Applications of and extensions to state-space models

David Allen Osthus
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

Part of the Statistics and Probability Commons

Recommended Citation
Osthus, David Allen, "Applications of and extensions to state-space models" (2015). Graduate Theses and Dissertations. 14623.
https://lib.dr.iastate.edu/etd/14623

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Applications of and extensions to state-space models

by

David Allen Osthus

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Petruța C. Caragea, Major Professor
Alicia Carriquiry
Jarad Niemi
Dan Nordman
Vivekananda Roy

Iowa State University
Ames, Iowa
2015
Copyright © David Allen Osthus, 2015. All rights reserved.
DEDICATION

To Grandma Osthus. Your unwavering support and encouragement meant the world to me.
# TABLE OF CONTENTS

**LIST OF FIGURES** ................................................................. vi

**LIST OF TABLES** ............................................................... xii

**ACKNOWLEDGEMENTS** ....................................................... xv

**ABSTRACT** ............................................................................. xvi

**CHAPTER 1. GENERAL INTRODUCTION** ................................. 1
  1.1 Introduction ........................................................................ 1
  1.2 Dissertation Organization ................................................ 3

**CHAPTER 2. DYNAMIC LINEAR MODELS FOR FORECASTING OF RADIATION BELT ELECTRONS AND LIMITATIONS ON PHYSICAL INTERPRETATION OF PREDICTIVE MODELS** .......... 5
  2.1 Introduction ........................................................................ 5
  2.2 Data Description ............................................................... 7
  2.3 Parameter Estimate Sensitivity to Training Data ................... 10
  2.4 Multi-Parameter Estimates and Sensitivity to Multicollinearity .... 15
  2.5 Dynamic Linear Models .................................................... 22
  2.6 Forecasting ......................................................................... 29
  2.7 Discussion .......................................................................... 34

**CHAPTER 3. MULTISCALE DYNAMIC LINEAR MODELS** ............ 37
  3.1 Introduction ........................................................................ 37
  3.2 Multiscale Motivating Example .......................................... 39
  3.3 Directed Graphical Models ............................................... 43
3.3.1 Dynamic Linear Models ........................................... 44
3.3.2 Multiscale Dynamic Linear Models ................................. 46
3.4 Directed Global Markov Property ...................................... 50
3.5 Estimation and Inference: Parameters Known .......................... 52
  3.5.1 Filtering .............................................................. 53
  3.5.2 Smoothing ............................................................ 58
  3.5.3 Forecasting ........................................................... 61
3.6 Estimation and Inference: Parameters Unknown ...................... 63
  3.6.1 DLM Forward-Filtering, Backward Sampling ...................... 64
  3.6.2 MSDLM Forward-Filtering, Backward Sampling .................. 65
3.7 Application: Space Weather ............................................ 67
3.8 Discussion ............................................................... 71

CHAPTER 4. FORECASTING SEASONAL INFLUENZA WITH A STATE-SPACE SIR MODEL ................................................. 73
  4.1 Introduction .............................................................. 73
  4.2 Data Description .......................................................... 76
  4.3 Susceptible-Infectious-Recovered (SIR) Model .......................... 78
  4.4 Dirichlet-Beta State-Space Model (DBSSM) ............................. 80
  4.5 Posterior Simulation and Forecasting .................................. 83
  4.6 Prior Specification of \( \phi \) ............................................. 85
    4.6.1 Specification of \([ \kappa ]\) and \([ \lambda ]\) .......................... 88
    4.6.2 Specification of \([ \theta_0 ]\) ........................................ 88
    4.6.3 Specification of \([ z | \theta_0 ]\) ..................................... 90
    4.6.4 Specification of \([ \rho | z, \theta_0 ]\) .................................. 92
    4.6.5 Specification of \([ \beta | \rho, z, \theta_0 ]\) ............................ 93
    4.6.6 Prior Predictive Distribution for \([ y_{1:T} ]\) ....................... 95
4.7 Forecasting Results ................................................. 97
  4.7.1 Forecasting Illustration ................................. 97
  4.7.2 Forecasting Comparison ............................ 103
4.8 Discussion ..................................................... 109

CHAPTER 5. GENERAL CONCLUSION ............................. 112
  5.1 General Discussion .................................... 112
  5.2 Directions for Future Work ............................. 112

APPENDIX A. MULTICOLLINEARITY ILLUSTRATION .......... 115

APPENDIX B. FILTERING, FORECASTING, AND SMOOTHING DETAILS ............................................. 119

APPENDIX C. PROOF OF THE DYNAMIC LINEAR MODEL FILTERING EQUATIONS (PROPOSITION 3.2) ............................. 122

APPENDIX D. PROOF OF THE MULTISCALE DYNAMIC LINEAR MODEL FILTERING EQUATIONS (PROPOSITION 3.3) ............. 124

APPENDIX E. PROOF OF THE DYNAMIC LINEAR MODEL SMOOTHING EQUATIONS (PROPOSITION 3.4) ....................... 127

APPENDIX F. PROOF OF THE MULTISCALE DYNAMIC LINEAR MODEL SMOOTHING EQUATIONS (PROPOSITION 3.5) ............. 129

APPENDIX G. PROOF OF THE DYNAMIC LINEAR MODEL FORECASTING RECURSIONS (PROPOSITION 3.6) .................... 133

APPENDIX H. PROOF OF THE MULTISCALE DYNAMIC LINEAR MODEL FORECASTING RECURSIONS (PROPOSITION 3.7) ............. 134

APPENDIX I. THE DYNAMIC LINEAR MODEL REPRESENTATION OF THE MULTISCALE DYNAMIC LINEAR MODEL ............................. 137

BIBLIOGRAPHY .................................................. 145
**LIST OF FIGURES**

| Figure 1.1 | Directed graph corresponding to state-space models. | 2 |
| Figure 1.2 | Venn diagram of various directed graphical models. | 3 |
| Figure 2.1 | (top) Log electron flux \((\text{cm}^2/\text{s/sr/kev})^{-1}\) versus solar wind speed (km/s). A triangular distribution is evident, with a solar wind dependent lower bound on log electron flux, but a solar wind independent upper bound on log electron flux. (bottom) Marginal histograms of solar wind speed (left) and log electron flux (right). The distribution of solar wind speed is unimodal and skewed right, ranging from 255.2 to 1103.0 km/s. The distribution of log electron flux is unimodal and roughly symmetric, ranging from -2.38 to 1.46 \((\text{cm}^2/\text{s/sr/kev})^{-1}\). | 9 |
| Figure 2.2 | Intercept (top) and slope (bottom) parameter estimates (points) with 95 percent confidence intervals (vertical segments) by year for Model D1. Parameter estimates are sensitive to the choice of training data and appear to vary systematically over time. | 12 |
| Figure 2.3 | Observed log electron flux (black), predicted log electron flux using training data from 1995 (red), 2000 (blue), and all years (green) for February and March in 1995 (top) and 2000 (bottom). | 14 |
Figure 2.4 Intercept (top) and slope (bottom) parameter estimates (points) with 95 percent confidence intervals (vertical segments) for models D1 and D2. Intercept estimates for models D1 and D2 are quite similar for each year, as a result of centering solar wind speed. Slope parameter estimates are considerably different in Model D1 and D2 as a result of multicollinearity.

Figure 2.5 Posterior means of smoothed states (DLM) and parameter estimates (points) with 95 percent confidence intervals (vertical segments) (D1) for intercept (top) and slope (bottom) parameters. The posterior means of the smoothed states line up very closely with the parameter estimates from Model D1.

Figure 3.1 Daily log electron flux from January 1st, 1990 - December 31st, 2009 (gray), monthly averages of log electron flux (black) and posterior means of $\theta_{0,t}$ (red).

Figure 3.2 Daily posterior means of $\theta_{0,t}$ (red) and monthly sunspot counts (black). Posterior means and sunspot counts are standardized for ease of comparison.

Figure 3.3 (top) Scatter plots of the posterior means of $\theta_{1,t}$ vs. $\theta_{0,t}$ for three selected months. (bottom) Daily log electron flux (gray) against time for July, 1996 - June 1997. The posterior means of the non-solar cycle specific intercept, $\theta_{0,t}$ (red), the solar cycle specific intercept, $\theta_{1,t}SSN_{t-1}$ (blue), and their sum, $\theta_{0,t} + \theta_{1,t}SSN_{t-1}$ (black) are plotted.

Figure 3.4 Directed acyclic graph.

Figure 3.5 The directed acyclic graphical structure for DLMs.

Figure 3.6 Generated time series from the local level DLM with $W = 0.1$, $V = 1$ (top) and $W = 2$, $V = 1$ (bottom). The black line represents the simulated observations, $y_{1:240}$, and the red line represents the latent states, $\theta_{0:240}$.
Figure 3.7 The directed acyclic graphical structure for MSDLMs. 47
Figure 3.8 Generated time series from the local level MSDLM with $S = 8$ and $T = 30$. The black line represents the simulated observations ($y_{s,t}$), the blue line represents the simulated coarse scale state ($\alpha_s$), and the red line represents the simulated coarse plus fine scale states ($\alpha_s + \beta_{s,t}$). $W_\alpha = 0.2$, $W_\beta = 0.01$, and $V = 1$. 49
Figure 3.9 Generated time series from the local level MSDLM with $S = 8$ and $T = 30$. The black line represents the simulated observations ($y_{s,t}$), the blue line represents the simulated coarse scale state ($\alpha_s$), and the red line represents the simulated coarse plus fine scale states ($\alpha_s + \beta_{s,t}$). $W_\alpha = 0.25$, $W_\beta = 0.5$, and $V = 1$. 50
Figure 3.10 $G$ 51
Figure 3.11 $G_m$ 51
Figure 3.12 $G$ 52
Figure 3.13 $G_{an(e\cup b\cup c)}$ 52
Figure 3.14 $(G_{an(e\cup b\cup c)})^m$ 52
Figure 3.15 DLM filtering (top) and one-step-ahead forecasts (bottom). Observations are points. Filtering means (top) and point forecasts (bottom) are black lines. The gray bands indicate 95% probability limits. 55
Figure 3.16 MSDLM filtering for only the coarse scale state (top), the coarse scale state plus the fine scale state (middle) and one-step-ahead forecasts (bottom). Observations are points. Filtering means (top and middle) and point forecasts (bottom) are black lines. Blue lines indicate the true coarse scale latent state, $\alpha_s$. Red lines indicate the true latent state, $\alpha_s + \beta_{s,t}$. Gray bands indicate 95% probability limits. 57
Figure 3.17 DLM smoothing distributions. Observations are points. The smoothing means and 95% probability limit are shown by the black line and gray bands, respectively. 58
Figure 3.18 MSDLM smoothing distributions for the coarse scale state (top) and coarse plus fine scale states (bottom). Observations are points. The smoothing means and 95% probability limit are shown by the black line and gray bands, respectively. The true, latent state for $\alpha_s$ is the blue line (top) and for $\alpha_s + \beta_{s,t}$ is the red line (bottom).

Figure 3.19 (top) Posterior means for $\alpha_{1,s}SSN_{s-1}$ (black), posterior means for $\theta_{0,t}$ from model 3.1 (red), and log electron flux observations (gray). (bottom) Posterior means for $\alpha_{1,s}SSN_{s-1} + \beta_{1,s,t}$ (black), posterior means for $\theta_{0,t}$ from model 3.1 (red), and log electron flux observations (gray).

Figure 3.20 (top) Posterior means for coarse scale intercept, $\alpha_{1,s}SSN_{s-1}$ (blue), posterior means for fine scale intercept, $\alpha_{1,s}SSN_{s-1} + \beta_{1,s,t}$ (red), and log electron flux (gray). (bottom) Posterior means (black) and 95% probability limits (gray band) for the fine scale intercept, $\alpha_{1,s}SSN_{s-1} + \beta_{1,s,t}$, posterior means for the fine scale intercept of model 3.1, $\theta_{0,t}$ (red), and log electron flux observations (gray).

Figure 4.1 ILI+ for influenza seasons 2002–2007 and 2010–2013. Weeks 1 and 35 roughly correspond to the beginning of October and the end of May, respectively.

Figure 4.2 Simulated SIR curve with $S(0) = 0.9$, $I(0) = 0.0002$, $R(0) = 0.0998$, $\beta = 2$, and $\gamma = 1.4$.

Figure 4.3 Three SIR curves simulated from three distinct sets of initial conditions and parameters. For all curves, $I(0) = 0.0002$. The black, red, and blue curves were simulated with $(S(0), \beta, \gamma) = (0.9, 2, 1.4), (0.65, 2.38, 1.145)$, and $(0.4, 3.124, 0.842)$, respectively.
Figure 4.4 Two solutions to the SIR model are displayed. The green solution (top) is $S(0) = 0.9$, $I(0) = 0.000172$, $R(0) = 0.099828$, $\beta = 2.21$, $\gamma = 1.7017$. The red solution (bottom) is $S(0) = 0.9$, $I(0) = 0.00018$, $R(0) = 0.09982$, $\beta = 0.39$, $\gamma = 0.078$. Points are 2010-2011 ILI+ observations for weeks one through nine. Grey lines are the ILI+ trajectories for years 2002-2007 and 2011-2013 for weeks 10 through 35.

Figure 4.5 Historical ILI+ on week 0 (vertical dashes) and the fitted Beta(1.62, 7084.10) prior.

Figure 4.6 The peak intensity (PI) on the y-axis versus the timing of peak intensity (PT) on the x-axis for years 2002-2007 and 2010-2013 (black points). 10,000 samples were drawn from the truncated normal distribution of equation 4.24 with $I(0) = 0.0002$ and plotted in grey.

Figure 4.7 The relationship between PI and $\rho$, for $S(0) = 0.9$ and $I(0) = 0.0002$ (black line). The grey vertical line represents the range of historical PIs. The grey horizontal line shows the range of corresponding $\rho$ values.

Figure 4.8 Five infectious curves simulated from the SIR model. For all curves, $S(0) = 0.9$, $I(0) = 0.0002$, and $\rho = 0.7$. The parameter $\beta$ varies from 0.25 to 4. Grey vertical lines denote the range of historically observed PTs.

Figure 4.9 120 SIR curves were simulated and the PT was recorded (points). For each of the of the six $I(0)$ and $\rho$ combinations, $log(\beta)$ was regressed on $log(PT)$. The expectation lines are overlaid.

Figure 4.10 The median (black line) and 95% prediction intervals (grey band) based on $M = 5000$ draws from the prior predictive distribution. Historical ILI+ observations are displayed for reference.
Figure 4.11 95% prediction intervals for \([y(t'+1:T)|y_{1:t'}]\) (red bands) and 95% credible intervals for \([I(0), I(1), \ldots, I(t')|y_{1:t'}]\) (blue bands). Medians (black lines) and ILI+ observations (points) are also displayed for the 2010-2011 (top) and 2003-2004 (bottom) nationwide, influenza seasons. The number at the top of each panel is \(t'\). All plots are based on 5000 simulations.

Figure 4.12 95% prediction intervals (line segments) and medians (points) for \([PI_{t'}|y_{1:t'}]\) (left) and \([PT_{t'}|y_{1:t'}]\) (right). Observed PI and PT for the nationwide 2010-2011 influenza season are displayed as red lines.

Figure 4.13 \(\hat{P}(PT > t'|y_{1:t'})\) versus \(t'\) for the 2010-2011 influenza season. The red point indicates the PT. The grey horizontal line represents \(I(PT > t')\).

Figure 4.14 Distribution of M1, M2, and M3 scores for the DBSSM (top) and SM2 (bottom).

Figure 4.15 \(\hat{P}(PT > t'|y_{1:t'})\) averaged over all scenarios versus standardized time. The grey horizontal line represents \(I(PT > t') = I(0 > \text{Standardized Time})\).

Figure 5.1 A possible DAG representation that incorporates two sources of surveillance data.

Figure A.1 Histograms of parameter estimates from Model S1 simulated data sets. The top row corresponds to the uncorrelated covariate scenario; the bottom row to the highly-linearly correlated covariate scenario. The columns correspond to the model that was fit (A, B, or C) and the estimated parameter (Beta 1 or 2). Red vertical lines indicate the true value of the parameter.
Figure A.2  Histograms of parameter estimates from Model S2 simulated data sets. The top row corresponds to the uncorrelated covariate scenario; the bottom row to the highly-linearly correlated covariate scenario. The columns correspond to the model that was fit (A, B, or C) and the estimated parameter (Beta 1 or 2). Red vertical lines indicate the true value of the parameter. ................................. 118

Figure C.1  The moralized graph of Figure 3.5 ................................. 122

Figure D.1  The moralized graph of Figure 3.7 ................................. 124
LIST OF TABLES

Table 2.1 Estimates for $\beta_1$ and 95% confidence intervals (CI) from Models D1, D3, D4, and D5 fit to 1996, 2000, and 2009. Correlations are between standardized solar wind speed and standardized ion density (Model D3) and standardized z-component of IMF (Model D4). Estimates for $\beta_1$ change appreciably between Model D1 and D3 when ion density is added to the model in 1996, as ion density is moderately correlated with solar wind speed. Little change to $\hat{\beta}_1$ is seen when the z-component of the IMF is added to the model, as it is weakly correlated to solar wind speed. 21

Table 2.2 Prediction efficiency (PE), mean absolute error (MAE), average prediction interval width (Width) and coverage for one day ahead forecasts for the persistence (P), REFM, and DLMforecast (DLM) models, by year. The number of days in each year where log electron flux was unobserved is presented in the column Missing. Almost half of all days with missing log electron fluxes occurred between 2006 and 2009. The DLMforecast model produces competitive forecasts with respect to both the persistence and REFM models, as seen by the similar PE and MAE values for all three models. The coverages are roughly equivalent for REFM and DLMforecast, while average prediction interval widths are generally smaller for DLMforecast. 33

Table 4.1 Health and Human Service regions. 76

Table 4.2 Polynomial regression of log($\beta$) on log(PT), log(I(0)), and log($\rho$). 95
Table 4.3  Percentage of scenarios each model was deemed best, by metric. Bold numbers indicate the largest percentage for each metric.
ACKNOWLEDGEMENTS

My sincerest thanks to Dr. Petruța Caragea, my major professor. I am grateful for her expertise, patience, guidance, and most importantly, her ability to humanize the at times overwhelming and discouraging research process. Our interactions always left me hopeful, inspired, and motivated.

Thanks to my numerous mentors and collaborators at Los Alamos National Laboratory, not the least of which include Dr. Dave Higdon, Dr. Brian Weaver, and Dr. Kyle Hickmann. The passion and excitement they have for their work is infectious.

Many thanks to my committee, Dr. Alicia Carriquiry, Dr. Jarad Niemi, Dr. Dan Nordman, and Dr. Vivekananda Roy, for their helpful questions, comments, and suggestions throughout this process.

I was fortunate to spend multiple enriching years working as a graduate research assistant under the advisement of Dr. Alicia Carriquiry, Dr. Sarah Nusser, and Dr. Wayne Fuller. I thank them for their mentorship. A special thanks to Dr. Fuller for so generously sharing his time, knowledge, stories, and wisdom. Through this research assistantship, I was afforded the opportunity to work with Dr. Bryan Stanfill. The gratitude I have for his friendship, scholarship, and candidness cannot be overstated.

The phrase no one achieves anything alone rang loud and true in graduate school. Thanks to all of my classmates and friends who valued communal success over individual achievement. Graduate school is hard. It would be even harder to do alone.

Finally, I extend my most heartfelt appreciation to my family, particularly my fiancée, Lynne, my mom, Tracie, my brother and sister-in-law, Scott and Tezra, and my grandparents, Bea, Betty, and Roger. They were with me at my worst, encouraged me at my lowest, and instilled in me a sense of confidence and self-worth I will never be able to repay.
ABSTRACT

State-space models have proven invaluable in the analysis of dynamic data, specifically time series data. They provide a natural and interpretable framework to learn about and describe dynamic processes. State-space models also provide a flexible framework for embedding prescriptive, mathematical models in ways that account for multiple sources of uncertainty. When considered within the more general directed graphical model formalism, state-space models can be reimagined and extended into arenas beyond the reach of traditional state-space models. In three papers, we consider various applications of and extensions to state-space models. All papers stem from collaborations with Los Alamos National Laboratory.

In the field of space weather forecasting, many modeling approaches have been developed in the last 25 years. These approaches attempt to make sense of the dynamic and not-well-understood relationships between electron flux intensities and relevant covariates. Many of these forecasting models possess inherent limitations because they are static in nature and thus are constrained to customized and narrow time windows. In Chapter 2, we discuss these limitations and present an alternate approach to space weather forecasting utilizing dynamic linear models (DLMs). Benefits of dynamic modeling when compared to static modeling are discussed and ground work is laid for future dynamic forecasting endeavors in space weather. This work was published in the journal Space Weather under research article number 10.1002/2014SW001057 in June 2014.

Multiscale modeling involves decomposing or explicitly modeling processes that arise at multiple scales. In Chapter 3, we extend the DLM into the multiscale arena with the presentation of the multiscale dynamic linear model (MSDLM). We present the MSDLM within the directed graphical model formalism. In so doing, we provide the necessary background to consider multiscale modeling generally. The MSDLM is a multiscale time series model that is interpretable, flexible, and coherently combines multiple scales of information in a princi-
pled, unified framework. Estimation and sampling procedures are presented. We illustrate the efficacy of the MSDLM by revisiting the problem of space weather forecasting discussed in Chapter 2.

Forecasting seasonal influenza in the U.S. is challenging and consequential. It is challenging because there is uncertainty in the form of the disease transmission process, the process is only partially observed, and those observations are noisy. It is consequential because influenza poses serious risks to both national security and public health. In Chapter 4, we propose a non-Gaussian, nonlinear state-space model that embeds a compartmental model (i.e., a set of nonlinear, ordinary differential equations) into the state equations. The state-space framework provides valuable flexibility to the deterministic disease transmission process while simultaneously allowing and accounting for uncertainties in the parameters, the process, and the measurement mechanism. Prior specification is discussed in detail. Forecasting metrics are proposed and compared with competing models.
CHAPTER 1. GENERAL INTRODUCTION

1.1 Introduction

A state-space model is a conditionally specified model consisting of two equations and an initialization,

\[
y_t | \theta_t \overset{d}{=} [y_t | \theta_t] \quad \text{(observation equation)} \quad (1.1a)
\]

\[
\theta_t | \theta_{t-1} \overset{d}{=} [\theta_t | \theta_{t-1}] \quad \text{(state equation)} \quad (1.1b)
\]

\[
\theta_0 \overset{d}{=} [\theta_0]. \quad \text{(initialization)} \quad (1.1c)
\]

The latent state vector \( \theta_{0:T} = (\theta_0, \theta_1, \ldots, \theta_T) \) is a first-order Markov chain and \( y_{1:T} = (y_1, y_2, \ldots, y_T) \) is an observable time series where \( y_t \) is assumed conditionally independent of \( y_s \) for \( s \neq t \), given \( \theta_t \) (e.g. Petris et al., 2009, Chapter 2.3). We adopt the short-hand notation where \([X|Y]\) is the conditional density of the random variable \( X \) given \( Y \). The joint distribution for \((\theta_{0:T}, y_{1:T})\) is then,

\[
[\theta_{0:T}, y_{1:T}] = \prod_{t=1}^{T} [y_t | \theta_t][\theta_t | \theta_{t-1}]. \quad (1.2)
\]

State-space models are natural choices for time series analysis because of how they represent the flow of information within a system. Figure 1.1 displays the directed graphical representation corresponding to state-space models. Nodes (circles) represent random variables and directed edges (arrows) represent dependencies. That is, notions of causality from one node to another are represented graphically by directed edges. For time series analysis, we naturally think of information going from the past to the future. This notion is captured in the latent states of Figure 1.1, with directed edges going from \( \theta_{t-1} \) to \( \theta_t \), where \( t \) is indexing time. State-space models assume the true state of the system is not observed directly. Rather what
is observed is a noisy measurement, \( y_t \), that is related to the latent system. This notion is graphically represented with a directed edge going from \( \theta_t \) to \( y_t \).

![Directed graph corresponding to state-space models.](image)

Figure 1.1: Directed graph corresponding to state-space models.

Two popular special cases of state-space models include dynamic linear models (DLMs) (e.g., West and Harrison, 1997) and hidden Markov models (HMMs) (e.g., Petris et al., 2009, Chapter 2.6). DLMs and HMMs are special cases of state-space models because they share a common graphical representation (i.e., the graphical representation displayed in Figure 1.1) and joint distribution factorization (equation 1.2). What differentiates DLMs from HMMs and other state-space models are the distributional and support assumptions made to equations 1.1a - 1.1c. For instance, HMMs are state-space models where \( \theta_t \) is discrete. DLMs, however, are linear, Gaussian state-space models. They are specified as

\[
\begin{align*}
[y_t | \theta_t] &\overset{d}{=} N(F_t\theta_t, V_t) & \text{(observation equation)} \quad (1.3a) \\
[\theta_t | \theta_{t-1}] &\overset{d}{=} N(G_t\theta_{t-1}, W_t) & \text{(state equation)} \quad (1.3b) \\
[\theta_0] &\overset{d}{=} N(m_0, C_0), & \text{(initialization)} \quad (1.3c)
\end{align*}
\]

where \( F_t \) and \( G_t \) are known matrices of appropriate dimensions, \( V_t \) and \( W_t \) are known variance-covariance matrices, and \( m_0 \) and \( C_0 \) are the assumed known mean and variance of the initialization, respectively.

Just as DLMs and HMMs are special cases of state-space models, state-space models themselves are special cases of more general graphical models. For instance, a state-space model is a special case of the more general directed tree model (e.g., Ferreira and Lee, 2007, Chapter 6), while the directed tree model is a special case of the more general directed graphical model.
(e.g. Jordan, 2004). What differentiates state-space models, directed tree models, and directed graphical models is the underlying graphical representation. Details about directed graphical models are discussed in Chapter 3 with the presentation of the multiscale dynamic linear model (MSDLM). The MSDLM is a special case of a directed graphical model but is not a special case of either a state-space model or a directed tree model. Figure 1.2 provides a Venn diagram organizing these different models from the most general directed graphical model to the most specialized DLM and HMM. Figure 1.2 is by no means exhaustive, nor is it intended to be.

![Venn diagram of various directed graphical models.](image)

1.2 Dissertation Organization

This dissertation is a collection of three manuscripts tied together by their connections to state-space models. Chapter 2 is an applied paper stemming from collaborations with the Statistical Sciences Group and the Space Science and Applications Group at Los Alamos National Laboratory (LANL). This work investigates the relationship between relativistic electron flux and solar wind speed in the Van Allen radiation belts. The DLM is introduced and discussed. Comparisons with competing static models are made and the benefits of dynamic modeling are articulated. This work was published in the journal *Space Weather* under research article number 10.1002/2014SW001057 in June 2014. In Chapter 3, we extend the DLM into the multiscale arena with the presentation of the MSDLM. This methodological paper provides a
brief primer on directed graphical models and introduces the MSDLM within this formalism. Estimation and sampling procedures are presented. The efficacy of the MSDLM is illustrated by revisiting the problem of space weather forecasting discussed in Chapter 2. In Chapter 4, we consider forecasting seasonal influenza in the United States. This applied paper, in collaboration with the Statistical Sciences Group and the Mathematical and Computational Epidemiology Group at LANL, embeds a set of nonlinear, ordinary differential equations into a flexible state-space modeling framework. Features of this nonlinear, non-Gaussian state-space model are discussed. Prior specification and forecasting results are presented. General conclusions and future research directions for all chapters are discussed in Chapter 5.
CHAPTER 2. DYNAMIC LINEAR MODELS FOR FORECASTING OF RADIATION BELT ELECTRONS AND LIMITATIONS ON PHYSICAL INTERPRETATION OF PREDICTIVE MODELS

A paper published in *Space Weather*, June 2014
Dave Osthus, Petruța Caragea, Dave Higdon, Steve Morley, Geoff Reeves, Brian Weaver

2.1 Introduction

Understanding the dynamics in the radiation belts is consequential. High-energy electrons can be hazardous to satellites. Thus, providing satellite operators with accurate forecasts of electron flux intensities with actionable lead times is also consequential. Forecasting models that simultaneously elucidate important relationships between electron flux intensities and solar drivers (e.g., solar wind, density, magnetic field) and provide reliable, accurate forecasts are desired. This dual objective is not easily accomplished, however, as model simplicity/interpretability is often in conflict with complexity/accuracy.

Historically, many forecasting models have been developed, leveraging solar wind drivers to forecast electron flux, with an emphasis on forecast accuracy. Baker et al. (1990) used daily averages of the general planetary geomagnetic index, the AE index, and solar wind speed (Vsw) to predict electron flux, separately, via a linear prediction filtering technique. Rigler et al. (2007) extended the single-input linear prediction technique to a multi-input linear prediction model. They forecast electron flux as a function of daily averages of Vsw, the magnitude of the interplanetary magnetic field (IMF), and solar wind plasma density. Both the single-input and multi-input linear prediction filter models relate a set of solar wind drivers (covariates) to electron flux (response) via a set of parameters. What makes these models linear prediction filters
is that the function relating the solar wind drivers to electron flux is linear in the parameters. Li et al. (2001) and Li (2004) utilize a nonlinear model to forecast electron flux. Namely, Li et al. (2001) and Li (2004) leverage the radial diffusion equations to relate covariates (location, solar wind speed, interplanetary magnetic field, season, and solar cycle) to electron flux via a set of parameters. Turner and Li (2008) consider a nonlinear prediction filter that utilizes one to two day lagged lower energy electron flux from GEO to forecast relativistic flux (multi-MeV) at GEO. In both the linear prediction filter models and the radial diffusion model, a single set of parameters relates solar wind drivers to electron flux. That is, both of these models are static models, because the parameters do not vary in time.

Alternatively, Rigler et al. (2004) proposed the use of an adaptive linear prediction model based on the Kalman filtering equations. This approach, the authors note, can improve upon forecasting accuracy of electron flux because the parameters relating solar wind drivers to electron flux, known as states in the Kalman filtering literature, are allowed to vary in time. Kondrashov et al. (2007) and Shprits et al. (2007) utilized Kalman filtering based models to forecast relativistic electron lifetimes in the outer radiation belt. Sakaguchi et al. (2013) used a multivariate autoregressive model, combined with a Kalman filtering approach, to jointly model daily averages of solar wind speed, dynamic pressure, the z-component of the IMF, and electron flux. All of these Kalman filtering based models just mentioned treated error parameters and in the case of Sakaguchi et al. (2013), auto-regressive parameters, as known quantities within the Kalman filtering framework, rather than estimating them within this framework.

Almost all of the aforementioned forecasting models settle upon a large set of covariates that are numerous in distinct solar wind drivers and/or have long time lags. These large sets of covariates can produce accurate forecasts for a specific time window. Often, though, insight into the particular relationships between solar wind drivers and electron flux levels are uninterpretable and nonunique, due to the complex, tangled relationships that often accompany large sets of covariates. Thus, the conclusions about the relationships between solar wind drivers and electron flux intensities can be misleading.

In this paper we develop a dynamic linear model (DLM) framework to forecast electron flux using solar wind drivers. The DLM framework is similar to the framework of Kalman
filtering based models. It uses a simple model, whose parameter values evolve over time, using a principled, Bayesian estimation approach. This strategy offers practical and computational advantages over approaches that use a static set of model parameters, long time lags for solar wind drivers, or a customized time window over which to estimate model parameters. Thus, the DLM framework offers a new possibility to accomplish the dual objective of increased scientific understanding of radiation belt dynamics and more accurate space weather forecasts. To the knowledge of the authors, both the Bayesian estimation approach and the emphasis on the DLM’s potential to accomplish these dual objectives is novel.

This paper proceeds as follows. In Section 2.2, we introduce and describe the data. Section 2.3 discusses the sensitivity of parameter estimation to the choice of training data. In Section 2.4, we introduce the term multicollinearity and discuss the undesired effects it has on modeling the relationship between solar wind speed and electron flux. Next, we present and discuss a class of models, dynamic linear models, that shows promise as a tool to investigate the relationship between solar wind drivers and electron flux (Section 2.5) as well as a tool to generate competitive forecasts (Section 2.6). We conclude with the discussion in Section 2.7.

2.2 Data Description

In this study, we use electron flux data from the Los Alamos National Laboratory (LANL) geosynchronous energetic particle instruments (Reeves et al., 2011) and solar wind data from the OMNI2 archive (King and Papitashvili, 2005), processed into daily averages to provide a uniform database spanning 20 years. The dates of the available data fall between September 22nd, 1989 and December 31st, 2009. However, in what follows, “all available data” will refer to the data from January 1st, 1990 to December 31st, 2009, unless otherwise specified. There are two reasons for this. In what follows, many parameter estimates and standard errors are presented by year. Because the available data for 1989 (100 days) is less the available data for each year between 1990 and 2009 (∼ 365 days), standard errors for parameter estimates in 1989 will be larger than standard errors for years 1990 - 2009, making it difficult to compare parameter uncertainty across years. The other reason is that the data in 1989 can be used to sensibly select initial state values for the DLM. This is discussed more in Section 2.5. For more
details about the collection and/or processing of these data, the reader is directed to Section 2 of Reeves et al. (2011) where the digital data are also available in the electronic supplement.

A scatter plot of solar wind speed and log electron flux, along with their respective histograms, are displayed in Figure 2.1. The means (standard deviations) of solar wind speed and log electron flux are 441.97 (100.33) km/s and -0.44 (0.79) (cm$^2$/s/sr/keV)$^{-1}$, respectively. A triangular shape is present in the scatter plot. That is, there appears to be a solar wind dependent lower bound to log electron flux, but solar wind independent upper bound for log electron flux. The triangular shape between solar wind speed and log electron flux has been presented and discussed in the literature (e.g., Reeves et al., 2011; Kellerman and Shprits, 2012). The histogram for solar wind speed is unimodal and positively skewed, while the histogram for log electron flux is unimodal and roughly symmetric.
Figure 2.1: (top) Log electron flux \((\text{cm}^2/\text{s}/\text{sr}/\text{keV})^{-1}\) versus solar wind speed (km/s). A triangular distribution is evident, with a solar wind dependent lower bound on log electron flux, but a solar wind independent upper bound on log electron flux. (bottom) Marginal histograms of solar wind speed (left) and log electron flux (right). The distribution of solar wind speed is unimodal and skewed right, ranging from 255.2 to 1103.0 km/s. The distribution of log electron flux is unimodal and roughly symmetric, ranging from -2.38 to 1.46 \((\text{cm}^2/\text{s}/\text{sr}/\text{keV})^{-1}\).

In the modeling that follows, we standardize the input solar wind velocity by mean-centering and scaling to the standard deviation. This facilitates convenient parameter interpretation in
a linear model and appealing statistical properties. Centering solar wind speed removes the
correlation between the intercept of a linear model and all other linear terms. It also removes
all nonessential multicollinearity between all linear terms (Marquardt and Snee, 1975). Why
this is beneficial will be discussed in Sections 2.3 and 2.4. Scaling the mean centered solar wind
speed helps with numerical stability in the sampling procedure described in Section 2.5. For
clarity, let $X_t$ represent solar wind speed on day $t$. Then the standardized solar wind speed on
day $t$, $V_t$, is the mean centered, standard deviation scaled version of solar wind speed. That is,$V_t = (X_t - 441.97)/100.33$. By construction, the mean of the standardized solar wind speed,$V_t$, is zero and the standard deviation is one.

2.3 Parameter Estimate Sensitivity to Training Data

It is well documented that changes in solar wind speed precede changes in electron flux by
one to three days (Baker et al., 1990; Li et al., 2005; Rigler et al., 2007; Reeves et al., 2011).
Historically, electron flux forecasting models have leveraged this relationship (Baker et al., 1990;
Li et al., 2001; Ukhorskiy et al., 2004; Sakaguchi et al., 2013). A simple model describing the
mean relationship between log electron flux and one day lagged solar wind speed by a straight
line is

$$F_t = \beta_0 + \beta_1 V_{t-1} + \epsilon_t, \quad \text{(D1)}$$

Here $F_t$ represents log electron flux on day $t$, $V_{t-1}$ is the standardized solar wind speed on
day $t - 1$, $\epsilon_t$ is a normally-distributed error term at time $t$ with mean zero and variance $\sigma^2$,
and $t = 1, 2, \ldots, T$. Furthermore, $\epsilon_t$ are assumed independent of all $\epsilon_s, s \neq t$, $\beta_0$ denotes the
intercept of the expectation line, and $\beta_1$ the slope. This is essentially the model one would
derive from a simple linear correlation.

An appealing feature of Model D1 is that the interpretations of $\beta_0$ and $\beta_1$ are simple and
intuitive. The interpretation of $\beta_0$ is the expected value of electron flux, $F_t$, when $V_{t-1} = 0$.
Recalling that $V_{t-1}$ is the standardized solar wind speed, $\beta_0$ is the average log electron flux
when the previous day’s solar wind speed is equal to the average solar wind speed, 442 km/s.
The interpretation of $\beta_1$ is the expected increase in $F_t$ when $V_{t-1}$ is increased by one unit.
Again recalling the standardization of solar wind speed, this interpretation is equivalent to saying that $\beta_1$ is the expected increase of log electron flux when the previous day’s solar wind speed is increased by approximately 100 km/s (i.e., one standard deviation).

Another appealing aspect of Model D1 is that parameter estimation is quite easy. The most frequently utilized estimation procedure to estimate $\beta_0$ and $\beta_1$ is ordinary least squares (OLS). A treatise on OLS can be found in Graybill (1976) and Neter et al. (1996). The OLS estimates for $\beta_0$ and $\beta_1$, denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively, are functions of the response, $F_t$, and the covariate, $V_t$. We refer to the data used to estimate $\beta_0$ and $\beta_1$ as training data. The choice of training data will affect the estimates for $\beta_0$ and $\beta_1$. In this section, we investigate the sensitivity of parameter estimates to the choice of training data by estimating $\beta_0$ and $\beta_1$ for each year of data, separately. Parameter estimates and 95% confidence intervals, using the normal distribution percentiles, are plotted in Figure 2.2, with the horizontal axis denoting the year.
Figure 2.2: Intercept (top) and slope (bottom) parameter estimates (points) with 95 percent confidence intervals (vertical segments) by year for Model D1. Parameter estimates are sensitive to the choice of training data and appear to vary systematically over time.

We note that intercept and slope estimates vary appreciably from year to year, indicating parameter estimates are sensitive to the choice of training data. For instance, $\hat{\beta}_0$ based on data from 1995 is -0.136 while $\hat{\beta}_0$ based on data from 2001 is -0.799. These estimates are significantly different from one another, as their respective confidence intervals do not overlap. This suggests that for the year 1995, a solar wind speed of approximately 442 km/s, on average, results in log electron flux of -0.136 (cm$^2$/s/sr/keV)$^{-1}$ the following day. But in 2001, the same solar
wind speed results in a log electron flux of only $-0.799 \ (\text{cm}^2/\text{s/sr/keV})^{-1}$ – over half an order of magnitude difference. Similar extremes can be seen by, for example, comparing $\hat{\beta}_1$ in 1996 and 2000.

To illustrate how these yearly parameter estimates predict flux differently, we compare predicted flux using training data from 1995, 2000, and all available years to observed flux in February and March of 1995 and 2000. The results are shown in Figure 2.3. As expected, predicted flux based on training data from 1995 reproduces the observed flux in 1995 very well. However, using training data from 2000 to predict flux in 1995 systematically under predicts flux. The opposite effect is seen when comparing predicted flux to observed flux in 2000 (bottom of Figure 2.3). That is, predicted flux based on training data from 2000 follows observed flux most closely, while predicted flux based on training data from 1995 systematically over predicts flux.
Figure 2.3: Observed log electron flux (black), predicted log electron flux using training data from 1995 (red), 2000 (blue), and all years (green) for February and March in 1995 (top) and 2000 (bottom).

It is often assumed that using a longer interval of training data will produce a more accurate prediction, but that is not necessarily true. Using longer training intervals in this type of model systematically under predicts high-flux years and over predicts low-flux years. Predicted flux based on training data for all available years is plotted in green in Figure 2.3. In 1995, predicted flux based on all years systematically under predicts flux, but not as severely as predictions based only on training data from 2000 (top of Figure 2.3). The opposite is seen in 2000 (bottom of Figure 2.3). This example illustrates that, when the relationship between solar wind speed and log electron flux is dynamic, a static model based on any given training period – even the entire available data set – may not provide the most accurate prediction for any given time.
period. This fact is particularly important for application to future years, after the model is published.

By fitting Model D1 separately for each year, we are able to detect that parameter estimates change over time. The time-varying functional relationship between log electron flux and solar wind speed has been noted in the literature (e.g., Reeves et al., 2013; Perry et al., 2010; Turner et al., 2011; Kellerman et al., 2013). Reeves et al. (2013) graphically examined the long-term variations between solar wind speed and log electron flux and concluded that the functional relationship between these two variables is not constant in time. Perry et al. (2010) compared the accuracy of models forecasting log electron flux between 1996 and 2008. They note that the worst forecasts occurred during solar maximum (second half of 1999 through first half of 2002), better forecasts during the declining years of solar minimum (second half of 2002 through 2008), and their best forecasts during the inclining years of solar minimum (1996 through first half of 1999). Turner et al. (2011) reviewed forecasting models for Earth’s outer radiation belt electrons other than those considered by Perry et al. (2010) but come to similar conclusions, namely that forecast accuracy is related to the solar cycle. Kellerman et al. (2013) examined forecast accuracy for their geosynchronous radiation-belt electron empirical prediction (GREEP) model between 1991 and 2009. They found forecast accuracy was highest during the descending phase of solar cycle 22 and the solar minimum between cycles 22 and 23. All of these studies suggest that the functional relationship between log electron flux and solar wind speed is time-varying in a systematic way. This behavior is important but not particularly surprising. Solar wind velocity is almost certainly not the only variable that determines radiation belt electron flux. We examine multiple variable static models in the next section but note here that even this simple example points to the potential value of dynamic predictive models over static ones.

2.4 Multi-Parameter Estimates and Sensitivity to Multicollinearity

It is hypothetically possible that choosing the “right” set of inputs (covariates) and the “right” functional form could remove the temporal variability in the parameter estimation.
However, in practice, adding more inputs can introduce a different set of problems – particularly when those inputs are correlated with one another.

Multicollinearity refers to the situation when two or more covariates in a regression model are correlated with each other (Neter et al., 1996). This is typically the case for solar wind inputs (solar wind speed, number density, pressure IMF, etc.) and it is almost always the case for multiple lags of the same input (e.g., 0 days, 1 day, etc.). Though not strictly a violation of any linear regression model assumptions, correlated covariates can pose challenges. One such challenge is that multi-parameter estimation can depend sensitively on the choice of inputs, possibly resulting in misleading conclusions regarding the physical interpretation of those parameters.

Estimated parameters tend to have large standard errors when covariates are highly correlated. This means estimated parameters can vary widely from sample to sample, diminishing the precision of the information we have about the true value of the parameter. This can result in many covariates individually being deemed statistically not significant (i.e., 95% confidence interval about a parameter estimate contains 0) when in fact there is a definite statistical relationship between the response variable and a group of covariates (inputs).

In addition to large standard errors, parameter estimates can be biased when a relevant covariate(s) is excluded from a model (“model misspecification”) and that covariate is correlated with another covariate included in the model. This can be challenging when one desires to attach physical meaning to the estimated parameters, as on average, this estimate is not equal to that of the true value of the parameter. Figure A.2 in Appendix A illustrates this point in a small simulation study.

In addition to bias, it is not clear how to interpret estimated coefficients in the presence of multicollinearity. Typically, one interprets an estimated (non-intercept) coefficient in a linear regression model as the change in the expected value of the response variable when the covariate is increased by one unit, holding all other covariates constant. In practice, when covariates are highly correlated, it may not be possible to increase one covariate while holding all others constant. For example, we cannot change solar wind pressure while holding velocity constant. Thus, the simple interpretation of the estimated coefficient measuring a marginal effect is likely
inappropriate, and it is not clear how to attach physical meaning to an estimated coefficient in the presence of multicollinearity.

For a brief illustration of how multicollinearity influences parameter estimation, the reader is directed to Appendix A. For a more detailed account of the effects of multicollinearity in the context of forecasting electron flux, the reader is directed to Rigler et al. (2007).

The concerns just mentioned about parameter estimation and interpretation in the presence of multicollinearity are relevant to model building in the context of forecasting electron flux. They are relevant because multiple linear regression models have been utilized to explore the relationship between log electron flux and relevant solar wind drivers (e.g., Baker et al., 1990; Rigler et al., 2007). A fairly standard multiple linear regression model relates log electron flux to lagged values of a covariate(s) (e.g., solar wind speed). Doing so is one attempt to account for the inherent temporal nature of these data. Using lagged versions of input variables to account for the temporal structure of the data, however, can lead to unintended complications in modeling. This is because lags of covariates are almost certainly correlated, and therefore such a model will suffer from multicollinearity issues.

To illustrate this, consider two models: Model D1 from the previous section and

\[ F_t = \beta_0 + \beta_1 V_{t-1} + \beta_2 V_{t-2} + \epsilon_t, \quad (D2) \]

where \( F_t, V_t, \) and \( \epsilon_t \) are defined as in Section 2.3. Model D2 describes the mean relationship between log electron flux and both one day lagged and two day lagged solar wind speed (\( V_{t-1} \) and \( V_{t-2} \), respectively). The estimated Pearson correlation coefficient between \( V_{t-1} \) and \( V_{t-2} \) for all the available data is 0.807, a relatively high correlation. As in Section 2.3, Model D2 was also fit to each year of data separately. Parameter estimates for \( \beta_0 \) and \( \beta_1 \) with 95% confidence intervals are shown in Figure 2.4 (estimates for \( \beta_2 \) not shown).
Figure 2.4 illustrates how parameter estimation for $\beta_1$ is sensitive to model specification by presenting side-by-side point estimates with associated 95% confidence intervals. For each
year, $\hat{\beta}_1$ is larger and has smaller 95% confidence intervals (smaller standard errors) for Model D1 than Model D2.

In both Model D1 and D2, the parameter $\beta_1$ specifies the change in electron flux when $V_{t-1}$ changes by 1 (i.e., actual solar wind velocity changes by one standard deviation – 100 km/s). Both use the same training data. But, the lower values of $\beta_1$ in Model D2 imply that electron flux is less sensitive to a change in the previous day’s solar wind speed when the speed two days earlier is included than when it is not. Is it true that Model D1 overestimates the dependence of flux on the previous day’s solar wind speed or does the correlation between solar wind speed on subsequent days just “split” that dependence among more variables?

To clarify the undesired consequences of multicollinearity, consider $\hat{\beta}_1$ in 2000. Under Model D1, $\hat{\beta}_1$ and 95% confidence interval is 0.243 (0.146, 0.310), while under Model D2, it is -0.059 (-0.158, 0.040). There are a couple of observations that are notable. The first is that these parameter estimates are significantly different from one another. Their 95% confidence intervals do not overlap. The second is that the physical interpretation one would attach to $\hat{\beta}_1$ under Model D1 is different than under Model D2. For Model D1, plausible values for $\beta_1$ are between 0.146 and 0.310 – positive values. Thus one would conclude that an increase in solar wind speed of any amount on day $t-1$ will on average result in an increase to log electron flux. For Model D2, plausible values for $\beta_1$ are between -0.158 and 0.040 – negative values, positive values, and zero. Thus one would conclude that an increase in solar wind speed of any amount on day $t-1$, holding solar wind speed constant on day $t-2$, could result in a decrease, increase, or no change to mean log electron flux. It is no longer clear what marginal effect one day lagged solar wind speed has on expected log electron flux.

The year 2000 is typical with respect to how the 95% confidence intervals about $\hat{\beta}_1$ can change our physical interpretation/understanding of how one day lagged solar wind speed marginally effects mean log electron flux. When Model D1 was fit to each year, the 95% confidence interval around $\hat{\beta}_1$ was strictly positive for all years. From this, we would conclude there is a significant, positive, linear relationship between one day lagged solar wind speed and the mean of log electron flux. When Model D2 was fit, this conclusion could only be made for 11 of the 20 years. That is, for 9 out of the 20 years, we would conclude there is insufficient
evidence to claim that an increase in solar wind speed on day \( t - 1 \), holding solar wind speed constant on day \( t - 2 \), would result in an increase to mean log electron flux on day \( t \). We note that since \( V_{t-1} \) and \( V_{t-2} \) are strongly correlated it is not possible to increase solar wind speed at day \( t - 1 \) while holding solar wind speed constant at day \( t - 2 \). This example illustrates why physical interpretation of estimated parameters is ambiguous when multicollinearity is present.

A linear model with correlated lags of a single solar wind driver (e.g., Model D2) is an example of multicollinearity. Multicollinearity, however, is more general than this. Any regression model where two or more covariates are correlated with one another is subject to multicollinearity. We emphasize that the correlation that leads to issues with multicollinearity is correlation between two or more covariates, not the correlation between a covariate and the response. Correlation between a covariate and the response is desired when modeling. Some covariates that are correlated with log relativistic electron flux include solar wind drivers (e.g., solar wind speed, ion density, z-component of IMF), geomagnetic indices (e.g., AE, Dst), lower energy electron flux, and lagged versions of these aforementioned drivers. Correlation between covariates, however, is undesired when modeling. This is because correlation between two or more covariates complicates one’s ability to untangle the marginal effects of a driver (e.g., a solar wind driver, a geomagnetic index, etc.) on flux from another driver. Attempts have been made to untangle these effects, however (e.g., Kellerman and Shprits, 2012). To illustrate that multicollinearity occurs simply when two or more covariates are correlated with each other, we considered four models: Model D1,

\[
F_t = \beta_0 + \beta_1 V_{t-1} + \beta_2 D_{t-1} + \epsilon_t, \tag{D3}
\]

\[
F_t = \beta_0 + \beta_1 V_{t-1} + \beta_2 Z_{t-1} + \epsilon_t, \tag{D4}
\]

\[
F_t = \beta_0 + \beta_1 V_{t-1} + \beta_2 D_{t-1} + \beta_3 Z_{t-1} + \epsilon_t, \tag{D5}
\]

where \( D_{t-1} \) is standardized ion density and \( Z_{t-1} \) is the standardized z-component of the IMF in the geocentric solar magnetospheric coordinate system, both on day \( t - 1 \). We fit Models D1, D3, D4, and D5 to the years 1996, 2000, and 2009 and present \( \hat{\beta}_1 \) and associated 95% confidence interval in Table 2.1. In Table 2.1, we also present the estimated Pearson correlation coefficient between standardized solar wind speed and standardized ion density (Model D3) and
standardized \( z \)-component of the IMF (Model D4). Table 2.1 reinforces what has previously been stated. That is, parameter estimates are prone to change when a correlated variable is added to a linear regression model. The amount of change and inflation to standard errors (and thus, 95% confidence intervals) is related to how correlated the covariates are. The estimate for \( \beta_1 \) in Models D3 and D5 is smaller than \( \hat{\beta}_1 \) in Model D1, as both these models include ion density. Ion density is mildly and negatively linearly correlated with solar wind speed \((-0.6567 \text{ in 1996 to } -0.3291 \text{ in 2000})\). The estimate for \( \beta_1 \) in Model D4, however, is roughly unchanged when compared to Model D1. This is because \( z \)-component of the IMF is at best weakly linearly correlated with solar wind speed \((-0.2072 \text{ in 1996 to 0.1292 in 2000})\). This illustrates that adding uncorrelated covariates to a model will not inherit the complications of multicollinearity. Namely, intuitive interpretations of model parameters are preserved.

Table 2.1: Estimates for \( \beta_1 \) and 95% confidence intervals (CI) from Models D1, D3, D4, and D5 fit to 1996, 2000, and 2009. Correlations are between standardized solar wind speed and standardized ion density (Model D3) and standardized \( z \)-component of IMF (Model D4). Estimates for \( \beta_1 \) change appreciably between Model D1 and D3 when ion density is added to the model in 1996, as ion density is moderately correlated with solar wind speed. Little change to \( \hat{\beta}_1 \) is seen when the \( z \)-component of the IMF is added to the model, as it is weakly correlated to solar wind speed.

<table>
<thead>
<tr>
<th>Year</th>
<th>Model</th>
<th>Estimate</th>
<th>95% CI</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>D1</td>
<td>0.7432</td>
<td>(0.6653, 0.8212)</td>
<td>-0.6567</td>
</tr>
<tr>
<td></td>
<td>D3</td>
<td>0.5769</td>
<td>(0.4768, 0.6770)</td>
<td>-0.2072</td>
</tr>
<tr>
<td></td>
<td>D4</td>
<td>0.7061</td>
<td>(0.6277, 0.7844)</td>
<td>-0.2072</td>
</tr>
<tr>
<td></td>
<td>D5</td>
<td>0.5650</td>
<td>(0.4661, 0.6638)</td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>D1</td>
<td>0.2429</td>
<td>(0.1758, 0.3099)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>D3</td>
<td>0.1418</td>
<td>(0.0780, 0.2056)</td>
<td>0.1292</td>
</tr>
<tr>
<td></td>
<td>D4</td>
<td>0.2556</td>
<td>(0.1888, 0.3224)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>D5</td>
<td>0.1545</td>
<td>(0.0909, 0.2180)</td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>D1</td>
<td>0.5209</td>
<td>(0.4288, 0.6131)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>D3</td>
<td>0.4639</td>
<td>(0.3561, 0.5717)</td>
<td>-0.5143</td>
</tr>
<tr>
<td></td>
<td>D4</td>
<td>0.5233</td>
<td>(0.4325, 0.6141)</td>
<td>0.0223</td>
</tr>
<tr>
<td></td>
<td>D5</td>
<td>0.4644</td>
<td>(0.3585, 0.5703)</td>
<td></td>
</tr>
</tbody>
</table>

Note that we do not suggest here that the true relationship between log electron flux and solar wind speed (or other solar drivers) is described by Models D1 - D5. These models are
simply used to illustrate the point that parameter estimation is dependent upon the model being fit. When new covariates are added to the model, especially in the context of multicollinearity, parameter estimates are prone to change. Model D1 - D5 are all relatively simple, but one could imagine how parameter estimates (and thus one’s physical interpretation of the underlying process) might change when more lags of a solar driver and/or more solar drivers are added. When many solar drivers/lags of solar drivers are correlated with each other, interpretation of the estimates attached to these variables is unclear. Thus, when model adequacy and predictive ability of two models is roughly equal, the simpler or more parsimonious model should be preferred – Ockham’s razor (e.g., Jefferys and Berger, 1992). We stress, however, that though the simpler model should be preferred for empirical prediction, that does not necessarily mean the model accurately represents the true physical process involved.

### 2.5 Dynamic Linear Models

As an alternative to static models that assume a constant functional relationship between covariates and response, such as Models D1 - D5, and multiple linear regression models subject to multicollinearity, such as Models D2 - D5, we propose an approach based on dynamic linear models (DLMs). DLMs are a large class of models that, among other features, allow for time-varying parameters. Thus, they do not need to assume a constant functional relationship between covariates (e.g., solar wind drivers) and response (electron flux). By embedding the temporal dependence within the functional relationship, DLMs alleviate the need for numerous lags of a covariate to account for the temporal nature of the data, producing cleaner parameter interpretation.

Generally, DLMs are defined by two equations and an initialization:

\[
F_t = M_t \beta_t + e_t \quad (2.1a)
\]

\[
\beta_t = G_t \beta_{t-1} + w_t \quad (2.1b)
\]

\[
\beta_0 \sim N(a_0, C_0). \quad (2.1c)
\]

Equation (2.1a) is known as the observation equation, (2.1b) as the state equation, and (2.1c) as the initialization of the state. In equation (2.1a), \( F_t \) is the observation (e.g., electron flux),
$M_t$ is a matrix of known constants (e.g., solar wind speed, ion density, etc.), $\beta_t$ is the state vector, and $e_t$ is the observation error, all at time $t$. Said another way, the observation at time $t$ is a linear combination of the state vector at time $t$, plus noise. It is assumed that $(e_t)_{t \geq 1}$ is an independent sequence of mean zero, Gaussian random variables with variance matrices $(E_t)_{t \geq 1}$, where $E_t$ can evolve in time or $E_t$ can be treated as static by imposing the constraint $E_t = E$ for all $t$. In equation (2.1b), $G_t$ is a known matrix of constants and $w_t$ is the state error, both at time $t$. Thus, the state vector at time $t$ is a linear combination of the state vector at time $t-1$, plus noise. It is assumed that $(w_t)_{t \geq 1}$ is an independent sequence of mean zero, Gaussian random variables with variance matrices $(W_t)_{t \geq 1}$, where $W_t$ can evolve in time or $W_t$ can be treated as static by imposing the constraint $W_t = W$ for all $t$. Independence is assumed between $(e_t)$ and $(w_t)$. In equation (2.1c), $a_0$ and $C_0$ are the assumed known prior mean and variance, respectively, for the initial state vector $\beta_0$. Finally, it is assumed that $\beta_0$ is independent of $(e_t)$ and $(w_t)$ (Petris et al., 2009).

Inference about the unobserved state vector $\beta_t$ is often of primary interest. Types of inference of interest are partitioned into three specific problems: filtering, forecasting, and smoothing. In his seminal work, Kalman addressed inference of an unobserved state vector by introducing the Kalman filter (Kalman, 1960). The Kalman filter is a popular estimation procedure for DLMs used to recursively estimate the unobserved state vector $\beta_t$, given all the data up to time $t$, $F_{1:t}$. The Kalman filter also provides a prescription for forecasting. For some $k \geq 1$, forecasting refers to computing the conditional distribution of $F_{t+k|1:t}$, the distribution of the yet to be observed observation at time $t + k$ given all the data up to time $t$. Kalman’s work with filtering and forecasting has been leveraged previously in the context of forecasting electron flux (Rigler et al., 2004; Sakaguchi et al., 2013) and modeling the dynamics of the radiation belt (Kondrashov et al., 2007; Shprits et al., 2007; Reeves et al., 2012).

In addition to the Kalman filter, however, there exists a lesser utilized estimation procedure often referred to as the Kalman smoother. The inference problem known as smoothing is a retrospective problem and is useful to retrospectively learn about the evolution of a system. Specifically, smoothing refers to inference about the conditional distribution of $\beta_t|F_{1:T}$. That is, the distribution of the unobserved state at time $t < T$, $\beta_t$, given all the available data, $F_{1:T}$. 
The necessary details to estimate the distribution $\beta_t$ for the inference problems of filtering, forecasting, and smoothing are presented in Appendix B.

For clarity, we emphasize that the Kalman filter and the Kalman smoother are not themselves models, but rather estimation procedures that have proven useful for certain classes of models. For example, both the Kalman filter and Kalman smoother are useful estimation procedures for DLMs, where a DLM is a model. For readers familiar with the extended Kalman filter (EKF), it is an estimation procedure useful for dynamic nonlinear models. More specifically, if $M_t \beta_t$ in equation (2.1a) were replaced with $h(\beta_t)$ and/or $G_t \beta_{t-1}$ in equation (2.1b) were replaced with $h(\beta_{t-1})$, where $h()$ represents a generic nonlinear function of the input, we would no longer have a dynamic linear model, but rather a dynamic nonlinear model. Thus, the EKF is the nonlinear analogue to the Kalman filter.

DLMs are a powerful class of models, largely because of the efficient state estimation procedures available for each of the three aforementioned state inference problems. DLMs also provide a high level of modeling flexibility absent from static models. That is, static models implicitly combine the effects of any potentially time-varying dynamics with the error terms or average over the effects in the parameter estimates. DLMs have the ability to capture these time-varying dynamics in the state vectors. Thus, the potentially contextually meaningful time-varying dynamics are not frivolously relegated to the error terms or lost in parameter estimation, but rather are captured in interpretable ways in the state vectors.

We believe DLMs are useful tools on two fronts. The smoothing distributions capture meaningful trends that can help inform the understanding of the time-varying dynamics while the filtering and forecasting distributions provide a powerful framework capable of producing competitive forecasts. For these reasons, we propose utilizing DLMs both as a forecasting tool and as a tool to help understand the time-varying dynamics between, generally, a response variable and covariates and, specifically here, log electron flux and solar wind drivers. Though some attention has been paid to the filtering and forecasting distributions in the literature, virtually none has been paid to the smoothing distributions.

To see how DLMs can be leveraged to learn about the underlying, time-varying dynamics between solar wind speed and log electron flux, let us first consider how DLMs are similar to
the static models introduced earlier. The static models discussed above can be expressed as specific cases of DLMs. For example, Model D1 is a DLM with $M_t = [1, V_{t-1}]$, $G_t = I_{2 \times 2}$, $E_t = E$, and $W_t = 0_{2 \times 2}$, where $I_{2 \times 2}$ is a two-by-two matrix of 1s on the diagonal and 0s on the off diagonal (an identity matrix) and $0_{2 \times 2}$ is a two-by-two matrix of 0s. In fact, all multiple linear regression models are DLMs, where $G_t = I_{p \times p}$, $W_t = 0_{p \times p}$, and $p$ is the length of the state vector. Model D1 can easily be extended to be a dynamic linear model by allowing $W_t$ to equal a valid variance-covariance matrix. Specifically, a simple time-varying DLM representation of Model D1 is

$$F_t = [1, V_{t-1}] \begin{bmatrix} \beta_{0,t} \\ \beta_{1,t} \end{bmatrix} + e_t$$

(2.2a)

$$\begin{bmatrix} \beta_{0,t} \\ \beta_{1,t} \end{bmatrix} = \begin{bmatrix} \beta_{0,t-1} \\ \beta_{1,t-1} \end{bmatrix} + \begin{bmatrix} w_{0,t} \\ w_{1,t} \end{bmatrix},$$

(2.2b)

where $e_t \sim N(0, E)$ and $(w_{0,t}, w_{1,t})' \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, W = \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{bmatrix}\right)$, where $a'$ means the transpose of $a$ and $MVN(\mu, \Sigma)$ is a multivariate normal distribution with mean $\mu$ and variance-covariance matrix $\Sigma$. Observation equation (2.2a) characterizes the mean relationship between log electron flux and one day lagged solar wind speed by a straight line with intercept $\beta_{0,t}$ and slope $\beta_{1,t}$, just like Model D1. Unlike Model D1, the intercept and slope are allowed to vary in time in the manner described by state equation (2.2b) – a random walk. That is, the slope and intercept at time $t$ are equal to the slope and intercept at time $t - 1$, plus the state error. The magnitude of the state error dictates how much the intercept and slope are allowed to change from time $t - 1$ to $t$. This imposes a smoothness to the solution. The magnitude of the state error is determined by $\sigma_0^2$ and $\sigma_1^2$. The closer either $\sigma_0^2$ and/or $\sigma_1^2$ is to zero, the more stable (smoother) in time the local linear relationship is, and the smaller the state error. The larger they become, the larger the state errors and the more unstable (less smooth) the local linear relationship becomes.

Figure 2.5 shows the results of the dynamic version of Model D1 (i.e., the DLM given by equations (2.2a) and (2.2b)). We first consider some features of the DLM before describing in
detail the derivation for space weather forecasting applications. Both models forecast electron flux, $F_t$, based only on the prior day's solar wind speed, $V_{t-1}$. We first note that the DLM captures the temporal variation of the fitting parameters but does not require us to pre-select any training period. Recall, however, that in a practical application, Model D1 does not actually vary from year to year but is fixed for all times based on the chosen training interval. Rather, the DLM parameters (states) update continuously based on past history of the input (speed) and output (flux) variables. Note also that in Model D1, even if we allow the training interval to update, the forecast will, at best, be based on the previous year's data. For the DLM, it will be based on all the data up to the previous day.
Figure 2.5: Posterior means of smoothed states (DLM) and parameter estimates (points) with 95 percent confidence intervals (vertical segments) (D1) for intercept (top) and slope (bottom) parameters. The posterior means of the smoothed states line up very closely with the parameter estimates from Model D1.

Next we note the similarity between Model D1 and its DLM representation in equations (2.2a) and (2.2b) with respect to parameter interpretation. The interpretation of \( \hat{\beta}_0 \) in Model
D1 was the expected log electron flux on day \( t \) when the observed solar wind speed on day \( t - 1 \) was \( \approx 442 \) km/s. The interpretation of the posterior mean of \( \beta_{0,t} \) in DLM representation of Model D1 is the expected log electron flux on day \( t \) when the observed solar wind speed on day \( t - 1 \) was \( \approx 442 \) km/s – the same intuitive interpretation. Notice, however, that if \( t \) corresponds to a day in 1994, then the posterior mean of \( \beta_{0,t} \) is roughly \( -0.1 \), while if \( t \) corresponds to a day in 2002, the posterior mean of \( \beta_{0,t} \) is roughly \( -1.0 \) – different by almost an entire order of magnitude. In other words, in 1994 and 2002 the same solar wind speed is predicted to result in electron fluxes that are an order of magnitude different! Thus, DLMs capture the time-varying parameters without having to pre-select any training period while simultaneously preserving the intuitive parameter interpretation Model D1 provides.

To perform the analysis resulting in Figure 2.5, selection of \( E \) and \( W \) must be addressed. In the literature, either a preliminary estimation of \( E \) and \( W \) or simply picking \( E \) and \( W \) has been done. We do not employ either of these approaches. Rather, we take a conjugate prior Bayesian approach. That is, we assign a probability distribution to the model parameters, rather than specific values. For more details on Bayesian approaches to DLMs, as well as a complete treatise on DLMs, we direct the reader to Petris et al. (2009) and West and Harrison (1997). The following prior distributions and hyperparameters (parameters of prior distributions) were chosen as noninformative and conjugate: \( E^{-1} \sim Gamma(1,1) \), \( W^{-1} \sim Wishart(I_{2\times2}, 10) \), \( a_0 = [-1.03, .34]' \), and \( C_0 = 
\begin{bmatrix}
.003 & -.001 \\
-.001 & .003
\end{bmatrix} \). A gamma distribution is a continuous probability distribution function with positive support. Thus, a gamma distribution is a common choice to model scalar quantities constrained to the positive real line (e.g., \( E \) and \( E^{-1} \)). The prior for \( E^{-1} \) is related to the range of \( F_t \). A Wishart distribution is a continuous probability distribution function defined over symmetric, non-negative definite matrices. Valid covariance matrices must be symmetric and non-negative definite. Thus, a Wishart distribution is a common choice to model covariance matrices, \( W \), or inverses thereof, \( W^{-1} \). The prior for \( W^{-1} \) is related to the range of the \( \hat{\beta}_0 \)s and \( \hat{\beta}_1 \)s in Figure 2.2. See Johnson et al. (1995) and Johnson et al. (2002) for more details on the gamma and Wishart distributions. The selection of \( a_0 \) is related to the anticipated intercept and slope between log electron
flux and one day lagged standardized solar wind speed on January 1st, 1990. Thus, $a_0$ is set equal to the OLS estimate for $\hat{\beta}_0$ and $\hat{\beta}_1$, using the data from September 22nd, 1989 through December 31st, 1989 as the training data. $C_0$ is set equal to the estimated variance-covariance matrix between $\hat{\beta}_0$ and $\hat{\beta}_1$ for the same training data.

The posterior distribution (the joint distribution of $E$, $W$, and $(\beta_{0,t}, \beta_{1,t}')_{t \geq 0}$, given the data, $(F_t)_{t \geq 1}$) is not available in closed form. Thus, a Markov chain Monte Carlo (MCMC) method, specifically Gibbs sampling, was used to obtain Monte Carlo samples from the posterior distribution. The details of this approach are beyond the scope of this paper, but also fairly standard in the literature now (e.g., Petris et al., 2009, Chap. 4). Essentially, MCMC algorithms allow us to draw samples from the posterior distribution (the distribution of interest). We can then perform inference on the drawn samples. The Gibbs sampler requires the specification of initial values for $E$ and $W$. Because of this, the first $B$ draws may not be considered draws from the posterior distribution. Thus, the first $B$ draws are excluded from the sample on which inference is performed. For MCMC methods, the first $B$ draws are referred to as the burn-in period, or simply, burn-in. For the DLM at hand, inference was performed on a sample of 2000 draws, after a burn-in of 500. In Figure 2.5, posterior means of the smoothed states are plotted with the results from Figure 2.2. Calculations were carried out using the $dlm$ package (Petris, 2010) within the statistical software $R$ (R Core Team, 2015) that accompany Petris et al. (2009).

### 2.6 Forecasting

As mentioned in Section 2.5, DLMs can produce competitive forecasts with relatively simple structure, compared to static models. To illustrate this, we compared three forecasting models with respect to their one day ahead forecasts. The first model was the persistence model. The persistence model forecast for log electron flux on day $t + 1$, $F_{t+1,\text{persistence}}$, is the observed log electron flux on day $t$, $F_t$. It is difficult to appreciably out-forecast the persistence model with respect to one day ahead forecasting, hence why it was selected (Perry et al., 2010). The second model is the Relativistic Electron Forecast Model (REFM). This model is based on the linear prediction filters of Baker et al. (1990) and is currently used by the Space Weather Prediction
The REFM one day ahead forecast is

\[ F_{t+1,REFM} = \hat{\beta}_0 + \hat{\beta}_1 V_t + \hat{\beta}_2 V_{t-1} + \ldots + \hat{\beta}_{30} V_{t-29} + F_t - F_{t,REFM}, \quad (2.3) \]

where \( \hat{\beta} \) is the estimate of \( \beta \). Note, the REFM is a model of the form of Model D2, but with 30 lags of solar wind speed, plus an offset. The offset is the difference between the observed log flux on day \( t \) and the forecast log electron flux on day \( t \), \( F_t - F_{t,REFM} \). Further note that, if all \( \hat{\beta} \)s are 0 and \( F_{t,REFM} \) is removed, then equation (2.3) reduces to the persistence model.

The third model, DLMforecast, is a DLM of the following form:

\[
F_{t+1,DLMforecast} = \beta_{0,t+1} + \beta_{1,t+1} V_t + f_{t+1} + e_t
\]

\[
\begin{bmatrix}
\beta_{0,t+1} \\
\beta_{1,t+1} \\
\phi_t
\end{bmatrix} =
\begin{bmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & \phi
\end{bmatrix}
\begin{bmatrix}
\beta_{0,t} \\
\beta_{1,t} \\
f_t
\end{bmatrix} +
\begin{bmatrix}
w_{0,t+1} \\
w_{1,t+1} \\
w_{f,t+1}
\end{bmatrix},
\quad (2.4b)
\]

where \( e_t \sim N(0, E) \) and \( (w_{0,t+1}, w_{1,t+1}, w_{f,t+1})' \sim MVN(0, diag(\sigma^2_W, \sigma^2_W, \sigma^2_f)) \), where \( diag(a) \) is a diagonal matrix with diagonal elements equal to the elements of the vector \( a \) and zero on the off-diagonals, and for clarity, we emphasize that \( f_{t+1} = \phi f_t \). The forecasts from the model DLMforecast are based on the forecasting distribution mentioned in Section 2.5.

Note the similarities between the REFM and equation (2.4a). First, each forecast for day \( t + 1 \) is a function of prior values of solar wind speed. In the REFM, it is a function of solar wind speed on days \( t \) through \( t - 29 \). In equation (2.4a), it is a function of solar wind speed on only day \( t \). Second, note that each model incorporates the previous value of observed log electron flux. In the REFM, it is incorporated through the offset term, \( F_t - F_{t,REFM} \). The term \( f_t \), an auto-regressive term of order 1, acts as the offset term in equation (2.4a). The term \( f_t \) is multiplied by the parameter \( \phi \) in state equation (2.4b). Thus log electron flux on day \( t \) enters into the forecast on day \( t + 1 \) up to a multiplicative constant.

A difference between these two models is the nature in which solar wind speed is related to log electron flux. In the REFM, the relationship between solar wind speed and log electron flux is static; \( \beta_0, \ldots, \beta_{30} \) are not indexed by time. In equation (2.4a), the relationship is dynamic; \( \beta_{0,t} \) and \( \beta_{1,t} \) are indexed by time. This allows us to reduce the number of covariates in the forecasting model (i.e., the number of lags) without appreciably sacrificing model fit.
To compare the one day ahead forecast accuracy of the three models, two metrics were used: prediction efficiency (PE) and mean absolute error (MAE). PE is defined as follows:

\[
PE = 1 - \frac{\text{MSE}_{\text{model}}}{\text{MSE}_{\text{reference}}},
\]

where

\[
\text{MSE}_{\text{model}} = N_{\text{pred}}^{-1} \sum_{t=1}^{N_{\text{pred}}} (F_{t,\text{model}} - F_t)^2,
\]

\[
\text{MSE}_{\text{reference}} = N_{\text{pred}}^{-1} \sum_{t=1}^{N_{\text{pred}}} (\bar{F} - F_t)^2,
\]

\[
\bar{F} = N^{-1} \sum_{t=1}^{N} F_t,
\]

\(F_{t,\text{model}}\) is the forecast of log electron flux from the model (persistence, REFM, or DLMforecast), \(N_{\text{pred}}\) is the number of days being forecasted, and MSE stands for mean squared error. We set \(\bar{F} = -0.44\), the average log electron flux between January 1st, 1990 through December 31st, 2009. PE = 1 corresponds to perfect forecasts. PE = 0 corresponds to forecasts that are no better or worse than using \(\bar{F}\) for the forecast, with respect to MSE. PE less than 0 corresponds to forecasts that are worse than using \(\bar{F}\), with respect to MSE. PE has no lower bound.

MAE is defined as follows:

\[
\text{MAE} = N_{\text{pred}}^{-1} \sum_{t=1}^{N_{\text{pred}}} |F_{t,\text{model}} - F_t|.
\]

MAE is nonnegative, with MAE = 0 corresponding to perfect forecasts. Larger values of MAE correspond to worse forecasts. An attractive feature of MAE is that it is interpretable. The value of MAE is the average absolute difference between the forecasts and observations, in units of log electron flux.

In addition to PE and MAE, we compared average prediction interval widths and coverage. Both the REFM and DLMforecast model are probabilistic models. In addition to forecasts, they can produce prediction intervals. Like confidence intervals, prediction intervals place bounds about levels of uncertainty of unknown quantities. With each forecast, we calculate the
nominal 95% prediction interval for each quantity. The 95% prediction interval should contain
the observed value of log electron flux 95% of the time. The proportion of the time the 95%
prediction interval actually contains the observed value of log electron flux is called coverage.
If two models have the same coverage, the model with smaller average prediction interval width
(upper limit minus lower limit) is preferred.

Average prediction interval width (Width) is calculated as

\[ \text{Width}_{\text{model}} = N^{-1}_{\text{pred}} \sum_{t=1}^{N_{\text{pred}}} (U_{t,\text{model}} - L_{t,\text{model}}), \]

where \( U_{t,\text{model}} \) and \( L_{t,\text{model}} \) are the upper and lower 95% prediction interval bounds for day \( t \)
and model REFM or DLMforecast, respectively. Coverage is calculated as

\[ \text{Coverage}_{\text{model}} = N^{-1}_{\text{pred}} \sum_{t=1}^{N_{\text{pred}}} I(L_{t,\text{model}} \leq F_t \leq U_{t,\text{model}}), \]

where \( I() \) is an indicator function that is equal to 1 if the condition is satisfied and 0 otherwise.

We note that the average prediction interval widths presented in Table 2.2 for the REFM are
underestimated. The calculated prediction interval widths would be appropriate for equation
(2.3) were the offset term removed. However, the offset term introduces an additional source
of uncertainty that is unaccounted for in the REFM prediction interval calculation.

The forecasting accuracy by year is presented in Table 2.2. Two years of data were used
as training data to estimate model parameters in both the REFM and DLMforecast model.
For the REFM, the model parameters are \( \beta_0, \ldots, \beta_{30} \) and estimated via OLS. In Model DLM-
forecast, the parameters are \( E, \sigma_W^2, \sigma_f^2, \text{ and } \phi \). The following prior distributions and hyperpa-
rameters were chosen for each model parameter: \( E^{-1} \sim \text{Gamma}(1,1), \sigma_W^{-2} \sim \text{Gamma}(1,1), \)
\( \sigma_f^{-2} \sim \text{Gamma}(1,1), \text{ and } \phi \sim N(1,1) \). Gibbs sampling was used to sample from the posterior
distribution. Posterior means were based on 1000 draws, after a burn-in of 200. Posterior
means for each parameter were used as their respective estimates. This approach deviates from
our approach in the previous section. That is, when working with the smoothing distributions,
we did not do a preliminary estimation of model parameters and subsequently treat those as
known values. In this section, that is what we do, as we are working with the forecasting
distributions. The states (\( \beta_{0,t}, \beta_{1,t}, \text{ and } f_t \)) are still dynamic and continuously evolving in the
DLMforecast model, but the parameters $E, \sigma_W^2, \sigma_f^2,$ and $\phi$ are treated as fixed, known values. The reasoning for this is discussed in Section 2.7.

Table 2.2: Prediction efficiency (PE), mean absolute error (MAE), average prediction interval width (Width) and coverage for one day ahead forecasts for the persistence (P), REFM, and DLMforecast (DLM) models, by year. The number of days in each year where log electron flux was unobserved is presented in the column Missing. Almost half of all days with missing log electron fluxes occurred between 2006 and 2009. The DLMforecast model produces competitive forecasts with respect to both the persistence and REFM models, as seen by the similar PE and MAE values for all three models. The coverages are roughly equivalent for REFM and DLMforecast, while average prediction interval widths are generally smaller for DLMforecast.

<table>
<thead>
<tr>
<th>Year</th>
<th>PE P</th>
<th>PE REFM</th>
<th>PE DLM</th>
<th>MAE P</th>
<th>MAE REFM</th>
<th>MAE DLM</th>
<th>Width (Coverage)</th>
<th>Width (Coverage)</th>
<th>Width (Coverage)</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>REFM</td>
<td>DLM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1992</td>
<td>0.658</td>
<td>0.683</td>
<td>0.661</td>
<td>0.327</td>
<td>0.321</td>
<td>0.345</td>
<td>2.255 (0.975)</td>
<td>2.032 (0.967)</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>1993</td>
<td>0.654</td>
<td>0.684</td>
<td>0.673</td>
<td>0.329</td>
<td>0.319</td>
<td>0.339</td>
<td>2.646 (0.992)</td>
<td>2.140 (0.955)</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>1994</td>
<td>0.749</td>
<td>0.789</td>
<td>0.763</td>
<td>0.275</td>
<td>0.268</td>
<td>0.286</td>
<td>2.777 (0.994)</td>
<td>2.287 (0.986)</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>1995</td>
<td>0.610</td>
<td>0.728</td>
<td>0.702</td>
<td>0.344</td>
<td>0.301</td>
<td>0.320</td>
<td>2.255 (0.978)</td>
<td>2.085 (0.972)</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>1996</td>
<td>0.755</td>
<td>0.822</td>
<td>0.809</td>
<td>0.271</td>
<td>0.244</td>
<td>0.249</td>
<td>1.926 (0.977)</td>
<td>1.906 (0.983)</td>
<td></td>
<td>11</td>
</tr>
<tr>
<td>1997</td>
<td>0.742</td>
<td>0.809</td>
<td>0.762</td>
<td>0.278</td>
<td>0.254</td>
<td>0.286</td>
<td>1.885 (0.981)</td>
<td>1.828 (0.964)</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>1998</td>
<td>0.719</td>
<td>0.772</td>
<td>0.705</td>
<td>0.300</td>
<td>0.284</td>
<td>0.320</td>
<td>1.965 (0.981)</td>
<td>1.778 (0.956)</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1999</td>
<td>0.680</td>
<td>0.754</td>
<td>0.719</td>
<td>0.312</td>
<td>0.284</td>
<td>0.311</td>
<td>2.007 (0.967)</td>
<td>1.907 (0.967)</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>2000</td>
<td>0.654</td>
<td>0.677</td>
<td>0.626</td>
<td>0.330</td>
<td>0.333</td>
<td>0.354</td>
<td>2.045 (0.969)</td>
<td>2.045 (0.956)</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>2001</td>
<td>0.643</td>
<td>0.684</td>
<td>0.630</td>
<td>0.339</td>
<td>0.323</td>
<td>0.349</td>
<td>2.211 (0.978)</td>
<td>2.115 (0.967)</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>2002</td>
<td>0.686</td>
<td>0.742</td>
<td>0.692</td>
<td>0.309</td>
<td>0.281</td>
<td>0.327</td>
<td>2.249 (0.980)</td>
<td>2.184 (0.978)</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>2003</td>
<td>0.643</td>
<td>0.667</td>
<td>0.703</td>
<td>0.320</td>
<td>0.317</td>
<td>0.304</td>
<td>2.179 (0.978)</td>
<td>2.314 (0.981)</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>2004</td>
<td>0.708</td>
<td>0.752</td>
<td>0.743</td>
<td>0.305</td>
<td>0.286</td>
<td>0.293</td>
<td>2.028 (0.975)</td>
<td>2.067 (0.978)</td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>2005</td>
<td>0.571</td>
<td>0.658</td>
<td>0.634</td>
<td>0.371</td>
<td>0.333</td>
<td>0.369</td>
<td>2.063 (0.949)</td>
<td>2.083 (0.952)</td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>2006</td>
<td>0.662</td>
<td>0.762</td>
<td>0.715</td>
<td>0.310</td>
<td>0.262</td>
<td>0.309</td>
<td>2.109 (0.967)</td>
<td>2.190 (0.974)</td>
<td></td>
<td>18</td>
</tr>
<tr>
<td>2007</td>
<td>0.679</td>
<td>0.769</td>
<td>0.740</td>
<td>0.302</td>
<td>0.274</td>
<td>0.282</td>
<td>2.169 (0.984)</td>
<td>2.270 (0.982)</td>
<td></td>
<td>26</td>
</tr>
<tr>
<td>2008</td>
<td>0.713</td>
<td>0.815</td>
<td>0.822</td>
<td>0.276</td>
<td>0.237</td>
<td>0.220</td>
<td>1.925 (0.981)</td>
<td>2.918 (0.981)</td>
<td></td>
<td>56</td>
</tr>
<tr>
<td>2009</td>
<td>0.888</td>
<td>0.906</td>
<td>0.872</td>
<td>0.157</td>
<td>0.158</td>
<td>0.176</td>
<td>1.773 (0.980)</td>
<td>1.855 (0.986)</td>
<td></td>
<td>14</td>
</tr>
</tbody>
</table>

The DLMforecast model produces competitive forecasts with both the persistence model and the REFM, with respect to PE and MAE. For most years, PE for the DLMforecast model is larger than PE for the persistence model, but smaller than the REFM model. Also for most years, all three models have an MAE between 0.30 and 0.35, suggesting the forecasts for each model are, on average, 0.30 to 0.35 orders of magnitude in absolute value off from the
observed electron flux. The coverage for REFM and DLMforecast are between 95% and 100% for almost all years, suggesting the actual coverage is slightly higher than the nominal 95%. In all years except for 2008, the average prediction interval widths are comparable between the REFM and the DLMforecast model. The average prediction interval width for the DLMforecast model in 2008 is unusually large because of the extreme number of days missing observed log electron flux in the data: 56 days. A discussion of how DLM forecasting handles data gaps in both flux and solar wind data streams is delayed until Section 2.7. The calculation for the prediction intervals for the DLMforecast model use the difference between forecasted and observed log electron flux to help reduce the uncertainty about the forecast. The average prediction interval width corresponding to the days with missing log electron flux in 2008 was 7.858, while the average prediction interval width corresponding to the days with observed log electron flux is 2.011. This accounts for the unusually large average prediction interval in 2008. The calculation for the prediction interval width for the REFM ignores this component because the offset piece is not built into a unified probabilistic model. We again note that the average prediction interval widths for the REFM are underestimates. By how much they underestimate the average prediction interval width is unclear. Thus, based on coverage and average prediction interval widths, a DLM-based forecasting approach may be an attractive alternative to the static approach used in the REFM.

2.7 Discussion

In this paper, we have argued that static models are not ideal for modeling daily averaged log electron flux, though they can produce useful forecasts. In Section 2.3, we showed that the relationship between log electron flux and solar wind speed systematically changes with time (Figure 2.2) and discussed why static models are unequipped to account for these changes.

In an attempt to account for the temporal nature of the data, some have modeled log electron flux as a function of numerous lags of solar wind speed or multiple solar wind parameters (e.g., REFM). In Section 2.4, we discussed how this approach almost certainly introduces multicollinearity. Multicollinearity makes interpretation of estimated parameters misleading, increases the standard errors of parameter estimates, and results in potentially biased param-
eter estimates. Thus, this approach to modeling the relationship between log electron flux and solar wind speed limits one’s ability to attach physical meaning to parameter estimates. To be fair, the REFM was not constructed with the intent of interpreting the relationship between solar wind speed and log electron flux, but rather to forecast log electron flux.

In Section 2.5, we presented and discussed dynamic linear models. Simple and interpretable models can be fit in the DLM framework that account for the time-varying functional relationship between log electron flux and solar wind speed. Because DLMs account for and adapt to the time-varying functional relationship between log electron flux and solar wind speed, one can describe their relationship more simply than with static models. This simplicity preserves the ability to attach physical meaning to the estimated states and has the potential to do so without sacrificing forecast accuracy.

The DLMforecast model uses 29 fewer model inputs than the REFM. The REFM does have preferable average PE (0.749) and average MAE (0.282) when compared to the DLMforecast model’s average PE (0.721) and average MAE (0.302). To be fair though, the DLMforecast was constructed for illustrative purposes. The DLMforecast demonstrates that by exchanging numerous model inputs for a dynamic model structure, one is still able to produce forecasts competitive with the REFM. In the case of the DLMforecast model, we were also able to avoid issues related to multicollinearity, as the DLMforecast model only includes one input. That is, the DLMforecast model is not a complete application of a multi-input dynamic linear model. Such a model is of interest but is left as the subject of future investigations.

Depending on the particular DLM, forecasting with DLMs can be operationally feasible, though a point of clarification is in order. The DLM can be partitioned into three pieces: data, states, and parameters. By way of example, for the DLM presented in equations (2.2a) and (2.2b), the data are \((F_t, V_t)_{t \geq 1}\), the states are \((\beta_{0,t}, \beta_{1,t})_{t \geq 0}\), and the parameters are \(E\) and \(W\). In the DLM framework, it is assumed that the data are known and the states are latent (unobservable) quantities. The parameters can either be assumed known or unknown. If we assume the parameters are known, generally, forecasting with DLMs would be feasible as the required calculations are computationally inexpensive. This is how we performed forecasting to calculate the quantities in Table 2.2 and is largely what has been done in the literature.
(see Section 2.1). If we assume the parameters are unknown, forecasting with DLMs can be computationally expensive and thus possibly unfeasible to do in real time. This is due to the MCMC techniques required to estimate the unknown parameters. The computational cost of running the MCMC depends on the number of parameters in the model, the form of the model, and the length of the data. An active area of research to make these computations more tractable is Sequential Monte Carlo or particle filtering (e.g., Petris et al., 2009, Chap. 5).

DLMs are robust to gaps in the response variable (log electron flux), as shown in Table 2.2. That is, DLMs can produce forecasts even when log electron flux is unobserved. If there are gaps in the covariate (e.g., solar wind speed), however, forecasts will not be produced automatically. One would need to input values for solar wind speed, similar to static models. This could be done by fitting a DLM to solar wind speed.

The purpose of the current study is to illustrate some of the inherent limitations in static forecasting models and some of the advantages of dynamic linear models. Previous studies of the dependence of geosynchronous electron fluxes on solar wind parameters have produced inconsistent and even contradictory results - even when the studies produced high correlations or prediction efficiencies. Our results suggest that the lack of consistency may be a function of applying static models to dynamic dependencies, the dependence of results on the choice of training data, and/or the presence of strong correlations among different solar wind inputs. These conclusions are not limited to forecasts of daily-averaged geosynchronous electron fluxes but, rather, are generally applicable to other data sets and models of interest to space weather forecasting. Similarly while our results confirm the dynamic nature of the relationship between geosynchronous MeV electron flux and solar wind velocity, they do not yet explain why that is the case. More complete application of multi-input dynamic linear models to the problem of radiation belt forecasting will be the subject of future investigations. Ultimately our twin goals are (1) improved understanding of the response of radiation belt electron fluxes to solar wind and geophysical drivers and (2) development of a forecast model with reproducible results and broad buy-in from the space weather community.
CHAPTER 3. MULTISCALE DYNAMIC LINEAR MODELS

3.1 Introduction

Probabilistic process models often need to incorporate information available at different scales, either arising naturally or as a consequence of data collection decisions. The coherent and principled modeling of information structured by scales is referred to as multiscale modeling. Temporal information is commonly structured by scales such as yearly, monthly, and/or daily. Scales are often nested. For example, days are nested within months.

The scale(s) of available information pertaining to a process or phenomenon will influence both the type of analysis performed and the conclusions drawn. For example, an analysis of multiple years of hourly temperature data at a single location in Iowa would reveal both a daily (higher midday) and annual (higher summer) cycle, while an analysis of daily temperature data would only reveal the annual cycle. As information at one scale may be relevant to another scale, there is a need to coherently integrate information from different scales.

Our interest in multiscale modeling is generated by several methodological factors. We desire to (i) integrate information from different scales in a principled, coherent manner, (ii) develop a multiscale model that is both parsimonious and interpretable and (iii) develop a multiscale model that explicitly captures relevant dynamics at each scale. Numerous methods exist for modeling multiscale processes. As is almost always the case, the appropriateness of a particular method is tied to both the objective(s) of the modeler and the application itself. In what follows, we discuss several often-employed techniques of multiscale modeling, followed by our proposed new approach.

Wavelet methods (e.g., Daubechies, 1992; Mallat, 1989) are useful when the main goal of multiscale modeling is the decomposition of a process into components organized by scales.
Wavelet methods use an infinite basis set that spans the space of continuous functions. What differentiates wavelet methods from, say, polynomial basis expansions is that the basis functions of wavelet methods are organized hierarchically. Coarser-scale components are considered first and finer-scale components are added until a sufficient number of wavelets have been included. Wavelet methods have been extensively used in image compression and signal analysis (e.g., Walter and Shen, 2000, and references therein), but they lack interpretability and a natural framework for forecasting.

Probabilistic tree models (e.g., Basseville et al., 1992) focus on interpretation and optimal prediction. These multiscale models are specified on graphical tree constructs where nodes of a given scale are conditionally independent given the immediately coarser scale. These models have applications in both engineering (e.g., Willsky, 2002) and spatial statistics (e.g., Huang et al., 2002). Probabilistic tree models are appealing because they admit a fast and efficient Kalman filter-like estimation procedure. While the assumption of conditional independence of one scale given the immediately coarser scale may be reasonable in many spatial applications, it is often less natural to assume for time series data.

Explicit multiscale time series models include the multiscale and hidden resolution time series models of Ferreira et al. (2006). Central to this method is the use of Jeffrey’s rule of conditioning to combine both a fine and coarse scale process in a probabilistically consistent manner capable of capturing relevant dynamics at multiple scales of resolution. This approach results in a new class of models for time series with a variety of autocorrelation structures based on a parsimonious parameterization and is naturally suitable to forecasting. However, the model suffers from a lack of interpretability due to nonlinearities in the posterior distributions at various scales, and efficient sampling methods are lacking.

We propose a novel multiscale model which we call the multiscale dynamic linear model (MSDLM). The MSDLM is rooted in the classical dynamic linear model (e.g., West and Harrison, 1997), but augmented to align with the methodological factors aforementioned. Like wavelets, the MSDLM can decompose a process into larger-scale trends and smaller-scale departures from larger-scale trends which can match underlying physical processes and admit interpretability. Unlike wavelets, the MSDLM provides a natural framework for forecasting.
Like multiscale models defined on trees, the MSDLM has a Kalman filter-like estimation procedure, enabling fast and efficient state estimation and sampling. Unlike multiscale tree models, the MSDLM does not assume conditional independence of a scale, given the immediately coarser scale. Rather, it assumes the first-order Markov property – the same conditional independence property assumed by DLMs – across the coarse scale and across the fine scale within a coarse scale. Like the multiscale time series models of Ferreira et al. (2006), the MSDLM is a class of flexible, parsimonious multiscale time series models. Unlike the model proposed by Ferreira et al. (2006), efficient sampling approaches do exist for MSDLMs and interpretation is provided in a manner similar to that of the classical DLM.

Since the MSDLM is a particular instance of a directed graphical model, we adopt a presentation within the directed graphical model formalism. Among the benefits of this approach we find that it provides a framework to design new multiscale systems, enables the modeler to draw inspiration from other graphical models, and it allows the user to leverage and exploit estimation techniques developed in other fields.

Chapter 3 is organized as follows. Section 3.2 presents a multiscale, motivating example. Section 3.3 provides a brief introduction to directed graphical models, where we present the DLM (Section 3.3.1) and introduce the MSDLM (Section 3.3.2). In Section 3.4 we present the directed global Markov property; a property used extensively for latent state estimation in both the DLM and the MSDLM. Sections 3.5 and 3.6 address state estimation and inference when parameters are assumed known and unknown, respectively. We apply the MSDLM to the motivating example of Section 3.2 in Section 3.7. We conclude and discuss future work in Section 3.8.

### 3.2 Multiscale Motivating Example

In 2014, Osthus et al. (2014) proposed the use of a dynamic linear model to describe the relationship between daily log electron flux and one day lagged solar wind speed in Earth’s radiation belts. The model of Osthus et al. (2014) is,
\[ y_t = \theta_{0,t} + \theta_{1,t} Vsw_{t-1} + v_t, \quad (3.1a) \]
\[ (\theta_{0,t}, \theta_{1,t})' = (\theta_{0,t-1}, \theta_{1,t-1})' + (w_{0,t}, w_{1,t})', \quad (3.1b) \]

where \( y_t \) is log electron flux, \( Vsw_t \) is standardized (mean centered and standard deviation scaled) solar wind speed, \( \theta_{0,t} \) is the latent intercept state, and \( \theta_{1,t} \) is the latent slope state, all on day \( t \). Additionally, \( v_t \overset{iid}{\sim} N(0,V) \) and \( (w_{0,t}, w_{1,t})' \overset{iid}{\sim} N(0,W) \), mutually independent of \( v_t \). Because of the centering in the standardization of \( Vsw_t \), the interpretation of \( \theta_{0,t} \) is approximately that of a moving average for log electron flux. This can be seen in Figure 3.1, where \( \theta_{0,t} \) roughly tracks the monthly means of log electron flux.

Figure 3.1: Daily log electron flux from January 1st, 1990 - December 31st, 2009 (gray), monthly averages of log electron flux (black) and posterior means of \( \theta_{0,t} \) (red).

As noted by Osthus et al. (2014), log electron flux can vary drastically at the daily scale while simultaneously varying systematically at an approximately 11 year scale corresponding to the solar cycle. Sunspot counts are indicators of the solar cycle. Figure 3.2 shows the monthly sunspot counts and the posterior means of \( \theta_{0,t} \), both standardized for ease of comparison. Figure 3.2 suggests sunspot information could be leveraged in the modeling of log electron flux.
Extending the model of Osthus et al. (2014) to incorporate sunspots, we fit the following model,

\[ y_t = \theta_{0,t} + \theta_{1,t} SSN_{t-1} + \theta_{2,t} Vsw_{t-1} + v_t \]  

\[ (\theta_{0,t}, \theta_{1,t}, \theta_{2,t})' = (\theta_{0,t-1}, \theta_{1,t-1}, \theta_{2,t-1})' + (w_{0,t}, w_{1,t}, w_{1,t})' \]  

where standardized daily sunspot counts, \( SSN_t \), have been produced by replicating standardized monthly sunspot counts at the daily level (i.e., all days within a month were assigned the same standardized sunspot count) in an effort to match the assumed common time indexing of the classical DLM. This replication, however, results in state unidentifiability. Because \( SSN_t \) is constant for each day within a month, there are essentially two intercepts in the dynamic regression of model 3.2, \( \theta_{0,t} \) and \( \theta_{1,t} SSN_{t-1} \). There is not, however, information available at a daily scale to uniquely partition the intercept \((\theta_{0,t} + \theta_{1,t} SSN_{t-1})\) into a non-solar cycle specific effect, \( \theta_{0,t} \), and a solar cycle specific effect, \( \theta_{1,t} SSN_{t-1} \), as the model appears to suggest. The measurable effect is the intercept, \( \theta_{0,t} + \theta_{1,t} SSN_{t-1} \). Evidence of this is shown in Figure 3.3. The top of Figure 3.3 plots \( \theta_{1,t} \) vs. \( \theta_{0,t} \) within three selected months and these time-indexed states are shown to be highly negatively correlated with one another. As with linear regression, issues related to interpretation and multicollinearity arise when estimates of parameters are
highly correlated. The top of Figure 3.3 displays what amounts to the dynamic regression version of multicollinearity. The bottom of Figure 3.3 plots the posterior means of $\theta_{0,t}$, $\theta_{1,t}SSN_{t-1}$, and $\theta_{0,t} + \theta_{1,t}SSN_{t-1}$ from July, 1996 through June, 1997. The $\theta_{0,t} + \theta_{1,t}SSN_{t-1}$ line effectively operates as a moving average quantity of the time series. The model decomposes the $\theta_{0,t} + \theta_{1,t}SSN_{t-1}$ line into two summable components, $\theta_{0,t}$ and $\theta_{1,t}SSN_{t-1}$. The interpretation of these components, separately, is however unclear due to the high correlation between them.

Figure 3.3: (top) Scatter plots of the posterior means of $\theta_{1,t}$ vs. $\theta_{0,t}$ for three selected months. (bottom) Daily log electron flux (gray) against time for July, 1996 - June 1997. The posterior means of the non-solar cycle specific intercept, $\theta_{0,t}$ (red), the solar cycle specific intercept, $\theta_{1,t}SSN_{t-1}$ (blue), and their sum, $\theta_{0,t} + \theta_{1,t}SSN_{t-1}$ (black) are plotted.

How to incorporate coarsely measured information about the slowly evolving solar cycle (i.e., monthly sunspot counts) into a model with a finer temporal index (day) is related to
the more general question of how to incorporate information collected at different scales into a unified modeling framework. As a way to answer this question, we develop and present the multiscale dynamic linear model (MSDLM) within the directed graphical model formalism.

3.3 Directed Graphical Models

We provide a brief primer on directed graphical models. For more details, we direct the reader to (Lauritzen, 1996, Chapters 2 and 3).

Let $\mathcal{G}(\mathcal{V}, \mathcal{E})$ be a graph where $\mathcal{V}$ is a finite set of vertices (nodes) and $\mathcal{E}$ is a subset of the set $\mathcal{V} \times \mathcal{V}$ ordered pairs of distinct vertices. An undirected edge is an ordered pair $(\nu_1, \nu_2)$ of distinct nodes in $\mathcal{V}$ such that $(\nu_1, \nu_2)$ and $(\nu_2, \nu_1)$ are in $\mathcal{E}$. A directed edge is an ordered pair $(\nu_1, \nu_2)$ of distinct nodes in $\mathcal{V}$ such that $(\nu_1, \nu_2)$ is in $\mathcal{E}$ but $(\nu_2, \nu_1)$ is not. Graphically, an undirected edge is represented by a line and a directed edge is represented by an arrow. A directed graph is a graph composed exclusively of directed edges. For illustration, Figure 3.4 displays a directed graph with $\mathcal{V} = \{a, b, c, d, e\}$ and $\mathcal{E} = \{(a, b), (b, e), (c, a), (c, b), (c, d), (d, e)\}$. 

Figure 3.4: Directed acyclic graph.

Consider a directed edge, with an arrow going from $\nu_1$ to $\nu_2$. Then, $\nu_1$ is said to be the parent of $\nu_2$ and $\nu_2$ is said to be the child of $\nu_1$. The set of parents of $\nu_2$ is denoted $\text{pa}(\nu_2)$ and the set of children of $\nu_1$ is denoted $\text{ch}(\nu_1)$. For example, in Figure 3.4, $\text{pa}(b) = \{a, c\}$ and $\text{ch}(b) = \{e\}$.

A path of length $n$ from $\nu_0$ to $\nu_n$ is a sequence $(\nu_0, \nu_1, \ldots, \nu_n)$ of distinct vertices in $\mathcal{V}$ such that $(\nu_{i-1}, \nu_i) \in \mathcal{E}$ for all $i = 1, 2, \ldots, n$. In Figure 3.4, $(c, a, b)$ is a path, but $(a, b, c)$ is not
a path, because the edge \((b,c) \notin \mathcal{E}\). A cycle is a path that starts and ends at the same node \((\nu_0 = \nu_n)\). A directed acyclic graph (DAG) is a directed graph with no cycles. Figure 3.4 is a DAG.

A directed graphical model is a family of probability distributions defined on a DAG (e.g., Jordan, 2004). Random variables are defined at each node and dependencies between random variables are specified by directed edges. For a directed acyclic graph \(\mathcal{G}(\mathcal{V}, \mathcal{E})\), let \(\mathbf{X} = \{X_\nu : \nu \in \mathcal{V}\}\) be a collection of random variables defined on the nodes in \(\mathcal{V}\). We adopt the notation where brackets around a random variable, (say, \(X\)), \([X]\), represents a short hand notation for the probability density (mass) function of \(X\). Given a collection of probability density (mass) functions \(\{[X_\nu|\text{pa}(X_\nu)] : \nu \in \mathcal{V}\}\) that integrate (sum) to one with respect to \(X_\nu\), the joint distribution of \(\mathbf{X}\) is defined as

\[
[\mathbf{X}] = \prod_{\nu \in \mathcal{V}} [X_\nu|\text{pa}(X_\nu)],
\]

where \([X_\nu|\text{pa}(X_\nu)]\) is the conditional distribution of \(X_\nu\) given the set of random variables defined on the parents of \(X_\nu\). As an illustration, the joint distribution corresponding to the directed graphical model of Figure 3.4 is

\[
[\mathbf{X}] = [X_a|X_c] [X_b|X_a, X_c] [X_c|X_c] [X_d|X_c] [X_e|X_b, X_d].
\]

### 3.3.1 Dynamic Linear Models

A useful directed graphical model for time series applications is the dynamic linear model (DLM) (West and Harrison, 1997). Consider a time series composed of observable scalars, \(y_t\), \(t = 1, 2, \ldots, T\). The underlying DAG structure for DLMs is shown in Figure 3.5.

![Figure 3.5: The directed acyclic graphical structure for DLMs.](image-url)
Conceptually, the DAG shown in Figure 3.5 captures temporal structure through the latent states, $\theta_t$. The latent state $\theta_t$ is dependent upon the previous state $\theta_{t-1}$, as indicated by directed edges. This is the graphical representation of the first-order Markov property. It is assumed we do not observe the system directly, but rather we observe a noisy estimate of the system, $y_t$.

The DLM coherently combines the latent states and observations through the following model specification:

$$y_t|\theta_t, \phi \sim N(F_t \theta_t, V_t),$$  \hspace{1cm} (3.5a)

$$\theta_t|\theta_{t-1}, \phi \sim N(G_t \theta_{t-1}, W_t),$$  \hspace{1cm} (3.5b)

$$\theta_0 \sim N(m_0, C_0).$$  \hspace{1cm} (3.5c)

Equation 3.5a is the observation equation, 3.5b is the state equation, and 3.5c is the initialization. States $\theta_t$ can be scalars or vectors. Quantities $F_t$ and $G_t$ are known matrices of appropriate dimensions. Hyperparameters $m_0$ and $C_0$ are considered known. $V_t$ is the observation variance and $W_t$ is the state (or evolution or process) variance, both at time $t$. $V_t$ and $W_t$ can be considered either known or unknown. Often the observation and state variances are assumed constant in time, where $V_t = V$ and $W_t = W$ for all $t = 1, 2, \ldots, T$. For notational convenience, we set $\phi = (V_1:T, W_1:T)$.

Let $y_{1:T} = (y_1, y_2, \ldots, y_T)$ and $\theta_{0:T} = (\theta_0, \theta_1, \ldots, \theta_T)$. Because the DLM is a special case of a directed graphical model, the joint distribution of all observations and latent states, $[y_{1:T}, \theta_{0:T}]$, of Figure 3.5 can be written following the prescription of equation 3.3. Specifically,

$$[y_{1:T}, \theta_{0:T}] = [\theta_0] \prod_{t=1}^{T} [y_t|\theta_t][\theta_t|\theta_{t-1}],$$  \hspace{1cm} (3.6)

where $pa(y_t) = \theta_t$, $pa(\theta_t) = \theta_{t-1}$ for $t \geq 1$, and $pa(\theta_0) = \emptyset$.

For illustration, consider the local level DLM,

$$y_t|\theta_t \sim N(\theta_t, V),$$

$$\theta_t|\theta_{t-1} \sim N(\theta_{t-1}, W),$$  \hspace{1cm} (3.7)

where $\theta_0 \sim N(0, W)$ and $V$ and $W$ are assumed known. Two time series were generated from the local level DLM under different choices of $V$ and $W$. They are plotted in Figure 3.6. In
the bottom of Figure 3.6, we see close agreement between the latent states and observations when $W$ is large relative to $V$ (i.e., when the signal-to-noise ratio, $W/V$, is large). When $W$ is small relative to $V$, however, we see poor agreement between the latent states and observations because the observations, $y_t$, are noisy measurements of the true, latent state, $\theta_t$.

![Figure 3.6: Generated time series from the local level DLM with $W = 0.1$, $V = 1$ (top) and $W = 2$, $V = 1$ (bottom). The black line represents the simulated observations, $y_{1:240}$, and the red line represents the latent states, $\theta_{0:240}$.](image)

3.3.2 Multiscale Dynamic Linear Models

We propose a novel directed graphical model, the multiscale dynamic linear model (MS-DLM), that can be viewed as the multiscale extension to the DLM. Consider a time series composed of observable scalars, $y_{s,t}$, indexed by a coarse scale time unit, $s = 1, 2, \ldots, S$, and a
fine scale time unit, \( t = 1, 2, \ldots, T \), nested within \( s \). An example of temporal nesting could be fine scale days \((t)\) nested within coarse scale months \((s)\). We consider only the situation where we have two scales, referred to as the coarse scale and the fine scale.

The MSDLM, like the DLM, is graphically represented by a DAG. It is shown in Figure 3.7. The observations, \( y_{s,t} \), are assumed to be observed at the fine scale. Latent states \( \beta_{s,t} \) are defined at the fine scale. Latent states \( \alpha_s \) are defined at the coarse scale. Information available at the coarse scale but not the fine scale is attached to coarse scale latent states, \( \alpha_s \). Information available at the fine scale is attached to both fine and coarse scale latent states, \( \beta_{s,t} \) and \( \alpha_s \). Observations \( y_{s,t} \) are dependent upon both coarse and fine scale states, as \( \text{pa}(y_{s,t}) = \{\alpha_s, \beta_{s,t}\} \).

![Diagram of the Directed Acyclic Graphical Structure for MSDLMs](image)

Conceptually, the DAG shown in Figure 3.7 captures large scale temporal structure in the coarse scale states, \( \alpha_s \), and captures small scale deviations from large scale structure in the fine scale states, \( \beta_{s,t} \). States are dependent on the previous state (i.e., \( \alpha_s \) depends on \( \alpha_{s-1} \) and \( \beta_{s,t} \) depends on \( \beta_{s,t-1} \)) as indicated by directed edges. Again, this is the graphical representation of the first-order Markov property. Within a coarse scale time unit, the fine scale graphical structure is that of the DLM (Figure 3.5). If observations and fine scale states were averaged over (in some sense) such that all temporal indexing occurred at the coarse scale \( s \), the graphical structure of the MSDLM would be similar to that of a DLM. Thus, there is temporal structure at the coarse scale and temporal structure at the fine scale, within the coarse scale.
The MSDLM is specified as
\[
y_{s,t}|\alpha_s, \beta_{s,t} \sim N(F_{\alpha,s,t}\alpha_s + F_{\beta,s,t}\beta_{s,t}, V_{s,t}), (3.8a)
\]
\[
\alpha_s|\alpha_{s-1} \sim N(G_{\alpha,s}\alpha_{s-1}, W_{\alpha,s}), (3.8b)
\]
\[
\alpha_0 \sim N(a_{\alpha,0}, A_{\alpha,0}), (3.8c)
\]
\[
\beta_{s,t}|\beta_{s,t-1} \sim N(G_{\beta,s,t}\beta_{s,t-1}, W_{\beta,s,t}) \text{ for } t > 1, (3.8d)
\]
\[
\beta_{s,1} \sim N(0, W_{\beta,s,1}) \text{ for all } s. (3.8e)
\]
Equation 3.8a is the observation equation, equations 3.8b and 3.8c are the coarse scale state equation and coarse scale initialization, respectively, and equations 3.8d and 3.8e are the fine scale state equation and fine scale initialization, respectively. States \(\alpha_s\) and \(\beta_{s,t}\) can be scalars or vectors. \(F_{\alpha,s,t}, F_{\beta,s,t}, G_{\alpha,s},\) and \(G_{\beta,s,t}\) are known matrices of appropriate dimensions. Hyperparameters \(a_{\alpha,0}\) and \(A_{\alpha,0}\) in the coarse scale initialization are considered known. We will consider \(W_{\alpha,s} = W_\alpha\) for all \(s\), \(W_{\beta,s,t} = W_{\beta,s}\) for all \(t\) within \(s\), and \(V_{s,t} = V\) for all \(s\) and \(t\).

The mean zero prior on \(\beta_{s,1}\) for all \(s\) admits the interpretation that fine scale states, \(\beta_{s,t}\), are deviations from coarse scale states, \(\alpha_s\).

Like the DLM, the MSDLM is a special case of a directed graphical model. Thus, the joint distribution of observations and latent states, \([y, \alpha, \beta]\), can also be written following the prescription of equation 3.3. Namely,
\[
[y, \alpha, \beta] = [a_0] \prod_{s=1}^{S} [\alpha_s|\alpha_{s-1}] \times \prod_{s=1}^{S} \left( \prod_{t=2}^{T} [\beta_{s,t}|\beta_{s,t-1}] \right) \times \prod_{s=1}^{S} \prod_{t=1}^{T} [y_{s,t}|\alpha_s, \beta_{s,t}], (3.9)
\]
where \(y = (y_1, y_2, \ldots, y_S)'\), \(y_s = (y_{s,1}, y_{s,2}, \ldots, y_{s,T})'\), \(\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_S)'\), \(\beta = (\beta_1, \beta_2, \ldots, \beta_S)'\), and \(\beta_s = (\beta_{s,1}, \beta_{s,2}, \ldots, \beta_{s,T})'\). Furthermore, \(pa(y_{s,t}) = \{\alpha_s, \beta_{s,t}\}\), \(pa(\alpha_s) = \alpha_{s-1}\) for \(s \geq 1\), \(pa(\alpha_0) = \emptyset,\) \(pa(\beta_{s,t}) = \beta_{s,t-1}\) for \(t \geq 2\), and \(pa(\beta_{s,1}) = \emptyset\) for all \(s\).

For illustration, consider the local level MSDLM,
\[
y_{s,t}|\alpha_s, \beta_{s,t} \sim N(\alpha_s + \beta_{s,t}, V),
\]
\[
\alpha_s|\alpha_{s-1} \sim N(\alpha_{s-1}, W_\alpha), \text{ for } s \geq 1, (3.10)
\]
\[
\beta_{s,t}|\beta_{s,t-1} \sim N(\beta_{s,t-1}, W_\beta), \text{ for } t \geq 2,
\]
where \(\alpha_0 \sim N(0, W_\alpha)\), \(\beta_{s,1} \sim N(0, W_\beta)\) for all \(s = 1, 2, \ldots, S\), and \(W_\alpha, W_\beta,\) and \(V\) are assumed known. Two time series were generated from the local level MSDLM under different choices.
of $V$, $W_\alpha$, and $W_\beta$. They are plotted in Figures 3.8 and 3.9 and illustrate the range of time series that can be generated by the local level MSDLM. Appreciable jumps at the coarse scale thresholds can be observed in Figure 3.9, while relatively smooth transitions across coarse scale time units can be seen in Figure 3.8. Large scale trend is governed by $\alpha_s$ and fine scale trend is governed by $\alpha_s + \beta_{s,t}$. The fine scale state, $\beta_{s,t}$, allows the model to make local adjustments within a coarse scale time unit, while the coarse scale state, $\alpha_s$, provides large scale stability across the time series. The partitioning of large scale stability and small scale adaptability makes the MSDLM quite flexible.

![Generated time series from the local level MSDLM](image)

Figure 3.8: Generated time series from the local level MSDLM with $S = 8$ and $T = 30$. The black line represents the simulated observations ($y_{s,t}$), the blue line represents the simulated coarse scale state ($\alpha_s$), and the red line represents the simulated coarse plus fine scale states ($\alpha_s + \beta_{s,t}$). $W_\alpha = 0.2$, $W_\beta = 0.01$, and $V = 1$. 
Figure 3.9: Generated time series from the local level MSDLM with $S = 8$ and $T = 30$. The black line represents the simulated observations ($y_{s,t}$), the blue line represents the simulated coarse scale state ($\alpha_s$), and the red line represents the simulated coarse plus fine scale states ($\alpha_s + \beta_{s,t}$). $W_\alpha = 0.25$, $W_\beta = 0.5$, and $V = 1$.

3.4 Directed Global Markov Property

Before addressing estimation for the MSDLM, we present the directed global Markov property (DGMP). For directed graphical models, the DGMP provides a sufficient condition for establishing conditional independence between two sets of random variables, given a third set of random variables. The DGMP will be utilized in the estimation procedures presented in Section 3.5.

Recall that a path from $\nu_0$ to $\nu_n$ is a sequence $(\nu_0, \nu_1, \ldots, \nu_n)$ of distinct vertices in $\mathcal{V}$ such that $(\nu_{i-1}, \nu_i) \in \mathcal{E}$ for all $i = 1, 2, \ldots, n$. A set $S \subseteq \mathcal{V}$ is said to be a $(\nu_1, \nu_2)$-separator if all paths from $\nu_1$ to $\nu_2$ intersect $S$. A set $S$ is said to separate $V_1$ from $V_2$ if $S$ is a $(\nu_1, \nu_2)$-separator for all $\nu_1 \in V_1$ and $\nu_2 \in V_2$. Consider Figure 3.4. There are three paths from $c$ to $e$: $(c, d, e)$,
(c, b, e), and (c, a, b, e). If \( S = \{b, d\} \), then \( S \) separates \( c \) from \( e \), as all paths from \( c \) to \( e \) intersect \( S \).

The vertices \( \nu_1 \) such that there exists a path from \( \nu_1 \) to \( \nu_2 \) and there does not exist a path from \( \nu_2 \) to \( \nu_1 \) are the ancestors of \( \nu_2 \), denoted \( \text{an}(\nu_2) \). For \( b \) in Figure 3.4, \( \text{an}(b) = \{a, c\} \) because there are paths from \( a \) to \( b \) and from \( c \) to \( b \) but not from \( b \) to \( a \) or \( b \) to \( c \). There are no paths from \( d \) or \( e \) to \( b \).

A DAG, \( \mathcal{G} \), has a corresponding moral graph, \( \mathcal{G}^m \). This moral graph is essentially an undirected “graphical cover” of the underlying DAG. To moralize a DAG, do the following:

1. For each \( \nu \in \mathcal{V} \) with two or more parents, draw an undirected edge between each pair of parents. This is called “marrying” parents.

2. Change all directed edges to undirected edges.

The resulting undirected graph is the moral graph, \( \mathcal{G}^m \), of the DAG, \( \mathcal{G} \). For illustration, consider the graphs in Figures 3.10 and 3.11. The graph of Figure 3.10 is a directed acyclic graph and its corresponding moral graph is shown in Figure 3.11. The bold lines in Figure 3.11 represent “marriages” between parents.

With the necessary terminology introduced, we present the DGMP as presented in Section 3.2.2 of Lauritzen (1996).

**Proposition 3.1** (directed global Markov property). Consider a joint probability distribution defined on a DAG, \( \mathcal{G}(\mathcal{V}, \mathcal{E}) \), that factorizes as in equation 3.3. Let \( V_1, V_2, \) and \( S \) be disjoint subsets of \( \mathcal{V} \). Let \( X_{V_1} = \{X_{\nu_1} : \nu_1 \in V_1\} \). Similarly define \( X_{V_2} \) and \( X_S \). Then \( X_{V_1} \) is condition-
ally independent of $X_{V_2}$, given $X_S$ whenever $V_1$ and $V_2$ are separated by $S$ in $(\mathcal{G}_{an(V_1 \cup V_2 \cup S)})^m$, the moral graph of the smallest ancestral set containing $V_1 \cup V_2 \cup S$.

Stated another way, Proposition 3.1 says if $V_1$ and $V_2$ are separated by $S$ in $(\mathcal{G}_{an(V_1 \cup V_2 \cup S)})^m$, then $[X_{V_1}, X_{V_2}, X_S] = [X_{V_1}, X_S]$.

For illustration, consider the question, is $X_e$ conditionally independent of $X_b$ given $X_c$ in Figure 3.12? By Proposition 3.1, if $e$ is separated from $b$ by $c$ in $(\mathcal{G}_{an(e \cup b \cup c)})^m$, then $[X_e | X_c, X_b] = [X_e | X_c]$. To answer this question, we first determine the smallest ancestral set containing $\{e, b, c\}$. Notice that $an(e) = \{a, b, c, g\}$, $an(b) = \{a\}$, and $an(c) = \{a, b\}$. Then, the smallest ancestral set containing $\{e, b, c\}$ is $\{a, b, c, e, g\}$. The subgraph induced by the smallest ancestral set containing $\{e, b, c\}$, $\mathcal{G}_{an(e \cup b \cup c)}$, is shown in Figure 3.13. To determine $(\mathcal{G}_{an(e \cup b \cup c)})^m$, we moralize the DAG in Figure 3.13. Figure 3.14 displays $(\mathcal{G}_{an(e \cup b \cup c)})^m$.

Given the moralized graph in Figure 3.14, it is immediately apparent that $e$ and $b$ are separated by $c$ in $(\mathcal{G}_{an(e \cup b \cup c)})^m$ because all paths from $b$ to $e$ intersect $c$. Thus by Proposition 3.1, $[X_e | X_c, X_b] = [X_e | X_c]$.

### 3.5 Estimation and Inference: Parameters Known

In this section we consider the problems of filtering (3.5.1), smoothing (3.5.2), and forecasting (3.5.3) with parameters known. The parameters of the DLM are $W_t$ and $V_t$ for all $t$ and the parameters of the MSDLM are $W_{\alpha,s}$, $W_{\beta,s,t}$, and $V_{s,t}$ for all $s$ and $t$. Within each section, we first present the problem and corresponding solution for the DLM as presented by Kalman
(1960), followed by the analogous problem and solution for the novel MSDLM. The estimation procedures for both the DLM and the MSDLM leverage the DGMP.

Using standard multivariate Gaussian results, it can be shown that \((\theta_{0:T}, y_{1:T})\) is jointly Gaussian for completely specified DLMs \((F_t, G_t, V_t, W_t, m_0, C_0)\) known for all \(t\). Thus, any marginals or conditionals will also be Gaussian. The same holds for completely specified MSDLMs \((F_{\alpha,s,t}, F_{\beta,s,t}, G_{\alpha,s}, G_{\beta,s,t}, V_{s,t}, W_{\alpha,s}, W_{\beta,s,t}, a_{\alpha,0}, A_{\alpha,0})\) known for all \(s,t\). Because Gaussian distributions are specified by their first two moments, determining the filtering, smoothing, and forecasting distributions reduces to determining their first two moments.

The filtering, smoothing, and forecasting results for the DLM follow that of Petris et al. (2009), Chapter 2.

### 3.5.1 Filtering

Filtering is a present analysis as the filtering distribution for the DLM (MSDLM) refers to the distribution of the latent state at time \(t\) \((s,t)\) given the observations up to time \(t\) \((s,t)\).

#### 3.5.1.1 DLM Filtering

In his seminal work, Kalman (1960) addressed the problem of filtering by presenting the Kalman filter. The Kalman filter is recursively specified. Given that \(\theta_{t-1|y_{1:t-1}} \sim N(m_{t-1}, C_{t-1})\), the Kalman filter provides a prescription for determining \(\theta_t|y_{1:t} \sim N(m_t, C_t)\). The Kalman filter is presented in Proposition 3.2.

**Proposition 3.2 (DLM Kalman filter).** Given \([\theta_{t-1|y_{1:t-1}}] \overset{d}{=} N(m_{t-1}, C_{t-1})\), the following statements hold:

(a) The one-step-ahead predictive distribution of \(\theta_t\) given \(y_{1:t-1}\) is Gaussian with

\[
a_t = E(\theta_t|y_{1:t-1}) = G_t m_{t-1},
\]

\[
R_t = Var(\theta_t|y_{1:t-1}) = G_t C_{t-1} G_t' + W_t.
\]
(b) The one-step-ahead forecasting distribution of \( y_t \) given \( y_{1:t-1} \) is Gaussian with

\[
\begin{align*}
    f_t &= E(y_t|y_{1:t-1}) = F_t a_t, \quad (3.12a) \\
    Q_t &= \text{Var}(y_t|y_{1:t-1}) = F_t R_t F_t' + V_t. \quad (3.12b)
\end{align*}
\]

(c) The filtering distribution of \( \theta_t \) given \( y_{1:t} \) is Gaussian with

\[
\begin{align*}
    m_t &= E(\theta_t|y_{1:t}) = a_t + R_t F_t' Q_t^{-1} (y_t - f_t), \quad (3.13a) \\
    C_t &= \text{Var}(\theta_t|y_{1:t}) = R_t - R_t F_t' Q_t^{-1} F_t R_t. \quad (3.13b)
\end{align*}
\]

**Proof.** The proof is in Appendix C.

Figure 3.15 shows a realization generated from a local level DLM with \( V = 1 \) and \( W = 0.1 \). The top of Figure 3.15 displays the filtering mean (black line) and 95% probability limits (gray band). The bottom of Figure 3.15 displays the one-step-ahead forecasts (black line) and 95% prediction limits (gray band). The 95% prediction limits for the one-step-ahead forecasts are wider than the 95% probability limits for the filtering distributions, as the forecasting variance for the local level DLM at time \( t \) is \( R_t + V \), while the filtering variance at time \( t \) is \( R_t - R_t Q_t^{-1} R_t \).
3.5.1.2 MSDLM Filtering

The MSDLM Kalman filter is a recursive specification for determining $[\alpha_s, \beta_{s,t}|y_{s,t}^*]$ given $[\alpha_s, \beta_{s,t-1}|y_{s,t-1}^*]$, where $y_{s,t}^* = (y_{1,1}, y_{1,2}, \ldots, y_{s,t})$. For convenience, in what follows we write $(\alpha'_s, \beta'_{s,t})'$ as $(\alpha_t, \beta_{s,t})'$. The MSDLM Kalman filter is presented in Proposition 3.3.

**Proposition 3.3** (MSDLM Kalman filter). Given

$$
\begin{pmatrix}
\alpha_s \\
\beta_{s,t-1}
\end{pmatrix}
| y_{s,t-1}^* \sim N
\begin{pmatrix}
\begin{bmatrix} a_{\alpha,s,t-1} \\ a_{\beta,s,t-1} \end{bmatrix} \\
A_{\alpha,s,t-1} \\
A'_{\alpha,s,t-1} \\
A_{\beta,s,t-1}
\end{bmatrix},

\begin{bmatrix}
A_{\alpha,s,t-1} & A_{\alpha\beta,s,t-1} \\
A'_{\alpha\beta,s,t-1} & A_{\beta,s,t-1}
\end{bmatrix}
$$

the following statements hold:

(a) For the one-step-ahead predictive distribution
(i) If \( t = 1 \), then \((\alpha_s, \beta_{s,t})'\) given \( y^*_{s-1,T} \) is \( N(b_{s,1}, B_{s,1}) \), where

\[
b_{s,1} = E((\alpha_s, \beta_{s,1})'|y^*_{s-1,T}) = (G_{\alpha,s}a_{\alpha,s-1,T}, 0)',
\]

\[
B_{s,1} = \text{Var}((\alpha_s, \beta_{s,1})'|y^*_{s-1,T}) = \begin{bmatrix}
G_{\alpha,s}A_{\alpha,s-1,T}G'_{\alpha,s} + W_{\alpha,s} & 0 \\
0 & W_{\beta,s,1}
\end{bmatrix}.
\]

(ii) If \( t \neq 1 \), then \((\alpha_s, \beta_{s,t})'\) given \( y^*_{s,t-1} \) is \( N(b_{s,t}, B_{s,t}) \), where

\[
b_{s,t} = E((\alpha_s, \beta_{s,t})'|y^*_{s,t-1}) = (a_{\alpha,s,t-1}, G\beta_{s,t}a_{\beta,s,t-1})',
\]

\[
B_{s,t} = \text{Var}((\alpha_s, \beta_{s,t})'|y^*_{s,t-1}) = \begin{bmatrix}
A_{\alpha,s,t-1} & A_{\alpha\beta,s,t-1}G'_{\beta,s,t} \\
G\beta_{s,t}A'_{\alpha\beta,s,t-1} & G\beta_{s,t}A_{\beta,s,t-1}G'_{\beta,s,t} + W_{\beta,s,t}
\end{bmatrix}.
\]

(b) The one-step-ahead forecasting distribution for \( y_{s,t} \) given \( y^*_{s,t-1} \) is \( N(d_{s,t}, D_{s,t}) \), with

\[
d_{s,t} = E(y_{s,t}|y^*_{s,t-1}) = F_{s,t}b_{s,t},
\]

\[
D_{s,t} = \text{Var}(y_{s,t}|y^*_{s,t-1}) = F_{s,t}B_{s,t}F'_{s,t} + V_{s,t},
\]

where \( F_{s,t} \equiv [F_{\alpha,s,t}, F_{\beta,s,t}] \).

(c) The filtering distribution of \((\alpha_s, \beta_{s,t})'\) given \( y^*_{s,t} \) is \( N(a_{s,t}, A_{s,t}) \), with

\[
a_{s,t} = E(\theta_{s,t}|y^*_{s,t}) = b_{s,t} + B_{s,t}F_{s,t}D_{s,t}^{-1}(y_{s,t} - d_{s,t}),
\]

\[
A_{s,t} = \text{Var}(\theta_{s,t}|y^*_{s,t}) = B_{s,t} - B_{s,t}F_{s,t}D_{s,t}^{-1}F_{s,t}B_{s,t}.
\]

Proof. The proof is provided in Appendix D. \(\Box\)

Figure 3.16 shows a realization generated from a local level MSDLM with \( V = 1 \), \( W_{\alpha} = 1 \), and \( W_{\beta} = .05 \). The top of Figure 3.16 displays the filtering mean and 95% probability limits for the coarse state \( \alpha_s \) as well as the true value of the state \( \alpha_s \). We see that the filtering distribution is variable at the beginning of a coarse scale time unit but eventually settles to a distribution. The middle of Figure 3.16 displays the filtering mean and 95% probability limits for the sum of the coarse and fine scale states, \( \alpha_s + \beta_{s,t} \), and the sum of the true states. The bottom of Figure 3.16 shows the one-step-ahead forecasts and 95% prediction limits.
Figure 3.16: MSDLM filtering for only the coarse scale state (top), the coarse scale state plus the fine scale state (middle) and one-step-ahead forecasts (bottom). Observations are points. Filtering means (top and middle) and point forecasts (bottom) are black lines. Blue lines indicate the true coarse scale latent state, $\alpha_s$. Red lines indicate the true latent state, $\alpha_s + \beta_{s,t}$. Gray bands indicate 95% probability limits.
3.5.2 Smoothing

Smoothing is a retrospective analysis as the smoothing distribution of the DLM (MSDLM) refers to the distribution of the latent state at time \( t \) \((s, t)\) given all the available observations up to time \( T \) \((S, T)\).

3.5.2.1 DLM Smoothing

Given \( \theta_{t+1}|y_{1:T} \sim N(s_{t+1}, S_{t+1}) \), the Kalman smoother provides a prescription for determining \( \theta_t|y_{1:T} \sim N(s_t, S_t) \). The Kalman smoother is presented in Proposition 3.4.

**Proposition 3.4** (DLM Kalman smoother). Given \( \theta_{t+1}|y_{1:T} \sim N(s_{t+1}, S_{t+1}) \), \( \theta_t|y_{1:T} \sim N(s_t, S_t) \), where

\[
\begin{align*}
    s_t &= m_t + C_t G_t^{t+1} R_{t+1}^{-1} (s_{t+1} - a_{t+1}), \\
    S_t &= C_t - C_t G_t^{t+1} R_{t+1}^{-1} (R_{t+1} - S_{t+1}) R_{t+1}^{-1} G_{t+1} C_t.
\end{align*}
\]

(3.18a) 
(3.18b)

**Proof.** The proof is in Appendix E.

Figure 3.17 displays the same simulated data as in Figure 3.15. The smoothing means and 95% probability limits are shown by the black line and gray band, respectively. The smoothing distribution has the effect of “smoothing out” the estimates for the latent states.

Figure 3.17: DLM smoothing distributions. Observations are points. The smoothing means and 95% probability limit are shown by the black line and gray bands, respectively.
3.5.2.2 MSDLM Smoothing

The MSDLM Kalman smoother is a recursive prescription for determining $[\alpha_s, \beta_{s,t}|y^*_S,T]$ given $[\alpha_s, \beta_{s,t+1}|y^*_S,T]$. The MSDLM Kalman smoother is presented in Proposition 3.5.

**Proposition 3.5** (MSDLM Kalman smoother). The MSDLM Kalman smoothing recursions are partitioned into two cases: 1) $t \neq T$ and 2) $s \neq S$ and $t = T$. Note that if $s = S$ and $t = T$, then $(\alpha_s, \beta_{S,T})'|y^*_S,T$ is known and is the filtering distribution.

(a) For $t \neq T$, given

$$ (\alpha_s, \beta_{s,t+1})'|y^*_S,T \sim N \left( \begin{bmatrix} h_{\alpha,s} \\ h_{\beta,s,t+1} \end{bmatrix}, \begin{bmatrix} H_{\alpha,s} & H_{\alpha \beta,s,t+1} \\ H'_{\alpha \beta,s,t+1} & H_{\beta,s,t+1} \end{bmatrix} \right), $$

then

$$ (\alpha_s, \beta_{s,t})'|y^*_S,T \sim N \left( \begin{bmatrix} h_{\alpha,s} \\ h_{\beta,s,t} \end{bmatrix}, \begin{bmatrix} H_{\alpha,s} & H_{\alpha \beta,s,t} \\ H'_{\alpha \beta,s,t} & H_{\beta,s,t} \end{bmatrix} \right), $$

where

$$ h_{\beta,s,t} = E(a_{\beta,s,t}|\alpha_s |y^*_S,T) + R_{\beta,s,t} (h_{\beta,s,t+1} - G_{\beta,s,t+1} E(a_{\beta,s,t}|\alpha_s |y^*_S,T)), \quad (3.19a) $$

$$ H_{\beta,s,t} = Q_{\beta,s,t} + H'_{\alpha \beta,s,t} H_{\alpha,s} H_{\alpha \beta,s,t}, \quad (3.19b) $$

$$ H_{\alpha \beta,s,t} = H_{\alpha,s} A_{\alpha,s,t}^{-1} A_{\alpha \beta,s,t}(I - R_{\beta,s,t} G_{\beta,s,t+1})', \quad (3.19c) $$

and

$$ E(a_{\beta,s,t}|\alpha_s |y^*_S,T) = a_{\beta,s,t} + A'_{\alpha \beta,s,t} A_{\alpha,s,t}^{-1} (h_{\alpha,s} - a_{\alpha,s,t}), $$

$$ Q_{\beta,s,t} = (I - R_{\beta,s,t} G_{\beta,s,t+1}) A_{\beta,s,t}|\alpha_s, $$

$$ R_{\beta,s,t} = A_{\beta,s,t}|\alpha_s G'_{\beta,s,t+1} (G_{\beta,s,t+1} A_{\beta,s,t}|\alpha_s A'_{\beta,s,t+1} G_{\beta,s,t+1} + W_{\beta,s,t+1})^{-1}, $$

$$ A_{\beta,s,t}|\alpha_s = A_{\beta,s,t} - A'_{\alpha \beta,s,t} A_{\alpha,s,t}^{-1} A_{\alpha \beta,s,t}. $$

(b) For $s \neq S$ and $t = T$, given

$$ (\alpha_{s+1}, \beta_{s+1})'|y^*_S,T \sim N \left( \begin{bmatrix} h_{\alpha,s+1} \\ h_{\beta,s+1} \\ H_{\alpha,s+1} \end{bmatrix}, \begin{bmatrix} H_{\alpha,s+1} & H_{\alpha \beta,s+1,1} \\ H'_{\alpha \beta,s+1,1} & H_{\beta,s+1,1} \end{bmatrix} \right), $$
then
\[(\alpha_s, \beta_{s,T})' | y^*_{s,T} \sim N \left( \begin{bmatrix} h_{\alpha,s} \\ h_{\beta,s,T} \end{bmatrix} , \begin{bmatrix} H_{\alpha,s} & H_{\alpha\beta,s,T} \\ H'_{\alpha\beta,s,T} & H_{\beta,s,T} \end{bmatrix} \right) , \]

where
\[
\begin{align*}
h_{\alpha,s} &= a_{\alpha,s,T} + R_{\alpha,s}(h_{\alpha,s+1} - G_{\alpha,s+1}a_{\alpha,s,T}), \\
h_{\beta,s,T} &= a_{\beta,s,T} + A'_{\alpha\beta,s,T}A_{\alpha,s,T}^{-1}(h_{\alpha,s} - a_{\alpha,s,T}), \\
H_{\alpha,s} &= Q_{\alpha,s} + R_{\alpha,s}H_{\alpha,s} + R'_{\alpha,s}, \\
H_{\beta,s,T} &= A_{\beta,s,T|\alpha_s} + A'_{\alpha\beta,s,T}A_{\alpha,s,T}^{-1}H_{\alpha,s}A_{\alpha,s,T}^{-1}A_{\alpha\beta,s,T}, \\
H_{\alpha\beta,s,T} &= H_{\alpha,s}A_{\alpha,s,T}^{-1}A_{\alpha\beta,s,T},
\end{align*}
\]

and
\[
\begin{align*}
Q_{\alpha,s} &= A_{\alpha,s,T} - R_{\alpha,s}G_{\alpha,s+1}A_{\alpha,s,T}, \\
R_{\alpha,s} &= A_{\alpha,s,T}G_{\alpha,s+1,B_{\alpha,s+1,1},1}, \\
A_{\beta,s,T|\alpha_s} &= A_{\beta,s,T} - A'_{\alpha\beta,s,T}A_{\alpha,s,T}^{-1}A_{\alpha\beta,s,T}.
\end{align*}
\]

**Proof.** The proof is provided in Appendix F.

Figure 3.18 displays the same simulated data as in Figure 3.16. The smoothing means and 95% probability limits are shown in the top of Figure 3.18 along with the true coarse scale state, \(\alpha_s\). The smoothing distribution for \(\alpha_s\) is constant for all fine scale time points, \(t\), within \(s\), as \(\alpha_s\) is conditioned on all the data. This is in contrast to the filtering distribution shown in the top of Figure 3.16, where \(\alpha_s\) is conditioned on the presently available data, \(y^*_{s,t}\). The true coarse plus fine scale state, \(\alpha_s + \beta_{s,t}\), is shown in the bottom of Figure 3.18. The smoothing distribution has the effect of smoothing the state estimates as compared to the filtering distribution of \(\alpha_s + \beta_{s,t}\), shown in the middle of Figure 3.18.
Figure 3.18: MSDLM smoothing distributions for the coarse scale state (top) and coarse plus fine scale states (bottom). Observations are points. The smoothing means and 95% probability limit are shown by the black line and gray bands, respectively. The true, latent state for $\alpha_s$ is the blue line (top) and for $\alpha_s + \beta_{s,t}$ is the red line (bottom).

3.5.3 Forecasting

Forecasting is a forward looking analysis as the forecasting distribution for the DLM (MSDLM) refers to the distribution of a future observation at time $t + k$ ($s + l, t + k$), given the observations up to time $t$ ($s, t$).

3.5.3.1 DLM Forecasting Recursions

Given $k \geq 1$ and $\theta_t | y_{1:t} \sim N(m_t, C_t)$, the DLM forecasting equations provide a prescription for determining $y_{t+k} | y_{1:t}$. The DLM forecasting recursions are presented in Proposition 3.6.
Proposition 3.6 (DLM forecasting recursions). Set $a_t(0) = m_t$ and $R_t(0) = C_t$. Given $[\theta_t | y_{1:t}] \sim N(a_t(0), R_t(0))$ and for $k \geq 1$, the following statements are true.

(a) The many-step-ahead predictive distribution for $\theta_{t+k}$ given $y_{1:t}$ is Gaussian with

$$a_t(k) = E(\theta_{t+k} | y_{1:t}) = G_{t+k}a_t(k-1),$$

$$R_t(k) = Var(\theta_{t+k} | y_{1:t}) = G_{t+k}G_t' + W_{t+k}. \quad (3.21a)$$

(b) The many-step-ahead forecasting distribution for $y_{t+k}$ given $y_{1:t}$ is Gaussian with

$$f_t(k) = E(y_{t+k} | y_{1:t}) = F_{t+k}a_t(k),$$

$$Q_t(k) = Var(y_{t+k} | y_{1:t}) = F_{t+k}R_t(k)F_t' + V_{t+k}. \quad (3.22b)$$

Proof. The proof is provided in Appendix G. \qed

3.5.3.2 MSDLM Forecasting Recursions

The MSDLM forecasting recursions are partitioned into two cases. The first case is forecasting within the same coarse scale time unit, $[(\alpha_s, \beta_{s,t+k})|y_{s,t}^*]$ for $k \geq 1$ and $t + k \leq T$. The second case is forecasting outside of the same coarse scale time unit, $[(\alpha_{s+l}, \beta_{s+l,t+k})|y_{s,t}^*]$ for $l \geq 1$ and $1 \leq t + k \leq T$. The MSDLM forecasting recursions are presented in Proposition 3.7.

Proposition 3.7 (MSDLM forecasting recursions). Given

$$\begin{pmatrix} \alpha_s \\ \beta_{s,t} \end{pmatrix} | y_{s,t}^* \sim N \begin{pmatrix} b_{\alpha,s} | y_{s,t}^* \\ b_{\beta,s,t} | y_{s,t}^* \end{pmatrix}, \begin{pmatrix} B_{\alpha,s} | y_{s,t}^* & B_{\alpha \beta,s,t} | y_{s,t}^* \\ B'_{\alpha \beta,s,t} | y_{s,t}^* & B_{\beta,s,t} | y_{s,t}^* \end{pmatrix}, \quad (3.23)$$

where $b_{\alpha,s} | y_{s,t}^* = a_{\alpha,s,t}$, $b_{\beta,s,t} | y_{s,t}^* = a_{\beta,s,t}$, $B_{\alpha,s} | y_{s,t}^* = A_{\alpha,s,t}$, $B_{\beta,s,t} | y_{s,t}^* = A_{\beta,s,t}$, and $B_{\alpha \beta,s,t} | y_{s,t}^* = A_{\alpha \beta,s,t}$ as defined in Proposition 3.3, the following statements hold.

(a) For the many-step-ahead predictive distribution
Proof. Thus the filtering and smoothing equations do not directly apply. In Section 3.6, it is our
(b) For the many-step-ahead forecasting distribution

\[ b_{s,t+k|y_{s,t}} = E((\alpha_s, \beta_s,t+k)^T|y_{s,t}^*) = (b_{\alpha_s|y_{s,t}^*}, G_{\beta_s,t+k} b_{y_{s,t}^*}) \]  \hspace{1cm} (3.24a)

\[ B_{s,t+k|y_{s,t}} = Var((\alpha_s, \beta_s,t+k)^T|y_{s,t}^*) \]

\[ = \begin{bmatrix}
B_{\alpha_s|y_{s,t}^*} & B_{\alpha\beta_s,t+k-1|y_{s,t}^*} G'_{\beta_s,t+k} \\
G_{\beta_s,t+k} B_{\alpha\beta_s,t+k-1|y_{s,t}^*} G'_{\beta_s,t+k} + W_{\beta,s,t+k} & G_{\beta_s,t+k}
\end{bmatrix}. \]  \hspace{1cm} (3.24b)

(ii) For \( l \geq 1 \) and \( 1 \leq t+k \leq T \), \((\alpha_s+l, \beta_s+l,t+k)^T\) given \( y_{s,t}^* \) is \( N(b_{s+l,t+k|y_{s,t}^*}, B_{s+l,t+k|y_{s,t}^*}) \),

\[ b_{s+l,t+k|y_{s,t}^*} = E((\alpha_s+l, \beta_s+l,t+k)^T|y_{s,t}^*) = (G_{\alpha,s+l} b_{\alpha,s+l-1|y_{s,t}^*}, 0) \]  \hspace{1cm} (3.25a)

\[ B_{s+l,t+k|y_{s,t}^*} = Var((\alpha_s+l, \beta_s+l,t+k)^T|y_{s,t}^*) \]

\[ = \begin{bmatrix}
G_{\alpha,s+l} B_{\alpha,s+l-1|y_{s,t}^*} G'_{\alpha,s+l} + W_{\alpha,s+l} & 0 \\
0 & B_{\beta,s+l,t+k|y_{s,t}^*}
\end{bmatrix}. \]  \hspace{1cm} (3.25b)

where

\[ B_{\beta,s+l,t+k|y_{s,t}^*} = \begin{cases} 
W_{\beta,s+l,t+k} & \text{if } t+k = 1 \\
G_{\beta,s+l,t+k} B_{\beta,s+l,t+k-1|y_{s,t}^*} G'_{\beta,s+l,t+k} + W_{\beta,s+l,t+k} & \text{if } 2 \leq t+k \leq T.
\end{cases} \]

(b) For the many-step-ahead forecasting distribution

For both case (i) where \( l = 0, k \geq 1 \), and \( t+k \leq T \), and case (ii) where \( l \geq 1 \) and \( 1 \leq t+k \leq T \), \( y_{s,t+l+k} \) given \( y_{s,t}^* \) is \( N(d_{s+l,t+k|y_{s,t}^*}, D_{s+l,t+k|y_{s,t}^*}) \),

\[ d_{s+l,t+k|y_{s,t}^*} = E(y_{s+l,t+k|y_{s,t}^*}) = F_{s+l,t+k} b_{s+l,t+k|y_{s,t}^*} \]  \hspace{1cm} (3.26a)

\[ D_{s+l,t+k|y_{s,t}^*} = Var(y_{s+l,t+k|y_{s,t}^*}) = F_{s+l,t+k} B_{s+l,t+k|y_{s,t}^*} F'_{s+l,t+k} + V_{s+l,t+k} \]  \hspace{1cm} (3.26b)

where \( F_{s+l,t+k} \equiv [F_{\alpha,s+l,t+k}, F_{\beta,s+l,t+k}] \).

Proof. The proof is provided in Appendix H. \qed

3.6 Estimation and Inference: Parameters Unknown

In almost all non-simulation settings, parameters of the DLM and MSDLM are unknown. Thus the filtering and smoothing equations do not directly apply. In Section 3.6, it is our
objective to compute the posterior distribution \( \theta_{0:T, \phi|y_{1:T}} \) for the DLM or \( [\alpha, \beta, \phi|y^{*}_{S,T}] \) for the MSDLM, where \( \phi \) generically represents the parameters of the model. These posterior distributions are not available in closed form, except in rare cases (e.g., Petris et al., 2009, Section 4.3). In general, summaries of the posterior distribution will be produced via simulation. Namely, we consider the Markov chain Monte Carlo (MCMC) sampling scheme, Gibbs sampling (e.g., Geman and Geman, 1984).

Gibbs sampling for the DLM (MSDLM) requires sampling from two distributions, \( [\theta_{0:T|\phi, y_{1:T}}] \) \( \left( [\alpha, \beta|\phi, y^*_{S,T}] \right) \) and \( [\phi|\theta_{0:T}, y_{1:T}] \) \( \left( [\phi|\alpha, \beta, y^*_{S,T}] \right) \). Sampling the parameter vector \( \phi \) from its full conditional distribution will be problem specific for both the DLM and MSDLM. For the DLM, a general, efficient algorithm exist for sampling the latent state vector, \( \theta_{0:T} \), from its full conditional distribution. This algorithm is known as the forward-filtering, backward sampling (FFBS) algorithm (Carter and Kohn, 1994; Frühwirth-Schnatter, 1994; Shephard, 1994) and is presented in Section 3.6.1. For the MSDLM, we propose a general, efficient algorithm for sampling the latent states, \( (\alpha, \beta) \), from its full conditional distribution. This algorithm can be viewed as the multiscale extension to the FFBS algorithm. Hence, it is referred to as the MSDLM FFBS algorithm and is presented in Section 3.6.2.

### 3.6.1 DLM Forward-Filtering, Backward Sampling

The FFBS algorithm samples \( \theta_{0:T} \) jointly from its full conditional distribution, \( [\theta_{0:T}|\phi, y_{1:T}] \). This is accomplished by sequentially sampling \( \theta_{t} \) from \( [\theta_{t}|\theta_{(t+1):T}, \phi, y_{1:T}] \) for \( t = T, T-1, \ldots, 1, 0 \) and noting that

\[
[\theta_{t}|\theta_{(t+1):T}, \phi, y_{1:T}] = [\theta_{t}|\theta_{t+1}, \phi, y_{1:t}]. \tag{3.27}
\]

The distributional equivalence of 3.27 can be verified by applying Proposition 3.1 to the appropriate moral graph of Figure 3.5. Furthermore, in equation E.2, it was shown that \( [\theta_{t}|\theta_{t+1}, \phi, y_{1:t}] \overset{d}{=} N(h_{t}, H_{t}) \), where

\[
h_{t} = m_{t} + C_{t}G_{t+1}R_{t+1}^{-1}(\theta_{t+1} - a_{t+1}), \tag{3.28a}
\]

\[
H_{t} = C_{t} - C_{t}G_{t+1}R_{t+1}^{-1}G_{t+1}C_{t}. \tag{3.28b}
\]
The FFBS algorithm for the DLM is summarized in Algorithm 1.

**Algorithm 1: FFBS algorithm.**

1. Run the Kalman filter of Proposition 3.2
2. Draw $\theta_T \sim N(m_T, C_T)$
3. For $t = T - 1, T - 2, \ldots, 1, 0$, draw $\theta_t \sim N(h_t, H_t)$

To sample from $[\theta_{0:T}, \phi|y_{1:T}]$, we embed the FFBS algorithm within a Gibbs sampler. In general, the distribution of $\phi$ given $\theta_{0:T}$ and $y_{1:T}$ will be problem specific. The general Gibbs sampler for the DLM is outlined in Algorithm 2.

**Algorithm 2: General Gibbs sampler for the DLM.**

1. Initialize: Set $\phi = \phi^{(0)}$
2. For $i = 1, \ldots, N$:
3. Draw $\theta_{0:T}^{(i)}$ from $[\theta_{0:T}|\phi = \phi^{(i-1)}, y_{1:T}]$ using the FFBS algorithm
4. Draw $\phi^{(i)}$ from $[\phi|\theta_{0:T} = \theta_{0:T}^{(i)}, y_{1:T}]$

### 3.6.2 MSDLM Forward-Filtering, Backward Sampling

The MSDLM FFBS algorithm is constructed analogously to the FFBS algorithm for the DLM. The MSDLM FFBS algorithm samples $(\alpha, \beta)$ jointly from its full conditional distribution, $[\alpha, \beta|\phi, y_{S,T}^\ast]$, and does so sequentially. To accomplish this, note the following distributional equivalencies which can be verified by applying Proposition 3.1 to the appropriate moral graphs of Figure 3.7:

$$[\alpha_0|\alpha_{1:S}, \beta_{S,T}^*, \phi, y_{S,T}^*] = [\alpha_0|\alpha_1, \phi],$$  

$$[\alpha_s|\alpha_{(s+1):S}, \beta_{s,T}^{**}, \phi, y_{S,T}^*] = [\alpha_s|\alpha_{s+1}, \phi, y_{s,T}^*]$$  

$$[\beta_{s,T}|\alpha_{s:S}, \beta_{s,T}^{**}, \phi, y_{S,T}^*] = [\beta_{s,T}|\alpha_s, \phi, y_{s,T}^*]$$  

$$[\beta_{s,t}|\alpha_{s:S}, \beta_{s,t}^{**}, \phi, y_{S,T}^*] = [\beta_{s,t}|\alpha_s, \beta_{s,t+1}, \phi, y_{s,t}^*]$$  

(for $t \neq T$),
where $\beta_{s,t}^{**} = \beta \setminus \beta_{s,t}^{*}$ and $\beta_{s,t}^{*}$ is defined analogously to $y_{s,t}^{*}$. Furthermore, from the proof of Proposition 3.5 we have that

\[
\begin{align*}
\left[ \alpha_s | \alpha_{s+1}, \phi, y_{s,T}^s \right] & \overset{F.5}{=} N(q_{\alpha,s}, Q_{\alpha,s}), \\
\left[ \beta_{s,T} | \alpha_s, \phi, y_{s,T}^s \right] & \overset{F.2}{=} N(a_{\beta,s,T|\alpha_s}, A_{\beta,s,T|\alpha_s}), \\
\left[ \beta_{s,t} | \alpha_s, \beta_{s,t+1}, \phi, y_{s,t}^s \right] & \overset{F.3}{=} N(q_{\beta,s,t}, Q_{\beta,s,t}).
\end{align*}
\]

By Result 1, we have that $\alpha_0 | \alpha_1, \phi \sim N(q_{\alpha,0}, Q_{\alpha,0})$ where

\[
\begin{align*}
q_{\alpha,0} &= a_{\alpha,0} + A_{\alpha,0}G_{\alpha,1}'(G_{\alpha,1}A_{\alpha,0}G_{\alpha,1}' + W_{\alpha,1})^{-1}(\alpha_1 - G_{\alpha,1}a_{\alpha,0}), \\
Q_{\alpha,0} &= A_{\alpha,0} - A_{\alpha,0}G_{\alpha,1}'(G_{\alpha,1}A_{\alpha,0}G_{\alpha,1}' + W_{\alpha,1})^{-1}G_{\alpha,1}A_{\alpha,0}.
\end{align*}
\]

The MSDLM FFBS algorithm is outlined in Algorithm 3.

**Algorithm 3**: MSDLM FFBS algorithm.

1. Run the MSDLM Kalman filter of Proposition 3.3
2. For $s = S, S - 1, \ldots, 1$:
3. \hspace{1em} For $t = T, T - 1, \ldots, 1$:
4. \hspace{2em} if $s = S$ and $t = T$ then
5. \hspace{3em} Draw $(\alpha_S, \beta_{S,T}) \sim N(a_{S,T}, A_{S,T})$;
6. \hspace{2em} end
7. \hspace{2em} if $s \neq S$ and $t = T$ then
8. \hspace{3em} Draw $\alpha_s \sim N(q_{\alpha,s}, Q_{\alpha,s})$;
9. \hspace{3em} Draw $\beta_{s,T} \sim N(a_{\beta,s,T|\alpha_s}, A_{\beta,s,T|\alpha_s})$;
10. \hspace{2em} end
11. \hspace{2em} if $t \neq T$ then
12. \hspace{3em} Draw $\beta_{s,t} \sim N(q_{\beta,s,t}, Q_{\beta,s,t})$;
13. \hspace{2em} end
14. Draw $\alpha_0 \sim N(q_{\alpha,0}, Q_{\alpha,0})$
To sample from $\left[\alpha, \beta, \phi | y_{S,T}^*\right]$, we embed the MSDLM FFBS in a Gibbs sampler. Like with the DLM, in general, the distribution of $\phi$ given $(\alpha, \beta)$ and $y_{S,T}^*$ will be problem specific. The general Gibbs sampler for the MSDLM is outlined in Algorithm 4.

**Algorithm 4:** General Gibbs algorithm for the MSDLM.

1. Initialize: Set $\phi = \phi^{(0)}$
2. For $i = 1, \ldots, N$:
   3. Draw $(\alpha, \beta)^{(i)}$ from $[(\alpha, \beta)|\phi = \phi^{(i-1)}, y_{S,T}^*]$ using MSDLM FFBS algorithm
   4. Draw $\phi^{(i)}$ from $[\phi|(\alpha, \beta) = (\alpha, \beta)^{(i)}, y_{S,T}^*]$

If we assume $V_{s,t} = V$, $W_{\alpha,s} = W_{\alpha}$, and $W_{\beta,s,t} = W_{\beta,s}$ and select the following priors,

\[
V^{-1} \sim \text{Gamma}(a_v, b_v) \quad (3.30a)
\]

\[
W_{\alpha}^{-1} \sim \text{Wishart}(\alpha_{\alpha}, B_{\alpha}) \quad (3.30b)
\]

\[
W_{\beta,s}^{-1} \overset{iid}{\sim} \text{Wishart}(\alpha_{\beta}, B_{\beta}), \quad (3.30c)
\]

where $E(V^{-1}) = \frac{a_v}{b_v}$, $E(W_{\alpha}^{-1}) = \alpha_{\alpha}B_{\alpha}$, and hyperparameters $a_v, b_v, \alpha_{\alpha}, B_{\alpha}, \alpha_{\beta},$ and $B_{\beta}$ assumed known, then it is true that

\[
V^{-1}|[- \sim \text{Gamma}(a_v + 0.5ST, b_v + 0.5 \sum_{s=1}^{S} \sum_{t=1}^{T} (y_{s,t} - F_{\alpha,s,t} \alpha_s - F_{\beta,s,t} \beta_{s,t})^2) \quad (3.31a)
\]

\[
W_{\alpha}^{-1}|[- \sim \text{Wishart}(\alpha_{\alpha} + 0.5S, B_{\alpha} + 0.5 \sum_{s=1}^{S} (\alpha_s - G_{\alpha,s}) (\alpha_s - G_{\alpha,s} \alpha_{s-1})') \quad (3.31b)
\]

\[
W_{\beta,s}^{-1}|[- \sim \text{Wishart}(\alpha_{\beta} + 0.5T, B_{\beta} + 0.5 \sum_{t=1}^{T} (\beta_{s,t} - G_{\beta,s,t}) (\beta_{s,t} - G_{\beta,s,t} \beta_{s,t-1})') \quad (3.31c)
\]

where $\beta_{s,0} = 0$ and $X[-$ is the full conditional distribution of $X$.

### 3.7 Application: Space Weather

We return here to the dynamic regression of log electron flux on solar wind speed and explore the incorporation of sunspot information in the MSDLM framework.
An MSDLM analogue to the DLM in equations 3.2a and 3.2b is presented in 3.32 as

\[ y_{s,t} | \alpha_s, \beta_{s,t} \sim \mathcal{N}(\alpha_s F_{\alpha,s,t} + \beta_{s,t} F_{\beta,s,t}, V) \]
\[ \sim \mathcal{N}((\alpha_{1,s} SSN_{s-1} + \beta_{1,s,t}) + (\alpha_{2,s} + \beta_{2,s,t}) Vsw_{s,t-1}, V), \quad (3.32a) \]
\[ \alpha_s | \alpha_{s-1} \sim \mathcal{N}(\alpha_{s-1}, W_\alpha), \quad (3.32b) \]
\[ \beta_{s,t} | \beta_{s,t-1} \sim \mathcal{N}(\beta_{s,t-1}, W_\beta), \quad (3.32c) \]

where \( \beta_{s,1} \sim \mathcal{N}(0, W_{\beta,s}) \) for all \( s \) and \( \alpha_0 \sim \mathcal{N}(0, I) \), where \( I \) is the identity matrix. Months are indexed by \( s \), and days are indexed by \( t \). Monthly centered sunspot information at month \( s \) corresponds to \( SSN_s \). Replicating sunspot counts at the daily scale is not necessary in the MSDLM framework as it was in model 3.2. This MSDLM formulation partitions both the intercept and slope into a large (coarse) scale effect and a small (fine) scale deviation from the large scale effect. For the intercept, the large scale effect is \( \alpha_{1,s} SSN_{s-1} \) and the deviation from this large scale effect is \( \beta_{1,s,t} \). For the slope, the large scale effect is \( \alpha_{2,s} \) and the deviation from this large scale effect is \( \beta_{2,s,t} \). The fine scale intercept is then \( \alpha_{1,s} SSN_{s-1} + \beta_{1,s,t} \), while the fine scale slope is \( \alpha_{2,s} + \beta_{2,s,t} \).

The MSDLM Gibbs sampler described in Algorithm 4 was used to draw samples from the posterior distribution. We assumed \( V_{s,t} = V, \sigma_{\alpha,s} = W_{\alpha}, \) and \( W_{\beta,s,t} = W_{\beta,s} \) and utilized the equations in 3.31 for the sampling of \( \phi \). Prior distributions were specified as \( W_{\alpha}^{-1} \sim \text{Wishart}(4, \mathbf{I}) \), \( W_{\beta,s}^{-1} \sim \text{iid \, Wishart}(4, \mathbf{I}) \), and \( V^{-1} \sim \text{Gamma}(1, 1) \). The MSDLM Gibbs sampler was run for 1000 iterations, after a burn-in of 100 iterations.
Figure 3.19: (top) Posterior means for $\alpha_{1,s}SSN_{s-1}$ (black), posterior means for $\theta_{0,t}$ from model 3.1 (red), and log electron flux observations (gray). (bottom) Posterior means for $\alpha_{1,s}SSN_{s-1} + \beta_{1,s,t}$ (black), posterior means for $\theta_{0,t}$ from model 3.1 (red), and log electron flux observations (gray).

Figure 3.19 shows the posterior means for both the coarse and fine scale intercept compared with the intercept from model 3.1. In the top of Figure 3.19, we see the coarse scale intercept, $\alpha_{1,s}SSN_{s-1}$, captures the large scale trend of the time series. That is, the coarse scale intercept is the mechanism that provides large scale stability across the time series. The MSDLM, through the fine scale states, allows for fine scale deviations from large scale trend. This provides the model a mechanism to locally adapt to the time series. Its effect can be seen in the bottom of Figure 3.19.

Consider a subset of the time series, plotted in Figure 3.20. As previously mentioned and can be seen in the top of Figure 3.20, the coarse scale intercept roughly tracks the large scale
trend of the time series. The fine scale intercept is locally anchored at the coarse scale intercept, but within a coarse scale time unit (i.e., a month), fine scale deviations from the large scale anchoring intercept occur. Again, this facilitates a refinement to the model fit, locally.

Figure 3.20: (top) Posterior means for coarse scale intercept, \( \alpha_{1,s}SSN_{s-1} \) (blue), posterior means for fine scale intercept, \( \alpha_{1,s}SSN_{s-1} + \beta_{1,s,t} \) (red), and log electron flux (gray). (bottom) Posterior means (black) and 95% probability limits (gray band) for the fine scale intercept, \( \alpha_{1,s}SSN_{s-1} + \beta_{1,s,t} \), posterior means for the fine scale intercept of model 3.1, \( \theta_{0,t} \) (red), and log electron flux observations (gray).

The bottom of Figure 3.20 shows the posterior means of the fine scale intercept, 95% probability limits, the posterior means for the fine scale intercept of model 3.1, and log electron flux observations. At the start of each coarse scale time unit, the MSDLM fine scale intercept
is roughly that of the fine scale intercept of model 3.1. However, refinement of fit occurs within coarse scale time units because of the MSDLM’s locally adaptable structure, namely $\beta_{s,t}$. Consider the final four months of the time series of Figure 3.20. We see that the level of log electron flux varies appreciably within a coarse scale time unit. The intercept of model 3.1 is unable to readily adapt to this local change, as the intercept of model 3.1 has to provide both large scale stability and small scale adaptability. In the MSDLM, these two roles are partitioned. The coarse scale state provides large scale stability while the small scale deviations from large scale states allow for small scale adaptability, providing a refined fit to the data.

### 3.8 Discussion

In this paper, we proposed a novel multiscale time series model, the MSDLM. We introduced and justified state estimation procedures that exploit local conditional independencies in the graphical structure of the MSDLM with the MSDLM Kalman filter and smoother. We also introduced the MSDLM FFBS algorithm and showed its utility when embedded within a Gibbs sampler.

The MSDLM extends the classical DLM into the multiscale arena within the directed graphical model formalism. In Appendix I, however, we show that there is a DLM representation for the measurement process of the MSDLM. The result of Appendix I articulates a few things. First, it emphasizes that graphical representations are not unique. That is, though the DLM and MSDLM have different graphical representations, if carefully specified, they can define equivalent likelihoods. Second, knowing that graphical representations are not unique amplifies the usefulness of graphical representations as a model building tool. That is, the graphical representation of the MSDLM intuitively describes the flow of information in a multiscale, temporal system. Describing the same flow of information can be done with the graphical representation of the DLM. However, it is far less intuitive how this is accomplished, as the multiscale features are not described by the graphical representation but rather primarily by the definitions of $G_t$ and $W_t$ in Appendix I.

An artifact of multiscale modeling is a blocking effect present when transitioning across the coarse scale. This blocking effect can be seen in the bottom of Figure 3.20 with the
“discontinuities” in the fine scale intercept, $\alpha_{1,s,SSN_{s-1}} + \beta_{1,s,t}$, as it transitions from time $(s,T)$ to time $(s+1,1)$. When considering multiscale spatial modeling, Huang and Cressie (1997) encountered a similar blocking effect. Their approach to alleviate this effect involved increasing the information flow between scales. For the MSDLM, drawing a directed edge from $\beta_{s,T}$ to $\alpha_{s+1}$ in Figure 3.7 might help lessen the blocking effect while maintaining the necessary directed acyclic graphical structure.

We focused on a multiscale time series model in this paper. Directed graphical models, however, provide a framework to consider multiscale modeling more generally. Applications include multiscale spatial modeling (e.g., Huang and Cressie, 1997) or multiscale spatio-temporal modeling. An appealing aspect of working with directed graphical models is that performing global inference and estimation can be executed by exploiting local conditional independencies. This reduces the dimensionality of the estimation problem, a common difficulty when working with multiscale processes.
CHAPTER 4. FORECASTING SEASONAL INFLUENZA WITH A STATE-SPACE SIR MODEL

4.1 Introduction

Influenza is volatile. Between 1976 and 2006, estimates of annual influenza-related deaths in the United States range from 3,000 to 49,000 people (Centers for Disease Control and Prevention, 2014b). Intervention strategies such as targeted vaccination campaigns (Harris et al., 2010) and public education efforts (Centers for Disease Control and Prevention, 2014a) exist to help mitigate and counteract the potentially severe effects of seasonal influenza. Despite these successful efforts, seasonal influenza persists and poses a serious risk to both national security and public health (Germann et al., 2006).

Disease surveillance systems play an integral role in public health preparedness against seasonal influenza. The Centers for Disease Control and Prevention’s (CDC) Outpatient Influenza-like Illness Surveillance Network (ILINet) collects data pertaining to the state of influenza in the United States. More than 2,900 outpatient healthcare providers from all 50 states, Puerto Rico, the District of Columbia, and the U.S. Virgin Islands participate and monitor influenza-like illness (ILI) (Centers for Disease Control and Prevention, 2015). ILINet is considered the gold standard for ILI surveillance (Nsoesie et al., 2013).

Compartmental models have been used to describe infectious disease transmission since the early 1900s (Ross, 1911). When coupled with disease surveillance data, compartmental models have proven to be invaluable tools for analyzing historical disease transmission dynamics (e.g., Mills et al., 2004), systematically discriminating between the transmission properties of various pathogens (e.g., Yang et al., 2015), and estimating meaningful functions of model parameters (e.g., Heffernan et al., 2005).
Recently, attention has turned from characterizing historical influenza outbreaks to forecasting them. Reliable forecasts with actionable lead times of meaningful outbreak metrics, such as the peak intensity (PI) and timing of the peak intensity (PT), would be valuable to public policy makers. Accurate forecasts of the anticipated overall impact of an outbreak would provide public health practitioners additional information when making decisions about resource allocation, intervention strategy implementation, and timely communications to the public (Nsoesie et al., 2014; Chretien et al., 2014).

Nsoesie et al. (2014) and Chretien et al. (2014) independently reviewed the influenza forecasting literature and noted a variety of forecasting and modeling approaches. Among others, these included statistical approaches (e.g., time series models, generalized linear models, classification models), compartmental modeling approaches (e.g., the Susceptible-Infectious-Recovered (SIR) model, the Susceptible-Exposed-Infectious-Recovered (SEIR) model, the Susceptible-Infectious-Recovered-Susceptible (SIRS) model), and agent-based modeling approaches (e.g., Mniszewski et al., 2008; Epstein, 2009).

Approaches that couple the strengths of compartmental models and statistical models have recently been used for forecasting influenza. They are referred to as data-assimilation or state-space modeling approaches and typically involve embedding a compartmental model into a probabilistic framework. This approach leverages the forecasting power of compartmental models while accounting for multiple sources of uncertainty in a probabilistically consistent manner.

Shaman and Karspeck (2012) demonstrated the promise of such an approach. They used a humidity-driven SIRS compartmental model to describe the disease transmission mechanism, while the ensemble adjustment Kalman filter (Anderson, 2001) was utilized to assimilate ILI data. They showed the potential near real-time value of this approach by forecasting the entirety of the influenza season given only a few weeks of ILI data. Shaman and Karspeck (2012) made a normality assumption for the likelihood, assumed no process error, and used external plug-in estimates for the observational noise. Like Shaman and Karspeck (2012), Dukic et al. (2012) embedded a compartmental model into a state-space modeling framework. Their approach, however, deviated from that of Shaman and Karspeck (2012) in various ways. Firstly, Dukic
et al. (2012) focused on tracking seasonal influenza, rather than sequentially forecasting the whole season. Secondly, Dukic et al. (2012) allowed for process error, which provided model flexibility and hedged against model misspecification. Thirdly, noting that the ILI data and the latent states of the SEIR model are all non-negative, Dukic et al. (2012) worked with the growth rate of the infectious population rather than the ILI data directly. Working with the growth rate, the support of which is the real line, Dukic et al. (2012) assumed normality for both the state and observation equations. Finally, rather than use plug-in estimates, Dukic et al. (2012) treated the observation variance as unknown and estimated it within their sampling procedure. Both Shaman and Karspeck (2012) and Dukic et al. (2012) adopted a simulation-based or Bayesian approach to estimation and specified prior distributions on the components of the parameter vector marginally, constraining the densities to the support of the parameters.

Our work is similar to that of Shaman and Karspeck (2012) and Dukic et al. (2012) in that we also embed a compartmental model into a state-space modeling framework and take a Bayesian approach to inference and forecasting. Unlike the work of Shaman and Karspeck (2012) and Dukic et al. (2012), we make distributional choices for both the state and observation equations that respect support constraints and do not require transformations out of the original scale. Furthermore, we work with the SIR model rather than the SEIR or SIRS models. The deterministic SIR model, though more restrictive than both the SEIR and SIRS models, is also more parsimonious. By embedding the SIR model in a state-space modeling framework and allowing for process and measurement error, our approach retains valuable flexibility for both model fitting and forecasting. Like Shaman and Karspeck (2012), our primary interest lies in forecasting, not tracking. Unlike Shaman and Karspeck (2012) and Dukic et al. (2012), we emphasize the care one should take when specifying a prior distribution for compartmental models, especially for the purposes of forecasting. By using the SIR model, we can leverage and exploit known analytical and readily estimated relationships between the latent parameters of the SIR model and functions of the observable data in the specification of the prior. We specify a prior on the components of the parameter vector hierarchically, not marginally. The hierarchical prior specification is facilitated by augmenting the parameter vector by carefully chosen latent
data, such as PI and PT, in a manner similar to data augmentation ideas popularized by Tanner and Wong (1987).

The layout of this chapter is as follows. In Section 4.2, we describe the data. Sections 4.3 and 4.4 describe the SIR model and state-space modeling approach, respectively. In Section 4.5, we discuss posterior simulation. Section 4.6 describes the prior specification procedure. Section 4.7 presents forecasting results. We finish with a conclusion and discussion of future work with Section 4.8.

### 4.2 Data Description

ILINet provides weekly estimates of the proportion of outpatient health care provider patients with ILI, where ILI is defined as a temperature of 100 degrees Fahrenheit or higher and a cough and/or sore throat without a known cause other than influenza (Centers for Disease Control and Prevention, 2015). Weekly estimates of ILI are available nationally and regionally, where regions correspond to Health and Human Service (HHS) regions defined in Table 4.1.

<table>
<thead>
<tr>
<th>Region</th>
<th>States</th>
</tr>
</thead>
<tbody>
<tr>
<td>HHS1</td>
<td>CT, ME, MA, NH, RI, VT</td>
</tr>
<tr>
<td>HHS2</td>
<td>NJ, NY, Puerto Rico, U.S. Virgin Islands</td>
</tr>
<tr>
<td>HHS3</td>
<td>DE, Washington D.C., MD, PA, VA, WV</td>
</tr>
<tr>
<td>HHS4</td>
<td>AL, FL, GA, KY, MS, NC, SC, TN</td>
</tr>
<tr>
<td>HHS5</td>
<td>IL, IN, MI, MN, OH, WI</td>
</tr>
<tr>
<td>HHS6</td>
<td>AR, LA, NM, OK, TX</td>
</tr>
<tr>
<td>HHS7</td>
<td>IA, KS, MO, NE</td>
</tr>
<tr>
<td>HHS8</td>
<td>CO, MT, ND, SD, UT, WY</td>
</tr>
<tr>
<td>HHS9</td>
<td>AZ, CA, HI, NV</td>
</tr>
<tr>
<td>HHS10</td>
<td>AK, ID, OR, WA</td>
</tr>
</tbody>
</table>

By virtue of its construction, ILI data will include patients with respiratory viruses other than influenza. To better approximate the proportion of ILINet patients with influenza, we follow the approach of Shaman et al. (2013) and couple ILINet data with National Respiratory and Enteric Virus Surveillance System (NREVSS) and US-based World Health Organization (WHO) Collaborating Laboratories data. Specifically, we multiply ILI data by the propor-
tion of ILI samples that tested positive for influenza, as reported by the NREVSS and WHO Collaborating Laboratories. We refer to this metric as ILI+, where ILI+ is a proportion. Furthermore, the NREVSS and WHO Collaborating Laboratories provide the proportion of ILI specimens testing positive for influenza types A and B. We compute ILIA+ and ILIB+ analogously to ILI+. In what follows, we work only with ILI+, ILIA+, and ILIB+. For more details on the ILI adjustment, the reader is directed to the Methods section of Shaman et al. (2013). For more information about the influenza testing procedure performed by the NREVSS and WHO Collaborating Laboratories, the reader is directed to Centers for Disease Control and Prevention (2015).

We consider ten influenza seasons, 2002–2007 and 2010–2013 where, for example, influenza season 2002 means 2002–2003. We omit the H1N1 dominant influenza seasons of 2008 and 2009 and focus only on forecasting seasonal influenza.

In what follows, epidemiology week 40 will be referred to as week 1. This roughly corresponds to the first week of October, a common choice for the start of the influenza season (e.g. Nsoesie et al., 2013). We define the duration of the influenza season to be 35 weeks, roughly corresponding to the end of May. For reference, nationwide ILI+ for 2002–2007 and 2010–2013 is shown in Figure 4.1.

![Figure 4.1: ILI+ for influenza seasons 2002-2007 and 2010-2013. Weeks 1 and 35 roughly correspond to the beginning of October and the end of May, respectively.](image-url)
4.3 Susceptible-Infectious-Recovered (SIR) Model

We consider a basic transmission model for a directly transmitted infectious disease, the SIR model (Kermack and McKendrick, 1927). Consider a closed population of individuals, partitioned into three compartments: susceptible (S), infectious (I), and recovered (R). At any time \( t = 0, 1, \ldots, T \), every individual is a member of exactly one of these three compartments. We denote the proportion of the population in the susceptible, infectious, and recovered compartments by \( S(t) \), \( I(t) \), and \( R(t) \), respectively, such that \( S(t) + I(t) + R(t) = 1 \) for all \( t \). Many assumptions accompany the SIR model, including a closed population (i.e., no new individuals enter or leave the population), the infection has zero latent period (i.e., an individual becomes infectious as soon as they become infected), recovering from the infection confers lifetime immunity, and the population homogeneously mixes. The appeal of using the SIR model lies in its parsimony, intuitive parameter interpretations, and well-understood behaviors.

The SIR model is described by the following set of nonlinear, ordinary differential equations:

\[
\frac{dS}{dt} = -\beta SI, \quad \frac{dI}{dt} = \beta SI - \gamma I, \quad \frac{dR}{dt} = \gamma I,
\]

where \( \beta > 0 \) is the disease transmission rate and \( \gamma > 0 \) is the recovery rate. Conceptually, susceptible individuals become infectious (i.e., move from the susceptible compartment to the infectious compartment), and then ultimately recover from the infection (i.e., move from the infectious compartment to the recovered compartment). The rates at which they move from one compartment to another are governed by the proportion of the population in each of these compartments, as well as the transmission and recovery rates associated with the disease. For illustration, a simulated SIR curve is shown in Figure 4.2 with initial conditions \( S(0) = 0.9 \), \( I(0) = 0.0002 \), \( R(0) = 0.0998 \), and parameters \( \beta = 2 \) and \( \gamma = 1.4 \). Note that an SIR curve is comprised of three trajectories – one for each compartment. ILI+ data provides an estimate of the infectious trajectory. To the best of our knowledge, data related to the susceptible or recovered trajectories either are not publicly available or do not exist.

Evans et al. (2005) showed that in a strict sense, the SIR model with demography is identifiable for all model parameters and initial conditions when incidence data is observed. That is, they showed that a change in one parameter cannot be compensated for by changes in
other parameters. However, Capaldi et al. (2012) pointed out that the strict definition of non-identifiability used by Evans et al. (2005) does not provide insight into the ease of parameter estimation. To illustrate the challenges in estimation articulated by Capaldi et al. (2012), Figure 4.3 plots three distinct SIR curves. These distinct SIR curves have effectively indistinguishable infectious trajectories. Thus, discriminating between these three SIR curves based solely on observations of the infectious curve (e.g., ILI+) would prove challenging.

Figure 4.3: Three SIR curves simulated from three distinct sets of initial conditions and parameters. For all curves, $I(0) = 0.0002$. The black, red, and blue curves were simulated with $(S(0), \beta, \gamma) = (0.9, 2, 1.4), (0.65, 2.38, 1.145)$, and $(0.4, 3.124, 0.842)$, respectively.

Capaldi et al. (2012) noted that estimating SIR model parameters when initial conditions are considered known is easier than estimating SIR model parameters and initial conditions.
simultaneously. Additionally, they emphasized estimating $S(0)$ and $\beta$ simultaneously is particularly challenging, as their estimates are highly correlated.

### 4.4 Dirichlet-Beta State-Space Model (DBSSM)

Various sources of uncertainty are considered when fitting the SIR model to ILI+ data. In addition to the uncertainty in the SIR model parameters and initial conditions, there is uncertainty in the form of the disease transmission mechanism (i.e., process error) as well as the ILI+ data (i.e., measurement error). To account for these various sources of uncertainty, we embed the deterministic SIR model in a state-space modeling framework. This is a flexible approach for capturing changes of the underlying disease transmission dynamics.

Consider two generic random variables $X$ and $Y$. We adopt the notation where brackets around a random variable $X$, $[X]$, represents a short hand notation for the probability density function of $X$. Similarly, $[X|Y]$ corresponds to the conditional distribution of $X$ given $Y$.

The state-space model, henceforth referred to as the Dirichlet-Beta state-space model (DBSSM), is defined as,

$$[y_t|\theta_t, \phi] \overset{d}{=} Beta(\lambda I(t), \lambda(1-I(t))),$$  \hfill (4.2a)  

$$[\theta_t|\theta_{t-1}, \phi] \overset{d}{=} Dirichlet(\kappa \alpha_S(t), \kappa \alpha_I(t), \kappa \alpha_R(t)),$$  \hfill (4.2b)

where $y_t$ is ILI+ data and $\theta_t = (S(t), I(t), R(t))^T$ represents the latent state of the system, all at week $t = 1, 2, \ldots, T$. Additionally, “$\overset{d}{=}$” means “equals in distribution,” $\phi = \{\theta_0, \rho, \beta, \kappa, \lambda\}$, $\kappa > 0$ and $\lambda > 0$ are scalars, and $\rho = \gamma/\beta$. It is worth noting that $\rho$ is the inverse of the basic reproductive number, $R_0$, where $R_0 = \beta/\gamma$. The DBSSM assumes $\theta_{0:T} = (\theta_0, \theta_1, \ldots, \theta_T)$ is a first-order Markov chain and for all $t \neq s$, $y_t$ is independent of $y_s$, given $\theta_t$. Furthermore,

$$\begin{pmatrix}
\alpha_S(t) \\
\alpha_I(t) \\
\alpha_R(t)
\end{pmatrix} = \begin{pmatrix}
S(t-1) + \frac{1}{6}[k_{S1}(t-1) + 2k_{S2}(t-1) + 2k_{S3}(t-1) + k_{S4}(t-1)] \\
I(t-1) + \frac{1}{6}[k_{I1}(t-1) + 2k_{I2}(t-1) + 2k_{I3}(t-1) + k_{I4}(t-1)] \\
R(t-1) + \frac{1}{6}[k_{R1}(t-1) + 2k_{R2}(t-1) + 2k_{R3}(t-1) + k_{R4}(t-1)]
\end{pmatrix}, \hfill (4.3)$$

where
\begin{align*}
k_{S1}(t) &= -\beta S(t)I(t) \\
k_{S2}(t) &= -\beta [S(t) + 0.5k_{S1}(t)][I(t) + 0.5k_{I1}(t)] \\
k_{S3}(t) &= -\beta [S(t) + 0.5k_{S2}(t)][I(t) + 0.5k_{I2}(t)] \\
k_{S4}(t) &= -\beta [S(t) + k_{S3}(t)][I(t) + k_{I3}(t)], \\
k_{I1}(t) &= \beta S(t)I(t) - \gamma I(t) \\
k_{I2}(t) &= \beta [S(t) + 0.5k_{S1}(t)][I(t) + 0.5k_{I1}(t)] - \gamma [I(t) + 0.5k_{I1}(t)] \\
k_{I3}(t) &= \beta [S(t) + 0.5k_{S2}(t)][I(t) + 0.5k_{I2}(t)] - \gamma [I(t) + 0.5k_{I2}(t)] \\
k_{I4}(t) &= \beta [S(t) + k_{S3}(t)][I(t) + k_{I3}(t)] - \gamma [I(t) + k_{I3}(t)], \\
k_{R1}(t) &= \gamma I(t) \\
k_{R2}(t) &= \gamma [I(t) + 0.5k_{I1}(t)] \\
k_{R3}(t) &= \gamma [I(t) + 0.5k_{I2}(t)] \\
k_{R4}(t) &= \gamma [I(t) + k_{I3}(t)].
\end{align*}

Equation 4.3 is the fourth order time step Runge-Kutta (RK4) approximation to the solution of the SIR model of equation 4.1. The RK4 method of equation 4.3 provides a prescription to propagate the SIR model forward in time. Equations 4.4a - 4.4c define the necessary details to execute the RK4 approximation. An alternative method to approximate the solution to the SIR model is Euler’s method (discussed in, for example, Butcher, 2000). Euler’s method, however, is known to be unstable whereas the RK4 method is stable for a wide range of SIR parameterizations and initial conditions while still being easy to implement.

The Beta and Dirichlet distributional choices in equations 4.2a and 4.2b, respectively, obey support constraints. For equation 4.2a, the observation \( y_t \in [0, 1] \). Beta distributions are natural choices for modeling data restricted to the \([0, 1]\) interval. If \( X \sim Beta(\eta_1, \eta_2) \) with
$\eta_1 > 0, \eta_2 > 0$, then

$$E(X) = \frac{\eta_1}{\eta_1 + \eta_2}$$  \hspace{1cm} (4.5a)$$

$$Var(X) = \frac{\eta_1 \eta_2}{(\eta_1 + \eta_2)^2 (\eta_1 + \eta_2 + 1)}.$$  \hspace{1cm} (4.5b)

For the observation equation of the DBSSM (equation 4.2a) with $\eta_1 = \lambda I(t)$ and $\eta_2 = \lambda (1 - I(t))$, we have

$$E(y_t|\theta_t, \phi) = I(t)$$  \hspace{1cm} (4.6a)$$

$$Var(y_t|\theta_t, \phi) = \frac{I(t)(1 - I(t))}{1 + \lambda}.$$  \hspace{1cm} (4.6b)

The conditional expectation of $y_t$ is unbiased for the latent state $I(t)$. The conditional variance of $y_t$ is a function of $I(t)$ and $\lambda$. The parameter $\lambda$ plays a role in controlling the conditional variance, but does not play a role in the conditional expectation. As $\lambda$ tends towards infinity, the conditional variance (i.e., the measurement error) of $y_t$ tends towards zero. Also note that for a fixed $\lambda$, the measurement error increases with increasing $I(t) \in [0, 0.5]$. This feature of measurement error varying with the magnitude of $I(t)$ is expected and attempts to mimic this error structure are present in the literature (e.g., the supporting information of Shaman and Karspeck, 2012). The observation equation of the DBSSM accounts for this measurement error structure naturally, circumventing ad hoc approaches.

For the state equation of the DBSSM (equation 4.2b), all elements of the latent state, $\theta_t$, must be non-negative (i.e., preserve positivity) and sum to unity (i.e., preserve balance). A Dirichlet distribution is a natural choice to model vector-valued data subject to positivity and balance. If $(X_1, X_2, \ldots, X_p) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_p)$ with $\alpha_j > 0$ for all $j = 1, 2, \ldots, p$, then

$$E(X_j) = \frac{\alpha_j}{\sum_{i=1}^{p} \alpha_i}$$  \hspace{1cm} (4.7a)$$

$$Var(X_j) = \frac{\alpha_j (\sum_{i=1}^{p} \alpha_i - \alpha_j)}{(\sum_{i=1}^{p} \alpha_i)^2 (1 + \sum_{i=1}^{p} \alpha_i)}.$$  \hspace{1cm} (4.7b)

An alternative parameterization of the Dirichlet distribution sets $\kappa = \sum_{j=1}^{p} \alpha_j$ and $\alpha_j^* = \kappa^{-1} \alpha_j$, implying $\sum_{j=1}^{p} \alpha_j^* = 1$. Then, for $(X_1, X_2, \ldots, X_p) \sim \text{Dirichlet}(\kappa \alpha_1^*, \kappa \alpha_2^*, \ldots, \kappa \alpha_p^*)$, the
expectation and variance for $X_j$ are expressed as,

$$E(X_j) = \alpha_j^*$$  \hspace{1cm} (4.8a)

$$\text{Var}(X_j) = \frac{\alpha_j^*(1 - \alpha_j^*)}{1 + \kappa}. \hspace{1cm} (4.8b)$$

For the state equation of the DBSSM, we set $\alpha_1^* = \alpha_S(t)$, $\alpha_2^* = \alpha_I(t)$, and $\alpha_3^* = \alpha_R(t)$. Based on equations 4.8a and 4.8b, we then have,

$$E(\theta_t|\theta_{t-1}, \phi) = (\alpha_S(t), \alpha_I(t), \alpha_R(t))'$$ \hspace{1cm} (4.9a)

$$\begin{bmatrix}
\text{Var}(S(t)|\theta_{t-1}, \phi) \\
\text{Var}(I(t)|\theta_{t-1}, \phi) \\
\text{Var}(R(t)|\theta_{t-1}, \phi)
\end{bmatrix} =
\begin{pmatrix}
\frac{\alpha_S(t)(1-\alpha_S(t))}{1+\kappa} \\
\frac{\alpha_I(t)(1-\alpha_I(t))}{1+\kappa} \\
\frac{\alpha_R(t)(1-\alpha_R(t))}{1+\kappa}
\end{pmatrix}. \hspace{1cm} (4.9b)$$

This alternative parameterization is useful for multiple reasons. First, $(\alpha_S(t), \alpha_I(t), \alpha_R(t))'$ in equation 4.3 is assumed to sum to one. It is then more natural to equate, for example, $\alpha_S(t)$ to $\alpha_1^*$ than to $\alpha_1$. Second, by separating the relative proportions of $(\alpha_1, \alpha_2, \alpha_3)$ from the sum of $(\alpha_1, \alpha_2, \alpha_3)$, we were able to isolate $\kappa$. In so doing, we were able to build the SIR model into the conditional mean structure of the process equation while separately controlling the conditional variance (i.e., process error) with the parameter $\kappa$. Like $\lambda$ in equation 4.6b, as $\kappa$ tends towards infinity, the process error becomes null.

### 4.5 Posterior Simulation and Forecasting

We take a Bayesian approach to inference and forecasting within the state-space modeling framework. For $t' \in 1, 2, \ldots, T$, our interest lies in estimating the joint posterior distribution of all latent states and the parameter vector given the observed data,

$$[\theta_1:t', \phi|y_1:t'] \propto [\phi] [y_1:t', \theta_1:t' | \phi] = [\phi] \prod_{t=1}^{t'} [y_t|\theta_t, \phi] [\theta_t|\theta_{t-1}, \phi], \hspace{1cm} (4.10)$$

where $[\phi]$ is the prior distribution for $\phi$. The equality of equation 4.10 holds because of the DBSSM’s conditional independence assumptions.

The posterior distribution of equation 4.10 is unavailable in closed form. Thus, simulation-based methods will be used to approximate 4.10. We use Markov chain Monte Carlo (MCMC)
methods, specifically Gibbs sampling (Geman and Geman, 1984), to simulate from the posterior distribution. This requires sampling iteratively from the full conditional distributions of each latent state and each parameter, conditioned on the most recent draws of all other latent states, parameters, and the observed data. To execute the sampling, we used the \texttt{rjags} package (Plummer, 2014) within the \texttt{R} programming language (R Core Team, 2015), which calls the software “Just Another Gibbs Sampler,” or JAGS (Plummer et al., 2003). JAGS queries a library of internal samplers ranging from efficient but specialized samplers to highly generic, but possibly inefficient base samplers. The base samplers of JAGS utilize slice sampling (e.g., Neal, 2003); the details of which are beyond the scope of this work.

Forecasts for the future observations, $y(t'+1)_T$, are based on the posterior predictive distribution,

$$
[y(t'+1)_T|y_{1:t'}] = \int \int [y(t'+1)_T, \theta_{1:T}, \phi|y_{1:t'}]d\theta_{1:T}d\phi
$$

$$
= \int \int [y(t'+1)_T, \theta(t'+1)_T|\theta_{1:t'}, \phi, y_{1:t'}][\theta_{1:t'}, \phi|y_{1:t'}]d\theta_{1:T}d\phi
$$

$$
= \int \int [y(t'+1)_T, \theta(t'+1)_T|\theta_{t'}, \phi][\theta_{1:t'}, \phi|y_{1:t'}]d\theta_{1:T}d\phi. \tag{4.11}
$$

The last line of equation 4.11 holds because of the assumed conditional independence structure of state-space models. The distribution $[\theta_{1:t'}, \phi|y_{1:t'}]$ of the last line of equation 4.11 is the posterior distribution of equation 4.10, and

$$
[y(t'+1)_T, \theta(t'+1)_T|\theta_{t'}, \phi] = \prod_{t=\tau+1}^{T} [y_t|\theta_t, \phi][\theta_t|\theta_{t-1}, \phi], \tag{4.12}
$$

where $[y_t|\theta_t, \phi]$ and $[\theta_t|\theta_{t-1}, \phi]$ are defined in equations 4.2a and 4.2b, respectively.

We are interested in sampling from the posterior distribution, $[\theta_{1:t'}, \phi|y_{1:t'}]$, and the posterior predictive distribution, $[y(t'+1)_T|y_{1:t'}]$, for $t' = 1, 2, \ldots T$. To sample from all $T$ posterior (predictive) distributions, we adopt a parallelized off-line or batch sampling approach. For state-space models, off-line inference can be computationally inefficient due to redundancies. It is redundant because, for example, the posterior output based on $y_{1:(t'-1)}$ is not used to evaluate the posterior distribution based on $y_{1:t'}$. Sequential Monte Carlo (SMC) methods, such as sequential importance sampling (e.g., Handschin and Mayne, 1969; Handschin, 1970) and particle filtering (e.g., Gordon et al., 1993; Kitagawa, 1996) have been developed to overcome
this redundancy. That is, SMC methods do utilize the posterior output based on $y_{1:(t'-1)}$ to evaluate the posterior distribution based on $y_{1:t'}$. There are known issues with some SMC methods. For example, some sequential importance sampling methods suffer from the problem of “sample impoverishment” or “particle degeneracy” (as discussed in, for example, Stroud et al., 2004; Cappé et al., 2007). However, techniques to overcome this issue (e.g., “sample rejuvenation”) have been presented (e.g., Gordon et al., 1993).

Our decision to use a parallelized off-line sampling approach rather than an SMC approach was motivated by multiple considerations. Firstly, JAGS is coded in C++ and the code runs quickly. Secondly, given the temporal resolution of the currently available ILI+ data, running a sampler in batch mode each time a new observation is observed amounts to running a sampler once per week. This is computationally viable. Thirdly, off-line posterior inference for each $t'$ is an “embarrassingly parallel problem.” That is, performing posterior inference on equation 4.10 and subsequently sampling from the posterior predictive distribution of equation 4.11 for each $t'$ is trivially parallelizable. Speed up in computing time over a serial approach is thus primarily limited by the number of available processing units. For all these reasons, we chose to adopt a parallelized off-line sampling approach. It should be noted, however, that freely available SMC software exists that takes advantage of fast C++ and Perl code for multi-core CPUs and GPUs, such as LibBi (Murray, 2013).

### 4.6 Prior Specification of $\phi$

For the purposes of forecasting, prior specification for compartmental models is challenging and consequential. It is challenging because the parameters of compartmental models have a nonlinear relationship to the data, gleaning intuition in this setting is difficult, and forecasting is necessarily done with incomplete data. It is consequential because, though a solution to a compartmental model may be consistent with observations early in the influenza season, propagating the model forward in time may result in an unreasonable forecast.

To illustrate this, consider the task of forecasting weeks 10 through 35 of the 2010-2011 influenza season, after observing weeks one through nine. The left of Figure 4.4 plots two solutions to the SIR model over the first nine weeks. Both solutions are consistent with these ILI+
observations. However, when the solutions are propagated forward in time, they take divergent trajectories. The forecasted trajectory in the top right of Figure 4.4 appears reasonable, in that it is consistent with historically observed ILI+ observations for weeks 10 through 35. The forecasted trajectory in the bottom right of Figure 4.4, however, appears unreasonable. This forecast suggests a peak intensity (PI) of 40% and the timing of the peak intensity (PT) on week 36. The PI of this forecast is over 15 times larger than the largest PI and occurs 12 weeks later than the latest PT observed between 2002-2007 and 2011-2013. Furthermore, this forecast suggests the duration of the 2010-2011 influenza season will be approximately two years. Historically, the duration of influenza seasons lasts less than a few months.

Figure 4.4: Two solutions to the SIR model are displayed. The green solution (top) is $S(0) = 0.9$, $I(0) = 0.000172$, $R(0) = 0.999828$, $\beta = 2.21$, $\gamma = 1.7017$. The red solution (bottom) is $S(0) = 0.9$, $I(0) = 0.00018$, $R(0) = 0.99982$, $\beta = 0.39$, $\gamma = 0.078$. Points are 2010-2011 ILI+ observations for weeks one through nine. Grey lines are the ILI+ trajectories for years 2002-2007 and 2011-2013 for weeks 10 through 35.
How we inform the DBSSM that the forecast in the top right of Figure 4.4 is reasonable while the forecast in the bottom right is unreasonable is accomplished via the prior specification. The goal of Section 4.6 is to specify a joint prior distribution on \( \phi \) such that the prior predictive distribution is consistent with historically observed ILI+ data, where the prior predictive distribution is,

\[
[y_{1:T}] = \int \int \left( \prod_{t=1}^{T} [y_t|\theta_t, \phi] [\theta_t|\theta_{t-1}, \phi] \right) [\phi] d\theta_{1:T}d\phi. \tag{4.13}
\]

To accomplish this goal, we borrow data augmentation ideas popularized by Tanner and Wong (1987). That is, we augment the parameter vector, \( \phi \), by the quantity \( z = (P_I, P_T) \). Following the terminology of Tanner and Wong (1987), we refer to \( z \) as the latent data. We augment \( \phi \) by \( z \) because specifying a joint prior distribution on \( (\phi, z) \) with desired properties is a simpler task than specifying a prior distribution on \( \phi \) alone with those same properties. Note that by specifying a joint distribution on \( (\phi, z) \), the distribution for \( \phi \) is,

\[
[\phi] = \int [\phi, z]dz. \tag{4.14}
\]

Substituting equation 4.14 in for \([\phi]\) in equation 4.13, the resulting prior predictive distribution is,

\[
[y_{1:T}] = \int \int \left( \prod_{t=1}^{T} [y_t|\theta_t, \phi] [\theta_t|\theta_{t-1}, \phi] \right) \left( \int [\phi, z]dz \right) d\theta_{1:T}d\phi. \tag{4.15}
\]

The joint prior distribution of \( (\phi, z) \) is hierarchically specified where \([\phi, z]\) is factorized into the product of conditional distributions via the product rule. Specifically, we factorize \([\phi, z]\) as follows:

\[
[\phi, z] = [\kappa, \lambda, \theta_0, z, \rho, \beta] = [\kappa] [\lambda|\kappa] [\theta_0|\lambda, \kappa] [z|\theta_0, \lambda, \kappa] [\rho|z, \theta_0, \lambda, \kappa] [\beta|\rho, z, \theta_0, \lambda, \kappa]. \tag{4.16}
\]

Furthermore, we assume

\[
[\lambda|\kappa] = [\lambda], \quad [z|\theta_0, \lambda, \kappa] = [z|\theta_0], \quad [\beta|\rho, z, \theta_0, \lambda, \kappa] = [\beta|\rho, z, \theta_0]. \tag{4.17}
\]

What remains is the specification of \([\kappa]\) and the conditional prior distributions of 4.17.
We illustrate how to specify these conditional prior distributions in practice by using historical nationwide ILI+ data for years 2002-2007 and 2010-2013 to derive hyperparameters. The resulting prior would subsequently be used for fitting and forecasting the 2014-2015 nationwide influenza season. We emphasize, however, that in Section 4.6 we are outlining a prior specification procedure. All hyperparameters derived from historical data are specific to this illustration. To avoid using the data twice, the historical data used to determine hyperparameters should not contain the data the prior will subsequently be used to model.

4.6.1 Specification of $[\kappa]$ and $[\lambda]$

The parameter $\kappa$ relaxes the determinism of the compartmental SIR model by allowing for process error. The parameter $\lambda$ is related to the magnitude of measurement error in the data collection mechanism, conditioned on the latent states. Relatively flat prior distributions with relatively large expected values were assigned to $\kappa$ and $\lambda$. These priors reflect an uncertain belief that the SIR model can adequately describe the infectious trajectory and that conditional on the process, there is a non-negligible but small amount of measurement error. The priors for $\kappa$ and $\lambda$ were specified as,

$$ [\kappa] \overset{d}{=} \text{Gamma} ( \text{shape} = 2, \text{rate} = 0.0001), \quad (4.18) $$

$$ [\lambda] \overset{d}{=} \text{Gamma} ( \text{shape} = 2, \text{rate} = 0.0001), \quad (4.19) $$

with $E(\kappa) = 2 \times 10^4$ and $Var(\kappa) = 2 \times 10^8$.

4.6.2 Specification of $[\theta_0]$

Capaldi et al. (2012) noted that estimating $S(0)$ and $\beta$ simultaneously is challenging due to the high correlation between their estimates. Furthermore, Figure 4.3 suggests that changes to $S(0)$ can be effectively compensated for by changes to other parameters with respect to the infectious curve. For these reasons, we choose to follow the precedent of Dukic et al. (2012) and Nsoesie et al. (2013) and treat $S(0)$ as fixed. In what follows, we assume 90% of the population is initially susceptible to seasonal influenza and assign the following degenerate distribution to
where “the degenerate distribution \( [S(0)] = 0.9 \)” is taken to mean \( S(0) = 0.9 \) with probability one.

ILI+ for \( t = 0, y_0 \), can be thought of as a noisy estimate for \( I(0) \). The vertical dashes of Figure 4.5 denote historical observations of \( y_0 \).

Historically, ILI+ on week 0 has been less than 0.06%. Thus, we expect \( I(0) \) to also fall in this range. We assign a Beta distribution to \( I(0) \), where the hyperparameters are estimated by fitting a Beta distribution to historical \( y_0 \) observations. For the purposes of forecasting the 2014-2015 influenza season, we define the prior on \( I(0) \) as,

\[
[I(0)|S(0)] = [I(0)] \overset{d}{=} \text{Beta}(1.62, 7084.10),
\]

where \( E(I(0)) = 0.00023 \) and \( Var(I(0)) = 0.00018^2 \). The fitted Beta distribution of equation 4.21 is displayed in Figure 4.5.

The prior distribution for \( R(0) \) is specified to preserve latent state balance. Thus, the following degenerate distribution was assigned to \( R(0) \) given \( S(0) \) and \( I(0) \),

\[
[R(0)|S(0), I(0)] = 1 - S(0) - I(0).
\]
Putting together equations 4.20 - 4.22, the joint prior specification for $\theta_0$ is,

$$[\theta_0] = [S(0)][I(0)][R(0)|S(0), I(0)].$$  (4.23)

### 4.6.3 Specification of $[z|\theta_0]$.

In the context of the SIR model, the Epidemic Threshold Theorem (ETT) provides a working definition for an epidemic (Weiss, 2013):

**Theorem 4.1** (Epidemic Threshold Theorem). For an SIR model defined as in equations 4.1,

1. If $S(0) \leq \rho$, then $I(t)$ decreases monotonically to zero as $t \to \infty$. This scenario is not referred to as an epidemic.

2. If $S(0) > \rho$, then $I(t)$ starts increasing, reaches its maximum, and then decreases to zero as $t \to \infty$. This scenario is referred to as an epidemic.

As can be seen in Figure 4.1, ILI+ from 2002-2007 and 2010-2013 exhibits the feature of an epidemic (i.e., $y_t$ increases, reaches a maximum, and then decreases). Given this, it is reasonable to assume this feature will be observed in the 2014-2015 season. Thus, the constraint $S(0) > \rho$ will be enforced in the prior specification of $(\phi, z)$.

Specifying a prior on the latent data, $z = (PI, PT)$, facilitates a prior specification on $\phi$. Figure 4.6 plots the bivariate distribution of historical PI vs. PT. Influenza seasons with peaks early in the season correspond to more intense peaks, while seasons with later peaks are often less intense. Explanations for this relationship have been posited in the literature (Towers et al., 2013).
Figure 4.6: The peak intensity (PI) on the y-axis versus the timing of peak intensity (PT) on the x-axis for years 2002-2007 and 2010-2013 (black points). 10,000 samples were drawn from the truncated normal distribution of equation 4.24 with $I(0) = 0.0002$ and plotted in grey.

A truncated, bivariate normal distribution was fit to the data (black points) in Figure 4.6. The bounds of the truncated normal distribution respect support constraints and enforce prior beliefs. Given that we expect an epidemic to occur, we enforce PI to be greater than or equal to $I(0)$ by setting the lower bound of PI to $I(0)$. PI is a proportion, thus the upper bound was set to 1. The lower and upper bound for PT was set to 1 and 35, respectively, aligning with the support of the influenza forecasting season. The full specification for $[z|θ₀]$ is,

$$[z|θ₀] \overset{d}{=} TN(μ = (0.0144, 17.9), Σ = \begin{pmatrix} 0.000036 & -0.0187 \\ -0.0187 & 16.09 \end{pmatrix}, lower = (I(0), 1), upper = (1, 35)),$$

(4.24)

where $TN(μ, Σ, lower, upper)$ stands for “truncated normal” distribution with mean $μ$, covariance matrix $Σ$, lower bound “lower”, and upper bound “upper.” 10,000 draws from the distribution of equation 4.24 are displayed in Figure 4.6.
4.6.4 Specification of $[\rho | z, \theta_0]$

Augmenting the parameter vector $\phi$ by the latent data $z$ in the prior specification allows us to exploit and leverage relationships between the parameters of the SIR model and functions of the observable ILI+ data. One such analytical relationship between PI and $\phi$ is noted in Section 2.2.7 of Weiss (2013):

$$\text{PI} = g(S(0), I(0), \rho) = I(0) + S(0) - \rho[\log(S(0)) + 1 - \log(\rho)].$$  \hspace{1cm} (4.25)

Equation 4.25 says that the observable quantity PI is a deterministic function of $S(0)$, $I(0)$, and $\rho$. For $S(0) = 0.9$ and $I(0) = 0.0002$, Figure 4.7 displays the relationship between PI and $\rho$. If we restrict our attention to epidemics, PI is a monotonically decreasing function in $\rho$. The grey vertical line of Figure 4.7 represents the range of historical PIs. The grey horizontal line shows the range of corresponding $\rho$ values. Thus, equation 4.25 provides valuable insight into the relationship between the observable quantity PI and the unobservable parameters of the SIR model.

![Figure 4.7: The relationship between PI and $\rho$, for $S(0) = 0.9$ and $I(0) = 0.0002$ (black line). The grey vertical line represents the range of historical PIs. The grey horizontal line shows the range of corresponding $\rho$ values.](image)

If we consider $S(0)$ and $I(0)$ known in equation 4.25, then $\text{PI} = g(\rho|\theta_0)$. For $\rho \in [0, S(0))$, $\rho = g^{-1}(\text{PI}, \theta_0)$ exists and is single-valued because the transformation $g$ is monotone over this
range. The specification for $[\rho|z, \theta_0]$ is then the degenerate distribution,

$$[\rho|z, \theta_0] = g^{-1}(\Pi, \theta_0). \quad (4.26)$$

### 4.6.5 Specification of $[\beta|\rho, z, \theta_0]$

To the best of our knowledge, PT does not correspond to a known analytical function of $\phi$. However, PT is related to the magnitude of $\beta$. Plotted in Figure 4.8 are five infectious curves simulated under the SIR model. For all five curves, $S(0) = 0.9$, $I(0) = 0.0002$, and $\rho = 0.7$. Consequently, $\Pi = 0.024$ for all curves. What varies for each curve is the value of $\beta$, ranging from 0.25 to 4. We see that $\beta$ and PT are inversely related. In Figure 4.8, if we consider week 1 the beginning of October, then the grey vertical lines represent the range of observed PTs between 2002-2007 and 2010-2013. For the five infectious curves plotted in Figure 4.8, only the $\beta = 2$ curve has a PT in the range of historically observed PTs. The $\beta = 1, 0.5,$ and 0.25 curves peak too late (PT too large) and the $\beta = 4$ curve peaks too early (PT too small).

![Figure 4.8](image-url)

**Figure 4.8**: Five infectious curves simulated from the SIR model. For all curves, $S(0) = 0.9$, $I(0) = 0.0002$, and $\rho = 0.7$. The parameter $\beta$ varies from 0.25 to 4. Grey vertical lines denote the range of historically observed PTs.
Estimating the relationship between PT and φ is of interest. For the full factorial design of $S(0) = 0.9$, $I(0) \in \{0.0002, 0.0004\}$, $\rho \in \{0.7, 0.75, 0.8\}$ and $\beta$ equal to 20 evenly spaced values between 1 and 4, we simulated SIR curves (120 in all) and recorded the PT for each curve. Figure 4.9 plots the relationship between $\log(\beta)$ and $\log(PT)$. After accounting for $I(0)$ and $\rho$, we see a highly linear relationship between $\log(\beta)$ and $\log(PT)$.

Figure 4.9: 120 SIR curves were simulated and the PT was recorded (points). For each of the six $I(0)$ and $\rho$ combinations, $\log(\beta)$ was regressed on $\log(PT)$. The expectation lines are overlaid.

Figure 4.9 suggests the relationship between PT and φ can be estimated via regression. To do so, we simulated SIR curves and recorded the PT of each curve for the full factorial design of $S(0) = 0.9$, 10 evenly spaced values of $\rho$ between 0.6 and 0.89, 35 evenly spaced values of $\beta$ between 0.75 and 4.5, and 15 evenly spaced values of $I(0)$ between 0.00001 and 0.001, for a total of 15,750 simulated SIR curves. We then regressed $\log(\beta)$ on a subset of a fourth degree polynomial interaction model using $\log(PT)$, $\log(I(0))$, and $\log(\rho)$ as covariates. Over 99%
of the variability in log(β) was explained by this regression \((R^2 = 0.9929)\) and the estimated MSE, \(\hat{\sigma}^2\), was 0.04207\(^2\). The estimated parameters for this regression, \(\hat{\delta}\), can be found in Table 4.2.

Because of the high estimated \(R^2\) and correspondingly small estimated MSE, we chose to ignore the variability in the estimated relationship between PT and \(\phi\). We specified the degenerate distribution \([\beta|\rho, z, \theta_0]\) as the mean of a log-normal,

\[
[\beta|\rho, z, \theta_0] = \exp(X\hat{\delta} + 0.5\hat{\sigma}^2),
\]

where \(X\) is a \(1 \times 17\) design matrix with columns corresponding to the rows of Table 4.2, \(\hat{\delta}\) is the vector of corresponding parameter estimates found in Table 4.2, and \(\hat{\sigma}^2 = 0.04207^2\).

### Table 4.2: Polynomial regression of log(β) on log(PT), log(I(0)), and log(ρ).

| Covariate | Parameter | Estimate | Std. Error | t value | Pr(>|t|) |
|-----------|-----------|----------|------------|---------|----------|
| Intercept | \(\delta_1\) | -49.7540 | 1.4653 | -33.95 | 0.0000 |
| log(PT)   | \(\delta_2\) | -0.9577 | 0.0072 | -132.17 | 0.0000 |
| log(PT)\(^2\) | \(\delta_3\) | -0.0065 | 0.0013 | -5.05 | 0.0000 |
| log(I(0)) | \(\delta_4\) | -9.4896 | 0.3327 | -28.52 | 0.0000 |
| log(I(0))\(^2\) | \(\delta_5\) | -0.3761 | 0.0184 | -20.43 | 0.0000 |
| log(\rho) | \(\delta_6\) | -590.0001 | 22.718 | -26.49 | 0.0000 |
| log(\rho)\(^2\) | \(\delta_7\) | -2537.6102 | 117.7408 | -21.55 | 0.0000 |
| log(\rho)\(^3\) | \(\delta_8\) | -4756.1828 | 259.2917 | -18.34 | 0.0000 |
| log(\rho)\(^4\) | \(\delta_9\) | -3265.2458 | 202.6393 | -16.11 | 0.0000 |
| log(I(0))\(*\)log(\rho) | \(\delta_{10}\) | -102.2665 | 5.0522 | -20.24 | 0.0000 |
| log(I(0))\(^2\)*log(\rho) | \(\delta_{11}\) | -4.0162 | 0.2793 | -14.38 | 0.0000 |
| log(I(0))*log(\rho)\(^2\) | \(\delta_{12}\) | -430.9596 | 26.7017 | -16.14 | 0.0000 |
| log(I(0))\(^2\)*log(\rho)\(^2\) | \(\delta_{13}\) | -16.7104 | 1.4755 | -11.33 | 0.0000 |
| log(I(0))*log(\rho)\(^3\) | \(\delta_{14}\) | -798.3443 | 58.7927 | -13.58 | 0.0000 |
| log(I(0))\(^2\)*log(\rho)\(^3\) | \(\delta_{15}\) | -30.6638 | 3.2478 | -9.44 | 0.0000 |
| log(I(0))*log(\rho)\(^4\) | \(\delta_{16}\) | -543.8857 | 45.9414 | -11.84 | 0.0000 |
| log(I(0))\(^2\)*log(\rho)\(^4\) | \(\delta_{17}\) | -20.7459 | 2.5373 | -8.18 | 0.0000 |

### 4.6.6 Prior Predictive Distribution for \([y_{1:T}]\)

With \([\phi, z]\) in equation 4.16 specified, we can sample from the prior predictive distribution for \([y_{1:T}]\) (equation 4.15). The prescription to do so is outlined in Algorithm 5.
**Algorithm 5:** Sampling $y_{1:T}$ from $[y_{1:T}]$ (equation 4.15)

1. **For** $m = 1, 2, \ldots, M$:
   2. Draw $\kappa^{(m)} \overset{d}{=} [\kappa]$ (equation 4.18)
   3. Draw $\lambda^{(m)} \overset{d}{=} [\lambda]$ (equation 4.19)
   4. Draw $S(0)^{(m)} \overset{d}{=} [S(0)]$ (equation 4.20)
   5. Draw $I(0)^{(m)} \overset{d}{=} [I(0)]$ (equation 4.21)
   6. Draw $R(0)^{(m)} \overset{d}{=} [R(0)|S(0)^{(m)}, I(0)^{(m)}]$ (equation 4.22)
   7. Set $\theta_0^{(m)} = (S(0)^{(m)}, I(0)^{(m)}, R(0)^{(m)})'$
   8. Draw $z^{(m)} \overset{d}{=} [z|\theta_0^{(m)}]$ (equation 4.24)
   9. Draw $\rho^{(m)} \overset{d}{=} [\rho|z^{(m)}, \theta_0^{(m)}]$ (equation 4.26)
   10. Draw $\beta^{(m)} \overset{d}{=} [\beta|\rho^{(m)}, z^{(m)}, \theta_0^{(m)}]$ (equation 4.27)
   11. Then $\phi^{(m)} = (\lambda^{(m)}, \kappa^{(m)}, \theta_0^{(m)}, \rho^{(m)}, \beta^{(m)})$ constitutes a draw from $[\phi]$  
   12. **For** $t = 1, 2, \ldots, T$:
       13. Draw $\theta_t^{(m)} \overset{d}{=} [\theta_t|\theta_{t-1}^{(m)}, \phi^{(m)}]$ (equation 4.2b)
       14. Draw $y_t^{(m)} \overset{d}{=} [y_t|\theta_t^{(m)}, \phi^{(m)}]$ (equation 4.2a)
       15. Then $y_{1:T}^{(m)} = (y_1^{(m)}, y_2^{(m)}, \ldots, y_T^{(m)})$ constitutes a draw from $[y_{1:T}]$ (equation 4.15)

For illustration, $M = 5000$ samples were drawn from the prior predictive distribution, $[y_{1:T}]$. For each $t = 1, 2, \ldots, T = 35$, we plot the median (black line) and 95% prediction intervals (grey bands), and overlaid the historical ILI+ data (colored lines) in Figure 4.10. The prior predictive distribution is visually consistent with historically observed ILINet data, as desired.
Figure 4.10: The median (black line) and 95% prediction intervals (grey band) based on $M = 5000$ draws from the prior predictive distribution. Historical ILI+ observations are displayed for reference.

The prior predictive distribution is the forecast for the entire upcoming influenza season prior to observing data for the upcoming influenza season. To be clear, Figure 4.10 displays the forecast with associated uncertainty for the 2014-2015 nationwide influenza season, prior to observing ILI+ data for 2014-2015.

4.7 Forecasting Results

4.7.1 Forecasting Illustration

For each $t' \in 1, 2, \ldots, T$, we simulated 62,500 draws from the posterior distribution, $[\theta_{1:t'}, \phi | y_{1:t'}]$, discarding the first 12,500 as burn-in and thinning the remaining 50,000 every tenth iteration. Posterior summaries are thus based on $M = 5000$ draws. Given $M$ draws from the posterior distribution, the prescription for simulating from the posterior predictive distribution, $[y_{(t'+1):T} | y_{1:t'}]$, is outlined in Algorithm 6.
Algorithm 6: Sampling $y_{(t'+1):T}$ from $[y_{(t'+1):T}|y_{1:t'}]$ (equation 4.11)

1. Given $M$ draws from $[\theta_{1:t'}, \phi|y_{1:t'}]$,
2. For $m = 1, 2, \ldots, M$:
3. For $t = t'+1, t'+2, \ldots, T$:
4. Draw $\theta_t^{(m)} \sim [\theta_t|\theta_{t-1}^{(m)}, \phi^{(m)}]$ (equation 4.2b)
5. Draw $y_t^{(m)} \sim [y_t|\theta_t^{(m)}, \phi^{(m)}]$ (equation 4.2a)
6. Then $y_{(t'+1):T}^{(m)} = (y_{t'+1}^{(m)}, y_{t'+2}^{(m)}, \ldots, y_T^{(m)})$ constitutes a draw from $[y_{(t'+1):T}|y_{1:t'}]$ (equation 4.11)

Percentiles of the posterior distribution for the latent infectious states, $[I(0), I(1), \ldots, I(t')|y_{1:t'}]$, and the posterior predictive distribution, $[y_{(t'+1):T}|y_{1:t'}]$, for the 2010-2011 and 2003-2004 nationwide, influenza seasons are shown in Figure 4.11.

For 2010-2011 and $t' = 1$ (top of Figure 4.11), the 95% prediction intervals capture the unknown future observations. There is, however, a large amount of uncertainty in the forecast. This is expected, as only one observation for the 2010-2011 influenza season has been observed. As more data is observed and assimilated into the model, possible forecast trajectories become constrained, resulting in tighter prediction intervals. It is worth noting the DBSSM is able to capture the bimodal feature of the 2010-2011 influenza season in the posterior credible distribution for the latent infectious states, even though the SIR model is used for the process model. This is possible because the DBSSM allows for process error in equation 4.2b. Though the deterministic SIR model may be restrictive for the purposes of modeling these data, embedding it in a state-space framework provides valuable flexibility.

For 2003-2004 and $t' = 1$ (bottom of Figure 4.11), even though the 95% prediction intervals are wide, they still do not capture the unknown future observations. The 2003-2004 influenza season exhibited a peak intensity and timing larger and earlier than all the historical influenza seasons used in the specification of the prior (2002, 2004-2007, 2010-2013). As more data is assimilated into the model though, forecasts improve. By $t' = 4$, the forecasts have adjusted to capture the unknown future observations. This is possible because of the data assimilation
framework of the DBSSM. Though prediction intervals tighten as a result of assimilating data, subsequent forecasts might not capture unknown future observations. For example, the DBSSM is unable to decline fast enough to capture the unknown future observations when $t' = 12$, indicating a systematic bias in the forecast.
Figure 4.11: 95% prediction intervals for $[y(t'+1:T)|y_{1:t'}]$ (red bands) and 95% credible intervals for $[I(0), I(1), \ldots, I(t')|y_{1:t'}]$ (blue bands). Medians (black lines) and ILI+ observations (points) are also displayed for the 2010-2011 (top) and 2003-2004 (bottom) nationwide, influenza seasons. The number at the top of each panel is $t'$. All plots are based on 5000 simulations.
Quantities of interest for forecasting can be computed based on draws from the posterior predictive distribution. Two such quantities are the timing of the peak intensity, $PT_t'$, and the peak intensity, $PI_t'$. For each $t' = 1, 2, \ldots, 34$ and $m = 1, 2, \ldots, 5000$, these quantities are computed as follows:

\[
PI_t^{(m)} = \max \left( y_1, y_2, \ldots, y_{t'}^{(m)}, y_{t'+1}, \ldots, y_{35}^{(m)} \right),
\]

(4.28)

\[
PT_t^{(m)} = \left\{ s \mid y_s^{(m)} = PI_t^{(m)} \right\} \text{ for } s = 1, 2, \ldots, 35.
\]

(4.29)

For $t' = 1, 2, \ldots, 34$, Figure 4.12 plots the median and 95% prediction interval for $[PI_t'|y_1:t']$ and $[PT_t'|y_1:t']$ for the 2010-2011 influenza season. For small $t'$, we see wide prediction interval widths for both $PI_t'$ and $PT_t'$, indicating a large amount of uncertainty in the forecasts when little data has been observed. As more data is observed, forecasts become more confident, as indicated by shrinking prediction interval widths. The ability to forecast with increased confidence is an advantage of data assimilating approaches, such as the DBSSM, over non-data assimilating approaches.
Another quantity of interest for forecasting is the probability the peak intensity has not yet been observed (i.e., \( P(PT > t'|y_{1:t'}) \)). Knowing if the worst of the influenza season has or has not occurred has implications for resource allocation and communication with the public. Thus an estimate of this probability is of practical importance. It is estimated as,

\[
\hat{P}(PT > t'|y_{1:t'}) = \frac{1}{M} \sum_{m=1}^{M} I(PT_{t'}^{(m)} > t'),
\]

(4.30)

where \( PT_{t'}^{(m)} \) is defined in equation 4.29 and \( I() \) is an indicator function that equals 1 if the condition is satisfied and 0 otherwise. For PT > \( t' \), a good forecast would estimate \( P(PT > t'|y_{1:t'}) \) close to one; for PT ≤ \( t' \), it would estimate \( P(PT > t'|y_{1:t'}) \) close to zero. For the 2010-2011 influenza season, Figure 4.13 plots \( \hat{P}(PT > t'|y_{1:t'}) \) for \( t' = 1, 2, \ldots, 34 \).
For $t' \leq 13$, $\hat{P}(PT > t'|y_{1:t'})$ is nearly one, indicating the DBSSM is correctly confident that the PT has not yet been observed. On week 18, the week of the peak intensity, the DBSSM believes there is roughly a 70% chance the peak intensity has not yet been observed, even though it has. One week after the peak, that percentage drops to just over 50%. By week 21, three weeks after the peak, $\hat{P}(PT > t'|y_{1:t'})$ is near zero where it stays for the remainder of the influenza season, indicating the DBSSM is correctly confident that the PT has been observed. For 2010-2011, the DBSSM was able to quickly recognize the peak timing had occurred.

Figure 4.13: $\hat{P}(PT > t'|y_{1:t'})$ versus $t'$ for the 2010-2011 influenza season. The red point indicates the PT. The grey horizontal line represents $I(PT > t')$.

4.7.2 Forecasting Comparison

It is our desire to compare the forecast accuracy of the DBSSM with other competing models. Accurately reproducing these competing models for the purposes of comparison, however, is challenging due to the varied and often times incomplete available model descriptions. Even comparing reported forecast accuracy metrics with competing models can be challenging, as forecasting methods, forecasting outcomes, and reported validation metrics vary widely in the literature (Chretien et al., 2014). For these reasons, we leave the comparison of the DBSSM’s forecasting accuracy with competing models to future work.
Some degree of comparison of forecast accuracy is needed, however. To judge the forecasting quality and justify the complexity of the DBSSM, we compare the DBSSM’s ability to forecast quantities related to PT and PI marginally, as well as jointly, to two straw man models. The first straw man model (SM1) is the prior predictive distribution of the DBSSM. The second straw man model (SM2) is that of Hickmann et al. (2014). For all historic influenza seasons not forecasted, SM2 computes the mean and standard deviation of ILI+ for each time point. A draw from SM2 amounts to independently drawing from a normal distribution at each time point with appropriate mean and standard deviation. Comparing the DBSSM forecasts to the forecasts of SM1 allows us to directly investigate whether or not assimilating data is beneficial to forecasting, relative to our prior. Comparing the DBSSM forecasts to the forecasts of SM2 allows us to directly investigate whether fitting the DBSSM is worth the effort. Note that neither SM1 nor SM2 assimilate data. The forecasts for SM1 and SM2 are produced on the basis of historical influenza seasons, not the current season.

Let PI be the observed peak intensity and PT be the week of the observed peak intensity. Define

\[
\begin{align*}
\text{PI}_{L,t'} &= q_{0.025}\left(\text{PI}_{t'}^{(1)}, \text{PI}_{t'}^{(2)}, \ldots, \text{PI}_{t'}^{(M)}\right), \\
\text{PI}_{U,t'} &= q_{0.975}\left(\text{PI}_{t'}^{(1)}, \text{PI}_{t'}^{(2)}, \ldots, \text{PI}_{t'}^{(M)}\right), \\
\text{PT}_{L,t'} &= q_{0.025}\left(\text{PT}_{t'}^{(1)}, \text{PT}_{t'}^{(2)}, \ldots, \text{PT}_{t'}^{(M)}\right), \\
\text{PT}_{U,t'} &= q_{0.975}\left(\text{PT}_{t'}^{(1)}, \text{PT}_{t'}^{(2)}, \ldots, \text{PT}_{t'}^{(M)}\right),
\end{align*}
\]

where \(q_a(x)\) returns the sample quantile of \(x\) for probability \(a \in [0, 1]\) and \(\text{PI}_{t'}^{(m)}\) and \(\text{PT}_{t'}^{(m)}\) are defined in equations 4.28 and 4.29, respectively.
We consider three metrics:

\[
M_1 = (T^*)^{-1} \sum_{t'=1}^{T^*} \frac{0.05 - \min((\text{PI}_{U,t'} - \text{PI}_{L,t'}), 0.05)}{0.05} I(\text{PI}_{L,t'} \leq \text{PI} \leq \text{PI}_{U,t'})
\]

\[
M_2 = (T^*)^{-1} \sum_{t'=1}^{T^*} \frac{34 - \min((\text{PT}_{U,t'} - \text{PT}_{L,t'}), 34)}{34} I(\text{PT}_{L,t'} \leq \text{PT} \leq \text{PT}_{U,t'})
\]

\[
M_3 = (T^*)^{-1} \sum_{t'=1}^{T^*} \frac{(34 \times 0.05) - \min((\text{PT}_{U,t'} - \text{PT}_{L,t'})(\text{PI}_{U,t'} - \text{PI}_{L,t'}), (34 \times 0.05))}{(34 \times 0.05)} I(\text{PT}_{L,t'} \leq \text{PT} \leq \text{PT}_{U,t'}) I(\text{PI}_{L,t'} \leq \text{PI} \leq \text{PI}_{U,t'})
\]

where \(T^* = \min(\text{PT} + 4, T - 1)\). For example, \(T^* = \min(18 + 4, 35 - 1) = 22\) for the nationwide 2010-2011 influenza season used for illustration in Section 4.7.1. \(M_1, M_2,\) and \(M_3\) sum to \(T^*\) rather than \(T\) because shortly after the PI and PT are observed, the DBSSM is able to confidently and accurately identify them (Figures 4.12 for reference). Summing to \(T\) would give an unfair advantage to the DBSSM over both SM1 and SM2. Also, by summing to \(T^*\), metrics \(M_1, M_2,\) and \(M_3\) focus on a model’s ability to forecast the PI and PT prior to observing them.

\(M_1\) assesses a model’s ability to forecast PI marginally, while \(M_2\) assesses a model’s ability to forecast PT marginally. \(M_3\) assesses a model’s ability to forecast PI and PT jointly. Each metric can be viewed as a score, where \(M_1, M_2,\) and \(M_3 \in [0, 1]\) and 1 is a perfect score. The score itself is a weighted combination of accuracy (coverage) and confidence (prediction interval width). \(M_1, M_2,\) and \(M_3\) are metrics that reward accurate forecasts but not inaccurate forecasts. Confident forecasts are rewarded more than unconfident forecasts, given they are accurate. Small prediction interval widths correspond to confident forecasts. We consider “small” prediction interval widths relative to “uninformatively large” interval widths. We define an interval width of 0.05 to be an uninformatively large interval for PI. Similarly, we define an interval width of 34 weeks to be an uninformatively large interval for PT. Consider \(t' = 1\) in Figure 4.12. The forecast for both PI and PT are accurate, as the observed PI and PT are contained in the 95% prediction intervals. They are, however, unconfident forecasts as the prediction intervals widths are large.

We consider 330 scenarios comprised of all year–region–type combinations of ten years (2002-2007 and 2010-2013), eleven geographic regions (HHS1 through HHS10 and nationwide),
and three sources of data corresponding to different influenza types (ILI+, ILIA+, and ILIB+). For each scenario, we compute M1, M2, and M3 for the DBSSM, SM1, and SM2. For each scenario and metric, we determine the best model, where best means the model with the corresponding metric closest to 1. The results are summarized in Table 4.3.

Table 4.3: Percentage of scenarios each model was deemed best, by metric. Bold numbers indicate the largest percentage for each metric.

<table>
<thead>
<tr>
<th>Model</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBSSM</td>
<td>60.3</td>
<td>38.8</td>
<td>54.8</td>
</tr>
<tr>
<td>SM1</td>
<td>5.5</td>
<td>0.0</td>
<td>6.1</td>
</tr>
<tr>
<td>SM2</td>
<td>34.2</td>
<td>61.2</td>
<td>39.1</td>
</tr>
</tbody>
</table>

The DBSSM is the preferred model for forecasting PI marginally and jointly with PT, as it was the best model in the highest percentage of scenarios for M1 and M3, respectively. SM2, however, is the preferred model with respect to forecasting PT marginally, as it was the best model in the highest percentage of scenarios for M2. SM1 is the least preferred model with respect to all considered metrics. Thus, based on the results of Table 4.3, we have strong evidence it is beneficial to assimilate data, relative to the prior of the DBSSM.

It would appear the DBSSM and SM2 are competitive with each other but good at different things. Examination of the distribution of M1, M2, and M3 scores, however, provide more insight into the differences between the DBSSM and SM2. The distribution of scores across all scenarios are displayed in Figure 4.14.
Figure 4.14: Distribution of M1, M2, and M3 scores for the DBSSM (top) and SM2 (bottom).

For SM2, we see a spike at zero and a clustering of scores generally above 0.5 for all metrics. Because SM2 does not assimilate data, the forecast made for every $t'$ is identical. This means the forecast for each scenario is either accurate or inaccurate, with inaccurate forecasts assigned a score of zero. The spike at zero in Figure 4.14 for SM2 is thus the marking of a non-data assimilating model.

The relatively high volume of zeros coupled with relatively high scores, however, is the marking of an over-confident, non-data assimilating model. Small prediction interval widths (i.e., confident forecasts), have the potential to yield high scores, but carry an increased likelihood of inaccuracy (i.e., scores of zero). For SM2, roughly 50%, 30%, and 60% percent of all scenarios were inaccurate for metrics M1, M2, and M3, respectively. Additionally, SM2’s non-zero scores were generally above 0.5, and were generally above 0.75 for M3.

Assimilating data hedges against the dichotomy of accuracy and inaccuracy. As can be seen, the DBSSM seldom has a score near zero, but also seldom has a score near one.
Finally, for each model and scenario, we estimate $P(PT > t'|y_{1:t'})$ as described in equation 4.30. Figure 4.15 displays $\hat{P}(PT > t'|y_{1:t'})$ averaged over all scenarios versus standardized time, where standardized time is $t' - PT$. Standardized time equal to zero corresponds to the week of the PT, negative values correspond to the number of weeks prior to the PT, and positive values correspond to the number of weeks after the PT. An ideal forecast corresponds to the grey lines of Figure 4.15.

Figure 4.15: $\hat{P}(PT > t'|y_{1:t'})$ averaged over all scenarios versus standardized time. The grey horizontal line represents $I(PT > t') = I(0 > \text{Standardized Time})$.

Figure 4.15 shows the superiority of the DBSSM over both SM1 and SM2 with respect to estimating $P(PT > t'|y_{1:t'})$. The DBSSM can correctly and confidently predict the peak intensity has not yet been observed prior to observing it. The same cannot be said for either SM1 or SM2. The DBSSM has difficulty recognizing the peak intensity has been observed on the week the peak intensity is observed. On average, for standardized time equal to zero, the DBSSM believes there is an 80% chance the peak intensity has not yet been observed, though it just was. However, the DBSSM is able to quickly recognize the PI has been observed one to two weeks after observing it. By three weeks after observing the peak intensity, the DBSSM is able to correctly and confidently predict the peak intensity has occurred. Knowing that the
worst of the influenza season has been observed shortly after observing it has practical value and is a strength of the DBSSM.

4.8 Discussion

In this chapter, we embed the SIR model in a state-space modeling framework. The deterministic SIR model inherits valuable flexibility via this embedding. The Beta and Dirichlet distributions of equations 4.2a and 4.2b, respectively, naturally obey support constraints without the need for transforming the data. Parallel sampling with JAGS proved computationally feasible, even for the forecasting analysis with 330 scenarios.

A major contribution of this work was the prior specification of Section 4.6. By augmenting $\phi$ by the latent data $z$ and hierarchically specifying the prior, we were able to exploit both known analytical and estimated relationships between $\phi$ and $z$. Incorporating prior information into the prior specification is a crucial component to forecasting seasonal influenza. Forecasts are most desired and valuable when data for the current influenza season are scarce. The prior distribution is the mechanism by which to incorporate information from historical influenza seasons.

The DBSSM is applicable to all seasons of influenza. The prior specification of Section 4.6 is, however, only applicable to seasonal influenza. Because the DBSSM is a data-assimilating model, we anticipate it would be able to adapt to and track non-seasonal influenza, such as the 2009-2010 H1N1 season. However, the forecasts produced by this model are likely to be suspect as the features of non-seasonal influenza years will deviate from the features built into the prior.

Forecasting seasonal influenza is in its infancy and numerous areas of future research exist. In this work, we used CDC ILI+ data exclusively. Though ILINet is considered the gold standard for ILI surveillance, it does have limitations. ILI+ data is publicly released on a one to two week delay and is subject to retrospective revisions, drawing into question its value for real-time forecasting (Nsoesie et al., 2013). Additionally, the most granular level of geographic resolution for ILI+ data is at the HHS region level. Non-negligible limitations of ILI+ data
must be overcome to someday produce reliable forecasts at the municipal level in real or near real-time.

Alternative disease surveillance systems exist and could be used to potentially alleviate the limitations of ILI+ data. Two alternative sources for surveillance system data include Google Flu Trends (Ginsberg et al., 2009) and Wikipedia (Generous et al., 2014). Both surveillance systems are available in near real-time. Google Flu Trends is available at the municipal level for over 100 municipalities, while it is hoped municipal level data will be possible with Wikipedia. These alternative surveillance systems are not without their own limitations, however. Both Google Flu Trends and Wikipedia serve as proxies for ILINet’s ILI data. That is, neither attempts to measure the influenza transmission mechanism. Rather, they both identify search queries that are correlated with ILINet’s ILI data. Furthermore, it is well documented that Google Flu Trends has a tendency to overestimate influenza activity (Lazer et al., 2014). Rather than select one disease surveillance system for influenza forecasting, multiple disease surveillance systems could be incorporated into a principled, probabilistic, data-assimilating model. This approach has the potential to leverage the accuracy of traditional surveillance systems and the timeliness and geographic resolution of alternative surveillance systems.

The SIR model was selected to exploit its well-understood structure in the prior specification. Other compartmental models, however, could have been selected. Embedding other compartmental models into the DBSSM should be straight-forward. How to related observable quantities of data to the parameter vector of other compartmental models for the purposes of forecasting is, however, an open question.

Finally, as more forecasting models are developed, the need for standard and meaningful forecasting metrics along with approaches to compare competing models will increase. Nsoesie et al. (2014) and Chretien et al. (2014) independently reviewed the current state of seasonal influenza forecasting. Some of their general conclusions and recommendations included the importance for studies to clearly define the predicted event, to clearly define accuracy measures for the predicted event, the need for best practices in forecasting studies, and the need to report head-to-head comparisons between competing forecasting approaches. We agree with these general conclusions and recommendations. When comparing forecasting models, the
determination of the “best” model will necessarily be relative to a forecasting metric or multiple forecasting metrics. Selecting meaningful metrics should not be taken lightly. Furthermore, Figure 4.14 suggests comparisons between forecasting methods should be broadly demonstrated, not just illustrated. This is because undesirable features of a forecasting approach (e.g., over-confidence) will not be apparent when applied to a single or even a few scenarios. General features of a forecasting model only become apparent when the model is applied to numerous scenarios.
CHAPTER 5. GENERAL CONCLUSION

5.1 General Discussion

State-space models, or more generally directed graphical models, provide a powerful and flexible framework for modeling dynamic data. In Chapter 2, we compared static modeling approaches to a dynamic modeling approach – the DLM. Advantages of utilizing this relatively straightforward linear, Gaussian state-space model for describing the relationship between relativistic electron flux and solar wind speed in the Van Allen radiation belts were presented and discussed. Chapter 3 introduced a multiscale time series model, the MSDLM, and presented it within the directed graphical model formalism. The MSDLM is a multiscale extension to the DLM. By presenting the DLM and the MSDLM in concert, intuitive connections between the two models were drawn. State estimation in the form of the MSDLM filtering, smoothing, and forecasting recursions were presented and justified within the directed graphical model formalism. In Chapter 4, forecasting influenza in the United States was investigated. A non-linear, non-Gaussian state-space model, the DBSSM, was developed by embedding a set of nonlinear, ordinary differential equations in the state equations. The DBSSM obeyed support constraints for the latent states and observations, avoiding transformations out of the original scale. An informative, hierarchical prior specification was presented for the DBSSM leveraging data-augmentation techniques. Forecasting metrics and comparisons to competing models were also presented.

5.2 Directions for Future Work

Numerous avenues for future work presented themselves throughout the writing of this dissertation. Some of those avenues are discussed below.
1. In Chapter 2, we illustrated how a dynamic modeling framework could improve forecasting. We demonstrated this by comparing the relatively simple DLM with one covariate to the REFM. Results were encouraging. The field of space weather forecasting is constantly interested in improving their predictive capabilities. Considering alternative and/or multiple covariates in the dynamic modeling framework is the next logical step. In addition to comparing various DLMs to the REFM, we could also compare the DLM to the MSDLM with respect to forecast accuracy.

2. A blocking effect was observed across coarse scale time units in the smoothing and filtering equations of the MSDLM, as presented in Chapter 3. This blocking effect might be minimized by including a directed edge from $\beta_{s,T}$ to $\alpha_{s+1}$ in Figure 3.7. Were these directed edges added, the MSDLM filtering, smoothing, and forecasting equations would have to be readdressed, as many of the conditional independencies would no longer hold under this updated directed graphical model.

3. To improve seasonal influenza forecasting capabilities, we must make use of the best available information. This will involve traditional and novel sources of spatially and temporally varying surveillance data. How to combine these varied and relevant sources of data into a principled model is an open question. Considering this problem within the directed graphical model formalism may be a promising approach. For example, a possible graphical representation for combining ILI+ and Google Flu Trends (GFT) is shown in Figure 5.1. The latent state of the disease at time $t$ is represented by $\theta_t$. ILI+$t$ and GFT$_t$ are conditionally independent given $\theta_t$. Because the graphical representation of Figure 5.1 is a DAG, the model defined on it would inherit all the conditional independence properties of directed graphical models discussed in Chapter 3. The directed graphical model formalism could also be used to model disease transmission systems that incorporate spatially varying information.
Figure 5.1: A possible DAG representation that incorporates two sources of surveillance data.

4. In Chapter 4, we chose to use the SIR compartmental model for its exploitable properties and its parsimony. Other compartmental models could have been used. Determining which compartmental model is most appropriate for modeling seasonal influenza and how to discriminate between them are open questions.
APPENDIX A. MULTICOLLINEARITY ILLUSTRATION

Let $X_{1,i}$ and $X_{2,i}$ represent two covariates and let $Y_i$ represent a response variable for unit $i$, $i = 1, 2, \ldots, N$. Consider two situations – one where $X_{1,i}$ and $X_{2,i}$ are uncorrelated (correlation of 0) and one where they are highly-linearly correlated (correlation of 0.99). For each situation, consider two data-generating models:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \epsilon_i$$  \hspace{1cm} \text{(Model S1)}

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \epsilon_i,$$  \hspace{1cm} \text{(Model S2)}

where $\epsilon_i \sim N(0, 1)$, independent and identically distributed, $\beta_0 = -0.5$, $\beta_1 = 1$, and $\beta_2 = 1$.

For data-generating Model S1 and Model S2, three models will be fit:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \epsilon_i$$  \hspace{1cm} \text{(Model A)}

$$Y_i = \beta_0 + \beta_2 X_{2,i} + \epsilon_i$$  \hspace{1cm} \text{(Model B)}

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \epsilon_i,$$  \hspace{1cm} \text{(Model C)}

where $\epsilon_i \sim N(0, \sigma^2)$, independent and identically distributed.

When Model S1 is the true model, Model A is the correctly specified model, Model B excludes the relevant covariate $X_1$ and includes an irrelevant covariate $X_2$, and Model C includes the relevant covariate $X_1$ but also includes an irrelevant covariate $X_2$. When Model S2 is the true model, both Model A and Model B exclude a relevant covariate, while Model C is the correctly specified model.

We simulate 1000 data sets of $N = 300$ under Model S1 and Model S2, for the uncorrelated and highly correlated covariate scenarios. Model A, Model B, and Model C are fit to each of
the simulated data sets. Histograms of parameter estimates for Model S1 and Model S2 are shown in Figures A.1 and A.2, respectively.

![Histograms of parameter estimates from Model S1 simulated data sets](image)

Figure A.1: Histograms of parameter estimates from Model S1 simulated data sets. The top row corresponds to the uncorrelated covariate scenario; the bottom row to the highly-linearly correlated covariate scenario. The columns correspond to the model that was fit (A, B, or C) and the estimated parameter (Beta 1 or 2). Red vertical lines indicate the true value of the parameter.

Figures A.1 and A.2 show that in the absence of multicollinearity, OLS parameter estimates are unbiased with relatively small standard errors (relative to the multicollinearity case). Figure A.1 shows that in the presence of multicollinearity, when Model B is fit to data generated from Model S1, \( \hat{\beta}_2 \) is centered near 1 not 0, the true value of \( \beta_2 \). Thus, \( \hat{\beta}_2 \) is a biased estimate. When
Model C is fit to data generated under Model S1, estimates for $\hat{\beta}_1$ and $\hat{\beta}_2$ are unbiased, but their standard errors are quite large relative to their uncorrelated counterpart. For example, 43 of the 1000 estimated $\beta_1$ parameters were less than 0.9 when the covariates are uncorrelated, but 410 of the 1000 estimated $\beta_1$ parameters were less than 0.9 when the covariates were highly correlated. Even though parameter estimates are unbiased when an irrelevant, but highly correlated covariate is added to the correct model, the parameter estimates for any one data set can deviate wildly from the true value.
Figure A.2: Histograms of parameter estimates from Model S2 simulated data sets. The top row corresponds to the uncorrelated covariate scenario; the bottom row to the highly-linearly correlated covariate scenario. The columns correspond to the model that was fit (A, B, or C) and the estimated parameter (Beta 1 or 2). Red vertical lines indicate the true value of the parameter.

Like Figure A.1, Figure A.2 shows that when a relevant covariate is excluded from the model, the resulting parameter estimate is biased in the presence of multicollinearity. When the true model is fit (Model C), the resulting parameter estimates are unbiased with large accompanying standard errors.
APPENDIX B. FILTERING, FORECASTING, AND SMOOTHING DETAILS

Consider the model described in (2.1a) - (2.1c). In the presentation of the filtering, forecasting, and smoothing equations that follow, it is assumed that $M_t$, $G_t$, $E_t$, $W_t$, $a_0$, and $C_0$ are known for all $t$, $t = 1, 2, \ldots, T$.

Filtering

Filtering refers to computing the conditional distribution of $\beta_t|F_{1:t}$ – the distribution of the state at time $t$ given all the data up to time $t$. Assume the filtering distribution of $\beta$ at time $t - 1$,

$$\beta_{t-1}|F_{1:t-1} \sim N(a_{t-1}, C_{t-1}),$$

is available. Notice that for $\beta_0$, this is true – equation (2.1c). Then, the filtering distribution for $\beta$ at time $t$ can be computed recursively. That is,

$$\beta_t|F_{1:t} \sim N(a_t, C_t),$$

where

$$a_t = b_t + B_t M_t' D_t^{-1} (F_t - d_t) \quad (B.1a)$$

$$C_t = B_t - B_t M_t' D_t^{-1} M_t B_t' \quad (B.1b)$$

and

$$d_t = M_t b_t \quad (B.2a)$$

$$D_t = M_t B_t M_t' + E_t \quad (B.2b)$$

$$b_t = G_t a_{t-1} \quad (B.2c)$$

$$B_t = G_t C_{t-1} G_t' + W_t. \quad (B.2d)$$
Forecasting

For some $k \geq 1$, forecasting refers to computing the conditional distribution of $F_{t+k|F_{1:t}}$ - the distribution of the observation at time $t + k$ given all the data up to time $t$. Assume the filtering distribution,

$$\beta_t|F_{1:t} \sim N(a_t, C_t),$$

is available. Then, the forecasting distribution for $F_{t+k}$ is

$$F_{t+k|F_{1:t}} \sim N(d_t(k), D_t(k)),$$

where

$$d_t(k) = M_{t+k}b_t(k) \quad (B.3a)$$
$$D_t(k) = M_{t+k}B_t(k)M_{t+k}' + E_{t+k} \quad (B.3b)$$

and

$$b_t(k) = G_{t+k}b_{t,k-1} \quad (B.4a)$$
$$B_t(k) = G_{t+k}B_{t,k-1}G_{t+k}' + W_{t+k} \quad (B.4b)$$
$$\beta_{t+k|F_{1:t}} \sim N(b_t(k), B_t(k)). \quad (B.4c)$$

To be clear, the point estimate forecast for $F_{t+k}$ is $d_t(k)$. Additionally, the forecasting distribution provides bounds of uncertainty about that point estimate. For example, a 95% prediction interval for $F_{t+k}$ is $d_t(k) \pm 1.96D_t(k)^{0.5}$.

Smoothing

Smoothing refers to the conditional distribution of $\beta_t|F_{1:T}$ - the distribution of the state at time $t$ given all the available data. Smoothing is a retrospective analysis and is of interest when one wants to retrospectively study the behavior of a system. Assume the smoothing distribution

$$\beta_{t+1|F_{1:T}} \sim N(s_{t+1}, S_{t+1})$$
is available for \( t < T \). Note that \( \beta_T | F_{1:T} \) is available, as it is the filtering distribution for \( \beta_T \).

Then, the smoothing distribution for \( \beta_t \) can be computed recursively. That is,

\[
\beta_t | F_{1:T} \sim N(s_t, S_t),
\]

where

\[
s_t = a_t + C_t G_{t+1}^{-1} B_{t+1}^{-1} (s_{t+1} - b_t) \tag{B.5a}
\]

\[
S_t = C_t - C_t G_{t+1}^{-1} B_{t+1}^{-1} (B_{t+1} - S_{t+1}) B_{t+1}^{-1} G_{t+1} C_t. \tag{B.5b}
\]
APPENDIX C. PROOF OF THE DYNAMIC LINEAR MODEL FILTERING EQUATIONS (PROPOSITION 3.2)

A result used in the proof of the filtering and smoothing equations for both the DLM and MSDLM:

**Result 1.** If \( y|\beta, X, V \sim N(X\beta, V) \) and \( \beta \sim N(m_0, C_0) \), then \( \beta|y, X, V \sim N(m_p, C_p) \), where

\[
m_p = m_0 + C_0X'(XC_0X' + V)^{-1}(y - Xm_0) \quad \text{and} \quad C_p = C_0 - C_0X'(XC_0X' + V)^{-1}XC_0.
\]

Some useful distributional equivalencies for the DLM Kalman filter are

\[
[y_t, \theta_t|\theta_{t-1}, y_{1:t-1}] = [y_t, \theta_t|\theta_{t-1}], \quad (C.1a)
\]

\[
[y_t|\theta_t, y_{1:t-1}] = [y_t|\theta_t]. \quad (C.1b)
\]

The distributional equivalencies of equations C.1 can be verified by applying Proposition 3.1 to the appropriate moral graphs of Figure 3.5. The moral graph corresponding to the DLM is shown in Figure C.1.

![Moralized graph of Figure 3.5](image-url)
Proof of the one-step-ahead predictive distribution of $\theta_t$ given $y_{1:t-1}$ (3.11)

\[ a_t = E(\theta_t|y_{1:t-1}) = E(E(\theta_t|\theta_{t-1}, y_{1:t-1})|y_{1:t-1}) \stackrel{3.1a}{=} E(E(\theta_t|\theta_{t-1})|y_{1:t-1}) \]
\[ \stackrel{3.5b}{=} E(G_t\theta_{t-1}|y_{1:t-1}) = G_t m_{t-1}, \]  
(C.2)

\[ R_t = Var(\theta_t|y_{1:t-1}) = E(Var(\theta_t|\theta_{t-1}, y_{1:t-1})|y_{1:t-1}) + Var(E(\theta_t|\theta_{t-1}, y_{1:t-1})|y_{1:t-1}) \]
\[ \stackrel{C.1a}{=} E(Var(\theta_t|\theta_{t-1})|y_{1:t-1}) + Var(E(\theta_t|\theta_{t-1})|y_{1:t-1}) \]
\[ \stackrel{3.5b}{=} E(W_t|y_{1:t-1}) + Var(G_t\theta_{t-1}|y_{1:t-1}) \]
\[ = G_tC_{t-1}G_t' + W_t. \]  
(C.3)

Proof of the one-step-ahead forecasting distribution of $y_t$ given $y_{1:t-1}$ (3.12)

\[ f_t = E(y_t|y_{1:t-1}) = E(E(y_t|\theta_t, y_{1:t-1})|y_{1:t-1}) \stackrel{C.1b}{=} E(E(y_t|\theta_t)|y_{1:t-1}) \]
\[ \stackrel{3.5a}{=} E(F_t\theta_t|y_{1:t-1}) = F_ta_t, \]  
(C.4)

\[ Q_t = Var(y_t|y_{1:t-1}) = E(Var(y_t|\theta_t, y_{1:t-1})|y_{1:t-1}) + Var(E(y_t|\theta_t, y_{1:t-1})|y_{1:t-1}) \]
\[ \stackrel{C.1b}{=} E(Var(y_t|\theta_t)|y_{1:t-1}) + Var(E(y_t|\theta_t)|y_{1:t-1}) \]
\[ \stackrel{3.5a}{=} E(V_t|y_{1:t-1}) + Var(F_t\theta_t|y_{1:t-1}) \]
\[ = F_tR_tF_t' + V_t. \]  
(C.5)

Proof of the filtering distribution of $\theta_t$ given $y_{1:t}$ (3.13)

Note that $[y_t, \theta_t|y_{1:t-1}] = [y_t|\theta_t, y_{1:t-1}] [\theta_t|y_{1:t-1}]$, where

\[ [y_t|\theta_t, y_{1:t-1}] \stackrel{C.1b}{=} [y_t|\theta_t] \overset{d}{=} N(F_t\theta_t, V_t), \]
\[ [\theta_t|y_{1:t-1}] \overset{d}{=} N(a_t, R_t). \]

By Result 1, $\theta_t|y_{1:t} \sim N(m_t, C_t)$, where

\[ m_t = a_t + R_tF_t'Q_t^{-1}(Y_t - f_t), \]
\[ C_t = R_t - R_tF_t'Q_t^{-1}F_tR_t. \]
APPENDIX D. PROOF OF THE MULTISCALE DYNAMIC LINEAR MODEL FILTERING EQUATIONS (PROPOSITION 3.3)

Some useful distributional equivalencies for the MSDLM Kalman filter are

\[
\begin{align*}
[\alpha_s, \beta_{s,1}|\alpha_{s-1}, y_{s-1,T}^*] &= [\alpha_s, \beta_{s,1}|\alpha_{s-1}], \\
[\alpha_s, \beta_{s,t}|\alpha_s, \beta_{s,t-1}, y_{s,t-1}^*] &= [\alpha_s, \beta_{s,t}|\alpha_s, \beta_{s,t-1}], \\
[y_{s,t}|\alpha_s, \beta_{s,t}, y_{s,t-1}^*] &= [y_{s,t}|\alpha_s, \beta_{s,t}].
\end{align*}
\]

The distributional equivalencies of equations D.1 can be verified by applying Proposition 3.1 to the appropriate moral graphs of Figure 3.7. The moral graph corresponding to the MSDLM is shown in Figure D.1.

\[
\begin{array}{c}
\text{Figure D.1: The moralized graph of Figure 3.7}
\end{array}
\]

Proof of the one-step-ahead predictive distribution (3.14 and 3.15)

(i) If \( t = 1 \), \((\alpha_s, \beta_{s,1})'\) given \( y_{s-1,T}^* \) is \( N(b_{s,1}, B_{s,1}) \), where
Let \( b_{s,1} = (b_{\alpha,s,1}, b_{\beta,s,1})' = E((\alpha_s, \beta_s,1)'|y_{s-1,T}^*) \)

\[
D_{1a} = E(E((\alpha_s, \beta_s,1)'|\alpha_{s-1})|y_{s-1,T}^*) = (G_{\alpha,s}a_{\alpha,s-1,T}, 0)',
\]

\[
B_{s,1} = \begin{bmatrix}
B_{\alpha,s,1} & B_{\alpha\beta,s,1} \\
B'_{\alpha\beta,s,1} & B_{\beta,s,1}
\end{bmatrix} = \text{Var}((\alpha_s, \beta_s,1)'|y_{s-1,T}^*) = D_{1a} + E(\text{Var}((\alpha_s, \beta_s,1)'|\alpha_{s-1})|y_{s-1,T}^*) + \text{Var}(E((\alpha_s, \beta_s,1)'|\alpha_{s-1})|y_{s-1,T}^*)
\]

\[
= \begin{bmatrix}
G_{\alpha,s}A_{\alpha,s-1,T}G'_{\alpha,s} + W_{\alpha,s} & 0 \\
0 & W_{\beta,s,1}
\end{bmatrix}.
\]

(ii) If \( t \neq 1 \), \((\alpha_s, \beta_s,t)' given y_{s,t-1}^* is N(b_{s,t}, B_{s,t}) \), where

\[
b_{s,t} = (b_{\alpha,s,t}, b_{\beta,s,t})' = E((\alpha_s, \beta_s,t)'|y_{s,t-1}^*) \]

\[
D_{1b} = E(E((\alpha_s, \beta_s,t)'|\alpha_{s-1}, \beta_{s-1})|y_{s,t-1}^*) = (a_{\alpha,s,t-1}, G_{\beta,s,t}a_{\beta,s,t-1})'
\]

\[
B_{s,t} = \begin{bmatrix}
B_{\alpha,s,t} & B_{\alpha\beta,s,t} \\
B'_{\alpha\beta,s,t} & B_{\beta,s,t}
\end{bmatrix} = \text{Var}((\alpha_s, \beta_s,t)'|y_{s,t-1}^*) = D_{1b} + E(\text{Var}((\alpha_s, \beta_s,t)'|\alpha_{s-1}, \beta_{s-1})|y_{s,t-1}^*) + \text{Var}(E((\alpha_s, \beta_s,t)'|\alpha_{s-1}, \beta_{s-1})|y_{s,t-1}^*)
\]

\[
= \begin{bmatrix}
A_{\alpha,s,t-1} & A_{\alpha\beta,s,t-1}G'_{\beta,s,t} \\
G_{\beta,s,t}A'_{\alpha\beta,s,t-1} & G_{\beta,s,t}A_{\beta,s,t-1}G'_{\beta,s,t} + W_{\beta,s,t}
\end{bmatrix}.
\]

Proof of the one-step-ahead forecasting distribution (3.16)

Let \( F_{s,t} \equiv [F_{\alpha,s,t}, F_{\beta,s,t}] \). Then the one-step-ahead forecast distribution for \( y_{s,t} \) given \( y_{s,t-1}^* \) is \( N(d_{s,t}, D_{s,t}) \), with

\[
d_{s,t} = E(y_{s,t}|y_{s,t-1}^*) \]

\[
D_{1c} = E(E(y_{s,t}|\alpha_{s,t}, \beta_{s,t})|y_{s,t-1}^*) = F_{s,t}b_{s,t},
\]

\[
D_{s,t} = \text{Var}(y_{s,t}|y_{s,t-1}^*) \]

\[
D_{1c} = E(\text{Var}(y_{s,t}|\alpha_{s,t}, \beta_{s,t})|y_{s,t-1}^*) + \text{Var}(E(y_{s,t}|\alpha_{s,t}, \beta_{s,t})|y_{s,t-1}^*) = F_{s,t}B_{s,t}F_{s,t}' + V_{s,t}.
\]

Proof of the filtering distribution (3.17)

Let \((\alpha_s, \beta_s,t)' \equiv \theta_{s,t} \). Then we have

\[
[y_{s,t,\theta_{s,t}}|y_{s,t-1}^*] = \begin{bmatrix}
y_{s,t,\theta_{s,t}} \\
\theta_{s,t} \\
y_{s,t-1}^*
\end{bmatrix},
\]

\[
[\theta_{s,t}|y_{s,t-1}^*] = \begin{bmatrix}
y_{s,t,\theta_{s,t}} \\
\theta_{s,t} \\
y_{s,t-1}^*
\end{bmatrix},
\]

\[
[D_{s,t}] = \begin{bmatrix}
D_{s,t} & 0 & 0 \\
0 & D_{s,t} & 0 \\
0 & 0 & D_{s,t}
\end{bmatrix},
\]

\[
[b_{s,t}] = \begin{bmatrix}
b_{s,t} \\
b_{s,t} \\
b_{s,t}
\end{bmatrix}.
\]
where

\[
\begin{align*}
[y_{s,t} | \theta_{s,t}, y^*_{s,t-1}] & \overset{d}{=} N(F_{s,t} \theta_{s,t}, V_{s,t}), \\
[\theta_{s,t} | y^*_{s,t-1}] & \overset{d}{=} N(b_{s,t}, B_{s,t}).
\end{align*}
\]

By Result 1, \( \theta_{s,t} | y^*_{s,t} \sim N(a_{s,t}, A_{s,t}) \), where

\[
\begin{align*}
a_{s,t} &= b_{s,t} + B_{s,t} F'_{s,t} D_{s,t}^{-1} (y_{s,t} - d_{s,t}), \\
A_{s,t} &= B_{s,t} - B_{s,t} F_{s,t} F'_{s,t} D_{s,t}^{-1} F_{s,t} B_{s,t}.
\end{align*}
\]
APPENDIX E. PROOF OF THE DYNAMIC LINEAR MODEL
SMOOTHING EQUATIONS (PROPOSITION 3.4)

Some useful distributional equivalencies for the DLM Kalman smoother are

\[
[\theta_t|\theta_{t+1}, y_{1:t}] = [\theta_t|\theta_{t+1}, y_{1:t}], \quad \text{(E.1a)}
\]

\[
[\theta_{t+1}|\theta_t, y_{1:t}] = [\theta_{t+1}|\theta_t]. \quad \text{(E.1b)}
\]

The distributional equivalencies in E.1 can be verified by applying Proposition 3.1 to the appropriate moral graphs of Figure 3.5.

Proof of the Kalman smoother (3.4)

Note that,

\[
[\theta_{t+1}|\theta_t, y_{1:t}] \stackrel{E.1b}{=} [\theta_{t+1}|\theta_t] \overset{d}{=} N(G_{t+1}\theta_t, W_{t+1}),
\]

\[
[\theta_t|y_{1:t}] \overset{d}{=} N(m_t, C_t). \quad \text{(the filtering distribution)}
\]

By Result 1,

\[
[\theta_t|\theta_{t+1}, y_{1:t}] \overset{d}{=} N(h_t, H_t), \quad \text{(E.2)}
\]

where

\[
h_t = m_t + C_t G'_{t+1} R_{t+1}^{-1} (\theta_{t+1} - a_{t+1})
\]

\[
H_t = C_t - C_t G'_{t+1} R_{t+1}^{-1} G_{t+1} C_t.
\]
The proof is finished by noting

\[ s_t = E(\theta_t | y_{1:T}) \quad E.1^a = E(E(\theta_t | \theta_{t+1}, y_{1:t}) | y_{1:T}) \]

\[ = E(m_t + C_t G_{t+1} R_{t+1}^{-1} (\theta_{t+1} - a_{t+1}) | y_{1:T}) \]

\[ = m_t + C_t G_{t+1} R_{t+1}^{-1} (s_{t+1} - a_{t+1}), \]

\[ S_t = Var(\theta_t | y_{1:T}) \quad E.1^a = E(Var(\theta_t | \theta_{t+1}, y_{1:t}) | y_{1:T}) + Var(E(\theta_t | \theta_{t+1}, y_{1:t}) | y_{1:T}) \]

\[ = E(C_t - C_t G_{t+1} R_{t+1}^{-1} G_{t+1} C_t | y_{1:T}) + Var(m_t + C_t G_{t+1} R_{t+1}^{-1} (\theta_{t+1} - a_{t+1}) | y_{1:T}) \]

\[ = C_t - C_t G_{t+1} R_{t+1}^{-1} (R_{t+1} - S_{t+1}) R_{t+1}^{-1} G_{t+1} C_t. \]
APPENDIX F. PROOF OF THE MULTISCALE DYNAMIC LINEAR MODEL SMOOTHING EQUATIONS (PROPOSITION 3.5)

Some useful distributional equivalencies of the MSDLM Kalman smoother are

\[
\begin{align*}
[\beta_{s,t} | \beta_{s,t+1}, \alpha_s, y^*_s] &= [\beta_{s,t} | \beta_{s,t+1}, \alpha_s, y^*_s], \\
[\beta_{s,t+1} | \beta_{s,t}, \alpha_s, y^*_s] &= [\beta_{s,t+1} | \beta_{s,t}, \alpha_s], \\
[\alpha_s | \alpha_{s+1}, y^*_s] &= [\alpha_s | \alpha_{s+1}, y^*_s], \\
[\beta_{s,T} | \alpha_s, y^*_s] &= [\beta_{s,T} | \alpha_s, y^*_s].
\end{align*}
\]

(F.1a) (F.1b) (F.1c) (F.1d)

The distributional equivalencies in F.1 can be verified by applying Proposition 3.1 to the appropriate moral graphs of Figure 3.7.

Proof of the MSDLM Kalman smoother (3.5)

(a) For \( t \neq T \),

\[
(\alpha_s, \beta_{s,t+1})' | y^*_S \sim N\left( \begin{bmatrix} h_{\alpha,s} \\ h_{\beta,s,t+1} \end{bmatrix}, \begin{bmatrix} H_{\alpha,s} & H_{\alpha\beta,s,t+1} \\ H'_{\alpha\beta,s,t+1} & H_{\beta,s,t+1} \end{bmatrix} \right).
\]

The distribution of \((\alpha_s, \beta_{s,t})' | y^*_S\) will be Gaussian. Thus determining this distribution reduces to determining the first two moments. The distribution of \(\alpha_s | y^*_S\) is \(N(h_{\alpha,s}, H_{\alpha,s})\). What remains is determining \(E(\beta_{s,t} | y^*_S), Var(\beta_{s,t} | y^*_S), \) and \(Cov(\alpha_s, \beta_{s,t} | y^*_S)\). Note that

\[
[\beta_{s,t+1} | \beta_{s,t}, \alpha_s, y^*_s] \overset{F.1b}{=} [\beta_{s,t+1} | \beta_{s,t}, \alpha_s] \overset{d}{=} N(G_{\beta,s,t+1} \beta_{s,t}, W_{\beta,s,t+1}), \\
[\beta_{s,t} | \alpha_s, y^*_s] \overset{d}{=} N(a_{\beta,s,t | \alpha_s}, A_{\beta,s,t | \alpha_s}),
\]

(F.2)
where
\[ a_{\beta,s,t|\alpha_s} = a_{\beta,s,t} + A'_{\alpha\beta,s,t}A^{-1}_{\alpha,s,t}(\alpha_s - a_{\alpha,s,t}), \]
\[ A_{\beta,s,t|\alpha_s} = A_{\beta,s,t} - A'_{\alpha\beta,s,t}A^{-1}_{\alpha,s,t}A_{\alpha,s,t}. \]

This is true by standard conditional Gaussian arguments applied to the filtering distribution of \((\alpha_s, \beta_{s,t})'|y_{s,t}^*.\)

By Result 1, we have
\[ [\beta_{s,t}|\beta_{s,t+1}, \alpha_s, y_{S,T}^*] \overset{d}{=} N(q_{\beta,s,t}, Q_{\beta,s,t}), \quad \text{(F.3)} \]

where
\[ q_{\beta,s,t} = a_{\beta,s,t|\alpha_s} + R_{\beta,s,t}(\beta_{s,t+1} - G_{\beta,s,t+1}a_{\beta,s,t|\alpha_s}), \]
\[ Q_{\beta,s,t} = (I - R_{\beta,s,t}G_{\beta,s,t+1})A_{\beta,s,t|\alpha_s}, \]
\[ R_{\beta,s,t} = A_{\beta,s,t|\alpha_s}G'_{\beta,s,t+1}(G_{\beta,s,t+1}A_{\beta,s,t|\alpha_s}G'_{\beta,s,t+1} + W_{\beta,s,t+1})^{-1}. \]

By F.1a, we have
\[ [\beta_{s,t}|\beta_{s,t+1}, \alpha_s, y_{S,T}^*] \overset{d}{=} N(q_{\beta,s,t}, Q_{\beta,s,t}). \quad \text{(F.4)} \]

The proof is completed by noting
\[
E(\beta_{s,t}|y_{S,T}^*) = E(E(\beta_{s,t}|\beta_{s,t+1}, \alpha_s, y_{S,T}^*)|y_{S,T}^*) \overset{F.4}{=} E(q_{\beta,s,t}|y_{S,T}^*)
\]
\[ = E(a_{\beta,s,t|\alpha_s}|y_{S,T}^*) + R_{\beta,s,t}(h_{\beta,s,t+1} - G_{\beta,s,t+1}E(a_{\beta,s,t|\alpha_s}|y_{S,T}^*)], \]

where
\[ E(a_{\beta,s,t|\alpha_s}|y_{S,T}^*) = a_{\beta,s,t} + A'_{\alpha\beta,s,t}A^{-1}_{\alpha,s,t}(h_{\alpha,s} - a_{\alpha,s,t}), \]
\[ H_{\alpha\beta,s,t} = \text{Cov}(\alpha_s, \beta_{s,t}|y_{S,T}^*) = \text{Cov}(\alpha_s, E(\beta_{s,t}|\beta_{s,t+1}, \alpha_s, y_{S,T}^*)|y_{S,T}^*) \overset{F.4}{=} \text{Cov}(\alpha_s, q_{\beta,s,t}|y_{S,T}^*)
\]
\[ = H_{\alpha,s}A^{-1}_{\alpha,s,t}A_{\alpha\beta,s,t}(I - R_{\beta,s,t}G_{\beta,s,t+1})', \]
and

\[ \text{Var}(\beta_{s,t}|y_{S,T}^*) = E(\text{Var}(\beta_{s,t}|\beta_{s,t+1}, \alpha_s, y_{S,T}^*)|y_{S,T}^*) + \text{Var}(E(\beta_{s,t}|\beta_{s,t+1}, \alpha_s, y_{S,T}^*)|y_{S,T}^*) \]

\[ \overset{\text{F.4}}{=} E(Q_{\beta,s,t}|y_{S,T}^*) + \text{Var}(q_{\beta,s,t}|y_{S,T}^*) = Q_{\beta,s,t} + H^\prime_{\alpha\beta,s,t}H_{\alpha,s}H_{\alpha\beta,s,t}. \]

(b) For \( s \neq S \) and \( t = T \), we have

\[ (\alpha_{s+1}, \beta_{s+1,1})'|y_{S,T}^* \sim N\left( \begin{bmatrix} h_{\alpha,s+1} \\ h_{\beta,s+1,1} \end{bmatrix}, \begin{bmatrix} H_{\alpha,s+1} & H_{\alpha\beta,s+1,1} \\ H_{\alpha\beta,s+1,1} & H_{\beta,s+1,1} \end{bmatrix} \right). \]

Note

\[ [\alpha_{s+1}|\alpha_s, y_{S,T}^*] = [\alpha_{s+1}|\alpha_s] \overset{d}{=} N(G_{\alpha,s+1}\alpha_s, W_{\alpha,s+1}), \]

\[ [\alpha_s|y_{S,T}^*] \overset{d}{=} N(a_{\alpha,s,T}, A_{\alpha,s,T}). \quad \text{(the filtering distribution)} \]

By Result 1, we have

\[ [\alpha_s|\alpha_{s+1}, y_{S,T}^*] \overset{d}{=} N(q_{\alpha,s}, Q_{\alpha,s}), \quad \text{(F.5)} \]

where

\[ q_{\alpha,s} = a_{\alpha,s,T} + R_{\alpha,s}(\alpha_{s+1} - G_{\alpha,s+1}a_{\alpha,s,T}), \]

\[ Q_{\alpha,s} = A_{\alpha,s,T} - R_{\alpha,s}G_{\alpha,s+1}A_{\alpha,s,T}, \]

\[ R_{\alpha,s} = A_{\alpha,s,T}G_{\alpha,s+1}B_{\alpha,s+1,1}^{-1}. \]

It then follows that

\[ h_{\alpha,s} = E(\alpha_s|y_{S,T}^*) \overset{\text{F.1c}}{=} E(E(\alpha_s|\alpha_{s+1}\alpha_s, y_{S,T}^*)|y_{S,T}^*) \overset{\text{F.5}}{=} E(q_{\alpha,s}|y_{S,T}^*) \]

\[ = a_{\alpha,s,T} + R_{\alpha,s}(h_{\alpha,s+1} - G_{\alpha,s+1}a_{\alpha,s,T}), \]

\[ H_{\alpha,s} = \text{Var}(\alpha_s|y_{S,T}^*) \overset{\text{F.1c}}{=} E(\text{Var}(\alpha_s|\alpha_{s+1}, y_{S,T}^*)|y_{S,T}^*) + \text{Var}(E(\alpha_s|\alpha_{s+1}, y_{S,T}^*)|y_{S,T}^*) \]

\[ \overset{\text{F.5}}{=} E(Q_{\alpha,s}|y_{S,T}^*) + \text{Var}(q_{\alpha,s}|y_{S,T}^*) = Q_{\alpha,s} + R_{\alpha,s}H_{\alpha,s+1}R^\prime_{\alpha,s}. \]

Next, note that

\[ [\beta_{s,T}|\alpha_s, y_{S,T}^*] \overset{d}{=} N(a_{\beta,s,T}|\alpha_s, A_{\beta,s,T}|\alpha_s), \quad \text{(F.6)} \]
where
\[ a_{\beta,s,T|\alpha_s} = a_{\beta,s,T} + A'_{\alpha\beta,s,T} A^{-1}_{\alpha,s,T} (\alpha_s - a_{\alpha,s,T}), \]
\[ A_{\beta,s,T|\alpha_s} = A_{\beta,s,T} - A'_{\alpha\beta,s,T} A^{-1}_{\alpha,s,T} A_{\alpha\beta,s,T}. \]

This is a straightforward application of properties of conditional normal distributions applied to the filtering distribution,
\[ \left( \begin{array}{c} \alpha_s \\ \beta_s,T \end{array} \right) \bigg| y^*_{s,T} \sim N \left( \begin{bmatrix} a_{\alpha,s,T} \\ a_{\beta,s,T} \end{bmatrix}, \begin{bmatrix} A_{\alpha,s,T} & A_{\alpha\beta,s,T} \\ A'_{\alpha\beta,s,T} & A_{\beta,s,T} \end{bmatrix} \right). \]

The proof is completed by noting that
\[ h_{\beta,s,T} = E(\beta_{s,T}|y^*_{S,T}) \overset{\text{F.1d}}{=} E(E(\beta_{s,T}|\alpha_s, y^*_{s,T})|y^*_{S,T}) \]
\[ = E(a_{\beta,s,T|\alpha_s}|y^*_{S,T}) = a_{\beta,s,T} + A'_{\alpha\beta,s,T} A^{-1}_{\alpha,s,T} (h_{\alpha,s} - a_{\alpha,s,T}), \]
\[ H_{\alpha\beta,s,T} = Cov(\alpha_s, \beta_{s,T}|y^*_{S,T}) \overset{\text{F.1d}}{=} Cov(\alpha_s, E(\beta_{s,T}|\alpha_s, y^*_{s,T})|y^*_{S,T}) \]
\[ = Cov(\alpha_s, a_{\beta,s,T|\alpha_s}|y^*_{S,T}) = H_{\alpha,s} A^{-1}_{\alpha,s,T} A_{\alpha\beta,s,T}; \]
\[ H_{\beta,s,T} = Var(\beta_{s,T}|y^*_{S,T}) \overset{\text{F.1d}}{=} E(Var(\beta_{s,T}|\alpha_s, y^*_{s,T})|y^*_{S,T}) + Var(E(\beta_{s,T}|\alpha_s, y^*_{s,T})|y^*_{S,T}) \]
\[ = E(A_{\beta,s,T|\alpha_s}|y^*_{S,T}) + Var(a_{\beta,s,T|\alpha_s}|y^*_{S,T}) \]
\[ = A_{\beta,s,T|\alpha_s} + A'_{\alpha\beta,s,T} A^{-1}_{\alpha,s,T} H_{\alpha,s} A^{-1}_{\alpha,s,T} A_{\alpha\beta,s,T}. \]
APPENDIX G. PROOF OF THE DYNAMIC LINEAR MODEL FORECASTING RECURSIONS (PROPOSITION 3.6)

For \( k \geq 1 \), some useful distributional equivalencies for the DLM forecasting recursions are

\[
[\theta_{t+k}|\theta_{t+k-1}, y_{1:t}] = [\theta_{t+k}|\theta_{t+k-1}],
\]

(G.1a)

\[
[y_{t+k}|\theta_{t+k}, y_{1:t}] = [y_{t+k}|\theta_{t+k}].
\]

(G.1b)

The distributional equivalencies in G.1 can be verified by applying Proposition 3.1 to the appropriate moral graphs of Figure 3.5.

Proof of the many-step-ahead predictive distribution of \( \theta_{t+k} \) given \( y_{1:t} \) (3.21)

\[
a_t(k) = E(\theta_{t+k}|y_{1:t}) \stackrel{G.1a}{=} E(E(\theta_{t+k}|\theta_{t+k-1})|y_{1:t}) = E(G_{t+k}\theta_{t+k-1}|y_{1:t}) = G_{t+k}a_t(k-1),
\]

\[
R_t(k) = Var(\theta_{t+k}|y_{1:t}) \stackrel{G.1a}{=} E(Var(\theta_{t+k}|\theta_{t+k-1})|y_{1:t}) + Var(E(\theta_{t+k}|\theta_{t+k-1})|y_{1:t})
\]

\[
= E(W_{t+k}|y_{1:t}) + Var(G_{t+k}\theta_{t+k-1}|y_{1:t}) = G_{t+k}R_t(k-1)G'_{t+k} + W_{t+k}.
\]

Proof of the many-step-ahead forecasting distribution of \( y_{t+k} \) given \( y_{1:t} \) (3.22)

\[
f_t(k) = E(y_{t+k}|y_{1:t}) \stackrel{G.1b}{=} E(E(y_{t+k}|\theta_{t+k})|y_{1:t}) = E(F_{t+k}\theta_{t+k}|y_{1:t}) = F_{t+k}a_t(k),
\]

\[
Q_t(k) = Var(y_{t+k}|y_{1:t}) \stackrel{G.1b}{=} E(Var(y_{t+k}|\theta_{t+k})|y_{1:t}) + Var(E(y_{t+k}|\theta_{t+k})|y_{1:t})
\]

\[
= E(V_{t+k}|y_{1:t}) + Var(F_{t+k}\theta_{t+k}|y_{1:t}) = F_{t+k}R_t(k)F'_{t+k} + V_{t+k}.
\]
APPENDIX H. PROOF OF THE MULTISCALE DYNAMIC LINEAR
MODEL FORECASTING RECURSIONS (PROPOSITION 3.7)

Some useful distributional equivalencies for the MSDLM forecasting recursions.
For \( k \geq 1 \) and \( t + k \leq T \),
\[
[\alpha_{s}, \beta_{s,t+k}| y_{s,t}^{*}] = [\alpha_{s}, \beta_{s,t+k}| y_{s,t}^{*}] = [\alpha_{s}, \beta_{s,t+k}| \alpha_{s}, \beta_{s,t+k-1}]. 
\] (H.1)

For \( l \geq 1 \) and \( 1 \leq t + k \leq T \),
\[
[\alpha_{s+l}, \beta_{s+l,t+k}| y_{s,t}^{*}] = [\alpha_{s+l}, \beta_{s+l,t+k}| \alpha_{s+l-1}]. 
\] (H.2)

For EITHER \( l = 0 \), \( k \geq 1 \), and \( t + k \leq T \) OR \( l \geq 1 \) and \( 1 \leq t + k \leq T \),
\[
[y_{s+l,t+k}| y_{s,t}^{*}, \alpha_{s+l}, \beta_{s+l,t+k}] = [y_{s+l,t+k}| \alpha_{s+l}, \beta_{s+l,t+k}]. 
\] (H.3)

The distributional equivalencies in H.1, H.2, and H.3 can be verified by applying Proposition 3.1 to the appropriate moral graphs of Figure 3.7.

Recall
\[
\begin{pmatrix}
\alpha_s \\
\beta_{s,t}
\end{pmatrix}
\left| y_{s,t}^{*}
\right. \sim N
\begin{pmatrix}
\left[ b_{\alpha,s} | y_{s,t}^{*} \\
 b_{\beta,s,t} | y_{s,t}^{*}
\right],
\left[ B_{\alpha,s,t} | y_{s,t}^{*} \\
 B'_{\alpha,s,t} | y_{s,t}^{*}
\right]
\end{pmatrix}.
\]

Proof of the many-step-ahead predictive distribution (3.24 and 3.25):
For both case (i) where \(s, t + k \leq T\), \((\alpha_s, \beta_{s, t+k})'\) given \(y_{s,t}^*\) is distributed as \(N(b_{s,t+k|y_{s,t}^*}, B_{s,t+k|y_{s,t}^*})\), with

\[
b_{s,t+k|y_{s,t}^*} = (b_{a,s|y_{s,t}^*}, b_{B_{s,t+k|y_{s,t}^*}})' = E((\alpha_{s,t+k})'|y_{s,t}^*)
\]

\[
H.1 \quad E(E((\alpha_{s,t+k})'|\alpha_{s,t+k-1})|y_{s,t}^*)
\]

\[
=b_{a,s|y_{s,t}^*}, G_{\beta_{s,t+k-1}|y_{s,t}^*} = \begin{bmatrix} B_{a,s|y_{s,t}^*} & B_{B_{s,t+k|y_{s,t}^*}} \\ B_{\alpha_{s,t+k}} & \beta_{s,t+k} \end{bmatrix}
\]

\[
H.1 \quad E(Var((\alpha_{s,t+k})'|\alpha_{s,t+k-1})|y_{s,t}^*) + Var(E((\alpha_{s,t+k})'|\alpha_{s,t+k-1})|y_{s,t}^*)
\]

\[
B_{s,t+k|y_{s,t}^*} = \begin{bmatrix} B_{a,s|y_{s,t}^*} & B_{\alpha_{s,t+k}} & G_{\beta_{s,t+k}} \\ B_{\alpha_{s,t+k}} & \beta_{s,t+k} & G_{\beta_{s,t+k}} \\ G_{\beta_{s,t+k}} & G_{\beta_{s,t+k}} & + W_{\beta_{s,t+k}} \end{bmatrix}
\]

(ii) For \(t + k \leq T\), \((\alpha_{s+t}, \beta_{s+t+k})'\) given \(y_{s,t}^*\) is distributed as \(N(b_{s+t+k|y_{s,t}^*}, B_{s+t+k|y_{s,t}^*})\), with

\[
b_{s+t+k|y_{s,t}^*} = (b_{a,s+t|y_{s,t}^*}, b_{B_{s+t+k|y_{s,t}^*}})' = E((\alpha_{s+t+k})'|y_{s,t}^*)
\]

\[
H.2 \quad E(E((\alpha_{s+t+k})'|\alpha_{s+t+k-1})|y_{s,t}^*)
\]

\[
=b_{a,s+t|y_{s,t}^*}, G_{\beta_{s+t+k-1}|y_{s,t}^*} = \begin{bmatrix} B_{a,s+t|y_{s,t}^*} & B_{B_{s+t+k|y_{s,t}^*}} \\ B_{\alpha_{s+t+k}} & \beta_{s+t+k} \end{bmatrix}
\]

\[
H.2 \quad E(Var((\alpha_{s+t+k})'|\alpha_{s+t+k-1})|y_{s,t}^*) + Var(E((\alpha_{s+t+k})'|\alpha_{s+t+k-1})|y_{s,t}^*)
\]

\[
B_{s+t+k|y_{s,t}^*} = \begin{bmatrix} G_{\alpha_{s+t}} B_{\alpha_{s+t-1}} & 0 \\ 0 & B_{\beta_{s+t+k}} \end{bmatrix}
\]

where

\[
B_{\beta_{s+t+k}} = \begin{cases} W_{\beta_{s+t+k}} & \text{if } t + k = 1 \\ G_{\beta_{s+t+k}} B_{\beta_{s+t+k-1}} & \text{if } 2 \leq t + k \leq T \end{cases}
\]

Proof of the many-step-ahead forecasting distribution (3.26):

For both case (i) where \(l = 0, k \geq 1\), and \(t + k \leq T\), and case (ii) where \(l \geq 1\) and \(1 \leq t + k \leq T\),
the many-step-ahead forecasting distribution for $y_{s+l,t+k}$ given $y_{s,t}^*$ is $N(d_{s+l,t+k|y_{s,t}^*}, D_{s+l,t+k|y_{s,t}^*})$, with

$$d_{s+l,t+k|y_{s,t}^*} = E(y_{s+l,t+k|y_{s,t}^*})$$

$$= E(E(y_{s+l,t+k|\alpha_{s+l}, \beta_{s+l,t+k}}|y_{s,t}^*)|y_{s,t}^*) = F_{s+l,t+k|y_{s,t}^*}$$

$$D_{s+l,t+k|y_{s,t}^*} = Var(y_{s+l,t+k|y_{s,t}^*})$$

$$= E(Var(y_{s+l,t+k|\alpha_{s+l}, \beta_{s+l,t+k}}|y_{s,t}^*) + Var(E(y_{s+l,t+k|\alpha_{s+l}, \beta_{s+l,t+k}}|y_{s,t}^*))$$

$$= F_{s+l,t+k|y_{s,t}^*} B_{s+l,t+k|y_{s,t}^*} F_{s+l,t+k|y_{s,t}^*} + V_{s+l,t+k},$$

where $F_{s+l,t+k} \equiv [F_{\alpha,s+l,t+k}, F_{\beta,s+l,t+k}].$
APPENDIX I. THE DYNAMIC LINEAR MODEL REPRESENTATION OF THE MULTISCALE DYNAMIC LINEAR MODEL

Consider the MSDLM of equation 3.8. Equivalently, the MSDLM can be written as,

\[ y_{s,t} = F_{\alpha,s,t}\alpha_s + F_{\beta,s,t}\beta_{s,t} + v_{s,t}, \]  \hspace{1cm} (I.1)

\[ \alpha_s = G_{\alpha,s}\alpha_{s-1} + w_{\alpha,s}, \quad (1 \leq s) \]  \hspace{1cm} (I.2)

\[ \beta_{s,t} = G_{\beta,s,t}\beta_{s,t-1} + w_{\beta,s,t}, \quad (2 \leq t) \]  \hspace{1cm} (I.3)

where \( \alpha_0 \sim N(\alpha_0, A_{\alpha,0}) \), \( \beta_{s,1} \sim N(0, W_{\beta,s,1}) \), \( v_{s,t} \sim N(0, V_{s,t}) \), \( w_{\alpha,s} \sim N(0, W_{\alpha,s}) \), and \( w_{\beta,s,t} \sim N(0, W_{\beta,s,t}) \). Furthermore, \( \alpha_s \) is \( p \times 1 \), \( \beta_{s,t} \) is \( q \times 1 \), and \( y_{s,t} \) is a scalar. Finally, \( \alpha_0 \), \( (\beta_{s,1}), (v_{s,t}), (w_{\alpha,s}), \) and \( (w_{\beta,s,t}) \) are assumed mutually independent.
Some results for the MSDLM:

\[ E(\alpha_s) = G_{\alpha,s}G_{\alpha,s-1} \ldots G_{\alpha,1}a_{\alpha,0} \quad (I.4) \]

\[ Var(\alpha_s) = G_{\alpha,s}Var(\alpha_{s-1})G_{\alpha,s} + W_{\alpha,s} \quad (I.5) \]

\[ E(\beta_{s,t}) = 0 \quad (I.6) \]

\[ Var(\beta_{s,1}) = W_{\beta,s,1} \quad (I.7) \]

\[ Var(\beta_{s,t}) = G_{\beta,s,t}Var(\beta_{s,t-1})G'_{\beta,s,t} + W_{\beta,s,t} \quad (2 \leq t) \quad (I.8) \]

\[ Cov(\alpha_s, \beta_{s,t}) = 0 \quad (I.9) \]

\[ Var((\alpha'_s, \beta'_{s,t})') = \begin{bmatrix} Var(\alpha_s) & 0 \\ 0 & Var(\beta_{s,t}) \end{bmatrix} \]

\[ E(y_{s,t}) = F_{\alpha,s,t}G_{\alpha,s}G_{\alpha,s-1} \ldots G_{\alpha,1}a_{\alpha,0} \quad (I.11) \]

\[ Var(y_{s,t}) = F_{\alpha,s,t}Var(\alpha_s)F'_{\alpha,s,t} + F_{\beta,s,t}Var(\beta_{s,t})F'_{\beta,s,t} + V_{s,t} \quad (I.12) \]

\[ Cov(y_{s,t}, y_{s',t'}) = F_{\alpha,s,t}Var(\alpha_s)(F_{\alpha,s,t'}G_{\alpha,s}G_{\alpha,s-1} \ldots G_{\alpha,1})' \quad (s < s') \quad (I.14) \]

Now consider the DLM of 3.5. Equivalently, the DLM can be written as,

\[ y_t = F_t \theta_t + v_t, \]

\[ \theta_t = G_t \theta_{t-1} + w_t, \]

where \( \theta_0 \sim N(m_0, C_0) \), \( v_t \sim N(0, V_t) \), and \( w_t \sim N(0, W_t) \). Furthermore, \( \alpha_0 \), \( (w_t) \), and \( (v_t) \) are assumed mutually independent.
Some results for the DLM:

\[ E(\theta_t) = G_tG_{t-1} \ldots G_1m_0 \]  \hspace{1cm} (I.15)

\[ Var(\theta_t) = G_tVar(\theta_{t-1})G'_t + W_t \]  \hspace{1cm} (I.16)

\[ Cov(\theta_t, \theta_{t'}) = Var(\theta_t)(G_tG'_{t-1} \ldots G_1)m_0' \]  \hspace{1cm} (I.17)

\[ E(y_t) = F_tG_tG_{t-1} \ldots G_1m_0 \]  \hspace{1cm} (I.18)

\[ Var(y_t) = F_tVar(\theta_t)F'_t + V_t \]  \hspace{1cm} (I.19)

\[ Cov(y_t, y_{t'}) = F_tVar(\theta_t)(F_tG_tG'_{t-1} \ldots G_1)m_0' \]  \hspace{1cm} (I.20)

The MSDLM can be expressed as a DLM. More precisely, for any MSDLM, it is possible to find a DLM whose measurement process \((y_t)\) has the same distribution as the measurement process of the MSDLM, \((y_{s,t})\). Recall that the joint distribution of all observations and latent states for the DLM and the MSDLM is multivariate normal for fully specified models (i.e., \(F_t, G_t, V_t, W_t, m_0\), and \(C_0\) known for DLMs and \(F_{\alpha,s,t}, F_{\beta,s,t}, G_{\alpha,s}, G_{\beta,s,t}, V_{s,t}, W_{\alpha,s}, W_{\beta,s,t}, a_{\alpha,0}\), and \(A_{\alpha,0}\) known for MSDLMs). Thus, for both the DLM and the MSDLM, the joint distribution for the measurement process is multivariate normal. To show there exists a DLM representation for the measurement process of any MSDLM, it suffices to show that,

\[ E(y_{s,t}) = E(y_{(s-1)T+t}), \]  \hspace{1cm} (I.21)

\[ Var(y_{s,t}) = Var(y_{(s-1)T+t}), \]  \hspace{1cm} (I.22)

\[ Cov(y_{s,t}, y_{s',t'}) = Cov(y_{(s-1)T+t}, y_{(s'-1)T+t'}), \]  \hspace{1cm} (I.23)

where \(y_{s,t}\) using the MSDLM time indexing equals \(y_{(s-1)T+t}\) using the DLM time indexing. For clarity, consider the 13th observation of a time series. For the MSDLM, if we assume \(T = 10\), the 13th observation is the third observation in the second coarse scale time unit and is represented as \(y_{2,3}\). For the DLM, it is simply \(y_{13} = y_{10(2-1)+3} = y_{(s-1)T+t} \).
For any fully specified MSDLM described by equations I.1 - I.3, its DLM representation is specified as follows:

\[ F_{(s-1)T+t} = [F_{\alpha,s,t}, F_{\beta,s,t}]_{1 \times (p+q)} \]  
\[ V_{(s-1)T+t} = V_{s,t} \]  
\[ G_{(s-1)T+t} = \begin{cases} 
G_{\alpha,s} & \text{if } t = 1 \\
0 & (p+q) \times (p+q) \\
I & \text{if } t \neq 1 \\
0 & G_{\beta,s,t} 
\end{cases} \]  
\[ W_{(s-1)T+t} = \begin{cases} 
W_{\alpha,s} & \text{if } t = 1 \\
0 & W_{\beta,s,t} \\
0 & (p+q) \times (p+q) \\
0 & (p+q) \times (p+q) 
\end{cases} \]  
\[ m'_0 = [a'_{\alpha,0}, 0']_{1 \times (p+q)} \]  
\[ C_0 = \begin{bmatrix} A_{\alpha,0} & 0 \\
0 & 0 \end{bmatrix}_{(p+q) \times (p+q)}, \]  

where \( \theta_{(s-1)T+t} \) is a \((p + q) \times 1\) vector and \( I \) is an identity matrix.

The proof that the DLM specification of equations I.24 - I.29 has the same measurement process distribution as the MSDLM follows with Results 2 - 4.

**Result 2.** For any fully specified MSDLM described by equations I.1 - I.3 and the DLM specified in equations I.24 - I.29, it is true that

\[ E(y_{(s-1)T+t}) = E(y_{s,t}). \]
Proof. Define

\[
G_{s,t}^* = G_{(s-1)T+t}G_{(s-1)T+t-1} \cdots G_{(s-1)T+2}G_{(s-1)T+1}
\]  \hspace{1cm} (I.30)

\[
\begin{bmatrix}
I & 0 \\
0 & G_{\beta,s,t}
\end{bmatrix}
\begin{bmatrix}
I & 0 \\
0 & G_{\beta,s,t-1}
\end{bmatrix}
\cdots
\begin{bmatrix}
I & 0 \\
0 & G_{\beta,s,2}
\end{bmatrix}
\begin{bmatrix}
G_{\alpha,s} & 0 \\
0 & 0
\end{bmatrix}
\]

\[
= \begin{bmatrix}
G_{\alpha,s} & 0 \\
0 & 0
\end{bmatrix}
\]  \hspace{1cm} (I.31)

Then,

\[
E(y_{(s-1)T+t}) = F_{(s-1)T+t}G_{(s-1)T+t}G_{(s-1)T+t-1} \cdots G_1m_0
\]

\[
= F_{(s-1)T+t}G_{s,t}^*G_{s-1,T}G_{s-2,T} \cdots G_{1,T}m_0
\]

\[
= [F_{\alpha,s,t}, F_{\beta,s,t}]
\begin{bmatrix}
G_{\alpha,s}G_{\alpha,s-1} \cdots G_{\alpha,1} & 0 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
a_{\alpha,0} \\
0
\end{bmatrix}
\]

\[
= F_{\alpha,s,t}G_{\alpha,s}G_{\alpha,s-1} \cdots G_{\alpha,1}a_{\alpha,0}
\]

\[
= E(y_{s,t})
\]

\[
\Box
\]

Result 3. For any fully specified MSDLM described by equations I.1 - I.3 and the DLM specified in equations I.24 - I.29, it is true that

\[
\text{Var}(y_{(s-1)T+t}) = \text{Var}(y_{s,t}).
\]
Proof. First note that

\[ \text{Var}(\theta_0) I.29 = \begin{bmatrix} A_{\alpha,0} & 0 \\ 0 & 0 \end{bmatrix}, \]

\[ \text{Var}(\theta_1) I.16 = G_1 \text{Var}(\theta_0) G_1' + W_1 \]

\[ I.26, I.27 \implies \begin{bmatrix} G_{\alpha,1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} A_{\alpha,0} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} G_{\alpha,1}' & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} W_{\alpha,1} & 0 \\ 0 & W_{\beta,1,1} \end{bmatrix} \]

\[ = \begin{bmatrix} G_{\alpha,1} A_{\alpha,0} G_{\alpha,1}' + W_{\alpha,1} & 0 \\ 0 & W_{\beta,1,1} \end{bmatrix} \]

\[ I.5, I.7 \implies \begin{bmatrix} V(\alpha_1) & 0 \\ 0 & V(\beta_1,1) \end{bmatrix} I.10 = V((\alpha_1', \beta_{1,1}')'). \]

Next, note the following cases:

(a) Case 1: \( t \neq T \)

Given \( \text{Var}(\theta_{(s-1)T+t}) = V((\alpha_s', \beta_{s,t}')') \), then

\[ \text{Var}(\theta_{(s-1)T+t+1}) I.16 = G_{(s-1)T+t+1} \text{Var}(\theta_{(s-1)T+t}) G_{(s-1)T+t+1}' + W_{(s-1)T+t+1} \]

\[ I.26, I.27 \implies \begin{bmatrix} I & 0 \\ 0 & G_{\beta,s,t+1} \end{bmatrix} \begin{bmatrix} \text{Var}(\alpha_s) & 0 \\ 0 & \text{Var}(\beta_{s,t}) \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & G_{\beta,s,t+1}' \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & W_{\beta,s,t+1} \end{bmatrix} \]

\[ = \begin{bmatrix} \text{Var}(\alpha_s) & 0 \\ 0 & G_{\beta,s,t+1} \text{Var}(\beta_{s,t}) G_{\beta,s,t+1}' + W_{\beta,s,t+1} \end{bmatrix} \]

\[ I.8 \implies \begin{bmatrix} \text{Var}(\alpha_s) & 0 \\ 0 & \text{Var}(\beta_{s,t+1}) \end{bmatrix} \]

\[ I.10 \implies \text{Var}((\alpha_s', \beta_{s,t+1}')'). \quad (I.32) \]
(b) Case 2: \( t = T \)

Given \( \text{Var}(\theta_{(s-1)T+T}) = V((\alpha'_s, \beta'_{s,T}')) \), then

\[
\text{Var}(\theta_{sT+1}) = \text{G}_{sT+1}\text{Var}(\theta_{(s-1)T+T})\text{G}_{sT+1}' + \text{W}_{sT+1}
\]

\[
\text{I.16} \quad \text{I.26, I.27} = \begin{bmatrix} G_{s,s+1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \text{Var}(\alpha_s) & 0 \\ 0 & \text{Var}(\beta_{s,T}) \end{bmatrix} \begin{bmatrix} G_{s,s+1}' \\ 0 \end{bmatrix} + \begin{bmatrix} W_{s,s+1} & 0 \\ 0 & W_{s,s+1,1} \end{bmatrix}
\]

\[
\text{I.33} \quad \text{I.10} = \text{Var}((\alpha'_{s+1}, \beta'_{s+1,1}')).
\]

From I.32 and I.33, for all \( s \) and \( t \), we have

\[
\text{Var}(\theta_{(s-1)T+t}) = \text{Var}((\alpha'_s, \beta'_{s,t}')).
\] (I.34)

The proof is finished by noting

\[
\text{Var}(y_{(s-1)T+t}) = \text{Var}(y_{s,t}).
\] (I.35)

\[
\text{Result 4.} \quad \text{For any fully specified MSDLM described by equations I.1 - I.3 and the DLM specified in equations I.24 - I.29, it is true that}
\]

\[
\text{Cov}(y_{(s-1)T+t}, y_{(s'-1)T+t'}) = \text{Cov}(y_{s,t}, y_{s',t'}).
\]
Proof. The proof is divided into two cases.

(a) Case 1: $t < t'$

\[
\text{Cov}(y_{(s-1)T+t}, y_{(s-1)T+t'}) = F_{(s-1)T+t} Var(\theta_{(s-1)T+t}) \times \left( F_{(s-1)T+t'} G_{(s-1)T+t'} G_{(s-1)T+t-1} \ldots G_{(s-1)T+t+1} \right)'
\]

\[
\overset{1.24,34,1.26}{=} [F_{\alpha,s,t}, F_{\beta,s,t}] \begin{bmatrix} Var(\alpha_s) & 0 \\ 0 & Var(\beta_s,t) \end{bmatrix} \times \\
\left( \begin{bmatrix} I & 0 \\ 0 & G_{\beta,s,t'} G_{\beta,s,t-1} \ldots G_{\beta,s,t+1} \end{bmatrix} \right)'
\]

\[
= F_{\alpha,s,t} Var(\alpha_s) F'_{\alpha,s,t} + F_{\beta,s,t} Var(\beta_s,t) (F_{\beta,s,t'} G_{\beta,s,t'} G_{\beta,s,t-1} \ldots G_{\beta,s,t+1})'
\]

\[
\overset{1.13}{=} = \text{Cov}(y_{s,t}, y_{s,t'}).
\]

(b) Case 2: $s < s'$

\[
\text{Cov}(y_{(s-1)T+t}, y_{(s'-1)T+t'}) = F_{(s-1)T+t} Var(\theta_{(s-1)T+t}) \times \left( F_{(s'-1)T+t'} G_{(s'-1)T+t'} G_{(s'-1)T+t-1} \ldots G_{(s-1)T+t+1} \right)'
\]

\[
\overset{1.24,34,1.26}{=} [F_{\alpha,s,t}, F_{\beta,s,t}] \begin{bmatrix} Var(\alpha_s) & 0 \\ 0 & Var(\beta_s,t) \end{bmatrix} \times \\
\left( \begin{bmatrix} G_{\alpha,s,t'} G_{\alpha,s-1} \ldots G_{\alpha,s+1} & 0 \\ 0 & 0 \end{bmatrix} \right)'
\]

\[
= F_{\alpha,s,t} Var(\alpha_s) (F_{\alpha,s,t'} G_{\alpha,s'} G_{\alpha,s'-1} \ldots G_{\alpha,s+1})'
\]

\[
\overset{1.14}{=} = \text{Cov}(y_{s,t}, y_{s',t'}).
\]

\[\Box\]

By Results 2 - 4, the proof is complete.
BIBLIOGRAPHY


