sigma_h, \nu_h)}{\sum_{m=1}^{H} \eta_m t(\mathbf{x}; \mu_m, \Sigma_m, \nu_m)}.\end{aligned}$$

Then, the conditional density of $\mathbf{x}_2$ given $\mathbf{x}_1$ is given by

$$p(\mathbf{x}_2|\mathbf{x}_1) = \frac{p(\mathbf{x}_1, \mathbf{x}_2)}{p(\mathbf{x}_1)} = \frac{\sum_{h=1}^{H} \eta_h t(\mathbf{x}; \mu_h, \Sigma_h, \nu_h)}{\sum_{h=1}^{H} \eta_h t(\mathbf{x}_1; \mu_{h,1}, \Sigma_{h,1}, \nu_h)} = \sum_{h=1}^{H} \eta_{h,2|1} t(\mathbf{x}_2; \mu_{h,2|1}, \Sigma_{h,2|1}, \nu_{h,2|1}),$$

with

$$\begin{aligned}\mu_{h,2|1} &= \mu_{h,2} + \Sigma_{h,21}\Sigma_{h,11}^{-1}(\mathbf{x}_1 - \mu_{h,1}), \\ \Sigma_{h,2|1} &= \frac{\nu_h + (\mathbf{x}_1 - \mu_{h,1})'\Sigma_{h,11}^{-1}(\mathbf{x}_1 - \mu_{h,1})}{\nu_h + d_1} \left(\Sigma_{h,22} - \Sigma_{h,21}\Sigma_{h,h,11}^{-1}\Sigma_{h,12}\right), \\ \nu_{h,2|1} &= \nu_h + d_1,\end{aligned}$$

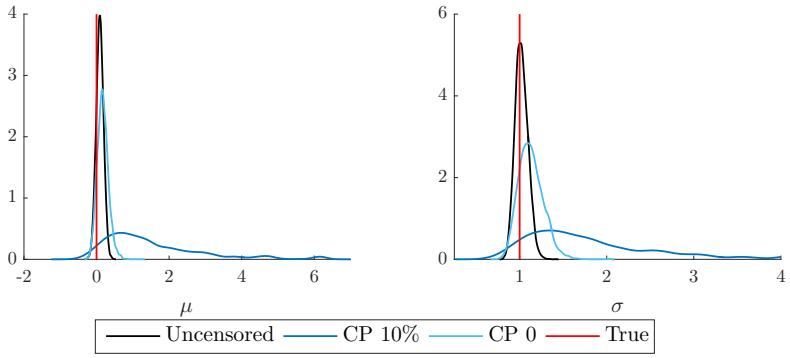
and with adjusted component probabilities

$$\eta_{h,2|1} = \mathbb{P}[\mathbf{z} = \mathbf{e}_h|\mathbf{x}] = \frac{\eta_h t(\mathbf{x}_1; \mu_{h,1}, \Sigma_{h,11}, \nu_h)}{\sum_{m=1}^{H} \eta_m t(\mathbf{x}_1; \mu_{m,1}, \Sigma_{m,11}, \nu_m)}.$$

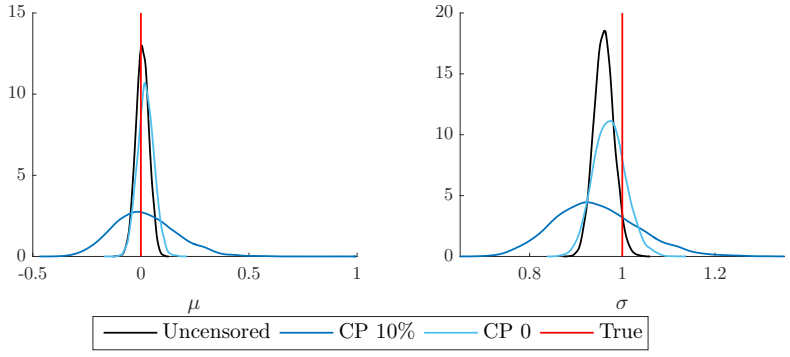
This implies that if we have obtained $q_{mit}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, a mixture of Student's t densities that approximates the joint censored posterior $p^{cp}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y})$, then we can use the M implied conditional mixtures of Student's t densities $q_{cmit}(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^{(i)})$ $(i = 1, \ldots, M)$, as candidate densities for $p^{cp}(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1^{(i)}, \mathbf{y})$ $(i = 1, \ldots, M)$. Hence, we only need one MitISEM approximation to obtain all the conditional candidate densities in our proposed Conditional MitISEM method.

Appendix 3.C Density estimates

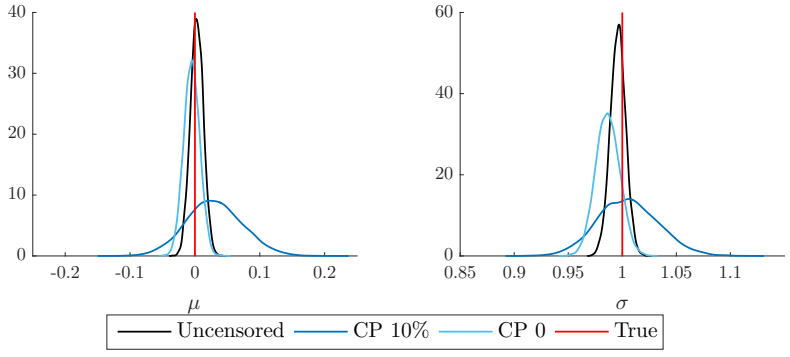
3.C.1 I.i.d.



(a) $T = 100$

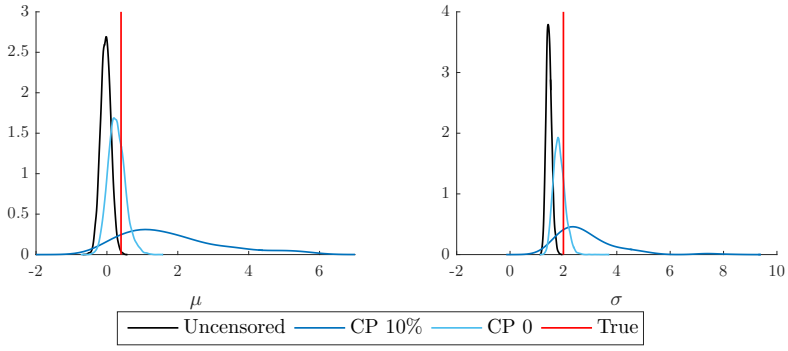


(b) $T = 1000$

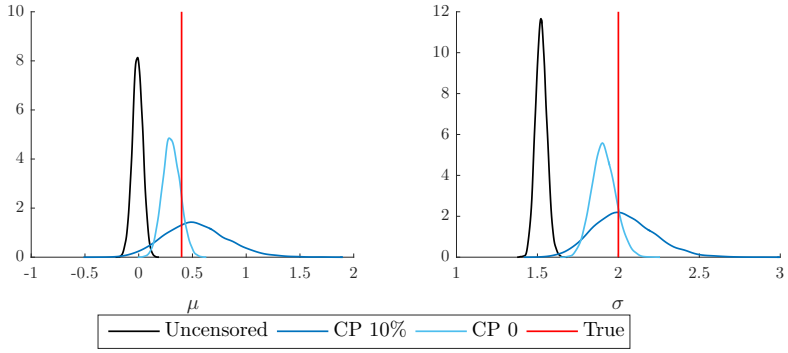


(c) $T = 10000$

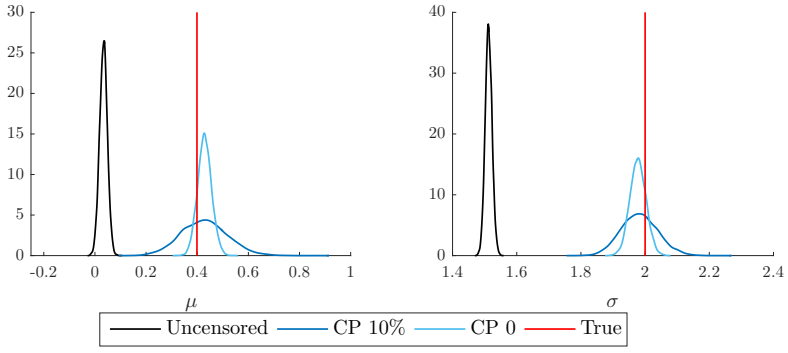
Figure 3.C.1: Estimation results in i.i.d. normal $N(\mu, \sigma^2)$ model for $T = 100, 1000, 10000$ observations from DGP of i.i.d. normal ($\sigma = 1$). Kernel density estimates of regular posterior and censored posterior (CP) with two different thresholds, at 0 (CP0) and at the 10% data percentile (CP10%) together with the true parameter values (corresponding to left tail).



(a) $T = 100$



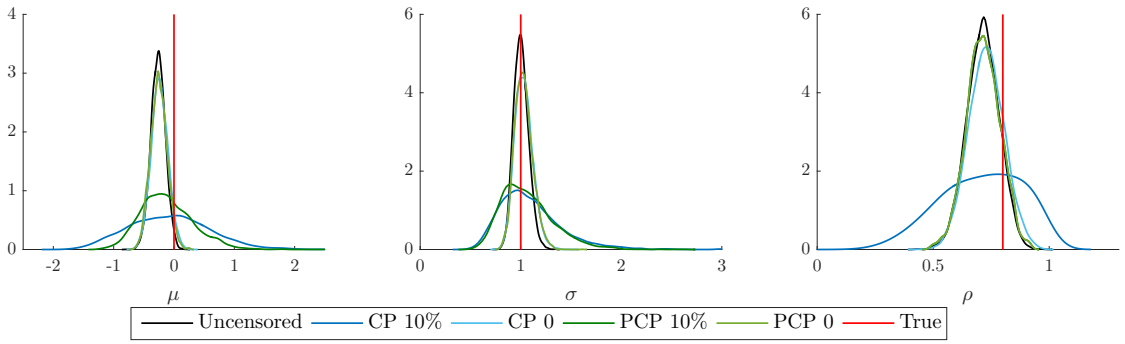
(b) $T = 1000$



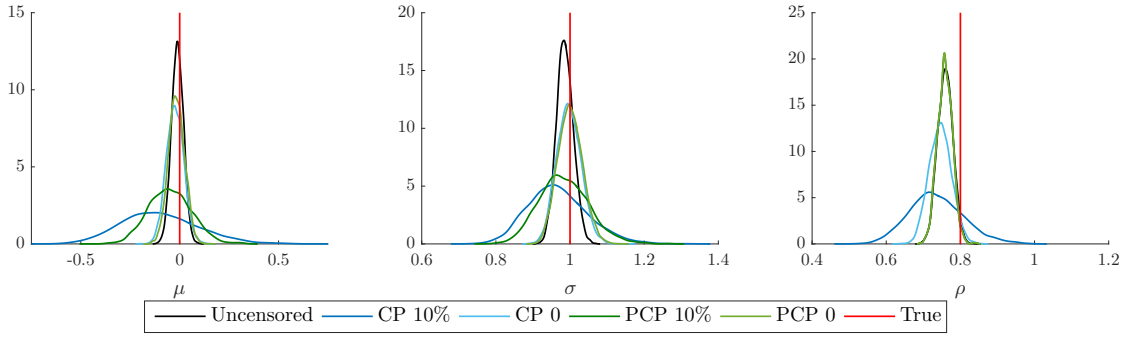
(c) $T = 10000$

Figure 3.C.2: Estimation results in i.i.d. normal $N(\mu, \sigma^2)$ model for $T = 100, 1000, 10000$ observations from DGP of i.i.d. split normal ($\sigma_1 = 1$, $\sigma_2 = 2$). Kernel density estimates of regular posterior and censored posterior (CP) with two different thresholds, at 0 (CP0) and at the 10% data percentile (CP10%) together with the true parameter values (corresponding to left tail).

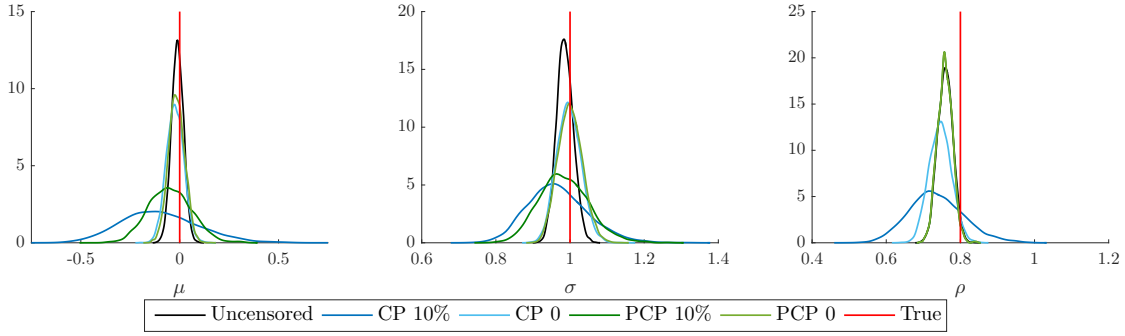
3.C.2 AR(1)



(a) $T = 100$

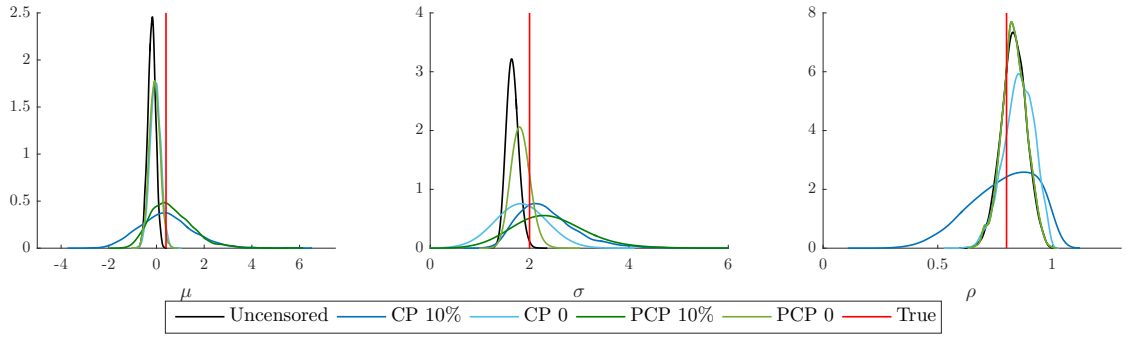


(b) $T = 1,000$

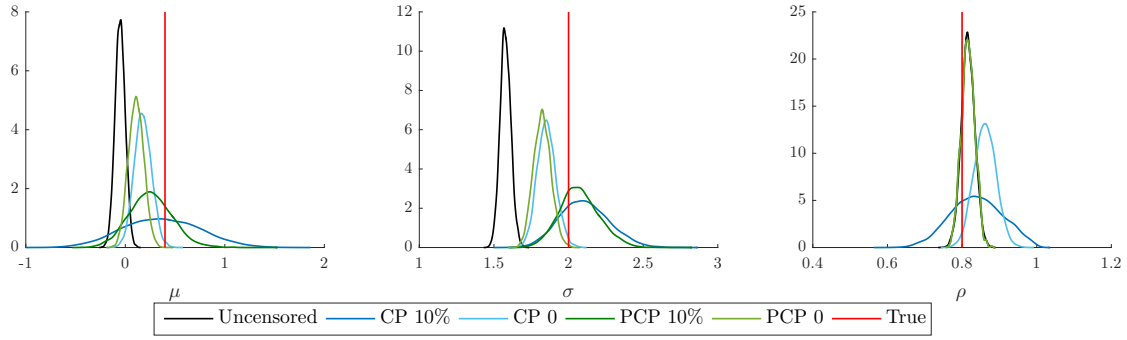


(c) $T = 10,000$

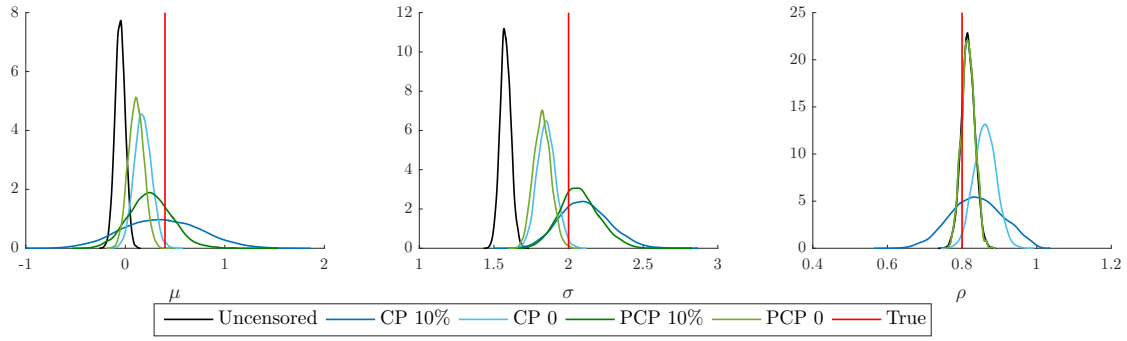
Figure 3.C.3: Symmetric (correctly specified) AR(1) mean zero split normal model with $\sigma_1 = 1$ and $\sigma_2 = 1$: kernel density estimates for the regular posterior, censored posterior (CP) and partially censored posterior (PCP) with different thresholds, at 0 (CP0, PCP0) and at the 10% data percentile (CP10%, PCP 10%) together with the true values for the left tail.



(a) $T = 100$



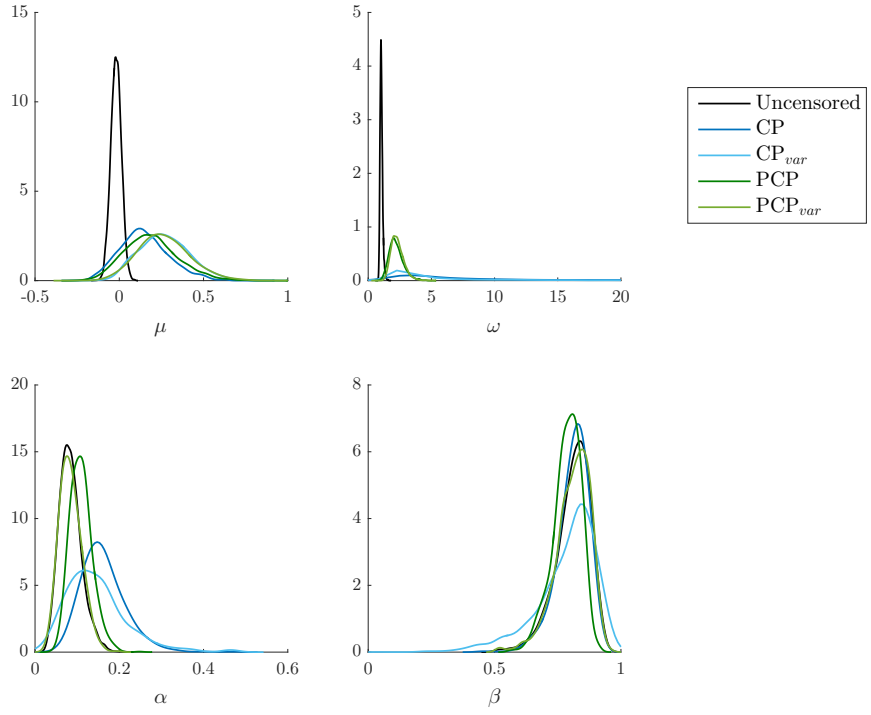
(b) $T = 1,000$



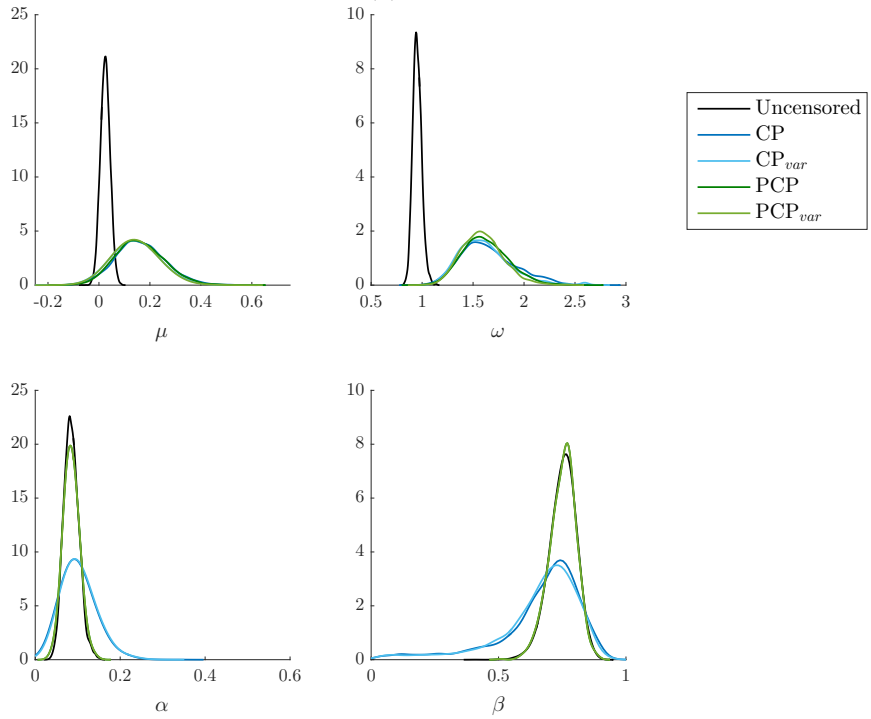
(c) $T = 10,000$

Figure 3.C.4: Asymmetric (misspecified) AR(1) mean zero split normal model with $\sigma_1 = 1$ and $\sigma_2 = 2$: kernel density estimates for the regular posterior, censored posterior (CP) and partially censored posterior (PCP) with different thresholds, at 0 (CP0, PCP0) and at the 10% data percentile (CP10%, PCP 10%) together with the true values for the left tail.

3.C.3 GARCH(1,1)



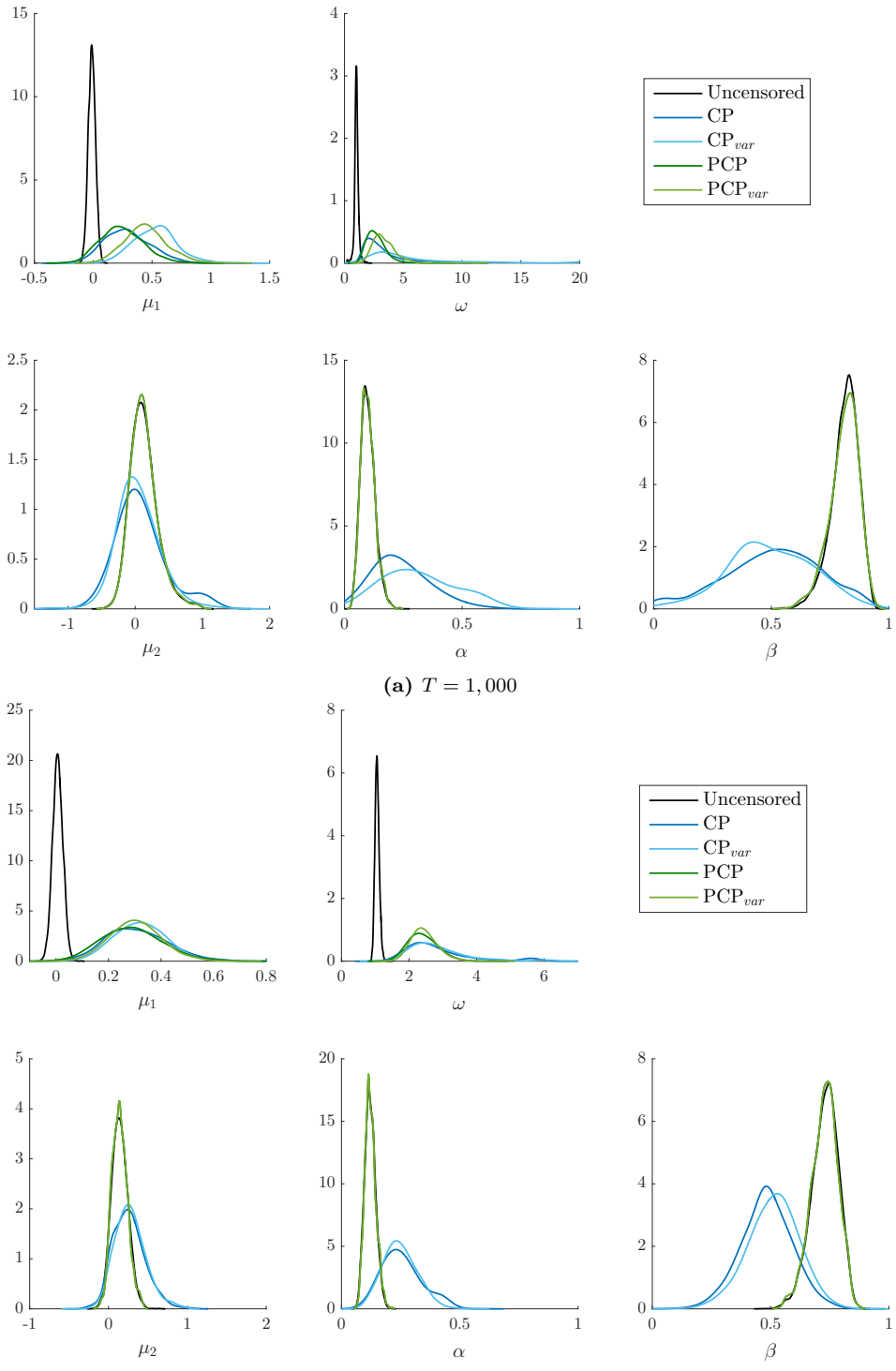
(a) $T = 1,000$



(b) $T = 2,500$

Figure 3.C.5: GARCH(1,1) mean zero split normal model with $\sigma_1 = 1$ and $\sigma_2 = 2$: kernel density estimates for the regular posterior, censored posterior (CP) and partially censored posterior (PCP) with different thresholds, at 0 (CP0, PCP0) and at the 10% data percentile (CP10%, PCP 10%).

3.C.4 AGARCH(1,1)

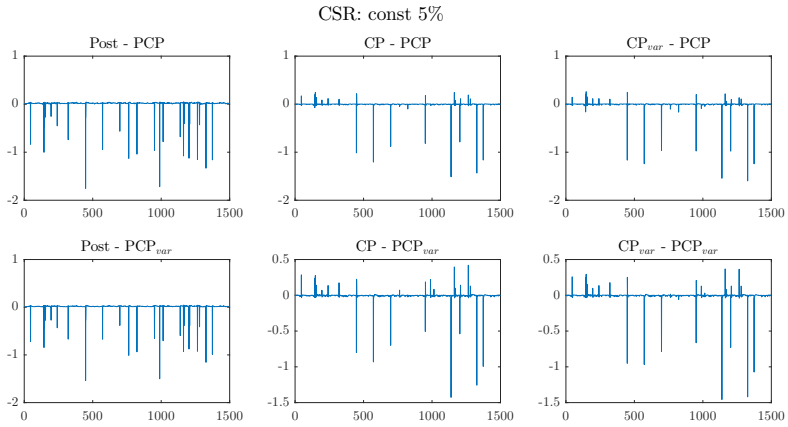


(a) $T = 1,000$

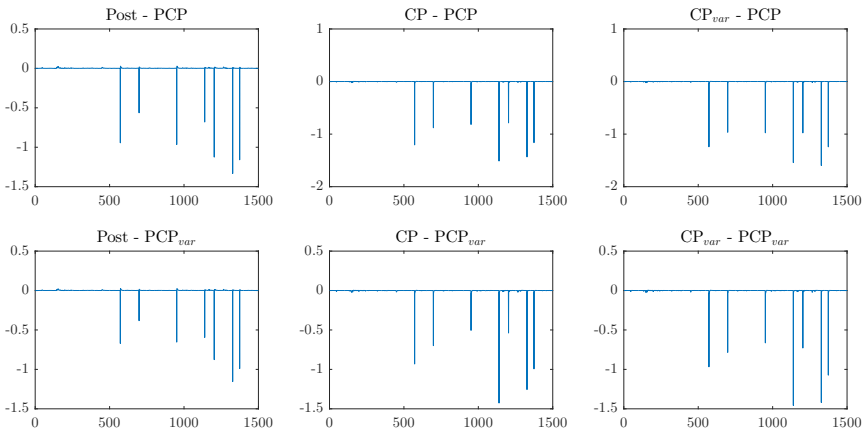
(b) $T = 2,500$

Figure 3.C.6: AGARCH(1,1) mean zero split normal model with $\sigma_1 = 1$ and $\sigma_2 = 2$: kernel density estimates for the regular posterior, censored posterior (CP) and partially censored posterior (PCP) with different thresholds, at 0 (CP0, PCP0) and at the 10% data percentile (CP10%, PCP 10%).

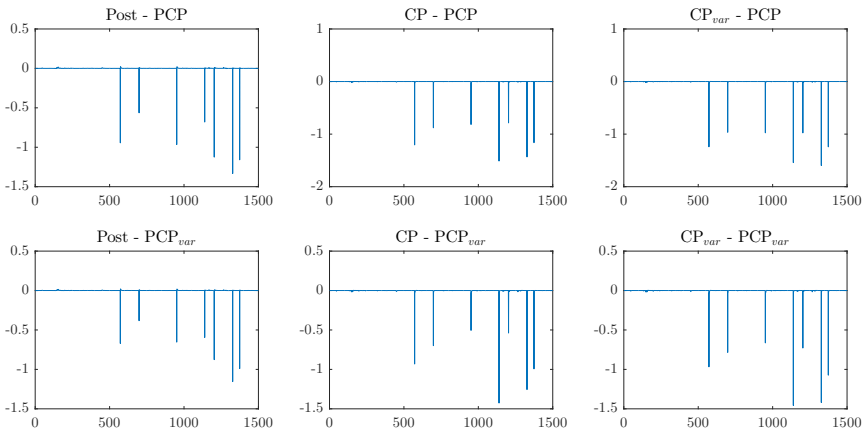
Appendix 3.D Loss differential plots



(a) Threshold for censored likelihood scoring rule = 5% percentile.

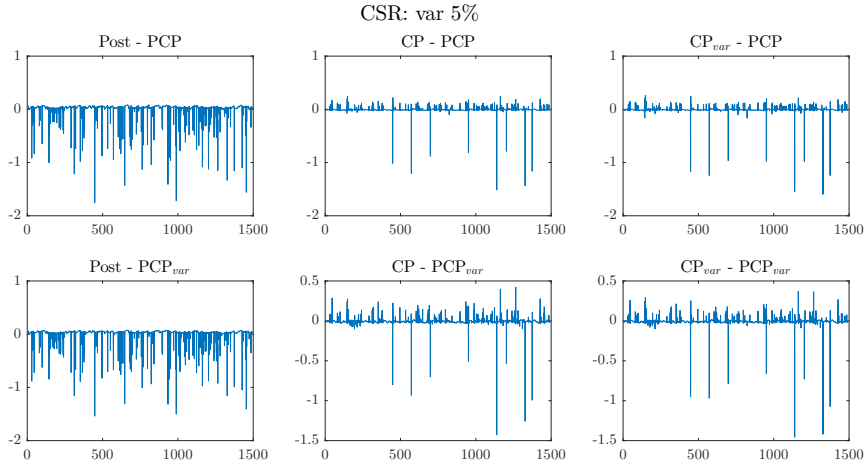


(b) Threshold for censored likelihood scoring rule = 1% percentile.

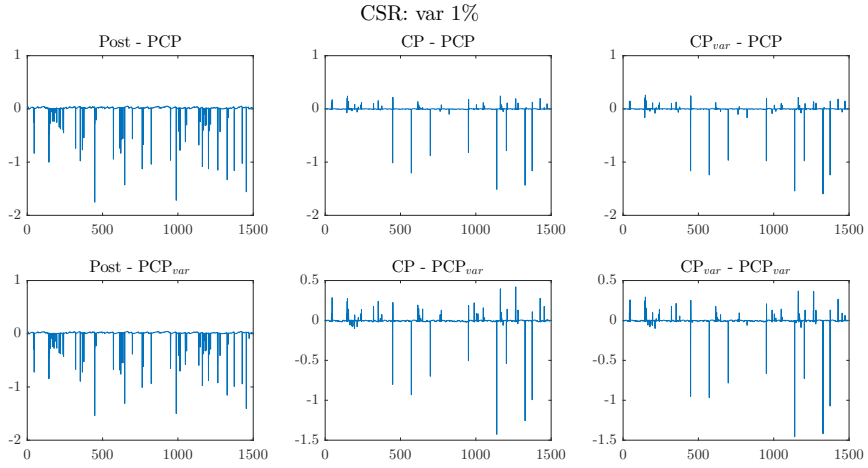


(c) Threshold for censored likelihood scoring rule = 0.5% percentile.

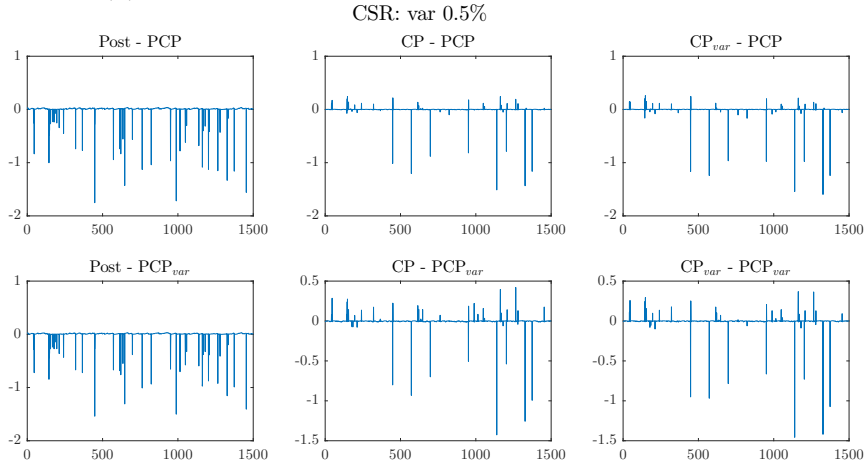
Figure 3.D.1: Empirical application to daily IBM logreturns: loss differentials based on the censored likelihood scoring rule with *time-constant* evaluation threshold (computed as the percentile of the empirical distribution). Negative values of the loss differential indicate that the Partially Censored Posterior (PCP) performs better than the alternative.



(a) Threshold for censored likelihood scoring rule = 5% percentile.



(b) Threshold for censored likelihood scoring rule = 1% percentile.



(c) Threshold for censored likelihood scoring rule = 0.5% percentile.

Figure 3.D.2: Empirical application to daily IBM logreturns: loss differentials based on the censored likelihood scoring rule with *time-varying* evaluation threshold (computed as the percentile of the estimated conditional distribution based on the ML estimator). Negative values of the loss differential indicate that the Partially Censored Posterior (PCP) performs better than the alternative.

Chapter 4

Semi-Complete Data Augmentation for Efficient State Space Model Fitting

The task of inference about a latent state governing the dynamics of the system under study given only the observed noisy data is ubiquitous in many contexts, e.g. in applied statistics, ecology, engineering or economics. A very intuitive way of describing such problems is provided by latent process models, also known as state space models (SSM), see Durbin and Koopman (2012) and West and Harrison (1997) for the Bayesian perspective. Such models are frequently used due to the combination of their natural separation of the different mechanisms acting on the system of interest: the (unobserved) underlying system process; and the observation process. Considering each distinct process separately simplifies the model specification process, and provides a very flexible modelling approach. This flexibility, however, typically comes at the price of substantially more complicated fitting of such models to data. For the general non-linear non-Gaussian SSM the associated likelihood is analytically intractable so that no closed-form solution is available to the optimal estimation problem. Only in certain circumstances the associated likelihood can be calculated explicitly: for linear Gaussian systems the Kalman filter provides the optimal state estimator; for hidden Markov models specified on a discrete state space the likelihood may be available in a closed-form (but may become infeasible for a large number of states). In this chapter we focus on models for which the likelihood is analytically intractable or for which it may be infeasible to compute explicitly.

Dominant approaches to intractable likelihood problems include: (i) numerical or

Monte Carlo integration to estimate the observed (or marginal) data likelihood; and (ii) data augmentation (DA), based on the complete (or joint) data likelihood of the observed and the imputed unobserved states, see Tanner and Wong (1987). The former group includes the sequential Monte Carlo (SMC) methods, see Doucet et al. (2001) for an extensive review, which can be used for parameter estimation within a standard Markov chain Monte Carlo (MCMC) algorithm (i.e. particle MCMC, see Andrieu et al., 2010). Provided the corresponding likelihood estimator is unbiased, the convergence to the correct posterior is guaranteed by the pseudo marginal theory (Beaumont, 2003; Andrieu and Roberts, 2009). In general, numerical integration provides a limited solution, feasible only for very low dimensional systems. The latter DA approach has become a standard method for inference for SSMs within a Bayesian framework, see Frühwirth-Schnatter (1994, 2004); Hobert (2011). DA relies on the true unknown states being treated as auxiliary variables and imputed within the MCMC algorithm. However, the general Bayesian DA approach implemented using standard "vanilla" MCMC algorithms may perform very poorly due to high correlation between the imputed states and/or parameters, see Hobert et al. (2011) and the references therein. This leads to the need for specialist, model-specific algorithms and related bespoke codes.

We propose a novel efficient model-fitting algorithm to circumvent these inefficiencies by combining DA with numerical integration in a Bayesian hybrid approach, where the associated standard "vanilla" algorithms perform substantially more efficiently. The underlying idea is to combine the "good" aspects of both methods by minimising the problems that arise for each, i.e. highly correlated latent states for DA and the curse of dimensionality for numerical integration. To this end, we utilise the structure of the unknown states which can be split into two types: auxiliary variables, which are imputed within the MCMC algorithm using DA; and "integrable" states, which are numerically integrated out within the likelihood expression. We specify the unknown states in such a way that the algorithm is efficient where the imputed states have limited/reduced correlation and the numerical integration is over a very low number of dimensions.

The structure of the chapter is a follows. Section 4.1 presents the general SSM specification and discusses the previous approaches to fit these general models to data. Section 4.2 introduces the proposed Semi-Complete Data Augmentation approach while Section 4.3 develops a general HMM-based approximation to the associated likelihood. We demonstrate the efficiency gains from the new method in Section 4.4, where we discuss two empirical applications relating to the abundance estimation for the ecological data,

and to the estimation of the stochastic volatility (SV) for financial data. Section 4.5 concludes with a discussion.

4.1 State space models

Consider a state space model of the form:

$$\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\theta} \sim p(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\theta}), \tag{4.1.1}$$

$$\mathbf{x}_{t+1}|\mathbf{x}_t, \boldsymbol{\theta} \sim p(\mathbf{x}_{t+1}|\mathbf{x}_t, \boldsymbol{\theta}), \tag{4.1.2}$$

$$\mathbf{x}_0|\boldsymbol{\theta} \sim p(\boldsymbol{\theta}), \tag{4.1.3}$$

for $t = 1, \ldots, T$, with $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_T)$. Here $\mathbf{y}_t \in \mathcal{Y}$ denotes a time series of observations (potentially multivariate, although in our examples they are univariate), $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ a series of latent states (with $\mathbf{x}_t = [x_{1,t}, \ldots, x_{D,t}]^T$ potentially multivariate, $x_{d,t} \in \mathcal{X}_d$) and $\boldsymbol{\theta}$ the model parameters for which we put a prior $p(\boldsymbol{\theta})$. T denotes the length of the time series and $D < \infty$ the dimension of the state $\mathbf{x}_t$. To simplify notation, below we use p as a general symbol for a probability mass function (pmf) or a probability density function (pdf), possibly conditional.

The system process describing the evolution of $\mathbf{x}_t$, the true (unobserved) state of the system over time is defined by the distribution (4.1.2). The observation process which generates $\mathbf{y}_t$, the observed data given the true underlying states, is specified by the distribution (4.1.1). This separation of the different mechanisms acting on the system of interest makes SSM a very intuitive and flexible description of time series data. Figure 4.1.1 graphically presents the dependencies between states and observations in the SSM. An extensive discussion of SSMs is provided by Durbin and Koopman (2012) and also Cappé et al. (2006), where this class of models is called hidden Markov models (HMM)[1]

Modelling flexibility of SSMs is, however, often offset with the issue of estimating $\boldsymbol{\theta}$, the associated model parameters. The *observed data likelihood* for the system (4.1.1)–(4.1.3) is given by

$$p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}) d\mathbf{x} = \int p(x_0|\boldsymbol{\theta}) \prod_{t=1}^{T} p(y_t|x_t, \boldsymbol{\theta}) p(x_t|x_{t-1}, \boldsymbol{\theta}) dx_0 dx_1 \ldots dx_T, \tag{4.1.4}$$

[1]The terminology is not fully consistent in this context: the term "HMM" is sometimes used only for SSMs with a finite state space, i.e. $\dim(\mathcal{X}_d) < \infty$. This convention is used by e.g. Zucchini et al. (2016).

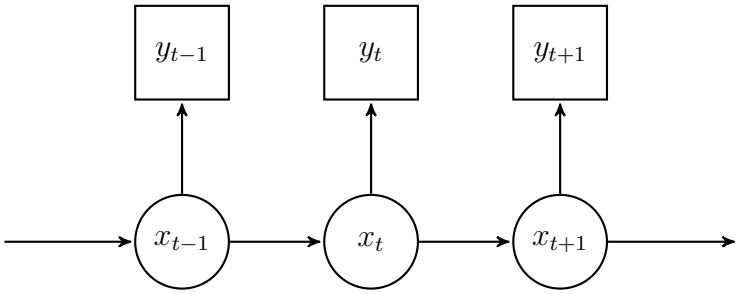


Figure 4.1.1: A graphical representation of the general first-order SSM.

and typically is not available in closed form. This is due to the integration over the latent variables, which is difficult to calculate, despite the tractability of the joint distribution of the data and the auxiliary variables $p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})$. The latter is often referred to as the *complete data likelihood*.

For models with discrete states the observed data likelihood is the likelihood of an HMM, where the states of the chain correspond to distinct values of the latent process, and the transition matrix can be derived from the transition equation (4.1.2). This likelihood can be efficiently calculated using the forward algorithm (see Zucchini et al., 2016). However, for systems with multiple processes there may be a very large number of possible states. This can lead to the approach being infeasible due to the curse of dimensionality. In addition, such an approach becomes infeasible even for simple systems, with e.g. only 2 processes, but with many potential state outcomes (i.e. when $\dim(\mathcal{X}_d)$ is "large").

To overcome the problem of the intractable likelihood, the standard DA technique is commonly adopted, see Tanner and Wong (1987); Frühwirth-Schnatter (1994, 2004); Hobert (2011). In DA the unknown states $\mathbf{x}$ are treated as auxiliary variables and imputed[2]. This way one can work with the closed-from complete data likelihood

$$p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}) = p(x_0|\boldsymbol{\theta}) \prod_{t=1}^{T} p(y_t|x_t, \boldsymbol{\theta}) p(x_t|x_{t-1}, \boldsymbol{\theta}).$$

In the Bayesian framework, the complete data likelihood is used to construct the joint posterior distribution of $\boldsymbol{\theta}$ and $\mathbf{x}$

$$p(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y}) \propto p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) = p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}).$$

Then an MCMC algorithm (or other) can be employed to draw from the joint posterior

[2]A similar idea underlies the expectation-maximisation algorithm of Dempster et al. (1977) in the classical framework.

distribution and the generated values of $\boldsymbol{\theta}$ are taken as a sample from the (marginal) posterior distribution of interest $p(\boldsymbol{\theta}|\mathbf{y})$. In practice the random walk Metropolis-Hastings (RW-MH) algorithm is often used and it acts as a "vanilla" MCMC algorithm (see Marin and Robert, 2007, Ch. 4).

DA is a powerful tool for dealing with intractable likelihoods, however it often results in posterior draws being highly correlated, indicating poor mixing and hence low efficiency of MCMC algorithms. This is particularly the case for SSMs models which impose a strong dependence structure on the latent variables and parameters. Single-update algorithms can perform especially poorly, nevertheless they are often used as they are easy to implement. An alternative approach based on block sampling, i.e. simultaneously updating the target distribution in multiple dimensions, can lead to an improved mixing. However, it requires defining an appropriate partition of the states and parameters into blocks and specifying an efficient proposal distributions for each block. These problems of the standard DA approach often result in specialist algorithms being developed for the purpose of efficient estimation of a given model. Consequently, bespoke codes need to be written dependent on model and data.

4.2 Semi-Complete Data Augmentation

To improve the efficiency of the standard DA approach, we propose to combine DA with numerical integration within a Bayesian hybrid framework, which we call *Semi-Complete Data Augmentation*. A key idea is to separate the latent state $\mathbf{x}$ into two components $\mathbf{x} = (\mathbf{x_{aug}}, \mathbf{x_{int}})$. We will refer to $\mathbf{x_{int}}$ and $\mathbf{x_{aug}}$ as the "integrated" states and the "augmented" states, respectively. The starting point for our method is to specify the *semi-complete data likelihood* (SCDL) $p(\mathbf{y}, \mathbf{x_{aug}}|\boldsymbol{\theta})$ as follows:

$$\begin{aligned} p(\mathbf{y}, \mathbf{x_{aug}}|\boldsymbol{\theta}) &= \int p(\mathbf{y}, \mathbf{x_{aug}}, \mathbf{x_{int}}|\boldsymbol{\theta}) d\mathbf{x_{int}} \\ &= \int p(\mathbf{y}|\mathbf{x_{aug}}, \mathbf{x_{int}}, \boldsymbol{\theta}) p(\mathbf{x_{aug}}, \mathbf{x_{int}}|\boldsymbol{\theta}) d\mathbf{x_{int}}. \end{aligned} \tag{4.2.1}$$

The joint posterior distribution of the parameters and augmented states can be then expressed as

$$\begin{aligned} p(\boldsymbol{\theta}, \mathbf{x_{aug}}|\mathbf{y}) &\propto p(\mathbf{y}, \mathbf{x_{aug}}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \\ &= p(\mathbf{y}|\mathbf{x_{aug}}, \boldsymbol{\theta}) p(\mathbf{x_{aug}}|\boldsymbol{\theta}) p(\boldsymbol{\theta}). \end{aligned}$$

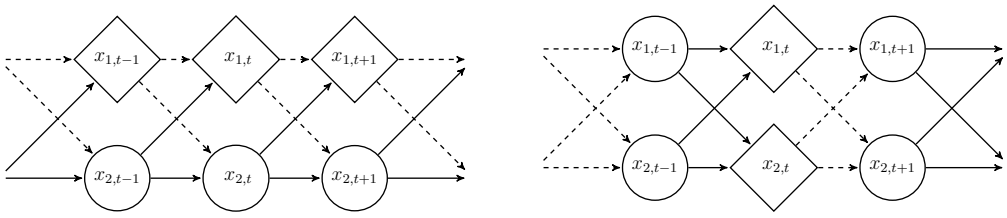
We note that our approach builds upon the work of King et al. (2016), who propose a Bayesian hybrid approach applied to the particular case of capture-recapture data. These authors define the "semi-complete" data likelihood as the product of a complete data likelihood component for the individuals observed within the study (related to $\mathbf{x_{aug}}$) and a marginal data likelihood component for the unobserved individuals (related to $\mathbf{x_{int}}$). We extend their approach to the general state space models framework and consider different schemes for specifying the semi-complete data likelihood in terms of defining $\mathbf{x_{aug}}$ and $\mathbf{x_{int}}$.

Specification of the auxiliary variables More precisely, consider a time series $\mathbf{x} = \{\mathbf{x}_t\}_{t=0}^{T}$ of length $T+1$, where the state at time t is D dimensional: $\mathbf{x}_t = \left[x_{1,t}, \ldots, x_{D,t}\right]^T$, for $t = 0, 1, \ldots, T$. We want to integrate out D_{int} dimensions of the state at time points T_{int}, where $D_{int} \subset \{1, \ldots, D\}$ are $T_{int} \subset \{0, 1, \ldots, T\}$ are "suitably" chosen subsets of dimension and time indices, respectively. Such a "suitable" specification of subsets D_{int} and T_{int} depends on the dependence structure of the model under consideration so that the implied integral can be efficiently calculated. For instance, it can be low dimensional or it can be reduced to a product of low-dimensional integrals. The compliments of both subsets are denoted D_{aug} and T_{aug}, respectively. We also denote T_{int}^{+} and T_{aug}^{+} the corresponding sets without the initial observations, i.e. excluding time $t = 0$. The "integrated" and "augmented" states are then defined by the partition of $\mathbf{x}$ into $\mathbf{x_{int}} = \{x_{d,t}\}_{d \in D_{int}, t \in T_{int}}$ and $\mathbf{x_{aug}} = \{x_{d,t}\}_{d \in D_{aug}, t \in T_{aug}}$, where we denote their corresponding elements at time t by $\mathbf{x_{int,t}} = \{x_{d,t}\}_{d \in D_{int}}$ and $\mathbf{x_{aug,t}} = \{x_{aug,t}\}_{d \in D_{int}}$, respectively. In particular, we give the following two examples of integration/augmentation schemes.

(a) *"Horizontal" integration:* e.g. for a $D = 2$ dimensional state we integrate out the second state at all time periods, so that $D_{int} = \{2\}$ (and hence $D_{aug} = \{1\}$), and $T_{int} = \{0, 1, \ldots, T\}$ (and hence $T_{aug} = T_{int}$), see Figure 4.2.1a. We use this scheme in the lapwings data application in Section 4.4.1.

(b) *"Vertical" integration:* e.g. all D states are integrated out at odd time periods $D_{int} = \{1, \ldots, D\}$ and $T_{int} = \{2t+1\}_{t=0}^{\lfloor T/2 \rfloor}$ (and hence $T_{aug} = \{2t\}_{t=0}^{\lfloor T/2 \rfloor}$), see Figure 4.2.1b. We use this scheme in the stochastic volatility (SV) model application in Section 4.4.2, for $D = 1$ dimensional state.

As we can see, in general T_{int} and T_{aug} do not need to be equal and their elements may not be consecutive numbers. However, we would like to iterate over both sets



(a) Horizontal integration. **(b) Vertical integration**

Figure 4.2.1: Two examples of an integration/augmentation scheme. Diamonds represent the imputed states, circles – the integrated states. Dashed lines used for the relations *from* the imputed (known) states.

using the same index. Therefore, we introduce two functions $\tau(t)$ and $a(t)$ such that the image of τ is T^+_{int} and the image of a covers T^+_{aug}, both defined on $1, 2, \ldots, |T^+_{int}|$. We require τ to be bijective and allow a to take values in the power set of T^+_{aug}. The latter characteristic means that $a(t)$ can take two or more values in T^+_{aug} but also no value (i.e. $a(t) = \emptyset$) and is required as we may associate multiple imputed stated with a single marginalised state[3]. For instance, we may modify the vertical integration scheme given in (b) so that states at two consecutive time points $3t + 1$, $3t + 2$ are imputed with the states at the preceding time point $3t$ being integrated out. With the introduced index functions the subsequent integrated and augmented states are given by $\ldots, \mathbf{x}_{int,\tau(t-1)}, \mathbf{x}_{int,\tau(t)}, \mathbf{x}_{int,\tau(t+1)}, \ldots$ and $\ldots, \mathbf{x}_{aug,a(t-1)}, \mathbf{x}_{aug,a(t)}, \mathbf{x}_{aug,a(t+1)}, \ldots$, respectively, for $t = 1, 2, \ldots, |T^+_{int}|$. In the two examples above we have $\tau(t) = t$ and $a(t) = t$ for the horizontal integration given in (a) and $\tau(t) = 2t + 1$ and $a(t) = 2t$ for the vertical integration in (b).

Additionally, we specify a function for observations $o(t)$ with a similar role to τ and a, i.e. allowing us to iterate over the set of observation indices $\{1, \ldots, T\}$ using the same index as to iterate over T_{int} and T_{aug}. Therefore, we want the image of $o(t)$ to cover $\{1, \ldots, T\}$, the whole set of indices of y_t, which may consist of elements from both T_{int} and T_{aug}. This means that we need to be able to assign multiple indices from $\{1, \ldots, T\}$ to the iterating variable t. To this end, we allow $o(t)$ to take values in the power set of $T_{int} \cup T_{aug}$. For illustration, consider vertical integration (b) together with conditionally independent observations $y_t|\mathbf{x}_t \sim p(y_t|\mathbf{x}_t)$. For $t = 1, 2, \ldots, |T^+_{int}|$ we consider states in two different time periods, i.e. at period $\tau(t) = 2t + 1$ for the integrated states and at period $a(t) = 2t$ for the imputed states, so that for each t we need to account for two different observations, $y_{\tau(t)}$ and $y_{a(t)}$. This means that $o(t) = \{2t, 2t + 1\}$. In the case of horizontal integration (a) $T_{int} = T_{aug}$ so we simply set $o(t) = t$.

[3]Below we discuss how the marginalised states can be related to states of a first order hidden Markov model.

In order to identify conditionally independent latent states to "integrate out", one can use the graphical structure of the problem: Figure 4.1.1 can be seen as an directed acyclic graph (DAG), for which the literature on Dynamic Bayesian Networks (see Murphy, 2002) provides insights regarding the impact of conditioning on a certain node (*d-separation*). In the context of particle filters Doucet et al. (2000a) note that the "tractable structure" of some state space models might by analytically marginalised out given imputed other nodes.

Rao-Blackwellisation We note that integrating out, or "marginalising out", some of the variables is a case of the general technique known as *Rao-Blackwellisation*, which relies on the Rao-Blackwell formula. Suppose that we are interested in a function f of two random vectors z_1 and z_2, and let $\hat{f}$ be an estimator of f. Then

$$\mathrm{Var}[\hat{f}(z_1, z_2)] = \mathrm{Var}[\underbrace{\mathbb{E}[\hat{f}(z_1, z_2)|z_2]}_{=:\hat{f}'}] + \underbrace{\mathbb{E}[\mathrm{Var}[\hat{f}(z_1, z_2)|z_2]]}_{(*)},$$

which implies that $\hat{f}'$ has the same expected value as $\hat{f}$ but a lower variance than $\hat{f}$ by an additive factor of $(*)$. Rao-Blackwellisation was introduced to the MCMC literature by Gelfand and Smith (1990) in their seminal paper on the Gibbs sampler to become a commonly applied tool for variance reduction of integral approximations. In general context of sampling schemes, Rao-Blackwellisation was further analysed by Casella and Robert (1996), whose approach was then used by Douc and Robert (2011) for improving efficiency of the MH algorithm and by Doucet et al. (2000b,a) to enhance particle filters. Durbin and Koopman (2012, Ch. 12) note that in the context of state space models z_2, i.e. the variable being integrated out, is not a *sufficient statistic*, hence the term "Rao-Blackwellisation" is not fully appropriate since the the Rao-Blackwell theorem concerns the case when z_2 is a sufficient statistic for z_1. The contribution of this paper is to employ the Rao-Blackwell principle for DA in the context of state space models.

Approximate marginal likelihood Recall that the joint posterior distribution over $\boldsymbol{\theta}$ and $\mathbf{x}_{\mathbf{aug}}$ can be expressed in terms of the SCDL

$$p(\boldsymbol{\theta}, \mathbf{x}_{\mathbf{aug}}|\mathbf{y}) \propto p(\mathbf{y}, \mathbf{x}_{\mathbf{aug}}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

However, the SCDL $p(\mathbf{y}, \mathbf{x}_{\mathbf{aug}}|\boldsymbol{\theta})$ may still be analytically intractable so that we need to estimate it using simulation-based techniques. Suppose we have a sample of length

N of unknown variables of interest (i.e. $\boldsymbol{\theta}$ and $\mathbf{x_{aug}}$). Here, N is a number of points used for integration: for a deterministic integration it it the number of grid points, for a stochastic, i.e. Monte Carlo (MC), integration it is a number of draws. We can use such a sample to compute $\hat{p}_N(\mathbf{y}, \mathbf{x_{aug}}|\boldsymbol{\theta})$, the N-sample estimate of the SCDL, and consequently to approximate the posterior distribution in the following way

$$\hat{p}_N(\boldsymbol{\theta}, \mathbf{x_{aug}}|\mathbf{y}) \propto \hat{p}_N(\mathbf{y}, \mathbf{x_{aug}}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

We set $\hat{p}_N(\mathbf{y}, \mathbf{x_{aug}}|\boldsymbol{\theta})$ such that

$$\hat{p}_N(\mathbf{y}, \mathbf{x_{aug}}|\boldsymbol{\theta}) \overset{N\to\infty}{\to} p(\mathbf{y}, \mathbf{x_{aug}}|\boldsymbol{\theta}),$$

so that

$$\hat{p}_N(\boldsymbol{\theta}, \mathbf{x_{aug}}|\mathbf{y}) \overset{N\to\infty}{\to} p(\boldsymbol{\theta}, \mathbf{x_{aug}}|\mathbf{y}).$$

Further properties of the resulting likelihood estimator depend in general on the approximation scheme, which in turn determine the properties of the corresponding MCMC algorithm. If $\mathbb{E}[\hat{p}_N(\mathbf{y}, \mathbf{x_{aug}}|\boldsymbol{\theta})] = p(\mathbf{y}, \mathbf{x_{aug}}|\boldsymbol{\theta})$ standard MH algorithms converge to $p(\boldsymbol{\theta}, \mathbf{x_{aug}}|\mathbf{y})$, which follows from the pseudo-marginal argument. The pseudo-marginal theory, originated by Beaumont (2003), further developed by Andrieu and Roberts (2009) and popularised by Andrieu et al. (2010) (who called their method the particle MCMC, PMCMC), guarantees that an MCMC scheme based on the unbiased (marginal) likelihood estimator converges to the exact posterior distribution[4]. Such an unbiased likelihood estimator is delivered by e.g. MC integration, in which the integral is evaluated at random points. Hence, whether the resulting MCMC algorithm is "exact approximate" or "just approximate" depends on whether the approximate likelihood is an unbiased estimator of the marginal likelihood.

For fixed points, such as in a quadrature, obtaining of an "exact approximate" algorithm is not guaranteed but the resulting approximation converges to the true value as $N \to \infty$. It means that a "just approximate" algorithm can be made arbitrarily close to the true integral by considering sufficiently many samples to construct the estimator. Additionally, we note that unbiased estimators might be characterised by large MC errors, particularly for a small number of samples, see e.g. Korattikara et al. (2014), Jacob and Thiery (2015). The choice between different likelihood approximation methods fits into the traditional discussion on the bias-variance trade-off. However, as pointed out by Robert (2016), especially from a Bayesian perspective unbiasedness

[4]PMCMC algorithms are thus called "exact approximate". Note that they are the extreme case of our approach with $\mathbf{x}_{int} = \mathbf{x}$.

is a "second order property"[5].

4.3 Approximations for MCMC sampling

Below we consider possible ways for obtaining an estimate $\hat{p}_N(\mathbf{y}, \mathbf{x_{aug}}|\boldsymbol{\theta})$. In particular, we focus on the case when it can be obtained as a product of one dimensional integrals. This assumption is less restrictive than it may appear at first: the choice of the auxiliary variables can often be made such that this condition is satisfied. There exist several methods to numerically estimate a single one dimensional integral including: (1) quadrature with fixed nodes; (2) quadrature with adaptive nodes; (3) stochastic (MC) integration. The two former approaches can be seen as "binning" of similar values of the integrated state vector within specified ranges ("bins"), which can then be interpreted as states of a (finite-dimensional) first-order HMM. In the context of bins of equal widths such an approach has been successfully applied e.g. by Langrock et al. (2012a,b); Langrock and King (2013). For the latter MC approach the resulting estimator of the complete data likelihood is unbiased $\mathbb{E}(\hat{p}_N(\mathbf{y}, \mathbf{x_{aug}}|\boldsymbol{\theta})) = p(\mathbf{y}, \mathbf{x_{aug}}|\boldsymbol{\theta})$. Hence, the pseudo-marginal argument guarantees that the chain generated with a standard MH algorithm (using the estimate $\hat{p}_N(\mathbf{y}, \mathbf{x_{aug}}|\boldsymbol{\theta})$) converging to the exact posterior distribution $p(\boldsymbol{\theta}, \mathbf{x_{aug}}|\mathbf{y})$ in this case; see Beaumont (2003), Andrieu and Roberts (2009) and Andrieu et al. (2010).

We note that in low dimensions all of these methods are feasible, however we focus on methods based on the two former approaches as they provide an intuitive interpretation in terms of state transition probabilities and conditional (augmented) observation distributions. There are two cases when such an approximation might be particularly useful. First, when the state vector is discrete but of a large size grouping of its elements into "bins" helps to reduce the size of the problem. Second, for continuous states any form of numerical integration basically reduces to splitting of the state space into "bins", which can then be further combined into larger groups to increase the efficiency of an algorithm.

[5]There are two obvious reasons for that. First, the concept of a bias is conditional on the true value of a parameter, which is unknown (Gelman, 2011). Second, unbiasedness cannot be achieved for most transformations of the model parameter vector and is not preserved under reparameterisation (Robert, 2016).

4.3.1 Approximation bins as hidden Markov model states

Below we discuss ways to specify the bins, or quadrature points: a deterministic one, with bins of a fixed size (but varying probability of occurring), and a stochastic one, with bins of a fixed probability (but varying size). To simplify the exposition, we assume that $\mathbf{x}_{int,\tau(t)}$ is univariate so we can write $x_{int,\tau(t)}$. For multivariate $\mathbf{x}_{int,\tau(t)}$ we may consider separate bins for each integrated state dimension $d \in D_{int}$ at time $\tau(t)$. We then interpret the bins as states of a latent (first-order) Markov process, which allows us to give the resulting integration/augmentation scheme an HMM embedding[6].

Fixed bins A straightforward approach to binning is via bins of a fixed size as it relates to a deterministic approximation of the likelihood with a quadrature and allows for a natural HMM interpretation. Discretising of the state space to perform numerical integration dates back to Kitagawa (1987) and was discussed in Zucchini et al. (2016). The associated approximate posterior distribution can be made arbitrarily accurate by increasing the number of bins (quadrature points).

The idea is to split the state space $\mathcal{X}_{int}$ of the state to be integrated out into B bins of length k (for integer-valued variables we assume $k \in \mathbb{N}$) and to consider e.g. the midpoints of the bins for integration. Then the values that fall in a given bin are approximated by the value of the midpoint of that bin. Such an approach is used by Langrock et al. (2012b) to efficiently approximate the likelihood for stochastic volatility models (with continuous bins) in a classical framework.

For infinitely dimensional states, either discrete or continuous, an "allowed integration range" needs to be specified. For instance, for a normal variable this means setting a lower and an upper bound for the integration b_0 and b_B, while for a Poisson variable only of an upper bound b_B since $b_0 = 0$ in this case. We divide the resulting domain into intervals as follows:

$$\underbrace{[b_0, \ldots, b_1)}_{\mathcal{B}_1,\text{ bin } 1}, \underbrace{[b_1, \ldots, b_2)}_{\mathcal{B}_2,\text{ bin } 2}, \ldots, \underbrace{[b_{j-1}, \ldots, b_j)}_{\mathcal{B}_j,\text{ bin } j}, \ldots, \underbrace{[b_{B-1}, \cdots, b_B)}_{\mathcal{B}_B,\text{ bin } B},$$
$$b_i - b_{i-1} = k, \qquad i = 1, \ldots, B.$$

For continuous variables $\mathcal{B}_i$ is simply a continuous interval of length k, while for discrete

[6]We note that from the perspective of the original process $\{\mathbf{x}\}$ the process we want to integrate out $\{\mathbf{x_{int}}\}$ will not be a Markov chain due to its potential dependence on the imputed states $\{\mathbf{x_{aug}}\}$. However, since we know the latter, conditioning on them can be understood as adopting a time-varying transition probabilities for $\{\mathbf{x_{int}}\}$, parametrised with relevant $\{\mathbf{x_{aug}}\}$.

variables it consists of k subsequent integers, e.g. for a Poisson variable we have $\mathcal{B}_i = \{ik, \dots, (i+1)k\}$. We specify the midpoints of the bins as $b_i^* = \frac{b_{i-1}+b_i}{2}$ (for integer-valued variables rounding is required for even k).

We then define $\{z_t\}$, $t \in 1, \dots, T^*$, as a B-state, discrete-time (not necessarily homogeneous) Markov chain[7] with transition probabilities $\gamma_{jk,t} = \mathbb{P}(z_t = k | z_{t-1} = j)$ defined as

$$\gamma_{jk,t} := \mathbb{P}(x_{int,\tau(t)} \in \mathcal{B}_k | x_{int,\tau(t-1)} \in \mathcal{B}_j, \mathbf{x}_{aug,a(t-1)}).$$

Then a transition of $z_{t-1} = j$ to $z_t = k$ is equivalent to $x_{int,\tau(t)}$ "falling into" bin k given $x_{int,\tau(t-1)}$ was in bin j and given $\mathbf{x}_{aug}$. For computationally intensive probabilities we can further approximate these as $\tilde{\gamma}^*_{jk,t} := p(b_k^* | b_j^*, \mathbf{x}_{aug,a(t-1)})$, which for discrete variables means $\mathbb{P}(x_{int,\tau(t)} = b_k^* | x_{int,\tau(t-1)} = b_j^*, \mathbf{x}_{aug,a(t-1)})$. To get the valid probability values (i.e. summing up to one) we normalise the transition probabilities as $\gamma^*_{jk,t} := \tilde{\gamma}^*_{jk,t} / \sum_{c=1}^{B} \tilde{\gamma}^*_{jc,t}$. Notice that this corresponds to treating the values in a bin uniformly. We can alternatively compute the transition probabilities between bins directly, by integrating with respect to the required ranges as follows

$$\mathbb{P}(x_{int,\tau(t)} \in \mathcal{B}_k | x_{int,\tau(t-1)} \in \mathcal{B}_j, \mathbf{x}_{aug,a(t-1)}) \propto \int_{\mathcal{B}_k} \int_{\mathcal{B}_j} p(x_{int,\tau(t)} | x_{int,\tau(t-1)}, \mathbf{x}_{aug,a(t-1)}) dx_{int,\tau(t-1)} dx_{int,\tau(t)}.$$

However, typically such an analytical integration will only be possible in simple cases, e.g. discrete variables. One can visualise this method by considering small squares of a bigger transition matrix instead of each of its element separately.

Adaptive bins An alternative approach to fixed width binning is to use adaptive intervals which do not require any limiting of the integration range. This can be done by transforming the variable of interest to the $[0, 1]$ range by applying a cdf. Then the bins can be specified on the $[0, 1]$ interval and their limits or midpoints can be transformed back to obtain the values needed for the approximation of the original variable of interest. In particular, quantiles of the distribution associated with the variable of interest can be used. Then instead of specifying the grid points we fix the probabilities for each bin, which previously needed to be determined. This means a

[7] Even though we hardly refer to $\{z_t\}$ explicitly later in the text, they are useful to understand the introduced construction relating the potentially continuously valued process of interest $x_{int,\tau(t)}$ to a finite state HMM z_t. Such an exposition is inspired by Langrock et al. (2012b, Section 2.2).

quantile determination problem which are needed e.g. to obtain the midpoint values used in conditioning.

Suppose that $x_{int,\tau(t)} \sim p(\vartheta_{\tau(t)})$, $\tau(t) \in T_{int}$, where $\vartheta_{\tau(t)}$ is a vector of possibly time varying parameters, with the corresponding cdf $F(\vartheta_{\tau(t)})$. Consider a vector of $B+1$ quantiles $\mathbf{q} = [q_0, q_2, \ldots, q_B]$. The corresponding B mid-quantiles, denoted $\mathbf{q}^* = [q_1^*, q_2^*, \ldots, q_B^*]$, are given by $q_i^* = \frac{q_{i-1}+q_i}{2}$ (for instance, $\mathbf{q} = [0.0, 0.1, 0.2, \ldots, 1.0]$ and $\mathbf{q}^* = [0.05, 0.15, \ldots, 0.95]$). For $F(\vartheta_t)$ continuous and strictly monotonically increasing (such as a normal cdf) the bin midpoints at time t are determined by the mid-quantiles as follows

$$b_i^* = F^{-1}\left(q_i^* | \vartheta_{\tau(t)}\right).$$

For discrete variables one can either use the generalized inverse distribution function, or use a continuous approximation to the associated discrete distribution. For instance for a Poisson variable with a large enough mean, the normal approximation could be adopted. We note in general, the adaptive approach can be easily implemented in any programming language or software for statistical computing.

4.3.2 Hidden Markov model likelihood

Having specified the states of the underlying Markov chain in Section 4.3.1, we aim to use them to approximate the joint SCDL (4.2.1) by embedding it into an HMM form (below, to ease the notation, we skip $\boldsymbol{\theta}$ in conditioning). We relate each state of the hidden Markov process with the relevant augmented states and observations. This imposes a time structure on the SCDL integral (4.4.16) with respect to the "integration time" and thus allows us to cast it into a likelihood of an HMM. Note that without any form of DA the likelihood can be simply decomposed by making use of the Markov property of the original state process $\mathbf{x}_t$, i.e. $p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x})d\mathbf{x} = \int \mathbf{x}_0 \prod_{t=1}^{T} p(y_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{x}_{t-1})d\mathbf{x}_0 d\mathbf{x}_1 \ldots d\mathbf{x}_T$.

Motivating example For illustration, consider the state specification from Figure 4.2.1a to which we add conditionally independent observations to result in an SSM presented in Figure 4.3.1. Such a system is representative for e.g. dynamic factor models (linear or nonlinear), with y_t multivariate, broadly applied in macroeconometrics and finance; it was also used by e.g. Abadi et al. (2010) to model population dynamics of little owls.

We specify $\mathbf{x_{aug}} = \{x_{1,t}\}_{t=0}^{T} =: \mathbf{x_1}$ (state 1) and $\mathbf{x_{int}} = \{x_{2,t}\}_{t=0}^{T} =: \mathbf{x_2}$ (state 2), which corresponds to the "horizontal" integration. Hence we put $T_{int} = T_{aug} = \{0, 1, \ldots, T\}$,

$\tau(t) = t$, $a(t) = t$ and $o(t) = t$. We denote $T^* = |T_{int}^+|$. Using the temporal dependence in this system, the SCDL $p(\mathbf{y}, \mathbf{x_{aug}})$ can be expressed as

$$\begin{aligned} p(\mathbf{y}, \mathbf{x_{aug}}) &= p(x_{1,0}) \prod_{t=1}^{T} p(y_t|x_{1,t}) p(x_{1,t}|x_{1,t-1}), \\ &= p(x_{1,0}) \prod_{t=1}^{T^*} p(y_{o(t)}|x_{1,a(t)}) p(x_{1,a(t)}|x_{1,a(t-1)}), \end{aligned}$$

which is not tractable without integrating out $\mathbf{x_2}$. Hence, we marginalise over $\mathbf{x_2}$ and aim at approximating the resulting integral using a quadrature based on B bins $\mathcal{B}_k$, $k = 1, \ldots, B$, as follows

$$\begin{aligned} p(\mathbf{y}, \mathbf{x_{aug}}) = \int \cdots \int p(x_{1,0}) p(x_{2,0}) \prod_{t=1}^{T^*} & p(y_{o(t)}|x_{1,a(t)}, x_{2,\tau(t)}) p(x_{1,a(t)}|x_{1,a(t-1)}, x_{2,\tau(t-1)}) \\ & p(x_{2,\tau(t)}|x_{1,a(t-1)}, x_{2,\tau(t-1)}) dx_{2,\tau(T^*)} \ldots dx_{2,\tau(1)} \\ \approx \sum_{k_1=1}^{B} \cdots \sum_{k_{T^*}=1}^{B} p(x_{1,0}) p(x_{2,0}) \prod_{t=1}^{T^*} & p(y_{o(t)}|x_{1,a(t)}, x_{2,\tau(t)} \in \mathcal{B}_{k_t}) \\ & p(x_{1,a(t)}|x_{1,a(t-1)}, x_{2,\tau(t-1)} \in \mathcal{B}_{k_{t-1}}) \\ & p(x_{2,\tau(t)} \in \mathcal{B}_{k_t}|x_{1,a(t-1)}, x_{2,\tau(t-1)} \in \mathcal{B}_{k_{t-1}}). \end{aligned} \tag{4.3.1}$$

The above approximation has a natural interpretation in terms of HMM by associating the events $x_{2,\tau(t)} \in \mathcal{B}_k$ with states of a hidden Markov process on B states. The transition matrix of this process is

$$\Gamma_t = \left[\mathbb{P}(x_{2,\tau(t)} \in \mathcal{B}_k^* | x_{1,a(t-1)}, x_{2,\tau(t-1)} \in \mathcal{B}_l^*) \right]_{k,l=1,\ldots,B}, \tag{4.3.2}$$

for $t \in 1, 2, \ldots, T^*$. Next to the transition matrix, we need to specify two matrices for the "augmented data": one for the augmented states $\mathbf{x}_{aug}$ and one for the real observations $\mathbf{y}$. This is different compared to standard HMMs in which only the latter is used. We specify the likelihood matrices for the augmented states and the observation as follows

$$P_t = \text{diag}\left(p(x_{1,a(t)}|x_{1,a(t-1)}, x_{2,\tau(t-1)} \in \mathcal{B}_l^*) \right)_{l=1,\ldots,B}, \tag{4.3.3}$$

$$Q_t = \text{diag}\left(p(y_{o(t)}|x_{1,a(t)}, x_{2,\tau(t)} \in \mathcal{B}_k^*) \right)_{k=1,\ldots,B}. \tag{4.3.4}$$

Notice that for both the integrated and augmented states the conditioning is with respect to their previous realisations, whilst for the observations it is with respect to the current values of both states. The quadrature based approximation to the SCDL (4.3.1) can be then approximated as

$$\hat{p}_B(\mathbf{y}, \mathbf{x_{aug}}) = p(x_{1,0})\mathbf{u}_0 \left(\prod_{t=1}^{T^*} P_{a(t)} \Gamma_{\tau(t)} Q_{\tau(t)} \right) \mathbf{1}, \tag{4.3.5}$$

where $\mathbf{u}_0 = \Big[\mathbb{P}(x_{2,0} \in \mathcal{B}_1) \quad \dots \quad \mathbb{P}(x_{2,0} \in \mathcal{B}_B) \Big]$ is the initial distribution of the Markov chain. Appendix 4.A.1 presents the underlying SSM and the details of the derivations.

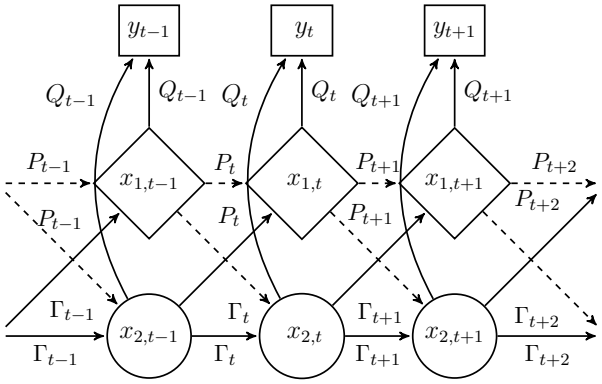


Figure 4.3.1: Illustration of combining DA and HMM structure. Conditionally independent observations added to the state specification from Figure 4.2.1a. Diamonds represent the imputed states, circles – the integrated states. Dashed lines used for the relations *from* the imputed (known) states.

General formulation The generic matrices of the HMM-based approximation have the form

$$\begin{aligned}
\Gamma_t &= \Big[\mathbb{P}(x_{int,\tau(t)} \in \mathcal{B}_k | x_{int,\tau(t-1)} \in \mathcal{B}_l, x_{aug,a(t-1)}) \Big]_{k,l=1,\dots,B}, \\
P_t &= \operatorname{diag}\left(p(x_{aug,a(t)} | x_{int,\tau(t-1)} \in \mathcal{B}_l) \right)_{l=1,\dots,B} \\
Q_t &= \operatorname{diag}\left(p(\mathbf{y}_{o(t)} | x_{int,\tau(t)} \in \mathcal{B}_k, \mathbf{x}_{aug,a(t)}) \right)_{k=1,\dots,B}
\end{aligned}$$

for $t \in 1, 2, \dots, T^*$ and lead to the following form of the HMM approximation

$$\hat{p}_B(\mathbf{y}, \mathbf{x_{aug}}) = p(x_{1,0})\mathbf{u}_0 Q_0 \left(\prod_{t=1}^{T^*} P_{a(t)} \Gamma_{\tau(t)} Q_{\tau(t)} \right) \mathbf{1}, \tag{4.3.6}$$

which differs from (4.3.5) by including $Q_0 := \text{diag}\left(p(y_{o(0)}|x_{int,\tau(0)} \in \mathcal{B}_k^*)\right)_{k=1,\dots,B}$, which allows for a dependence of some observations on the initial state of the Markov process[8]. We require $\tau(t) \geq \max\{a(t)\}$ and $o(t) \subset \tau(t) \cup a(t)$, which is natural given the real dependencies in the original SSM (4.1.1)–(4.1.3).

4.4 Applications

In this section we consider applications of the proposed SCDA method and assess their performance. We consider two case studies with distinctively different features resulting in different integration schemes. The first application involves the dataset on the Northern lapwing (*Vanellus vanellus*), which has been extensively analysed in statistical ecology, see Besbeas et al. (2002), Brooks et al. (2004), King et al. (2008), and the references therein. We adopt the integrated population modelling approach of Besbeas et al. (2002), to be explained below, however our main focus is on modelling the abundance of the species via a state space model with discrete states. The second application relates to the well-known stochastic volatility model (SV), which is a popular tool to model time-varying volatility especially for financial time series, see Taylor (1994), Ghysels et al. (1996) or Shephard (1996). Further, we demonstrate how the SCDA framework can be easily adjusted to accommodate more complex properties of financial time series such as SV in the mean of Koopman and Uspensky (2002) or leverage effects, see Jungbacker and Koopman (2007).

Algorithm tuning In each case study we are interested in comparing the performance of the standard DA approach with that of the SCDA. To guarantee the between-method comparability, for each method we perform the estimation using a "vanilla" RW-MH (single-update) algorithm. We tune each sampler so that the acceptance rates for each element of the parameter vector θ and the average acceptance rates for each of the imputed states are "reasonable", i.e. between $20 - 40\%$.

Such a range corresponds to the seminal results of Gelman et al. (1996) and Roberts and Rosenthal (2001). The former authors prove that the asymptotically (as the dimension of the state space diverges to infinity) optimal mean acceptance rate is equal to 0.234 for a target distribution consisting of i.i.d. components and a normal proposal distribution of the same dimension as the target. Hence, they do not consider single-state updates (i.e. one-dimensional increments), for which the later authors deliver the optimal

[8]The SV model example in Appendix 4.C demonstrates the role of Q_0.

acceptance rate of 0.44 for a normal proposal distribution (see also Rosenthal, 2011). Generally, a mean acceptance rate of 20–40% is considered to deliver a well-performing chain.

Effective sample size Since the samples generated by MCMC algorithms are not independent, standard convergence results for independent MC sampling do not apply; in particular, the standard variance estimator (i.e. the sample empirical variance) cannot be used to measure the variance of the empirical average delivered by an MCMC algorithm. The stochastic dependence in the (stationary) Markov chain $X_1, X_2, \ldots$ results in the associated asymptotic variance σ^2_{MCMC} taking account of the covariance in the Markov chain

$$\begin{aligned}\sigma^2_{\text{MCMC}} &= \text{Var}[X_i] + 2\sum_{k=1}^{\infty} \text{Cov}[X_i, X_{i+k}] \\ &= \text{Var}[X_i] \underbrace{\left(1 + 2\sum_{k=1}^{\infty} \rho(k)\right)}_{\text{IF}},\end{aligned} \tag{4.4.1}$$

where $\rho(k)$ is the kth order serial correlation (Geyer, 2011). The term in the parentheses in (4.4.1) is referred to as the *autocorrelation function* (Geyer, 2011), *(integrated) autocorrelation time* (Robert and Casella, 2004, Ch. 12.3.5) or *inefficiency factor* (IF, Pitt et al., 2012), which is the name we use. High values of autocorrelation, typically reported for MCMC sampling, lead to the standard variance estimator underestimating the true variance σ^2_{MCMC}.

A common measure for assessing the deterioration in the sampling efficiency due to the draws autocorrelation is the *effective sample size* (ESS) defined as

$$\text{ESS} = \frac{M}{IF}, \tag{4.4.2}$$

where M is the sample size (Robert and Casella, 2004, Ch. 12.3.5). It indicates what the size of i.i.d. sample would be, had it the same variance as the MCMC sample. Equivalently, the IF gives the factor by which the "nominal" MCMC sample size would need to be increased in order to achieve the same accuracy as an i.i.d. sampling.

In practice, one typically cannot compute the IF directly and needs to estimate it instead. As noted by Robert and Casella (2004, Ch. 12.3.5) estimation of IF is a "delicate issue", as it contains an infinite sum. A possible solution to this problem

is to set a cut-off value K for the autocorrelation terms being summed up: $\widehat{IF} = 1+2\sum_{k=1}^{K}\hat{\rho}(k)$. The choice of K poses the risk of subjectiveness; setting K to the lowest lag at which $\hat{\rho}(k)$ become insignificant seems to be a reasonable solution suggested by e.g. Kass et al. (1998) or Pitt et al. (2012) and this is the approach we take here.

4.4.1 Ecological model: lapwings data

We consider a time series of observations relating to census data (abundance index) of adult British lapwings (*Vanellus vanellus*), which we denote by $\mathbf{y} = (y_1, \ldots, y_T)$. The lapwings dataset plays an important role in statistical ecology and has served as an illustration in several handbooks (see King, 2011; King et al., 2010) and papers (e.g. Besbeas et al., 2002) in this field. It was also used as an example of a complex statistical model by e.g. Goudie et al. (2018). We provide the details of this dataset in Appendix 4.B. Figure 4.4.1 presents the data on the index of lapwings as well as on the normalised frost days, used as a covariate to describe the survival process. The latter is based on the number of days below freezing between April of year t and March of year $t+1$, inclusive and is a proxy for harshness of winter, which can affect the survival probability of wild birds more by lengthy cold periods rather than by low average temperature.

Figure 4.4.1: Lapwings census data and normalised frost days.

The counts are only estimates of the true unknown population size, which is assumed to change over time according to a first order Markov process. The latent population is related to two times series: for first-years and adults, which we denote $\mathbf{N_1} = (N_{1,1}, \ldots, N_{1,T})$ and $\mathbf{N_a} = (N_{a,1}, \ldots, N_{a,T})$, respectively. Hence, the latent state is given by $\mathbf{x} = \{\mathbf{N_1}, \mathbf{N_a}\}$. Following Besbeas et al. (2002) we model the count data

via the following state space model:

$$y_t|N_{a,t},\boldsymbol{\theta} \sim \mathcal{N}(N_{a,t},\sigma_y^2), \tag{4.4.3}$$
$$N_{1,t+1}|N_{a,t},\boldsymbol{\theta} \sim \mathcal{P}(N_{a,t}\rho_t\phi_{1,t}), \tag{4.4.4}$$
$$N_{a,t+1}|N_{1,t},N_{a,t},\boldsymbol{\theta} \sim \mathcal{B}\big((N_{1,t}+N_{a,t}),\phi_{a,t}\big), \tag{4.4.5}$$
$$N_{1,0} \sim \mathcal{NB}(r_{1,0},p_{1,0}), \tag{4.4.6}$$
$$N_{a,0} \sim \mathcal{NB}(r_{a,0},p_{a,0}), \tag{4.4.7}$$

for $t = 1,\ldots,T$, where $\mathcal{N}$, $\mathcal{P}$, $\mathcal{B}$ and $\mathcal{NB}$ stand for normal, Poisson, binomial and negative binomial distributions, respectively. The model is parametrised by the time-varying productivity rate ρ_t, and time-varying survival rates $\phi_{1,t}$ and $\phi_{a,t}$, for first-years and adults, respectively, while $a_{i,0}$ and $p_{i,0}$ are hyperparameters of the prior distribution on the initial state value $N_{i,0}$, $i \in \{1,a\}$.

Following Besbeas et al. (2002), we assume the following functional forms for the model time varying parameters

$$\begin{aligned}
\operatorname{logit}\phi_{1,t} &= \log\left(\frac{\phi_{1,t}}{1-\phi_{1,t}}\right) = \alpha_1 + \beta_1 f_t,\\
\operatorname{logit}\phi_{a,t} &= \log\left(\frac{\phi_{a,t}}{1-\phi_{a,t}}\right) = \alpha_a + \beta_a f_t,\\
\log\rho_t &= \alpha_\rho + \beta_\rho \tilde{t},
\end{aligned}$$

where f_t denotes the normalised value of frost days $fdays$ in year t and $\tilde{t}$ the normalised time index. As explained by King (2011), we introduce normalisation of f_t and $\tilde{t}$ to improve the mixing of the Markov chain and to facilitate the interpretation of the regression parameters.

To improve the estimation, Besbeas et al. (2002) propose using an additional source of information provided by the ring-recovery (RR) data, independent from the count series. The RR model shares with the SSM the survival parameters $\phi_{1,t}$ and $\phi_{a,t}$ but it does not involve the productivity rate ρ_t. Instead, the RR models includes the common time-varying recovery rate λ_t (denoting the probability that a bird which dies in year t is recovered), specified to be of the form

$$\operatorname{logit}\lambda_t = \log\left(\frac{\lambda_t}{1-\lambda_t}\right) = \alpha_\lambda + \beta_\lambda \tilde{t}_t.$$

Combining both models results in the so-called *integrated model*, which is parametrised

by the regression parameters and the variance of the observation error. We refer to Besbeas et al. (2002) for a more detailed description of the integrated model. The model parameters are collected in a vector $\boldsymbol{\theta} = (\alpha_1, \alpha_a, \alpha_\rho, \alpha_\lambda, \beta_1, \beta_a, \beta_\rho, \beta_\lambda, \sigma_y^2)^T$.

Finally, to complete the Bayesian specification of the model, we set independent vague priors being normal $\mathcal{N}(0, 100)$ for the logistic regression coefficient α_i and β_i, $i \in \{1, a, \rho, \lambda\}$, while for the observation variance σ_y^2 a conjugate inverse gamma $\Gamma^{-1}(a_y, b_y)$ with $a_y = 0.001 = b_y$ is used. For the initial states, we set the following values for the hyperparameters: for first-years $r_{1,0} = 4$ and $p_{1,0} = 0.98$ so that the prior mean and variance of 1-years are roughly 200 and $10,000$, respectively; for adults $r_{a,0} = 111$ and $p_{a,0} = 0.9$, so that the prior mean and variance adults are roughly $1,000$ and $10,000$, respectively.

System (4.4.3)–(4.4.7) is non-Gaussian and nonlinear with the associated likelihood unavailable in a closed form. It could be analysed using the normal approximation, which has an advantage that the Kalman filtering and smoothing techniques can be employed, see Besbeas et al. (2002). However, we aim at estimation of the original model, in which case the standard approach has been a DA approach. The problem with the standard DA approach is that it may lead to poorly mixing MCMC algorithms as demonstrated by King (2011). To this end, we first analyse the dependence structure in the model to select most promising states to integrate out.

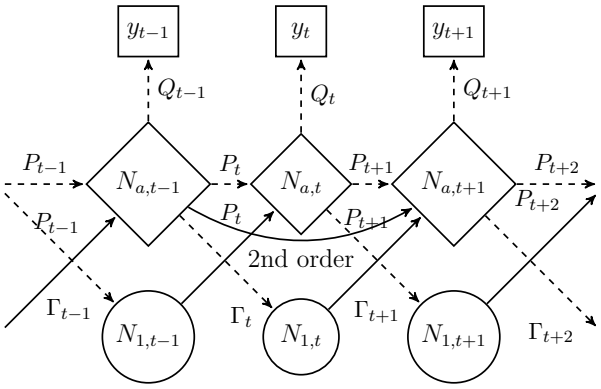


Figure 4.4.2: Lapwings data: combining DA and HMM structure. Diamonds represent the imputed nodes, squares – the data, circles – the unknown variables. Integrating out $\mathbf{N_1}$ leads to a second order HMM on $\mathbf{N_a}$. Dashed lines used for the relations *from* the imputed (known) states.

The two-dimensional state $[N_1, N_a]_{t=1}^T$ follows the first-order Markov process with a non-trivial transition kernel. We can notice that first-year birds in t only feed into adults in $t+1$. However, adults in t contribute to both the number of first-years and adults in $t+1$, as well as the observed estimate y_t. This suggests that reducing the strength of the dependence structure can be obtained by integrating out $\mathbf{N_1}$, while

imputing $\mathbf{N_a}$. This corresponds to the *horizontal* integration scheme with $\mathbf{x_{int}} = \mathbf{N_1}$ and $\mathbf{x_{aug}} = \mathbf{N_a}$. The resulting modified dependence structure is presented in Figure 4.4.2. Marginalising over $\mathbf{N_1}$ allows us to simplify the analysis as we only need to consider $\mathbf{N_a}$ which now follows a second-order Markov process. A similar second order structure in this context has also been noted by Besbeas and Morgan (2018).

Hidden Markov Model approximation The resulting SCDL for the augmented data set $(\mathbf{y}^T, \mathbf{N_a}^T)^T$ is given by

$$p(\mathbf{y}, \mathbf{N_a}|\boldsymbol{\theta}) = p(\mathbf{y}|\mathbf{N_a}, \boldsymbol{\theta})p(\mathbf{N_a}|\boldsymbol{\theta}), \tag{4.4.8}$$

which is still intractable. Hence, we employ an HMM-based approximation to (4.4.8) discussed in Section 4.3. Since $N_{1,t}$ follows a Poisson distribution, we only need to specify a truncation value N^* for the maximum population size for first-years for a fixed bin approach (i.e. we set $b_B = N^*$, with b_0 naturally being equal to 0). Since the observations $\mathbf{y}$ are conditionally independent from $\mathbf{N_1}$ given $\mathbf{N_a}$, integrating out of $\mathbf{N_1}$ can be done only for the second term on the right hand side (4.4.8), to obtain the marginal pmf for $\mathbf{N_a}$. Below, to ease the notation, we omit $\boldsymbol{\theta}$ in the conditioning. The marginal pmf of $\mathbf{N_a}$ is given by

$$\begin{aligned} p(\mathbf{N_a}) =& p(N_{a,0}, N_{a,1}, \ldots, N_{a,T}) \\ =& \sum_{\mathbf{N_1}} p(N_{a,0}), p(N_{1,0})p(N_{1,1}|N_{a,0}) \\ & \times p(N_{a,1}|N_{a,0}, N_{1,0}) \ldots p(N_{1,T}|N_{a,T-1})\, p(N_{a,T}|N_{a,T-1}, N_{1,T-1}) \end{aligned} \tag{4.4.9}$$

and we want next to approximate the elements of this multiple sum. To simplify the exposition below we consider the "exact" approximation with the bin size equal to one, which is possible due to the discreet nature of the integrated state, and in which the approximation error is only due to N^*, the upper limit of the allowed integration range. A typical element of the sum in (4.4.9) can be approximated as (for $t \geq 2$)

$$\begin{aligned} p(N_{a,t}|\mathbf{N}_{a,0:t-1}) &= \sum_{k=0}^{N^*} \mathbb{P}(N_{1,t-1} = k|\mathbf{N}_{a,0:t-1}) p(N_{a,t}\mathbf{N}_{a,0:t-1}, N_{1,t-1} = k), \\ &= \sum_{k=0}^{N^*} \underbrace{\mathbb{P}(N_{1,t-1} = k|N_{a,t-2})}_{=:u_{k,t-1}} \underbrace{p(N_{a,t}|N_{a,t-1}, N_{1,t-1} = k)}_{=:p_{k,t}}. \end{aligned} \tag{4.4.10}$$

In (4.4.10) $p_{k,t}$ denotes the conditional pmf of $N_{a,t}$ given $N_{1,t-1} = k$ and $N_{a,t-1}$ for which

$$p_{k,t} = p(N_{a,t}|N_{a,t-1}, N_{1,t-1} = k) \equiv \mathcal{B}\big((N_{a,t-1} + k), \phi_{a,t-1}\big).$$

Further, $u_{k,t}$ denotes the "quasi-unconditional"[9] probability of $N_{1,t} = k$. These unconditional probabilities of the hidden states can be derived as

$$\begin{aligned} u_{k,t} &= \mathbb{P}(N_{1,t} = k|N_{a,t-1}) \\ &= \sum_{l=0}^{N^*} \mathbb{P}(N_{1,t-1} = l|\mathbf{N}_{a,0:t-1})\mathbb{P}(N_{1,t} = k|N_{1,t-1} = l, \mathbf{N}_{a,0:t-1}) \\ &= \sum_{l=0}^{N^*} \underbrace{\mathbb{P}(N_{1,t-1} = l|N_{a,t-2})}_{=u_{l,t-1}} \underbrace{\mathbb{P}(N_{1,t} = k|N_{a,t-1})}_{=:\gamma_{lk,t}}, \end{aligned}$$

which we collect in a vector $\mathbf{u}_t = \Big[u_{k,t}\Big]_{k=1}^{N^*}$. In general, the unconditional probabilities of an HMM are related to each other via the transition probabilities $\gamma_{lk,t}$ (i.e. conditional probabilities) as $\mathbf{u}_t = \mathbf{u}_{t-1}\Gamma_t$ with $\Gamma_t = \Big[\gamma_{lk,t}\Big]_{l,k=1}^{N^*}$. Here we have $\gamma_{lk,t} = \mathbb{P}(N_{1,t} = k|N_{1,t-1} = l, \mathbf{N}_{a,0:t-1})$, but since in the model $N_{1,t}$'s are mutually independent given $N_{a,t-1}$ we can simplify the transition probabilities to

$$\gamma_{lk,t} = \mathbb{P}(N_{1,t} = k|N_{a,t-1}) \equiv \mathcal{P}(N_{a,t}\rho_t\phi_{1,t})$$

for $k = 0, \ldots, N^* - 1$, while for $k = N^*$ we need $\gamma_{lk,t} = 1 - \sum_{j=0}^{N^*-1} \gamma_{lj,t}$ to ensure a valid probability distribution. This means that the time varying state transition matrix Γ_t takes a simple form

$$\Gamma_t = \begin{bmatrix} \gamma_{1,t} & \ldots & \gamma_{N^*-1,t} & \gamma_{N^*,t} \end{bmatrix},$$

i.e. with each column equal to $\gamma_{k,t} = \gamma_{lk,t}\mathbf{1}$.

[9]By "quasi-unconditional" we mean unconditional in the sense of the Markov structure, i.e. previous latent states $N_{1,t-1}, N_{1,t-2}, dots$, but not in terms of the imputed values $\mathbf{N_a}$, which we treat as known, and the parameter vector $\boldsymbol{\theta}$ (see Zucchini et al., 2016, p.16, 32).

Finally, we can conveniently express (4.4.10) using matrix notation as

$$p(N_{a,t}|\mathbf{N}_{a,0:t-1}) = \underbrace{\begin{bmatrix} \gamma_{11,t-1} & \dots & \gamma_{1N^*,t-1} \\ \vdots & \ddots & \vdots \\ \gamma_{11,t-1} & \dots & \gamma_{1N^*,t-1} \end{bmatrix}}_{=\Gamma_{t-1}} \underbrace{\begin{bmatrix} p_{1,t} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & p_{N^*,t} \end{bmatrix}}_{=:P_t} \underbrace{\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}}_{\mathbf{1}} = \Gamma_{t-1} P_t \mathbf{1}.$$

Combining (4.4.9) and (4.4.10) yields the HMM form for the joint pmf of the imputed states

$$p(\mathbf{N_a}) = \mathbf{u}_0 p(N_{a,0}) \left(\prod_{t=1}^{T} P_t \Gamma_t \right) \mathbf{1},$$

where $\mathbf{u}_0 = \begin{bmatrix} p(N_{1,0}) = 0 & \dots & p(N_{1,0}) = N^* \end{bmatrix}^T$ is the initial state distribution.

As stated above, the real observations y_t, conditionally on $N_{a,t}$, are independent of $N_{1,t}$ so that the observation matrix becomes a scaled identity matrix

$$Q_t = p(y_t|N_{a,t})\mathbb{I} = \mathcal{N}(y_t|N_{a,t}, \sigma_y^2)\mathbb{I}.$$

Finally, the approximation to the SCDL (4.4.8) can be expressed as

$$p(\mathbf{y}, \mathbf{N_a}|\boldsymbol{\theta}) = p(\mathbf{y}|\mathbf{N_a}) p(\mathbf{N_a}) = \mathbf{u}_0 p(N_{a,0}) \left(\prod_{t=1}^{T} P_t \Gamma_t Q_t \right) \mathbf{1}.$$

State acceptance rate Let the current state of a Markov chain be $\mathbf{N_a^{(j)}} = \{N_{a,t}^{(j)}\}_{t=1}^{T}$ and consider updating of its t'th component. Let the proposed value be $N_{a,t}^{(\bullet)}$, with $\mathbf{N_a^{(\bullet)}} = \{N_{a,1}^{(j)}, \dots, N_{a,t-1}^{(j)}, N_{a,t}^{(\bullet)}, N_{a,t+1}^{(j)}, \dots, N_{a,T}^{(j)}\}$. The move is accepted with the probability $1 \wedge a(\mathbf{N_a^{(j)}}, \mathbf{N_a^{(\bullet)}})$, where $a(\mathbf{N_a^{(j)}}, \mathbf{N_a^{(\bullet)}})$ is the acceptance rate. Since we use a single update MH-RW with a symmetric (uniform) proposal distribution, the proposal terms required for the ratio cancel in the acceptance rate, which then can be further simplified as follows

$$\begin{aligned} a(\mathbf{N_a^{(j)}}, \mathbf{N_a^{(\bullet)}}) &= \frac{p(\mathbf{y}, \mathbf{N_a^{(\bullet)}}|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathbf{y}, \mathbf{N_a^{(j)}}|\boldsymbol{\theta}) p(\boldsymbol{\theta})} = \frac{p(\mathbf{y}, \mathbf{N_a^{(\bullet)}}|\boldsymbol{\theta})}{p(\mathbf{y}, \mathbf{N_a^{(j)}}|\boldsymbol{\theta})} \\ &= \frac{p(y_t|N_t^{(\bullet)}) \mathbf{1}^T \Gamma_{t-1}^{(\bullet)} P_t^{(\bullet)} \Gamma_t^{(\bullet)} P_{t+1}^{(\bullet)} \Gamma_{t+1}^{(\bullet)} P_{t+2}^{(\bullet)} \mathbf{1}}{p(y_t|N_t^{(j)}) \mathbf{1}^T \Gamma_{t-1}^{(j)} P_t^{(j)} \Gamma_t^{(j)} P_{t+1}^{(j)} \Gamma_{t+1}^{(j)} P_{t+2}^{(j)} \mathbf{1}}, \end{aligned}$$

Method	Absolute time	Relative time
DA	1203.67	1.00
Adapt10	978.27	0.81
Adapt20	1067.61	0.89
Adapt30	1024.87	0.85
Bin10	1022.32	0.85
Bin20	1060.41	0.88
Bin30	1135.83	0.94
Exact	2855.16	2.37

Table 4.4.1: Lapwings data: absolute (in seconds) and relative (wrt the full DA) computation times for $M = 100,000$ posterior draws after a burn-in of $10,000$.

where the superscripts $^{(j)}$ and $^{(\bullet)}$ refer to values computed based on the current state of the Markov chain, $N_{a,t}^{(j)}$, and on the proposed value $N_{a,t}^{(\bullet)}$, respectively. Hence, due to the second-order structure we need five elements of the vector $\mathbf{N_a}$ when updating $N_{a,t}$ (i.e. $N_{a,t-2}$, $N_{a,t-1}$, $N_{a,t}$, $N_{a,t+1}$ and $N_{a,t+2}$), while all other terms cancel in the acceptance probability as they are conditionally independent.

Results We compare the performance of the standard DA approach, in which we impute $\boldsymbol{\theta}$, $\mathbf{N_1}$ and $\mathbf{N_a}$, with that of the SCDA, in which we impute $\boldsymbol{\theta}$ and $\mathbf{N_a}$. As already mentioned above, for comparability we use a "vanilla" MH RW algorithm for the estimation of the integrated model. In particular, we use a discrete uniform Metropolis RW algorithm to perform single-step updates of the states and normal Metropolis RW to sample the logistic regression coefficients. For the observation variance we use a Gibbs update with the conditional distribution being of the form

$$\sigma_y^2|\mathbf{N_a} \sim \Gamma{-}1\left(a_y + \frac{T}{2}, b_y + \frac{1}{2}\sum_{t=1}^{T}(y_t - N_{a,t})^2\right).$$

For the SCDA we first consider the "exact" integration used in the derivations above, in which the only influence on the posterior is the upper limit of the admissible integration range which we set $b_B = 679$. This choice of the upper bound is based on the results for first-years from previous studies and from preliminary runs of the full DA. We further consider a number of approximate schemes based on fixed and adaptive intervals (with 10, 20 and 30 bins in each case). For adaptive bins we use a normal approximation to the Poisson distribution as mentioned in Section 4.3.1. Each time we draw $M = 100,000$ draws after a burn-in of $10,000$.

Table 4.4.1 summarises computation time for each of the schemes. As expected, the exact method is the slowest (2.5 times than the full DA approach) as each integration is based on summing of 680 elements. All the approximate schemes are faster (10–20%) than the DA approach thanks to their efficient implementation based on vectorised computations with relatively few elements to be summed every iteration. Tables 4.4.2 and 4.4.3 present the results for $\boldsymbol{\theta}$ and for selected elements of $\mathbf{N_a}$, respectively, and we report posterior means and standard deviations as well as ESSs (as defined in (4.4.2)) and ESSs per second. Figure 4.4.3 illustrates the posterior means and 95% credible intervals (CI) for the adult population comparing the accuracy of the full DA with that of the SCDA methods (separately for the adaptive intervals and fixed bins). We can see that all the methods deliver virtually the same posterior means and comparable 95% symmetric CI, with only the fixed bin case with 10 bins deviating slightly from all other methods. Interestingly, 10 adaptive bins give very comparable estimates to the other approaches in this case, indicating an increased accuracy of the adaptive approach.

Our results demonstrate the efficiency of the proposed SCDA approach: all the SCDA-based schemes, except the one based on 10 fixed bins, outperform the full DA approach by delivering much higher (up to 4 times) ESSs and ESSs per second. This can be also seen in Figures 4.4.4 and 4.4.5 which show the autocorrelation (ACF) plots for the SSM parameters (except for Gibbs-updated σ_y^2) and for the selected elements of $\mathbf{N_a}$, respectively. In most of the illustrated cases the ACF plots for all the SCDA variants are much flatter that these for the full DA approach.

Method		α_1	α_a	α_ρ	α_λ	β_1	β_a	β_ρ	β_λ	σ_y^2
DA	Mean	0.547	1.574	-1.189	-4.578	-0.164	-0.240	-0.348	-0.364	30180.443
	(Std)	(0.068)	(0.071)	(0.091)	(0.035)	(0.062)	(0.039)	(0.043)	(0.040)	(8890.540)
	ESS	685.018	124.003	111.731	1089.194	1050.268	389.651	105.958	8205.778	1245.440
[1203.67 s]	ESS/sec.	0.569	0.103	0.093	0.905	0.873	0.324	0.088	6.817	1.035
Adapt10	Mean	0.547	1.564	-1.180	-4.580	-0.163	-0.239	-0.350	-0.364	30355.448
	(Std)	(0.068)	(0.070)	(0.092)	(0.035)	(0.061)	(0.040)	(0.040)	(0.040)	(8928.277)
	ESS	1490.019	390.223	316.009	**3035.051**	2777.217	527.023	126.022	7491.584	1852.490
[978.27 s]	ESS/sec.	1.523	**0.399**	**0.323**	**3.102**	**2.839**	0.539	0.129	7.658	1.894
Adapt20	Mean	0.544	1.564	-1.173	-4.581	-0.162	-0.238	-0.342	-0.363	30002.520
	(Std)	(0.069)	(0.072)	(0.094)	(0.035)	(0.060)	(0.039)	(0.039)	(0.040)	(8759.372)
	ESS	1359.221	395.695	324.918	2720.786	2685.188	586.964	243.425	8212.358	2074.915
[1067.62 s]	ESS/sec.	1.273	0.371	0.304	2.548	2.515	**0.550**	0.228	7.692	1.944
Adapt30	Mean	0.542	1.561	-1.166	-4.581	-0.162	-0.241	-0.339	-0.363	30311.546
	(Std)	(0.069)	(0.071)	(0.092)	(0.036)	(0.061)	(0.039)	(0.040)	(0.040)	(8888.626)
	ESS	1438.464	322.169	243.433	2736.107	2471.447	563.625	195.736	7146.030	2129.241
[1024.87 s]	ESS/sec.	1.404	0.314	0.238	2.670	2.411	0.550	0.191	6.973	2.078
Bin10	Mean	0.512	1.441	-1.044	-4.599	-0.207	-0.205	-0.348	-0.353	29992.001
	(Std)	(0.070)	(0.055)	(0.063)	(0.034)	(0.050)	(0.039)	(0.022)	(0.040)	(8837.189)
	ESS	942.247	34.191	37.276	562.131	181.066	104.744	**282.356**	8770.878	1627.515
[1022.32 s]	ESS/sec.	0.922	0.033	0.036	0.550	0.177	0.102	**0.276**	**8.579**	1.592
Bin20	Mean	0.546	1.570	-1.179	-4.579	-0.170	-0.240	-0.343	-0.364	30156.695
	(Std)	(0.069)	(0.069)	(0.090)	(0.035)	(0.061)	(0.039)	(0.040)	(0.040)	(8802.537)
	ESS	1250.068	269.489	210.552	2328.072	2566.054	525.402	139.107	8582.549	2269.829
[1060.41 s]	ESS/sec.	1.179	0.254	0.199	2.195	2.420	0.495	0.131	8.094	2.141
Bin30	Mean	0.545	1.562	-1.170	-4.580	-0.162	-0.240	-0.342	-0.363	30012.541
	(Std)	(0.069)	(0.073)	(0.095)	(0.035)	(0.061)	(0.039)	(0.040)	(0.040)	(8698.313)
	ESS	**1758.336**	**438.518**	**329.308**	2901.919	**2873.035**	501.803	207.643	7613.013	2705.886
[1135.83 s]	ESS/sec.	**1.548**	0.386	0.290	2.555	2.529	0.442	0.183	6.703	**2.382**
Exact	Mean	0.545	1.564	-1.175	-4.580	-0.162	-0.240	-0.345	-0.363	30063.430
	(Std)	(0.068)	(0.069)	(0.090)	(0.035)	(0.060)	(0.039)	(0.042)	(0.040)	(8770.776)
	ESS	1631.604	433.106	361.747	3058.624	2658.813	**720.514**	191.434	**8859.552**	**2734.263**
[2855.16 s]	ESS/sec.	0.571	0.152	0.127	1.071	0.931	0.252	0.067	3.103	0.958

ESS: at lag equal to the lowest order at which sample autocorrelation is not significant.

Computing times (in seconds) in square brackets.

Table 4.4.2: Posterior means, standard deviations and effective sample sizes (ESS) of the model parameters for $M = 100,000$ posterior draws after a burn-in of $10,000$ for the lapwings data. The highest ESS and ESS/sec. for each parameter in bold.

Method		Na_4	Na_8	Na_{12}	Na_{16}	Na_{20}	Na_{24}	Na_{28}	Na_{32}	Na_{36}	Na_{min}	Na_{max}
DA	Mean	1083.511	1325.452	1674.379	1935.843	1614.735	1264.606	1174.201	964.850	776.152	1113.758	1083.511
	(Std)	(26.311)	(43.084)	(52.062)	(67.986)	(53.974)	(53.679)	(49.459)	(59.570)	(72.066)	(50.975)	(26.311)
	ESS	459.890	179.262	120.533	134.145	147.491	154.218	59.186	61.888	68.193	55.390	459.890
[1203.67 s]	ESS/sec.	0.382	0.149	0.100	0.111	0.123	0.128	0.049	0.051	0.057	0.046	0.382
Adapt10	Mean	1083.463	1326.451	1681.573	1947.344	1621.635	1268.140	1174.992	962.144	770.630	1113.567	1083.463
	(Std)	(26.252)	(43.505)	(51.511)	(68.960)	(53.635)	(56.490)	(49.993)	(56.982)	(66.910)	(51.062)	(26.252)
	ESS	879.048	**393.504**	317.925	335.006	294.857	234.322	42.972	46.077	51.279	41.776	879.048
[978.27 s]	ESS/sec.	**0.899**	**0.402**	0.325	0.342	0.301	0.240	0.044	0.047	0.052	0.043	**0.899**
Adapt20	Mean	1081.745	1320.731	1670.036	1934.081	1615.289	1268.925	1181.088	973.608	785.422	1121.175	1081.745
	(Std)	(27.448)	(44.994)	(52.415)	(67.959)	(51.419)	(53.695)	(46.369)	(51.319)	(61.671)	(46.408)	(27.448)
	ESS	792.495	300.795	262.727	413.404	298.921	310.999	173.526	150.006	**160.730**	167.407	792.495
[1067.62 s]	ESS/sec.	0.742	0.282	0.246	0.387	0.280	0.291	0.163	0.141	0.151	0.157	0.742
Adapt30	Mean	1081.738	1319.287	1670.942	1938.897	1617.392	1268.431	1184.041	978.162	791.521	1124.325	1081.738
	(Std)	(27.373)	(45.491)	(51.544)	(65.247)	(53.622)	(54.570)	(48.404)	(56.590)	(68.242)	(49.640)	(27.373)
	ESS	596.757	278.204	246.717	434.153	326.282	305.667	180.570	154.889	163.795	**170.926**	596.757
[1024.87 s]	ESS/sec.	0.582	0.271	0.241	0.424	0.318	0.298	**0.176**	0.151	**0.160**	**0.167**	0.582
Bin10	Mean	1075.915	1343.027	1699.239	1979.991	1671.882	1313.004	1194.677	954.958	733.686	1140.264	1075.915
	(Std)	(26.815)	(43.436)	(50.665)	(62.415)	(48.444)	(54.649)	(42.543)	(39.704)	(41.444)	(43.179)	(26.815)
	ESS	868.244	310.552	327.451	243.190	172.939	108.027	97.045	**162.952**	91.937	87.733	868.244
[1022.32 s]	ESS/sec.	0.849	0.304	0.320	0.238	0.169	0.106	0.095	**0.159**	0.090	0.086	0.849
Bin20	Mean	1079.800	1319.959	1671.116	1939.171	1619.882	1270.619	1183.223	976.198	788.919	1123.716	1079.800
	(Std)	(25.975)	(44.138)	(51.979)	(67.014)	(52.053)	(54.001)	(46.782)	(53.316)	(64.386)	(47.502)	(25.975)
	ESS	785.417	279.161	225.312	331.389	344.148	328.106	73.864	57.640	63.944	67.105	785.417
[1060.41 s]	ESS/sec.	0.741	0.263	0.212	0.313	0.325	**0.309**	0.070	0.054	0.060	0.063	0.741
Bin30	Mean	1079.485	1320.178	1671.219	1936.264	1615.844	1268.040	1181.895	975.230	787.549	1121.970	1079.485
	(Std)	(25.719)	(43.238)	(47.997)	(62.926)	(50.327)	(53.921)	(46.464)	(53.458)	(65.120)	(47.102)	(25.719)
	ESS	**911.428**	369.874	**373.588**	**504.546**	346.512	246.825	111.292	89.719	98.283	102.004	**911.428**
[1135.83 s]	ESS/sec.	0.802	0.326	**0.329**	**0.444**	**0.305**	0.217	0.098	0.079	0.087	0.090	0.802
Exact	Mean	1083.134	1324.134	1675.323	1939.629	1615.882	1265.869	1176.687	968.356	780.241	1116.332	1083.134
	(Std)	(27.191)	(44.555)	(52.061)	(67.385)	(53.858)	(54.547)	(46.090)	(54.115)	(66.988)	(47.247)	(27.191)
	ESS	902.234	349.462	293.269	365.968	**402.009**	**418.141**	**196.821**	121.888	116.693	167.980	902.234
[2855.16 s]	ESS/sec.	0.316	0.122	0.103	0.128	0.141	0.146	0.069	0.043	0.041	0.059	0.316

ESS: at lag equal to the lowest order at which sample autocorrelation is not significant.

Na_{min}/Na_{max}: corresponding to the lowest/highest ESS for the DA method.

Computing times (in seconds) in square brackets.

Table 4.4.3: Posterior means, standard deviations and effective sample sizes (ESS) of the model parameters for $M = 100,000$ posterior draws after a burn-in of $10,000$ for the lapwings data. The highest ESS and ESS/sec. for each parameter in bold.

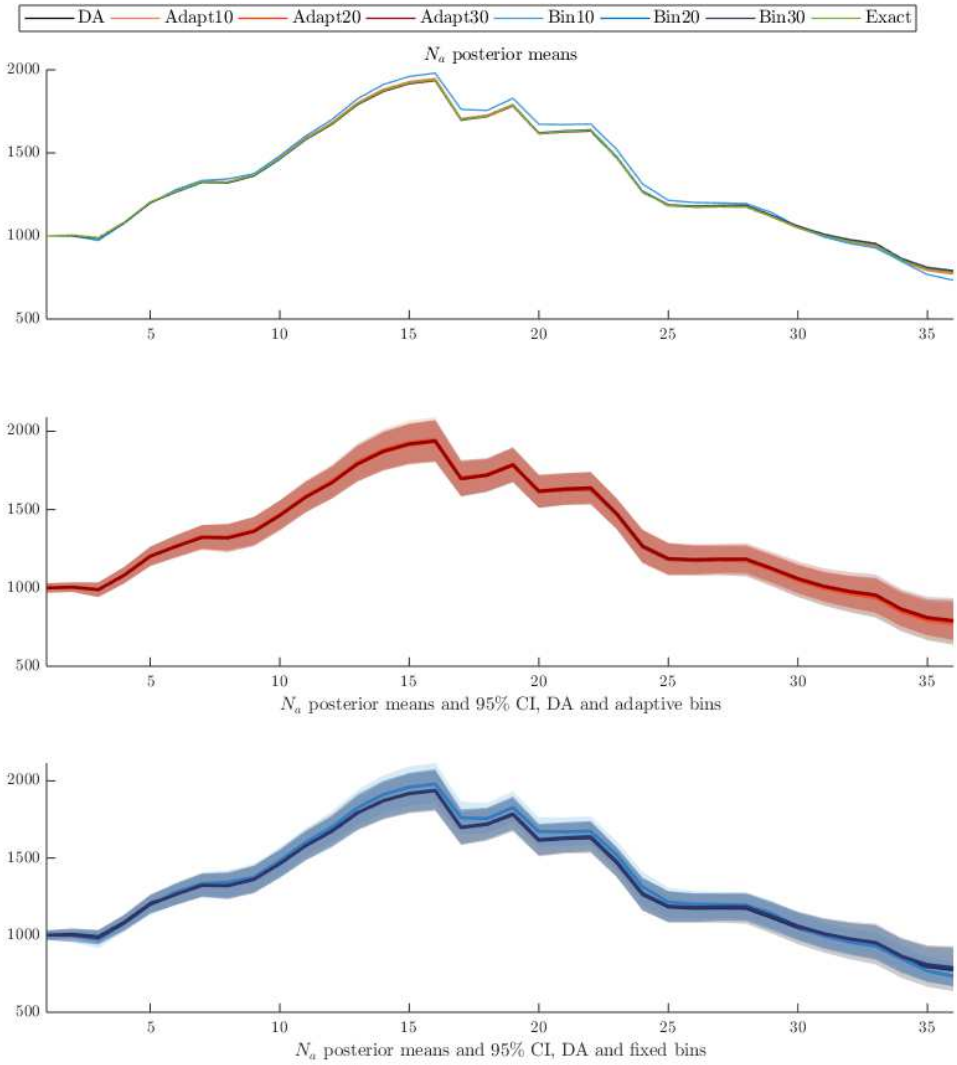


Figure 4.4.3: Lapwings data: the posterior means and 95% CI for the adult population.

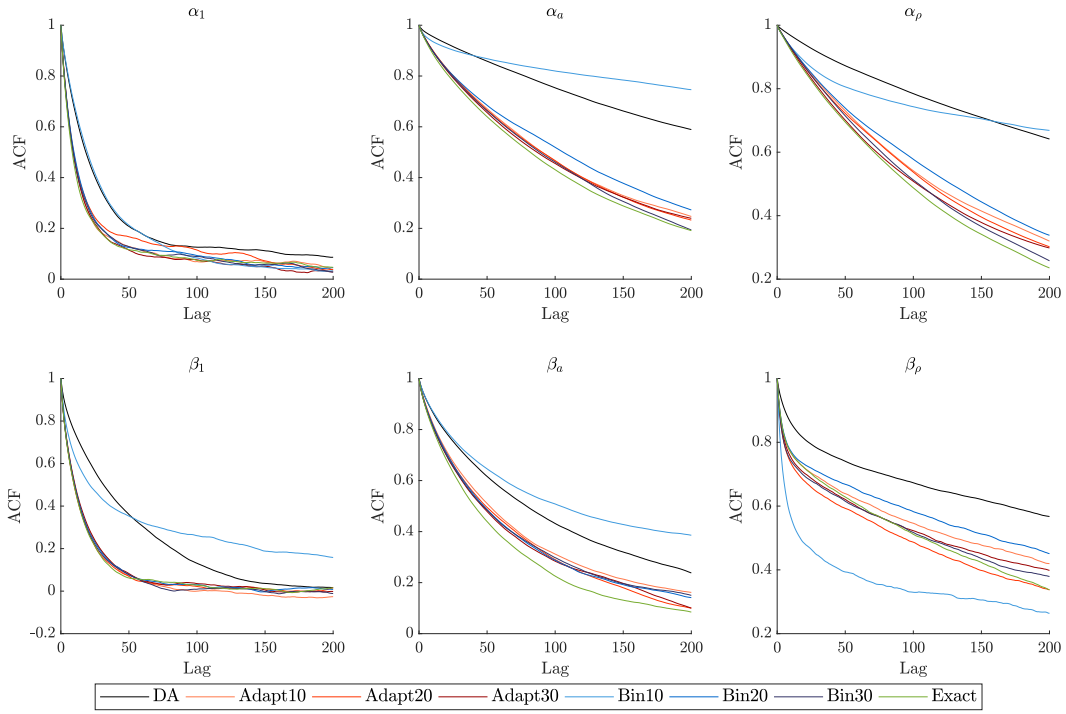


Figure 4.4.4: Lapwings data: ACF plots for the SSM parameters.

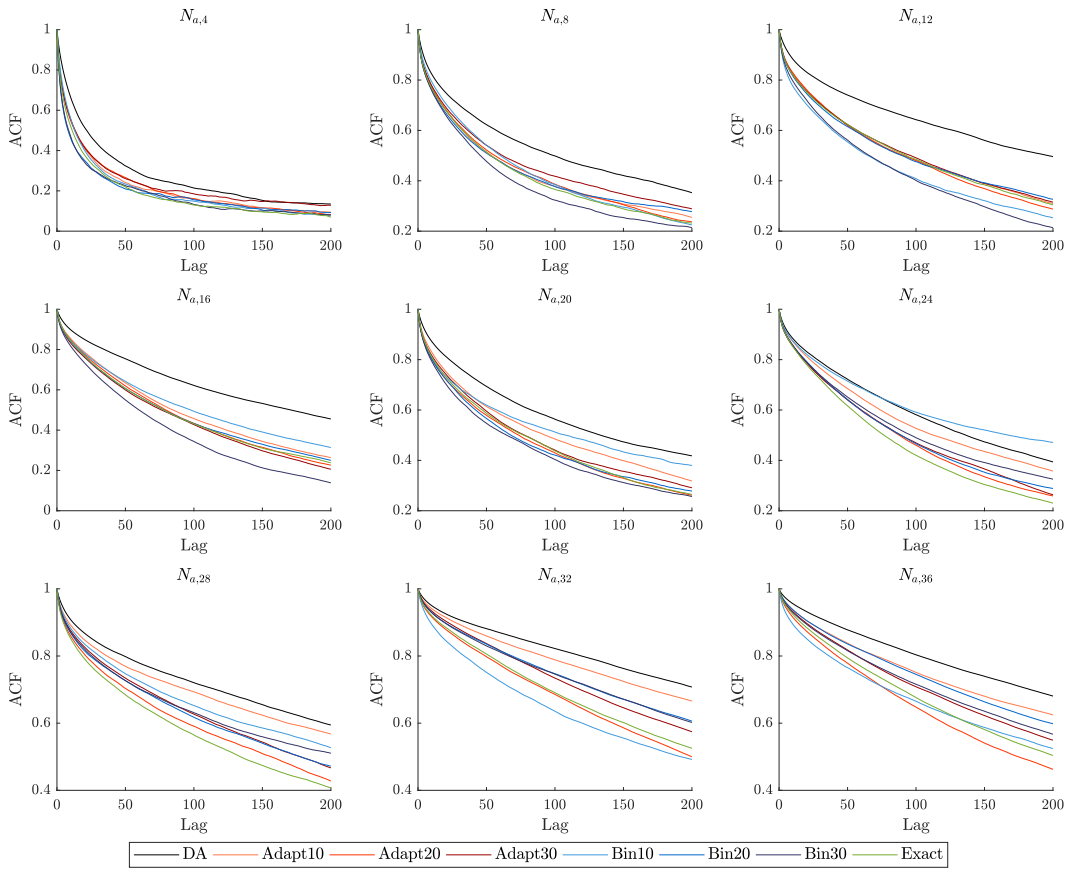


Figure 4.4.5: Lapwings data: ACF plots for the adult population.

4.4.2 Financial model: stochastic volatility

As our second illustration we consider the SV model in its basic form given by

$$y_t|h_t, \boldsymbol{\theta} \sim \mathcal{N}\left(0, \exp(h_t)\right), \tag{4.4.11}$$

$$h_{t+1}|h_t \sim \mathcal{N}\left(\mu + \phi(h_t - \mu), \sigma^2\right), \tag{4.4.12}$$

$$h_0 \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{1-\phi^2}\right), \tag{4.4.13}$$

for $t = 1, \ldots, T$. We adopt the prior specification of Kim et al. (1998)

$$\begin{aligned} \mu &\sim \mathcal{N}(0, \sigma^2_{\mu 0}), \\ \frac{\phi+1}{2} &\sim \mathcal{B}(\alpha_{\phi 0}, \beta_{\phi 0}), \\ \sigma^2 &\sim \mathcal{IG}(\alpha_{\sigma^2 0}, \beta_{\sigma^2 0}), \end{aligned}$$

with $\sigma^2_{\mu 0} = 10$, $\alpha_{\phi 0} = 20$, $\beta_{\phi 0} = 1.5$, $\alpha_{\sigma^2 0} = 5/2$, $\beta_{\sigma^2 0} = 0.05/2$. Estimation of the SV model has been considered as a challenging problem due to the intractable likelihood

$$p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}, \mathbf{h}) d\mathbf{h} = \int p(h_0) \prod_{t=1}^{T} p(y_t|h_t) p(h_t|h_{t-1}) dh_0 dh_1 \ldots dh_T. \tag{4.4.14}$$

Some of the previous approaches to tackle this issue include standard DA approach, in which the latent volatilities are imputed in an MCMC scheme, see Kim et al. (1998), Omori et al. (2007). Then, the augmented likelihood can be expressed in a closed form as

$$p(\mathbf{y}, \mathbf{h}|\boldsymbol{\theta}) = p(h_0) \prod_{t=1}^{T} p(y_t|h_t) p(h_t|h_{t-1}).$$

An alternative approach is provided by Fridman and Harris (1998) or Langrock et al. (2012b) who propose numerical integration of the latent states. In particular, Langrock et al. (2012b) approximate (4.4.14) using an HMM by discretising the state space of the SV model. They consider a form of numerical integration of the latent states based on a grid of B equally sized intervals (bins) $B_i = [b_{i-1}, b_i)$, $i = 1, \ldots, B$, with the corresponding representative points b_i^* (e.g. the midpoints). The range of the admissible values for the demeaned volatility, b_0 and b_B is set e.g. to $\pm 5\sigma_h$, with σ_h being the stationary (unconditional) standard deviation of the logvolatility process.

This leads to an approximation of (4.4.14) via

$$p(\mathbf{y}|\boldsymbol{\theta}) \approx \mathbf{u}_0 \prod_{t=1}^{T} \Gamma Q_t \mathbf{1},$$

where $\Gamma = \left[\gamma_{i,j}\right]_{i,j=1,\ldots,B}$, with

$$\begin{aligned}
\gamma_{i,j} &= \mathbb{P}(h_t - \mu \in B_j | h_{t-1} - \mu = b_i^*) \\
&= \Phi\left(\frac{b_j - \phi b_i^*}{\sigma}\right) - \Phi\left(\frac{b_{j-1} - \phi b_i^*}{\sigma}\right), \\
Q_t &= \operatorname{diag}\left(\varphi\left(\frac{y_t}{\exp((\mu + b_i^*)/2)}\right)\right)_{i=1,\ldots,B},
\end{aligned}$$

where Φ and φ denote the cdf and the pdf of the standard normal density, respectively. Notice that the transition probabilities are time-constant so that the underlying Markov chain is homogeneous. Sandmann and Koopman (1998) point out that for the SV model such a form of numerical integration might not be always suitable since a fixed grid cannot efficiently capture different scales of volatility (periods of low and high volatility). We address this issue by suggesting an adaptive and more efficient HMM-based approximation as an alternative to the fixed bins used by Langrock et al. (2012b).

Finally, we note that for μ and σ^2 Gibbs updates can be performed based on full conditional densities, see Kim et al. (1998). Furthermore, numerous enhancements for sampling of the hidden states has been devised, Kim et al. (1998), Omori et al. (2007) and Bos (2011) for an overview. However, our aim is to provide a general framework requiring only "vanilla" type updates (based on a RW–MH algorithm) and hence we consider the standard full DA as a comparison benchmark.

Dependence structure and SCDL The basic SV model specification concerns a single one-dimensional state on the real line, which is saliently different from the lapwings case. The sampling inefficiency in the current case originates from a high persistence of the logvolatility process. In order to break this dependence, we propose to impute $\mathbf{h_{2T}}$, the even states and to integrate out $\mathbf{h_{2T+1}}$, the odd ones. This corresponds to the *vertical* integration scheme with $\mathbf{x_{int}} = \mathbf{h_{2T+1}}$ and $\mathbf{x_{aug}} = \mathbf{h_{2T}}$. Without loss of generality we assume that T is odd so that h_T is integrated out; if T is even then we add one extra integration based on uniformly distributed h_{T+1}. We denote $T^* = \frac{T-1}{2}$

and skip $\boldsymbol{\theta}$ in conditioning for simplicity. The exact SCDL is given by

$$p(y, \mathbf{h_{2T}}) = p(h_0) \int p(h_1|h_0)p(y_1|h_0) \prod_{t=1}^{T^*} p(y_{2t+1}|h_{2t+1})p(h_{2t+1}|h_{2t}) \\ p(y_{2t}|h_{2t})p(h_{2t}|h_{2t-1})dh_1 \dots dh_T, \tag{4.4.15}$$

and conditioning on the even states allows us to split (4.4.15) into a product $T^* + 1$ of integrals

$$p(y, \mathbf{h_{2T}}) = \underbrace{p(h_0)}_{=:C_0} \underbrace{\int p(h_1|h_0)p(y_1|h_0)dh_1}_{=:D_0} \\ \times \prod_{t=1}^{T^*} \underbrace{p(y_{2t}|h_{2t})}_{=:C_t} \underbrace{\int p(y_{2t+1}|h_{2t+1})p(h_{2t+1}|h_{2t})p(h_{2t}|h_{2t-1})dh_{2t+1}}_{=:D_t}. \tag{4.4.16}$$

Since the integrals in (4.4.16) are conditionally independent, it can be expressed as

$$p(y, \mathbf{h_{2T}}) = C_0 D_0 \prod_{t=1}^{T^*} C_t D_t = \prod_{t=0}^{T^*} C_t D_t,$$

which block structure is helpful for visualising the MH update scheme as we present below.

We denote by $\mathbf{h_{2T}}^{(j)} = \{h_0^{(j)}, h_2^{(j)}, \dots, h_{2t+2}^{(j)}, \dots, h_T^{(j)}\}$ the current sequence of the imputed states and suppose that a single RW MH step for h_{2t+2} results in the proposed sequence $\mathbf{h_{2T}}^{(\bullet)}$ with the element $h_{2t+2}^{(j)}$ replaced by the candidate $h_{2t+2}^{(\bullet)}$. Since the proposal distribution is symmetric and thus the proposal terms cancel out, the state acceptance rate is given by

$$a(h_{2t+2}^{(\bullet)}, h_{2t+2}^{(j)}) = \frac{p(y, h_{2T}^{(\bullet)}|\boldsymbol{\theta})}{p(y, h_{2T}^{(j)}|\boldsymbol{\theta})} = \frac{C_t^{(\bullet)} D_t^{(\bullet)} D_{t+1}^{(\bullet)}}{C_t^{(j)} D_t^{(j)} D_{t+1}^{(j)}}, \tag{4.4.17}$$

where $^{(\bullet)}$ and $^{(j)}$ refer to the blocks evaluated on the proposed and the current variable, respectively (either the imputed state here or the parameter vector below).

For a single step RW MH update of $\boldsymbol{\theta}$, given $h_{2T}^{(j)}$ and y:

$$a(\boldsymbol{\theta}^{(j)}, \boldsymbol{\theta}^{(\bullet)}) = \frac{p(y, \mathbf{h_{2T}}^{(j)}|\boldsymbol{\theta}^{(\bullet)})p(\boldsymbol{\theta}^{(\bullet)})}{p(y, \mathbf{h_{2T}}^{(j)}|\boldsymbol{\theta}^{(j)})p(\boldsymbol{\theta}^{(j)})} = \frac{p(\boldsymbol{\theta}^{(\bullet)}) \prod_{t=0}^{T^*} D_t^{(\bullet)}}{p(\boldsymbol{\theta}^{(j)}) \prod_{t=0}^{T^*} D_t(j)}. \tag{4.4.18}$$

Hidden Markov model approximation In practice the integrals D_t cannot be evaluated analytically and a form of numerical approximation needs to be adopted. We first propose to approximate each integral using a B-state HMM structure with fixed bins. This approach follows Langrock et al. (2012b) and consists in relating $z_t = k$, the Markov chain being in state k, to the event of $h_{2t+1} - \mu \in \mathcal{B}_k$, the demeaned volatility in an odd time period $2t+1$ falling into the kth bin B_k. Falling into bin B_k can be specified as e.g. lying in the interval $[b_{k-1}, b_k)$ or being equal to this interval's midpoint $b_k^* = \frac{b_{k-1}+b_k}{2}$. We take equally spaced bins, each of length λ. In particular, we consider approximation of the following form

$$D_t \approx \hat{D}_t = \sum_{k=1}^{B} p(y_{2t+1}|h_{2t+1} - \mu = b_k^*) p(h_{2t+2}|h_{2t+1} - \mu = b_k^*) p(h_{2t+1} - \mu \in B_i|h_{2t}). \tag{4.4.19}$$

The last term in (4.4.19) can be approximated as

$$p(h_{2t+1} - \mu \in B_k|h_{2t}) \approx \Phi\left(\frac{b_k - \phi(h_{2t} - \mu)}{\sigma}\right) - \Phi\left(\frac{b_{k-1} - \phi(h_{2t} - \mu)}{\sigma}\right),$$

which is adopted in Langrock et al. (2012b), or using a simpler midpoint approximation

$$p(h_{2t+1} - \mu \in B_k|h_{2t}) \approx \lambda\varphi\left(\frac{b_k^* - \phi(h_{2t} - \mu)}{\sigma}\right),$$

which we adopt in our application due to computing time. Then the state acceptance rate (4.4.17) is approximated as

$$a(h_{2t+2}^{(\bullet)}, h_{2t+2}^{(j)}) \approx \frac{\varphi\left(\frac{y_{2t+1}}{\exp(h_{2t+2}^{(\bullet)}/2)}\right) \sum_{k=1}^{B} \varphi\left(\frac{y_{2t+1}}{\exp((b_k^*+\mu)/2)}\right) \varphi\left(\frac{h_{2t+2}^{(*)} - \mu - \phi b_k^*}{\sigma}\right) \varphi\left(\frac{b_k^* - \phi(h_{2t}-\mu)}{\sigma}\right)}{\phi\left(\frac{y_{2t+1}}{\exp(h_{2t+2}^{(j)}/2)}\right) \sum_{k=1}^{B} \varphi\left(\frac{y_{2t+1}}{\exp((b_k^*+\mu)/2)}\right) \varphi\left(\frac{h_{2t+2}^{(j)} - \mu - \phi b_k^*}{\sigma}\right) \varphi\left(\frac{b_k^* - \phi(h_{2t}-\mu)}{\sigma}\right)}$$
$$\times \frac{\sum_{k=1}^{B} \varphi\left(\frac{y_{2t+3}}{\exp((b_k^*+\mu)/2)}\right) \varphi\left(\frac{h_{2t+4}^{(j)} - \mu - \phi b_k^*}{\sigma}\right) \varphi\left(\frac{b_k^* - \phi(h_{2t+2}^{(*)}-\mu)}{\sigma}\right)}{\sum_{k=1}^{B} \varphi\left(\frac{y_{2t+3}}{\exp((b_k^*+\mu)/2)}\right) \varphi\left(\frac{h_{2t+4}^{(j)} - \mu - \phi b_k^*}{\sigma}\right) \varphi\left(\frac{b_k^* - \phi(h_{2t+2}^{(j)}-\mu)}{\sigma}\right)},$$

while for the parameter acceptance rate (4.4.18) we obtain

$$a(\boldsymbol{\theta}^{(j)}, \boldsymbol{\theta}^{(\bullet)}) \approx \frac{p(\boldsymbol{\theta}^{(\bullet)}) \prod_{t=0}^{T^*} \hat{D}_t^{(\bullet)}}{p(\boldsymbol{\theta}^{(j)}) \prod_{t=0}^{T^*} \hat{D}_t^{(j)}},$$

where

$$\hat{D}_t^{(\bullet)} = \sum_{k=1}^{B} \varphi\left(\frac{y_{2t+1}}{\exp((b_k^* + \mu^{(\bullet)})/2)}\right) \varphi\left(\frac{b_k^* - \phi^{(\bullet)}(h_{2t}^{(j)} - \mu^{(\bullet)})}{\sigma^{(\bullet)}}\right) \varphi\left(\frac{h_{2t+2}^{(j)} - \mu^{(\bullet)} - \phi^{(\bullet)} b_k^*}{\sigma^{(\bullet)}}\right),$$

$$\hat{D}_t^{(j)} = \sum_{k=1}^{B} \varphi\left(\frac{y_{2t+1}}{\exp((b_k^* + \mu^{(j)})/2)}\right) \varphi\left(\frac{b_k^* - \phi(h_{2t}^{(j)} - \mu^{(j)})}{\sigma^{(j)}}\right) \varphi\left(\frac{h_{2t+2}^{(j)} - \mu^{(j)} - \phi^{(j)} b_k^*}{\sigma^{(j)}}\right).$$

Adaptive HMM-based approximation An alternative approach to the approximation task is to use adaptive intervals. In particular, quantiles corresponding to intervals of equal probability can be used. Then, instead of specifying the grid points, we fix the probabilities for each bin, which previously needed to be determined. Thus, we face a quantile determination problem, as these are needed to obtain the midpoint values (used in conditioning). Consider a vector of quantiles $\mathbf{q} = [q_0, q_1, \ldots, q_B]$ together with their midpoints $\mathbf{q}^* = [q_1^*, q_1^*, \ldots, q_B^*]$ given by $q_k^* = \frac{q_{k-1}+q_k}{2}$. Then the bin midpoints at time $2t+1$ determined by the mid-quantiles are given by

$$\beta_{k,2t+1}^* = \phi(h_{2t} - \mu) + \sigma \cdot \Phi^{-1}(q_k^*), \qquad k = 1, \ldots, B,$$

where h_{2t} the imputed volatility for the previous time period. This means that

$$\gamma_{k,t} = p(h_{2t+1} - \mu \in \mathcal{B}_{k,2t+1} | h_{2t}, \boldsymbol{\theta}) = \frac{1}{B}$$

and we approximate D_t as

$$D_t \approx \sum_{k=1}^{B} \varphi\left(\frac{y_{2t+1}}{\exp((\beta_{k,2t+1}^* + \mu)/2)}\right) \varphi\left(\frac{h_{2t+2} - \mu - \phi\beta_{k,2t+1}^*}{\sigma}\right) \cdot \frac{1}{B},$$

where the constant transition probabilities from an imputed state cancel out in the acceptance ratios.

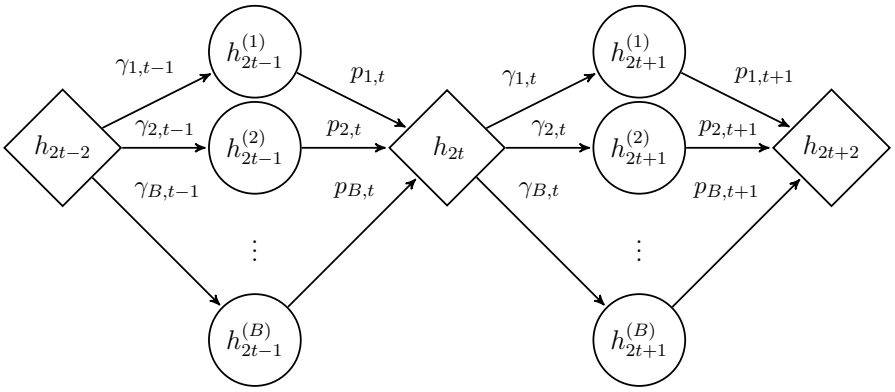


Figure 4.4.6: SV model: combining DA and the HMM-based integration. Diamonds represent the imputed states, circles – the states being integrated out. $h_t^{(k)}$ denotes $h_t \in \mathcal{B}_k$. The graph presents a single imputation problem of h_{2t} with the associated integrations.

Extensions of the basic SV model

SV in the mean The proposed SCDA scheme easily extends to more complex models, e.g. the popular Stochastic Volatility in the Mean (SVM) model of Koopman and Uspensky (2002) (see also Chan, 2017). Its basic specification is given by

$$y_t|h_t, \boldsymbol{\theta} \sim \mathcal{N}\left(\beta \exp(h_t), \exp(h_t)\right), \tag{4.4.20}$$

$$h_{t+1}|h_t \sim \mathcal{N}\left(\mu + \phi(h_t - \mu), \sigma^2\right), \tag{4.4.21}$$

$$h_0 \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{1-\phi^2}\right), \tag{4.4.22}$$

for $t = 1, \ldots, T$. Hence, the latent volatility process h_t influences both the conditional variance and the conditional mean of the observation series y_t, which is additionally controlled by a scaling parameter β. For the volatility parameters μ, ϕ and σ^2 we adopt the prior specification as for the standard SV model, while for the mean-scaling parameter we specify $\beta \sim \mathcal{N}(0, \sigma_{\beta_0}^2)$, with $\sigma_{\beta_0}^2 = 10$.

SV with leverage logreturns to the current value of the volatility process. This effect is typically modelled as a negative correlation between the last period logreturns and the current value of volatility. The motivation behind the leverage effect is that the volatility in financial markets may adapt differently to positive and negative

shocks/news (affecting logreturns), where large negative shocks are likely to increase the volatility. The SV model with leverage (SVL) has been frequently analysed in the literature, see Jungbacker and Koopman (2007), Meyer and Yu (2000), Yu (2005), Durbin and Koopman (2012, Section 9.5.5.) or Zucchini et al. (2016, Section 20.2.3). For convenience, we rewrite the basic SV model (4.4.11)–(4.4.13) as

$$\begin{aligned} y_t &= \exp(h_t/2)\varepsilon_t, & \varepsilon_t &\sim \mathcal{N}(0,1), \\ h_{t+1} &= \mu + \phi(h_t - \mu) + \eta_t, & \eta_t &\sim \mathcal{N}\left(0, \sigma^2\right), \\ h_1 &\sim \mathcal{N}\left(\mu, \frac{\sigma^2}{1-\phi^2}\right), \end{aligned}$$

for $t = 1, \ldots, T$. The only difference between the SVL model and the basic specification of the SV model is that now the error terms ε_t and η_t are assumed to be correlated: $\text{corr}[\varepsilon_t, \eta_t] = \rho \neq 0$, with ρ typically estimated to be negative[10]. This apparently slight modification has, however, substantial effect on the dependence structure in the model (see Figure 4.4.7) and hence the conditional distribution of h_t. To derive the latter several reformulations of the model have been proposed (Jungbacker and Koopman, 2007 or Meyer and Yu, 2000), however we will use the treatment provided by Zucchini et al. (2016, Section 20.2.3). These authors use the basic regression lemma for normal variables to show that

$$h_t | h_{t-1}, y_{t-1}, \mu, \phi, \sigma^2, \rho \sim \mathcal{N}\left(\mu + \phi(h_{t-1} - \mu) + \frac{\rho\sigma y_{t-1}}{\exp(h_{t-1}/2)}, \sigma^2(1-\rho^2)\right) \quad (4.4.23)$$

(Appendix 4.C provides the details of the derivation). Formulation (4.4.23) is particularly convenient for "reusing" the derived integration scheme for the basic SV model, as we only need to adjust the transition probabilities in the approximation to C_t.

Modifications to the HMM-based approximation The proposed HMM-based approximation to SCDL can be easily adapted to allow for both extensions by simply modifying the components of the matrices Γ_t, P_t and Q_t specified in (4.3.2)–(4.3.4). Notice that for the SVM model the dependence structure of the state is the same as for the basic SV model, hence the core of the integration/imputation scheme remains unchanged. What needs to be adjusted is the observation density, which can be done in a straightforward manner. The modification for the SVL model requires adjusting of the transition probabilities and the pdfs of the augmented states. Appendix (4.A.3)

[10]For this reason we also need to initialise the state vector one period later, i.e. at h_1. This can be easily understood from (4.4.23), where h_t is conditioned on y_{t-1}, among others.

presents the required modifications for the largest model, allowing for both SV in the mean and for the leverage effect (which we refer to as the SVML model).

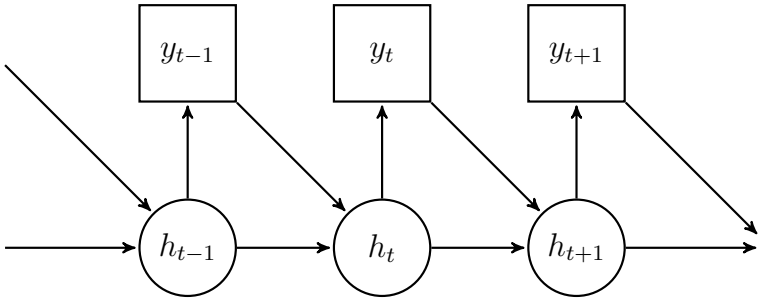


Figure 4.4.7: SV model with leverage: modified dependence structure due to feedback from the logreturns y_{t-1} to logvolatilities h_t.

Application To illustrate the SCDA approach based on vertical integration we consider daily log-returns of the IBM stock from 4th January 2000 to 29th December 2017 (4527 observations). The data are illustrated in Figure 4.4.8. We consider the basic SV model as well as its extended version, i.e. the SVML model. For both models we use adaptive intervals based on 10, 20 and 30 quantiles, while for the SV model we also consider fixed bins based on 20 and 30 intervals. The reason for the latter is that fixed bins turned out to be infeasible for the SVML model (if we want to keep a reasonable number of bins, say below 50) while for the SV model we needed to specify minimum 20 intervals to obtain stable results. For fixed bin we set the integration range to ± 4 (i.e. $b_0 = -4$ and $b_B = 4$). The obtained posterior means for the imputed volatilities suggest that this choice was sufficient, as the estimated mean of the state ranges roughly from -1 to 3 (Figure 4.4.9). For each model and method we simulate 50,000 draws after a burn-in of 10,000.

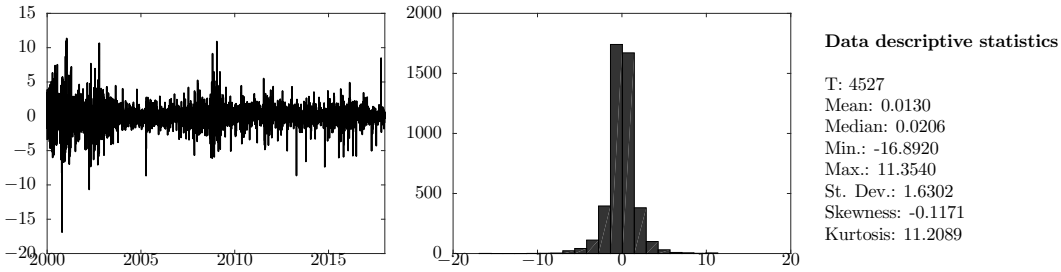


Figure 4.4.8: SV model: IBM series, 4527 observations from 4th January 2000 to 29th December 2017.

Tables 4.4.4 and 4.4.5 present the parameter estimation results for the SV model and SVML model, respectively. Tables 4.4.6 and 4.4.7 report the results for selected

volatilities for SV and SVML, repetitively. We can see that for both models all the methods deliver comparable posterior means and standard deviations of parameters and standard deviations. A good agreement of the HMM-based schemes with the benchmark DA approach demonstrates that the developed methods provide a close approximation to the exact semi-complete data posterior. Interestingly, as few as 10 *adaptive* bins suffice to provide accurate estimates, which contrasts with *minimum* 50 fixed bins considered by Langrock et al. (2012b). This demonstrates the flexibility of the adaptive bins used within the SCDA scheme. Figure 4.4.9 illustrates that the estimates (posterior means) of the volatilities from the SV model obtained by the methods considered are very close to each other.

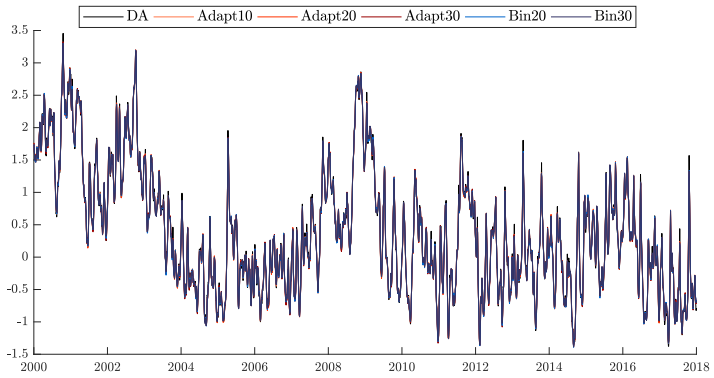


Figure 4.4.9: SV model: posterior means for the imputed volatilities. For illustration, for the SCDA methods the volatilities at odd time period are intrapolated between even time points.

Tables 4.4.4–4.4.7 further reveal that the proposed vertical integration scheme breaks the strong dependence between subsequent states to achieve the desired improvement in mixing. The ESS for model parameters obtained with the SCDA methods are typically higher than for the full DA approach. The only exception is the β parameter of the SVML for which all the methods exhibit excellent mixing with the DA approach slightly outperforming the HMM-based approximations. This high efficiency in the estimations of β is related to the presence of this parameter only in the observation equation hence being less affected by the high autocorrelation of the state process. On the other hand, the second extra parameter of the SVML model, i.e. the leverage parameter ρ, is hard to estimate efficiently. For this parameter the SCDA turns out particularly useful in improving the mixing with the corresponding ESS values being up to 4.5 higher than for the benchmark DA. Figure 4.4.10 displays the ACF plots for the parameters for the SV and SVML model, while Figure 4.4.11 for the selected volatilities. As suggested by the ESS reported in Tables 4.4.4–4.4.7, in the majority of the cases we observe

much quicker decays in the autocorrelations for the SCDA algorithm compared to the "vanilla" DA approach.

Finally, we note however that the computing times are higher for the SCDA approaches, with the computations for the adaptive case based on 10 bins taking roughly 17 times and 7 times longer than for full DA for the basic SV model and the SVML model, respectively. This suggests that the resulting gains in mixing may not necessarily be worth the extra computational cost. However, given the very simple structure of the basic SV model and not much more complex one of the SVML model, this is hardly surprising. We expect the SCDA approach to be more beneficial for more complex models, with even more involved dependence structure and relatively slower computation time for the benchmark DA approach. This can be already partly seen from shorter *relative* (to DA) computing times for the SCDA methods for the SVML compared to these for the SV model. For instance, the proposed integration scheme for the SV model could be particularly useful for a dynamic factor model with double stochastic volatility (where both the observation and the factor disturbances are subject to stochastic volatility). Due to the complex dependence structure as well matrix computations involved, the standard DA can be expected to perform relatively poorly and be time consuming to run. Then, there are several possibilities how to specify the augmentation-integration scheme, e.g. to fully integrate one of the SV processes; or interweave between every-second state of both SV processes (e.g. to integrate odd states for one SV process and even states for another SV process). We leave these extensions for further research as our current goal is to illustrate the generalisability of the SCDA approach using the important "building block" of many econometric models.

Method		μ	ϕ	σ^2
DA	Mean	0.376	0.962	0.081
	(Std)	(0.115)	(0.006)	(0.011)
[111.83 s]	ESS	3201.695	114.259	63.427
Adapt10	Mean	0.382	0.962	0.086
	(Std)	(0.116)	(0.006)	(0.013)
[1980.48 s]	ESS	5398.200	249.402	135.917
Adapt20	Mean	0.379	0.961	0.085
	(Std)	(0.115)	(0.006)	(0.013)
[2290.29 s]	ESS	5440.155	279.273	143.247
Adapt30	Mean	0.376	0.961	0.084
	(Std)	(0.115)	(0.006)	(0.012)
[2566.66 s]	ESS	**5784.093**	120.823	60.656
Bin20	Mean	0.381	0.962	0.082
	(Std)	(0.116)	(0.006)	(0.012)
[1683.02 s]	ESS	4504.012	239.698	147.409
Bin30	Mean	0.378	0.962	0.081
	(Std)	(0.115)	(0.006)	(0.012)
[2056.75 s]	ESS	5484.200	**301.272**	**167.720**

ESS: at lag equal to the lowest order at which sample autocorrelation is not significant.

Computing times (in seconds) in square brackets.

Table 4.4.4: SV model: posterior means, standard deviations and effective sample sizes (ESS) of the model parameters for $M = 50,000$ posterior draws after a burn-in of $10,000$.

Method		μ	ϕ	σ^2	β	ρ
DA	Mean	0.389	0.960	0.085	0.005	-0.286
	(Std)	(0.115)	(0.006)	(0.011)	(0.009)	(0.047)
[202.57 s]	ESS	2510.153	119.283	49.404	**7375.378**	147.507
Adapt10	Mean	0.377	0.963	0.081	0.006	-0.289
	(Std)	(0.112)	(0.006)	(0.011)	(0.009)	(0.044)
[1511.49 s]	ESS	**5879.322**	340.407	171.999	6969.391	**681.943**
Adapt20	Mean	0.375	0.961	0.084	0.006	-0.293
	(Std)	(0.115)	(0.006)	(0.012)	(0.009)	(0.045)
[2039.16 s]	ESS	4835.772	**432.012**	**187.508**	6692.492	513.298
Adapt30	Mean	0.373	0.960	0.084	0.006	-0.292
	(Std)	(0.113)	(0.006)	(0.013)	(0.009)	(0.048)
[2497.22 s]	ESS	5445.668	239.0242	149.680	7185.875	552.179

ESS: at lag equal to the lowest order at which sample autocorrelation is not significant.

Computing times (in seconds) in square brackets.

Table 4.4.5: SVML model: posterior means, standard deviations and effective sample sizes (ESS) of the model parameters for $M = 50,000$ posterior draws after a burn-in of $10,000$.

Method		h_{450}	h_{950}	h_{1450}	h_{1950}	h_{2450}	h_{2950}	h_{3450}	h_{3950}	h_{4450}
DA	Mean	0.974	0.259	-0.009	-0.006	0.257	1.083	0.011	0.733	-0.782
	(Std)	(0.444)	(0.417)	(0.404)	(0.492)	(0.466)	(0.449)	(0.485)	(0.457)	(0.454)
[111.83 s]	ESS	458.198	487.795	573.470	301.690	394.748	501.153	300.990	376.820	428.370
Adapt10	Mean	1.002	0.213	0.0104	-0.051	0.243	1.064	-0.022	0.753	-0.781
	(Std)	(0.434)	(0.431)	(0.432)	(0.492)	(0.469)	(0.445)	(0.479)	(0.453)	(0.482)
[1980.48 s]	ESS	1251.076	1489.877	1438.211	1245.904	1045.616	**1509.202**	**1425.722**	1338.145	1299.627
Adapt20	Mean	0.978	0.228	0.020	-0.040	0.241	1.059	-0.014	0.742	-0.805
	(Std)	(0.434)	(0.447)	(0.4358)	(0.504)	(0.467)	(0.442)	(0.489)	(0.448)	(0.467)
[2290.29 s]	ESS	**1666.989**	1223.563	1435.324	**1330.146**	1285.580	1458.985	1157.446	1288.562	1402.586
Adapt30	Mean	0.961	0.225	0.014	-0.052	0.240	1.101	0.014	0.745	-0.782
	(Std)	(0.439)	(0.438)	(0.431)	(0.499)	(0.457)	(0.442)	(0.482)	(0.436)	(0.474)
[2566.66 s]	ESS	1583.096	1638.283	1521.382	1270.313	**1410.302**	1607.990	1359.351	**1457.938**	1331.350
Bin20	Mean	0.979	0.207	0.018	-0.048	0.251	1.069	0.010	0.756	-0.808
	(Std)	(0.437)	(0.428)	(0.428)	(0.490)	(0.471)	(0.441)	(0.480)	(0.448)	(0.482)
[1683.02 s]	ESS	1419.504	1243.523	**1632.920**	1223.331	1179.196	1455.774	1395.415	1217.547	1171.629
Bin30	Mean	0.961	0.228	0.008	-0.048	0.238	1.075	-0.011	0.731	-0.819
	(Std)	(0.424)	(0.424)	(0.428)	(0.501)	(0.464)	(0.445)	(0.4866)	(0.4366)	(0.4644)
[2056.75 s]	ESS	1152.304	**1759.833**	1366.974	1150.885	1395.970	1422.304	1292.074	1222.696	**1539.078**

ESS: at lag equal to the lowest order at which sample autocorrelation is not significant.

Computing times (in seconds) in square brackets.

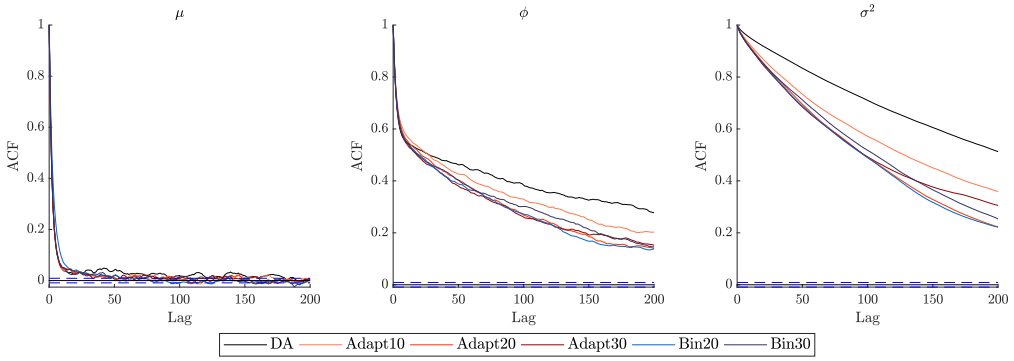
Table 4.4.6: SV model: posterior means, standard deviations and effective sample sizes (ESS) of the latent volatilities for $M = 50,000$ posterior draws after a burn-in of $10,000$.

Method		h_{450}	h_{950}	h_{1450}	h_{1950}	h_{2450}	h_{2950}	h_{3450}	h_{3950}	h_{4450}
DA	Mean	0.961	0.478	-0.163	0.011	0.448	1.026	-0.010	0.659	-0.963
	(Std)	(0.429)	(0.405)	(0.449)	(0.517)	(0.407)	(0.439)	(0.479)	(0.431)	(0.506)
[202.57 s]	ESS	650.696	420.645	540.354	276.558	621.079	633.797	506.926	348.977	346.512
Adapt10	Mean	0.955	0.454	-0.198	-0.022	0.413	1.024	-0.004	0.682	-0.984
	(Std)	(0.416)	(0.402)	(0.423)	(0.481)	(0.408)	(0.408)	(0.435)	(0.447)	(0.468)
[1511.49 s]	ESS	1550.904	1192.579	1183.693	1199.115	**1548.657**	**1860.399**	1277.721	1139.830	1089.324
Adapt20	Mean	0.920	0.470	-0.127	0.008	0.419	1.016	-0.055	0.677	-0.929
	(Std)	(0.400)	(0.405)	(0.450)	(0.489)	(0.4190)	(0.421)	(0.463)	(0.435)	(0.493)
[2039.16 s]	ESS	1708.410	1280.039	1152.046	1008.572	1493.396	1621.910	1352.884	1231.148	970.933
Adapt30	Mean	0.915	0.409	-0.136	-0.034	0.402	1.030	-0.053	0.697	-0.948
	(Std)	(0.404)	(0.398)	(0.439)	(0.486)	(0.403)	(0.430)	(0.438)	(0.451)	(0.483)
[2497.22 s]	ESS	**1860.153**	**1333.922**	**1846.106**	**1226.611**	1486.479	1549.930	**1353.398**	**1322.118**	**1205.346**

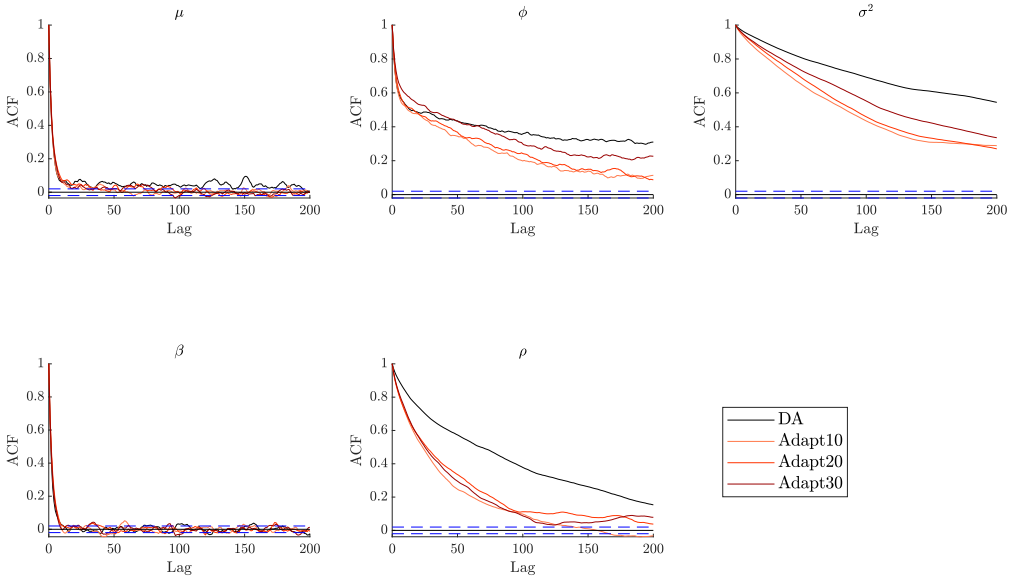
ESS: at lag equal to the lowest order at which sample autocorrelation is not significant.

Computing times (in seconds) in square brackets.

Table 4.4.7: SVML model: posterior means, standard deviations and effective sample sizes (ESS) of the latent volatilities for $M = 50,000$ posterior draws after a burn-in of $10,000$.

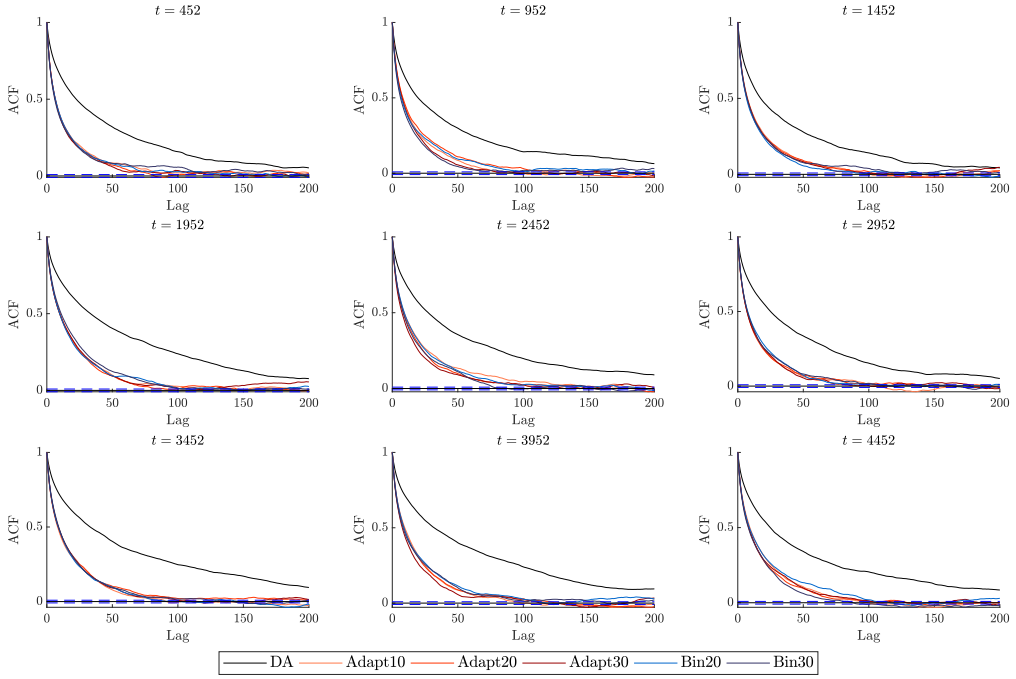


(a) SV model.

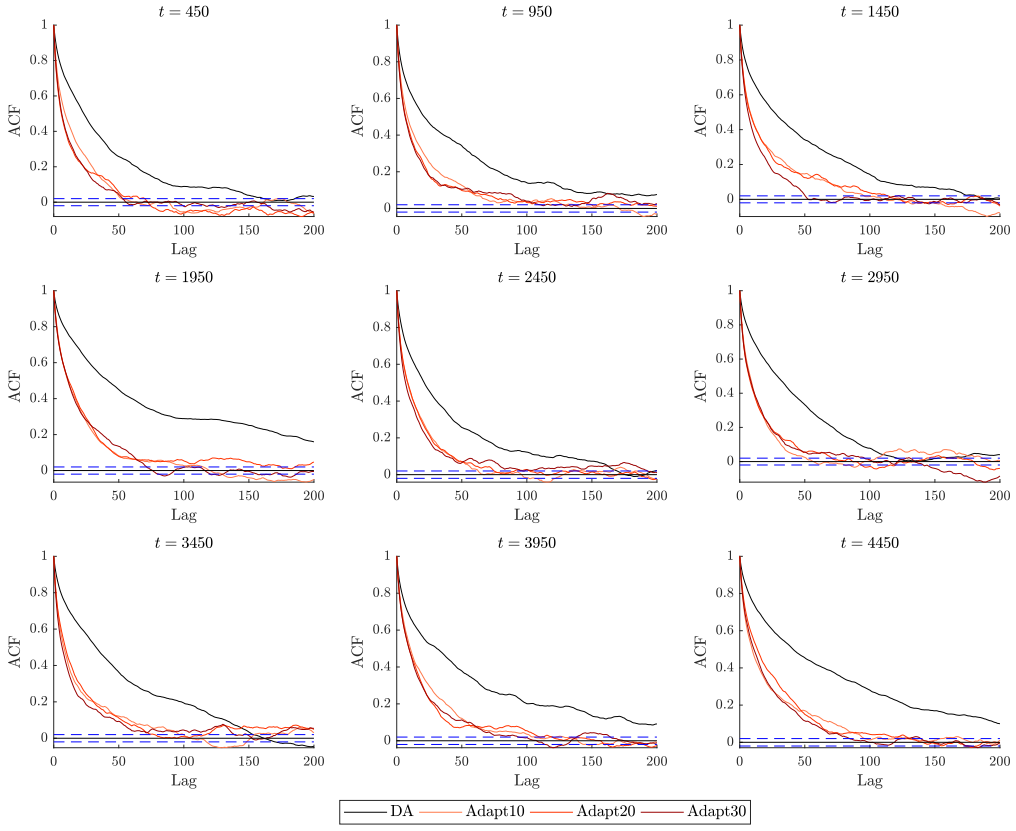


(b) SVML model

Figure 4.4.10: SV and SVML model: ACF plots for parameters.



(a) SV model.



(b) SVML model

Figure 4.4.11: SV and SVML model: ACF plots for selected volatilities.

4.5 Discussion

We have presented a new estimation method for state space models, called Semi-Complete Data Augmentation, designed to increase the efficiency of "vanilla" MCMC algorithms. The main idea behind the introduced approach is to combine data augmentation with numerical integration, where the latter aims at reducing the dependence between the imputed auxiliary variables. This concept relates to general Rao-Blackwellisation methods, however we do not require the resulting conditional distribution (given the imputed states) to be available in a closed-form (i.e. to be analytically integrable), nor the imputed auxiliary variables to be sufficient statistics for the marginalised variables.

We propose integration schemes based on the insights from hidden Markov models in the sense that we specify new transition probabilities between redefined states, to be numerically integrated out, conditionally on the auxiliary variables. Further efficiency gains can be obtained by "binning", i.e. approximating similar values of the marginalised state with e.g. a single mid-value. This results in an approximation to the semi-complete data likelihood and we note that for continuous states such an approximation is a natural starting point for our approach (as in principle for any MC based analysis). We describe two types of "binning": "fixed bins" based a pre-specified grid and "adaptive bins" based on e.g. quantiles of the relevant distribution. The latter remove the problem of specifying the "essential domain" required for fixed bins, considered by e.g. Kitagawa (1987) and Langrock et al. (2012b). Adaptive bins are also more suited for problems with highly varying integration ranges such as in the SV model with leverage and SV in the mean (SVML), for which fixed bins are unlikely to be efficient (see Sandmann and Koopman, 1998). Moreover, a specific approximation accuracy typically can be achieved by using fewer adaptive bins than fixed bins, which – given similar computing times for both approaches based on the same number of bins – means the adaptive bins require less computing time to attain an appropriate precision.

The two empirical studies considered demonstrate the gains from applying the SCDA approach compared to the general "vanilla" MCMC algorithm. For the lapwings data model the efficiency gains are substantial, not only in terms of higher effective sample sizes compared to the standard DA technique but also when taking into account the computing time (ESS/sec.). For the SV and SVML models, SCDA boosts the mixing of the Markov chains, however at the cost of an increased computing time. Nevertheless, for larger models with more complex dependency structure, such as a dynamic factor

model with double stochastic volatility, the proposed SCDA method is likely to become much more profitable – also in terms of increased ESS/sec – as discussed in Section 4.4.2. We leave these extensions for further research, focusing in the current work on illustrating the generalisability of the SCDA algorithm with the important "building block" of numerous (econometric) models.

The split of the latent states into "auxiliary" and "integrated" variables is model-dependent and should be specified in such a way that the algorithm is efficient. This is a non-unique choice and multiple approaches may be applied – the efficiency of these will be dependent on both the model and data. On the one hand, the imputed states aim to have reduced correlation, to improve mixing of MCMC algorithms; on the other hand, the numerical integration is over a very low number of dimensions, which in many cases is feasible due to conditional independence of the integration problems. To identify such conditionally independent latent states investigating of the underlying graphical structure of an SSM can be useful (cf. the concept of *d-separation* in Bayesian Networks). In general, high dimensional integration remains a challenging problem, which we leave for further research, noting that insights from the SMC samplers (Del Moral et al., 2006) could be useful in this context.

The proposed methodology naturally leads to several topics for further research. First, we aim to investigate error bounds due to using approximate approach. This should allow us to quantify the demonstrated higher usefulness of adaptive bins compared to fixed bins. Second, making use of automated methods for identifying correlation structure would make applying of the SCDA approach to new models easier and potentially more efficient. The latter can be the case if the model at hand is complex and/or there are no "natural candidates" for the ingratiated states. Finally, we expect the increased computing time recorded for the SV-type models to be reduced through parallelisation methods. Since updating a given state is associated with conditioning on only two other states, the previous one and the next one (see Figure 4.4.6), it is possible to update every second augmented state in parallel. In principle, such an approach could be also adopted in the lapwings application, however there only every fourth state updating could be used due to the second-order dependence in the associated HMM-representation.

Appendix 4.A Specification details of the HMM approximations

In this section we present how the general formulation of the HMM-based approximation to the SCDL can be applied for the examples discussed in Sections 4.3 and 2.2.

4.A.1 Motivating example from Section 4.3.2

The SSM from Figure 4.3.1 is given by

$$\begin{aligned} y_t|x_{1,t}, x_{2,t} &\sim p(x_{1,t}, x_{2,t}), \\ x_{1,t+1}|x_{1,t}, x_{2,t} &\sim p(x_{1,t}, x_{2,t}), \\ x_{2,t+1}|x_{1,t}, x_{2,t} &\sim p(x_{1,t}, x_{2,t}), \\ x_{i,0} &\sim p(x_{i,0}), i = 1, 2 \end{aligned}$$

and we aim at imputing $x_{1,t}$ and integrating out $x_{2,t}$. Since in this model $T_{int} = T_{aug} = \{0, 1, \ldots, T\}$, so that the index functions $\tau(t)$, $a(t)$ and $o(t)$ are simply identities, we skip them below to simplify the exposition. The marginal distribution[11] of the imputed state $x_{1,t}$ can be approximated as

$$\begin{aligned} p(x_{1,t}|\mathbf{x}_{1,0:t-1}) &\approx \sum_{j=1}^{B} \mathbb{P}(x_{2,t-1} \in \mathcal{B}_j|\mathbf{x}_{1,0:t-1}) p(x_{1,t}|\mathbf{x}_{1,0:t-1}, x_{2,t-1} \in \mathcal{B}_j), \\ &= \sum_{j=1}^{B} \underbrace{\mathbb{P}(x_{2,t-1} \in \mathcal{B}_j|\mathbf{x}_{1,0:t-2})}_{=:u_{j,t-1}} \underbrace{p(x_{j,t}|x_{1,t-1}, x_{2,t-1} \in \mathcal{B}_j)}_{=:p_{j,t}}, \end{aligned} \tag{4.A.1}$$

where $p_{j,t}$ is the likelihood of the augmented state at t given the imputed state at $t-1$ was in the jth bin (and previous realisations of x_{aug} but these are treated as known) and $u_{j,t-1}$ is the unconditional probability of the hidden process $x_{2,t}$ falling into the jth bin at $t-1$. This unconditional probability can be expressed as

$$u_{k,t} = \mathbb{P}(x_{2,t} \in \mathcal{B}_k|\mathbf{x}_{1,0:t-1}) = \sum_{l=j}^{B} \underbrace{\mathbb{P}(x_{2,t-1} \in \mathcal{B}_j|\mathbf{x}_{1,0:t-2})}_{=u_{j,t-1}} \underbrace{\mathbb{P}(x_{2,t} \in \mathcal{B}_k|\mathbf{x}_{1,0:t-1}, x_{2,t-1} \in \mathcal{B}_j)}_{=:\gamma_{jk,t}}, \tag{4.A.2}$$

[11] Marginal in the sense of the Markov structure, not the augmented states which we treat as known.

which leads us to the standard result in HMM that the unconditional distributions in subsequent periods are related via the transition matrix $\Gamma_t = [\gamma_{jk,t}]_{k,j=1,\dots,T}$ as follows (see Zucchini et al., 2016, p.16, 32)

$$\mathbf{u}_t = \mathbf{u}_{t-1}\Gamma_t.$$

Next, the observations are conditionally independent, hence we have

$$p(y_t|\mathbf{x}_{1,0:t}) \approx \sum_{k=1}^{B} u_{k,t} \underbrace{p(y_t|x_{1,t}, x_{2,t} \in \mathcal{B}_k)}_{=:q_{k,t}}, \tag{4.A.3}$$

with $q_{k,t}$ denoting the likelihood of the observation at t given the hidden state in the same period t falling into bin k.

Comparing (4.A.2) and (4.A.3) shows that the distributions of the same period t augmented states and "real" observations are conditioned on the latent states from different periods, i.e. $t-1$ and t, respectively. This is a consequence of the general dependence structure in SSMs. The transition matrix at t captures this change of the underlying state so that combining of all there parts (4.A.1), (4.A.2) and (4.A.3) results in

$$p(y_t, x_{1,t}|\mathbf{x}_{1,0:t-1}) \approx \sum_{j=1}^{B}\sum_{k=1}^{B} u_{j,t-1}p_{j,t}\gamma_{jk,t}q_{k,t}.$$

To compute the HMM-based approximation to the SCDL we consider *forward probabilities* $\mathbf{f}\alpha_t$ of the imputed states $x_{1,t}$ and observations y_t (see Zucchini et al., 2016, Sec. 2.3.2) defined as

$$\begin{aligned}\alpha_t &= p(x_{1,0})\mathbf{u}_0 \prod_{s=1}^{t} P_s\Gamma_s Q_s, \qquad t = 1, 2, \dots T,\\ \alpha_0 &= p(x_{1,0})\mathbf{u}_0 Q_0,\end{aligned}$$

with $\mathbf{u}_0 = \begin{bmatrix}\mathbb{P}(x_{2,0} \in \mathcal{B}_1) & \dots & \mathbb{P}(x_{2,0} \in \mathcal{B}_B)\end{bmatrix}$ being the initial distribution of the latent state and $Q_0 = \mathbb{I}$. It follows from this definition that the forward probabilities can be expressed recursively as

$$\alpha_t = \alpha_{t-1}P_t\Gamma_t Q_t$$

so that the required approximation to the SCDL being given by

$$\hat{p}_B(\mathbf{y}, \mathbf{x_1}) = p(x_{1,0})\mathbf{u}_0\alpha_t\mathbb{I}.$$

Notice that the transition matrix Γ_t is a full matrix, however in some cases, e.g. the lapwing population model, the transition matrix can take a simpler form e.g. it is "column-wise constant": $\gamma_{lk,t} = \mathbb{P}(x_{2,t} = k|\mathbf{x}_{1,0:t-1}),\ \forall l$ (each row is the same). On the other hand, the augmented observation matrix and the "real" observation matrix have a diagonal forms $P_t = \text{diag}\left(p_{j,t}\right)_{j=1,\ldots,B}$ and $Q_t = \text{diag}\left(q_{k,y}\right)_{j=k,\ldots,B}$, respectively. Using the notation introduced in Section 4.3.2 we can write

$$\hat{p}_B(\mathbf{y}, \mathbf{x_1}) = p(x_{1,0})\mathbf{u}_0 Q_0 \prod_{t=1}^{T^*} \left(P_{\tau(t)}\Gamma_{\tau(t)}Q_{\tau(t)}\right)\mathbf{1}.$$

We can verify the above results be explicitly calculating

$$\begin{aligned}
P_t\Gamma_t Q_t &= \begin{bmatrix} p_{1,t} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & p_{B,t} \end{bmatrix} \begin{bmatrix} \gamma_{11,t} & \ldots & \gamma_{1B,t} \\ \vdots & \ddots & \vdots \\ \gamma_{B1,t} & \ldots & \gamma_{BB,t} \end{bmatrix} \begin{bmatrix} q_{11,t} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & q_{BB,t} \end{bmatrix} \\
&= \begin{bmatrix} p_1\gamma_{11,t} & \ldots & p_1\gamma_{1B,t} \\ \vdots & \ddots & \vdots \\ p_B\gamma_{B1,t} & \ldots & p_B\gamma_{BB,t} \end{bmatrix} \begin{bmatrix} q_{1,t} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & q_{B,t} \end{bmatrix}, \\
\alpha_{t-1}P_t\Gamma_t Q_t\mathbf{1} &= \begin{bmatrix} \alpha_{1,t-1} & \ldots & \alpha_{B,t-1} \end{bmatrix} \begin{bmatrix} p_1\gamma_{11,t}q_{1,t} & \ldots & p_1\gamma_{1B,t}q_{B,t} \\ \vdots & \ddots & \vdots \\ p_B\gamma_{B1,t}q_{1,t} & \ldots & p_B\gamma_{BB,t}q_{B,t} \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \\
&= \begin{bmatrix} \underbrace{\sum_{j=1}^{B} \alpha_{l,t-1}p_j\gamma_{j1,t}q_{11,t}}_{=\alpha_{1,t}} & \ldots & \underbrace{\sum_{j=1}^{B} \alpha_{l,t-1}p_j\gamma_{jB,t}q_{B,t}}_{=\alpha_{B,t}} \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = \alpha_t\mathbf{1}
\end{aligned}$$

and expressing

$$\hat{p}_B(\mathbf{y}, \mathbf{x_1}) = \sum_{k_1}^{B} \cdots \sum_{k_T}^{B} p(x_{1,0})\mathbf{u}_0 \prod_{t=1}^{T} p_{k_{t-1},t}\gamma_{k_{t-1}k_t,t}q_{k_t,t}.$$

4.A.2 Lapwing population model

The approximation for the lapwings model is a special case of scheme used for the general model discussed in the Section 4.A.1, with the transition matrix Γ_t having equal rows. The hidden Markov chain is here given as $\{z_t\} = \{N_{1,t}\}$ for $t = 0, \ldots, T$ and we again set $T_{int} = T_{aug} = \{0, 1, \ldots, T\}$, so that the index functions $\tau(t)$, $a(t)$ and $o(t)$ are simply identities, The transition matrix has the form

$$\begin{aligned}\Gamma_t &= \begin{bmatrix} \mathbb{P}(N_{1,t} = b_1^* | N_{1,t-1} = b_1^*, \mathbf{N}_{a,0:t-1}) & \ldots & \mathbb{P}(N_{1,t} = b_B^* | N_{1,t-1} = b_1^*, \mathbf{N}_{a,0:t-1}) \\ \vdots & \ddots & \vdots \\ \mathbb{P}(N_{1,t} = b_1^* | N_{1,t-1} = b_B^*, \mathbf{N}_{a,0:t-1}) & \ldots & \mathbb{P}(N_{1,t} = b_B^* | N_{1,t-1} = b_B^*, \mathbf{N}_{a,0:t-1}) \end{bmatrix} \\ &= \begin{bmatrix} \mathbb{P}(N_{1,t} = b_1^* | N_{a,t-1}) & \ldots & \mathbb{P}(N_{1,t} = b_B^* | N_{a,t-1}) \\ \vdots & \ddots & \vdots \\ \mathbb{P}(N_{1,t} = b_1^* | N_{a,t-1}) & \ldots & \mathbb{P}(N_{1,t} = b_B^* | N_{a,t-1}) \end{bmatrix},\end{aligned}$$

with $b_k^* = k$, for $k = 0, \ldots, N^*$. We can see that for each column of Γ_t its elements are the same. For the augmented observation matrix P_t we have

$$P_t = \operatorname{diag}\left(p(N_{a,t} | \mathbf{N}_{a,0:t-1}, N_{1,t-1} = b_j^*)\right)_{j=1,\ldots,B},$$

so P_t and Γ_t condition on the same hidden state. The observation matrix has a simple form

$$Q_t = p(y_t | N_{a,t}) \mathbb{I}.$$

Inserting Q_t, P_t and Q_t in (4.3.6) leads to

$$\hat{p}_B(y, \mathbf{N_a}) = p(h_0) \mathbf{u}_0 Q_0 \prod_{t=1}^{T^*} P_t \Gamma_t Q_t \mathbf{1}, \tag{4.A.4}$$

where $\mathbf{u}_0 = \begin{bmatrix} \mathbb{P}(N_{1,0} \in \mathcal{B}_k) & \ldots & \mathbb{P}(N_{1,0} \in \mathcal{B}_B) \end{bmatrix}$ and $Q_0 = \mathbb{I}$. Then (4.A.4) is an HMM-based approximation to (4.4.8) converging to its true value in $B \to \infty$ and $b_B \to \infty$.

4.A.3 SV model

Basic SV model The SCDL for the basic SV model can be expressed as

$$\begin{aligned} p(y, \mathbf{h_T}|\boldsymbol{\theta}) =& p(h_0) \int p(h_1|h_0)p(y_1|h_0) \\ & \prod_{t=1}^{T^*} p(h_{2t+1}|h_{2t})p(y_{2t+1}|h_{2t+1})p(h_{2t}|h_{2t-1})p(y_{2t}|h_{2t})\, dh_1 \dots dh_{T^*}, \end{aligned} \tag{4.A.5}$$

where $T^* = \frac{T-1}{2}$ (we assume T being odd). Since we impute volatilities at even time periods the Markov chain is given by $\{z_t\} = \{h_{2t+1}\}$ for $t = 1, \dots, T^*$ and its transition matrix has the form

$$\begin{aligned} \Gamma_t &= \begin{bmatrix} \mathbb{P}(h_{2t+1} \in \mathcal{B}_1|h_{2t-1} \in \mathcal{B}_1, h_{2t}) & \dots & \mathbb{P}(h_{2t+1} \in \mathcal{B}_B|h_{2t-1} \in \mathcal{B}_1, h_{2t}) \\ \vdots & \ddots & \vdots \\ \mathbb{P}(h_{2t+1} \in \mathcal{B}_1|h_{2t-1} \in \mathcal{B}_B, h_{2t}) & \dots & \mathbb{P}(h_{2t+1} \in \mathcal{B}_B|h_{2t-1} \in \mathcal{B}_B, h_{2t}) \end{bmatrix} \\ &= \begin{bmatrix} \mathbb{P}(h_{2t+1} \in \mathcal{B}_1|h_{2t}) & \dots & \mathbb{P}(h_{2t+1} \in \mathcal{B}_B|h_{2t}) \\ \vdots & \ddots & \vdots \\ \mathbb{P}(h_{2t+1} \in \mathcal{B}_1|h_{2t}) & \dots & \mathbb{P}(h_{2t+1} \in \mathcal{B}_B|h_{2t}) \end{bmatrix}. \end{aligned}$$

We can see that the rows of Γ_t are the same, which means that the hidden states are conditionally independent given the imputed states. For the augmented observation matrix P_t we have

$$P_t = \operatorname{diag}\left(p(h_{2t}|h_{2t-1} \in \mathcal{B}_j)\right)_{j=1,\dots,B},$$

so P_t and Γ_t condition on the same hidden state. The observation matrix has the form

$$Q_t = \operatorname{diag}\left(p(y_{2t}, y_{2t+1}|h_{2t+1} \in \mathcal{B}_j, h_{2t})\right)_{j=1,\dots,B}.$$

Inserting Q_t, P_t and Q_t in (4.3.6) with $\tau(t) = 2t+1$, $a(t) = 2t$ and $o(t) = \{2t, 2t+1\}$ leads to

$$\hat{p}_B(y, \mathbf{h_T}) = p(h_0)\mathbf{u}_0 Q_0 \prod_{t=1}^{T^*} P_t \Gamma_t Q_t \mathbf{1}, \tag{4.A.6}$$

where $\mathbf{u}_0 = \begin{bmatrix} \mathbb{P}(h_1 \in \mathcal{B}_k|h_0) & \dots & \mathbb{P}(h_1 \in \mathcal{B}_B|h_0) \end{bmatrix}$ and $Q_0 = \operatorname{diag}\left(y_1|h_1 \in \mathcal{B}_k\right)_{k=1,\dots,B}$. Then (4.A.6) is an HMM-based approximation to (4.A.5) converging to its true value in $B \to \infty$ and $b_0 \to -\infty$, $b_B \to \infty$.

SVML model For the SVML model we only need to adjust the matrices P_t and Q_t as the dependence structure of the observations remains unchanged

$$\Gamma_t = \begin{bmatrix} \mathbb{P}(h_{2t+1} \in \mathcal{B}_1 | h_{2t}, y_{2t}) & \dots & \mathbb{P}(h_{2t+1} \in \mathcal{B}_B | h_{2t}, y_{2t}) \\ \vdots & \ddots & \vdots \\ \mathbb{P}(h_{2t+1} \in \mathcal{B}_1 | h_{2t}, y_{2t}) & \dots & \mathbb{P}(h_{2t+1} \in \mathcal{B}_B | h_{2t}, y_{2t}) \end{bmatrix},$$
$$P_t = \operatorname{diag}\left(p(h_{2t} | h_{2t-1} \in \mathcal{B}_j, y_{2t-1})\right)_{j=1,\dots,B}.$$

Appendix 4.B Lapwings dataset

The lapwings dataset plays an important role in statistical ecology and has served as an illustration in several handbooks (see King, 2011; King et al., 2010) and papers (e.g. Besbeas et al., 2002) in this field. It was also used as an example of a complex statistical model by e.g. Goudie et al. (2018). One of the main reasons for such a particular interest in this species is a sharp decline in its population in recent years: its European population is considered as *near threatened* by International Union for Conservation of Nature (2018), while in Britain in particular it has been moved to the *red list* of species of conservation concern, see The Royal Society for the Protection of Birds (2018) (i.e. of the highest conservation priority, with species needing urgent action) from the *amber list* (mentioned by previous literature, see Besbeas et al., 2002; Brooks et al., 2004). The state of its population is crucial as it serves as an indicator species for other farmland birds, giving us an insight into the dynamics of similar bird species.

We follow the approach of Besbeas et al. (2002) and use three datasets for the lapwings application: the count census data for the population index, the weather data on the number of frost days, and the ring-recovery data. Combining independent sources of data underlies the integrated population modelling (IPM) framework and allows for a more precise parameter estimation. This is due to the survival parameters $\alpha_i, \beta_i, i \in 1, a$ being common to the state space model for the census data and to the ring-recovery model

Census data The census data are derived from the Common Birds Census (CBC) of the British Trust for Ornithology, which recently has been replaced by the Breeding Bird Survey. The dataset is constructed as annual estimates of the number of breeding female lapwings based on annual counts made at a number of sites around the UK.

Since only a small fraction of sites are surveyed each year, the index can be seen as a proxy for the total population size. For comparability, we use the same time span as Brooks et al. (2004) and King (2011), i.e. from 1965 to 1998. The choice of the starting year is there motivated by the fact that in earlier years the index protocol was being standardised. Finally we note that year 1965 is associated with time index $t = 3$, for consistency with the ring-recovery data (to be discussed below) which start in 1963.

Weather data For bird species there is a natural relationship between the survival probabilities and the weather conditions, most importantly winter severity. Following Besbeas et al. (2002) we measure this factor for year t by the number of days between April of year t and March of year $(t+1)$ inclusive in which the temperature in Central England fell below freezing and denote it by $fdays_t$. We further normalise $fdays_t$ to obtain f_t which we use as a regressor in the logistic regression for the survival probabilities. As noted by King (2011), normalisation of covariates is done to improve the mixing of the sampling scheme and to facilitate the interpretation of the parameters of the logistic regression (intercept and slope).

Ring–recovery data Ring–recovery studies aim at estimating demographic parameters of the population under consideration including first-year survival probabilities, adult survival probabilities and mortality probabilities (referred to as 'recovery' probabilities). These studies consist in marking individuals (e.g. with a ring or a tag) at the beginning of period t and then releasing them. In subsequent periods $t+1, t+2, \ldots$ the number of dead animals is recorded, where it is assumed that any recovery of a dead animal is immediate. For lapwings, the ringed birds are chicks ("fist-years") and a "period" corresponds to a "bird year" i.e. 12 months from April to March. We analyse the ring-recovery data for the releases from 1963 to 1997, with the recoveries up to 1998.

Ring-recovery data are stored in an array, an example of which is provided in Table 4.B.1. The first column corresponds to the number of ringed animals in a given year R_t, $t = 1, \ldots, T$, and the subsequent columns report the number of recovered rings $m_{t,s}$ (i.e. animals found dead) in each following year s, $s = 1, \ldots, S$. Obviously, $m_{t,s} = 0$ for $t > s$. Finally, there is an additional $(S+1)$th column, with the entries $m_{t,S+1}$ providing the number of individuals ringed in year t but never seen again (their rings are not recovered), $m_{t,S+1} = R_t - \sum_{s=1}^{S} m_{t,s}$.

The parameters of interest are $\phi_{1,s}$, $\phi_{a,s}$ and λ_s. The former two are the conditional probabilities of survival until year $s+1$ of a first-year and an adult, respectively, given

Year of Ringing	Number Ringed	Year of Recovery										
		1964	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974
1963	1147	14	4	1	2	1	0	1	1	0	0	0
1964	1285		20	3	4	0	1	1	0	0	0	0
1965	1106			10	1	2	2	0	2	2	1	1
1966	1615				9	7	4	2	1	1	0	0
1967	1618					12	1	6	2	0	0	1
1968	2120						9	6	4	0	2	2
1969	2003							10	8	5	3	1
1970	1963								8	3	2	0
1971	2463									4	1	1
1972	3092										7	2
1973	3442											15

Table 4.B.1: A fragment of Ring-Recovery Data for lapwings for the years 1963-1973, table from King (2011).

such an individual is alive in year s. The latter one is the conditional probability of ring recovery in year s given an individual dies in year s. Let $\mathbf{v} = \{v_s\}_{s=1}^{S-1}$ denote a vector of a variable $v_s \in \{\phi_{1,s}, \phi_{a,s}, \lambda_s\}$. Then each row $\mathbf{m}_t$ of the m-array is multinomially distributed: $\mathbf{m}_t \sim \mathcal{MN}(R_i, \mathbf{q}_i)$ ($\mathcal{MN}$ denotes the multinomial distribution), where $\mathbf{q}_i$ are the multinomial cell probabilities specified for $s = 1, \ldots, S$ as[12]

$$q_{t,s} = \begin{cases} 0, & t > s, \\ (1 - \phi_{1,t})\lambda_s & t = s, \\ \phi_{1,t}\lambda_s(1 - \phi_{a,s-1}) \prod_{k=1}^{j-2} \phi_{a,k}, & t > s \end{cases}$$

ans for $s = S + 1$ as $q_{t,s} = 1 - \sum_{s=1}^{S} q_{t,s}$.

The likelihood of the m-array is then given by

$$p(\mathbf{f}m|\phi_{\mathbf{1}}, \phi_{\mathbf{a}}, \lambda) \propto \prod_{t=1}^{T} \prod_{s=1}^{S+1} q_{t,s}^{m_{t,s}}.$$

The array $\mathbf{m} = [m_{t,s}]_{t=1,\ldots,T}^{s=1,\ldots,S+1}$ is a sufficient statistic for ring-recovery data.

Appendix 4.C Conditional state distribution for the SV model with leverage

Following Zucchini et al. (2016), we aim at deriving the conditional distribution of h_{t+1} given $\boldsymbol{\theta}$, h_t and y_t. Below, we skip $\boldsymbol{\theta}$ in the conditioning to simplify notation. Since

[12]For $j - 2 < t$ we put $\prod_{k=1}^{j-2} := 1$.

$y_t = \exp(h_t/2)\varepsilon_t$, after demeaning (by $\mu + \phi(h_t - \mu)$), this is the distribution of η_t given h_t and ε_t.

The distribution of $\eta_t|\varepsilon_t$ can be obtained using the basic result from multivariate normal regression, which we recall below for convenience:

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix} \right) \Rightarrow x|y \sim \mathcal{N}\left(\mu_x + \frac{\sigma_{xy}}{\sigma_y^2}(y - \mu_y), \sigma_x^2 - \frac{\sigma_{xy}^2}{\sigma_y^2} \right).$$

Hence, we obtain

$$\eta_t|\varepsilon_t \sim \mathcal{N}\left(0 + \frac{\rho\sigma}{1}(y - 0), \sigma^2 - \frac{\rho^2\sigma^2}{1} \right) = \mathcal{N}(\rho\sigma\varepsilon_t, \sigma^2(1 - \rho^2))$$

so that

$$h_{t+1}|h_t, \varepsilon_t \sim \mathcal{N}\big(\mu + \rho(h_t - \mu) + \rho\sigma\varepsilon_t, \sigma^2(1 - \rho^2)\big).$$

Finally, we can express the latter in terms of the actual observation y_t rather than the unobserved disturbance ε_t. For the basic SV model this becomes

$$h_{t+1}|h_t, \varepsilon_t \sim \mathcal{N}\left(\mu + \rho(h_t - \mu) + \rho\sigma\frac{y_t}{\exp(h_t/2)}, \sigma^2(1 - \rho^2) \right),$$

which is the result reported in Section 4.4.2, while for the SVM we have

$$h_{t+1}|h_t, \varepsilon_t \sim \mathcal{N}\left(\mu + \rho(h_t - \mu) + \rho\sigma\frac{y_t - \beta\exp(h_t)}{\exp(h_t/2)}, \sigma^2(1 - \rho^2) \right).$$

Chapter 5

Forecast Density Combinations of Dynamic Models and Data Driven Portfolio Strategies

The problem of asset allocation has long been an area of intense interest for both investment practitioners and academics. A well established approach to portfolio construction relies on the series of seminal papers by Fama and French (Fama and French, 1992, 1993, 2015). The traditional factor models proposed by them rely on macro or firm specific factors to explain expected pay-offs of financial assets. More complex factor models have also been broadly used, e.g. the dynamic factor model with stochastic volatility of Aguilar and West (2000). However, return forecasts from (static or dynamic) models do not directly lead to a practical policy tool for investors, i.e. to a decision which portfolio strategy to follow. A standard practice in portfolio management is based on realised returns from different portfolio strategies so that the best performing one is selected. Such an approach is thus not backed by any modelling rigour and obviously cannot provide any uncertainty quantification. It would be preferable, therefore, to incorporate a particular portfolio strategy in the modelling framework. However, this typically requires a specific model-based strategy such as mean-variance optimization, see e.g. Winkler and Barry (1975), and a specific utility function for the investor, see e.g. Aguilar and West (2000). We propose to overcome these shortcomings by directly connecting forecasts from an "appropriately specified set of models" with a set of "data-driven portfolio strategies", without the need to specify a separate scoring function like a utility or loss function. We note that data-driven portfolio strategies have also been analysed by Garlappi et al. (2006) and DeMiguel et al. (2007), yet our

approach differs from theirs as we aim at considering sets of models as well as strategies.

To specify the above-mentioned "appropriate set of models", we start with a scrupulous investigation of typical stylised facts of the time series of monthly returns of 10 US industries over the period 1926M7–2015M6. The findings of this analysis allow us to define a general class of models, with different short and long-run dynamics, which extends the factor-augmented vector autoregressive model of Bernanke et al. (2005). The "data-driven portfolio strategies" refer to the basic practice in financial investment that one invests in the "winner" industry and goes short in the "loser" industry, corresponding to the industries with the highest and lowest cumulative returns in past periods. I.e. one aims to take advantage of a positive or a negative "momentum" in returns of particular industries. Combining of models and strategies in a single modelling framework is achieved by using a mixture of alternative models and alternative portfolio strategies, which we represent in probabilistic terms as a density combination of model forecasts and strategy returns. The combination weights are defined through feedback mechanisms that enable learning, to allow for cross-correlation and correlation over time. Our approach can be seen as an extension of the mixture of experts analysis of Jacobs et al. (1991); Jordan and Jacobs (1994); Jordan and Xu (1995); Peng et al. (1996). Further, we allow for model and strategy incompleteness. This enables us to study misspecification effects through diagnostic analysis of economic results and posterior residuals. This, to the best of our knowledge, novel methodology provides dynamic asset-allocations using a learning period for optimal weights at every decision period.

The proposed *Forecast Density Combination* (FDC) scheme generalised the approach of Billio et al. (2013) by including sets of model forecasts and strategy returns. This, together with a fully Bayesian inference over the resulting model, allows us for the quantification of uncertainty from multiple sources, which is important from a risk management perspective. In other words, our approach supplies an investor with policy recommendations about different portfolio scenarios in which the returns uncertainty is explicitly incorporated. Hence, an investor has full information about his/her portfolio, including e.g. the Value-at-Risk estimates, which is not provided by merely standard point forecasts.

Bayesian inference over the proposed FDC model is based on its representation as a nonlinear non-Gaussian state space model. The complexity of the system under consideration brings a challenge in terms of estimation efficiency and robustness as well as the amount of computing time, particularly in the case of a large number of models and strategies. To overcome these difficulties, we introduce a novel nonlinear, non-

Gaussian filter called *MFilter*, which is embedded in the density combination procedure. MFilter is based on mixtures of Student's t distribution obtained with the MitISEM algorithm proposed by Hoogerheide et al. (2012) and further developed in Baştürk et al. (2016) and Baştürk et al. (2017).

To validate the practical usefulness of the developed FDC method, we investigate its performance using the data on 10 US industry returns over the period 1926M7–2015M6. We draw three main conclusions. First, we obtain evidence that averaging over density combinations of sets of model forecasts and strategy returns pays off in terms of expected return and risk features. Forecasts from model sets help to improve expected return while incorporating strategy sets in forecasting helps to reduce risk features. Basic model structures and strategies with fixed weights perform worse in terms of expected return and Sharpe ratio. Second, we demonstrate that the dynamic patterns of the weights in these combinations differ in tranquil and more volatile periods. For this reason, even basic learning mechanisms adopted for the weight estimation should be useful for obtaining a portfolio bespoke to the current market trends. Third, there exist adverse effects of misspecification of the model and strategy set on the results. Diagnostic learning based on the posterior residual patterns as well as on the related economic implications can lead to improved modelling and policy making. We emphasize that our empirical results are conditional upon the employed information set which consists of the dataset and the selected model and strategy sets. The results of our empirical analysis contain informative signals about the scenarios resulting from alternative portfolio policies, which can be useful for practitioners such as investment companies.

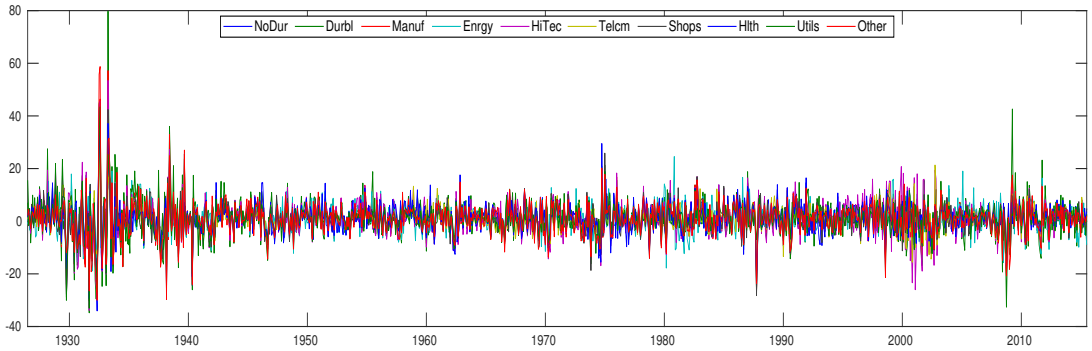
The remainder of this chapter is organised as follows. In Section 5.1 we first discuss stylised facts about 10 US industry returns. This naturally leads us to the specification of the general dynamic model capturing these features. Section 5.2 introduces the concept of data-driven portfolio strategies and discusses several potential choices of such strategies. We then describe in Section 5.3 how to combine strategy returns with model forecasts in the extended FDC scheme and we introduce the computational tool for its inference (MFilter). Section 5.4 illustrates the usefulness the developed FDC method based on an empirical application to 10 US industry returns. Section 5.5 concludes with a discussion. We provide additional results in Appendices 5.A–5.D.

5.1 Modelling of US industry returns based on stylised facts

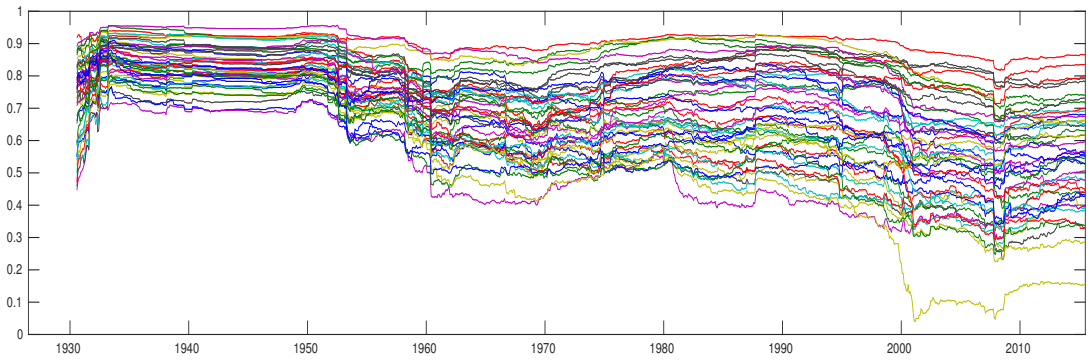
Traditional factor models rely on macro or firm specific factors to explain expected pay-offs of financial assets, see Fama and French (1992, 1993, 2015). However, several stylised facts about asset returns which we discuss in detail below, such as a stationary auto-regressive pattern, strong time-varying cross-sectional correlations between series and volatility clustering, suggest that these traditional non-dynamic models may not be flexible enough to accurately explain the data. Therefore, various specifications of dynamic factor models (DFMs) have been proposed in the literature. Allowing for different long and short-run dynamics of returns, the DFMs have been shown useful in capturing some properties of return series, see Ng et al. (1992), Quintana et al. (1995), Aguilar and West (2000) and Han (2006), among several others. The basic specification of a DFM has been later incorporated in the vector autoregressive model (VAR) framework to form a class of factor-augmented VAR model (FAVAR), see Bernanke et al. (2005) and Stock and Watson (2005). FAVAR have been applied for portfolio construction in e.g. Aguilar and West (2000), Talih and Hengartner (2005), Engle and Colacito (2006), Carvalho et al. (2011) and Zhou et al. (2014).

To specify a more convenient and flexible model structure, we first carry out a careful analysis of the time series of monthly returns of 10 US industries over the period 1926M7–2015M6, which amounts to 1069 observations on the vector of returns[1]. Figure 5.1.1a illustrates monthly returns of 10 industries, where the industries are abbreviated as follows: "NoDur" for consumer non-durables, "Durbl" for consumer durables, "Manuf" for manufacturing, "Enrgy" for oil, gas, and coal extraction and products, "HiTec" for business equipment, "Telcm" for telephone and television transmission, "Shops" for wholesale, retail, and some services, "Hlth" for health care, medical equipment, and drugs, "Utils" for utilities, "Other" for other industries. Next, 45 pairwise correlations of the 10 industry returns are presented in Figure 5.1.1b, while 4 principal components are shown in Figure 5.1.1c. We compute both pairwise correlations and principal components based on moving windows with 240 monthly observations. For the initialisation, we use the first 50 observations as the initial sample (so that the first correlations are computed for the returns over the period 1926M7–1930M8), which we then expand until observation 240 (1946M6).

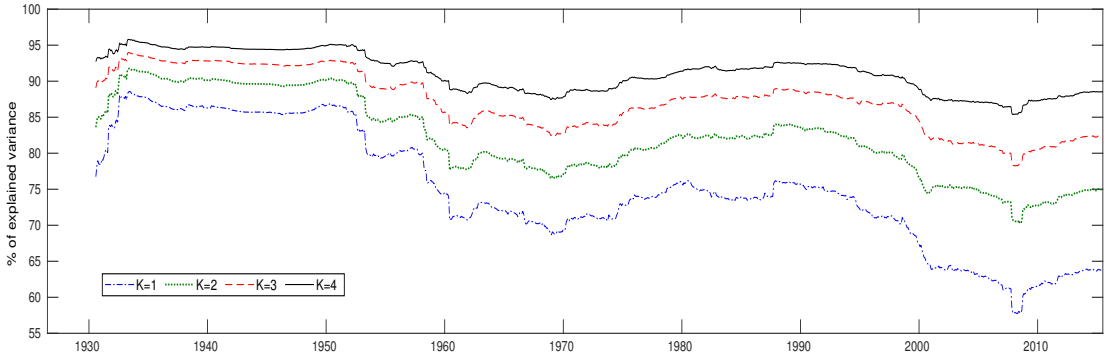
[1]The industry returns are constructed by equally weighting all the stock returns in a specific industry, which is similar to Moskowitz and Grinblatt (1999). The data are retrieved from http://mba.tuck.dartmouth.edu/pages/faculty/ken.french on 24/10/2015.



(a) Monthly percentage returns.



(b) Pairwise correlations.



(c) Percentage of explained variation by PCA.

Figure 5.1.1: Features of 10 US industry portfolios 1926M7–2015M6.

Figures 5.1.1a–5.1.1c imply at least four stylised facts. In the top figure, all the return series exhibit a stationary autoregressive pattern and clear volatility clustering. Strong cross-sectional correlations between the returns are visible in the middle figure, with correlations being heavily time-varying. The bottom figure indicates that the total variation in the series is well captured with as few as one to four principal components. However, there is a time-varying pattern in the percentage of explained variation.

Given these typical data features, we consider several dynamic models with distinct short and long-run dynamics and allow for disturbance distributions. All the models considered in our analysis are members, or combinations of members, of the FAVAR class extended to include stochastic volatility (SV) of the idiosyncratic disturbances (FAVAR-SV) and can be expressed in the following form

$$\begin{aligned} \mathbf{y}_t &= \boldsymbol{\beta}\mathbf{x}_t + \Lambda\mathbf{f}_t + \boldsymbol{\varepsilon}_t, & \boldsymbol{\varepsilon}_t &\sim \mathcal{N}(0, \Sigma_t), \\ \mathbf{f}_t &= \phi_1\mathbf{f}_{t-1} + \cdots + \phi_L\mathbf{f}_{t-L} + \boldsymbol{\eta}_t, & \boldsymbol{\eta}_t &\sim \mathcal{N}(0, \mathrm{Q}), \end{aligned} \tag{5.1.1}$$

where the dependent variable $\mathbf{y}_t = (y_{1,t}, \ldots, y_{N,t})'$ is the $N \times 1$ vector of industrial portfolio returns, with $y_{i,t}$ denoting the return from industry i at time t, and the time series runs from $t = 1, \ldots, T$. The $C \times 1$ vector of predetermined variables $\mathbf{x}_t$ may contain explanatory variables as well as lagged dependent variables. The $K \times 1$ vector $\mathbf{f}_t$ contains unobservable factors, where ϕ_j for $j = 1, \ldots L$ is a $K \times K$ matrix of autoregressive coefficients at lag j. Λ is an $N \times K$ matrix of factor loadings. In addition, we define a time-varying variance-covariance matrix for the idiosyncratic disturbances, Σ_t, and a fixed covariance matrix for the factor disturbances, Q. In all specifications Σ_t is a diagonal matrix[2].

Different short and long-run dynamic behaviour of member models of the FAVAR-SV class is obtained by specifying different assumptions regarding the predetermined variables $\mathbf{x}_t$, the factor structure $\mathbf{f}_t$, the idiosyncratic disturbances and the factor disturbances. The basic DFM assumes $\boldsymbol{\beta} = 0_{(N \times C)}$, a normal distribution for the idiosyncratic disturbances and the factor disturbances with time-invariant variance-covariance matrices. Another basic model is the VAR model, which is obtained by letting $\Lambda = 0_{(N \times K)}$, defining $\mathbf{x}_t$ as the lagged dependent variable and a time-invariant variance-covariance matrix of the disturbances. The standard SV model results from setting $\boldsymbol{\beta} = 0$ and $\Lambda = 0$. We provide more details on the model specification in Ap-

[2]We have also estimated models with a Student's t distribution and/or a time varying covariance matrix, Q_t. Both extensions led to overfitting and poor empirical and forecasting results. We have therefore skipped these models in our final analysis. Particularly for the latter case, we acknowledge that the MCMC sampler can be improved, see Kastner et al. (2017).

pendix 5.A together with the corresponding prior specification and Bayesian estimation procedures.

In our empirical analysis in Section 5.4, we compare the performance of alternative combinations of models for forecasting and portfolio analysis. We start with exploring the contribution of each of the three basic models (DFM, VAR and SV) separately as well as in combination. As the next step, we investigate combinations of more flexible models, DFM-SV and VAR-SV. Finally, the general FAVAR-SV class is investigated.

As a final remark we emphasise that the general model (5.1.1) is not identified without further parameter restrictions due to both the factors $\mathbf{f}_t$ and the loading matrix Λ being unknown. This can be seen by expressing the factor part of the observation process in (5.1.1) as

$$\mathbf{f}_t\Lambda = \mathbf{f}_t\mathrm{R}\mathrm{R}^{-1}\Lambda,$$

in which the left hand side and the right hand side are observationally equivalent for any $K \times K$ invertible matrix R. Such a matrix R has K^2 free parameters hence at least K^2 restrictions are needed for the model to be identified, see Geweke and Zhou (1996), Lopes and West (2004), Bai and Peng (2015) and Frühwirth-Schnatter and Lopes (2018). For all the models we consider, we follow the identification scheme of Lopes and West (2004) and assume diagonal covariance matrices. We refer to Chan et al. (2018) and Kaufmann and Schumacher (2017) for more recent treatment of identification for this class of models.

5.2 Data-driven portfolio strategies

The dynamic models specified in Section 5.1 provide a flexible and useful tool for describing and forecasting financial returns. However, as pointed out in the introduction, such forecasts do not directly lead to a practical policy tool for investors, i.e. to a decision which portfolio strategy to follow. Below we discuss how model-based predictions can be directly connected with portfolio strategies. We focus on data-driven portfolio strategies which have an advantage of not depending on a particular scoring function, such as a utility or loss function. We note that data-driven portfolio strategies have also been analysed by Garlappi et al. (2006) and DeMiguel et al. (2007), yet our approach differs from theirs as we consider sets of models as well as strategies.

Standard momentum As a benchmark data-driven portfolio strategy, we consider the so-called standard industry momentum (SM), discussed e.g. by Jegadeesh and Titman (1993), Chan et al. (1996) and Jegadeesh and Titman (2001). It is not based on any model structure but directly makes use of typical momentum patterns in a return time series. The practice is that one invests in the "winner" industry and goes short in the "loser" industry, which correspond to the industries with the highest and lowest *cumulative returns* in, say, the past 12 periods. The selected momentum breakpoints correspond to e.g. 90% and 10% quantiles for the industries in a portfolio, and these values can be adjusted for alternative momentum strategies. The economic motivation behind this strategy is to capture market trends in industry returns.

Next, we consider two portfolio strategies sharing the main concept of SM strategy but being directly connected with in-sample forecasts from a model or a set of models. We note that our approach can be generalised to a broader selection of model-based portfolio strategies, such as those analyses in Gruber and West (2017).

Model based momentum The model-based momentum (MM) strategy is based on the *fitted industry returns* in the past period from one of the models or a set of models from Section 5.1. It prescribes to go long in the industry with the highest fitted returns and go short in the industry with the lowest fitted returns. With ten industries under consideration, this corresponds to 90% and 10% quantiles of fitted returns. The MM strategy in this case is similar to the SM strategy where the portfolio return $\tilde{r}_{t+1}$ is now given as the weighted sum:

$$\tilde{r}_{t+1} = \sum_{n=1}^{N} \tilde{y}_{n,t+1} \omega_{n,t}, \tag{5.2.1}$$

where $\tilde{y}_{n,t+1}$ is a draw from the one-period-ahead forecast distribution of the nth industry's return $y_{n,t+1}$[3]. The weights are given as

$$\omega_{n,t} = \begin{cases} 1 & \text{if } \bar{y}_{n,t} = \max_n\{\bar{y}_{1,t}, \dots, \bar{y}_{N,t}\}, \\ -1 & \text{if } \bar{y}_{n,t} = \min_n\{\bar{y}_{1,t}, \dots, \bar{y}_{N,t}\}, \\ 0 & \text{otherwise,} \end{cases} \tag{5.2.2}$$

where $\bar{y}_{n,t}$ is the average of the fitted mean returns of the nth industry over last 12 periods, including time t.

[3]Note that here we specify a draw from the one-period ahead forecast distribution of the portfolio return. Realised returns can also be calculated alternatively using observed returns instead of $\tilde{y}_{n,t+1}$.

To our knowledge, such a model-based momentum strategy has not been considered in the literature, even though it is a natural extension of the SM strategy. We emphasize that since the weights are a nonlinear function of random variables $\bar{y}_{n,t}$, our Bayesian inference procedure allows us to fully take into account the underlying model and parameter uncertainty.

Residual based momentum Next, we consider a model-based residual momentum (RM) strategy. To construct a portfolio based on this strategy, we use the fitted asset returns from the past period and invest in the assets with the highest *unexpected* returns and go short in assets with the lowest *unexpected* returns. Unexpected returns in this strategy correspond to the *model residuals* at the investment decision time. This strategy can be seen as an extension of the approach of Blitz et al. (2011), who sort the returns based on past 12 residuals from the Fama-French factor model. The assets with unexpectedly high or low residuals are given a positive or negative weight, respectively. The proposed RM strategy follows the same intuition but we do restrict our analysis to the Fama-French factor model and allow for any model specification from Section 5.1. Similarly to the MM strategy, the constructed industry portfolio is a weighted sum of 10 industry returns, with the weights now computed as

$$\omega_{n,t} = \begin{cases} 1 & \text{if } \bar{\varepsilon}_{n,t} = \max_n\{\bar{\varepsilon}_{1,t}, \ldots, \bar{\varepsilon}_{N,t}\}, \\ -1 & \text{if } \bar{\varepsilon}_{n,t} = \min_n\{\bar{\varepsilon}_{1,t}, \ldots, \bar{\varepsilon}_{N,t}\}, \\ 0 & \text{otherwise,} \end{cases} \tag{5.2.3}$$

where $\bar{\varepsilon}_{n,t}$ is the average of the residuals for the nth industry return over last 12 periods, including time t. The difference between the MM and RM strategies is the use of fitted residuals indicating unexpected returns in the latter case. This can be interpreted as an error correction mechanism in which portfolio weights adjust according to the deviation of the last periods' industry returns from the fitted industry return distribution.

The two proposed model-based strategies can be seen as complimentary policies, targeted at the explained and the unexplained parts of the returns, respectively. More precisely, MM follows the market trends explained by the systematic component, such as common factors, while RM builds on return patterns that relate to the unexplained component, i.e. RM can serve as a "correction mechanism" when the underlying model of returns fails to represent all market dynamics. Hence, these two strategies, when conditioned on each model, or a set of models, and combined, span a space of plausible profitable policies to follow. Moreover, they have the advantage of providing an

economic intuition of capturing estimated market trends.

Equally weighted portfolios In the proposed FDC scheme, as discussed in the next section, the model-strategy pairs are combined with the combination weights which need to be inferred. As a simpler approach we can consider a similar combination of models and strategies but with all the weights being equal. This results in an equally weighted portfolio of combined models and strategies and can serve as an additional benchmark to the SM strategy. We note that a portfolio constructed this way differs from a model-and-strategy-free equal weight portfolio, which as a non-model-based approach does not provide a measure of uncertainty. In our model-based equally weighted strategy, we allocate an equal weight $\frac{1}{M \times S}$ to each portfolio resulting from a model-strategy pair specified in (5.2.2) and (5.2.3) and we borrow at the risk-free rate in the sense that the 1-month Treasury bill rate gets weight -1. Since the portfolio weights in (5.2.2) and (5.2.3) sum up to 0, the equally weighted portfolio weights also sum up to 0. The purpose of considering this equally weighted portfolios is to identify the importance of time-variation in model and portfolio strategy performances.

Remark on minimum variance strategy We have experimented with the minimum variance (MV) strategy, given that it is widely used in applications and directly relates to the forecasts of asset returns, volatilities and co-volatilities. However, we have decided not to explicitly include the results of the MV strategy in our empirical exercise since the realised returns from this strategy were turned out to be unstable for all models. This can be attributed to the estimation uncertainty and potential ill-conditioning in variance-covariance matrix estimates, see also Michaud (1989). To fairly include the results of this strategy, one would require more structured or "sparse" variance-covariance matrix estimation as in Kaufmann and Schumacher (2017). This is left as a topic for further research.

5.3 Weights estimation of Forecast Density Combination

To combine models from the general class specified in Section 5.1 with the data-driven strategies discussed in Section 5.2 we build upon the approach of Billio et al. (2013) for predictive densities combination. The proposed FDC scheme extends the one of the cited authors by explicitly incorporating portfolio strategies in the analysis. It also

relates to the literature on dynamic prediction pools proposed in Geweke and Amisano (2010b), Waggoner and Zha (2012) and Del Negro et al. (2016), however we make use of a different law of motion for the combination weights. Aastveit et al. (2018b) provide a survey on the evolution of the density combination approach to forecasting in economics.

The rationale behind the proposed methodology is the common practice in macroeconomic and financial forecasting of using a weighted combination of forecasts from many sources, e.g. models, experts and/or large micro-data sets. One then deals with three groups of variables: forecasts from different models, weights to combine these, and the variable of interest which is forecasted. The density combination approach gives this practice a probabilistic foundation by using three types of densities:

FDC.1 forecast densities for different models,
FDC.2 a weight density,
FDC.3 a combination density.

This allows us for the quantification of uncertainty related to the properties of the implied distributions. In our case, we are mostly interested in mean returns, volatilities and risk of large losses.

Our focus in this section is threefold. First, we explain the time-line of model estimation and portfolio construction, in which we distinguish four specific periods. Each of these periods implies different return variables of a portfolio strategy.We note that in the standard FDC approach the forecast densities from different models are combined to form a single forecast density of the observed variable of interest (e.g. GDP growth or inflation) in some optimal way. In our case, we deal with several constructed return variables and we discuss how and when to use their densities in different periods of the process timeline. Second, we show how the proposed FDC of model forecasts and strategy returns can be represented as a nonlinear, non-Gaussian state space model (SSM). In the general case, such an SSM is analytically intractable and needs to be analysed using simulation-based methods. Therefore, in the third step, we adopt numerical methods based on Bayesian sampling-based filtering to conduct inference about the resulting system. Given the computational complexity of the problem, we introduce a novel, efficient and robust filtering method, which we call the *MFilter*. This leads to a substantial reduction in computation time and allows us to parallelise the computations. We refer to Appendix 5.B for technical details of the algorithm.

5.3.1 Timeline of model estimation, construction and portfolio holding

Extending the approach of Billio et al. (2013) to allow for the joint analysis of models and strategies requires us to decide what subsamples of the whole dataset are used for certain purposes. I.e. we need to set up a timeline for (1) the model estimation and forecasting, (2) industry portfolios construction, (3) combining of the models and strategies in our FDC scheme and, finally, (4) the actual portfolio holding over a fixed period of time (to yield the realised return). Figure 5.3.1 presents our timeline specification for these four periods.

In the first two periods, `[t0-t2]` indicated at the top of Figure 5.3.1, the M models under consideration are estimated annually in the month of June using the preceding 240 monthly observations. This results in a set of fitted returns and the corresponding residuals denoted by $\tilde{y}_{n,m,t}$ and $\tilde{\varepsilon}_{n,m,t}$, respectively, for $m = 1, \ldots, M$ models and $n = 1, \ldots, N$ industries.

In the second period, `[t1-t2]` in Figure 5.3.1, we use the fitted returns and residuals to form S investment strategies for each model. For each model-strategy pair we then form the weights which are based on the implied portfolio performances in the last 12 months, including June. Such a strategy formation is similar to Jegadeesh and Titman (1993) and Fama and French (1993), as we construct industry weights $\omega_{n,m,s,t}$ for industry n, model m and strategy s at time t, at the end of a *skip month*, July[4]. Using (5.2.1), we specify **FDC.1** by expressing the one-period-ahead forecast of the portfolio return from strategy s and model m at time $t+1$ as

$$\tilde{r}_{m,s,t+1} = \sum_{n=1}^{N} \tilde{y}_{n,m,t+1} \omega_{n,m,s,t}. \tag{5.3.1}$$

We re-emphasize that our extension of the FDC approach includes an important difference compared to the standard one. In the latter case one compares the one-period-ahead forecast distribution of return, $\tilde{r}_{m,s,t+1}$, with the density of the variable of interest which is observable. In our case, we define the variable of interest, r_t, as the actual return obtained from investing one unit in the asset with maximum return and disinvesting from the asset with minimum return. This is not observed ex ante. We define

[4]In the literature, the skip month is often used to remove market micro-structure effects, see Asness et al. (2013). Our empirical results are robust to using the month of June for obtaining forecasts and keeping July as the skip month. The portfolio is held for 12 months starting from August every year.

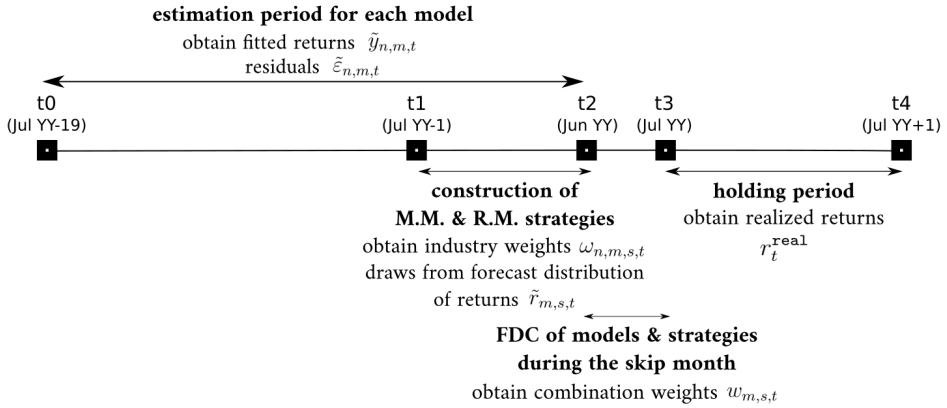


Figure 5.3.1: Time-line of model estimation, strategy construction, FDC, portfolio holding period and realised return. 'YY' indicates the year of a portfolio decision.

this as the *full information return* under the constraint that portfolio weights sum up to 0. That is, it is based on a strategy that goes long in the asset with the highest return, and goes short in the asset with the lowest return. Therefore, we compute the full-information return as

$$r_t = \max_n\{y_{n,t}\} - \min_n\{y_{n,t}\}. \tag{5.3.2}$$

In the third period, `[t2-t3]` at the bottom of Figure 5.3.1, our Bayesian FDC approach approximates the distribution of (5.3.2) with the distribution of (5.3.1) (in the sense of minimizing the Kullback-Leibler divergence) in order to construct densities which are the basis for the combination approach and obtain the combination weights $w_{m,s,t}$ for each model-strategy pair. We explain details of this step in the next subsections.

In the fourth period, `[t3-t4]` in Figure 5.3.1, we evaluate the actual returns, denoted by $r^{\texttt{real}}_{m,s,t+12}$ using the sets of models and strategies. In addition, we evaluate and obtain the combined realised return, $r^{\texttt{real}}_{t+12}$, over a holding period of 12 months as follows:

$$r^{\texttt{real}}_{m,s,t+12} = \sum_{t'=t+1}^{t+12} r^{\texttt{real}}_{m,s,t'} = \sum_{t'=t+1}^{t+12}\sum_{n=1}^{N} y_{n,t'}\omega_{n,m,s,t}, \tag{5.3.3}$$

$$r^{\texttt{real}}_{t+12} = \sum_{t'=t+1}^{t+12}\sum_{m=1}^{M}\sum_{s=1}^{S} r^{\texttt{real}}_{m,s,t'} w_{m,s,t}, \tag{5.3.4}$$

where $y_{n,t'}$ are the realised returns for each industry, $\omega_{n,m,s,t}$ is the weight of industry n given model m and strategy s, $w_{m,s,t}$ is the weight of the combination of model m and strategy s; both types of weights are determined at time t. Realised returns in (5.3.3) and (5.3.4) are then used to assess the risk-return features of all the models, strategies

and the combination of these.

5.3.2 Density combinations of model forecasts and strategy returns

Below we explain our procedure for the third period ([t2-t3] in Figure 5.3.1), i.e. how the FDC approach approximates the distribution of the the full-information return (5.3.2) with the distributions of one-period-ahead forecast (5.3.1) using the returns generated from our sets of models and strategies. We express the forecast combination model of the full-information return (5.3.2) as

$$\begin{aligned} p(r_t|I) &= \int\int p(r_t, \mathbf{w}_t, \tilde{\mathbf{r}}_t|I) d\mathbf{w}_t d\tilde{\mathbf{r}}_t \\ &= \int\int p(r_t|\mathbf{w}_t, \tilde{\mathbf{r}}_t) p(\mathbf{w}_t) p(\tilde{\mathbf{r}}_t|I) d\mathbf{w}_t d\tilde{\mathbf{r}}_t, \end{aligned} \tag{5.3.5}$$

where $\mathbf{w}_t$ and $\tilde{\mathbf{r}}_t$ are the $M \times S$ matrices consisting of weights $w_{m,s,t}$ and forecasts $\tilde{r}_{m,s,t}$, respectively. The conditional density $p(r_t|\mathbf{w}_t, \tilde{\mathbf{r}}_t)$ depends on the weights and the forecasts, which are distributed according to $p(\mathbf{w}_t)$ and $p(\tilde{\mathbf{r}}_t|I)$, respectively, with the latter density being the joint forecast density of all M models and S strategies. Note that integrals are thus of dimension $M \times S$.

Regarding the combination density **FDC.3**, for convenience, we specify it as a normal density. We note, however, that different specifications of the combination density are possible, but we leave this topic for further research. The chosen normal specification implies that the model connecting the $M \times S$ forecasts from the different sources, $\tilde{r}_{m,s,t}$ with the full-information return r_t is given by

$$r_t = \sum_{m=1}^{M} \sum_{s=1}^{S} \tilde{r}_{m,s,t} w_{m,s,t} + \varepsilon_t, \qquad \varepsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2), \;\; t = 1, \ldots, T. \tag{5.3.6}$$

We note two fundamental features of the model in (5.3.6). First, the matrix of weights $w_{m,s,t}$ for M models and S strategies consists of *(latent) random variables* so that we can model and estimate their uncertainty. Note that one can also evaluate the correlations between the weights of the different model-strategy pairs. Second, we include an error term ε_t which is an indication that *model incompleteness* can be modelled and evaluated. Hence, next to Bayesian learning, (5.3.6) also allows for Bayesian diagnostic analysis of misspecification. Note that for $\varepsilon_t \to 0$ the density $p(r_t|\mathbf{w}_t, \tilde{\mathbf{r}}_t)$ approaches a delta Dirac distribution. These two features make the proposed approach more general

	$\int\int p(r_t\|\mathbf{w}_t,\tilde{\mathbf{r}}_t)p(\mathbf{w}_t)p(\tilde{\mathbf{r}}_t\|I)d\mathbf{w}_t d\tilde{\mathbf{r}}_t$	
FDC.3	Combination density $r_t \sim \mathcal{N}\left(\sum_{m=1}^{M}\sum_{s=1}^{S}\tilde{r}_{m,s,t}w_{m,s,t}, \sigma_\varepsilon^2\right)$	Measurement equation $r_t = \sum_{m=1}^{M}\sum_{s=1}^{S}\tilde{r}_{m,s,t}w_{m,s,t} + \varepsilon_t,$ $\varepsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma_\varepsilon^2)$
FDC.2	Weight density	Link function
	$w_{m,s,t} = \frac{\exp(x_{m,s,t})}{\sum_{m=1}^{M}\sum_{s=1}^{S}\exp(x_{m,s,t})},$ $m = 1,\ldots,M,\ \ s = 1,\ldots,S.$	
	Markov process $\mathbf{x}_t \sim \mathcal{N}\left(\mathbf{x}_{t-1} + h(z_t), \sigma_\eta^2 I_{M\times S}\right),$	Transition equation $\mathbf{x}_t = \mathbf{x}_{t-1} + h(z_t) + \boldsymbol{\eta}_t,$ $\boldsymbol{\eta}_t \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma_\eta^2 I_{M\times S}\right),$
	where $\mathbf{x}_t$ is the $M \times S$ vector of latent states $x_{m,s,t}$, $I_{M\times S}$ is the identity matrix and z_t may be included to capture (observed) economic variables believed to help explain $\mathbf{x}_t$.	
FDC.1	Forecast densities	Time-varying coefficients
	$\tilde{r}_{m,s,t+1} = \sum_{n=1}^{N}\tilde{y}_{n,m,t+1}\omega_{n,m,s,t}.$	

Table 5.3.1: FDC as a nonlinear state space model.

than Bayesian model averaging where the weights are posterior probabilities that are fixed and the true model is assumed to be in the model set.

Following Billio et al. (2013), we implicitly define the weight density **FDC.2** using a link function of latent states $\mathbf{x}_t$, which we choose to be the multivariate logistic transform. The process for $\mathbf{x_t}$ is provided in Table 5.3.1, which also summarises the remaining densities in the FDC scheme.

The final step of the procedure for the third period is the estimation of the combination model (5.3.5). This is non-trivial because typically the associated likelihood is analytically intractable. A possible solution to this problem, proposed by Billio et al. (2013), is to represent the combination model as a nonlinear, non-Gaussian SSM, as shown in Table 5.3.1. Hence, the FDC approach is related to filtering methods from the literature on nonlinear state space modelling and inference. We illustrate this connection in the next subsection.

5.3.3 MFilter

Given the nonlinear, non-Gaussian SSM structure of the FDC scheme, our goal is to perform on-line inference over the optimal underlying weights[5]. We stress that even though our task is the same as in the sequential Monte Carlo (SMC) literature, our approach is noticeably different. Moreover, we argue that our MFilter shares some similarities with the literature on importance sampling for SSM based on smoothing, e.g. EIS of Richard and Zhang (2007) and Liesenfeld and Richard (2003), or NAIS of Koopman et al. (2015). We show through a set of simulation studies that the proposed MFilter outperfomrs (in terms of the approximation quality and computing time) other nonlinear, non-Gaussian filters such as the Bootstrap Particle Filter (BPF) of Gordon et al. (1993) and the Auxiliary Particle Filter (APF) of Pitt and Shephard (1999).

The combination scheme in Table 5.3.1 admits the general SSM representation:

$$r_t \sim p(r_t|\boldsymbol{\alpha}_t), \tag{5.3.7}$$

$$\boldsymbol{\alpha}_t \sim p(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}), \tag{5.3.8}$$

in which (5.3.7) and (5.3.8) describe the measurement process of the "optimal return" r_t from equation (5.3.2) (treated as "the dependent observation"), and the transition process of the extended state, respectively. We assume the initial state distribution $\boldsymbol{\alpha}_0 \sim p(\boldsymbol{\alpha}_0)$. The extended state consists of the latent combination weights $\mathbf{w}_t$ and the parameters of the system $\boldsymbol{\theta}$, i.e. $\boldsymbol{\alpha}_t = (\mathbf{w}_t^T, \boldsymbol{\theta}^T)^T$. The vector of model parameters $\boldsymbol{\theta}$ contains the measure of model-strategy set incompleteness σ_ε^2 and potentially also appropriately specified learning parameters.

We are interested in $p(\boldsymbol{\alpha}_t|r_{1:t})$, the marginal distribution of the posterior distribution of the state, called *filtering distribution* and given by

$$p(\boldsymbol{\alpha}_t|r_{1:t}) = \int p(\boldsymbol{\alpha}_{0:t}|r_{1:t}) d\boldsymbol{\alpha}_{0:t-1}. \tag{5.3.9}$$

Our novel filtering approach is summarised as follows. First, the MFilter modifies the particle filtering methods by not requiring a resampling step. Second, it extends smoothing-based importance sampling methods by using an on-line sequential procedure for inference. Third, we use mixtures of Student's t distribution as the importance density, to allow for more flexibility and robustness in the approximation compared to the more restrictive exponential class.

[5]Optimal in the sense of minimizing the Kullback-Leibler divergence between (5.3.2) and (5.3.1).

We start with providing links between the MFilter and particle filters (PFs). PFs are SMC methods based on a *recursive* formula for (5.3.9), which expresses $p(\boldsymbol{\alpha}_t|r_{1:t})$ as a function (potentially time-varying) of $p(\boldsymbol{\alpha}_{t-1}|r_{1:t-1})$ and r_t. Then the computations are carried out in two steps: prediction and updating. The former step relates to the way we sample draws at time t and the latter provides an IS correction for not using the true target density for sampling. Importantly, propagation of the particles brings about the necessity of *resampling*, as the sequential importance sampling is bound to lead to weight degeneracy. The consequence of the weight degeneracy problem is that finally only one particle carries the full weight. Not only might the resampling step be time consuming but it also introduces additional MC variation[6].

We avoid the propagation step by replacing it by an independent sampling step in each time period t. To this end we relate to the literature on the smoothing-based importance sampling for SSM, e.g. efficient importance sampling of Richard and Zhang (2007) and Liesenfeld and Richard (2003), or numerically accelerated importance sampling of Koopman et al. (2015). These methods are based on obtaining a good approximation to the smoothing density at each time period t and drawing from each $p(\boldsymbol{\alpha}_t|r_{1:t})$ independently. However, they are designed for an off-line analysis, i.e. they are based on a sample of a fixed size, while our primary goal is on-line tracking based on filtering, i.e. inference over a state space of increasing dimension. We make use of independent sampling in a *sequential* way using a very *flexible* approximation density based on mixtures of Student's t densities.

In order to specify our filtering method, we expres (5.3.9) as

$$p(\boldsymbol{\alpha}_t|r_{1:t}) \propto p(r_t|\boldsymbol{\alpha}_t)p(\boldsymbol{\alpha}_t|r_{1:t-1}),$$

which presents the key Bayesian idea, where the posterior distribution of the current state $\boldsymbol{\alpha}_t$ given all the available data $r_{1:t}$ is proportional to the prior $p(\boldsymbol{\alpha}_t)$ updated by the likelihood $p(r_t|\boldsymbol{\alpha}_t)$, where we condition upon $r_{1:t-1}$. The likelihood involves only the most recent observation r_t due to the sequential structure of the SSM. Even though we do not want to perform propagation of importance densities in the usual way of filtering procedures, we still need to keep track of the sequential structure of the SSM. We achieve this by putting a *hierarchical prior* on $\boldsymbol{\alpha}_t$, based on the empirical distribution of $\boldsymbol{\alpha}_{t-1}$ as follows

$$p(\boldsymbol{\alpha}_t|r_{1:t}, \boldsymbol{\alpha}_{t-1}) \propto p(r_t|\boldsymbol{\alpha}_t)p(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1})p(\boldsymbol{\alpha}_{t-1}). \tag{5.3.10}$$

[6]It also leads to *path degeneracy*, which is particularly problematic in the context of smoothing and in the MCMC sampling based on Particle MCMC, see Andrieu et al. (2010) and Lindsten et al. (2014).

Suppose that we have a sample $\{\boldsymbol{\alpha}_{t-1}^{(i)}\}_{i=1}^{M}$ from the previous time period $t-1$ so that we can approximate $p(\boldsymbol{\alpha}_{t-1})$ as $p(\boldsymbol{\alpha}_{t-1}) \approx \frac{1}{M}\sum_{i=1}^{M} \delta_{\boldsymbol{\alpha}_{t-1}^{(i)}}(\boldsymbol{\alpha}_{t-1})$, where $\delta_a(\cdot)$ is the Dirac measure at a. Then, (5.3.10) becomes:

$$p(\boldsymbol{\alpha}_t|r_{1:t}, \boldsymbol{\alpha}_{t-1}) \propto\!\!\!\!\sim \frac{1}{M} p(r_t|\boldsymbol{\alpha}_t) \sum_{i=1}^{M} p(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}^{(i)}) \delta_{\boldsymbol{\alpha}_{t-1}^{(i)}}(\boldsymbol{\alpha}_{t-1}), \tag{5.3.11}$$

where $\propto\!\!\!\!\sim$ means "approximately proportional to". Typically we cannot draw from (5.3.11) directly and we need to resort to sampling techniques such as importance sampling.

The choice of the proposal density is crucial for the performance of any importance sampling scheme and it has received considerable attention in the SMC literature, see Doucet et al. (2001), Liu (2001), Kunsch (2005) and Creal (2012). In the MFilter algorithm we base our approximation of (5.3.11) on the *Mixture of t by Importance Sampling weighted Expectation-Maximization* (MitISEM) algorithm proposed by Hoogerheide et al. (2012) and developed in Baştürk et al. (2016). It has been shown to be able to effectively approximate complex, non-elliptical distributions thanks to two main features of this algorithm: the class of importance distributions (mixtures of multivariate Student's t distributions), and their joint optimization (with the expectation-maximization, EM, algorithm). The former allows to closely track distributions of nonstandard shape (multimodal, skewed). The latter is iteratively carried out with the objective of minimizing the Kullback-Leibler divergence between the unknown true target distribution and the candidate density.

Robustness and flexibility in constructing approximations are particularly important from the filtering perspective in econometrics. For instance, stochastic volatility of many time series demonstrates itself via volatility clustering and it might be hard to efficiently capture periods of low and high volatility using standard approaches based on a single density approximation. Furthermore, especially in macro-econometrics one often observes breaks in time series which usually are very challenging to filter. We refer to the latter issue in the later part of this section.

Employing the basic MitISEM algorithm to approximate (5.3.11) means targeting the marginal posterior density $p(\boldsymbol{\alpha}_t|r_{0:1}, \boldsymbol{\alpha}_{t-1})$ with a categorical prior $\mathcal{C}(\{\boldsymbol{\alpha}_{t-1}^{(i)}\}_{i=1}^{M})$ (with equal weights). Hence, drawing from such a posterior density requires sampling the prior hyperparameters from the categorical distribution being the equally weighted sample of $\{\boldsymbol{\alpha}_{t-1}^{(i)}\}_{i=1}^{M}$. In practice, this means adopting *hierarchical Bayesian modelling*, in which at the first stage we draw $\boldsymbol{\alpha}_{t-1} \sim \mathcal{C}(\{\boldsymbol{\alpha}_{t-1}^{(i)}\}_{i=1}^{M})$, and at the second stage we

draw $\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1} \sim g_t^{(H)}(\boldsymbol{\alpha}_t)$, where $g_t^{(H)}(\boldsymbol{\alpha}_t)$ is the final approximation being a mixture of H Student's t densities. The resulting sample $\{\boldsymbol{\alpha}_t^{(j)}\}_{j=1}^N$ becomes the empirical prior for the next time period's analysis.

Importantly, the MitISEM algorithm requires only candidate draws and IS weights, so it can simultaneously deal with several target densities. Suppose that at time t a separate target density is specified based on each draw $\boldsymbol{\alpha}_{t-1}^{(j)}$, $j = 1, \ldots, M$, obtained in the previous time period, i.e.

$$p(r_t|\boldsymbol{\alpha}_t, \tilde{r}_t)p(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}^{(j)}).$$

Then we use MitISEM to construct a single approximation for these multiple targets for each time period t by minimizing the *average of the Kullback-Leibler divergences* between the target densities and the candidate density. In this setting the target for $\boldsymbol{\alpha}_t$ depends on $\boldsymbol{\alpha}_{t-1}^{(j)}$ but the candidate does not. This specific application of MitISEM for the purpose of quick filtering constitutes the core of the MFilter algorithm. In our situation the target density of $\boldsymbol{\alpha}_t$ given $\boldsymbol{\alpha}_{t-1}$ does not crucially depend on the particular value of $\boldsymbol{\alpha}_{t-1}$, so that conditioning on the mean, variance and other characteristics of the distribution of $\boldsymbol{\alpha}_{t-1}$ suffices here. We provide the details of the algorithm in Appendix 5.B. Note that computational efficiency gains are feasible by making use of GPU and parallel computing.

Validation and importance for typical features of economic time series Monte Carlo (MC) experiments reported in Appendix 5.C demonstrate a good statistical performance of the MFilter. To illustrate its economic relevance, we compare below the performance of the MFilter and two other filters, BPF and APF, on an experiment with structural breaks in the time series. We examine two cases of structural breaks in AR(1) models and we use the finite mixture scheme in Table 5.3.1 with the logistic weight specification, so that the measurement equation is nonlinear in the state process.

We simulate the following five return series with different persistence, which play the role of the draws $\tilde{r}_t$ from the forecast densities

$$\tilde{r}_{1,t} = \frac{k}{10} + \frac{k}{10}\tilde{r}_{1,t-1} + \eta_t, \quad \eta_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1), \quad k = 1, \ldots, 5.$$

Next, we create the measurement series r_t as a series switching between the generated series $\tilde{r}_{i,t}$, $i = 1, \ldots, 5$. We then compare the MFilter with the BPF and APF for two different cases, varying in the number of breaks in the series, as described below. The

	Case 1		Case 2	
Model	MSE	Time	MSE	Time
KF	1.000	0.007	1.000	0.007
BPF	0.052	58.483	0.202	58.483
APF	0.081	68.015	0.077	68.015
MFilter	**0.039**	**40.676**	**0.067**	**41.180**

first case has a single break/switch while the second case has two breaks/switches to emulate crisis periods.

Case 1 One switch at $t = 101$ from $\tilde{r}_1$ to $\tilde{r}_5$:

$$r_t = \begin{cases} \tilde{r}_{1,t} + \varepsilon_t & \text{for } t = 1, 2, \ldots, 100, \\ \tilde{r}_{5,t} + \varepsilon_t & \text{for } t = 101, 102, \ldots, 200, \end{cases}$$

where $\varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ with $\sigma_\varepsilon = 0.05$.

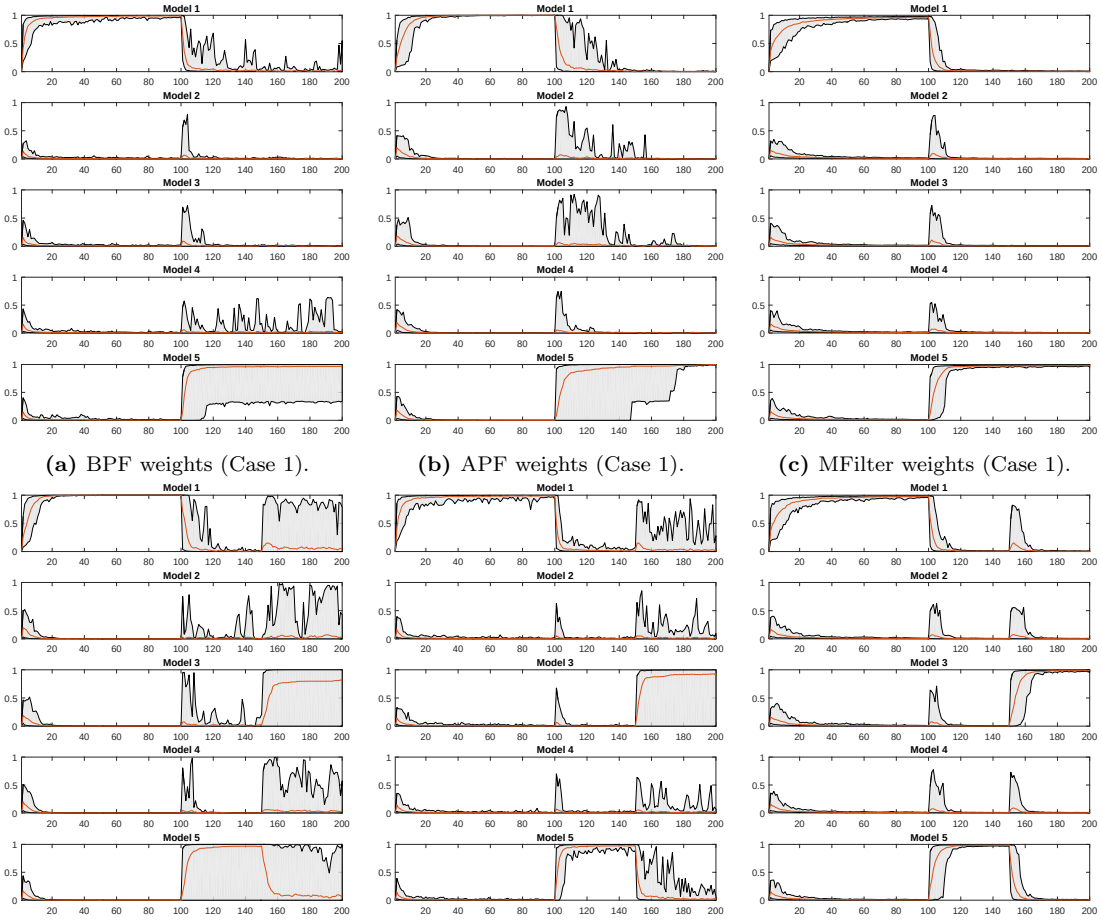
Case 2 Two switches at $t = 101$ ($\tilde{r}_1 \to \tilde{r}_5$) and $t = 151$ ($\tilde{r}_5 \to \tilde{r}_3$):

$$r_t = \begin{cases} \tilde{r}_{1,t} + \varepsilon_t & \text{for } t = 1, 2, \ldots, 100, \\ \tilde{r}_{5,t} + \varepsilon_t & \text{for } t = 101, 102, \ldots, 150, \\ \tilde{r}_{3,t} + \varepsilon_t & \text{for } t = 151, 152, \ldots, 200, \end{cases}$$

where $\varepsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$ with $\sigma_\varepsilon = 0.05$.

We compare the performance of the BPF, APF and MFilter in a small MC experiment of $R = 100$ replications. Table 5.3.2a presents a comparison of different filters for structural breaks in AR(1) models based on the Mean Squared Error (MSE), where the error is the difference between the estimated state and the true state $r_t - \varepsilon_t$, for two different experiments. In both Case 1 and Case 2 the MSE is lowest for the MFilter. This can be contributed to the fact that it is more precise in adapting after the shift(s), even though it requires a bit more time in adapting at the beginning of the sample. The MFilter importance density adapts quickly at each time period after the break(s).

We next compare the weights obtained by APF and MFilter visually. Figures 5.3.2a–5.3.2c show the model weights for Case 1. The switch in the data generating process from Model 1 to Model 5 makes it difficult for the BPF and APF to adjust quickly and one can see that the MFilter is faster in picking up the "break" due to the updated



(a) BPF weights (Case 1). **(b)** APF weights (Case 1). **(c)** MFilter weights (Case 1).

(d) BPF weights (Case 2). **(e)** APF weights (Case 2). **(f)** MFilter weights (Case 2).

Figure 5.3.2: Filtered model probability weights (red lines) using the Bootstrap Particle Filter (BPF), the Auxiliary Particle Filter (APF), and our MFilter together with the 95% credibility region (gray area) for models 1 to 5 (different rows). Top (case 1): the true model has state $\tilde{r}_{1,t} = 0.1 + 0.1\tilde{r}_{1,t-1} + \eta_t$, $\eta_t \sim \mathcal{N}(0,1)$ for $t = 1, \ldots, 100$, and model $\tilde{r}_{5,t} = 0.5 + 0.5\tilde{r}_{5,t-1} + \eta_t$, $\eta_t \sim \mathcal{N}(0,1)$ for $t = 101, \ldots, 200$; bottom (case 2): the true model has state $\tilde{r}_{1,t} = 0.1 + 0.1\tilde{r}_{1,t-1} + \eta_t$, $\eta_t \sim \mathcal{N}(0,1)$ for $t = 1, \ldots, 100$, model $\tilde{r}_{5,t} = 0.5 + 0.5\tilde{r}_{5,t-1} + \eta_t$, $\eta_t \sim \mathcal{N}(0,1)$ for $t = 101, \ldots, 150$ and model $\tilde{r}_{3,t} = 0.3 + 0.3\tilde{r}_{3,t-1} + \eta_t$, $\eta_t \sim \mathcal{N}(0,1)$ for $t = 151, \ldots, 200$.

candidate at each time period. Figures 5.3.2d–5.3.2f illustrate Case 2, in which there are two switches in the data generating process, first from Model 1 to Model 5, and then further to Model 3. The MFilter is the fastest in picking up the "breaks" (particularly the second one) which again can be contributed to the updated candidate at each time period.

5.4 Empirical application

The aim of our empirical analysis is to answer three central questions. First, whether averaging over sets of models and strategies in FDC pays off in terms of improving expected returns and risk measures. Second, what insights can be gained from the dynamic patterns of the combination weights. For instance, one can expect different patterns in quiet and in more volatile periods or learning effects which lead to improving the set of models and strategies. Third, what effect can misspecification of the model-strategy set have on the results. More specifically, we are interested in whether or not we can identify "bad" models and strategies and how removing of "bad" models and strategies affects the results. Moreover, we would like to use diagnostic learning, economic information and/or posterior residual analysis to improve modelling and strategy choice. We note that the second problem relates to learning through updating available past information, natural in a Bayesian setting. The third goal deals with the robustness of our results with respect to misspecification.

In our final analysis we consider eight sets of models from the general FAVAR-SV class from Section 5.1, reported in Table 5.4.1. For the DFM and the FAVAR model we consider $K = 1, \dots 4$ factors and $L = 1, 2$ lags. For the all models, Bayesian inference is performed with 5000 burn-in and 5000 posterior draws. We combine these models with two data-driven strategies from Section 5.2, namely MM and RM, using the FDC framework. As a benchmark for comparison we use the SM strategy and, for FDCs based on sets of models and strategies in Subsection 5.4.2, additionally the equally weighted portfolio. We focus on the distributions of realised returns $r_t^{\texttt{real}}$, see (5.3.4), stemming from different combinations, which we compare using four indicators: the mean and three risk measures being volatility, Sharpe Ratio and the largest loss. We first analyse the mean returns and risk measures for the realised returns obtained using several model-strategy pairs individually, which we compare with the performance of the benchmark SM strategy. Next, we report the time-varying performances of FDCs using sets of models and strategies.

Abbreviation	*Description*
SV	Basic SV model
VAR-N	VAR with one lag and normally distributed errors
VAR-SV	VAR with one lag with stochastic volatility in errors
DFM-N	DFM with normally distributed idiosyncratic errors
DFM-SV	DFM with stochastic volatility in idiosyncratic errors
DFM-SV2	DFM with stochastic volatility in idiosyncratic and latent errors
FAVAR-SV	FAVAR with SV in idiosyncratic errors
FAVAR-SV2	FAVAR with SV in idiosyncratic and latent errors

Table 5.4.1: Analysed models sorted by increasing complexity.

5.4.1 FDCs using individual models and strategies

We first discuss "small" FDCs, in which we combine a single model with a single strategy. As a preliminary step we consider the performance of FDCs using three basic individual models: the VAR model with normal disturbances (VAR-N), the standard SV model (SV) and the DFM with $K = 4$ factors and $L = 2$ lags (DFM(4,2)). These three models belong to the FAVAR-SV(4,2) class. We then move to a broader range of 86 model-strategy pairs, also discussed individually.

Basic models The results for three basic models are presented in Table 5.4.2 and compared with the results of the baseline SM strategy. We can draw three conclusions from this exercise. First, there does not exist a clear winning model-strategy combination in terms of all four indicators. Moreover, the performance of alternative model-strategy combinations based on distinct indicators is noticeably different. Second, for each indicator there is a model-strategy combination which dominates the benchmark SM strategy, with the SV model combined with RM outperforming the SM strategy in terms of all four indicators. Clearly, it pays off to make use of a particular econometric model with a stochastic volatility component combined with an effective model-based strategy. Third, there is one model-strategy combination which clearly performs worst: the DFM-N(4,2) model in combination with the MM strategy is the only combination that yields a negative average return. This may be caused by a type of model misspecification which is particularly detrimental for the MM strategy, but a more detailed examination would be required to pin down the specific reason for this very poor performance.

	MM				RM			
Model	Mean	Vol.	SR	LL	Mean	Vol.	SR	LL
VAR-N	0.02	**5.0**	0.005	**-24.1**	**0.09**	5.8	0.015	-35.0
SV	**0.10**	**5.1**	**0.019**	-34.7	**0.11**	**5.6**	**0.019**	**-26.0**
DFM-N(4,2)	-0.05	**5.5**	-0.009	-27.4	**0.12**	**5.4**	**0.022**	-31.1
	SM							
Model	Mean	Vol.	SR	LL				
–	0.09	5.7	0.016	-26.2				

Table 5.4.2: Mean returns and risk measures for the realised return densities from individual models-strategy combinations. Bold values: an "equal or better" value compared to the benchmark of SM. SM results reported in a single row as this strategy is not based on any model.

All specifications As mentioned above, for each of three DFM models and two FAVAR models we consider 8 different specifications corresponding to $K = 1, \ldots, 4$ factors and $L = 1, 2$ lags in the factor equation. This results in 40 combinations of DFM and FAVAR models to be estimated. In addition, we estimate the SV model and two VAR models for which we restrict the dynamics to the case of one lag. Given 10 data series, the VAR(1) delivers already very flexible dynamic patterns (shown by their implied moving averages). For each of these 43 specifications, we construct portfolios based on the MM strategy and the RM strategy. Hence, we obtain 86 specifications of model-strategy combinations.

		MM				RM			
Model	(K, L)	Mean	Vol.	SR	LL	Mean	Vol.	SR	LL
VAR-N	–	0.02	**5.0**	0.005	**-24.1**	**0.09**	5.8	0.015	-35.0
SV	–	**0.10**	**5.1**	0.019	-34.7	**0.11**	**5.6**	0.019	**-26.0**
VAR-SV	–	**0.12**	**4.5**	**0.028**	-20.2	**0.13**	5.8	**0.021**	-37.4
DFM-N	(1,1)	-0.04	**4.9**	-0.009	**-20.0**	**0.13**	**5.7**	**0.023**	-34.4
DFM-N	(4,2)	-0.05	**5.5**	-0.009	-27.4	**0.12**	**5.4**	**0.022**	-31.1
DFM-SV	(1,1)	0.04	**5.0**	**0.007**	**-20.0**	**0.11**	5.8	0.019	-37.1
DFM-SV	(4,2)	**0.12**	**5.4**	**0.023**	**-21.7**	0.06	**5.4**	0.011	-31.1
DFM-SV2	(1,1)	0.07	**4.6**	0.014	**-18.2**	0.06	**5.5**	0.010	-37.4
DFM-SV2	(4,2)	0.07	**5.7**	0.013	-32.3	0.00	**5.2**	0.000	-37.4
FAVAR-SV	(1,1)	0.08	**4.6**	0.018	**-18.3**	0.06	**5.5**	0.011	-37.4
FAVAR-SV	(4,2)	0.08	**5.7**	0.015	-32.3	0.02	**5.2**	0.005	-37.4
FAVAR-SV2	(1,1)	**0.09**	**4.6**	0.019	**-18.3**	0.06	**5.5**	0.011	-37.4
FAVAR-SV2	(4,2)	0.08	**5.7**	0.014	-32.3	0.03	**5.2**	0.005	-37.4

Table 5.4.3: Mean returns and risk measures (volatility [Vol.], Sharpe Ratio [SR], and the largest loss [LL]) for the realised return densities from the selection of model-strategy pairs, with models from Section 5.1 and strategies being MM and RM. Measures from the SM strategy: mean 0.09, volatility 5.7, Sharpe ratio 0.02 and largest loss -26.2. Bold values: an "equal or better" value compared to SM. K: the number of factors, L: the number of lags.

Table 5.4.3 presents a selection of the results on the properties of the realised re-

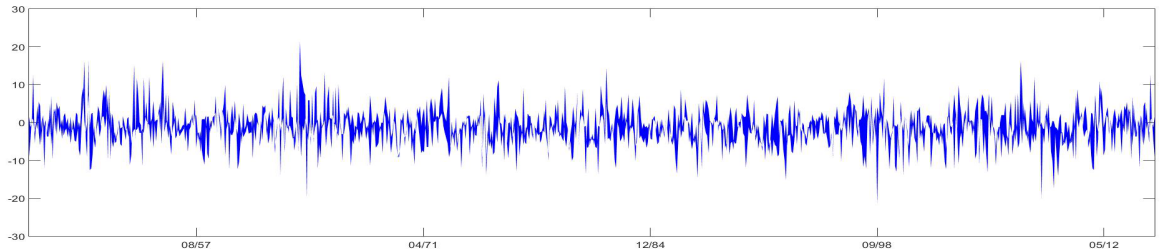
turns from these 86 combinations, which in detail are shown in Table 5.D.1 in Appendix 5.D. We can see that mean realised returns differ substantially over alternative model-strategy specifications, with the MM strategy giving poor results for simple VAR-N and DFM-N models. For both strategies (MM and RM) highly complex model, DFM-SV2 and FAVAR-SV2, do not necessarily lead to higher mean returns compared to less complex specifications of DFM-SV and FAVAR-SV. This suggests that the SV component in factor residuals (SV2) may mostly lead to over-fitting rather than to better out-of-sample performance. However, allowing for SV in the observation error for the VAR model and DFM leads to substantially better results (for both strategies) compared to i.i.d. disturbances. It is also noteworthy that the choice of the number of factors and lags in the factor models strongly influences the results for DFMs and FAVAR models. Further, mean returns from some model and strategy combinations, e.g. DFM-SV(4,2)-MM and DFM-SV(1,1)-RM (or FAVAR-SV(3,1)-MM and FAVAR-SV(2,2)-RM in Table 5.D.1), are equal or higher than those from the SM strategy. In summary, there exist noticeable differences in the performance of the two strategies: in general, more complex model structures such as DFM-SV and FAVAR-SV are well-suited for the MM strategy, the RM strategy already leads to relatively high mean returns when adopted with simpler models. Apparently, using the latter strategy implies *learning from past errors* which can compensate for the lack of model complexity.

We next compare the volatility of realised returns. Table 5.4.3 reveals that the differences in realised return volatilities between model-strategy combinations are less pronounced than the differences in mean realised returns. The volatilities obtained from each model-strategy combination are also close to the volatility from the SM strategy. An interesting observation stems from comparing both model-based strategies: given the same model class, MM generally leads to a lower volatility compared to RM. However, this difference is sensitive to the choice of the number of factors and lags in the factor models.

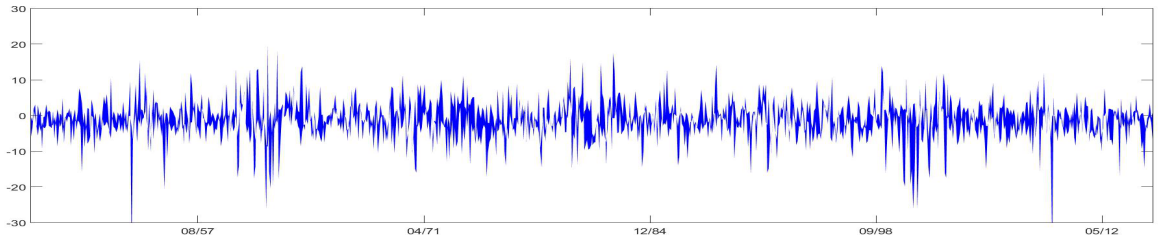
Regarding the remaining risk indicators, we note that the results for Sharpe ratios exhibit a similar pattern to those on mean returns, hence our conclusions on mean returns listed above hold also in this case. As far as largest losses are concerned, both overly simple models and overly complex ones (like DFM-SV2 and FAVAR-SV2) lead to excessive large losses compared to the models with a well-balanced parametrisation. Contrary to its superior results in terms of mean returns, the SV model leads to substantial risk of loss when combined with the MM strategy. Nevertheless, for all models except SV, the largest loss is substantially lower for the MM strategy than for the RM strategy. A complex model like FAVAR-SV combined with MM leads to a very small

extreme loss. Clearly, the strategy choice crucially matters for the risk of returns.

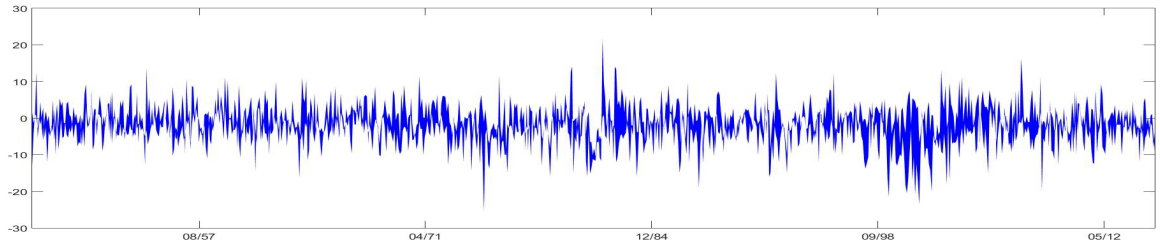
An important advantage of Bayesian inference is that it provides a complete distribution of the realised returns from a specific model-strategy combination, at any time point t. Hence, we can easily compute credible intervals (CI) for the realised returns as well as for the four measures discussed above. Due to space considerations we cannot present the intervals from all the model-strategy pairs, but for illustration we consider three selected model and portfolio strategies. Figure 5.4.1 presents the 99% CI of the realised returns from DFM(1,1) with MM, DFM-SV(4,1) with MM and DFM-SV(3,2) with RM. These intervals of returns are relatively tight for all three model and strategy combinations. In addition, even the worst-performing model and strategy combination, DFM(1,1) and MM, has very high returns in some periods. Similarly, the better performing strategies, DFM-SV(3,2) with RM and DFM-SV(4,1) with MM, lead to extreme losses in some periods. We find these observations to be valid for all the models we consider.



(a) DFM (K = 1, L = 1) with MM.



(b) DFM-SV (K = 3, L = 2) with RM.



(c) DFM-SV (K = 4, L = 1) with MM.

Figure 5.4.1: 99% CI for realised return for three selected model-strategy pairs.

The analysis of realised return from different model structures and investment strate-

gies leads to several general conclusions. Overall, the results on the mean and risk of returns are sensitive to model and strategy choices.Using MM in combination with very simple models which do not fit well to the data, like VAR-N and DFM-N, results in poor average realised returns. Complex models, like DFM-SV2 and FAVAR-SV2, tend to overfit and do not lead to better results compared to slightly simpler models like DFM-SV and FAVAR-SV. RM generates reasonable returns for simple models such as VAR-N, SV and DFM-N. However, similarly to MM, it also does not preform well for overly complex models, DFM-SV2 and FAVAR-SV2. As far as risk measures are concerned, the MM strategy performs reasonably well for almost all models. The exceptions are some members of the DFM class (for which there exists a sensitivity to the number of factors and lags) and the SV model which generates returns subject to a very high risk. Interestingly, the latter performed quite well when combined with the RM strategy. However, such reasonable outcomes from the SV-RM pair cannot conceal the underperformance of RM in terms of risk for all the remaining models. Apart from diverse performance on average, we also notice substantial time variation in the behaviour of individual model-strategy combination. This lack of a universal, time-independent "winner" among both models and strategies is one of the main reasons for combining sets of models and strategies. We explore such time-varying FDCs in the next subsection.

5.4.2 FDCs using sets of models and strategies

The analysis of the time-varying performances of FDCs using sets of models and strategies is carried out in three stages. We start with the basic model structures used in the first part of the previous subsection, VAR-N, SV and DFM-N(4,2), which we consider as a set and combine with the set of MM and RM strategies using the FDC scheme. We aim to disentangle the contribution of each component from the final outcome. Next, we investigate whether a combination of two more flexible models can outperform this combination of three basic models. Finally, we explore whether it is effective to choose only one model but with a very flexible parametric structure, in combination with the set of MM and RM strategies. To this end, we consider the FAVAR-SV(1-4,1-2) model and optimise it with respect to the number of factors and lag. Table 5.4.4 presents the main outcomes.

Combination of three basic models and two strategies The top panel of Table 5.4.4 showns that a FDC of three basic models and two strategies leads to improved

risk features compared to individual models combined with individual strategies. The risk for this combination, as measured by volatility and the largest loss, is typically lower than for the individual models. Including strategies in the combination seems to be crucial for such an improvement in terms of risk. Second, the FDC in question does not give substantially better mean returns than its component models individually. This can be attributed to an excessive weight of the "bad" DFM-N(4,2) model. Moreover, as shown in Figure 5.4.2a, to be discussed in more detail below, there does not exist any strong learning about the weight of this "bad" model, DFM-N(4,2), in the sense that this weight remains substantial over time. We conclude that, for our data and model-strategy set, the learning mechanism for the combination weights does not effectively lower combination weights of a poor performing model over time.

Combinations of two flexible models and two strategies The analysis of the combination of three basic models and two strategies leads to rather diverse results, therefore in the next step we aim to explore a smaller set of more flexible models. We thus consider a set of models belonging to the VAR-SV and DFM-SV classes. For the latter model the FDC is optimised over the number of factors $K = 1, \ldots, 4$ and the number of lags $L = 1, 2$. We refer to the optimised DFM-SV model as DFM-SV(1-4,1-2). The middle panel in Table 5.4.4 presents the results of the FDC of this set of models and the two strategies. We can conclude that the set of two flexible models and two strategies leads to better results than the set of three basic models and two strategies. Note that the results for the FDC of the set of two strategies and individual models indicate that VAR-SV has good mean return features but less so for risk features while for the case of the model DFM-SV(1-4, 1-2) the opposite holds. Thus, if an investor is interested in the joint behaviour of expected return and risk, then averaging over a set of flexible models and strategies is beneficial.

Combination of models from a single very flexible class and two strategies In the third panel in Table 5.4.4 we report the properties of the realised returns obtained using one model with a very flexible parametric structure, specified as FAVAR-SV and optimised over 4 factors and 2 lags, combined with both investment strategies. In total there are 16 forecast densities, which are summarised in Table 5.D.2 in Appendix 5.D. Our conclusion is that choosing a set of one very flexible model and two strategies implies better mean return but also more risk than the set of two flexible models and two strategies.

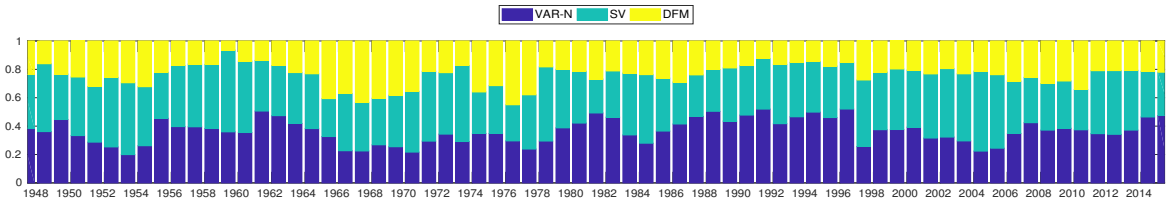
Model	Strategy	Mean	Vol.	SR	LL
Combination of three basic models and two strategies					
VAR-N & SV & DFM-N(4,2)	MM & RM	**0.10** (0.01,0.18)	**3.9** (3.6,4.2)	**0.025** (0.002,0.047)	**-23.0** (-28.8,-17.5)
Combination of two strategies per component model					
VAR-N	MM & RM	**0.09** (-0.03,0.20)	**4.7** (4.0.4,5)	**0.019** (-0.007,0.043)	-32.6 (-35.6,-20.9)
SV	MM & RM	**0.13** (-0.02,0.28)	**4.3** (3.9,4.6)	**0.032** (-0.005,0.065)	**-22.2** (-29.9,-16.1)
DFM-N(4,2)	MM & RM	0.03 (-0.12,0.17)	**4.3** (4.0,4.7)	0.006 (-0.028,0.041)	**-24.4** (-31.1,-16.8)
Combination of two flexible models and two strategies					
VAR-SV & DFM-SV(1-4,1-2)	MM & RM	**0.15** (0.08, 0.22)	**3.7** (3.5, 3.9)	**0.041** (0.021, 0.061)	**-21.6** (-26.4, -16.4)
Combination of two strategies per component model					
VAR-SV	MM & RM	**0.23** (0.11, 0.35)	**4.5** (4.2, 4.9)	**0.051** (0.024, 0.080)	-37.2 (-37.3, -36.8)
DFM-SV(1-4,1-2)	MM & RM	0.06 (0.00, 0.12)	**3.4** (3.2, 3.5)	**0.018** (0.000, 0.036)	**-14.4** (-20.1, -11.0)
Combination of models from a single very flexible class and two strategies					
FAVAR-SV(1-4, 1-2)	MM & RM	**0.18** (0.14, 0.22)	**4.5** (4.5, 4.6)	**0.039** (0.031, 0.048)	-34.8 (-35.0, -34.6)
Benchmark strategies and combination of models and strategies					
–	SM	0.09	5.7	0.016	-26.2
VAR-N, SV, DFM-N(4,2) (equal weight)	MM & RM (equal weight)	0.07 (-0.01,0.13)	**3.5** (3.3,3.8)	**0.018** (-0.002,0.038)	**-21.4** (-26.4,-16.2)
VAR-SV,DFM-SV(1-4,1-2) (equal weight)	MM & RM (equal weight)	0.07 (0.03,0.11)	**3.3** (3.2,3.4)	**0.022** (0.01,0.033)	**-13.7** (-17.8,-10.9)
FAVAR-SV(1-4, 1-2) (equal weight)	MM & RM (equal weight)	0.05 (0.02,0.07)	**3.6** (3.5,3.7)	0.013 (0.005,0.021)	**-21.6** (-24.6,-19.5)

Table 5.4.4: Mean returns and risk measures (volatility [Vol.], Sharpe Ratio [SR], and the largest loss [LL]) for the realised return densities from different sets of models and strategies. Equal weight denotes equally weighted models and strategies. Bold values: an "equal or better" value compared to the benchmark of SM. 90% credible intervals in parentheses.

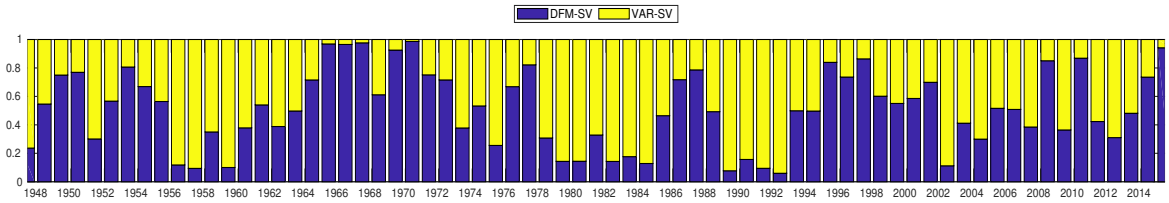
Equal weights The bottom panel of Table 5.4.4 shows the results for the equally weighted portfolio. Equal weights perform – for our data and model-strategy set – worse than time-varying weights in terms of mean returns and Sharpe ratios, see the first, second and third panel. Equal weights also perform worse than SM in terms of mean returns. For some models our FDC procedure performs slightly worse than equal weights in terms of the volatility and the largest loss, however we note that the portfolio optimisation underlying our FDC aims at maximising the return (and not e.g. minimising the volatility). Equal weights lead to smaller variance and lower loss than the benchmark SM strategy. Overall, the choice of the model set remains important, in both cases of equal weights and time-varying weights. Therefore, a sensible a priori model selection and/or an a posteriori trimming of models can be beneficial.

Credible intervals As already discussed in the previous subsection, the chosen Bayesian framework provides us with complete densities of the realised returns and of the implied measures for each model-strategy specification. Hence, we also report in Table 5.4.4 the 90% CI for each of the four criteria. It can be seen that these intervals become smaller going from the set of three basic models, through the set of two more flexible models, to the set of one very flexible model and even to the equal weights case for the FAVAR-SV(1-4,1-2) model. Thus, a very flexible model structure and a priori restrictions on the parameters (fixed weights) lead to more accurate estimation results. This information is a relevant input to the decision-making process of an investment manager.

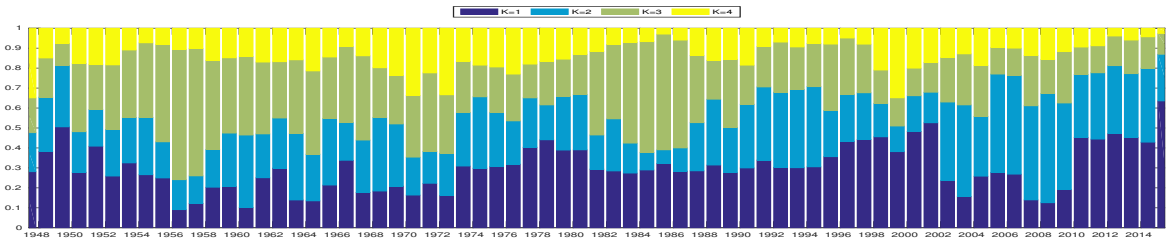
Learning about weight dynamics and their uncertainty The evolution over time of the distribution of the realised returns, including the mean returns and risk measures, is certainly important for understanding the process optimal asset allocation. In the FDC scheme the realised returns are obtained as a time-varying mixture of model-strategy pairs, therefore we are interested in the investigation of model weights and strategy weights in these combinations. Figures 5.4.2 and 5.4.3, presenting posterior means of the model weights and strategy weights, respectively, reveal a considerable time variation in the FDC weights for different sets of specifications. The weights of the basic models in Figure 5.4.2a, the two flexible models in Figure 5.4.2b and models from a single very flexible class FAVAR-SV (with models with the same number of factors but different number of lags being treated together) in Figure 5.4.2c show a clear time variation. This suggests that different data features (such as autocorrelation, cross-correlation and time-varying volatility) are better (or worse) captured at certain time



(a) Combination of VAR-N, SV, DFM(4,2) and two strategies MM and RM.

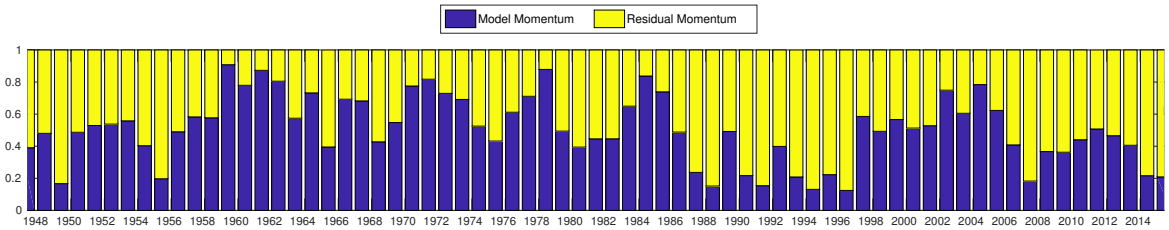


(b) Combination of DFM-SV(1-4,1-2), VAR-SV and two strategies MM and RM.

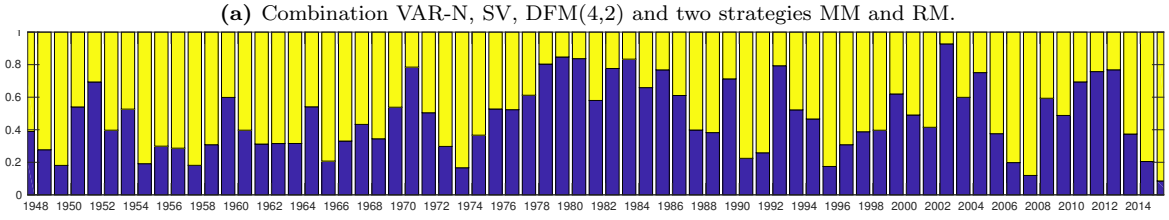


(c) Combination of 8 FAVAR-SV models and two strategies MM and RM. Models with the same number of factors but different number of lags are treated together.

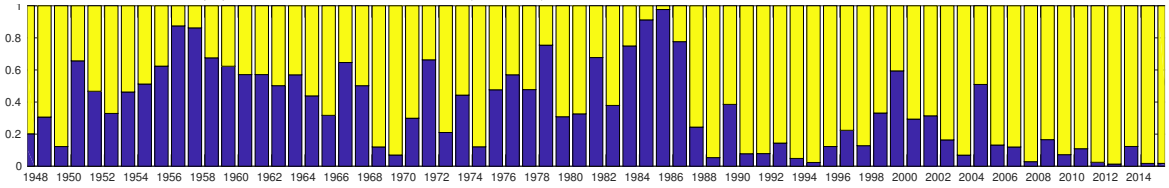
Figure 5.4.2: Model weights (posterior means) of from FDCs with different sets of models and strategies.



(a) Combination VAR-N, SV, DFM(4,2) and two strategies MM and RM.



(b) Combination DFM-SV(1-4,1-2), VAR-SV and two strategies MM and RM.



(c) Combination 8 FAVAR-SV models and two strategies MM and RM

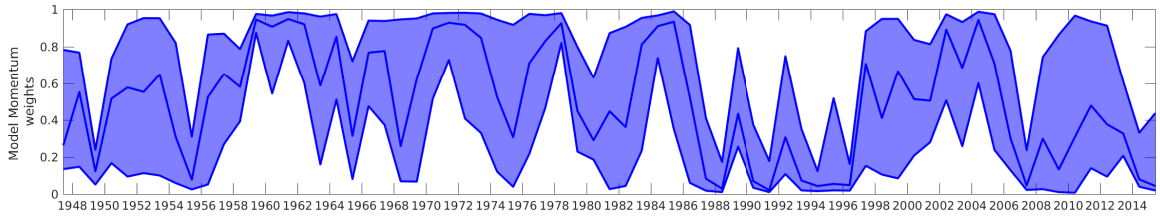
Figure 5.4.3: Strategy weights (posterior means) of from FDCs with different sets of models and strategies.

periods by particular models in these combinations, so these models become more (or less) relevant in given periods. For instance, the weights per number of factors in the FAVAR-SV combination reveal that in the most recent periods models and strategy combinations with a single factor have higher weights than in earlier periods. This finding is in line with the relatively low canonical correlations between returns at the end of the sample compared to the beginning of the sample, as shown in Figure 5.1.1b. However, using a single model with only one factor is not sufficient to provide optimal portfolio allocation in the 2010s, for which model diversification is still required. Similarly, posterior means of strategy weights are subject to a substantial time-variation, see Figure 5.4.3. Interestingly, these fluctuations are more pronounced than those for the model weights. A plausible reason for this is that there is a fundamental difference between the strategies, while the models in the FAVAR-SV case are all nested in the same class.

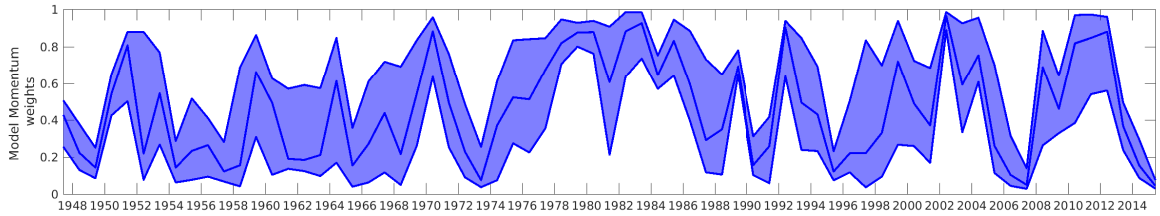
We present the uncertainty in the strategy weights in Figure 5.4.5 based on their 60% CI. The set of the three basic models (Figure 5.4.4a) generally leads to higher uncertainty, i.e. wider CI, than the two flexible models (Figure 5.4.4b). The differences between the two flexible models (Figure 5.4.4b) and one model FAVAR-SV(1-4,1-2) (Figure 5.4.4c) are less pronounced, however for the latter case the importance of the RM strategy at the beginning of the recent financial crisis is confirmed by relatively low uncertainty in strategy weights.

It is also interesting and relevant for policy recommendations to investigate the behaviour of the RM weights versus the MM weights. For the best performing model-strategy combination, i.e. VAR-SV and DFM-SV(1-4,1-2), we observe that the MM strategy remains important also during the recent crisis period. It is noteworthy that for both, the three basic models and the one very flexible model, the RM strategy is vital, particularly, around the 1990s and at the beginning of the recent financial crisis around 2008. This suggests that the RM strategy performs better in volatile periods. Moreover, the RM strategy may be more robust against misspecification, but a closer investigation of this matter is left as a topic for further research.

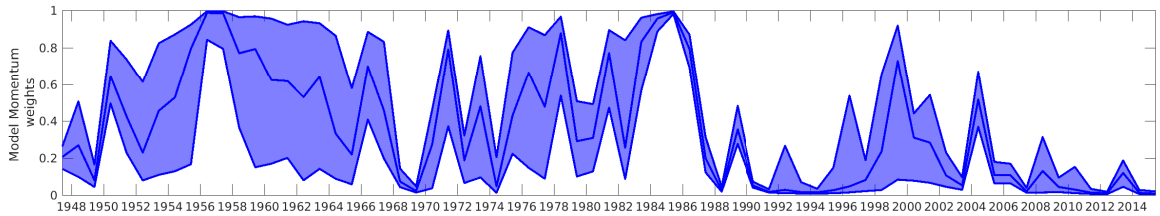
Our findings on the strategy weights relate to Jegadeesh and Titman (2001) and Blitz et al. (2011). Jegadeesh and Titman (2001) show that the momentum effect, indicated by the MM strategy, is apparent before and shortly after the 1990s. Our results confirm this observation. Blitz et al. (2011) find that the RM strategy is less affected by the market sentiments compared to MM during the financial crisis of 2008. This is in line with the increased weights for RM in Figure 5.4.3 and its tight CI in Figure 5.4.4c, which indicate good performance of the RM strategy around 2008. One explanation



(a) Combination of VAR-N, SV and DFM(4,2) and two strategies MM and RM.



(b) Combination of DFM-SV(1-4,1-2), VAR-SV and two strategies MM and RM.



(c) Combination of 8 FAVAR-SV models and two strategies MM and RM.

Figure 5.4.4: Strategy weights (posterior means) and 60% credible intervals for FDCs with different sets of models and strategies.

for this result is that the RM strategy is intended to take advantage of large residuals (in absolute sense).

A more detailed analysis of the dynamics and learning of weights is left for further research. However, it can already be seen that the dynamic patterns in model and strategy weights are very relevant pieces of information for portfolio analysis. Moreover, the construction of the currently used learning mechanism in the FDC framework does not allow it to unambiguously assign very low weights to "bad" models. For this purpose one would need a stronger feedback in the learning scheme. For instance, diagnostic information about posterior residual behaviour and poor economic performance may be useful as a complimentary source of information.

Misspecification and diagnostic learning The model and strategy sets that we consider are potentially misspecified. An important issue is therefore how to measure the degree of this misspecification. In the statistical literature there exist several diagnostic tests and methodologies to determine the correct number of relevant components, see McLachlan and Peel (2004, Ch. 6), Frühwirth-Schnatter (2006, Ch. 4) and for a recent analysis Baştürk et al. (2018). In this chapter, we follow two approaches. The motivation for the first one stems from the discussed literature in economics and finance. This approach uses economic interpretation of the results delivered by a set of models and strategies for trimming of this initial set. For instance, one can account for the effect of a "bad" model on the total returns, like in the case of the above mentioned DFM(4,2). The second approach follows from taking a forecasting approach. We extend the interpretation of σ_ε^2 from section 5.3.2 to forecast errors and take the standard deviation of the forecast residuals as a measure of incompleteness. Clearly, even when the model set is perfectly specified then this measure will be non-zero due to forecast errors. However, it serves as useful *relative* measure for comparing the performance of alternative sets of models and strategies.

Figure 5.4.4 presents the standard deviations of the forecast residuals from the three considered types of FDCs. Standard deviations from the set of three basic models and two strategies are generally higher than these from the other two considered sets. This confirms our earlier conclusion about the better fit of the flexible models compared to the basic models. Regarding the former, the comparison of both flexible mixtures (VAR-SV and DFM-SV(1-4,1-2) vs FAVAR-SV(1-4,1-2)) does not lead to a clear conclusion about a specification more robust against model incompleteness. It is interesting to observe that periods with high standard deviations of residuals, such as the period 2009–2012, can be caused by high volatility in the data as well as by

a misspecified model and strategy set. Interestingly, these periods also correspond to relatively high variations in the weights of strategies. This suggests that in the case of a misspecified model-strategy set and/or a highly volatile period, it is hard to tell which model-strategy combination should be followed in order to improve expected return and risk measures. However, in periods with low standard deviations, which may be due to low volatility in the markets as well as a more complete model-strategy set, it is easier to identify a single "winning" strategy.

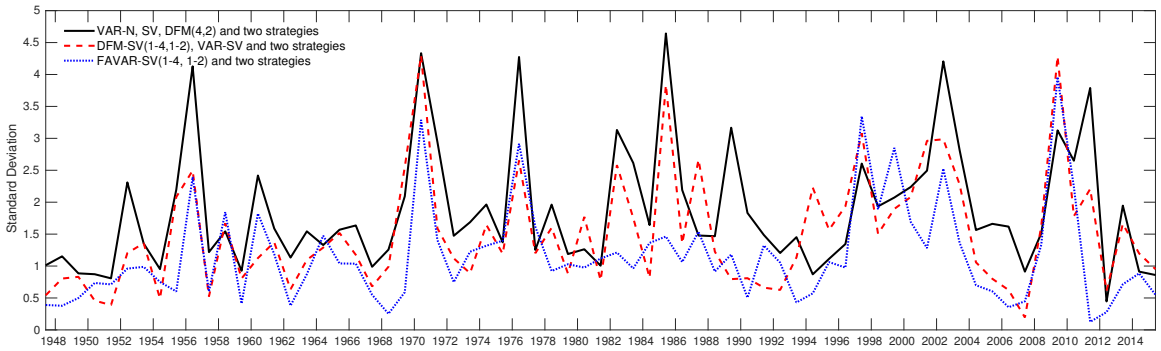


Figure 5.4.5: Model and strategy incompleteness measure (standard deviation of residuals).

5.5 Conclusions

We have introduced a dynamic asset-allocation approach specified as a forecast density combination of a set of models and momentum strategies in which portfolios are updated at every decision period based on learning about their past performance. To allow for efficient and robust Bayesian estimation of the resulting nonlinear state space model, we have introduced a novel non-linear filter based on the MitISEM alogirthm of Hoogerheide et al. (2012). We have demonstrated that the proposed M-filter leads to substantial gains in accuracy and computational speed.

Our extensive empirical study based on over 80 years of returns of ten US industries has revealed several implications for asset allocation. In volatile periods with substantial shocks it is profitable to make use of models that capture short-run properties through stochastic volatility components. On the other hand, in quiet periods relatively less complex models receive substantial weights. Regarding the two momentum strategies, we have found that in volatile periods a residual momentum strategy which "learns" from past forecast errors has higher weights compared to the simple model-based momentum strategy. Thus, a time-varying equity momentum strategy leads to better

performance over time. In particular, a set of models with different long and short-run dynamics together with a set of investment strategies improve two key risk measures, volatility and largest loss, of realized returns.

There are several opportunities to extend this line of research. One can consider a larger data set of industries and use more (economic) prior information in the model selection and in the formulation of (informative) prior distributions. This information may be particularly helpful when one intends to include mean-variance optimization methods in the analysis. Another possible extension is to assess alternative sets of combination weights of models and strategies, see also Johnstone (2012). Further, analysis of the behaviour of stocks within an industry is relevant for a more detailed portfolio analysis.

Our findings can be beneficial for practitioners, e.g. an investment company, in setting up their portfolio strategy. Note that the information set upon which we condition our FDC, i.e. US industrial returns between 1926M7 and 2015M6 and the proposed set of dynamic models and data driven portfolio strategies, is available to any professional institution. Conditionally upon this information set, adopting our FDCs with sets of models and strategies improves the properties of mean return and risk of a portfolio compared to using single models and strategies, including the standard momentum strategy. In most cases, the latter yields lower mean returns and more risk. Importantly, learning weights of the FDCs of sets of models and strategies should be carefully incorporated in exploring alternative scenarios of portfolio strategies. The proposed time-varying weights of the set of two flexible models and two strategies outperform the equally weighted combination of models and strategies in the sense of better return and risk trade-off. Compared to fixed weights, these gains are rather pronounced in volatile periods. Finally, we note that how a trader deals with the proposed recommendations obviously depends on his/her preferences regarding the return-risk trade-off. We do emphasize, however, that our proposed data-driven FDC approach for combining models and strategies does not require a fully specified utility function that is particular for a trader.

Appendix 5.A Models within the FAVAR-SV class

In this section we describe different model structures used in Section 5.1 resulting from the general formulation (5.1.1).

5.A.1 Linear and Gaussian dynamic factor model

The linear Gaussian DFM is a special case of model (5.1.1) with $\boldsymbol{\beta} = 0$ and a diagonal Σ matrix:

$$\begin{aligned} \mathbf{y}_t &= \Lambda \mathbf{f}_t + \boldsymbol{\varepsilon}_t, & \boldsymbol{\varepsilon}_t &\sim \mathcal{N}(0, \Sigma), \\ \mathbf{f}_t &= \phi_1 \mathbf{f}_{t-1} + \cdots + \phi_L \mathbf{f}_{t-L} + \boldsymbol{\eta}_t, & \boldsymbol{\eta}_t &\sim \mathcal{N}(0, \mathrm{Q}), \end{aligned} \tag{5.A.1}$$

which is a linear and Gaussian DFM. We estimate this model with the following priors.

1) For the diagonal elements of Σ we set independent Inverse Gamma ($\mathcal{IG}$) priors

$$\underline{\sigma}^2_{\varepsilon,ii} \sim \mathcal{IG}\left(\frac{v_i}{2}, \frac{s_i}{2}\right),$$

   where we set $v_i = 2$ and $s_i = 5$ for $i = 1, \ldots, N$.
2) For the loading parameters we specify normal priors, $\underline{\Lambda} \sim \mathcal{N}(\underline{\boldsymbol{\mu}}, \underline{\mathbf{C}})$, where $\underline{\boldsymbol{\mu}} = 0$ and $\underline{\mathbf{C}} = \mathrm{I}$.
3) The priors for the autoregressive parameters $\Phi = [\phi_1, \ldots, \phi_L]$ and latent errors variance Q are diffuse conjugate Normal-Wishart:

$$\underline{\boldsymbol{\Phi}}|\mathrm{Q} \sim \mathcal{N}(0, \mathrm{Q} \otimes \Omega_0), \quad \underline{\mathrm{Q}} \sim \mathcal{IW}(\mathrm{Q}_0, N + K + 2),$$

   where $\boldsymbol{\Phi} = \mathrm{vec}(\Phi)$ consists of the elements of Φ stacked in a column vector of length $L \times K^2$, where L is the number of lags of the latent factor and K is the number of factors. As in Bernanke et al. (2005) we set the prior to express the beliefs that parameters on longer lags are more likely to be zero, in the spirit of the Minnesota prior. The diagonal elements of Q_0 are set to the residual variances of the corresponding univariate autoregressions, $\hat{\sigma}^2_{\eta,kk}$ for $k = 1, \ldots, K$. The diagonal elements of Ω_0 are set on k lagged jth variable in ith equation equals $\sigma_i^2 / k\sigma_j^2$.

Defining $\Lambda_i = (\lambda_{i,1}, \ldots, \lambda_{i,k})$, for $i = 1, \ldots, N$, we can specify the following Gibbs sampling scheme.

1) The full conditional posterior for the elements of Σ reduces to a set of N independent inverse-gamma distributions with

$$\overline{\sigma}^2_{\varepsilon,ii} \sim \mathcal{IG}\left(\frac{v_i + T}{2}, \frac{v_i s_i^2 + d_i}{2}\right),$$

   where $d_i = \sum_{t=1}^{T} (y_{it} - \Lambda_i f_{it})(y_{it} - \Lambda_i f_{it})'$, $i = 1, \ldots, N$.

2) The draws of the loading parameters which satisfy the related restrictions are generated as follows.
   a) For $i = 1, \ldots, K$, draw $\overline{\Lambda}_i \sim \mathcal{N}(\overline{m}_i, \overline{\mathrm{C}}_i)\mathbf{I}(\lambda_{ii} > 0)$, where $\overline{m}_i = \overline{\mathrm{C}}_i(\underline{\mathrm{C}}_i^{-1}\underline{\mu}_i + \sigma_{\varepsilon,ii}^{-2} f_i' y_i)$ and $\overline{\mathrm{C}}_i^{-1} = \underline{\mathrm{C}}_i^{-1} + \sigma_{\varepsilon,ii}^{-2} f_i' f_i$
   b) For $i = K+1, \ldots, N$ draw $\overline{\Lambda}_i \sim \mathcal{N}(\overline{m}_i, \overline{\mathrm{C}}_i)$ where $\overline{m}_i = \overline{\mathrm{C}}_i(\underline{\mathrm{C}}_i^{-1}\underline{\mu}_i + \sigma_{\varepsilon,ii}^{-2} f_i' y_i)$ and $\overline{\mathrm{C}}_i^{-1} = \underline{\mathrm{C}}_i^{-1} + \sigma_{\varepsilon,ii}^{-2} f_i' f_i$.
3) The posterior of $\mathbf{\Phi}$ and Q follows from the standard VAR form that we adopt, which can be estimated equation by equation to yield the following simulation scheme.
   a) Draw $\overline{\mathrm{Q}}$ from $\mathcal{IW}(\hat{\mathrm{Q}}, T + K + N + 2)$, where $\hat{\mathrm{Q}} = \underline{\mathrm{Q}} + \hat{\Gamma}'\hat{\Gamma} + \hat{\Phi}'[\Omega_0 + (\hat{\mathrm{F}}_t'\hat{\mathrm{F}}_t)^{-1}]^{-1}\hat{\Phi}$ and $\hat{\Gamma}$ is the matrix of OLS residuals.
   b) Draw $\overline{\mathbf{\Phi}}$ from the conditional normal distribution of the form:

$$\overline{\mathbf{\Phi}} \sim \mathcal{N}(\mathrm{vec}(\tilde{\Phi}), \mathrm{Q} \otimes \tilde{\Omega}), \tag{5.A.2}$$

   where $\tilde{\Phi} = \tilde{\Omega}(\hat{\mathbf{f}}_{t-1}'\hat{\mathbf{f}}_{t-1})\hat{\Phi}$ and $\tilde{\Omega} = (\Omega_0^{-1} + \hat{\mathbf{f}}_{t-1}'\hat{\mathbf{f}}_{t-1})^{-1}$.
5) Draw the latent states $\mathbf{f}_t$ using the FF-BS algorithm as described in Carter and Kohn (1994).
6) Go to step 1.

5.A.2 Linear dynamic factor model with stochastic volatility)

We obtain the DFM-SV by setting $\boldsymbol{\beta} = 0$ in equation (5.1.1):

$$\begin{aligned} \mathbf{y}_t &= \Lambda \mathbf{f}_t + \boldsymbol{\varepsilon}_t, & \boldsymbol{\varepsilon}_t &\sim \mathcal{N}(0, \Sigma_t), \\ \mathbf{f}_t &= \phi_1 \mathbf{f}_{t-1} + \cdots + \phi_L \mathbf{f}_{t-L} + \boldsymbol{\eta}_t, & \boldsymbol{\eta}_t &\sim \mathcal{N}(0, \mathrm{Q}), \end{aligned} \tag{5.A.3}$$

and specifying a time-varying variance-covariance matrix:

$$\Sigma_t = \begin{pmatrix} \sigma_{11,t}^2 & 0 & \ldots & 0 \\ 0 & \sigma_{22,t}^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_{NN,t}^2 \end{pmatrix}, \; i = 1, \ldots, N. \tag{5.A.4}$$

We assume that the log volatilities $h_{it} = \log(\sigma_{ii,t}^2)$ follow a stationary and mean reverting process

$$h_{it} = \mu_i + \psi_i h_{it-1} + \zeta_t, \quad \zeta_t \sim \mathcal{N}(0, \gamma_{ii}), \quad \psi_i \sim \mathcal{U}(-1, 1), \qquad p(\mu_i) \propto 1.$$

Starting from equation (5.A.3) and rearranging, we get $\boldsymbol{\varepsilon}_t = \mathbf{y}_t - \Lambda \mathbf{f}_t = \mathbf{y}_t^*$. Taking the squares plus an offset constant we obtain

$$\begin{aligned} \mathbf{y}_t^{**} &= \log\left((\mathbf{y}_t^*)^2 + \bar{c}\right), \\ \mathbf{y}_t^{**} &= 2\mathbf{h}_t + \mathbf{e}_t, \\ \mathbf{h}_t &= \boldsymbol{\mu} + \boldsymbol{\psi}\mathbf{h}_{t-1} + \xi_t, \quad \xi_t \sim \mathcal{N}(0, \Gamma), \end{aligned} \tag{5.A.5}$$

where $\mathbf{e}_t = \log(\boldsymbol{\varepsilon}_t)$ follows the $\chi^2(1)$ distribution. Therefore, the standard Kalman filter and smoother cannot be adopted, see Carter and Kohn (1994). To solve this problem Kim et al. (1998) employ a data augmentation approach and introduce a new state variable $\mathbf{s}_{1:T} = \{s_1, \ldots, s_t\}$, so that the linear, non-Gaussian state space model (5.A.5) can be rewritten as conditionally linear Gaussian. Then, the distribution of $\mathbf{e}_t$ can be approximated as

$$\mathbf{e}_t \approx \sum_{j=1}^{7} q_j \mathcal{N}(\tau_j - 1.2704, \nu_j^2),$$

where τ_j, ν_j^2 and q_j for $j = 1, \ldots, 7$ are constants specified in Kim et al. (1998). Conditionally on the state $s_{t+1} = j$, the errors $\mathbf{e}_t$ can be sampled as

$$\begin{aligned} \mathbf{e}_t | s_{t+1} = j &\sim \mathcal{N}(\tau_j - 1.2704, \nu_j^2), \\ \Pr(s_{t+1} = j) &= q_j. \end{aligned}$$

The sequence of states s_t is drawn using

$$\Pr(s_t = j | \mathbf{y}_t^{**}, \mathbf{h}_t) \propto q_j f_{\mathcal{N}}(\mathbf{y}_t^{**} | 2\mathbf{h}_t + \tau_j - 1.2704, \nu_j^2), \tag{5.A.6}$$

where $f_{\mathcal{N}}(\cdot)$ denotes the kernel of a normal density and $j = 1, \ldots, 7$, $t = 1, \ldots, T$. Conditionally on $\mathbf{s}_{1:T}$ the model is linear Gaussian and the algorithm of Carter and Kohn (1994) can be used.

The priors remains as described before, with the only difference related to the SV parameters, $\boldsymbol{\mu}$, $\boldsymbol{\psi}$ and variance of the errors Γ. For the two former we specify

$$\begin{aligned} \begin{bmatrix} \mu_i \\ \psi_i \end{bmatrix} &\sim \mathcal{N}\left(\begin{bmatrix} \underline{\mathrm{m}}_{\mu_i} \\ \underline{\mathrm{m}}_{\psi_i} \end{bmatrix}, \begin{bmatrix} \underline{\mathrm{V}}_{\mu_i} & 0 \\ 0 & \underline{\mathrm{V}}_{\psi_i} \end{bmatrix} \right), \\ |\psi_i| &< 1, \end{aligned}$$

while for γ_{ii}^{-2} we put $\gamma_{ii}^{-2} \sim \mathcal{G}(1/\underline{\mathrm{k}}_{\gamma}, 1)$. For the hyperparameters we follow Pettenuzzo

and Ravazzolo (2016) and set $\underline{\mathrm{k}}_{\gamma} = 0.01$, $\underline{\mathrm{m}}_{\mu_i} = 0$, $\underline{\mathrm{m}}_{\psi_i} = 0.95$, $\underline{\mathrm{V}}_{\mu_i} = 10$ and $\underline{\mathrm{V}}_{\psi_i} = 1.0e^{-06}$. These values imply a strong autocorrelation structure for h_{it}, which is typical for financial time series.

For this model, the Gibbs sampling steps are as follows.

1) Initialize $\mathbf{f}_t^{(0)}$, $\mathbf{h}_t^{(0)}$, $\Lambda_t^{(0)}, \Sigma^{(0)}, \mathrm{Q}^{(0)}$.
2) Draw latent factors $\mathbf{f}_t$ from $p(\mathbf{f}_t|, \Lambda, \mathrm{Q}, \Sigma_t, \mathbf{h}_t)$ using the FF-BS algorithm described in Carter and Kohn (1994).
3) Conditionally on $\mathbf{h}_t$ and Λ, draw the indicator variable s_t for the mixture according to (5.A.6).
4) Draw a sequence of stochastic volatilities $\mathbf{h}_t$, $t = 1, \ldots, T$ from $p(\mathbf{h}_t|\Lambda, \mathbf{f}_t, s_t, \boldsymbol{\mu}, \boldsymbol{\psi})$ from the conditional linear and Gaussian system using the method of Carter and Kohn (1994).
5) Draw the stochastic volatility variances γ_{ii}^2 from $p(\gamma_{ii}^2|h_{it}, \mu_i, \psi_i)$ from the following posterior:

$$\overline{\gamma}_{ii}^{-2} \sim \mathcal{G}\left(\left[\frac{\underline{\mathrm{k}}_{\gamma} + \sum_{t=1}^{T-1}(h_{it+1} - \mu_i - \psi_i h_{it})^2}{t}\right]^{-1}, T\right).$$

6) Draw the SV parameters jointly

$$\begin{bmatrix}\overline{\mu}_i \\ \overline{\psi}_i\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\bar{m}_{\mu_i} \\ \bar{m}_{\psi_i}\end{bmatrix}, \bar{V}_{(\mu_i,\psi_i)}\right) \times |\psi_i| < 1,$$

where

$$\bar{V}_{(\mu_i,\psi_i)} = \begin{bmatrix}\underline{\mathrm{V}}_{\mu_i}^{-1} & 0 \\ 0 & \underline{\mathrm{V}}_{\psi_i}^{-1}\end{bmatrix} + \bar{\gamma}_{ii}^{-2}\sum_{t=1}^{T-1}\begin{bmatrix}1 & h_{it} \\ h_{it} & h_{it}^2\end{bmatrix}$$

and

$$\begin{bmatrix}\bar{m}_{\mu_i} \\ \bar{m}_{\psi_i}\end{bmatrix} = \bar{V}_{(\mu_i,\psi_i)}\left(\begin{bmatrix}\underline{\mathrm{V}}_{\mu_i}^{-1} & 0 \\ 0 & \underline{\mathrm{V}}_{\psi_i}^{-1}\end{bmatrix}\begin{bmatrix}\underline{\mathrm{m}}_{\mu_i} \\ \underline{\mathrm{m}}_{\psi_i}\end{bmatrix} + \bar{\gamma}_{ii}^{-2}\sum_{t=1}^{T-1}\begin{bmatrix}1 \\ h_{it}\end{bmatrix}h_{it+1}\right).$$

7) Go to step 2.

5.A.3 Linear dynamic factor model with two stochastic volatility components

We obtain the DFM model with two stochastic volatilities by assuming $\boldsymbol{\beta} = 0$ in equation (5.1.1) and by defining the following time-varying covariance matrices for the idiosyncratic and latent errors:

$$\begin{aligned} \mathbf{y}_t &= \Lambda \mathbf{f}_t + \boldsymbol{\varepsilon}_t, & \boldsymbol{\varepsilon}_t &\sim \mathcal{N}(0, \Sigma_t), \\ \mathbf{f}_t &= \phi_1 \mathbf{f}_{t-1} + \cdots + \phi_L \mathbf{f}_{t-L} + \boldsymbol{\eta}_t, & \boldsymbol{\eta}_t &\sim \mathcal{N}(0, \mathrm{Q}_t), \end{aligned} \tag{5.A.7}$$

with the idiosyncratic errors defined as in equation (5.A.4) and latent error variances is given by

$$\mathrm{Q}_t = \begin{pmatrix} \eta^2_{11,t} & 0 & \dots & 0 \\ 0 & \eta^2_{22,t} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \eta^2_{KK,t} \end{pmatrix}, i = 1, \dots, K, \tag{5.A.8}$$

where log volatilities $k_{it} = \log(\eta^2_{ii,t})$ follow a stationary and mean reverting process:

$$k_{it} = \omega_i + \beta_i k_{it-1} + \xi_{it}, \qquad \xi_{it} \sim \mathcal{N}(0, \sigma^2_{\xi_i}).$$

The estimation of this model proceeds as before with an added step in the Gibbs sampler to extract the latent time-varying variance.

5.A.4 Factor augmented VAR models with stochastic volatility components

Assuming in equation (5.1.1) $\boldsymbol{\beta} \neq 0$ and a time-varying variance-covariance matrix for the idiosyncratic and latent errors we obtain the FAVAR-SV2 model given by

$$\begin{aligned} \mathbf{y}_t &= \beta \mathbf{x}_t + \Lambda \mathbf{f}_t + \boldsymbol{\varepsilon}_t, & \boldsymbol{\varepsilon}_t &\sim \mathcal{N}(0, \Sigma_t), \\ \mathbf{f}_t &= \phi_1 \mathbf{f}_{t-1} + \cdots + \phi_L \mathbf{f}_{t-L} + \boldsymbol{\eta}_t, & \boldsymbol{\eta}_t &\sim \mathcal{N}(0, \mathrm{Q}_t). \end{aligned} \tag{5.A.9}$$

The FAVAR model extends the state equation by defining $\mathbf{x}_t$ as a vector of the lagged dependent variables. This leads to a VAR form in the state equation of (5.A.9)

$$\begin{bmatrix} \mathbf{f}_t \\ \mathbf{x}_t \end{bmatrix} = \tilde{\Phi}_1 \begin{bmatrix} \mathbf{f}_{t-1} \\ \mathbf{x}_{t-1} \end{bmatrix} + \cdots + \tilde{\Phi}_L \begin{bmatrix} \mathbf{f}_{t-L} \\ \mathbf{x}_{t-L} \end{bmatrix} + \tilde{\boldsymbol{\varepsilon}}_t,$$

see also Stock and Watson (2005). Conditionally on the latent states, the estimation of the VAR parameters $\boldsymbol{\beta}$ is similar to that of the univariate linear regression model, hence Bayesian inference is standard. The two proposed FAVAR models are defined by a stochastic volatility component in the idiosyncratic disturbances (FAVAR-SV) and a stochastic volatility components in the idiosyncratic and latent disturbances (FAVAR-SV2). Note that the FAVAR-SV (and FAVAR-SV2) model simplifies to a DFM model in Section 5.A.1 when $\beta = 0$, a VAR model if factor coefficient $\Lambda = 0$, and a stochastic volatility model when both $\beta = 0$ and $\Lambda = 0$. Hence DFM, VAR and SV models listed together constitute parts of the FAVAR-SV (and FAVAR-SV2) models. We refer to the earlier sections of this appendix for the inference on the SV components conditionally on the remaining parameters.

Appendix 5.B MFilter algorithm

Below we present the details of the recursion for the proposed MFilter from Section 5.3. In the description we treat the estimated parameter vector θ as known and we omit it for the sake of notation. For a detailed discussion of the general MitISEM procedure we refer to Hoogerheide et al. (2012).

1) **Initialization.** Draw $\boldsymbol{\alpha}_0^{(j)} \sim p(\boldsymbol{\alpha}_0)$ for $j = 1, \ldots, M$.
2) **Recursion.** For $t = 1, \ldots, T$ construct the candidate density $g_t(\boldsymbol{\alpha}_t)$ using the MitISEM algorithm as follows.
   a) **Initialization.** Simulate draws $\boldsymbol{\alpha}_t^{(j)}$, $j = 1, \ldots, M$, from a 'naive' candidate distribution with density $g_t^{(0)}(\cdot)$ (e.g. a Student's t distribution with $v = 5$ degrees of freedom).
   Compute the corresponding IS weights:
   $$\tilde{w}_t^{(j)} = \frac{p(r_t|\boldsymbol{\alpha}_t^{(j)})p(\boldsymbol{\alpha}_t^{(j)}|\boldsymbol{\alpha}_{t-1}^{(j)})}{g_t^{(0)}(\boldsymbol{\alpha}_t^{(j)})},$$
   where the target density kernel has the form $p(r_t|\boldsymbol{\alpha}_t)p(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}^{(j)})$, and normalize them to $w_t^{(j)}$.
   b) **Adaptation.** Use the draws $\boldsymbol{\alpha}_t^{(j)}$ and the weights $\tilde{w}_t^{(j)}$ from the naive distribution $g_t^{(0)}(\cdot)$ to IS estimate the mean and covariance matrix of the target distribution. Use these estimates as the mode and the scale matrix of the Student's t adapted density $g_t^{(a)}(\cdot)$. Draw a sample $\boldsymbol{\alpha}_t^{(j)}$ from $g_t^{(t)}(\cdot)$ and compute the IS weights for this sample.
   c) Apply the the **IS weighted EM (ISEM) algorithm** given the sample from $g_t^{(a)}(\cdot)$ and the corresponding IS weights. The output consists of the new candidate density with $h = 1$ component $g_t^{(H)}(\cdot)$ with the optimized parameters. Draw a new sample $\boldsymbol{\alpha}_t^{(j)}$ from this candidate, compute the corresponding IS weights. Calculate the coefficient of variation $\mathrm{CV}^{(H)}$ of the normalized weights $w_t^{(j)}$, $j = 1, \ldots, M$.
   d) **Iterate on the number of mixture components.** Given the current mixture of h components $g_t^{(H)}(\cdot)$ add the next component to the mixture in the following way.
      d.1) Use a chosen fraction (e.g. 0.1) of the draws $\boldsymbol{\alpha}_t^{(j)}$ from the current mixture corresponding to the highest IS weights to IS estimate the mean and variance. Use these parameters as the starting mode and scale parameters for the new mixture component, μ_{h+1} and Σ_{h+1}.

d.2) Update the mixture probabilities: assign the starting value for the new component probability η_{h+1} (e.q. 0.1) and multiply the old mixture probabilities $\eta_1, \ldots, \eta_h$ by $(1 - \eta_{h+1})$. Set the number of degrees of freedom for the new component ν_h to a specified fixed value (e.g. 5).

d.3) Given the starting parameters of the new mixture, adapt the candidate for the model parameters by performing ISEM based on the draws from the previous mixture $g_t^{(H)}(\cdot)$ and the corresponding weights.

d.4) Draw $\boldsymbol{\alpha}_t^{(j)}$ from the new mixture $g_t^{(h+1)}(\cdot)$ and evaluate the corresponding normalized importance weights $w_t^{(j)}$, $j = 1, \ldots, M$.

d.5) Calculate the coefficient of variation $\text{CV}^{(h+1)}$ of the normalized weights $w_t^{(j)}$, $j = 1, \ldots, M$.

e) **Assess convergence** of the candidate density's quality by inspecting whether the relative change between $\text{CV}^{(H)}$ and $\text{CV}^{(h+1)}$ is greater than the chosen threshold (e.g. 0.01) and return to step d) unless the algorithm has converged.

3) **Draws**. Draw $\boldsymbol{\alpha}_t^{(j)}$ from the constructed density $g_t^{(H)}(\boldsymbol{\alpha}_t^{(j)}|\boldsymbol{\alpha}_{t-1}^{(j)})$ and approximate $\text{E}[h_t(\boldsymbol{\alpha}_t)|r_{1:T}]$ by:

$$\hat{h}(\boldsymbol{\alpha}_t) = \sum_{j=1}^{M} w_t^{(j)} h(\boldsymbol{\alpha}_t^{(j)}).$$

4) **Likelihood Approximation**. The approximation of the log likelihood function is given by:

$$\log \hat{p}(r_{1:T}) = \sum_{t=1}^{T} \log \left(\frac{1}{M} \sum_{j=1}^{M} \tilde{w}_t^{(j)} \right).$$

Appendix 5.C Simulation results for MFilter

Below we report simulation results confirming a good statistical performance of the MFilter. In all the examples we are interested in the estimation of the target function $h_t(\boldsymbol{\alpha}_t^{(j)}) = \boldsymbol{\alpha}_t$ that is the posterior mean of the latent state. We compare four filters, the Kalman filter (KF), the Bootstrap Particle Filter (BPF), the Auxiliary Particle filter (APF) and the MFilter. All the Monte Carlo (MC) experiments presented in this section are based on R = 100 replications with T = 100 observations each. For the BPF, APF and MFilter we use $M = 50,000$ particles. In the MFilter the particles correspond to draws from the proposal density.

To quantify the performance of the filters we consider three measures: loglikelihood bias (LLB), absolute deviation (Bias) and variability (Var), with the two latter measures defined as

$$\text{Bias} = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{R}\sum_{i=1}^{R}|\tilde{\alpha}_{t,i} - \alpha_{t,i}|\right),$$
$$\text{Var} = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{R}\sum_{i=1}^{R}(\tilde{\alpha}_{t,i} - \alpha_{t,i})^2\right),$$

where $\tilde{\alpha}_{t,i}$ is the estimated posterior mean of the state at time t from the ith replication. We present the results with respect to the KF (applied to the original or the transformed model) and report LLB only for the local level model and the dynamic factor models (where the loglikelihood is available in a closed form).

5.C.1 Local level model

The first model we consider is a standard local level model:

$$\begin{aligned} y_t &= \alpha_t + \varepsilon_t, & \varepsilon_t &\sim \mathcal{N}(0, \sigma_\varepsilon^2), \\ \alpha_t &= \alpha_{t-1} + \eta_t, & \eta_t &\sim \mathcal{N}(0, \sigma_\eta^2), \end{aligned} \tag{5.C.1}$$

which is a linear and Gaussian model often used as benchmark for comparing filtering methods. In this case KF provides the sequential state distribution in analytical form and is the optimal filter.

In the simulations experiments, we fix the latent state variance at $\sigma_\eta^2 = 0.1$ and we define four different levels for the state variance σ_ε^2, corresponding to four levels of the Noise to Signal Ratio (NtS): 0.1, 0.5, 1 and 2.5. We note that the exact likelihood of

the model in equation (5.C.1) can be calculated using the KF, and we can compare the exact likelihood of this model with the remaining non-linear filters. This allows to assess the degree of the likelihood bias in the non-linear filters, including the proposed MFilter.

Table 5.C.1 reports the results for the model in equation (5.C.1). KF filter is the best filter, as expected, in terms of the minimum bias and the smallest computing time. The results of the non-linear filters, however, are in line with those of KF in terms of the bias measures. The proposed MFilter performs similarly to the BPF and the APF but has a lower bias in the estimate likelihood especially for smallest NtS ratio of 0.1. In all cases the computing time is lower then the BPF and APF.

NtS	*0.1*			*0.5*			*Time*	
Model	LLB	Bias	Var	LLB	Bias	Var	0.1	0.5
KF	0.00	1.00	1.00	0.00	1.00	1.00	0.01	0.01
BPF	-48.93	1.22	1.48	-19.43	1.26	1.62	33.71	35.55
APF	-13.87	1.00	1.00	-9.56	1.01	1.02	35.54	37.67
MFilter	-10.40	1.00	1.01	-9.52	1.01	1.02	12.83	12.81
NtS	*1*			*2.5*			*Time*	
Model	LLB	Bias	Var	LLB	Bias	Var	1	2.5
KF	0.00	1.00	1.00	0.00	1.00	1.00	0.01	0.01
BPF	-37.85	1.31	1.71	-21.16	1.43	2.04	35.22	34.53
APF	-10.43	1.00	1.00	-9.05	1.00	1.00	37.29	35.72
MFilter	-10.18	1.01	1.01	-9.39	1.00	1.01	12.67	12.13

Table 5.C.1: MC results for the linear and Gaussian model (5.C.1). Loglikelihood Bias (LLB), absolute deviation (Bias) and variability (Var) with respect to the KF. Final column: computing time in seconds for different NtS.

5.C.2 Stochastic volatility model

The second model is the SV model of Kim et al. (1998) given by

$$\begin{aligned} y_t &= \exp\left(\alpha_t/2\right)\varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2), \\ \alpha_t &= \mu + \phi\alpha_{t-1} + \eta_t, \quad \eta_t \sim \mathcal{N}(0, \sigma_\eta^2), \end{aligned} \tag{5.C.2}$$

where η_t and ε_t are independent and y_t is the observed series. Due to the non-linear structure of the observation equation the analytical form for filtering and predictive densities do not exist in this model.

In the simulations, we fix the autoregressive parameter ϕ to 0.90, 0.95, and 0.98, which are in line with the values found in other studies, see for example Aguilar and West

(2000). For each value of ϕ we consider four values of σ_η^2, that correspond to different coefficient of variation (CV) of the volatility $h = \sigma_\eta^2 \exp(\alpha_t)$ defined as:

$$\text{CV} = \frac{\text{Var}(h)}{\text{E}(h)^2} = \exp\left(\frac{\sigma_\eta^2}{1-\phi^2}\right) - 1.$$

The CV takes values 0.1, 0.5, 1, and 2.5 where high values indicate more strength of the volatility process and low values indicate that the volatility is close to a constant.

Table 5.C.2 reports the results for the SV model of equation (5.C.2) with $\phi = 0.98$ and different values of σ_η^2 that corresponds to CV $= 0.1, 0.5, 1, 2$. In all cases the KF is the worst filter due to being a linear and Gaussian filter. The MFilter performs similarly to the BPF and the APF in term of bias and estimation variability. In this model the computational speed is comparable between the three non-linear filters, namely BPF, APF and MFilter.

CV	*0.1*		*0.5*		*Time*	
Model	Bias	Var	Bias	Var	0.1	0.5
KF	1.00	1.00	1.00	1.00	0.01	0.01
BPF	0.24	0.10	0.31	0.12	13.82	13.99
APF	0.25	0.10	0.31	0.13	14.58	14.66
MFilter	0.26	0.10	0.31	0.14	14.15	12.67
CV	*1.0*		*2.5*		*Time*	
Model	Bias	Var	Bias	Var	1	2.5
KF	1.00	1.00	1.00	1.00	0.01	0.01
BPF	0.32	0.12	0.29	0.11	13.98	13.88
APF	0.31	0.13	0.29	0.11	14.61	14.70
MFilter	0.30	0.13	0.28	0.11	13.54	12.96

Table 5.C.2: MC results for the SV model (5.C.2) with $\phi = 0.98$ and CV $= 0.1, 0.5, 1, 2.5$. Absolute deviation (Bias) and variability (Var) with respect to the KF. Final column: computing time in seconds with different CV.

5.C.3 Dynamic factor model

The last model we examine is a DFM given by

$$\begin{aligned} \mathbf{y}_t &= \Lambda \mathbf{f}_t + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(0, \Sigma), \\ \mathbf{f}_t &= \Phi_1 \mathbf{f}_{t-1} + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim \mathcal{N}(0, \mathrm{Q}), \end{aligned} \tag{5.C.3}$$

where $\mathbf{y}_t$ is a $N \times 1$ vector of time series,the $K \times 1$ vector $\mathbf{f}_t$ contains unobservable factors with one lag where Φ_1 is a $K \times K$ matrix of autoregressive coefficients, Λ is an

$N \times K$ matrix of factor loadings. Finally, $\boldsymbol{\varepsilon}_t$ is an $N \times 1$ i.i.d. vector of idiosyncratic disturbances and $\boldsymbol{\eta}_t$ is an $K \times 1$ i.i.d. vector of latent disturbances.

The model in equation (5.C.3) is linear and Gaussian and as in equation (5.C.1) the KF is the optimal filter. As before we compare the performance of the non-linear filters against the KF for different number of factors.

Table 5.C.3 reports the results for the model in equation (5.C.3) for $R = 100$ MC replication, $N = 20$ series and $K = 2, 4, 6, 10$ factors. In all simulation experiments the following simulation setting is used: Λ is a $N \times K$ matrix with zeros on the $K \times (K-1)/2$ upper-diagonal elements and the remaining elements being i.i.d. standard normal; Φ_1 is a diagonal matrix with elements being i.i.d. uniform on $[0, 1)$; Σ is a diagonal matrix with elements being i.i.d. uniform on $[0, 2.5]$; Q is a diagonal matrix for which the simulations are performed in two steps. First, we simulate $\tilde{\mathrm{Q}} = \Psi\Psi'$ where Ψ is a $K \times K$ upper triangular matrix with elements simulated from independent uniform distributions in $[0, 1]$. The diagonal elements of Q are defined as the diagonal elements of $0.1 \times \tilde{\mathrm{Q}}^{-1}$. We note that our general conclusions hold under different parameter values such as $\Phi_1 = 0.9$ as well as for different specifications for the non-zero elements of Λ.

Due to the linear Gaussian model structure in equation (5.C.3), the KF leads to the best results in terms of the speed and accuracy, but the non-linear filters are in line with the KF. The MFilter performs better than both the BPF and APF, with substantially lower bias and variance. The MFilter has also the lowest likelihood bias compared to the other nonlinear non-Gaussian filters. For all the filters the computing time increases with the number of factors. In all the cases, however, the MFilter requires less computing time than the BPF and APF.

Factors	*2*			*4*			*Time*	
Model	LLB	Bias	Var	LLB	Bias	Var	2	4
KF	0	1	1	0	1	1	0.01	0.01
BPF	-77.42	1.15	1.33	-145.49	1.15	1.32	708.79	811.73
APF	-39.98	1.03	1.05	-164.80	1.05	1.05	836.69	878.13
MFilter	-23.23	1.01	1.02	-23.39	1.00	1.01	106.33	138.18
Factors	*6*			*10*			*Time*	
Model	LLB	Bias	Var	LLB	Bias	Var	6	10
KF	0.00	1.00	1.00	0.00	1.00	1.00	0.02	0.02
BPF	-193.74	1.16	1.31	-333.33	1.27	1.65	861.10	897.86
APF	-309.26	1.07	1.12	-568.18	1.08	1.18	953.72	1011.21
MFilter	-16.97	1.03	1.03	-112.68	1.02	1.03	213.20	402.82

Table 5.C.3: MC results for the DFM with $N = 20$ and $K = 2, 4, 6, 10$ latent factors. Loglikelihood Bias (LLB), absolute deviation and variability with respect to the KF. Final column: computing time in seconds with $K = 2, 4, 6, 10$.

Appendix 5.D Additional empirical results

5.D.1 Individual model-strategy pairs

Table 5.D.1 presents detailed results on the properties of the realised returns from 86 individual combinations of models and strategies, which are summarised in the main text in Table 5.4.3.

5.D.2 Combinations of model-strategy pairs

Table 5.D.2 presents the detailed results for the third stage of the analysis in Subsection 5.4.2, i.e. for the combination of very flexible FAVAR-SV models and the two strategies (MM and RM).

		MM				RM			
	(K, L)	Mean	Vol.	SR	LL	Mean	Vol.	SR	LL
VAR-N	–	0.02	**5.0**	0.005	**-24.1**	**0.09**	5.8	0.015	-35.0
SV	–	**0.10**	**5.1**	0.019	-34.7	**0.11**	**5.6**	0.019	**-26.0**
VAR-SV	–	**0.12**	**4.5**	**0.028**	-20.2	**0.13**	5.8	**0.021**	-37.4
DFM-N	(1,1)	-0.04	**4.9**	-0.009	**-20.0**	**0.13**	**5.7**	**0.023**	-34.4
	(1,2)	-0.04	**4.9**	-0.009	**-20.0**	**0.13**	**5.7**	**0.022**	-34.4
	(2,1)	-0.13	**5.2**	-0.024	**-25.4**	**0.10**	**5.6**	0.017	-34.0
	(2,2)	-0.11	**5.2**	-0.020	**-24.2**	**0.10**	**5.6**	0.017	-34.1
	(3,1)	-0.14	**5.4**	-0.027	**-23.7**	**0.09**	**5.5**	0.017	-33.7
	(3,2)	-0.08	**5.4**	-0.016	**-23.3**	0.08	**5.4**	0.015	-33.1
	(4,1)	-0.07	**5.5**	-0.013	-26.7	**0.10**	**5.4**	0.018	-31.3
	(4,2)	-0.05	**5.5**	-0.009	-27.4	**0.12**	**5.4**	**0.022**	-31.1
DFM-SV	(1,1)	0.04	**5.0**	**0.007**	**-20.0**	**0.11**	5.8	0.019	-37.1
	(1,2)	0.04	**5.0**	**0.008**	**-20.0**	**0.10**	5.8	0.018	-37.1
	(2,1)	-0.04	**5.2**	-0.009	**-22.0**	**0.15**	**5.7**	**0.026**	-36.3
	(2,2)	-0.05	**5.2**	-0.009	**-22.0**	**0.15**	**5.7**	**0.027**	-36.6
	(3,1)	0.00	**5.2**	0.000	**-21.2**	**0.14**	**5.4**	**0.026**	-33.0
	(3,2)	0.03	**5.2**	0.005	**-20.8**	**0.16**	**5.4**	**0.030**	-32.8
	(4,1)	**0.12**	**5.4**	**0.023**	**-20.8**	0.05	**5.4**	0.009	-31.8
	(4,2)	**0.12**	**5.4**	**0.023**	**-21.7**	0.06	**5.4**	0.011	-31.1
DFM-SV2	(1,1)	0.07	**4.6**	0.014	**-18.2**	0.06	**5.5**	0.010	-37.4
	(1,2)	0.07	**4.6**	0.014	**-18.2**	0.06	**5.5**	0.010	-37.4
	(2,1)	-0.01	**4.8**	-0.002	**-22.8**	0.08	**5.5**	0.015	-37.4
	(2,2)	-0.02	**4.8**	-0.003	**-22.8**	**0.09**	**5.5**	0.016	-37.4
	(3,1)	0.02	**5.0**	0.005	-27.1	-0.02	**5.5**	-0.003	-37.4
	(3,2)	0.03	**5.0**	0.006	-27.1	-0.02	**5.5**	-0.003	-37.4
	(4,1)	0.07	**5.7**	0.013	-32.3	0.00	**5.2**	0.000	-37.4
	(4,2)	0.07	**5.7**	0.013	-32.3	0.00	**5.2**	0.000	-37.4
FAVAR-SV	(1,1)	0.08	**4.6**	0.018	**-18.3**	0.06	**5.5**	0.011	-37.4
	(1,2)	0.08	**4.6**	0.018	**-18.3**	0.06	**5.5**	0.011	-37.4
	(2,1)	-0.03	**4.9**	-0.005	**-23.1**	0.08	**5.5**	0.015	-37.4
	(2,2)	-0.03	**4.9**	-0.006	**-23.5**	**0.09**	**5.5**	0.016	-37.4
	(3,1)	**0.09**	**5.0**	0.018	**-25.3**	-0.02	**5.5**	-0.005	-37.4
	(3,2)	0.08	**5.0**	0.017	**-25.7**	-0.02	**5.5**	-0.004	-37.4
	(4,1)	0.08	**5.7**	0.014	-32.3	0.03	**5.2**	0.005	-37.4
	(4,2)	0.08	**5.7**	0.015	-32.3	0.02	**5.2**	0.005	-37.4
FAVAR-SV2	(1,1)	**0.09**	**4.6**	0.019	**-18.3**	0.06	**5.5**	0.011	-37.4
	(1,2)	0.08	**4.6**	0.018	**-18.3**	0.06	**5.5**	0.011	-37.4
	(2,1)	-0.03	**4.9**	-0.005	**-23.5**	**0.09**	**5.5**	0.016	-37.4
	(2,2)	-0.03	**4.9**	-0.005	**-23.8**	0.08	**5.5**	0.015	-37.4
	(3,1)	0.08	**5.0**	0.017	**-25.6**	-0.03	**5.5**	-0.005	-37.4
	(3,2)	0.08	**5.0**	0.017	**-25.3**	-0.02	**5.5**	-0.004	-37.4
	(4,1)	0.08	**5.7**	0.014	-32.3	0.03	**5.2**	0.005	-37.4
	(4,2)	0.08	**5.7**	0.014	-32.3	0.03	**5.2**	0.005	-37.4

Table 5.D.1: Mean returns and risk measures (volatility [Vol.], Sharpe Ratio [SR], and the largest loss [LL]) for the realised return densities from all model-strategy pairs, with models from Section 5.1 and strategies being MM and RM. Measures from the SM strategy: mean 0.09, volatility 5.7, Sharpe ratio 0.02 and largest loss -26.2. Bold values: an 'equal or better' value compared to SM. K: the number of factors, L: the number of lags.

Model	Strategy	Mean	Vol.	SR	LL
Combination of component models and two strategies					
FAVAR-SV(1-4, 1-2)	MM & RM	0.18 (0.14, 0.22)	4.5 (4.5, 4.6)	0.039 (0.031, 0.048)	-34.8 (-35.0, -34.6)
Combination of two strategies per component model					
FAVAR-SV(1, 1)	MM & RM	0.11 (0.02, 0.19)	4.5 (4.4, 4.6)	0.024 (0.004, 0.042)	-33.8 (-34.0, -33.1)
FAVAR-SV(1, 2)	MM & RM	0.11 (0.02, 0.19)	4.5 (4.4, 4.6)	0.023 (0.004, 0.042)	-34.2 (-34.4, -33.6)
FAVAR-SV(2, 1)	MM & RM	0.14 (0.05, 0.22)	5.1 (5.0, 5.2)	0.027 (0.010, 0.043)	-37.1 (-37.2, -36.9)
FAVAR-SV(2, 2)	MM & RM	0.14 (0.05, 0.22)	5.1 (5.0, 5.2)	0.027 (0.010, 0.044)	-37.1 (-37.2, -36.8)
FAVAR-SV(3, 1)	MM & RM	0.15 (0.07, 0.25)	4.7 (4.5, 4.9)	0.033 (0.014, 0.054)	-34.1 (-34.3, -34)
FAVAR-SV(3, 2)	MM & RM	0.14 (0.05, 0.25)	4.7 (4.6, 4.9)	0.031 (0.011, 0.052)	-34.4 (-34.5, -34.2)
FAVAR-SV(4, 1)	MM & RM	0.11 (0.02, 0.20)	5.1 (5.0, 5.2)	0.022 (0.004, 0.040)	-31.3 (-31.8, -31.1)
FAVAR-SV(4, 2)	MM & RM	0.12 (0.03, 0.21)	5.1 (5.0, 5.2)	0.023 (0.005, 0.040)	-31.5 (-32.4, -31.3)

Table 5.D.2: Mean returns and risk measures (volatility [Vol.], Sharpe Ratio [SR], and the largest loss [LL]) for the realised return densities from combinations of flexible parametric models and two investment strategies. Top panel: results for the combination consisting of flexible models FAVAR-SV(1-4,1-2) and two investment strategies (MM, RM). Bottom panel: results for the combination of two investment strategies combined with each component model separately. Measures from the SM strategy: mean 0.09, volatility 5.7, Sharpe ratio 0.02 and largest loss -26.2. Bold values: an 'equal or better' value compared to SM. 90% CI in parentheses.

Chapter 6

Summaries

6.1 English summary

This thesis investigates Bayesian inference over time series models with the emphasis put on applications in economics and finance. We adopt simulation-based techniques which are necessary in any nontrivial problem in this setting. The main motivation behind the presented research is to increase the efficiency and accuracy of these computationally intensive methods in several different contexts. One of the main topics addressed is efficient and precise risk estimation, or rare event analysis. Another problem studied in this thesis is the efficiency of various sampling algorithms, in particular importance sampling (IS) and Markov chain Monte Carlo (MCMC) algorithms. Finally, we address the issue of forecasting, from a single model as well as from a combination of models.

In Chapter 2 we present an accurate and efficient method for Bayesian estimation of two financial risk measures, Value-at-Risk and Expected Shortfall, for a given volatility model. We obtain precise forecasts of the tail of the distribution of returns not only for the 10-days-ahead horizon required by the Basel Committee but even for long horizons, like one-month or one-year ahead. The key insight behind our proposed IS based approach is the sequential construction of marginal and conditional importance densities for consecutive periods. By oversampling the extremely negative scenarios and giving them lower importance weights, we achieve a much higher precision in characterising the properties of the left tail.

In Chapter 3 we introduce a novel approach to inference for a specific region of the predictive distribution. An important domain of application is accurate prediction

of financial risk measures, where the area of interest is the left tail of the predictive density of logreturns. Our proposed approach originates from the Bayesian approach to parameter estimation and time series forecasting, however it is robust in the sense that it provides a more accurate estimation of the predictive density in the region of interest in case of misspecification. The main contribution of this chapter is the novel concept of the partially censored posterior, where the set of model parameters is partitioned into two subsets: for the first subset of parameters we consider the standard marginal posterior, for the second subset of parameters (that are particularly related to the region of interest) we consider the conditional censored posterior. This approach yields more precise parameter estimation than a fully censored posterior for all parameters, and has more focus on the region of interest than a standard Bayesian approach.

In Chapter 4 we develop a novel efficient model-fitting algorithm for state space models. This flexible class of models is challenging due to their substantially more complicated fitting to data as the associated likelihood is typically analytically intractable. For the general case a Bayesian data augmentation approach is often employed, however, standard "vanilla" updating MCMC algorithms may perform very poorly in that case. This it due to high correlation between the imputed states and/or parameters and leads to the need for specialist algorithms. A Semi-Complete Data Augmentation algorithm circumvents the inefficiencies of the previous approaches by combining data augmentation with numerical integration in a Bayesian hybrid approach. This approach permits the use of standard "vanilla" updating algorithms that perform considerably better than the traditional approach in terms of considerably improved mixing and hence lower autocorrelation.

In Chapter 5 we propose a novel dynamic asset allocation approach in which model-based forecasts are directly combined with a set of data driven portfolio strategies, without the necessity to define a utility or other scoring function. The resulting dynamic asset-allocation model is specified as a combination of return distributions stemming from multiple pairs of models and strategies. The combination weights are defined through feedback mechanisms that enable learning, to allow for cross-correlation and correlation over time. To increase the efficiency and robustness of the simulations we introduce a new nonlinear filter based on mixtures of Student's t distributions. Diagnostic analysis of posterior residuals gives insight into the model and strategy incompleteness or misspecification.

6.2 Nederlandse samenvatting

Dit proefschrift onderzoekt de Bayesiaanse inferentie over tijdreeksmodellen met de nadruk op economische en financiële toepassingen. We passen simulatietechnieken toe. De belangrijkste motivatie achter het gepresenteerde onderzoek is om de efficiëntie en nauwkeurigheid van deze rekenintensieve methoden te verhogen. Een van de belangrijkste onderwerpen die worden behandeld is een efficiënte en nauwkeurige risicoschatting, of een analyse van zeldzame gebeurtenissen. Een ander probleem dat in dit proefschrift wordt bestudeerd, is de efficiëntie van verschillende simulatie algoritmen, met name Importance Sampling (IS) en Markov-keten Monte Carlo (MCMC) algoritmen. Ten slotte behandelen we het probleem van sta:s:sche voorspellingen, zowel met behulp van een enkel model als van een combinatie van modellen.

In Hoofdstuk 2 presenteren we een nauwkeurige en efficiënte methode voor de Bayesiaanse schatting van twee financiële risicomaatstaven, Value-at-Risk en Expected Shortfall, voor een gegeven volatiliteitsmodel. We krijgen nauwkeurige voorspellingen van de staart van de verdeling van de rendementen, niet alleen voor de horizon van 10 dagen vooruit die het Bazels Comité nodig heeft, maar zelfs voor een lange horizon, zoals een maand of een jaar vooruit. Het belangrijkste inzicht achter onze voorgestelde op IS gebaseerde aanpak is de sequentiële constructie van marginale en conditionele "importance" dichtheden voor 'opeenvolgende perioden. Door de extreem negatieve scenario's te vervangen en deze lage belangrijkheidsgewichten te geven, bereiken we een veel hogere precisie bij het karakteriseren van de eigenschappen van de linkerstaart.

In Hoofdstuk 3 introduceren we een nieuwe benadering van inferentie voor een specifieke regio van de voorspellende verdeling. Een belangrijke toepassing is het nauwkeurig voorspellen van financiële risicomaatstaven, waarbij het aandachtsgebied is de linker staart van de verdeling van logreturns. Wij volgen de Bayesiaanse benadering van parameterschatting en tijdreeksvoorspelling. De methode is robuust in de zin dat het een nauwkeuriger schatting geeft van de voorspellende dichtheid in het gebied. Maar de belangrijkste bijdrage van dit hoofdstuk is het nieuwe concept van de gedeeltelijk gecensureerde posterior, waarbij de set modelparameters is verdeeld in twee subsets: voor de eerste subset van parameters beschouwen we de standaard marginale posterior, voor de tweede subset van parameters beschouwen we de voorwaardelijke gecensureerde posterior. Deze benadering levert nauwkeuriger schattingen van de parameters op, nauwkeuriger dan met een volledig gecensureerde posterior voor alle parameters. Daarnaast heeft deze methode meer aandacht voor het "importance" gebied dan een standaard Bayesiaanse benadering.

In Hoofdstuk 4 ontwikkelen we een nieuw efficiënt algoritme voor state space-modellen. Deze flexibele klasse van modellen is een uitdaging vanwege hun aanzienlijk gecompliceerdere aanpassing aan gegevens, omdat de bijbehorende loglikelihood doorgaans geen analytische oplossing heeft. Voor het algemene geval wordt vaak een Bayesiaanse methode voor data augmentation gebruikt, maar standaard MCMC algoritmen kunnen in dat geval zeer slecht presteren, voornamelijk vanwege de hoge correlatie tussen de geïmputeerde states. Dit leidt tot de noodzaak om gespecialiseerde algoritmen te ontwikkelen. Het voorgestelde Semi-Complete Data Augmentation-algoritme omzeilt de inefficiënties van de eerdere benaderingen door data-augmentatie te combineren met numerieke integratie in een Bayesiaanse hybride aanpak. Met deze aanpak kunnen standaard algoritmen worden toegepast voor het bijwerken van de geïmputeerde states die aanzienlijk beter presteren dan de traditionele aanpak.

In Hoofdstuk 5 stellen we een nieuwe benadering voor van dynamische activaspreiding waarbij model-prognoses direct worden gecombineerd met een reeks van portfoliostrategieën, zonder de noodzaak om een utiliteits- of andere score-functie te definiëren. Het resulterende dynamische model wordt gespecificeerd als een combinatie van rendementverdelingen die afkomstig zijn van meerdere paren van modellen en strategieën. De combinatiegewichten worden gedefinieerd via feedback-mechanismen en worden steeds aangepast. Om de efficiëntie en robuustheid van de simulaties te vergroten, introduceren we een nieuw niet-lineair filter op basis van een mix van Student t verdelingen. Diagnostische analyse van de residuen geeft inzicht in de incompleetheid of verkeerde specificatie van het model.

Bibliography

Aastveit, K. A., L. Hoogerheide, J. Mitchell, and H. K. van Dijk (2018a), "Structure and Workings of Density Forecast Combinations in Economics." Unpublished manuscript.

Aastveit, K. A., J. Mitchell, F. Ravazzolo, and H. K. van Dijk (2018b), "The Evolution of Forecast Density Combinations in Economics." To appear in Oxford Research Encyclopedia of Economics and Finance.

Abadi, F., O. Gimenez, B. Ullrich, R. Arlettaz, and M. Schaub (2010), "Estimation of Immigration Rate using Integrated Population Models." *Journal of Applied Ecology*, 393–400.

Aguilar, O. and M. West (2000), "Bayesian Dynamic Factor Models and Portfolio Allocation." *Journal of Business & Economic Statistics*, 18, 338–357.

Amisano, G. and R. Giacomini (2007), "Comparing Density Forecasts via Weighted Likelihood Ratio Tests." *Journal of Business and Economic Statistics*, 25, 177–190.

Andrieu, C., A. Doucet, and R. Holenstein (2010), "Particle Markov Chain Monte Carlo Methods." *Journal of the Royal Statistical Society Series B*, 72, 269–342.

Andrieu, C. and G. Roberts (2009), "The Pseudo-Marginal Approach for Efficient Monte Carlo Computations." *Annals of Statistics*, 37, 697–725.

Artzner, P., F. Delbaen, J. M. Eber, and D. Heath (1999), "Coherent Measures of Risk." *Mathematical Finance*, 9, 203–228.

Asness, C. S., T. J. Moskowitz, and L. H. Pedersen (2013), "Value and Momentum Everywhere." *The Journal of Finance*, 68, 929–985.

Baştürk, N., A. Borowska, S. Grassi, L. Hoogerheide, and H. K. van Dijk (2018), "Forecast Density Combinations of Dynamic Models and Data Driven Portfolio Strategies." *Journal of Econometrics*, 210, 170–186.

Baştürk, N., S. Grassi, L. Hoogerheide, A. Opschoor, and H. K. van Dijk (2017), "The R Package MitISEM: Efficient and Robust Simulation Procedures for Bayesian Inference." *Journal of Statistical Software, Articles*, 79, 1–40.

Bai, J. and W. P. Peng (2015), "Identification and Bayesian Estimation of Dynamic Factor Models." *Journal of Business & Economic Statistics*, 33, 221–240.

Basel Committee on Banking Supervision (1995), "An Internal Model-based Approach to Market Risk Capital Requirements." *The Bank for International Settlements, Basel, Switzerland*.

Baştürk, N., S. Grassi, L. Hoogerheide, and H. K. van Dijk (2016), "Parallelization Experience with Four Canonical Econometric Models Using ParMitISEM." *Econometrics*, 4, 1–11.

Baştürk, N., L. F. Hoogerheide, and H. K. van Dijk (2018), "Bayes Estimates of Multimodal Density Features in DNA and Economic Data." Technical report, Tinbergen Institute.

Beaumont, M. (2003), "Estimation of Population Growth or Decline in Genetically Monitored Populations." *Genetics*, 164, 1139–1160.

Bernanke, S. B., J. Boivin, and P. Eliasz (2005), "Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach." *The Quarterly Journal of Economics*, 120, 387–422.

Besbeas, P., S. N. Freeman, B. J. T. Morgan, and E. A. Catchpole (2002), "Integrating Mark–Recapture–Recovery and Census Data to Estimate Animal Abundance and Demographic Parameters." *Biometrics*, 58, 540–547.

Besbeas, P. and B. J. T. Morgan (2018), "Exact Inference for Integrated Population Modelling." Technical report.

Billio, M., R. Casarin, F. Ravazzolo, and H. K. van Dijk (2013), "Time-Varying Combinations of Predictive Densities using Nonlinear Filtering." *Journal of Econometrics*, 177, 213–232.

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe (2017), "Variational Inference: A Review for Statisticians." *Journal of the American Statistical Association*, 112, 859–877.

Blitz, D., J. Huij, and M. Martens (2011), "Residual Momentum." *Journal of Empirical Finance*, 18, 506–521.

Bollerslev, T. (1986), "Generalised Autoregressive Conditional Heteroskedasticity." *Journal of Econometrics*, 51, 307–327.

Bollerslev, T. (2008), "Glossary to ARCH (GARCH)." Technical Report 2008-49, CREATES Research Paper.

Bos, C. (2011), "Relating Stochastic Volatility Estimation Methods." Technical Report 11-049/4, Tinbergen Institute.

Brooks, S. P., R. King, and B. J. T. Morgan (2004), "A Bayesian Approach to Combining Animal Abundance and Demographic Data." *Animal Biodiversity and Conservation*, 27, 515–529.

Cappé, O., E. Moulines, and T. Ryden (2006), *Inference in Hidden Markov Models*. Springer Series in Statistics, Springer New York.

Carter, C. and R. Kohn (1994), "On Gibbs Sampling for State Space Models." *Biometrika*, 81, 541–553.

Carvalho, C. M., H. F. Lopes, and O. Aguilar (2011), "Dynamic Stock Selection Strategies: A Structured Factor Model Framework." In *Bayesian Statistics 9* (J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West, eds.), 1–21, Oxford University Press, Oxford.

Casella, G. and C. P. Robert (1996), "Rao-Blackwellisation of Sampling Schemes." *Biometrika*, 83, 81–94.

Chan, J., R. Leon-Gonzalez, and R. W. Strachan (2018), "Invariant Inference and Efficient Computation in the Static Factor Model." *Journal of the American Statistical Association*, 113, 819–828.

Chan, J. C. C. (2017), "The Stochastic Volatility in Mean Model with Time-Varying Parameters: An Application to Inflation Modeling." *Journal of Business & Economic Statistics*, 35, 17–28.

Chan, L. K. C., N. Jegadeesh, and J. Lakonishok (1996), "Momentum Strategies." *The Journal of Finance*, 51, 1681–1713.

Christoffersen, P. F., F. X. Diebold, and T. Schuermann (1998), "Horizon Problems and Extreme Events in Financial Risk Management." *Economic Policy Review*, 109–118.

Cox, D. R. (1981), "Statistical Analysis of Time Series: Some Recent Developments." *Scandinavian Journal of Statistics*, 8, 93–115.

Creal, D. (2012), "A Survey of Sequential Monte Carlo Methods for Economics and Finance." *Econometric Reviews*, 31, 245–296.

Creal, D., S. J. Koopman, and A. Lucas. (2013), "Generalized Autoregressive Score Models with Applications." *Journal of Applied Econometrics*, 28, 777–795.

Cross Validated (2017), "Why should I be Bayesian when my model is wrong?" `https://stats.stackexchange.com/questions/274815/why-should-i-be-bayesian-when-my-model-is-wrong`. Accessed: 2017-07-18.

Daníelsson, J. and J. P. Zigrand (2006), "On Time-Scaling of Risk and the Square-Root-of-Time Rule." *Journal of Banking & Finance*, 30, 2701–2713.

De Roon, F. and P. Karehnke (2016), "A Simple Skewed Distribution with Asset Pricing Applications." *Review of Finance*, 1–29.

Del Moral, P., A. Doucet, and A. Jasra (2006), "Sequential Monte Carlo Samplers." *Journal of the Royal Statistical Society: Series B*, 68, 411–436.

Del Negro, M., R. B. Hasegawa, and F. Schorfheide (2016), "Dynamic Prediction Pools: An Investigation of Financial Frictions and Forecasting Performance." *Journal of Econometrics*, 192, 391–405.

DeMiguel, V., L. Garlappi, and R. Uppal (2007), "Optimal Versus Naive Diversification: How Inefficient is the 1/N Portfolio Strategy?" *The Review of Financial Studies*, 22, 1915–1953.

Dempster, A.P., N. M. Laird, and D. B. Rubin (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society Series B*, 39, 1–38.

Diebold, F. X., A. Hickman, A. Inoue, and T. Schuermann (1997), "Converting 1-Day Volatility to h-Day Volatitlity: Scaling by $\sqrt{h}$ is Worse Than You Think." Technical Report 97–34, Wharton Financial Institutions Center Working Papers.

Diebold, F. X. and R. S. Mariano (1995), "Comparing Predictive Accuracy." *Journal of Business and Economic Statistic*, 13, 253–263.

Diks, C., V. Panchenko, and D. van Dijk (2011), "Likelihood-based Scoring Rules for Comparing Density Forecasts in Tails." *Journal of Econometrics*, 163, 215–230.

Douc, R. and C. P. Robert (2011), "A Vanilla Rao–Blackwellization of Metropolis–Hastings Algorithms." *The Annals of Statistics*, 39, 261–277.

Doucet, A., N. de Freitas, and N. Gordon, eds. (2001), *Sequential Monte Carlo Methods in Practice*. Springer.

Doucet, A., N. De Freitas, K. Murphy, and S. Russell (2000a), "Rao-Blackwellised Particle Filtering for Dynamic Bayesian Networks." In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, 176–183.

Doucet, A., S. Godsill, and C. Andrieu (2000b), "On Sequential Monte Carlo Sampling Methods for Bayesian Filtering." *Statistics and Computing*, 10, 197–208.

Durbin, J. and S. J. Koopman (2012), *Time Series Analysis by State Space Methods: Second Edition*. Oxford Statistical Science Series, OUP Oxford.

Embrechts, P., R. Kaufmann, and P. Patie (2005), "Strategic Long-Term Financial Risks: Single Risk Factors." *Computational Optimization and Applications*, 32, 61–90.

Engle, R. and R. Colacito (2006), "Testing and Valuing Dynamic Correlations for Asset Allocation." *Journal of Business & Economic Statistics*, 24, 238–253.

Engle, R. F. (1982), "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of the United Kingdom Inflation." *Econometrica*, 50, 987–1007.

Engle, R. F. (2009), "The Risk That Risk Will Change." *Journal of Investment Management*, 7, 24–28.

Engle, R. F. and V. K. Ng (1993), "Measuring and Testing the Impact of News on Volatility." *Journal of Finance*, 48, 1749–1778.

Fama, E. F. and K. R. French (1992), "The Cross-Section of Expected Stock Returns." *The Journal of Finance*, 47, 427–465.

Fama, E. F. and K. R. French (1993), "Common Risk Factors in the Returns on Stocks and Bonds." *Journal of Financial Economics*, 33, 3–56.

Fama, E. F. and K. R. French (2015), "A Five-Factor Asset Pricing Model." *Journal of Financial Economics*, 116, 1–22.

Fridman, M. and L. Harris (1998), "A Maximum Likelihood Approach for non-Gaussian Stochastic Volatility Models." *Journal of Business & Economic Statistics*, 87, 284–291.

Frühwirth-Schnatter, S. (1994), "Data Augmentation and Dynamic Linear Models." *Journal of Time Series Analysis*, 15, 183–202.

Frühwirth-Schnatter, S. (2004), "Efficient Bayesian Parameter Estimation." In *State Space and Unobserved Component Models: Theory and Applications* (A. C. Harvey, S. J. Koopman, and N. Shephard, eds.), chapter 7, 123–151, Cambridge University Press.

Frühwirth-Schnatter, S. (2006), *Finite Mixture and Markov Switching Models*. Springer Verlag.

Frühwirth-Schnatter, S. and H. F. Lopes (2018), "Sparse Bayesian Factor Analysis when the Number of Factors is Unknown." ArXiv preprint 1804.04231.

Garlappi, L., R. Uppal, and T. Wang (2006), "Portfolio Selection with Parameter and Model Uncertainty: A Multi-prior Approach." *The Review of Financial Studies*, 20, 41–81.

Gatarek, L. T., L. F. Hoogerheide, K. Hooning, and H. K. van Dijk (2014), "Censored Posterior and Predictive Likelihood in Bayesian Left-tail Prediction for Accurate Value at Risk Estimation." Technical Report TI 2013-060/III, Tinbergen Institute Discussion Paper.

Gelfand, A. and A. Smith (1990), "Sampling Based Approaches to Calculating Marginal Densities." *Journal of the American Statistical Association*, 85, 398–409.

Gelman, A. (2011), "The Bias-Variance Tradeoff." `https://andrewgelman.com/2011/10/15/the-bias-variance-tradeoff/`. Accessed: 2018-10-17.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013), *Bayesian Data Analysis: Third Edition*. Chapman & Hall/CRC Texts in Statistical Science, Taylor & Francis.

Gelman, A., G. O. Roberts, and W. R. Gilks (1996), "Efficient Metropolis Jumping Rule." In *Bayesian Statistics 5: Proceedings of the Fifth Valencia International Meeting* (S. Brooks, J. Galin A. Gelman, and X. L. Meng, eds.), 599–607, Oxford University Press.

Geweke, J. (1977), "The Dynamic Factor Analysis of Economic Time Series." In *Latent Variables in Socio-Economic Models* (D. J. Aigner and A. S. Goldberger, eds.), North-Holland.

Geweke, J. (1989), "Bayesian Inference in Econometric Models using Monte Carlo Integration." *Econometrica*, 57, 1317–1739.

Geweke, J. and G. Amisano (2010a), "Comparing and Evaluating Bayesian Predictive Distributions of Asset Returns." *International Journal of Forecasting*, 26, 216–230.

Geweke, J. and G. Amisano (2010b), "Optimal Prediction Pools." *Journal of Econometrics*, 164, 130–141.

Geweke, J. and G. Amisano (2012), "Prediction with Misspecified Models." *The American Economic Review*, 102, 482–486.

Geweke, J. and G. Zhou (1996), "Measuring the Pricing Error of the Arbitrage Pricing Theory." *Review of Financial Studies*, 9, 557–587.

Geyer, C. J. (2011), "Introduction to Markov Chain Monte Carlo." In *Handbook of Markov Chain Monte Carlo* (S. Brooks, J. Galin A. Gelman, and X. L. Meng, eds.), chapter 4, 3–48, Chapman and Hall/CRC.

Ghysels, E., A. C. Harvey, and E. Renault (1996), "Stochastic volatility." In *Handbook of Statistics* (G. S. Maddala and C. R. Rao, eds.), volume 14, 119–191, Elsevier.

Glasserman, P., P. Heidelberger, and P. Shahabuddin (1999), *Importance Sampling and Stratification for Value-at-Risk*. IBM Thomas J. Watson Research Division.

Glasserman, P., P. Heidelberger, and P. Shahabuddin (2000), "Variance Reduction Techniques for Estimating Value-at-Risk." *Management Science*, 46, 1349–1364.

Glasserman, P., P. Heidelberger, and P. Shahabuddin (2002), "Portfolio Value-at-Risk with Heavy-Tailed Risk Factors." *Mathematical Finance*, 12, 239–269.

Glasserman, P. and J. Li (2005), "Importance Sampling for Portfolio Credit Risk." *Management science*, 51, 1643–1656.

Gordon, N. J., D. J. Salmond, and A. F. M. Smith (1993), "Novel Approach to Nonlinear/non-Gaussian Bayesian State Estimation." In *IEE Proceedings F (Radar and Signal Processing)*, volume 140, 107–113, IET.

Goudie, R. J. B., A. M. Presanis, D. Lunn, D. De Angelis, and L. Wernisch (2018), "Joining and Splitting Models with Markov Melding." *Bayesian Analysis*.

Green, P. J., K. Latuszyński, M. Pereyra, and C. P. Robert (2015), "Bayesian computation: a summary of the current state, and samples backwards and forwards." *Statistics and Computing*, 25, 835–862.

Gruber, L. F. and M. West (2017), "Bayesian Online Variable Selection and Scalable Multivariate Volatility Forecasting in Simultaneous Graphical Dynamic Linear Models." *Econometrics and Statistics*, 3, 3–22.

Hall, S. G. and J. Mitchell (2007), "Combining Density Forecasts." *International Journal of Forecasting*, 23, 1–13.

Hammersley, J. M. and D. C. Handscomb (1964), *Monte Carlo Methods*. Methuen.

Han, Y. (2006), "Asset Allocation with a High Dimensional Latent Factor Stochastic Volatility Model." *Review of Financial Studies*, 19, 237–271.

Hobert, J. P. (2011), "The Data Augmentation Algorithm: Theory and Methodology." In *Handbook of Markov Chain Monte Carlo* (S. Brooks, J. Galin A. Gelman, and X. L. Meng, eds.), chapter 10, 253–294, CRC Press.

Hobert, J. P., V. Royand, and C. P. Robert (2011), "Improving the Convergence Properties of the Data Augmentation Algorithm with an Application to Bayesian Mixture Modeling." *Statistical Science*, 26, 332–351.

Hoogerheide, L. F., J. F. Kaashoek, and H. K. van Dijk (2007), "On the Shape of Posterior Densities and Credible Sets in Instrumental Variable Regression Models with Reduced Rank: an Application of Flexible Sampling Methods using Neural Networks." *Journal of Econometrics*, 139, 154–180.

Hoogerheide, L. F., A. Opschoor, and H. K. van Dijk (2012), "A Class of Adaptive Importance Sampling Weighted EM Algorithms for Efficient and Robust Posterior and Predictive Simulation." *Journal of Econometrics*, 171, 101–120.

Hoogerheide, L. F. and H. K. van Dijk (2010), "Bayesian Forecasting of Value at Risk and Expected Shortfall using Adaptive Importance Sampling." *International Journal of Forecasting*, 26, 231–247.

International Union for Conservation of Nature (2018), "The IUCN Red List of Threatened Species: lapwing." http://www.iucnredlist.org/details/22693949/0. Accessed: 2018-08-05.

Jacob, P. E. and A. H. Thiery (2015), "On Nonnegative Unbiased Estimators." *The Annals of Statistics*, 43, 769–784.

Jacobs, R. A., M. I. Jordan, S. J. Nowlan, and G. E. Hinton (1991), "Adaptive Mixtures of Local Experts." *Journal of Neural Computation*, 3, 79–87.

Jegadeesh, N. and S. Titman (1993), "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency." *The Journal of Finance*, 48, 65–91.

Jegadeesh, N. and S. Titman (2001), "Profitability of Momentum Strategies: An Evaluation of Alternative Explanations." *The Journal of Finance*, 56, 699–720.

Jelsma, H. and K. Lasak (2016), "Forecasting Volatility Using Long Memory Dynamics: How Effective Is the Use of a Realised Measure?" mimeo.

Johnstone, D. J. (2012), "Log-optimal Economic Evaluation of Probability Forecasts." *Journal of the Royal Statistical Society: Series A*, 175, 661–689.

Jordan, M. I. and R. A. Jacobs (1994), "Hierarchical Mixtures of Experts and the EM Algorithm." *Journal Neural Computation*, 6, 181–214.

Jordan, M. I. and L. Xu (1995), "Convergence Results for the EM Approach to Mixtures of Experts Architectures." *Neural Networks*, 8, 1409–1431.

Jungbacker, B. and S. J. Koopman (2007), "Monte Carlo Estimation for Nonlinear Non-Gaussian State Space Models." *Biometrika*, 94, 827–839.

Kahn, H. and A. W. Marshal (1953), "Methods of Reducing Sample Size in Monte Carlo Computations." *Journal of the Operational Research Society of America*, 46, 263–271.

Kahn, H. and A. Marshall (1953), "Methods of Reducing Sample Size in Monte Carlo Computations." *Journal of the Operations Research Society of America*, 1, 263–278.

Kass, R. E., B. P. Carlin, A. Gelman, and R. M. Neal (1998), "Markov Chain Monte Carlo in Practice: A Roundtable Discussion." *The American Statistician*, 52, 93–100.

Kastner, G., S. Frühwirth-Schnatter, and H. F. Lopes (2017), "Efficient Bayesian Inference for Multivariate Factor Stochastic Volatility Models." *Journal of Computational and Graphical Statistics*, 26, 905–917.

Kaufmann, S. and C. Schumacher (2017), "Identifying Relevant and Irrelevant Variables in Sparse Factor Models." *Journal of Applied Econometrics*, 32, 1123–1144.

Kim, S., N. Shephard, and S. Chib (1998), "Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models." *The Review of Economic Studies*, 65, 361–393.

King, R. (2011), "Statistical Ecology." In *Handbook of Markov Chain Monte Carlo* (S. Brooks, J. Galin A. Gelman, and X. L. Meng, eds.), chapter 17, 419–447, Chapman and Hall/CRC.

King, R., S. P. Brooks, C. Mazzetta, S. N. Freeman, and B. J. T. Morgan (2008), "Identifying and Diagnosing Population Declines: a Bayesian Assessment of Lapwings in the UK." *Journal of the Royal Statistical Society: Series C*, 57, 609–632.

King, R., B. T. McClintock, D. Kidney, and D. Borchers (2016), "Capture–recapture Abundance Estimation using a Semi-complete Data Likelihood Approach." *The Annals of Applied Statistics*, 10, 264–285.

King, R., B. Morgan, O. Gimenez, and S. Brooks (2010), *Bayesian Analysis for Population Ecology*. Chapman and Hall/CRC.

Kitagawa, G. (1987), "Non-Gaussian State-Space Modeling of Nonstationary Time Series." *Journal of the American Statistical Association*, 82, 1032–1041.

Kleijn, B. J. K. and A. W. van der Vaart (2006), "Misspecification in Infinite-Dimensional Bayesian Statistics." *The Annals of Statistics*, 34, 837–877.

Kloek, T. and H. K. van Dijk (1978), "Bayesian Estimates of Equation System Parameters: an Application of Integration by Monte Carlo." *Econometrica*, 46, 1–20.

Koopman, S. J., A. Lucas, and M. Scharth (2015), "Numerically Accelerated Importance Sampling for Nonlinear Non-Gaussian State Space Models." *Journal of Business and Economic Statistics*, 33, 114–127.

Koopman, S. J. and E. Hol Uspensky (2002), "The Stochastic Volatility in Mean Model: Empirical Evidence from International Stock Markets." *Journal of Applied Econometrics*, 17, 667–689.

Korattikara, A., Y. Chen, and M. Welling (2014), "Austerity in MCMC Land: Cutting the Metropolis-Hastings Budget." In *Proceedings of the 31st International Conference on International Conference on Machine Learning-Volume 32*, 181–189.

Kullback, S. and R. A. Leibler (1951), "On Information and Sufficiency." *The Annals of Mathematical Statistics*, 1, 79–86.

Kunsch, H. R. (2005), "Recursive Monte Carlo Filters: Algorithms and Theoretical Analysis." *Annals of Statistics*, 33, 1983–2021.

Langrock, R. and R. King (2013), "Maximum Likelihood Estimation of Mark–Recapture–Recovery Models in the Presence of Continuous Covariates." *The Annals of Applied Statistics*, 7, 1709–1732.

Langrock, R., R. King, J. Matthiopoulos, L. Thomas, D. Fortin, and J. M. Morales (2012a), "Flexible and Practical Modeling of Animal Telemetry Data: Hidden Markov Models and Extensions." *Ecology*, 93, 2336–2342.

Langrock, R., I. L. MacDonald, and W. Zucchini (2012b), "Some Nonstandard Stochastic Volatility Models and their Estimation using Structured Hidden Markov Models." *Journal of Empirical Finance*, 147–161.

Liesenfeld, R. and J. F. Richard (2003), "Univariate and Multivariate Stochastic Volatility Models: Estimation and Diagnostics." *Journal of Empirical Finance*, 10, 505–531.

Lindsten, F., M. Jordan, and T. B. Schön (2014), "Particle Gibbs with Ancestor Sampling." *Journal of Machine Learning Research*, 15, 2145–2184.

Liu, J. S. (2001), *Monte Carlo Strategies in Scientific Computing*. Springer Verlag, New York, USA.

Lopes, H. F. and M. West (2004), "Bayesian Model Assessment in Factor Analysis." *Statistica Sinica*, 14, 41–68.

Marin, J. M., P. Pudlo, C. P. Robert, and R. J. Ryder (2012), "Approximate Bayesian Computational Methods." *Statistics and Computing*, 22, 1167–1180.

Marin, J. M. and C. Robert (2007), *Bayesian Core: a Practical Approach to Computational Bayesian Statistics*. Springer Science & Business Media.

Marshall, A. W. (1956), "The Use of Multi-Stage Sampling Schemes in Monte Carlo Computations." In *Symposium on Monte Carlo Methods* (M. Meyer, ed.), 123–140, Wiley.

McLachlan, G. and D. Peel (2004), *Finite Mixture Models*. John Wiley & Sons.

McNeil, A. J. and R. Frey (2000), "Estimation of Tail-Related Risk Measures for Heteroscedastic Financial Time Series: an Extreme Value Approach." *Journal of Empirical Finance*, 7, 271–300.

McNeil, A. J., R. Frey, and P. Embrechts (2015), *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press.

McNeil, A. J. and J. Wendin (2007), "Bayesian Inference for Generalized Linear Mixed Models of Portfolio Credit Risk." *Journal of Empirical Finance*, 14, 131–149.

Meyer, R. and J. Yu (2000), "BUGS for a Bayesian Analysis of Stochastic Volatility Models." *Econometrics Journal*, 3, 198–215.

Michaud, R. O. (1989), "The Markowitz Optimization Enigma: Is 'Optimized' Optimal?" *Financial Analysts Journal*, 45, 31–42.

Moskowitz, T. J. and M. Grinblatt (1999), "Do Industries Explain Momentum?" *The Journal of Finance*, 54, 1249–1290.

Müller, U. K. (2013), "Misspecification in Infinite-Dimensional Bayesian Statistics." *Econometrica*, 81, 1805–1849.

Murphy, K. P. (2002), *Dynamic Bayesian Networks: Representation, Inference and Learning*. Ph.D. thesis, University of California, Berkeley.

Ng, S., R. F. Engle, and M. Rothschild (1992), "A Multi-Dynamic-Factor Model for Stock Returns." *Journal of Econometrics*, 52, 245–266.

Omori, Y., S. Chib, N. Shephard, and J. Nakajima (2007), "Stochastic Volatility with Leverage: Fast and Efficient Likelihood Inference." *Journal of Econometrics*, 140, 425–449.

Opschoor, A., D. van Dijk, and M. van der Wel (2016), "Combining Density Forecasts using Focused Scoring Rules." *Tinbergen Institute Discussion Paper*, 14-090/III.

Peel, D. and G. McLachlan (2000), "Robust Mixture Modeling using the t-Distribution." *Statistics and Computing*, 10, 339–348.

Peng, F., R. A. Jacobs, and M. A. Tanner (1996), "Bayesian Inference in Mixtures-of-Experts and Hierarchical Mixtures-of-Experts Models with an Application to Speech Recognition." *Journal of the American Statistical Association*, 91, 953–960.

Pettenuzzo, D. and F. Ravazzolo (2016), "Optimal Portfolio Choice under Decision-Based Model Combinations." *Journal of Applied Econometrics*, 31, 1312–1332.

Pitt, M. K. and N. Shephard (1999), "Filtering via Simulation: Auxiliary Particle Filter." *Journal of the American Statistical Association*, 94, 590–599.

Pitt, M. K., R. S. Silva, P. Giordani, and R. Kohn (2012), "On Some Properties of Markov Chain Monte Carlo Simulation Methods Based on the Particle Filter." *Journal of Econometrics*, 171, 134–151.

Quintana, J. M., V. K. Chopra, and B. H. Putnam (1995), "Global Asset Allocation: Stretching Returns by Shrinking Forecasts." In *Proceedings of the ASA Section on Bayesian Statistical Science*, 199–205.

Richard, J. F. and W. Zhang (2007), "Efficient High-dimensional Importance Sampling." *Journal of Econometrics*, 141, 1385–1411.

Robert, C. (2007), *The Bayesian Choice: from Decision-Theoretic Foundations to Computational Implementation: Second Edition*. Springer Texts in Statistics.

Robert, C. (2017), "Why should I be Bayesian when my model is wrong?" `https://xianblog.wordpress.com/2017/05/09/why-should-i-be-bayesian-when-my-model-is-wrong/`. Accessed: 18 July 2017.

Robert, C. P. (2016), "Exact, unbiased, what else?!" https://xianblog.wordpress.com/2016/04/13/exact-unbiased-what-else/. Accessed: 2018-10-17.

Robert, C. P. and G. Casella (2004), *Monte Carlo Statistical Methods: Second Edition*. Springer Texts in Statistics.

Robert, C. P. and G. Casella (2011), "A Short History of Markov Chain Monte Carlo: Subjective Recollections from Incomplete Data." *Statistical Science*, 26, 102–115.

Roberts, G. O. and J. S. Rosenthal (2001), "Optimal Scaling for Various Metropolis-Hastings Algorithms." *Statistical Science*, 16, 351–367.

Rosenthal, J. S. (2011), "Optimal Proposal Distributions and Adaptive MCMC." In *Handbook of Markov Chain Monte Carlo* (S. Brooks, J. Galin A. Gelman, and X. L. Meng, eds.), chapter 4, 93–111, Chapman and Hall/CRC.

Roth, M. (2013), "On the Multivariate t Distribution." Technical Report LiTH-ISY-R-3059, Automatic Control Group at Linköpings Universitet.

Sandmann, G. and S. J. Koopman (1998), "Estimation of Stochastic Volatility Models via Monte Carlo Maximum Likelihood." *Journal of Econometrics*, 87, 271–301.

Shephard, N. (1996), "Statistical Aspects of ARCH and Stochastic Volatility." In *Time Series Models in Econometrics, Finance and Other Fields* (D. R. Cox, D. V. Hinkley, and O. E. Barndorff-Neilsen, eds.), 1–67, Chapman & Hall.

Stock, J. H. and M. W. Watson (2002), "Forecasting using Principal Components from a Large Number of Predictors." *Journal of the American Statistical Association*, 97, 1167–1179.

Stock, J. H. and W. M. Watson (2005), "Implications of Dynamic Factor Models for VAR Analysis." Technical report, NBER Working Paper No. 11467.

Svensén, M. and C. M. Bishop (2005), "Robust Bayesian Mixture Modeling." *Neurocomputing*, 64, 339–348.

Talih, M. and N. Hengartner (2005), "Structural Learning with Time-varying Components: Tracking the Cross-section of Financial Time Series." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 321–341.

Tanner, M. A. and W. H. Wong (1987), "The Calculation of Posterior Distributions by Data Augmentation." *Journal of the American Statistical Association*, 82, 528–540.

Taylor, S. J. (1994), "Modeling Stochastic Volatility: A Review and Comparative Study." *Mathematical Finance*, 4, 183–204.

The Royal Society for the Protection of Birds (2018), "The Red List of Conservation Concern: lapwing." https://www.rspb.org.uk/birds-and-wildlife/wildlife-guides/bird-a-z/lapwing/. Accessed: 2018-08-05.

The Volatility Laboratory (2012), "V-Lab: Long Run Value at Risk Documentation." `https://vlab.stern.nyu.edu/doc/4?topic=apps`. Accessed: 30 November 2016.

Waggoner, D. F. and T. Zha (2012), "Confronting Model Misspecification in Macroeconomics." *Journal of Econometrics*, 171, 167–184.

West, M. and J. Harrison (1997), *Bayesian Forecasting and Dynamic Models: Second Edition*. Springer-Verlag.

Winkler, R. L. and C. B. Barry (1975), "A Bayesian Model for Portfolio Selection and Revision." *The Journal of Finance*, 30, 179–192.

Yu, J. (2005), "On Leverage in a Stochastic Volatility Models." *Journal of Econometrics*, 127, 165–178.

Zeevi, A. J. and R. Meir (1997), "Density Estimation through Convex Combinations of Densities; Approximation and Estimation Bounds." *Neural Networks*, 10, 99–106.

Zellner, A. (1996), *An Introduction to Bayesian Inference in Econometrics*. Wiley.

Zhou, X., J. Nakajima, and M. West (2014), "Bayesian Forecasting and Portfolio Decisions Using Dynamic Dependent Sparse Factor Models." *International Journal of Forecasting*, 30, 963–980.

Zucchini, W., I. L. MacDonald, and R. Langrock (2016), *Hidden Markov Models for Time Series: An Introduction Using R, Second Edition*. Monographs on Statistics and Applied Probability 150, CRC Press.

The Tinbergen Institute is the Institute for Economic Research, which was founded in 1987 by the Faculties of Economics and Econometrics of the Erasmus University Rotterdam, University of Amsterdam and VU University Amsterdam. The Institute is named after the late Professor Jan Tinbergen, Dutch Nobel Prize laureate in economics in 1969. The Tinbergen Institute is located in Amsterdam and Rotterdam. The following books recently appeared in the Tinbergen Institute Research Series:

692 G. DE OLIVEIRA, *Coercion and Integration*
693 S. CHAN, *Wake Me up before you CoCo: Implications of Contingent Convertible Capital for Financial Regulation*
694 P. GAL, *Essays on the role of frictions for firms, sectors and the macroeconomy*
695 Z. FAN, *Essays on International Portfolio Choice and Asset Pricing under Financial Contagion*
696 H. ZHANG, *Dealing with Health and Health Care System Challenges in China: Assessing Health Determinants and Health Care Reforms*
697 M. VAN LENT, *Essays on Intrinsic Motivation of Students and Workers*
698 R.W. POLDERMANS, *Accuracy of Method of Moments Based Inference*
699 J.E. LUSTENHOUWER, *Monetary and Fiscal Policy under Bounded Rationality and Heterogeneous Expectations*
700 W. HUANG, *Trading and Clearing in Modern Times*
701 N. DE GROOT, *Evaluating Labor Market Policy in the Netherlands*
702 R.E.F. VAN MAURIK, *The Economics of Pension Reforms*
703 I. AYDOGAN, *Decisions from Experience and from Description: Beliefs and Probability Weighting*
704 T.B. CHILD, *Political Economy of Development, Conflict, and Business Networks*
705 O. HERLEM, *Three Stories on Influence*
706 J.D. ZHENG, *Social Identity and Social Preferences: An Empirical Exploration*
707 B.A. LOERAKKER, *On the Role of Bonding, Emotional Leadership, and Partner Choice in Games of Cooperation and Conflict*
708 L. ZIEGLER, *Social Networks, Marital Sorting and Job Matching. Three Essays in Labor Economics*
709 M.O. HOYER, *Social Preferences and Emotions in Repeated Interactions*
710 N. GHEBRIHIWET, *Multinational Firms, Technology Transfer, and FDI Policy*
711 H.FANG, *Multivariate Density Forecast Evaluation and Nonparametric Granger Causality Testing*
712 Y. KANTOR, *Urban Form and the Labor Market*

713 R.M. TEULINGS, *Untangling Gravity*
714 K.J.VAN WILGENBURG, *Beliefs, Preferences and Health Insurance Behavior*
715 L. SWART, *Less Now or More Later? Essays on the Measurement of Time Preferences in Economic Experiments*
716 D. NIBBERING, *The Gains from Dimensionality*
717 V. HOORNWEG, *A Tradeoff in Econometrics*
718 S. KUCINSKAS, *Essays in Financial Economics*
719 O. FURTUNA, *Fiscal Austerity and Risk Sharing in Advanced Economies*
720 E. JAKUCIONYTE, *The Macroeconomic Consequences of Carry Trade Gone Wrong and Borrower Protection*
721 M. LI, *Essays on Time Series Models with Unobserved Components and Their Applications*
722 N. CIURILĂ, *Risk Sharing Properties and Labor Supply Disincentives of Pay-As-You-Go Pension Systems*
723 N.M. BOSCH, *Empirical Studies on Tax Incentives and Labour Market Behaviour*
724 S.D. JAGAU, *Listen to the Sirens: Understanding Psychological Mechanisms with Theory and Experimental Tests*
725 S. ALBRECHT, *Empirical Studies in Labour and Migration Economics*
726 Y.ZHU, *On the Effects of CEO Compensation*
727 S. XIA, *Essays on Markets for CEOs and Financial Analysts*
728 I. SAKALAUSKAITE, *Essays on Malpractice in Finance*
729 M.M. GARDBERG, *Financial Integration and Global Imbalances*
730 U. THŰMMEL, *Of Machines and Men: Optimal Redistributive Policies under Technological Change*
731 B.J.L. KEIJSERS, *Essays in Applied Time Series Analysis*
732 G. CIMINELLI, *Essays on Macroeconomic Policies after the Crisis*
733 Z.M. LI, *Econometric Analysis of High-frequency Market Microstructure*
734 C.M. OOSTERVEEN, *Education Design Matters*
735 S.C. BARENDSE, *In and Outside the Tails: Making and Evaluating Forecasts*
736 S. SÓVÁGÓ, *Where to Go Next? Essays on the Economics of School Choice*
737 M. HENNEQUIN, *Expectations and Bubbles in Asset Market Experiments*
738 M.W. ADLER, *The Economics of Roads: Congestion, Public Transit and Accident Management*
739 R.J. DÖTTLING, *Essays in Financial Economics*
740 E.S. ZWIERS, *About Family and Fate: Childhood Circumstances and Human Capital Formation*
741 Y.M. KUTLUAY, *The Value of (Avoiding) Malaria*