

Bayesian Emulation for Sequential Modeling, Inference and Decision Analysis

by

Kaoru Irie

Department of Statistical Science
Duke University

Date: _____

Approved:

Mike West, Supervisor

David L. Banks

Surya T. Tokdar

Andrew Cron

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2016

ABSTRACT

Bayesian Emulation for Sequential Modeling, Inference and
Decision Analysis

by

Kaoru Irie

Department of Statistical Science
Duke University

Date: _____

Approved:

Mike West, Supervisor

David L. Banks

Surya T. Tokdar

Andrew Cron

An abstract of a dissertation submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University

2016

Copyright © 2016 by Kaoru Irie
All rights reserved

Abstract

Advances in three related areas of state-space modeling, sequential Bayesian learning, and decision analysis are addressed, with core statistical challenges of scalability and associated dynamic sparsity. Research presented in these three areas emphasizes the common theme of Bayesian model emulation: solving challenging analysis/computational problems using creative model emulators. This idea defines theoretical and applied advances in non-linear, non-Gaussian state-space modeling, dynamic sparsity, decision analysis and statistical computation, across linked contexts of multivariate time series and dynamic networks studies. Examples and applications in financial time series and portfolio analysis, macroeconomics and internet studies from computational advertising demonstrate the utility of the core methodological innovations.

Chapter 1 summarizes the three areas/problems and the key idea of Bayesian emulation in those areas. Chapter 2 discusses the sequential analysis of latent threshold models with use of emulating models that allow for analytical filtering to enhance the efficiency of posterior sampling. Chapter 3 examines a novel emulation approach in decision analysis, using synthetic statistical models to define computational approaches to solving optimization problems, and evaluating its performance in the context of sequential portfolio optimization. Chapter 4 develops sets of efficient sub-models for counts/flows of streaming data on networks, and exploits them in emulation of more structured inter-dependent network flow models, with studies of

internet data in e-commerce. Chapter 5 reviews and summarizes these research advances, and adds pointers to potential future directions.

To my family

Contents

Abstract	iv
List of Tables	xi
List of Figures	xii
List of Abbreviations and Symbols	xvii
Acknowledgements	xix
1 Introduction	1
2 Sequential Analysis and Bayesian Model Emulation for Dynamic Latent Threshold Models	4
2.1 Introduction	4
2.2 Model	5
2.2.1 Latent Threshold Models	5
2.2.2 Challenges to Sequential Analysis	6
2.3 Sequential Importance Sampling and Model Emulation	7
2.3.1 SIS and Emulating Models	7
2.3.2 Choices of LTM Emulators	8
2.3.3 Comparison of Emulators	11
2.4 Sequential Learning with Fixed Parameters and Volatilities	13
2.4.1 Particle Learning for AR(1) Parameters	13
2.4.2 Auxiliary Particle Filter for Latent Threshold	14

2.4.3	Volatility Models	14
2.4.4	Sequential Posterior Update	15
2.5	Application: US Macroeconomic Study	18
2.5.1	Latent Thresholded TV-VAR Models	19
2.5.2	Posterior Analysis	22
2.5.3	Comparison	28
2.6	Summary Comments	31
2.7	Appendix: Technical Details of SMC for LTMs	34
2.7.1	Forward Filtering with Emulators	34
2.7.2	Particle Learning	35
2.7.3	Auxiliary Particle Filter	36
2.7.4	Extension to Stochastic Volatility	37
2.8	Appendix: Supplemental Figures on Parameter Learning	38
3	Bayesian Emulation in Decision Analysis: Sequential, Multi-Step Portfolio Optimization	42
3.1	Emulation by Synthetic Models in General Decision Analysis	42
3.2	Introduction: Sequential Portfolio Optimization	44
3.3	Statistical Model Emulation of Expected Loss Functions	48
3.3.1	Settings and Notation	48
3.3.2	Dynamic Linear Models for Quadratic Loss	49
3.3.3	FFBS for Posterior Modes	51
3.4	Laplace State Space Models and Implied Loss Function	52
3.4.1	Synthetic Models and EM Algorithm	53
3.4.2	Another Laplace Factor for Non-Negativity Constraint	56
3.5	Marginalization of Loss Functions	58
3.5.1	Joint and Marginal Loss Functions	58

3.5.2	Marginal Laplace Loss Function and Mode Searching	60
3.6	Application: FX Commodity Dataset	62
3.6.1	Dataset	62
3.6.2	Models for Prediction	62
3.6.3	Evaluation by Cumulative Returns	63
3.6.4	Results	64
3.7	Summary Comments	72
3.8	Appendix: Forward Filtering with Multiple Observations	74
3.8.1	Forward Filtering	74
3.8.2	Two Observational Equations with the Common State	74
3.8.3	Computational Efficiency	77
3.9	Appendix: Exact Shrinkage in EM Algorithm	78
3.10	Appendix: Posterior Marginal Mode of Laplace Models	79
3.10.1	Gibbs Sampler and Maximization for Marginal Models	79
3.10.2	Sum-to-One Constraint	80
3.10.3	Realization of Exact Zero Weights and Transitions	82
3.11	Appendix: Dynamic Dependence Network Models	83
3.12	Appendix: Supplemental Analysis	86
3.12.1	Choice of Tuning Parameters	86
3.12.2	Additional Figures for Profiled and Marginal Portfolios	89
4	Emulation of Dynamic Gravity Model by Bayesian Dynamic Flow Models	91
4.1	Introduction	91
4.2	Bayesian Dynamic Flow Models	93
4.2.1	Forward Filtering and Backward Sampling	94
4.2.2	Marginal Likelihoods and Optimal Discount Factor	97

4.2.3	Generalized Gamma Random Walk and Power Discount	99
4.2.4	Extension to Multinomial-Dirichlet State-Space Models	99
4.3	Dynamic Gravity Models	102
4.3.1	Identification	104
4.3.2	Identification by a Reference Flow	105
4.3.3	Identification by Geometric Means	106
4.3.4	Computational Problems in Practice	107
4.3.5	Alternative Identification Strategy under Sparsity	109
4.4	Dependent Gravity Model and MCMC	110
4.5	Application: Analysis of FoxNews Website Access Records	112
4.5.1	Study of Internet Traffic Flow	112
4.5.2	Context and Data	113
4.5.3	BDFM Analysis of FoxNews Data	116
4.5.4	DGM Analysis of FoxNews Data	121
4.5.5	Comparison Across Days	126
4.6	Summary Comments	128
4.7	Appendix: FFBS for Poisson-Gamma Models	132
4.7.1	Forward Filtering	132
4.7.2	Backward Sampling	134
5	Concluding Remarks	135
	Bibliography	142
	Biography	148

List of Tables

2.1	Log-Marginal Likelihoods of emulators.	28
2.2	Multistep RMSFEs for 3 variables with MC errors.	32
3.1	List of currencies, commodities and indeces.	62
3.2	Parental sets used in prediction.	85

List of Figures

2.1	US macroeconomic indices. From top to bottom, the inflation, unemployment rate and short-term nominal interest rate in US from 1963 to 2011. This is the same dataset as used in Nakajima and West (2013a).	18
2.2	Posterior estimate of B_{1t} and the inclusion probability. Each element of the 9 plots corresponds with the result of (i, j) -element of B_{1t} . In the upper 3×3 frames, the solid line and dotted lines represent on-line updates of the posterior median and 90% intervals of (i, j) -element of B_{1t} for $t = 1:196$, respectively. The lower 3×3 frames show the corresponding values of each the inclusion probability defined by $Pr[\beta_{jt} > d_j]$ for each of B_{1t}	24
2.3	Posterior estimate of thresholds d for B_{1t} . Shown are the posterior medians and 90% intervals of thresholds d for each (i, j) -element of B_{1t}	25
2.4	Posterior trajectories for $(a_{21t}, a_{31t}, a_{32t})$ and the corresponding inclusion probabilities.	26
2.5	Stochastic volatilities (h_{1t}, h_{2t}, h_{3t}) . The left column shows the results on the log-scale, and the right column shows them on the original scale (standard deviation).	27
2.6	Posterior trajectories for intercepts (c_{1t}, c_{2t}, c_{3t}) and the corresponding inclusion probabilities.	27
2.7	1-step/4-step ahead predictive distributions with observation. The summaries of time trajectories of sequentially updated predictive distributions for the three series are shown for 1-step (top) and 4-step (bottom) ahead forecasting.	29
2.8	Comparison of log marginal likelihoods and posterior model probabilities. Log marginal likelihoods (Top) and posterior model probabilities (Bottom) for the three emulators and vanilla SMC. Red: DLM, Blue: Shrinkage, Green: LLTM, Pink: Vanilla SMC.	30

2.9	ESS results in analyses of DLM, Shrinkage, LLTM and vanilla SMC emulation methods. The ESS is scaled percentage, i.e., $100 \times ESS/N$, where N is the number of particles. The 3 columns represent results on each of the 3 univariate time series LTMs. For each, the upper 4 rows show time trajectories of the ESS measures, while the lower 4 rows show the resulting histograms of ESS measures aggregated over the time period.	33
2.10	Time trajectories of posteriors for elements of B_{2t} and the corresponding inclusion probabilities.	38
2.11	Time trajectories of posteriors for elements of B_{3t} and the corresponding inclusion probabilities.	39
2.12	On-line posteriors of Φ for B_{1t}	40
2.13	On-line posteriors of Σ for B_{1t}	40
2.14	On-line posteriors of α for B_{1t}	41
3.1	The statistical approach to loss minimization. There are two interpretations for this diagram. If the problem of interest has the specific loss function to be minimized, it can be transformed into the synthetic model, shown by the arrow from top to bottom, so that the computational method of statistical inference can solve the original problem. The other interpretation is that, for statisticians, the problem itself is defined as the statistical models, then transformed back into the form of optimization of an expected loss function, indicated by the arrow from bottom to top, in order to make use of the knowledge and techniques of statistical modeling.	44
3.2	Sequence of portfolio optimization problems. At each time t , the posterior and forecast distribution is updated with additional information x_t . Based on the prediction of r_{t+1} (and r_{t+i} for $i \geq 1$), the loss function of w_{t+1} is defined. The output at time t is thus the portfolio for the next time, w_{t+1} . This process is repeated at every time point, yielding the sequence of portfolio vectors, $\{w_t\}_{t=1:T}$	45
3.3	Optimal portfolios of DLM loss function and Markowitz method. From left to right, top to bottom: portfolio weights of a quadratic loss function with $\lambda_t = 1, 100$ and 10000 with $(\alpha_t, \beta_t) = (100, 1)$ and Markowitz-type portfolio.	65
3.4	Cumulative returns of DLM and Markowitz portfolios. Top: cumulative returns with no transaction cost. Bottom: with 0.1% transaction cost. Four cumulative returns of DLM portfolios with $\lambda_t = 1$ (red), 100 (blue), 10000 (green) and Markowitz portfolio (grey) are shown in both pictures.	66

3.5	Standard deviations of DLM and Markowitz portfolios. Four standard deviations of DLM portfolios of $\lambda_t = 1$ (red), 100 (blue), 10000 (green) and Markowitz portfolio (grey) are shown in both pictures with that of minimum risk portfolio (black).	66
3.6	Optimal portfolio of the Laplace model. From top to bottom, left to right: portfolio of Laplace loss function with $(\lambda_t, \gamma_t) = (100, 100)$, $(100, 1)$, $(1000, 100)$, $(1000, 1)$ with $(\alpha_t, \beta_t) = (100, 1)$	68
3.7	A Laplace portfolio with the number of active weights. Top: portfolio of Laplace loss function with $(\alpha_t, \beta_t, \lambda_t, \gamma_t) = (1, 100, 100, 100)$. Bottom: number of non-zero weights in the portfolio.	68
3.8	Cumulative returns of Laplace portfolios for $\lambda_t = 100$. While the weights for transaction costs ($ w_t - w_{t-1} $) are fixed as $\lambda_t = 100$ in addition to $(\alpha_t, \beta_t) = (100, 1)$, those for sparsity in portfolio ($ w_t $) are $\gamma_t = 10, 100$ or 1000 (red, blue and green, respectively). For comparison, the cumulative returns of Gaussian portfolio with $\gamma_t = 100$ (dotted black) and Markowitz (grey) are shown.	69
3.9	Cumulative returns of Laplace portfolios for $\gamma_t = 100$. Conversely to Figure 3.8, we fix $\gamma_t = 100$ and change the value of λ_t to 10, 100 and 1000.	70
3.10	Portfolios of profiled and marginal Laplace models. Top: portfolio of profiled Laplace loss function with $\lambda_t = 100$. Bottom: portfolio of marginal Laplace loss function with $\lambda_t = 100$	71
3.11	Cumulative returns of profiled and marginal Laplace portfolios. No transaction cost is used ($\delta = 0$). Four portfolios are distinguished by red lines for profiled (joint) loss functions, blue lines for marginal ones, solid lines for $\lambda_t = 100$ and dotted lines for $\lambda_t = 10$	72
3.12	Convergence diagnoses for $w_{1:3,t+1}$ at $t = 1$. From the top row to the bottom one, the correlograms, sample paths and estimated marginal densities are shown.	81
3.13	Convergence diagnoses for $\tau_{1:3,t+1}$ at $t = 1$. The sample paths shown in the second row are re-scaled to be $10^{-4}\tau_{it}$	81
3.14	Cumulative returns for $\lambda_t = 10000$. Top: cumulative returns with no transaction cost. Bottom: with 0.1% transaction cost. Four cumulative returns of DLM portfolios of $\lambda_t = 1$ (red), 100 (blue), 10000 (green) and Markowitz portfolio (grey) are shown in both pictures.	87

3.15	Standard deviation for $\lambda_t = 10000$. The black line shows the standard deviation of the minimum risk portfolio.	87
3.16	Cumulative returns for $(\alpha_t, \beta_t) = (0.01, 1)$. Top: cumulative returns with no transaction cost. Bottom: with 0.1% transaction cost. Four cumulative returns of DLM portfolios of $\alpha_t = 0.0001$ (red), 0.01 (blue), 0.1 (green), 1 (pink) and Markowitz portfolio (grey) are shown in both pictures.	88
3.17	Standard deviation for $(\alpha_t, \beta_t) = (0.01, 1)$	88
3.18	Profiled and marginal Portfolios for $\lambda_t = 10$. Top: profiled, Bottom: marginal.	89
3.19	Profiled and marginal portfolios for $\lambda_t = 1000$. Top: profiled, Bottom: marginal.	90
4.1	Network schematic and notation for flows at time t . The circular numbered nodes represent the domains at the current time point t . Each arrow $i \rightarrow j$ shows the flow of y_{ijt} visitors.	115
4.2	The standardized values of the marginal log-posteriors of discount factors δ_i for the in-flows to FoxNews nodes $i = 1:22$ (top right, reading across rows), for the period 09:05am-09:55am on February 23, 2015.	117
4.3	BDFM-based inference over time t on in-flows to domain $i = 10$ (Leisure). <i>Upper</i> : data y_{0Xt} (circles) with one-step ahead forecast means and 95% intervals. <i>Center</i> : trajectory of mean and 95% intervals from on-line posteriors $p(\phi_{0Xt} y_{0X,1:t})$ plotted against t . <i>Lower</i> : revised trajectory under full retrospective posterior $p(\phi_{0Xt} y_{0X,1:T})$	119
4.4	BDFM-based inference over time t on transitions from domain $i = 1$ (Homepage) to $j = 2$ (Politics). <i>Upper</i> : data y_{12t} (plus signs) with one-step ahead forecast means and 95% intervals. <i>Center</i> : trajectory of mean and 95% intervals from on-line posteriors of ϕ_{12t} . <i>Lower</i> : revised trajectory under full retrospective posterior.	119
4.5	Retrospective mean and 95% CI of trajectories of transition probability θ_{11t} (staying at Homepage) from analysis on data collected from each of the six mornings.	120
4.6	Retrospective mean and 95% CI of trajectories of transition probability θ_{10t} (Homepage \rightarrow External) from analysis on each of the six mornings.	120
4.7	Retrospective mean and 95% CI of trajectories of transition probabilities θ_{15t} (Homepage \rightarrow Entertainment) for each of the six mornings.	121

4.8	DGM-based smoothed trajectory of baseline level process $\mu_{1:T}$. In this and following figures, the dashed lines indicate 95% intervals about the displayed posterior mean trajectory.	122
4.9	DGM-based smoothed trajectories of node-specific outflows $\alpha_{i,1:T}$	123
4.10	DGM-based smoothed trajectories of node-specific inflows $\beta_{j,1:T}$	123
4.11	<i>Upper</i> : DGM-based smoothed trajectories of transition affinities, Home-page \rightarrow Opinion. <i>Lower</i> : Bayesian credible values corresponding to the affinity trajectories.	125
4.12	<i>Upper</i> : DGM-based smoothed trajectories of transition affinities, Home-page \rightarrow Science. <i>Lower</i> : Corresponding Bayesian credible values.	125
4.13	DGM-based inference on baseline flow level trajectories for all six days, with 95% credible intervals. The red trajectories correspond to the 09:05-09:55am time window, and the blue trajectories correspond to the 01:05-01:55 p.m. time window.	127
4.14	Posterior trajectories of $\mu_{1:T}$ with the common restriction.	128
4.15	DGM-based inference on $\gamma_{0,5,1:T}$ for all six days. The 12 retrospective posterior trajectories of the affinity effect, $\gamma_{0,5,t}$, which corresponds to the flow from External to Entertainment.	129

List of Abbreviations and Symbols

Symbols

$N(x \mu, \Sigma)$	Normal distribution with mean μ and variance matrix Σ , or its density evaluated at x .
$Ga(x r, c)$	Gamma distribution with shape r and rate c (mean r/c) with normalizing constant $\Gamma(r)$ if $c = 1$.
$Be(x a, b)$	Beta distribution with mean $a/(a+b)$ with norm. const. $B(a, b)$.
$Po(x \mu)$	Poisson distribution with mean (rate) μ .
$Multi(x n, \theta)$	Multinomial distribution with probability vector θ and population n .
$Dir(x q)$	Dirichlet distribution with concentration parameters q .
\mathcal{D}_t	Information available at time t .
$s : t$	Indices $s, s + 1, \dots, t - 1, t$.

Abbreviations

APF	Auxiliary Particle Filter.
AR(p)	Autoregressive process with order p .
BDFM	Bayesian Dynamic Flow Model.
DDN	Dynamic Dependent Network (models).
DGM	Dynamic Gravity Model.
DLM	Dynamic Linear Model.
GIG	Generalized Inverse Gaussian (distribution).
LTM	Latent Threshold Model.

MCMC	Markov Chain Monte Carlo.
PL	Particle Learning.
SMC	Sequential Monte Carlo.
SVD	Singular Value Decomposition (of a matrix).
TV-VAR(p)	Time-Varying Vector Autoregressive Model with order p .

Acknowledgements

First, I would like to sincerely thank my advisor Mike West. His guidance, help and encouragement have been, and will be, undoubtedly essential to my study and career development. The discussion with him about research, that was always full of novel ideas and insights, was definitely the integral part of my academic life at Duke, which led to the realization of this thesis.

I thank the faculty and staff at Department of Statistical Science and appreciate their work and effort that create this great research environment. I thank my Ph.D. and Preliminary Exam Committee Members, David Banks, Surya Tokdar, Andrew Cron, Li Ma, Sayan Mukherjee and Galen Reeves for their time and support. My further thanks go to David and Li for their personal advice and help for my research and academic career. I also thank my fellow graduate students for their daily discussion about Bayesian statistics in Old Chem 017 that kept me motivated.

I am deeply grateful to many people and institutes off campus who supported my research. Yasuhiro Omori brought me to Bayesian research and strongly recommended that I study at Duke. The Nakajima Foundation and their generous scholarship financially supported my four years. The BEST Award, funded by the BEST Foundation, provided financial support of my research in summer 2015. Max-Point kindly offered me many research opportunities on applied data analysis.

Finally, I thank my family members, Minoru, Masae and Atsushi, for their love and respect.

Introduction

This thesis concerns advances in three related areas of state-space modeling, sequential Bayesian analysis, and decisions. Innovations in these areas are linked in several ways, including a key theme of Bayesian model emulation: solving challenging analysis/computational problems using creative model emulators.

- Chapter 2, Inference: “*Sequential Analysis and Bayesian Model Emulation for Dynamic Latent Threshold Models*”

This research focuses on filtering and forecasting of time series using latent threshold models—an approach to dynamic sparsity via state-dependent dynamic variable selection. On-line filtering of time-evolving posterior and predictive distributions in these models is computationally challenging. My new idea is to build analysis “emulators”—synthetic models that are more computationally tractable—using ideas from auxiliary particle filtering. Sequential importance sampling is used to adapt analyses of emulators to the original latent threshold model. Several choices of emulating models are examined in studies of synthetic data and a serious applied problem of macroeconomic time

series forecasting, evaluating accuracy of prediction and efficiency in posterior approximation. An interesting finding is that one of the best emulators is, in fact, that based on direct, adaptive “soft shrinkage” of state variables, while the target models are based on underlying “hard shrinkage” of latent states.

This chapter is based on Irie and West (2016b).

- Chapter 3, Decision: *“Bayesian Emulation in Decision Analysis: Sequential, Multi-Step Portfolio Optimization”*

In the context of sequential forecasting and portfolio optimization, I introduce a novel approach to Bayesian analysis based on mapping a specified loss function minimization problem to that of finding the mode of a posterior distribution in a “synthetic” statistical model. Computational methods for exploring distributions can then be applied to solve the original optimization problem. I do this in the context of novel portfolio utility functions that extend traditional Markowitz-type methods to multiple-step ahead investments with explicit penalties for transaction costs. Various forms introduced and explored include sparsity-inducing penalties on portfolio turnover, which yield interesting classes of synthetic statistical models of state-space forms with non-Gaussian structure. The resulting computational problems are addressed using combinations of EM, MCMC and analytic filtering and smoothing. Significant practical benefits in application to financial portfolios are realized in applied studies of FX, commodity and stock index time series, based on sequential forecasting using customized dynamic dependency network models.

This chapter is based on Irie and West (2016a).

- Chapter 4, Modeling: *“Emulation of Dynamic Gravity Model by Bayesian Dynamic Flow Models”*

This research concerns models and analyses of counts/flows of traffic into, out of and between nodes in a time-evolving network. The applied/motivating context and case study involves time series of flows between domains within a web site. New approaches based on dynamic Poisson-Gamma (and Multinomial-Dirichlet) models are invented, using ideas from stochastic volatility modeling. Sequential analysis is analytically tractable in forward filtering and efficient backward sampling, and scalable due to conditional decoupling across nodes that yet maintains the full complexity of dynamic network flows. Again using the idea of emulation, this flexible framework maps to a parallel model of greater complexity— a so-called gravity model of flow dynamics that aims to more incisively infer node-specific and node-pair patterns and their changes. The overall analysis can inherently exploit parallel computation, so is scalable to large networks.

This chapter is based on “*Bayesian Dynamic Modeling and Analysis of Streaming Network Data*” (Chen et al., 2015).

Final discussion in Chapter 5 provides summary review of the thesis and comments on open research areas. At a high-level, this thesis defines theoretical and applied advances in non-linear, non-Gaussian state-space modeling, dynamic sparsity, decision analysis and statistical computation, across linked contexts of multivariate time series and dynamic networks studies. Examples and applications in financial time series and portfolio analysis, macroeconomics and internet studies from computational advertising demonstrate the utility of the core methodological innovations.

All computations are implemented in Ox (Doornik, 2007).

Sequential Analysis and Bayesian Model Emulation for Dynamic Latent Threshold Models

2.1 Introduction

As time series analysis faces challenges of increasing dimensions of time series recordings, the need to induce dynamic sparsity into model structures— in the sense of both global and local shrinkage of parameters— is increasingly important. To meet this need, dynamic latent thresholding (Nakajima and West, 2013a) defines a general approach to adaptive, dynamic sparsity modeling. Recent applications of LTMs (e.g. Nakajima and West, 2013b, 2015; Zhou et al., 2014) demonstrate insightful interpretations and increased predictive accuracy relative to standard models. To date, however, analysis and model fitting have involved intensive MCMC methods that are restricted to batch processing of historical data. On-line analysis for sequentially observed/streaming data is increasing in importance, in areas such as economics (e.g. Carvalho and Lopes, 2007; Lopes and Tsay, 2011) and finance, but also now moving into business/IT areas that are generating time series challenges with increasingly high-dimensional dynamic data streams. In response, our work here concerns *se-*

quential Bayesian learning and forecasting in LTMs.

The inherent nonlinearities of LTMs obviate the use of particle learning (Carvalho et al., 2010) though allow for variants of auxiliary particle filtering (Liu and West, 2001) to address fixed parameter learning in state space models. To address particle degeneracy, we exploit the theoretical model structure by introducing *model emulators* to enhance existing sequential Monte Carlo (SMC) methods using adaptive sequential importance sampling (SIS, e.g., Liu and Chen, 1998).

We review the specification of LTMs in Section 2.2, and then discuss sequential analysis and the approach of model emulation in Section 2.3. Here we discuss several candidate LTM emulators and measures for comparative evaluation of emulation approaches. The full procedure of sequential analysis is outlined Section 2.4, covering emulation-based SMC methods for dynamic states combined with the crucial issue of fixed parameter learning. Technical support material appears in the appendix. Section 2.5 considers a topical macroeconomic application based on LTM extensions of time-varying vector autoregressive (TV-VAR) models. The applied analyses bear out the utility of the Bayesian model emulation/SMC approach. Some summary comments are given in Section 2.6.

2.2 Model

2.2.1 Latent Threshold Models

In a univariate latent threshold model (LTM, Nakajima and West, 2013a), scalar observations y_t ($t = 1:T$) are represented as

$$y_t = (x_t \circ s_t)' \beta_t + \epsilon_t, \quad \epsilon_t \sim N(0, v_t), \quad \beta_t = (\beta_{1t}, \dots, \beta_{kt})', \quad (2.1)$$

$$s_{jt} = I[|\beta_{jt}| > d_j], \quad s_t = (s_{1t}, \dots, s_{kt})', \quad (2.2)$$

$$\beta_t = \mu + \Phi(\beta_{t-1} - \mu) + \omega_t, \quad \omega_t \sim N(0, \Sigma), \quad (2.3)$$

where x_t is a $k \times 1$ -vector of predictors known at time t , v_t is the observational variance, $d = (d_1, \dots, d_k)'$ are threshold parameters, and the error and innovation sequences, ϵ_t and ω_s , are independent over time and mutually independent. The \circ symbol represents the element-wise product, and $I(\cdot)$ is the indicator function. This is a non-linear (in state β_t) Gaussian state space model. The latent state follows the stationary vector autoregressive model of order 1 with $\mu = (\mu_1, \dots, \mu_k)'$, $\Phi = \text{diag}(\phi_1, \dots, \phi_k)$ and $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_k^2)$; this VAR(1) is thus a set of independent univariate AR(1) processes

$$\beta_{jt} = \mu_j + \phi_j(\beta_{j(t-1)} - \mu_j) + w_{jt}, \quad w_{jt} \sim N(0, \sigma_j^2), \quad (2.4)$$

independently over $j = 1:k$ and where $0 < \phi_j < 1$ for each j . LTMs represent dynamically adaptive variable selection models in which the effects of individual predictors may be apparent for some periods of time but practically irrelevant at other periods, based on the trajectory of the latent states and whether they lie outside/inside the threshold regions. Multiple applications show that this can substantially improve predictive modelling (e.g. Nakajima and West, 2013a,b, 2015; Zhou et al., 2014)

2.2.2 Challenges to Sequential Analysis

The challenges of on-line analysis can be seen immediately. Suppose we are standing at time t with a prior for the latent state β_t together with the fixed parameters (μ, Φ, Σ, d) and error variance v_t , all conditional on current information \mathcal{D}_{t-1} . The filtering step updates this prior to the posterior given, in addition, y_t . Whatever the form of the prior, the resulting posterior $p(\cdot | \mathcal{D}_t)$ is very complicated due to the significant non-linearities in the conditional likelihood function $p(y_t | \beta_t, v_t, d)$ from eqn. (2.1). This involves truncations in each β_{jt} dimension based on the threshold indicators s_{jt} , and a complicated form in the threshold parameters as well, so that the implied posterior is not at all manageable analytically. These issues are exacerbated

for models with higher-dimensional state vectors.

However, the fact that the likelihood function can be trivially evaluated at any set of values of (β_t, d, v_t) indicates the potential for particular methods, i.e. some form of SMC approach. It is also clear, however, that the dependence of the model on a set of fixed parameters (μ, Φ, Σ, d) in addition to the time-evolving states will present challenges; we do not, for example, have access to existing efficient SMC methods such as particle learning (Carvalho et al., 2010) that rely on “nice” analytic structure. Hence a more holistic approach to customizing SMC methods to the LTM context is mandated.

2.3 Sequential Importance Sampling and Model Emulation

This section assumes that constant model parameters and error variances are known, so focusing on SMC for filtering only on the time-evolving states β_t . Extension to include parameter learning then follows in Section 2.4.

2.3.1 SIS and Emulating Models

Particle-based methods of sequential important sampling (SIS, e.g., Liu and Chen, 1998; West, 1993b) represent the time $t-1$ posterior in terms of a discrete distribution on N particles β_t^i with weights w_{t-1}^i , i.e.,

$$p(\beta_{t-1}|\mathcal{D}_{t-1}) = \sum_{i=1:N} w_{t-1}^i \delta_{\beta_{t-1}^i}(\beta_{t-1}), \quad w_{t-1}^i > 0, \quad \sum_{i=1:N} w_{t-1}^i = 1,$$

where $\delta_b(\beta)$ is the k -dimensional Dirac delta function of β at point b . Under the VAR(1) evolution of eqn. (2.3), the conditional prior for β_t and the “auxiliary” index i is $p(\beta_t, i|\mathcal{D}_{t-1}) \propto p(\beta_t|\beta_{t-1}^i)w_{t-1}^i$ where the first term is the conditional normal p.d.f. of the state evolution. The resulting posterior is then

$$p(\beta_t, i|\mathcal{D}_t) \propto p(y_t|\beta_t)p(\beta_t|\beta_{t-1}^i)w_{t-1}^i, \quad (2.5)$$

with likelihood function $p(y_t|\beta_t)$ from eqn. (2.1).

We consider SIS methods that maximally exploit the analytic form of the prior here, each having proposal distributions of the form

$$\tilde{p}(\beta_t, i | \mathcal{D}_t) \propto \tilde{p}(y_t | \beta_t, \beta_{t-1}^i) p(\beta_t | \beta_{t-1}^i) w_{t-1}^i, \quad (2.6)$$

where $\tilde{p}(y_t | \beta_t, \beta_{t-1}^i)$ represents an approximation to the exact likelihood function based on the local region of the state space for β_t consistent with the value of prior particle β_{t-1}^i . So $\tilde{p}(\cdot|\cdot)$ may depend on β_{t-1}^i and model parameters as well as β_t ; different choices of $\tilde{p}(\cdot|\cdot)$ define different *emulators* of the LTM updates. If we now sample a pair (β_t^i, i) from the emulating density eqn. (2.6), importance reweighting is based on the updated/posterior weights give by the ratio of eqn. (2.5) to (2.6), i.e.,

$$w_t^i \propto \frac{p(\beta_t^i, i | \mathcal{D}_t)}{\tilde{p}(\beta_t^i, i | \mathcal{D}_t)} \propto \frac{p(y_t | \beta_t^i)}{\tilde{p}(y_t | \beta_t^i, \beta_{t-1}^i)}.$$

A core example is traditional auxiliary particle filtering (APF, e.g., Pitt and Shephard, 1999; Liu and West, 2001; Lopes et al., 2010) in which $\tilde{p}(y_t | \beta_t, \beta_{t-1}^i) = p(y_t | \hat{\beta}_t^i)$ for some choice of a point-estimate $\hat{\beta}_t^i$, such as the conditional prior mean $\hat{\beta}_t^i = \Phi \beta_{t-1}^i$. In this case, it is easy to directly sample (β_t^i, i) from eqn. (2.6) by composition: sample the auxiliary index i from the the set of N with weights $w_{t-1|t}^i \propto p(y_t | \hat{\beta}_t^i) w_{t-1}^i$, then draw the state from the evolution model/prior $p(\beta_t | \beta_{t-1}^i)$. We will refer to this as vanilla SMC, or simply SMC.

2.3.2 Choices of LTM Emulators

The first point is to simply note that any emulating model that replaces the s_{jt} in eqn. (2.1) with estimates or approximating values known at time t creates a model that is conditionally linear and Gaussian for the one-step filtering update, and thus moves us into a dynamic linear model with analytically trivial (Kalman filter-style) analysis. We consider several such emulators to be used in parallel. That is, at time

$t - 1$ use a conditionally Gaussian dynamic linear model (DLM) where $\tilde{p}(\cdot|\cdot)$ comes from

$$y_t = (x_t \circ q_t)' \beta_t + \epsilon_t, \quad \epsilon_t \sim N(0, v_t),$$

where q_t does not involve β_t but may be dependent on the past states β_{t-1} as well as other model parameters including thresholds d . Then sampled particles will be reweighted using the resulting approximation with indicators q_t^i based on the sampled auxiliary indicators. Efficiency of sampling, in terms of variability of resulting weights and the need to avoid degeneracy, depends on the choice of the forms of q_t .

Emulator 1: Dynamic Linear Model (DLM)

An easy first choice is to set $q_t = (1, \dots, 1)'$ for all t , resulting in the unthresholded DLM. This is equivalent to setting each $d_i = 0$ in the LTM, and follows the use of this as a global approximating model to define MCMC methods in non-sequential batch analysis of LTMs (Nakajima and West, 2013a). Computation and generation of proposal values is trivial, and this choice of emulator will be increasingly effective in cases when the data suggests generally low probabilities of thresholding for all predictors. In practical cases requiring substantial levels of dynamic sparsity, there will tend to be over-dispersion in posteriors for states and greater potential for particle weight degeneracy in the resulting SIS.

Emulator 2: Linearized LTM

A standard “extended Kalman filtering” approach uses a linear (in β_t) approximation to the mean term of y_t to define an approximating, analytically tractable DLM (e.g., West and Harrison, 1997, sect 13.2). That is, at any chosen value $\beta_t = c_t$ for some $c = (c_{1t}, \dots, c_{kt})'$, take $q_{jt} = I[|c_{jt}| > d_j]$ to define a linearized latent threshold model (LLTM). Here c_t can depend on anything but β_t , its value being known at time $t - 1$ based on existing state particles and parameter values. Particle-specific vectors c_t^i

customize a local approximation to each current particle, such as $c_t^i = \alpha + \Phi\beta_{t-1}^i$ with the intercept in AR(1) process $\alpha = (I - \Phi)\mu$ as used in the vanilla SMC using the auxiliary particle strategy. This strategy can be expected to improve on vanilla SMC since, in contrast to that standard approach, the LLTM strategy more thoroughly exploits the analytic form of the model, drawing particles β_t^i from the analytically available conditional (normal) posterior defined in the linearized model.

A potential drawback of this emulator is persistence in thresholding; use of $(I - \Phi)\mu + \Phi\beta_{t-1}$ at time t tends to imply the same thresholding pattern as that at time $t - 1$, being potentially under-adaptive in cases of more sudden change in elements of the state.

Emulator 3: Shrinkage

It is natural to consider alternative emulators that modify the “hard-thresholded” LLTM emulator above with a “soft-thresholding” extension where q_t contains dynamically updated thresholding probabilities. A natural choice is to take the expected value of the original indicator function s_{it} in the current on-line posterior of β_{it} . This results in

$$\begin{aligned} q_{jt} &= Pr[s_{jt} = 1 \mid \beta_{j(t-1)}] \\ &= 1 - \Phi \left[(d_j - \alpha_j - \phi_j\beta_{j(t-1)})/\sigma_j \right] + \Phi \left[-(d_j + \alpha_j + \phi_j\beta_{j(t-1)})/\sigma_j \right], \end{aligned}$$

where $\Phi[\cdot]$ is the standard normal c.d.f. and $\alpha_j = (1 - \phi_j)\mu_j$. Refer to this as the *shrinkage emulator* since $0 < q_{jt} < 1$. That is, at each time we simply plug-in the one-step ahead predictive probability $s_{jt} = 1$ for each state element j , so acting to more aggressively push towards zero those elements for which the conditional state distribution $p(\beta_t|\beta_{t-1})$ suggests a high probability of thresholding. Again as with all emulators, this will be applied at each particle β_{t-1}^i in the SMC analysis outlined in Section 2.3.1.

2.3.3 Comparison of Emulators

We are interested in exploring the comparative performance of emulators across situations. An holistic comparison from a Bayesian decision theoretic viewpoint is generated by taking a broader view of the definition of “model”, one that includes the specification of the emulator. Define the extended model $M = \{M_0, E, N\}$ as the triplet of elements: M_0 , the exact, or “target” model, here a specific LTM; the emulator E ; and a specific number of particles N for the SMC. We have four emulating methods E : vanilla SMC, DLM, LLTM and the shrinkage emulator; we explicitly include the standard SMC as a benchmark. The number of particles, N must be taken into account to evaluate its effect on the comparisons, and to explore potential insights relevant to its choice in practice. Denoting the set of models compared by $\mathcal{M} = \{M_1, \dots, M_m\}$, we make comparisons of all $M \in \mathcal{M}$ using the following metrics.

Marginal Likelihood. On sequential analysis of any $t = 1:T$ observations, the time T marginal likelihood values

$$p(y_{1:T}|M) = \prod_{t=1}^T p(y_t|\mathcal{D}_{t-1}, M), \quad \text{where } M \in \mathcal{M}, \quad (2.7)$$

score cumulative one-step forecasting performance (and, of course, are used to define conditional– time T – posterior model probabilities $p(M|\mathcal{D}_T)$, if desired.) In our SMC-based analyses of the LTM, the compositional one-step components of marginal likelihood are directly approximated sequentially over $t = 1:T$ via

$$p(y_t|\mathcal{D}_{t-1}, M) = \sum_{i=1}^N w_{t-1}^i N(y_t|(x_t \circ s_t^i)' \beta_t^i, v_t),$$

where (s_t^i, β_t^i) is the particle which evolves from the current posterior particle $(s_{t-1}^i, \beta_{t-1}^i)$ following the state equations. If the particles were resampled at $t - 1$,

then, of course, $w_{t-1}^i = 1/N$.

Since it directly underlies posterior model probabilities, higher marginal likelihood should be preferred. However, as noted above and as is explicit in eqn. (2.7), this essentially scores models based only on one-step ahead forecasting accuracy; other metrics should also be considered.

Empirical Forecast Errors. In our macroeconomic example— as in other applications of LTMs— key interest lies in forecasting multiple steps ahead. Hence, following Nakajima and West (2013a), we include evaluations of h -step ahead empirical forecasting accuracy, in terms of root mean squared forecast errors (RMSFE), for $h \geq 1$ up to some horizon. For any $h \geq 1$, the RMSFE of h -step ahead predictions across period $t = t_0:t_1$ is denoted by $R_{h,t_0:t_1}$ and defined as

$$R_{h,t_0:t_1}^2 = \sum_{t=t_0:t_1} (y_{t+h} - \hat{y}_{t+h})^2 / (t_1 - t_0 + 1)$$

where \hat{y}_{t+h} is the forecast mean. SMC methods trivially yield forecasts in terms of Monte Carlo samples: at any time t , sampling current state and parameter particles underlies projection of those samples through the evolution equations over $R = 1:h$ steps ahead, and then Monte Carlo approximation of the predictive means. In models with lagged values of y_t as predictors— as in latent thresholded time-varying autoregressive models such as in Section 2.5— this is extended to generate samples of each y_{t+r} at each $r = 1:h$ to be saved and used as predictors for the next step $r + 1$. Then, $p(y_{t+h}|\mathcal{D}_t, M)$ can be simulated with the posterior particles at time t .

Effective Sample Size. Effective Sample Size (ESS, Liu, 1996) is an often considered a measure of efficiency in SIS, and importance sampling more generally. At each time t , the current ESS measure $M_{t,ESS} = 1 / \sum_{i=1:N} (w_t^i)^2$ relates to variation in the resampling weights w_t^i , so defines a metric relevant to the current accuracy

of approximation of the posterior for state and parameters. Larger values on the range $M_{t,ESS} \in [1, N]$ indicate more uniform weights, whereas smaller values indicate substantial variation and particle “degeneration”. As discussed in Gruber and West (2016), there is also direct relationship between ESS and the Kullback-Leibler divergence of the Monte Carlo approximating posterior from the exact/target posterior, derived from the Monte Carlo/discrete version of Kullback-Leibler based on the entropy of the weights relative to uniformity, $\sum_{i=1:N} w_t^i \log(Nw_t^i)$ (e.g. West, 1993a).

2.4 Sequential Learning with Fixed Parameters and Volatilities

We now discuss completion of the overall sequential analysis that includes learning on the fixed model parameters $(\alpha, \Phi, \Sigma, d)$, after re-parameterizing μ into $\alpha = (I - \Phi)\mu$, as well as conditional error variances v_t under specific volatility models. For these we utilize best-practices based on particle learning (PL) and auxiliary particle filtering (APF) overlaid on the SIS emulation approach for states.

2.4.1 Particle Learning for AR(1) Parameters

The parameters $(\alpha, \Phi, \Sigma) = \{\alpha_j, \phi_j, \sigma_j^2; j = 1:k\}$, appear only in the state evolution model of eqns. (2.3,2.4) that define a set of conditionally independent AR(1) models. We adopt the traditionally used priors for these AR(1) parameters: independently over $j = 1:k$. These are either normal/gamma priors for the $(\alpha_j, \phi_j, 1/\sigma_j^2)$, or modified forms in which the conditional normal prior for any one or more of the ϕ_j are truncated to $(0, 1)$ or $(-1, 1)$ to enforce stationarity. In these cases, particle learning (Carvalho et al., 2010; Lopes et al., 2010) applies to these parameters since, conditional on values of the latent state process β_* over times $0:t$, there exist conditional sufficient statistics \mathcal{S}_t for (α, Φ, Σ) that are trivially updated from time $t - 1$ to t based on values of β_t, β_{t-1} and y_t . That is, their time t full conditional posterior reduces to $p(\alpha, \Phi, \Sigma | \beta_{0:t}, d, \mathcal{D}_t) = p(\alpha, \Phi, \Sigma | \mathcal{S}_t)$ where \mathcal{S}_t is a trivially updated sum-

mary defining a set of k conditional normal/gamma, or truncated normal/gamma posteriors across the $j = 1:k$ latent state processes. These conditional posteriors above are easy to sample, and the PL approach is enabled. See Section 2.7.2 for full technical details.

2.4.2 Auxiliary Particle Filter for Latent Threshold

The mathematical form of the conditional likelihood function in thresholds d obviates the possibility of PL for these parameters. We therefore adopt the standard APF method for fixed parameters, using kernel approximations to each of the k marginal posterior densities for each d_j at each time (Liu and West, 2001) to regenerate particles for weighting in the SIS analysis. The specific form of the kernel APF uses non-Gaussian kernels customized to the threshold context, as detailed in Section 2.7.3.

2.4.3 Volatility Models

In most applications of LTMs. we will utilize some form of stochastic volatility model for the error variances v_t , the popular choices being the simple gamma/beta random walk model (West and Harrison, 1997) and the widely-used stationary log-AR(1) model; several examples of each appear in prior MCMC-based analyses of LTM (Nakajima and West, 2013a,b, 2015; Zhou et al., 2014). The simple gamma/beta random walk model has the advantage that the emulator-based SIS method for latent states β_t is directly extended to include particles for the v_t with little change. Use of the log-AR(1) volatility model, which is often desirable in view of the ability to constrain to stationarity, introduces v_t as a new latent state along with the additional fixed parameters of the AR(1) model for $\log(v_t)$. However, the SIS approach is trivially extended to include efficient resampling of v_t particles, and this can be directly coupled with an extension of the PL method to now include this additional

fixed parameters. Full technical details appear in Section 2.7.4.

2.4.4 Sequential Posterior Update

In the following, we use the log-AR(1) model to model the latent observational variance v_t sequence;

$$v_t = e^{h_t}, \quad h_t = (1 - \phi_h)\mu_h + \phi_h h_{t-1} + \eta_t, \quad \eta_t \sim N(0, \sigma_h^2), \quad (2.8)$$

where $0 < \phi_h < 1$. Denote the intercept by $\alpha_h = (1 - \phi_h)\mu_h$. The three additional parameters, α_h , ϕ_h and σ_h^2 , have sufficient statistics conditional on values of the variance series, this enables the use of particle learning.

In total, the variables to be sampled at time t are $\{\beta_t, h_t, d\} \cup \{\alpha_j, \phi_j, \sigma_j^2\}_{j \in \{1:k, h\}}$. We group them into three categories; state variables $\theta_t = \{\beta_t, h_t\}$, parameters with sufficient statistics $\vartheta = \{\alpha_j, \phi_j, \sigma_j^2\}_{j \in \{1:k, h\}}$, and latent thresholds d . For the latent thresholds, there are no sufficient statistics and hence there is a need for density estimation. Denote the set of sufficient statistics of ϑ by \mathcal{S}_t ; this is a function of $\{\theta_t, \theta_{t-1}, \mathcal{S}_{t-1}\}$. Finally, summarize these parameters at time t by $\Theta_t = \{\theta_t, \mathcal{S}_t, \vartheta, d\}$. and the sequential analysis in the following is still valid for them.

The sequential update at time t starts with the on-line posterior at time $t - 1$ as the prior, approximated by particles as

$$p(\Theta_{t-1} | \mathcal{D}_{t-1}) = \sum_{i=1}^N w_{t-1}^i p_{t-1}^i(d) \delta_{(\theta_{t-1}^i, \mathcal{S}_{t-1}^i, \vartheta^i)}(\theta_{t-1}, \mathcal{S}_{t-1}, \vartheta), \quad (2.9)$$

where $w_{t-1}^i = 1/N$ for all i . The definition of kernel function, $p_{t-1}^i(d)$, is given in 2.7.3. With a little modification to the calculation in Section 2.3, the joint distribution of the implied time t posterior can be decomposed as

$$p(\Theta_t, i | \mathcal{D}_t) \propto p(\vartheta | \mathcal{S}_t) p(\mathcal{S}_t | \theta_t, \theta_{t-1}^i, \mathcal{S}_{t-1}^i) p(y_t | \theta_t, \vartheta^i, d) p(\theta_t | \theta_{t-1}^i, \vartheta^i) p_{t-1}^i(d) w_{t-1}^i, \quad (2.10)$$

where $p(y_t|\theta, \vartheta^i, d)$ is the likelihood in the observational equation and $p(\beta_t|\beta_{t-1}^i, \vartheta^i)$ is the prior in the state transition specified in eqns. (2.1,2.3). $p(\vartheta|\mathcal{S}_t)$ is the posterior of fixed parameters and $p(\mathcal{S}_t|\theta_t, \theta_{t-1}^i, \mathcal{S}_{t-1}^i)$ defines the deterministic update of sufficient statistics; the details of these terms are discussed in 2.7.2. The distribution of $\{\Theta_t, i\}$ in eqn. (2.10) is the target distribution to sample. Next, consider the proposal density used in SIS,

$$\tilde{p}(\Theta_t, i|\mathcal{D}_t) \propto p(\vartheta|\mathcal{S}_t)p(\mathcal{S}_t|\theta_t, \theta_{t-1}^i, \mathcal{S}_{t-1}^i)\tilde{p}^i(y_t|\beta_t)p(\theta_t|\theta_{t-1}^i, \vartheta^i)p_{t-1}^i(d)w_{t-1}^i, \quad (2.11)$$

where \tilde{p}^i means the likelihood, posterior and forecast densities of the emulator model, each of which may depend on i -th particle and $(\hat{h}_t^i, \theta_{t-1}^i, \vartheta^i, \hat{d}^i)$. Note that the product of the likelihood and prior is re-written as the product of the posterior and marginal likelihood,

$$\tilde{p}^i(y_t|\beta_t)p(\theta_t|\theta_{t-1}^i, \vartheta^i) = \tilde{p}^i(\beta_t|\mathcal{D}_t)\tilde{p}^i(y_t|\mathcal{D}_{t-1})p(h_t|h_{t-1}^i, \vartheta^i).$$

To be precise, the likelihood of emulator, $\tilde{p}^i(y_t|\beta_t)$, means $\tilde{p}(y_t|\beta_t, \hat{h}_t^i, \vartheta^i, \hat{d}^i)$ and is defined by

$$y_t = (x_t \circ q_t^i)' \beta_t + N(0, v_t^i), \quad q_t^i = q_t(\beta_{t-1}^i, \hat{d}^i, \vartheta^i), \quad v_t^i = e^{\hat{h}_t^i}. \quad (2.12)$$

Note that this emulator model, to be used to sample the candidate particles, does not include h_t and d , but \hat{h}_t^i and \hat{d}^i that are available prior to sampling θ_t . Note also that, as mentioned in Section 2.3, posterior and forecast densities, $\tilde{p}^i(\beta_t|\mathcal{D}_t)$ and $\tilde{p}^i(y_t|\mathcal{D}_{t-1})$, are Gaussian and derived analytically by forward filtering. Therefore, the former density is used to sample state variables and the latter is absorbed in the weights of particles that defines the new weights for propagation as $w_{(t-1)|t}^i \propto \tilde{p}^i(y_t|\mathcal{D}_{t-1})w_{t-1}^i$. For notational clarity, rename the auxiliary particle index i as i_0 (this is the particle at time $t-1$ used in sampling the current particle at t) and introduce i_1 for the index of a current particle to be sampled. Then, new particles

$\{\Theta_t^{i_1}\}_{i_1=1:N}$ from the posterior of emulator \tilde{p} need to be resampled with weights proportional to the ratio of the target and proposal densities

$$w_t^{i_1} \propto \frac{p(\Theta_t^{i_1}|\mathcal{D}_t)}{\tilde{p}(\Theta_t^{i_1}|\mathcal{D}_t)} \propto \frac{p(y_t|\theta_t^{i_1}, \vartheta^{i_0}, d^{i_1})}{\tilde{p}(y_t|\beta_t^{i_1}, \hat{h}_t^{i_0}, \vartheta^{i_0}, \hat{d}^{i_0})}. \quad (2.13)$$

This procedure corrects the bias caused by the use of the emulator. The sampling procedure is modified as follows:

1. (Propagate) For $i_1 = 1:N$, repeat the following.

(i) Sample an auxiliary index i_0 from $\{1, \dots, N\}$ with probability $w_{t-1|t}^{i_0} \propto \tilde{p}^{i_0}(y_t|\mathcal{D}_{t-1})w_{t-1}^{i_0}$. Call this index $i_0 = i_0(i_1)$.

The observational variance in the emulator is replaced by $\hat{v}_t = \exp\{\hat{h}_t^{i_0}\}$, and $\hat{h}_t^{i_0} = \alpha_h^{i_0} + \phi_h^{i_0} h_{t-1}^{i_0}$. Similarly, the latent threshold parameter involved in the computation has the particle-dependent value, \hat{d}^{i_0} , that is defined based on the theory of APF in (2.7.3). Conditional on those variables, the forecast density in the weight is computed by forward filtering in (2.7.1).

(ii) Sample d^{i_1} from $p^{i_0}(d|\mathcal{D}_{t-1})$.

The elements of this vector are sampled independently from gamma distributions given in (2.7.3).

(iii) Sample $\beta_t^{i_1}$ from $\tilde{p}^{i_0}(\beta_t|\mathcal{D}_t)$ and $h_t^{i_1}$ from $N(h_t|\alpha_h^{i_0} + \phi_h^{i_0} h_{t-1}^{i_0}, (\sigma_h^{i_0})^2)$. The former distribution is Gaussian and computed by forward filtering as discussed in (2.7.1).

(iv) Construct $\mathcal{S}_t^{i_1}$ from $\mathcal{S}_{t-1}^{i_0}$, $\theta_t^{i_1}$, and $\theta_{t-1}^{i_0}$. See (2.7.2).

(v) Sample ϑ^{i_1} from $p(\vartheta|\mathcal{S}_t^{i_1})$.

Each element of ϑ , i.e. $\{\alpha_j, \phi_j, \sigma_j^2\}$ for $j = \{1:k, h\}$, is sampled independently from the normal-inverse gamma distribution given in (2.7.2).

2. (Resampling) Resample $\{\Theta_t^{i_1}, \Theta_{t-1}^{i_0(i_1)}\}_{i_1=1:N}$ with weights $w_t^{i_1}$.

This weight is proportional to the ratio of two likelihoods: LTM $N(y_t | (x_t \circ s_t^{i_1})' \beta_t^{i_1}, v_t^{i_1})$ and Emulator $N(y_t | (x_t \circ q_t^{i_0(i_1)})' \beta_t^{i_1}, \hat{v}_t^{i_0(i_1)})$, where $v_t^{i_1} = \exp\{h_t^{i_1}\}$ and $\hat{v}_t^{i_0(i_1)} = \exp\{\hat{h}_t^{i_0(i_1)}\}$.

2.5 Application: US Macroeconomic Study

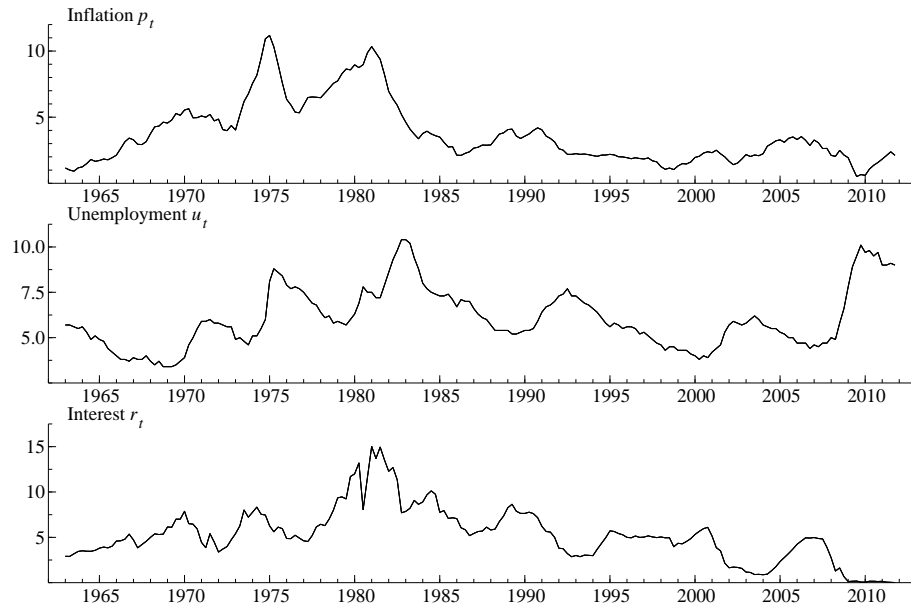


FIGURE 2.1: US macroeconomic indices. From top to bottom, the inflation, unemployment rate and short-term nominal interest rate in US from 1963 to 2011. This is the same dataset as used in Nakajima and West (2013a).

The US macroeconomic time series of Nakajima and West (2013a) has $m = 3$ series of quarterly inflation, unemployment and nominal short-term interest rates from 1963/Q1 to 2011/Q4 ($T = 196$); see Figure 2.1. The inflation rate is the annual percentage change in a chain-weighted GDP price index, the unemployment rate is seasonally adjusted (all workers over 16), and the interest rate is the yield on three-month Treasury bills. These and related macroeconomic series are of central interest in economic policy studies; similar series are widely studied in national/central

banking analyses in multiple countries, where models to forecast several steps ahead underlie impulse response analyses that feed into policy discussions. Improved sequential analysis for forecasting is a key interest; we study this here using LTM extensions of the class of time-varying vector autoregressive models that have become increasingly adopted for such macroeconomic studies (e.g. Cogley and Sargent, 2005; Primiceri, 2005; Koop et al., 2009; Koop and Korobilis, 2010; Korobilis, 2011; Koop and Korobilis, 2013; Nakajima and West, 2013a). Modeling assumptions and details follow the latter reference, and sequential analysis results are compared with the MCMC results of that paper.

2.5.1 Latent Thresholded TV-VAR Models

Model Structure

The m -dimensional column vector time series $Y_t = (y_{1t}, \dots, y_{mt})'$ follows the TV-VAR model of order p given by

$$(I - A_t)Y_t = c_t + B_{1t}Y_{t-1} + \dots + B_{pt}Y_{t-p} + N(0, \Lambda_t^{-1}),$$

$$\Lambda_t = \text{diag}(e^{-h_{1t}}, \dots, e^{-h_{mt}}), \quad (2.14)$$

where $c_t = (c_{1t}, \dots, c_{mt})'$, the B_{lt} are $m \times m$ -matrices of time-varying coefficients at lags $l = 1:p$, and A_t is a strict lower triangular matrix whose non-zero entries represent contemporaneous and time-varying dependencies among the univariate series y_{jt} . In this study, $m = 3$ with y_{1t} = inflation rate, y_{2t} = unemployment rate, and y_{3t} = interest rate in quarter t .

The general development in Sections 2.2–2.4 applied to a univariate time series y_t . Our development here uses that directly for each of these three univariate series $y_t = y_{jt}$, for each $j = 1, 2, 3$ in turn and in parallel. This arises as the multivariate model *decouples* the three for model fitting and sequential learning, while *recoupling* them to properly infer cross-series relationships for forecasting. This yields efficiencies in

sequential computation as it does in the original MCMC analysis, and is enabled by the triangular form of A_t , i.e.,

$$A_t = \begin{bmatrix} 0 & 0 & 0 \\ a_{21t} & 0 & 0 \\ a_{31t} & a_{32t} & 0 \end{bmatrix}.$$

The implied reduced model form is

$$Y_t = c_t^* + B_{1t}^* Y_{t-1} + \cdots + B_{pt}^* Y_{t-p} + N(0, \Omega_t^{-1}), \quad \Omega_t = (I - A_t') \Lambda_t (I - A_t), \quad (2.15)$$

where $c_t^* = (I - A_t)^{-1} c_t$ and $B_{jt}^* = (I - A_t)^{-1} B_{jt}$ for $j = 1:p$. The precision matrix Ω_t is a ‘‘Cholesky-style’’ time-varying volatility matrix with the flexibility to model different patterns of change over time in contemporaneous relationships. The univariate series follow conditionally independent dynamic regression models in the *compositional* form (Zhou et al., 2014)

$$y_{1t} = c_{1t} + b'_{11t} Y_{t-1} + \cdots + b'_{1pt} Y_{t-p} + N(0, e^{-h_{1t}}), \quad (2.16)$$

$$y_{jt} = c_{jt} + b'_{j1t} Y_{t-1} + \cdots + b'_{jpt} Y_{t-p} + \sum_{r=1}^{j-1} a_{jrt} y_{rt} + N(0, e^{-h_{jt}}), \quad \text{for } j \geq 2,$$

where b'_{jt} is the j -th row of matrix B_{jt} . The univariate LT-AR(1) model of Section 2.2 is applied independently to each of the scalar elements of c_t and each B_{jt} . For each series j , map to the notation of Section 2.2 with: (i) y_t is substituted by y_{jt} ; (ii) β'_t is given by $(c_{jt}, b'_{j1t}, \dots, b'_{jpt}, a_{j1t}, \dots, a_{j(j-1)t})$ where the last terms a_* appear only for $j > 1$; (iii) x'_t is given by the (known at time t) vector $(1, Y'_{t-1}, \dots, Y'_{t-p}, y_{1t}, \dots, y_{(j-1)t})'$ where the last terms y_* appear only for $j > 1$; and (iv) v_t is the volatility $e^{-h_{jt}}$.

Decoupled SIS and Recoupled Forecasting Analyses

With the full multivariate model decomposed into these m univariate, independent LTMs, analysis runs on each in parallel, applying the SIS approach with the same

Monte Carlo sample size N across series. At each time t , with Θ_{jt} representing the set of current states, volatilities and fixed model parameters in analysis of series j , these parallel SIS analyses will generate posterior samples $\{\Theta_{jt}^i; i = 1:N\}$ that can be directly combined to recouple across series $j = 1:m$ for inference on Ω_t , if desired. Forecasting is via simulation and most efficiently recouples by exploiting the compositional representation of the full joint distribution given in eqn. (2.16). This is basically as used in MCMC analysis when forecasting ahead (Nakajima and West, 2013a).

At time t , forecast ahead to time $t+T$, ($T > 0$), by repeating the following procedure independently (and in parallel) over Monte Carlo particles $i = 1:N$. Beginning at $s = 1$, proceed sequentially to each step-ahead $s = 1:T$ by recursively simulating the model equations as follows.

- First, at $s = 1$ simulating values of Y_{t+1} is based wholly on the model and posterior conditional on the observed data \mathcal{D}_t . However, for $s > 1$ we will have available forecast samples $Y_{t+1:t+s-1}^i = \{Y_{t+1}^i, \dots, Y_{t+s-1}^i\}$, ($i = 1:N$), to augment the observed data \mathcal{D}_t .
- To generate samples for Y_{t+s} :
 1. First, evolve all N particles in the states and volatilities to “move” through the one-step AR(1) evolutions of each, in each of the $j = 1:m$ models.
 2. Set $j = 1$ and draw a sample $y_{1,t+s}^i$ from the first model in eqn. (2.16) with all required states and parameters set at value i from the current particles, and conditioning on already sampled values $Y_{t+1:t+s-1}^i$.
 3. For $j > 1$, draw a sample $y_{j,t+s}^i$ from the j -th model in eqn. (2.16) with all required states and parameters set at i from the current particles, and conditioning on the current values of all $y_{r,t+s}^i$ just simulated for $r =$

1:($j - 1$) as well as the already sampled $Y_{t+1:t+s-1}^i$.

This results in a full posterior predictive sample of the entire trajectory of the m -dimensional series over steps ahead $s = 1:T$, i.e., the full predictive sample $\{Y_{t+1}^i, \dots, Y_{t+T}^i; i = 1:N\}$.

Priors

Priors are modified versions of those in Nakajima and West (2013a). Some differences arise due to the interest in maintaining conditional conjugacy as required in the PL component of the SIS; this, for example, the ϕ_i are assigned normal priors. Otherwise, the priors and hyperparameters are set to be as consistent with Nakajima and West (2013a) as possible: $\beta_{i0} \sim N(0, 2)$ (initial distribution of state variables); $(\alpha_i, \phi_i, \sigma_i^{-2}) \sim N(\alpha_i|0, 27.938\sigma_i^2) N(\phi_i|0.95, 22.8\sigma_i^2) G(\sigma_i^{-2}|20, 0.01)$ for $i \leq mp + 1$ (AR(1) parameters for c and B); $(\alpha_i, \phi_i, \sigma_i^{-2}) \sim N(\alpha_i|0, 1.47\sigma_i^2) N(\phi_i|0, 1.2\sigma_i^2) G(\sigma_i^{-2}|2, 0.01)$ for $i > mp + 1$ (AR(1) parameters for A); and $(\alpha_{h,i}, \phi_{h,i}, \sigma_{h,i}^{-2}) \sim N(\alpha_{h,i}, \phi_{h,i}|m_{h,i}, \sigma_{h,i}^2 C_{h,i}) G(\sigma_{h,i}^{-2}|2, 0.01)$ for $i = 1:m$ (AR(1) parameters for h) where

$$m_{h,i} = \begin{bmatrix} -0.23 \\ 0.95 \end{bmatrix}, \quad C_{h,i} = \begin{bmatrix} 24.95 & 5.4 \\ 5.4 & 1.2 \end{bmatrix}.$$

These priors have the same means (or mode for ϕ_i) and variances as those in Nakajima and West (2013a). Note that the priors of d_i and $(\beta_{i0}, \gamma_i, d_i)$ are used only in sampling the initial particles. This samples the former from $d_i \sim [0, \mu_i + 3\sigma_i]$ and the latter from its stationary distribution. The discount factors in the kernel density of (d, γ) are set 0.97.

2.5.2 Posterior Analysis

This subsection presents some results on parameter learning using the Shrinkage emulator, with $N = 50,000$ particles for each regression, at each time (150,000

particles in total). The results here naturally show differences to those based on MCMC in Nakajima and West (2013a), due inherently to the sequential analysis here in which posteriors are based only on historical data. That said, there is a good degree of concordance and they show the similar patterns in posterior distributions and thresholding that are meaningful and interpretable in the context of macroeconomic study.

Figure 2.2 shows the posterior median and 90% intervals of B_{1t} for $t = 1:T$ and its inclusion probability (i.e., probability that the variable is not thresholded). From this figure, the analysis confirms the basic characteristics that are known by the existing analyses by TV-VAR in macroeconomic applications: some of the diagonal, self-autoregressive elements are positively active, while the off-diagonal, cross-autoregressive terms have their posterior mass around zero, meaning they contribute little to the model. This is more apparent in the plots of inclusion probabilities which show the model thresholded all the entries of the matrix except for $(1, 1)$ -element, $b_{11,1t}$, which is significantly positive. The corresponding threshold parameters estimated by APF are plotted in Figure 2.3. In this figure, it is seen that the posterior of thresholds become small to include the state variable more for $(1, 1)$ -entry, but unchanged for the others. From these results, it seems that the first variable, inflation, has strong and significant autocorrelation of order 1. The other terms, on the other hand, are completely thresholded and ignored in the model. In contrast, $(1, 2)$ -entry is included in the model with more than 80% posterior probability until 1980, but later excluded from the model, losing all the inclusion probability. This is the typical, local sparsity in time series, in addition to “significance” and “insignificance” in the usual variable selection problem; the model allows cases where the state variables are “sometime significant, sometime not.”

The other state variables in B_{2t} and B_{3t} are shown in the appendix: Figures 2.10 and 2.11 in Section 2.8. In contrast to the first-lag matrix, these evidence higher

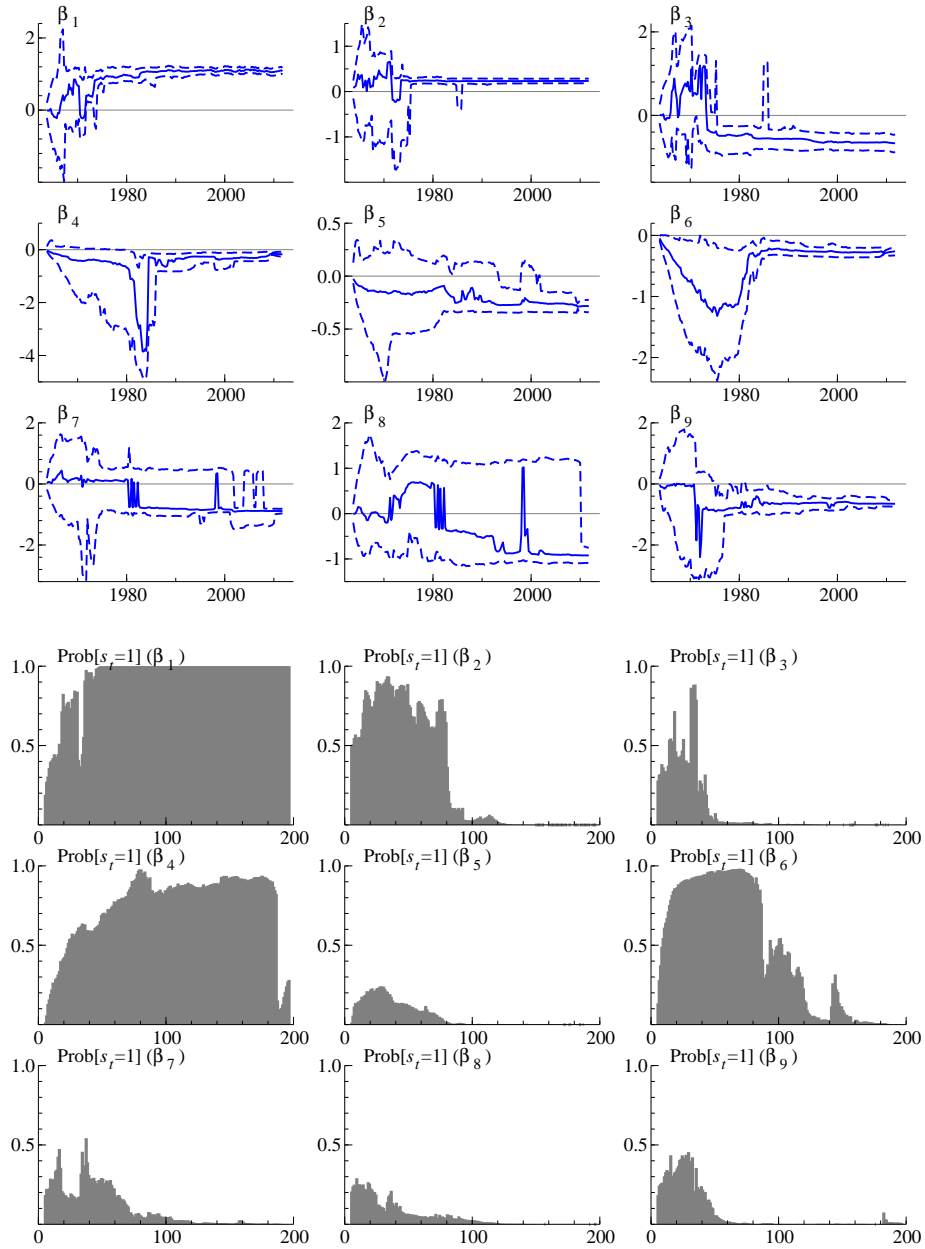


FIGURE 2.2: Posterior estimate of B_{1t} and the inclusion probability. Each element of the 9 plots corresponds with the result of (i, j) -element of B_{1t} . In the upper 3×3 frames, the solid line and dotted lines represent on-line updates of the posterior median and 90% intervals of (i, j) -element of B_{1t} for $t = 1:196$, respectively. The lower 3×3 frames show the corresponding values of each the inclusion probability defined by $Pr[|\beta_{jt}| > d_j]$ for each of B_{1t} .

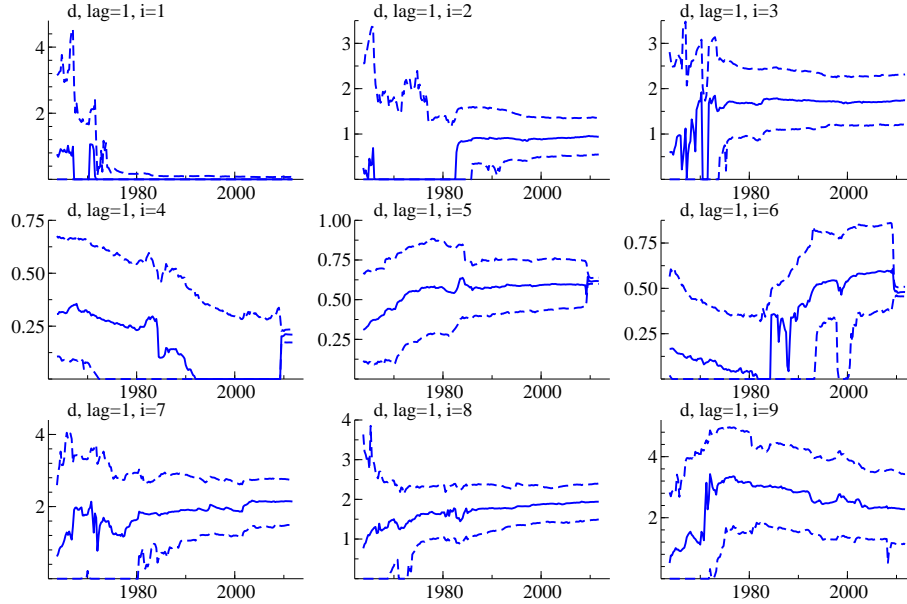


FIGURE 2.3: Posterior estimate of thresholds d for B_{1t} . Shown are the posterior medians and 90% intervals of thresholds d for each (i, j) -element of B_{1t} .

levels of sparsity, having fewer entries with inclusion probability higher than 50%.

This analysis also highlights the advantage of particle learning in the on-line posteriors of AR(1) parameters. Section 2.8 collects some of those results; for parameters (α, Φ, Σ) underlying B_{1t} , see Figures 2.12, 2.13 and 2.14. Note that the diffuse priors are gradually shaped into the concentrated posteriors as they adapt to the incoming stream of data.

In addition to the TV-VAR coefficients, the simultaneous regressive coefficient, or simultaneous correlations, denoted by A_t , are expected to be significant. Figure 2.4 shows the result of posterior analysis of A_t , the simultaneous effect among the three variables. All the three variables are always significant in the sense of the posterior inclusion probabilities. Here, the effect from inflation to interest rate, a_{21t} in the first row in Figure 2.4, is clear in the location of the posterior density. The spike in 2010 is obvious and has potential interpretation of increased dependence between inflation and unemployment after the financial crisis. Also, the posterior on the link between

inflation and interest rate, a_{31} in the third row in Figure 2.4, almost always favors positive values, except for the last few years. This sudden decline of the posterior median in the last few years is meaningful, reflecting the impact of the financial crisis that yields interest rates that are stable at low values as they were more and more controlled by central bank policies.

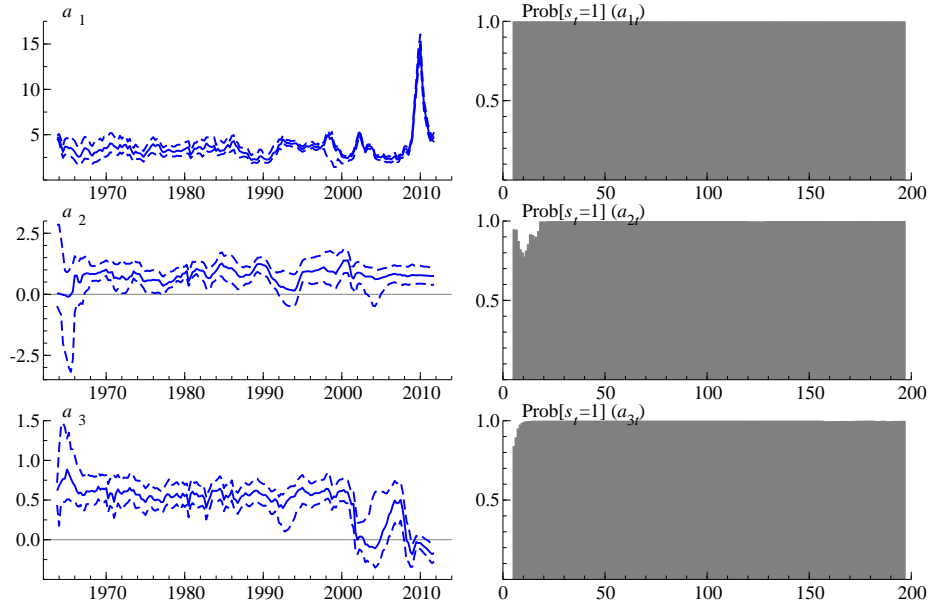


FIGURE 2.4: Posterior trajectories for $(a_{21t}, a_{31t}, a_{32t})$ and the corresponding inclusion probabilities.

Figure 2.5 plots the posteriors of the stochastic volatilities. Their dynamics are apparently significant, which shows the appropriateness of modeling time-varying volatilities in this context. Some differences relative to a global MCMC analysis are partly explained by the different results on the intercepts shown in Figure 2.6; the different values and thresholding of c_t will naturally impact to reduce/inflate of the variance of the error terms.

The predictive accuracy of the estimated model can be seen in the 1-step and 4-step ahead predictive distributions shown in Figure 2.7. The posterior medians track the actual observations and the credible interval correctly quantify the uncertainty.

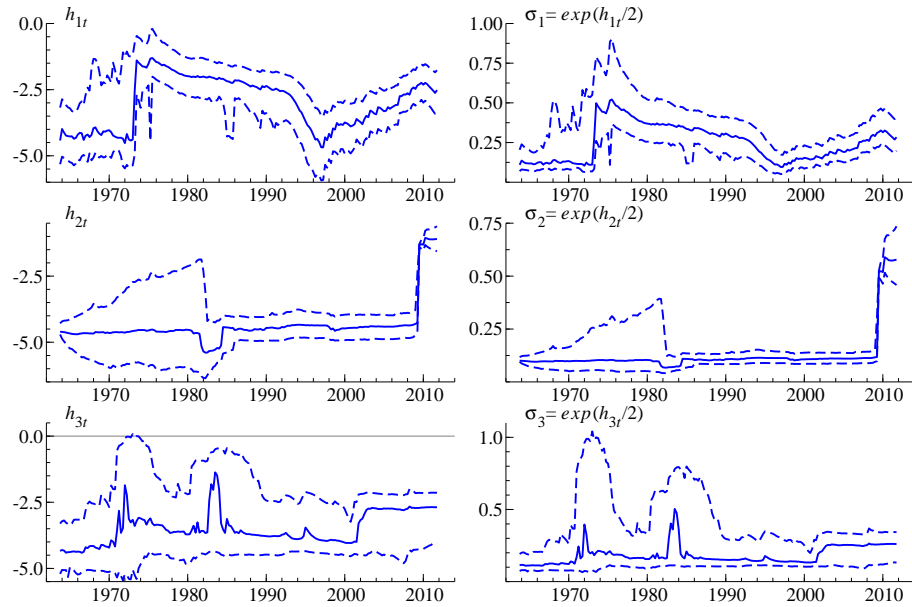


FIGURE 2.5: Stochastic volatilities (h_{1t}, h_{2t}, h_{3t}). The left column shows the results on the log-scale, and the right column shows them on the original scale (standard deviation).

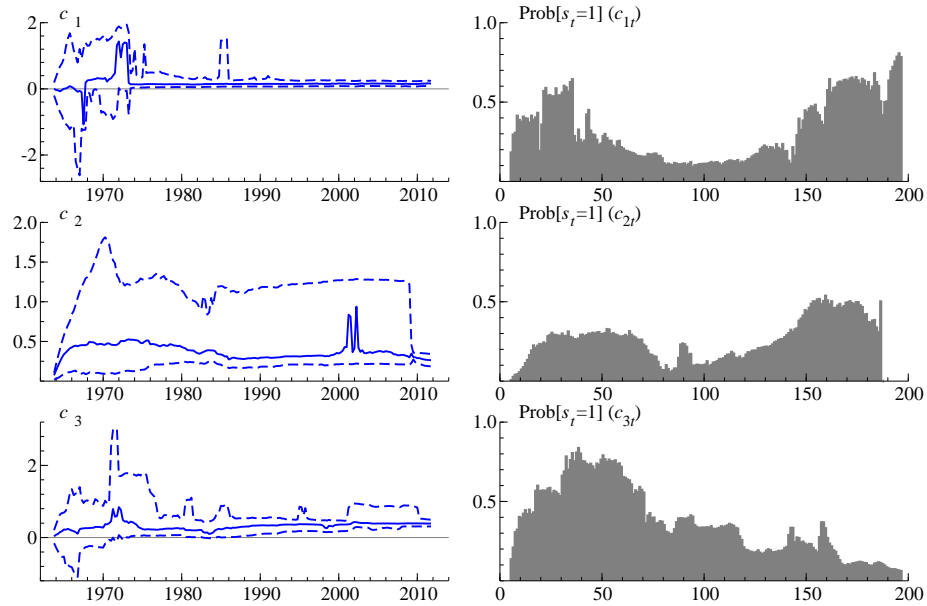


FIGURE 2.6: Posterior trajectories for intercepts (c_{1t}, c_{2t}, c_{3t}) and the corresponding inclusion probabilities.

In the first half of the time series, 4-step ahead predictions become challenged because of the lack of data and diffuse posteriors, and succeed in stabilizing in the last half. This result is summarized and compared with those of the other emulators in the next subsection.

Overall, our sequential method is able to represent both global and local sparsity and, to some extent, successfully recover the well-known finding in LTMs from the dataset, under almost the same conditions as MCMC in terms of, for example, prior settings.

2.5.3 Comparison

In this section, the three emulators and vanilla SMC are compared using the metrics introduced in Section 2.3.3. The number of particles is still $N = 50,000$ for each variable and emulator for fair comparison.

First, the marginal likelihoods and model probabilities are calculated after obtaining the particles from the sequentially updated posteriors using each method. As seen in Table 2.1 and Figure 2.8, the LLTM emulator has the largest marginal likelihood, followed by the shrinkage emulator. In particular, both emulators are still successful in the sense of marginal likelihood between 1991 and 2011, the period which includes the time of financial crisis and in which any model or emulator is challenged by the series of unusual events and corresponding observations. This result shows that LLTM and Shrinkage emulators totally outperform SMC, while a bad choice of emulator, such as the DLM emulator in the presence of sparsity, leads

Table 2.1: Log-Marginal Likelihoods of emulators.

log ML	DLM	Shrinkage	LLTM	SMC
Full Sample	-1222.4	-727.88	-566.97	-1080.0
1996-2011	-377.78	-146.97	-102.57	-290.16

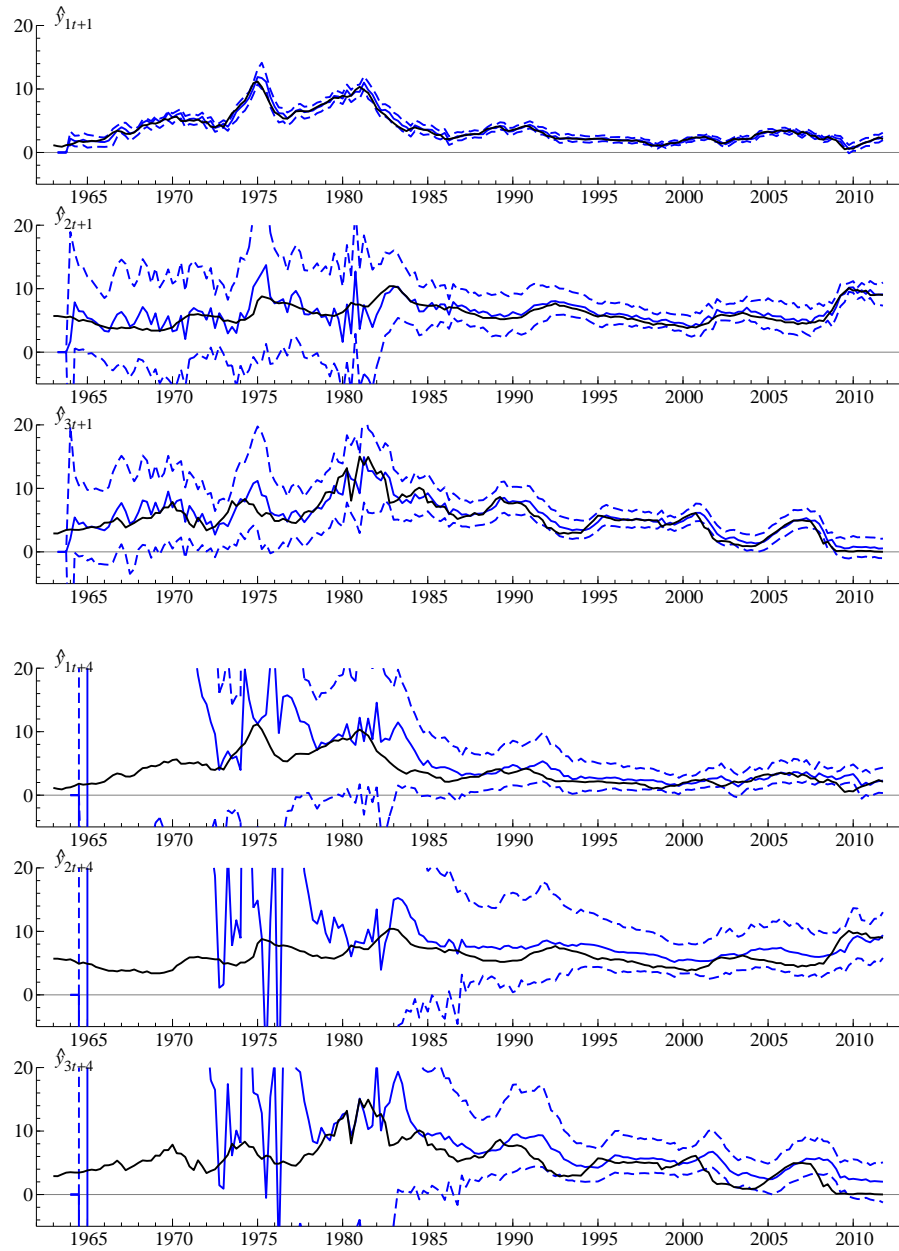


FIGURE 2.7: 1-step/4-step ahead predictive distributions with observation. The summaries of time trajectories of sequentially updated predictive distributions for the three series are shown for 1-step (top) and 4-step (bottom) ahead forecasting.

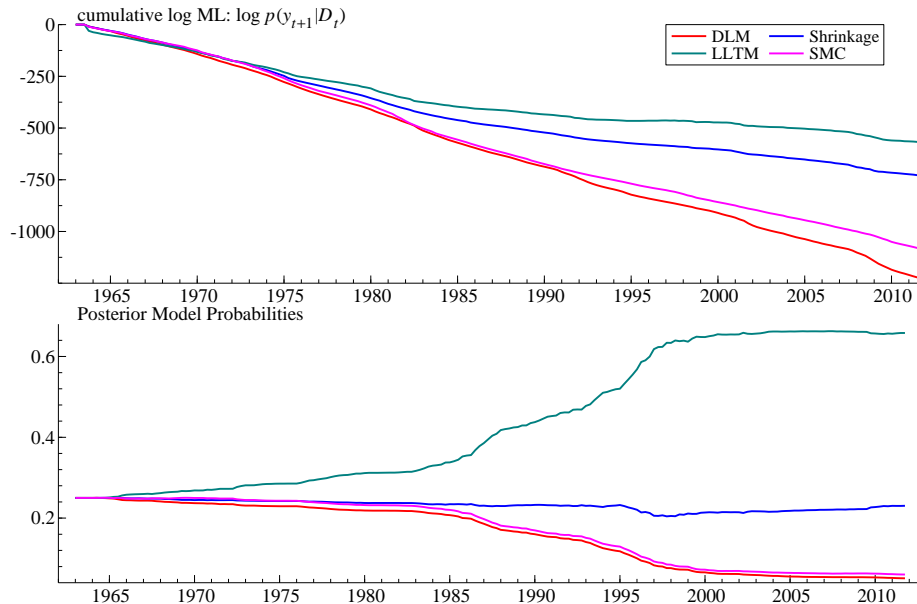


FIGURE 2.8: Comparison of log marginal likelihoods and posterior model probabilities. Log marginal likelihoods (Top) and posterior model probabilities (Bottom) for the three emulators and vanilla SMC. Red: DLM, Blue: Shrinkage, Green: LLTM, Pink: Vanilla SMC.

to less reliable analysis.

Since the marginal likelihood considers one-step ahead prediction only, the multi-step forecast errors are considered by RMSFE. Following Nakajima and West (2013a), the period for forecasting starts from 1996 Q4 and ends at 2011 Q4. At each time point of the period t based on current data \mathcal{D}_t , predictive analysis then simulates the 1:4-step ahead forecasts $(Y_{t+1}, \dots, Y_{t+4})$. Estimated forecast distributions and RMSFE measures are calculated based on these predictive particles, the results of which are summarized in Table 2.2. As expected, MCMC shows better prediction than all the sequential analyses, while both Shrinkage and LLTM achieve similar accuracy of prediction to MCMC in the sense of RMSFE. These results conform the validity and utility of the proposed sequential method in prediction.

In addition to predictive performance, for the three univariate submodels, it is of interest to consider ESS as a measure of particle degeneracy and the efficacy of

sampling. The time series and histograms of ESS in Figure 2.9 show that shrinkage/LLTM outperform vanilla SMC quite substantially. In fact, this is the main advantage of using emulators; the high ESS shows its efficiency in sampling and reduced particle degeneracy. Also, note that the particles of the DLM emulator seriously degenerate, indicated by very low ESS. This is because DLM includes all the parameters in the model, while as observed in Section 2.5.2, there is a high-degree of sparsity in LTM with our dataset, which leads to the discrepancy between the simple DLM emulator and the original LTM.

Though some emulators show higher ESS than vanilla SMC, the degeneracy is still seen in the period when ESS becomes below, for example, 50%. In practice, one can avoid this potential degeneracy by running off-line MCMC to refresh the particles and then come back to the on-line sequential analysis. ESS is again useful in detecting the timing of need for such intervention and refresh.

2.6 Summary Comments

In this chapter, to realize sequential analysis in LTMs, we introduce the emulating approach and proposed Shrinkage emulator as a good approximation of LTMs. By combining this idea with PL and APF, practically useful sequential analysis of posteriors in LTMs, including those of the fixed parameters, is fully possible. In the macroeconomic application, this method gives interpretable posterior summaries that have all the characteristics expected and are comparable in many respects to the “gold standard” results based on repeated (moving window) analysis using MCMC, an approach that is infeasible in problems of higher dimensions and high incoming data rates. The results of different emulators and vanilla SMC are compared by marginal likelihood, ESS and RMSFE, concluding that the usage of Shrinkage and LLTM emulators should be recommended in terms of model fitting, efficient sampling and accurate prediction.

Table 2.2: Multistep RMSFEs for 3 variables with MC errors.

	1-step ahead forecast				
	DLM	Shrinkage	LLTM	SMC	MCMC
p_{t+1}	1.242 (1.017)	0.298 (0.298)	0.278 (0.261)	0.294 (0.742)	0.264
u_{t+1}	2.677 (3.883)	0.558 (1.042)	0.342 (0.555)	2.550 (8.426)	0.308
r_{t+1}	3.253 (3.839)	0.721 (0.831)	0.602 (0.555)	1.986 (5.694)	0.477
	2-step				
	DLM	Shrinkage	LLTM	SMC	MCMC
p_{t+2}	2.164 (2.101)	0.518 (0.564)	0.439 (0.417)	0.475 (1.421)	0.393
u_{t+2}	4.990 (7.754)	0.885 (1.436)	0.624 (0.800)	1.193 (10.184)	0.539
r_{t+2}	2.833 (5.344)	1.214 (1.143)	1.067 (0.838)	1.650 (7.404)	0.839
	3-step				
	DLM	Shrinkage	LLTM	SMC	MCMC
p_{t+3}	2.071 (3.790)	0.711 (0.835)	0.574 (0.565)	0.820 (14.993)	0.552
u_{t+3}	5.221 (10.388)	1.249 (1.832)	0.918 (1.004)	6.391 (183.005)	0.840
r_{t+3}	4.779 (9.648)	1.668 (1.434)	1.521 (1.064)	4.926 (124.833)	1.147
	4-step				
	DLM	Shrinkage	LLTM	SMC	MCMC
p_{t+4}	2.555 (5.550)	0.881 (1.078)	0.717 (0.710)	1.831 (32.385)	0.726
u_{t+4}	8.278 (23.277)	1.427 (2.143)	1.182 (1.192)	12.872 (268.381)	1.121
r_{t+4}	5.665 (18.606)	2.137 (1.744)	1.881 (1.282)	9.949 (196.232)	1.431

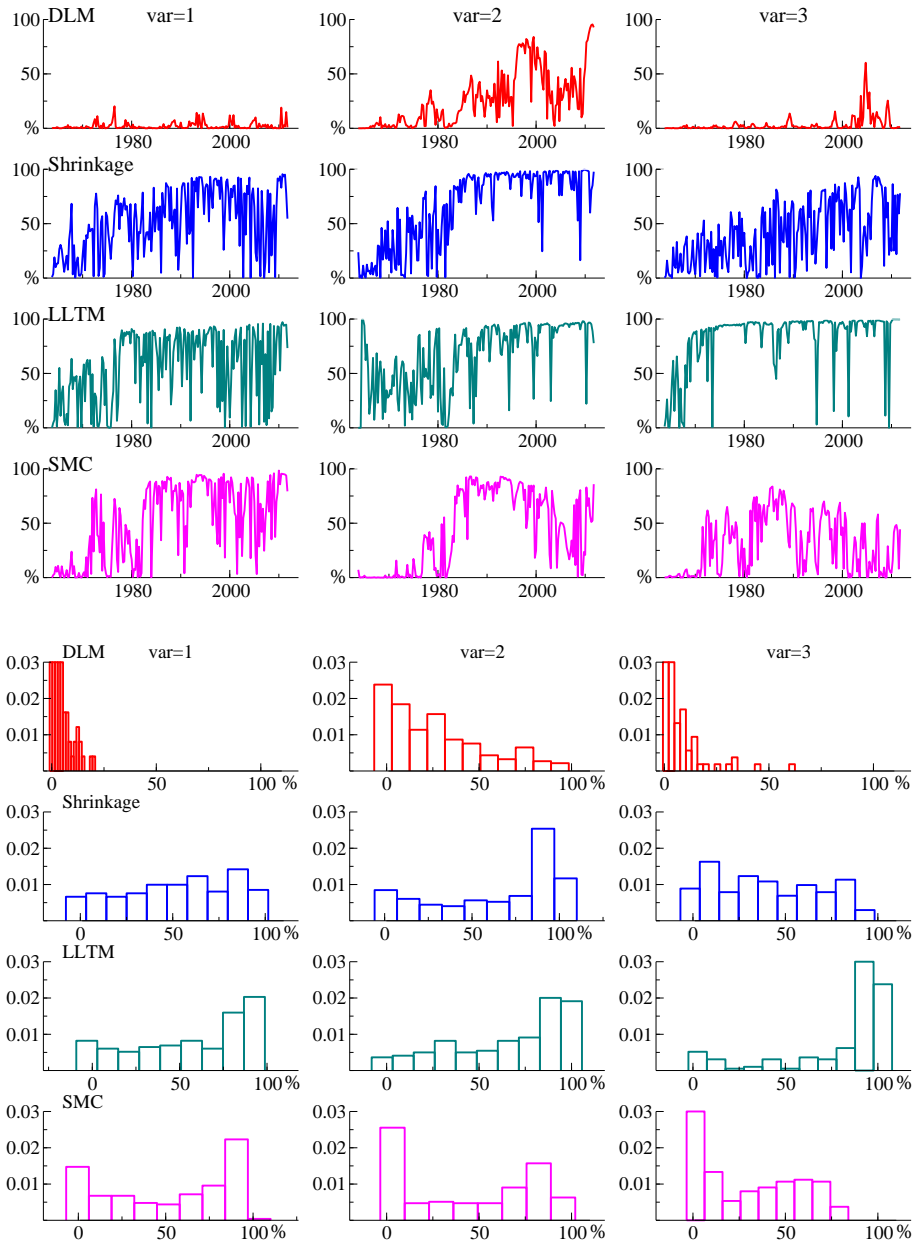


FIGURE 2.9: ESS results in analyses of DLM, Shrinkage, LLTM and vanilla SMC emulation methods. The ESS is scaled percentage, i.e., $100 \times ESS/N$, where N is the number of particles. The 3 columns represent results on each of the 3 univariate time series LTMs. For each, the upper 4 rows show time trajectories of the ESS measures, while the lower 4 rows show the resulting histograms of ESS measures aggregated over the time period.

2.7 Appendix: Technical Details of SMC for LTMs

2.7.1 Forward Filtering with Emulators

Throughout Section 4, it has been emphasized that forward filtering is available with emulators. In this section, the statement above is reviewed with the basics of forward filtering in DLMS (West and Harrison, 1997; Prado and West, 2010).

As specified in Section 4.1, any emulator of LTMs has the following form,

$$y_t = F_t' \beta_t + N(0, v_t), \quad F_t = x_t \circ q_t \quad (2.17)$$

$$\beta_t = \alpha + \Phi \beta_{t-1} + N(0, \Sigma), \quad (2.18)$$

where q_t is free from β_t but may be dependent on $(\beta_{t-1}, \alpha, \Phi, \Sigma, d)$. This lack of dependence of q_t on β_t makes the observational equation linear in β_t , so the emulator becomes a DLM. Then, the forward filtering for this model, conditional on $\{F_t, v_t, \alpha, \Phi, \Sigma\}$, is conducted as follows:

- Posterior at $t - 1$: $p(\beta_{t-1} | \mathcal{D}_{t-1}) = N(m_{t-1}, C_{t-1})$.

- Prior: $p(\beta_t | \mathcal{D}_{t-1}) = N(a_t, R_t)$,

where $a_t = \alpha + \Phi m_{t-1}$ and $R_t^i = \Phi C_{t-1} \Phi' + \Sigma$.

- Forecast: $p(y_t | \mathcal{D}_{t-1}) = N(f_t, q_t)$,

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

- Posterior: $p(\beta_t | \mathcal{D}_t) = N(m_t, C_t)$,

where $m_t = a_t + A_t e_t$, $C_t = R_t - A_t A_t' / q_t$, $e_t = y_t - f_t$ and $A_t = R_t F_t / q_t$.

In the sequential analysis of Section 2.4.4, the forecast and posterior distributions above become necessary; the former is used in evaluating the forecast density as the part of mixture weights and the latter is the distribution from which we sample the candidates of state variables β_t .

The computations above rely on two assumptions: the initial posterior is normally distributed and all the other parameters are given. Fortunately, when sampling β_t during the sequential learning from $t - 1$ to t , both assumptions are satisfied. To see this, remember that we start the sequential update by sampling auxiliary index i and the associated particle at $t - 1$. This fixes β_{t-1} to be β_{t-1}^i in the subsequent steps of sequential learning, meaning that β_{t-1} follows the degenerate normal distribution with $m_{t-1} = \beta_{t-1}^i$ and $C_{t-1} = 0$. Also, this auxiliary index i allows us to condition the distributions of interest on h_{t-1}^i , $\{\alpha_j^i, \phi_j^i, (\sigma_j^i)^2\}_{j=1:k,h}$ and d^i ; these variables define $\{F_t, v_t, \alpha, \Phi, \Sigma\}$ and enable us to focus on sampling of β_t and to apply the theory of filtering.

2.7.2 Particle Learning

The evolution model of eqns. (2.3,2.4) is a set of k conditionally independent AR(1) processes with transition p.d.f.s $N(\beta_{jt}|\alpha_j + \phi_j\beta_{j,t-1}, 1/w_j)$ for $j = 1:k$, where $\alpha_j = (1 - \phi_j)\mu_j$ and $w_j = 1/\sigma_j^2$. Consider first the traditional normal/gamma priors for (α_j, ϕ_j, w_j) assumed independent over j . Take these priors as

$$p(\alpha_j, \phi_j, w_j|\mathcal{D}_0) = N(\alpha_j, \phi_j|m_{j0}, C_{j0}/w_j) Ga(w_j|a_{j0}/2, b_{j0}/2)$$

for specified prior parameters $\mathcal{S}_0 = \{m_{j0}, C_{j0}, a_{j0}, b_{j0}; j = 1:k\}$. Standard theory shows that, for all $t > 0$ and maintaining independence across j ,

$$p(\alpha_j, \phi_j, w_j|\mathcal{D}_t) = p(\alpha_j, \phi_j, w_j|\mathcal{S}_t) = N(\alpha_j, \phi_j|m_{jt}, C_{jt}/w_j) Ga(w_j|a_{jt}/2, b_{jt}/2)$$

where the set $\mathcal{S}_t = \{m_{jt}, C_{jt}, a_{jt}, b_{jt}; j = 1:k\}$ is updated from \mathcal{S}_{t-1} via

$$\begin{aligned} m_{jt} &= m_{j,t-1} + A_{jt}e_{jt} & C_{jt} &= C_{j,t-1} - q_{jt}A_{jt}A'_{jt} \\ a_{jt} &= a_{j,t-1} + 1 & b_{jt} &= b_{j,t-1} + e_{jt}^2/q_{jt}, \end{aligned}$$

and

$$\begin{aligned} G_{jt} &= [1, \beta_{j,t-1}] & e_{jt} &= \beta_{jt} - G'_{jt}m_{j,t-1} \\ q_{jt} &= 1 + G'_{jt}C_{j,t-1}G_{jt} & A_{jt} &= C_{j,t-1}F_t/q_{jt}. \end{aligned}$$

The above analysis holds with a minor modification in cases when one or more of the priors adds a stationarity constraint (e.g. Prado and Lopes, 2013). That is, if the prior for any one ϕ_j is normal truncated to $(0, 1)$, then the posterior at any time t maintains the above form— based on the sufficient summaries in \mathcal{S}_t — but subject to the truncation. The conditional posteriors can still be easily sampled in this case.

2.7.3 Auxiliary Particle Filter

Unlike AR(1) parameters in the previous section, latent thresholds $d = (d_1, \dots, d_k)'$ have no sufficient statistics because of the complexity of the mean term in eqns. (2.1) and (2.2).

Instead of sampling from its full conditional based on particle learning, we estimate the marginal posterior density of d at time $t - 1$ by the particles, and generate the next particle from the estimated distribution. With the assumption of independence, the estimated density of d is written as

$$p(d|\mathcal{D}_{t-1}) = \sum_{i=1:N} w_{t-1}^i \prod_{j=1}^k G(d_j|a_{jt}^i, b_{jt}^i). \quad (2.19)$$

To construct this density, we need to estimate the univariate density of each threshold d_j with gamma kernels. The two parameters in the kernels, (a_{jt}^i, b_{jt}^i) , are chosen in the same way as in Liu and West (2001), in order for the mean and variance of the kernel to match those of those of sample analogue, m_j and V_j , computed by particles $\{d_j^i\}_{i=1:N}$ with the technique of shrinkage. Therefore, the kernel parameters are then determined by

$$a_{jt}^i = \frac{(\alpha d_j^i + (1 - \alpha)m_j)^2}{(1 - \alpha^2)V_j}, \quad b_{jt}^i = \frac{(\alpha d_j^i + (1 - \alpha)m_j)}{(1 - \alpha^2)V_j}, \quad (2.20)$$

where α is the tuning parameter for shrinkage and set to be 0.985.

The independence and distributional assumptions might be viewed as too restrictive, but the successful results of the empirical analysis in Section 2.5 suggest that these assumptions do not affect the estimation significantly. From the theoretical point of view, it can be seen that the dependencies among elements of d , which are ignored in the prior of d under the independence assumption, is also captured in mixture weights $w_{t-1|t}^i$. Also, the exponential decay in the tail of gamma density has already been observed in empirical studies in (Nakajima and West, 2013a) and is appropriate especially for modeling the distribution of latent thresholds.

2.7.4 Extension to Stochastic Volatility

The log-AR(1) stochastic volatility model in eqn. (2.8) introduces additional state variable $\{h_t\}_{t=1:T}$ and three parameters $\{\alpha_h, \phi_h, \sigma_h^2\}$ where $\alpha_h = (1 - \phi_h)\mu_h$. These parameters can be sampled by the methods discussed already in the previous sections. First, the sequential update of h_t is explained in Section 2.4.4; the new state variable is sampled from its prior $N(\alpha_h + \phi_h h_{t-1}, \sigma_h^2)$ conditional on $(h_{t-1}, \alpha_h, \phi_h, \sigma_h^2)$. Next, the AR(1) coefficient and variance, $(\alpha_h, \phi_h, \sigma_h^2)$, have the sufficient statistics and posterior conjugacy with normal-inverse gamma prior, similar to that for (α, Φ, Σ) in (2.7.2). Thus, the sampling procedure in (2.7.2) is valid for $(\alpha_h, \phi_h, \sigma_h^2)$ by replacing subscript j by h , and β_{jt} by h_t .

2.8 Appendix: Supplemental Figures on Parameter Learning

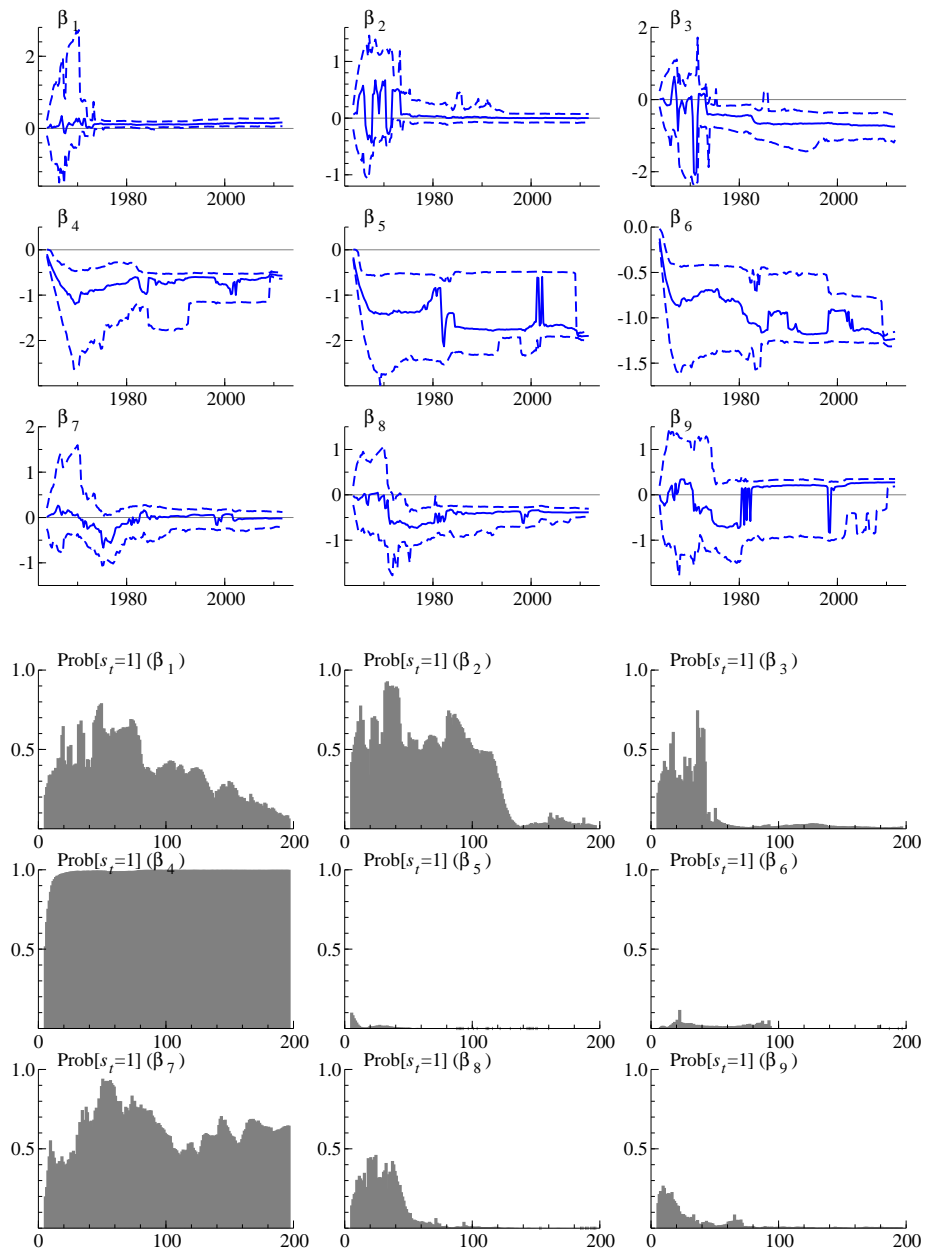


FIGURE 2.10: Time trajectories of posteriors for elements of B_{2t} and the corresponding inclusion probabilities.

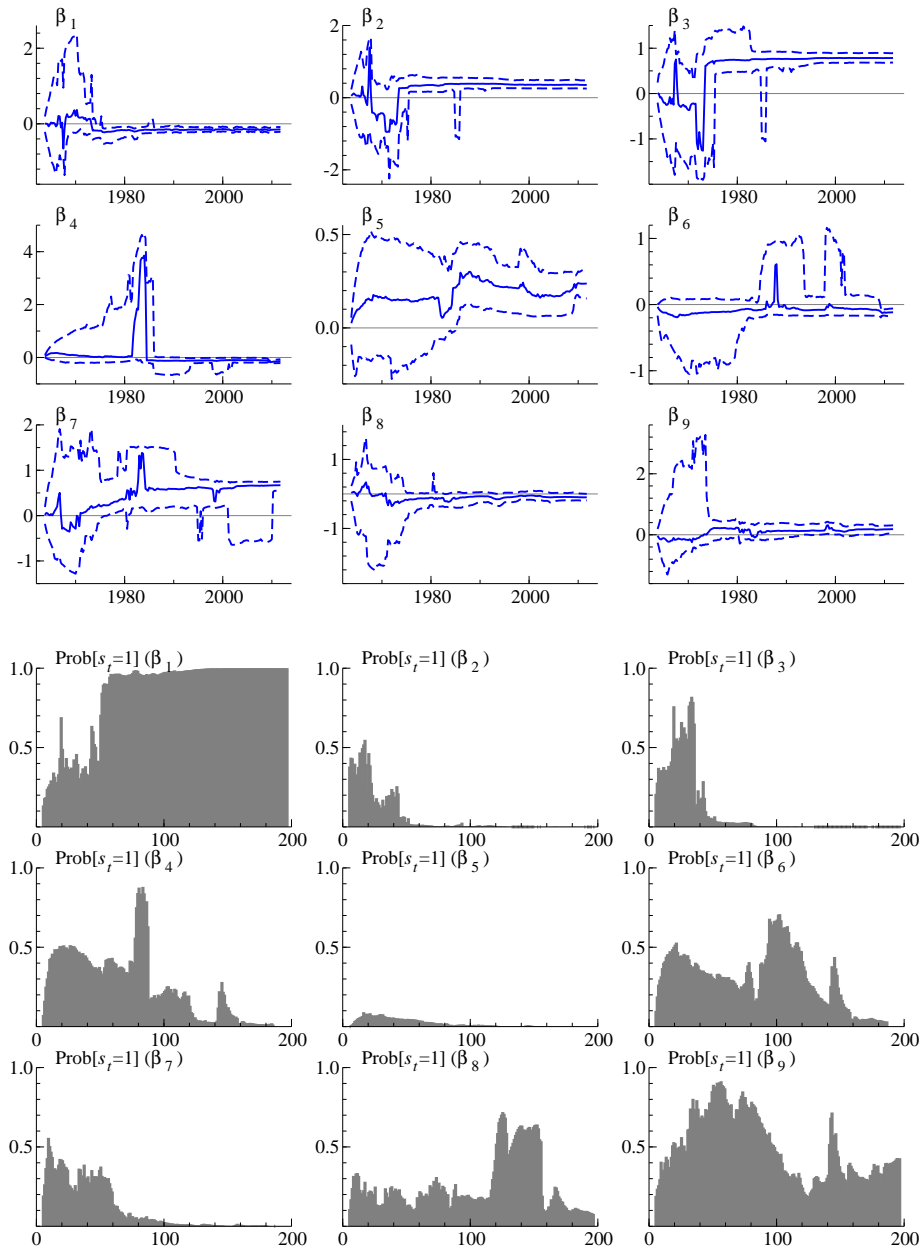


FIGURE 2.11: Time trajectories of posteriors for elements of B_{3t} and the corresponding inclusion probabilities.

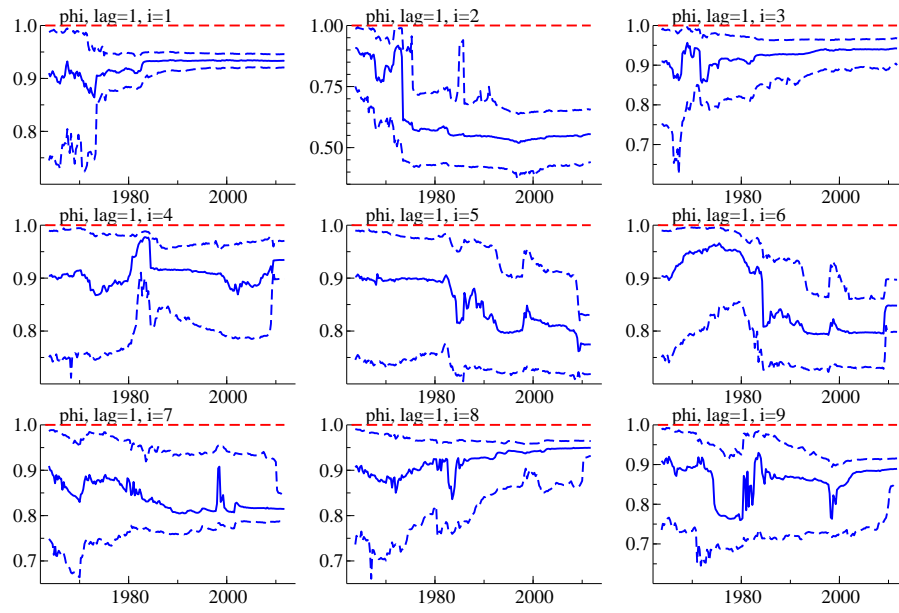


FIGURE 2.12: On-line posteriors of Φ for B_{1t} .

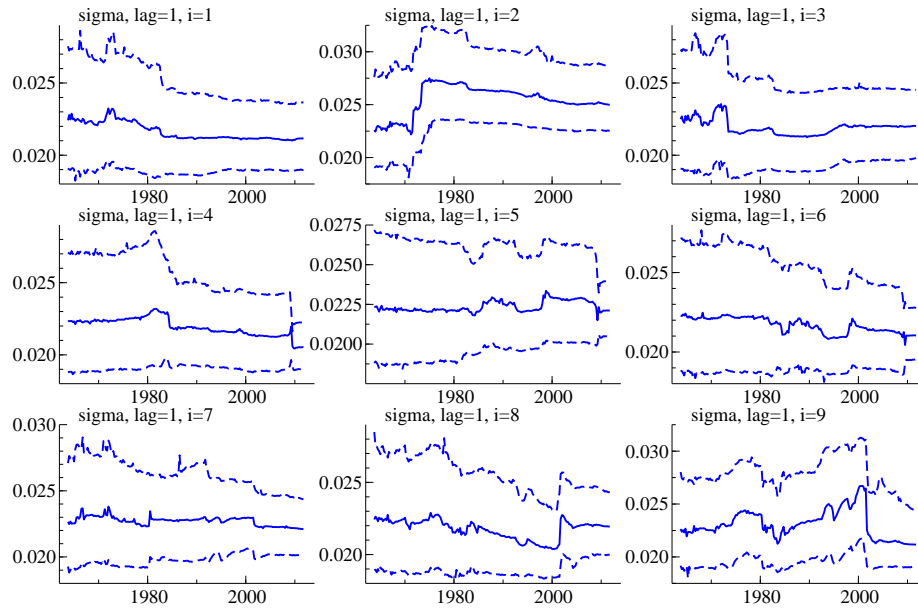


FIGURE 2.13: On-line posteriors of Σ for B_{1t} .

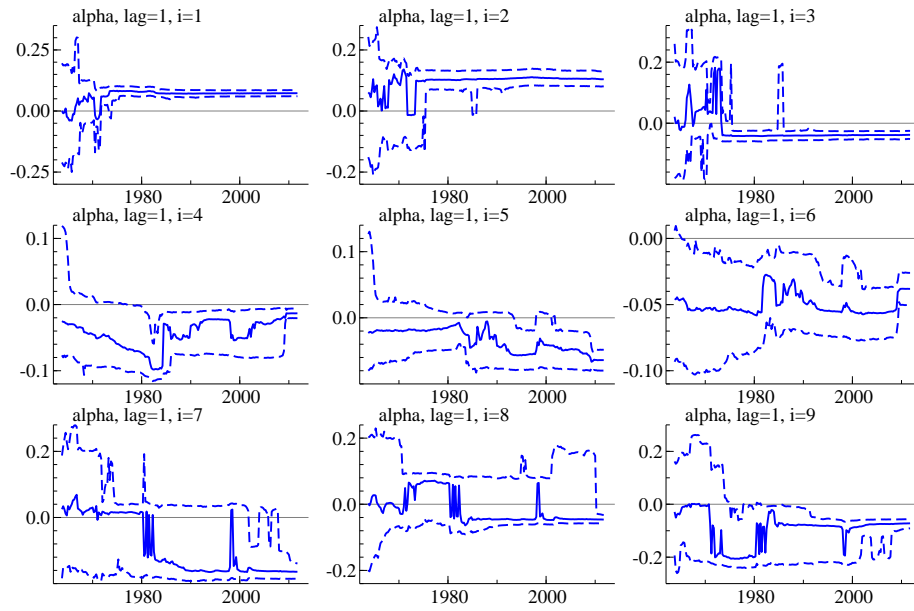


FIGURE 2.14: On-line posteriors of α for B_{1t} .

Bayesian Emulation in Decision Analysis: Sequential, Multi-Step Portfolio Optimization

3.1 Emulation by Synthetic Models in General Decision Analysis

The topic of this chapter is decision analysis after conducting posterior and predictive analysis. Given predictive information, we discuss loss (or utility) functions and methodology to compute the resulting optimal Bayesian decisions. The role of emulation arises here, in quite novel ways, as the resulting optimization problems can be cast as problems of searching for modes in posterior distributions in completely synthetic statistical/probabilistic models. That is, we propose Bayesian statistical model emulators for solving decision problems. For a certain class of loss functions $L(w)$ of decision w , there exists a corresponding probability density defined by

$$p(w) \propto \exp \left\{ -\frac{1}{2}L(w) \right\}, \quad \text{if } \int p(w)dw < \infty.$$

Therefore, the loss minimization problem is now equivalent to the computation of the mode of density $p(w)$. This simple idea is seen and well-known in, for example, ordinary least squares and maximum likelihood in the Gaussian linear model; minimizing

$L(\beta) = (y - X\beta)'(y - X\beta)$ in β is equivalent to computing the mode of a posterior distribution (with non-informative prior) defined by likelihood $y = X\beta + N(0, \sigma^2 I)$.

This classical idea has also been examined in detail in Bayesian decision theory and its application to complex decision making problems, as surveyed by Müller (1999). The following methodological development and application include Müller et al. (2004) and Amzal et al. (2006), in which the loss function, as the synthetic model, is integrated with statistical models for inference, hence the optimal decision can be simulated from a “synthetic” posterior using methods such as MCMC and simulated annealing.

This research takes the same approach and advantage in portfolio optimization problems. It can be regarded as a special case of the existing research cited above in the sense that the decision making process is independent of inference, so that the optimization can be conducted separately after the computation needed for posterior and predictive analysis of the model, as illustrated in Figure 3.2 in the next section. On the other hand, this simplification allows for a variety of choices in computational methodology for optimization such as analytical FFBS and EM methods, in addition to MCMC, exploiting the statistical features of the synthetic models. In addition to its computational advantages, this approach can also help decision makers to define the appropriate loss functions that match their personal or corporate preferences. In any specific decision problem, where one has to consider multiple aspects of preferences to be expressed in mathematical form, it is easier, at least for the applied statisticians, to propose statistical models that have the required mathematical properties. Figure 3.1 depicts these two advantages in using a synthetic model. The transformation of the original loss minimization problem to the analysis of synthetic models enables not only the computational methods for Bayesian time-series analysis but also the translation of well-known properties of the parallel, emulating statistical model into the decision analysis problem.

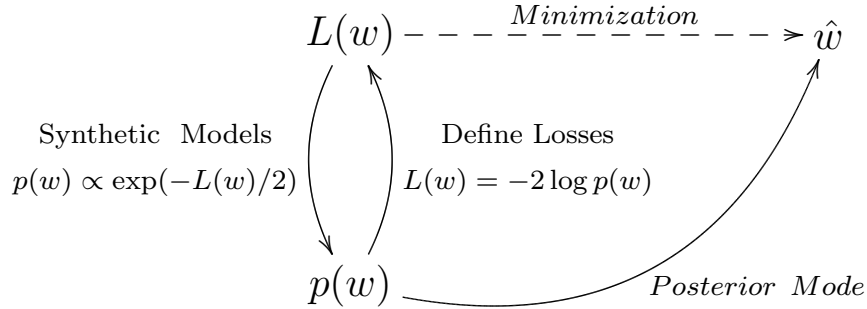


FIGURE 3.1: The statistical approach to loss minimization. There are two interpretations for this diagram. If the problem of interest has the specific loss function to be minimized, it can be transformed into the synthetic model, shown by the arrow from top to bottom, so that the computational method of statistical inference can solve the original problem. The other interpretation is that, for statisticians, the problem itself is defined as the statistical models, then transformed back into the form of optimization of an expected loss function, indicated by the arrow from bottom to top, in order to make use of the knowledge and techniques of statistical modeling.

3.2 Introduction: Sequential Portfolio Optimization

Portfolio optimization, where the investors update their portfolio based on their prediction of financial asset returns, has been an important problem in statistics and financial econometrics. This problem concerns both predictions by Bayesian statistical model(s) and decision making based on those predictions. On modeling and predictive analysis, recent advances in portfolio study emphasize the aspect of sequential analysis that discusses the computational feasibility in updating the current predictions by incorporating daily or more frequently observed stream of data into the models. Figure 3.2 shows the procedure of sequential portfolio optimization, conducted simultaneously with posterior analysis and prediction. To make a decision on portfolio choice in a timely manner, the preceding research of sequential analysis avoids the use of MCMC methods to save the computational time. The examples of MCMC-free modeling and computation include the use of massive numbers of analytically tractable DLMS to make predictions by model averaging technique (Zhou et al., 2014), advanced computational methodologies such as parallel computing by

GPUs (Gruber and West, 2016), and non-linear stochastic volatility model with SMC methods (Johannes et al., 2014).

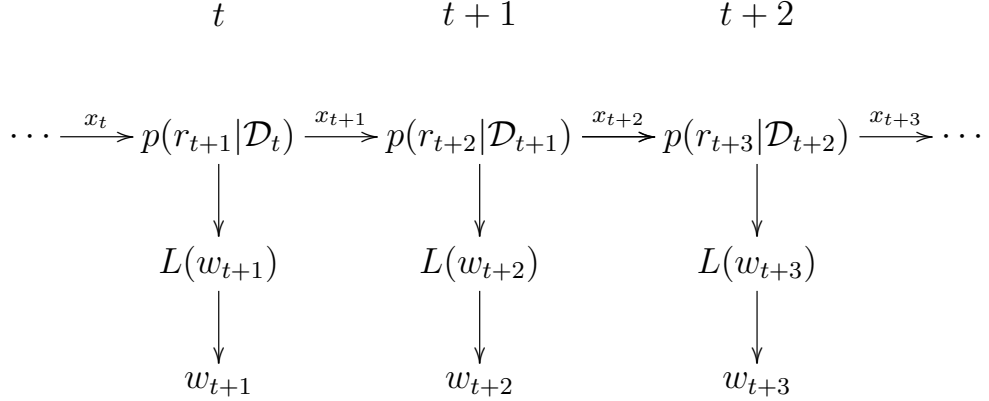


FIGURE 3.2: Sequence of portfolio optimization problems. At each time t , the posterior and forecast distribution is updated with additional information x_t . Based on the prediction of r_{t+1} (and r_{t+i} for $i \geq 1$), the loss function of w_{t+1} is defined. The output at time t is thus the portfolio for the next time, w_{t+1} . This process is repeated at every time point, yielding the sequence of portfolio vectors, $\{w_t\}_{t=1:T}$.

However, in comparison to modeling and forecasting, less attention has been paid in the statistical research literatures to the aspect of decision problem: the appropriate functional form of the loss function and the methodology for optimization. To this problem, the mean-variance optimization (Markowitz, 1952, 1968) is known as a simple but feasible and practical approach and has been applied very broadly for decades; in the Bayesian forecasting literature, most recent works include copious historical references (e.g. Zhou et al., 2014; Gruber and West, 2016). Denote the optimal portfolio at time $t = 1$, that is computed at $t = 0$, by the k -dimensional vector w_1 . This is defined as the minimizer of the following expected loss function of w_1 :

$$L(w_1) = w_1' K_1^{-1} w_1 \quad \text{subject to} \quad 1_k' w_1 = 1 \quad \text{and} \quad E[r_1' w_1] = m_1, \quad (3.1)$$

where K_1 is the predictive precision matrix ($k \times k$ -matrix), m_t is the target total return (scaler) and r_1 is the observed return (k -dimensional vector). Solving this

problem is not costly in most cases; the analytical form of the optimal portfolio is available and its numerical computation involves one inversion of $k \times k$ -matrix only.

In practice, however, the simple functional form in eqn. (3.1) might not describe precisely the personal preference on more profitable and stable portfolios. The more carefully the loss function is specified, the more profit and stability can be expected from the optimal portfolio as desired, which motivates further extension of this loss function. One crucial improvement of this loss function is to consider the multiple-step ahead predictions, rather than the one-step ahead prediction only. From this perspective, the loss function in eqn. (3.1) should include not only $w_1'K_1^{-1}w_1$ but also $w_t'K_t^{-1}w_t$ for $t = 2, \dots, h$ for some horizon h , which allows investors to weight their uncertainties and utilities about outcomes and risk over several, forthcoming time periods. Another improvement is motivated by the switching cost of portfolios, which we call transaction costs in this research, to smooth the dynamics of portfolios. Though there is theoretical and practical evidence that explicitly incorporating transaction costs in the loss function contributes to more profitable portfolios we rather consider this factor important as the realization of the psychological cost of individuals for the volatile portfolios. In the loss function, transaction costs are measured by distances between consecutive portfolios and can have a significant impact on the realized portfolios. While these two aspects of investment—multiple-step ahead predictions and transaction costs—can improve the quality of loss functions sufficiently for the practical application, in return for this improvement, the methodology to solve the resulting optimization problems are relatively under-developed.

In this research, we propose novel loss functions for the portfolio optimization problem based on multiple-steps ahead forecasting and transaction cost and, to solve it, offer a feasible and computationally efficient method. As stated in Section 3.1, we address the latter using novel Bayesian ideas to “translate” the optimization problem into an estimation problem in statistics. The resulting expected loss functions are,

in fact, the state space models in Bayesian time series analysis that are DLMS in the simplest case. This transformation also enables “modeling” the loss function with desirable characteristics exploiting insights generated from the statistical models. For these models, we can make use of the existing methods of computation, including analytical forward filtering and backward smoothing (FFBS), EM algorithm, MCMC, and their combination, to find the optimal portfolio weights as the posterior mode. These advances in modeling and computation are applied in Section 3.3 to portfolios that can be solved using computation by FFBS. Its extension to modified loss functions that include absolute distance metrics for transaction costs is discussed in Section 3.4, with the development of the tailored EM algorithm.

In addition to the above-mentioned modeling and computational advantages in practice, the statistical approach to the portfolio problem highlights classical but fundamental questions about definitions of loss functions in this context. Since we only need w_1 at time $t = 0$, the loss function should be dependent solely on w_1 , and it is obtained by minimizing the “joint” loss function of (w_1, w_2, \dots, w_h) , or “profiling out” redundant variables (w_2, \dots, w_h) . On the other hand, once the statistical approach is taken, it makes more sense for Bayesians to marginalize the redundant variables out in the synthetic model and transform the marginal model back to the form of loss function again. This “marginalization” and its connection to “profiling” is an important topic in statistical research today; Polson and Scott (2015) shows theoretical connection between the two approaches in shrinkage models, which is, unfortunately, not directly applicable to our portfolio problem. We, instead, propose methodology for computing the marginal model in Section 3.5, then discuss its difference from the joint model through the extensive application to real datasets in Section 3.6.

The performance of new loss functions introduced in this research, based on implementation using the Bayesian emulation concept and resulting synthetic statistical

models, is examined through application to foreign exchange rate forecasting and portfolio construction in Section 3.6. Cumulative returns of portfolios are monitored with and without discounts by transaction cost in order to evaluate the performance and sensitivity of different portfolio strategies. In the presence of transaction costs, the optimal portfolio strategy based on our loss functions consistently outperforms the mean-variance loss function with one-step forecasting. The marginalization approach is also applied and we discuss and compare the results with those achieved under the joint profiling method.

3.3 Statistical Model Emulation of Expected Loss Functions

3.3.1 Settings and Notation

The general sequential analysis consists of estimation, prediction and optimization; one receives a new observation at time t , processes this datum and updates his or her model (estimation), forecasts the future observations by using the updated model (prediction), optimizes his or her action at time $t + 1$ (optimization), and moves on to the next observation at time $t + 1$ (see Figure 3.2). In the context of portfolio analysis, the observation is the asset prices or returns and the decision made by investors is to change the portfolio allocation based on the prediction of returns at the next set of time points of interest. Suppose we have the model for the time series of asset prices p_t ($k \times 1$ vector, k is the number of assets of consideration), or return rates r_t defined by $r_{it} = p_{it}/p_{i(t-1)} - 1$. At time t , with the set of information updated as $\mathcal{D}_t = \{r_t\} \cup \mathcal{D}_{t-1}$, Bayesian analysis of this model provides the predictive means and variances: $f_h = E[r_{t+h}|\mathcal{D}_t]$ and $K_h^{-1} = V[r_{t+h}|\mathcal{D}_t]$. Note that both f_h and K_h^{-1} depend on t , but this subscript is abbreviated in the following. For simplicity, and without loss of generality, set the current time point to be $t = 0$ and abbreviate \mathcal{D}_0 in the conditional expectation; this means the same argument can be made at any t -th repetition of the sequential analysis by replacing subscript 0 by t , 1 by $t + 1$,

and so on. Also, note that the model for r_t and (f_h, K_h^{-1}) are assumed to be given for necessary t and h .

Denote the portfolio weights at time t by w_t . Now, the portfolio optimization problem is specified as the minimization of the expected loss function, $L(w_1)$. To incorporate the uncertainty about the future into the loss function appropriately, we define the joint loss function, $L(w_1, \dots, w_h)$, for some horizon h , which implicitly defines the target loss function $L(w_1)$ after minimization in (w_2, \dots, w_h) .

3.3.2 Dynamic Linear Models for Quadratic Loss

Next, we need to specify the functional form of the joint loss function. In this research, there are three (or four in Section 3.5) components which the loss function comprises; mean squared error (MSE), risk and transaction cost.

- MSE: $(m_t - f'_t w_t)^2$.

This is the mean squared error of predictive return $f'_t w_t$ against target return m_t . The target return, m_t , has to be provided by investors prior to the analysis to reflect their preference on the loss function. Relative to the other components in the loss function, the contribution of this MSE to the entire loss is determined by weight parameter α_t in eqn. (3.2). When $\alpha_t = 0$, it realizes the hard constraint on the target return, $f'_t w_t = m_t$, as in the traditional Markowitz approach. For this reason, we call this term a “soft constraint” on the target return, and control the strength of this constraint by weight α_t . An obvious drawback in using this soft constraint is the penalty on the excess return; the portfolio is penalized even if it achieves the target return, i.e. $f'_t w_t \geq m_t$.

- Risk: $w_t'K_t^{-1}w_t$.

Risk is the term used here for– explicitly– the variance of predictive return, i.e., $V[w_t'r_t]$. The corresponding weight is denoted by β_t .

- Transaction cost: $(w_t - w_{t-1})'(w_t - w_{t-1})$.

This represents the loss caused by the distance between w_t and w_{t-1} . It is weighted by λ_t .

The loss function we consider is the weighted sum of the three factors above, i.e.

$$L(w_{1:h}) = \sum_{t=1}^h \left\{ \alpha_t^{-1}(m_t - f_t'w_t)^2 + \beta_t^{-1}w_t'K_t^{-1}w_t + \lambda_t^{-1}(w_t - w_{t-1})'W_t^{-1}(w_t - w_{t-1}) \right\}, \quad (3.2)$$

where α_t , β_t and λ_t are also given prior to decision making in order to balance their impacts on the loss function. This quadratic functional form is motivated as the “expected loss function” of Bayesian decision theory (e.g. Berger, 1985, Section 1.3); the squared error of target/realized return, $(r_t'w_t - m_t)^2$, gives our loss function in eqn. (3.2) with $\alpha_t = \beta_t$ as the expectation in $p(r_{1:h}|\mathcal{D}_0)$.

Now, we claim that this has the corresponding “synthetic” model,

$$\begin{aligned} m_t &= f_t'w_t + N(0, \alpha_t), \\ z_t &= w_t + N(0, \beta_t K_t), \\ w_t &= w_{t-1} + N(0, \lambda_t W_t), \end{aligned} \quad (3.3)$$

which is a DLM with state variance $W_t = I_k$ and initial portfolio w_0 when observing $z_t = 0_k$. The synthetic DLM must have the initial prior, $w_0 \sim N(\mu_0, C_0)$, and this is automatically set as $C_0 = 0$ and $\mu_0 = w_0$ since w_0 is given in the process of sequential decision making. Denote the two densities of the time t likelihood and prior in this

model by $p(m_t, z_t|w_t)$ and $p(w_t|w_{t-1})$. The correspondence between the loss function in eqn. (3.2) and the model in eqn. (3.3) is now clear by observing that

$$\begin{aligned}
 e^{-\frac{1}{2}L(w_{1:h})} &\propto \prod_{t=1}^k p(m_t, z_t|w_t)p(w_t|w_{t-1}) \\
 &\propto p(w_{1:h}|m_{1:h}, z_{1:h}).
 \end{aligned}
 \tag{3.4}$$

Therefore, finding the minimizer of $L(w_{1:h})$ is equivalent to calculating the posterior mode in the synthetic DLM. After computing the posterior mode as $w_{1:h}$, take w_1 , the first sub-vector only, and simply ignore the rest. That is the solution of the original optimization problem.

Sum-to-One Constraint

In the synthetic model, the sum-to-one constraint is realized by the degenerate covariance matrix, $W_t = I_k - 1_k 1_k'/k$. This covariance matrix implies, almost surely, that $1_k'w_t = 1_k'w_0$ for all t . Consequently, as long as the initial portfolio satisfies the sum-to-one constraint, all the subsequent portfolios satisfy this constraint as well. In fact, conversely, conditioning the original synthetic model by $1_k'w_t = 1_k'w_{t-1}$ implies the degenerate covariance, so this approach is justified in the probabilistic sense. Thus, the optimal portfolio under this constraint is the posterior mode of the model defined by this covariance matrix.

3.3.3 FFBS for Posterior Modes

To compute the posterior mode of DLMs efficiently, the theory of FFBS (e.g. West and Harrison, 1997; Prado and West, 2010), plays a crucial role. With given hyperparameters, FFBS becomes analytic, with the backward smoothing step (rather than backward sampling or simulation) being analytic— as well as the forward filtering step. This leads to effectively trivial computation of the full, smoothed posterior distributions for the $w_{1:h}$ based on any synthetic data over times $1 : h$. Thus, we can

compute the joint posterior mean $w_{1:h}$ and simply note that, due to the joint normal structure, this implies the value of the posterior mode of $p(w_1|m_{1:h}, z_{1:h})$. The details of computation by FFBS and its efficacy are discussed in the appendix: Section 3.8.1. The use of FFBS in computation is more efficient than working directly on the quadratic loss function in eqn. (3.2) without any advanced theory of optimization. The closed form of the optimizer involves the inverse of $kh \times kh$ matrix that is the function of (K_t, f_t, m_t) , the direct computation of which becomes intense, for example, in the analysis of stock prices where k is large. We touch the details of computational efficiency in Section 3.8.3.

In application, the sum-to-one constraint causes the degeneracy of the variance matrix. This degeneracy can be inherited to other related matrices that have to be inverted in FFBS. To avoid this problem, we use the generalized inverse based on the singular value decomposition to compute the necessary inverse matrices. The resulted posterior mode is still valid in the synthetic model as the posterior mode conditional on the sum-to-one constraint.

3.4 Laplace State Space Models and Implied Loss Function

The quadratic loss function in Section 3.3 can smooth the dynamics of portfolios over time as desired, but it is still impossible for that loss function to have portfolios remain unchanged in some period of time either partially or wholly. In times of stable economy and asset prices, it is rather desirable that the optimal portfolio keeps its weights for some assets to be the same as those at the previous time, i.e. $w_{it} = w_{i,t-1}$ for some i . This property can be realized by adding penalty terms in the loss function or, in the Bayesian synthetic models, by using shrinkage priors on state variables.

Motivated by the ideas above, we modify the loss function as

$$L(w_{1:h}) = \sum_{t=1}^h \left\{ \alpha_t^{-1} (m_t - f_t' w_t)^2 + \beta_t^{-1} w_t' K_t^{-1} w_t + 2\lambda_t^{-1} 1_k' |w_t - w_{t-1}| \right\}. \quad (3.5)$$

The third term for transaction costs is now the sum of absolute changes of asset weights.

3.4.1 Synthetic Models and EM Algorithm

The synthetic state-space model is

$$m_t \sim N(f_t' w_t, \alpha_t), \quad (3.6)$$

$$z_t \sim N(w_t, \beta_t K_t), \quad (3.7)$$

$$w_{it} - w_{i(t-1)} : iid \sim L(\lambda_t^{-1}), \quad (3.8)$$

where $L(\lambda_t^{-1})$ means the Laplace distribution with parameter λ_t^{-1} , the density of which is given by

$$p(w_{it}|w_{i(t-1)}) = \frac{\lambda_t^{-1}}{2} e^{-\lambda_t^{-1} |w_{it} - w_{i(t-1)}|}.$$

In fact, this is the state-space version of a Bayesian lasso model (Park and Casella, 2008; Figueiredo, 2003), or might be interpreted as use of a fused lasso priors (Liu et al., 2014), in which the posterior modes of parameters (in our case, $w_{it} - w_{i,t-1}$) can exactly be zeros. This property of the model clearly addresses our preference on sparsity in portfolio switching.

In computation, note that the density of the state evolution has the form of a scale mixture of normals (Andrews and Mallows, 1974 and West, 1987) with mixing parameter τ_t as

$$p(w_t|w_{t-1}) = \int N(w_t|w_{t-1}, \text{diag}(\tau_t)) \prod_{i=1}^k G(\tau_{it}|1, \lambda_t^{-2}/2) d\tau_t,$$

so the augmented model is

$$\begin{aligned}
m_t &= f_t' w_t + N(0, \alpha_t), \\
z_t &= w_t + N(0, \beta_t K_t), \\
w_t &= w_{t-1} + N(0, W_t), \quad W_t = \text{diag}(\tau_t), \quad \tau_{it} : iid \sim G(1, \lambda_t^{-2}/2).
\end{aligned}
\tag{3.9}$$

Conditional on $\tau_{1:h}$, the model becomes a DLM, and FFBS for this conditional model gives the posterior mode of $p(w_{1:h}|\tau_{1:h}, m_{1:h}, z_{1:h})$. This is not the posterior mode of interest that should maximize $p(w_{1:h}|m_{1:h}, z_{1:h})$. To marginalize $\tau_{1:h}$ out, we can use the EM algorithm (Dempster et al., 1977) combined with FFBS at the maximization step. In this algorithm, instead of working on the original loss function in eqn. (3.5) and the non-Gaussian synthetic model in eqn. (3.6), we can use the augmented model in eqn. (3.9) that is analytically tractable in posterior analysis, and iteratively maximize the posterior of this model and update the latent variables. In other words, we first define the objective function in eqn. (3.2) with parametrization of $W_t = \text{diag}(\tau_t)$. $w_{1:h}$ is the control variables and $\tau_{1:h}$ is considered to be the latent parameters. Note that this is the exponential part of the augmented model in eqn. (3.9). EM methods claims that the iterative minimization of the expectation of this objective function ensure the convergence of the sequence of solutions to the minimizer of the original loss function. The computation by this algorithm proceeds as follows:

EM Algorithm for Laplace Loss/Model

1. (Initialization): Set $w_t^{(0)}$ arbitrarily. We recommend to use the solution of the quadratic loss function optimization in Section 3.3, as it is reliable as the approximation of the target solution while being easy to compute.
2. For $s = 1:S$, repeat the following two steps. At s -th iteration,
 - (a) (Expectation): Replace τ_t in $W_t = \text{diag}(\tau_t)$ of the objective function in

eqn. (3.2) by

$$\tau_{it}^{(s)} = \lambda_t^2 \left| w_{it}^{(s-1)} - w_{i(t-1)}^{(s-1)} \right|,$$

and define $W_t^{(s)} = \text{diag} \left(\tau_t^{(s)} \right)$.

This computation comes from the conditional expectation of the objective function or, essentially, $E \left[W_t^{-1} \mid y_{1:h}, w_{1:h}^{(s-1)} \right]$.

- (b) (Optimization): Implement FFBS for Model in eqn. (3.9) to find the optimizer $w_t^{(s)}$ with replacing covariance matrix by $W_t^{(s)}$.

This is equivalent to solving the objective function in eqn. (3.2) with $W_t = W_t^{(s)}$.

3. For a sufficiently large S , we can use $w_{1:h}^{(S)}$ as the approximate posterior mode of $p(w_{1:h} \mid y_{1:h})$.

Note that the algorithm above does not give exact zeros, i.e. $w_{it} = w_{i,t-1}$, though the value of w_{it} can be very similar to that of $w_{i,t-1}$ numerically. One can have an additional step at each iteration to set $w_{it} = w_{i,t-1}$ for some i if they are sufficiently close to one another. See Section 3.9 for details.

Sum-to-One Constraint in Laplace Models

It is worth exploring the degenerate multivariate lasso model to have the built-in sum-to-one constraint in the synthetic model. However, to the extent we know, there has been no research on such models. For example, Eltoft et al. (2006) develops the multivariate Laplace distribution with one mixing random variable, but limited to the non-degenerate case.

To impose the sum-to-one constraint on the optimizer, we take the same approach in Section 3.3.2 for the augmented model by replacing the covariance matrix in the

synthetic model by $W_t = \text{diag}(\tau_t) - \tau_t \tau_t' / \mathbf{1}'_k \tau_t$. As in the Gaussian case, this degenerate covariance matrix implies $\mathbf{1}'_k w_t = \mathbf{1}'_k w_0$ for all t almost surely.

3.4.2 Another Laplace Factor for Non-Negativity Constraint

In the practice of personal investments, it is sometimes of interest to take long positions only, i.e. to have portfolio weights always non-negative. The hard constraint on this non-negativity condition is directly addressed by adding kh inequalities, i.e. $w_{jt} \geq 0$ for $j = 1:k$ and $t = 1:h$, which complicates the computation if k is large. To promote computational simplicity, a partial, soft constraint is considered here with the introduction of an additional absolute term, the sum of $|w_{jt}|$, which is originally intended for another shrinkage of weights toward zeros. This additional shrinkage penalizes portfolios with negative weights only under the sum-to-one constraint. To see this, note that $\mathbf{1}'_k |w_t| > 1$ if the portfolio has negative weights, while $\mathbf{1}'_k |w_t| = 1$ if all the weights are non-negative. Again, this does not always guarantee the non-negativity condition theoretically, but, in practice, it is rare that the non-negativity constraint is violated under this specification. We see this in Section 3.6 with application.

The new loss function based on this idea is

$$L(w_{1:h}) = \sum_{t=1}^h \left\{ \alpha_t^{-1} (m_t - f_t' w_t)^2 + \beta_t^{-1} w_t' K_t^{-1} w_t + 2\gamma_t^{-1} \mathbf{1}'_k |w_t| + 2\lambda_t^{-1} \mathbf{1}'_k |w_t - w_{t-1}| \right\},$$

with additional tuning parameter γ_t . The synthetic model with augmentation is

$$\begin{aligned} m_t &= f_t' w_t + N(0, \alpha_t), \\ z_t &= w_t + N(0, \beta_t K_t), \\ u_t &= w_t + N(0, \Phi_t), \quad \Phi_t = \text{diag}(\phi_t), \quad \phi_{it} : iid \sim G(1, \gamma_t^{-2}/2), \\ w_t &= w_{t-1} + N(0, W_t), \quad W_t = \text{diag}(\tau_t), \quad \tau_{it} : iid \sim G(1, \lambda_t^{-2}/2), \end{aligned} \tag{3.10}$$

with phantom observations, $z_t = u_t = 0$, and another mixing parameter ϕ_t for $|w_t|$, where $\tau_{1:h}$ and $\phi_{1:h}$ are mutually independent. For this model, the EM algorithm introduced in Section 3.4.1 is still applicable with additional expectation step on ϕ_t . The algorithm is modified as follows:

EM Algorithm for Full Laplace Loss/Model

1. (Initialization): Set $w_t^{(0)}$ arbitrarily.
2. For $s = 1:S$, repeat the following three steps.
 - (a) (Expectation 1): Update τ_t in the objective function by

$$\tau_{it}^{(s+1)} = \lambda_t^2 \left| w_{it}^{(s)} - w_{i(t-1)}^{(s)} \right|,$$

and replace W_t in the synthetic model by $W_t^{(s+1)} = \text{diag} \left(\tau_t^{(s+1)} \right)$.

- (b) (Expectation 2): Update ϕ_t in the objective function by

$$\phi_{it}^{(s+1)} = \gamma_t^2 \left| w_{it}^{(s)} \right|,$$

and replace Φ_t in the synthetic model $\Phi_t^{(s+1)} = \text{diag} \left(\phi_t^{(s+1)} \right)$.

- (c) (Optimization): Implement FFBS for the model in eqn. (3.10) to find the optimizer $w_t^{(s+1)}$.

3. $w_{1:h}^{(S)}$ are the estimate of posterior mode of $p(w_{1:h} | m_{1:h}, z_{1:h}, u_{1:h})$.

Similarly to the EM method for the former Laplace model, the original algorithm yields neither $w_{it} = 0$ nor $w_{it} = w_{i,t-1}$. To realize this exact sparsity in portfolio switching, we can take the additional step at the end of each iteration, as discussed in details in Section 3.9.

3.5 Marginalization of Loss Functions

3.5.1 Joint and Marginal Loss Functions

Under the setting of sequential portfolio optimization, the decision problem at time $t = 0$ concerns only the portfolio weights at the next time point, i.e. w_1 . The other weights observed in the further future, $w_{2:h}$, do not have to be derived at $t = 0$ but later. For example, we can obtain desirable w_2 with more information at time $t = 1$ when observing new prices r_1 and updating our predictive distribution. Likewise, we should decide w_3 , knowing r_2 at $t = 2$, and continue this process sequentially. Put differently from the viewpoint of decision theory, it is proper to set up and work on the loss function of w_1 that is independent of $w_{2:h}$. In contrast, the joint loss function, $L(w_{1:h})$, is useful and easy to be specified in application to incorporate the multiple-step ahead predictions into the decision making process. In Section 3.3 and 3.4, the joint loss function is directly minimized in $w_{1:h}$, then w_1 is taken out from the entire vector $w_{1:h}$, ignoring the rest $w_{2:h}$. This is justified as the minimizer of the following proper loss function,

$$L(w_1) = \min_{w_{2:h}} L(w_{1:h}). \quad (3.11)$$

Once the loss function is converted to the synthetic model, the definition of the loss function above is equivalent to profiling the joint model $p(w_{1:h})$ to obtain the optimal w_1 . However, this is not the only way to obtain the optimizing w_1 from the joint loss function. From this viewpoint of synthetic models, it makes more sense for Bayesians (e.g. Polson and Scott, 2015) to marginalize out the nuisance variables to define the marginal synthetic model, $p(w_1)$, then induce the loss function from this model. Given the joint loss function $L(w_{1:h})$, the following steps define the resulting marginal loss function $L^*(w_1)$:

1. Find the corresponding statistical model, i.e. likelihood $p(y_t|w_t)$ and prior

$p(w_t|w_{t-1})$, from the relation

$$p(w_{1:h}|y_{1:h}) \propto \prod_{t=1}^h p(y_t|w_t)p(w_t|w_{t-1}) \propto \exp \left\{ -\frac{1}{2}L(w_{1:h}) \right\},$$

with the appropriate definition of y_t (e.g., in Section 3.4, the Laplace models set $y_t = \{m_t, z_t\}$ with $z_t = 0$) and other necessary variables.

2. Take the marginal posterior as

$$p(w_1|y_{1:h}) = \int p(w_{1:h}|y_{1:h})dw_{2:h}.$$

3. Define the loss function by

$$L^*(w_1) = -2 \log \{p(w_1|y_{1:h})\}.$$

We denote the marginal loss function by $L^*(\cdot)$ to distinguish it from $L(\cdot)$ in eqn. (3.11) that is based on profiling of the joint loss function.

The objective of this section is to provide the methodologies for minimization of marginal loss functions, in order to address the difference between the two approaches: the profiled loss function $L(w_1)$ and marginal loss function $L^*(w_1)$.

Even if the profiled and marginal loss functions share the common joint loss function in their definitions, their solutions, \hat{w}_1 and w_1^* , could be different in general. One significant exception is the case of the quadratic loss functions and synthetic DLMs in Section 3.3. If the synthetic model defined in step 2 is a DLM as in eqn. (3.9) (or, if the joint loss function is quadratic as in eqn. (3.2)), the optimal portfolio obtained from the marginal loss function is, in fact, exactly the same as that of the profiled one, since the joint posterior of state variables in the emulating synthetic DLM is known to be the normal distribution and the marginal mean/mode of a normal distribution equals the corresponding element of the joint mean/mode.

Thus, to see the difference between the marginal and profiled loss functions, we need to examine non-Gaussian emulating synthetic models, and the Laplace-type loss function discussed in Section 3.4 is suitable for this consideration.

3.5.2 Marginal Laplace Loss Function and Mode Searching

Consider the Laplace joint loss function in eqn. (3.5) or the statistical model in eqn. (3.9) with the sum-to-one constraint. The density function to be maximized is not the joint posterior, $p(w_{1:h}|\mathcal{D}_h)$, but the marginal one, $p(w_1|\mathcal{D}_h)$, where $\mathcal{D}_h = \{m_{1:h}, z_{1:h}\}$. The parameters in the augmented synthetic model before marginalization are grouped naturally into two groups: control variables w_1 and nuisance parameters $(w_{2:h}, \tau_{1:h})$. We discuss the computational procedure to integrate out the latter and maximize the marginal loss function of the former.

Unfortunately, unlike the minimization of the joint loss function in Section 3.4.1, the direct application of EM methods for the current problem is not straightforward. To see this difficulty, note that the original EM idea consists of the E-step and the M-step. The M-step is easy to be implemented even in the marginal case, since the expected objective function is quadratic in w_1 . However, in the E-step we are required to, but unable to, derive the analytical expression of the expected logarithm of the loss function in terms of $p(w_{2:h}, \tau_{1:h}|w_1, \mathcal{D}_h)$, where w_1 is the tentative solution at an iteration of the algorithm. This conditional distribution is not well known, while we need the moments of this distribution, such as $E[W_2^{-1}]$ and $E[W_2^{-1}w_2]$.

To solve or avoid this problem, there are several approaches based on the existing literature, including the extension of EM methods and the direct approximation of the target density. For example, the complex expectations in the E-step can be replaced by simulation-based approximation so that we are still able to use the EM method. We have explored and evaluated a range of methods, including rejection sampling, MCMC and mean-field approximation (Ghahramani, 1995), but they all

suffer from numerical bias and computational inefficiency. Other than EM methods, as the approximating technique for the posterior density, integrated nested Laplace approximation (INLA, Rue et al., 2009) is widely known for its accuracy of approximation in contexts that provide some analytic structure in terms of conditional Gaussianity. This method is, unfortunately, computationally infeasible in our case because of the curse of dimensionality. Yet, the analytical form of the model allows for the easy implementation of MCMC, and we utilize it to define direct numerical/analytic approximation of the marginal posterior.

The target, marginal density is approximated by its sample analogue

$$\hat{p}(w_1|\mathcal{D}_h) = \frac{1}{S} \sum_{i=1}^S p\left(w_1 \mid \tau_{1:h}^{(i)}, \mathcal{D}_h\right), \quad (3.12)$$

with S particles, where $\{\tau_{1:h}^{(i)}\}_{i=1:S}$ are sampled by MCMC with $\{w_{1:h}^{(i)}\}_{i=1:S}$. The details of this MCMC and the diagnosis of its convergence in the real-data application in Section 3.6 are discussed in Section 3.10. Most importantly, the density in the right-hand-side is the mixture of normals; the conditional density $p\left(w_1 \mid \tau_{1:h}^{(i)}, \mathcal{D}_h\right)$ is the retrospective marginal posterior distribution of the corresponding emulating DLM, i.e., the normal distribution with known mean and variance computed by FFBS. Thus, the problem reduces to the optimization of the mixture of normal densities. We propose to use a simple updating rule based on the first order derivative of the target density. With a reliable initial value, the solution is iteratively updated by the first order condition to converge to the optimal point. The drawback of this method is the speed of convergence, so one has to choose an initial value which is sufficiently close to the actual solution. The candidates for such an appropriate initial value include sampled w_{1t} , the means of the mixture components and the previous optimal portfolio. For the details of the mixture representation, its optimization and MCMC, see Section 3.10.

3.6 Application: FX Commodity Dataset

3.6.1 Dataset

The proposed loss functions, portfolio strategies and computational methodologies are applied to dynamic portfolio choice in analysis and sequential forecasting of daily financial returns time series focused on a series of 10 FX series, 2 US stock market index series, and 2 commodity series, over a long recent period of years. Table 3.1 shows the list of the thirteen variables ($k = 13$) used in this analysis. The series of prices is recorded daily from August 8, 2000, to December 30, 2011. The whole time series is used for the sequential learning of the model, while the prediction and portfolio optimization starts from January 1, 2009 ($T = 761$).

Table 3.1: List of currencies, commodities and indeces.

Names	Acronyms	Names	Acronyms
Australian Dollar	AUD	Swiss Franc	CHF
Euro	EUR	British Pound	GBP
Japanese Yen	JPY	New Zealand Dollar	NZD
Canadian Dollar	CAD	Norwegian Kroner	NOK
South African Rand	ZAR	Oil price	OIL
Gold	GLD	Nasdaq index	NSD
S&P index	S&P		

3.6.2 Models for Prediction

The model used for this analysis is a TV-VAR(2) (e.g. Primiceri, 2005; Nakajima, 2011) with the covariance modeled by a dynamic dependence network structure (DDN, Zhao et al., 2016) to integrate the conditional independence relationships of assets— and their dynamics over time— into the model. Prior to the main posterior and predictive analyses, we take the first 500 observations as the training dataset to estimate the simultaneous correlations of returns with the TV-VAR model with the full covariance matrix. This preliminary analysis determines the conditional

independence structure assumed in the main analysis by ignoring the insignificant correlations.

In computation, the multivariate TV-VAR model is decomposed into univariate sub-DLMs that are processed independently and in parallel. The DDN framework couples these parallel univariate models to define a flexible global but dynamic model representing time-varying multivariate volatility pattern across the series. Analytical forward filtering can be applied to each submodel to obtain its on-line posterior and predictive distribution. Then, the multiple-step ahead forecast means and variances, $\{f_{t+i}, K_{t+i}^{-1}\}_{i=1:h}$, are computed by Monte Carlo, simulating many particles from these predictive distributions and coupling these samples across series to define coherent multivariate forecast. The number of particles used in this simulation-based method is 50000 at each t . The more detailed description of the model, settings and methods of forecasting is given in Section 3.11.

Based on the predictive information, the loss function is set up and the optimization problems are solved in the ways introduced in Section 3.3, 3.4 and 3.5, in addition to the mean-variance optimization in eqn. (3.1). We evaluate progressively revised cumulative returns for the comparison of different portfolios/loss functions.

3.6.3 Evaluation by Cumulative Returns

For weights $w_{1:s}$, realized returns $r_{1:s}$ and transaction cost δ , we define cumulative return R_s at time s by

$$R_s = \prod_{t=1}^s \left[(r_t + 1_k)' w_t - \delta 1_k' |w_t - w_{t-1}| \right] - 1. \quad (3.13)$$

In the following examples, we compute the resulting series of cumulative returns with and without transaction costs; $\delta = 0$ and 0.001, respectively. It is expected by definition that the Markowitz approach yields more cumulative return with no transaction cost, while the portfolios with λ_t works better with nonzero δ .

3.6.4 Results

In the following, $h = 5$; up to 5-day-ahead predictions are considered. The values of the tuning parameters are changed to see their effect on portfolios. We first fix $\alpha_t = 100$, $\beta_t = 1$, $\lambda_t = 100$ and $\gamma_t = 100$ for all t , and then change one of them. The target returns are always set at $m_t = 0.0005$ for all t . Approximately, daily 0.05% returns for a year (261 days) lead to 13.9% expected annual return.

Note that, the larger λ_t is, the less weight we put on the transaction term. As λ_t becomes larger, we expect more volatile portfolios with less penalty from transaction costs, and vice versa.

Gaussian Models

Figure 3.3 compares the portfolio weight vectors of the quadratic loss functions in eqn. (3.3) with different tuning parameters, computed by FFBS as discussed in Section 3.3, and the Markowitz-type portfolio obtained by the mean-variance optimization defined in eqn. (3.1). From the comparison with the Markowitz portfolio, it is clear that the transaction cost in our loss function significantly affects the smoothness of portfolio dynamics. We also confirm that large λ_t makes the portfolio be more volatile and imitate the Markowitz portfolio.

In Figure 3.4, the cumulative returns from the portfolios in Figure 3.3 are shown for different transaction costs. Again, as expected in Section 3.6.3, the Markowitz-type method is totally outperformed by our portfolios in the presence of transaction costs. Obviously, the smaller λ_t is, the more robust to transaction costs the cumulative return becomes, as seen in the downward shift of the cumulative return curve in the result of $\lambda_t = 10000$. One interesting finding here is that, during 2009, or right after the initial phase of the economic crisis, the Markowitz rule performs as well as our portfolios do. However, after that period, especially during late 2010 and early 2011, the loss functions with moderate λ_t yield more profit by tracking the

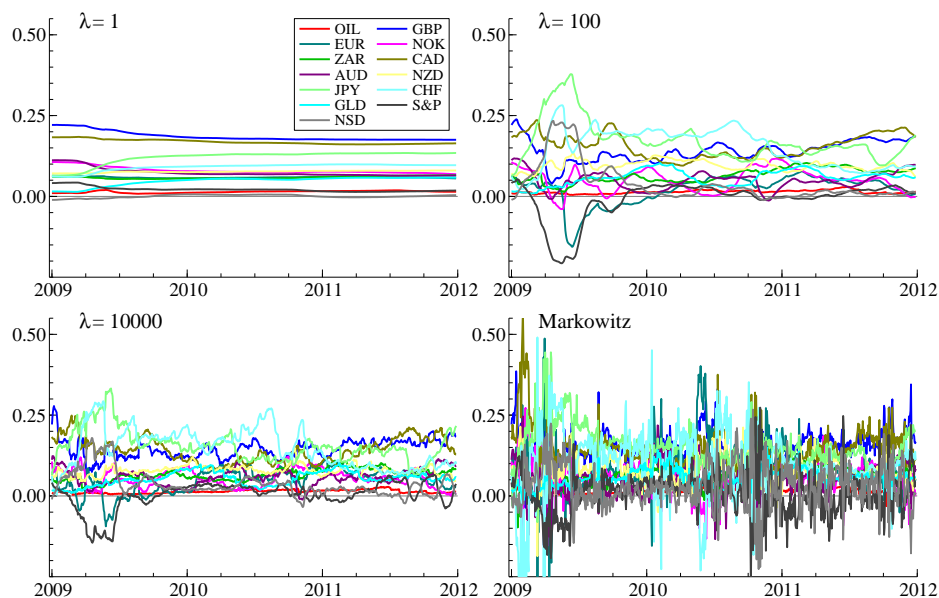


FIGURE 3.3: Optimal portfolios of DLM loss function and Markowitz method. From left to right, top to bottom: portfolio weights of a quadratic loss function with $\lambda_t = 1, 100$ and 10000 with $(\alpha_t, \beta_t) = (100, 1)$ and Markowitz-type portfolio.

ideal allocation with less switching, saving not only the transaction costs but also the variance of portfolios.

Figure 3.5 shows the realized standard deviation of portfolios, i.e. $\sqrt{w_t' K_t^{-1} w_t}$. The standard deviation of the minimum risk portfolio is defined by $(1_k' K_t^{-1} 1_k)^{-1/2}$ and this is, in fact, the lower bound of standard deviations of the other portfolios. The portfolios with less penalty on switching ($\lambda_t = 10000$) have almost as small standard deviations as the minimum risk portfolio. The other portfolios, such as $\lambda_t = 100$, have smaller standard deviation in most period than that of the Markowitz method. This means that we do not have to inflate the risk of our portfolios in return to the smoothness in their dynamics and robustness to transaction costs. While large λ_t leads to small portfolio risk in general, interestingly, the relation between the parameter λ_t and the risk is not linear; we see a period in the middle of 2009 in which the standard deviation of $\lambda_t = 100$ is larger than that of $\lambda_t = 1$. It is worth

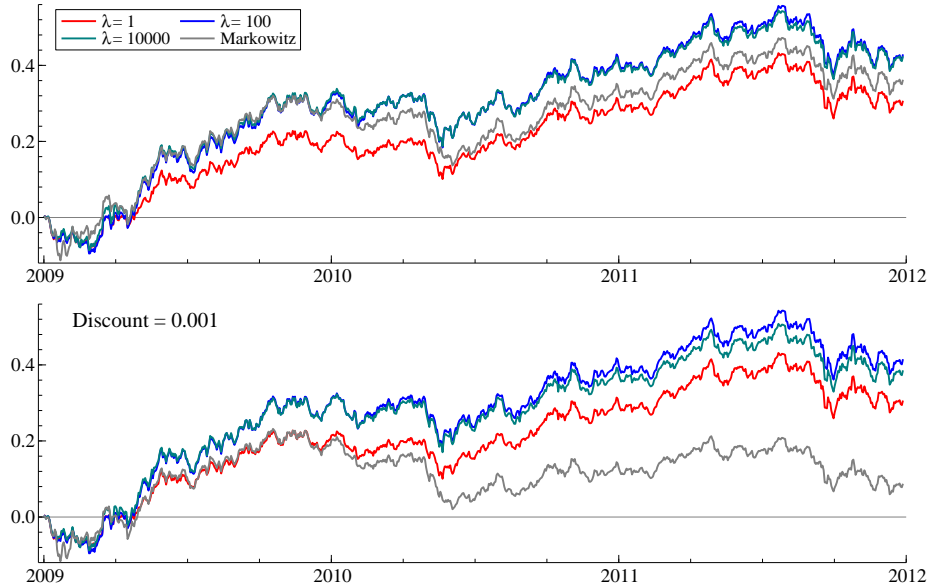


FIGURE 3.4: Cumulative returns of DLM and Markowitz portfolios. Top: cumulative returns with no transaction cost. Bottom: with 0.1% transaction cost. Four cumulative returns of DLM portfolios with $\lambda_t = 1$ (red), 100 (blue), 10000 (green) and Markowitz portfolio (grey) are shown in both pictures.

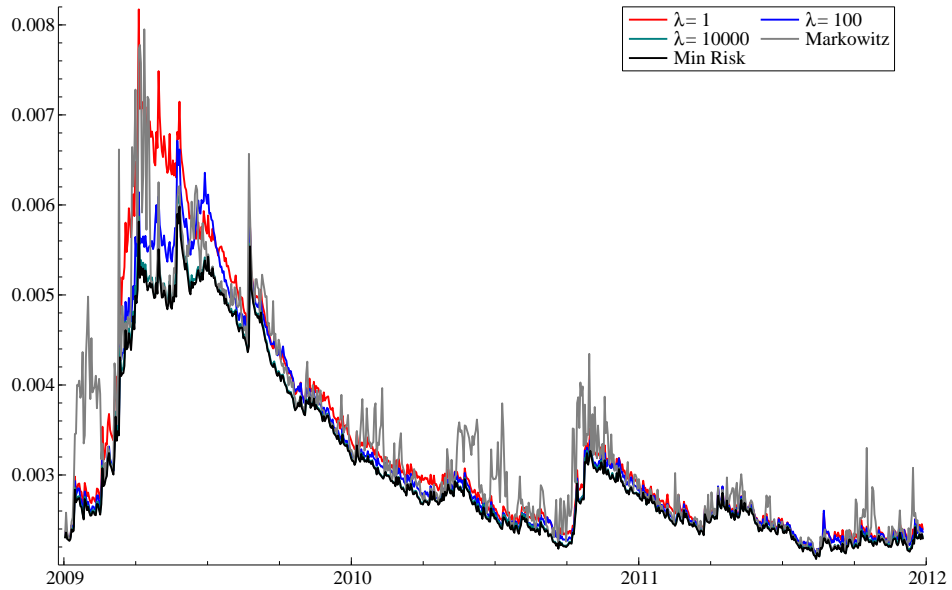


FIGURE 3.5: Standard deviations of DLM and Markowitz portfolios. Four standard deviations of DLM portfolios of $\lambda_t = 1$ (red), 100 (blue), 10000 (green) and Markowitz portfolio (grey) are shown in both pictures with that of minimum risk portfolio (black).

mentioning that a portfolio becomes almost constant over time if transaction cost is heavily imposed as in $\lambda_t = 1$, which might explain why its standard deviance can be temporally small.

Laplace Models

In this section, we see the performance of the full Laplace loss function in Section 3.4.2, which involves shrinkage of weights represented by $1'_k|w_t|$. The Laplace loss function in Section 3.4.1 is presented in the next subsection with its marginal equivalent discussed in Section 3.5.

Figure 3.6 shows the optimal weights obtained from Laplace loss functions with different tuning parameters. As expected, we see from this figure the two shrinkage effects on the difference of consecutive weights and weights themselves. First, the hard shrinkage of weights difference leads to less switching in portfolio allocation over time, seen in the figure as the stepwise increase and decrease of portfolio weights. These zero changes are also observed even in the case of larger λ_t , where the portfolio becomes volatile and similar to that of Markowitz. Next, the hard shrinkage of weights themselves, related to the penalty on short positions, makes the weights in Figure 3.6 almost always non-negative as expected in Section 3.4.2. In addition, it works to enforce shrinkage and some weights become zeros exactly. Figure 3.7 shows another result with $(\alpha_t, \beta_t, \lambda_t, \gamma_t) = (1, 100, 100, 100)$ in which we are more ambitious to achieve the target return. Ultimately, such a portfolio is known to be “degenerate,” having all the weights on a single asset. Reflecting this aspect, the shrinkage from $|w_t|$ pushes the weights of irrelevant assets to zeros. The figure also shows the number of nonzero weights over time out of thirteen assets, where we see the portfolio does not always use all the thirteen assets, but some of them. This exact shrinkage and its meaning are crucial in practice, especially when the number of assets in consideration is large and our portfolio should consist of not all but

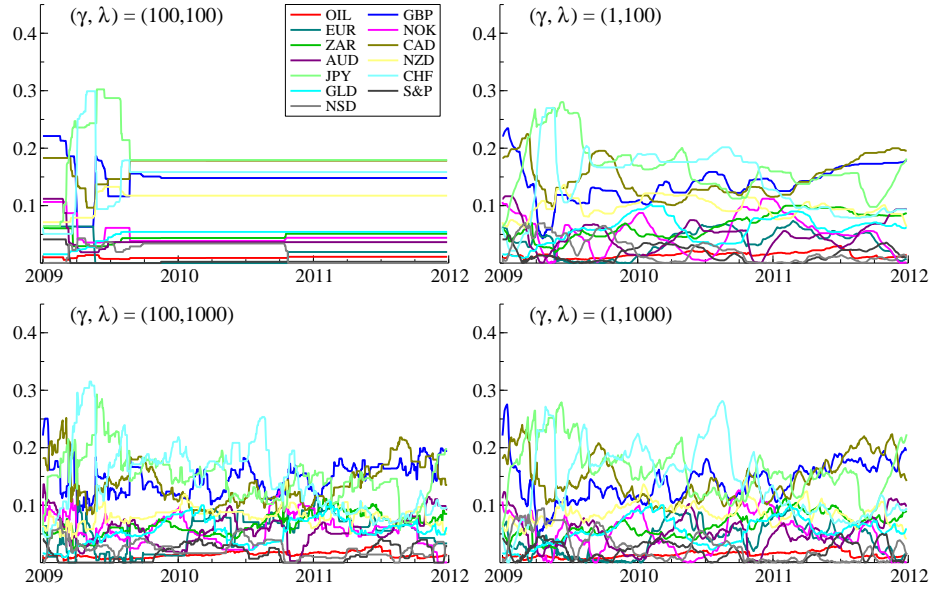


FIGURE 3.6: Optimal portfolio of the Laplace model. From top to bottom, left to right: portfolio of Laplace loss function with $(\lambda_t, \gamma_t) = (100, 100)$, $(100, 1)$, $(1000, 100)$, $(1000, 1)$ with $(\alpha_t, \beta_t) = (100, 1)$.

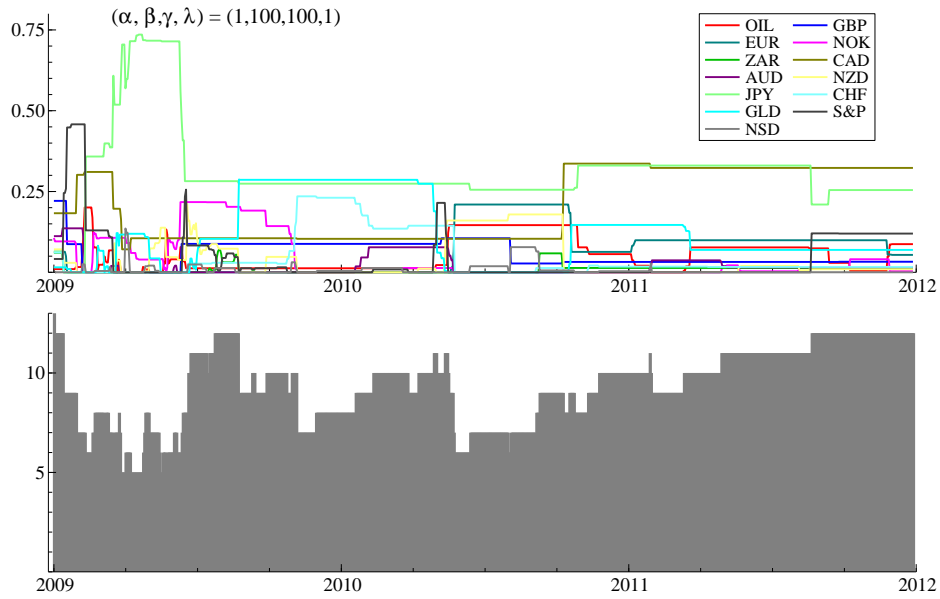


FIGURE 3.7: A Laplace portfolio with the number of active weights. Top: portfolio of Laplace loss function with $(\alpha_t, \beta_t, \lambda_t, \gamma_t) = (1, 100, 100, 100)$. Bottom: number of non-zero weights in the portfolio.

several assets that are really necessary.

These two types of shrinkage are balanced by two tuning parameters: λ_t and γ_t . Large λ_t makes the effect of $|w_t|$ relatively ignorable, meaning that we prefer the persistency of portfolios and expect to see more stepwise allocation switch in our portfolio. Conversely, if γ_t is sufficiently large, then we appreciate the non-negativity and sparsity of portfolio more, resulting in dynamically switching portfolio weights with fewer assets, since we allow for more switching by discounting the sparsity effect from $|w_t - w_{t-1}|$ by relatively small λ_t .

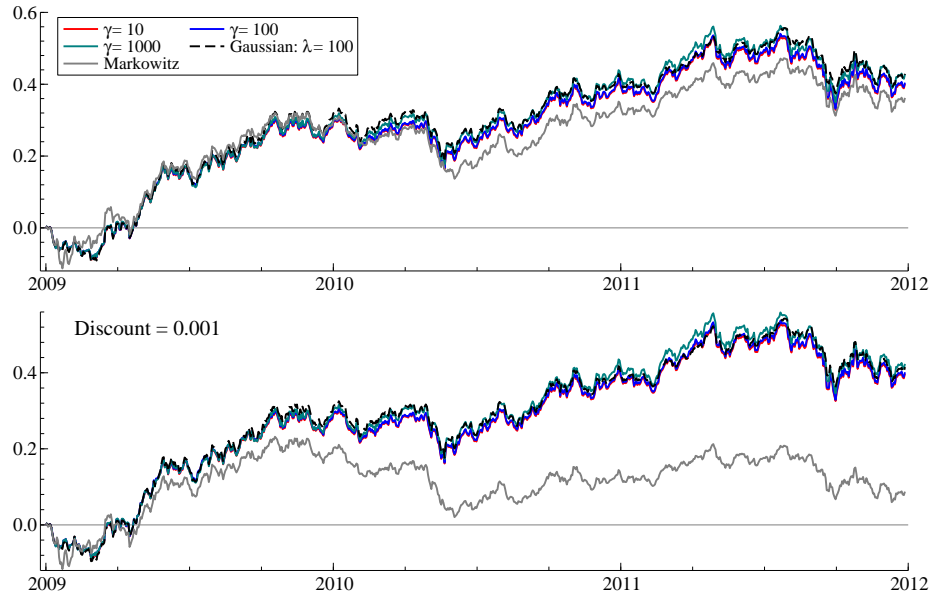


FIGURE 3.8: Cumulative returns of Laplace portfolios for $\lambda_t = 100$. While the weights for transaction costs ($|w_t - w_{t-1}|$) are fixed as $\lambda_t = 100$ in addition to $(\alpha_t, \beta_t) = (100, 1)$, those for sparsity in portfolio ($|w_t|$) are $\gamma_t = 10, 100$ or 1000 (red, blue and green, respectively). For comparison, the cumulative returns of Gaussian portfolio with $\gamma_t = 100$ (dotted black) and Markowitz (grey) are shown.

The cumulative returns are shown in Figure 3.8 and 3.9. All the results show the strong robustness to the transaction costs, thanks to the hard shrinkage on portfolio switching. With λ_t and γ_t being chosen appropriately, the portfolio of the Laplace loss function can outperform both those of DLM loss functions and Markowitz-type

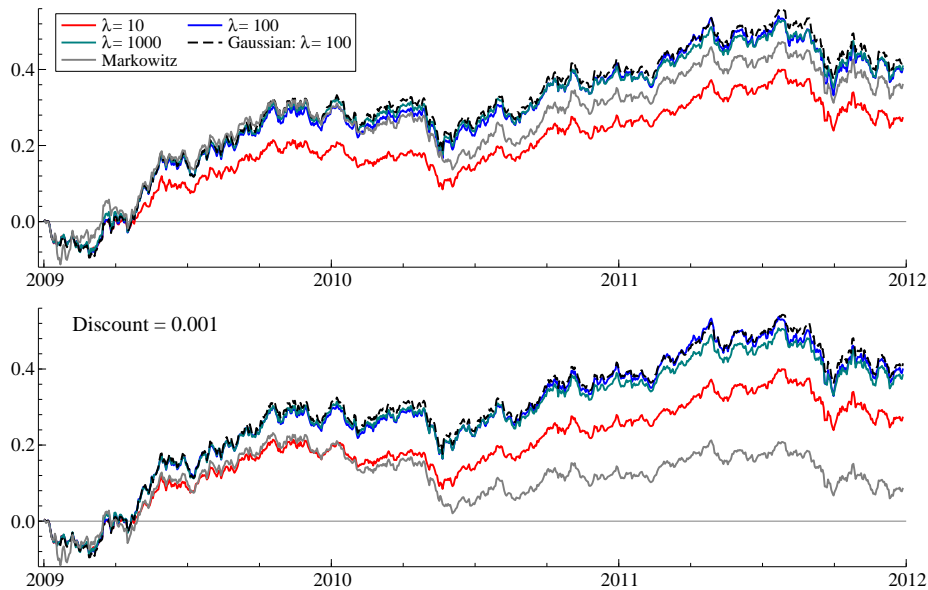


FIGURE 3.9: Cumulative returns of Laplace portfolios for $\gamma_t = 100$. Conversely to Figure 3.8, we fix $\gamma_t = 100$ and change the value of λ_t to 10, 100 and 1000.

method with and without transaction costs. In Figure 3.8, we see the robustness of cumulative returns to the value of γ_t , while in Figure 3.9 the choice of λ_t affects the performance of portfolios significantly.

Marginal Loss Approach

Figure 3.10 shows the optimal weights of the profiled Laplace loss function in Section 3.4.1 and the marginal Laplace loss function in Section 3.5 where $\gamma_t = \infty$ and $|w_t|$ terms are ignored. The patterns of portfolio switching in both models are almost the same. For example, both portfolios consist mainly of JPY in 2009, but later favor GBP and CAD. Yet, the difference between these two portfolios are clear in the shrinkage on weights difference; the profiled loss function leads to a longer period in which some of the weights are totally unchanged, while the marginal loss function allows for flexibility in portfolio dynamics as the quadratic loss function does. It is not that the marginal portfolio has no shrinkage at all; the stepwise weights shift

exists in the marginal portfolios, though they last for a short period of time and this is not clearly seen in the figure.

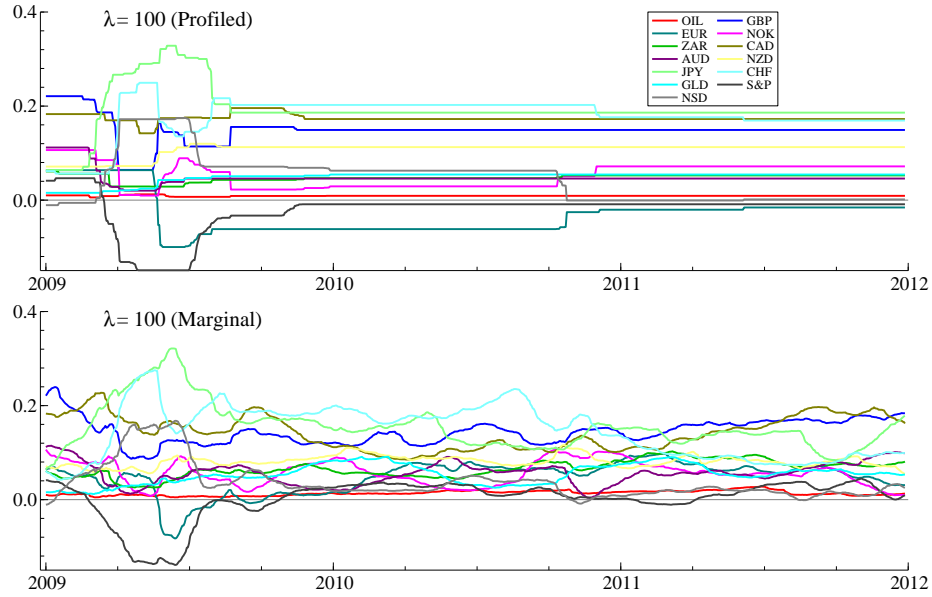


FIGURE 3.10: Portfolios of profiled and marginal Laplace models. Top: portfolio of profiled Laplace loss function with $\lambda_t = 100$. Bottom: portfolio of marginal Laplace loss function with $\lambda_t = 100$.

The cumulative returns without transaction costs are presented in Figure 3.11. While the profiled portfolio with $\lambda_t = 100$ has the highest cumulative return in this group, the marginal portfolio is relatively “robust” to the value of λ_t . The reason that the marginal portfolio outperforms the profiled one when $\lambda_t = 10$ is that the profiled portfolio becomes completely constant because of the strong effect of transaction costs, while the marginal portfolio remain persistent but dynamic.

This marginal strategy, which is insensitive to the tuning parameter λ_t , could be a conservative approach to the optimization problem when an investor is afraid of the misspecification of his or her loss function. Even if one uses too large or too small λ_t , the optimal portfolios of marginal models do not suffer from a serious failure as those of the joint models do.

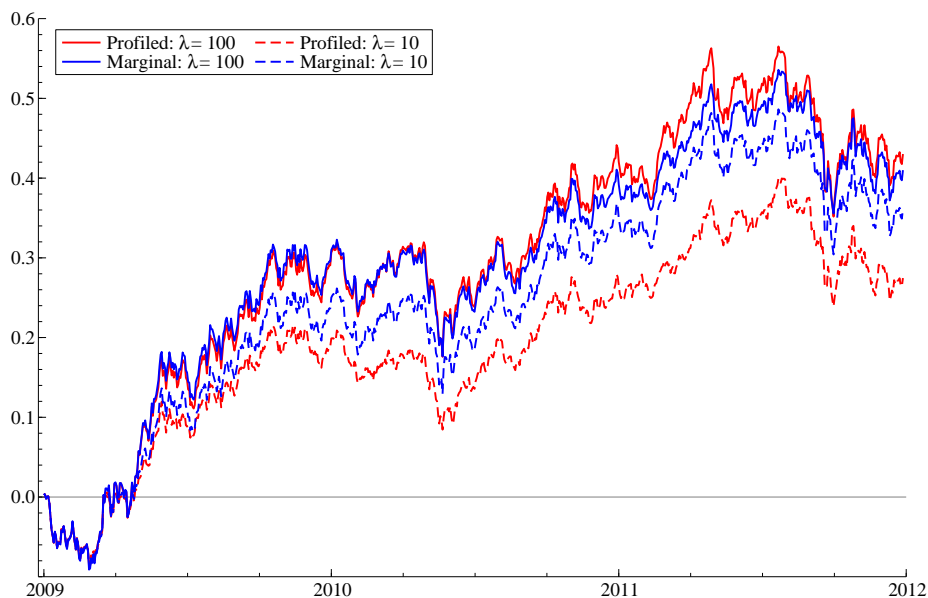


FIGURE 3.11: Cumulative returns of profiled and marginal Laplace portfolios. No transaction cost is used ($\delta = 0$). Four portfolios are distinguished by red lines for profiled (joint) loss functions, blue lines for marginal ones, solid lines for $\lambda_t = 100$ and dotted lines for $\lambda_t = 10$.

3.7 Summary Comments

In this chapter, the new loss functions formed by the multi-step ahead predictions and transaction costs are proposed and the computational methods tailored to those loss functions are developed in detail. The application shows empirical evidence that the portfolio strategy based on the proposed loss functions consistently outperforms the mean-variance optimization with and without transaction costs. By choosing appropriate loss functions and tuning parameters, it is possible to customize the portfolio to avoid suffering from a large loss of returns in the economically challenging period during/after the recession and to achieve reasonable profit in the stable economy after 2011.

The key theme of this chapter, or the “translation” of loss functions to synthetic models, can lead to the further development and extension of this type of research

in the future. The idea of synthetic models allow us to start from the statistical models to construct the loss function, as we make good use of the sparseness in the Laplace models in Section 3.4.1 and define the marginal loss function in Section 3.5. By this approach, many promising loss functions, which have not been tested in the field of finance, might be proposed for evaluation in the future. For example, the asymmetric penalty for the target return, in which we penalize the excess return less (or, ideally, not at all), can be formed through the skewed distributions. Also, as the loss functions are sophisticated in this way, more technical methods of estimation in statistics, such as simulation based approach, are expected to be useful in this optimization problem.

3.8 Appendix: Forward Filtering with Multiple Observations

3.8.1 Forward Filtering

The DLM with random walk state evolution is written as

$$\begin{aligned}y_t &= F_t' \theta_t + v_t, & v_t &\sim N(0, V_t), \\ \theta_t &= \theta_{t-1} + w_t, & w_t &\sim N(0, W_t),\end{aligned}$$

where y_t is $k \times 1$ observational vector and θ_t is $p \times 1$ state vector. All the necessary matrices, including F_t , V_t and W_t , are assumed to be known; this assumption is satisfied in the synthetic models of interest. Define the update of information at time t by $\mathcal{D}_t = \{y_t\} \cup \mathcal{D}_{t-1}$.

Forward Filtering:

- Posterior at $t - 1$: $p(\theta_{t-1} | \mathcal{D}_{t-1}) = N(m_{t-1}, C_{t-1})$.

- Prior at t : $p(\theta_t | \mathcal{D}_{t-1}) = N(a_t, R_t)$,

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

- Forecast at t : $p(y_t | \mathcal{D}_{t-1}) = N(f_t, Q_t)$,

where $f_t = F_t' m_{t-1}$ and $Q_t = F_t' R_t F_t + V_t$.

- Posterior at t : $p(\theta_t | \mathcal{D}_t) = N(m_t, C_t)$,

where $m_t = a_t + A_t(y_t - f_t)$, $C_t = R_t - A_t Q_t A_t'$ and $A_t = R_t F_t Q_t^{-1}$.

See West and Harrison (1997) for the proof and more details.

3.8.2 Two Observational Equations with the Common State

Consider the DLM with two observational equations with the common state vector,

$$\begin{aligned}y_{1t} &= F_{1t}' \theta_t + v_{1t}, & v_{1t} &\sim N(0, V_{1t}), \\ y_{2t} &= F_{2t}' \theta_t + v_{2t}, & v_{2t} &\sim N(0, V_{2t}), \\ \theta_t &= \theta_{t-1} + w_t, & w_t &\sim N(0, W_t),\end{aligned}$$

where y_{1t} and y_{2t} are the k_1 - and k_2 -dimensional vectors, respectively, and mutually independent. The set of information is, in this context, $\mathcal{D}_t = \{y_{1t}, y_{2t}\} \cup \mathcal{D}_{t-1}$, and the on-line posterior of θ_t is defined accordingly. To compute the posterior, we can still use the forward filtering by stacking the vectors and matrices as

$$y_t = \begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} \quad (k \times 1), \quad F_t = \begin{bmatrix} F_{1t} & F_{2t} \end{bmatrix} \quad (p \times k), \quad V_t = \begin{bmatrix} V_{1t} & 0 \\ 0 & V_{2t} \end{bmatrix} \quad (k \times k),$$

where $k = k_1 + k_2$. Despite its theoretical simplicity, this approach could be computationally burdensome with large k , especially in inverting $k \times k$ -matrix Q_t .

To soften this problem, the method of filtering is decomposed into two steps by processing y_{1t} and y_{2t} to make the size of inverse matrices to be k_1 and k_2 . This method first ignores the second observation and work on the following submodel

$$\begin{aligned} y_{1t} &= F'_{1t}\theta_t + v_{1t}, & v_{1t} &\sim N(0, V_{1t}), \\ \theta_t &= \theta_{t-1} + w_t, & w_t &\sim N(0, W_t), \end{aligned}$$

to obtain the half-updated temporal posterior

$$p(\theta_t|y_{1t}, \mathcal{D}_{t-1}) = N(m_{1t}, C_{1t}).$$

This is, in fact, considered the prior distribution when integrating the second observation into the current posterior, since the full on-line posterior has the form

$$\begin{aligned} p(\theta_t|\mathcal{D}_t) &\propto p(y_{2t}|\theta_t, y_{1t}, \mathcal{D}_{t-1})p(\theta_t|y_{1t}, \mathcal{D}_{t-1}) \\ &= p(y_{2t}|\theta_t)p(\theta_t|y_{1t}, \mathcal{D}_{t-1}), \end{aligned}$$

where the second equation is justified by the the conditional independence of observations. This implies that the forward filtering for DLM with a single observation can be used for computing the above distribution by using $N(m_{1t}, C_{1t})$ as “prior at t ” for the second observation. Call this repeated use of forward filtering “sequential forward filtering.” The algorithm is summarized as follows:

Sequential Forward Filtering:

- Posterior at $t - 1$: $p(\theta_{t-1}|\mathcal{D}_{t-1}) = N(m_{t-1}, C_{t-1})$.
- Stage 1: Process y_{1t} and implement forward filtering.
 - Prior at t : $p(\theta_t|\mathcal{D}_{t-1}) = N(a_{1t}, R_{1t})$.
 - Forecast at t : $p(y_{1t}|\mathcal{D}_{t-1}) = N(f_{1t}, Q_{1t})$.
 - Posterior at t : $p(\theta_t|y_{1t}, \mathcal{D}_{t-1}) = N(m_{1t}, C_{1t})$.
- Stage 2: Use the posterior to be the prior to process y_{2t} .
 - Prior at t : Define the prior by $N(m_{1t}, C_{2t})$ as
$$p(\theta_t|y_{1t}, \mathcal{D}_{t-1}) = N(a_{2t}, R_{2t}) = N(m_{1t}, C_{2t}).$$

Implement forward filtering:

- Forecast at t : $p(y_{2t}|y_{1t}, \mathcal{D}_{t-1}) = N(f_{2t}, Q_{2t})$.
- Posterior at t : $p(\theta_t|y_{2t}, y_{1t}, \mathcal{D}_{t-1}) = N(m_{2t}, C_{2t})$.
- On-line posterior: $p(\theta_t|\mathcal{D}_t) = N(m_t, C_t) = N(m_{2t}, C_{2t})$.

Again, the matrix inversion needed in this procedure is for $k_1 \times k_1$ - and $k_2 \times k_2$ -matrices. Considering the cost of the direct inversion of $k \times k$ -matrix that is roughly $\mathcal{O}(k^3)$, this sequential approach has the great advantage in computation.

Note that this is easily extended to the case of N observations with the single common state vector. In our portfolio research, $y_{1t} = m_t$, $y_{2t} = z_t = 0$ in eqns. (3.3,3.9) and $y_{3t} = u_t = 0$ in eqn. (3.10). Another advantage of this approach can be seen in this application as the easiness to add and drop the new component of the model. It is also useful when we have some missing observation in some period. This flexibility is helpful in simplifying the code for computations.

It is also worth mentioning that the multiple observations in DLMS do not affect the retrospective analysis; we compute $p(\theta_{1:T}|\mathcal{D}_T)$ and its marginal distributions completely in the same way as we do in the single observation case. Denote the on-line prior and posterior by $N(a_t, R_t)$ and $N(m_t, C_t)$. With the single observation, the marginal and conditional recursive posteriors are as follows:

Backward Smoothing/Sampling:

- $p(\theta_{t-1}|\mathcal{D}_T) = N(a_t(-1), R_t(-1)),$

where $a_t(-1) = m_{t-1} + B_{t-1}(m_t - a_t)$, $R_t(-1) = C_{t-1} - B_{t-1}[C_t - R_t]B'_{t-1}$ and $B_{t-1} = C_{t-1}R_t^{-1}$.

- $p(\theta_{t-1}|\theta_t, \mathcal{D}_T) = N(h_{t-1}, H_{t-1}),$

where $h_{t-1} = m_{t-1} + B_{t-1}(\theta_t - a_t)$ and $H_{t-1} = C_{t-1} - B_{t-1}R_tB'_{t-1}$.

As seen in the list of distributions above, the vectors and matrices needed in this analysis are all about the state variables, not observations. Therefore, even in the case of multiple observations, one can still use the same backward smoothing or sampling after he or she has finished the sequential forward filtering by keeping the on-line posterior mean and variance (m_t, C_t) . In the sequential forward filtering, the on-line prior mean and variance (a_t, R_t) are obtained as (a_{1t}, R_{1t}) .

3.8.3 Computational Efficiency

We emphasized the computational efficiency of FFBS in Section 3.3.3, but, of course, there are many methodological advances in the area of optimization and matrix manipulation that are potentially applicable for our study. For example, we pointed out the difficulty in inverting $kh \times kh$ -matrix in the optimal portfolio, but the sparsity of this matrix can be exploited to contribute to more efficient computation. The general computational algorithm for this type of matrix, called block-banded matrices

(in our case, with lag 1), is developed by Asif and Moura (2005), in which the inversion of the $kh \times kh$ matrix is reduced to those of $k \times k$ submatrices. Similarly, FFBS used in our method only needs the inverses of submatrices, from which we can confirm the computational efficiency of our method and conclude that the sequential FFBS is at least as efficient as the optimization method with the advanced matrix calculation technique.

3.9 Appendix: Exact Shrinkage in EM Algorithm

Section 3.4.1 shows the rough sketch of the EM methods used for the Laplace models, but its Expectation step needs more detailed explanation for implementation.

Suppose we have $\tau_{it}^{(s)} = \lambda_t^2 |w_{it}^{(s)} - w_{i(t-1)}^{(s)}| = 0$ at s -th iteration. Then, the expectation of the objective function diverges unless $w_{it} - w_{i(t-1)} = 0$ in which the corresponding term that involves τ_{it} is erased before expectation. This means that any decision that does not satisfy $w_{it} - w_{i(t-1)} = 0$ cannot be the optimal point of the objective function. Hence, the difference at the next iteration must be zero, i.e. $w_{it}^{(s+1)} - w_{i(t-1)}^{(s+1)} = 0$. In computation, if we encounter this situation, we use $\tau_{it} = 0$ and the SVD-based generalized inverse matrix of degenerated W_t . By this trick, we can numerically ensure that $w_{it}^{(s+1)} - w_{i(t-1)}^{(s+1)} = 0$ in the algorithm.

Note that the algorithm itself does not yield $w_{it} - w_{i(t-1)} = 0$ which we appreciate as the sparsity in portfolio switching. In the actual computation in Section 3.6, we insert (as noted in the section) an additional step where we set $w_{it}^{(s+1)} - w_{i(t-1)}^{(s+1)} = 0$ if $|w_{it}^{(s+1)} - w_{i(t-1)}^{(s+1)}| < \epsilon$ for some pre-specified threshold ϵ (we used $\epsilon = 0.0001$). We may implement this additional step only if the monotonicity of the objective function is guaranteed so that the convergence of series is still assured. Similarly, for the full Laplace model in eqn. (3.10), we use two thresholds for both $w_t - w_{t-1}$ and w_t . Because of this thresholding, the portfolio weights are classified into three groups:

changed, unchanged and thresholded to zero. The application of this thresholding might violate the sum-to-one constraint, so the thresholded weights must be added to the other active weights (see Section 3.10.3). We have to be careful to avoid breaking the monotonicity of the objective function by this additional step.

3.10 Appendix: Posterior Marginal Mode of Laplace Models

3.10.1 Gibbs Sampler and Maximization for Marginal Models

To construct the approximate density in eqn. (3.12), we need to sample from the (joint) posterior of the model in eqn. (3.9). The Gibbs sampler for our model is the same as that of lasso regression in Park and Casella (2008) with re-parametrization and the slight modification by FFBS.

Gibbs Sampler for Laplace State-Space Models:

Suppose $t = 0$. At each iteration $i \in 1:S$,

1. Sample j -th element of vector τ_t from

$$\tau_{jt}^{(i)} | w_{1:h}^{(i-1)} \sim GIG(p = 1/2, a = \lambda_t^{-2}, b = (w_{jt} - w_{j(t-1)})^2),$$

independently across j and t . This is the Generalized Inverse Gaussian distribution with density defined by

$$p(x|p, a, b) \propto x^{p-1} \exp\{(ax + b/x)/2\}.$$

2. Sample $w_{1:h}^{(i)} | \tau_{1:h}^{(i)}$ by Backward Sampling.
3. For the later use, record the marginal mean and variances, $(m_1^{(i)}, C_1^{(i)})$, in $p(w_1 | \tau_{1:h}^{(i)}, \mathcal{D}_h) = N(w_1 | m_1^{(i)}, C_1^{(i)})$.

In application of Section 3.6, we sampled 2000 particles after 1000 burn-in. Figures 3.12 and 3.13 show some convergence diagnoses for $w_{1:3,t+1}$ and $\tau_{1:3,t+1}$ at $t = 1$

(this means we are at $t = 1$ and the loss functions has $w_{(t+1):(t+h)}$). The chain is mixing well, and this is true for the other variables and time points.

With the samples $\{w_{1:h}^{(i)}, \tau_{1:h}^{(i)}\}$ and the by-product $\{m_1^{(i)}, C_1^{(i)}\}$, the estimate becomes

$$\hat{p}(w_1|\mathcal{D}_h) = \frac{1}{S} \sum_{i=1}^S N(w_1|m_1^{(i)}, C_1^{(i)}).$$

Here, note that we know the normality of $p(w_1|\tau_{1:h}^{(i)}, \mathcal{D}_h) = N(w_1|m_1^{(i)}, C_1^{(i)})$ by the theory of FFBS.

The next step is the maximization of this mixture of normals. The maximizer of the mixture density must satisfy the first order condition,

$$w_1 = \left\{ \frac{1}{S} \sum_{i=1}^S N(w_1|m_i, C_i) C_i^{-1} \right\}^{-1} \left\{ \frac{1}{S} \sum_{i=1}^S N(w_1|m_i, C_i) C_i^{-1} m_i \right\} \quad (3.14)$$

with $m_i = m_1^{(i)}$ and $C_i = C_1^{(i)}$. We use this equation for the iterative optimization, so that the solution in the s -th step $w_1^{(s)}$ is obtained from the equation above with $w_1 = w_1^{(s)}$ in the left-hand-side and $w_1 = w_1^{(s-1)}$ in the right. In application, we have 100 iterations of this update.

3.10.2 Sum-to-One Constraint

Note that, if the sum-to-one constraint is imposed on the original model by setting $W_t = \text{diag}(\tau_t) - \tau_t \tau_t' / \mathbf{1}_k' \tau_t$, then the full conditional of τ_t is not necessarily a GIG distribution as in the previous section. To sample τ_t in such a case, we use an independent Metropolis-Hastings algorithm with this GIG distribution as a proposal distribution that is based on the model with $W_t = \text{diag}(\tau_t)$. Though the degenerate normal distribution has no proper density, we define its approximation based on the idea of SVD-based generalized inverse, by using $W_t = \text{diag}(\tau_t) - \tau_t \tau_t' / (\mathbf{1}_k' \tau_t + c)$ when evaluating the proposal density. The constant c should be very small and we use

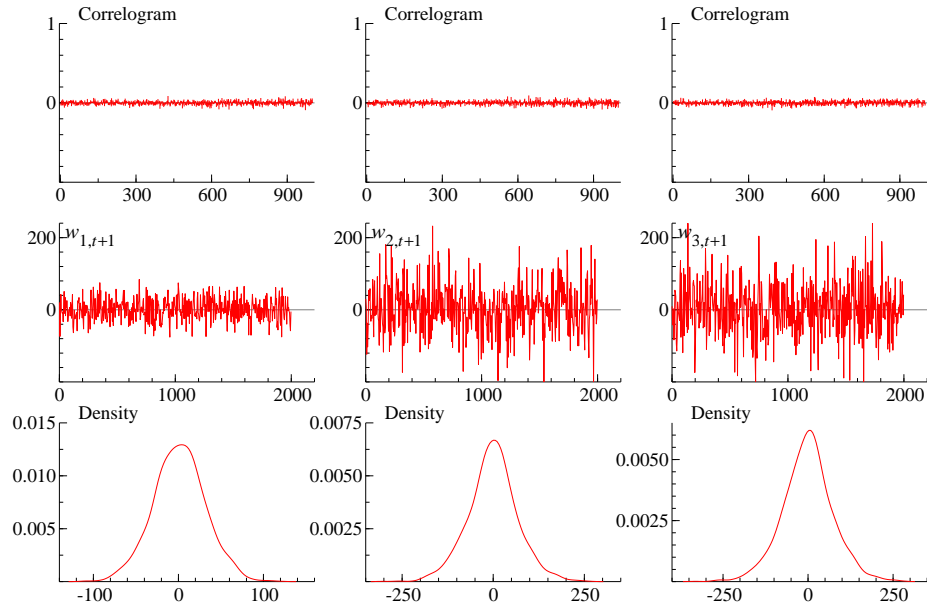


FIGURE 3.12: Convergence diagnoses for $w_{1:3,t+1}$ at $t = 1$. From the top row to the bottom one, the correlograms, sample paths and estimated marginal densities are shown.

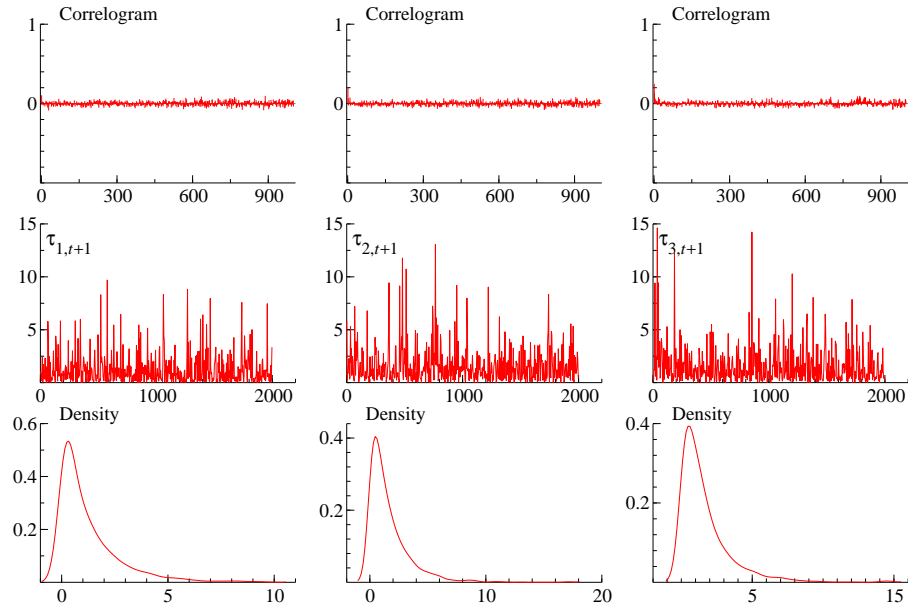


FIGURE 3.13: Convergence diagnoses for $\tau_{1:3,t+1}$ at $t = 1$. The sample paths shown in the second row are re-scaled to be $10^{-4}\tau_{it}$.

$c = 10^{-9}$ in our study in Section 3.6. The acceptance probability of proposed sample τ_{it}^{new} conditional on previous sample τ_{it}^{old} and the other parameters is

$$\sqrt{\frac{c + \mathbf{1}'_k \tau_t^{new}}{c + \mathbf{1}'_k \tau_t^{old}}},$$

where $\tau_t^{new} = (\tau_{1t}, \dots, \tau_{it}^{new}, \dots, \tau_{kt})'$ and $\tau_t^{old} = (\tau_{1t}, \dots, \tau_{it}^{old}, \dots, \tau_{kt})'$. We observed that this acceptance probability is almost always high enough, about 98%.

Also, because of the sum-to-one constraint, the sampled C_i , hence the matrix that is inverted in eqn. (3.14), are rank-deficient. We use the generalized inverse based on singular value decomposition in FFBS of Gibbs Sampler and optimization. This usage of generalized inverse ensures the sum-to-one condition at every iteration of optimization.

3.10.3 Realization of Exact Zero Weights and Transitions

In the iterations of the Gibbs sampler and optimization in Section 3.10.1, there is no step where w_1 can shrink to the old portfolio w_0 . One can make the ad-hoc “correction” to this output of the optimization as follows. Denote the original maximizer obtained from the optimization by w_1 . We want the additional constraint that requires $w_{i1} = w_{i0}$ if $|w_{i1} - w_{i0}| < \epsilon$. This destroys the sum-to-one constraint, so we need another correction after implementing this constraint. Suppose there are $k_1 \leq k$ elements in w_1 that satisfies $|w_{i1} - w_{i0}| \geq \epsilon$, denote the set of those elements by \mathcal{S} and set $k_0 = k - k_1$. Define the difference between the two portfolios by $\Delta = \sum_{i \notin \mathcal{S}} (w_{i1} - w_{i0})$. Then, we “correct” w_1 to w_1^* so that it satisfies both the sum-to-one constraint and exact shrinkage. We do this by setting $w_{i1}^* = w_{i0}$ if $i \notin \mathcal{S}$ and $w_{i1}^* = w_{i1} - \Delta/k_1$ otherwise. The threshold used in Section 3.6 is $\epsilon = 0.0001$.

3.11 Appendix: Dynamic Dependence Network Models

In this appendix, the model used in analyzing and predicting the log prices is explained. It is largely based on the work by Zhao et al. (2016) and called Dynamic Dependence Network Models (DDNMs). Denote the $k \times 1$ vector of asset prices by y_t . This follows the TV-VAR(2) as

$$y_t \sim N(\Phi_t x_t, \Sigma_t),$$

where Φ_t is the matrix of dynamic regression coefficients and $x_t = (1, y'_{t-1}, y'_{t-2})'$. The stochastic volatility matrix, Σ_t , is supposed to have the following Cholesky decomposition (or, if necessary, the singular value decomposition),

$$\Sigma_t = A_t D_t A_t', \quad (3.15)$$

where $D_t = \text{diag}(v_{1t}, \dots, v_{kt})$ and A_t is the lower triangular matrix with diagonal ones. Note that the inverse of A_t is also lower triangular, so write $A_t^{-1} = I - \Gamma_t$ where Γ_t is the lower triangular matrix with diagonal zeros. Then, the model is rewritten as

$$y_t = \Phi_t x_t + \Gamma_t y_t + N(0, D_t).$$

Each j -th element of the vector follows

$$y_{jt} = x'_t \phi_{jt} + y'_{pa(j),t} \gamma_{jt} + N(0, v_{jt}), \quad (3.16)$$

where ϕ'_{jt} is the j -th row of matrix Φ_t , $pa(j)$ is the parental set defined as the indices of j -th row of Γ_t with non-zero elements, and $y_{pa(j),t}$ and γ_{jt} are the corresponding sub-vectors with $|pa(j)|$ elements of y_t and j -th row of Γ_t . The rest of the model has the same structure as that of DLMS. The state parameters are (ϕ_{jt}, γ_{jt}) and these are assumed to follow Gaussian random walks with the discount factor for the state innovations. The observational variance v_{jt} is modeled by stochastic volatility,

i.e., the traditional gamma-beta random walk based on another specified (volatility) discount factor. See Zhao et al. (2016) and West and Harrison (1997) for more details.

Given the parental sets $pa(j)$, the model is estimated efficiently through independent analyses of the univariate submodels in Equation (3.16). This strategy allows us to implement forward filtering in each submodels to construct analytically the online posterior of state parameters and one-step forward predictive distribution of observations. In the computation of h -step ahead forecasts, however, the analytical expression is no longer obvious because of the future observation y_{t+i} ($1 \leq i \leq h$) in the predictor x_{t+h} that is unknown at time t . For this reason, we use the simulation-based method to generate simulated predictions. In other words, we generate sufficiently many random variables from

$$p(y_{(t+1):(t+h)}|\mathcal{D}_t) = \prod_{i=1}^h p(y_{t+i}|\mathcal{D}_{t+i-1}),$$

sequentially from y_{t+1} to y_{t+h} .

The observational variable y_t in the model above is usually the log prices, $y_t = \log p_t$. In contrast, the forecast we need in portfolio choice is on the return scale. The return from i -th asset is defined by $r_{it} = (p_{it} - p_{i(t-1)})/p_{i(t-1)}$. Similarly, the return of the h -step ahead forecast at time t is $r_{i(t+h)|t} = (p_{i(t+h)} - p_{it})/p_{it}$. To obtain the moments of these h -step ahead returns, we can simply transform the sampled prices into returns $r_{i(t+h)|t}$ and then calculate the sample analogues of the moments of returns.

Before starting the posterior analysis and prediction, we have to fix the parental sets. To do this, we take the naive thresholding approach with the estimate of Γ_t in a specified training period. First, we use the first 500 observation of the time series as a training set and estimate the model with full Γ_t matrix with $pa(j) = 1:(j-1)$.

Table 3.2: Parental sets used in prediction.

Parent j	$pa(j)$
OIL	\emptyset
GBP	\emptyset
EUR	\emptyset
NOK	EUR
ZAR	GBP NOK
CAD	\emptyset
AUD	NOK CAD
NZD	AUD
JPY	GBP EUR CAD AUD
CHF	\emptyset
GLD	GBP ZAR CAD CHF
S&P	GBP EUR NOK CAD AUD NZD
NSD	AUD JPY CHF S&P

Then, the Cholesky decomposition of the posterior mean of Γ_T is computed, where T here is the end point of the training dataset. Finally, we set the threshold d , and define the parental set $pa(j)$ by setting $i \in pa(j)$ for each $j \geq 2$ and $i \leq j - 1$ if the (i, j) -element of the estimated Γ_t is above the threshold. Note that $pa(1) = \emptyset$.

The choice of the threshold is crucial. The larger the threshold is, the closer to the full covariance our model becomes. The threshold controls the number of conditional independence assumptions, affecting the uncertainty in the resulting predictive information. After trying several values of the threshold, we decided to use $d = 0.2$ to choose the parental sets. The resulting parental sets are displayed in Table 3.2.

All the results presented in Section 3.6 are based on the predictions obtained with 500 training time series, 0.97 for discount factors of stochastic volatilities, 0.98 for those of state evolutions, 50000 simulation to obtain the predictive means and variances.

3.12 Appendix: Supplemental Analysis

3.12.1 Choice of Tuning Parameters

In the main analysis of the quadratic loss functions, we used $(\alpha_t, \beta_t) = (0.01, 1)$ for all t and the three values of λ_t . Now we change these tuning parameters to see their effects on the resulting portfolios.

Similarity to Markowitz Rule

Fixing $\lambda_t = 10000$ and $\beta_t = 1$, we try $\alpha_t = 0.0001, 0.001, 0.01, 0.1$, and 1 . With the strong discount of the transaction cost, the portfolio is expected to be similar to Markowitz portfolio. The choice of small α_t relative to (β_t, λ_t) means the loss from MSE has more impact on the decision. Remember that this includes the hard constraint on the achievement of the target return as the limiting case where $\alpha_t \rightarrow 0$.

Figures 3.14 and 3.15 show the cumulative return and standard deviation of those portfolios. With no transaction cost, the cumulative returns are almost the same, while our portfolios still possess their robustness to the transaction costs. The standard deviation with values of α_t is larger than the lower bound but less than that of Markowitz rule. This indicates that having the soft shrinkage on MSE with the moderate value of α_t can achieve the intermediate state between the minimum-risk portfolio and the profitable one by the hard constraint on the target return.

Balance between MSE and Variance

In contrast to the main analysis, consider a large discount on the variance term and more weight on the MSE term by $\alpha_t = 0.01$ and $\beta_t = 1$. In other words, we now pursue more profit by achieving the target return with more risk. The results are shown in Figures 3.16 and 3.17. Unfortunately, these over-ambitious strategies were not successful in this example, as seen in less return and the inflated risk.

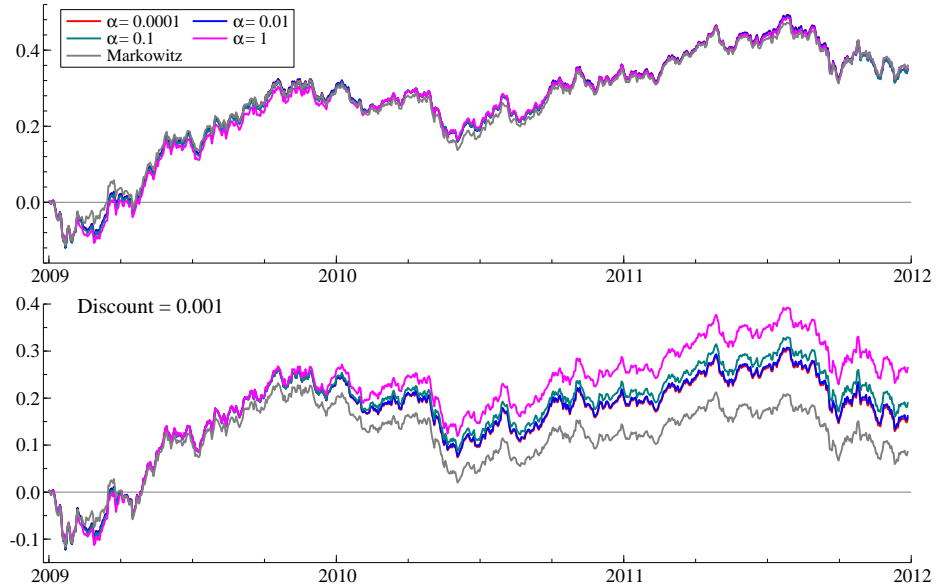


FIGURE 3.14: Cumulative returns for $\lambda_t = 10000$. Top: cumulative returns with no transaction cost. Bottom: with 0.1% transaction cost. Four cumulative returns of DLM portfolios of $\lambda_t = 1$ (red), 100 (blue), 10000 (green) and Markowitz portfolio (grey) are shown in both pictures.

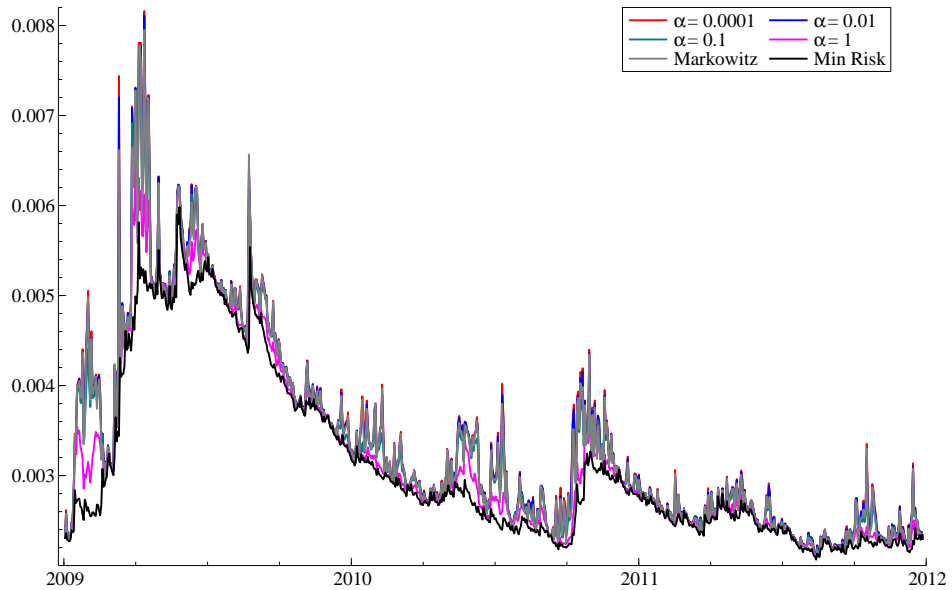


FIGURE 3.15: Standard deviation for $\lambda_t = 10000$. The black line shows the standard deviation of the minimum risk portfolio.

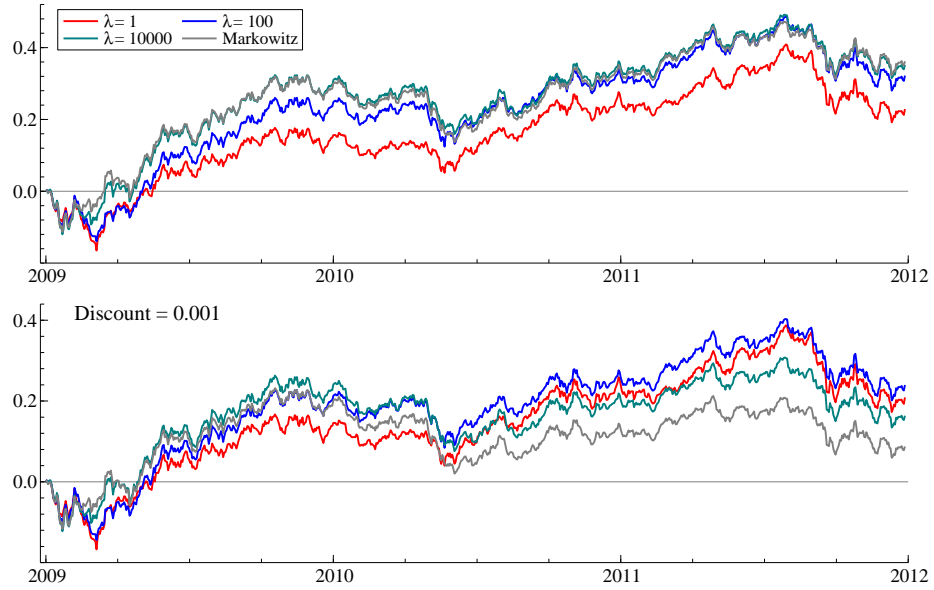


FIGURE 3.16: Cumulative returns for $(\alpha_t, \beta_t) = (0.01, 1)$. Top: cumulative returns with no transaction cost. Bottom: with 0.1% transaction cost. Four cumulative returns of DLM portfolios of $\alpha_t = 0.0001$ (red), 0.01 (blue), 0.1 (green), 1 (pink) and Markowitz portfolio (grey) are shown in both pictures.

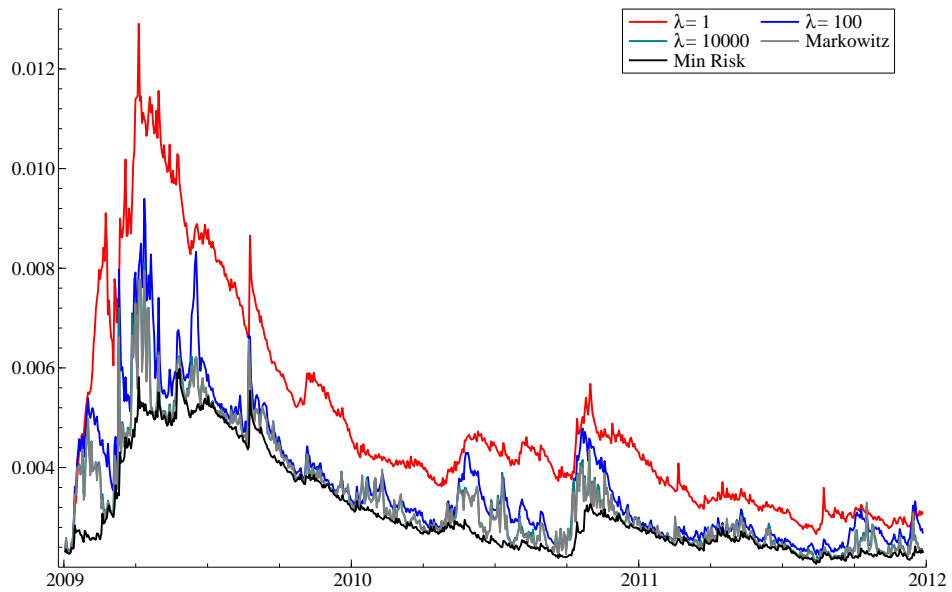


FIGURE 3.17: Standard deviation for $(\alpha_t, \beta_t) = (0.01, 1)$.

Use of Prior on Tuning Parameters

The difficulty of tuning the fixed parameters motivates us to put a prior on $(\alpha_t, \beta_t, \lambda_t)$. Though the functional form of the loss function becomes unclear, the use of priors in the synthetic model is useful especially in the case where one cannot write down his or her preference in the form of the tuning parameters. The computation of the posterior mode is feasible by, for example, the EM methods with the gamma priors. Yet, one has to make the “initial guess” on the values of his or her tuning parameters to set the hyperparameters of those priors.

3.12.2 Additional Figures for Profiled and Marginal Portfolios

Figures 3.18 and 3.19 show the profiled and marginal portfolios with the smaller and larger values of λ_t than in the main analysis.

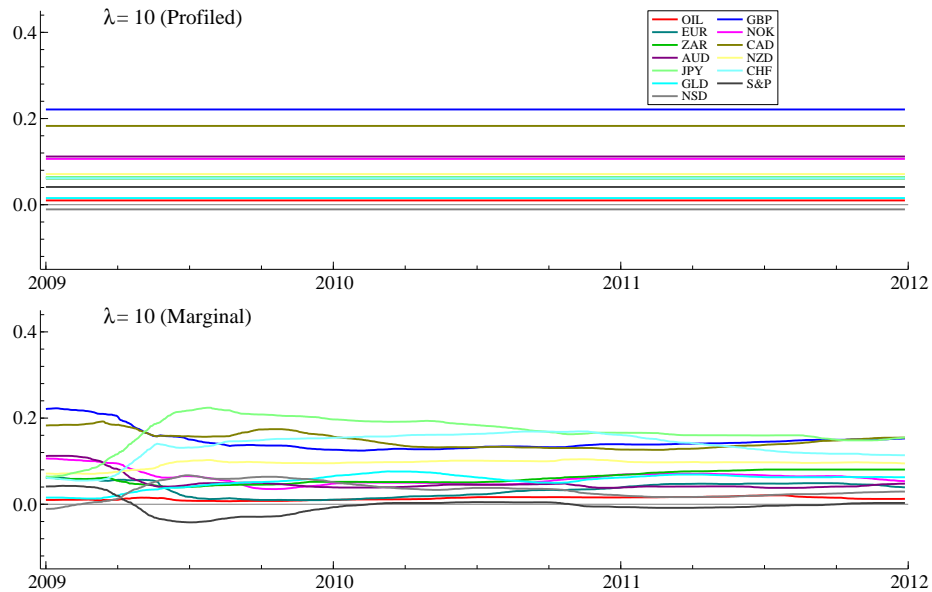


FIGURE 3.18: Profiled and marginal Portfolios for $\lambda_t = 10$. Top: profiled, Bottom: marginal.

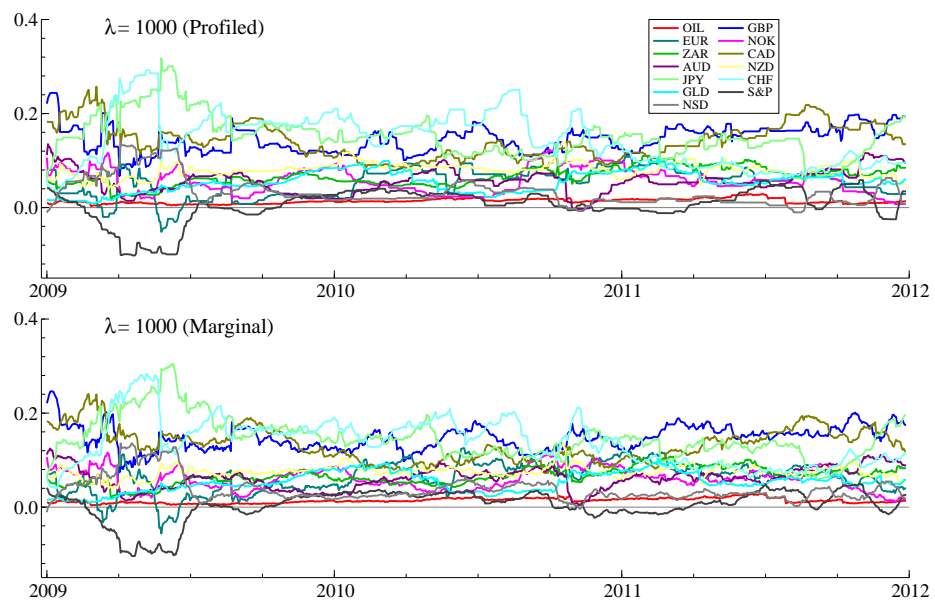


FIGURE 3.19: Profiled and marginal portfolios for $\lambda_t = 1000$. Top: profiled, Bottom: marginal.

Emulation of Dynamic Gravity Model by Bayesian Dynamic Flow Models

4.1 Introduction

Traffic flow count data in networks arise in many applications, such as automobile or aviation transportation, certain directed social network contexts, and internet studies. This increasing access to streaming data on large and evolving networks drives interest in formal models to quantify stochasticity and structure of latent processes underlying observable data streams. A key challenge is to define relevant statistical models that yield computationally efficient and scalable methods for streaming data, and that lead to sound statistical methods for monitoring, short-term prediction, single sample inference, and multi-sample comparisons across contexts.

We contribute to this area with an applied study that builds on two methodological advances. First, we develop a flexible, adaptive (non-stationary and non-Gaussian) state-space model for streaming count data. The analysis is explicitly designed to be computationally efficient for on-line data analysis; it scales quadratically in the number of network nodes, and is inherently parallelizable so enables

distributed implementation for streaming data on increasingly large networks. We achieve this by (a) theory-based decoupling of models for individual network node-pair flows, and (b) adapting a computationally trivial univariate stochastic volatility model (e.g. West and Harrison, 1997, Section 10.8) to apply to latent rates of underlying flows. This *Bayesian dynamic flow model* (BDFM) is developed in Section 4.2 with its underlying model theory and Bayesian analysis.

Second, we use the BDFM as an *emulator* of a *dynamic gravity model* (DGM). The DGM represents flows between network nodes with time-varying random processes for node-specific main effects and node-pair interaction effects. This is fundamentally interesting for understanding dynamics in network flows and node-node interactions. However, the fully Bayesian analyses of specific classes of gravity models involve significant computational demands inherent in the use of MCMC and related methods. In any realistic dynamic extension appropriate for scalable, on-line analysis of streaming network flow data, such methods must be avoided. Emulation analysis uses the unconstrained BDFM as a surrogate model, mapping posterior distributions to those of a coupled DGM. This exploits the statistical flexibility/adaptability and computational efficiency/scalability of the BDFM to create posterior inferences on the more highly structured and substantively focused DGM whose analysis is otherwise far more challenging. This is developed in Section 4.3.

Section 4.5.4 concerns application of this emulation analysis in a study of past browser traffic on FoxNews websites. We focus on flows of visitors to a set of domains of webpages in a structurally well-defined but dynamic/evolving network. The analysis by BDFM and DGM reveals the dynamic aspects of network structure, including overall traffic population, popular/unpopular domains and outstanding flows in the network. Summary comments conclude this chapter in Section 4.6.

4.2 Bayesian Dynamic Flow Models

We define a state-space model for the univariate series of counts, $\{y_t\}_{t=1:T}$, where $y_t \in \mathbb{Z}_+$. We assume conditional Poisson sampling models for observations at each time point; $y_t|\phi_t \sim Po(\phi_t)$. The question is how the state evolution of ϕ_t , or the prior of ϕ_t , should be modeled in order to appropriately allow for temporal changes while also providing analytic tractability. One approach to this goal is to have a model that yields marginal prior gamma distributions for ϕ_t at each time, so defining conjugate prior-posterior updates. In other words, we require the prior, or $p(\phi_t|\mathcal{D}_{t-1})$, to be gamma distributed, and define the state evolution, or the conditional distribution $p(\phi_t|\phi_{t-1}, \mathcal{D}_{t-1})$, so that it matches this property.

The effort to find such an ideal state-space model dates back to Smith (1979), where he proposes “power discounting” to ensure that the current marginal posterior distribution becomes the conjugate prior for the likelihood of the next observation. Though this idea is general and applicable to other types of state-space models, it lacks a probability model justification; the state evolution is not defined explicitly. This research is reviewed in Section 4.2.3 as a special case of our state-space models. However, the standard gamma-beta discount volatility model as used in earlier chapters is such a formal model, and is adopted here.

Denote by \mathcal{D}_t the set of information available at time t that includes $y_{1:t}$. Suppose the posterior distribution at time $t - 1$ is $p(\phi_{t-1}|\mathcal{D}_{t-1}) = Ga(\phi_{t-1}|r_{t-1}, c_{t-1})$. Then, conditional on \mathcal{D}_{t-1} and ϕ_{t-1} , the state evolution at time t is defined by

$$\phi_t = \phi_{t-1}\eta_t/\delta, \tag{4.1}$$

where $\eta_t \in (0, 1)$ is the stochastic, multiplicative contamination term that is conditionally independent of ϕ_{t-1} and distributed as

$$\eta_t|\mathcal{D}_{t-1} \sim Be(\delta r_{t-1}, (1 - \delta)r_{t-1}),$$

and $\delta \in (0, 1)$. This constant value, δ , is called the discount factor and interpreted as an “information decay rate” controlling parameter, defining the level of variation in the random “shock” η_t to the latest state ϕ_{t-1} . For now, assume this value is fixed prior to the analysis. Note that the contamination term is defined conditional on \mathcal{D}_{t-1} and the result of the posterior analysis at the previous time point. By definition,

$$E[\phi_t | \phi_{t-1}, \mathcal{D}_{t-1}] = \phi_{t-1},$$

so that the model has the form of a multiplicative random walk. A lower value of δ leads to a more diffuse beta innovation distribution for η_t and more flexibility for on-line posteriors to adapt to volatile observations, while a value closer to one indicates a steady, stable evolution. This random walk nature of the model allows for changes, but does not anticipate specific directional changes.

The gamma random walk model for positive state variables was first introduced in modeling the dynamic variance parameter (or stochastic volatility) in DLMS (as a special case of matrix-beta discount formulation proposed by Uhlig (1994, 1997)), and is discussed in detail by West and Harrison (1997) and Prado and West (2010). Some contents covered in this section are also discussed in Section 10.4.8 of West and Harrison and Section 4.3.7 of Prado and West.

4.2.1 *Forward Filtering and Backward Sampling*

The goal of the posterior analysis is to sequentially revise the following probability distributions:

- Forecast distribution: $p(y_t | \mathcal{D}_{t-1})$.
- On-line posterior: $p(\phi_t | \mathcal{D}_t)$.
- Retrospective posterior: $p(\phi_{1:T} | \mathcal{D}_T)$.

Key interest lies in the retrospective posterior as it represents our best understanding on the parameters of interest, utilizing all the information available for use. Yet, the other two distributions are necessary and need to be computed prior to the retrospective analysis. Moreover, the forecast distribution is closely related to the marginal likelihood, naturally central to model comparison and choice of discount factor, as seen in Section 4.2.2. Though our main interest is in the retrospective analysis in this research, we do note and stress that the on-line filtering and forecast distributions are critical for monitoring data as it is processed, as well as for predictive purposes.

Forward Filtering

Assume the prior is set as $\phi_0 \sim Ga(r_0, c_0)$. The forecast distribution and on-line posterior are obtained recursively as follows.

1. Posterior at $t - 1$:

$$\phi_{t-1} | \mathcal{D}_{t-1} \sim Ga(r_{t-1}, c_{t-1}).$$

2. Prior at t :

$$\phi_t | \mathcal{D}_{t-1} \sim Ga(\delta r_{t-1}, \delta c_{t-1}).$$

3. Forecasting for t :

$$y_t | \mathcal{D}_{t-1} \sim NB(k_t, p_t), \quad Pr[y_t = y] = \binom{y + k_t - 1}{y} p_t^{k_t} q_t^y, \quad (4.2)$$

where

$$k_t = \delta r_{t-1}, \quad p_t = \frac{\delta c_{t-1}}{1 + \delta c_{t-1}}.$$

4. Posterior at t :

$$\phi_t | \mathcal{D}_t \sim Ga(r_t, c_t)$$

where

$$r_t = \delta r_{t-1} + y_t, \quad c_t = \delta c_{t-1} + 1.$$

The proof is given in Section 4.7.1. The updating rule for the gamma parameters of the on-line posterior is similar to that of the Poisson-gamma model with no time-dependence; if $\phi_t = \phi$ for all t and $\phi \sim Ga(r_0, c_0)$, then $p(\phi|\mathcal{D}_t) = Ga(r_t, c_t)$ where $r_t = r_{t-1} + y_t$ and $c_t = c_{t-1} + 1$. This static model is a special case of our dynamic model with $\delta \rightarrow 1$. It also clarifies the effect of the discount factor; it literally discounts the past information, r_{t-1} and c_{t-1} , against the new observation. In addition, regarding sensitivity to prior specification, this updating rule shows that the prior choice has little impact. Suppose we have a non-informative prior in the sense that c_0 is extremely small to realize the large variance in prior of ϕ_0 . Once we observe counts at time $t = 1$, however, this uncertainty is “washed away” immediately as $c_1 = \delta c_0 + 1$. With the adjustment of the location by $r_1 = \delta r_0 + y_1$, the variance of on-line posterior has to be moderate, no matter how diffuse the rate parameter is in the initial distribution.

This adaptation of the existing gamma-beta volatility model for Poisson rates was first examined by Harvey and Fernandes (1989a,b) and Shephard (1994) and applied it to the study of political science by Brandt et al. (2000) and Brandt and Williams (2001). Their non-Bayesian analysis emphasizes some part of the results of filtering above, especially the forecast distribution, and they use it as the likelihood to derive point estimates of the hyperparameters. We, from the Bayesian point of view, understand the results above as posterior distributions; we can explicitly evaluate distributions of state variables which are integrated out in non-Bayesian analysis as nuisance parameters.

The next step is to derive the retrospective posterior, $p(\phi_{1:T}|\mathcal{D}_T)$. However, unlike the three distributions we obtained in filtering, the retrospective posterior is not a well-known distribution by itself. For this reason, we take the simulation-based approach, in that we sample many particles from the target, retrospective posterior distribution to represent its density and moments by the histogram and sample analog. To sample from $p(\phi_{1:T}|\mathcal{D}_T)$, we take the compositional form of this joint distribution, based on the decomposition

$$p(\phi_1, \dots, \phi_T|\mathcal{D}_T) = p(\phi_T|\mathcal{D}_T) \prod_{t=1}^{T-1} p(\phi_t|\phi_{t+1}, \mathcal{D}_T),$$

by which the problem reduces to sampling from $p(\phi_t|\phi_{t+1}, \mathcal{D}_T)$. Given ϕ_{t+1} , we sample ϕ_t from $p(\phi_t|\phi_{t+1}, \mathcal{D}_T)$ and repeat this procedure inductively by starting from ϕ_T so that the sampled $\{\phi_1, \dots, \phi_T\}$ follows $p(\phi_{1:T}|\mathcal{D}_T)$. Note that the initial marginal distribution we first sample from, $p(\phi_T|\mathcal{D}_T)$, is actually the on-line posterior at time $t = T$, that is the gamma distribution with known parameters.

Backward Sampling

For $p(\phi_t|\phi_{t+1}, \mathcal{D}_T)$, the following distributional equation holds:

$$\phi_t = \delta\phi_{t+1} + \epsilon_t, \quad \epsilon_t \sim G((1 - \delta)r_t, c_t).$$

The detailed derivation is given in Section 4.7.2. The equation above tells us that we may sample ϵ_t from $G((1 - \delta)r_t, c_t)$ and set $\phi_t = \delta\phi_{t+1} + \epsilon_t$. Here, the discount factor serves as a smoothness parameter; the larger δ is, the smoother the latent process becomes; precisely, $\phi_t \rightarrow \phi_{t+1}$ and $\epsilon_t \rightarrow 0$ almost surely as $\delta \rightarrow 1$.

4.2.2 Marginal Likelihoods and Optimal Discount Factor

The choice of discount factor impacts the posterior analysis greatly. The formal rule derived from the theory of Bayesian model selection in choosing the best discount

factor is based on the posterior of δ , i.e. $p(\delta|\mathcal{D}_T)$. That is, the discount factor which maximizes this posterior is regarded as the optimal discount factor. The posterior is decomposed as $p(\delta|\mathcal{D}_T) \propto p(y_{1:T}|\delta)p(\delta)$ and, if we use the non-informative prior, or the constant prior defined as $p(\delta) \propto 1$, the target posterior is proportional to the marginal likelihood, $p(y_{1:T}|\delta)$. For fixed δ , the form of this density is analytically known as

$$p(y_{1:T}|\delta) = \prod_{t=1}^T p(y_t|D_{t-1}, \delta) = \prod_{t=1}^T NB(y_t|k_t, p_t),$$

where (k_t, p_t) are the parameters of one-step forecast distributions in eqn. (4.2). It is not easy to maximize this function in δ directly since the second derivative of the log density shows this function is not concave. Still, we can compute the value of this density across a discrete grid of values of δ to choose the best discount factor among the pre-specified finite set.

The process of choosing the discount factor is summarized as follows:

1. Define the finite set of discount factors $\Delta \subset (0, 1]$.
E.g., $\Delta = \{0.900, 0.901, \dots, 0.998, 0.999\}$.
2. For each $\delta \in \Delta$, compute $p(y_{1:T}|\delta)$ (or $\log p(y_{1:T}|\delta)$).
3. Add the prior information $p(\delta)$ to compute $p(\delta|\mathcal{D}_T)$ in case a non-constant prior is used.
4. Identify the posterior mode for δ .

The logarithm form is useful in computation. This is simplified as

$$\begin{aligned} \log p(y_{t+1}|\mathcal{D}_t, \delta) = \text{const.} + 1[y_{t+1} \neq 0] \sum_{n=1}^{y_{t+1}} \log(y_{t+1} + \delta c_t - n) \\ + \delta r_t \log(\delta c_t) - (\delta r_t + y_{t+1}) \log(1 + \delta c_t). \end{aligned} \quad (4.3)$$

4.2.3 Generalized Gamma Random Walk and Power Discount

We can extend the proposed gamma random walk as

$$\eta_t | \mathcal{D}_{t-1} \sim Be(\alpha_t, \beta_t), \quad \alpha_t + \beta_t = r_{t-1}.$$

Our model sets $\alpha_t = \delta r_{t-1}$ and $\beta_t = (1 - \delta)r_{t-1}$. For this extended model, the same filtering and smoothing is valid. The key feature is the constraint on the sum of two parameters, $\alpha_t + \beta_t = r_{t-1}$, which erases the term $\phi_{t-1}^{r_{t-1} - (\alpha_t + \beta_t)}$ in computing $p(\phi_t, \phi_{t-1} | \mathcal{D}_{t-1})$ and realizes the conjugacy.

This model includes power discounting for generalized state-space models proposed by Smith (1979). It defines the prior at t through the posterior at $t - 1$ as

$$p(\phi_t | \mathcal{D}_{t-1}) \propto \{p_{\phi_{t-1} | \mathcal{D}_{t-1}}(\phi_t)\}^\delta,$$

where $p_{\phi_{t-1} | \mathcal{D}_{t-1}}(\phi_t)$ in the right-hand-side means the density of $p(\phi_{t-1} | \mathcal{D}_{t-1})$ evaluated at ϕ_t . This is generally applicable to a wide class of state-space models, but, in the context of Poisson-gamma models, it means

$$p(\phi_t | \mathcal{D}_{t-1}) = Ga(\delta(r_{t-1} - 1) + 1, \delta c_{t-1}),$$

where the shape parameter is slightly different from ours. This indirect definition of the state evolution is interpreted as a special case of the extended gamma random walk with $\alpha_t = \delta(r_{t-1} - 1) + 1$ and $\beta_t = (1 - \delta)(r_{t-1} - 1)$. The problem with this approach is that the model is not well-defined in the sense of probability theory if $r_{t-1} \leq 1$ and $\beta_t \leq 0$. This problem might arise when the process of counts is sparse, yielding $y_t = 0$ at many time points. It is also required to have $r_0 > 1$ in the initial prior. In such a case, the use of power discounting is limited.

4.2.4 Extension to Multinomial-Dirichlet State-Space Models

Now we consider the multivariate observation of counts, defined by $y_t = \{y_{jt}\}_{j=1:J}$. While the individual count series can be usefully modeled by Poisson-Gamma mod-

els independently, the data might have additional information about the sum of outcomes as $\sum_{j=1}^J y_{jt} = n_t$. In such a case, the basic underlying sampling models is a multinomial distribution, $Multi(y_t|n_t, \theta_t)$ with θ_t on $(J - 1)$ -dimensional simplex,

$$p(y_t|\theta_t) = \binom{n_t}{y_{1t}, \dots, y_{Jt}} \prod_{j=1}^J \theta_{jt}^{y_{jt}}.$$

The conjugate prior for this likelihood is the Dirichlet distribution, $Dir(\theta_t|q_t)$, the density of which is

$$p(\theta_t|q_t) = \frac{\Gamma(q_t)}{\prod_{j=1}^J \Gamma(q_{jt})} \prod_{j=1}^J \theta_{jt}^{q_{jt}-1}, \quad \text{where } q_t = \sum_{j=1}^J q_{jt}.$$

Our interest is again in the state evolution of θ_t that preserves Dirichlet distribution as its marginal. For this purpose, we re-parametrize the model by the well-known decomposition of Dirichlet random quantities into gamma variables; if $\phi_{jt} \sim Ga(r_{jt}, c_t)$ independently across j , and if θ_t is defined as the ratio

$$\theta_{jt} = \frac{\phi_{jt}}{\sum_{j=1}^J \phi_{jt}},$$

then $\theta_t \sim Dir(q_t)$ where $q_{jt} = r_{jt} / \sum_{j=1}^J r_{jt}$. With this characteristic, we propose to define a multinomial-Dirichlet state-space model implicitly through multiple, independent Poisson-gamma models as constructed in the previous section. Denote the on-line posterior of each Poisson-gamma model by $\phi_{jt}|\mathcal{D}_t \sim Ga(r_{jt}, c_{jt})$ with discount factor δ_j . In order to recover the Dirichlet distribution, we need to set $c_{j0} = c_0$ and $\delta_j = \delta$ for all j , so that we have all the subsequent rate parameters to be common across j ; $c_{jt} = c_t$ for all j at any t . This common rate parameter/discount factor among observations implies our informational gain from known n_{it} against the use of completely independent Poisson-gamma models.

The resulting on-line posterior and forecast distribution are summarized below.

Forward Filtering of Multinomial-Dirichlet State-Space Models

With the prior $\theta_0 \sim Dir(q_0)$, the filtering proceeds as follows:

1. Posterior at $t - 1$:

$$\theta_{t-1} | \mathcal{D}_{t-1} \sim Dir(q_{t-1}).$$

2. Prior at t :

$$\theta_t | \mathcal{D}_{t-1} \sim Dir(\delta q_{t-1}).$$

3. Forecasting for t :

$$y_t | \mathcal{D}_{t-1} \sim MP(q_{t-1}),$$

known as multivariate Polya distribution, whose density is

$$p(y_t | \mathcal{D}_{t-1}, \delta) = \frac{n_t!}{\prod_{j=1}^J y_{jt}!} \frac{\Gamma(\delta q_{\cdot t})}{\prod_{j=1}^J \Gamma(\delta q_{jt})} \frac{\prod_{j=1}^J \Gamma(y_{jt} + \delta q_{jt})}{\Gamma(n_t + \delta q_{\cdot t})},$$

where $q_{\cdot t} = \sum_{j=1}^J q_{jt}$.

4. Posterior at t :

$$q_t | \mathcal{D}_t \sim Dir(q_t),$$

where $q_t = \delta q_{t-1} + y_t$.

For the evaluation of marginal likelihoods and posterior model probabilities, the forecast density is simplified as

$$p(y_t | \mathcal{D}_{t-1}, \delta) = \frac{n_t!}{\prod_{j=1}^J y_{jt}!} \frac{\prod_{j \in J_+} \prod_{m=1}^{y_{jt}} (y_{jt} + \delta q_{jt} - m)}{\prod_{m=1}^{n_t} (n_t + \delta q_{\cdot t} - m)},$$

where $J_+ = \{ j \in 1:J \mid y_{jt} > 0 \}$, the set of elements that has non-zero counts. The log-density is

$$\log p(y_t | \mathcal{D}_{t-1}, \delta) = \text{const.} + \sum_{j \in J_+} \sum_{m=1}^{y_{jt}} \log(y_{jt} + \delta q_{jt} - m) - \sum_{m=1}^{n_t} \log(n_t + \delta q_{.t} - m). \quad (4.4)$$

In filtering, it is hard to derive $p(\theta_t | \theta_{t-1}, \mathcal{D}_{t-1})$ explicitly since the state evolution of θ_t is implicitly defined through independent gamma random walks. Unfortunately, this aspect becomes problematic in the retrospective analysis. Alternatively, we can simply keep the latent gamma parameters (r_{jt}, c_t) , sample ϕ_{jt} from its retrospective posterior, and re-construct Dirichlet parameters θ_t . The problem of this approach is the dependence on the choice of c_0 , which originally does not exist in multinomial-Dirichlet models. However, as discussed in Section 4.2.1, the effect of c_0 on the posteriors is limited. The resulting retrospective posterior of θ_t is expected to be robust to the choice of c_0 as it should be in the original multinomial-Dirichlet structure that does not have that rate parameter.

4.3 Dynamic Gravity Models

Consider the time series of a square matrix of counts: $Y_t = \{y_{ijt}\}_{i,j=1:I}$ where $y_{ijt} \in \mathbb{Z}_+$. In application, y_{ijt} represents the flow from node i to j at time t on the network. In modeling this type of observation, it is natural to consider the dependence between two flows, or observations on two different edges in the network, that share the common origin i or the destination j . One framework that has been developed for static networks is that of *gravity models*, in which the likelihoods are independent Poisson distributions conditional on rate parameters that capture the network dependence in the form of Analysis of Variance (ANOVA). To be precise, and making the novel extension of traditional, static gravity models to a dynamic context with potentially time-varying parameters, the dynamic gravity model (DGM)

is written as

$$y_{ijt} | \phi_{ijt} \sim Po(\phi_{ijt}), \quad (4.5)$$

independently across i, j and t , and

$$\phi_{ijt} = \mu_t \alpha_{it} \beta_{jt} \gamma_{ijt}, \quad (4.6)$$

where $\{\mu_t, \alpha_{it}, \beta_{jt}, \gamma_{ijt}\}$ are called overall effect, origin effect of i , destination effect of j and individual effect, or affinity, of flow from i to j , respectively. Models of this and more elaborate forms have been popular in transportation studies (e.g. West, 1994; Sen and Smith, 1995) where the interaction term is typically structured as a function of physical distance between nodes; there the “gravity model” terminology relates to the role of small distances in defining large interactions and hence “attraction” of traffic from node i to node j . We refer to the γ_{ijt} interactions as “affinities.” In dissecting the network flow activity, we are most interested in questions about which affinities are greater than one (j attracts flow from i over and above the main effects), or less than one (j is relatively unattractive to i), or not significantly different to one (neutral). Critically, affinities are time-varying, and any identified patterns of variation over time may be related to interpretable events or network changes.

In the first fully Bayesian approach to gravity models using MCMC methods, West (1994) developed such models in the static case; i.e., with no dynamics in the model parameters, and applied the model to a large transportation flow network. Congdon (2000) explored a similar approach in studies of patient flows to hospitals. Analysis via MCMC is computationally very demanding, and the burden increases quadratically in I , and inherently non-sequentially. More recently, Jandarov et al. (2014) studied such models for spread of infectious diseases, and used Gaussian process approximations for approximate inference rather than full MCMC or other computational methods.

We share the spirit of this latter work, in using the simply and efficiently implemented BDFM as a path to fitting the gravity model—now extended to time-varying effect parameter processes. However, we do not constrain the affinity parameters γ_{ijt} as a function of covariates of any kind, simply treating the DGM as a dynamic, random effects model. This leads to a *direct* parameter mapping between the BDFM to the DGM; as a result, the trivially generated simulations from the full posterior of the BDFM are mapped directly to full posterior samples from the DGM, providing immediate access to inference on main effect and affinity processes over time.

While the models and analyses developed here represent new applications of Bayesian dynamic modeling ideas, there are at least conceptual intersections with studies of monitoring, inference, and forecasting traffic flows in areas. Some of the developments here may well extend to apply to such areas, including, for example, inference in origin-destination analysis (e.g. Tebaldi and West, 1998) and physical traffic studies (e.g. Tebaldi et al., 2002; Queen and Albers, 2009; Anacleto et al., 2013a,b).

4.3.1 Identification

In emulating DGM by BDFMs, specifically, we *independently* apply the Poisson-gamma model for each flow y_{ijt} , or the multinomial-Dirichlet model for the set of flows, say $(y_{i1t}, \dots, y_{iIt})$ if conditioned by $n_{it} = \sum_{j=1:I} y_{ijt}$, in order to obtain particles $\{\phi_{ijt}\}$ from the retrospective posterior and map them to those of DGM. Note that we can sample these particles in parallel and the computational complexity is at most order $\mathcal{O}(T)$, while we completely ignore the dependency of flows on the network in applying independent BDFMs. The mapping of those particles to the gravity model components, $\{\mu_t, \alpha_{it}, \beta_{jt}, \gamma_{ijt}\}$, can be done in a deterministic way, without adding any significant computations, as shown in this and subsequent subsections. This enables the construction of posterior distributions of gravity model components

based on simulation in a feasible way under the scalable network. We expect that, though we ignore the dependence of nodes in BDFM emulators, we are able to see dependencies in the implied DGM posteriors in terms of the “significant” values of μ_t , α_{it} and β_{jt} , recovering the degree of dependency through their posterior distributions.

In this subsection, the theory to convert the Poisson mean parameters to the components of gravity models, or mapping BDFMs to DGM, is explained. As we saw, these two representations of parameters are tied in eqn. (4.6) and on the log-scale,

$$f_{ijt} = m_t + a_{it} + b_{jt} + g_{ijt}, \quad (4.7)$$

where $f_{ijt} = \log \phi_{ijt}$, $m_t = \log \mu_t$, $a_{it} = \log \alpha_{it}$, $b_{jt} = \log \beta_{jt}$ and $g_{ijt} = \log \gamma_{ijt}$. At each time t , the number of parameters, $\{\phi_{ijt}\}_{i,j=1:I}$, in the emulator is I^2 , while the DGM representation has $\{\mu_t, \{\alpha_{it}\}_{i=1:I}, \{\beta_{jt}\}_{j=1:I}, \{\gamma_{ijt}\}_{i,j=1:I}\}$, where the number of elements is $(I+1)^2$. Thus, the DGM is over-parameterized against the emulators from which we sample particles. To map the parameters of emulators to those of DGM *uniquely*, we need additional constraints on parameters and decrease the number of free parameters, or degrees of freedom.

4.3.2 Identification by a Reference Flow

In the additive representation in eqn. (4.7), the similarity to a traditional linear ANOVA model is evident; this suggests the use of the traditional methods for identification. One straightforward set of constraints for identification, used frequently in ANOVA models, is to set some parameters to be constant as the reference level. For example,

$$\alpha_{1t} = 1, \quad \beta_{1t} = 1, \quad \gamma_{1jt} = 1, \quad \gamma_{i1t} = 1$$

for $i, j = 1:I$. By these constraints, $1 + 1 + (2I - 1) = 2I + 1$ parameters become constrained, and the total number of free parameters in the model is now I^2 . The

one-to-one mapping is obtained by the following equations,

$$\begin{aligned}\phi_{11t} &= \mu_t, \\ \phi_{i1t} &= \mu_t \alpha_{it}, & i = 2 : I \\ \phi_{1jt} &= \mu_t \beta_{jt}, & j = 2 : I \\ \phi_{ijt} &= \mu_t \alpha_{it} \beta_{jt} \gamma_{ijt}, & i, j = 2 : I,\end{aligned}$$

or, equivalently,

$$\begin{aligned}\mu_t &= \phi_{11t}, \\ \alpha_{it} &= \phi_{i1t} / \mu_t, & i = 2 : I \\ \beta_{jt} &= \phi_{1jt} / \mu_t, & j = 2 : I \\ \gamma_{ijt} &= \phi_{ijt} / \mu_t \alpha_{it} \beta_{jt}, & i, j = 2 : I,\end{aligned}$$

in addition to $\alpha_{1t} = \beta_{1t} = \gamma_{11t} = \gamma_{1jt} = 1$. This transformation is easy to implement, but it makes the interpretation of each parameter more complicated. For instance, the overall level, μ_t , is defined only by ϕ_{11t} . This means the flow from domain 1 to 1 is now considered the reference level for the other flows, i.e., all the other parameters, α_{it} , β_{jt} and γ_{ijt} , are understood as the deviation from this reference. Therefore, the order of domains, or the choice of the reference flow, is crucial in interpretation. Plus, of course, once the order of domains is changed, the posteriors of the DGM parameters become different.

4.3.3 Identification by Geometric Means

Another set of constraints consists of the geometric means of DGM parameters as

$$\begin{aligned}\sum_{i=1}^I a_{it} = 0, & \quad i = 1:I; & \sum_{j=1}^I b_{jt} = 0, & \quad j = 1:I; \\ \sum_{j=1}^I g_{ijt} = 0, & \quad i = 1:I; & \sum_{i=1}^I g_{ijt} = 0, & \quad j = 1:(I-1).\end{aligned}$$

In total, we have $1 + 1 + I + (I - 1) = 2I + 1$ equations, and the degrees of freedom decreases by $2I + 1$ and becomes I^2 . Note that, even for $j = I$, these equations imply the same sum-to-zero constraint as

$$\sum_{i=1}^I g_{iI} = \sum_{i=1}^I \sum_{j=1}^{I-1} \{ -g_{ijt} \} = \sum_{j=1}^{I-1} \left\{ - \sum_{i=1}^I g_{ijt} \right\} = \sum_{j=1}^{I-1} \{ -0 \} = 0.$$

Next, use those constraints in the definition of DGM parameters in eqn. (4.7), taking the sum in i, j and both to obtain the equations relating ϕ_{ijt} to DGM parameters as follows:

$$\sum_{i=1}^I \sum_{j=1}^I f_{ijt} = I^2 m_t, \quad \sum_{j=1}^I f_{ijt} = I m_t + I a_{it}, \quad \sum_{i=1}^I f_{ijt} = I m_t + I b_{jt}.$$

Thus, DGM parameters are obtained recursively by

$$m_t = \frac{1}{I^2} \sum_{i=1}^I \sum_{j=1}^I f_{ijt}, \quad a_{it} = \frac{1}{I} \sum_{j=1}^I f_{ijt} - m_t,$$

$$b_{jt} = \frac{1}{I} \sum_{i=1}^I f_{ijt} - m_t, \quad g_{ijt} = f_{ijt} - m_t - a_{it} - b_{jt}.$$

This transformation has several advantages in interpretation. First, thanks to the sum-to-zero constraints, the signs and magnitudes of domain-specific effects, a_{it} and b_{jt} , mean how they deviate from the average of all the domains. Second, unlike the previous transformation, this is invariant to the order of domains. For these reasons, this transformation is preferred.

4.3.4 Computational Problems in Practice

To implement the transformation above, we need to take the logarithm of the Poisson rate parameters as in eqn. (4.7), and this might be problematic in an applied study. The problem arises in the case where some of the flows are “sparse”, i.e. $y_{ijt} = 0$ for

many t in the flow from i to j . In the extreme case where $y_{ijt} = 0$ for all t , the posterior of the Poisson-gamma model at $t = T$ is $Ga(\delta^T r_0, \delta^T c_0 + (1 - \delta^{T-1})/(1 - \delta))$. For large T , this distribution is nearly degenerate at zero, having almost all probability mass in the neighborhood of the origin. In computation, generating random variables from the gamma distribution with small shape parameter is not straightforward. While we can generate the log scale particles by using the importance sampling method of Liu et al. (2013), once it is converted to the exponential scale it become numerically zero.

This is not desirable if the flow might become active again after a long period of sparsity. In such a case, we implicitly assume that the model should preserve the uncertainty in its posterior and avoid being overly confident. This aspect of our belief is realized in the Poisson-gamma model by the additional constraint on filtering as

$$r_{ijt} = r \quad \text{if} \quad \delta r_{ij,t-1} + y_{ijt} < r, \quad (4.8)$$

for some r . In other words, when the scale parameter of the on-line posterior becomes below the threshold r , we abandon the original updating rule but keep the shape parameter on the threshold. This exceptional treatment is understood as an intervention on the process defined by the model, and included in the set of information \mathcal{D}_t . In Section 4.5, we use $r = 0.1$.

More importantly, another technical problem with this mapping arises in cases of sparse flows, in which the very small values of concentration of the posterior for ϕ_{ijt} unduly impacts the resulting overall mean and/or origin or destination means. To see this, suppose $y_{ijt} = 0$ for many pairs of (i, j) and we sampled very small ϕ_{ijt} for those (i, j) . On taking logarithms, $f_{ijt} = \log \phi_{ijt}$ tends to be negative and large. Since the definition of m_t is the average of all the f_{ijt} , those large and negative values of f_{ijt} make the overall level $\mu_t = e^{m_t}$ extremely small. With this small μ_t , the other parameters give little insight. For example, the affinity of sparse flow γ_{ijt} has to be

extremely large to counter the smallness of $\{\mu_t, \alpha_{it}, \beta_{jt}\}$ to balance them to the value of ϕ_{ijt} .

While one can imagine model extensions to address this, at a practical level it suffices to adjust the mapping as is typically done in related problems of log-linear models of contingency tables with structural zeros. This is implemented by simply restricting the summations in the identifiability constraints to those node pairs for which $y_{ijt} > d$, for some small d , and adjusting the divisors to count the numbers of terms in each summation. With this adjustment, very small ϕ_{ijt} appropriately lead to very small affinities γ_{ijt} , so the latter then represent very sparse flows. We will see this part in detail in the next subsection.

4.3.5 *Alternative Identification Strategy under Sparsity*

The DGM can account for zero counts on a certain edge in different ways, but, for our purpose of analysis, it is meaningful to let γ_{ijt} be very small, so that the other parameters can be interpreted as the levels of “active” flow. To allow for this interpretation, we restrict the decomposition of Poisson rate parameters into DGM components by adding several constraints (Bishop et al., 1975, chap. 5). First, we assume that a flow $\{y_{ijt}\}_{t=1:T}$ is sparse if $\sum_{t=1:T} y_{ijt} < d$ where d is the pre-specified threshold that should be a small value. In Section 4.5, this threshold is set to be $d = 10$. Define $s_{ij} = 0$ if the flow from i to j is sparse, and $s_{ij} = 1$ otherwise. Next, we modify the identification constraints as

$$\begin{aligned} \sum_{i=1}^I s_{i \cdot} a_{it} &= 0, \quad i = 1:I; & \sum_{j=1}^I s_{\cdot j} b_{jt} &= 0, \quad j = 1:I; \\ \sum_{j=1}^I s_{ij} g_{ijt} &= 0, \quad i = 1:I; & \sum_{i=1}^I s_{ij} g_{ijt} &= 0, \quad j = 1:(I-1), \end{aligned}$$

where $s_{i \cdot} = \sum_{j=1}^I s_{ij}$ and $s_{\cdot j} = \sum_{i=1}^I s_{ij}$. The indicator variables introduced above allow us to exclude Poisson rate parameters of sparse flows from the definition of the

overall mean and the average of origin/desination effects. The set of parameters, μ_t , α_{it} and β_{jt} , can then be computed by solving the linear equations

$$\begin{aligned} \sum_{i,j=1:I} s_{ij} f_{ijt} &= s_{..} m_t + \sum_{i=1}^I s_{.j} a_{it} + \sum_{j=1}^I s_i b_{jt}, \\ \sum_{j=1:I} s_{ij} f_{ijt} &= s_i m_t + s_i a_{it} + \sum_{j=1}^I s_{ij} b_{jt}, \quad i = 1:I, \\ \sum_{i=1:I} s_{ij} f_{ijt} &= s_{.j} m_t + \sum_{j=1}^I s_{ij} a_{it} + s_{.j} b_{jt}, \quad j = 1:I, \end{aligned}$$

where $s_{..} = \sum_{i,j=1:I} s_{ij}$. This linear system is solvable in DGM components if $s_i > 0$ and $s_{.j} > 0$, and we assume this is true in most cases. Once these three types of parameters are obtained, the interaction effects are defined as $g_{ijt} = f_{ijt} - m_t - a_{it} - b_{jt}$. If a flow is sparse, its log-rate f_{ijt} is negatively large, making its log-affinity g_{ijt} to be negative and large as well, and γ_{ijt} to become dominant in ϕ_{ijt} .

4.4 Dependent Gravity Model and MCMC

In Section 4.3, the direct analysis of the DGM is mentioned with its difficulty in computation under scalable datasets. Here, one of those approaches is examined in detail to see its specification, difficulty, and problem.

We now put priors on the four types of state variables in eqn. (4.6) that comprise the state-space model. Let $\varphi_t \in \{\mu_t, \alpha_{it}, \beta_{it}, \gamma_{ijt}\}$. A natural choice for this latent process is the log-normal AR(1) process,

$$\log \varphi_t = \mu_\varphi + \phi_\varphi (\log \varphi_{t-1} - \mu_\varphi) + N(0, \sigma_\varphi^2),$$

as in the static generalized linear model. Alternatively, the same gamma random walk can be used for each component as

$$\varphi_t = \varphi_{t-1} \eta_t^\varphi / \delta^\varphi, \tag{4.9}$$

with the appropriate distribution for η_t^φ . The advantage of this model is the availability of forward filtering in Gibbs sampling, i.e., conditional on the other DGM components, all the state variables of interest, $\varphi_{1:T}$, can be sampled simultaneously. To take an example, the full conditional posterior of $\alpha_{i,1:T}$ for some i depends on the likelihoods of y_{ijt} for $j = 1:I$, so its kernel is

$$\prod_{j=1:I} Po(y_{ijt}|\phi_{ijt}) \propto \alpha_{it}^{\sum_{j=1:I} y_{ijt}} \exp \left\{ -\alpha_{it} \sum_{j=1:I} \mu_t \beta_{jt} \gamma_{ijt} \right\},$$

and this likelihood contributes to the shape and rate parameters of the on-line posterior by $\sum_{j=1:I} y_{ijt}$ and $\sum_{j=1:I} \mu_t \beta_{jt} \gamma_{ijt}$, respectively. Note also that the retrospective analysis is exactly the same as Poisson-gamma BDFMs. The full analysis by MCMC proceeds as follows.

MCMC for Dependent DGM:

The parameters of interest are the state variables $\varphi_t \in \{\mu_t, \alpha_{1:I,t}, \beta_{1:I,t}, \gamma_{1:I,1:I,t}\}$.

1. Set priors, $\varphi_0 \sim G(r_0^\varphi, c_0^\varphi)$, and discount factors δ^φ .
2. At each iteration of MCMC, for each DGM component φ_t conditional on the others, implement FFBS. In filtering, $\varphi_t \sim G(r_t^\varphi, c_t^\varphi)$, where

- $\varphi_t = \mu_t$:

$$r_t^\mu = \delta^\mu r_{t-1}^\mu + \sum_{i=1}^I \sum_{j=1}^I y_{ijt}, \quad c_t^\mu = \delta^\mu c_{t-1}^\mu + \sum_{i=1}^I \sum_{j=1}^I \alpha_{it} \beta_{jt} \gamma_{ijt}.$$

- $\varphi_t = \alpha_{it}$:

$$r_t^{\alpha_i} = \delta^{\alpha_i} r_{t-1}^{\alpha_i} + \sum_{j=1}^I y_{ijt}, \quad c_t^{\alpha_i} = \delta^{\alpha_i} c_{t-1}^{\alpha_i} + \sum_{j=1}^I \mu_t \beta_{jt} \gamma_{ijt}.$$

- $\varphi_t = \beta_{jt}$:

$$r_t^{\beta_j} = \delta^{\beta_j} r_{t-1}^{\beta_j} + \sum_{i=1}^I y_{ijt}, \quad c_t^{\beta_j} = \delta^{\beta_j} c_{t-1}^{\beta_j} + \sum_{i=1}^I \mu_t \alpha_{it} \gamma_{ijt}.$$

- $\varphi_t = \gamma_{ijt}$:

$$r_t^{\gamma_{ij}} = \delta^{\gamma_{ij}} r_{t-1}^{\gamma_{ij}} + y_{ijt}, \quad c_t^{\gamma_{ij}} = \delta^{\gamma_{ij}} c_{t-1}^{\gamma_{ij}} + \mu_t \alpha_{it} \beta_{jt}.$$

Sampling of α_{it} , β_{jt} and γ_{ijt} can be done in parallel across i , j and (i, j) , respectively.

3. The sampled parameters are not constrained. That is, the model is augmented with an excess number of parameters. For identification, modify the original particles as

- $g_{ijt}^* = g_{ijt} - g_{i\cdot t} - g_{\cdot jt} + g_{\cdot\cdot t}$,
- $a_{it}^* = [a_{it} + g_{i\cdot t}] - \frac{1}{I} \sum_{i'=1}^I [a_{i't} + g_{i'\cdot t}]$,
- $b_{jt}^* = [b_{jt} + g_{\cdot jt}] - \frac{1}{I} \sum_{j'=1}^I [b_{j't} + g_{\cdot j't}]$,
- $m_t^* = m_t - g_{\cdot\cdot t} + \frac{1}{I} \sum_{i=1}^I [a_{it} - g_{i\cdot t}] + \frac{1}{I} \sum_{j=1}^I [b_{jt} - g_{\cdot jt}]$,

where $a_{\cdot t}$, $b_{\cdot t}$, $g_{\cdot jt}$, $g_{i\cdot t}$ and $g_{\cdot\cdot t}$ are the means in i and/or j . The new particles above satisfy the constraints for identification. This transformation is easily adjusted to the case of identification under the constraints by sparse flow discussed in Section 4.3.4.

4.5 Application: Analysis of FoxNews Website Access Records

4.5.1 Study of Internet Traffic Flow

We focus on flows of visitors to a set of domains of webpages in a structurally well-defined but dynamic/evolving network. The network is a subset of domains in the

FoxNews website, monitored to generate streaming data on visitors to each domain over time. We use the term “domain” loosely, to refer to collections of webpages that correspond to categories of interest, as defined by Fox News.

On-line advertisers are interested in a host of statistical issues related to traffic flow and domain content. The field has become quite sophisticated, employing complex recommender systems (Koren et al., 2009), sentiment analysis (Pang and Lee, 2008), text mining (Soriano et al., 2013), and other methods (Agarwal et al., 2010; Taddy, 2013). However, basic questions of understanding and characterizing traffic across domains have not received the attention they require.

Trajectories of users, and groups of users, can give important information about browsing intent and the evolution of purchase interest, and thus ultimately affect ad placement decisions. As pages within a website are updated, questions arise as to whether browsing traffic patterns change as a result. To begin to address this statistically, we need to understand stochastic variation in past browser traffic so that comparisons can be made of incoming traffic streams against recent statistical “norms,” and significant deviations from short-term predictions based on current dynamic patterns can be identified.

4.5.2 Context and Data

Modeling internet traffic flow is a Big Data problem. We necessarily focus on smaller defined networks for which there is contextual knowledge to guide interpretation. Our context is traffic flow among “domains” of the FoxNews website. Besides the home/landing page, other domains include Politics, Entertainment, Travel, Science, and similar broad categories of news content. The contents of the domains change, usually on a daily basis at midnight, but more rapidly when some noteworthy event occurs. MaxPoint places ads on pages in these FoxNews domains, and thus can track flows of anonymized users as they move through its pages.

Data

The dataset contains FoxNews website visit data during 09:00-10:00am and 01:00-02:00pm EST on each of six days, February 23-24, March 2-3 and 9-10, 2015. These days are Mondays or Tuesdays. Since the FoxNews website structure changes often, with new pages being added and old pages being archived, the analysis aggregates webpages into groups specified by the host domain `www.foxnews.com`, and the set of first url paths after the host domain, including examples such as e.g. `www.foxnews.com/politics/*` and `www.foxnews.com/US/*`. These classify all the pages into 22 domains: Homepage, Politics, US, Opinion, Entertainment, Technology, Science, Health, Travel, Leisure, World News, Sports, Shows, Weather, Category, Latino, Story, On-Air, Video, National News, Magazine, and Other.

The dataset includes anonymized users (browsers) from nearly every time zone on the planet. In order to study time-of-day effects, such as, say, a tendency to browse news in the morning and entertainment in the afternoon, it is necessary to stratify by time zone. Here we focus on users in the Eastern North America time zone; those are the most numerous, and the two time windows used in this study were chosen with the expectation that different browsing patterns might occur at those times.

We aggregate data to counts in half-minute intervals to form a time series with 120 time points showing domain occupancy, flows from each domain, and flows into each domain for each period. Several details are relevant. First, in each half minute interval, if the record shows the same user in two or more domains, then each of her/his moves is counted in the flow data into each of these domains. Second, if the user refreshes the same page multiple times spanning more than one time interval, then s/he is counted as simply staying in that domain; this can be done as the web browsing tool performs automatic refresh. Importantly, if a user stays in the same domain for more than five minutes, we assume s/he is no longer active, and

is counted as leaving the FoxNews site. If such a user later appears in one domain, s/he counts as in-flow from outside the FoxNews site. Finally, we cannot track user information either before or after the one-hour observation window, so there is a form of censoring; we thus restrict attention to the period 09:05 – 09:55am and 01:05 – 01:55pm, consisting of uncensored flows, using the first 5 minutes of data informally to define priors. Thus the series runs from $t = 1:T$ with $T = 110$ in each time period.

Network Structure and Notation

Referring to sites external to the FoxNews website as node 0, we have 23 network nodes; the $I = 22$ actual domains and “External”, indexed as $i = 0:I$. At each time $t = 1:T$, define n_{it} as the number of occupants of node i , and define y_{ijt} as the flow count from node i to j , including the in-flows y_{0it} and out-flows y_{i0t} relative to the External domain. The flows outside Foxnews, y_{00t} , are unobservable and missing. The schematic chart in Figure 4.1 reflects our notation.

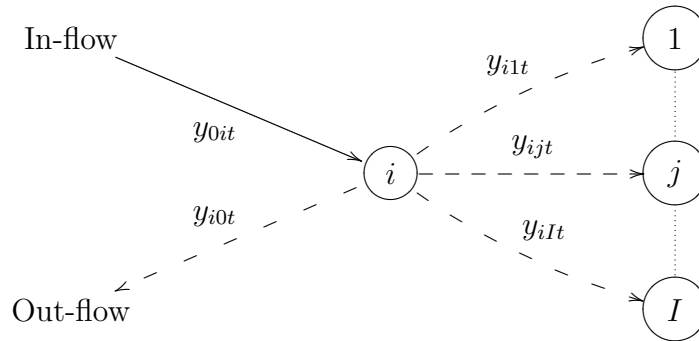


FIGURE 4.1: Network schematic and notation for flows at time t . The circular numbered nodes represent the domains at the current time point t . Each arrow $i \rightarrow j$ shows the flow of y_{ijt} visitors.

4.5.3 BDFM Analysis of FoxNews Data

Consider now the set of flows departing from node i . Before observing the flow at time t , the population on this domain is known as the sum of flows incoming into i at the previous time point $t - 1$, denoted by $n_{it} = \sum_{j=0}^I y_{ij,t-1}$. Thus, the subsequent move of visitors at t must be constrained by this known number of population, which justifies the multinomial likelihood

$$y_{i,0:I,t} \text{ ind} : \sim \text{Multi}(n_{it}, \theta_{it}) \quad i = 1:I.$$

We can model the evolution of θ_{it} by introducing the latent rate parameters $\phi_{i,0:I,t}$ that recovers θ_{it} as their ratios as discussed in Section 4.2.4. In contrast, there is no available information on the population in the external websites that is considered in modeling the in-flow y_{0jt} for $j = 1:I$. Naturally, the Poisson-gamma models in Section 4.2 are used for the in-flows as

$$y_{0jt} \text{ ind} : \sim \text{Poi}(\phi_{0jt}) \quad j = 1:I.$$

The state variables are the rate parameters and follow gamma random walk model in eqn. (4.1), associated with discount factors δ_i for θ_{it} (or $\phi_{i,0:I,t}$) and δ_{0jt} for ϕ_{0jt} .

The posterior analysis is conducted separately for each of the 12 datasets of the six days. We analyze the morning of the first day (February 23 2015) for illustration of parameter learning before the comparison across days. In each dataset and analysis, the gamma priors for the in-flow rates are $\phi_{i0}|y_{0i0} \sim Ga(r_{i0}, c_{i0})$ with $c_{i0} = 1$ and $r_{i0} = c_{i0}z_i$ where z_i is the point estimate based on the in-flows of the first 5 minutes (prior to $t = 1$). The Dirichlet prior for the transition probabilities is $\theta_{i,0:I,0}|y_{i,0:I,0} \sim \text{Dir}(q_{i,0:I,0})$ where each q_{ij0} is a simple point estimate based on the first 5 minutes in-flows but rescaled so that $q_{i0:I0}$ becomes a probability vector. The priors for the underlying, unconstrained node-node flow rates are then $\phi_{ij0}|y_{0ij0} \sim Ga(r_{ij0}, c_{i0})$

with $c_{i0} = 1$ and $r_{ij0} = q_{ij0}$. The threshold for the shape parameters is set to be $r = 0.1$.

Priors for each discount factor are $Be(19, 1)$ distributions truncated to $(0.9, 1)$ with grid 0.001; reanalysis using uniform priors on this range led to little in the way of noticeable differences, as the marginal likelihoods at $t = T$ dominate. The prior truncation ensures some degree of smoothness. Running the models in parallel across discrete grids of discount factors and evaluating the marginal likelihoods in eqns. (4.3,4.4) at each time point t gives the marginal posteriors of discount factors. Across all in-flows, this suggested $\delta_i \approx 0.9$ as a posterior mode. Figure 4.2 plots posteriors truncated at 0.9 for the discounts δ_i in the transition flow models. Some nodes exhibit higher volatility in flows to other nodes over time than others, requiring and hence favoring smaller discount factors; these are particularly associated with

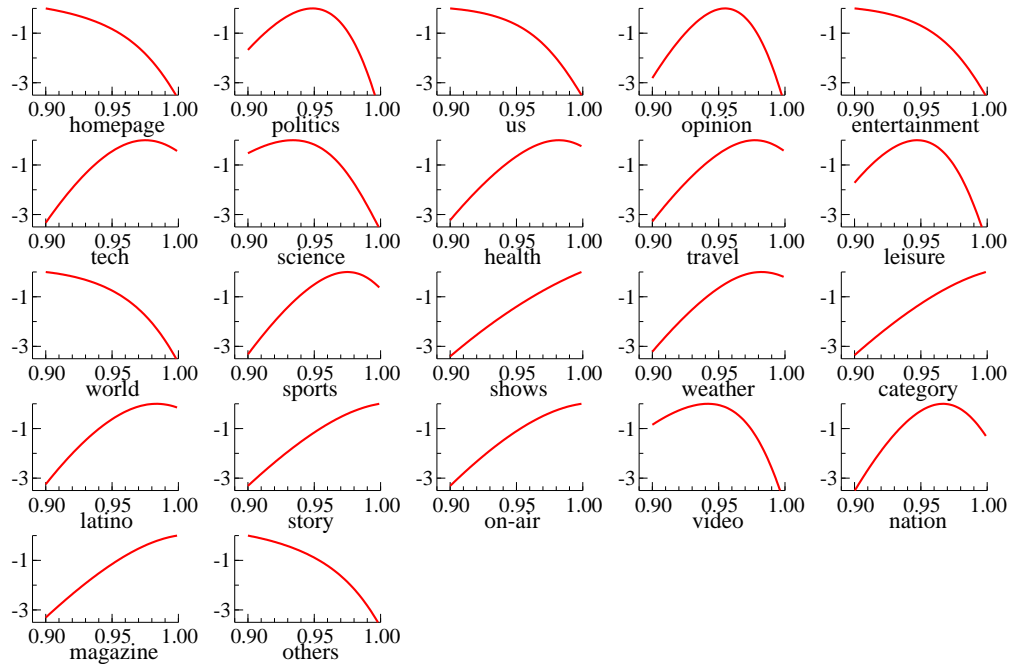


FIGURE 4.2: The standardized values of the marginal log-posteriors of discount factors δ_i for the in-flows to FoxNews nodes $i = 1:22$ (top right, reading across rows), for the period 09:05am-09:55am on February 23, 2015.

nodes representing domains with high flow counts (e.g., in-flows from External to Homepage).

Some summary inferences on selected model components are reported, based on models with discount factors fixed at their posterior modes. Figure 4.3 gives one example of learning about in-flow rates, in this case the flow from all nodes to node 10 (the Leisure domain). The figure exemplifies sequential learning about the flow rate together with its retrospectively updated trajectory and a visual assessment of one-step ahead forecasting aligned with the data.

A similar display in Figure 4.4 highlights the same aspects of the analysis, now with an example of flow between two network nodes (from Homepage to the Politics domain). It shows the rate between two nodes together with the retrospective smoothing for full inference on the trajectory and one-step ahead forecast summaries.

Homepage is the most popular single domain on FoxNews, so the transition probabilities from Homepage to other domains are of particular interest. Figure 4.5 shows that most Homepage visitors stay on Homepage for a while. Of those that leave, many exit the FoxNews site entirely. Across all six days, the probability of staying on the Homepage decreases over the course of the 50 minute morning period.

On examples of transition probabilities, Figure 4.6 shows that the probability of people leaving the FoxNews website from Homepage increases in this 50 minute window for each of the six mornings. Note that there are significant day effects; e.g., visitors are more likely to leave FoxNews on the morning of March 9th compared to the other mornings. More detailed insights, based on the gravity model, are noted in the next section.

As an illustration of a more detailed analysis of a very specific flow, consider Figure 4.7. Among the visitors who leave the Homepage for other FoxNews domains, Entertainment is generally the most popular destination. For the six datasets collected during the morning, we see large differences in transition probabilities; in

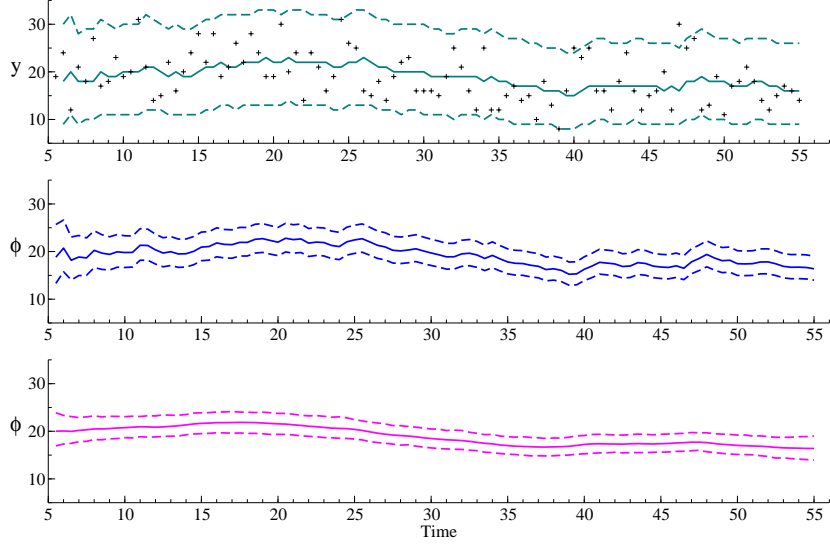


FIGURE 4.3: BDFM-based inference over time t on in-flows to domain $i = 10$ (Leisure). *Upper:* data y_{0Xt} (circles) with one-step ahead forecast means and 95% intervals. *Center:* trajectory of mean and 95% intervals from on-line posteriors $p(\phi_{0Xt}|y_{0X,1:t})$ plotted against t . *Lower:* revised trajectory under full retrospective posterior $p(\phi_{0Xt}|y_{0X,1:T})$.

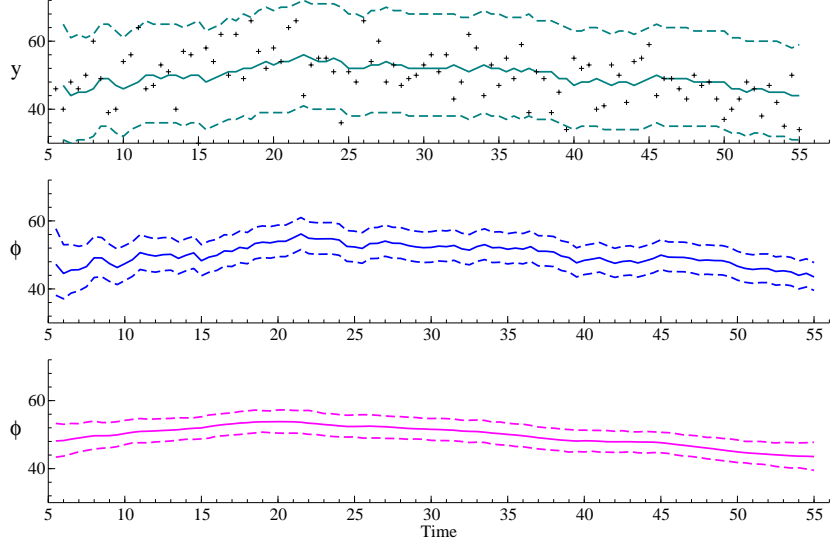


FIGURE 4.4: BDFM-based inference over time t on transitions from domain $i = 1$ (Homepage) to $j = 2$ (Politics). *Upper:* data y_{12t} (plus signs) with one-step ahead forecast means and 95% intervals. *Center:* trajectory of mean and 95% intervals from on-line posteriors of ϕ_{12t} . *Lower:* revised trajectory under full retrospective posterior.

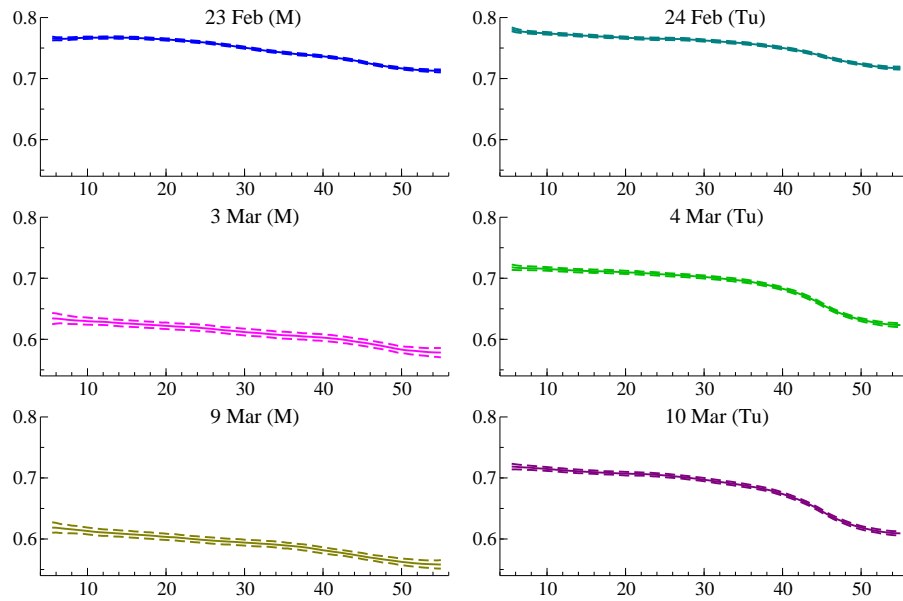


FIGURE 4.5: Retrospective mean and 95% CI of trajectories of transition probability θ_{11t} (staying at Homepage) from analysis on data collected from each of the six mornings.

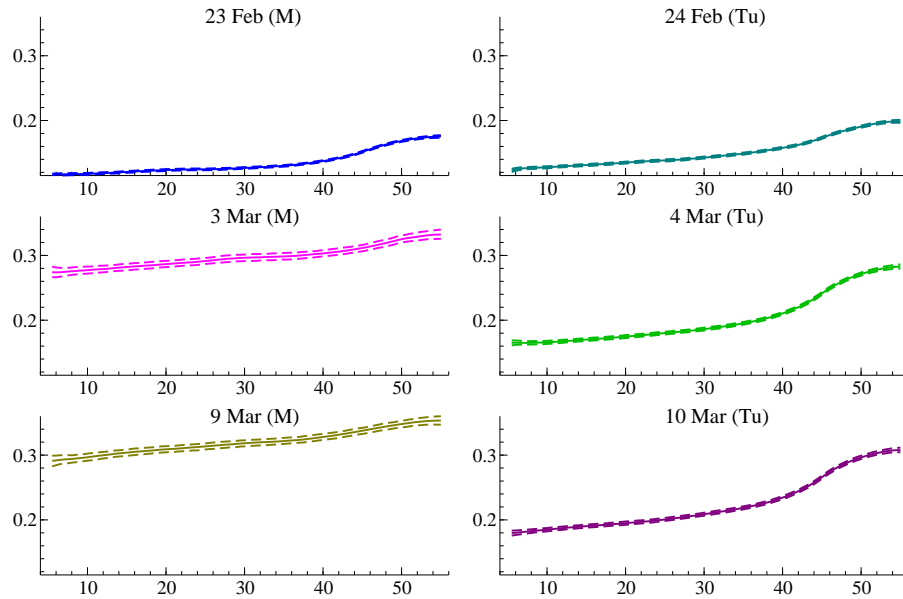


FIGURE 4.6: Retrospective mean and 95% CI of trajectories of transition probability θ_{10t} (Homepage \rightarrow External) from analysis on each of the six mornings.

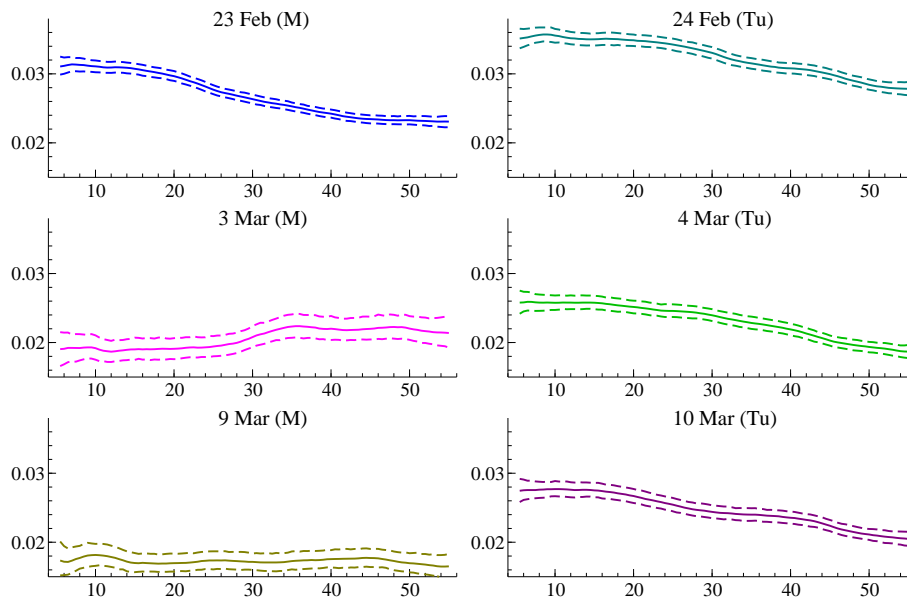


FIGURE 4.7: Retrospective mean and 95% CI of trajectories of transition probabilities θ_{15t} (Homepage \rightarrow Entertainment) for each of the six mornings.

particular February 23 and 24 have larger rates than the other days. It is noteworthy that the Academy Awards ceremony was held on the night of February 22, which may have driven this uptick.

4.5.4 DGM Analysis of FoxNews Data

We now consider DGM as the more nuanced model that structures flow rates in terms node-specific main effects and node-node interaction terms. With the sampled rate parameters $\{\phi_{ijt}\}_{i,j=1:I}$, it is straightforward to apply the mapping from BDFMs to DGM in Section 4.3. The rate parameter of the missing flow, ϕ_{00t} , is naturally excluded from the mapping procedure by setting $s_{00} = 0$.

February 23 2015, 09:00-10:00am

We first apply the gravity model decomposition to the morning data on February 23rd. Each flow is analyzed by the dynamic Poisson-Gamma and multinomial-Dirichlet models independently, under the same settings used in Section 4.5.3. The

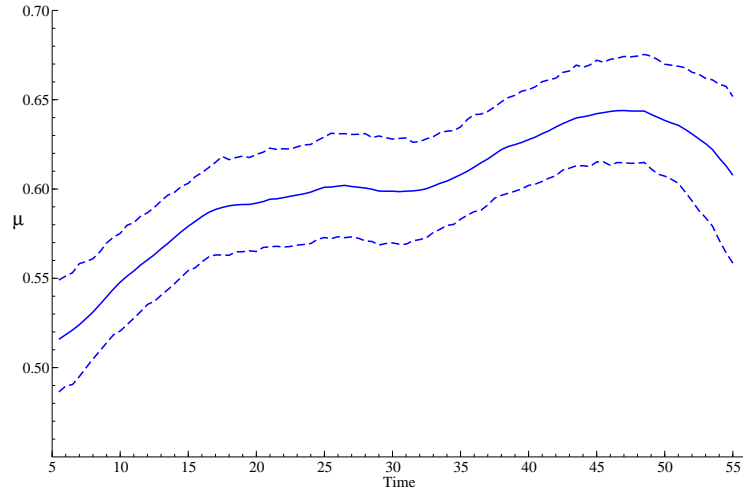


FIGURE 4.8: DGM-based smoothed trajectory of baseline level process $\mu_{1:T}$. In this and following figures, the dashed lines indicate 95% intervals about the displayed posterior mean trajectory.

particles of Poisson rate parameters sampled from their posteriors are decomposed into the gravity model components, which allows us to construct the posteriors for the gravity model parameters.

Figure 4.8 shows the retrospective posterior of the baseline level μ_t . Its posterior mean is almost stable around 0.60 throughout the fifty minute period and its fluctuation is at most 0.10, or a 10% increase of its effect to the total access counts.

The origin and destination effects are shown in Figures 4.9 and 4.10. The posteriors for origin effects show that large-scale domains, such as Homepage (domain 1), have higher values of α_{it} , while domains with low or zero flows, such as Shows (domain 13), naturally have lower values. The two graphs show similar patterns in many domains, but differences are also apparent. In particular, the posterior analysis for several domains, such as Health (domain 8) and Video (domain 19), shows “significance” in their origin effects but “insignificance” in their destination effects (i.e., their 95% Bayesian credible intervals of the destination effects contain one, but this is not case for the origin effects). These distinctions between the two effects show the roles of

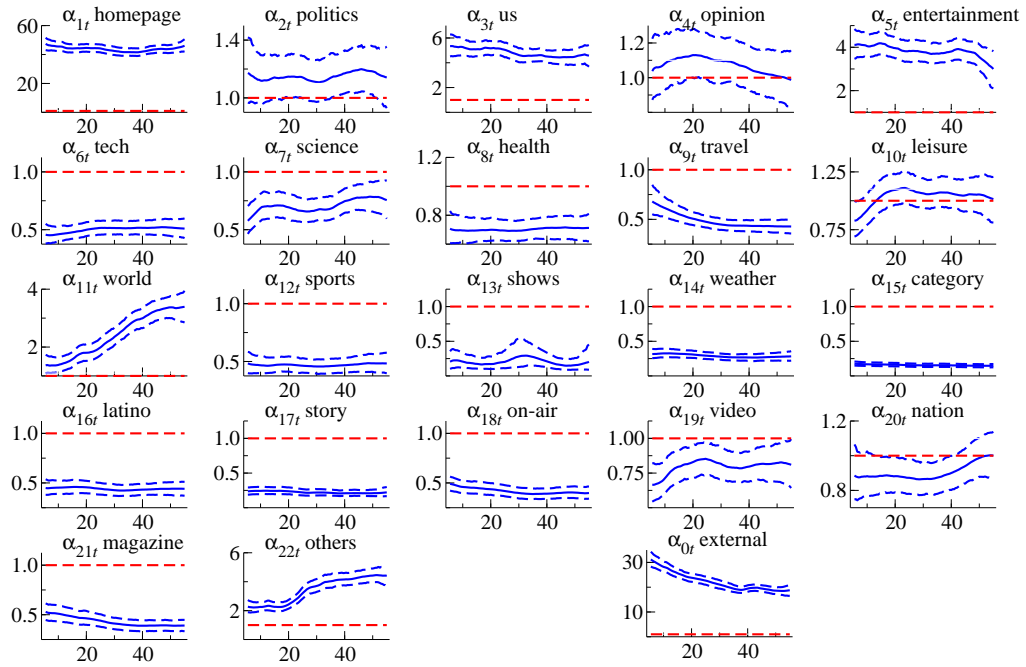


FIGURE 4.9: DGM-based smoothed trajectories of node-specific outflows $\alpha_{i,1:T}$.

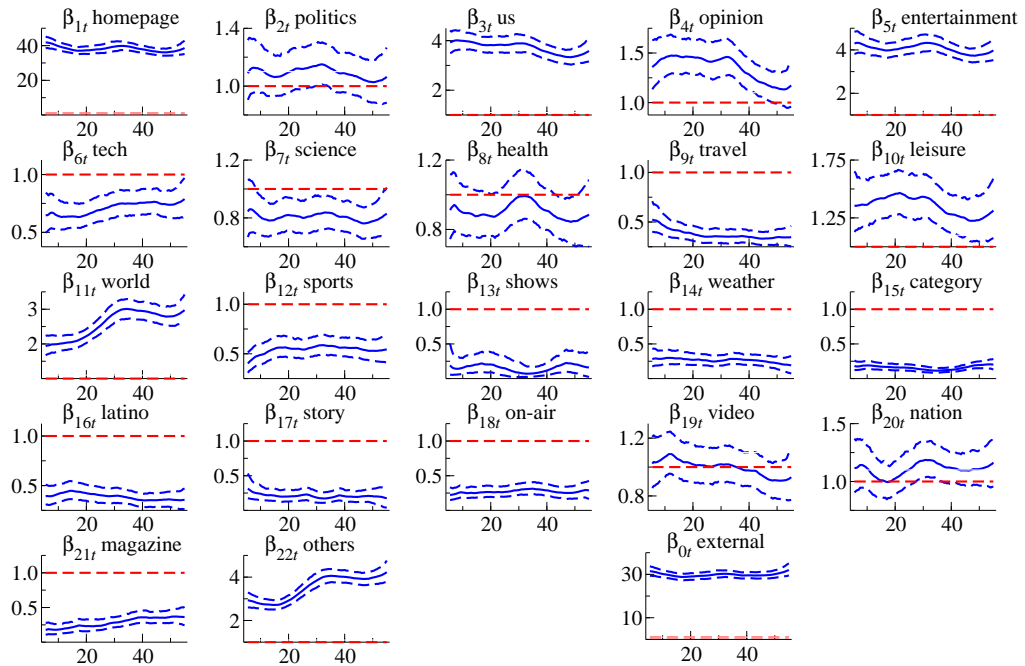


FIGURE 4.10: DGM-based smoothed trajectories of node-specific inflows $\beta_{j,1:T}$.

α_{it} and β_{jt} ; they represent the common factors across the origin and destination of the flows, which is confirmed by the differences of the two results. These effects are also essential in order to capture the scale of the domains, by having similar patterns for α_{it} and β_{jt} when $i = j$.

For the affinity effects γ_{ijt} , we have $(I + 1)^2 - 1$ parameters (one for each pair of nodes except the unobserved External \rightarrow External flow) at each time t . The number of effects becomes massive for large I . Even in this example for illustration, $I = 22$, the number of γ_{ijt} for fixed t is 528, so it is impossible to examine all the results here. For this reason, we pick up a few affinity effects that may interest readers in terms of interpretation. For affinity γ_{ijt} with retrospective posterior c.d.f $\Phi_{ijt}(\gamma)$, we use the Bayesian credible value $p_{ijt} = \min\{\Phi_{ijt}(1), 1 - \Phi_{ijt}(1)\}$ as a simple numerical measure of deviation from the “neutral” value of 1. This highlights the practical relevance of the affinity effect and its changes over time.

First, we focus on the flows from Homepage (domain 1). Those flows are crucial in understanding the user’s preference from the aggregated data since Homepage is usually the landing page for visitors. Flows from it must have information on which domain the user wants to access first, which would be an important finding in advertisement and marketing. Figure 4.11 shows that the affinity trajectory and credible values for flows from Homepage (domain 1) to Opinion (domain 4) are entirely greater than 1. This implies that visitors to the FoxNews landing page tend to access articles in the Opinion domain (during this time period). In contrast, Figure 4.12 displays the affinity trajectory and credible values for the flow from Homepage to Science (domain 7). The affinity effect is significantly negative for much of the time interval, implying that at the beginning of the hour, people on the FoxNews landing page tend not to check the websites within the Science domain. But, as time passes, the trajectory gradually increases to 1 and becomes insignificant. This change of significance is clear in the Bayesian credible values, which are almost

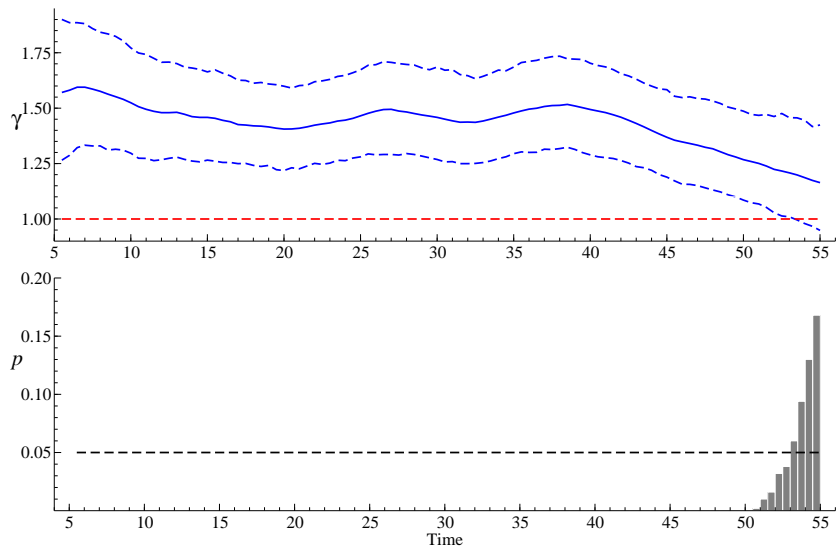


FIGURE 4.11: *Upper:* DGM-based smoothed trajectories of transition affinities, Home-page \rightarrow Opinion. *Lower:* Bayesian credible values corresponding to the affinity trajectories.

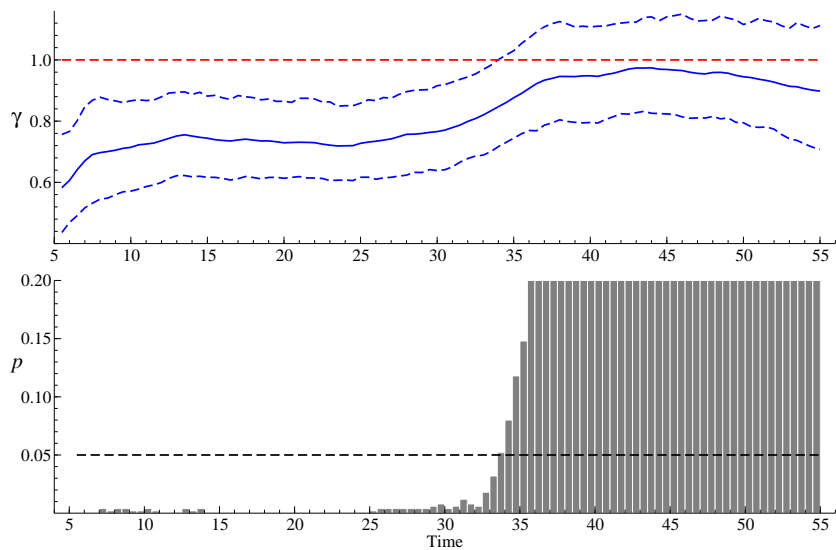


FIGURE 4.12: *Upper:* DGM-based smoothed trajectories of transition affinities, Home-page \rightarrow Science. *Lower:* Corresponding Bayesian credible values.

zero at first but suddenly begin to increase at around 9:40-9:45am and quickly exceed the reference line 0.05.

The change in significance of the Homepage \rightarrow Opinion affinities shows that the DGM has the flexibility to detect time-varying sparsity in parameters. This phenomenon is peculiar to time series analysis and modeling, and the example demonstrates that such sparsity actually exists in the dataset. It also has important implications for on-line advertising, as it shows that users have different interests at different times of the day.

4.5.5 Comparison Across Days

The FoxNews dataset covers both the morning (09:00-10:00am) and the afternoon (01:00-02:00pm) period on each of the 6 days. We have already discussed a range of comparisons, and differences, across days for the morning periods in Section 4.5.3. Moving to the DGM, we now explore additional features concerning time-of-day effects as well as day-to-day variation. This is based on running the coupled BDFM-DGM analysis separately on each time period/day.

Figure 4.13 shows the DGM trajectories for the retrospective baseline parameter process $\mu_{1:T}$ for each of the 12 fifty-minute intervals. Trajectories are similar across days but for notable differences between February 24 and March 9. On February 24, the afternoon flow is significantly lower than the morning flow, while the morning flow that day is much larger than across other days. One plausible reason is increased morning traffic in response to discussions following the Academy Awards ceremony, with a resulting lull in the afternoon traffic. The reverse happens on March 3, 4 and 9 where, although the morning traffic seems typical, the afternoon traffic is unusually high. March 3 was the day on which FoxNews posted an article concerning Hillary Clinton's use of her personal email account for all correspondence during her tenure as Secretary of State. It is plausible that this led to larger than usual afternoon

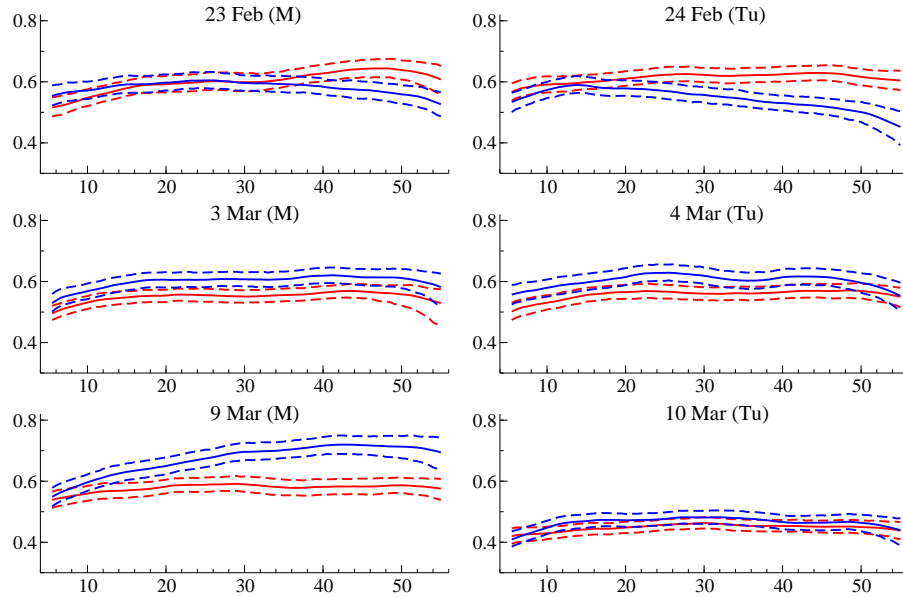


FIGURE 4.13: DGM-based inference on baseline flow level trajectories for all six days, with 95% credible intervals. The red trajectories correspond to the 09:05-09:55am time window, and the blue trajectories correspond to the 01:05-01:55 p.m. time window.

traffic flows as the controversy unfolded.

A “fair” comparison of the different datasets is difficult since the DGM can provide different results for the same dataset if the identification strategy is changed. The comparison above is based on the identification with constraints from Section 4.3.5 that are applied to each dataset independently. Since restriction indicators s_{ij} are defined by observations $y_{ij,1:T}$, the different datasets may lead to different restrictions even under the same identification rule. This problem motivates another possible approach to comparison that uses a common restriction: defining s_{ij} based on the first dataset and using the same indicators for the others. Figure 4.14 shows the posteriors of $\mu_{1:T}$ obtained with the common restriction (defined by the morning data on 23rd February, so the result of this time/date is the reference line for the other datasets). Though they are different from the posteriors in Figure 4.13 in terms of their scales, the basic relation between morning and afternoon within a

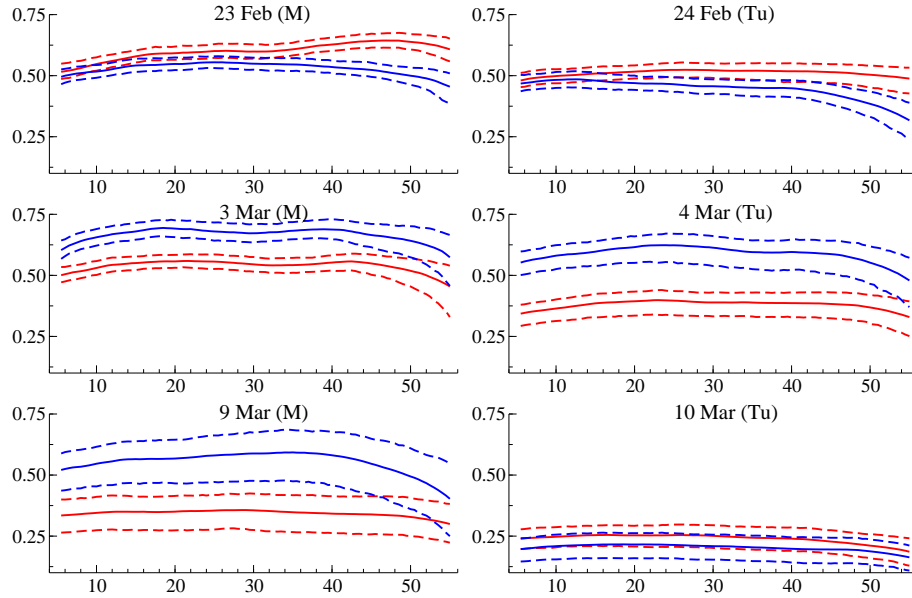


FIGURE 4.14: Posterior trajectories of $\mu_{1:T}$ with the common restriction.

day is preserved. This representation emphasizes the increase of the population of visitors in the afternoon of 3rd March, that matches the effect of the political article posted on that date.

One of the advantages of the DGM representation is that it allows an easy path to check these speculative explanations. For example, as seen in Figure 4.15, the examination of the affinity effect on the incoming flow into the Entertainment domain, $\gamma_{0,5,t}$, shows the unusual popularity of this domain that continues until the morning of February 24. The observed affinity effects in the end of February that are much larger than those in early March supports our reasoning that ties the increase of traffic to the Academy Awards ceremony.

4.6 Summary Comments

The BDFM framework is adaptive to time-varying rates of flows within dynamic networks and able to coherently quantify non-stationary changes in within- and into-/out of- network flow rate processes. The sequential analysis of this Bayesian

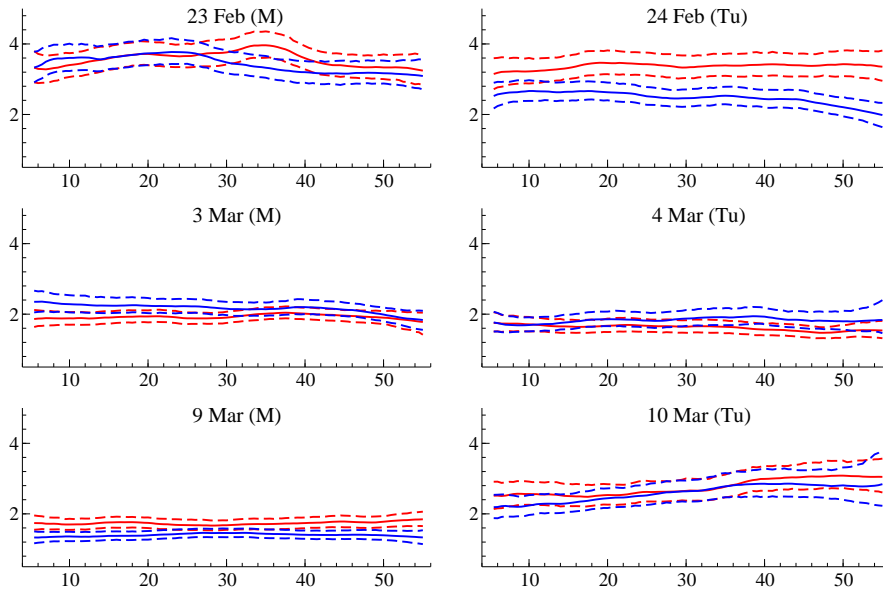


FIGURE 4.15: DGM-based inference on $\gamma_{0,5,1:T}$ for all six days. The 12 retrospective posterior trajectories of the affinity effect, $\gamma_{0,5,t}$, which corresponds to the flow from External to Entertainment.

dynamic flow model is fast and efficient; computational demands scale linearly in time and quadratically in node number. Importantly, this almost semi-parametric approach generates a parallelizable analysis yielding full posterior distributions for underlying rate parameter process parameters across nodes and pairs of nodes in a scalable manner. While the model inherently reflects the complexity of interactions among the full set of nodes, and their changes over time, the BDFM approach “decouples” analysis to individual nodes and pairs of nodes, and “recouples” them (via the map from decoupled gamma to recoupled Dirichlet posteriors over time) for formal inferences. Our analysis of the FoxNews network time series datasets shows the utility of the BDFM in generating initial inferences on flow rate processes, in highlighting differences across days and in generating potential practical “leads”. On the latter, for example, it is immediately clear from the BDFM results that most visitors go to just one domain, rather than traversing to multiple domains. This has poten-

tial decision implications for computational advertising, and also likely highlights a difference between on-line news consumers and traditional newspaper readers.

The Bayesian emulation “map” from the BDFM to the dynamic gravity model represents a modelling/computational strategy of increasing interest in many areas, and especially in emerging Big Data applications. That is, we fit a flexible, adaptive model in a set of (conditionally) decoupled analyses, and then directly map posterior samples to the more substantively interesting and interpretable parameter processes in a model, the DGM, that is otherwise challenging to fit. Applied to the FoxNews flow data, we see that this indicates “time-varying sparsity” in node-node interaction effects over time, nicely captured by the DGM. Our use of Bayesian credible values over time is one nice way of focusing attention on this, allowing us to highlight the “significance” of inferred DGM interactions across time. Interestingly, many of the interaction effects (affinities) appear significant at some points in time but not in others. A number of the specific node-node inferences mentioned in the application section highlight additional results of substantive interest, some of which are initially unexpected. These include, for example, the sustained positive affinity of Opinion for Homepage, but a similarly sustained but negative affinity of Science for Homepage. Additionally, comparisons across different times of the day identified and quantified patterns related to anomalous flows corresponding to identifiable news events that appear to have driven traffic to specific nodes on the FoxNews site.

From this point, we see opportunities to now develop these models as a basis to characterize the stochastic dynamics of website flows, and hence feed into modeling and decision analysis that addresses the needs to respond to changing patterns in computational advertising. An ability to rapidly signal potential anomalies in a small subset of domains in real-time will be of huge interest in this field. We also note, as remarked in the introduction to the chapter, the potential connections with problems of physical traffic flows, origin-destination problems, and other kinds of

dynamic network studies including social networks, capital flows between financial institutions, electrical power grids, and others.

More immediately, some of the evident questions arising from the current study concern the overlay of the “unbiased” inferences about changes and structure in network flows with substantive covariate information. In many applications, including computational advertising but also capital and transportation flows, there are useful covariates that could inform the analysis. Our perspective here has been more exploratory, aiming to define a formal basis for effectively characterizing non-stationary stochastic dynamics in flow data. A next step is to overlay any particular application with covariate information as descriptive/explanatory as we exemplified with some vignettes from the FoxNews study. At a more predictive level, the DGM is naturally extensible to incorporate covariates— in main effects and/or interaction terms— so that some consideration of how to extend the flexible, computational efficient and scalable BDFM-DGM in that direction is warranted.

4.7 Appendix: FFBS for Poisson-Gamma Models

4.7.1 Forward Filtering

The derivation of the prior, forecast and on-line posterior distributions are explained.

To compute the prior $p(\phi_t|\mathcal{D}_{t-1})$, use the method of change of variables for ϕ_t . First, note that $\eta_t = \delta\phi_t/\phi_{t-1}$, so the Jacobian is $d\eta_t/d\phi_t = \delta/\phi_{t-1}$. Also, since $0 < \eta_t < 1$, the support of the distribution of ϕ_t is $\phi_{t-1} > \delta\phi_t$. Thus, the conditional distribution of ϕ_t is

$$\begin{aligned} p(\phi_t|\phi_{t-1}, \mathcal{D}_{t-1}) &= p(\eta_t|\mathcal{D}_{t-1}) \left| \frac{d\eta_t}{d\phi_t} \right| \\ &= \frac{1}{B(\delta r_{t-1}, (1-\delta)r_{t-1})} \left(\frac{\delta\phi_t}{\phi_{t-1}} \right)^{\delta r_{t-1}-1} \left(1 - \frac{\delta\phi_t}{\phi_{t-1}} \right)^{(1-\delta)r_{t-1}-1} \frac{\delta}{\phi_{t-1}} \\ &\propto \phi_{t-1}^{-r_{t-1}+1} \phi_t^{\delta r_{t-1}-1} (\phi_{t-1} - \delta\phi_t)^{(1-\delta)r_{t-1}-1}, \end{aligned}$$

where the last expression above is obtained by ignoring the constant terms which contain neither ϕ_t nor ϕ_{t-1} . With the prior at $t-1$, the prior at t is calculated by

$$\begin{aligned} p(\phi_t|\mathcal{D}_{t-1}) &= \int_{\{0 < \phi_t < \delta\phi_{t-1}\}} p(\phi_t|\phi_{t-1}, \mathcal{D}_{t-1}) p(\phi_{t-1}|\mathcal{D}_{t-1}) d\phi_{t-1} \\ &\propto \int_{\{0 < \phi_t < \delta\phi_{t-1}\}} \left(\phi_{t-1}^{-r_{t-1}+1} \phi_t^{\delta r_{t-1}-1} (\phi_{t-1} - \delta\phi_t)^{(1-\delta)r_{t-1}-1} \right) \\ &\quad \times \left(\phi_{t-1}^{r_{t-1}-1} e^{-c_{t-1}\phi_{t-1}} \right) d\phi_{t-1} \\ &\propto \phi_t^{\delta r_{t-1}-1} \int_{\{0 < \phi_t < \delta\phi_{t-1}\}} (\phi_{t-1} - \delta\phi_t)^{(1-\delta)r_{t-1}-1} e^{-c_{t-1}\phi_{t-1}} d\phi_{t-1}, \end{aligned}$$

and, with another change of variable in that $\theta = \phi_{t-1} - \delta\phi_t$, $d\theta/d\phi_{t-1} = 1$ and

$\theta \in (0, \infty)$ (since, as mentioned above, $\eta_t < 1$ implies $\phi_{t-1} > \delta\phi_t$), we continue as

$$\begin{aligned} p(\phi_t|\mathcal{D}_{t-1}) &\propto \phi_t^{\delta r_{t-1}-1} \int_0^\infty \theta^{(1-\delta)r_{t-1}-1} e^{-c_{t-1}\theta} e^{-\delta c_{t-1}\phi_t} d\theta \\ &= \phi_t^{\delta r_{t-1}-1} e^{-\delta c_{t-1}\phi_t} \frac{\Gamma((1-\delta)r_{t-1})}{(\delta c_{t-1}^{1-\delta r_{t-1}})} \\ &\propto \phi_t^{\delta r_{t-1}-1} e^{-\delta c_{t-1}\phi_t}, \end{aligned}$$

which is exactly the kernel of gamma density with shape δr_{t-1} and rate δc_{t-1} .

The forecast distribution is obtained in a similar way by integrating the joint distribution of (y_t, ϕ_t) as

$$\begin{aligned} p(y_t|\mathcal{D}_{t-1}) &= \int p(y_t|\phi_t)p(\phi_t|\mathcal{D}_{t-1})d\phi_t \\ &= \int \left(\frac{\phi_t^{y_t}}{y_t!} e^{-\phi_t} \right) \left(\frac{(\delta c_{t-1})^{\delta r_{t-1}}}{\Gamma(\delta r_{t-1})} \phi_t^{\delta r_{t-1}-1} e^{-\delta c_{t-1}\phi_t} \right) d\phi_t \\ &\propto \frac{1}{y_t!} \int \phi_t^{y_t+\delta r_{t-1}-1} e^{-(\delta c_{t-1}+1)\phi_t} d\phi_t \\ &\propto \frac{1}{y_t!} \frac{\Gamma(y_t + \delta r_{t-1})}{(\delta c_{t-1} + 1)^{y_t+\delta r_{t-1}}} \\ &\propto \frac{\Gamma(y_t + \delta r_{t-1})}{\Gamma(y_t + 1)} \left(\frac{1}{\delta c_{t-1} + 1} \right)^{y_t}, \end{aligned}$$

which is the negative binomial distribution as desired.

The on-line posterior can be obtained from the joint distribution as

$$\begin{aligned} p(\phi_t|\mathcal{D}_t) &\propto p(y_t, \phi_t|\mathcal{D}_{t-1}) \\ &= p(y_t|\phi_t)p(\phi_t|\mathcal{D}_{t-1}) \\ &= \left(\frac{\phi_t^{y_t}}{y_t!} e^{-\phi_t} \right) \left(\frac{(\delta c_{t-1})^{\delta r_{t-1}}}{\Gamma(\delta r_{t-1})} \phi_t^{\delta r_{t-1}-1} e^{-\delta c_{t-1}\phi_t} \right) \\ &\propto \phi_t^{\delta r_{t-1}+y_t-1} e^{-(\delta c_{t-1}+1)\phi_t}, \end{aligned}$$

which is the kernel of gamma distribution with shape $r_t = \delta r_{t-1} + y_t$ and rate $c_t = \delta c_{t-1} + 1$.

4.7.2 Backward Sampling

Based on the result of filtering, the details of backward sampling are explained.

First, note that the conditional independence structure in this state-space model simplifies the target distribution as $p(\phi_t|\phi_{t+1}, \mathcal{D}_T) = p(\phi_t|\phi_{t+1}, \mathcal{D}_t)$, i.e., the observations after t are independent of ϕ_t conditional on ϕ_{t+1} . Then,

$$\begin{aligned}
 p(\phi_t|\phi_{t+1}, \mathcal{D}_T) &\propto p(\phi_{t+1}|\phi_t, \mathcal{D}_t)p(\phi_t|\mathcal{D}_t) \\
 &= \frac{1}{B(\delta r_t, (1-\delta)r_t)} \left(\frac{\delta\phi_{t+1}}{\phi_t}\right)^{\delta r_t - 1} \left(1 - \frac{\delta\phi_{t+1}}{\phi_t}\right)^{(1-\delta)r_t - 1} \frac{\delta}{\phi_t} \\
 &\quad \times \frac{c_t^{r_t}}{\Gamma(r_t)} \phi_t^{r_t - 1} e^{-c_t\phi_t} \\
 &\propto (\phi_t - \delta\phi_{t+1})^{(1-\delta)r_t - 1} e^{-c_t\phi_t} \\
 &\propto (\phi_t - \delta\phi_{t+1})^{(1-\delta)r_t - 1} e^{-c_t(\phi_t - \delta\phi_{t+1})},
 \end{aligned}$$

and, by the change of variables with $\epsilon_t = \phi_t - \delta\phi_{t+1}$, it follows that

$$\epsilon_t \sim G((1-\delta)r_t, c_t),$$

which proves the statement in Section 4.2.1.

Concluding Remarks

This dissertation has discussed research advances in three areas of Bayesian time series analysis: sequential modeling, inference and decision making. In each area, we have considered modern and challenging statistical problems of modeling/analyzing multivariate time series in the presence of high-dimensionality and sparsity. Building on the cross-cutting concept of Bayesian model emulation, the thesis demonstrates substantial advances in Bayesian methodology to overcome challenging problems raised in addressing practical use in both academic and industrial applications.

Each research area has multiple themes and directions that are open to future research.

Chapter 2: Inference

- Though the two recommended LTM emulators perform well in predictive accuracy, the central problem of particle degeneracy– the main issue facing all SMC approaches– leaves room for improvements. The combination of SMC with MCMC, mentioned in Chapter 2, can help to resolve this, by allowing for periodic “refreshing” of particle samples using off-line MCMC methods, that

then seed a new period of SMC analysis.

- There is no unique definition of the LTM emulator. This research developed and explored several, and left open questions of “optimal” emulators. For example, one idea would be to develop some kind of model discrepancy measure, such as based on Kullback-Leibler divergences between target and emulating posteriors (at some time point or points), aiming to define “optimal” choices. This idea could be applied, instead, to emulation of sampling distributions at some time points, and/or one-step ahead predictive distributions. How to begin to develop this concept is an open question, but the ideas seem natural.
- The idea and methodology of Bayesian model emulation in time series is not restricted to LTMs, and could be explored and developed in other contexts of non-linear state-space models, such as Markov switching models (increasingly popular in macroeconomics), or others.

Chapter 3: Decisions

- The idea of emulation will yield many synthetic models in various kinds of decision problems. In our example, the asymmetric version of MSE that penalizes excess returns less (if at all!) than losses can be derived from the generalized hyperbolic skewed-t distribution in the synthetic model. This distribution, as well as the Laplace distribution, has the mixture form (e.g. Hu and Kercheval, 2008) that enables modeling of the target return as

$$m_t = f_t'w_t + N(\xi_t\rho_t, \xi_t), \quad \xi_t^{-1} \sim Ga(\nu_t/2, \nu_t/2),$$

where (ν_t, ρ_t) are the pre-specified degree of freedom and skewness parameters. Related EM methods allow for numerical search for posterior modes in the synthetic/emulating model.

Another example is to group similar assets and consider common shrinkage toward that group. Such a grouping/classification might be done in the inference step based on the emulating statistical model. For the s -th group of assets, defined by an index subset $\mathcal{S}_s \subset 1:k$, we could then make use of a common mixing parameter for the Laplace distribution, i.e.,

$$w_{s,t} = w_{s,t-1} + N(0, \tau_{s,t} W_{s,t}), \quad \tau_{s,t} \sim Ga(1, \lambda_{s,t}^{-2}/2),$$

where $w_{s,t} = \{w_{it}\}_{i \in \mathcal{S}_s}$, the sub-vector of w_t , and with $W_{s,t}$ as the pre-specified covariance matrix (typically $W_{s,t} = I$.) This distribution is known as the multivariate version of Laplace distribution (Eltoft et al., 2006). The EM method is available and open for development for solution in this grouped context, and compatible with the other mixing parameters.

Again, it should be emphasized that the core idea of mapping a hard, computational optimization problem to an emulating, purely synthetic statistical model exploration program, is very general strategy and likely of interest to explore in many future areas. Some opportunities may exist in exploring discussions of currently topical optimization problems in applied mathematics and engineering fields, for example, that may be open (technically) to approaches based on new Bayesian emulation ideas.

- It is of interest in both theory and practice to explore what classes of loss functions might in fact yield parallel synthetic models. Presumably, not all expected loss functions can be converted to probabilistic models. To see this, consider the new MSE loss function in a portfolio example that does not penalize excess returns at all. One example is the truncated quadratic loss function defined as

$$(m_t - f_t' w_t)^2 \mathbb{1}[f_t' w_t \leq m_t].$$

Any monotonic transformation of the loss function above results in the function of w_t that is constant on $\{ w_t \mid f'_t w_t > m_t \}$. This cannot yield a parallel, synthetic probability density because its integral is not finite. Instead, it might be viewed as an “improper likelihood,” that could yield a proper posterior distribution, assuming a proper (synthetic) prior. Though the philosophical question on the appropriateness of this type of analysis remains, it may be practically useful unless it loses the well-defined posterior and the unique posterior mode.

- The difference between the profiled and marginal approaches was examined empirically but not fully theoretically. In multi-step optimization— in portfolio analysis and other areas, such as dynamic control in engineering, or in policy decision making in macroeconomics— this issue is central and critical. The two approaches simply lead to different loss functions; they just represent different preferences. Yet, it is worth understanding that difference theoretically, since the theory on the difference between those loss functions can help the decision maker to choose one of them that really matches his or her personal preference.

Theoretical studies of relationships between joint and marginal modes are of interest here. In simple lasso regression,

$$y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I_n)$$

with known σ^2 , we can find the marginal density, $p(\beta_1|y)$, analytically with β partitioned as β_1 and β_2 . The extension of this type of analysis to more elaborate problems may reveal insights based on differences in results, including the relevance to resulting differences in portfolios based on joint versus marginal optimization, as in Section 3.5 and 3.6.

In more general contexts, theoretical considerations of the marginal loss function might be addressed in terms of nuisance parameters for inference, e.g.

using ideas from semi-parametric estimation.

- More extensive application to higher-dimensional portfolio investment problems is a key interest— from applied perspectives in financial analysis, and also from research perspectives in terms of computational feasibility. With advances in research in dynamic models and forecasting geared at scaling forecasting models to many more series— such as in Gruber and West (2016) and our work with LTMs— the ability to generate sensitive short-term predictive models is advancing, so promoting a need for more active research in scaling the decision analysis to enable banks of portfolios to be run in real-time against multiple utility functions. Significant advances in the abilities to (i) customize multi-step utility functions to context and increasing numbers of assets, potentially with structure to group assets within classes, and (ii) implement fast and reliable portfolio optimization to help advise allocation decisions, promotes and highlights the core ideas underlying our Bayesian emulation research presented here, and the need for active development to scale and implement the idea more broadly.

Chapter 4: Modeling

- The beautiful mathematical property of the mapping from BDFMs to DGM is largely dependent on the relatively simple parametrization of DGM. This is no longer available if one introduces covariates and regression forms in one of the DGM components, e.g., $\log \gamma_{ijt} = x'_{ijt} \lambda_{ijt}$. It is of great practical interest to have covariates associated with each node or edge/flow in the model in many areas.

Naturally, the parallelizable simulation-based methods are needed in sampling from the retrospective posteriors of this type of model. There exists some re-

search on sequential analysis of the Poisson state-space models (Aktekin et al., 2016), but the use of covariates in such models is very limited— at least in terms of easily accessible, published research. The idea of Bayesian model emulation might help the development of new sequential analysis methods for extended DGM and broader classes of models, including specifically classes of dynamic generalized linear models.

- In practice, covariates are often categorical or can be reasonably discretized. The proposed emulation by BDFMs is valid for this case by dividing the nodes into several sub-groups based on the covariates. For example, the FoxNews dataset has information on the time zone associated with each visit. For time zone s , we might define a new node (i, s) and flow y_{ijst} to reflect this. For the Poisson rate ϕ_{ijst} , the DGM would then correspond to a 3-way ANOVA, adding the time-zone specific effect and the interaction to the origin/destination.

Each of the three main areas of research represented in this thesis has its own origin, structure and problems, but they are tied together through the core idea of defining analysis solutions via emulation: emulating a target model, or a target decision function, depending on context. While it is easier to define the model that directly addresses the problem/data/preference of interest, such an approach is highly likely to result in a non-linear, non-Gaussian state-space model that is mathematically complex and overly parametrized, e.g. LTMs in Chapter 2, Laplace models in Chapter 3, and DGMs in Chapter 4. The emulator(s) should be “simple” statistical models associated with the well-known computational methodologies, that are thoughtfully developed so as to address the core goals of the target context remaining amenable to efficient solution.

The emulators used in this dissertation are all DLMS—Gaussian linear state-space

models—and their variants, which are theoretically complete and well-understood (e.g. West and Harrison, 1997). It is worth reiterating that posterior analysis using simulation via FFBS is central to computation in many situations, even though the models are no longer Gaussian and linear. Despite computation advances that enable use of increasingly large-scale and complex MCMC methods, the fact that many statistical challenges are still emerging and unsolved implies the importance of revisiting simpler but analytically tractable models and potential tools, as exemplified by this core role of DLM/FFBS analysis in all three areas of this thesis. As a result of this experience, and defining a position for future research and application, I believe that the spirit of Bayesian emulation will be seen in the future in various areas of statistical research and application.

Bibliography

- Agarwal, D., Agrawal, R., Khanna, R., and Kota, N. (2010), “Estimating rates of rare events with multiple hierarchies through scalable long-linear models,” *KDD’10 Proceedings of the 16th ACM SIGKDD*, pp. 213–222.
- Aktekin, T., Polson, N. G., and Soyer, R. (2016), “Sequential Bayesian analysis of multivariate Poisson count data,” *arXiv preprint arXiv:1602.01445*.
- Amzal, B., Bois, F. Y., Parent, E., and Robert, C. P. (2006), “Bayesian-optimal design via interacting particle systems,” *Journal of the American Statistical Association*, 101, 773–785.
- Anacleto, O., Queen, C., and Albers, C. J. (2013a), “Forecasting multivariate road traffic flows using Bayesian dynamic graphical models, splines and other traffic variables,” *Australian & New Zealand Journal of Statistics*, 55, 69–86.
- Anacleto, O., Queen, C., and Albers, C. J. (2013b), “Multivariate forecasting of road traffic flows in the presence of heteroscedasticity and measurement errors,” *Journal of the Royal Statistical Society (Series C, Applied Statistics)*, 62, 251–270.
- Andrews, D. F. and Mallows, C. L. (1974), “Scale mixtures of normal distributions,” *Journal of the Royal Statistical Society (Series B: Methodological)*, 36, 99–102.
- Asif, A. and Moura, J. M. (2005), “Block matrices with L-block-banded inverse: Inversion algorithms,” *IEEE Transactions on Signal Processing*, 53, 630–642.
- Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis*, Springer.
- Bishop, Y., Fienberg, S. E., and Holland, P. (1975), *Discrete Multivariate Analysis: Theory and Practice*, The MIT Press.
- Brandt, P. T. and Williams, J. T. (2001), “A linear Poisson autoregressive model: The Poisson AR (p) model,” *Political Analysis*, 9, 164–184.
- Brandt, P. T., Williams, J. T., Fordham, B. O., and Pollins, B. (2000), “Dynamic modeling for persistent event-count time series,” *American Journal of Political Science*, pp. 823–843.

- Carvalho, C. M. and Lopes, H. F. (2007), “Simulation-based sequential analysis of Markov switching stochastic volatility models,” *Computational Statistics & Data Analysis*, 51, 4526–4542.
- Carvalho, C. M., Johannes, M. S., Lopes, H. F., and Polson, N. G. (2010), “Particle learning and smoothing,” *Statistical Science*, 25, 88–106.
- Chen, X., Irie, K., Banks, D., Haslinger, R., Thomas, J., and West, M. (2015), “Bayesian dynamic modeling and analysis of streaming network data,” *Technical Report, Duke University*.
- Cogley, T. and Sargent, T. J. (2005), “Drifts and volatilities: Monetary policies and outcomes in the post WWII U.S.” *Review of Economic Dynamics*, 8, 262–302.
- Congdon, P. (2000), “A Bayesian approach to prediction using the gravity model, with an application to patient flow modeling,” *Geographical Analysis*, 32, 205–224.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38.
- Doornik, J. A. (2007), *Object-Oriented Matrix Programming Using Ox, 3rd ed.*, London: Timberlake Consultants Press and Oxford, 3rd edn.
- Eltoft, T., Kim, T., and Lee, T.-W. (2006), “On the multivariate Laplace distribution,” *IEEE Signal Processing Letters*, 13, 300–303.
- Figueiredo, M. A. T. (2003), “Adaptive sparseness for supervised learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25, 1150–1159.
- Ghahramani, Z. (1995), “Factorial learning and the EM algorithm,” in *Advances in neural information processing systems*, pp. 617–624, Morgan Kaufman.
- Gruber, L. F. and West, M. (2016), “GPU-accelerated Bayesian learning in simultaneous graphical dynamic linear models,” *Bayesian Analysis*, 11, 125–149, Advance Publication, 2 March 2015.
- Harvey, A. and Fernandes, C. (1989a), “Time series models for insurance claims,” *Journal of the Institute of Actuaries*, 116, 513–528.
- Harvey, A. C. and Fernandes, C. (1989b), “Time series models for count or qualitative observations,” *Journal of Business & Economic Statistics*, 7, 407–417.
- Hu, W. and Kercheval, A. (2008), “The skewed t distribution for portfolio credit risk,” *Advances in econometrics*, 22, 55–83.
- Irie, K. and West, M. (2016a), “Bayesian emulation for multi-step portfolio decisions,” *Technical Report, Duke University*.

- Irie, K. and West, M. (2016b), “Bayesian emulation for sequential analysis of dynamic latent threshold models,” *Technical Report, Duke University*.
- Jandarov, R., Haran, M., Bjornstad, O. N., and Grenfell, B. T. (2014), “Emulating a gravity model to infer the spatiotemporal dynamics of an infectious disease,” *Journal of the Royal Statistical Society (Series C, Applied Statistics)*, 63, 423–444.
- Johannes, M., Korteweg, A., and Polson, N. (2014), “Sequential learning, predictability, and optimal portfolio returns,” *Journal of Finance*, 69, 611–644.
- Koop, G. and Korobilis, D. (2010), “Bayesian multivariate time series methods for empirical macroeconomics,” *Foundations and Trends in Econometrics*, 3, 267–358.
- Koop, G. and Korobilis, D. (2013), “Large time-varying parameter VARs,” *Journal of Econometrics*, 177, 185–198.
- Koop, G., Leon-Gonzalez, R., and Strachan, R. W. (2009), “On the evolution of the monetary policy transmission mechanism,” *Journal of Economic Dynamics and Control*, 33, 997–1017.
- Koren, R., Bell, R., and Volinsky, C. (2009), “Matrix factorization techniques for recommender systems,” *Computer*, 8, 30–37.
- Korobilis, D. (2011), “VAR forecasting using Bayesian variable selection,” *Journal of Applied Econometrics*.
- Liu, C., Martin, R., and Syring, N. (2013), “Simulating from a gamma distribution with small shape parameter,” *arXiv preprint arXiv:1302.1884*.
- Liu, F., Chakraborty, S., Li, F., Liu, Y., Lozano, A. C., et al. (2014), “Bayesian regularization via graph Laplacian,” *Bayesian Analysis*, 9, 449–474.
- Liu, J. and West, M. (2001), “Combined parameter and state estimation in simulation-based filtering,” in *Sequential Monte Carlo Methods in Practice*, eds. A. Doucet, J. F. G. D. Freitas, and N. J. Gordon, pp. 197–217, Springer.
- Liu, J. S. (1996), “Metropolized independent sampling with comparisons to rejection sampling and importance sampling,” *Statistics and Computing*, 6, 113–119.
- Liu, J. S. and Chen, R. (1998), “Sequential Monte Carlo methods for dynamic systems,” *Journal of the American Statistical Association*, 93, 1032–1044.
- Lopes, H. F. and Tsay, R. S. (2011), “Particle filters and Bayesian inference in financial econometrics,” *Journal of Forecasting*, 30, 168–209.

- Lopes, H. F., Carvalho, C. M., Johannes, M., and Polson, N. G. (2010), “Particle learning for sequential Bayesian computation,” in *Bayesian Statistics 9*, eds. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, pp. 175–96, Oxford University Press.
- Markowitz, H. (1952), “Portfolio Selection,” *The Journal of Finance*, 7, 77–91.
- Markowitz, H. M. (1968), *Portfolio Selection: Efficient Diversification of Investments*, Yale University Press.
- Müller, P. (1999), “Simulation based optimal design,” in *Bayesian Statistics 6*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, pp. 459–474, Oxford University Press.
- Müller, P., Sansó, B., and De Iorio, M. (2004), “Optimal Bayesian design by inhomogeneous Markov chain simulation,” *Journal of the American Statistical Association*, 99, 788–798.
- Nakajima, J. (2011), “Time-varying parameter VAR model with stochastic volatility: An overview of methodology and empirical applications,” Tech. rep., Institute for Monetary and Economic Studies, Bank of Japan.
- Nakajima, J. and West, M. (2013a), “Bayesian analysis of latent threshold dynamic models,” *Journal of Business and Economic Statistics*, 31, 151–164.
- Nakajima, J. and West, M. (2013b), “Bayesian dynamic factor models: Latent threshold approach,” *Journal of Financial Econometrics*, 11, 116–153.
- Nakajima, J. and West, M. (2015), “Dynamic network signal processing using latent threshold models,” *Digital Signal Processing*, 47, 5–16.
- Pang, B. and Lee, L. (2008), “Opinion mining and sentiment analysis,” *Foundations and Trends in Information Retrieval*, 2, 1–135.
- Park, T. and Casella, G. (2008), “The Bayesian lasso,” *Journal of the American Statistical Association*, 103, 681–686.
- Pitt, M. K. and Shephard, N. (1999), “Filtering via simulation: Auxiliary variable particle filter,” *Journal of the American Statistical Association*, 94, 590–599.
- Polson, N. G. and Scott, J. G. (2015), “Mixtures, envelopes and hierarchical duality,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.
- Prado, R. and Lopes, H. F. (2013), “Sequential parameter learning and filtering in structured autoregressive state-space models,” *Statistics and Computing*, 23, 43–57.

- Prado, R. and West, M. (2010), *Time Series: Modeling, Computation and Inference*, Chapman and Hall/CRC Press.
- Primiceri, G. E. (2005), “Time varying structural vector autoregressions and monetary policy,” *The Review of Economic Studies*, 72, 821–852.
- Queen, C. M. and Albers, C. J. (2009), “Intervention and causality: Forecasting traffic flows using a dynamic Bayesian network,” *Journal of the American Statistical Association*, 104, 669–681.
- Rue, H., Martino, S., and Chopin, N. (2009), “Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations,” *Journal of the Royal Statistical Society (Series B: Methodological)*, 71, 319–392.
- Sen, A. and Smith, T. (1995), *Gravity Models of Spatial Interaction Behavior*, Springer.
- Shephard, N. (1994), “Local scale models: State space alternative to integrated GARCH processes,” *Journal of Econometrics*, 60, 181–202.
- Smith, J. Q. (1979), “A generalization of the Bayesian steady forecasting model,” *Journal of the Royal Statistical Society, Series B (Methodological)*, 375–387.
- Soriano, J., Au, T., and Banks, D. (2013), “Text mining in computational advertising,” *Statistical Analysis and Data Mining*, 6, 273–285.
- Taddy, M. (2013), “Multinomial inverse regression for text analysis,” *Journal of the American Statistical Association*, 108, 755–770.
- Tebaldi, C. and West, M. (1998), “Bayesian inference on network traffic using link count data (with discussion),” *Journal of the American Statistical Association*, 93, 557–576.
- Tebaldi, C., West, M., and Karr, A. F. (2002), “Statistical analyses of freeway traffic flows,” *Journal of Forecasting*, 21, 39–68.
- Uhlig, H. (1994), “On singular Wishart and singular multivariate beta distributions,” *The Annals of Statistics*, pp. 395–405.
- Uhlig, H. (1997), “Bayesian vector autoregressions with stochastic volatility,” *Econometrica: Journal of the Econometric Society*, pp. 59–73.
- West, M. (1987), “On scale mixtures of normal distributions,” *Biometrika*, 74, 646–648.
- West, M. (1993a), “Approximating posterior distributions by mixtures,” *Journal of the Royal Statistical Society (Ser. B)*, 54, 553–568.

- West, M. (1993b), “Mixture models, Monte Carlo, Bayesian updating and dynamic models,” *Computing Science and Statistics*, 24, 325–333.
- West, M. (1994), “Statistical inference for gravity models in transportation flow forecasting,” Discussion Paper 94-20, Institute of Statistics & Decision Sciences, Duke University (June 1994). Also available as NISS Technical Report #60, US National Institute of Statistical Sciences.
- West, M. and Harrison, P. J. (1997), *Bayesian Forecasting and Dynamic Models*, Springer Verlag, 2nd edn.
- Zhao, Z. Y., Xie, M., and West, M. (2016), “Dynamic dependence networks: Financial time series forecasting and portfolio decisions (*with discussion*),” *Applied Stochastic Models in Business and Industry*, To appear.
- Zhou, X., Nakajima, J., and West, M. (2014), “Bayesian forecasting and portfolio decisions using dynamic dependent sparse factor sparse,” *International Journal of Forecasting*, 30, 963–980.

Biography

Kaoru Irie was born in Tochigi, Japan. He earned his B.A. and M.A. in Economics from University of Tokyo in 2010 and 2012, and his M.S. in Statistical Science from Duke University in 2014. He will earn his Ph.D. in Statistical Science from Duke University in 2016. In June of 2016, he will be joining Graduate School of Economics at University of Tokyo as an Assistant Professor.