Bayesian Multiregression Dynamic Models with Applications in Finance and Business

by

Zoey Yi Zhao

Department of Statistical Science
Duke University

Date: ____________________

Approved:

____
Mike West, Supervisor

____
Surya Tapas Tokdar

____
Li Ma

____
Jonathan R. M. Hosking

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
2015

Abstract

Bayesian Multiregression Dynamic Models with Applications in Finance and Business

by

Zoey Yi Zhao

Department of Statistical Science
Duke University

Date: ____________________

Approved:

____
Mike West, Supervisor

____
Surya Tapas Tokdar

____
Li Ma

____
Jonathan R. M. Hosking

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
2015

Abstract

This thesis discusses novel developments in Bayesian analytics for high-dimensional multivariate time series. The focus is on the class of multiregression dynamic models (MDMs), which can be decomposed into sets of univariate models processed in parallel yet coupled for forecasting and decision making. Parallel processing greatly speeds up the computations and vastly expands the range of time series to which the analysis can be applied.

I begin by defining a new sparse representation of the dependence between the components of a multivariate time series. Using this representation, innovations involve sparse dynamic dependence networks, idiosyncrasies in time-varying autoregressive lag structures, and flexibility of discounting methods for stochastic volatilities.

For exploration of the model space, I define a variant of the Shotgun Stochastic Search (SSS) algorithm. Under the parallelizable framework, this new SSS algorithm allows the stochastic search to move in each dimension simultaneously at each iteration, and thus it moves much faster to high probability regions of model space than does traditional SSS.

For the assessment of model uncertainty in MDMs, I propose an innovative method that converts model uncertainties from the multivariate context to the univariate context using Bayesian Model Averaging and power discounting techniques. I show that this approach can succeed in effectively capturing time-varying model

uncertainties on various model parameters, while also identifying practically superior predictive and lucrative models in financial studies.

Finally I introduce common state coupled DLMs/MDMs (CSCDLMs/CSCMDMs), a new class of models for multivariate time series. These models are related to the established class of dynamic linear models, but include both common and series-specific state vectors and incorporate multivariate stochastic volatility. Bayesian analytics are developed including sequential updating, using a novel forward-filtering-backward-sampling scheme. Online and analytic learning of observation variances is achieved by an approximation method using variance discounting. This method results in faster computation for sequential step-ahead forecasting than MCMC, satisfying the requirement of speed for real-world applications.

A motivating example is the problem of short-term prediction of electricity demand in a "Smart Grid" scenario. Previous models do not enable either time-varying, correlated structure or online learning of the covariance structure of the state and observational evolution noise vectors. I address these issues by using a CSCMDM and applying a variance discounting method for learning correlation structure. Experimental results on a real data set, including comparisons with previous models, validate the effectiveness of the new framework.

To my family

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Symbols

$\mathrm{E}(\cdot)$	Expected value of a random variable
$\mathrm{V}(\cdot)$	Variance (also referred to as covariance) matrix of a random vector
$\mathrm{C}(\cdot,\cdot)$	Covariance vector or matrix of two random vectors (variables)
$\mathrm{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$	Normal distribution with mean vector $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$
pa(j)	Parental set of series j
s : t	Indices $s, s+1, \cdots, t$

Abbreviations

DLM	Dynamic linear model
MDM	Multiregression dynamic model
SSS	Shotgun stochastic search
BMA	Bayesian model averaging
CSCDLM	Common state coupled dynamic linear model
CSCMDM	Common state coupled multiregression dynamic model
FFBS	Forward-filtering, backward-sampling
MCMC	Markov chain Monte Carlo
TV-VAR	Time-varying vector autoregressive (model)

Acknowledgements

First, I would like to thank my advisor, Mike West, for his insightful guidance, invaluable advice, and constant support throughout my years at the department of statistical science at Duke. Working with Mike has been one of the luckiest experience in my life - pleasantly challenging, and surprisingly rewarding. I am most grateful for the insights, knowledge and skills in statistics he shared with me throughout my graduate studies. As a humorous mentor and a role model, Mike not only influenced me in development of research skills, but also in professional career development, work ethic, and collaborative spirit.

Thanks also go to my thesis committee members, Jonathan Hosking, Surya Tokdar, Li Ma, and prelim exam committee member Francesco Bianchi for their suggestions. I would like to further thank Jon for his invaluable contributions as a collaborator in our joint project, and his mentorship during my summer internship at IBM T.J. Watson Research Center. Thanks also go to Fan Li for her guidance as my first-year advisor.

I thank my friends and colleagues for their timely help, inspiring discussions, and pleasant company during my time at Duke. Thanks go to my friends at Duke: Junchi, Yuan, Lan, Shaan, Mengyang, Brian, Shih-Han, Yun, Doug, James, Monika, Jacopo, Nicole, Christoph, Quanli, Kai, Mary Beth, Andrew, Tommy and many others. I would also like to thank my manager at BEAM Bayesian Efficient Asset Management, José Quintana, for sharing his insights and experience helpful not only to my PhD studies, but also to my professional career. I am grateful to my colleagues at IBM and Morgan Stanley, Fei Liu, Hongxia Yang, Yasuo Amemiya, Gangqiang Xia - thank you for providing me guidance, support, and encouragement for my research and work experience during my PhD studies.

Most importantly, I would like to thank my dearest love, Ryan Yu Liu, for coming into my life at the best age, supporting and encouraging my research and career, starting to build our own family, and picturing a bright future in front of us.

Finally, I would also like to express my deepest gratitude to my parents, Jinglong Zhao, Shuping (Pingping) Li, and my aunt Cindy Zhao for their unconditional support, advice and love, for their beyond-distance company throughout my ups and downs. Special thanks to my mom, Pingping, for her constant encouragement, especially during the time when I applied for graduate school, without which I may not have been able to start my beautiful journey at Duke.

1

Introduction

1.1 Multiregression Dynamic Models

Multiregression dynamic models (MDMs) (Queen and Smith, 1993) are multivariate non-Gaussian DLMs which preserve certain conditional independence structures related to causality over time. A MDM essentially decouples dynamic modeling of multivariate time series into sets of univariate dynamic models processed in parallel, and recouple them for forecasting and decisions. Parallel processing efficiently scales down the computational time for multivariate modeling of higher dimensional series, and significantly alleviates the computation burdens of MCMC for various types of multivariate time series models, such as the Latent Threshold Models in Nakajima and West (2013). The $1/k$-step-ahead ($k \geqslant 1$) predictive covariance structures in MDMs are develped in Queen et al. (2008), but the results on k-step-ahead forecasts only apply to the case when regression predictors $\boldsymbol{x}_{t+k}$ are known at time t, thus excluding the vector autoregressive MDMs. We address this issue by a Monte Carlo method.

Approaches to modeling stochastic volatility are crucial to research in forecasting and portfolio management, and in practical applications in industry. Theories and applications of various approaches of stochastic volatility modeling have been concurrently constructed (e.g. Chib et al., 2006; Doz and Renault, 2006; Zhou et al., 2014). From a completely new perspective on stochastic volatility modeling, I consider the decomposition in MDMs as essentially a Cholesky decomposition of stochastic covariance matrices. The conditional dependence network allows us to factor the stochastic modeling of covariance structure into univariate DLMs with predictor set including simutaneous endogenous variables. This concords with the reformulation of MDMs by Lopes et al. (2012) from the perspective of modeling the Cholesky roots of time-varying covariance matrices.

MDMs essentially offer a closed-form, conjugate analysis of time-varying covariance structures in multivariate state space models. Previous work has limited closed-form developments in modeling time-varying covariances, often using an approximation of some kind (e.g. Triantafyllopoulos, 2007). Nakajima and West (2013) reformulates with an evolution model for the covariance matrix and MCMC methods for model fitting; AR(1) models are used for the elements of Cholesky decompositions of observation covariance matrices. However, this is computationally challenging when applied to data of higher dimensions. With a conjugate analysis, MDMs avoid the computational burden of MCMC and enable various patterns of stochastic volatility modeling.

As illustrated by Carvalho and West (2007), sparse graphical modeling has been recognized as of benefit in inducing robust estimation of model parameters and enhancing of computational feasibility. Since the conditional dependence structure can be represented by a directed acyclic graph (DAG), MDMs are essentially a class of graphical models for multivariate time series. By decoupling a multivariate dynamic model to a set of univariate dynamic models, MDMs more naturally allow flexible sparse graphical modeling; this underlies my approach. One of my innovations is establishing the concept of "sparsity" in conditional dependence networks and developing a method to enable dynamic modeling of these sparse networks under an inherently parallelizable framework.

MDMs also enrich the class of stochastic evolution models by providing the flexibility of discounting for state and observation noise structures. Traditional multivariate Gaussian and non-Gaussian DLMs with Wishart matrix discounting, as specified in Chapter 10 of Prado and West (2010), essentially discounts historical cumulative information of square errors by a discount factor of β at each consecutive time point to reflect the effects of variation over time. However, as pointed out in Prado and West (2010), maintaining a valid model over time requires β to fall in the interval $((m-2)/(m-1), 1]$, where m is the dimension of the time series vector. When m increases, however, β quickly converges to 1 and thus limits practical utility. This limitation affects the capability of the Wishart discounting model when approaching high-dimensional data. MDMs avoid this problem, decomposing the multivariate DLMs into sets of univariate DLMs, where no lower bound on β is required. Additionally, unlike the Wishart discounting model, MDMs allow for series-specific discount factors for state and observation noise, and thus engender a richer class of stochastic evolution models.

In this dissertation, MDMs are first reviewed as multivariate dynamic linear models, also known as state space models. This context also illustrates how the MDM idea enriches the classes of multivariate dynamic models from the perspective of Cholesky stochastic volatility modeling (West and Harrison, 1997; Prado and West, 2010). I develop sequential updating, on-line forecasting and decisions under this dynamic conditional dependence network, broadening the current theory and methods of Bayesian dynamic models. The methodology is then incorporated into time-varying vector autoregressive (TV-VAR) models (Primiceri, 2005; Prado and West, 2010), time-varying covariance matrix processes (Quintana and West, 1987; Bollerslev et al., 1992; Aguilar and West, 2000; Shephard, 2005; Carvalho and West, 2007). Applications in time series analysis in financial econometrics and electricity demand forecasting illustrate the impacts of MDMs in terms of enhancements in forecasting performance, practical uses, and computational efficiency.

Dynamic compositional networks require model structure search and uncertainty analysis, which are addressed in Chapters 3 and 4. Chapter 3 introduces a new algorithmic approach to identify high posterior probability MDMs, utilized in parallel for each univariate DLM. The parallelization efficiently scales up computation for multivariate modeling for more than a small number of dimensions, and significantly alleviates the computational burden of MCMC for various types of multivariate time series models, such as the model in Nakajima and West (2013).

Queen (1994) and Anacleto et al. (2013) apply MDMs to different types of multivariate time series problems, such as forecasting of brand sales and traffic flows. While these previous works have illustrated the effectiveness of MDMs in inference and forecasting of multivariate time series, another step further is to address online model structure search and uncertainty analysis. Bayesian model averaging (BMA) serves as a common method to deal with model uncertainty and has a well-known, superior performance in out-of-sample predictions, which can provide numurous practical benefits. In our portfolio application, BMA enables the forecasts of financial time series to adapt to potential regime switches and market shocks, and contributes to constructing robustly lucrative portfolio management strategies.

My innovative BMA is customized to MDM by taking advantage of the decomposable feature to address uncertainties about various types of model parameters. It enables time-varying AR lag structures, conditional dependence networks, and noise evolution discount factors. It essentially transforms the problem of model averaging from the multivariate to coupled univariate contexts. Analytic forms and recursive algorithms are developed to calculate the one-step-ahead forecast means and covariance matrices analytically. Multi-step-ahead forecast moments have complicated analytic forms due to correlated multi-step-ahead lag structure, and thus are calculated by simulation. Additionally, when applying standard BMA on a particular dataset, it is well known that the posterior probability of only one model will converge to one. This dominance of one single model limits the adaptability and potential gains from the averaging process in the long term. Following West and Harrison (1989) and Xie (2013), power discounting techniques are used for adaptively processing new information, defining a dynamic extension of BMA of significant practical utility.

1.2 Common State Coupled Dynamic Linear Model

A new model context– the class of common state coupled dynamic linear models (CSCDLMs) is proposed in this thesis. My interest lies in m univariate series $y_{jt},\ j = 1 : m$, over time t, each following a DLM with its own, individual state vector elements together with an additional state vector that is shared across all series. The motivating context is a problem with large m and multiple common predictors with an assumedly common state vector– so that information about common effects is shared across series– but also with series-specific effects. When assuming the observational noise terms are independent across series and all regression vectors are known at $t-1$, for all $j = 1, \cdots, m$, the CSCDLM is a special class of dynamic hierarchical models (Gamerman and Migon, 1993). One example of the latter– a key example– is when the series-specific effects arise via a compositional DLM representation of multivariate stochastic volatility, a flexible and parsimonious approach to modeling dynamic patterns of variance and precision matrices. As I relax the assumption to correlated, time-varying observation covariance matrix, as well as incorporation of random regression vectors, this is a new area that no previous work has touched, to my knowledge.

Beyond this motivation, this development appears to be breaking important new ground in providing the ability to scale-up computations to higher numbers of related time series, since the forward-filtering analysis is conditionally decoupled to permit analytic updates and forecasting. It follows naturally that this also is inherited by backward-sampling analyses; the backward sampling results generated from the new model and assumptions are detailed as direct extensions of standard theory.

The extension to variance learning is critical for application, as observational error variances in univariate DLMs are usually unknown in reality. Unfortunately, traditional variance discounting– with updated gamma posteriors– cannot be immediately applied to the volatilities. Quintana and West (1987) introduce matrix-variate DLMs, which are matrix-variate linear state space models allowing for covariance estimation, and propose a Bayesian conjugate analysis with an inverse Wishart distribution, further developed in Quintana and West (1988) and Uhlig (1994). This was extended to graphical models in Carvalho and West (2007) and Wang and West (2009). However, the conjugate variance discounting only applies to the common components DLM, when each scalar series y_{jt} has precisely the same set of predictors. Otherwise, there is no closed form analysis.

Reformulating with alternative evolution models are amenable to using Markov chain Monte Carlo (MCMC) methods for model fitting (Sims and Zha, 1998; Primiceri, 2005; Sims and Zha, 2006; Nakajima and West, 2013). However, this kind of method is computationally too challenging for applications requiring fast computation, and inapplicable to sequential analysis.

I develop a modest approximation for MDMs using variance discounting to enable closed-form, sequential updating and filtering and thus avoid MCMC. Results of comparison on posterior model parameters with MCMC analysis justify adequacy of the approximation, as the two methods produce variance estimates close to each other. This approximation method maintains the nice conditional independence structure and the main results and methods, yielding a compositional decoupling that allows sequential analysis of the set of univariate DLMs in parallel. Thus, this method can serve as a building block of larger models of increasing practical interest in business and econometrics, such as dynamic factor models and hierarchical models.

1.3 Outline

This thesis discusses novel developments in Bayesian analytics for the class of multiregression dynamic models. My focus is on increasingly high-dimensional multivariate time series; I aim to enable analysis that can be decomposed into sets of univariate models processed in parallel yet coupled for forecasting and decisions.

Chapter 2 discusses these novel developments in Bayesian analytics for the class of multiregression dynamic models. The innovation involves sparse conditional dependence networks, idiosyncrasies in time-varying auto-regressive lag structures, and flexibility of discounting methods for stochastic volatilities. Conditional dependence networks provide an inherently parallelizable framework under which I develop sequential updating, and online forecasting and decisions, so broadening the current theory and methods of Bayesian dynamic models.

In Chapter 3 I develop a novel Shotgun Stochastic Search (SSS) algorithm. Under the parallelizable framework, this new SSS algorithm allows the stochastic search to move in each dimension simultaneously at each iteration, and much faster than traditional SSS to regions of high probability. These innovations in parallelization resolve issues of computational tractability for increasingly high-dimensional data. The new SSS approach is applied in the context of financial time series for predictive portfolio analysis. It has shown success not only in exploring the space of high posterior probability models, but also in identifying practically lucrative models in this financial application.

Chapter 4 proposes an innovative method to address the uncertainty analysis by transforming the model uncertainties from multivariate context to univariate context using BMA and power discounting techniques. The examples show that this can succeed in effectively capturing time-varying model uncertainties on various model parameters, while also identifying practically superior predictive and lucrative models in financial applications.

Chapter 5 introduces a new class of state space models for multivariate time series– the class of CSCDLMs. Bayesian analytics are developed including the key sequential updating– FFBS scheme. Online and analytic learning of observation variances are achieved by an approximation method using variance discounting. This method results in fast computation for sequential filtering and step-ahead forecasting than approaches using MCMC, satisfying the requirement of speed-up for real-world applications. A key motivating example is the problem of short-term prediction of electricity demand in a Smart Grid scenario. I first include a method to allow time-varying state covariance matrices through the use of discount factors, and then apply the CSCDLM and variance discounting method. Experimental results on the Smart Grid project data set, including comparisons with the previous method, validate the effectiveness of the proposed methods in this applied study.

Chapter 6 concludes the dissertation and comments on potential future work.

2

Multiregression Dynamic Models

2.1 Sparse Conditional Dependence Network

Under the framework of standard multivariate DLMs, MDMs allow for direct inference on conditional independence and causal structure across a multivariate series. Write $\boldsymbol{y}_t$ for the $m \times 1$-vector time series, $t = 1, 2, \cdots, T$, $\mathcal{D}_0$ for all initial prior information, $\mathcal{D}_t = \{\mathcal{D}_{t-1}, \boldsymbol{y}_t\}$ for information sets sequentially updated by received observations. The full one-step-ahead predictive density for $\boldsymbol{y}_t$, conditional on all the information before time t, can be written as the product of a set of conditional pdfs, i.e.

$$p(\boldsymbol{y}_t|\mathcal{D}_{t-1}) = \prod_{j=1:m-1} p(y_{jt}|\boldsymbol{y}_{j+1:m,t}, \mathcal{D}_{t-1})p(y_{mt}|\mathcal{D}_{t-1}). \tag{2.1}$$

Thus, a natural, direct route to specifying a set of univariate DLMs that cohere is via a "triangular" model that defines the joint distribution by composition as in (2.1). This triangular model is such that, for each series j, its predictors may include both the endogenous variables $\boldsymbol{y}_{j+1:m}$ and other exogenous and lagged endogenous variables. This requires a specific ordering of the time series, which we simply take $1, 2, \cdots, m$ without loss of generality.

Imposing sparsity on the conditional dependence network across multiple series is of increasing interest due to a variety of benefits, such as computational efficiency and stability for higher-dimensional data. When fitting the set of m univariate DLMs implied above, I consider specific subsets of compositional endogenous variables, denoted by parental set, $pa(j) \subseteq \{j+1:m\}$, for each j (and with $pa(m)$ empty). This defines a set of independent univariate DLMs, with time-varying dependence structure preserved through the time-varying coefficients of the predictors in the parental sets. The parental sets reflect the conditional dependence structure between endogenous variables. For instance, $j+1 \notin pa(j)$ means $y_{j+1,t} \perp\!\!\!\perp y_{jt} | \boldsymbol{y}_{pa(j),t}$. Sparsity structure is constructed by selecting "appropriate" parental sets, generally with small sizes. In fact, the smaller the sizes of parental sets, the more sparse the conditional dependence structure is.

2.2 Model Framework

I define triangular equations for the set of univariate, compositional regressions for elements of the column $m-$vector $\boldsymbol{y}_t$ in the context of an overall set of univariate DLMs. Define $\mathcal{P} = \{pa(j), j = 1:m\}$ for all the parental sets. For each $j = 1:m$, write $pa(j) \in \{j+1:m\}$ for the index set of parents of y_{jt} in the triangular system, noting that $pa(m) = \varnothing$, the empty set. Note that this allows for sparsity in that $pa(j)$ may be a few (or no) elements, while it can also be the full set of potential parent variables. Then, over $j = 1:m$, the coupled models are defined as follows:

$$
\begin{aligned}
y_{jt} &= \boldsymbol{x}'_{jt}\boldsymbol{\phi}_{jt} + \boldsymbol{\gamma}'_{jt}\boldsymbol{y}_{pa(j),t} + \nu_{jt}, \\
&= \boldsymbol{F}'_{jt}\boldsymbol{\theta}_{jt} + \nu_{jt},
\end{aligned}
$$

where:

- $\boldsymbol{\gamma}_{jt}$ is the vector of dynamic regression coefficients γ_{jht}, $h \in pa(j)$, with dimension $p_{j\gamma} = |pa(j)|$;

- $\boldsymbol{x}_{jt}$ is a known column vector of predictors or constants, with corresponding dynamic regression coefficients in the column state vector $\boldsymbol{\phi}_{jt}$, each of dimension $p_{j\phi}$;

- the full state and regression vectors, each of dimension $p_j = p_{j\phi} + p_{j\gamma}$, are

$$\boldsymbol{\theta}_{jt} = \begin{pmatrix} \boldsymbol{\phi}_{jt} \\ \boldsymbol{\gamma}_{jt} \end{pmatrix} \quad \text{and} \quad \boldsymbol{F}_{jt} = \begin{pmatrix} \boldsymbol{x}_{jt} \\ \boldsymbol{y}_{pa(j),t} \end{pmatrix};$$

- the $\boldsymbol{\theta}_{jt}$ follow some specified state evolutions models, such as vector random walks, TV-VAR models, etc.

- I assume conjugate DLM structure for learning, including use of discounting for states and observation variances.

For notational convenience, set $\widetilde{\gamma}_{jht} = 0$, for $h \notin pa(j)$, and $\widetilde{\boldsymbol{\gamma}}_{j,pa(j),t} = \boldsymbol{\gamma}_{jt}$. Collect the effective coefficients $\boldsymbol{\gamma}_{jt}$ along with the implicit zero values in the matrix

$$\boldsymbol{\Gamma}_t = \begin{pmatrix} 0 & \widetilde{\gamma}_{12t} & \widetilde{\gamma}_{13t} & \cdots & \widetilde{\gamma}_{1mt} \\ 0 & 0 & \widetilde{\gamma}_{23t} & \cdots & \widetilde{\gamma}_{2mt} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & & & 0 & \widetilde{\gamma}_{m-1,mt} \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix} \tag{2.2}$$

such that

$$(\boldsymbol{I} - \boldsymbol{\Gamma}_t)\boldsymbol{y}_t = \boldsymbol{\mu}_t + \boldsymbol{\nu}_t,$$

where $\boldsymbol{\mu}_t = (\mu_{1t}, \ldots, \mu_{mt})'$ with $\mu_{jt} = \boldsymbol{x}'_{jt}\boldsymbol{\phi}_{jt}$. Under the independent normal model $\boldsymbol{\nu}_t \sim N(\boldsymbol{0}, \boldsymbol{\Lambda}_t^{-1})$ with precision matrix $\boldsymbol{\Lambda}_t = \text{diag}(\lambda_{1t}, \ldots, \lambda_{mt})$, we then have $\boldsymbol{y}_t \sim N((\boldsymbol{I} - \boldsymbol{\Gamma}_t)^{-1}\boldsymbol{\mu}_t, \boldsymbol{\Omega}_t^{-1})$ where

$$\boldsymbol{\Omega}_t = (\boldsymbol{I} - \boldsymbol{\Gamma}_t)'\boldsymbol{\Lambda}_t(\boldsymbol{I} - \boldsymbol{\Gamma}_t).$$

That is, $(\boldsymbol{I} - \boldsymbol{\Gamma}_t)'$ is the lower triangular (left) Cholesky component of the precision matrix of $\boldsymbol{y}_t$ conditional on $\boldsymbol{\mu}_t, \boldsymbol{\Gamma}_t$, subject to scaling of its rows by the square roots of

λ_j. We can also represent the network structure of a MDM by defining the indicator matrix $\boldsymbol{Z}$ such that $\zeta_{jh} = 0$ for $h \notin pa(j)$, otherwise $\zeta_{jh} = 1$, as the following:

$$\boldsymbol{Z} = \begin{pmatrix} 0 & \zeta_{12} & \zeta_{13} & \cdots & \zeta_{1m} \\ 0 & 0 & \zeta_{23} & \cdots & \zeta_{2m} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & & & 0 & \zeta_{m-1,m} \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}$$

2.3 Forward-Filtering and Predictions: Discount Models

2.3.1 Summary of univariate models

Assume a specified state evolution for each of the $\boldsymbol{\theta}_{jt}$, and with corresponding variance discounting for stochastic volatility using discount factors β_j for each of the observation precisions λ_j. The usual conjugate updating and forecasting equations then arise, with normal-inverse gamma priors and posteriors for states and variances at each stage. Assuming the priors at $t = 0$ are independent across the m series, the sequential updating applies in parallel. I follow the usual DLM notation (West and Harrison, 1997; Prado and West, 2010), where $(\boldsymbol{\theta}, \lambda) \sim NIG(\boldsymbol{\mu}, \boldsymbol{\Sigma}, n, d)$ denote that $\boldsymbol{\theta} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}/\lambda)$, and $\lambda \sim Gamma(n, d)$. At time t, we have the following.

Posterior at time $t-1$:

$$(\boldsymbol{\theta}_{j,t-1}, \lambda_{j,t-1} | \mathcal{D}_{t-1}) \sim NIG(\boldsymbol{m}_{j,t-1}, \boldsymbol{C}_{j,t-1}, n_{j,t-1}/2, d_{j,t-1}/2)$$

with $s_{j,t-1} = d_{j,t-1}/n_{j,t-1}$ being the usual point estimates of residual variances $1/\lambda_{jt}$.

One-step-ahead prior at time $t-1$:

$$(\boldsymbol{\theta}_{jt}, \lambda_{jt} | \mathcal{D}_{t-1}) \sim NIG(\boldsymbol{a}_{jt}, \boldsymbol{R}_{jt}, r_{jt}/2, h_{jt}/2)$$

where $\boldsymbol{a}_{jt}, \boldsymbol{R}_{jt}$ arise from the state evolution step, and $r_{jt} = \beta_j n_{j,t-1}$ and $h_{jt} = \beta_j d_{j,t-1}$. Note that the residual variances estimates are $s_{j,t-1} = d_{j,t-1}/n_{j,t-1} \equiv h_{jt}/r_{jt}$.

One-step-ahead forecasts at time $t-1$:

$$(y_{jt}|\boldsymbol{y}_{pa(j),t}, \mathcal{D}_{t-1}) \sim T_{r_{jt}}(f_{jt}(\boldsymbol{y}_{pa(j),t}), q_{jt}(\boldsymbol{y}_{pa(j),t}))$$

where $f_{jt}(\boldsymbol{y}_{pa(j),t}) = \boldsymbol{F'}_{jt}\boldsymbol{a}_{jt}$ and $q_{jt}(\boldsymbol{y}_{pa(j),t}) = s_{j,t-1} + \boldsymbol{F'}_{jt}\boldsymbol{R}_{jt}\boldsymbol{F}_{jt}$. Write

$$\boldsymbol{a}_{jt} = \begin{pmatrix} \boldsymbol{a}_{j\phi t} \\ \boldsymbol{a}_{j\gamma t} \end{pmatrix} \quad \text{and} \quad \boldsymbol{R}_{jt} = \begin{pmatrix} \boldsymbol{R}_{j\phi t} & \boldsymbol{R}_{j\phi\gamma t} \\ \boldsymbol{R}'_{j\phi\gamma t} & \boldsymbol{R}_{j\gamma t} \end{pmatrix}.$$

Then

$$\begin{aligned} f_{jt}(\boldsymbol{y}_{pa(j),t}) &= \boldsymbol{x}'_{jt}\boldsymbol{a}_{j\phi t} + \boldsymbol{y}'_{pa(j),t}\boldsymbol{a}_{j\gamma t}, \\ q_{jt}(\boldsymbol{y}_{pa(j),t}) &= s_{j,t-1} + \boldsymbol{y}'_{pa(j),t}\boldsymbol{R}_{j\gamma t}\boldsymbol{y}_{pa(j),t} + 2\boldsymbol{y}'_{pa(j),t}\boldsymbol{R}'_{j\phi\gamma t}\boldsymbol{x}_{jt} + \boldsymbol{x}'_{jt}\boldsymbol{R}_{j\phi t}\boldsymbol{x}_{jt}. \end{aligned} \tag{2.3}$$

k-step-ahead prior at time t:

$$(\boldsymbol{\theta}_{j,t+k}, \lambda_{j,t+k}|\mathcal{D}_t) \sim NIG(\boldsymbol{a}_{jt}(k), \boldsymbol{R}_{jt}(k), r_{jt}(k)/2, h_{jt}(k)/2)$$

where $\boldsymbol{a}_{jt}(k), \boldsymbol{R}_{jt}(k)$ are updated inductively from $\boldsymbol{a}_{jt}(k-1), \boldsymbol{R}_{jt}(k-1)$, and $r_{jt}(k) = \beta_j n_{j,t}, h_{jt}(k) = \beta_j d_{j,t}$. Note again that the residual variances estimates stay the same as the value at time t.

k-step-ahead forecasts at time t: when $\boldsymbol{x}_{j,t+k}$ is a known column vector of predictors or constants at time t, we have the following:

$$(y_{j,t+k}|\boldsymbol{y}_{pa(j),t+k}, \mathcal{D}_t) \sim T_{r_{jt}}(f_{j,t+k}(\boldsymbol{y}_{pa(j),t+k}), q_{j,t+k}(\boldsymbol{y}_{pa(j),t+k})) \tag{2.4}$$

where $f_{j,t+k}(\boldsymbol{y}_{pa(j),t+k}) = \boldsymbol{F'}_{j,t+k}\boldsymbol{a}_{j,t}(k)$ and $q_{j,t+k}(\boldsymbol{y}_{pa(j),t+k})) = s_{j,t-1} + \boldsymbol{F'}_{j,t+k}\boldsymbol{R}_{j,t}(k)\boldsymbol{F}_{j,t+k}$. Write

$$\boldsymbol{a}_{j,t}(k) = \begin{pmatrix} \boldsymbol{a}_{j\phi t}(k) \\ \boldsymbol{a}_{j\gamma t}(k) \end{pmatrix} \quad \text{and} \quad \boldsymbol{R}_{j,t}(k) = \begin{pmatrix} \boldsymbol{R}_{j\phi t}(k) & \boldsymbol{R}_{j\phi\gamma t}(k) \\ \boldsymbol{R}_{j\phi\gamma t}(k)' & \boldsymbol{R}_{j\gamma t}(k). \end{pmatrix}$$

Then

$$
\begin{aligned}
f_{j,t+k}(\boldsymbol{y}_{pa(j),t+k}) =& \boldsymbol{x}'_{j,t+k}\boldsymbol{a}_{j\phi t}(k) + \boldsymbol{y}'_{pa(j),t+k}\boldsymbol{a}_{j\gamma t}(k) \\
q_{j,t+k}(\boldsymbol{y}_{pa(j),t+k})) =& s_{j,t-1} + \boldsymbol{y}'_{pa(j),t+k}\boldsymbol{R}_{j\gamma t}(k)\boldsymbol{y}_{pa(j),t+k} + 2\boldsymbol{y}'_{pa(j),t+k}\boldsymbol{R}_{j\phi\gamma t}(k)'\boldsymbol{x}_{j,t+k} \\
&+ \boldsymbol{x}'_{j,t+k}\boldsymbol{R}_{j\phi t}(k)\boldsymbol{x}_{j,t+k}.
\end{aligned}
\tag{2.5}
$$

2.3.2 Joint predictive density

The full 1-step-ahead predictive density for $\boldsymbol{y}_t$ is

$$
p(\boldsymbol{y}_t|\mathcal{D}_{t-1}) = \prod_{j=1:m} p(y_{jt}|\boldsymbol{y}_{pa(j),t}, \mathcal{D}_{t-1}). \tag{2.6}
$$

Since the univariate conditional distribution of $y_{jt}|\boldsymbol{y}_{pa(j),t}, \mathcal{D}_{t-1}$ is a T distribution, characteristics of (2.6) are easily evaluated based on observed data. The observed value of (2.6) also feeds into computation of model marginal likelihoods for model assessment, such as for exploring data-supported values of discount factors, choice of parental index sets, assumptions about structure underlying the model components μ_{jt}, and so forth.

2.3.3 Joint predictive mean, variance matrix and precision matrix

For portfolio applications based on mean-variance optimizations and trade-offs, the moments

$$
\boldsymbol{f}_t = E(\boldsymbol{y}_t|\mathcal{D}_{t-1}), \quad \boldsymbol{Q}_t = V(\boldsymbol{y}_t|\mathcal{D}_{t-1}) \quad \text{and precision} \quad \boldsymbol{K}_t = V(\boldsymbol{y}_t|\mathcal{D}_{t-1})^{-1}
$$

under (2.6) are required. Assume that, for all j, t, the degrees-of-freedom parameters r_{jt} exceed 2, so that these moments all exist. The recursive form of the compositional model means that we must compute these via a sequential algorithm that recognizes the appearance of contemporaneous values $\boldsymbol{y}_{pa(j),t}$ in the conditioning of forecasts for y_{jt}. The computations for one-step-ahead predictions are performed below; k-step-ahead computations are similar, and so are omitted here. For each $j = 1 : m-1$, denote the mean vector, variance matrix and precision matrix of the $(m-j)-$vector $\boldsymbol{y}_{j+1:m,t}$ by

$$\begin{aligned}\boldsymbol{f}_{j+1:m,t} &= E(\boldsymbol{y}_{j+1:m,t}|\mathcal{D}_{t-1}), \quad \boldsymbol{Q}_{j+1:m,t} = V(\boldsymbol{y}_{j+1:m,t}|\mathcal{D}_{t-1}),\\ \boldsymbol{K}_{j+1:m,t} &= V(\boldsymbol{y}_{j+1:m,t}|\mathcal{D}_{t-1})^{-1}.\end{aligned} \tag{2.7}$$

Note that at $j = m-1$ these values reduce to scalars. We proceed as follows to sequentially compute elements of the required $\boldsymbol{f}_t$, $\boldsymbol{Q}_t$ and $\boldsymbol{K}_t$.

Joint predictive mean, variance matrix

i. Start at $j = m$: compute the univariate mean, variance of y_{mt},

$$f_{mt} = f_{mt}(\varnothing), \quad q_{mt} = \frac{r_{mt}}{(r_{mt}-2)} q_{mt}(\varnothing)$$

using the implied simplified forms of (2.3) with all terms in $\boldsymbol{y}_{pa(m)}$ set to zero. Insert f_{mt} as the $m-$the element of $\boldsymbol{f}_t$ and q_{mt} as the $(m,m)-$element of $\boldsymbol{Q}_t$.

ii. For $j = m-1, m-2, \ldots, 1$ in turn:

- At this point, we have already computed the values of the moments of (5.25) from the previous steps. These are used in the following calculations. First, write $\boldsymbol{f}_{pa(j),t}$ for the subvector of means in $\boldsymbol{f}_{j+1:m,t}$ on elements in $pa(j)$ only, and $\boldsymbol{Q}_{pa(j),t}$ for the corresponding submatrix of $\boldsymbol{Q}_{j+1:m,t}$. Then, the marginal mean f_{jt} and variance q_{jt} of y_{jt} are computed as follows.

$$\begin{aligned} f_{jt} &= \boldsymbol{x}'_{jt}\boldsymbol{a}_{j\phi t} + \boldsymbol{f}'_{pa(j),t}\boldsymbol{a}_{j\gamma t},\\ q_{jt} &= \frac{r_{jt}}{(r_{jt}-2)}(s_{j,t-1} + u_{jt}) + \boldsymbol{a}'_{j\gamma t}\boldsymbol{Q}_{pa(j),t}\boldsymbol{a}_{j\gamma t}\end{aligned} \tag{2.8}$$

  where

$$u_{jt} = \boldsymbol{f}'_{pa(j),t}\boldsymbol{R}_{j\gamma t}\boldsymbol{f}_{pa(j),t} + \mathrm{tr}(\boldsymbol{R}_{j\gamma t}\boldsymbol{Q}_{pa(j),t}) + 2\boldsymbol{x}'_{jt}\boldsymbol{R}_{j\phi\gamma t}\boldsymbol{f}_{pa(j),t} + \boldsymbol{x}'_{jt}\boldsymbol{R}_{j\phi t}\boldsymbol{x}_{jt}.$$

Insert f_{jt} as the $j-$th element of $\boldsymbol{f}_t$ and q_{jt} as the $(j,j)-$element of $\boldsymbol{Q}_t$, respectively.

- Compute the covariance vector $C(y_{jt}, \boldsymbol{y}_{j+1:m,t})|\mathcal{D}_{t-1}$ as follows. Write $\boldsymbol{a}_{j+1:m\gamma t}$ for the $(m-j)-$vector that extends $\boldsymbol{a}_{j\gamma t}$ with zeros for elements $h \notin pa(j)$. Then

$$C(y_{jt}, \boldsymbol{y}_{j+1:m,t}|\mathcal{D}_{t-1}) = \boldsymbol{Q}_{j+1:m,t}\boldsymbol{a}_{j+1:m\gamma t} = \boldsymbol{q}_{j,j+1:m,t}.$$

  Insert element h of this vector as the $(j, h+j)$ and $(h+j, j)$ entries of $\boldsymbol{Q}_t$, for $h = 1 : m-j$.

iii. End: at this point, $j = 1$ and we have filled in all elements of the $m-$vector $\boldsymbol{f}_t$, $m \times m$ matrix $\boldsymbol{Q}_t$.

Joint predictive precision matrix

The precision matrix directly reflects the conditional dependence structure between variables and thus plays a crucial component in various types of analysis. In our application, for example, it has a determining effect on the portfolio allocation weights. To compute the precision matrix at each time t, we avoid direct inversion of the covariance matrix by utilizing interim products, the covariance vectors $\boldsymbol{q}_{j,j+1:m,t}$ already calculated above. This block-wise inversion decreases the computational instability and reduces complexity to $O(m^2)$, especially critical under sparse conditional dependence networks in higher dimensions. The computation of $\boldsymbol{K}_t$ is as follows.

i. Start at $j = m$: compute the conditional precision of y_{mt}, namely $K_{m:m,t} = 1/q_{mt}$.

ii. For $j = m-1, m-2, \ldots, 1$ in turn:

- Given the computed $\boldsymbol{q}_{j,j+1:m,t}$, compute the precision matrix $\boldsymbol{K}_{j:m,t}$ via

$$\boldsymbol{K}_{j:m,t} = \begin{pmatrix} k_{jt} & \boldsymbol{h}_{jt} \\ \boldsymbol{h}'_{jt} & \boldsymbol{H}_{j,t} \end{pmatrix}$$

with entries

$$\begin{aligned} k_{jt}^{-1} &= q_{jt} - \boldsymbol{q}_{j,j+1:m,t}\boldsymbol{K}_{j+1:m,t}\boldsymbol{q}'_{j,j+1:m,t}, \\ \boldsymbol{h}_{jt} &= -k_{jt}\boldsymbol{q}_{j,j+1:m,t}\boldsymbol{K}_{j+1:m,t}, \\ \boldsymbol{H}_{j,t} &= \boldsymbol{K}_{j+1:m,t} + k_{jt}^{-1}\boldsymbol{h}'_{jt}\boldsymbol{h}_{jt}. \end{aligned}$$

Notice that k_{jt} is a scalar, so this recursive computation avoids matrix inversions.

iii. End: at this point, $j = 1$ and we have filled in all elements of the precision matrix $\boldsymbol{K}_t$, i.e. $\boldsymbol{K}_{1:m,t}$.

2.4 Multiregression Dynamic VAR Models

We now consider the MDM in multivariate time series analysis using time-varying parameter vector autoregressive (TV-VAR) models. For the $m \times 1$-vector time series $\boldsymbol{y}_t$, $(t = 1, 2, \cdots)$, the TV-VAR model consists of m coupled univariate time-varying autoregressive (TVAR) DLMs. Denote the autoregressive order of series j to be $p_{j\lambda}$. For each $j \in \{1, \cdots, m\}$, the model is a univariate TVAR($p_{j\lambda}$) DLM:

$$y_{jt} = c_{jt} + \sum_{i=1}^{p_{j\lambda}} \boldsymbol{\phi}'_{jit}\boldsymbol{y}_{t-i} + \boldsymbol{\gamma}'_{jt}\boldsymbol{y}_{pa(j),t} + \nu_{jt}, \tag{2.9}$$

where c_{jt} is the time-varying intercept, $\boldsymbol{\phi}_{jit}$ the vector containing time-varying coefficients at lag i, $(i = 1, \cdots, p_{j\lambda})$, ν_{jt} the observation noise. The model is essentially a special case of a MDM with intercept and AR lagged variables as predictors. Let $p_\lambda = \max(p_{1\lambda}, \cdots, p_{m\lambda})$. Write $\tilde{\boldsymbol{\phi}}_{jit}$ for the $p_\lambda \times 1$-vector that extends $\boldsymbol{\phi}_{jit}$ with zeros for the elements of subscript larger than $p_{j\lambda}$. Then (2.9) can be written in the vector form

$$(\boldsymbol{I} - \boldsymbol{\Gamma}_t)\boldsymbol{y}_t = \boldsymbol{c}_t + \boldsymbol{\phi}_{1t}\boldsymbol{y}_{t-1} + \cdots + \boldsymbol{\phi}_{p_\lambda t}\boldsymbol{y}_{t-p_\lambda} + \boldsymbol{\nu}_t,$$

where $\boldsymbol{\phi}_{it} = (\tilde{\boldsymbol{\phi}}_{1it}, \cdots, \tilde{\boldsymbol{\phi}}_{mit})^T$, and $\boldsymbol{\Gamma}_t$ has the same form as in (5.13). In the portfolio application, I consider the simple but common VAR model where the autoregressive predictor variables of series j only contain its own lags and the intercept:

$$y_{jt} = c_{jt} + \sum_{i=1}^{p_{j\lambda}} \phi_{jit} y_{j,t-i} + \boldsymbol{\gamma}'_{jt}\boldsymbol{y}_{pa(j),t} + \nu_{jt}, \tag{2.10}$$

where ϕ_{jit} is the time-varying coefficient of series j at lag i, $(i = 1, \cdots, p_{j\lambda})$.

Most of the forward-filtering and 1-step-ahead forecast results in Section 2.3.1 hold for the multiregression VAR models. The exception is (2.5), on k-step-ahead forecasts, since the predictors $\boldsymbol{x}_{j,t+k}$ in the VAR model are not known at time t.

3

Shotgun Stochastic Model Search

3.1 Introduction

This chapter proposes a new algorithmic approach to fast exploration of the high-dimensional MDM model space.

Traditional methods via MCMC are known for demanding computation and difficulties in addressing model uncertainty. Hans et al. (2007) introduced shotgun stochastic search (SSS), now a traditional method for rapidly exploring regions of high-dimensional model spaces in regression. Jones et al. (2005) viewed stochastic computation in Gaussian graphical models from the perspective of high-dimensional model search, with a particular interest in the scalability with dimension of Markov Chain Monte Carlo (MCMC) and other stochastic search methods via examples range from low (12-20) to moderate (150) dimensions and exhibited the superiority of SSS over MCMC.

I am motivated by the same goal and the desire to parallelize computation. Taking advantage of the compositional network structure of MDMs, I propose a parallel algorithmic approach to utilize SSS for each component univariate DLM, simultaneously. Parallel computation efficiently scales computation for multivariate modeling to more than a small number of dimensions, and significantly alleviates the computational burden of MCMC for various types of multivariate time series models.

For simplicity, the method is introduced under the framework of local trend MDMs, namely:

$$\begin{aligned} y_{jt} &= \phi_{jt} + \boldsymbol{\gamma}'_{jt}\boldsymbol{y}_{pa(j),t} + \nu_{jt}, \quad \nu_{jt} \sim N(0, \lambda_{jt}) \\ &= \boldsymbol{F}'_{jt}\boldsymbol{\theta}_{jt} + \nu_{jt}, \\ \boldsymbol{\theta}_{jt} &= \boldsymbol{\theta}_{j,t-1} + \boldsymbol{w}_{jt}, \quad \boldsymbol{w}_{jt} \sim N(0, \boldsymbol{W}_{jt}) \end{aligned} \tag{3.1}$$

where $\boldsymbol{W}_{jt} = (1-\delta_j)\boldsymbol{C}_{j,t-1}/\delta_j$, for $j = 1, \cdots, m$. Notice that ϕ_{jt} is the time-varying intercept and $\boldsymbol{F}_{jt} = (1, \boldsymbol{y}'_{pa(j),t})'$. The state evolution models are vector random walks. Sequential updating and forecasting are described in detail in West and Harrison (1997).

Let $\mathcal{P}$ represent the conditional dependence network $\{pa(j), j = 1, \cdots, m\}$. Denote the discount factor set by $\boldsymbol{S} = (\boldsymbol{S}_1, \cdots, \boldsymbol{S}_m)$, where $\boldsymbol{S}_j = (\delta_j, \beta_j)$, for $j = 1, \cdots, m$. Let $\mathcal{D}_t(j) = \{\mathcal{D}_0, \boldsymbol{y}_{j:m,1:t}\}$ denote the information set that comprises all the information of series $j, j+1, \cdots, m$ available at t.

In the application, with $m = 13$, the data are daily prices of 9 exchange rates, 2 commodities, and 2 U.S. stock indices in dollars over a time-span of more than 5 years. I discuss the portfolio results to illustrate the practical efficacy of MDMs in generating profitable strategies, and compare different model results to illustrate the benefits of parsimony of parental sets and flexibility of variable-wise discount factors.

Section 3.2 discusses priors and likelihoods in MDMs. The important concept of neighbors is defined in Section 3.3. Section 3.4 constructs the algorithm for the new SSS. Section 3.5 details a financial time series application, and Section 3.6 exhibits the results. Section 3.7 provides summary comments on MDMs and scalability of stochastic search computation, and suggests further research directions including broader questions of sparsity modeling and model uncertainty.

3.2 Priors and Likelihoods

3.2.1 Priors

To stochastically or deterministically explore the space of MDMs, I specify priors for model parameters, parental sets and discount factors.

The interest in sparse models favors priors on the parental sets $pa(j)$ in the compositional network that can penalize large values of $p_{j\gamma} = \#pa(j)$. I use the traditional Bayesian Bernoulli prior, specified in terms of a common inclusion/exclusion probability across all series. The parental sets are encouraged to have sparsity under this prior, via a Bernoulli prior on each variable inclusion indicator variable, with parameter ρ. Thus, parental set $pa(j)$ with $p_{j\gamma}$ elements has prior probability $P(pa(j)) = \rho^{p_{j\gamma}}(1-\rho)^{m-j-p_{j\gamma}}$, and an expected size of $\rho \times (m-j)$, which provides a crude quantitative measure of selecting the ρ. It is worth noticing that ρ has to be sufficiently small to favor sparse parental sets. Smaller ρ values more strongly penalize complexity and induce sparsity of the resulting compositional network.

Additionally, let $\boldsymbol{\Psi} = (\boldsymbol{\psi}_1, \cdots, \boldsymbol{\psi}_K)$ denote the K possible discount factors (δ_j, β_j), i.e. $\boldsymbol{S}_j = \boldsymbol{\psi}_{n_j} = (\delta_j, \beta_j)$, for $j = 1, \cdots, m$. where $n_j \in \{1, 2, \cdots, K\}$. I assume independent priors for each $\boldsymbol{S}_j$ with equal probability on every element in $\boldsymbol{\Psi}$. Naturally, I suppose that the prior on discount factors $\boldsymbol{S}$ is independent of $\mathcal{P}$, i.e. $P(\boldsymbol{S}|\mathcal{P}) = P(\boldsymbol{S})$.

3.2.2 Likelihoods and posteriors

With $p(\boldsymbol{y}_{j,1:t}|pa(j), \boldsymbol{S}_j) = \prod_{\tau=1}^{T} p(y_{j\tau}|\boldsymbol{y}_{pa(j),\tau}, \boldsymbol{S}, \mathcal{D}_{\tau-1})$, the likelihood function over $\mathcal{P}, \boldsymbol{S}$ can be expressed as

$$p(\boldsymbol{y}_{1:t}|\mathcal{P}, \boldsymbol{S}) = \prod_{j=1}^{m} p(\boldsymbol{y}_{j,1:t}|pa(j), \boldsymbol{S}_j) = \prod_{j=1}^{m}\prod_{\tau=1}^{T} p(y_{j\tau}|\boldsymbol{y}_{pa(j),\tau}, \boldsymbol{S}_j, \mathcal{D}_{\tau-1}) \qquad (3.2)$$

where $p(y_{j\tau}|\boldsymbol{y}_{pa(j),\tau}, \boldsymbol{S}_j, \mathcal{D}_{\tau-1})$ is calculated via equation (2.6). Thus, the posterior probability of a model with parameters $(\mathcal{P}, \boldsymbol{S})$ is

$$P(\mathcal{P}, \boldsymbol{S}|\boldsymbol{y}_{1:t}) \propto P(\boldsymbol{S})p(\boldsymbol{y}_{1:t}|\mathcal{P}, \boldsymbol{S}) \prod_{j=1}^{m} \rho^{p_{j\gamma}}(1-\rho)^{m-j-p_{j\gamma}}. \tag{3.3}$$

This posterior can be decomposed as a product of the posterior probabilities of the univariate models with parameters being the parental sets $pa(j)$ and the corresponding discount factors $\boldsymbol{S}_j$, according to equations (3.2) and (3.3). That is,

$$P(\mathcal{P}, \boldsymbol{S}|\boldsymbol{y}_{1:t}) = \prod_{j=1}^{m} P(pa(j), \boldsymbol{S}_j|\boldsymbol{y}_{j,1:t}), \tag{3.4}$$

where

$$P(pa(j), \boldsymbol{S}_j|\boldsymbol{y}_{j,1:t}) \propto p(\boldsymbol{y}_{j,1:t}|pa(j), \boldsymbol{S}_j)\rho^{p_{j\gamma}}(1-\rho)^{m-j-p_{j\gamma}}P(\boldsymbol{S}_j). \tag{3.5}$$

This decomposition indicates that stochastic search for $\mathcal{P}$ and $\boldsymbol{S}$ of high posterior probability can be paralleled by searching $pa(j), \boldsymbol{S}_j$ separately. That is, for each j, I can apply stochastic search with respect to the defined set of candidate predictors, the parental set $pa(j)$, and discount factors $\boldsymbol{S}_j$ and score each DLM by the posterior model probabilities $P(pa(j), \boldsymbol{S}_j|\boldsymbol{y}_{j,1:t})$.

Notice that although the order of the variables may have substantial impact on the accuracy and stability of the forecasts, I focus on illustrating the impact of variable-wise discount factors and sparse parental sets on forecasting. Future research related to the ordering of the variables can be an interesting and new area to explore.

To identify regions of high posterior probability over both parental sets and discount factors of MDMs, I propose a new shotgun stochastic search approach. From the above, given the discount factors $\boldsymbol{S}$, a compositional network $\mathcal{P}$ defines a MDM, $M \triangleq (\boldsymbol{S}, \mathcal{P})$, called "models" for convenience. Obviously, the space of M is high-dimensional. I calculate the number of all possible models, which is the product of the number of possible discount factors combinations and the number of possible parental sets $\mathcal{P}$. It is easy to show that the former is m^K, while the latter is $2^{m(m-1)/2}$, so the total number of possible models is $m^K \times 2^{m(m-1)/2}$. This implies that when m is more than a small number, enumerating all possible models is an unrealistically huge computation task.

SSS is a fast and popular tool in exploring higher-dimensional graphical models. A connection between graphs and the parental sets can be easily built since both graphs and parental sets are essentially defined by sets of included predictors. However, I generalize SSS from just searching the inclusion of variables, i.e., parental sets, to searching more parameters, i.e. both parental sets and different sets of discount factors. In this setting, I am able to identify both good discount factors and parental sets simultaneously. The following discusses the two major components of SSS in detail: the choice of neighborhood and the model move (sampling) strategy.

3.3 Choices of Neighborhoods and Sampling Moves

The neighborhood of a MDM M, denoted by $nbd(M)$, is defined to contain two parts: the parental set neighborhood and discount factors $\boldsymbol{S}$ for each variable. In other words, a model in the neighborhood of the current MDM is such a model that its parental set is a neighbor of the current parental sets $\mathcal{P}$, and its discount factors for each series can be any element in $\boldsymbol{\Psi}$. For notation, I have $nbd(pa(j))$ as the neighborhood of a parental set, and $pa(j)$, $nbd(\mathcal{P})$ as the neighborhood of the set of parental sets, $\mathcal{P}$. Then

$$nbd(\mathcal{P}) \triangleq \{\{Z_1, \cdots, Z_m\}, \text{ where }, Z_j \in nbd(pa(j)), j = 1, \cdots, m\}.$$

Define $nbd(pa(j))$ to be all subsets of $\{j+1 : m\}$ with one variable change to $pa(j)$. For the current MDM M, for each $j = 1, \cdots, m$, suppose $pa(j)$ has $p_{j\gamma}$ elements (i.e., having $p_{j\gamma}$ non-zero elements in the jth row of the upper triangular part of $m \times m$ matrix $\boldsymbol{Z}$). The neighborhood has three elements, i.e. $nbd(pa(j)) = \{pa(j)^{+}, pa(j)^{\circ}, pa(j)^{-}\}$, where $pa(j)^{+}$ is a set containing parental sets for variable j with $p_{j\gamma}+1$ variables, called the "addition" moves, and the newly added element has to belong to $\{j+1:m\}\backslash pa(j)$. The "replacement" moves are denoted by $pa(j)^{\circ}$, a set containing parental sets for variable j with $p_{j\gamma}$ variables, with the newly deleted element belonging to $pa(j)$, the newly added variable belonging to $\{j+1:m\}\backslash pa(j)$. Finally, the deletion moves are denoted by $pa(j)^{-}$, a set containing parental sets for variable j with $p_{j\gamma}-1$ variables, with the newly deleted element belonging to $pa(j)$.

When $p_{j\gamma}$ is not small, $|pa(j)^{\circ}|$ is much larger than $max(|pa(j)^{+}|, |pa(j)^{-}|)$, and thus the posterior probability of models with a parental set of series j in $pa(j)^{\circ}$ cumulate to larger values than that in $pa(j)^{+}$ or $pa(j)^{-}$. To move across dimension effectively, I break sampling into two parts: three models with parental sets $pa(j)^{+}_{*}, pa(j)^{\circ}_{*}$ and $pa(j)^{-}_{*}$ are sampled from $pa(j)^{+}, pa(j)^{\circ}$ and $pa(j)^{-}$ respectively, and then one of these three models is selected.

This neighborhood for parental sets, $nbd(\mathcal{P})$, is substantially different from that of graphical models in Jones et al. (2005), in which the neighborhood contains all the models with the graph of one-edge change to the current one. Using graphical notation, each MDM network can be represented a graph with m nodes and the edges between nodes i and k if $i \in pa(k)$ or $k \in pa(i)$. Our definition of neighborhood include graphs with $(m-1)$-edge changes to the current one, which allows the stochastic search to move faster to high probability regions of model space than that of Jones et al. (2005). Define $M(j)$ to be the model parameter for series j of the current model M, i.e. $(pa(j), \boldsymbol{S}_j)$. Now define the neighborhood of $M(j)$ as

$$nbd(M(j)) = \{(pa(j)_{*}, \boldsymbol{S}_{j*}) : pa(j)_{*} \in nbd(pa(j)), \boldsymbol{S}_{j*} \in \boldsymbol{\Psi}\}.$$

Corresponding to the three parts of $nbd(pa(j))$, $nbd(M(j))$ also consists of three parts, i.e. $nbd(M(j)) = \{M(j)^+, M(j)^\circ, M(j)^-\}$, where

$$\begin{aligned} M(j)^+ &= \{(pa(j)_*, \boldsymbol{S}_{j*}) : pa(j)_* \in pa(j)^+, \boldsymbol{S}_{j*} \in \boldsymbol{\Psi}\}, \\ M(j)^\circ &= \{(pa(j)_*, \boldsymbol{S}_{j*}) : pa(j)_* \in pa(j)^\circ, \boldsymbol{S}_{j*} \in \boldsymbol{\Psi}\}, \\ M(j)^- &= \{(pa(j)_*, \boldsymbol{S}_{j*}) : pa(j)_* \in pa(j)^-, \boldsymbol{S}_{j*} \in \boldsymbol{\Psi}\}. \end{aligned}$$

3.4 MDM Shotgun Stochastic Search Schema

The following algorithm outlines the new MDM SSS Schema, which is attractive because step 2, containing most of the computational burden, can be easily parallelized. Let $S(M(j))$ denote the corresponding (unnormalized) score for model $M(j)$. Given a starting model $M^{[0]}$, iterate in $i = 1, \cdots, I$ with the following steps:

1. For $j = 1, \cdots, m$, construct $nbd(M^{[i]}(j)) = \{M^{[i]}(j)^+, M^{[i]}(j)^\circ, M^{[i]}(j)^-\}$.

2. In parallel, compute the score, $S(X)$, for all $X \in nbd(M^{[i]}(j))$, for all $j = 1, \cdots, m$.

3. In parallel, sample $M(j)^+_*, M(j)^\circ_*$ and $M(j)^-_*$ from $M^{[i]}(j)^+, M^{[i]}(j)^\circ, M^{[i]}(j)^-$ respectively for each j, with probabilities proportional to the corresponding $S(X)^\eta$, normalized within each set, where η is an annealing parameter.

4. Propose $M(j)^{[i+1]}$ from $\{M(j)^+_*, M(j)^\circ_*, M(j)^-_*\}$ with probability proportional to $S(X)^\eta$, normalized within this set. Update the next model to be $M^{[i+1]} = \{M(j)^{[i+1]}, j = 1, \cdots, m\}$.

The score in the application section I used is proportional to the posterior probability of the model $M(j)$ for series j in formula (3.5): $P(M(j))p(\boldsymbol{y}_{j,1:t}|M(j))$. However, notice that the score can be any other measure that can quantify the probability of the model candidates for the next move.

In the following application, for each sampled model $M^{[i]}$, I compute the 1-step-ahead predictive means and variances of $\boldsymbol{y}_t$ under this model at each time point t according to Section 2.3.3. Based on these forecasts, Bayesian decisions on portfolio allocation are made, and quantities of practical interests, such as the cumulative returns and risks, are calculated.

3.5 Application to Financial Time Series

This applied forecasting and financial portfolio study aims to illustrate that sparsity on parental sets and flexibility in having variable-wise discount factors in MDMs can contribute to significant improvements in model fit, forecasting and Bayesian decisions in a practical context. The study focuses on one-day-ahead predictions and portfolio decisions.

3.5.1 Dynamic portfolio allocation

Bayesian forecasting models and Bayesian decision analysis in asset allocation problems have been well recognized and broadly applied for decades. I follow previous work by Quintana (1992), Aguilar and West (2000), and Quintana et al. (2003) in utilizing Bayesian decision theory via Markowitz portfolio optimization. Forecasting the mean and covariance of future returns is critical in mean-variance portfolio optimization methods. Dynamic models allow for parameter change and uncertainty formally taken into account in sequential investment decisions. A main goal here is to examine and illustrate the efficiency of MDMs in forecasting and portfolio decision making. To use a model performance measurement of practical relevance to investors, I focus on comparison of realized portfolio performance.

For portfolio decisions, at time $t-1$, given the historical information $\mathcal{D}_{t-1}$, I can obtain the mean and variance matrix $(\boldsymbol{f}_t, \boldsymbol{Q}_t)$ of the predictive distribution of the vector of next-period returns. Quintana et al. (2003) point out that, given $(\boldsymbol{f}_t, \boldsymbol{Q}_t)$, the investor's decision is simply to select the vector of portfolio weights $\boldsymbol{w}_t$ that maximizes his/her utility function subject to some constraints on $\boldsymbol{w}_t$. In this thesis this utility function is selected to be opposite of risk, such that I minimize the ex ante (at time t given $\mathcal{D}_{t-1}$) "risk", represented by the one-step-ahead portfolio variance $\boldsymbol{w}_t'\boldsymbol{Q}_t\boldsymbol{w}_t$, subject to the constraints $\boldsymbol{w}_t'\mathbf{1} = 1$ and some other constraints. Denote the precision matrix by $\boldsymbol{K}_t = \boldsymbol{Q}_t^{-1}$. Three commonly used allocation rules are as follows.

1. Target return, minimum risk: for some specified (at time $t-1$) target value r_t, optimize the portfolio weights by minimizing the ex ante portfolio variance among the restricted portfolios whose 1-step-ahead expectation is $\boldsymbol{w}_t'\boldsymbol{f}_t = r_t$. The investor decision problem reduces to choosing the vector of portfolio weights $\boldsymbol{w}_t$ to minimize $\boldsymbol{w}_t'\boldsymbol{Q}_t\boldsymbol{w}_t$, subject to constraints $\boldsymbol{w}_t'\boldsymbol{f}_t = r_t$ and $\boldsymbol{w}_t'\mathbf{1} = 1$. As summarized in Chapter 10 of Prado and West (2010), the solution for the weights vector in this case is $\boldsymbol{w}_t = \boldsymbol{K}_t(u_t\boldsymbol{f}_t + z_t\mathbf{1})$, where $u_t = \boldsymbol{h}_t'\boldsymbol{g}_t$, $\boldsymbol{h}_t = \boldsymbol{K}_t\mathbf{1}$, $z_t = -\boldsymbol{f}_t'\boldsymbol{K}_t\boldsymbol{g}_t$, $\boldsymbol{g}_t = (r_t\mathbf{1} - \boldsymbol{f}_t)/e_t$, and $e_t = \mathbf{1}'\boldsymbol{h}_t \times (\boldsymbol{f}_t'\boldsymbol{K}_t\boldsymbol{f}_t) - (\boldsymbol{f}_t'\boldsymbol{h}_t)^2$.

2. Target return, minimum risk with nonnegative constraint: given the target return, minimize the risk with nonnegative constraints added to the portfolio weights, i.e. at each time t, minimize portfolio variance $\boldsymbol{w}_t'\boldsymbol{Q}_t\boldsymbol{w}_t$, subject to constraints $\boldsymbol{w}_t(j) \geqslant 0, j \in \{1, \cdots, m\}$, $\boldsymbol{w}_t'\boldsymbol{f}_t = r_t$ and $\boldsymbol{w}_t'\mathbf{1} = 1$. No analytical form can be found in this non-linear optimization problem, so I use the quadratic programming package in MATLAB to solve the optimal weights for this rule.

3. Target margin, minimum risk with zero correlation with a benchmark asset: given the target margin r_t over a benchmark asset with s_t as the ex ante mean of one-step-ahead return, optimize the portfolio weights by minimizing the ex ante risk among the restricted portfolios whose ex ante correlation with a benchmark asset is zero. Let x_t, s_t, v_t denote the return at time t, ex ante

predictive return and ex ante predictive variance given $\boldsymbol{x}_{1:t-1}$ of the benchmark asset, and write the ex ante mean and covariance of $\boldsymbol{y}_t$ and x_t jointly as follows:

$$E\begin{pmatrix} \boldsymbol{y}_t \\ x_t \end{pmatrix} = \begin{pmatrix} \boldsymbol{f}_t \\ r_t + s_t \end{pmatrix}, \quad V\begin{pmatrix} \boldsymbol{y}_t \\ x_t \end{pmatrix} = \begin{pmatrix} \boldsymbol{Q}_t & \boldsymbol{q}_t \\ \boldsymbol{q}_t' & v_t \end{pmatrix},$$

where $\boldsymbol{q}_t = C(\boldsymbol{y}_t, x_t|\mathcal{D}_{t-1}, \boldsymbol{x}_{1:t-1})$. Then this optimization problem is reduced to solving $\boldsymbol{w}_t$, at each time t, that minimizes $\boldsymbol{w}_t'\boldsymbol{Q}_t\boldsymbol{w}_t$ subject to the constraints $C(\boldsymbol{w}_t'\boldsymbol{y}_t, x_t) = 0$, $\boldsymbol{w}_t'\boldsymbol{f}_t = r_t + s_t$ and $\boldsymbol{w}_t'\boldsymbol{1} = 1$. As derived by Zhou et al. (2014), the solution for the weights vector in this case is $\boldsymbol{w}_t = \boldsymbol{K}_t\boldsymbol{B}_t'(\boldsymbol{B}_t\boldsymbol{K}_t\boldsymbol{B}_t')^{-1}\boldsymbol{a}_t$, where $\boldsymbol{B}_t = [\boldsymbol{f}_t; \boldsymbol{q}_t; \boldsymbol{1}]'$, $\boldsymbol{a}_t = (r_t, 0, 1)'$.

The portfolio constructed under the rule of “target return, minimum risk” is based on modern portfolio theory introduced by Markowitz (1959), which I call “target portfolio” for convenience. It points out that investing is a tradeoff between risk and expected return, and quantifies the risk increased by the increase in expected return.

The rule of “target return, minimum risk with nonnegative constraint” is another common rule that constrains the strategy to be long-only, relevant to long-only funds or markets, or investors with short-averse preference. This type of portfolio, called here as “target constraint portfolio”, essentially bears larger minimum risk than that of the target portfolio. As exemplified below, it empirically realizes a larger cumulative return as a trade-off of the larger risk than that of the target portfolio.

The third rule is a commonly used decoupling strategy, a risk diversification tool with respect to the risk in the benchmark asset in portfolio management. The benchmark can be any asset, such as a stock index, a commodity, or foreign exchange futures. I choose the benchmark asset to be the S&P 500 index, since profitable investment strategies that are zero-correlated with this major market index are of broad, practical interest. This is especially true during the crises such as that experienced over late 2007-2009.

Due to their determining effects on conditional dependence structure, covariance matrix elements can play key roles in portfolio decisions. As pointed out by Carvalho and West (2007), in higher-dimensional portfolios the optimal weights can be very volatile over time, induced by the uncertainty in the estimation of covariance matrices. Selecting a sparse compositional network is crucial for a MDM to avoid unnecessary parameters, collinearity issues, and estimation instability of covariance matrices. Aiming to decrease the variance of $\boldsymbol{w}_t$, I apply large-scale SSS over parental sets, which essentially attempts to capture accurate conditional dependence structures among assets. If the compositional network should be much more sparse than what is actually specified, then unnecessary parameters would be estimated, and collinearity issues could arise and cause instability of covariance matrices. The following example explicitly illustrates the benefits of sparsity in terms of decreasing portfolio risk and improving profitability, compared with the "full" model, whose parental sets $pa(j)$ are all equal to $\{j+1:m\}$, for $j=1,\cdots,m$.

3.5.2 Data

This study analyzes $m = 13$ financial time series, which are daily closing prices of 9 exchange rates, 2 commodities, and 2 U.S. stock indices in dollars from April 2006 through December 2011, consisting of 1,489 days in total. Table 3.1 shows the series and corresponding abbreviations. The data cover the major liquid benchmark securities across three asset classes to approximately reflect the global macroeconomic environment. To scale each asset to a similar level, daily returns instead of daily prices are analyzed. The daily return for asset j at time t is computed as $y_{jt} = 100(P_{jt}/P_{j,t-1} - 1)$, where P_{jt} is the price of asset j at time point t. Figure 3.1 shows the daily returns of the 13 time series over the given period. I perform one-step-ahead forecasting and Bayesian decision-making for the whole selected period.

Table 3.1: Asset price series

Abbreviation	Currency/Commodity/Index
CHF	Swiss Franc
EUR	Euro
NSD	NASDAQ Composite Index
S&P	S&P 500 Index
NOK	Norwegian Krone
GBP	British Pound
AUD	Australian Dollar
NZD	New Zealand Dollar
ZAR	South African Rand
GOL	Gold
CAD	Canadian Dollar
JPY	Japanese Yen
OIL	Crude Oil

3.5.3 Parameter set

I analyze the data using the new SSS under the MDM model framework as specified in (3.1). I consider every combination of the parameters in the first two rows of Table 3.2, a total of $K = 25$ combinations as the value set $\boldsymbol{\Psi}$, from which discount factors $\boldsymbol{S}_j$ for each variable j are chosen, recalling that δ discounts regressions and β discounts volatility. To bear practical relevance to the real market, the range of δ and β is chosen to be 0.975 to 0.995; see Xie (2013). I believe that δ is strictly smaller than 1 since the impact of historical information should be discounted instead of being magnified or constant when updating the current posterior distribution of the variance parameter processes.

The order I take for the series is shown in Table 3.3. The daily target return is chosen to be 0.30%, corresponding to a (20-day) monthly target return of approximately 6%.

For all models of the form (3.1), I use shrinkage initial priors at $t = 0$:

$$(\boldsymbol{\theta}_{j0}, \lambda_{j0}|\mathcal{D}_0) \sim NIG(\boldsymbol{m}_{j0}, \boldsymbol{C}_{j0}, n_{j0}/2, d_{j0}/2),$$

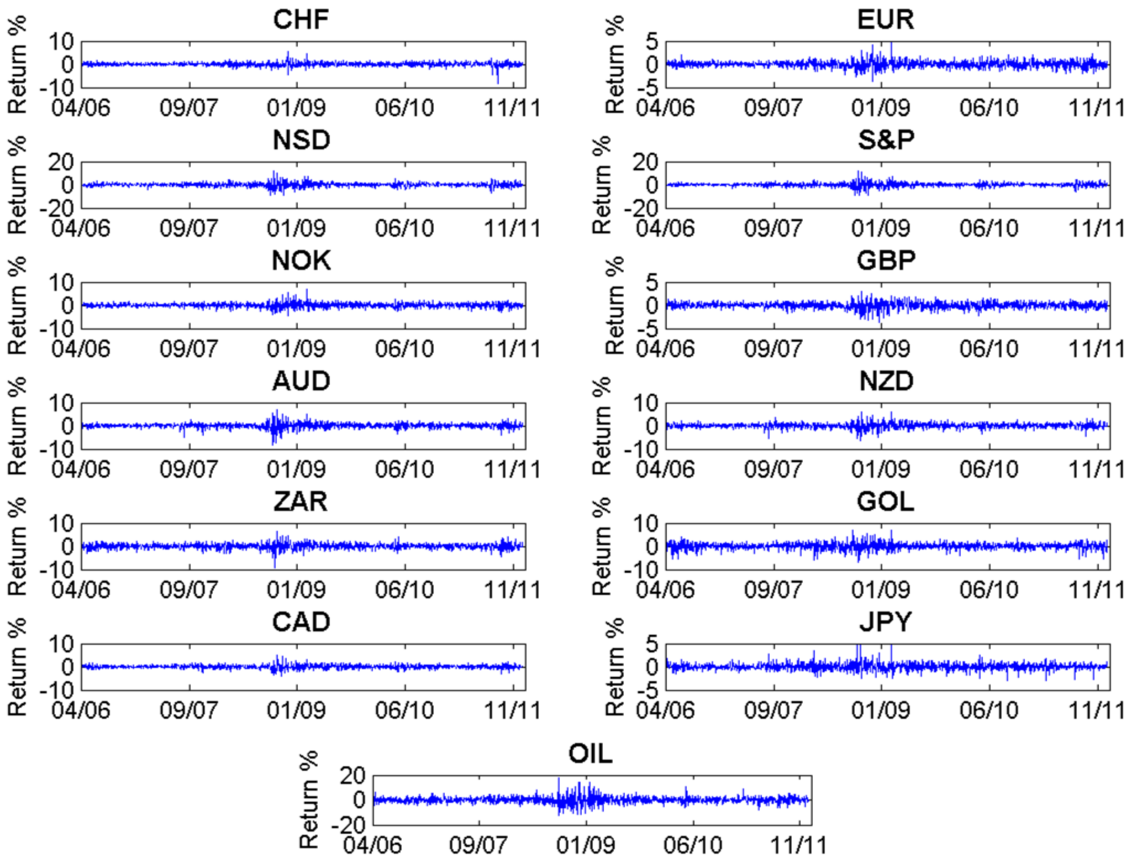


FIGURE 3.1: Daily returns data of the 13 time series. The data set consists of 1,489 daily returns for each of these 13 assets, over the period of 04/14/2006 to 12/30/2011

where $\boldsymbol{m}_{jt} = (0, \cdots, 0), \boldsymbol{C}_{jt} = 0.1\boldsymbol{I}$, and $n_{j0} = m$, for $j = 1, \cdots, m$. The prior residual variance estimate is $s_{j0} = d_{j0}/n_{j0} = 0.1$, and thus $d_{j0} = 0.1m$.

3.6 Results

3.6.1 Model search

The SSS algorithm for the MDM models is run for 10,000 iterations with the starting parental set for each series as the empty set. At each iteration for each series, all the possible models in the neighborhood of the current model are evaluated by unnormalized posterior probabilities. The computing time mainly depends on the speed of the calculation of posterior probabilities of univariate DLMs in the neighborhood.

Table 3.2: List of parameter values used for SSS

Parameters	Values
δ	0.975, 0.980, 0.985, 0.990, 0.995
β	0.975, 0.980, 0.985, 0.990, 0.995
ρ	0.1
η	1
Target Return	0.0030

In fact, at each iteration i, given the current model $M^{[i]} = (\mathcal{P}^{[i]}, S^{[i]})$, the number of univariate DLMs in the neighborhood is

$$K \times \sum_{j=1}^{m} [m + p_{j\gamma}(m - p_{j\gamma})] = K \times (\frac{m^3}{4} + m^2 - \sum_{j=1}^{m} (p_{j\gamma} - \frac{m}{2})^2),$$

where the minimum is $K \times m^2 = 25 \times 13^2 = 4,225$ when $p_{j\gamma} = 0$ for all j and the maximum is 15,075. In general, this number is of the order of $O(m^3)$.

Given the specified order of variables, the total number of possible univariate DLMs for all models is $K \times (2^m - 1)$. Thus, when m is relatively small, I can calculate all of them in parallel and save them in advance. Then when implementing SSS, at each iteration, I can directly extract the corresponding posterior probability of each model in the neighborhood from the saved values. However, when m increases to extraordinarily large, since the number of univariate series needed to be calculated at each iteration is $O(m^3)$, if the number of iteration is also on a smaller scale of m, I should not calculate all the univariate DLMs in advance. In the current example with $m = 13$, I calculated all the 204,775 possible univariate DLMs in parallel in advance.

After running the SSS algorithm with 276 iterations, I obtain the "optimal" model in terms of cumulative return for the target constraint portfolio, which is as high as 234.0% over the given period; after 900 iterations, I find the "optimal" model in terms of cumulative return for both the target portfolio and the benchmark neutral

Table 3.3: Variable-wise discount factors for the optimal model for the target constraint portfolio

Asset	CHF	EUR	NSD	S&P	NOK	GBP	AUD
δ	0.9800	0.9800	0.9900	0.9950	0.9800	0.9850	0.9950
β	0.9750	0.9800	0.9750	0.9750	0.9800	0.9750	0.9750
Asset	NZD	ZAR	GOL	CAD	JPY	OIL	
δ	0.9900	0.9950	0.9850	0.9850	0.9950	0.9900	
β	0.9750	0.9750	0.9750	0.9750	0.9750	0.9800	

Table 3.4: Variable-wise discount factors for the optimal model of target portfolio and benchmark neutral portfolio

Asset	CHF	EUR	NSD	S&P	NOK	GBP	AUD
δ	0.9800	0.9750	0.9850	0.9950	0.9800	0.9850	0.9950
β	0.9750	0.9750	0.9750	0.9750	0.9800	0.9750	0.9750
Asset	NZD	ZAR	GOL	CAD	JPY	OIL	
δ	0.9900	0.9900	0.9850	0.9950	0.9950	0.9950	
β	0.9750	0.9750	0.9750	0.9750	0.9750	0.9750	

portfolio, which has cumulative returns of 151.1% and 149.0% over the given period. The high speed of locating the most lucrative model illustrates the efficiency of SSS. The parental sets and discount factors for each series under the "optimal" model for the target constraint portfolio is shown in Figure 3.2, Table 3.3 and Table 3.4.

To test the robustness of the new SSS algorithm, I run 2,000 iterations of SSS 6 times, each time beginning from different starting points, i.e. parental sets, and examine the efficiency of the algorithm in finding profitable models. The starting parental sets chosen are $\mathcal{P}_k, k = 0,1,2,3,4,5$ such that for $k = 0,1,2,3,4$, $\mathcal{P}_k = \{pa(j) : p_{\gamma j} = k \text{ if } j = 1,\cdots,m-k; pa(j) = j+1:q, \text{ if } j = m-k+1,\cdots,m\}$, where the k elements in the corresponding $pa(j)$ are uniformly drawn from $j+1:m$; the last set of parental sets is the full set, i.e. $\mathcal{P}_5 = \{pa(j) = j+1:q, j = 1,\cdots,m\}$. The results suggest that for 6 runs, SSS finds the "optimal" models under the 3 portfolio decision rules within the first 1,000 iterations; the "optimal" models found

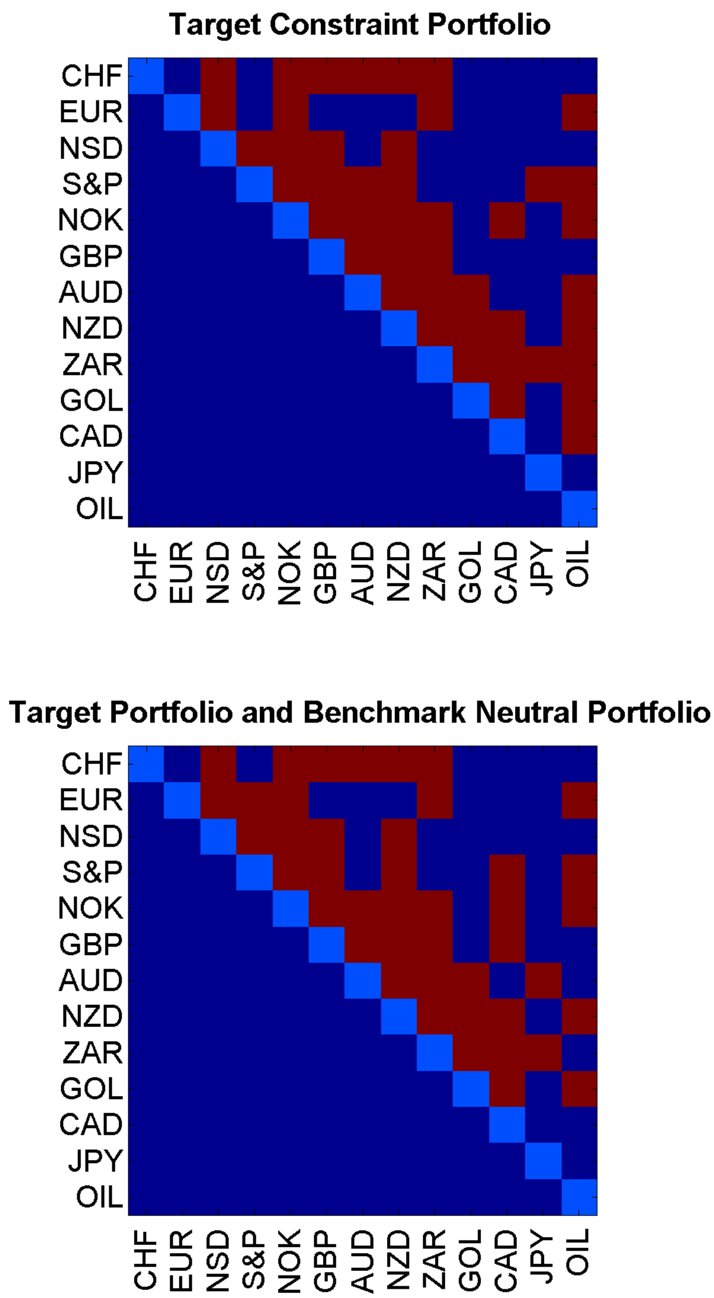


FIGURE 3.2: Parental sets of the optimal model

realize similarly high cumulative return over the given period– all above 200% for the target portfolio, around 150% for the target portfolio and benchmark neutral portfolio. These results exhibit the robustness of the new SSS in finding good models with regard to different starting points in the high-dimensional model space.

3.6.2 Model comparison on parental set sparsity

This section aims to illustrate the advantage of sparsity of parental sets in terms of ratios of cumulative return and risk. For each type of portfolio strategy, I compare two models: (1) the optimal model found by SSS, which has a sparse compositional network, and (2) the model of full parental sets, i.e., $pa(j) = j + 1 : m$ for all $j = 1, \cdots, m$. To control the impact of other parameters, I use the same state and observation discount factors for the latter with the former. A couple of important quantities that can reflect the impact on predictions and portfolio decisions are examined in detail in the model comparison by the following two measures.

1. The accuracy of forecasts of asset returns and resulting portfolio decisions, in terms of realized cumulative returns.

2. The projected risk, i.e. the forecast variance of the portfolio return.

I assume that there are no transaction costs for portfolio rebalancing, or that such transaction costs are negligible in terms of influence on the computation of cumulative returns. The comparisons of cumulative return between the portfolios computed by model (1) and (2) for each type of portfolio strategy are shown in Figure 3.3. As can be clearly seen, the model with sparse structure outperforms the model without sparsity in terms of cumulative returns thoughout the given period.

Figure 3.4 demonstrates the ex-ante portfolio risk ratios of model (2) relative to model (1) for each type of portfolio. These ratios are larger than 1 for 68.1%, 56.0% and 54.6% of the sample period, respectively for the target constraint, target and

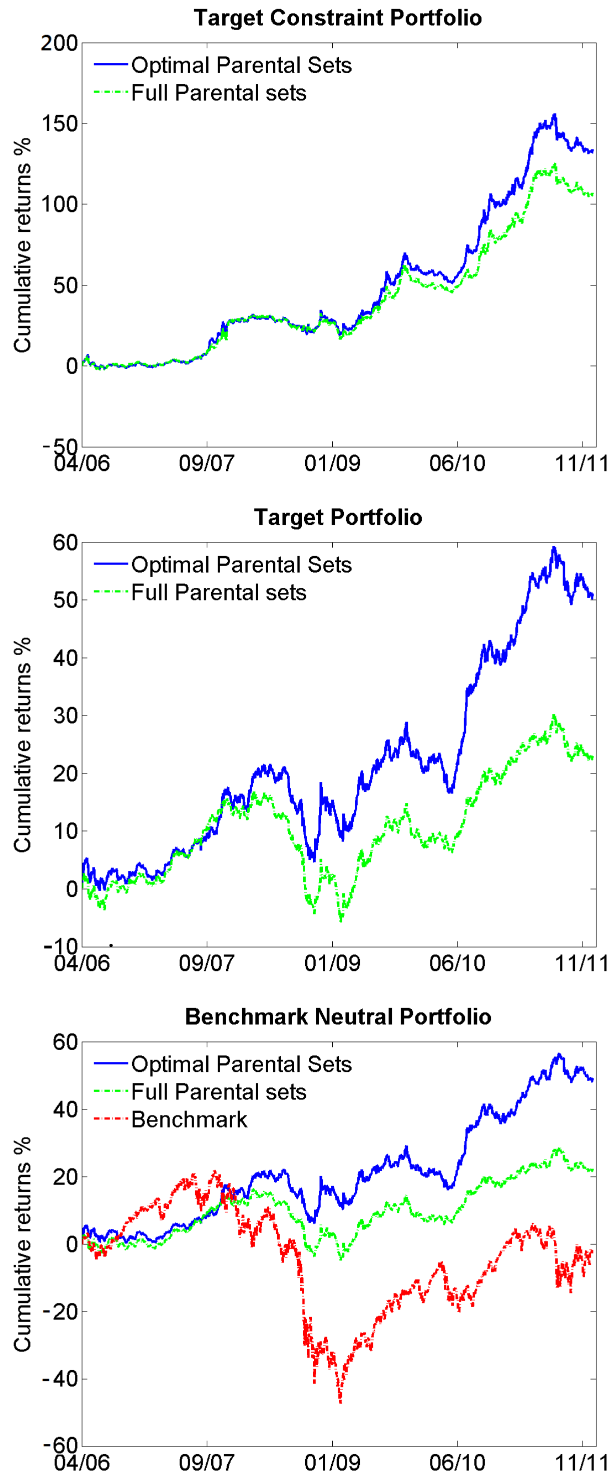


FIGURE 3.3: Portfolio cumulative return of the optimal MDM parental sets and full parental sets

benchmark portfolio. Thus the results suggest that sparse compositional network indeed leads to lower risks for each of the 3 types of portfolios.

This example demonstrates the importance of appropriate compositional network structure and the fact that the resulting parameter parsimony can indeed contribute to portfolio forecasts and resulting decisions by improving returns and decreasing risks.

3.6.3 Variable-wise discount factors

Evidence in support of allowing discount factors to vary across series is shown in detail. For the target constraint portfolio, I set the parental sets $\mathcal{P}$ to be the same as in model (1) in Section 3.6.2. Then if the discount factors (δ_j, β_j) are constrained to be the same across series, i.e.

$$(\delta_1, \beta_1) = (\delta_2, \beta_2) = \cdots = (\delta_m, \beta_m),$$

there are 25 combinations of (δ_j, β_j) and thus 25 models under which I forecast the returns and make portfolio decisions. For the target constraint portfolio, the highest cumulative return is 169.8% out of these 25 models, which is significantly lower than the "optimal" portfolio with variable-wise discount factors by 37%. Similar results can also be found for target portfolio and benchmark neutral portfolio, and thus are omitted here.

More justification for allowing discount factors to be series-specific can be found in the scatter plots in Figure 3.5 and Figure 3.6. The two horizontal axes are discount factors δ_j and β_j, and the vertical axis is log posterior probability, i.e. $P(\delta_j, \beta_j, pa(j)|\boldsymbol{y}_{j,1:T})$. I add a tiny disturbance, a zero-mean normal noise of standard deviation 0.00001, to each dimension of each point in the figure, so that the density of points can be viewed as roughly proportional to the marginal posterior probability of discount factors (δ_j, β_j) within the space of the 10,000 models visited

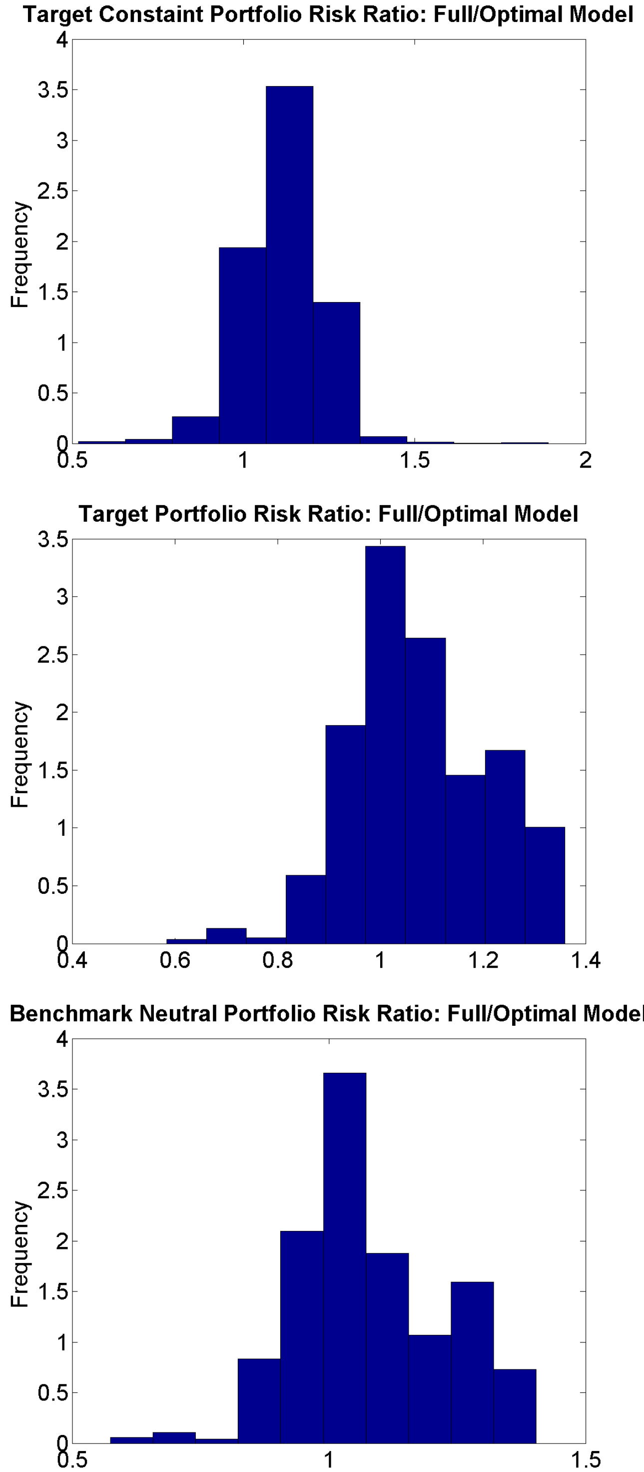


FIGURE 3.4: Ex ante (at time t given $\mathcal{D}_{t-1}$) ratio of portfolio risk under the full model relative to the "optimal" model of each type of portfolio. The discount factors for each series under the two portfolios are identical

by the SSS. It is obviously seen that the "optimal" combinations of δ_j and β_j– in terms of both joint and marginal posterior probability– differ across assets, which indicates support for the flexible, series-specific structure of discount factors under the MDM framework.

3.6.4 Probability and return

Figure 3.7 shows the average risk, i.e. the ex ante standard deviation of the portfolio's return through the given period. Apart from the three axes, the fourth dimension is the color of each point, which shows the sparsity of each model. As can be seen in the figure, some models are of high posterior probabilities, high returns and low risks; some models are of high posterior probabilities, yet low returns and high risks; some models are of low posterior probabilities, yet high returns and low risks. These figures clearly illustrate the fact that higher posterior probability does not necessarily lead to higher investment performance. From the perspective of a practical investor, the main interest is focused on the latter instead of the former. This practical perspective suggests a new perspective on model parameters, differing from the traditional probability-oriented approach. That is, the model parameters can be treated as decision variables and be selected by Bayesian decision theory based on practical utility functions.

3.7 Summary: Discussion and Future Research

By decomposing a multivariate DLM into sets of univariate DLMs that can be analyzed in parallel, I find that MDMs, a rich class of multivariate DLMs, allow for sparse conditional dependence network, and flexibility of discounting historical information in updating stochastic volatility. The compositional network represented by parental sets provides a natural framework under which I develop not only sequential updating, on-line forecasts and decisions that broadens the current theory

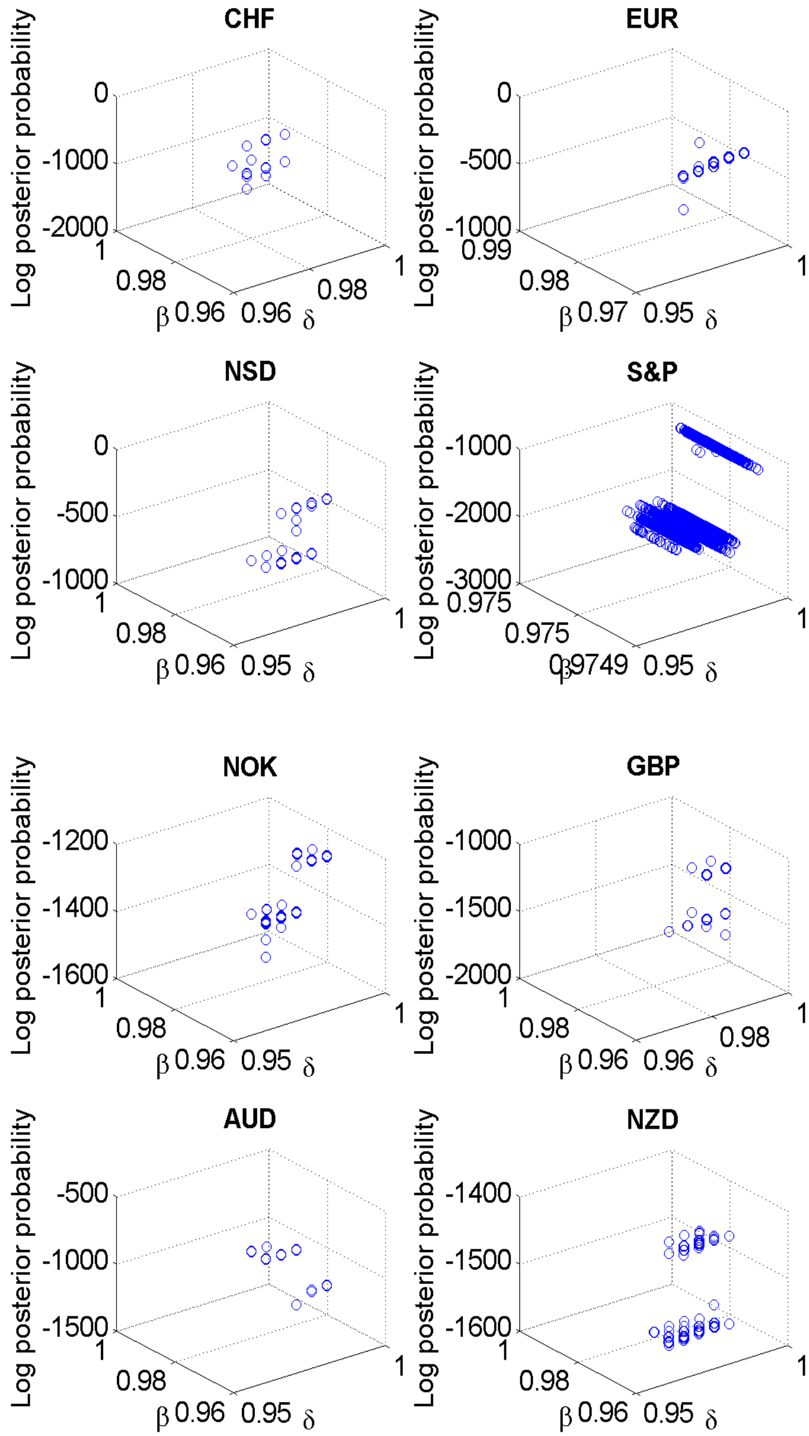


FIGURE 3.5: Scatter plots of the discount factors (δ, β), and their log posterior probability of CHF, EUR, NSD, S&P, NOK, GBP, AUD, NZD

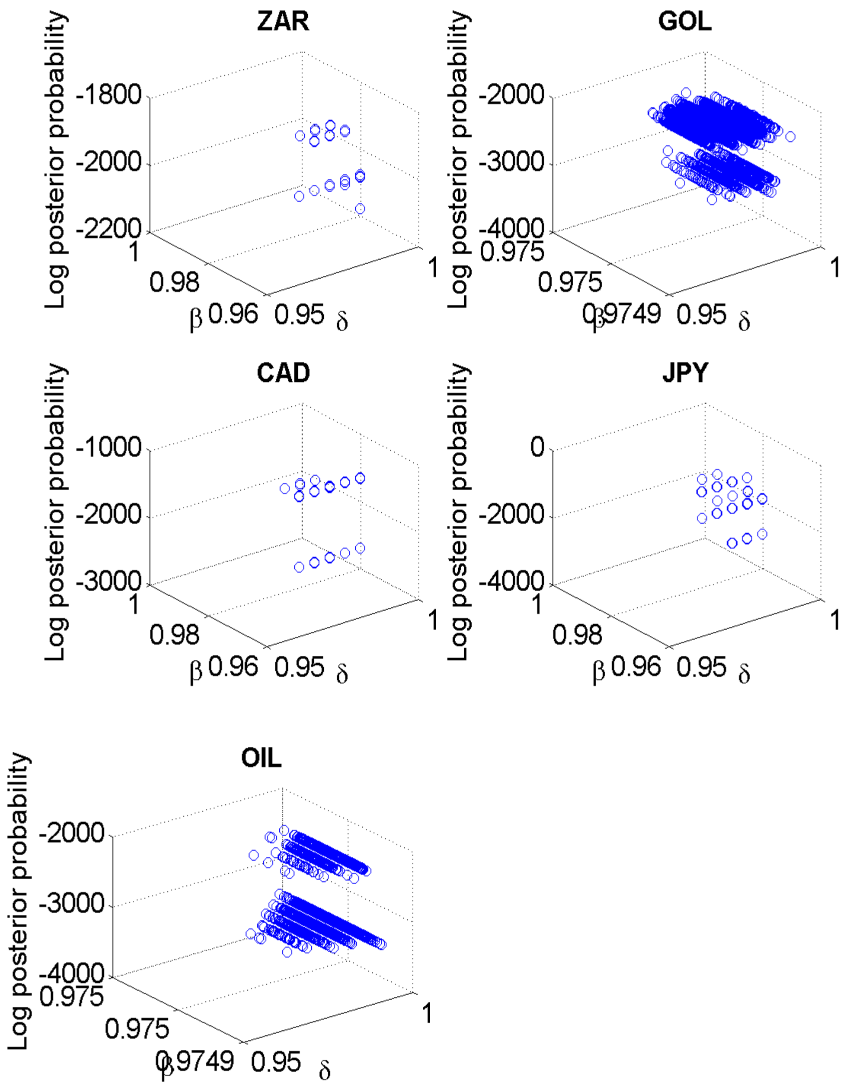


FIGURE 3.6: Scatter plots of the discount factors (δ, β), and their log posterior probability of ZAR, GOL, CAD, JPY, OIL

and methods of Bayesian dynamic models, but also parallel algorithms to enable the previously unfeasible computation. The local trend model utilized in my application can be easily extended to time-varying vector autoregressive (TV-VAR) models, as noted in the following chapter.

I have shown that SSS can succeed in not only exploring large MDM space, but

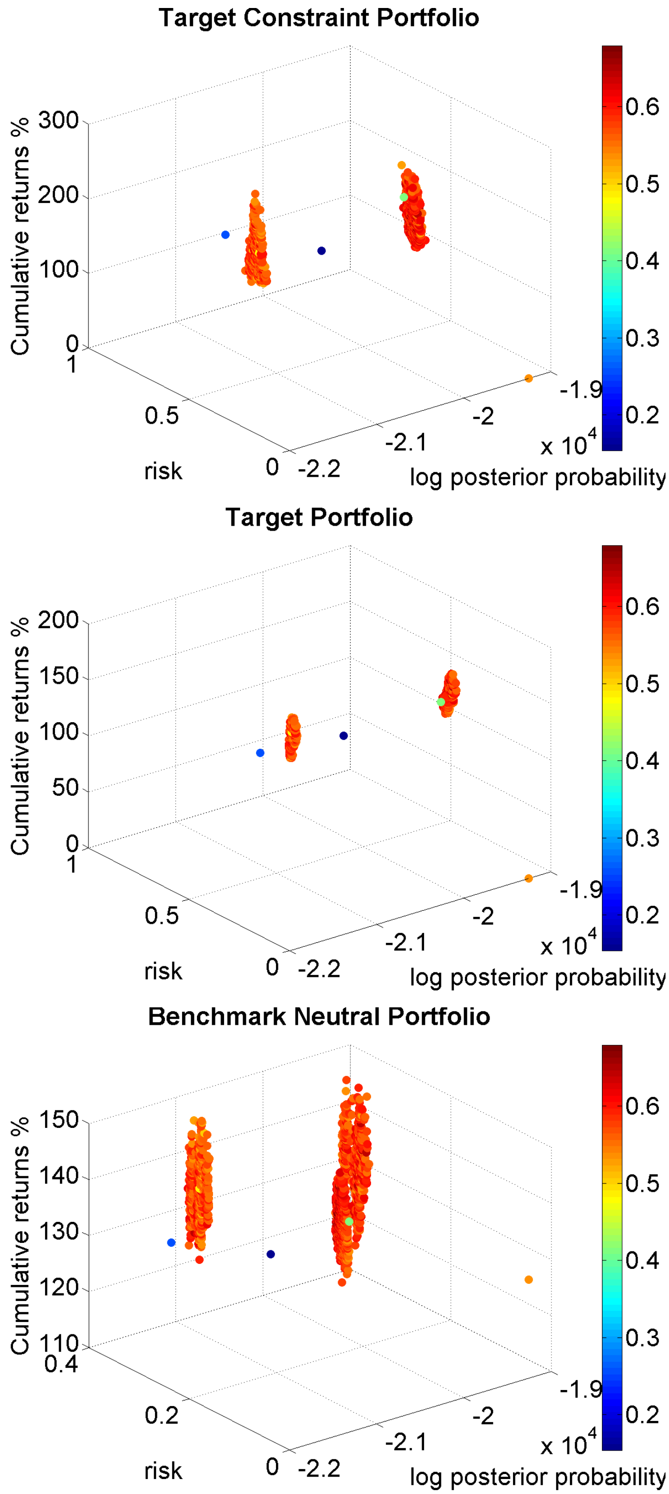


FIGURE 3.7: Scatterplot of the 10,000 models visited by SSS

also in locating practically lucrative models in financial application. Moreover, under the parallelizable framework, the new SSS algorithm allows the stochastic search to move in each dimension simultaneously at each iteration, and thus moves faster to high probability model space than does the original SSS of Jones et al. (2005).

While this chapter develops the framework of MDMs, several research questions are motivated and remain open for future research. One indication from the above analysis is that fitting MDMs using local move methods is feasible for higher-dimensional time series data, at least up to a few hundreds of series, since the number of univariate DLMs in the neighborhood is $O(m^3)$. When m increases further, parallel computation by GPU can still enable computation for SSS at each move, which mainly involves the calculation of $O(m^3)$ univariate models in the neighborhood. Hence, to extend the method to higher dimensional series, future work might include the GPU implementation for analysis of large numbers of univariate DLMs in parallel.

The results of the application indicate that higher posterior probability models do not necessarily lead to higher investment performance. Therefore, further research might include treating the discount factors as decision variables and using Bayesian decision theory to select their values. The modelling context of multi-process, class I (West and Harrison, 1997) provides a natural framework to start the exploration in this area.

The above results illustrate significant improvements in forecasts and decisions in MDMs with particular model parameters. Another step is to construct robustly lucrative strategies that can choose appropriate model parameters adapting to potential regime switch and market shocks. Under various possible multi-process modeling contexts, new interesting theory and empirical analysis in exploring model mixing and averaging of MDM models might be developed, especially on sequential model selection procedures that address uncertainty about parental sets, discount factors as well as the ordering of variables. I discuss the model averaging of MDMs with detailed development and applications in the next chapter.

4

Extended Bayesian Model Averaging

4.1 Introduction

While Chapter 3 has illustrated the effectiveness of some particular MDMs in inference and forecasting of multivariate time series, another step further is to address online model structure search and uncertainty analysis. BMA serves as a common method to deal with model uncertainty and is well-known for superior performance in out-of-sample predictions, which can provide numerous practical benefits.

As discussed by West and Harrison (1997), issues of model uncertainty in the context of DLMs can be addressed by two types of multi-process models. Let $\boldsymbol{\eta}$ be any defining parameters of the model which are possibly subject to uncertainty and $\mathcal{A}$ be the set of all possible values for $\boldsymbol{\eta}$. The model with parameter values of $\boldsymbol{\eta}_0$ at time t is $M_t(\boldsymbol{\eta}_0)$. The two types of multi-process contexts are defined as follows:

I For some $\boldsymbol{\eta}_0 \in \mathcal{A}$, $M_t(\boldsymbol{\eta}_0)$ holds for all t.

II For some sequence of values $\boldsymbol{\eta}_t \in \mathcal{A}, (t = 1, 2, \cdots)$, $M_t(\boldsymbol{\eta}_t)$ holds at time t.

In the multi-process, class I modeling context, I develop new Bayesian analytics for forecasting with model mixing and averaging of MDMs. I also utilize power discounting techniques for adaptively processing new information, which effectively shortens the average span of relevance historical information.

BMA on MDMs is customized to take advantage of the decomposable feature of MDMs. It essentially transforms the problem of model averaging from multivariate to univariate contexts. Specifically, for each series, I consider the mixture of multiple candidate univariate DLMs whose posterior probabilities over time are sequentially updated in parallel. I also naturally design parallel algorithms across all series to update univariate DLMs so that tractable computation of online learning of model parameters is achieved.

Section 4.2 introduces priors and likelihoods for model parameters and reviews the model framework of MDMs. Section 4.3 introduces an extended BMA method and outlines Bayesian computation for the first two $1/k$-step-ahead predictive moments. Finally, an application is detailed using the MDM with extended model averaging, in Section 4.4. In Section 4.5, I discuss forecasting and portfolio results and illustrate the practical efficacy of MDMs in generating profitable strategies in a financial time series sequential investment decision context. Section 4.6 provides summary comments and indications of potential future research.

4.2 Priors and Likelihoods for Model Parameters

I specify notation as follows. Let $\mathcal{L}$ be the autoregressive lag order set of a MDM, i.e. $\mathcal{L} \triangleq \{p_{j\lambda}, j = 1, \cdots, m\}$, and assume that l is the maximum possible autoregressive lag order, i.e. $p_{j\lambda} \in L \triangleq \{1, \cdots, l\}$, for all $j = 1, \cdots, m$. Define $M(j)$ to be the submodel for series j under M, i.e. $(pa(j), \boldsymbol{S}_j, p_{j\lambda})$. From the model specification of (2.9), $(\boldsymbol{S}, \mathcal{L}, \mathcal{P})$ defines a multiregression dynamic VAR model, denoted by $M \triangleq (\boldsymbol{S}, \mathcal{L}, \mathcal{P})$.

4.2.1 Priors

In this subsection I specify the prior distributions for the three main model parameters, AR lag order, parental sets, state discount factors and observation discount factors. Due to the interest in sparse models, I use the same prior on the parental sets $pa(j)$ as specified in Section 3.2.1, i.e. Bernoulli prior. Assume identically independent priors for each $\boldsymbol{S}_j$ with discrete uniform distribution on $\boldsymbol{\Psi}$ and $p_{j\lambda}$ on L, and that $\boldsymbol{S}$, $\mathcal{P}$ and $\mathcal{L}$ are mutually independent.

4.2.2 Likelihoods and posteriors

Notice that the compositional network structure of MDM implies that

$$p(y_{jt}|\mathcal{D}_{t-1}) = p(y_{jt}|\mathcal{D}_{t-1}(j)). \tag{4.1}$$

Thus at each time t, the likelihood of M can be written as the product of likelihoods of the $M(j)$, namely

$$p(\boldsymbol{y}_{1:t}|\mathcal{P},\boldsymbol{S},\mathcal{L}) = \prod_{j=1}^{m} p(y_{j,1:t}|pa(j),\boldsymbol{S}_j,p_{j\lambda}) = \prod_{j=1}^{m}\prod_{\tau=1}^{T} p(y_{j\tau}|\boldsymbol{y}_{pa(j),\tau},\boldsymbol{S}_j,p_{j\lambda},\mathcal{D}_{\tau-1}(j))$$

where $p(y_{j\tau}|\boldsymbol{y}_{pa(j),\tau},\boldsymbol{S}_j,p_{j\lambda},\mathcal{D}_{\tau-1}(j))$ is calculated via equation (2.6). Thus, the posterior probability of a model M is proportional to

$$P(\boldsymbol{S})P(\mathcal{L})p(\boldsymbol{y}_{1:t}|\mathcal{P},\boldsymbol{S},\mathcal{L})\prod_{j=1}^{m}\rho^{p_{j\gamma}}(1-\rho)^{m-j-p_{j\gamma}},$$

which can be decoupled as the product of the posteriors of each $M(j)$, i.e.

$$P(M|\mathcal{D}_t) \propto \prod_{j=1}^{m} P(M(j)|y_{j,1:t}) = \prod_{j=1}^{m} P(\boldsymbol{S}_j,p_{j\lambda},pa(j)|y_{j,1:t}),$$

where

$$P(\boldsymbol{S}_j,p_{j\lambda},pa(j)|y_{j,1:t}) \propto p(y_{j,1:t}|pa(j),\boldsymbol{S}_j,p_{j\lambda})\rho^{p_{j\gamma}}(1-\rho)^{m-j-p_{j\gamma}}P(\boldsymbol{S}_j).$$

This decomposition indicates that BMA on M can be decomposed by recursively averaging over $M(j)$ conditional on the averaged results for $M(k)$, $k = j+1, \cdots, m$. That is, for each $j = m, m-1, \cdots, 1$ in turn, I can apply model averaging with respect to $M(j)$ and then combine the averaged results, as discussed in the following section in detail.

4.3 Univariate Model Averaging for Multivariate Time Series

Due to the compositional and decomposable structure of MDMs, a m-dimensional multivariate MDM is defined by the m univariate DLMs. For each series $y_{j,1:T}$, I consider the case where multiple models are considered and there is uncertainty about models. That is, each time series follows a multi-process, class I model, where a single DLM with a certain set of parameters holds at all time points, but uncertainty of which model is "true " exists. According to West and Harrison (1997), the basic averaging theory underlying mixture models is used to calculate time-varying posterior probabilities of each candidate model, which are the weights assigned to different parameter values. The weighted average parameter values across models reflect the changes in model uncertainty over time. I can also assess how sensitive model forecasts are to each of the parameters.

The space for M is high dimensional with even a small m, since it increases exponentially with the increase of m. The number of possible models is the product of the numbers of possible combination of volatility discount factors, autoregressive lag orders, and parental sets, which is $m^{K+l} \times 2^{m(m-1)/2}$. This means when m is more than a small number, enumerating all possible models is impossible. I propose an innovative method to address this by transforming the model uncertainties from multivariate to univariate contexts. This method is natural due to the decomposable structure of MDMs and reduces the computation of model averaging to analyzing $\sum_{j=1}^{m} K \times l \times 2^{m-j} = Kl(2^m - 1)$ univariate DLMs and then combining the results of the m series.

Starting with an initial prior density $p(M|\mathcal{D}_0)$ for model M, information is sequentially processed to provide inferences about model M via the posterior $P(M|\mathcal{D}_t)$ at time t, using Bayes' theorem,

$$P(M|\mathcal{D}_t) \propto p(\boldsymbol{y}_t|M, \mathcal{D}_{t-1})P(M|\mathcal{D}_{t-1}). \tag{4.2}$$

For each series j, let $\boldsymbol{M}_j = \{M_{ji}, i = 1 : m_j\}$ denote the union set of all possible parameter values of $M(j)$, which has m_j elements. At time t, let $p_{jt}(i) = P(M(j) = M_{ji}|\mathcal{D}_t(j))$, and $\boldsymbol{X}_t$ be any vector of random quantities of interest. I thus have the following posterior probabilities for any set of parameter values M_{ji} on the univariate dimensions:

$$p_{jt}(i) \propto p(y_{jt}|M_{ji}, \boldsymbol{y}_{j+1:m,t}, \mathcal{D}_{t-1}(j))p_{j,t-1}(i),$$

for all $i = 1 : m_j$ and $j = 1 : m$. To make inference about $\boldsymbol{X}_t$ within model j but without reference to any particular value of M_{ji}, the required marginal posterior densities are now simply

$$p(\boldsymbol{X}_t|\mathcal{D}_t(j)) = \sum_{i=1}^{m_j} p(\boldsymbol{X}_t|M_{ji}, \mathcal{D}_t(j))p_{jt}(i). \tag{4.3}$$

4.3.1 Extended BMA by power discounting

Following West and Harrison (1989, p. 445) and Xie (2013), I introduce model probability discounting for adaptively processing new information. The model posterior probabilities at time t contains all the historical information from time 1 to $t-1$, but the goal is to place more emphasis on more recent data, since more recent information usually has more relevance to making inferences on current and future quantities. Probability discounting is essentially, at any time t, updating the prior probabilities of any model by raising the power of the posterior probability of this model at $t-1$ to a certain value strictly smaller than 1. This power, denoted by

α, reflects the degree of information discount per unit of time. For example, the daily discount of an α equal to 0.9857 would lead to a quarter of market information discount of the information from a month ago, since 0.9857^{20} is 75%, where I assume one month comprises 20 business days.

With model probability discounting, the model posterior probability is sequentially updated as

$$p_{jt}(i) \propto p(y_{jt}|M_{ji}, \boldsymbol{y}_{j+1:m,t}, \mathcal{D}_{t-1}(j))p_{j,t-1}(i)^{\alpha}.$$

Hence the log ratio of posterior probabilities between two models, M_{ji}, M_{jk}, is:

$$\begin{aligned} log\frac{p_{jt}(i)}{p_{jt}(k)} &= \alpha log\frac{p_{j,t-1}(i)}{p_{j,t-1}(k)} + log\frac{P(y_{jt}|M_{ji}, \boldsymbol{y}_{j+1:m,t}, \mathcal{D}_{t-1}(j))}{P(y_{jt}|M_{jk},, \boldsymbol{y}_{j+1:m,t}\mathcal{D}_{t-1}(j))} \\ &= \sum_{s=1}^{t} \alpha^{t-s} log\frac{P(y_{js}|M_{ji}, \boldsymbol{y}_{j+1:m,t}, \mathcal{D}_{s-1}(j))}{P(y_{js}|M_{jk}, \boldsymbol{y}_{j+1:m,t}, \mathcal{D}_{s-1}(j))}. \end{aligned}$$

The joint predictive density under α is

$$\begin{aligned} p(\boldsymbol{y}_t|\alpha, \mathcal{D}_{t-1}) &= \prod_{j=1}^{m} p(y_{jt}|\boldsymbol{y}_{j+1:m,t}, \alpha, \mathcal{D}_{t-1}(j)) \\ &= \prod_{j=1}^{m} \sum_{i=1}^{m_j} p(y_{jt}|M_{ji}, \boldsymbol{y}_{j+1:m,t}, \mathcal{D}_{t-1}(j))p_{j,t-1}(i)^{\alpha} / \sum_{i=1}^{m_j} p_{j,t-1}(i)^{\alpha}. \end{aligned}$$

Starting with an initial prior density $p(\alpha|\mathcal{D}_0)$ for α, information is sequentially processed to provide inferences about α via the posterior $p(\alpha|\mathcal{D}_t)$ at time t,

$$p(\alpha|\mathcal{D}_t) \propto p(\alpha|\mathcal{D}_{t-1})p(\boldsymbol{y}_t|\alpha, \mathcal{D}_{t-1}).$$

In the next section I develop analytic forms for 1-step-ahead joint predictive mean, covariance and precision matrix for MDMs following multi-process, class I. However, analytic forms for k-step-ahead ($k > 1$) predictive moments can be similarly deduced like 1-step-ahead forecasts but are mathematically complicated. In the application example, simulation method is used to calculate the k-step-ahead moments of y_{jt} and hence $\boldsymbol{y}_t$.

4.3.2 1-step-ahead joint predictive mean, variance matrix

I develop an innovative recursive method to calculate model averaging results for the 1-step-ahead joint predictive mean, covariance matrix and precision matrix. Following Section 4.3, one key point to highlight here is that this method exploits the compositional structure of a MDM by translating model averaging from multivariate to univariate contexts. From (4.3) it follows that all posterior distributions for 1-step-ahead predictive distributions are calculated as discrete mixtures. Some basic inference about such mixtures is as follows. I first give notation for conditional expectations and variances as

$$\begin{aligned} \mathrm{E}(y_{jt}|\mathcal{D}_{t-1}) &= f_{jt}, \quad \mathrm{V}(y_{jt}|\mathcal{D}_{t-1}) = q_{jt}, \\ \mathrm{E}(y_{jt}|M_{ji}, \mathcal{D}_{t-1}(j)) &= f_{jit}, \quad \mathrm{V}(y_{jt}|M_{ji}, \mathcal{D}_{t-1}(j)) = q_{jit}. \end{aligned}$$

Due to (4.1), I have

$$\mathrm{E}(y_{jt}|\mathcal{D}_{t-1}) = \mathrm{E}(y_{jt}|\mathcal{D}_{t-1}(j)), \quad \mathrm{V}(y_{jt}|\mathcal{D}_{t-1}) = \mathrm{V}(y_{jt}|\mathcal{D}_{t-1}(j)).$$

Then

$$\begin{aligned} f_{jt} &= \mathrm{E}(y_{jt}|\mathcal{D}_{t-1}(j)) = \sum_{i=1}^{m_j} p_{jt}(i) f_{jit}, \\ q_{jt} &= \mathrm{V}(y_{jt}|\mathcal{D}_{t-1}(j)) = \sum_{j=1}^{m_j} p_{jt}(i)[(f_{jit} - f_{jt})^2 + q_{jit}]. \end{aligned} \tag{4.4}$$

The 1-step-ahead covariance of y_{jt} and y_{kt} can be calculated using the following theorem.

Theorem 4.3.1. *Under the modeling context of multi-process, class I, for $j < k$, the 1-step-ahead covariance of y_{jt} and y_{kt} of an MDM is*

$$C(y_{jt}, y_{kt}|\mathcal{D}_{t-1}) = \sum_{i=1}^{m_j} C(y_{jt}, y_{kt}|M_{ji}, \mathcal{D}_{t-1}) p_{jt}(i).$$

Proof. Since $j < k$, I have

$$p(y_{kt}|M_{ji}, \mathcal{D}_{t-1}) = p(y_{kt}|\mathcal{D}_{t-1}),$$

and thus

$$\mathrm{E}(y_{kt}|M_{ji}, \mathcal{D}_{t-1}) = \mathrm{E}(y_{kt}|\mathcal{D}_{t-1}), \ \forall i \in \{1, \cdots, m_j\}.$$

Therefore, the covariance of series j and series k is

$$\begin{aligned}
\mathrm{C}(y_{jt}, y_{kt}|\mathcal{D}_{t-1}) &= \mathrm{E}(y_{jt}y_{kt}|\mathcal{D}_{t-1}) - \mathrm{E}(y_{jt}|\mathcal{D}_{t-1})\mathrm{E}(y_{kt}|\mathcal{D}_{t-1}) \\
&= \sum_{i=1}^{m_j} [\mathrm{E}(y_{jt}y_{kt}|M_{ji}, \mathcal{D}_{t-1}) - \mathrm{E}(y_{jt}|M_{ji}, \mathcal{D}_{t-1})\mathrm{E}(y_{kt}|\mathcal{D}_{t-1})]p_{jt}(i) \\
&= \sum_{i=1}^{m_j} [\mathrm{E}(y_{jt}y_{kt}|M_{ji}, \mathcal{D}_{t-1}) - \mathrm{E}(y_{jt}|M_{ji}, \mathcal{D}_{t-1})\mathrm{E}(y_{kt}|M_{ji}, \mathcal{D}_{t-1})]p_{jt}(i) \\
&= \sum_{i=1}^{m_j} \mathrm{C}(y_{jt}, y_{kt}|M_{ji}, \mathcal{D}_{t-1})p_{jt}(i).
\end{aligned} \tag{4.5}$$

□

Then, for each series $j \in \{1, \cdots, m\}$, the joint predictive mean $\boldsymbol{f}_t = \mathrm{E}(\boldsymbol{y}_t|\mathcal{D}_{t-1})$ and covariance matrix $\boldsymbol{Q}_t = \mathrm{V}(\boldsymbol{y}_t|\mathcal{D}_{t-1})$ at each time t can be calculated recursively from $j = m$ to 1 with the following steps.

i. Start at $j = m$: with $M(j) = i$, this defines $M_{ji} = (pa^{[i]}(j), \boldsymbol{S}_j^{[i]}, p_{j\lambda}^{[i]})$, $pa^{[i]}(j) = \varnothing$, $i \in \{1, 2, \cdots, m_j\}$ at $j = m$. Compute the univariate mean and variance of y_{mt} under each M_{ji}, namely

$$f_{jit} = f_{mit}(\varnothing), \quad q_{jit} = \frac{r_{mit}}{(r_{mit} - 2)} q_{mit}(\varnothing)$$

using the implied simplified forms of (5.27) with all terms in $\boldsymbol{y}_{pa(m)}$ set to zero. Then calculate the marginal predictive mean and variance for $y_{mt}|\mathcal{D}_{t-1}$ using (4.4). Insert f_{mt} as the $m-$the element of $\boldsymbol{f}_t$ and q_{mt} as the $(m, m)-$element of $\boldsymbol{Q}_t$.

ii. For $j = m-1, m-2, \ldots, 1$ in turn: With $M_{ji} = (pa^{[i]}(j), \boldsymbol{S}_j^{[i]}, p_{j\lambda}^{[i]}), i = 1, \cdots, m_j$, calculate the following quantities in parallel:

$$\begin{aligned}
f_{jit} &= \mathrm{E}[\mathrm{E}(\boldsymbol{x}'_{jit}\boldsymbol{a}_{ji\phi t} + \boldsymbol{y}'_{pa^{[i]}(j),t}\boldsymbol{a}_{ji\gamma t}|\mathcal{D}_{t-1}, \boldsymbol{y}_{pa^{[i]}(j),t}, M_{ji})|M_{ji}, \mathcal{D}_{t-1}], \\
&= \boldsymbol{x}'_{jit}\boldsymbol{a}_{ji\phi t} + \boldsymbol{f}'_{pa^{[i]}(j),t}\boldsymbol{a}_{ji\gamma t}, \\
q_{jit} &= \frac{r_{jit}}{(r_{jit}-2)}(s_{ji,t-1} + u_{jit}) + \boldsymbol{a}'_{ji\gamma t}\boldsymbol{Q}_{pa^{[i]}(j),t}\boldsymbol{a}_{ji\gamma t},
\end{aligned}$$

where u_{jit} is

$$\boldsymbol{f}'_{pa^{[i]}(j),t}\boldsymbol{R}_{ji\gamma t}\boldsymbol{f}_{pa^{[i]}(j),t} + \mathrm{tr}(\boldsymbol{R}_{ji\gamma t}\boldsymbol{Q}_{pa^{[i]}(j),t}) + 2\boldsymbol{x}'_{jit}\boldsymbol{R}_{ji\phi\gamma t}\boldsymbol{f}_{pa^{[i]}(j),t} + \boldsymbol{x}'_{jit}\boldsymbol{R}_{ji\phi t}\boldsymbol{x}_{jit}.$$

These two moments are obtained from the conditional distribution given by (2.3). Then calculate f_{jt} and q_{jt} according to (4.4), and insert f_{jt} as the $j-$th element of $\boldsymbol{f}_t$ and q_{jt} as the $(j,j)-$element of $\boldsymbol{Q}_t$, respectively. Write $\boldsymbol{a}_{j+1:m,i,\gamma t}$ for the $(m-j)-$vector that extends $\boldsymbol{a}_{ji\gamma t}$ with zeros for elements $h \notin pa^{[i]}(j)$. Then

$$\mathrm{C}(y_{jt}, \boldsymbol{y}_{j+1:m,t}|M_{ji}, \mathcal{D}_{t-1}) \quad = \quad \boldsymbol{Q}_{j+1:m,t}\boldsymbol{a}_{j+1:m,i,\gamma t}.$$

Then calculate $\mathrm{C}(y_{jt}, \boldsymbol{y}_{j+1:m,t}|\mathcal{D}_{t-1}) = \boldsymbol{q}_{j,j+1:m,t}$ according to (4.5), and insert element h of this vector as the $(j, h+j)$ and $(h+j, j)$ entries of $\boldsymbol{Q}_t$, for $h = 1 : m-j$.

iii. End: at this point, $j = 1$ and I have filled in all elements of the $m-$vector $\boldsymbol{f}_t$, $m \times m$ matrix $\boldsymbol{Q}_t$. The precision matrix can be calculated similarly as I discussed in Section 2.3.3.

4.3.3 k-step-ahead joint predictive moments– outline of Bayesian computation

At each time t, the k-step-ahead predictive mean, covariance matrix are calculated based on the k-step-ahead prior and forecasts summarized in Section 2.3.1. However, unlike the 1-step-ahead case, it is easy to prove that $C(\boldsymbol{y}_{pa(j),t+h}, \boldsymbol{y}_{j,t+h-p_{j\lambda}:t+h-1})$ is non-zero in many cases so I cannot use the recursive method developed in Section

4.3.2 to obtain its analytic form. Therefore, I generate Monte Carlo simulations to calculate the joint predictive moments for $\boldsymbol{y}_{t+h}$. The computational procedure is as follows. Let nmc be the sample size. At each time t, for each iteration $i = 1 : nmc$, and in the order of $j = m, m-1, \cdots, 1$ in turn, proceed with the following steps.

1. Sample index $i^* \in \{1, \cdots, m_j\}$ with probability $P(\text{index} = i^*) = p_{jt}(i^*)$.

2. For $h = 1$, draw $y^{(i)}_{j,t+1}$ according to the pdf $p(y_{j,t+1}|\boldsymbol{y}_{pa^{[i^*]}(j),t+h}, M_{ji^*}, \mathcal{D}_t)$, a T distribution as specified in (2.4). Append the sampled value to the vector of predictors $x^{(i)}_{j,t+2}$, to serve as the first lag for the next forecast $y^{(i)}_{j,t+2}$.

3. For $h = 2, \cdots, k$ in turn, according to (2.4), sample the h-step-ahead value $y^{(i)}_{j,t+h}$ from the conditional predictive distribution

$$p(y^{(i)}_{j,t+h}|\boldsymbol{y}_{pa^{[i^*]}(j),t+h}, \boldsymbol{x}^{(i)}_{j,t+h}, \mathcal{D}_t).$$

   Append the sampled value to the predictor vector $x^{(i)}_{j,t+h+1}$, to serve as the corresponding lag for the next few forecasts $y^{(i)}_{j,t+h+1:t+k}$.

4. In the order of $h = 1, 2, \cdots, k$, compute the sample mean and variance of $\{y^{(i)}_{j,t+h}, i = 1 : nmc\}$ as the predictive moments of $y_{j,t+h}$, $f_{jt}(h)$ and $q_{jt}(h)$. Insert $f_{jt}(h)$ as the j-th element of $\boldsymbol{f}_t(h)$ and $q_{jt}(h)$ as the (j, j) element of $\boldsymbol{Q}_t(h)$. Compute the sample covariance of $\{y^{(i)}_{j,t+h}, i = 1 : nmc\}$ and $\{y^{(i)}_{j+1:m,t+h}, i = 1 : nmc\}$ as the predictive covariance $C(y_{j,t+h}, y_{j,j+1:m,t+h}|\mathcal{D}_t)$; insert element s of this vector as the (j, s) and (s, j) entries of $\boldsymbol{Q}_t(h)$, for $s = j + 1 : m$.

Notice that when $j = m$, I have $pa^{[i^*]}(j) = \varnothing$ for all $i^* \in \{1, \cdots, m_j\}$.

The sequence of simulation steps above can be clearly shown in the picture below. For each iteration i, start from the first row, calculate each column from column 1 to k, and then move to the next row, until finishing by computing the mth row.

$$\begin{array}{ccccccc} y^{(i)}_{m,t+1} & \rightarrow & y^{(i)}_{m,t+2} & \rightarrow & \cdots & \rightarrow & y^{(i)}_{m,t+k} \\ y^{(i)}_{m-1,t+1} & \rightarrow & y^{(i)}_{m-1,t+2} & \rightarrow & \cdots & \rightarrow & y^{(i)}_{m-1,t+k} \\ \vdots & & \vdots & & & & \vdots \\ y^{(i)}_{1,t+1} & \rightarrow & y^{(i)}_{1,t+2} & \rightarrow & \cdots & \rightarrow & y^{(i)}_{1,t+k} \end{array}$$

4.3.4 k-step-ahead predictive return moments

Since the k-step-ahead (for $k > 1$) predictive distribution is not a standard distribution, such as normal or T, analytic forms of forecast moments are not available. Thus I also use Monte Carlo simulations to calculate the forecast mean and variance matrix of k-step-ahead forecasts. This can be done by adding another step after step 4 above: calculating the sample predictive mean and variance matrix for the series. In the following application in portfolio decisions, I forecast the first two moments of the predictive returns at each time t, then make decisions on portfolio weights allocation, and calculate quantities of practical interests such as the cumulative returns and risks.

4.4 Application to Financial Time Series

This portfolio study in financial time series illustrates the time-varying patterns of model parameters and the benefits of power discounting in forecast accuracy and Bayesian decisions.

4.4.1 Dynamic portfolio allocation

Following the dynamic portfolio allocation process and rules introduced in Section 3.5.1, the analysis in this chapter also uses traditional Bayesian mean-variance optimization (Markowitz, 1959) subject to constraints. In this section, I use extended BMA over a pool of models instead of any particular MDM. Since forecast means and variances are critical in mean-variance portfolio optimization methods, extended

BMA, with well-known robustness in capturing model uncertainty, can be expected to lead to practically more attractive sequential investment decisions.

The application uses time-varying autoregressive MDMs under extended BMA in forecasting and portfolio decision making. I again focus on measurements of practical relevance, i.e. realized investment performance. The three portfolio construction rules are the same as in Section 3.5.1, namely target, target constraint, and benchmark neutral rules. Since extended BMA allows adaptive learning of uncertainties in conditional dependence networks through time, different sparsity patterns on parental sets can be captured to avoid the collinearity issues.

4.4.2 Data

This analysis uses the same $m = 13$ financial time series introduced in Section 3.5.2, but the difference is that log of daily closing prices are modeled instead of returns. Hang (2011) illustrated that modeling prices can gain more predictive power than modeling returns.

Another difference is that the sample period is from 1st August, 2000 through 31st December, 2011, consisting of 2,979 days in total, double the time period of the previous study. I divide this period into two datasets: (1) from 1st August 2000 to 14th April 2006 (1,489 log-prices) as the training dataset to estimate the parameters, and (2) from 17th April 2006 to 31st December 2011 (1,490 log-prices) to evaluate out-of-sample performance of portfolio decisions.

The specific series as in Table 3.1. Figure 4.1 presents the daily log-prices of the 13 time series over the whole sample period. The vertical line in the middle of each plot divides the data into the training and testing period.

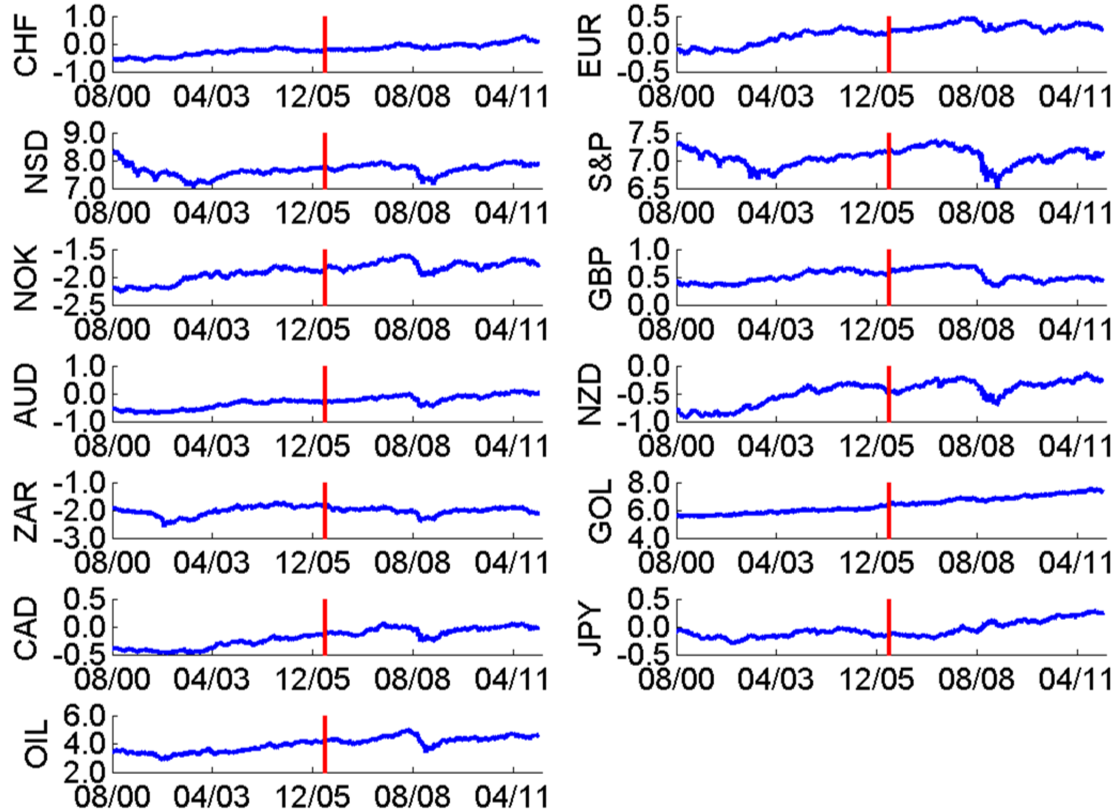


FIGURE 4.1: Daily log-prices of the 13 time series. The data set consists of 2,979 daily closing prices for each of these 13 assets, over the period of 01/08/2000 to 12/30/2011. The vertical red lines demarcate the first period of training data from the following period of test data

4.4.3 Model and parameter sets

I analyze the data using extended BMA on multiregression dynamic VAR models. The candidate models are MDMs with observation equation specified by (2.10). The state evolution models are vector random walks, whose form is a local trend model. Specifically,

$$\begin{aligned} y_{jt} &= \boldsymbol{F}'_{jit}\boldsymbol{\theta}_{jit} + \nu_{jit}, \quad \nu_{jit} \sim N(0, \lambda_{jit}), \\ \boldsymbol{\theta}_{jit} &= \boldsymbol{\theta}_{j,i,t-1} + \boldsymbol{w}_{jit}, \quad \boldsymbol{w}_{jit} \sim N(0, \boldsymbol{W}_{jit}), \end{aligned} \tag{4.6}$$

where $\boldsymbol{W}_{jit} = (1-\delta_{ji})\boldsymbol{C}_{j,i,t-1}/\delta_{ji}$, for $j = 1, \cdots, m$. The sequential updating of the distribution of $\boldsymbol{\theta}_{jit}$ is described in detail by West and Harrison (1997).

I calculate 1-step-ahead and 5-step-ahead predictive moments and make portfolio decisions during the testing period. The analytic forms of 1-day-ahead predictive moments are obtained as discussed in Section 4.3.2. The Monte Carlo method introduced in Section 4.3.3 is used to calculate the 5-day-ahead predictive moments for the log-prices, where 10,000 iterations have been generated at each time t. The 1-day-ahead and 5-day-ahead predictive moments of price returns are also calculated by the Monte Carlo method. The computing time mainly depends on the speed of the computation of two parts - the forward-filtering of the univaraite candidate DLMs, and the simulations to calculate 5-day-ahead moments.

Given the specified order of variables, the total number of possible univariate DLMs for all models is $l \times K \times (2^m - 1)$. Thus, since m is only 13, I calculate the forward-filtering results of all possible models over the training period in advance in parallel. I then select a subset of these models as candidates to apply our extended BMA to the testing data to make out-of-sample predictions.

The models selected are those with posterior probability larger than a threshold, so that only a medium number of models would be included in the averaging process for out-of-sample prediction. In fact, choice of this threshold is of subtlety since extremely small value can result in inclusion of too many models for averaging in terms of computation burden, which would rapidly increase as the dimension m increases. However, on the contrary, a large threshold may eliminate some essentially influential models so that out-of-sample prediction might degrade.

I consider every combination of the parameters in the first two rows of Table 4.1. This includes a total of $K = 25$ combinations as the value set $\boldsymbol{\Psi}$, from which discount factors $\boldsymbol{S}_j$ for each series j can be choson, where δ discounts regressions and β discounts state volatility.

To bear a practical relevance to the real market, the range of δ and β is chosen to be 0.975 to 0.995, and the range of α to be 0.950 to 1.000. Empirical evidence can be found in Xie (2013). I believe that δ, β, α are smaller than 1 since the historical information should be discounted instead of being magnified or constant when updating the current posterior distribution of the variance parameters and model probabilities.

For all candidate models of the form (4.6), I use shrinkage initial priors at $t = 0$:

$$(\boldsymbol{\theta}_{ji0}, \lambda_{ji0}|\mathcal{D}_0) \sim NIG(\boldsymbol{m}_{ji0}, \boldsymbol{C}_{ji0}, n_{ji0}/2, d_{ji0}/2)$$

where $\boldsymbol{m}_{jit} = (0, \cdots, 0)', \boldsymbol{C}_{jit} = \boldsymbol{I}$, and $n_{ji0} = m$ and $d_{ji0} = m$, for $j = 1, \cdots, m$. Note that the residual variances estimate is $s_{ji0} = d_{ji0}/n_{ji0} = 1$ for each j. The threshold for model selection is 0.001.

For portfolio decisions, the weekly (5-day) target return is chosen to be 0.5%, corresponding to an annual target return of 30%.

Table 4.1: List of parameter values used for model averaging

Parameters	Values
δ	0.975, 0.980, 0.985, 0.990, 0.995
β	0.975, 0.980, 0.985, 0.990, 0.995
α	0.950, 0.955, 0.960, 0.965, 0.970, 0.975, 0.980, 0.985, 0.990, 0.995, 1.000
l	1, 2
ρ	0.3
Threshold	0.001
nmc	10,000
Target Return	0.5%

4.5 Results

4.5.1 Model probability discounting

Posterior Probability

Figure 4.2 illustrates the marginal posterior probabilities of α, $\{P(\alpha|\mathcal{D}_t) : t = 1, 2, \cdots, T\}$, over the training period and testing period. Note that the models with no model probability discounting, i.e. $\alpha = 1$, are clearly dominated by those with discounting, i.e. $\alpha < 1$, for having obviously lower posterior probability throughout the whole sample period. This dominance confirms the advantage of model discounting in better adapting to new information. Notice that the posterior probabilities of $\alpha < 1$ stay relatively stable over time with insignificant differences between each other. Thus I need extra evidence for choosing an appropriate α.

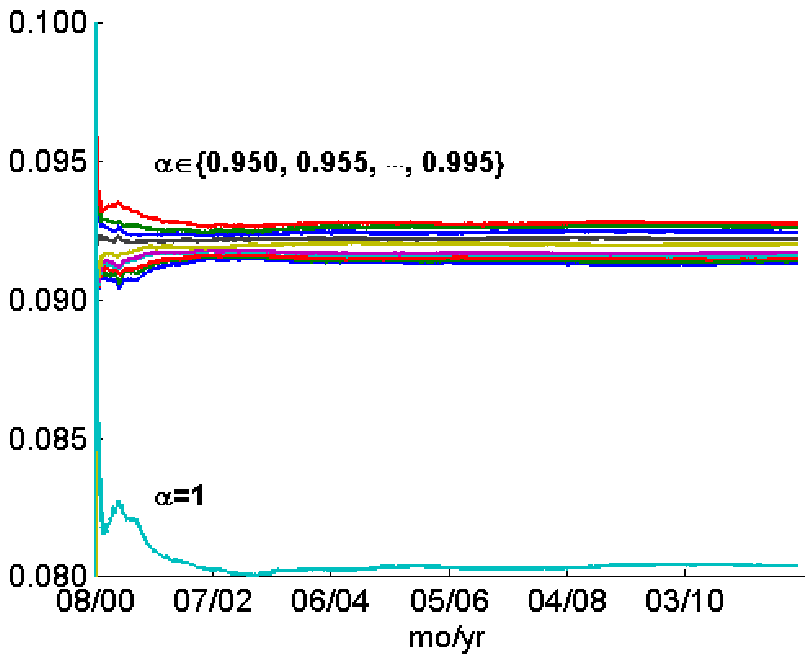
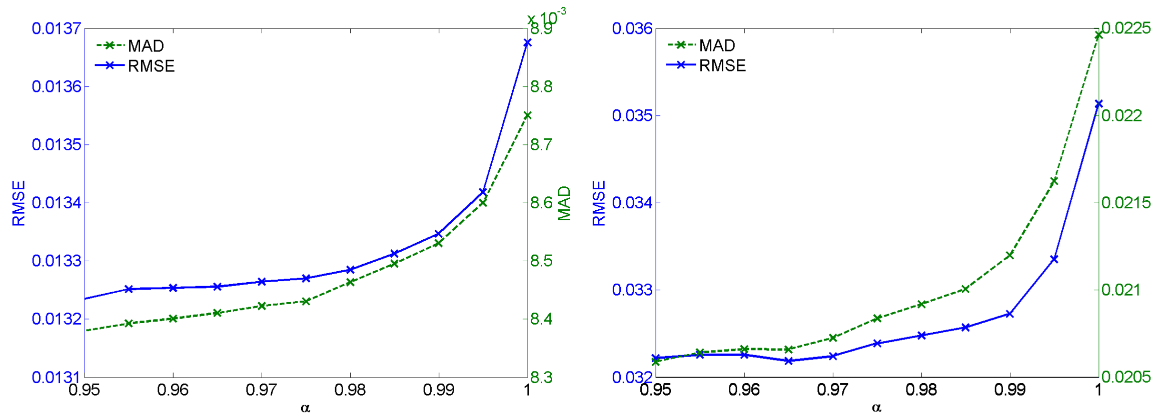


FIGURE 4.2: Posterior probabilities of model discount factor α, $\{P(\alpha|\boldsymbol{y}_{1:t}) : t = 1, 2, \cdots, T\}$, over the training data period and testing data period

Forecast accuracy

Figure 4.3 shows the forecast accuracy measured by out-of-sample RMSE and MAD of the 1-step-ahead and 5-step-ahead forecasts under different α values during the testing period. As shown in both figures, higher α leads to lower prediction accuracy. In most of the following discussions, I take $\alpha = 0.980$ as an example to illustrate the practical relevance of model averaging, since both the forecast accuracy and model posterior probabilities are relatively high at this power discounting level.



(a) 1-step-ahead (b) 5-step-ahead

FIGURE 4.3: Out-of-sample prediction accuracy measured by RMSE and MAD of 1-step-ahead and 5-step-ahead forecasts under different model discount factors

Let $e_{jt}(k), t = 1, 6, \cdots, T$, denote the standardized k-step-ahead forecast errors, i.e.

$$e_{jt}(k) = \frac{f_{jt}(k) - y_{j,t+k}}{\sqrt{q_{jt}(k)}}.$$

The upper subplot of Figure 4.4 shows the trajectories of log prices, 5-day-ahead forecast mean, and 90% credit interval of OIL. The lower subplots of Figure 4.4 illustrate the 5-day-ahead standardized forecast errors of log prices of Oil, i.e. $e_{13t}(5)$, and its QQ plot. As can be seen from the normal QQ plot, the shape of the QQ curve suggests that the standardized errors follow a distribution slightly more fat-tailed than the standard normal distribution. This justifies our assumption that the observation noise follows a mixture T-distribution. Moreover, it approximately lies on the line $y = x$, supportive of a reasonable goodness-of-fit. Table 4.2 shows the percentage of real values falling into the 90% credible interval of the 5-step-ahead predictive distribution for each asset. The 90% credible intervals generally provide a good predictive reference of the empirical coverage of data values. A good example is that of the 90% credible intervals for OIL, with 90.2% coverage.

Table 4.2: Coverage is the percentage of real values falling into the 90% credible interval of the 5-step-ahead predictive distribution for each asset

Asset	CHF	EUR	NSD	S&P	NOK	GBP	AUD
Coverage	79.8%	73.4%	89.6%	89.2%	77.8%	82.2%	75.4%
Asset	NZD	ZAR	GOL	CAD	JPY	OIL	
Coverage	82.5%	84.5%	87.5%	86.5%	87.5%	90.2%	

4.5.2 Time-varying auto-regressive lag structures

Figure 4.5 shows the trajectories of average lag order for each asset and the posterior probability of AR lag order equal to 2 over the testing period, i.e. $E(p_{j\lambda}|\mathcal{D}_t)$ and $P(p_{j\lambda} = 2|\mathcal{D}_t)$. It can be seen that all series demonstrate model uncertainties on AR order over time. The difference of average AR order across assets can be significant.

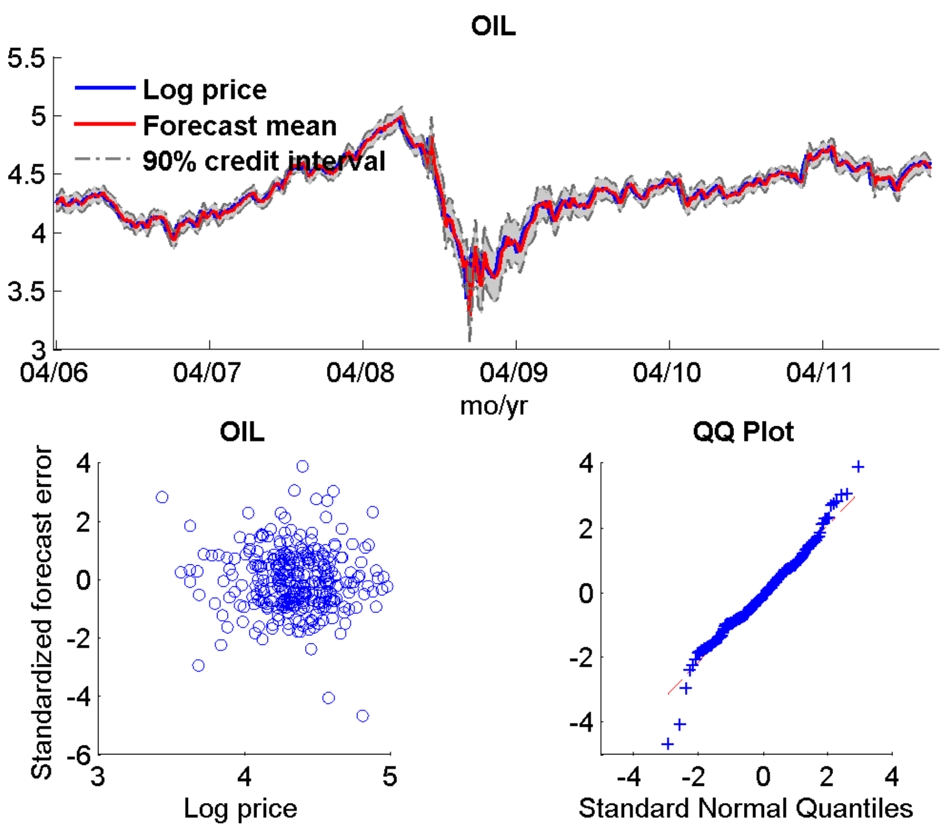


FIGURE 4.4: Log price, 5-day-ahead predictive mean, and their difference, i.e. forecast error, versus log price

For example, CHF has relatively stable, high posterior probabilities of AR order equal to 1 during most of the test period, while S&P has very volatile probabilities of AR order equal to 1 throughout the testing period.

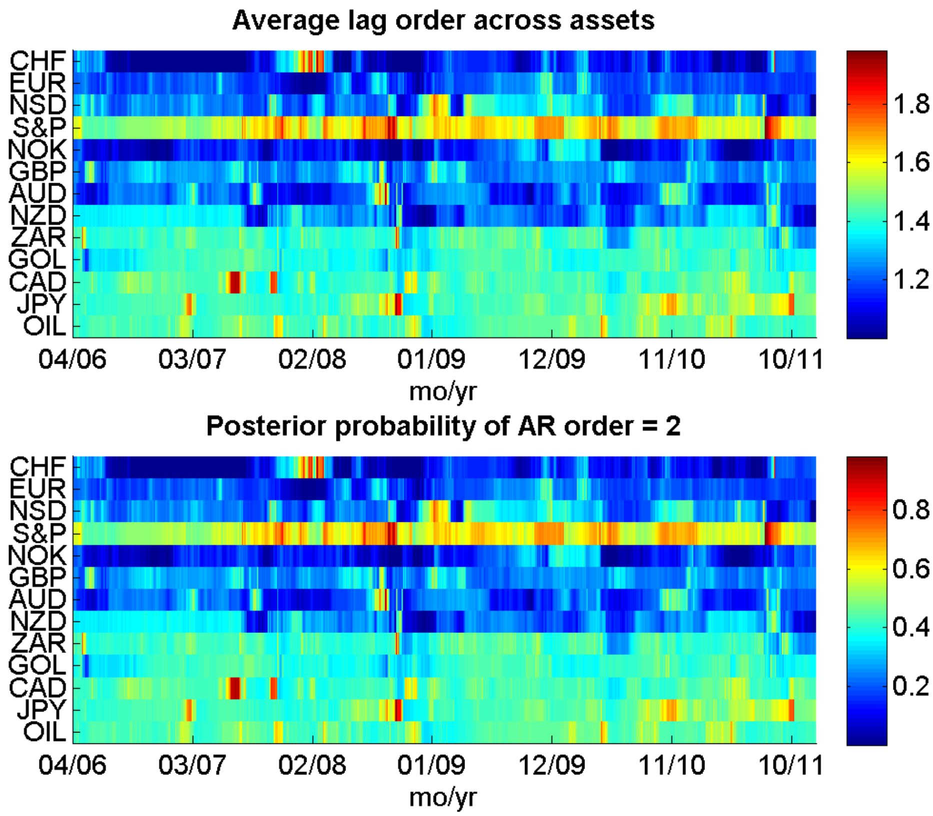


FIGURE 4.5: Bayesian model averaged lag order for each asset (upper frame), and posterior probability of AR order = 2 (lower frame)

4.5.3 Series-specific time-varying discounting methods for stochastic volatilities

Figure 4.6 plots the trajectories of average observation and state discount factors for each asset over the testing period, i.e. $E(\delta_j|\mathcal{D}_t)$ and $E(\beta_j|\mathcal{D}_t)$. As can be seen, there is shifting of model probabilities over time across both state and observation discount factors. For instance, I can observe one striking decline of state discount factors, β_j for almost all the 13 assets from the September of 2008 onwards, which coincides with the onset of the global financial crisis. This is the time when global financial market began to crash and price volatilities of almost all assets surged to historically high levels. In addition to this time-varying feature, it can also be seen that before the 2008 crisis, idiosyncratic volatility dominates so that the levels of the β_j show larger differences than those after the crisis when systematic volatility dominates.

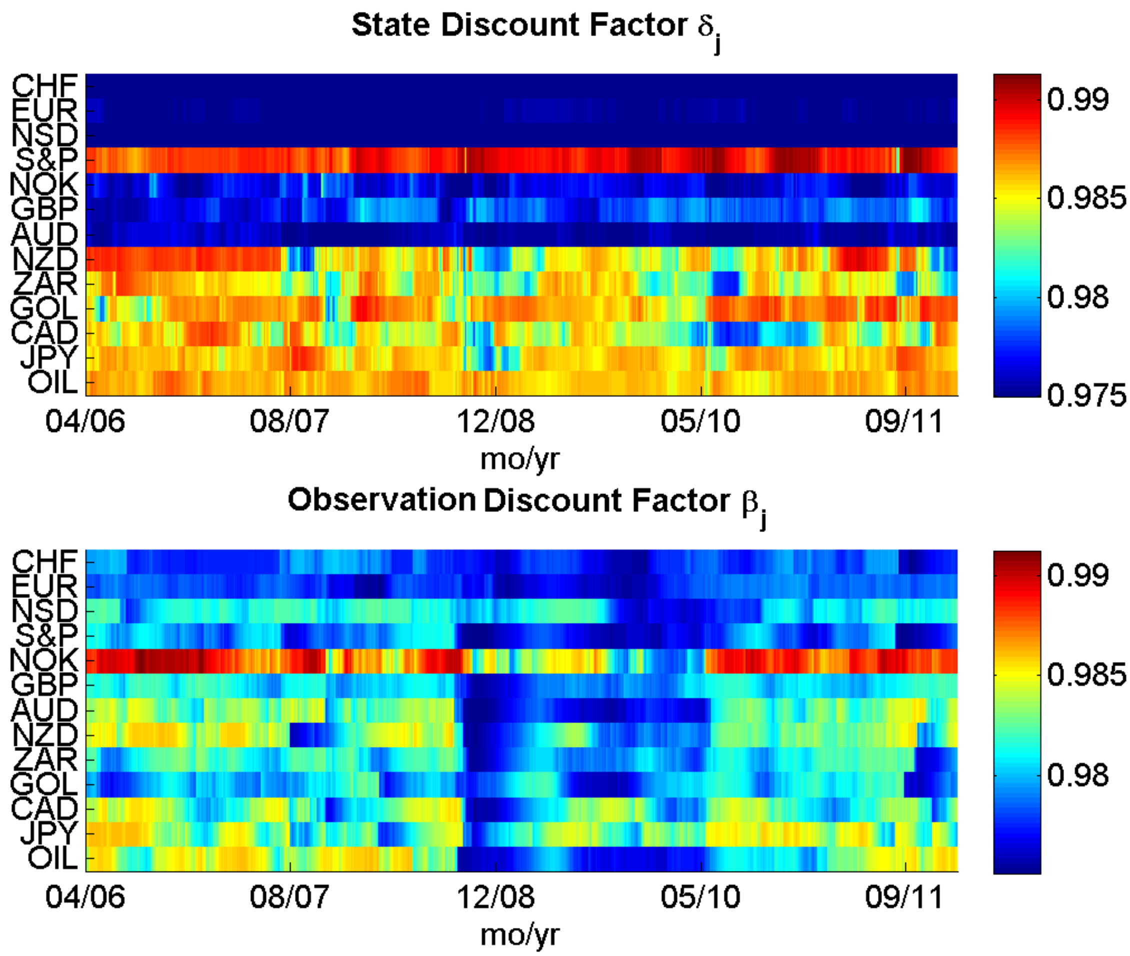


FIGURE 4.6: Summary posterior inferences on observation and state discount factors for each asset

4.5.4 Dynamic sparse conditional dependence networks

Trajectories of parent inclusion posterior probabilities, $P(\zeta_{ji} = 1|\mathcal{D}_t)$, $j = 1, \cdots, m$; $i = j+1, \cdots, m$; $t = 1, \cdots, T$, for each asset j are shown in Figures 4.7, 4.8 and

4.9. As can be seen, model uncertainties on parental sets demonstrate time-varying and series-specific patterns. For example, the posterior probability of S&P belonging to NSD's parent set is high throughout the whole sample period. This relationship is in accord with the common market sense that the two stock indices are highly correlated. Similarly, NZD is AUD's parent with posterior probability close to 1 throughout the whole sample period, while the two currencies have a well-known high correlati on. Notice that model averaging allows time-varying parents for each asset.

Figure 4.10 shows the trajectories of the average size of parental set of each asset, i.e. $E(\sum_{i=j+1}^{m} \zeta_{ji} | \mathcal{D}_t)$. Compared with the highly-volatile parent inclusion probabilities over time, the average sizes of parent sets are more stable for most of the 13 assets. The numbers on the right of this figure are the sample means of the posterior means of the parental set size of each asset over the testing period, which are all less than the maximum possible parental size $m-j$. This small volatility and average size of parent set recognizes collinearity of different assets, and thus supports the choice of a sparse prior for parent inclusion.

4.5.5 Portfolio analysis

Performance measures

As shown in numerous previous works, there is no single conclusive measure to evaluate the performance of dynamic portfolio allocation strategies. I consider popular performance measures which are often adopted in the literature, including cumulative return (CR), mean realized return (MRR), realized risk (RR), projected risk (PR), Sharpe ratio (SR) and projected Sharpe ratio (PSR).

Cumulative return. Denote RR_t to be the realized return of the portfolio at time t. Then $CR_t = \prod_{\tau=1}^{t} (1 + RR_\tau)$.

Mean realized return. $MRR = \sum_{t=1}^{T} RR_t/T$, where T is the total number of

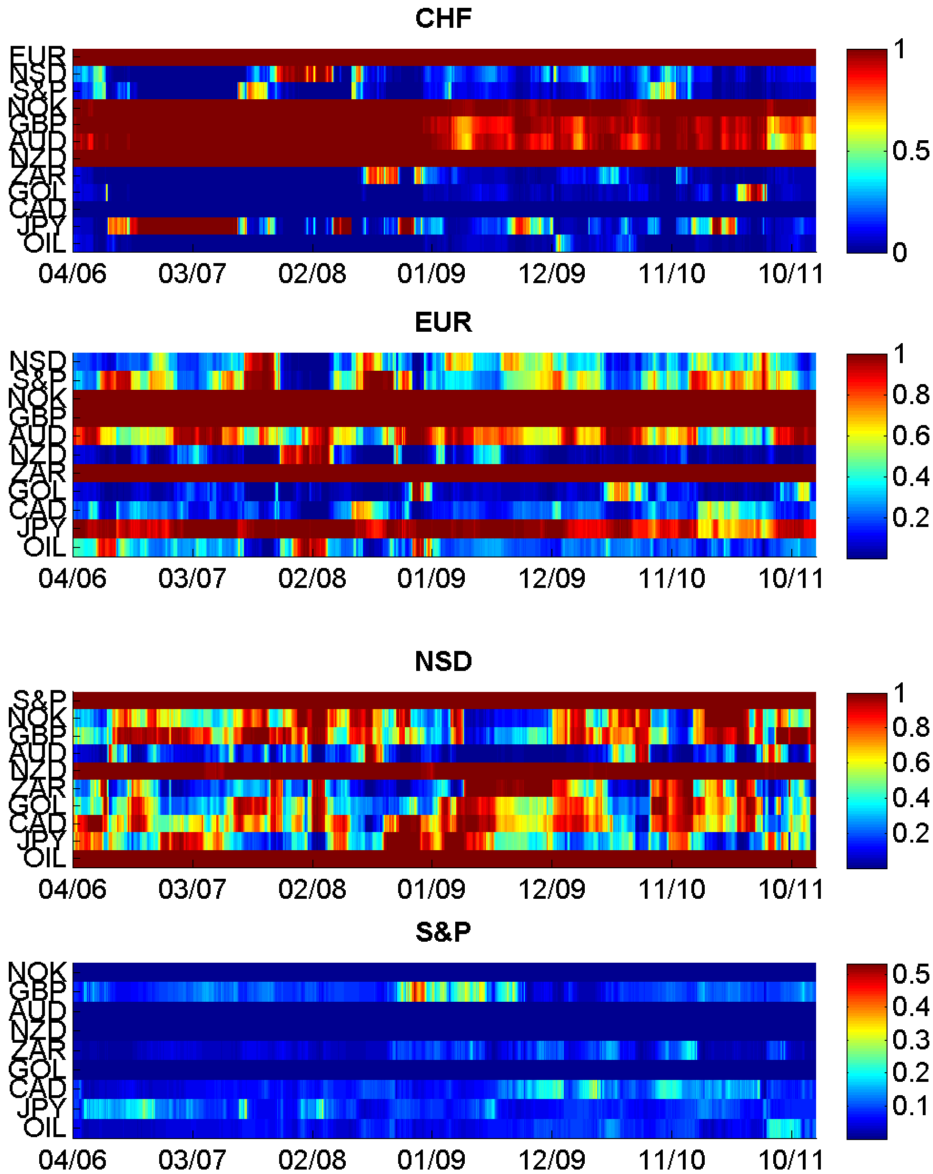


FIGURE 4.7: Parent inclusion posterior probabilities for each asset over the testing period

trading days.

Risk. Risk is defined as the standard deviation of $RR_{1:T}$.

Projected Risk. This is the $1/k$-step-head predictive standard deviations of the portfolio returns $PR_t = \sqrt{\boldsymbol{w}_t' \boldsymbol{f}_t \boldsymbol{w}_t}$.

Sharpe ratio. This measure is the ratio of annualized MRR over the annualized

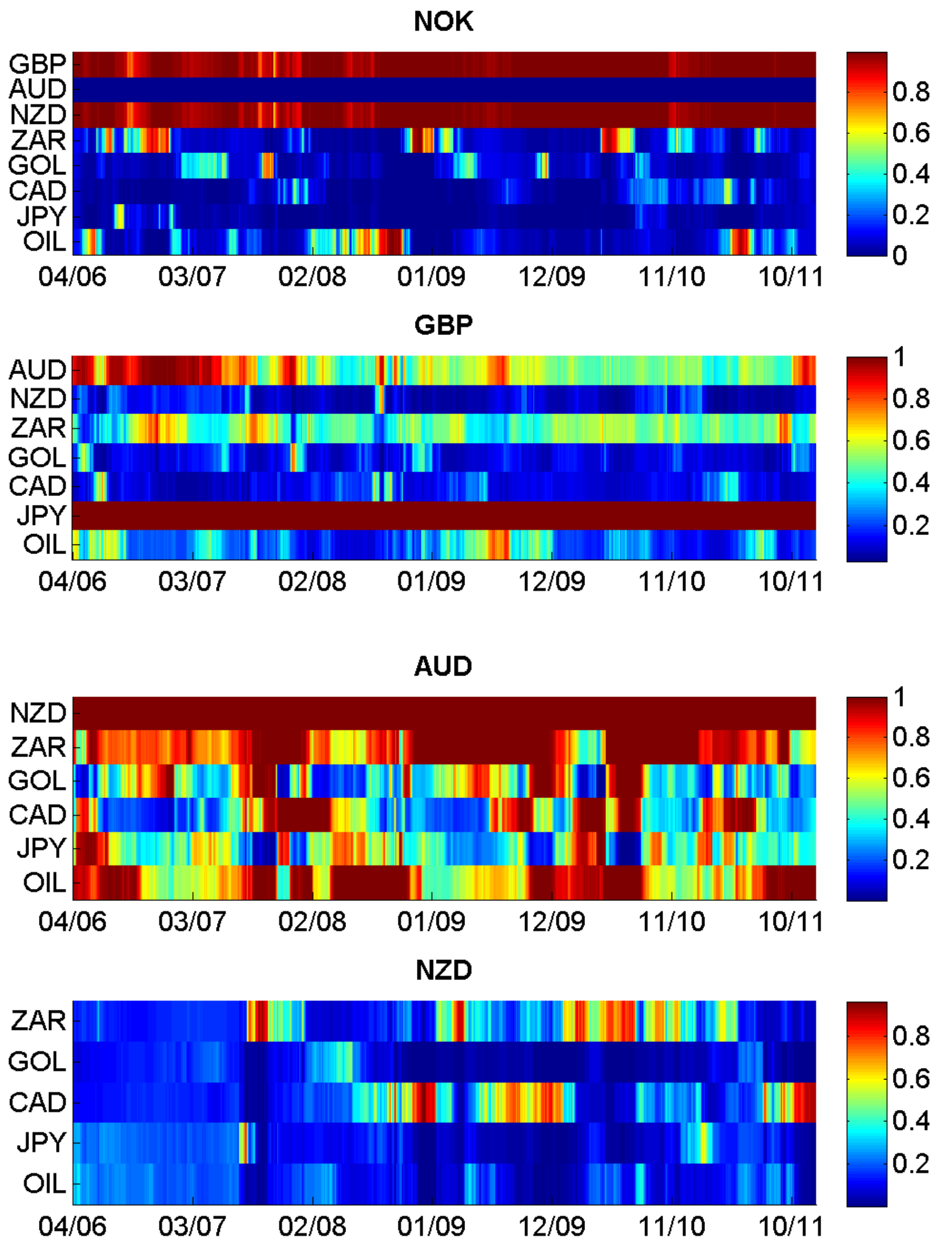


FIGURE 4.8: Parent inclusion posterior probabilities for each asset over the testing period

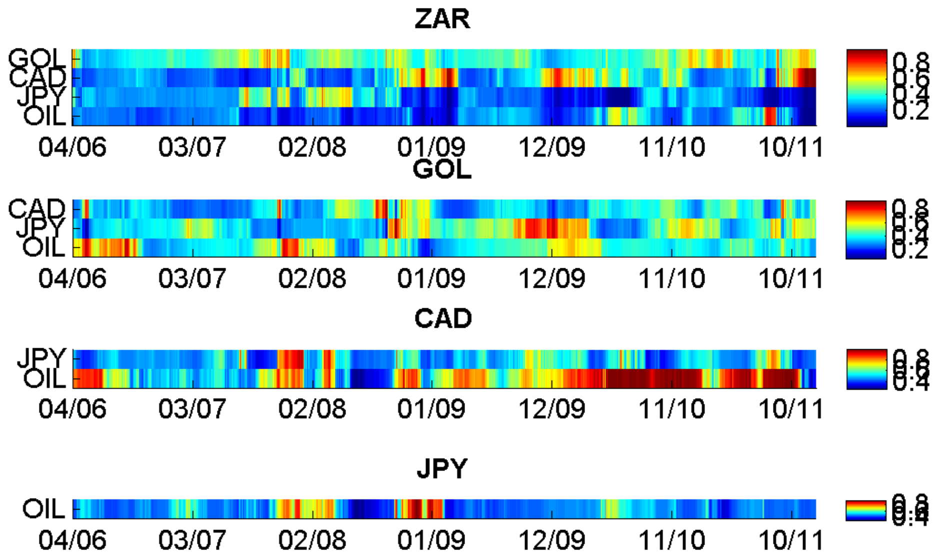


FIGURE 4.9: Parent inclusion posterior probabilities for each asset over the testing period

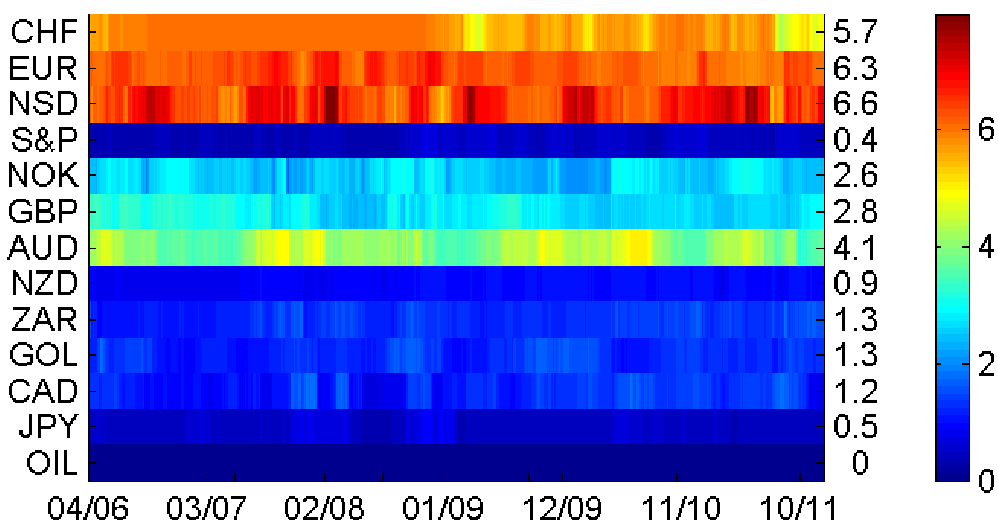


FIGURE 4.10: Average parental set size across assets when $\alpha = 0.980$

risk, i.e. $SR = MRR/Risk$.

Projected Sharpe ratio. This is the ratio of the 1/k-step-ahead predictive mean return and PR of the portfolio at time t. $PSR_t = \boldsymbol{w}_t'\boldsymbol{f}_t/PR_t$.

Performance comparison with benchmark index

Professional money managers, known as commodity trading advisors (CTA), typically monitor managed futures accounts, which generally have exposure to a number of markets such as stocks, derivatives, commodities, energy, agriculture and currency. The Newedge CTA Index is designed to track the largest 20 (by AUM) CTAs and be representative of the managed futures space. Since the portfolio of the 13 assets is comparable to that of CTAs, I compare the performance of my above portfolios with the benchmark index of CTAs.

1-step-ahead forecasts

Set daily target return to be 0.1%, which corresponds to a 30% annual return. Figure 4.11 shows the cumulative returns, Sharpe ratios and risks of the three types of portfolios based on 1-day-ahead forecasts under different α and the corresponding values of the CTA index over the testing period. It can be seen that the performance measures of the target and benchmark neutral portfolios are close to each other, while the target constraint portfolio has different performance from them. Compared with the CTA index, target and benckmark neutral portfolios have larger CRs and SRs, while the risks are also higher as the trade-off of returns.

5-step-ahead forecasts

Now consider 5-step-ahead forecasts and decisions, and set 5-day target return to be 0.5%, which also corresponds to a 30% annual return. Figure 4.12 shows the three performance measures of the three types of portfolios based on 5-day-ahead forecasts under different α's and the corresonding values of the CTA index over the testing

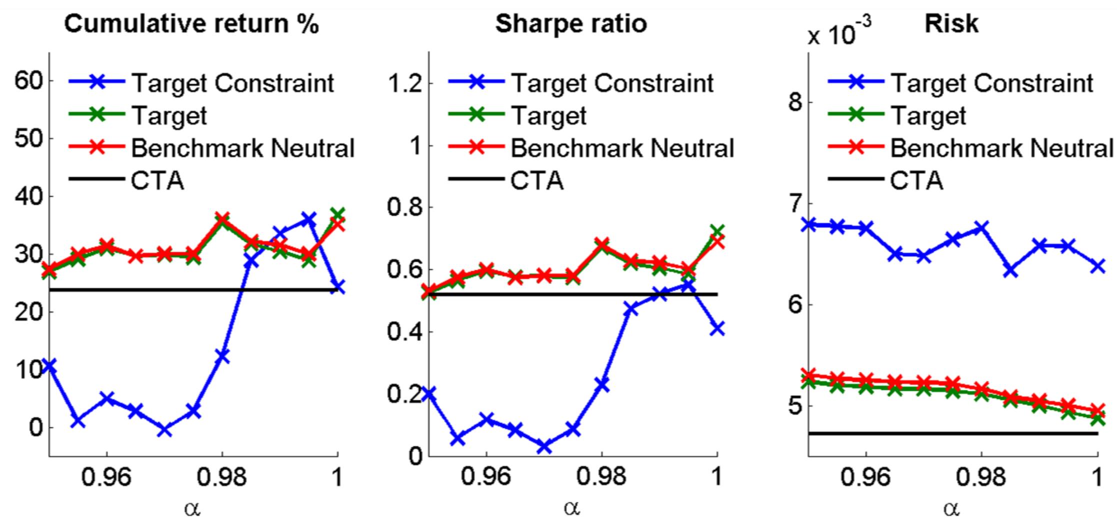


FIGURE 4.11: Portfolio CRs, SRs and risks based on 1-day-ahead forecasts under different α values

period. It is evident that when $0.960 \leqslant \alpha \leqslant 0.980$, the values of CRs and SRs lie in a higher range than that of α elsewhere for all three types of portfolios, while the risk is only slightly higher. This implies that the model without power discount is clearly dominated by those with it in terms of portfolio performance. Compared with the CTA index, all three types of portfolios with $0.960 \leqslant \alpha \leqslant 0.980$ have larger CRs and SRs, while the risks are also larger but better-compensated in exchange for returns. Again, if I take $\alpha = 0.980$ as an example, a clearer performance comparison between the CTA index and the three types of portfolios is shown in Table 4.3. These comparisons demonstrate the practical benefits resulted from our constructed portfolios. Figure 4.13 plots the trajectories of cumulative returns, projected risks, model Sharpe ratios from portfolios using the three types of allocation rules. All of the three portfolios have higher cumulative returns than that of the CTA during most of the testing period. The target constraint portfolio empirically realize a largest cumulative return as a trade-off of having the largest risk among the three types of portfolios, yet it also has the least projected Sharpe ratios during most of the testing period.

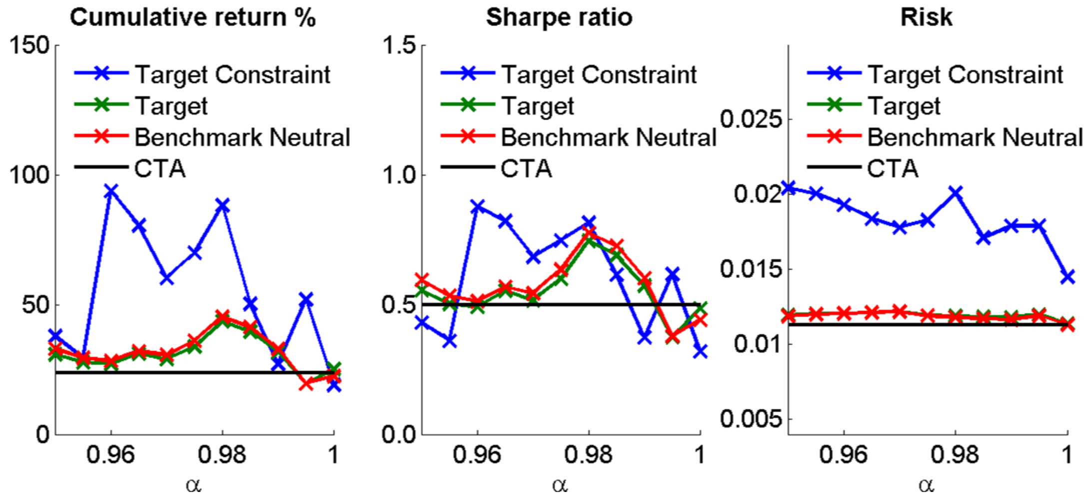


FIGURE 4.12: Portfolio CRs, SRs and risks based on 5-day-ahead forecasts under different α values

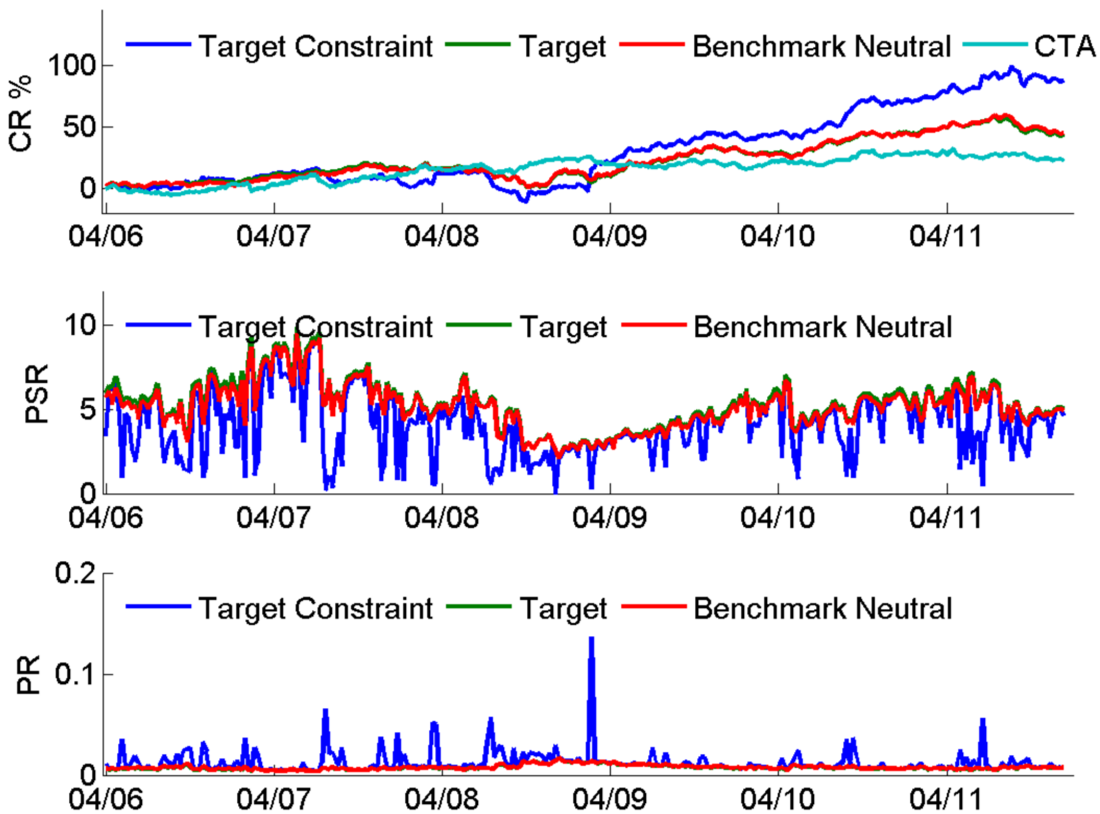


FIGURE 4.13: Portfolio CRs, PRs and PSRs based on 5-day-ahead forecasts over the testing period, when $\alpha = 0.980$

Table 4.3: Out-of-sample portfolio performances on 5-day returns under three allocation rules using extended BMA on MDM models with $\alpha = 0.980$, and the performance of the Newedge CTA Index. Sharpe ratios are annualized

	CTA Index	Target	Target Constraint	Benchmark Neutral
MRR	0.0008	0.0023	0.0012	0.0013
Risk	0.0113	0.0201	0.0119	0.0118
Sharpe ratio	0.4977	0.8150	0.7451	0.7764

Figures 4.14 and 4.15 plot the trajectories of the CRs, PSRs, and PRs of the three types of portfolio rules under models with different AR lag orders. The three models are (1) the model with AR order being only 1, (2) the model with AR order being only 2, and (3) averaged model over models (1) and (2) by extended BMA. From Figure 4.15, model (3) has more volatile PSRs and PRs than the other two models. Model (3) slightly dominates in terms of cumulative returns over the target and benchmark neutral portfolio, and also outperforms model (1) for the target constraint portfolio. However, it is dominated by model (2) for the target constraint portfolio, though the difference is much smaller than its difference with model (1). For the benchmark neutral portfolio, it is clear that the CRs of the benchmark asset, S&P 500, are dominated by all the three models regardless of AR order for most of the testing period.

4.6 Summary

In this chapter, I introduce and explore a novel approach to address model uncertainty issues in MDMs by transforming them from multivariate to univariate contexts, with normal Bayesian model mixing and averaging of MDMs. The sequential model selection procedure addresses uncertainty about a variety of model parameters including conditional dependence networks, discount factors, AR lag structures and so on. The conditional dependence network, represented by parental sets, provides a natural parallelizable framework under which I design parallel algorithms to enable

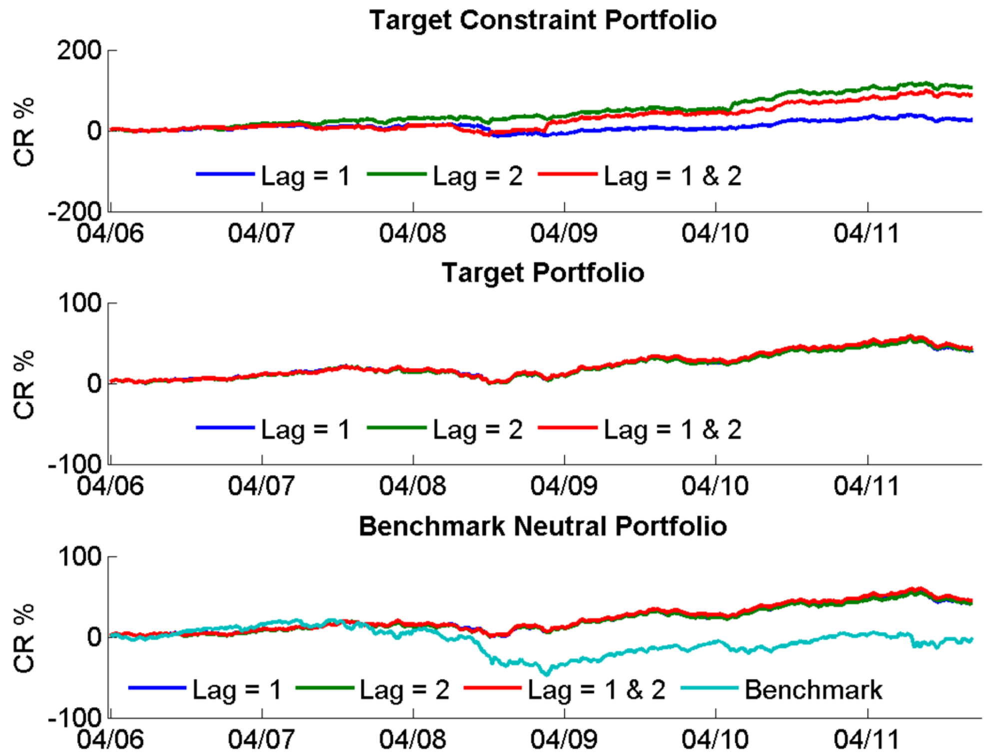


FIGURE 4.14: Portfolio CRs based on 5-day-ahead forecasts under different lag orders and portfolio rules, with the same $\alpha = 0.980$

previously unfeasible computation. The results of extended BMA suggest that there are indeed time-varying patterns of auto-regressive lag structures, discount factors for stochastic volatilities, and conditional dependence networks. Moreover, the decomposable property of MDMs enables all of these features as not only time-varying, but also series-specific, providing flexibility in modeling individual series. Additionally, the results of utilizing power discounting on model probabilities succeeds in not only scoring higher posterior probability, but also locating more practically lucrative models in the financial application, demonstrating both analysis feasibility and resulting forecasting and decision outcomes.

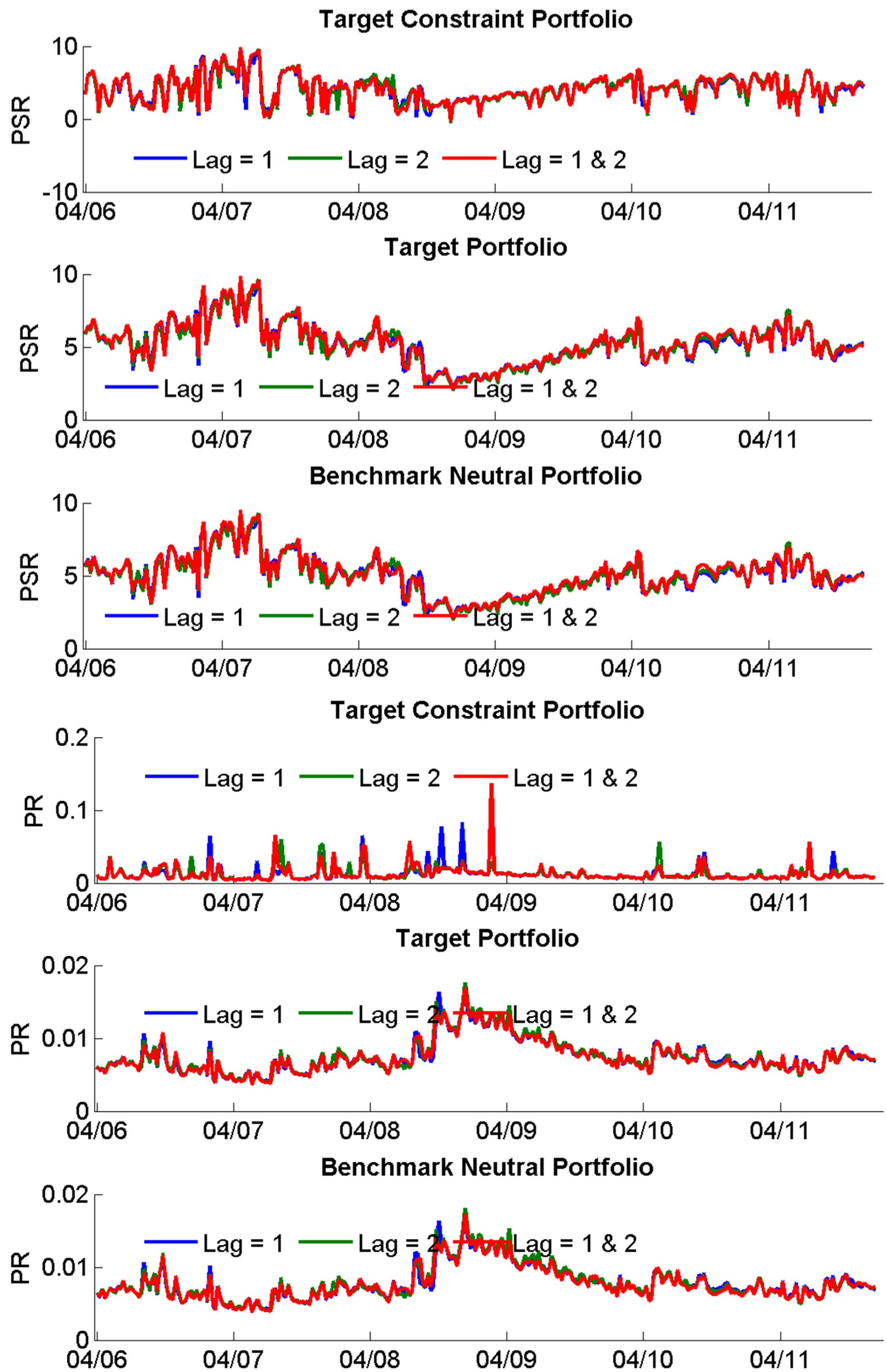


FIGURE 4.15: Portfolio PSRs (upper frame) and PRs (lower frame) based on 5-day-ahead forecasts under different lag orders and portfolio rules, with the same $\alpha = 0.980$

5

Common State Coupled Dynamic Linear Models

5.1 Introduction

A new model context– common state coupled dynamic linear model (CSCDLM) is proposed in this chapter. Interest lies in m univariate series $y_{jt},\ j = 1:m$, over time t, each following a DLM with its own, individual state vector elements together with an additional state vector that is shared across all series. One motivating context is a problem with large m and multiple common predictors with an assumedly common state vector–so that information about common effects is shared across series– but also with series-specific effects. Previous works often assume the observational noise to be independent across series and all regression vectors are known at $t-1$. A key example motivates us when the series-specific effects arise via a compositional DLM representation of multivariate stochastic volatility. I thus relax the assumption to allow correlated, time-varying observation variance matrices, as well as to incorporate random regression vectors.

A modest approximation using variance discounting enables closed-form, sequential updating and filtering, and thus avoids MCMC. The highlight here is that observational error variances are nuisance parameters, on which the full posterior inference is of little practical importance. An analytically tractable approximation using plug-in estimates, is of immense practical utility. For instance, it enables us to rerun analyses much faster and do proper, sequential step-ahead forecasting needed in real world application. A comparison with MCMC analysis in the slightly modified model justifies adequacy of the approximation.

A core motivating example is the problem of short-term prediction of electricity demand in "Smart Grid" scenarios in which consumers receive information about the current and expected future price of electricity. Previous works include well-known features such as in different levels of seasonality (yearly, weekly, daily) and weather dependence such as Harvey and Koopman (1993); Smith and Kohn (2002); Dordonnat et al. (2008). Hosking et al. (2013) is the first paper to add the new feature of price incentives into the model. They used multivariate linear Guassian state space models constructed with evolution noise vectors that are independent and have constant variance, with variance parameters estimated by maximum likelihood. These models do not enable either time-varying, correlated structure or online learning of the covariance structure of the state and observational evolution noise vectors. I would like to have a fully online and sequential learning method to construct time-varying, correlated state and observation covariance matrices, and address this here.

The model developed by Hosking et al. (2013) is a multivariate model based on a data vector containing all the load observations for a day. It essentially approximates the daily load curve by 12 periodic spline basis functions, which simplifies the model and provides consistency over changes of time granularity of observations. The daily load curve is modeled as a weighted sum of basis functions, and the "weights" are the basis function coefficients for that day. The influence of explanatory factors is modeled by their effects on these basis function coefficients. The coefficients of the basis functions are shared across dimensions of load values on the same day, thus are "common" states to the dependent variables. In order to introduce multivariate Cholesky stochastic volatility to this model, I combine series-specific states with the common states on each univariate dimension of the observation equation.

I begin with preliminaries and notation for the CSCDLMs, and describe the forward-filtering, backward-smoothing (FFBS) theorems in Section 5.2. Then a novel extension of CSCDLMs to incorporate multivariate stochastic volatility for observation covariance matrices is introduced in Section 5.3. Here I develop extensions of the theory for Bayesian analytics in sequential updating and forecasting. Section 5.4 proposes a custom method of approximating learning of unknown observation variances with online and analytic updates. Section 5.5 describes a fully Bayesian analysis and computation including FFBS for states and variances. Section 5.6 shows how the common state coupled stochastic volatility structure can be naturally embedded in a broad class of time series models, and develops a detailed analysis of a residential electricity demand data set for illustration of the effectiveness and utility of the new CSCDLMs. Section 5.7 concludes. Appendix 5.A provides additional proofs of theorems related to the analyses in this chapter.

5.2 Coupled Univariate DLMs with Idiosyncratic and Common States

Suppose $\boldsymbol{y}_t = (y_{1t}, \ldots, y_{mt})'$ is a vector time series and each univariate time series y_{jt}, $(j = 1 : m)$, follows a DLM defined via the following observation and state equations:

$$y_{jt} = \boldsymbol{F}'_{jt}\boldsymbol{\theta}_{jt} + \boldsymbol{H}'_{jt}\boldsymbol{\theta}_t + \nu_{jt}, \qquad \nu_{jt} \sim N(0, v_{jt}), \tag{5.1}$$

$$\boldsymbol{\theta}_{jt} = \boldsymbol{G}_{jt}\boldsymbol{\theta}_{j,t-1} + \boldsymbol{\omega}_{jt}, \qquad \boldsymbol{\omega}_{jt} \sim N(\boldsymbol{0}, \boldsymbol{W}_{jt}), \tag{5.2}$$

$$\boldsymbol{\theta}_t = \boldsymbol{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \qquad \boldsymbol{\omega}_t \sim N(\boldsymbol{0}, \boldsymbol{W}_t), \tag{5.3}$$

over $t = 1, 2, \ldots$. The elements here– and some relevant terminology and assumptions– are as follows.

- $\boldsymbol{\theta}_{jt}$ is the p_j-dimensional state vector specific to series j, referred to as the series$-j$ individual or idiosyncratic state vector.

- $\boldsymbol{F}_{jt}$ is the p_j-dimensional idiosyncratic regression vector for series j, known at time t.

- ν_{jt} is the series j observation error at time t having variance v_{jt}; these errors are independent over series j and across time t, with $\nu_{it} \perp\!\!\!\perp \nu_{js}$ for all series $i \neq j$ and all times t, s, and are also conditionally independent of current and all past state vectors.

- $\boldsymbol{G}_{jt}$ is the known $p_j \times p_j$ idiosyncratic state evolution matrix for series j at time t.

- $\boldsymbol{\omega}_{jt}$ is a normally distributed, p_j-dimensional idiosyncratic evolution noise vector with variance matrix $\boldsymbol{W}_{jt}$ known at time t; these evolution error terms are independent over time, with $\boldsymbol{\omega}_{it} \perp\!\!\!\perp \boldsymbol{\omega}_{js}$ for all series $i \neq j$ and times $t \neq s$, and are also conditionally independent of current and past observation errors.

- $\boldsymbol{H}_{jt}$ is a $p-$dimensional regression vector for series j, known at time t; some of the elements may be the same across series $j = 1 : m$ while others may differ, but the elements have the same interpretations (and scales) across series.

- $\boldsymbol{\theta}_t$ is the common (across series $j = 1 : m$) $p-$dimensional state vector defining the regression parameters to link $\boldsymbol{H}_{jt}$ to y_{jt} at time t.

- $\boldsymbol{G}_t$ is the known, $p \times p$ common state evolution matrix for the time variation in $\boldsymbol{\theta}_t$; for our current development, I assume $\boldsymbol{G}_t$ is non-singular for all t, as is the case in most common model contexts (West and Harrison, 1997; Prado and West, 2010).

- $\boldsymbol{\omega}_t$ is a normally distributed, $p-$dimensional common state evolution noise vector; with variance matrix $\boldsymbol{W}_t$ known at time t; these common evolution error terms are independent over time, with $\boldsymbol{\omega}_t \perp\!\!\!\perp \boldsymbol{\omega}_s$ for all $t \neq s$, and are also conditionally independent of current and past observation errors.

The full specification of the model requires details of the multivariate normal distributions for the full set of state evolution errors $\boldsymbol{\omega}_{1:m,t},\ \boldsymbol{\omega}_t$, which are central and critical to the modeling innovations here. This is also coupled with structured forms of prior/posterior distributions for the full set of state vectors, and these are now developed in the usual forward-sequential analysis context.

5.2.1 Conditional independence of idiosyncratic states given common states

Define $\boldsymbol{\Theta}_t$ to be the column vector stack of all state vectors– idiosyncratic and common– at each time t, and $\boldsymbol{\Omega}_t$ the corresponding column vector stack of all state evolution noise vectors. That is,

$$\boldsymbol{\Theta}_t = \begin{pmatrix} \boldsymbol{\theta}_{1t} \\ \vdots \\ \boldsymbol{\theta}_{mt} \\ \boldsymbol{\theta}_t \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Omega}_t = \begin{pmatrix} \boldsymbol{\omega}_{1t} \\ \vdots \\ \boldsymbol{\omega}_{mt} \\ \boldsymbol{\omega}_t \end{pmatrix},$$

each of dimension $p + \sum_{j=1:m} p_j$.

Suppose I am standing at time $t-1$ and have available all past data $\boldsymbol{y}_{1:t-1}$. Assume the following time $t-1$ posterior for the full set of states.

Assumption 1. Assume that

$$p(\boldsymbol{\Theta}_{t-1}|\boldsymbol{y}_{1:t-1}) = p(\boldsymbol{\theta}_{t-1}|\boldsymbol{y}_{1:t-1}) \prod_{j=1:m} p(\boldsymbol{\theta}_{j,t-1}|\boldsymbol{\theta}_{t-1}, \boldsymbol{y}_{1:t-1}) \tag{5.4}$$

with the following components:

- $(\boldsymbol{\theta}_{t-1}|\boldsymbol{y}_{1:t-1}) \sim N(\boldsymbol{m}_{t-1}, \boldsymbol{C}_{t-1})$ for some known moments $\boldsymbol{m}_{t-1}, \boldsymbol{C}_{t-1}$.

- For each $j = 1 : m$,

$$(\boldsymbol{\theta}_{j,t-1}|\boldsymbol{\theta}_{t-1}, \boldsymbol{y}_{1:t-1}) \sim N(\boldsymbol{m}_{j,t-1} + \boldsymbol{B}_{j,t-1}(\boldsymbol{\theta}_{t-1} - \boldsymbol{m}_{t-1}), \boldsymbol{U}_{j,t-1}) \tag{5.5}$$

  for a known p_j-vector mean $\boldsymbol{m}_{j,t-1}$, a known $p_j \times p$ regression matrix $\boldsymbol{B}_{j,t-1}$, and a known conditional (or residual) $p_j \times p_j$ variance matrix $\boldsymbol{U}_{j,t-1}$.

Under Assumption 1, $p(\boldsymbol{\Theta}_t|\boldsymbol{y}_{1:t-1})$ is evidently multivariate normal. The marginal time $t-1$ posterior for each idiosyncratic $\boldsymbol{\theta}_{j,t-1}$ vector is

$$(\boldsymbol{\theta}_{j,t-1}|\boldsymbol{y}_{1:t-1}) \sim N(\boldsymbol{m}_{j,t-1}, \boldsymbol{C}_{j,t-1}) \quad \text{with} \quad \boldsymbol{C}_{j,t-1} = \boldsymbol{B}_{j,t-1}\boldsymbol{C}_{t-1}\boldsymbol{B}'_{j,t-1} + \boldsymbol{U}_{j,t-1}.$$

This uses an extension of the standard notation for parameters in DLMs (West and Harrison, 1997). Note, however, that the $\boldsymbol{\theta}_{j,t-1}$ are dependent across series; they are conditionally independent when conditioning on the common state vector, but there are marginal dependencies induced by that common linkage. Specifically, $C(\boldsymbol{\theta}_{i,t-1}, \boldsymbol{\theta}_{j,t-1}|\boldsymbol{y}_{1:t-1}) = \boldsymbol{B}_{i,t-1}\boldsymbol{C}_{t-1}\boldsymbol{B}'_{j,t-1}$ for each $i \neq j$. In full, the time $t-1$ posterior variance matrix is

$$V(\boldsymbol{\Theta}_{t-1}|\boldsymbol{y}_{1:t-1}) = \boldsymbol{B}_{t-1}\boldsymbol{C}_{t-1}\boldsymbol{B}'_{t-1} + \boldsymbol{U}_{t-1} \tag{5.6}$$

where

$$\boldsymbol{B}_{t-1} = \begin{pmatrix} \boldsymbol{B}_{1,t-1} \\ \vdots \\ \boldsymbol{B}_{m,t-1} \\ \boldsymbol{I} \end{pmatrix} \quad \text{and} \quad \boldsymbol{U}_{t-1} = \begin{pmatrix} \boldsymbol{U}_{1,t-1} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{U}_{2,t-1} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{U}_{m,t-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \end{pmatrix}. \tag{5.7}$$

The corresponding precision matrix is $V(\boldsymbol{\Theta}_{t-1}|\boldsymbol{y}_{1:t-1})$ with inverse given by the precision matrix

$$\begin{pmatrix} \boldsymbol{U}_{1,t-1}^{-1} & \boldsymbol{0} & \cdots & \boldsymbol{0} & -\boldsymbol{U}_{1,t-1}^{-1}\boldsymbol{B}_{1,t-1} \\ \boldsymbol{0} & \boldsymbol{U}_{2,t-1}^{-1} & \cdots & \boldsymbol{0} & -\boldsymbol{U}_{2,t-1}^{-1}\boldsymbol{B}_{2,t-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{U}_{m,t-1}^{-1} & -\boldsymbol{U}_{m,t-1}^{-1}\boldsymbol{B}_{m,t-1} \\ -\boldsymbol{B}_{1,t-1}'\boldsymbol{U}_{1,t-1}^{-1} & -\boldsymbol{B}_{2,t-1}'\boldsymbol{U}_{2,t-1}^{-1} & \cdots & -\boldsymbol{B}_{m,t-1}'\boldsymbol{U}_{m,t-1}^{-1} & \boldsymbol{A}_{t-1} \end{pmatrix}$$

where

$$\boldsymbol{A}_{t-1} = V(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}_{1:m,t-1}, \boldsymbol{y}_{1:t-1})^{-1} = \boldsymbol{C}_{t-1}^{-1} + \sum_{j=1:m} \boldsymbol{B}_{j,t-1}'\boldsymbol{U}_{j,t-1}^{-1}\boldsymbol{B}_{j,t-1}.$$

This form of the precision matrix here is an immediate consequence of– and can be directly "read-off"– the conditional structure of the posterior. The precision neatly highlights the conditional independences inherent in the time $t-1$ posterior through the blocks of zeros.

5.2.2 Maintaining conditional independence through evolution

The basic idea should now be obvious: I aim to formulate the model so that I can run analytic sequential updating for each of the univariate series in parallel. I aim to do this for the DLMs *conditional* on the common state vector $\boldsymbol{\theta}_t$, and couple that with a marginal update for the common state vector following marginalization over the idiosyncratic states.

Since the time t observations y_{jt} are conditionally independent over series $j = 1:m$, then I will have independent updating of the idiosyncratic states *conditional* on the common state if I can maintain the conditional independence structure of equation (5.4) following evolution through equations (5.2) and (5.3). To do this requires coupling of the evolution error terms to induce the "right" form of contemporaneous, cross-series dependencies, and the critical– necessary and sufficient– structure for this is defined as follows.

Assumption 2. At time $t-1$, specify the variance matrix of the full set of evolution noise terms as

$$V(\boldsymbol{\Omega}_t|\boldsymbol{y}_{1:t-1}) = \boldsymbol{D}_t\boldsymbol{W}_t\boldsymbol{D}_t' + \boldsymbol{S}_t \tag{5.8}$$

where

$$\boldsymbol{D}_t = \begin{pmatrix} \boldsymbol{D}_{1t} \\ \vdots \\ \boldsymbol{D}_{m,t} \\ \boldsymbol{I} \end{pmatrix} \quad \text{and} \quad \boldsymbol{S}_t = \begin{pmatrix} \boldsymbol{S}_{1t} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{S}_{2t} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{S}_{mt} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \end{pmatrix} \tag{5.9}$$

with the following components:

- $\boldsymbol{W}_t$ is a $p \times p$ variance matrix.
- For each $j = 1:m$,
  - $\boldsymbol{S}_{jt}$ is a $p_j \times p_j$ variance matrix, and
  - $\boldsymbol{D}_{jt} = \boldsymbol{G}_{jt}\boldsymbol{B}_{j,t-1}\boldsymbol{G}_t^{-1}$.

Under Assumption 2, note the correspondence of the form of the distributions of the evolution error with that of the time $t-1$ state posterior. That is,

$$p(\boldsymbol{\Omega}_t|\boldsymbol{y}_{1:t-1}) = p(\boldsymbol{\omega}_t|\boldsymbol{y}_{1:t-1}) \prod_{j=1:m} p(\boldsymbol{\omega}_{jt}|\boldsymbol{\omega}_t, \boldsymbol{y}_{1:t-1})$$

with the following components:

- $(\boldsymbol{\omega}_t|\boldsymbol{y}_{1:t-1}) \sim N(\mathbf{0}, \boldsymbol{W}_t)$.
- For each $j = 1:m$, idiosyncratic evolution noise terms depend on the common evolution noise term via $(\boldsymbol{\omega}_{jt}|\boldsymbol{\omega}_t, \boldsymbol{y}_{1:t-1}) \sim N(\boldsymbol{D}_{jt}\boldsymbol{\omega}_t, \boldsymbol{S}_{jt})$, while being conditionally independent.
- The resulting margins for idiosyncratic evolution noise terms are $(\boldsymbol{\omega}_{jt}|\boldsymbol{y}_{1:t-1}) \sim N(\mathbf{0}, \boldsymbol{W}_{jt})$ with $\boldsymbol{W}_{jt} = \boldsymbol{D}_{jt}\boldsymbol{W}_t\boldsymbol{D}_{jt}' + \boldsymbol{S}_{jt}$, for each $j = 1:m$.

I am now in a position to evolve the full state from the time $t-1$ posterior $p(\boldsymbol{\Theta}_{t-1}|\boldsymbol{y}_{1:t-1})$ to the implied time t prior $p(\boldsymbol{\Theta}_t|\boldsymbol{y}_{1:t-1})$. The result is as follows.

Theorem 5.2.1. *Under Assumptions 1 and 2, the time t prior for the full state vector maintains the conditional independence structure and has pdf*

$$p(\boldsymbol{\Theta}_t|\boldsymbol{y}_{1:t-1}) = p(\boldsymbol{\theta}_t|\boldsymbol{y}_{1:t-1}) \prod_{j=1:m} p(\boldsymbol{\theta}_{jt}|\boldsymbol{\theta}_t, \boldsymbol{y}_{1:t-1}) \tag{5.10}$$

with the following components:

- $(\boldsymbol{\theta}_t|\boldsymbol{y}_{1:t-1}) \sim N(\boldsymbol{a}_t, \boldsymbol{R}_t)$ *with* $\boldsymbol{a}_t = \boldsymbol{G}_t\boldsymbol{m}_{t-1}$ *and* $\boldsymbol{R}_t = \boldsymbol{G}_t\boldsymbol{C}_{t-1}\boldsymbol{G}_t' + \boldsymbol{W}_t$.
- *For each* $j = 1:m$,

$$(\boldsymbol{\theta}_{jt}|\boldsymbol{\theta}_t, \boldsymbol{y}_{1:t-1}) \sim N(\boldsymbol{a}_{jt} + \boldsymbol{D}_{jt}(\boldsymbol{\theta}_t - \boldsymbol{a}_t), \boldsymbol{V}_{jt}), \tag{5.11}$$

  with

  - *marginal mean vectors* $\boldsymbol{a}_{jt} = \boldsymbol{G}_{jt}\boldsymbol{m}_{j,t-1}$;
  - *residual variance matrices* $\boldsymbol{V}_{jt} = \boldsymbol{S}_{jt} + \boldsymbol{G}_{jt}\boldsymbol{U}_{j,t-1}\boldsymbol{G}_{jt}'$, *and*
  - *regression matrices* $\boldsymbol{D}_{jt} = \boldsymbol{G}_{jt}\boldsymbol{B}_{j,t-1}\boldsymbol{G}_t^{-1}$ *as already defined.*

The corresponding variance matrix on the full set of states $\boldsymbol{\Theta}_t$ is

$$V(\boldsymbol{\Theta}_t|\boldsymbol{y}_{1:t-1}) = \boldsymbol{D}_t\boldsymbol{R}_t\boldsymbol{D}_t' + \boldsymbol{V}_t \tag{5.12}$$

with $\boldsymbol{D}_t$, $\boldsymbol{R}_t$ already defined and

$$\boldsymbol{V}_t = \begin{pmatrix} \boldsymbol{V}_{1t} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{V}_{2t} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{V}_{mt} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Note that the implied marginal priors for individual state vectors are $(\boldsymbol{\theta}_{jt}|\boldsymbol{y}_{1:t-1}) \sim N(\boldsymbol{a}_{jt}, \boldsymbol{R}_{jt})$ *where* $\boldsymbol{R}_{jt} = \boldsymbol{D}_{jt}\boldsymbol{R}_t\boldsymbol{D}'_{jt} + \boldsymbol{V}_{jt}$. *Marginal dependencies induced by the common state are reflected in covariance matrices* $C(\boldsymbol{\theta}_{it}, \boldsymbol{\theta}_{jt}|\boldsymbol{y}_{1:t-1}) = \boldsymbol{D}_{it}\boldsymbol{R}_t\boldsymbol{D}'_{jt}$ *for each* $i \neq j$.

Proof of Theorem 5.2.1. From the model specification it is immediate that the time t state prior is multivariate normal. I have $\boldsymbol{\Theta}_t = \boldsymbol{G}_{*t}\boldsymbol{\Theta}_{t-1} + \boldsymbol{\Omega}_t$ where $\boldsymbol{G}_{*t} = \text{diag}[\boldsymbol{G}_{1t}, \ldots, \boldsymbol{G}_{mt}, \boldsymbol{G}_t]$. Under Assumption 1, $(\boldsymbol{\Theta}_{t-1}|\boldsymbol{y}_{1:t-1})$ is normal with mean vector that is the vector stack of the $\boldsymbol{m}_{1:m,t}$ and $\boldsymbol{m}_t$, and variance matrix in equation (5.6). Under Assumption 2, $\boldsymbol{\Omega}_t$ is zero mean normal with variance matrix in equation (5.8). Substituting these components into the implied mean vector $E(\boldsymbol{\Theta}_t|\boldsymbol{y}_{1:t-1}) = \boldsymbol{G}_{*t}E(\boldsymbol{\Theta}_{t-1}|\boldsymbol{y}_{1:t-1})$ and variance matrix

$$V(\boldsymbol{\Theta}_t|\boldsymbol{y}_{1:t-1}) = \boldsymbol{G}_{*t}V(\boldsymbol{\Theta}_{t-1}|\boldsymbol{y}_{1:t-1})\boldsymbol{G}'_{*t} + V(\boldsymbol{\Omega}_t|\boldsymbol{y}_{1:t-1})$$

yields, after some linear algebra whose details can be found in Appendix 5.A, the moments as detailed in the theorem statement. The structured form of the variance matrix evidently corresponds precisely to the conditional linear regressions of each $\boldsymbol{\theta}_{jt}$ on $\boldsymbol{\theta}_t$ and the associated conditional independences. □

Corollary 5.2.1. *Assuming* $\boldsymbol{F}_{jt}$ *is known at time* $t-1$, *the joint predictive 1-step-ahead mean and variance* $E(\boldsymbol{y}_t|\mathcal{D}_{t-1})$ *and* $V(\boldsymbol{y}_t|\mathcal{D}_{t-1})$, *denoted by* $\boldsymbol{f}_t$ *and* $\boldsymbol{Q}_t$, *are*

$$\begin{aligned} \boldsymbol{f}_t &= E(\boldsymbol{y}_t|\mathcal{D}_{t-1}) = \boldsymbol{\Pi}_t\boldsymbol{a}^*_t, \\ \boldsymbol{Q}_t &= V(\boldsymbol{y}_t|\mathcal{D}_{t-1}) = \boldsymbol{\Pi}_tV(\boldsymbol{\Theta}_t|\mathcal{D}_{t-1})\boldsymbol{\Pi}'_t + \boldsymbol{N}_t, \end{aligned}$$

where

$$\boldsymbol{\Pi}_t = \begin{pmatrix} \boldsymbol{F}'_{1t} & 0 & \cdots & 0 & \boldsymbol{H}'_{1t} \\ 0 & \boldsymbol{F}'_{2t} & \cdots & 0 & \boldsymbol{H}'_{2t} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{F}'_{mt} & \boldsymbol{H}'_{mt} \end{pmatrix}, \tag{5.13}$$

and $\boldsymbol{a}_t^{'*} = (\boldsymbol{a}_{1t}', \boldsymbol{a}_{2t}', \cdots, \boldsymbol{a}_{mt}', \boldsymbol{a}_t')$, $\boldsymbol{N}_t = diag(v_{1t}, \cdots, v_{mt})$. *Specifically,*

$$\begin{aligned} E(y_{jt}|\mathcal{D}_{t-1}) &= \boldsymbol{F}_{jt}'\boldsymbol{a}_{jt} + \boldsymbol{H}_{jt}'\boldsymbol{a}_t. \\ V(y_{jt}|\mathcal{D}_{t-1}) &= (\boldsymbol{F}_{jt}'\boldsymbol{D}_{jt} + \boldsymbol{H}_{jt}')\boldsymbol{R}_t(\boldsymbol{D}_{jt}'\boldsymbol{F}_{jt} + \boldsymbol{H}_{jt}) + \boldsymbol{F}_{jt}'\boldsymbol{V}_{jt}\boldsymbol{F}_{jt} + v_{jt}. \end{aligned} \tag{5.14}$$

5.2.3 Conditional independence through updates

The key implication of the above is the following.

Theorem 5.2.2. *Under Assumptions 1 and 2, the update to the time t posterior for the full state vector maintains the conditional independence structure and normal form of the prior.*

Proof and details of Theorem 5.2.2. First, conditional on $\boldsymbol{\theta}_t$, marginalization of $p(\boldsymbol{y}_t|\boldsymbol{\Theta}_t)$ over the idiosyncratic states $\boldsymbol{\theta}_{1:m,t}$ with respect to the *conditional* prior $\prod_{j=1:m} p(\boldsymbol{\theta}_{jt}|\boldsymbol{\theta}_t, \boldsymbol{y}_{1:t-1})$ leads to a multivariate normal $p(\boldsymbol{y}_t|\boldsymbol{\theta}_t, \boldsymbol{y}_{1:t-1})$ with a mean vector that is a linear regression on $\boldsymbol{\theta}_t$ and a variance matrix that does not depend on $\boldsymbol{\theta}_t$. Hence, the normal prior $(\boldsymbol{\theta}_t|\boldsymbol{y}_{1:t-1}) \sim N(\boldsymbol{a}_t, \boldsymbol{R}_t)$ is updated to a normal posterior $(\boldsymbol{\theta}_t|\boldsymbol{y}_{1:t}) \sim N(\boldsymbol{m}_t, \boldsymbol{C}_t)$ with easily computed moments.

Second, conditional on $\boldsymbol{\theta}_t$, for each $j = 1 : m$ I have conditionally independent, normal priors $p(\boldsymbol{\theta}_{jt}|\boldsymbol{\theta}_t, \boldsymbol{y}_{1:t-1})$ with linear regression on $\boldsymbol{\theta}_t$ and constant residual variances. Also conditional on $\boldsymbol{\theta}_t$, the data break down into conditionally independent observations y_{jt} that, from the models of equation (5.1), define m component likelihood functions in parallel. The linear, normal forms of these in each $\boldsymbol{\theta}_{jt}$ lead to standard normal:normal updates to conditionally independent posteriors $p(\boldsymbol{\theta}_{jt}|\boldsymbol{\theta}_t, \boldsymbol{y}_{1:t-1}, y_{jt})$, $j = 1 : m$. Each of these is multivariate normal with linear regression on the common state $\boldsymbol{\theta}_t$ defining the mean vector, and a variance matrix that depends on j, t but not $\boldsymbol{\theta}_t$. That is, independent conditional posteriors $(\boldsymbol{\theta}_{jt}|\boldsymbol{\theta}_t, \boldsymbol{y}_{1:t-1}, y_{jt})$ have the distribution $N(\boldsymbol{m}_{jt} + \boldsymbol{B}_{jt}(\boldsymbol{\theta}_t - \boldsymbol{m}_t), \boldsymbol{U}_{jt})$ with easily computed elements $\boldsymbol{m}_{jt}, \boldsymbol{B}_{jt}$ and $\boldsymbol{U}_{jt}$. The result follows.

$\square$

Corollary 5.2.2. *Assuming $\boldsymbol{F}_{jt}$ is known at time $t-1$, the form of $\boldsymbol{m}_t$ and $\boldsymbol{C}_t$ are given below. With terms defined in Corollary 5.2.1, establish the joint distribution:*

$$\begin{pmatrix}\boldsymbol{\theta}_t \\ \boldsymbol{y}_t\end{pmatrix} |\mathcal{D}_{t-1} \sim N\left(\begin{pmatrix}\boldsymbol{a}_t \\ \boldsymbol{f}_t\end{pmatrix}, \quad \begin{pmatrix}\boldsymbol{R}_t & \boldsymbol{R}_{\theta yt} \\ \boldsymbol{R}'_{\theta yt} & \boldsymbol{Q}_t\end{pmatrix}\right).$$

where $\boldsymbol{R}_{\theta yt} = C(\boldsymbol{\theta}_t, \boldsymbol{y}_t|\mathcal{D}_{t-1}) \triangleq (\boldsymbol{R}_{1\theta yt}, \cdots, \boldsymbol{R}_{m\theta yt})$. It follows that

$$\boldsymbol{R}_{j\theta yt} = C(\boldsymbol{\theta}_t, y_{jt}|\mathcal{D}_{t-1}) = \boldsymbol{R}_t(\boldsymbol{D}'_{jt}\boldsymbol{F}_{jt} + \boldsymbol{H}_{jt}).$$

The standard normal theory immediately gives the required conditional distribution

$$\left(\boldsymbol{\theta}_t|\boldsymbol{y}_t, \mathcal{D}_{t-1}\right) \sim N\left(\boldsymbol{m}_t, \boldsymbol{C}_t\right),$$

where $\boldsymbol{m}_t$ and $\boldsymbol{C}_t$ are as follows:

$$\begin{aligned}\boldsymbol{m}_t &= \boldsymbol{a}_t + \boldsymbol{R}_{\theta yt}\boldsymbol{Q}_t^{-1}(\boldsymbol{y}_t - \boldsymbol{f}_t), \\ \boldsymbol{C}_t &= \boldsymbol{R}_t - \boldsymbol{R}_{\theta yt}\boldsymbol{Q}_t^{-1}\boldsymbol{R}'_{\theta yt}\end{aligned} \tag{5.15}$$

Theorem 5.2.3. *Assume that at time $t = 0$ the initial state distribution $p(\boldsymbol{\Theta}_0|\mathcal{D}_0)$ has the form defined by Assumption 1 at $t = 0$ (and by convention take $\boldsymbol{y}_{1:t-1}$ at $t = 0$ to be all prior information at that time.) Then the forms of time $t-1$ priors (of Assumption 1, Theorem 5.2.1) and prior-to-posterior updating over all times $t > 0$ (of Theorem 5.2.2) follow. Specifically, $\boldsymbol{m}_{jt}, \boldsymbol{B}_{jt}$ and $\boldsymbol{U}_{jt}$ are as follows:*

$$\begin{aligned}\boldsymbol{U}_{jt} &= (\boldsymbol{F}_{jt}\boldsymbol{F}'_{jt}/v_{jt} + \boldsymbol{V}_{jt}^{-1})^{-1}, \\ \boldsymbol{B}_{jt} &= \boldsymbol{U}_{jt}(\boldsymbol{V}_{jt}^{-1}\boldsymbol{D}_{jt} - \boldsymbol{F}_{jt}\boldsymbol{H}'_{jt}/v_{jt}), \\ \boldsymbol{m}_{jt} &= \boldsymbol{B}_{jt}\boldsymbol{m}_t + \boldsymbol{U}_{jt}(y_{jt}/v_{jt}\boldsymbol{F}_{jt} + \boldsymbol{V}_{jt}^{-1}(\boldsymbol{a}_{jt} - \boldsymbol{D}_{jt}\boldsymbol{a}_t)).\end{aligned}$$

Proof of Theorem 5.2.3. Details can be found in Appendix 5.A. $\square$

Theorem 5.2.3 indicates that I can run parallel, conditionally independent sequential learning on all idiosyncratic states in a closed form, Kalman filter-style format, coupled with corresponding closed-form filtering on the marginal prior:posterior distributions for the common states.

5.2.4 Filtering recurrences

Retrospective analysis is of importance for dynamic models widely applied in various situations. The goal here is to make inference about past state vectors of time t based on data before, at, and after time t, i.e. the smoothed distributions of both common state vector $\boldsymbol{\theta}_{t-k}|\mathcal{D}_t$ and series-specific state vector $\boldsymbol{\theta}_{j,t-k}|\mathcal{D}_t$, $t = 1, \cdots, T, j = 1, \cdots, m$. This section concentrates on the derivation of such retrospective distributions and sequential updating procedures. Using the standard theory of multivariate normal conditional distributions (Gamerman and Migon, 1993; West and Harrison, 1997), the main result is as the following.

Theorem 5.2.4. *Define*

$$\boldsymbol{A}_t = \boldsymbol{G}_{t+1}\boldsymbol{C}_t, \text{ and } \boldsymbol{A}_{jt} = \boldsymbol{G}_{j,t+1}\boldsymbol{C}_{jt}, j = 1, \cdots, m.$$

Then for all k, $(1 \leqslant k < t)$ the filtered marginal distributions are

$$(\boldsymbol{\theta}_{t-k}|\mathcal{D}_t) \sim N[\boldsymbol{a}_t(-k), \boldsymbol{R}_t(-k)] \text{ and } (\boldsymbol{\theta}_{j,t-k}|\mathcal{D}_t) \sim N[\boldsymbol{a}_{jt}(-k), \boldsymbol{R}_{jt}(-k)]$$

where

$$\boldsymbol{a}_t(-k) = \boldsymbol{m}_{t-k} + \boldsymbol{A}_t'\boldsymbol{R}_t^{-1}[\boldsymbol{a}_t(-k+1) - \boldsymbol{a}_{t-k+1}],$$
$$\boldsymbol{a}_{jt}(-k) = \boldsymbol{m}_{j,t-k} + \boldsymbol{A}_{jt}'\boldsymbol{R}_{jt}^{-1}[\boldsymbol{a}_{jt}(-k+1) - \boldsymbol{a}_{j,t-k+1}],$$

and

$$\boldsymbol{R}_t(-k) = \boldsymbol{C}_{t-k} + \boldsymbol{A}_t'\boldsymbol{R}_t^{-1}[\boldsymbol{R}_t(-k+1) - \boldsymbol{R}_{t-k+1}]\boldsymbol{R}_t^{-1}\boldsymbol{A}_t,$$
$$\boldsymbol{R}_{jt}(-k) = \boldsymbol{C}_{j,t-k} + \boldsymbol{A}_{jt}'\boldsymbol{R}_{jt}^{-1}[\boldsymbol{R}_{jt}(-k+1) - \boldsymbol{R}_{j,t-k+1}]\boldsymbol{R}_{jt}^{-1}\boldsymbol{A}_{jt}$$

with starting values

$$\boldsymbol{a}_t(0) = \boldsymbol{m}_t, \quad \boldsymbol{a}_{jt}(0) = \boldsymbol{m}_{jt},$$
$$\boldsymbol{R}_t(0) = \boldsymbol{C}_t, \quad \boldsymbol{R}_{jt}(0) = \boldsymbol{C}_{jt}.$$

and where as usual,

$$\boldsymbol{a}_{t-k}(1) = \boldsymbol{a}_{t-k+1}, \quad \boldsymbol{a}_{j,t-k}(0) = \boldsymbol{a}_{j,t-k+1},$$
$$\boldsymbol{R}_{t-k}(1) = \boldsymbol{R}_{t-k+1}, \quad \boldsymbol{R}_{j,t-k}(1) = \boldsymbol{R}_{j,t-k+1}.$$

Proof of Theorem 5.2.4. The proof is similar to West and Harrison (1997, p. 114-116). □

5.2.5 Efficient MCMC: FFBS algorithm

I derive an important backward sampling theorem which further leads to an efficient MCMC method in sampling the key conditional distribution of the states $p(\boldsymbol{\Theta}_t|\mathcal{D}_T)$. Carter and Kohn (1994) and Frühwirth-Schnatter (1994) provide the basic and original development of the renowned FFBS algorithm. I construct a modified FFBS algorithm for CSCDLM utilizing the conditional independent structure of the idiosyncratic states given the common states, which yields a compositional decoupling allowing us to sequentially sample the set of series-specific states in parallel.

Theorem 5.2.5 (Backward-Sampling)**.**

$$p(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{1:m,t}|\boldsymbol{\theta}_{t+1}, \boldsymbol{\theta}_{1:m,t+1}, \mathcal{D}_t) = p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, \mathcal{D}_t)\prod_{j=1}^{m} p(\boldsymbol{\theta}_{jt}|\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}, \boldsymbol{\theta}_{j,t+1}, \mathcal{D}_t),$$

where

$$(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, \mathcal{D}_t) \sim N[\boldsymbol{h}_t(-1), \boldsymbol{H}_t(-1)] \text{ and } (\boldsymbol{\theta}_{jt}|\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}, \boldsymbol{\theta}_{j,t+1}, \mathcal{D}_t) \sim N[\boldsymbol{h}_{jt}(-1), \boldsymbol{H}_{jt}(-1)]$$

with

$$\boldsymbol{h}_t(-1) = \boldsymbol{m}_t + \boldsymbol{A}_t'\boldsymbol{R}_{t+1}^{-1}(\boldsymbol{\theta}_{t+1} - \boldsymbol{a}_{t+1}),$$
$$\boldsymbol{H}_t(-1) = \boldsymbol{C}_t - \boldsymbol{A}_t'\boldsymbol{R}_{t+1}^{-1}\boldsymbol{A}_t,$$
$$\boldsymbol{h}_{jt}(-1) = \boldsymbol{G}_{j,t+1}^{-1}\boldsymbol{\theta}_{j,t+1} - \boldsymbol{B}_{jt}\boldsymbol{G}_{t+1}^{-1}\boldsymbol{\theta}_{t+1} + \boldsymbol{B}_{jt}\boldsymbol{\theta}_t,$$
$$\boldsymbol{H}_{jt}(-1) = \boldsymbol{G}_{j,t+1}^{-1}(\boldsymbol{S}_{j,t+1} - \boldsymbol{S}_{j,t+1}\boldsymbol{V}_{j,t+1}^{-1}\boldsymbol{S}_{j,t+1})\boldsymbol{G}_{j,t+1}'^{-1}.$$

Proof of Theorem 5.2.5. Details can be found in Appendix 5.A.

Specifically in the case of $\boldsymbol{S}_{j,t+1} = \boldsymbol{G}_{j,t+1}\boldsymbol{U}_{jt}\boldsymbol{G}'_{j,t+1}(\delta_j^{-1} - 1)$, I have $\boldsymbol{H}_{jt}(-1) = (1 - \delta_j)\boldsymbol{U}_{jt}$. □

5.2.6 Practically central special cases

Discount specifications

While there are many forms of discount specification of state evolution variances, the simplest and natural specifications here provide a first example. Given a common state discount factor $\delta \in (0, 1)$ and a set of idiosyncratic state discount factors $\delta_j \in (0, 1),\ j = 1 : m$, the relevant specifications are as follows:

- $\boldsymbol{W}_t = \boldsymbol{G}_t\boldsymbol{C}_{t-1}\boldsymbol{G}'_t(\delta^{-1} - 1)$ so that $\boldsymbol{R}_t = \boldsymbol{G}_t\boldsymbol{C}_{t-1}\boldsymbol{G}'_t/\delta$.
- For each $j = 1 : m$, $\boldsymbol{S}_{jt} = \boldsymbol{G}_{jt}\boldsymbol{U}_{j,t-1}\boldsymbol{G}'_{jt}(\delta_j^{-1} - 1)$ so that $\boldsymbol{V}_{jt} = \boldsymbol{G}_{jt}\boldsymbol{U}_{j,t-1}\boldsymbol{G}'_{jt}/\delta_j$.

Random walk evolutions

Many practical models have $\boldsymbol{G}_t = \boldsymbol{I}$ and/or, for each $j = 1 : m$, $\boldsymbol{G}_{jt} = \boldsymbol{I}$ for all t. Assuming all $\boldsymbol{G}_t, \boldsymbol{G}_{jt}$ are identity and under the discount specifications of variances, I have particular simplifications to the evolution of time $t - 1$ posterior to time t prior; that is, the key matrices reduce to

$$\boldsymbol{D}_{jt} = \boldsymbol{B}_{j,t-1} \quad \text{and} \quad \boldsymbol{V}_{jt} = \boldsymbol{U}_{j,t-1}/\delta_j.$$

5.3 Extension: Common State Coupled MDM

One of the important assumptions previous sections depend on is that the observational noise vector ν_{jt} is independent across series j, i.e. $\boldsymbol{\nu}_t$ has a diagonal covariance matrix. Another assumption related to this is that the regression vector $\boldsymbol{F}_{jt}$ is known at $t-1$, for all $j = 1, \cdots, m$, in order to obtain closed-form analytics of 1-step-ahead forecast of $\boldsymbol{y}_t$ at time $t - 1$. In fact, in this case the CSCDLM is a special class of dynamic hierarchical models (Gamerman and Migon, 1993) with the following structure:

$$\boldsymbol{y}_t = \boldsymbol{\Pi}_t \boldsymbol{\Theta}_t + \boldsymbol{\nu}_t, \quad \boldsymbol{\nu}_t \sim N(\boldsymbol{0}, \boldsymbol{\Lambda}_t^{-1}),$$

$$\boldsymbol{\Theta}_t = \boldsymbol{D}_t \boldsymbol{\theta}_t + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim N(\boldsymbol{0}, \boldsymbol{V}_t),$$

$$\boldsymbol{\theta}_t = \boldsymbol{G}_t \boldsymbol{\theta}_{t-1} + \boldsymbol{w}_t, \quad \boldsymbol{w}_t \sim N(\boldsymbol{0}, \boldsymbol{W}_t),$$

where $\boldsymbol{\Lambda}_t$ is the diagonal precision matrix for the observational noise vector $\boldsymbol{\nu}_t$. I now extend this model to multiregression dynamic models (MDM) by relaxing the two assumptions above to allow contemporaneous, volatile dependencies.

5.3.1 Compositional representation

The model with the relaxed assumption on observational volatilities can be written as

$$\boldsymbol{y}_t = \tilde{\boldsymbol{\Pi}}_t \tilde{\boldsymbol{\Theta}}_t + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_t), \tag{5.16}$$

where $\boldsymbol{\Sigma}_t$ is a volatility matrix,

$$\tilde{\boldsymbol{\Pi}}_t = \begin{pmatrix} \boldsymbol{x}_{1t}' & 0 & \cdots & 0 & \boldsymbol{H}_{1t}' \\ 0 & \boldsymbol{x}_{2t}' & \cdots & 0 & \boldsymbol{H}_{2t}' \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{x}_{mt}' & \boldsymbol{H}_{mt}' \end{pmatrix}, \quad \tilde{\boldsymbol{\Theta}}_t = \begin{pmatrix} \boldsymbol{\phi}_{1t} \\ \vdots \\ \boldsymbol{\phi}_{mt} \\ \boldsymbol{\theta}_t \end{pmatrix},$$

and $\boldsymbol{x}_{jt}$ is the exogenous idiosyncratic regression vector for series j. Factorize the precision matrices as

$$\boldsymbol{\Sigma}_t^{-1} = (\boldsymbol{I} - \boldsymbol{\Gamma}_t)' \boldsymbol{\Lambda}_t (\boldsymbol{I} - \boldsymbol{\Gamma}_t),$$

where

$$\boldsymbol{\Gamma}_t = \begin{pmatrix} 0 & \tilde{\gamma}_{12,t} & \tilde{\gamma}_{13,t} & \cdots & \tilde{\gamma}_{1m,t} \\ 0 & 0 & \tilde{\gamma}_{23,t} & \cdots & \tilde{\gamma}_{2m,t} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}$$

and

$$\boldsymbol{\Lambda}_t = \mathrm{diag}(1/v_{1t}, \ldots, 1/v_{mt})$$

with $v_{jt} > 0$ for each $j = 1 : m$. For notational convenience, set $\tilde{\gamma}_{jht} = 0$, for $h \notin pa(j)$, and $\tilde{\boldsymbol{\gamma}}_{j,pa(j),t} = \boldsymbol{\gamma}_{jt}$. The $pa(j)$ denotes the parental set of series j, and I have $pa(j) \subseteq \{j+1 : m\}$, for each j (and with $pa(m)$ empty). Then I can collect the effective coefficients $\boldsymbol{\gamma}_{jt}$ along with the implicit zero values in the matrix $\boldsymbol{\Gamma}_t$. The main conceptual and practical point is to properly identify $(\boldsymbol{I} - \boldsymbol{\Gamma}_t)$ as defining the time t relationships among the residuals in $\boldsymbol{\nu}_t$. Here $(\boldsymbol{I} - \boldsymbol{\Gamma}_t)$ is the (unit diagonal) Cholesky component of $\boldsymbol{\Sigma}_t^{-1}$. I then have the factor representation

$$\boldsymbol{\nu}_t = (\boldsymbol{I} - \boldsymbol{\Gamma}_t)\boldsymbol{\epsilon}_t \tag{5.17}$$

where $\boldsymbol{\nu}_t \sim N(\boldsymbol{0}, \boldsymbol{\Lambda}_t^{-1})$ i.e., having independent elements $\nu_{jt} \sim N(0, v_{jt}),\ j = 1 : m$.

Then, premultiplying equation (5.16) by $(\boldsymbol{I} - \boldsymbol{\Gamma}_t)$ and rearranging yields

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{\Gamma}_t \boldsymbol{y}_t + (\boldsymbol{I} - \boldsymbol{\Gamma}_t)\tilde{\boldsymbol{\Pi}}_t \tilde{\boldsymbol{\Theta}}_t + \boldsymbol{\nu}_t, \quad \boldsymbol{\nu}_t \sim N(\boldsymbol{0}, \boldsymbol{\Lambda}_t^{-1}) \\ &= \boldsymbol{\Gamma}_t(\boldsymbol{y}_t - \tilde{\boldsymbol{\Pi}}_t \tilde{\boldsymbol{\Theta}}_t) + \tilde{\boldsymbol{\Pi}}_t \tilde{\boldsymbol{\Theta}}_t + \boldsymbol{\nu}_t. \end{aligned} \tag{5.18}$$

I am essentially making the regression vector contain a random and unknown part, $\boldsymbol{y}_{pa(j),t}$, at time $t-1$, where $pa(j)$ is the specific subsets of compositional endogenous variables I consider. In terms of the univariate series this shows that, for $j = 1 : m-1$, I have DLMs conditional on all parent series at time t, as follows

$$\begin{aligned} y_{jt} &= \boldsymbol{H}'_{jt}\boldsymbol{\theta}_t + \boldsymbol{x}'_{jt}\boldsymbol{\phi}_{jt} + (\boldsymbol{y}_{pa(j),t} - \boldsymbol{H}'_{pa(j),t}\boldsymbol{\theta}_t - \sum_{k \in pa(j)} \boldsymbol{x}'_{kt}\boldsymbol{\phi}_{kt})'\boldsymbol{\gamma}_{jt} + \nu_{jt} \\ &= \boldsymbol{H}'_{jt}\boldsymbol{\theta}_t + \boldsymbol{x}'_{jt}\boldsymbol{\phi}_{jt} + \boldsymbol{\epsilon}'_{pa(j),t}\boldsymbol{\gamma}_{jt} + \nu_{jt}, \end{aligned} \tag{5.19}$$

where ϵ_{kt} is the actual observation errors corresponding to series k, i.e.

$$\epsilon_{kt} = y_{kt} - \boldsymbol{H}'_{kt}\boldsymbol{\theta}_t - \boldsymbol{x}'_{kt}\boldsymbol{\phi}_{kt}.$$

5.3.2 Decoupled univariate DLMs

Evidently at $j = m$ I have the original DLM for y_{mt} and $\nu_{mt} = \epsilon_{mt}$. For $j < m$, I have augmented DLMs in which the actual observation errors ϵ_{kt} of parents $k \in pa(j)$ of series j are theoretically implied as conditional predictors for y_{jt}. Hence, *if these values were known*, I would have standard DLMs: for each $j < m$, the series-specific regression vector is $(\boldsymbol{x}'_{jt}, \boldsymbol{\epsilon}'_{jt})'$ and the augmented series-specific state vector $\boldsymbol{\theta}_{jt}$ is $(\boldsymbol{\phi}'_{jt}, \boldsymbol{\gamma}'_{jt})'$. I can see how to immediately induce natural models of time variation for the cross-series dependency structure by using some form of DLM for each of the additional state vectors $\boldsymbol{\gamma}_{jt}$. Simple random walk models (using a discount formulation) are an obvious first step.

The immediate problem is that I cannot do this directly, since the ϵ_{kt} are unknown. This means that I must consider some form of approximation. I am not currently interested in computationally demanding sequential Monte Carlo (SMC) simulation-based methods (such as particle filtering); I may consider these approaches later as part of a broader research project and, perhaps, as an alternative approach to assess adequacy of more direct analytic approximations. First, the interest here is in more immediate, and simpler analytic approximations that can be implemented efficiently.

5.3.3 A first approximation/model reformulation

An obvious first approximation is to plug in conditional prior means $\boldsymbol{a}_t$ as estimates of the $\boldsymbol{\theta}_t$ and $\boldsymbol{a}_{k\phi t}$ as estimates of $\boldsymbol{\phi}_{kt}$ for the series-specific predictors in equation (5.19). Then, over $j = 1:m$, the coupled models are of the same form as (5.1). The full state and regression vectors specific to series j, each of dimension $p_j = p_{j\phi} + p_{j\gamma}$, are

$$\boldsymbol{\theta}_{jt} = \begin{pmatrix} \boldsymbol{\phi}_{jt} \\ \boldsymbol{\gamma}_{jt} \end{pmatrix} \quad \text{and} \quad \boldsymbol{F}_{jt} = \begin{pmatrix} \boldsymbol{x}_{jt} \\ \boldsymbol{y}_{pa(j),t} - \boldsymbol{H}'_{pa(j),t}\boldsymbol{a}_t - \sum_{k \in pa(j)} \boldsymbol{x}'_{kt}\boldsymbol{a}_{k\phi t} \end{pmatrix}.$$

In the special case of $\boldsymbol{x}_{jt} = \mathbf{0}$, the full series-specific state and regression vector is simplified to the following

$$\boldsymbol{\theta}_{jt} = \boldsymbol{\gamma}_{jt} \quad \text{and} \quad \boldsymbol{F}_{jt} = \boldsymbol{y}_{pa(j),t} - \boldsymbol{H}'_{pa(j),t}\boldsymbol{a}_t.$$

A key point to highlight here is that because of the contemporaneous dependence between y_{jt} and its parent $\boldsymbol{y}_{pa(j),t}$, the conditional and marginal distribution of $\boldsymbol{y}_t, \boldsymbol{\theta}_t | \mathcal{D}_{t-1}$ are as the following:

$$\begin{aligned}
p(\boldsymbol{\theta}_t | \mathcal{D}_{t-1}) &\sim \text{Normal distribution}, \\
p(\boldsymbol{y}_t | \mathcal{D}_{t-1}) &\sim \text{Non-normal distribution}, \\
p(\boldsymbol{y}_t | \boldsymbol{\theta}_t, \mathcal{D}_{t-1}) &\sim \text{Non-normal distribution}, \\
p(\boldsymbol{\theta}_t | \boldsymbol{y}_t, \mathcal{D}_{t-1}) &\sim \text{Normal distribution}.
\end{aligned}$$

5.3.4 A refinement

For future work, I can explore a modification that adjusts for the expected additional variation the above approximation induces. That is, a modification of the form

$$y_{jt} = \boldsymbol{H}'_{jt}\boldsymbol{\theta}_t + \boldsymbol{F}'_{jt}\boldsymbol{\theta}_{jt} + \xi_{jt} + \nu_{jt},$$

for each $j < m$, where I consider adding new, zero-mean error terms ξ_{jt} representing approximation error. We cannot do this exactly while also maintaining the closed-form analytic updating, but a simple approach that captures some of the quantitative error expected might be valuable. The "obvious" assumption to try would be to take the ξ_{jt} to be conditionally independent across series j (to maintain the decoupling) with zero means, and variances w_{jt}/v_{jt} for some chosen w_{jt} values (based on some approximate analytic reasoning). It is clear that some simple analytic approximations can be derived to generate "appropriate" values of w_{jt}, while the $1/v_{jt}$ terms mean that I maintain the conjugate updating.

5.3.5 An alternative model specification

An alternative model that *immediately* yields closed-form analytic updating in decoupled univariate DLMs is defined by applying the contemporaneous compositional regression to the data series directly, rather than to the observational errors. The latter is more natural, but this alternative approach should be considered as it has the benefit of generating the closed-form analysis without approximations. The practical performance, in terms of relative improvements in short-term forecasting and other metrics/utilities of interest, should be the guide to comparison and choice of approach, eventually. In the first stages, both approaches can be explored and compared as programming one basically defines code for both.

The specification simply replaces equation (5.18) with

$$\boldsymbol{y}_t = \boldsymbol{\Gamma}_t \boldsymbol{y}_t + \tilde{\boldsymbol{\Pi}}_t \tilde{\boldsymbol{\Theta}}_t + \boldsymbol{\nu}_t. \tag{5.20}$$

That is, there is now no $(\boldsymbol{I} - \boldsymbol{\Gamma}_t)$ premultiplying $\tilde{\boldsymbol{\Pi}}_t$ in the underlying dynamic regressions for each of the univariate series. Again, this maintains the original DLM for y_{mt} and then, for each $j > 1$, I have:

$$\begin{aligned} y_{jt} &= \sum_{k=j+1:m} \tilde{\gamma}_{jkt} y_{kt} + \boldsymbol{H}_{jt}\boldsymbol{\theta}_t + \boldsymbol{x}'_{jt}\boldsymbol{\theta}_{jt} + \nu_{jt}, \\ &= \boldsymbol{y}'_{j+1:m,t}\tilde{\boldsymbol{\gamma}}_{jt} + \boldsymbol{H}_{jt}\boldsymbol{\theta}_t + \boldsymbol{x}'_{jt}\boldsymbol{\theta}_{jt} + \nu_{jt}, \end{aligned} \tag{5.21}$$

where $\tilde{\boldsymbol{\gamma}}_{jt} = (\tilde{\gamma}_{j,j+1,t}, \ldots, \tilde{\gamma}_{jmt})'$ and $\boldsymbol{y}_{j+1:m,t} = (\boldsymbol{y}_{j+1,t}, \ldots, \boldsymbol{y}_{mt})'$. This is basically the same model but with parents y_{kt} replacing the (unknown) errors ϵ_{kt} in the augmented dynamic regressions. I would therefore expect to see more "action" in the $\tilde{\gamma}_{jkt}$ coefficients relative to the direct formulation above.

In this alternative model, I maintain the interpretation of $(\boldsymbol{I} - \boldsymbol{\Gamma}_t)$ as the (unit diagonal) Cholesky component of $\boldsymbol{\Omega}_t$, while the underlying dynamic regressions are changed. That is, I can deduce that

$$\boldsymbol{y}_t = (\boldsymbol{I} - \boldsymbol{\Gamma}_t)^{-1} \tilde{\boldsymbol{\Pi}}_t \tilde{\boldsymbol{\Theta}}_t + \boldsymbol{\nu}_t \tag{5.22}$$

showing the same form as the original model of equation (5.16) but with $\tilde{\mathbf{\Pi}}_t$ altered to $(\boldsymbol{I} - \mathbf{\Gamma}_t)^{-1}\tilde{\mathbf{\Pi}}_t$.

Write $\mathbf{\Psi}_t = (\boldsymbol{I} - \mathbf{\Gamma}_t)^{-1}$ so that $\mathbf{\Psi}_t$ is upper triangular with a unit diagonal and upper off-diagonal elements ψ_{jkt}. The latter have the interpretation as factor loadings: each ϵ_{jt} is a linear combination of the independent errors ν_{kt} with loadings (i.e., weights) ψ_{jkt} (and, of course, zero weights for $k \leqslant j$). The key/only difference between this model and the original is that the implied state vectors are modified as follows:

$$y_{jt} = \boldsymbol{H}'_{jt}\boldsymbol{\theta}^*_t + \boldsymbol{x}'_{jt}\boldsymbol{\theta}^*_{jt} + \nu_{jt}$$

where

$$\boldsymbol{\theta}^*_t = \boldsymbol{\theta}_t + \sum_{k=j+1:m} \psi_{jkt}\boldsymbol{\theta}_t \quad \text{and} \quad \boldsymbol{\theta}^*_{jt} = \boldsymbol{\theta}_{jt} + \sum_{k=j+1:m} \psi_{jkt}\boldsymbol{\theta}_{kt}.$$

This basic model has each series regressed on its parents, whereas what I am really trying to do with the multi-regression component is model co-volatilities. Therefore, I use the "obvious model reformation" in Section 5.3.3 for our application example where I use the prior (given $\mathcal{D}_{t-1}$) means of $\boldsymbol{\theta}_t$ to yield a DLM.

5.3.6 Forward-filtering, backward-sampling

Theorem 5.3.1. *Theorem 5.2.1, 5.2.2, 5.2.3, 5.2.4 still apply to CSCMDMs.*

Proof of Theorem 5.3.1. Therom 5.2.1, 5.2.3, 5.2.4 easily apply and thus I omit the proof. Now I prove that Theorem 5.2.2 also applies, simply from

$$\begin{aligned} p(\boldsymbol{\theta}_t|\mathcal{D}_t) &\propto p(\boldsymbol{\theta}_t|\mathcal{D}_{t-1})p(\boldsymbol{y}_t|\boldsymbol{\theta}_t, \mathcal{D}_{t-1}) \\ &\propto p(\boldsymbol{\theta}_t|\mathcal{D}_{t-1})\prod_{j=1}^{m} p(y_{jt}|\boldsymbol{y}_{pa(j),t}, \boldsymbol{\theta}_t, \mathcal{D}_{t-1}). \end{aligned}$$

Given the assumption of conditional independence $\boldsymbol{\theta}_{jt} \perp \boldsymbol{y}_{pa(j),t}|(\boldsymbol{\theta}_t, \mathcal{D}_{t-1})$, it follows that

$$p(y_{jt}|\boldsymbol{y}_{pa(j),t}, \boldsymbol{\theta}_t, \mathcal{D}_{t-1}) = \int p(y_{jt}|\boldsymbol{y}_{pa(j),t}, \boldsymbol{\theta}_{jt}, \boldsymbol{\theta}_t, \mathcal{D}_{t-1}) p(\boldsymbol{\theta}_{jt}|\boldsymbol{y}_{pa(j),t}, \boldsymbol{\theta}_t, \mathcal{D}_{t-1}) d\boldsymbol{\theta}_{jt}$$

$$= \int p(y_{jt}|\boldsymbol{y}_{pa(j),t}, \boldsymbol{\theta}_{jt}, \boldsymbol{\theta}_t, \mathcal{D}_{t-1}) p(\boldsymbol{\theta}_{jt}|\boldsymbol{\theta}_t, \mathcal{D}_{t-1}) d\boldsymbol{\theta}_{jt}.$$

Thus $(y_{jt}|\boldsymbol{y}_{pa(j),t}, \boldsymbol{\theta}_t, \mathcal{D}_{t-1})$ has a normal distribution

$$(y_{jt}|\boldsymbol{y}_{pa(j),t}, \boldsymbol{\theta}_t, \mathcal{D}_{t-1}) \sim N(\boldsymbol{H}'_{jt}\boldsymbol{\theta}_t + \boldsymbol{F}'_{jt}(\boldsymbol{a}_{jt} + \boldsymbol{D}_{jt}(\boldsymbol{\theta}_t - \boldsymbol{a}_t)), v_{jt} + \boldsymbol{F}'_{jt}\boldsymbol{V}_{jt}\boldsymbol{F}_{jt}).$$

Therefore, $(\boldsymbol{\theta}_t|\mathcal{D}_{t-1}) \sim N(\boldsymbol{a}_t, \boldsymbol{R}_t)$ is updated to posterior $(\boldsymbol{\theta}_t|\mathcal{D}_t) \sim N(\boldsymbol{m}_t, \boldsymbol{C}_t)$. Notice that since $\boldsymbol{y}_{pa(j),t}$ is a subvector of $\boldsymbol{F}_{jt}$, even though $(y_{jt}|\boldsymbol{y}_{pa(j),t}, \boldsymbol{\theta}_t, \mathcal{D}_{t-1})$ has a normal distribution, both its mean and variance involve $\boldsymbol{y}_{pa(j),t}$. This conditional dependence of y_{jt} and its parents breaks the joint normality of $\boldsymbol{y}_t|\mathcal{D}_{t-1}$. Due to the non-Gaussian distribution of $(\boldsymbol{y}_t|\mathcal{D}_{t-1})$, $\boldsymbol{m}_t$ and $\boldsymbol{C}_t$ no longer has the same form as (5.15). However, from the pdf of $\boldsymbol{\theta}_t$ and $\boldsymbol{y}_t$, I can obtain the form of $\boldsymbol{m}_t$ and $\boldsymbol{C}_t$ as follows:

$$\begin{aligned}
\boldsymbol{m}_t &= \boldsymbol{C}_t[\boldsymbol{R}_t^{-1}\boldsymbol{a}_t + \sum_{j=1}^{m} y_{jt}/v_{jt}(\boldsymbol{B}'_{jt}\boldsymbol{F}_{jt} + \boldsymbol{H}_{jt}) - (\boldsymbol{D}'_{jt} - \boldsymbol{B}'_{jt})\boldsymbol{V}_{jt}^{-1}(\boldsymbol{a}_{jt} - \boldsymbol{D}_{jt}\boldsymbol{a}_t)], \\
\boldsymbol{C}_t &= (\boldsymbol{R}_t^{-1} + \sum_{j=1}^{m}(\boldsymbol{N}_{jt} - \boldsymbol{B}'_{jt}\boldsymbol{U}_{jt}^{-1}\boldsymbol{B}_{jt}))^{-1},
\end{aligned} \tag{5.23}$$

where

$$\boldsymbol{N}_{jt} = \boldsymbol{H}_{jt}\boldsymbol{H}'_{jt}/v_{jt} + \boldsymbol{D}'_{jt}\boldsymbol{V}_{jt}^{-1}\boldsymbol{D}_{jt}.$$

Details of calculation are left in Appendix 5.A. □

To improve the computational efficiency of (5.23) by avoiding directly inverting $p \times p$ matrices, I can sequentially compute the conditional distributions of $(\boldsymbol{\theta}_t|\boldsymbol{y}_{j:m,t}, \mathcal{D}_{t-1})$ to obtain the required $\boldsymbol{m}_t$, $\boldsymbol{C}_t$. Assuming

$$(\boldsymbol{\theta}_t|\boldsymbol{y}_{j:m,t}, \mathcal{D}_{t-1}) \sim N(\boldsymbol{m}_t^{(j)}, \boldsymbol{C}_t^{(j)}), \text{ and } (y_{jt}|\boldsymbol{y}_{j+1:m,t}, \mathcal{D}_{t-1}) \sim N(f_{jt}^{(j+1)}, q_{jt}^{(j+1)}),$$

I proceed as follows.

i. Start at $j = m$: I have $f_{mt}^{(m+1)} = f_{mt}(\varnothing), q_{mt}^{(m+1)} = q_{mt}(\varnothing)$. The mean and variance of $(\boldsymbol{\theta}_t|\boldsymbol{y}_{mt}, \mathcal{D}_{t-1})$ are

$$\boldsymbol{m}_t^{(m)} = \boldsymbol{a}_t + \boldsymbol{R}_t(\boldsymbol{D}'_{mt}\boldsymbol{F}_{mt} + \boldsymbol{H}_{mt})(y_{mt} - f_{mt}^{(m+1)})/q_{mt}^{(m+1)},$$

$$\boldsymbol{C}_t^{(m)} = \boldsymbol{R}_t - \boldsymbol{R}_t(\boldsymbol{D}'_{mt}\boldsymbol{F}_{mt} + \boldsymbol{H}_{mt})(\boldsymbol{D}'_{mt}\boldsymbol{F}_{mt} + \boldsymbol{H}_{mt})'\boldsymbol{R}_t/q_{mt}^{(m+1)},$$

using the simplified forms of (5.14) with all terms in $\boldsymbol{y}_{pa(m),t}$, which is in $\boldsymbol{F}_{mt}$, set to zero.

ii. For $j = m-1, \cdots, 1$ in turn: First, the conditional mean $f_{jt}^{(j+1)}$ and variance $q_{jt}^{(j+1)}$ of $(y_{jt}|y_{j+1:m,t}, \mathcal{D}_{t-1})$ are computed as

$$f_{jt}^{(j+1)} = \boldsymbol{H}'_{jt}\boldsymbol{m}_t^{(j+1)} + \boldsymbol{F}'_{jt}[\boldsymbol{a}_{jt} + \boldsymbol{D}_{jt}(\boldsymbol{m}_t^{(j+1)} - \boldsymbol{a}_t)],$$

$$q_{jt}^{(j+1)} = \boldsymbol{H}'_{jt}\boldsymbol{C}_t^{(j+1)}\boldsymbol{H}_{jt} + \boldsymbol{F}'_{jt}(\boldsymbol{V}_{jt} + \boldsymbol{D}_{jt}\boldsymbol{C}_t^{(j+1)}\boldsymbol{D}'_{jt})\boldsymbol{F}_{jt} + 2\boldsymbol{H}'_{jt}\boldsymbol{C}_t^{(j+1)}\boldsymbol{D}'_{jt}\boldsymbol{F}_{jt} + v_{jt}.$$

Then establish the joint distribution of $\begin{pmatrix} \boldsymbol{\theta}_t \\ y_{jt} \end{pmatrix} |\boldsymbol{y}_{j+1:m,t}, \mathcal{D}_{t-1}$ as

$$N\left(\begin{pmatrix} \boldsymbol{m}_t^{(j+1)} \\ f_{jt}^{(j+1)} \end{pmatrix}, \begin{pmatrix} \boldsymbol{C}_t^{(j+1)} & \boldsymbol{C}_t^{(j+1)}(\boldsymbol{H}_{jt} + \boldsymbol{D}'_{jt}\boldsymbol{F}_{jt}) \\ (\boldsymbol{H}_{jt} + \boldsymbol{D}'_{jt}\boldsymbol{F}_{jt})'\boldsymbol{C}_t^{(j+1)} & q_{jt}^{(j+1)} \end{pmatrix}\right). \tag{5.24}$$

Then according to conditional normal theory, I have

$$\boldsymbol{m}_t^{(j)} = \boldsymbol{m}_t^{(j+1)} + \boldsymbol{C}_t^{(j+1)}(\boldsymbol{H}_{jt} + \boldsymbol{D}'_{jt}\boldsymbol{F}_{jt})(y_{jt} - f_{jt}^{(j+1)})/q_{jt}^{(j+1)},$$

$$\boldsymbol{C}_t^{(j)} = \boldsymbol{C}_t^{(j+1)} - \boldsymbol{C}_t^{(j+1)}(\boldsymbol{H}_{jt} + \boldsymbol{D}'_{jt}\boldsymbol{F}_{jt})(\boldsymbol{H}_{jt} + \boldsymbol{D}'_{jt}\boldsymbol{F}_{jt})'\boldsymbol{C}_t^{(j+1)}/q_{jt}^{(j+1)}.$$

iii. End: at this point, $j = 1$ and I have $\boldsymbol{m}_t^{(1)} = \boldsymbol{m}_t, \boldsymbol{C}_t^{(1)} = \boldsymbol{C}_t$.

Note that since $q_{jt}^{(j+1)}$ for $j = 1, \cdots, m$ are scalars, I manage to avoid any matrix inversions in the sequential algorithm above.

The complexity of this sequential algorithm is $O(p^2 \sum_{j=1}^m p_j)$ and that of direct inversion of $\boldsymbol{C}_t$ is $O(p^{2+\epsilon} + p\sum_{j=1}^m p_j^2), \epsilon > 0.3$. When p is large, $p^\epsilon > \sum_{j=1}^m p_j$, the sequential algorithms has less computing burden than that of the direct inversion. Otherwise the direct inversion should be adopted.

5.3.7 *Joint predictive mean, variance matrix and precision matrix*

I aim to obtain analytic forms of joint predictive 1-step-ahead mean and variance, denoted by

$$\boldsymbol{f}_t = E(\boldsymbol{y}_t|\mathcal{D}_{t-1}) \quad \text{and} \quad \boldsymbol{Q}_t = V(\boldsymbol{y}_t|\mathcal{D}_{t-1}).$$

Although the FFBS results hold for CSCMDM, the results of (5.14) on the joint predictive moments fail to apply to CSCMDM. Notice that the predictive distribution of $\boldsymbol{y}_t|\mathcal{D}_{t-1}$ is non-Gaussian as it involves distributions of polynomials of normal random vectors. Thus the analytic forms are ungracefully complicated.

Alternative methods include directly approximating these values by Monte Carlo simulations. Let nmc be the number of iterations. For $t = 2, \cdots, T$ and in the order of $j = m, m-1, \cdots, 1$, I draw i-th sample $(y_{jt}^{[i]}|\boldsymbol{y}_{j+1:m,t}^{[i]}, \mathcal{D}_{t-1})$ from the normal distribution specified in (5.24) for each $i = 1, \cdots, nmc$.

Instead of this direct simulation method, whose speed highly depends on computer capacities, I approximate the analytic form of $\boldsymbol{f}_t$ and $\boldsymbol{Q}_t$ by the following fast algorithm, which sequentially computes elements of the required $\boldsymbol{f}_t$ and $\boldsymbol{Q}_t$.

i. Start at $j = m$: compute the univariate mean ,variance of y_{mt}, i.e. f_{mt}, q_{mt}, using the simplified forms of (5.14) with all terms in $\boldsymbol{y}_{pa(m)}$, which is in $\boldsymbol{F}_{mt}$, set to zero. Insert f_{mt} as the 1st element of $\boldsymbol{f}_t$ and q_{mt} as the $(1,1)-$element of $\boldsymbol{Q}_t$.

ii. For $j = m-1, \cdots, 1$ in turn: denote the mean vector, variance matrix and precision matrix of the $(m-j)-$vector $\boldsymbol{y}_{j+1:m,t}$ by

$$\boldsymbol{f}_{j+1:m,t} = E(\boldsymbol{y}_{j+1:m,t}|\mathcal{D}_{t-1}) \quad \text{and} \quad \boldsymbol{Q}_{j+1:m,t} = V(\boldsymbol{y}_{j+1:m,t}|\mathcal{D}_{t-1}). \tag{5.25}$$

Note that at $j = m-1$ these reduce to scalars.

- At this point, I have already computed the values of the moments of (5.25) from the previous steps. These are used in the following calculations. First, write $\boldsymbol{f}_{pa(j),t}$ for the subvector of means in $\boldsymbol{f}_{j+1:m,t}$ on elements in $pa(j)$ only, and $\boldsymbol{Q}_{pa(j),t}$ for the corresponding submatrix of $\boldsymbol{Q}_{j+1:m,t}$. Then, the marginal mean f_{jt} and variance q_{jt} of y_{jt} are computed as follows.

$$f_{jt} = \boldsymbol{H}'_{jt}\boldsymbol{a}_t + \boldsymbol{x}'_{jt}\boldsymbol{a}_{j\phi t} + E(\begin{pmatrix} \boldsymbol{x}_{jt} \\ \boldsymbol{y}_{pa(j),t} \end{pmatrix}' \boldsymbol{D}_{jt}\boldsymbol{m}_t^{(j-1)}|\mathcal{D}_{t-1}) - \begin{pmatrix} \boldsymbol{x}_{jt} \\ \boldsymbol{f}_{pa(j),t} \end{pmatrix}' \boldsymbol{D}_{jt}\boldsymbol{a}_t \tag{5.26}$$

As the analytic form of $E(\begin{pmatrix} \boldsymbol{x}_{jt} \\ \boldsymbol{y}_{pa(j),t} \end{pmatrix}' \boldsymbol{D}_{jt}\boldsymbol{m}_t^{(j-1)}|\mathcal{D}_{t-1})$ is complex, I approximate it by $\begin{pmatrix} \boldsymbol{x}_{jt} \\ \boldsymbol{f}_{pa(j),t} \end{pmatrix}' \boldsymbol{D}_{jt}\boldsymbol{a}_t$, which is equivalent to assuming $\boldsymbol{y}_{pa(j),t} \perp \boldsymbol{m}_t^{(j-1)}|\mathcal{D}_{t-1}$ just for the purpose of this approximation, and similarly for the computation of the variance. The predictives then have the following approximate forms:

$$\begin{aligned} f_{jt} &= \boldsymbol{H}'_{jt}\boldsymbol{a}_t + \boldsymbol{x}'_{jt}\boldsymbol{a}_{j\phi t}, \\ q_{jt} &= v_{jt} + u_{jt} + 2\boldsymbol{H}'_{jt}\boldsymbol{R}_t\boldsymbol{D}'_{jt}\begin{pmatrix} \boldsymbol{x}_{jt} \\ \boldsymbol{0} \end{pmatrix} + \boldsymbol{H}'_{jt}\boldsymbol{R}_t\boldsymbol{H}_{jt} + \boldsymbol{a}'_{j\gamma t}\boldsymbol{Q}_{pa(j),t}\boldsymbol{a}_{j\gamma t} \end{aligned} \tag{5.27}$$

where

$$u_{jt} = \mathrm{tr}(\boldsymbol{R}_{j\gamma t}\boldsymbol{Q}_{pa(j),t}) + \boldsymbol{x}'_{jt}\boldsymbol{R}_{j\phi t}\boldsymbol{x}_{jt}.$$

Insert f_{jt} as the $j-$th element of $\boldsymbol{f}_t$ and q_{jt} as the $(j,j)-$element of $\boldsymbol{Q}_t$, respectively.

- Compute the covariance vector $C(y_{jt}, \boldsymbol{y}_{j+1:m,t}|\mathcal{D}_{t-1})$ as follows. Write $\boldsymbol{a}_{j+1:m\gamma t}$ for the $(j-1)-$vector that extends $\boldsymbol{a}_{j\gamma t}$ with zeros for elements $h \notin pa(j)$. Then, under the same assumption that $\boldsymbol{y}_{pa(j),t} \perp \boldsymbol{m}_t^{(j-1)}|\mathcal{D}_{t-1}$, I have

$$C(y_{jt}, \boldsymbol{y}_{j+1:m,t}|\mathcal{D}_{t-1}) = \boldsymbol{Q}_{j+1:m,t}\boldsymbol{a}_{j+1:m\gamma t} = \boldsymbol{q}_{j,j+1:m,t}.$$

Insert element h of this vector as the (j,h) and (h,j) entries of $\boldsymbol{Q}_t$, for $h = j+1:m$.

iii. End: at this point, $j = 1$ and I have filled in all elements of the $m-$vector $\boldsymbol{f}_t$, $m \times m$ matrix $\boldsymbol{Q}_t$.

The forecast results on our real data example show the efficiency of this approximation by comparison between this sequential algorithm and the direct simulation method.

5.4 CSCDLMs with Variance Learning: Approximate Analysis

5.4.1 Unknown variance

The results derived so far apply under the assumption that the variance matrices are known. It is rarely the case in practice and thus appropriate methods must be developed to allow for variance estimation.

A conjugate analysis is possible if all variances $\boldsymbol{W}_{jt}, v_{jt}$ and $\boldsymbol{W}_t$ are scaled by σ^2, the unknown factor, with an inverse gamma prior distribution. The model is as in model (5.1)– (5.3) with variances multiplied by σ^2 with the addition of a prior distribution $\sigma^2|\mathcal{D}_{t-1} \sim IG(n_{t-1}/2, d_{t-1}/2)$. The updated distribution of σ^2 is $IG(n_t/2, d_t/2)$ where $n_t = n_{t-1} + m$ and $d_t = d_{t-1} + (\boldsymbol{y}_t - \boldsymbol{f}_t)'\boldsymbol{Q}_t(\boldsymbol{y}_t - \boldsymbol{f}_t)$. The results of Theorem 5.2.1– 5.2.4 remain valid except for the multiplication of all variances by σ^2. The respective unconditional distributions (after integrating out σ^2) are obtained by the replacement of normal by Student t-distributions with n_t degrees of freedom and σ^2 by its estimate d_t/n_t in the result is conditional on $\boldsymbol{D}_t$. A variance discount factor can be used to downweight prior information in n_{t-1} and d_{t-1} if σ^2 is also allowed to vary with time.

However, I am more interested in a more flexible stochastic volatility model that allows time-varying, series-specific observational variances. Unfortunately, conjugate analysis does not hold when the σ^2 above is extended to change across series, i.e. σ_j^2. Alternatively, MCMC analysis can be used to obtain a full posterior samples of both states and variance parameters, whose details can be found in Section 5.5.

5.4.2 Variance learning: An approximate analysis

Our goal is to enable closed-form, sequential updating and filtering that include variance learning in Common State Coupled DLMs. I know that extension of the usual conjugate analysis is not available, so let's look at an obvious approximation in the model using variance discounting for the v_{jt}. One key point here is that these are nuisance parameters and I do not really care so much about full posterior inference on them; a "sensible" and viable approximation using plug-in estimates of the v_{jt} – rather than full posterior sampling etc– will have much more practical utility.

CSCDLM with known v_{jt}

Now back to the standard CSCDLM, each univariate time series y_{jt}, $(j = 1 : m)$, follows a DLM defined via equations (5.1)– (5.3). Recap the key evolution and update steps equations (5.4), posteriors at $t-1$, and (5.5), priors at t, all conditional on the common state and, for now, known values of all the v_{jt} up to time t. Then marginalizing over the common as well as individual state vectors, the predictives in each series at t are:

$$(y_{jt}|v_{jt}, \boldsymbol{y}_{1:t-1}) \sim N(f_{jt}, v_{jt} + u_{jt}) \tag{5.28}$$

where

- $f_{jt} = \boldsymbol{F}'_{jt}\boldsymbol{a}_{jt} + \boldsymbol{H}'_{jt}\boldsymbol{a}_t$, and
- $u_{jt} = \boldsymbol{F}'_{jt}\boldsymbol{V}_{jt}\boldsymbol{F}_{jt} + (\boldsymbol{H}_{jt} + \boldsymbol{D}'_{jt}\boldsymbol{F}_{jt})'\boldsymbol{R}_t(\boldsymbol{H}_{jt} + \boldsymbol{D}'_{jt}\boldsymbol{F}_{jt})$,

and I now make dependence on v_{jt} explicit in the conditioning. While the y_{jt} are dependent across j due to the common state, the v_{jt} only appear in the marginal predictive for series j, and the only appearance is as the addition to the base, structural variance u_{jt}.

Variance discount model and updating point-estimates

With a variance discount model for stochastic volatilities v_{jt}, the usual conjugate analysis would have a time $t-1$ posterior

$$(v_{j,t-1}^{-1}|\boldsymbol{y}_{1:t-1}) \sim Ga(n_{j,t-1}/2, n_{j,t-1}s_{j,t-1}/2)$$

with estimate (harmonic mean) $s_{j,t-1}$. Evolution to time t discounts the degree-of-freedom to $\beta_j n_{j,t-1}$ based on discount factor β_j, with the harmonic mean estimate being unchanged. This then updates to the time t posterior.

In the current model, this updating would be conjugate if the joint predictive density, which defines the time t likelihood function for the v_{jt}, were approximated by a product of terms across $j = 1 : m$, with the $j-th$ term being conjugate to the inverse gamma prior. The first point in this approximation is to ignore cross-series information about the v_{jt}; that is, only y_{jt} provides time t information on v_{jt}, and not the other y_{it} for $i \neq j$. This seems a reasonable approximation in general; it would be exactly the case if the common state vector were known. The second point in the approximation is key, and uses an obvious modification of the marginal predictive in series j of equation (5.28). That is, simply replace this as defining the likelihood for v_{jt} with the form

$$(y_{jt}|v_{jt}, \boldsymbol{y}_{1:t-1}) \sim N(f_{jt}, v_{jt}q_{jt}) \quad \text{where} \quad q_{jt} = 1 + u_{jt}/s_{j,t-1}.$$

This simply approximates the exact conditional variance $v_{jt}+u_{jt} = v_{jt}(1+u_{jt}/v_{jt})$ by replacing v_{jt} in the second term with its prior estimate $s_{j,t-1}$. This will be increasingly accurate as an approximation for large degrees-of-freedom $n_{j,t-1}$.

Under this approximation, the time t posterior for v_{jt} is

$$(v_{jt}^{-1}|\boldsymbol{y}_{1:t}) \sim Ga(n_{jt}/2, n_{jt}s_{jt}/2)$$

with

$$n_{jt} = \beta_j n_{j,t-1} + 1 \quad \text{and} \quad s_{jt} = \alpha_{jt}s_{j,t-1} + (1-\alpha_{jt})e_{jt}^2/q_{jt}$$

where $\alpha_{jt} = \beta_j n_{j,t-1}/(1+\beta_j n_{j,t-1})$ and e_{jt} is the 1-step ahead forecast error $e_{jt} = y_{jt} - f_{jt}$.

Conditional posteriors at t: The point now is to simply use the updated point estimates s_{jt} of v_{jt} for the parallel, conditionally independent updates of series-specific state vectors. Using these values, the priors of equation (5.5) update to posteriors $(\boldsymbol{\theta}_t|\boldsymbol{y}_{1:t}) \sim N(\boldsymbol{m}_t, \boldsymbol{C}_t)$ and, for $j = 1:m$,

$$(\boldsymbol{\theta}_{jt}|\boldsymbol{\theta}_t, \boldsymbol{y}_{1:t-1}, y_{jt}) \sim N(\boldsymbol{m}_{jt} + \boldsymbol{B}_{jt}(\boldsymbol{\theta}_t - \boldsymbol{m}_t), \boldsymbol{U}_{jt})$$

where the updated values $\boldsymbol{m}_t, \boldsymbol{C}_t$ and, for $j = 1:m$, the $\boldsymbol{m}_{jt}, \boldsymbol{B}_{jt}, \boldsymbol{U}_{jt}$ are evaluated using the usual equations now based on the plug-in estimate s_{jt} of v_{jt}.

This completes the time t analysis. Forecasting ahead more than one step from time t will be based on the current, plug-in estimates s_{jt} of v_{jt}.

I just need the same modification to details of forecasting for cases in which $\boldsymbol{F}_{jt}$ contains some of the contemporaneous values $\boldsymbol{y}_{pa(j),t}$ as in the multiregression model.

5.5 Bayesian Analysis and Computation

I refer to the framework here as a *Common State Coupled Multiregression Dynamic Model (CSCMDM)*. Bayesian analysis using MCMC methods for CSCMDM is developed in this section, building on nowadays standard simulation methods in a set of computational "modules" such as Gibbs sampling, FFBS.

In the context of observing data $\{\boldsymbol{y}_{1:T}, \boldsymbol{x}_{1:T}, \boldsymbol{H}_{1:t}\}$ over a time interval $1:T$, the full set of latent state process parameters and model hyper-parameters for inclusion in the posterior analysis are as follows:

- The common state vector $\boldsymbol{\theta}_{0:T}$ including the uncertain initial values at $t = 0$;
- The state vector specific to series j, $\boldsymbol{\theta}_{j,0:T}$, including their uncertain initial values at $t = 0$;

- The stochastic volatility processes $\boldsymbol{v}_{1:m,1:T}$.

The key analysis components for each of the sets of parameters and states are as follows. In each, I simply note the states or parameters being sampled, implicitly conditional on all other states and parameters if not explicitly stated.

Common state process $\boldsymbol{\theta}_{0:T}$ and state vector specific to series j, $\boldsymbol{\theta}_{j,0:T}$: Conditional on model hyper-parameters, volatility process states and the data $\mathcal{D}_T$, CSCMDMs reduce to standard CSCDLMs with known variances. Resampling the full sets of states is then easily obtained by a direct extension of the standard FFBS algorithm (e.g. Prado and West, 2010). Importantly, this is an efficient algorithm that regenerates full trajectories of these latent states over $0:T$ at each iterate of the overall MCMC. These new values are then conditioned upon for resampling of other components.

FFBS algorithm: I wish to sample a full set of state vectors including $\boldsymbol{\Theta}_T \triangleq \{\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_T\}$ and $\boldsymbol{\Theta}_{jT} \triangleq \{\boldsymbol{\theta}_{j0}, \boldsymbol{\theta}_{j1}, \cdots, \boldsymbol{\theta}_{jT}\}$, $j = 1, \cdots, m$, from the full, multivariate normal posterior distribution $p(\boldsymbol{\Theta}_T, \boldsymbol{\Theta}_{1:m,T}|\mathcal{D}_T)$. Note that exploiting the Markov structure of the evolution equation of the DLM, I may write

$$p(\boldsymbol{\Theta}_T, \boldsymbol{\Theta}_{1:m,T}|\mathcal{D}_T) = p(\boldsymbol{\theta}_T, \boldsymbol{\theta}_{1:m,T}|\mathcal{D}_T) \prod_{t=1}^{T} p(\boldsymbol{\theta}_{T-t}, \boldsymbol{\theta}_{1:m,T-t}|\boldsymbol{\theta}_{T-t+1}, \boldsymbol{\theta}_{1:m,T-t+1}, \mathcal{D}_{T-t}).$$

According to Theorem 5.2.5, I may sample the entire $\boldsymbol{\Theta}_T, \boldsymbol{\Theta}_{1:m,T}$ by sequentially simulating the individual state vectors $\boldsymbol{\theta}_{0:T}, \boldsymbol{\theta}_{1:m,0:T}$ as follows:

(1) Sample $\boldsymbol{\theta}_T$ from $(\boldsymbol{\theta}_T|\mathcal{D}_T) \sim N(\boldsymbol{m}_T, \boldsymbol{C}_T)$.

(2) For each $j = 1, \cdots, m$, sample $\boldsymbol{\theta}_{jT}$ from $(\boldsymbol{\theta}_{jT}|\boldsymbol{\theta}_T, \mathcal{D}_T) \sim N(\boldsymbol{m}_{jT} + \boldsymbol{B}_{jT}(\boldsymbol{\theta}_T - \boldsymbol{m}_T), \boldsymbol{U}_{jT})$, where the conditioning value of $\boldsymbol{\theta}_T$ is the value just sampled.

(3) For each $t = T-1, T-2, \cdots, 1, 0$, proceed to the following two steps:

1. Sample $\boldsymbol{\theta}_t$ from $(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, \mathcal{D}_t) \sim N[\boldsymbol{h}_t(-1), \boldsymbol{H}_t(-1)]$, where the conditioning value of $\boldsymbol{\theta}_{t+1}$ is the value just sampled.

2. Sample $\boldsymbol{\theta}_{jt}$ from $(\boldsymbol{\theta}_{jt}|\boldsymbol{\theta}_{t+1}, \boldsymbol{\theta}_t, \boldsymbol{\theta}_{j,t+1}, \mathcal{D}_t) \sim N[\boldsymbol{h}_{jt}(-1), \boldsymbol{H}_{jt}(-1)]$ where the conditioning value of $\boldsymbol{\theta}_{t+1}, \boldsymbol{\theta}_t, \boldsymbol{\theta}_{j,t+1}$ are the values just sampled.

Filtering with discounted variance v_{jt}: Following prior work of West and Harrison (1997), I now adopt standard discount variance models. Sampling the full sequence of variances from the joint posterior then consists of forward-filtering and backward-sampling. The resulting model class is sufficient for our current purposes but can clearly be extended to more elaborate submodels in new applications if context demands. I sample the observational variance $v_{1:m,1:T}$ from the full posterior distribution $p(v_{1:m,1:T}|\boldsymbol{\Theta}_T, \boldsymbol{\Theta}_{1:m,T}, \mathcal{D}_T)$. Define $\nu_{jt} = y_{jt} - \boldsymbol{H}'_{jt}\boldsymbol{\theta}_t - \boldsymbol{F}'_{jt}\boldsymbol{\theta}_{jt}$, and $\tilde{\mathcal{D}}_{jt} = \{\boldsymbol{\nu}_{j,1:t}, \boldsymbol{\Theta}_T, \boldsymbol{\Theta}_{1:m,T}, m_{j0}, \boldsymbol{B}_{j0}, \boldsymbol{U}_{j0}\}$. Note that since $(\nu_{jt}|\boldsymbol{\Theta}_T, \boldsymbol{\Theta}_{1:m,T}, m_{j0}, \boldsymbol{B}_{j0}, \boldsymbol{U}_{j0})$ are conditionally independent across j, I can obtain the filtered discounted variance v_{jt} in parallel. This means that across j, the volatility processes, $(v_{j,1:T}|\boldsymbol{\Theta}_T, \boldsymbol{\Theta}_{1:m,T}, \mathcal{D}_T)$, are conditionally independent in the posteriors given all the other model quantities, so this applies in parallel defining efficient sampling of full trajectories at each stage of the overall iterative MCMC. I sample the conditional posteriors for each of the observational volatility processes $\boldsymbol{v}_{j,1:T}, (j = 1 : m)$ using the standard discounted variance techniques for univariate stochastic volatility models (West and Harrison, 1997). Given values of these volatility trajectories, the volatilities of each observation noise of the CSCMDM are directly instantiated. For $j = 1, \cdots, m$, in parallel, proceed with the following steps:

(1) Forward filtering: for $t = 1, \cdots, T$, update recurrence relationships as follows:

$$n_{jt} = \beta_j n_{j,t-1} + 1, \quad d_{jt} = \beta_j d_{j,t-1} + \nu_{jt}^2, \quad s_{jt} = d_{jt}/n_{jt}.$$

(2) Backward sampling:

1. Sample v_{jT} from $(v_{jT}|\mathbf{\Theta}_T, \mathbf{\Theta}_{1:m,T}, \mathcal{D}_T) \sim IG(n_{jT}/2, d_{jT}/2)$.
2. For each $t = T-1, T-2, \cdots, 1, 0$, sample $\eta_{jt} \sim G[(1-\beta_j)n_{jt}/2, d_{jt}/2]$, and let $v_{jt}^{-1} = \eta_{jt} + \beta_j v_{j,t+1}^{-1}$ where the conditioning value of $v_{j,t+1}$ is the value just sampled.

5.6 Application: Smart Grid Forecasting

Electricity demand forecasting has been a popular topic. A number of univariate time series models for intraday load forecasting are reviewed and compared in Taylor et al. (2006), yet most of these single-equation approaches suffer from over-parametrization. Multivariate Bayesian dynamic modeling has become increasingly common in electricity demand forecasting. For example, Harvey and Koopman (1993), Cottet and Smith (2003), Dordonnat et al. (2008) involve studies that assess dynamic relationships between various features and load values, typically focusing on seasonality effects, trend effects and temperature effects. As mentioned before, Hosking et al. (2013) was the first to evaluate the dynamic price incentive effect on load values in the Smart Grid context, which is also a motivating example for our modeling context. I thus focus on the comparison between their model and our models using the same data.

5.6.1 Data

I evaluate the modeling methodology using data from the OlyPen project (Hammerstrom et al., 2007), which was introduced in Section 5.1. The data contains electricity usage for a period of 1 year from 120 residential customers in Olympic Peninsula, Washington. The customers were divided into four groups, each of which comprised about 30 customers, and were provided with different price schedules. Two groups had constant prices and one had time-varying prices set by a fixed schedule. The other group is of particular interest to us, the real-time-price (RTP) group, which received dynamic prices every 15 minutes. These prices were real-time instantaneous prices, rather than short-term price forecasts, which was unavailable to the customer. Therefore, it is unclear that the RTP group received sufficient information so as to shift their demand to hours of relatively low prices to reduce costs. However, customers are conceivably able to potentially infer future prices from historical relationship between price and demand cycles, and plan their electricity usage accordingly. More details about the data sets and data preparation can be found in Hosking et al. (2013).

I use hourly data of price and usage for the RTP customer group for a time span of 11 months from May 1, 2006 to April 30, 2007, out of which 2.0% of the intervals are missing. Some assumptions about missing values are detailed in Appendix 5.A. Aggregate group level data, the average usage per customer, is modeled as of more interest to the overall supply planning and also to eliminate idiosyncrasies of abnormal customer behaviour. The price curve over the entire period is shown in the upper plot of Figure 5.1, while the load curve for a typical week, March 25 - 31, is displayed in the lower left plot in the same figure. The daily load curve pattern is obviously characterized by the bimodality caused by morning and evening peaks. Weekdays tend to peak at higher levels than that of weekends. The lower right plot in Figure 5.1 illustrates daily load curves, averaged over the weekends or weekdays of summer and winter months. I use the same hourly temperature data as that of Hosking et al. (2013), which has no missing values.

5.6.2 Model structure

I use the CSCMDM with stochastic volatility as described above. The common states are selected based on (Hosking et al., 2013). Their model is basically a standard multivariate DLM in which the observed daily load curve is represented in terms of a set of periodic smoothing-spline basis functions, with the basis function coefficients

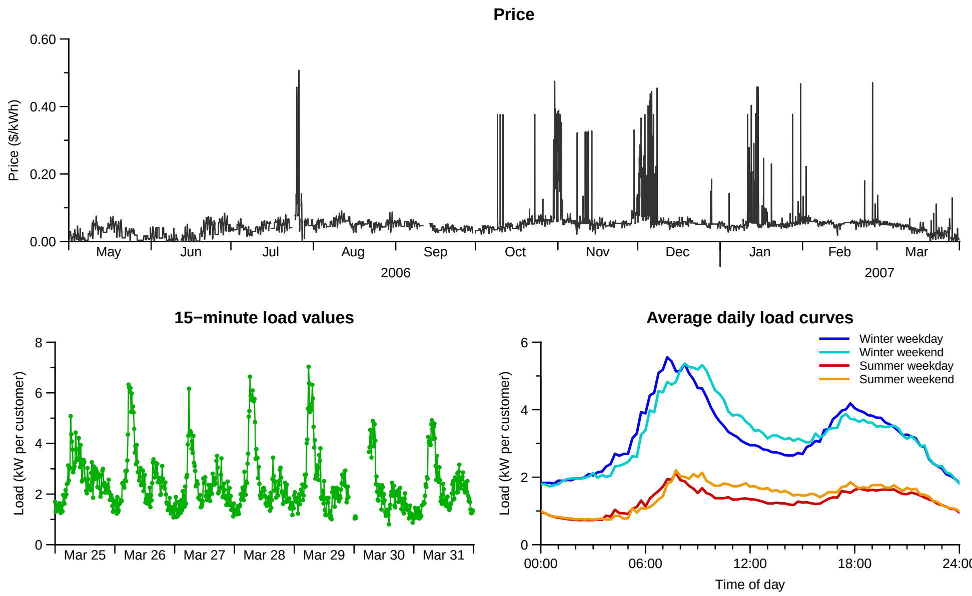


FIGURE 5.1: The upper figure is the price of electricity as seen by the RTP customer group; on the lower left is electricity usage by the RTP customer group in 15-minute intervals over a 1-week period in March 2007; the lower right figure is the daily load curves for the RTP customer groups in the OlyPen project, averaged over the weekends or weekdays of summer (May 2006 to September 2006) and winter (October 2006 to March 2007)

evolving according to a linear Gaussian state-space model that incorporates level shifts, day of the week adjustments, and weather effects, as well as the dynamic price-incentive effects.

The details of the model specification (Hosking et al., 2013) is as follows. Assume $\boldsymbol{y}_t$ is a $m = 24$ dimensional vector which contains the 24 hourly residential electricity load data points at day t, and y_{jt} is the load value of $(24-j)$-th hour at day t. Then

$$\boldsymbol{y}_t = \boldsymbol{F}\boldsymbol{\alpha}_t + \boldsymbol{v}_t,$$

where

- $\boldsymbol{y}_t$ is the $m \times 1$ load values for day t;
- $\boldsymbol{F}$ is the $m \times B$ matrix of spline coefficients (spline design matrix), where $B = 12$;

– $\boldsymbol{\alpha}_t$ is the $B \times 1$ basis function coefficients for day t;

– $\boldsymbol{v}_t \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{V})$, independently with diagonal $\boldsymbol{V}$.

This is equivalent to

$$y_{jt} = \sum_{b=1}^{B} \alpha_{bt} f_{jb} + v_{jt},$$

where B is the number of basis functions, f_{jb}, $j = 1, \cdots, m$ are the values of the bth basis function during day t; and v_{jt} $j = 1, \cdots, m$ are zero-mean noise terms. The coefficients of the spline functions, the states, follow some state evolution model that incorporates level shifts, day of the week adjustments, temperature effects, as well as the dynamic price-incentive effects caused by the two-way communication of price/load information of Smart Grid. This model has structure similar to that in Harvey and Koopman (1993), with the addition of price response effects. The extensions assume

$$\boldsymbol{\alpha}_t = \boldsymbol{\mu}_t + \boldsymbol{\delta}_t^{(1)} + \tilde{\boldsymbol{X}}_t \boldsymbol{\beta} + \boldsymbol{E}\tilde{\boldsymbol{p}}_t + \boldsymbol{w}_t^{(\alpha)} \tag{5.29}$$

where

– $\boldsymbol{\alpha}_t$ are the basis function coefficients for day t;

– $\boldsymbol{\mu}_t$ is the local level at day t for each basis function coefficient, and evolves as a random walk: $\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1} + \boldsymbol{w}_t^{(\mu)}$, with $\boldsymbol{w}_t^{(\mu)} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{W}^{(\mu)})$;

– $\boldsymbol{\delta}_t^{(\cdot)}$ is the day-of-week effect for s days before day t, evolving as $\boldsymbol{\delta}_t^{(1)} = \boldsymbol{\delta}_{t-1}^{(7)} + \boldsymbol{w}_t^{(\delta)}$, with $\boldsymbol{w}_t^{(\delta)} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{W}^{(\delta)})$; $\boldsymbol{\delta}_t^{(s)} = \boldsymbol{\delta}_{t-1}^{(s-1)}$, $s = 2, \ldots, 7$ and $\sum_{s=1}^{7} \boldsymbol{\delta}_t^{(s)} = \boldsymbol{0}$;

– $\tilde{\boldsymbol{X}}_t$ are the regressors (weather variables) for day t. Two variables are included in the model: cooling degrees (excess over 18°C) and heating degrees (deficiency below 15°C), each sharing one common coefficient for each basis function (i.e., $\beta_b^{(l)}$ in (5.30) is the same for all $b = 1, \cdots, B$);

– $\boldsymbol{\beta}$ are the regression coefficients;

– $\tilde{\boldsymbol{p}}_t$are the price variables for day t;

– $\boldsymbol{E}$ is the matrix of price elasticities, $[\epsilon_{bc}]$, with two distinct elements: a single simultaneous elasticity ϵ_S for all the diagonal elements and a single cross-elasticity ϵ_X for all the off-diagonal elements.

– $\boldsymbol{w}_t^{(\alpha)} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{W}^{(\alpha)})$ independently, where $\boldsymbol{W}^{(\alpha)}$ is diagonal.

Note that $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{p}}$ are integrated over the basis functions, e.g. if $\boldsymbol{p}_t$ is the vector of hourly prices for day t, then $\tilde{\boldsymbol{p}}_t = \boldsymbol{F}^{\mathrm{T}}\boldsymbol{p}_t$ is a vector whose jth element can be interpreted as the price of the jth basis function unit of electricity.

This Hosking et al. (2013) model can be written as a standard DLM with a $p = 88$ dimensional state vector $\boldsymbol{\theta}_t$ containing $\boldsymbol{\beta}_t^{(r)}$, $r = 1, \cdots, R$. That is,

$$\boldsymbol{y}_t = \boldsymbol{H}'_t\boldsymbol{\theta}_t + \boldsymbol{\nu}_t, \quad \boldsymbol{\nu}_t \sim N(0, \boldsymbol{U}_t),$$

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \boldsymbol{w}_t, \quad \boldsymbol{w}_t \sim N(0, \boldsymbol{W}_t),$$

where $\boldsymbol{H}_t$ is a 88×24 regression matrix as follows:

$$\boldsymbol{H}_t = [\boldsymbol{F}\boldsymbol{K}_t^{(1)}, \cdots, \boldsymbol{F}\boldsymbol{K}_t^{(R)}].$$

$$\boldsymbol{y}_t = \sum_{r=1}^{R} \boldsymbol{F}\boldsymbol{K}_t^{(r)}\boldsymbol{\theta}_t^{(r)} + \boldsymbol{\nu}_t, \quad \boldsymbol{\nu}_t \sim N(0, \boldsymbol{U}_t),$$

$$\boldsymbol{\theta}_t^{(r)} = \boldsymbol{\theta}_{t-1}^{(r)} + \boldsymbol{w}_t^{(r)}, \quad \boldsymbol{w}_t^{(r)} \sim N(0, \boldsymbol{W}^{(r)}),$$

and

$$\boldsymbol{\alpha}_t = \sum_{r=1}^{R} \boldsymbol{K}_t^{(r)}\boldsymbol{\theta}_t^{(r)} + \boldsymbol{w}_t^{(\alpha)},$$

with

$$\boldsymbol{\theta}_t^T = (\boldsymbol{\theta}_t^{(1)T}, \cdots, \boldsymbol{\theta}_t^{(R)T}), \quad \boldsymbol{w}_t^T = (\boldsymbol{w}_t^{(1)T}, \cdots, \boldsymbol{w}_t^{(R)T}).$$

My extension of this model represents the temperature and price effects as dynamic components with random-walk dependence, while Hosking et al. (2013) treat them as static. However, as no day between October 3, 2006 and March 31, 2007 has cooling degrees, I treat the corresponding state in this period as static. Thus, the extended model for the basis function coefficients $\boldsymbol{\alpha}_t = (\alpha_{t1}, \cdots, \alpha_{Bt})^T$ is

$$\alpha_{bt} = \mu_{bt} + \delta_{bt}^{(0)} + \sum_{l=1}^{L} \beta_{bt}^{(l)} \tilde{X}_{bt}^{(l)} + \sum_{c=1}^{B} \epsilon_{bct} \tilde{P}_{tc} + w_{bt}^{(\alpha)}, \text{ for } b = 1, \cdots, B, \tag{5.30}$$

with random-walk evolutions

$$\begin{aligned}
\mu_{bt} &= \mu_{t-1,b} + w_{bt}^{(\mu)}, \\
\delta_{bt}^{(0)} &= \delta_{b,t-1}^{(6)} + w_{bt}^{(\delta^{(0)})}, \\
\delta_{bt}^{(s)} &= \delta_{b,t-1}^{(s-1)} + w_{bt}^{(\delta^{(s)})}, \text{ for } s = 1, \cdots, 6.
\end{aligned}$$

Here, μ_{bt}, $\tilde{X}_{bt}^{(l)}$, $\tilde{P}_{tc}$, and $w_{bt}^{(\alpha)}$ are the same as in (5.29), and the following terms are different:

- $\delta_{b,t-1}^{(s)}$ is the day-of-week effect for s days before day t, $s = 0, 1, \cdots, 6$, and $\sum_{s=0}^{6} \delta_{b,t-1}^{(s)} = 0$ for $b = 1, \cdots, B$.
- ϵ_{bct} $(b, c = 1, \cdots, B)$ is the time-varying elasticity of demand with respect to price, with two distinct elements at each time t : a single simultaneous elasticity ϵ_{St} for all the diagonal elements and a single cross-elasticity ϵ_{Xt} for all the off-diagonal elements.
- $\beta_{bt}^{(l)}$ is the time-varying coefficient of $\tilde{X}_{bt}^{(l)}$.
- $(w_{1t}^{(\delta^{(0)})} \cdots w_{1t}^{(\delta^{(6)})}, \cdots, w_{Bt}^{(\delta^{(0)})} \cdots w_{Bt}^{(\delta^{(6)})}) \sim \mathrm{N}(\mathbf{0}, \boldsymbol{W}^{(\delta)})$ and $(w_{1t}^{(\mu)}, \cdots, w_{Bt}^{(\mu)}) \sim \mathrm{N}(\mathbf{0}, \boldsymbol{W}^{(\mu)})$ independently, where $\boldsymbol{W}^{(\delta)}$ and $\boldsymbol{W}^{(\mu)}$ are diagonal.

Let y_{jt} be the load value of $(24 - j)$-th hour at day t. Each of the basis function $f_{jb}, b = 1, \cdots, B$ is a periodic cubic spline with arbitrarily chosen knots to achieve a good approximation to the typical daily load curves of the Olypen data set. Here B is set to 12. The 12 spline functions f_{jb} with $j \in \{1, 2, \cdots, 24\}$ representing hours of day are plotted in Figure 5.2.

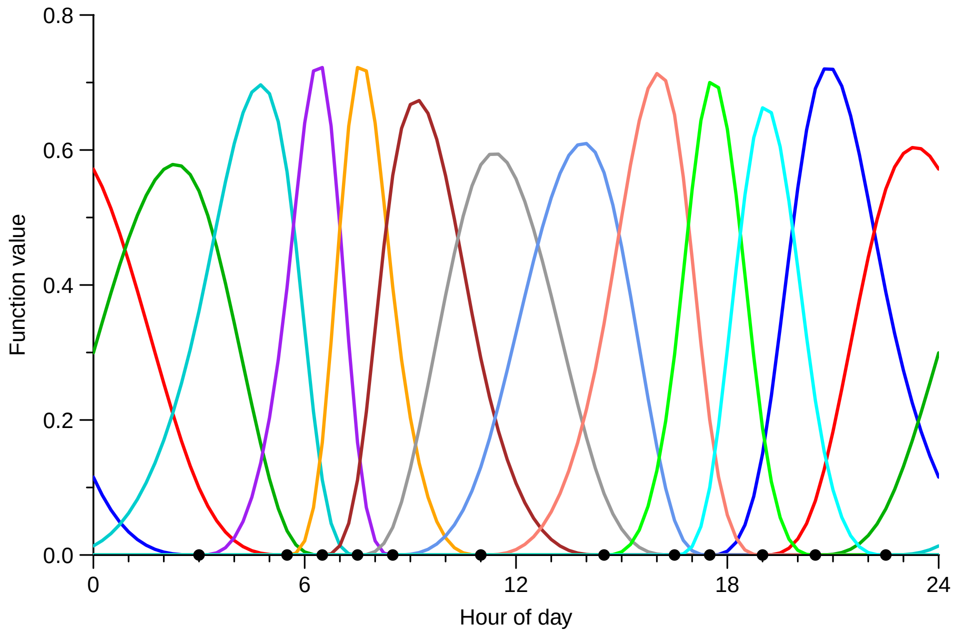


FIGURE 5.2: Spline function of time

I consider the following seven methods from the class of DLMs to illustrate the model performance of CSCDLM compared with the standard DLM of Hosking et al. (2013). For the first three methods, I divide the whole sample period into two datasets: (1) from May 1,2006 to October 15, 2006 (168 days) to estimate the parameters and (2) from October 16, 2006 to March 31, 2007 (167 days) to evaluate the 1-day-ahead forecast accuracy of the different models.

M_1: CSCDLM with specified non-empty parental sets except the first hour, as specified in (5.31), with 1-day-ahead forecasts using my fast algorithm to approximate analytic forms and the observation variance approximation method. My model has the same common state vector as their model, but is built under the framework of CSCMDM with individual state vectors $\boldsymbol{\theta}_{jt}$. The idiosyncratic regression vector for series j is taken to be the contemporaneous values from other series of the same day, i.e. $\boldsymbol{F}_{jt} = \boldsymbol{y}_{pa(j),t}$. Since y_{jt} is naturally ordered according to time, I take the parental set to be the last hour, i.e.

$$\{pa(j) = \{j+1\} \text{ for } j = 1, \cdots, 23, pa(m) = \varnothing\}. \tag{5.31}$$

The practical meaning behind this is that from the second hour on, each hourly load is dependent on the load of previous hour, which allows capturing of the same-day covariance between hourly loads.

A random walk state evolution is assumed for the common state and all idiosyncratic states, i.e. $\boldsymbol{G}_t = \boldsymbol{I}_p, \boldsymbol{G}_{jt} = \boldsymbol{I}_{p_j}, j = 1, \cdots, m$. One simple and natural discount specification of state evolution variances is adopted in this example. The covariance matrix for the common state is a block diagonal matrix $\boldsymbol{W}_t$ with diagonal matrix $\boldsymbol{W}_t^{(\mu)}, \boldsymbol{W}_t^{(\delta)}, \boldsymbol{W}_t^{(\beta)}$ as block diagonal elements, corresponding to the local level effects, day-of-week effects and temperature/price incentive effects, and similar notation is adopted for $\boldsymbol{C}_t$ and $\boldsymbol{R}_t$. Denote the common state discount factors as $\delta^{(\mu)}, \delta^{(\delta)}, \delta^{(\beta)} \in (0,1)$ and the set of idiosyncratic state discount factors $\delta_j \in (0,1), j = 1, \cdots, m$. The state noise variance matrices are:

- For the common states, denoting $* \in \{\mu, \delta, \beta\}$, $\boldsymbol{W}_t^{(*)}(i,j)$ is

  $$\begin{cases} \boldsymbol{C}_{t-1}^{(*)}(i,i)(1/\delta^{(*)} - 1), \text{ i.e. } \boldsymbol{R}_t^{(*)}(i,i) = \boldsymbol{C}_{t-1}^{(*)}(i,i)/\delta^{(*)}, & \text{if } i = j, \\ 0, \text{ i.e. } \boldsymbol{R}_t^{(*)}(i,j) = \boldsymbol{C}_{t-1}^{(*)}(i,j), & \text{if } i \neq j. \end{cases}$$

  This discounting technique enables time-varying state variances while also shrinks the prior absolute correlation between the common state variables at time $t+1$ from that of the posterior at time t.

- For states specific to each series $j = 1, \cdots, m$, $\boldsymbol{S}_{jt} = \boldsymbol{U}_{j,t-1}(\delta_j^{-1} - 1)$ so that $\boldsymbol{V}_{jt} = \boldsymbol{U}_{j,t-1}/\delta_j$.

Diffuse priors are used for initial common states $\boldsymbol{\theta}_0 \sim N(\boldsymbol{m}_0, \boldsymbol{C}_0)$ and states specific to series $j = 1, \cdots, m$, $\boldsymbol{\theta}_{j0} \sim N(\boldsymbol{m}_{j0}, \boldsymbol{C}_{j0})$, where $\boldsymbol{C}_0 = 10^3\boldsymbol{I}$, $\boldsymbol{C}_{j0} = \boldsymbol{I}$, $\boldsymbol{m}_0 = \boldsymbol{0}$, $\boldsymbol{m}_{j0} = (0.5, \cdots, 0.5)$. Additionally, parameters associated with the observation variance discount model are $n_0 = 24, s_0 = 0.25$.

I conduct grid search over state and observation variance discount factors using the data in period (1) as training data and select the set of values minimizing in-sample 1-step-ahead forecast accuracy measured by mean absolute percentage error; the set of values are $\delta^{(\mu)} = \delta^{(\beta)} = 0.96, \delta^{(\delta)} = 0.9999$. I set $\delta_j = \beta_j = 0.99, j = 1, \cdots, m$, after finding that the forecasting results are not sensitive to these parameters. Fixing the state discount factors to these values, I test this model using the out-of-sample data. I also clone the MCMC to adapt to the variance discounting model, and then run the approximate analysis and compare with MCMC results.

M_2: CSCDLM with empty parental sets, with 1-step-ahead forecasting using observation variance approximation. The state and observation discount factors are the same as M_1.

M_3: Standard DLM as specified in Hosking et al. (2013), using MLEs as estimates of variance parameters.

M_4: Same model settings as M_1, but using MCMC with the FFBS algorithm on both state and observation variances to compute posterior distributions of model parameters.

M_5: Same model settings as M_2, but using MCMC with the FFBS algorithm on both state and observation variances to compute posterior distributions of model parameters.

M_6: Same model settings as M_3, but using the standard FFBS theorem (West and Harrison, 1997) to compute posterior distributions of state parameters.

M_7: Same model settings as M_1, but using Monte Carlo simulations with $nmc = 10,000$ iterations at each time t to compute 1-day-ahead predictive distributions of $\boldsymbol{y}_t, t = 1, \cdots, T$.

5.6.3 Results

The following comparisons between CSCMDM M_1 and CSCDLM M_2 are made to illustrate the impact of the inclusion of appropriate parental sets for each series which induces stochastic covariances. The former has non-empty parental sets for all the series except the first one– allowing correlated observation noise, while the latter has empty parental sets for all series– assuming observation noises are independent across series. M_3 is considered the benchmark standard DLM model to compare with the new CSCDLM, i.e. M_2, and CSCMDM, i.e. M_1.

Forecasting performance

The 1-day-ahead forecasting performance of the four models, M_1, M_2, M_3, M_7, is investigated for the out-of-sample observations in period (2) and summarized in Table 5.1. Accuracy is measured by mean absolute percentage error (MAPE), computed for the untransformed load values. Uncertainty is measured by 95% credible interval (CI) coverage percentage, the percentage of real load values falling into the 95% credible intervals.

Both the forecast accuracy and coverage percentages of M_7 are very close to those of M_1. The mean absolute percentage difference between the 1-step-ahead mean of these two models is 2.3%. This close forecast performance provides practical and

Table 5.1: Comparison of out-of-sample forecast performance between the four models

Model	Parental set	MAPE %	Coverage %
M_1	Lag 1	12.38	96.7
M_2	None	12.44	95.9
M_3	None	12.48	96.7
M_7	Lag 1	12.39	96.9

Table 5.2: 1-day-ahead forecast accuracy by month for the whole period June 2006 to March 2007

	June	**July**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**	**Jan**	**Feb**	**Mar**	**Total**
M_3	15.79	15.78	12.40	14.30	15.15	13.06	11.54	11.33	11.65	14.07	13.51
M_2	19.95	13.72	12.14	14.48	15.01	12.86	11.77	11.22	11.92	13.91	13.70
M_1	16.61	13.77	11.82	14.10	14.83	12.88	11.49	11.59	11.61	13.83	13.25

empirical support that the approximation of 1-step-ahead forecast analytic form is empirically very accurate.

It can be seen that M_1 has a better overall forecast performance than M_3 with a 0.1% higher accuracy and same CI coverage ratios. There is 0.08% improvement in accuracy of M_1 over M_2, suggesting the benefits of including non-empty parental sets. A more detailed forecast accuracy comparison between M_1, M_2, and M_3 is summarized in Table 5.2 and Figure 5.3. M_1 outperforms M_3 on weekends, Monday, Wednesday and Friday. Out of the 11 months, I only see 2 months, January and June, in which M_1 was dominated by M_3. Comparing the three models by day of week, M_1 is the most accurate model for Monday and Sunday, while M_3 wins on Tuesday, Thursday, Friday and Saturday. In terms of accuracy by hours, M_1 outperforms the other two models for thirteen hours, and M_1, M_2 dominate for seven hours, and four hours respectively.

The in-sample fitting accuracy of M_4, M_5, M_6, is examined for all the observations from August 2006 through March 2007, excluding the first three months as a burn-in period. The in-sample fit of the load values is calculated from the posterior mean of

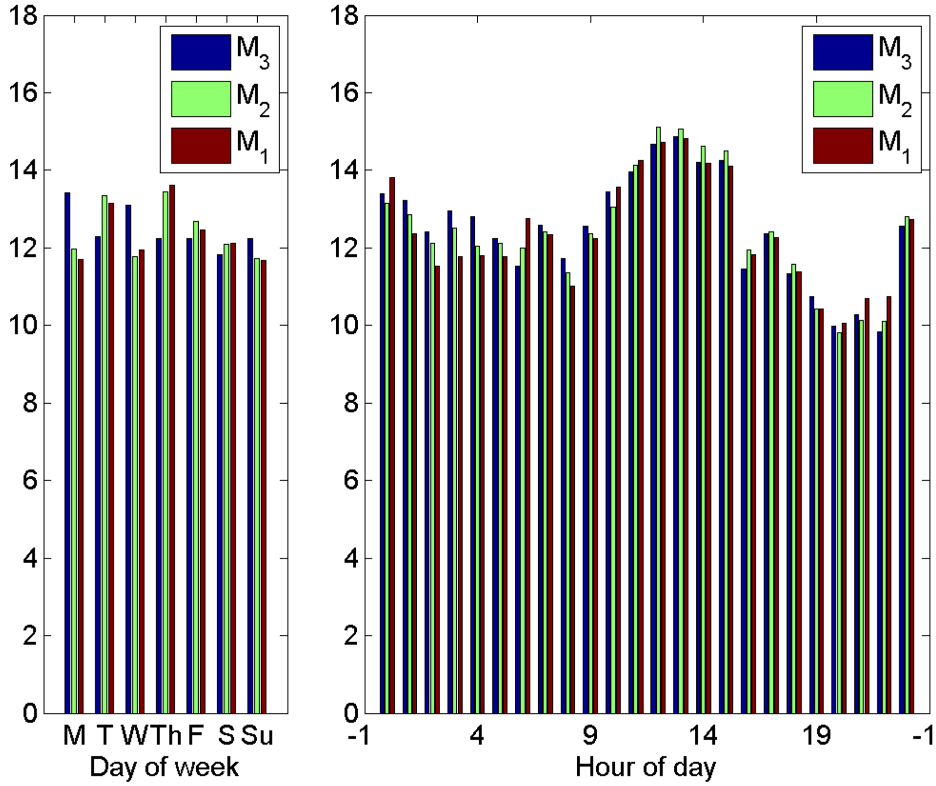


FIGURE 5.3: 1-day-ahead forecast accuracy by day type and by hour of M_1 and M_3, for the months from August 2006 to March 2007, separately for each day of the week and each hour of the day

the backward-smoothed state parameters, i.e. $\boldsymbol{\Theta}_T|\mathcal{D}_T, \boldsymbol{\Theta}_{1:m,T}|\mathcal{D}_T$. From Table 5.3, M_4 is the most accurate model out of the three with a 0.45% higher accuracy than M_6.

Table 5.3: Comparison of in-sample forecast performance between the three models

Model	Parental set	MAPE %
M_4	Lag 1	9.60
M_5	None	10.82
M_6	None	10.05

Figure 5.4 demonstrates the backward-filtered posterior means of series-specific stochastic volatilities $\sqrt{v_{j,1:T}}|\mathcal{D}_T$, for $j = 1 : 24$, that is, the volatility elements excluding the common and individual states. The upper figure shows the results of M_4, and the lower figure shows that of M_1. Both plots exhibit that different series, i.e. hours of the day, display idiosyncrasies in observation variance levels throughout the whole period. And different series slightly vary with time with a gradual decrease from a relatively high value at the beginning of the period to a lower level where each series tend to remain steady towards the end of the period. It can been seen that the two methods produce variance estimates close to each other throughout the whole period excluding the first three months– the burn-in period. In fact, this closeness of the posterior observation variances of M_1 and MCMC analysis, i.e. M_4, justifies the adequacy of the approximation of variance updates.

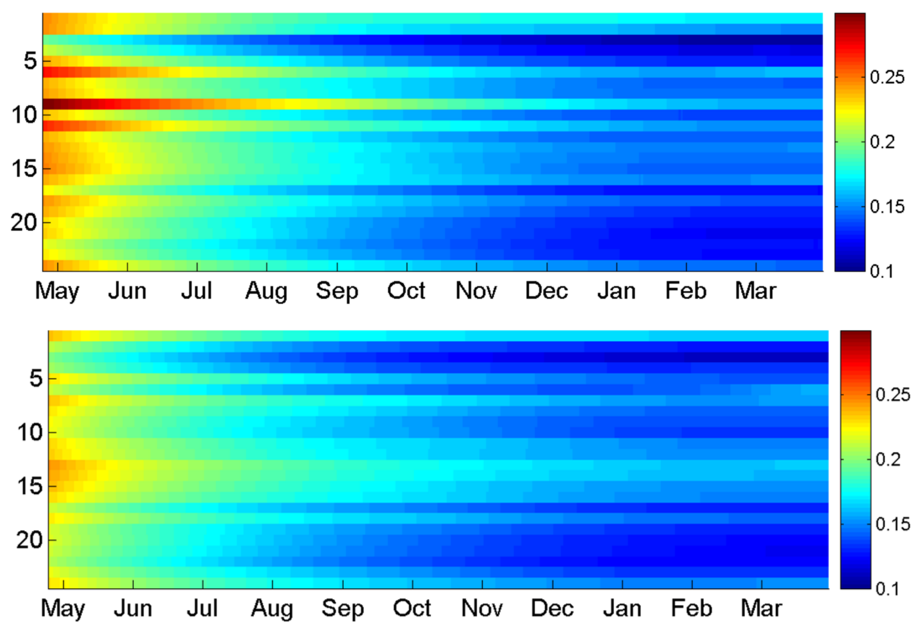


FIGURE 5.4: Backward-filtered posterior means of series-specific stochastic volatility of M_4 (in the upper figure) and M_1 (in the lower figure)

Signal extraction: Time-varying coefficients

After comparing different models on forecast accuracy, I now focus on the estimates of the feature effects of M_1. Figure 5.5 exhibits the posterior means and 90% credible intervals of the forward filtered series-specific states $\{\boldsymbol{\theta}_{jt}|\mathcal{D}_{t-1} : t = 1, \cdots, T, j = 1, \cdots, m\}$ of M_1. All the 24 idiosyncratic states estimates change over time, reflecting the time-varying stochastic volatilities of observation noise. The idiosyncratic state estimates are significantly positive throughout almost the whole period for 13 out of 23 hours, when $j \in \{2, 5, 12, 13, 14, 15, 16, 17, 18, 20, 22, 23, 24\}$, indicating positive correlations between these hours and their adjacently previous hours. The series-specific states are significantly positive for about half of the whole period when $j \in \{3, 4, 7, 8, 10, 19, 21\}$. Only when $j \in \{6, 9, 11\}$ do the state estimates change signs over time. These findings provide strong support for allowing correlated and time-varying structure of the observation noise vectors.

The temperature effects are significant for the entire period. Figure 5.6 displays the estimated temperature effects trajectories over the whole sample period of M_1. For the first couple of months, the estimates are unstable with wide credible intervals. Their variation stabilizes with narrower credible intervals by the end of June, when the temperature has reached both cooling degrees and heating degrees a significant number of times. The heating effects are greater than the cooling effects for the entire period, which concords with the findings of Hosking et al. (2013). The coefficients of heating degrees have gradually increased from the August, the end of summer, to January, the middle of winter, with gradually declining temperature in the same period. It makes sense that the time-varying heating coefficients show an obviously upward trend in the period when electric heating becomes more intensively used.

Figure 5.7 shows the estimated price elasticities of M_1 varying with time. The price is initially low for the first couple of months, so the price incentive effects is

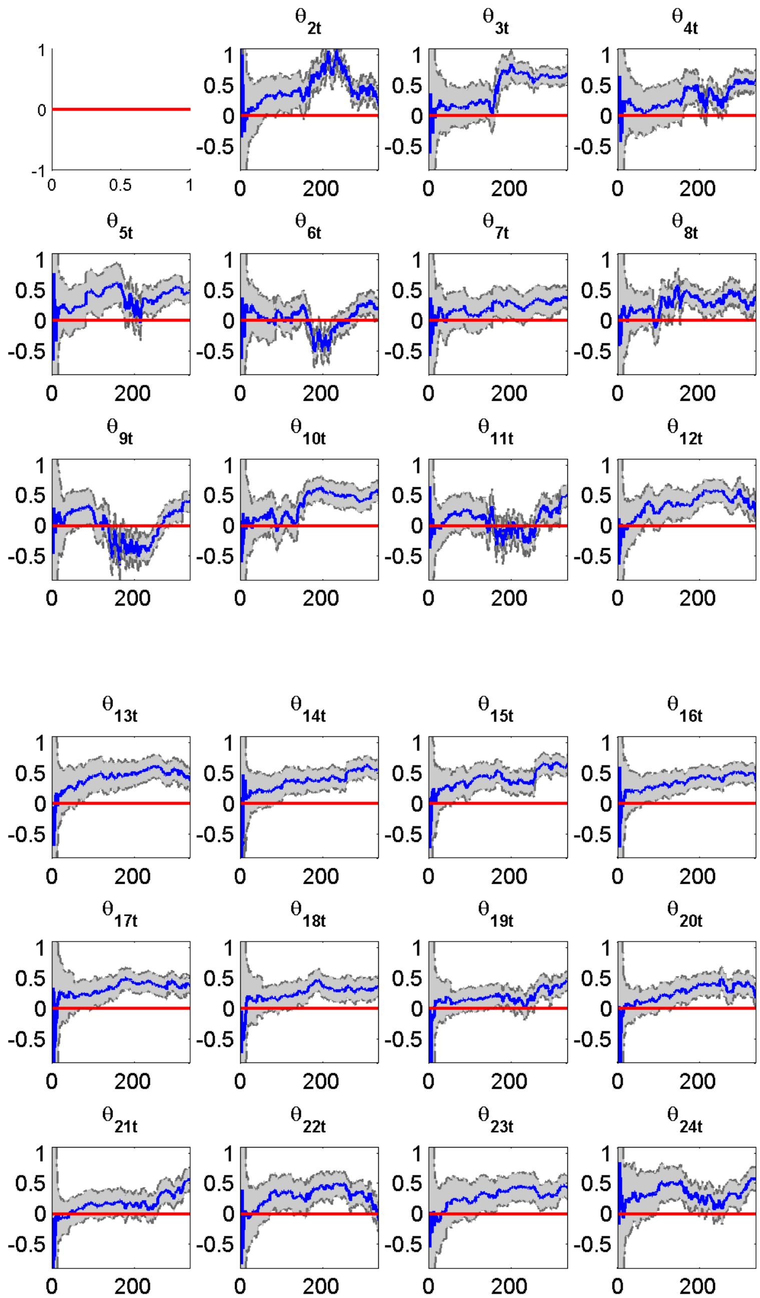


FIGURE 5.5: Posterior means of series-specific state $\boldsymbol{\theta}_{jt}|\mathcal{D}_t$ and 90% credible intervals

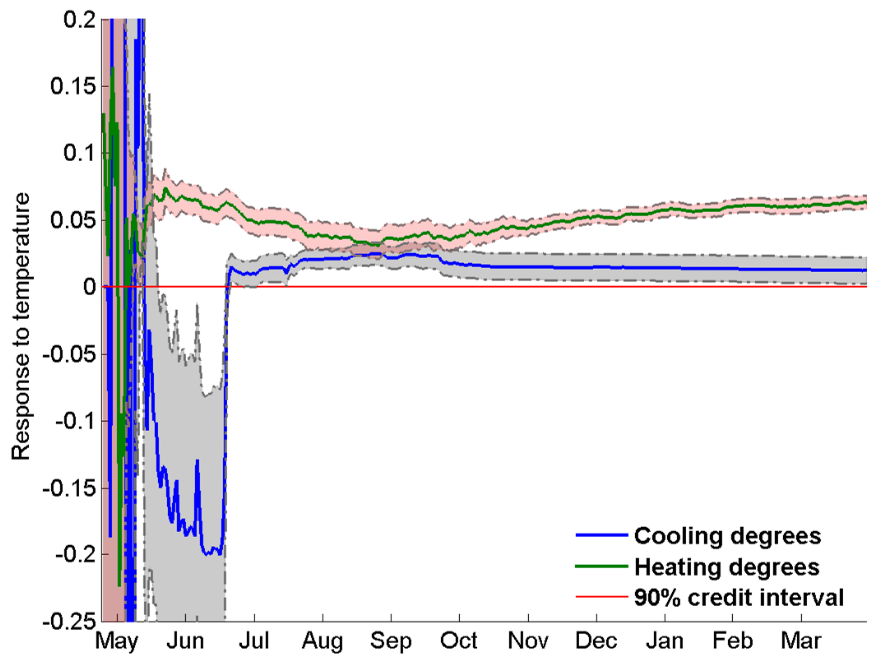


FIGURE 5.6: M_1 posterior means of common state with respect to temperature, with 90% credible intervals

highly volatile and may be unreliable in this period. Sudden emergence of high prices in July is associated with a pronounced shift in the level of estimated simultaneous elasticities. Starting in October, high prices become more frequent, both the cross and simultaneous elasticities become less volatile and vary within a small range. The simultaneous elasticity effects are significantly negative and the cross-elasticity effects are significantly positive throughout the period. These results are consistent with the results of Hosking et al. (2013), meaning that consumers have shifted demand from times of high price of electricity to times of lower price, and indicates that they were able to infer future prices in order to reduce their costs by load shifting even though they only have access to current prices.

In-sample diagnostics

The model analysis relies on a few modeling assumptions, among which our main focus is the assumed conditional independence across series for the observation er-

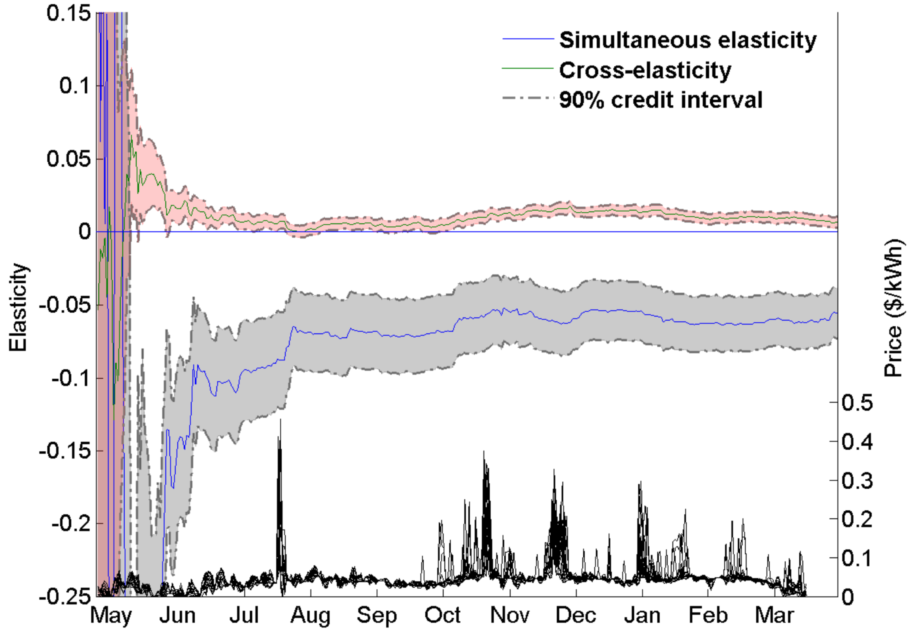


FIGURE 5.7: M_1 posterior means of common state with respect to price elasticities with 90% credible intervals and price data

rors $\boldsymbol{\nu}_{jt}$ in eqn (5.1). Conditional independence would justify our goal of capturing stochastic volatility by including contemporaneous conditional dependence network structures.

I assess consistency of the data with the assumed conditional independence of $\boldsymbol{\nu}_{jt}$ by comparing M_4 and M_5 on the posterior estimate of correlation matrix corresponding to the posterior samples of the standardised errors: $\tilde{\nu}_{jt} = \nu_{jt}/\sqrt{v_{jt}}$. At each MCMC iteration I save the current sampled values of each $\tilde{\nu}_{jt}$, building up a full posterior sample of all $m \times T$ error terms. Define $\hat{\boldsymbol{C}} = \sum_{i=1}^{N} \boldsymbol{C}_i/N$ to be the posterior estimate of the correlation matrix of standardized observation error vector $\tilde{\boldsymbol{\nu}}_t = (\tilde{\nu}_{1t}, \cdots, \tilde{\nu}_{mt})$, where $\boldsymbol{C}_i$ is the correlation matrix of $\tilde{\boldsymbol{\nu}}_t, t \in \{1:T\}$ in the $i-$th posterior sample, and N is the total number of MCMC iterations in M_4 and M_5. Given our assumption in (5.1) about v_{jt}, $\tilde{\boldsymbol{\nu}}_t$ is independent over series j and across time t and thus $\hat{\boldsymbol{C}}$ is supposed to be close to identity matrix. The left image in Figure 5.8 shows the image of error correlation matrix $\hat{\boldsymbol{C}}$ for M_4, which has almost all off-diagonal values close to zero, demonstrating close concordance with the assumption of conditional independence. However, as shown in the right image, $\hat{\boldsymbol{C}}$ of M_5 has substantial positive correlation between adjacent hours, and the absolute correlations between 98.8% of all pair of hours, especially the adjacent hours, are larger than that in the left. This comparison provides strong evidence for adding the assumption of correlated observational disturbance, as suggested by the data. It also distinctly validates the efficacy of including appropriate parental sets to allow for correlated observation errors.

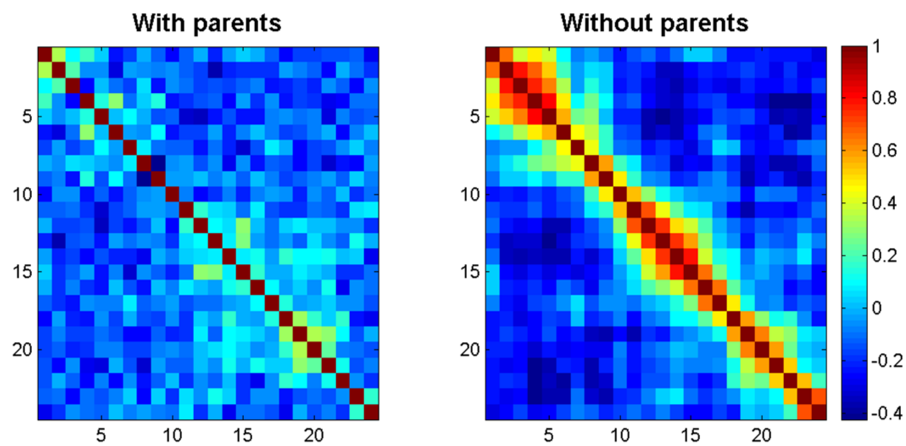


FIGURE 5.8: Posterior estimate of correlation matrix of standardised errrors $\tilde{\nu}_{jt}$'s of M_4 and M_5

5.7 Summary

I have developed a state space model for multivariate time series with both common and series-specific states, and extended the model to incorporating multivariate stochastic volatility. Bayesian analytics, including FFBS schemes, are developed. An approximation method using variance discounting has been established to enable online and analytic learning of observation variances. Even though it requires some modest approximation, this method defines computation of sequential step-ahead forecasting simply unavailable using MCMC methods, and appears to be of real practical value.

This method is applied to short-term prediction of electricity demand in a Smart Grid scenario. The model shares features of that in Hammerstrom et al. (2007), yet is adapted to include online learning of multivariate stochastic volatility in observation and state variances. I evaluate model performance using OlyPen data (Hammerstrom et al., 2007), and compare the out-of-sample prediction results with Hosking et al. (2013). I conclude that this method, allowing volatilities and correlations in observation disturbances, leads to improvements in general forecasting performance of the daily load curve for residential electricity usage. I compare with results using MCMC analysis adapted to the variance discounting model, and validate the efficacy of the new model and methodology.

Appendix 5.A: Additional Proofs of Theorems and Assumptions

Additional proofs of theorems

Proof of Theorem 5.2.1. First, I prove that

$$(\boldsymbol{\theta}_{jt}|\boldsymbol{\theta}_t, \mathcal{D}_{t-1}) \sim N(\boldsymbol{a}_{jt} + \boldsymbol{D}_{jt}(\boldsymbol{\theta}_t - \boldsymbol{a}_t), \boldsymbol{V}_{jt}).$$

Note that

$$\begin{aligned} C(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{jt}|\mathcal{D}_{t-1}) &= C(\boldsymbol{G}_t\boldsymbol{\theta}_{t-1} + \omega_t, \boldsymbol{G}_{jt}\boldsymbol{\theta}_{j,t-1} + \omega_{jt}|\mathcal{D}_{t-1}) \\ &= C(\boldsymbol{G}_{jt}\boldsymbol{\theta}_{j,t-1}, \boldsymbol{G}_t\boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1}) + C(\omega_{jt}, \omega_t|\mathcal{D}_{t-1}) \\ &= \boldsymbol{G}_{jt}\boldsymbol{B}_{j,t-1}\boldsymbol{C}_{t-1}\boldsymbol{G}_t' + \boldsymbol{D}_{jt}\boldsymbol{W}_t \\ &= \boldsymbol{D}_{jt}\boldsymbol{R}_t. \end{aligned}$$

Hence,

$$E(\boldsymbol{\theta}_{jt}|\boldsymbol{\theta}_t, \mathcal{D}_{t-1}) = \boldsymbol{a}_{jt} + \boldsymbol{D}_{jt}\boldsymbol{R}_t\boldsymbol{R}_t^{-1}(\boldsymbol{\theta}_t - \boldsymbol{a}_t) = \boldsymbol{a}_{jt} + \boldsymbol{D}_{jt}(\boldsymbol{\theta}_t - \boldsymbol{a}_t),$$

$$\begin{aligned} V(\boldsymbol{\theta}_{jt}|\boldsymbol{\theta}_t, \mathcal{D}_{t-1}) &= \boldsymbol{G}_{jt}\boldsymbol{C}_{j,t-1}\boldsymbol{G}'_{jt} + \boldsymbol{D}_{jt}\boldsymbol{W}_t\boldsymbol{D}'_{jt} + \boldsymbol{S}_{jt} - \boldsymbol{D}_{jt}\boldsymbol{R}_t\boldsymbol{R}_t^{-1}\boldsymbol{R}_t\boldsymbol{D}'_{jt} \\ &= \boldsymbol{G}_{jt}\boldsymbol{C}_{j,t-1}\boldsymbol{G}'_{jt} - \boldsymbol{D}_{jt}\boldsymbol{G}_t\boldsymbol{C}_{t-1}\boldsymbol{G}'_t\boldsymbol{D}'_{jt} + \boldsymbol{S}_{jt} \\ &= \boldsymbol{G}_{jt}\boldsymbol{U}_{j,t-1}\boldsymbol{G}'_{jt} + \boldsymbol{S}_{jt} = \boldsymbol{V}_{jt}. \end{aligned}$$

The rest of the proof easily follows. □

Proof of Corollary 5.2.2.

$$\begin{aligned} p(\boldsymbol{\theta}_t|\mathcal{D}_t) \propto exp\{-\Big[\sum_{j=1}^m (\boldsymbol{\theta}'_t\boldsymbol{N}_{jt}\boldsymbol{\theta}_t - 2(y_{jt}/\nu_{jt}\boldsymbol{H}'_{jt} - (\boldsymbol{a}_{jt} - \boldsymbol{D}_{jt}\boldsymbol{a}_t)'\boldsymbol{V}_{jt}^{-1}\boldsymbol{D}_{jt})\boldsymbol{\theta}_t \\ - (\boldsymbol{L}_{jt}\boldsymbol{\theta}_t + \boldsymbol{M}_{jt})'\boldsymbol{U}_{jt}(\boldsymbol{L}_{jt}\boldsymbol{\theta}_t + \boldsymbol{M}_{jt})) + (\boldsymbol{\theta}_t - \boldsymbol{a}_t)'\boldsymbol{R}_t^{-1}(\boldsymbol{\theta}_t - \boldsymbol{a}_t)\Big]/2\}, \end{aligned}$$

Therefore, it follows that

$$\begin{aligned} \boldsymbol{C}_t &= (\boldsymbol{R}_t^{-1} + \sum_{j=1}^m (\boldsymbol{N}_{jt} - \boldsymbol{L}'_{jt}\boldsymbol{U}_{jt}\boldsymbol{L}_{jt}))^{-1} \\ &= (\boldsymbol{R}_t^{-1} + \sum_{j=1}^m (\boldsymbol{N}_{jt} - \boldsymbol{B}'_{jt}\boldsymbol{U}_{jt}^{-1}\boldsymbol{B}_{jt}))^{-1} \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{L}_{jt} &= \boldsymbol{V}_{jt}^{-1}\boldsymbol{D}_{jt} - \boldsymbol{F}_{jt}\boldsymbol{H}'_{jt}/v_{jt} = \boldsymbol{U}_{jt}^{-1}\boldsymbol{B}_{jt}, \\ \boldsymbol{M}_{jt} &= \frac{y_{jt}}{v_{jt}}\boldsymbol{F}_{jt} + \boldsymbol{V}_{jt}^{-1}(\boldsymbol{a}_{jt} - \boldsymbol{D}_{jt}\boldsymbol{a}_t). \end{aligned}$$

Therefore, I have

$$\boldsymbol{C}_t^{-1} = \boldsymbol{R}_t^{-1} + \sum_{j=1}^m (\boldsymbol{N}_{jt} - \boldsymbol{B}'_{jt}\boldsymbol{U}_{jt}^{-1}\boldsymbol{B}_{jt}),$$

$$\boldsymbol{A}_t = V(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{1:m,t}, \mathcal{D}_t)^{-1} = \boldsymbol{R}_t^{-1} + \sum_{j=1}^m \boldsymbol{N}_{jt}.$$

And then

$$\boldsymbol{m}_t = \boldsymbol{C}_t\big[\boldsymbol{R}_t^{-1}\boldsymbol{a}_t + \sum_{j=1}^{m}\big(\boldsymbol{B}'_{jt}\boldsymbol{M}_{jt} + y_{jt}/v_{jt}\boldsymbol{H}_{jt} - \boldsymbol{D}'_{jt}\boldsymbol{V}_{jt}^{-1}(\boldsymbol{a}_{jt} - \boldsymbol{D}_{jt}\boldsymbol{a}_t)\big)\big]$$

$$= \boldsymbol{C}_t\big[\boldsymbol{R}_t^{-1}\boldsymbol{a}_t + \sum_{j=1}^{m}\boldsymbol{D}'_{jt}\boldsymbol{V}_{jt}^{-1}\boldsymbol{U}_{jt}\boldsymbol{F}_{jt}y_{jt}/v_{jt} - \boldsymbol{H}_{jt}\boldsymbol{F}'_{jt}\boldsymbol{U}_{jt}\boldsymbol{F}_{jt}y_{jt}/v_{jt}^2 + y_{jt}\boldsymbol{H}_{jt}/v_{jt}$$

$$- (\boldsymbol{D}'_{jt}\boldsymbol{V}_{jt}^{-1}\boldsymbol{U}_{jt}\boldsymbol{F}_{jt}\boldsymbol{F}'_{jt}/v_{jt} + \boldsymbol{H}_{jt}\boldsymbol{F}'_{jt}\boldsymbol{U}_{jt}\boldsymbol{V}_{jt}^{-1}/v_{jt})(\boldsymbol{a}_{jt} - \boldsymbol{D}_{jt}\boldsymbol{a}_t)\big]$$

$$= \boldsymbol{C}_t\big[\boldsymbol{R}_t^{-1}\boldsymbol{a}_t + \sum_{j=1}^{m}\boldsymbol{B}'_{jt}\boldsymbol{F}_{jt}y_{jt}/v_{jt} + y_{jt}\boldsymbol{H}_{jt}/v_{jt}$$

$$- (\boldsymbol{D}'_{jt}\boldsymbol{V}_{jt}^{-1}\boldsymbol{U}_{jt}\boldsymbol{F}_{jt}\boldsymbol{F}'_{jt}/v_{jt} + \boldsymbol{H}_{jt}\boldsymbol{F}'_{jt}\boldsymbol{U}_{jt}\boldsymbol{V}_{jt}^{-1}/v_{jt})(\boldsymbol{a}_{jt} - \boldsymbol{D}_{jt}\boldsymbol{a}_t)\big]$$

$$= \boldsymbol{C}_t\big[\boldsymbol{R}_t^{-1}\boldsymbol{a}_t + \sum_{j=1}^{m} y_{jt}/v_{jt}(\boldsymbol{B}'_{jt}\boldsymbol{F}_{jt} + \boldsymbol{H}_{jt}) - (\boldsymbol{D}'_{jt} - \boldsymbol{B}'_{jt})\boldsymbol{V}_{jt}^{-1}(\boldsymbol{a}_{jt} - \boldsymbol{D}_{jt}\boldsymbol{a}_t)\big].$$

Thus I have

$$\boldsymbol{C}_t^{-1}\boldsymbol{m}_t = \boldsymbol{R}_t^{-1}\boldsymbol{a}_t + \sum_{j=1}^{m} y_{jt}/v_{jt}(\boldsymbol{B}'_{jt}\boldsymbol{F}_{jt} + \boldsymbol{H}_{jt}) - (\boldsymbol{D}'_{jt} - \boldsymbol{B}'_{jt})\boldsymbol{V}_{jt}^{-1}(\boldsymbol{a}_{jt} - \boldsymbol{D}_{jt}\boldsymbol{a}_t).$$

$\square$

Proof of Theorem 5.2.3.

$$p(\boldsymbol{\theta}_{1t}, \cdots, \boldsymbol{\theta}_{mt}|\boldsymbol{\theta}_t, \mathcal{D}_t) \propto p(\boldsymbol{\theta}_{1t}, \cdots, \boldsymbol{\theta}_{mt}, \boldsymbol{\theta}_t|\mathcal{D}_t),$$

$$\propto p(\boldsymbol{y}_t|\boldsymbol{\theta}_t, \boldsymbol{\theta}_{jt}, \mathcal{D}_{t-1})p(\boldsymbol{\theta}_t|\mathcal{D}_{t-1})\prod_{j=1}^{m} p(\boldsymbol{\theta}_{jt}|\boldsymbol{\theta}_t, \mathcal{D}_t),$$

Therefore, I have the conditional independence of $\boldsymbol{\theta}_{1t}, \cdots, \boldsymbol{\theta}_{mt}|\boldsymbol{\theta}_t$ and for all j,

$$\begin{aligned}
p(\boldsymbol{\theta}_{jt}|\boldsymbol{\theta}_t, \mathcal{D}_t) &\propto p(\boldsymbol{\theta}_{jt}, \boldsymbol{\theta}_t|\mathcal{D}_t), \\
&\propto p(\boldsymbol{y}_t|\boldsymbol{\theta}_t, \boldsymbol{\theta}_{jt}, \mathcal{D}_{t-1})p(\boldsymbol{\theta}_t|\mathcal{D}_{t-1})p(\boldsymbol{\theta}_{jt}|\boldsymbol{\theta}_t, \mathcal{D}_t), \\
&\propto exp\Big\{ -\frac{1}{2}[(y_{jt} - \boldsymbol{F}'_{jt}\boldsymbol{\theta}_{jt} - \boldsymbol{H}'_{jt}\boldsymbol{\theta}_t)^2/v_{jt} \\
&\quad + (\boldsymbol{\theta}_{jt} - (\boldsymbol{a}_{jt} + \boldsymbol{D}_{jt}(\boldsymbol{\theta}_t - \boldsymbol{a}_t))'\boldsymbol{V}_{jt}^{-1}(\boldsymbol{\theta}_{jt} - (\boldsymbol{a}_{jt} + \boldsymbol{D}_{jt}(\boldsymbol{\theta}_t - \boldsymbol{a}_t)))]\Big\}, \\
&\propto exp\Big\{ -\frac{1}{2}\big[\boldsymbol{\theta}'_{jt}(\boldsymbol{F}_{jt}\boldsymbol{F}_{jt}/v_{jt} + \boldsymbol{V}_{jt}^{-1})\boldsymbol{\theta}_{jt} \\
&\quad - 2[y_{jt}\boldsymbol{F}'_{jt}/v_{jt} - \boldsymbol{H}'_{jt}\boldsymbol{\theta}_t\boldsymbol{F}'_{jt}/v_{jt} + (\boldsymbol{a}'_{jt} + (\boldsymbol{\theta}'_t - \boldsymbol{a}'_t)\boldsymbol{D}'_{jt})\boldsymbol{V}_{jt}^{-1}]\boldsymbol{\theta}_{jt}\big]\Big\}.
\end{aligned}$$

Therefore, I have

$$(\boldsymbol{\theta}_{jt}|\boldsymbol{\theta}_t, \mathcal{D}_t) \sim N(\boldsymbol{m}_{jt} + \boldsymbol{B}_{jt}(\boldsymbol{\theta}_t - \boldsymbol{m}_t), \boldsymbol{U}_{jt}),$$

where

$$\begin{aligned}
\boldsymbol{m}_{jt} + \boldsymbol{B}_{jt}(\boldsymbol{\theta}_t - \boldsymbol{m}_t) = \boldsymbol{U}_{jt}[(&\boldsymbol{V}_{jt}^{-1}\boldsymbol{D}_{jt} - \boldsymbol{F}_{jt}\boldsymbol{H}'_{jt}/v_{jt})\boldsymbol{\theta}_t + \frac{y_{jt}}{v_{jt}}\boldsymbol{F}_{jt} - \boldsymbol{V}_{jt}^{-1}\boldsymbol{D}_{jt}\boldsymbol{a}_t \\
&+ \boldsymbol{V}_{jt}^{-1}\boldsymbol{a}_{jt}],
\end{aligned}$$

with

$$\begin{aligned}
\boldsymbol{U}_{jt} &= (\boldsymbol{F}_{jt}\boldsymbol{F}'_{jt}/v_{jt} + \boldsymbol{V}_{jt}^{-1})^{-1}, \\
\boldsymbol{B}_{jt} &= \boldsymbol{U}_{jt}(\boldsymbol{V}_{jt}^{-1}\boldsymbol{D}_{jt} - \boldsymbol{F}_{jt}\boldsymbol{H}'_{jt}/v_{jt}),
\end{aligned}$$

so that

$$\boldsymbol{m}_{jt} = \boldsymbol{B}_{jt}\boldsymbol{m}_t + \boldsymbol{U}_{jt}(\frac{y_{jt}}{v_{jt}}\boldsymbol{F}_{jt} + \boldsymbol{V}_{jt}^{-1}(\boldsymbol{a}_{jt} - \boldsymbol{D}_{jt}\boldsymbol{a}_t)).$$

$\square$

Proof of Theorem 5.2.5. I first prove that

$$p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, \boldsymbol{\theta}_{1:m,t+1}, \mathcal{D}_t) = p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, \mathcal{D}_t),$$

which is equivalent to

$$\boldsymbol{\theta}_t \perp \boldsymbol{\theta}_{j,t+1} | \boldsymbol{\theta}_{t+1}, \mathcal{D}_t, \text{ for } j = 1, \cdots, m.$$

It follows that

$$C(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{j,t+1} | \boldsymbol{\theta}_{t+1}, \mathcal{D}_t) = C(E(\boldsymbol{\theta}_{j,t+1} | \boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}, \mathcal{D}_t), E(\boldsymbol{\theta}_t | \boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}, \mathcal{D}_t) | \boldsymbol{\theta}_{t+1}, \mathcal{D}_t)$$

where

$$\begin{aligned}
E(\boldsymbol{\theta}_{j,t+1} | \boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}, \mathcal{D}_t) &= E(E(\boldsymbol{\theta}_{j,t+1} | \boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}, \boldsymbol{\theta}_{jt}, \mathcal{D}_t) | \boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}, \mathcal{D}_t) \\
&= E(\boldsymbol{G}_{j,t+1} \boldsymbol{\theta}_{jt} + \boldsymbol{D}_{j,t+1} (\boldsymbol{\theta}_{t+1} - \boldsymbol{G}_{t+1} \boldsymbol{\theta}_t) | \boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}, \mathcal{D}_t) \\
&= E(\boldsymbol{G}_{j,t+1} [\boldsymbol{m}_{jt} + \boldsymbol{B}_{jt} (\boldsymbol{\theta}_t - \boldsymbol{m}_t)] \\
&\quad + \boldsymbol{D}_{j,t+1} (\boldsymbol{\theta}_{t+1} - \boldsymbol{G}_{t+1} \boldsymbol{\theta}_t) | \boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}, \mathcal{D}_t) \\
&= E(\boldsymbol{G}_{j,t+1} (\boldsymbol{m}_{jt} - \boldsymbol{B}_{jt} \boldsymbol{m}_t) + \boldsymbol{D}_{j,t+1} \boldsymbol{\theta}_{t+1} | \boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}, \mathcal{D}_t)
\end{aligned}$$

Therefore,

$$C(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{j,t+1} | \boldsymbol{\theta}_{t+1}, \mathcal{D}_t) = \mathbf{0},$$

which means

$$\boldsymbol{\theta}_t \perp \boldsymbol{\theta}_{j,t+1} | \boldsymbol{\theta}_{t+1}, \mathcal{D}_t, \text{ for } j = 1, \cdots, m.$$

First, the proof of $(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t+1}, \mathcal{D}_t) \sim N[\boldsymbol{h}_t(-1), \boldsymbol{H}_t(-1)]$ can be easily shown using conditional normal theory (West and Harrison, 1997, p. 115).

Secondly, I need to show that $(\boldsymbol{\theta}_{jt} | \boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}, \boldsymbol{\theta}_{j,t+1}, \mathcal{D}_t) \sim N[\boldsymbol{h}_{jt}(-1), \boldsymbol{H}_{jt}(-1)]$. The expectation $E(\boldsymbol{\theta}_{jt} | \boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}, \boldsymbol{\theta}_{j,t+1}, \mathcal{D}_t)$ is

$$E[\boldsymbol{G}_{j,t+1}^{-1} (\boldsymbol{\theta}_{j,t+1} - \boldsymbol{D}_{j,t+1} (\boldsymbol{\theta}_{t+1} - \boldsymbol{G}_{t+1} \boldsymbol{\theta}_t)) | \boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}, \boldsymbol{\theta}_{j,t+1}, \mathcal{D}_t] = \boldsymbol{h}_{jt}(-1).$$

Let $\boldsymbol{\epsilon}_{j,t+1} = \boldsymbol{w}_{j,t+1} - \boldsymbol{D}_{j,t+1} \boldsymbol{w}_{t+1}$. It follows that $\boldsymbol{\epsilon}_{j,t+1} \perp \boldsymbol{\theta}_t | (\boldsymbol{\theta}_{t+1}, \mathcal{D}_t)$ and $\boldsymbol{\theta}_{j,t+1} \perp$

$\boldsymbol{\theta}_t|(\boldsymbol{\theta}_{t+1}, \mathcal{D}_t)$. Then the covariance is

$$
\begin{aligned}
C(\boldsymbol{\theta}_{jt}|\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}, \boldsymbol{\theta}_{j,t+1}, \mathcal{D}_t) &= C[\boldsymbol{G}_{j,t+1}^{-1}\boldsymbol{w}_{j,t+1}|\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}, \boldsymbol{\theta}_{j,t+1}, \mathcal{D}_t] \\
&= \boldsymbol{G}_{j,t+1}^{-1} C(\boldsymbol{\epsilon}_{j,t+1}|\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}, \boldsymbol{\theta}_{j,t+1}, \mathcal{D}_t)\boldsymbol{G}_{j,t+1}^{\prime -1} \\
&= \boldsymbol{G}_{j,t+1}^{-1} C(\boldsymbol{\epsilon}_{j,t+1}|\boldsymbol{\theta}_{t+1}, \boldsymbol{\theta}_{j,t+1}, \mathcal{D}_t)\boldsymbol{G}_{j,t+1}^{\prime -1} \\
&= \boldsymbol{H}_{jt}(-1).
\end{aligned}
$$

Specifically in the case of $\boldsymbol{S}_{j,t+1} = \boldsymbol{G}_{j,t+1}\boldsymbol{U}_{jt}\boldsymbol{G}_{j,t+1}^{\prime}(\delta_j^{-1} - 1)$, I have $\boldsymbol{H}_{jt}(-1) = (1 - \delta_j)\boldsymbol{U}_{jt}$. $\square$

Assumptions for missing values

- If $\boldsymbol{F}_{jt}$ or y_{jt} is partially or completely missing, I assume $\boldsymbol{U}_{jt} = \boldsymbol{V}_{jt}, \boldsymbol{B}_{jt} = \boldsymbol{D}_{jt}$, which means $(\boldsymbol{\theta}_{jt}|\boldsymbol{\theta}_t, \mathcal{D}_t)$ follows the same distribution as $(\boldsymbol{\theta}_{jt}|\boldsymbol{\theta}_t, \mathcal{D}_{t-1})$.
- If $\boldsymbol{F}_{jt}$ is partially or completely missing, I assume $(y_{jt}|\boldsymbol{y}_{pa(j),t}, \mathcal{D}_{t-1})$ has the same distribution as $(y_{j,t-1}|\mathcal{D}_{t-1})$.

6

Concluding Remarks

It is increasingly understood that sparsity network modeling is essentially appealing and almost necessary in high-dimensional data analysis. Thus the concept of sparsity in conditional dependence networks of MDMs is proposed, and the new area of developing adaptive approaches to identifying sparse network structures is explored. The first step I take is to stochastically search over network model space as discussed in Chapter 3. Novel SSS methods achieve success in identifying combinations of conditional networks and discount factors, and resulting models with attractive forecasting accuracy leading to profitable portfolio decisions.

BMA methods investigated in Chapter 4 provide robust and accurate out-of-sample predictive performance. The goal is to enable conditional dependence network structures– as well as other model parameters– to dynamically adapt to the new information arriving over time. Power discounting techniques are utilized and facilitate the adaptation to information change. A study of financial time series forecasting and portfolio analysis is conducted in Chapter 4, and shows that extended BMA on MDMs results in significant practical improvements in predictions and decisions.

Another indication from the example in Chapter 4 is that extended BMA for MDMs is feasible for higher-dimensional time series data, based on parallelization. If GPU computing is developed, significant increases in scales are likely to be available. GPU implementation that allows efficient online learning of model parameters serves as a comprehensive future research area to advance this overall research agenda.

Notice that although the order of the variables may have substantial impact on the accuracy and stability of the forecasts, this thesis only focuses on illustrating the impact of AR lag order, series-specific discount factors and sparse parental sets on forecasting in MDMs. Future research related to the ordering of the variables might define an additional next-stage research agenda.

Viewing the MDM from a completely new perspective - a reformulation of multivariate dynamic Cholesky volatility– I finally extend it to be coupled with common states shared across series. With creative analytic approximation, this model enables closed-form, online learning in the context of realistic, practical assumptions about volatility processes. When applied to short-term prediction of electricity demand in a Smart Grid scenario, this method achieves improvements in general forecasting performance of the daily load curve. The comparison with MCMC validates the efficacy of the analysis and overall approach. My new model framework provides additional advantages. First, the model enables intraday forecasts utilizing the same day load curve information, thus enhancing the usability of end users. Second, the online method is relatively robust to choices of tuning parameters.

In the era of big data, multiregression dynamic models for multivariate time series analysis is still a currently underdeveloped yet promising area. The approach essentially decouples multivariate state space models into sets of univariate models to be processed in parallel, and recouples the results for forecasting and decisions. The parallel structure underlying MDMs is the key and critical feature of the approach, and underlies the significant advances of this thesis. I anticipate the concept and aspects of the technical innovation here will underlie further research advances in scaling and applying multivariate dynamic Bayesian models.

Bibliography

Aguilar, O. and West, M. (2000), “Bayesian dynamic factor models and variance matrix discounting for portfolio allocation,” *Journal of Business and Economic Statistics*, 18, 338–357.

Anacleto, O., Queen, C. M., and Albers, C. J. (2013), “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, Issue 2, 251–270.

Bollerslev, T., Chou, R. Y., and Kroner, K. F. (1992), “ARCH modeling in finance,” *Journal of Econometrics*, 52, 5–59.

Carter, C. K. and Kohn, R. (1994), “On Gibbs sampling for state space models,” *Biometrika*, 81, 541–553.

Carvalho, C. M. and West, M. (2007), “Dynamic matrix-variate graphical models,” *Bayesian Analysis*, 2, 69–98.

Chib, S., Nardari, F., and Shephard, N. (2006), “Analysis of high dimensional multivariate stochastic volatility models,” *Journal of Econometrics*, 134, 341–371.

Cottet, R. and Smith, M. (2003), “Bayesian modeling and forecasting of intraday electricity load,” *Journal of the American Statistical Association*, 98, 839–849.

Dordonnat, V., Koopman, S., Ooms, M., Dessertaine, A., and Collet, J. (2008), “An hourly periodic state space model for modelling French national electricity load,” *International Journal of Forecasting*, 24, 566–587.

Doz, C. and Renault, E. (2006), “Factor stochastic volatility in mean models: A GMM approach,” *Econometric Reviews*, 25, 275–309.

Frühwirth-Schnatter, S. (1994), “Data augmentation and dynamic linear models,” *Journal of Time Series Analysis*, 15, 183 – 202.

Gamerman, D. and Migon, H. S. (1993), “Dynamic hierarchical models,” *Journal of the Royal Statistical Society. Series B (Methodological)*, 55, 629–642.

Hammerstrom, D. J., Ambrosio, R., Carlon, T., DeSteese, J., Horst, G., Kajfasz, R., Kiesling, L., Michie, P., Pratt, R., Yao, M., Brous, J., Chassin, D., Guttromson, R., Jarvegren, O., Katipamula, S., Le, N., Oliver, T., and Thompson, S. (2007), "Pacific northwest gridwise testbed demonstration projects. Part I. Olympic peninsula project," Report pnnl-17167, Pacific Northwest National Laboratory, Richland, Washington.

Hang, D. (2011), "High-frequency Bayesian modeling and analysis of stochastic volatility in finance," Duke University, Department of Statistical Science: Undergraduate Thesis.

Hans, C., Dobra, A., and West, M. (2007), "Shotgun stochastic search for "large p" regression," *Journal of the American Statistical Association*, 102, 507–516.

Harvey, A. and Koopman, S. (1993), "Forecasting hourly electricity demand using time-varying splines," *Journal of the American Statistical Association*, 88, 1228–1236.

Hosking, J. R. M., Natarajan, R., Ghosh, S., Subramanian, S., and Zhang, X. (2013), "Short-term forecasting of the daily load curve for residential electricity usage in the Smart Grid," *Applied Stochastic Models in Business and Industry*, 29, 604–620.

Jones, B., Carvalho, C. M., Dobra, A., Hans, C., Carter, C., and West, M. (2005), "Experiments in stochastic computation for high-dimensional graphical models," *Statistical Science*, 20, 388–400.

Lopes, H. F., McCulloch, R. E., and Tsay, R. (2012), "Cholesky stochastic volatility models for high-dimensional time series," Tech. rep., University of Chicago, Booth Business School.

Markowitz, H. (1959), *Portfolio Selection: Efficient Diversification of Investments*, John Wiley & Sons, Inc., New York Chapman & Hall, Limited, London.

Nakajima, J. and West, M. (2013), "Bayesian dynamic factor models: Latent threshold approach," *Journal of Financial Econometrics*, 11, 116–153.

Prado, R. and West, M. (2010), *Time Series Modeling, Computation, and Inference*, Chapman and Hall/CRC, Taylor and Francis Group.

Primiceri, G. E. (2005), "Time varying structural vector autoregressions and monetary policy," *Review of Economic Studies*, 72, 821–852.

Queen, C. M. (1994), "Using the multiregression dynamic model to forecast brand sales in a competitive product market," *Journal of the Royal Statistical Society. Series D (The Statistician)*, 43, No. 1, 87–98.

Queen, C. M. and Smith, J. Q. (1993), "Multiregression dynamic models," *Journal of the Royal Statistical Society. Series B (Methodological)*, 55, No. 4, 849–870.

Queen, C. M., Wright, B. J., and Albers, C. J. (2008), "Forecast covariances in the linear multiregression dynamic model," *Journal of Forecasting*, 27, 175–191.

Quintana, J. M. (1992), "Optimal portfolios of forward currency contracts," in *Bayesian Statistics IV*, eds. J. Berger, J. Bernardo, A. Dawid, and A. Smith, pp. 753–762, Oxford University Press.

Quintana, J. M. and West, M. (1987), "An analysis of international exchange rates using multivariate DLM's," *The Statistician*, 36, 275–281.

Quintana, J. M. and West, M. (1988), "Time series analysis of compositional data," in *Bayesian Statistics 3*, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, pp. 747–756, Oxford University Press.

Quintana, J. M., Lourdes, V., and Aguilar, O.and Liu, J. (2003), "Global gambling," in *Bayesian Statistics VII*, eds. J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West, pp. 349–367, Oxford University Press.

Shephard, N. (2005), *Stochastic Volatility: Selected Readings*, Oxford University Press.

Sims, C. A. and Zha, T. (1998), "Bayesian methods for dynamic multivariate models," *International Economic Review*, pp. 949–968.

Sims, C. A. and Zha, T. (2006), "Were there regime switches in U.S. monetary policy?" *American Economic Review*, 96, 54–81.

Smith, M. and Kohn, R. (2002), "Parsimonious covariance matrix estimation for longitudinal data," *Journal of the American Statistical Association*, 97, No. 460, 1141–1153.

Taylor, J. W., De Menezes, L. M., and McSharry, P. E. (2006), "A comparison of univariate methods for forecasting electricity demand up to a day ahead," *International Journal of Forecasting*, 22, 1–16.

Triantafyllopoulos, K. (2007), "Covariance estimation for multivariate conditionally Gaussian dynamic linear models," *Journal of Forecasting*, 26(8), 551–569.

Uhlig, H. (1994), "On singular Wishart and singular multivariate beta distributions," *The Annals of Statistics*, 22, No.1, 395–405.

Wang, H. and West, M. (2009), "Bayesian analysis of matrix normal graphical models," *Biometrika*, 96, 821–834.

West, M. and Harrison, P. J. (1989), *Bayesian Forecasting and Dynamic Models*, Springer, 1st edn.

West, M. and Harrison, P. J. (1997), *Bayesian Forecasting and Dynamic Models*, Springer, 2nd edn.

Xie, M. (2013), "Discount-weighted Bayesian model averaging for portfolio decisions in matrix variate dynamic linear models," Duke University, Department of Statistical Science: Undergraduate Thesis.

Zhou, X., Nakajima, J., and West, M. (2014), "Dynamic dependent factor models: improving forecasts and portfolio decisions in financial time series," *International Journal of Forecasting*, 30, 963–980.

Biography

Zoey Yi Zhao was born in October 1988 in Beijing, China. She received her bachelor's degree in Mathematics and Applied Mathematics from Peking University in July 2011. In December 2014 she earned a Master's degree in Statistical Science en route to her Ph.D at Duke University. She graduated with a Doctor of Philosophy in Statistical Science under the supervision of Professor Mike West in January 2015.