Bayesian modeling and computation with latent variables

by

Matthew Simpson

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics & Economics

Program of Study Committee:

Jarad Niemi, Major Professor

Gray Calhoun

Alicia Carriquiry

Brent Kreider

Vivekananda Roy

Iowa State University

Ames, Iowa

2015

DEDICATION

For Nora. There is light at the end of the tunnel.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank by former and current major professors both, Dr. Alyson Wilson and Dr. Jarad Niemi, for allowing me to pursue the lines of research that interested me, make mistakes, and learn along the way. Their guidance through this process was cruicial and I feel grateful for having this freedom throughout graduate school.

And of course I would like to thank my mom for believing in me and supporting me through a decade of collegiate education.

CHAPTER 1. INTRODUCTION

This dissertation is a collection of papers in the large and varied field of Bayesian statistics and econometrics. The Bayesian method has proven to be a powerful technique for combining data and prior knowledge to answer scientific questions when the appropriate model can be constructed and the posterior distribution can be computed, but there are always limits to our ability to perform both tasks. This dissertation attempts to improve our collective abilities to overcome both obstacles largely by using latent variable techniques.

1.1 Interweaving in dynamic linear models

Historically the largest impediment to Bayesian statistics was computation. Due to the work of Reverend Thomas Bayes and Pierre-Simon Laplace the statistics and mathematics communities have known about Bayes' rule and the Bayesian method for approaching statistical inference for a couple hundred years, but it was treated more as a theoretical curiosity than a practically applicable method of inference with the rise of Fisherian and Neyman-Pearson schools of statistical inference. It was only with the Markov chain Monte Carlo (MCMC) revolution in the late twentieth century that Bayesian statistics began to be seen as a method one could actually use rather than just talk about.

The central idea of MCMC is to construct a Markov chain on the model's parameter space that converges in distribution to the posterior distribution of the model we are interested in. While it is easy to construct a Markov chain that is guaranteed to converge to the target distribution eventually, it has always been much harder to guarantee *quick* convergence and a vast literature exists exploring the various way to construct and improve these chains. One method of constructing an appropriate Markov chain is called data augmentation or more literally state-space expansion. This method works by creating additional parameters for the model, often called missing data or augmented data, but we can think of them as latent variables. To complete the data augmentation method we must construct a Markov chain on the larger parameter space. In many problems there exists natural missing data and the so called data augmentation algorithm represents a drastic speed up compared to any easily implementable Markov chain that lives in the original parameter space.

Data augmentation algorithms are still often plagued with slow convergence and a large literature developed around speeding of these algorithms. A relatively new method in this literature, called interweaving, uses two or more data augmentations and "weaves" them together inside a larger Markov chain on the expanded state-space. Chapter 2 of this dissertation applies this method to a class of time series models called dynamic linear models (DLMs). These models are linear, Gaussian state-space models MCMC algorithms constructed to compute their posterior can often be slow to converge. In order to apply the ideas of interweaving to these models I construct new data augmentations and stumble across a limitation of interweaving along the way.

1.2 Latent representation of group and treatment means

As with all new methods, there were some initial limitations to how the interweaving methods could be applied to DLMs. Chapter 3 provides an example of how to overcome these methods in the context of analyzing an economic experiment. This experiment consists of several treatments, each with about several replications, and each replication consists of 35 periods. A hierarchical DLM for the response variable is natural in this setting. I construct such a model for a single treatment of the experiment. The entire treatment has a mean that evolves over time and each replication of the treatment has a deviation from that mean that independently evolves over time. This allows us to think about treatment and replication level evolutions separately while still allowing for shrinkage between the replication level means.

In order to use the interweaving algorithms it took a little creativity in order to apply them. The model I construct does not have a square observation level matrix, F_t, so the interweaving methods I constructed in Chapter 2 do not directly apply. Instead of augmenting F_t, which is one method of getting them to work, I instead applied the methods of Chapter 2 to the model conditional on one of the parameters being fixed. Then everything can be put together in a larger Gibbs sampler which alternates between drawing the fixed parameter and drawing the the rest of the parameters through the interweaving steps.

1.3 Modeling treatment effects using latent variables

A crucial area in econometrics is causal inference and, in particular, program evaluation. Public policy programs are implemented every day without random controls and it is challenging to evaluate their consequences. In the simplest cast most programs allow any eligible individual to participate. This causes problems for trying to evaluate the efficacy of the program because individuals who choose to participate are often systematically different from individuals who choose not to participate, and these differences are usually at least partially unobservable.

A modeling language for causal inference has been developed in the social sciences for dealing with such problems, centered on the notion of a potential outcome. We think of each individual as having two potential outcomes – one if they participated in the program and one if they did not. The outcomes could be any response variable of interest – income, education, nutrition, etc. One of these outcomes we observe directly while the other is purely hypothetical, but we need to learn about this hypothetical outcome in order to learn about whether and how much the treatment improved or harmed the individual's situation. The basic idea, then, is to model the relationship between the observed outcome and the missing hypothetical, often called the missing counterfactual.

One approach to causal inference in this framework is called partial identification. The idea is to construct a data model that is fully parameterized so that the parameters driving the missing counterfactual are unidentified. Then, relate the unidentified parameters back to the identified parameters by bounding them or some function of them. Estimates of identified parameters then allow us to bound unidentified parameters and, more important, treatment effects – i.e. the difference between what would happen to an individual if they were on the program and what would happen to them if they were not on the program.

Often it is difficult to construct complicated models and perform partial identification in frequentist settings due to the difficulty understanding the variation in set estimators in order to construct confidence intervals. In the Bayesian context, computing posteriors in partially identified models is fairly straightforward. There is some difficulty with MCMC for parameters which are unidentified in the likelihood, but these are often surmountable.

This is the subject of Chapter 4 – an extension to Bayesian partial identification methods that only forces a particular constraint to hold some fraction of the time. Rather, each constraint holds with some probability which can be adjusted in order to represent how plausible we think it is. In order to construct priors capturing this notion in a hierarchical setting I ultimately have to resort to creating latent variables which determine the distribution of certain probabilities. The approach is then applied to the effect of the National School Lunch Program on whether or not a child from an income eligible household is food secure.

Throughout this dissertation, the common thread is using latent variables to construct better models and improve computation. This theme appears over and over again – in the context of data augmentation algorithms which are emphasized in Chapter 2 but are used in all three chapters, and in the context of constructing appropriate models in Chapters 3 and 4.

CHAPTER 2. INTERWEAVING MARKOV CHAIN MONTE CARLO STRATEGIES FOR EFFICIENT ESTIMATION OF DYNAMIC LINEAR MODELS

A paper under revision for *The Journal of Computational and Graphical Statistics*

Abstract

In dynamic linear models (DLMs) with unknown fixed parameters, a standard Markov chain Monte Carlo (MCMC) sampling strategy is to alternate sampling of latent states conditional on fixed parameters and sampling of fixed parameters conditional on latent states. In some regions of the parameter space, this standard data augmentation (DA) algorithm can be inefficient. To improve efficiency, we seek to employ the interweaving strategies of Yu and Meng (2011) that combine separate DAs by weaving them together. For this, we introduce a number of novel alternative DAs for a general class of DLMs: the scaled errors, wrongly-scaled errors, and wrongly-scaled disturbances. With the latent states and the less commonly used scaled disturbances, this yields five unique DAs to employ in MCMC algorithms. Each DA implies a unique MCMC sampling strategy and they can be combined into interweaving or alternating strategies that improve MCMC efficiency. We assess the strategies using the local level DLM and demonstrate that several strategies improve efficiency relative to the standard approach, the most efficient being either interweaving or alternating the scaled errors and scaled disturbances.

2.1 Introduction

The Data Augmentation (DA) algorithm of Tanner and Wong (1987) and the closely related Expectation Maximization (EM) algorithm of Dempster et al. (1977) have become widely used strategies for computing posterior distributions and maximum likelihood estimates, with a long history of using ideas from the EM literature to inform the construction of DA algorithms and vice versa (Meng and Van Dyk, 1997; Van Dyk and Meng, 2010). While useful, DA and EM algorithms often suffer from slow convergence s a large literature has grown up around various possible improvements to both algorithms (Meng and Van Dyk, 1997, 1999; Liu and Wu, 1999; Hobert and Marchev, 2008; Yu and Meng, 2011), though much of the work on constructing improved algorithms has focused on hierarchical models (Gelfand et al., 1995; Roberts and Sahu, 1997; Meng and Van Dyk, 1998; Van Dyk and Meng, 2001; Bernardo et al., 2003; Papaspiliopoulos et al., 2007; Papaspiliopoulos and Roberts, 2008). Despite some similarities with some hierarchical models, relatively little attention has been paid to time series models. Exceptions include (Pitt and Shephard, 1999; Frühwirth-Schnatter and Sögner, 2003; Frühwirth-Schnatter and Wagner, 2006) in the DA literature and(Van Dyk and Tang, 2003) in the EM literature.

We seek to improve DA schemes in dynamic linear models (DLMs), i.e. linear Gaussian state-space models. The standard DA scheme uses the latent states and alternates between drawing from the full conditional distributions of the latent states and the model parameters (Frühwirth-Schnatter, 1994; Carter and Kohn, 1994). The existing literature on improving DA algorithms in time series models tends to focus on non-Gaussian state-space models — particularly the stochastic volatility model and models based on it (Shephard, 1996; Frühwirth-Schnatter and Sögner, 2003; Roberts et al., 2004; Bos and Shephard, 2006; Strickland et al., 2008; Frühwirth-Schnatter and Sögner, 2008; Kastner and Frühwirth-Schnatter, 2014), but a few work with the class of DLMs we consider (Frühwirth-Schnatter, 2004). One recent development in the DA literature is an “interweaving” strategy for using two separate DAs in a single algorithm (Yu and Meng, 2011). This strategy draws on the strengths of both underlying

DA algorithms in order to construct an MCMC algorithm which is at least as efficient as the worst of the two DA algorithms and typically at least as efficient as the best. We implement interweaving algorithms in a general class of DLMs and in order to do so we introduce several new DAs for this class of models. We also show under some assumptions that no *practical* sufficient augmentation (centered augmentation) exists for the DLM, which restricts the sort of interweaving algorithms we can construct. Using the local level model, we fit the model to simulated data using a variety of the MCMC strategies we discuss in order to assess their relative performance.

The rest of the paper is organized as follows. In Section 2.2 we review the DA literature while in Section 2.3 we introduce the dynamic linear model and discuss the subclass of DLMs we consider. Section 2.4 explores several possible DAs for our class of DLMs and shows that any sufficient augmentation is likely to be difficult to use. Section 2.5 discusses the various MCMC strategies available for the DLM while Section 2.6 applies these algorithms to the local level model. Finally, Section 2.7 discusses these results and suggests directions for further research. In addition, several additional sections serve to supplement the main body of the paper. Section 2.A contains a derivation of the marginal model for the data in a class of DLMs, Section 2.B contains a proof of Lemma 1, while Section 2.C explicitly constructs the wrongly-scaled DAs. Section 2.D shows the full conditional distributions of each block of parameters in the DLM under a variety of parameterizations while Section 2.E shows how to draw from the full conditional of the latent states using the mixed Cholesky factorization algorithm. Next, Section 2.F shows how to use some of the DAs we introduce when F_t is not invertible while Sections 2.G and 2.H show how to draw from some of the difficult full conditional distributions that appear under certain paramaterizations. Section 2.I shows that certain classes of interweaving algorithms are equivalent for the DLM and Section 2.J introduces another class of interweaving algorithms that is also equivalent to certain algorithms discussed in the main body. Finally Section 2.K uses the behavior of the posterior to help explain how the various MCMC algorithms perform while Section 2.L contains additional plots to supplement those covered Section 2.6.3.

2.2 Variations of data augmentation

Suppose $p(\phi|y)$ is a probability density, for example the posterior distribution of some parameter ϕ given data y. Then a DA algorithm adds a DA θ with joint distribution $p(\phi, \theta|y)$ such that $\int_\Theta p(\phi, \theta|y)d\theta = p(\phi|y)$. The DA algorithm is a Gibbs sampler for (ϕ, θ), except we focus attention on the marginal chain for ϕ. In this DA algorithm, the $k + 1$'st state of ϕ is obtained from the k'th state as follows (we implicitly condition on the data y in all algorithms and only superscript the previous and new draws of the model parameters of interest):

Algorithm: DA. *Data Augmentation*

$$[\theta|\phi^{(k)}] \quad \rightarrow \quad [\phi^{(k+1)}|\theta]$$

where $[\theta|\phi^{(k)}]$ means a draw of θ from $p(\theta|\phi^{(k)}, y)$ and $[\phi^{(k+1)}|\theta]$ means a draw from $p(\phi|\theta, y)$. The DA need not be interesting in any scientific sense — it can be viewed purely as a computational construct.

2.2.1 Reparameterization and alternating DAs

One well known method of improving mixing and convergence in MCMC samplers as well as convergence in EM algorithms is reparameterization of the model (see Papaspiliopoulos et al. (2007) and references therein). The DA θ is called a *sufficient augmentation* (SA) for the model parameter ϕ if $p(y|\theta, \phi) = p(y|\theta)$. Similarly θ is called an *ancillary augmentation* (AA) for ϕ if $p(\theta|\phi) = p(\theta)$. An SA is sometimes called a centered augmentation or centered parameterization in the literature while an AA is sometimes called a non-centered augmentation or non-centered parameterization. Like Yu and Meng (2011) we prefer the SA and AA terminology because it suggests a connection with Basu's theorem (Basu, 1955), which we will return to in Section 2.2.2.

A key reason behind the emphasis on SAs and AAs is that typically when the DA algorithm based on the SA has nice mixing and convergence properties, the DA algorithm based on the AA has poor mixing and convergence properties and vice-versa. This property suggests combining the two such DA algorithms to construct an improved sampler. One intuitive approach is to alternate between the two augmentations within a Gibbs sampler (Papaspiliopoulos et al., 2007). Suppose we have a second distinct DA γ such that $\int_\Gamma p(\phi,\gamma|y)d\gamma = p(\phi|y)$, then the alternating algorithm for sampling from $p(\phi|y)$ is as follows:

Algorithm: Alt. *Alternating Algorithm*

$$[\theta|\phi^{(k)}] \quad \rightarrow \quad [\phi|\theta] \quad \rightarrow \quad [\gamma|\phi] \quad \rightarrow \quad [\phi^{(k+1)}|\gamma].$$

One iteration of the alternating algorithm consists of one iteration of the DA algorithm based on θ to obtain an intermediate value of ϕ, followed by one iteration of the DA algorithm based on γ.

When ϕ and θ are highly dependent in their joint posterior, the draws from $p(\theta|\phi,y)$ and $p(\phi|\theta,y)$ will hardly move the chain in Algorithm DA, resulting in high autocorrelation. In an alternating algorithm, there are essentially two chances to substantially move the chain – one using θ and the other using γ. Often at least one of θ and γ has low dependence with ϕ, resulting in a chain that mixes well.

2.2.2 Interweaving: an alternative to alternating

Another option is to *interweave* the two DAs together (Yu and Meng, 2011). A global interweaving strategy (GIS) is an MCMC algorithm that obtains $\phi^{(k+1)}$ from $\phi^{(k)}$ as follows:

Algorithm: GIS. *Global Interweaving Strategy*

$$[\theta|\phi^{(k)}] \quad \rightarrow \quad [\gamma|\theta] \quad \rightarrow \quad [\phi^{(k+1)}|\gamma].$$

The GIS algorithm obtains the next iteration of the parameter ϕ in three steps: 1) draw θ conditional on $\phi^{(k)}$, 2) draw γ conditional on θ, and 3) draw $\phi^{(k+1)}$ conditional on γ. This looks similar to the usual DA algorithm except a second DA is “weaved” in between the draw of the first DA and of the parameter.

The second step of the GIS algorithm is often accomplished by sampling $\phi|\theta$ and then $\gamma|\theta,\phi$. If we expand this out, then the GIS algorithm becomes:

Algorithm: eGIS. *Expanded GIS*

$$[\theta|\phi^{(k)}] \quad \rightarrow \quad [\phi|\theta] \quad \rightarrow \quad [\gamma|\theta,\phi] \quad \rightarrow \quad [\phi^{(k+1)}|\gamma].$$

In addition, γ and θ are often, but not always, one-to-one transformations of each other conditional on (ϕ,y), i.e. $\gamma = M(\theta;\phi,y)$ where $M(.;\phi,y)$ is a one-to-one function, and thus $[\gamma|\theta,\phi]$

is deterministic. The key difference between Algorithm GIS and Algorithm Alt can be seen in step three of Algorithm eGIS: instead of drawing from $p(\gamma|\phi, y)$, the GIS algorithm draws from $p(\gamma|\theta, \phi, y)$, connecting the two DAs together while the alternating algorithm keeps them separate.

Yu and Meng (2011) call a GIS approach where one of the DAs is an SA and the other is an AA an ancillary sufficient interweaving strategy (ASIS). They show that the GIS algorithm has a geometric rate of convergence no worse than the worst of the two underlying DA algorithms and in some cases better than the the corresponding alternating algorithm. In particular, their Theorem 1 suggests that the weaker the dependence between the two DAs in the posterior, the more efficient the GIS algorithm. With *a posteriori* independent DAs, the GIS algorithm obtains iid draws from ϕ's posterior. This helps motivate their focus on ASIS and the choice of terminology — conditional on the model parameter, an SA and an AA are independent under the conditions of Basu's theorem (Basu, 1955), which suggests that the dependence between the two DAs will be limited in the posterior. In fact, when the prior on ϕ is nice in some sense, Yu and Meng (2011) show that the ASIS algorithm is the same as the optimal parameter expanded data augmentation (PX-DA) algorithm (Liu and Wu, 1999), which is closely related to marginal and conditional augmentation (Meng and Van Dyk, 1999; Hobert and Marchev, 2008).

In addition to the GIS, it is possible to define a componentwise interweaving strategy (CIS) that interweaves within specific steps of a Gibbs sampler as well. A CIS algorithm for $\phi = (\phi_1, \phi_2)$ essentially employs interweaving for each block of ϕ separately, e.g.

Algorithm: CIS. *Componentwise Interweaving Strategy*

$$\begin{array}{lllllll}
[\theta_1|\phi_1^{(k)}, \phi_2^{(k)}] & \rightarrow & [\gamma_1|\phi_2^{(k)}, \theta_1] & \rightarrow & [\phi_1^{(k+1)}|\phi_2^{(k)}, \gamma_1] & \rightarrow \\
[\theta_2|\phi_1^{(k+1)}, \phi_2^{(k)}, \gamma_1] & \rightarrow & [\gamma_2|\phi_1^{(k+1)}, \theta_2] & \rightarrow & [\phi_2^{(k+1)}|\phi_1^{(k+1)}, \gamma_2] &
\end{array}$$

where θ_i and γ_i are distinct data augmentations for $i = 1, 2$, but potentially $\gamma_1 = \theta_2$ or $\gamma_2 = \theta_1$. The first row draws ϕ_1 conditional on ϕ_2 using interweaving in a Gibbs step, while the second row does the same for ϕ_2 conditional on ϕ_1. The algorithm can easily be extended to greater than two blocks within ϕ. The main attraction of CIS is that it is often easier to find an AA–SA pair of DAs for ϕ_1 conditional on ϕ_2 and another pair for ϕ_2 conditional on ϕ_1 than it is to find and AA–SA pair for $\phi = (\phi_1, \phi_2)$ jointly.

2.3 Dynamic linear models

The general dynamic linear model is well studied (West and Harrison, 1999; Petris et al., 2009; Prado and West, 2010) and is defined as

$$\begin{aligned} y_t &= F_t\theta_t + v_t & v_t &\overset{ind}{\sim} N_k(0, V_t) & &\text{(observation equation)}\\ \theta_t &= G_t\theta_{t-1} + w_t & w_t &\overset{ind}{\sim} N_p(0, W_t) & &\text{(system equation)} \end{aligned}$$

where $N_d(\mu, \Sigma)$ is a d-dimensional multivariate normal distribution with mean μ and covariance Σ and the observation errors, v_t for $t = 1, 2, \cdots, T$, and system disturbances, w_t for $t = 1, 2, \cdots, T$, are independent. The observed data are $y \equiv y_{1:T} \equiv (y_1', y_2', \cdots, y_T')'$ while the latent states are $\theta \equiv \theta_{0:T} \equiv (\theta_0', \theta_1', \cdots, \theta_T')'$. For each $t = 1, 2, \cdots, T$, F_t is a $k \times p$ matrix and G_t is a $p \times p$ matrix. Let ϕ denote the vector of unknown parameters in the model. Then possibly F_t, G_t, V_t, and W_t are all functions of ϕ for $t = 1, 2, \cdots, T$.

The subclass of DLMs we will focus on sets $V_t = V$ and $W_t = W$ and treats F_t and G_t as known for all t. Our results can be extended when V_t or W_t is time-varying or when F_t or G_t depend on unknown parameters, but we ignore those cases for simplicity. As a result $\phi = (V, W)$ is our unknown parameter and we can write the model as

$$y_t|\theta, V, W \overset{ind}{\sim} N_k(F_t\theta_t, V) \qquad\qquad \theta_t|\theta_{0:t-1}, V, W \sim N_p(G_t\theta_{t-1}, W) \tag{2.1}$$

for $t = 1, 2, \cdots T$. We use the standard conditionally conjugate priors, that is θ_0, V, and W independent with $\theta_0 \sim N_p(m_0, C_0)$, $V \sim IW(\Lambda_V, \lambda_V)$ and $W \sim IW(\Lambda_W, \lambda_W)$ where m_0, C_0, Λ_V, λ_V, Λ_W, and λ_W are known hyperparameters and $IW(\Lambda, \lambda)$ denotes the inverse Wishart distribution with degrees of freedom λ and positive definite scale matrix Λ.

The latent states can be integrated out to obtain the marginal model for the y:

$$y|V, W \overset{ind}{\sim} N_{Tk}(D\tilde{m}, \tilde{V} + \tilde{W} + \tilde{C}). \tag{2.2}$$

where $\tilde{V} = I_T \otimes V$, D is block diagonal with elements $D_1, \ldots, D_T$,

$$\tilde{W}_{Tk\times Tk} = \begin{bmatrix} K_1'F_1' & K_2'F_2' & \cdots K_T'F_T' \end{bmatrix}' W \begin{bmatrix} K_1'F_1' & K_2'F_2' & \cdots K_T'F_T' \end{bmatrix},$$
$$\tilde{C}_{Tk\times Tk} = \begin{bmatrix} H_1'F_1' & H_2'F_2' & \cdots H_T'F_T' \end{bmatrix}' C_0 \begin{bmatrix} H_1'F_1' & H_2'F_2' & \cdots H_T'F_T' \end{bmatrix},$$

$\tilde{m}_{Tp\times 1} = (m_0', m_0', \cdots m_0')'$, and D_t, K_t, and H_t are functions of the F_t's and G_t's for $t = 1, 2, \ldots, T$. A derivation of this distribution is in Section 2.A.

2.4 Augmenting the DLM

The standard definition of the DLM includes the standard DA used in estimation of the DLM, θ. We now introduce one data augmentation that is known, the scaled disturbances, and three other novel augmentations: scaled errors, wrongly-scaled disturbances, and wrongly-scaled errors. The primary purpose of these augmentations is for use in interweaving algorithms, but each DA will also implicitly define a DA algorithm.

A natural way to create new DAs is by reparameterizing old DAs. Papaspiliopoulos et al. (2007) note that typically the standard augmentation results in an SA for the parameter ϕ. All that would be necessary for an ASIS algorithm, then, is to construct an AA for ϕ. We immediately run into a problem because the standard DA for a DLM is θ but in equation (2.4) V is in the observation equation so that θ is not an SA for (V, W) while W is in the system equation so that θ is not an AA for (V, W) either. In order to find an SA we need to somehow move V from the observation equation to the system equation and similarly to find an AA we need to somehow move W from the system equation to the observation equation.

As Papaspiliopoulos et al. (2007) suggests, we can construct a pivotal quantity in order to find an ancillary augmentation, e.g. by appropriately centering and scaling a random variable. Notice from equation (2.4) that if we hold V constant then θ is an SA for W conditional on V, i.e. for $W|V$. Similarly θ is an AA for $V|W$. This suggests that if we center and scale θ_t by W appropriately for all t we will have an ancillary augmentation for V and W jointly, thus creating the *scaled disturbances* (SDs).

2.4.1 The scaled disturbances

To define the scaled disturbances let L_W denote the Cholesky decomposition of W, i.e. the lower triangle matrix L_W such that $L_W L_W' = W$. Then we will define the scaled disturbances $\gamma \equiv \gamma_{0:T} \equiv (\gamma_0', \gamma_1', \cdots, \gamma_T')'$ by $\gamma_0 = \theta_0$ and $\gamma_t = L_W^{-1}(\theta_t - G_t\theta_{t-1})$ for $t = 1, 2, \cdots, T$. There are actually $p!$ different versions of the scaled disturbances depending on how we order the elements of θ_t (Meng and Van Dyk, 1998) but we make no attempt to determine which ordering should be used. The reverse transformation is defined recursively by $\theta_0(\gamma, L_W) = \gamma_0$ and $\theta_t(\gamma, L_W) = L_W\gamma_t + G_t\theta_{t-1}(\gamma, L_W)$ for $t = 1, 2, \cdots, T$. Under the scaled disturbance parameterization we can write the model as

$$y_t|\gamma, V, W \overset{ind}{\sim} N_k\left(F_t\theta_t(\gamma, L_W), V\right), \qquad \gamma_t \overset{iid}{\sim} N_p(0, I_p) \tag{2.3}$$

for $t = 1, 2, \cdots, T$ where I_p is the $p \times p$ identity matrix. Neither V nor W are in the system equation so the scaled disturbances are an AA for (V, W). The SDs are well known — the disturbance smoother of Koopman (1993) finds the conditional posterior of the scaled disturbances given the parameter and Frühwirth-Schnatter (2004) uses the SDs in a dynamic regression model with stationary regression coefficients.

2.4.2 The scaled errors

The scaled disturbances immediately suggest our first novel augmentation called the *scaled errors* (SEs), i.e. $v_t = y_t - F_t\theta_t$ appropriately scaled by V. Let L_V denote the Cholesky decomposition of V so that $L_V L_V' = V$, then we can define a version of the scaled errors as $\psi_t = L_V^{-1}(y_t - F_t\theta_t)$ for $t = 1, 2, \cdots, T$ and $\psi_0 = \theta_0$. This time there are $k!$ versions of the scaled errors depending on how y_t is ordered.

Assuming F_t is invertible for all t (see Section and Simpson (2014) for examples of how to relax this restriction), then $\theta_t = F_t^{-1}(y_t - L_V\psi_t)$ for $t = 1, 2, \cdots, T$ while $\theta_0 = \psi_0$. Define $\mu_1 = L_V\psi_1 + F_1G_1\psi_0$ and $\mu_t = L_V\psi_t + F_tG_tF_{t-1}^{-1}(y_{t-1} - L_V\psi_{t-1})$ for $t = 2, 3, \cdots, T$. Then the scaled error parameterization is

$$y_t|V, W, \psi, y_{1:t-1} \sim N_p(\mu_t, F_tWF_t'), \qquad \psi_t \overset{iid}{\sim} N_p(0, I_k)$$

for $t = 1, 2, \cdots, T$ where I_k is the $k \times k$ identity matrix. Since neither V nor W are in the system equation, we immediately see that the scaled errors are an AA for (V, W). However, both V and W are in the observation equation so that ψ is not an SA for $V|W$ nor for $W|V$.

2.4.3 The "wrongly-scaled" DAs

Two other novel augmentations can be obtained by scaling the SD and SE by the "wrong" variance so long as F_t is square, i.e. that V and W have the same dimension. Define $\tilde{\gamma}_t = L_V^{-1}(\theta_t - G_t\theta_{t-1})$ and $\tilde{\psi}_t = L_W^{-1}(y_t - \theta_t)$ for $t = 1, 2, \cdots, T$ and $\tilde{\psi}_0 = \tilde{\gamma}_0 = \theta_0$. Then the *wrongly-scaled disturbances* (WSDs) are $\tilde{\gamma} \equiv \tilde{\gamma}_{0:T} \equiv (\tilde{\gamma}_0', \tilde{\gamma}_1', \cdots, \tilde{\gamma}_T')'$ and the *wrongly-scaled errors* (WSEs) are $\tilde{\psi} \equiv \tilde{\psi}_{0:T} \equiv (\tilde{\psi}_0', \tilde{\psi}_1', \cdots, \tilde{\psi}_T')'$.

We can write the model in terms of $\tilde{\gamma}$ as

$$y_t|\tilde{\gamma}, V, W \overset{ind}{\sim} N_p\left(F_t\theta_t(\tilde{\gamma}, L_V), V\right), \qquad \tilde{\gamma}_t \overset{ind}{\sim} N_p(0, L_V^{-1}W(L_V^{-1})')$$

for $t = 1, 2, \cdots, T$ where $\theta_t(\tilde{\gamma}, L_V)$ denotes the transformation from $\tilde{\gamma}$ to θ defined by the wrongly-scaled disturbances. Since L_V is the Cholesky decomposition of V, the observation equation does not contain W, so $\tilde{\gamma}$ is an SA for $W|V$. Since W and L_V are both in the system equation, $\tilde{\gamma}$ is not an AA for $V|W$ nor for $W|V$.

Similarly, we can write the model in terms of $\tilde{\psi}$ as

$$y_t|V, W, \tilde{\psi}, y_{1:t-1} \sim N_p(\tilde{\mu}_t, F_tWF_t'), \qquad \tilde{\psi}_t \overset{iid}{\sim} N_p(0, L_W^{-1}V(L_W^{-1})')$$

for $t = 1, 2, \cdots, T$ where we define $\tilde{\mu}_1 = L_W\tilde{\psi}_1 - F_1G_1\tilde{\psi}_0$ and for $t = 2, 3, \cdots, T$ $\tilde{\mu}_t = L_W\tilde{\psi}_t - F_tG_tF_{t-1}^{-1}(y_{t-1} - L_W\tilde{\psi}_{t-1})$. Since $\tilde{\mu}_t$ only depends on W and not on V, V is absent from the observation equation and thus $\tilde{\psi}$ is an SA for $V|W$. Once again, since both W and V are in the system equation $\tilde{\psi}$ is not an AA for either V or W.

2.4.4 The elusive search for a sufficient augmentation

Next we would like to find a sufficient augmentation in order to construct an ASIS for sampling from the posterior distribution. The following lemma suggests that this may be difficult if not impossible.

Lemma 1. *Suppose η is an SA for the DLM such that conditional on ϕ, η and y are jointly normally distributed, that is*

$$\begin{bmatrix}\eta\\ y\end{bmatrix}\Bigg|\phi \sim N\left(\begin{bmatrix}\alpha_\eta\\ D\tilde{m}\end{bmatrix},\begin{bmatrix}\Omega_\eta & \Omega_{y,\eta}'\\ \Omega_{y,\eta} & \tilde{V}+\tilde{W}+\tilde{C}\end{bmatrix}\right).$$

Let $A = \Omega_{y,\eta}'\Omega_\eta^{-1}$ and $\Sigma = \tilde{V}+\tilde{W}+\tilde{C}-A\Omega_\eta A'$. Then A, Σ, and α_η are constants with respect to ϕ and if $A'A$ is invertible, then

$$\begin{aligned}p(\phi|\eta,y) \propto & p(y|\eta,\phi)p(\eta|\phi)p(\phi) \propto p(\eta|\phi)p(\phi)\\ &\propto p(\phi)|(A'A)^{-1}A'(\tilde{V}+\tilde{W}+\tilde{C}-\Sigma)A(A'A)^{-1}|^{-1/2}\\ &\quad\times\exp\left[-\frac{1}{2}(\eta-\alpha_\eta)'[(A'A)^{-1}A'(\tilde{V}+\tilde{W}+\tilde{C}-\Sigma)A(A'A)^{-1}]^{-1}(\eta-\alpha_\eta)\right].\end{aligned}$$

A proof of this lemma is in Section 2.B. The posterior density we wish to sample from comes from equation (2.5) and is similar to $p(\phi|\eta,y)$ except less complicated. So what this lemma shows is that in order to use an SA in a GIS algorithm, we probably need to obtain draws from a density that is just as hard to sample from as the posterior density we are already trying to approximate. This does not mean that the full conditional posterior density of the parameters given a SA has to be difficult to draw from. Rather it means that if we can draw from that density we could probably draw from the target posterior — perhaps using the same technology. This result brings to mind Van Dyk and Meng (2001)'s contention that there is an art to constructing data augmentation algorithms. Our goal is to find an MCMC algorithm that has nice convergence and mixing properties and is also easy to implement, and this second criteria is much more difficult to quantify.

2.5 MCMC strategies for the DLM

This section briefly discusses how to construct various MCMC algorithms for approximating the posterior distribution of the DLM. We focus on *what* to do, not *why*. Derivations of the relevant full conditional distributions are available in Section 2.C. We occasionally come across a full conditional density that is difficult to sample from — the details about why this happens and how to overcome it are in the Sectons 2.G and 2.H.

2.5.1 Base algorithms

Using any of the DAs introduced in Section 2.4, we can construct several DA algorithms which we call *base algorithms* to distinguish them from the alternating and interweaving algorithms we will construct later. We will call the standard DA algorithm (Frühwirth-Schnatter, 1994; Carter and Kohn, 1994) using θ the *state sampler*. In order to construct this sampler, we need to draw from two densities — $p(\theta|V,W,y)$ and $p(V,W|\theta,y)$. In their conditional posterior, V and W are independent with

$$V|\theta,y \sim IW\left(\Lambda_V + \sum_{t=1}^T v_t v_t', \lambda_V + T\right), \qquad W|\theta,y \sim IW\left(\Lambda_W + \sum_{t=1}^T w_t w_t', \lambda_W + T\right),$$

where $v_t = y_t - F_t\theta_t$, and $w_t = \theta_t - G_t\theta_{t-1}$.

The density $p(\theta|V,W,y)$ is multivariate normal and any algorithm that obtains a random draw from it is called a simulation smoother in the literature. The most commonly used smoother, FFBS, uses the Kalman filter (Frühwirth-Schnatter, 1994; Carter and Kohn, 1994), but other examples are Koopman (1993) and De Jong and Shephard (1995). The smoothers introduced in McCausland et al. (2011) and Rue (2001), dubbed “all without a loop” smoothers by Kastner and Frühwirth-Schnatter (2014) exploit the tridiagonal structure of θ’s precision matrix in order to speed up the computation of its Cholesky factor. The method of Rue (2001) computes this Cholesky fact and samples from the density in separate steps, and is called the Cholesky factor algorithm (CFA). On the other hand McCausland et al. (2011) mixes these two steps together in a backward sampling structure, so we call it the mixed Cholesky factor algorithm (MCFA). We use the MCFA for drawing θ and include the details of the algorithm in the context of the DLM in Section 2.E.

Putting the pieces together, the state sampler is the following DA algorithm:

Algorithm: State. *State Sampler*

$$[\theta|V^{(k)},W^{(k)}] \quad \rightarrow \quad [V^{(k+1)},W^{(k+1)}|\theta]$$

where the first step uses the MCFA and the second step is the independent inverse Wishart draws defined above. As we will show in Section 2.6, the Markov chain constructed using the state sampler can mix poorly in some regions of the parameter space. For example, in a dynamic regression through the origin with stationary regression coefficient, if the variance of the latent states is too small relative to the variance of the data, mixing will be poor for W (Frühwirth-Schnatter, 2004).

Next, we can use γ in order to construct a second DA algorithm called the *scaled disturbance sampler*. In the smoothing step we need to obtain a draw from $p(\gamma|V, W, y)$. This density is also Gaussian but has a more complex precision matrix, so in order to obtain a draw from it we use the MCFA to draw from $p(\theta|V, W, y)$, then transform from θ to γ. The density $p(V, W|\gamma, y)$ is rather complicated and does not appear easy to draw from, but it is easy to show that $V|W, \gamma, y \sim IW\left(\Lambda_V + \sum_{t=1}^{T} v_t v_t', \lambda_V + T\right)$ where $v_t = y_t - F_t\theta_t$ and θ_t is a function of γ and W. However, it is not easy to draw from $p(W|\gamma, y)$ so we abandon drawing V and W jointly. The density $p(W|V, \gamma, y)$ is simpler and, at least in the local level model, can be sampled from with tolerable efficiency. As a result Algorithm SD, the scaled disturbance sampler, has three steps instead of the usual two.

Algorithm: SD. *Scaled Disturbance Sampler*

$$[\theta|V^{(k)}, W^{(k)}] \quad \rightarrow \quad [V^{(k+1)}|W^{(k)}, \theta] \quad \rightarrow \quad [\gamma|V^{(k+1)}, W^{(k)}, \theta] \quad \rightarrow \quad [W^{(k+1)}|V^{(k+1)}, \gamma]$$

The first and second steps are the draws as in Algorithm State while the third step is a transformation from θ to γ. The last step is the difficult one. When W is a scalar a tolerably efficient rejection sampling algorithm can be constructed, but in models where W is a matrix it is not clear whether drawing from $p(W|V, \gamma, y)$ can be accomplished efficiently. Section 2.G has more detail as well as an algorithm for drawing from $P(W|V, \gamma, y)$ in the local level model when V and W are scalars.

The DA algorithm based on the scaled errors is called the *scaled error sampler* (Algorithm SE) and is similar to the scaled disturbance sampler with a couple of key differences. First, the simulation smoothing step in the scaled error sampler can be accomplished directly with the MCFA because the precision matrix of the conditional posterior of ψ retains the necessary tridiagonal structure. Second, the full conditional distribution of W is the familiar inverse Wishart density and the full conditional of V is the complicated density. The density of $V|W, \psi, y$ is in the same class as that of $W|V, \gamma, y$. In fact there is a strong symmetry here — the joint conditional posterior of (V, W) given γ is from the same family of densities as that of (W, V) given ψ so that V and W essentially switch places when we condition on the scaled errors instead of the scaled disturbances.

Algorithm: SE. *Scaled Error Sampler*

$$[\psi|V^{(k)}, W^{(k)}] \quad \rightarrow \quad [V^{(k+1)}|W^{(k)}, \psi] \quad \rightarrow \quad [W^{(k+1)}|V^{(k+1)}, \psi]$$

The first step uses the MCFA directly for ψ while the third step is the same inverse Wishart draw for W as in Algorithm State. The second step contains the difficult draw.

In addition, we can construct DA algorithms based on the wrongly-scaled disturbances and errors – the *wrongly-scaled disturbance sampler* and the *wrongly-scaled error sampler*. In Section 2.6 we show that these samplers perform poorly, so the construction of these algorithms is left to Section 2.C, though the wrongly-scaled DAs will ultimately be helpful in the construction of certain interweaving algorithms in Section 2.5.4.

2.5.2 Alternating algorithms

Using the full conditional distributions defined in Section 2.5.1, we can construct several alternating algorithms based on any two of the DA algorithms. The algorithms have the form of Algorithm Alt on page 9. For example, the *State-SD alternating sampler* which alternates between the states and the scaled disturbances, obtains the $k+1$'st iteration of (V, W) from the k'th as follows:

$$\begin{aligned}
&[\theta|V^{(k)}, W^{(k)}] \rightarrow [V^{(k+0.5)}, W^{(k+0.5)}|\theta] \rightarrow \\
&[\gamma|V^{(k+0.5)}, W^{(k+0.5)}] \rightarrow [V^{(k+1)}|W^{(k+0.5)}, \gamma] \rightarrow [W^{(k+1)}|V^{(k+1)}, \gamma].
\end{aligned}$$

The first line is an iteration of the state sampler while the second line is an iteration of the scaled disturbance sampler. No work is necessary to link up the two iterations — we simply plug in the values of V and W obtained from the state sampler iteration into the draw of γ from step one of the scaled disturbance sampler iteration. Each other alternating algorithm is analogous and can be constructed without complication. The order in which the base algorithms are used within an alternating algorithm could in principle affect the convergence properties of the algorithm, but typically is not important.

The naming convention we use for these algorithms is to list each DA in the order in which they appear in the alternating sampler, separated by hyphens. We shorten the scaled disturbances to "SD", the scaled errors to "SE", and the wrongly-scaled version of each to "WSD" and "WSE" respectively. So for example, the alternating sampler which alternates between the scaled disturbances and the wrongly-scaled disturbances, in that order, we call *SD-WSD Alt*.

2.5.3 GIS algorithms

We can use the various DAs of Section 2.4 to construct interweaving algorithms as well. We will start with Algorithm eGIS on page 9. Given the full conditional distributions listed in Section 2.5.1, the only additional ingredients we need are the definitions of the various available DAs in order to perform the one-to-one transformations from any one DA to another. For example, in the *State-SD GIS sampler* we obtain $(V^{(k+1)}, W^{(k+1)})$ from $(V^{(k)}, W^{(k)})$ as follows:

$$
\begin{aligned}
&[\theta|V^{(k)}, W^{(k)}] \to [W^{(k+0.5)}, V^{(k+0.5)}|\theta] \to \\
&[\gamma|V^{(k+0.5)}, W^{(k+0.5)}, \theta] \to [V^{(k+1)}|W^{(k+0.5)}, \gamma] \to [W^{(k+1)}|V^{(k+1)}, \gamma].
\end{aligned}
$$

In the first step of the second line we transform θ to γ by means of the defining equations for γ: $\gamma_0 = \theta_0$ and $\gamma_t = L_W^{-1}(\theta_t - G_t\theta_{t-1})$ for $t = 1, 2, \cdots, T$ where L_W is the Cholesky decomposition of W.

There are often some small improvements that can be made simply by thinking clearly about what the GIS algorithm is doing. For example in the above version of the State-SD GIS sampler, the draw of V in step two of line one and the draw of V in step two of line two are redundant — they come from the same distribution and only the last one is ever used in later steps. The resulting State-SD GIS sampler is as follows:

Algorithm: State-SD GIS. *State-Scaled Disturbance GIS Sampler*

$$
[\theta|V^{(k)}, W^{(k)}] \quad \to \quad [V^{(k+1)}, W^{(k+0.5)}|\theta] \quad \to \quad [\gamma|V^{(k+1)}, W^{(k+0.5)}, \theta] \quad \to \quad [W^{(k+1)}|V^{(k+1)}, \gamma].
$$

The first two steps are both steps of Algorithm State, the third step simply transforms from θ to γ, and the final step is a the difficult draw from Algorithm SD.

The naming convention for GIS algorithms is similar to that of alternating algorithms — DAs appear in the name in the order that they appear in the algorithm, separated by hyphens, e.g. a GIS algorithm based on the states, scaled disturbances, and scaled errors in that order would be called the *State-SD-SE GIS sampler*. There is no additional difficulty encountered by using a GIS with greater than two DAs and like alternating algorithms, the performance of GIS algorithms may depend on the order in which the DAs are used, but in our experience this tends to make no difference, which is consistent with what Yu and Meng (2011) report.

2.5.4 CIS algorithms

Next we consider CIS algorithms which have the form of Algorithm CIS on page 10. The advantage of using CIS is that it is sometimes possible to find an AA-SA pair of DAs for each part of the parameter vector even when no such pair of DAs exist for the entire vector. From Section 2.4, we know that the scaled disturbances and the wrongly-scaled disturbances form an AA-SA pair for $W|V$ while the scaled errors and the wrongly-scaled errors form an AA-SA pair for $V|W$. A CIS sampler based on these AA-SA pairs obtains $(V^{(k+1)}, W^{(k+1)})$ from $(V^{(k)}, W^{(k)})$ as follows:

$$[\psi|V^{(k)}, W^{(k)}] \to [V^{(k+0.5)}|W^{(k)}, \psi] \to [\tilde{\psi}|V^{(k+0.5)}, W^{(k)}, \psi] \to [V^{(k+1)}|W^{(k)}, \tilde{\psi}] \to$$
$$[\tilde{\gamma}|V^{(k+1)}, W^{(k)}, \tilde{\psi}] \to [W^{(k+0.5)}|V^{(k+1)}, \tilde{\gamma}] \to [\gamma|V^{(k+1)}, W^{(k+0.5)}, \tilde{\gamma}] \to [W^{(k+1)}|V^{(k+1)}, \gamma].$$

The first line is essentially a Gibbs step for drawing V that interweaves between ψ and $\tilde{\psi}$ while the second line is essentially a Gibbs step for drawing W that interweaves between γ and $\tilde{\gamma}$. In the second line we use the SA before the AA in order to minimize the number of transformations we have to make in every iteration.

Notice that each time one of the wrongly-scaled DAs appears in the CIS sampler, it would make no difference if the states were used instead because $p(V|W, \tilde{\psi}, y) = p(V|W, \theta, y)$ and $p(W|V, \tilde{\gamma}, y) = p(W|V, \theta, y)$, despite the fact that the states are not an SA for $V|W$. Using this we obtain a slightly different version of the CIS sampler in Algorithm CIS:

Algorithm: CIS. *Componentwise Interweaving Sampler*

$$[\psi|V^{(k)}, W^{(k)}] \to [V^{(k+0.5)}|W^{(k)}, \psi] \to [\psi|V^{(k+0.5)}, W^{(k)}, \theta] \to [V^{(k+1)}|W^{(k)}, \theta] \to$$
$$[W^{(k+0.5)}|V^{(k+1)}, \theta] \to [\gamma|V^{(k+1)}, W^{(k+0.5)}, \theta] \to [W^{(k+1)}|V^{(k+1)}, \gamma].$$

We show in Section 2.I that this algorithm is equivalent to SD-SE GIS in a certain sense so that we expect the mixing and convergence properties of the two algorithms to be very similar, and we confirm this in the local level model in Section 2.6. So ease of implementation and computational cost per iteration are the only real considerations involved in choosing between the two algorithms.

In our original definition of the CIS sampler for the DLM we used the scaled disturbances as the AA for W and the scaled errors and the AA for V. We could have reversed this or used the same AA for both V and W since both the scaled errors and scaled disturbances are AAs for (V, W), or we can have used θ as the AA for V. In each of these cases, the resulting algorithm would reduce to either the state sampler or a *partial CIS* algorithm, also introduced by Yu and Meng (2011). Section 2.J discusses partial CIS algorithms in general and in the DLM. In the next section we will characterize the efficiency of the various available samplers in the local level model, both in terms of computational cost and in terms of the mixing and convergence of the Markov chain.

2.6 Application: The local level model

2.6.1 The local level model and its DAs

The local level model (LLM) is a DLM with univariate data y_t for $t = 1, 2, \cdots, T$ and a univariate latent state θ_t for $t = 0, 1, \cdots, T$. In the general DLM notation, $F_t = 1 = G_t = 1$ for all t while V and W are scalar. We can write the model as

$$y_t|\theta, V, W \overset{ind}{\sim} N(\theta_t, V), \qquad\qquad \theta_t|\theta_{0:t-1}, V, W \sim N(\theta_{t-1}, W)$$

for $t = 1, 2, \cdots, T$. The priors on (θ_0, V, W) from Section 2.3 become $\theta_0 \sim N(m_0, C_0)$, $V \sim IG(\alpha_V, \beta_V)$ and $W \sim IG(\alpha_W, \beta_W)$ with θ_0, V and W mutually independent where $IG(\alpha, \beta)$ is the inverse gamma distribution with shape parameter α and rate parameter β. Commonly in this model W is called the signal, V is called the noise, and $R = W/V$ is called the signal-to-noise ratio.

We can define the various DAs from Section 2.4 in the context of the local level model. The latent states are simply θ. From the states we obtain the scaled disturbances as $\gamma_0 = \theta_0$

and $\gamma_t = (\theta_t - \theta_{t-1})/\sqrt{W}$ for $t = 1, 2, \cdots, T$. Similarly, the scaled errors are $\psi_0 = \theta_0$ and $\psi_t = (y_t - \theta_t)/\sqrt{V}$ for $t = 1, 2, \cdots, T$. The wrongly-scaled disturbances are then $\tilde{\gamma}_0 = \theta_0$ and $\tilde{\gamma}_t = (\theta_t - \theta_{t-1})/\sqrt{V}$ while the wrongly-scaled errors are $\tilde{\psi}_0 = \theta_0$ with $\tilde{\psi}_t = (y_t - \theta_t)/\sqrt{W}$, both for $t = 1, 2, \cdots, T$.

Most of the full conditional distributions required to construct each of the MCMC samplers in Section 2.5 for the LLM follow straightforwardly from the general case and their derivations can be found in Secton 2.D. For all algorithms, we use the MCFA to draw the DA except in the case of γ, where we use MCFA to draw θ and then transform to γ. For V and W, their draws are either an inverse gamma draw or a draw from a difficult full conditional. In Secton 2.D we derive the difficult density in detail and in Secton 2.G we show how to obtain random draws from it.

2.6.2 Simulation setup

We simulated data from the local level model using a factorial design with V and W each taking the values $10^{i/2}$ where $i = -4, -3, \ldots, 4$ and with T taking the values $10, 100, 1000$. Then for each dataset, we fit the local level model using a variety of the algorithms discussed in this paper. We used the same rule for constructing priors for each model: $\theta_0 \sim N(0, 10^7)$, $V \sim IG(5, 4V^*)$, and $W \sim IG(5, 4W^*)$, mutually independent where (V^*, W^*) are the true values of V and W used to simulate the time series. Note that the prior means are equal to the true values of V and W, so both the prior and likelihood and thus the posterior roughly agree about the likely values of V and W. The behavior of each of these samplers depends on where in the parameter space the posterior distribution puts most of its mass and this prior allows us to highlight that.

For each dataset and sampler we obtained $n = 6500$ draws and threw away the first 500. The chains were started at the true values used to simulate the time series, so we can examine the behavior of the chains to determine how well they mix but not how quickly they converge. Define the effective sample proportion for a scalar component of the chain as the effective sample size (ESS) (Gelman et al., 2013) of the component divided by the actual sample size n ($ESP = ESS/n$). When $ESP = 1$ the Markov chain is behaving as if it obtains iid draws

Table 2.1: Rule of thumb for when each sampler has a high ESP for each variable as a function of the true signal-to-noise ratio, $R^* = W^*/V^*$. The bottom panel of the table applies to both the interweaving and alternating algorithms. Note that as the length of the time series increases, the farther away from one R^* has to be for a given sampler to have a high ESP.

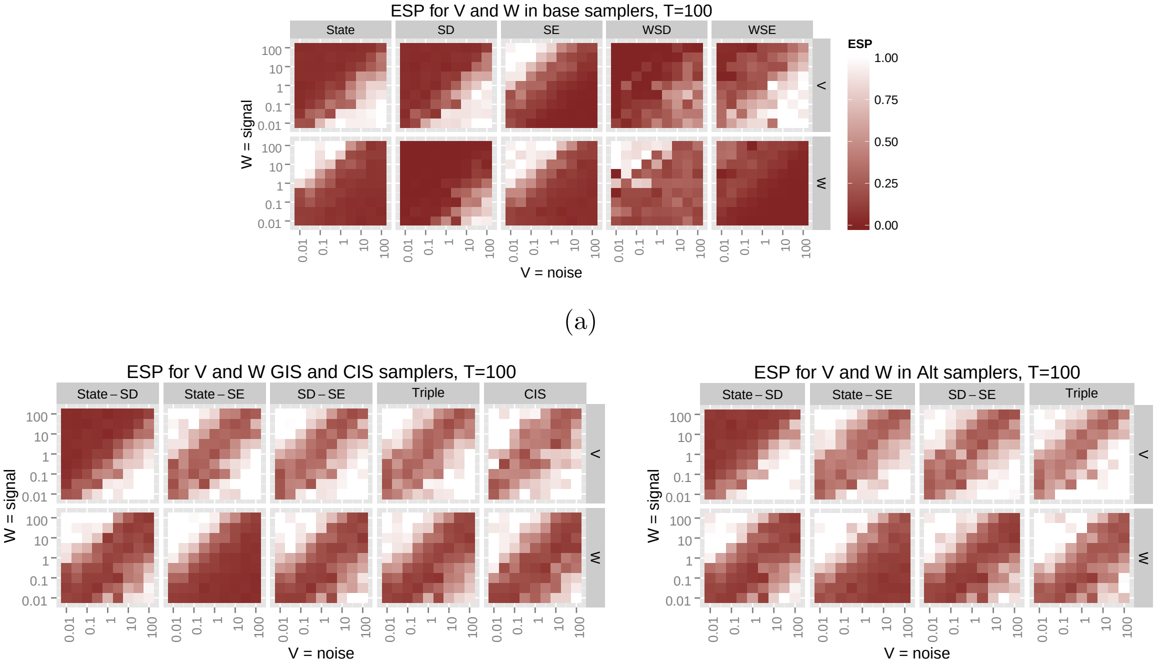
	State	SD	SE	WSD	WSE
V	$R^* < 1$	$R^* < 1$	$R^* > 1$	$R^* < 1$	$R^* < 1$
W	$R^* > 1$	$R^* < 1$	$R^* > 1$	$R^* > 1$	$R^* > 1$
	State-SD	State-SE	SD-SE	Triple	CIS
V	$R^* < 1$	$R^* \not\approx 1$	$R^* \not\approx 1$	$R^* \not\approx 1$	$R^* \not\approx 1$
W	$R^* \not\approx 1$	$R^* > 1$	$R^* \not\approx 1$	$R^* \not\approx 1$	$R^* \not\approx 1$

from the posterior. It is possible that $ESP > 1$ if the draws are negatively correlated and this occasionally happens in our simulations, but we round the ESP down to one for plotting purposes.

2.6.3 Simulation results

Figure 2.1a contains plots of ESP for V and W in each chain of each base samplers for $T = 100$ — the $T = 10$ and $T = 1000$ plots are similar and can be found in Section 2.L. Table 2.1 summarizes the results for the base samplers on the top. Let $R^* = V^*/W^*$ denote the true signal-to-noise ratio. The State sampler tends to have a low ESP for V and high ESP for W when $R^* > 1$ with the behavior switched when $R^* < 1$. The SD sampler has low ESP for both V and W when $R^* > 1$ while the SE sampler has low ESP for both when $R^* < 1$ and in particular for V. We omit the results here, but as T increases, in all samplers the region of the parameter space with high ESP shrinks and in the low ESP regions, ESP drops closer to zero. In Section 2.K, we discuss how the pattern of correlations between various quantities in the posterior distribution determines the pattern of ESPs we see in Figure 2.1.

We fit the LLM to the simulated datasets using several GIS samplers and a CIS sampler as well. Since the wrongly-scaled samplers behaved similarly to the state sampler and neither of the underlying DAs were a SA for V and W jointly, we ignored them in the construction of the GIS samplers. Instead, we constructed the State-SD, State-SE, SD-SE, and Triple (State-SD-SE) GIS samplers, as well as the CIS sampler. Figure 2.1b has plots of ESP for each of the GIS and CIS algorithms while Figure 2.1c has plots of ESP for each of the Alt algorithms.



(a)

(b)

(c)

Figure 2.1: Effective sample proportion in the posterior sampler for a time series of length $T = 100$, for V and W in the each sampler. Figure 2.1a contains ESP for V and W for the base samplers, Figure 2.1b contains ESP in the GIS and CIS samplers, and Figure 2.1c contains ESP in the Alt samplers. X and Y axes indicate the true values of V and W respectively for the simulated data. Note that the signal-to-noise ratio is constant moving up any diagonal. In the upper left the signal is high, in the lower right the noise is high.

Table 2.1 summarizes the results on the bottom.

Essentially, each GIS and Alt algorithm has high ESP when at least one of the base algorithms has high ESP. For example, the State-SD GIS and Alt algorithms have high ESP for W except for a narrow band where R^* is near one while ESP is high for W in the state sampler when $R^* > 1$ and in the SD sampler when $R^* < 1$. Similarly in the State-SD GIS and Alt algorithms, mixing for V is identical to the State and SD samplers since neither sampler improves on the other in any region of the parameter space. Both the State-SD GIS and Alt algorithms take advantage of the fact that the state and SD DA algorithms make up a "beauty and the beast" pair for W and thus improves mixing in the marginal chain for W. However, GIS without an SA-AA pair does not appear to improve on Alt. In Section 2.5.4 we noted that the CIS and the SD-SE GIS algorithms consist of the same steps, just rearranged. This suggests that they should perform similarly and in fact the SD-SE GIS algorithm behaves essentially identically to the CIS and Triple GIS algorithms. We also include simulations with differing sizes of T using these samplers in Section 2.L. Like with the base samplers, increasing the length of the time series worsens ESP for both V and W in all samplers and in particular shrinks the area of the parameter space in which ESP is high.

In Section 2.L we also compare the time required to adequately characterize the posterior distribution between various algorithms, taking into account both mixing and computational time. GIS and Alt perform essentially identical in this respect, though there is good reason to expect GIS to sometimes be more efficient.

2.7 Discussion

In order to explore reparameterizing the DA and apply the interweaving strategies of Yu and Meng (2011) in dynamic linear models, we start with two DAs, the latent states and the scaled disturbances, and introduce three new DAs for the DLM: the scaled errors, the wrongly-scaled disturbances, and the wrongly-scaled errors. Using these DAs, we construct several alternating algorithms and GIS algorithms and a CIS algorithm. We also find under some assumptions that any SA for a general class of DLMs yields a full conditional distribution for the model parameters that is as difficult to sample from as the target posterior. With the available DAs, we construct each possible DA algorithm, several GIS algorithms and their corresponding Alt algorithms, and a CIS algorithm for the general DLM and test these algorithms in the local level model using a simulation study. We find that the true signal-to-noise ratio, $R^* = V^*/W^*$, is important for determining when each algorithm performs well, and in addition that there appears to be no substantive difference in mixing between a GIS algorithm an its corresponding Alt algorithm. In fact, the three best performing algorithms under most circumstances are the SD-SE GIS algorithm, the SD-SE Alt algorithm and the CIS algorithm. The only caveat is that for very long time series the GIS version of an algorithm will start to become relatively efficient.

The importance of the true signal-to-noise ratio in DLMs to the mixing and convergence properties of various MCMC algorithms has been anticipated in the literature. In the AR(1) plus noise model, Pitt and Shephard (1999) find that the signal-to-noise ratio along with the

AR(1) coefficient determine the convergence rate of a Gibbs sampler. In addition, they find that the convergence rate decreases as the length of the time series increases, which is consistent with our empirical findings in the local level model. When Frühwirth-Schnatter (2004) study the dynamic regression model with a stationary AR(1) process on the regression coefficient, they use both the states and the scaled disturbances (non-centered disturbances) and several other DAs motivated by some results for Gibbs samplers in the hierarchical model literature. When examining the behavior of the resulting DA algorithms, they find that the relative behavior of the SD sampler and the State sampler depends on a function of the true signal-to-noise ratio that also depends on the true value of the autocorrelation parameter and the distribution of the covariate. In addition none of the other DA algorithms they consider are more efficient than both the state sampler and the SD sampler at the same time. Given this previous work, it is likely that in the general DLM the signal-to-noise ratio will in some way determine how well each algorithm performs even if we do not know the precise manner in which it affects mixing and convergence behavior. This is probably consequence of the relevance of the Bayesian fraction of missing information and the related EM fraction of missing information to the performance of the DA and EM algorithms (see Van Dyk and Meng (2001) for a good explication of both concepts).

A major computational bottleneck in most of our algorithms occurs when we have to draw from $p(W|V,\gamma,y)$, $p(V|W,\psi,y)$, $p(V|W,\tilde{\gamma},y)$ or $p(W|V,\tilde{\psi},y)$. The densities $p(W|V,\gamma,y)$ and $p(V|W,\psi,y)$ have the form

$$p(x) \propto x^{-\alpha-1}\exp\left[-ax+b\sqrt{x}-c/x\right],$$

while the densities $p(W|V,\tilde{\psi},y)$ and $p(V|W,\tilde{\gamma},y)$ have the form

$$p(x) \propto x^{-\alpha-1}\exp\left[-ax+b/\sqrt{x}-c/x\right]$$

where $\alpha, a, c > 0$ and $b \in \Re$. When $b = 0$ we have a special case of the generalized inverse Gaussian (GIG) distribution, so perhaps the methods used to speed up draws from the GIG can be used here (Jørgensen, 1982; Dagpunar, 1989; Devroye, 2012). On the other hand, it might be worth putting effort into drawing V and W jointly. Using the scaled disturbances, the conditional distribution of V given W is inverse gamma in the LLM and inverse Wishart in the general DLM, so it is easy to derive the marginal density $p(W|\gamma, y)$ up to a proportionality constant. In our LLM example, this density turns out to be very difficult to sample from and in particular, it is not easy to come up with a generally good approximation for rejection sampling or for a Metropolis step. The problem could be solved by a more judicious choice of priors — we chose inverse Wishart priors for V and W partially because they are standard and partially because their conditional conjugacy with the states is computationally convenient, but outside of the state sampler there may be a more convenient prior. In addition, there are well known inferential problems with the inverse Wishart prior in the hierarchical model literature, e.g. Gelman (2006) and Alvarez-Castro et al. (2014), though it is unclear whether this transfers over to DLMs or more generally any time series model. An alternative is to use the conditionally conjugate prior conditional on the scaled disturbances, or whichever DA we prefer. In the LLM, the conditionally conjugate prior for $\sqrt{W}$ using the scaled disturbances as the DA is a Gaussian distribution — strictly speaking this prior is on $\pm\sqrt{W}$. If we use this prior for $\pm\sqrt{V}$ as well, the V step in the scaled disturbance sampler becomes a draw from the generalized inverse Gaussian distribution. This prior has been used by Frühwirth-Schnatter and Wagner (2011) and Frühwirth-Schnatter and Tüchler (2008) to speed up computation while using the scaled disturbances in hierarchical models and by Frühwirth-Schnatter and Wagner (2010) for time series models with a DA similar to the scaled disturbances. We omit the results here, but using this prior on both variances does not alter our mixing results for any of the MCMC samplers. There is a trade-off in computation time to consider — for example when using the scaled disturbances, the draw of $W|V, \gamma, y$ is sped up by using the Gaussian prior on $\pm\sqrt{W}$ since it becomes a Gaussian draw while the $V|W, \gamma, y$ is slower since it becomes a generalized inverse Gaussian draw instead of an inverse gamma. The gains outweigh the costs, at least in the local level model.

In the general DLM, however, it is unclear whether this will hold because of the additional complications stemming from V and W being matrices. The conditionally conjugate prior for W given γ is a normal distribution on $\pm L_W$, or in the case of V given ψ, a normal distribution on $\pm\, L_V$. But the full conditional for the other covariance matrix becomes a matrix analogue of the generalized inverse Gaussian distribution, which appears difficult to sample from. So no matter which conditionally conjugate prior is used under the scaled errors or scaled disturbances, one of V or W's full conditionals will be intractable. This is not a problem for the DA algorithms necessarily – you have the freedom to use the inverse Wishart prior for V and the normal prior for $\pm L_W$ in the scaled disturbance sampler, for example. But in any interweaving or alternating algorithm each covariance matrix needs to be drawn from two full conditionals – one given each of the DAs used in the algorithm, yielding at least one intractable full conditional. A Metropolis step is probably a tolerable solution to the problem, though the details of how to best accomplish this will likely have to be determined on a case by case basis.

APPENDIX

2.A Marginal model of the DLM

The class of DLMs we consider is

$$y_t|\theta, V, W \overset{ind}{\sim} N_k(F_t\theta_t, V) \qquad\qquad \theta_t|\theta_{0:t-1}, V, W \sim N_p(G_t\theta_{t-1}, W) \tag{2.4}$$

for $t = 1, 2, \cdots T$ where V and W are unknown covariance matrices. Define $v_t = y_t - F_t\theta_t$ and $w_t = \theta_t - G_t\theta_{t-1}$. Then we can rewrite the model by recursive substitution:

$$y_t = v_t + F_t\left(w_t + G_tw_{t-1} + G_tG_{t-1}w_{t-2} + ... + G_tG_{t-1}\cdots G_2w_1 + G_tG_{t-1}\cdots G_1\theta_0\right).$$

Then conditional on $\phi = (V, W)$ each y_t is a linear combination of normal random variables. After marginalizing out θ, $y = (y_1', y_2', \ldots, y_T')$ has a normal distribution such that $\mathrm{E}[y_t|\phi] = F_tH_tm_0$,

$$\mathrm{Var}[y_t|\phi] = V + F_t(K_tWK_t' + H_tC_0H_t')F_t', \quad \text{and} \quad \mathrm{Cov}[y_s, y_t|\phi] = F_s(K_sWK_t' + H_sC_0H_t')F_t',$$

where $H_t = G_tG_{t-1}\cdots G_1$ and $K_t = I_p + G_t + G_tG_{t-1} + \cdots + G_tG_{t-1}\cdots G_2$. Next define $D_t = F_tG_tG_{t-1}\cdots G_1$. Then let $\tilde{V} = I_T \otimes V$ and D be block diagonal with elements $D_1, \ldots, D_T$,

$$\begin{aligned}\tilde{W}_{Tk\times Tk} &= \begin{bmatrix}K_1'F_1' & K_2'F_2' & \cdots K_T'F_T'\end{bmatrix}' W \begin{bmatrix}K_1'F_1' & K_2'F_2' & \cdots K_T'F_T'\end{bmatrix},\\ \tilde{C}_{Tk\times Tk} &= \begin{bmatrix}H_1'F_1' & H_2'F_2' & \cdots H_T'F_T'\end{bmatrix}' C_0 \begin{bmatrix}H_1'F_1' & H_2'F_2' & \cdots H_T'F_T'\end{bmatrix},\end{aligned}$$

and $\tilde{m}_{Tp\times 1} = (m_0', m_0', \cdots m_0')'$. Now we have the data model for y without any data augmentation:

$$y|V, W \overset{ind}{\sim} N_{Tk}(D\tilde{m}, \tilde{V} + \tilde{W} + \tilde{C}). \tag{2.5}$$

2.B Proof of lemma 1

First the normality assumption implies

$$\begin{aligned} y|\eta,\phi &\sim N(D\tilde{m}+\Omega'_{y,\eta}\Omega_\eta^{-1}(\eta-\alpha_\eta), \tilde{V}+\tilde{W}+\tilde{C}-\Omega'_{y,\eta}\Omega_\eta^{-1}\Omega_{y,\eta}) \\ \eta|\phi &\sim N(\alpha_\eta,\Omega_\eta). \end{aligned}$$

Now for η to be a sufficient augmentation we need $D\tilde{m}+\Omega'_{y,\eta}\Omega_\eta^{-1}(\eta-\alpha_\eta)$ and $\tilde{V}+\tilde{W}+\tilde{C}-\Omega'_{y,\eta}\Omega_\eta^{-1}\Omega_{y,\eta}$ to be functionally independent of ϕ. This requires that

$$D\tilde{m}-\Omega'_{y,\eta}\Omega_\eta^{-1}\alpha_\eta+\Omega'_{y,\eta}\Omega_\eta^{-1}\eta = b+A\eta$$

where $A=\Omega'_{y,\eta}\Omega_\eta^{-1}$ and $b=D\tilde{m}-A\alpha_\eta$ must both be free of ϕ. As a result $A\alpha_\eta$ is also free of ϕ and thus so is α_η.

Then using the second equation, we now require Σ free of ϕ where $\Sigma=\tilde{V}+\tilde{W}+\tilde{C}-A\Omega_\eta A'$. This ensures that $\Omega_{\eta,y}$ is not the zero matrix since $\tilde{V}+\tilde{W}+\tilde{C}$ is not free of ϕ. Rearranging we have $A\Omega_\eta A'=\tilde{V}+\tilde{W}+\tilde{C}-\Sigma$. Consider $\tilde{\eta}=A\eta$, which is also a sufficient augmentation since it is just a linear transformation by a constant matrix. Then we have

$$\begin{aligned} y|\tilde{\eta},\phi &\sim N(b+A\eta,\Sigma) \\ \tilde{\eta}|\phi &\sim N(A\alpha_\eta, A\Omega_\eta A') \end{aligned}$$

in other words

$$\begin{aligned} y|\tilde{\eta},\phi &\sim N(b+\tilde{\eta},\Sigma) \\ \tilde{\eta}|\phi &\sim N(A\alpha_\eta,\tilde{V}+\tilde{W}+\tilde{C}-\Sigma). \end{aligned}$$

Thus the posterior density of ϕ given $\tilde{\eta}$ can be written as

$$\begin{aligned} p(\phi|\tilde{\eta},y) &\propto p(y|\tilde{\eta},\phi)p(\tilde{\eta}|\phi)p(\phi) \propto p(\tilde{\eta}|\phi)p(\phi) \\ &\propto p(\phi)|\tilde{V}+\tilde{W}+\tilde{C}-\Sigma|^{-1/2}\exp\left[-\frac{1}{2}(\tilde{\eta}-A\alpha_\eta)'(\tilde{V}+\tilde{W}+\tilde{C}-\Sigma)^{-1}(\tilde{\eta}-A\alpha_\eta)\right]. \end{aligned}$$

Now given that $A'A$ is invertible and the properties of multivariate normal distributions, the density of $p(\phi|\eta,y)$ follows from $\eta=(A'A)^{-1}A'\tilde{\eta}$.

2.C Construction of the wrongly-scaled DA algorithms

The wrongly-scaled DA algorithms are close analogues to their correctly scaled cousins. Starting with the *wrongly-scaled disturbance sampler* (Algorithm WSD), the simulation smoothing step to draw from $p(\tilde{\gamma}|V, W, y)$ is similar to that of the scaled disturbance sampler — the density is Gaussian, but the precision matrix is not tridiagonal, so we draw θ using the MCFA and transform to obtain a draw of $\tilde{\gamma}$. The density of $V, W|\tilde{\gamma}, y$ is too complicated to draw from directly, as was the case when we used the scaled disturbances. In this case, the full conditional distribution of W is the same as its distribution when we condition on the states while the density of $V|\tilde{\gamma}, y$ is once again difficult to draw from. The density of $V|W, \tilde{\gamma}, y$ is easier to work with, at least in the local level model example in Section 6.

Algorithm: WSD. *Wrongly-Scaled Disturbance Sampler*

1. *Use MCFA to draw* $\theta \sim p(\theta|V, W, y)$.
2. *Transform* θ *to* $\tilde{\gamma}$.
3. *Draw* $V \sim p(V|W, \tilde{\gamma}, y)$.
4. *Draw* $W \sim IW\left(\Lambda_W + \sum_{t=1}^{T} w_t w_t', \lambda_W + T\right)$.

Now the third step is difficult and we demonstrate how to accomplish it in the local level model in Section 2.F. We could switch the order in which V and W are drawn in this algorithm so that we can draw W before transforming θ to $\tilde{\gamma}$. This would make each iteration slightly cheaper and probably would not affect the mixing and convergence properties of the algorithm, however we are more interested in comparing the mixing and convergence properties of the various samplers, so we always sample V before W when we cannot sample them jointly.

The *wrongly-scaled error sampler* (Algorithm WSE) is closely related to both the wrongly-scaled disturbance sampler and the scaled error sampler. The density of $\tilde{\psi}|V, W, y$ is Gaussian with a tridiagonal precision matrix, so the simulation smoothing step can be accomplished using the MCFA. The density $p(V, W|\tilde{\psi}, y)$ is from the same class as $p(W, V|\tilde{\gamma}, y)$ so that V and W essentially switch places when we condition on $\tilde{\psi}$ instead of $\tilde{\gamma}$. In particular, $V|W, \tilde{\psi}, y$

has an inverse Wishart density and the density of $W|V,\tilde{\psi},y$ is from the same class as that of $V|W,\tilde{\gamma},y$.

Algorithm: WSE. *Wrongly-Scaled Error Sampler*

1. *Use MCFA to draw* $\tilde{\psi} \sim p(\theta|V,W,y)$.
2. *Draw* $V \sim IW\left(\Lambda_V + \sum_{t=1}^T v_t v_t', \lambda_V + T\right)$.
3. *Draw* $W \sim p(W|V,\tilde{\psi},y)$

The constructions of Algorithms WSD and WSE in the local level model example from Section 6 require $p(W|V,\tilde{\psi},y)$ and $p(V|W,\tilde{\gamma},y)$ respectively. Both densities have the form $p(x) \propto x^{-\alpha-1}\exp\left[-ax + b/\sqrt{x} - c/x\right]$, which is closely related to the difficult density from the correctly scaled samplers. For $p(V|W,\tilde{\gamma},y)$ we show in Section 2.C that $\alpha = \alpha_V$, $a = a_{\tilde{\gamma}} \equiv \frac{1}{2W}\sum_{t=1}^T \tilde{\gamma}_t^2$, $b = b_{\tilde{\gamma}} \equiv \sum_{t=1}^T (y_t - \tilde{\gamma}_0)\sum_{s=1}^t \tilde{\gamma}_s$, and $c = c_{\tilde{\gamma}} \equiv \beta_V + \frac{1}{2}\sum_{t=1}^T (y_t - \tilde{\gamma}_0)^2$ while for $p(W|V,\tilde{\psi},y)$ we show that $\alpha = \alpha_W$, $a = a_{\tilde{\psi}} \equiv \frac{1}{2V}\sum_{t=1}^T \tilde{\psi}_t^2$, $b = b_{\tilde{\psi}} \equiv \sum_{t=1}^T \mathcal{L}\tilde{y}_t \mathcal{L}\tilde{\psi}_t$, and $c = c_{\tilde{\psi}} \equiv \beta_W + \frac{1}{2}\sum_{t=1}^T \mathcal{L}\tilde{y}_t^2$. This density is harder to sample from because adaptive rejection sampling does not work very well, so we construct a rejection sampler on the log scale using a t approximation in Section 2.G.

2.D Full conditional distributions in the general DLM for various DAs

The class of DLMs we consider is defined as follows:

$$y_t = F_t\theta_t + v_t \qquad v_t \overset{ind}{\sim} N_k(0,V) \qquad \text{(observation equation)} \qquad (2.6)$$

$$\theta_t = G_t\theta_{t-1} + w_t \qquad w_t \overset{ind}{\sim} N_p(0,W) \qquad \text{(system equation)} \qquad (2.7)$$

for $t = 1, 2, \cdots T$ with the priors $\theta_0 \sim N_p(m_0, C_0)$, $V \sim IW(\Lambda_V, \lambda_V)$ and $W \sim IW(\Lambda_W, \lambda_W)$ with (θ_0, V, W) mutually independent. Then the full joint distribution of (V, W, θ, y) is

$$\begin{aligned} p(V, W, \theta, y) \propto & \exp\left[-\frac{1}{2}(\theta_0 - m_0)' C_0^{-1}(\theta_0 - m_0)\right] \\ & \times |V|^{-(\lambda_V + k + T + 2)/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_V V^{-1}\right)\right] \exp\left[-\frac{1}{2}\sum_{t=1}^{T}(y_t - F_t\theta_t)' V^{-1}(y_t - F_t\theta_t)\right] \\ & \times |W|^{-(\lambda_W + p + T + 2)/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_W W^{-1}\right)\right] \exp\left[-\frac{1}{2}\sum_{t=1}^{T}(\theta_t - G_t\theta_{t-1})' W^{-1}(\theta_t - G_t\theta_{t-1})\right] \end{aligned} \tag{2.8}$$

where tr(.) is the matrix trace operator.

In the following subsections, we provide derivations of the full conditional distributions for when using states, scaled disturbances or scaled errors as the data augmentation.

2.D.1 States

With the usual DA, the full conditional distributions can be derived from equation (2.8). First, the full conditional distribution of θ is as follows:

$$\begin{aligned} p(\theta|V, W, y) \propto p(V, W, \theta, y) \propto & \exp\left[-\frac{1}{2}(\theta_0 - m_0)' C_0^{-1}(\theta_0 - m_0)\right] \\ \times \exp\left[-\frac{1}{2}\sum_{t=1}^{T}(y_t - F_t\theta_t)' V^{-1}(y_t - F_t\theta_t)\right] & \exp\left[-\frac{1}{2}\sum_{t=1}^{T}(\theta_t - G_t\theta_{t-1})' W^{-1}(\theta_t - G_t\theta_{t-1})\right]. \end{aligned}$$

It turns out that this density is Gaussian. In Section 2.E, we show how to use the mixed Cholesky factorization algorithm (MCFA) in order to efficiently determine and draw from this distribution.

The full conditional of (V, W) is:

$$\begin{aligned} p(V, W|\theta, y) \propto p(V, W, \theta, y) \propto & |V|^{-(\lambda_V + k + T + 2)/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_V V^{-1}\right)\right] \exp\left[-\frac{1}{2}\sum_{t=1}^{T}(y_t - F_t\theta_t)' V^{-1}(y_t - F_t\theta_t)\right] \\ & \times |W|^{-(\lambda_W + p + T + 2)/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_W W^{-1}\right)\right] \exp\left[-\frac{1}{2}\sum_{t=1}^{T}(\theta_t - G_t\theta_{t-1})' W^{-1}(\theta_t - G_t\theta_{t-1})\right] \\ & \propto |V|^{-(\lambda_V + k + T + 2)/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\left(\Lambda_V + \sum_{t=1}^{T}(y_t - F_t\theta_t)(y_t - F_t\theta_t)'\right) V^{-1}\right)\right] \\ & \times |W|^{-(\lambda_W + p + T + 2)/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\left(\Lambda_W + \sum_{t=1}^{T}(\theta_t - G_t\theta_{t-1})(\theta_t - G_t\theta_{t-1})'\right) W^{-1}\right)\right]. \end{aligned}$$

In other words, V and W are conditionally independent given y and θ with

$$V|\theta, y \sim IW\left(\Lambda_V + \sum_{t=1}^{T} v_t v_t', \lambda_V + T\right), \qquad W|\theta, y \sim IW\left(\Lambda_W + \sum_{t=1}^{T} w_t w_t', \lambda_W + T\right)$$

where $v_t = y_t - F_t\theta_t$ and $w_t = \theta_t - G_t\theta_{t-1}$.

In the local level model, the priors on V and W become $V \sim IG(\alpha_V, \beta_V)$ and $W \sim IG(\alpha_W, \beta_W)$. The full conditionals then become

$$V|\theta, y \sim IG\left(\alpha_V + T/2, \beta_V + \sum_{t=1}^{T}(y_t - \theta_t)^2/2\right), \quad W|\theta, y \sim IG\left(\alpha_W + T/2, \beta_W + \sum_{t=1}^{T}(\theta_t - \theta_{t-1})^2/2\right).$$

2.D.2 Scaled disturbances

Let L_W denote the Cholesky decomposition of W, i.e. the lower triangle matrix L_W such that $L_W L_W' = W$. Then the scaled disturbances are $\gamma = \gamma_{0:T} = (\gamma_0', \gamma_1', \cdots, \gamma_T')'$ defined by $\gamma_0 = \theta_0$ and $\gamma_t = L_W^{-1}(\theta_t - G_t\theta_{t-1})$ for $t = 1, 2, \cdots, T$. The reverse transformation is defined recursively by $\theta_0 = \gamma_0$ and $\theta_t = L_W\gamma_t + G_t\theta_{t-1}$ for $t = 1, 2, \cdots, T$. Then the Jacobian is block lower triangular with the identity matrix and T copies of L_W along the diagonal blocks, so $|J| = |L_W|^T = |W|^{T/2}$. From equation (2.8) we can write the full joint distribution of (V, W, γ, y) as

$$\begin{aligned} p(V, W, \gamma, y) \propto & \exp\left[-\frac{1}{2}(\gamma_0 - m_0)'C_0^{-1}(\gamma_0 - m_0)\right] \exp\left[-\frac{1}{2}\gamma_t'\gamma_t\right] \\ & \times |W|^{-(\lambda_W+p+2)/2}|V|^{-(\lambda_V+k+T+2)/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_W W^{-1}\right)\right] \\ & \times \exp\left[-\frac{1}{2}\left(\operatorname{tr}\left(\Lambda_V V^{-1}\right) + \sum_{t=1}^{T}\left[y_t - F_t\theta_t(\gamma, W)\right]' V^{-1}\left[y_t - F_t\theta_t(\gamma, W)\right]\right)\right]. \end{aligned} \tag{2.9}$$

where $\theta_t(\gamma, W)$ denotes the recursive back transformation defined by the scaled disturbances. The full conditional distribution of γ is then

$$\begin{aligned} p(\gamma|V, W, y) \propto p(V, W, \gamma, y) \propto & \exp\left[-\frac{1}{2}(\gamma_0 - m_0)'C_0^{-1}(\gamma_0 - m_0)\right] \exp\left[-\frac{1}{2}\gamma_t'\gamma_t\right] \\ & \times \exp\left[-\frac{1}{2}\left(\sum_{t=1}^{T}\left[y_t - F_t\theta_t(\gamma, W)\right]' V^{-1}\left[y_t - F_t\theta_t(\gamma, W)\right]\right)\right]. \end{aligned}$$

This density is Gaussian, but difficult to draw from. We use the MCFA to draw from $\theta|V, W, y$ instead, then transform from θ to γ using the definition of γ.

Under this parameterization, the full conditional distribution of (V, W) is

$$p(V, W, |\gamma, y) \propto p(V, W, \gamma, y)|W|^{-(\lambda_W+p+2)/2}|V|^{-(\lambda_V+k+T+2)/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_W W^{-1}\right)\right]$$
$$\times \exp\left[-\frac{1}{2}\left(\operatorname{tr}\left(\Lambda_V V^{-1}\right) + \sum_{t=1}^{T}\left[y_t - F_t\theta_t(\gamma, W)\right]' V^{-1}\left[y_t - F_t\theta_t(\gamma, W)\right]\right)\right].$$

The back transformation from θ to γ sets $\theta_0 = \gamma_0$ and for $t = 1, 2, \cdots, T$

$$\begin{aligned}\theta_t &= L_W\gamma_t + G_t\theta_{t-1}\\ &= L_W\gamma_t + \sum_{s=0}^{t-2} G_tG_{t-1}\dots G_{t-s}L_W\gamma_{t-s-1} + G_tG_{t-1}\dots G_1\gamma_0\\ &= \sum_{s=0}^{t-1}\tilde{G}_{s,t}L_W\gamma_{t-s} + \tilde{G}_{t,t}\gamma_0\end{aligned}$$

where $\tilde{G}_{s,t} = G_tG_{t-1}\cdots G_{t-s+1}$ for $s > 0$ and $\tilde{G}_{0,t} = I_p$, the $p \times p$ identity matrix.. Then we can rewrite the conditional distribution of (V, W) as

$$p(V, W, |\gamma, y) \propto p(V, W, \gamma, y) \propto |W|^{-(\lambda_W+p+2)/2}|V|^{-(\lambda_V+k+T+2)/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_W W^{-1}\right)\right] \exp\left[-\frac{1}{2}\left(\operatorname{tr}\left(\Lambda_V V^{-1}\right)\right.\right.$$
$$\times \exp\left[-\frac{1}{2}\left(\sum_{t=1}^{T}\left[y_t - F_t\sum_{s=0}^{t}\tilde{G}_{s,t}L_W\gamma_{t-s} - F_t\tilde{G}_{t,t}\gamma_0\right]' V^{-1}\left[y_t - F_t\sum_{s=0}^{t-1}\tilde{G}_{s,t}L_W\gamma_{t-s} - F_t\tilde{G}_{t,t}\gamma_0\right]\right)\right].$$

This density is fairly complicated, so we resort to the full conditionals of V and W separately. The full conditional of V is familiar:

$$p(V|W, \gamma, y) \propto p(V, W|\gamma, y) \propto |V|^{-(\lambda_V+k+T+2)/2} \times \exp\left[-\frac{1}{2}\left(\operatorname{tr}\left[\Lambda_V + \sum_{t=1}^{T} v_tv_t'\right]V^{-1}\right)\right]$$

where $v_t = y_t - F_t\sum_{s=0}^{t}\tilde{G}_{s,t}L_W\gamma_{t-s} - F_t\tilde{G}_{t,t}\gamma_0 = y_t - F_t\theta_t$. This implies that

$$V|W, \gamma, y \sim IW\left(\Lambda_V + \sum_{t=1}^{T} v_tv_t', \lambda_V + T\right)$$

which is the same distribution as for $V|\theta, y$. In the local level model this reduces to

$$V|W, \gamma, y \sim IG\left(\alpha_V + T/2, \beta_V + \sum_{t=1}^{T}(y_t - \theta_t(\gamma))^2/2\right)$$

which is again the same density if we conditioned on θ.

The full conditional density of W is more complicated:

$$p(W|V, \gamma, y) \propto p(V, W, \gamma, y) \propto |W|^{-(\lambda_W+p+2)/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_W W^{-1}\right)\right]$$
$$\times \exp\left[-\frac{1}{2}\left(\sum_{t=1}^{T}\left[y_t - F_t\sum_{s=0}^{t}\tilde{G}_{s,t}L_W\gamma_{t-s} - F_t\tilde{G}_{t,t}\gamma_0\right]' V^{-1}\left[y_t - F_t\sum_{s=0}^{t-1}\tilde{G}_{s,t}L_W\gamma_{t-s} - F_t\tilde{G}_{t,t}\gamma_0\right]\right)\right].$$

In the local level model, the density is even simpler:

$$p(W|V,\gamma,y) \propto W^{-\alpha_W-1}\exp\left[-\frac{1}{W}\beta_W\right]\exp\left[-\frac{1}{2}\left(\sum_{t=1}^{T}\left[y_t-\sum_{s=0}^{t}\gamma_{t-s}\sqrt{W}\right]'V^{-1}\left[y_t-\sum_{s=0}^{t-1}\gamma_{t-s}\sqrt{W}\right]\right)\right]$$
$$\propto W^{-\alpha_W-1}\exp\left[-a_\gamma W+b_\gamma\sqrt{W}-\frac{\beta_W}{W}\right].$$

where $a_\gamma=\sum_{t=1}^{T}(\sum_{s=1}^{t}\gamma_j)^2/2V$ and $b_\gamma=\sum_{t=1}^{T}(y_t-\gamma_0)(\sum_{s=1}^{t}\gamma_j)/V$. In Section 2.G we show how to efficiently obtain a random draw from this density.

2.D.3 Scaled errors

Let L_V denote the Cholesky decomposition of V, that is $L_VL_V'=V$, then we can define the scaled errors as $\psi_t=L_V^{-1}(y_t-F_t\theta_t)$ for $t=1,2,\cdots,T$ and $\psi_0=\theta_0$. Here we assume that $k=p$ and that F_t is invertible for all t. Then the back transformation is $\theta_t=F_t^{-1}(y_t-L_V\psi_t)$ for $t=1,2,\cdots,T$ and $\theta_0=\psi_0$. The Jacobian of this transformation is block diagonal with a single copy of the identity matrix along with the $F_t^{-1}L_V$'s along the diagonal, so $|J|=(\prod_{t=1}^{T}|F_t|^{-1})|V|^{T/2}$. Then from equation (2.8) we can write the joint distribution of (V,W,ψ,y) as

$$p(V,W,\psi,y)\propto\exp\left[-\frac{1}{2}(\psi_0-m_0)'C_0^{-1}(\psi_0-m_0)\right]\exp\left[-\frac{1}{2}\sum_{t=1}^{T}\psi_t'\psi_t\right]$$
$$\times|V|^{-(\lambda_V+p+2)/2}\exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_VV^{-1}\right)\right]\times|W|^{-(\lambda_W+p+T+2)/2}$$
$$\exp\left[-\frac{1}{2}\left(\operatorname{tr}\left(\Lambda_WW^{-1}\right)+\sum_{t=1}^{T}(y_t-\mu_t)'(F_tWF_t')^{-1}(y_t-\mu_t)\right)\right]\tag{2.10}$$

where we define $\mu_1=L_V\psi_1+F_1G_1\psi_0$ and for $t=2,3,\cdots,T$, $\mu_t=L_V\psi_t+F_tG_tF_{t-1}^{-1}(y_{t-1}-L_V\psi_{t-1})$. The $|F_t|^{-1}$'s have been absorbed into the normalizing constant, but if they depended on some unknown parameter then we could not do this and as a result would have to take them into account in the Gibbs step or steps for the model parameters.

The full conditional distribution of ψ is

$$p(V,W,\psi,y)\propto\exp\left[-\frac{1}{2}(\psi_0-m_0)'C_0^{-1}(\psi_0-m_0)\right]\exp\left[-\frac{1}{2}\sum_{t=1}^{T}\psi_t'\psi_t\right]$$
$$\exp\left[-\frac{1}{2}\left(\sum_{t=1}^{T}(y_t-\mu_t)'(F_tWF_t')^{-1}(y_t-\mu_t)\right)\right]$$

where note that μ_t depends on ψ. This density is Gaussian and like with γ, we can use the MCFA from Section 2.E to draw from the full conditional of θ and then transform from θ to ψ. However it turns out the precision matrix of ψ's full conditional distribution has the necessary block tridiagonal structure, so we use the MCFA directly on ψ.

The full conditional distribution of (V, W) is complicated, like the case of the scaled disturbances, so we find the full conditionals of V and W separately instead. The full conditional of W is

$$p(W|V,\psi,y) \propto |W|^{-(\lambda_W+p+T+2)/2} \exp\left[-\frac{1}{2}\left(\mathrm{tr}\left(\left[\Lambda_W + \sum_{t=1}^{T} F_t^{-1}(y_t-\mu_t)(y_t-\mu_t)'(F_t^{-1})'\right]W^{-1}\right)\right)\right],$$

in other words

$$W|V,\psi,y \sim IW\left(\Lambda_W + \sum_{t=1}^{T} w_t w_t', \lambda_W + T\right)$$

where $w_t = F_t^{-1}(y_t - \mu_t) = \theta_t - G_t\theta_{t-1}$. In the local level model, this becomes

$$W|V,\psi,y \sim IG\left(\alpha_W + T/2, \beta_W + \sum_{t=1}^{T}(\theta_t(\psi) - \theta_{t-1}(\psi))^2/2\right).$$

The full conditional distribution of V is more complicated:

$$p(V|W,\psi,y) \propto p(V,W,\psi,y) \propto |V|^{-(\lambda_V+p+2)/2} \exp\left[-\frac{1}{2}\,\mathrm{tr}\left(\Lambda_V V^{-1} + \sum_{t=1}^{T}(y_t-\mu_t)'(F_t W F_t')^{-1}(y_t-\mu_t)\right)\right]$$

with μ_t a function of V, defined above. In the local level model with an $IG(\alpha_V, \beta_V)$ prior on V, this density is simpler:

$$p(V|W,\psi,y) \propto V^{-\alpha_V-1} \exp\left[-\frac{\beta_V}{V} + \frac{1}{W}\sum_{t=1}^{T}(y_t-\mu_t)'(y_t-\mu_t)\right]$$

where $\mu_1 = \sqrt{V}\psi_1 + \psi_0$ and for $t = 2, 3, \cdots, T$, $\mu_t = \sqrt{V}(\psi_t - \psi_{t-1}) + y_{t-1}$. Thus

$$p(V|W,\psi,y) \propto V^{-\alpha_V-1} \exp\left[-a_\psi V + b_\psi\sqrt{V} - \frac{\beta_V}{V}\right]$$

where $a_\psi = \sum_{t=1}^{T}(\mathcal{L}\psi_t)^2/2W$ and $b_\psi = \sum_{t=1}^{T}(\mathcal{L}\psi_t\mathcal{L}y_t)/W$, and we define $\mathcal{L}y_t = y_t - y_{t-1}$ for $t = 2, 3, \cdots, T$, $\mathcal{L}y_1 = y_1 - \psi_0$, $\mathcal{L}\psi_t = \psi_t - \psi_{t-1}$ for $t = 2, 3, ..., T$ and $\mathcal{L}\psi_1 = \psi_1 - 0$. In other words, the form of $p(V|W,\psi,y)$ is the same as $p(W|V,\gamma,y)$. The general form of these two densities is $p(x) \propto x^{-\alpha-1}\exp\left[-ax + b\sqrt{x} - c/x\right]$. In Section 2.G we show how to efficiently sample from this distribution.

2.D.4 The wrongly-scaled disturbances

The wrongly-scaled disturbances are defined as $\tilde{\gamma} = \tilde{\gamma}_{0:T} = (\tilde{\gamma}_0', \tilde{\gamma}_1', \cdots, \tilde{\gamma}_T')'$. The wrongly-scaled disturbances are related to the scaled disturbances by $\tilde{\gamma}_t = L_V^{-1} L_W \gamma_t$ for $t = 1, 2, \cdots, T$ and $\tilde{\gamma}_0 = \gamma_0$. The reverse transformation is $\gamma_t = L_W^{-1} L_V \tilde{\gamma}_t$ and the Jacobian is block diagonal with a copy of the identity matrix and T copies of $L_W^{-1} L_V$ along the diagonal. Thus $|J| = |L_W|^{-T} |L_V|^T = |W|^{-T/2} |V|^{T/2}$. Then from equation (2.9) we can write the joint distribution of $(V, W, \tilde{\gamma}, y)$ as

$$\begin{aligned} p(V, W, \tilde{\gamma}, y) \propto & \exp\left[-\frac{1}{2}(\tilde{\gamma}_0 - m_0)' C_0^{-1} (\tilde{\gamma}_0 - m_0)\right] |V|^{-(\lambda_V + p + 2)/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_V V^{-1}\right)\right] \\ & \times \exp\left[-\frac{1}{2}\sum_{t=1}^{T} (y_t - F_t\theta_t(\tilde{\gamma}, L_V))' V^{-1} (y_t - F_t\theta_t(\tilde{\gamma}, L_V))\right] \\ & \times |W|^{-(\lambda_W + p + T + 2)/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_W W^{-1}\right)\right] \exp\left[-\frac{1}{2}\sum_{t=1}^{T} \tilde{\gamma}_t' (L_V^{-1} W (L_V^{-1})')^{-1} \tilde{\gamma}_t\right] \end{aligned} \tag{2.11}$$

where $\theta_t(\tilde{\gamma}, L_V)$ denotes the transformation from $\tilde{\gamma}$ to θ defined by the wrongly-scaled disturbances.

Now from equation (2.11), we can write the full conditional density of $\tilde{\gamma}$ as

$$\begin{aligned} p(\tilde{\gamma}|V, W, y) \propto & \exp\left[-\frac{1}{2}(\tilde{\gamma}_0 - m_0)' C_0^{-1} (\tilde{\gamma}_0 - m_0)\right] \exp\left[-\frac{1}{2}\sum_{t=1}^{T} \tilde{\gamma}_t' (L_V^{-1} W (L_V^{-1})')^{-1} \tilde{\gamma}_t\right] \\ & \times \exp\left[-\frac{1}{2}\sum_{t=1}^{T} (y_t - F_t\theta_t(\tilde{\gamma}, L_V))' V^{-1} (y_t - F_t\theta_t(\tilde{\gamma}, L_V))\right]. \end{aligned}$$

This density is Gaussian but difficult to draw from, so we use the MCFA to draw $\theta|V, W, y$ instead, then transform from θ to $\tilde{\gamma}$.

Then full conditional density of (V, W) is complicated, but their separate full conditionals are easier to work with. The full conditional density of W is

$$p(W|V, \tilde{\gamma}, y) \propto |W|^{-(\lambda_W + p + T + 2)/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\left[\Lambda_W + \sum_{t=1}^{T} L_V \tilde{\gamma}_t \tilde{\gamma}_t' L_V'\right] W^{-1}\right)\right],$$

i.e.

$$W|V, \tilde{\gamma}, y \sim IW\left(\Lambda_W + \sum_{t=1}^{T} w_t w_t', \lambda_W + T\right)$$

where $w_t = L_V\tilde{\gamma}_t = \theta_t - G_t\theta_{t-1}$. In the local level model, this density becomes

$$W|V,\tilde{\gamma},y \sim IG\left(\alpha_W + T/2, \beta_W + \sum_{t=1}^T(\theta_t(\tilde{\gamma}) - \theta_{t-1}(\tilde{\gamma}))^2/2\right).$$

The full conditional density of V is more complicated, from equation (2.11):

$$\begin{aligned}p(V|W,\tilde{\gamma},y) \propto& |V|^{-(\lambda_V+p+2)/2}\exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_V V^{-1}\right)\right]\exp\left[-\frac{1}{2}\sum_{t=1}^T\tilde{\gamma}_t'(L_V^{-1}W(L_V^{-1})')^{-1}\tilde{\gamma}_t\right]\\ &\times\exp\left[-\frac{1}{2}\sum_{t=1}^T(y_t - F_t\theta_t(\tilde{\gamma},L_V))'\,V^{-1}\,(y_t - F_t\theta_t(\tilde{\gamma},L_V))\right].\end{aligned}$$

In the local level model with an $IG(\alpha_V,\beta_V)$ prior on V, this density becomes simpler. Since in that case $\theta_t = \sqrt{V}\sum_{s=1}^t\tilde{\gamma}_s + \tilde{\gamma}_0$, we have

$$p(V|W,\tilde{\gamma},y) \propto V^{-\alpha_V-1}\exp\left[-a_{\tilde{\gamma}}V + b_{\tilde{\gamma}}/\sqrt{V} - c_{\tilde{\gamma}}/V\right]$$

where $a_{\tilde{\gamma}} = \frac{1}{2W}\sum_{t=1}^T\tilde{\gamma}_t^2$, $b_{\tilde{\gamma}} = \sum_{t=1}^T(y_t - \tilde{\gamma}_0)\sum_{s=1}^t\tilde{\gamma}_s$, and $c_{\tilde{\gamma}} = \beta_V + \frac{1}{2}\sum_{t=1}^T(y_t - \tilde{\gamma}_0)^2$. We show in Section 2.H how to efficiently obtain a random draw from this density.

2.D.5 The wrongly-scaled errors

The wrongly-scaled errors are denoted by $\tilde{\psi} = \tilde{\psi}_{0:T} = (\tilde{\psi}_0', \tilde{\psi}_1', \cdots, \tilde{\psi}_T')'$. They are related to the scaled errors by $\tilde{\psi}_t = L_W^{-1}L_V\psi_t$ for $t = 1, 2, \cdots, T$ and $\tilde{\psi}_0 = \psi_0$. Then $\psi_t = L_V^{-1}L_W\tilde{\psi}_t$ and the Jacobian is block diagonal with a copy of the identical matrix and T copies of $L_V^{-1}L_W$ along the diagonal. So $|J| = |V|^{-T/2}|W|^{T/2}$ and from equation (2.10) we can write the joint distribution of $(V,W,\tilde{\psi},y)$ as

$$\begin{aligned}p(V,W,\tilde{\psi},y) \propto& \exp\left[-\frac{1}{2}(\tilde{\psi}_0 - m_0)'C_0^{-1}(\tilde{\psi}_0 - m_0)\right]\\ &\times|V|^{-(\lambda_V+p+T+2)/2}\exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_V V^{-1}\right)\right]\exp\left[-\frac{1}{2}\sum_{t=1}^T\tilde{\psi}_t'(L_W^{-1}V(L_W^{-1})')^{-1}\tilde{\psi}_t\right]\\ &\times|W|^{-(\lambda_W+p+2)/2}\exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_W W^{-1}\right)\right]\exp\left[-\frac{1}{2}\sum_{t=1}^T(y_t - \tilde{\mu}_t)'(F_tWF_t')^{-1}(y_t - \tilde{\mu}_t)\right]\end{aligned} \tag{2.12}$$

where we define $\tilde{\mu}_1 = L_W\tilde{\psi}_1 - F_1G_1\tilde{\psi}_0$ and for $t = 2, 3, \cdots, T$ $\tilde{\mu}_t = L_W\tilde{\psi}_t - F_tG_tF_{t-1}^{-1}(y_{t-1} - L_W\tilde{\psi}_{t-1})$.

From equation (2.12) the full conditional distribution of $\tilde{\psi}$ is

$$p(\tilde{\psi}|V,W,y) \propto \exp\left[-\frac{1}{2}(\tilde{\psi}_0 - m_0)'C_0^{-1}(\tilde{\psi}_0 - m_0)\right]\exp\left[-\frac{1}{2}\sum_{t=1}^{T}\tilde{\psi}_t'(L_W^{-1}V(L_W^{-1})')^{-1}\tilde{\psi}_t\right]$$
$$\times\exp\left[-\frac{1}{2}\sum_{t=1}^{T}(y_t-\tilde{\mu}_t)'(F_tWF_t')^{-1}(y_t-\tilde{\mu}_t)\right].$$

This density is again Gaussian and it can be shown that the precision matrix is tridiagonal, so the MCFA can be directly applied. The full conditional density of V is the familiar inverse Wishart:

$$p(V|W,\tilde{\psi},y) \propto |V|^{-(\lambda_V+p+T+2)/2}\exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_V V^{-1}\right)\right]\exp\left[-\frac{1}{2}\sum_{t=1}^{T}\tilde{\psi}_t'(L_W^{-1}V(L_W^{-1})')^{-1}\tilde{\psi}_t\right].$$

So $V|W,\tilde{\psi},y \sim IW\left(\Lambda_V + \sum_{t=1}^{T}v_tv_t', \lambda_V + T\right)$ where $v_t = L_W\tilde{\psi}_t = y_t - F_t\theta_t$. In the local level model, this becomes

$$V|W,\tilde{\psi},y \sim IG\left(\alpha_V + T/2, \beta_V + \sum_{t=1}^{T}(y_t - \theta_t(\tilde{\psi}))^2/2\right).$$

The full conditional density of W is more complicated, but has the same form as the full conditional density of V given $\tilde{\gamma}$:

$$p(W|V,\tilde{\psi},y) \propto |W|^{-(\lambda_W+p+2)/2}\exp\left[-\frac{1}{2}\sum_{t=1}^{T}\tilde{\psi}_t'(L_W^{-1}V(L_W^{-1})')^{-1}\tilde{\psi}_t\right]$$
$$\times\exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_W W^{-1}\right)\right]\exp\left[-\frac{1}{2}\sum_{t=1}^{T}(y_t-\tilde{\mu}_t)'(F_tWF_t')^{-1}(y_t-\tilde{\mu}_t)\right].$$

In the case of the local level model with a $IG(\alpha_W,\beta_W)$ prior on W, this density simplifies to

$$p(W|V,\tilde{\psi},y) \propto W^{-\alpha_W-1}\exp\left[-a_{\tilde{\psi}}W + b_{\tilde{\psi}}/\sqrt{W} - c_{\tilde{\psi}}/W\right]$$

where $a_{\tilde{\psi}} = \frac{1}{2V}\sum_{t=1}^{T}\tilde{\psi}_t^2$, $b_{\tilde{\psi}} = \sum_{t=1}^{T}\mathcal{L}\tilde{y}_t\mathcal{L}\tilde{\psi}_t$, and $c_{\tilde{\psi}} = \beta_W + \frac{1}{2}\sum_{t=1}^{T}\mathcal{L}\tilde{y}_t^2$. Here we define $\mathcal{L}y_t = y_t - y_{t-1}$ for $t = 2,3,\cdots,T$ while $\mathcal{L}y_1 = y_1 - \tilde{\psi}_0$, and $\mathcal{L}\tilde{\psi}_t = \tilde{\psi}_t - \tilde{\psi}_{t-1}$ for $t = 2,3,\cdots,T$ while $\mathcal{L}\tilde{\psi}_1 = \tilde{\psi}_1 - 0$. This is the same family of densities as $p(V|W,\tilde{\gamma},y)$, and in Section 2.H we show how to efficiently obtain random draws.

2.E Mixed Cholesky Factorization Algorithm (MCFA) for simulation smoothing

Traditionally in DLMs, forward filtering, backward sampling (FFBS) is used in order to draw from the latent states $\theta_{0:T}$. This requires running the Kalman filter in order to determine the marginal distribution of θ_T, then drawing $\theta_t|\theta_{t+1:T}$ for $t = T-1, T-2, \cdots, 1$ Carter and Kohn (1994); Frühwirth-Schnatter (1994). The mixed Cholesky factorization algorithm (MCFA) determines the joint distribution of $\theta_{0:T}$ and draws from it using a backward sampling step as in FFBS. The idea comes from Rue (2001), which introduces a Cholesky factorization algorithm (CFA) for drawing from a Gaussian Markov random field and notes that the conditional distribution of $\theta_{0:T}$ given $y_{1:T}$ in a Gaussian linear statespace model is a special case. The algorithm exploits the fact that the full conditional distribution of $\theta_{0:T}$ is Gaussian with a block tridiagonal precision matrix in order to quickly compute its Cholesky decomposition. McCausland et al. (2011) improves the idea by implicitly computing this Cholesky decomposition through a backward sampling strategy, starting with sampling from the marginal distribution of θ_T.

Suppose our model is as follows:

$$y_t = F_t\theta_t + v_t$$

$$\theta_t = G_t\theta_{t-1} + w_t$$

with $v_t \overset{ind}{\sim} N(0, V_t)$ independent of $w_t \overset{ind}{\sim} N(0, W_t)$ for $t = 1, 2, \cdots, T$ and $\theta_0 \sim N(m_0, C_0)$. This is the usual DLM except now we allow for time dependent variances for illustrative purposes. Then $(y_{1:T}, \theta_{0:T})$ is joint Gaussian conditional on $(V_{1:T}, W_{1:T})$ (in this section, everything is conditional on $V_{1:T}$ and $W_{1:T}$, so we will not make this conditioning explicit). So we can write $p(\theta_{0:T}|y_{1:T})$ as

$$\log p(\theta_{0:T}|y_{1:T}) = -\frac{1}{2}g(\theta_{0:T}, y_{1:T}) + K$$

where K is some constant with respect to $\theta_{0:T}$ and

$$g(\theta_{0:T}, y_{1:T}) = \theta_{0:T}'\Omega\theta_{0:T} - 2a'\theta_{0:T}.$$

However, we also have

$$\log p(\theta_{0:T}|y_{1:T}) = \log p(\theta_{0:T}, y_{1:T}) - \log p(y_{1:T}).$$

This means that

$$\begin{aligned}
g(\theta_{0:T}, y_{1:T}) = (\theta_0 - m_0)C_0^{-1}(\theta_0 - m_0) + K'\\
+ \sum_{t=1}^{T}(y_t - F_t\theta_t)'V_t^{-1}(y_t - F_t\theta_t)\\
+ \sum_{t=1}^{T}(\theta_t - G_t\theta_{t-1})'W_t^{-1}(\theta_t - G_t\theta_{t-1}).
\end{aligned}$$

where K' is another constant that doesn't depend on $\theta_{0:T}$.

So now we can identify blocks of Ω with the cross product terms of the θ_t's and blocks of a with the single product terms. Specifically, Ω is a banded diagonal matrix with

$$\Omega = \begin{bmatrix}
\Omega_{00} & \Omega_{01} & 0 & \ddots & 0 & 0\\
\Omega_{10} & \Omega_{11} & \Omega_{12} & \ddots & 0 & 0\\
0 & \Omega_{21} & \Omega_{22} & \ddots & 0 & 0\\
\ddots & \ddots & \ddots & \ddots & \ddots & \ddots\\
0 & 0 & 0 & \ddots & \Omega_{T-1,T-1} & \Omega_{T-1,T}\\
0 & 0 & 0 & \ddots & \Omega_{T,T-1} & \Omega_{TT}
\end{bmatrix}$$

and $\omega = (\omega_0', \omega_1', \cdots, \omega_T')$ where the Ω_{st}'s and ω_t's defined below:

$$\begin{aligned}
\Omega_{00} &= C_0^{-1} + G_1'W_1^{-1}G_1 & &\\
\Omega_{tt} &= F_t'V_t^{-1}F_t + W_t^{-1} + G_{t+1}'W_{t+1}^{-1}G_{t+1} & &\text{for } \ t = 1, 2, \cdots T-1\\
\Omega_{TT} &= F_T'V_T^{-1}F_T + W_T^{-1} & &\\
\Omega_{t,t-1} &= -W_t^{-1}G_t & &\text{for } \ t = 1, 2, \cdots T\\
\Omega_{t-1,t} &= -G_t'W_t^{-1} = \Omega_{t,t-1}' & &\text{for } \ t = 1, 2, \cdots T\\
\omega_0 &= C_0^{-1}m_0 & &\\
\omega_t &= F_t'V_t^{-1}y_t & &\text{for } \ t = 1, 2, \cdots T.
\end{aligned}$$

Together, Ω and a determine the Gaussian distribution from which $\theta_{0:T}$ should be drawn. Rue (2001) shows how to take advantage of the sparsity of Ω in order to quickly compute its Cholesky factorization and in order to find the mean vector from ω and this factorization. McCausland et al. (2011) shows that instead of computing these quantities directly, you can draw

θ_T and $\theta_t|\theta_{t+1:T}$ iteratively, which ultimately reduces the number of linear algebra operations which must be performed and typically speeds up the computation despite taking advantage of essentially the same mathematical technology.

The resulting algorithm requires a couple more intermediate quantities. Let $\Sigma_0 = \Omega_{00}^{-1}$, $\Sigma_t = (\Omega_{tt} - \Omega_{t,t-1}\Sigma_{t-1}\Omega_{t-1,t})^{-1}$ for $t = 1, 2, \cdots, T$, $h_0 = \Sigma_0\omega_0$, and $h_t = \Sigma_t(\omega_t - \Omega_{t,t-1}h_{t-1})$ for $t = 1, 2, \cdots, T$. Then

$$\begin{aligned}
\theta_T &\sim N(h_T, \Sigma_T) \\
\theta_{t|t+1:T} &\sim N(h_t - \Sigma_t\Omega_{t,t+1}\theta_{t+1}, \Sigma_t) \qquad\qquad \text{for} \;\; t = T-1, T-2, \cdots, 0.
\end{aligned}$$

McCausland et al. (2011) shows how to quickly compute the required linear algebra operations and finds that this method is often faster than simply doing the Cholesky factorization. This algorithm can also be applied to drawing the scaled errors, $\psi_{0:T}$, and the wrongly-scaled errors, $\tilde{\psi}_{0:T}$.

2.F Further augmentation for non-invertible F_t

Throughout the paper we assumed that F_t is square and invertible for all t which made the construction of the SE sampler and other samplers that use the scaled errors easier. However, most DLMs do not have F_t's which are square, let alone invertible. The samplers we constructed can still be used in this case with one tweak: an additional DA is required in order to ensure that F_t is square and invertible for all t. The basic strategy is to add elements to y_t or θ_t or both until F_t is invertible, then add an additional step to the sampler in order to draw the new augmentation. A second issue is that often G_t or F_t or both depend on some unknown parameter which must also be sampled from in the various MCMC samplers. The second case is easily dealt with simply by adding another sampling step for the unknown parameters in F_t and G_t. The following example illustrates how to deal with the first case. See Simpson (2014) for another example.

Consider the dynamic regression model

$$\begin{aligned} y_t &= \alpha_t + x_t\beta_t + v_t \\ \alpha_t &= \alpha_{t-1} + w_{1,t} \\ \beta_t &= \beta_{t-1} + w_{2,t} \end{aligned}$$

for $t = 1, 2, \cdots, T$ with $v_{1:T}$ independent of $w_{1:T} = (w_1', w_2', \cdots, w_T')'$ where $w_t = (w_{1,t}, w_{2,t})'$, $v_t \overset{iid}{\sim} N(0, V)$ and $w_t \overset{iid}{\sim} N_2(0, W)$. Here the latent state in period t is $\theta_t = (\alpha_t, \beta_t)'$. The problem is that $F_t = [1, x_t]$ is neither square nor invertible. But notice that the matrix

$$F_t^* = \begin{bmatrix} 1 & x_t \\ 0 & 1 \end{bmatrix}$$

is invertible. Now we add an additional DA z_t to y_t to construct $y_t^* = (y_t, z_t)'$ so that now the model is

$$\begin{aligned} y_t^* &= F_t^*\theta_t + v_t^* \\ \theta_t &= \theta_{t-1} + w_t \end{aligned}$$

where $v_t^* = (v_t, u_t)$ where $u_{1:T}$ is independent of $(v_{1:T}, w_{1:T})$ and $u_t \overset{iid}{\sim} N(0, 1)$. By construction $v_t^* \overset{iid}{\sim} N_2(0, V^*)$ where V^* is a diagonal matrix with the vector $(V, 1)$ along the diagonal and the full conditional distribution of z_t is $N(\beta_t, 1)$. Then we define the scaled errors as $\psi_0 = \theta_0$ and $\psi_t = L_{V^*}^{-1}(y_t^* - F_t^*\theta_t)$. Let $z = z_{1:T}$ and $y^* = y_{1:T}^*$ for brevity.

In terms of θ, the likelihood is

$$\begin{aligned} p(y, z, \theta|V, W) \propto& |V^*|^{-T/2} \exp\left[-\frac{1}{2}\sum_{t=1}^T (y_t^* - F_t^*\theta_t)'(V^*)^{-1}(y_t^* - F_t^*\theta_t)\right] \\ &\times |W|^{-T/2} \exp\left[-\frac{1}{2}\sum_{t=1}^T (\theta_t - \theta_{t-1})'W^{-1}(\theta_t - \theta_{t-1})\right] \\ \propto& V^{-T/2} \exp\left[-\frac{1}{2V}\sum_{t=1}^T (y_t - \alpha_t - x_t\beta_t)^2\right] \exp\left[-\frac{1}{2}\sum_{t=1}^t (z_t - \beta_t)^2\right] \\ &\times |W|^{-T/2} \exp\left[-\frac{1}{2}\sum_{t=1}^T (\theta_t - \theta_{t-1})'W^{-1}(\theta_t - \theta_{t-1})\right] \end{aligned}$$

Then by transforming to ψ, the back transformation is $\theta_t = (F_t^*)^{-1}(y_t^* - L_{V^*}\psi_t)$ so the Jacobian is block diagonal with T copies of $(F_t^*)^{-1}L_{V^*}$ along with a single copy of the identity matrix along the diagonal. So the determinant of the Jacobian is $|J| = |V^*|^{T/2}$ and the likelihood can be written in terms of ψ as

$$p(y, z, \theta|V, W) \propto \exp\left[-\frac{1}{2}\sum_{t=1}^{T}\psi_t'\psi_t\right]|W|^{-T/2}\exp\left[-\frac{1}{2}\sum_{t=1}^{T}(y_t^* - \mu_t)'(F_t^*W(F_t^*)')^{-1}(y_t^* - \mu_t)\right]. \tag{2.13}$$

where we define $\mu_1 = L_{V^*}\psi_1 + F_1^*\psi_0$ and for $t = 2, 3, \cdots, T$, $\mu_t = L_{V^*}\psi_t + F_t^*(F_{t-1}^*)^{-1}(y_{t-1}^* - L_{V^*}\psi_{t-1})$.

Now in order to construct a sampler that uses ψ, we simply add a new step to sampler to draw z from its full conditional just before transforming to ψ. In the GIS and alternating algorithms, we now have to draw an updated z every time we change the DA. When using the states, $z_t|V, W, \theta, y \overset{iid}{\sim} N(\beta_t, 1)$, so it is easiest to transform to θ before drawing z. So for example in the SD-SE GIS sampler with V, W, α_0, and β_0 independent in the prior, an $IG(\alpha_V, \beta_V)$ prior on V, and an $IW(\Lambda_W, \lambda_W)$ prior on W, the algorithm becomes

Algorithm: SD-SE GIS for dynamic regression. *Scaled Disturbance-Scaled Error GIS Sampler for the dynamic regression model*

1. *Use the MCFA to sample* $\theta \sim p(\theta|V, W, y)$.
2. *Sample* $V \sim IG\left(\alpha_V + T/2, \beta_V + \frac{1}{2}\sum_{t=1}^{T}(y_t - \alpha_t - \beta_t)^2\right)$.
3. *Transform* θ *to* γ.
4. *Sample* $W \sim p(W|V, \gamma, y)$.
5. *Transform* γ *to* θ.
6. *Sample* $z_t \overset{iid}{\sim} N(\beta_t, 1)$ *and form* y^*.
7. *Transform* θ *to* ψ.
8. *Sample* $V \sim p(V|W, \psi, y^*)$.
9. *Sample* $W \sim IW\left(\Lambda_W + \sum_{t=1}^{T} w_t w_t', \lambda_W + T\right)$.

Step 8 is particularly tricky since V is a component of V^*, and V^* has the same density $p(V|W,\psi,y)$ that shows up in the usual case of the scaled disturbances, except now the lower right diagonal element is set to one. So while we can write down the various algorithms in the non-invertible F case, the density $p(V|W,\psi,y^*)$ is tricky to work with. In step 8 V is drawn conditional on y^*, but another option is to draw V conditional on y but not on z. This would require integrating z out of the likelihood, equation (2.13). It is not clear which of these is easier or faster, though it is likely that the changing the prior for V and W will have an impact.

2.G Efficiently drawing from $p(W|V,\gamma,y)$ and $p(V|W,\psi,y)$ in the LLM

From Section 2.D.2, the full conditional distribution of W given γ is

$$p(W|V,\gamma,y) \propto p(V,W,\gamma,y) \propto |W|^{-(\lambda_W+p+2)/2}\exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_W W^{-1}\right)\right]$$
$$\times\exp\left[-\frac{1}{2}\left(\sum_{t=1}^T\left[y_t - F_t\sum_{s=0}^t\tilde{G}_{s,t}L_W\gamma_{t-s} - F_t\tilde{G}_{t,t}\gamma_0\right]'V^{-1}\left[y_t - F_t\sum_{s=0}^{t-1}\tilde{G}_{s,t}L_W\gamma_{t-s} - F_t\tilde{G}_{t,t}\gamma_0\right]\right)\right]$$

where L_W is the Cholesky factor of W defined so that $L_W L_W' = W$. We can write this density as

$$\begin{aligned}p(W|V,\gamma,y) \propto & |W|^{-(\lambda_W+p+2)/2}\exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_W W^{-1}\right)\right]\\ & \times\exp\left[-\frac{1}{2}\left(\operatorname{vec}(L_W)'A_W\operatorname{vec}(L_W) - 2B_W\operatorname{vec}(L_W)\right)\right]\end{aligned}$$

where

$$A_W = \sum_{t=1}^T\sum_{s=0}^t\left(\gamma_{t-s}\gamma_{t-s}'\otimes\tilde{G}_{s,t}'F_t'V^{-1}F_t\tilde{G}_{s,t}\right)$$

and

$$B_W = \sum_{t=1}^T\sum_{s=0}^t\left(\gamma_{t-s}'\otimes(y_t - F_t\tilde{G}_{t,t}\gamma_0)'V^{-1}F_t\tilde{G}_{s,t}\right)$$

can be found using the properties of the vec and tr operators.

Similarly from Section 2.D.3, the full conditional distribution of V given ψ is

$$p(V|W,\psi,y) \propto p(V,W,\psi,y) \propto |V|^{-(\lambda_V+p+2)/2}\exp\left[-\frac{1}{2}\left(\operatorname{tr}\left(\Lambda_V V^{-1}\right) + \sum_{t=1}^T(y_t-\mu_t)'(F_tWF_t')^{-1}(y_t-\mu_t)\right)\right]$$

where $\mu_1 = L_V\psi_1 + F_1G_1\psi_0$ and for $t = 2, 3, \cdots, T$, $\mu_t = L_V\psi_t + F_tG_tF_{t-1}^{-1}(y_{t-1} - L_V\psi_{t-1})$. This density can be written in a familiar form:

$$p(V|W,\psi,y) \propto p(V,W,\psi,y) \propto |V|^{-(\lambda_V+p+2)/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_V V^{-1}\right)\right]$$
$$\times \exp\left[-\frac{1}{2}\Big(\operatorname{vec}(L_V)'A_V\operatorname{vec}(L_V) - 2B_V\operatorname{vec}(L_V)\Big)\right]$$

where

$$\begin{aligned}A_V =&\sum_{t=1}^T \psi_t\psi_t' \otimes (F_tWF_t')^{-1} + \sum_{t=2}^T \psi_{t-1}\psi_{t-1}' \otimes (G_tF_{t-1}^{-1})'W^{-1}G_tF_{t-1}^{-1}\\ &- \sum_{t=2}^T \psi_t\psi_{t-1}' \otimes (WF_t')^{-1}G_tF_{t-1}^{-1} - \sum_{t=2}^T \psi_{t-1}\psi_t' \otimes (G_tF_{t-1}^{-1})'(F_tW)^{-1}\end{aligned}$$

and

$$\begin{aligned}B_V =&\psi_1' \otimes (y_1 + F_1G_1\psi_0)'(F_1WF_1')^{-1} + \sum_{t=2}^T \psi_t' \otimes (y_t - F_tG_tF_{t-1}^{-1}y_{t-1})'(F_tWF_t')^{-1}\\ &- \sum_{t=2}^T \psi_{t-1}' \otimes (y_t - F_tG_tF_{t-1}^{-1}y_{t-1})'(WF_t')^{-1}G_tF_{t-1}^{-1}\end{aligned}$$

can again be found using the properties of the vec and tr operators. Both of these densities are of the form

$$p(X) \propto |X|^{-(\lambda+p+2)/2} \exp\left[-\frac{1}{2}\Big(\operatorname{tr}(\Lambda X^{-1}) + \operatorname{vec}(L_X)'A\operatorname{vec}(L_X) - 2B\operatorname{vec}(L_X)\Big)\right]$$

where X is a $p \times p$ symmetric and positive definite random matrix, L_X is the Cholesky factor of X so that $L_XL_X' = X$, $\lambda > 0$, Λ is a $p \times p$ symmetric and positive definite matrix, A is a $p^2 \times p^2$ matrix, and B is a $1 \times p^2$ matrix.

The complexity of this density is caused by the interaction between the inverse Wishart prior and the augmented data likelihood in terms of the scaled disturbances for W or for the scaled errors for V. In the local level model, the density still is not a known form and is difficult to sample from, but sampling from it is possible. In this case the log density is

$$\log p(x) = -(\alpha + 1)\log x - ax + b\sqrt{x} - c/x + C$$

for $x > 0$ where C is some constant, $\alpha > 0$ and $c > 0$ are the hyperparameters for x, and $a > 0$ and $b \in \Re$ are parameters that depend on the data, y, the relevant data augmentation (ψ or γ), and the other variable (W or V). We provide two different rejection sampling strategies below that work well under different circumstances, and combine them into a single strategy.

2.G.1 Adaptive rejection sampling

One nice strategy is to use adaptive rejection sampling, e.g. Gilks and Wild (1992). This requires $\log p(x)$ to be concave, which is easy enough to check. The second derivative of $\log p(x)$ is:

$$\frac{\partial^2 \log p(x)}{\partial x^2} = -\frac{1}{4} b x^{-3/2} + (\alpha+1)x^{-2} - 2cx^{-3}.$$

Then we have

$$\frac{\partial^2 \log p(x)}{\partial x^2} < 0 \qquad \Longleftrightarrow \qquad -\frac{b}{4}x^{3/2} + (\alpha+1)x - 2c < 0$$

which would imply that $\log p(x)$ is concave. We can maximize the left hand side of the last equation very easily. When $b \leq 0$ the max occurs at $x = \infty$ such that $LHS > 0$, but when $b > 0$:

$$\frac{\partial LHS}{\partial x} = -\frac{3}{8}bx^{1/2} + \alpha + 1 = 0 \qquad \Longrightarrow \qquad x^{max} = \frac{(\alpha+1)^2}{b^2}\frac{64}{9}.$$

Then we have

$$LHS \leq LHS|_{x=x^{max}} = \frac{(\alpha+1)^3}{b^2}\frac{64}{27} - 2c$$

so that

$$LHS|_{x=x^{max}} < 0 \Longleftrightarrow \frac{(\alpha+1)^3}{b^2}\frac{64}{27} < 2c \qquad \Longleftrightarrow \qquad b > \left(\frac{(\alpha+1)^3}{c}\right)^{1/2}\frac{4\sqrt{2}}{3\sqrt{3}}.$$

This last condition is necessary and sufficient for $\log p(x)$ to be globally (for $x > 0$) concave since $b < 0$ forces $LHS > 0$ for some x. When the condition is satisfied, we can use adaptive rejection sampling — which is already implemented in the `R` package `ars` (Rodriguez, 2009). We input the initial evaluations of $\log p(x)$ at the mode x^{mode} and at $2x^{mode}$ and $0.5x^{mode}$ in order to get the algorithm going.

2.G.2 Rejection sampling on the log scale

When $b \leq \left(\frac{(\alpha+1)^3}{c}\right)^{1/2}\frac{4\sqrt{2}}{3\sqrt{3}}$, which happens often — especially for small T — we need to rely on a different method to sample from $p(x)$. A naive approach would be to construct a normal or t approximation to $p(x)$ and use that as a proposal in a rejection sampler. It turns out that this is often very inefficient, but for $z = \log(x)$ the approach works well. Note that

$$p_z(z) = p_x(e^z)e^z$$

so that we can write the log density of z as (dropping the subscripts):

$$\log p(z) = -ae^z + be^{z/2} - \alpha z - ce^{-z}.$$

The mode of this density z^{mode} can be easily found numerically, and the second derivative is:

$$\frac{\partial^2 \log p(z)}{\partial z^2} = -ae^z + \frac{b}{4}e^{z/2} - ce^{-z}.$$

The t approximation then uses the proposal distribution p

$$t_v\left(z^{mode}, \left[-\left.\frac{\partial^2 \log p(z)}{\partial z^2}\right|_{z=z^{mode}}\right]^{-1}\right).$$

In practice choosing degrees of freedom $v = 1$ works very well over the region of the parameter space where adaptive rejection sampling cannot be used. We can easily use this method when adaptive rejection sampling does not work, then transform z back to x. It remains to check that the tails of t distribution dominate the tails of our target distribution. Let $\log q(z)$ denote the log density of the proposal distribution. Then we need

$$\log p(z) - \log q(z) \leq M$$

for some constant M, i.e.

$$-ae^z + be^{z/2} - \alpha z - ce^{-z} - \left(\frac{v+1}{2}\right)\log\left[1 + \frac{1}{v}\left(\frac{z-\mu}{\sigma}\right)^2\right] \leq M$$

where $a > 0$, $c > 0$, $\alpha > 0$, $v > 0$, $\sigma > 0$, and $b, \mu \in \Re$. We can rewrite the LHS as

$$e^{z/2}(b - ae^{z/2}) - \alpha z - ce^{-z} - \left(\frac{v+1}{2}\right)\log\left[1 + \frac{1}{v}\left(\frac{z-\mu}{\sigma}\right)^2\right].$$

So as $z \to \infty$ this quantity goes to $-\infty$ since the first term will eventually become negative no matter the value of b, and all other terms are always negative. Similarly as $z \to -\infty$ this quantity goes to $-\infty$. Now pick any interval (z_1, z_2) such that outside of the interval, $LHS < \epsilon$. Since treated as a function of z the LHS is clearly continuous, it attains a maximum on this interval, and thus is bounded.

2.G.3 Intelligently choosing a rejection sampler

In practice, adaptive rejection sampling is relatively efficient for $p_x(x)$ but inefficient for $p_z(z)$ — so much so that rejection sampling with the t approximation for $p_z(z)$ is more efficient. To minimize computation time, it is best to use adaptive rejection sampling for $p_x(x)$ when the concavity condition is satisfied. When it is not, the t approximation works well enough.

2.H Efficiently drawing from $p(W|V, \tilde{\gamma}, y)$ and $p(V|W, \tilde{\psi}, y)$ in the LLM

Both the density of $\log(W)|V, \tilde{\gamma}, y$ and the density of $\log(V)|W, \tilde{\psi}, y$ have the following form:

$$p(z) \propto \exp\left[-\alpha z - ae^{-z} + be^{-z/2} - ce^{z}\right].$$

where $\alpha > 0$, $a > 0$, $c > 0$, and $b \in \Re$. The log density is:

$$\log p(z) = -\alpha z - ae^{-z} + be^{-z/2} - ce^{z} + C$$

where C is some constant. We only provide one strategy for rejection sampling from this density: the t approximation. Similar reasoning to the previous subsection above shows that we can use a t distribution as a proposal in a rejection sampler for this density. Now we choose the location parameter by maximizing $\log p(z)$ in z numerically to find the mode, z^{mode}. Next the second derivative of $\log p(z)$ is given by

$$\frac{\partial^2 \log p(z)}{\partial z^2} = -ae^{-z} + \frac{b}{4}e^{-z/2} - ce^{z}.$$

We then set the scale parameter to be

$$-\left[\left.\frac{\partial^2 \log p(z)}{\partial z^2}\right|_{z=z^{mode}}\right]^{-1}$$

as in the normal approximation, and the degrees of freedom parameter to $v = 1$. This rejection sampler is tolerably efficient for our purposes, but there is much room for improvement.

2.I Equivalence of CIS and GIS in the DLM

The CIS algorithm consists of the following steps:

$$[\psi|V^{(k)},W^{(k)}] \to [V^{(k+0.5)}|W^{(k)},\psi] \to [\tilde{\psi}|V^{(k+0.5)},W^{(k)},\psi] \to [V^{(k+1)}|W^{(k)},\tilde{\psi}] \to$$
$$[\tilde{\gamma}|V^{(k+1)},W^{(k)},\tilde{\psi}] \to [W^{(k+0.5)}|V^{(k+1)},\tilde{\gamma}] \to [\gamma|V^{(k+1)},W^{(k+0.5)},\tilde{\gamma}] \to [W^{(k+1)}|V^{(k+1)},\gamma].$$

In the fourth step of line one and the second step of line two, each of those densities would be unchanged if we conditioned on θ instead of $\tilde{\psi}$ on the first line or $\tilde{\gamma}$ on the second line. So the CIS algorithm above is equivalent to the following:

$$[\psi|V^{(k)},W^{(k)}] \to [V^{(k+0.5)}|W^{(k)},\psi] \to [\theta|V^{(k+0.5)},W^{(k)},\psi] \to [V^{(k+1)}|W^{(k)},\theta] \to$$
$$[W^{(k+0.5)}|V^{(k+1)},\theta] \to [\gamma|V^{(k+1)},W^{(k+0.5)},\theta] \to [W^{(k+1)}|V^{(k+1)},\gamma].$$

Now since V and W are conditionally independent given θ and y, the last step of line one and the first step of line 2 can be switched:

$$[\psi|V^{(k)},W^{(k)}] \to [V^{(k+0.5)}|W^{(k)},\psi] \to [\theta|V^{(k+0.5)},W^{(k)},\psi] \to [W^{(k+0.5)}|V^{(k+0.5)},\theta] \to$$
$$[V^{(k+1)}|W^{(k+0.5)},\theta] \to [\gamma|V^{(k+1)},W^{(k+0.5)},\theta] \to [W^{(k+1)}|V^{(k+1)},\gamma].$$

Next V's conditional density is the same whether we condition on θ or γ, so we can do the V step between the γ step and the W step in line two. Similarly we can move the W step to between the V step and the θ step in line one. This yields:

$$[\psi|V^{(k)},W^{(k)}] \to [V^{(k+0.5)}|W^{(k)},\psi] \to [W^{(k+0.5)}|V^{(k+0.5)},\psi] \to$$
$$[\gamma|V^{(k+0.5)},W^{(k+0.5)},\psi] \to [V^{(k+1)}|W^{(k+0.5)},\gamma] \to [W^{(k+1)}|V^{(k+1)},\gamma].$$

This is actually a SE-SD GIS algorithm, so the CIS sampler we started with is equivalent to SE-SD GIS. Since we do not expect the order in which the DAs appear in a GIS algorithm to matter, CIS should have the same mixing and convergence properties as the SD-SE GIS algorithm we constructed.

2.J Partial CIS algorithms in the DLM

In addition to the GIS and CIS algorithms discussed in the main body of the article, Yu and Meng (2011) also introduce *partial CIS* algorithms. While a CIS algorithm interweaves in separate Gibbs steps for each sub-vector of the parameter, a partial CIS algorithm has a usual Gibbs step for at least one of the parameter vectors. For example, suppose that the model parameter is $\phi = (\phi_1, \phi_2)$, and γ_1, γ_2, and θ are available DAs. Then a partial CIS algorithm using these DAs is

Algorithm: partial CIS. *Partial Componentwise Interweaving Strategy*

$$[\gamma_1|\phi_1^{(k)}, \phi_2^{(k)}] \rightarrow [\phi_1^{(k+0.5)}|\phi_2^{(k)}, \gamma_1] \rightarrow [\gamma_2|\phi_1^{(k+0.5)}, \phi_2^{(k)}, \gamma_1] \rightarrow [\phi_1^{(k+1)}|\phi_2^{(k)}, \gamma_2] \rightarrow$$
$$[\theta|\phi_1^{(k+1)}, \phi_2^{(k)}, \gamma_2] \rightarrow [\phi_2^{(k+1)}|\phi_1^{(k+1)}, \theta].$$

The first line is an interweaving step for ϕ_1 while the second line is a standard Gibbs step for ϕ_2. Partial CIS algorithms are easier to construct than full CIS algorithms at the cost of slower convergence Yu and Meng (2011).

In the DLM we can construct two partial CIS algorithms using the wrongly-scaled DAs in much the same way they were used to construct the full CIS algorithm. The first algorithm interweaves for W using the scaled disturbances, γ, and the wrongly-scaled disturbances, $\tilde{\gamma}$:

$$[\theta|V^{(k)}, W^{(k)}] \rightarrow [V^{(k+1)}|W^{(k)}, \theta] \rightarrow$$
$$[W^{(k+0.5)}|V^{(k+1)}, \theta] \rightarrow [\gamma|V^{(k+1)}, W^{(k+0.5)}, \theta] \rightarrow [W^{(k+1)}|V^{(k+1)}, \gamma].$$

As in the construction of the full CIS algorithm, we use θ instead of $\tilde{\gamma}$ in the second line since $p(W|V, \tilde{\gamma}) = p(W|V, \theta)$. Using an argument similar to that used in Section 2.I, we can show that this partial CIS algorithm is equivalent to the SD-State GIS algorithm.

Analogously, we can use the scaled errors, ψ, and the wrongly-scaled errors, $\tilde{\psi}$, to construct a partial CIS algorithm that interweaves for V:

$$[\psi|V^{(k)}, W^{(k)}] \rightarrow [V^{(k+0.5)}|W^{(k)}, \psi] \rightarrow [\theta|V^{(k+0.5)}, W^{(k)}, \psi] \rightarrow [V^{(k+1)}|W^{(k)}, \theta] \rightarrow$$
$$[W^{(k+1)}|V^{(k+1)}, \theta].$$

This algorithm is equivalent to the SE-State GIS algorithm.

2.K Using posterior correlations to understand patterns of ESP

Most of the patterns in Figures 2.L.1, 2.L.2, and 2.L.3 in the next section can be explained by Figure 2.K.1, which contains the estimated posterior correlations between various functions of parameters estimated using the simulations from the Triple-Alt sampler for a time series with $T = 100$. We omit a similar analysis for $T = 10$ and $T = 1000$. The state sampler consists of two steps — a draw of θ given V and W, and a draw of (V, W) given θ. From Section 2.D.1 we have that conditional on θ, V and W are independent in the posterior and each has an inverse gamma distribution that depends on the states only through the second parameter:

$$b_V \equiv \beta_V + \sum_{t=1}^{T}(y_t - \theta_t)^2/2 \qquad\qquad b_W \equiv \beta_W + \sum_{t=1}^{T}(\theta_t - \theta_{t-1})^2/2.$$

So we can view (b_V, b_W) as the data augmentation instead of θ and thus the state sampler is

$$[b_V, b_W|V^{(k)}, W^{(k)}] \to [V^{(k+1)}, W^{(k+1)}|b_V, b_W].$$

Thus the dependence between (V, W) and (b_V, b_W) in the posterior will determine how much the state sampler moves in a given iteration and, in particular, it is possible that V and W have very different serial dependence from each other since we are drawing them jointly. When the dependence between V and b_V is high, the (V, W) step will hardly move V even if it drastically moves W since V and W are independent. However, the (b_V, b_W) step may move both elements a moderate amount since they both depend on (V, W).

In Figure 2.K.1 we see that the posterior correlation between V and b_V is high in magnitude and positive when $R^* > 1$ while the posterior correlation between V and b_W is moderate to low and negative. When R^* is large enough though, the posterior correlation between V and b_W evaporates. Similarly when $R^* < 1$ the posterior correlation between W and b_W is high and positive and the posterior correlation between W and b_V is high and negative. Again as R^* becomes large enough the correlation between W and b_V goes to zero. So when $R^* > 1$, the draw of (b_V, b_W) is unlikely to move b_V much since b_V is so highly correlated with V and essentially uncorrelated with b_W, but b_W is essentially uncorrelated with W and negatively correlated with V so b_W is likely to move a fair amount. Furthermore the draw of V is highly correlated with b_V while the draw of W is essentially independent of b_W (and the draws of V and W are independent conditional on b_V and b_W). Thus when $R^* > 1$ we should expect high serial dependence for V and low serial dependence for W, and so low ESP for V and high ESP for W, which is exactly what we see in Figure 2.L.2. By similar reasoning when $R^* < 1$, we should expect low serial dependence for V and high serial dependence for W and thus high ESP for V and low ESP for W, which can also be seen in Figure 2.L.2.

For the SD sampler, things are a bit more complicated. The draw of $V|W,\gamma$ still depends on b_V since it is the same inverse gamma draw as in the state sampler, but the draw of $W|V,\gamma$ now depends on a_γ and b_γ defined in Section 2.D.2 as

$$a_\gamma \equiv \frac{1}{2V}\sum_{t=1}^{T}\left(\sum_{j=1}^{t}\gamma_j\right)^2 \qquad\qquad b_\gamma \equiv \frac{1}{V}\sum_{t=1}^{T}(y_t-\gamma_0)\left(\sum_{j=1}^{t}\gamma_j\right).$$

So the dependence between V and b_V determines how much the chain moves in the V step, and the dependence between W and (a_γ, b_γ) determines how much it moves in the W step. The dependence between (V,W) and γ determines how much the chain moves in the DA step, but we can view this step instead as a draw of b_V in which case the dependence between W and b_V determines how much the chain moves in that step. So if any one of these steps has high dependence, we should expect every element of the chain, and (V,W) in particular, to have high serial dependence in the chain. The SE sampler is analogous to the SD sampler except with b_W, a_ψ and b_ψ where

$$a_\psi = \frac{1}{2W}\sum_{t=1}^{T}(\mathcal{L}\psi_t)^2 \qquad\qquad b_\psi = \frac{1}{W}\sum_{t=1}^{T}(\mathcal{L}\psi_t\mathcal{L}y_t).$$

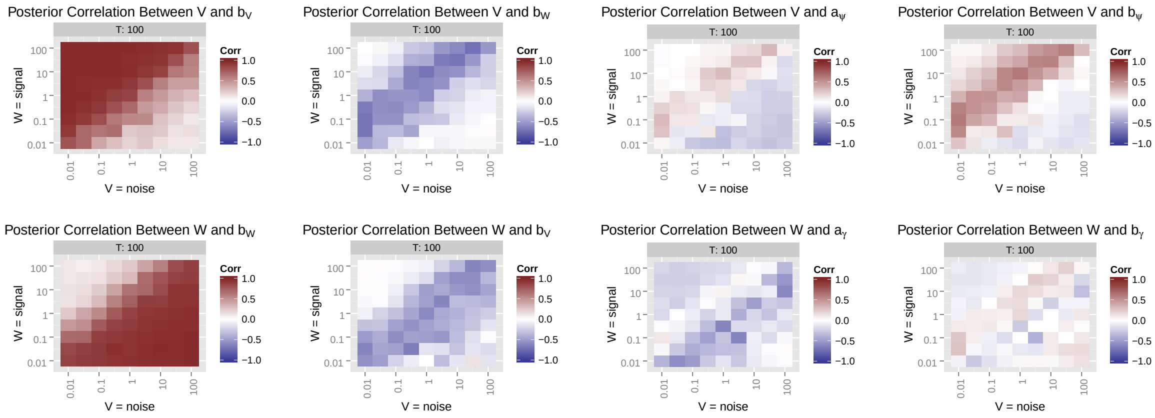


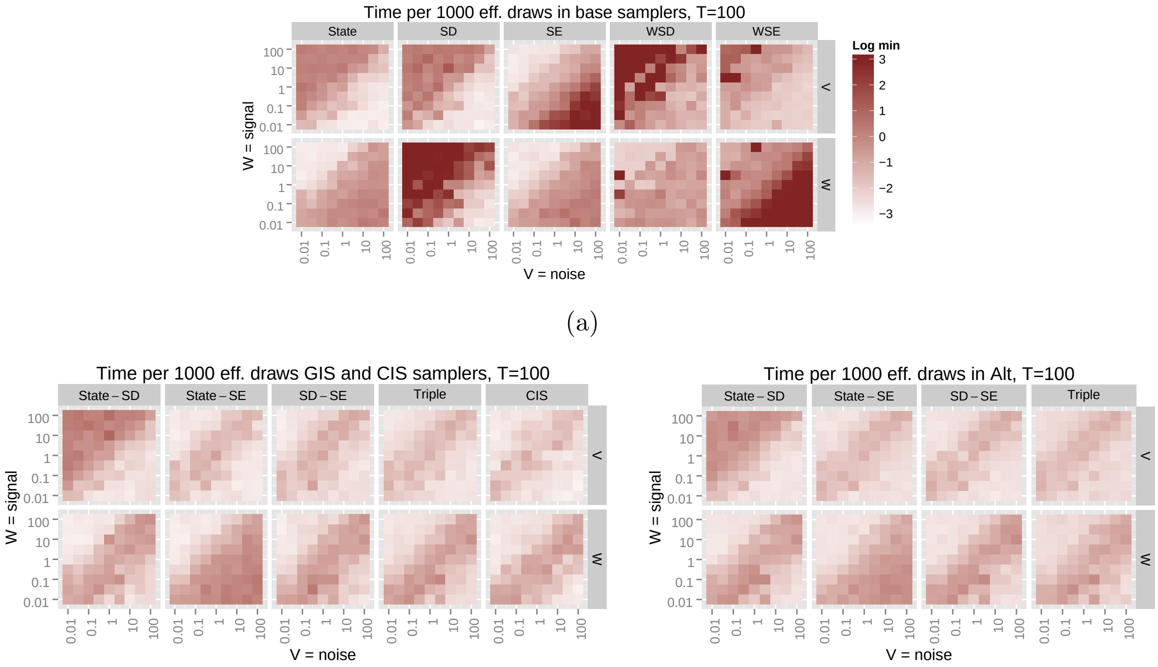
Figure 2.K.1: Posterior correlation between V or W and b_V, b_W, a_γ, b_γ, a_ψ or b_ψ. X and Y axes indicate the true values of V and W respectively for the simulated data with $T = 100$.

In order to analyze the SD sampler, first suppose $R^* > 1$. Then from Figure 2.K.1 b_V has high correlation with V and low correlation with W, so the draw of b_V should not move the chain much. Next, the draw of V should again not move the chain much because of the high correlation between V and b_V. Finally the draw of W has a fair chance to move the chain because it has low correlation with both a_γ and b_γ. But this has little impact on b_V and thus the entire chain since b_V is so highly correlated with V but hardly correlated with W. So when $R^* > 1$, we should expect high serial dependence and low ESP for V. We should also expect similar behavior for W since the entire chain is hardly moving so W's hyperparameters are hardly moving. This is roughly what we see in Figure 2.L.2, though this reasoning does not allow us to predict which of V and W will have lower ESP. When $R^* < 1$ the posterior correlation in each of the steps is broken, though in the W step the correlation between W and both a_γ and b_γ becomes negative and somewhat high in magnitude. Here we should not expect less serial dependence in V or W, but we should perhaps expect higher ESP's since negatively correlated draws decrease Monte Carlo standard error. Indeed, we see ESP's near one for both variances in Figure 2.L.2. The SE sampler is analogous to the SD sampler and a similar analysis applies — the posterior correlations between V or W and b_W, a_ψ or b_ψ in Figure 2.K.1 roughly predict the ESP of the SE sampler in Figure 2.L.2. When one or more of the correlations are high, ESPs for V and W are low while when all of the correlations are low, both ESPs are high. We omit a similar analysis of the wrongly-scaled samplers for brevity, but note that their behavior will allows us to predict the behavior of the CIS sampler.

2.K.1 Computational time

From a practical standpoint a more important question than how well the chain mixes is the full computational time required to adequately characterize the target posterior distribution. In order to investigate this, we compute the natural log of the average time in minutes required for each sampler to achieve an effective sample size of 1000 — in other words the log minutes per 1000 effective draws. All simulations were performed on a server with Intel Xeon X5675 3.07 GHz processors. While different systems will yield different absolute times, the relative times should be similar. Figure 2.K.2 contains plots of the log minutes per 1000 effective draws for both V and W and for each of the samplers.

For $T = 100$ the pattern we saw for ESP also appears for log minutes per 1000 effective



(a)

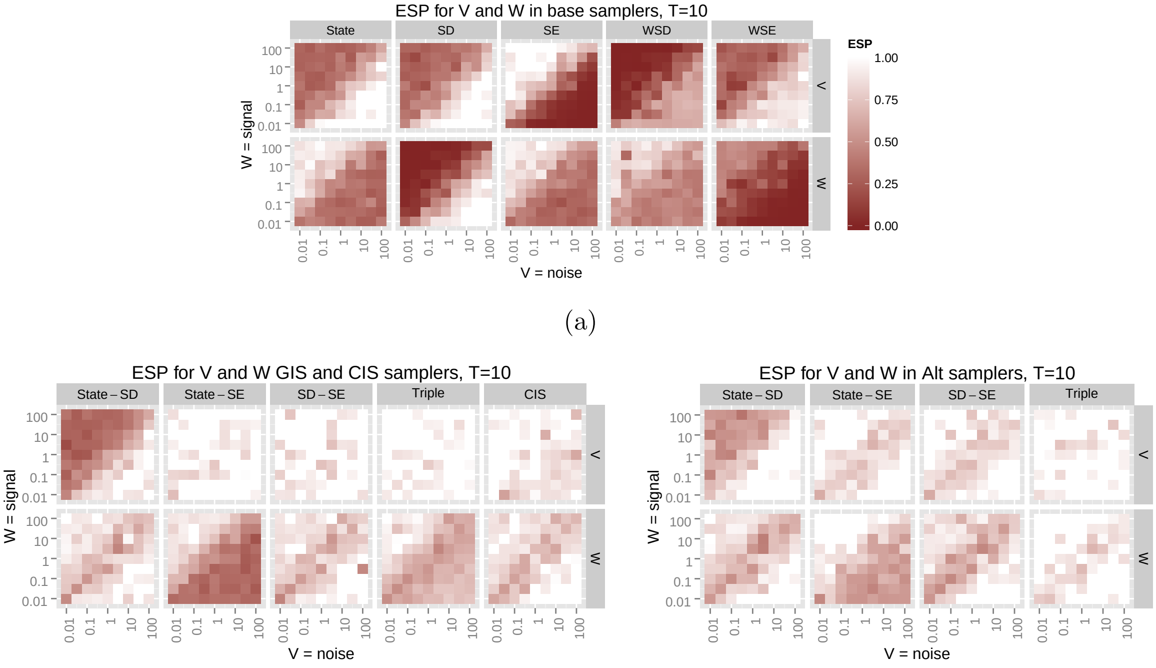
(b)

(c)

Figure 2.K.2: Log of the time in minutes per 1000 effective draws in the posterior sampler for V and W, for $T = 100$ in each sampler. Figure 2.K.2a contains the base samplers, Figure 2.K.2b contains the GIS and CIS samplers, while Figure 2.K.2c contains the Alt samplers. Log times larger than three log min are rounded down to three for plotting purposes.

draws. The State sampler becomes slow to reach 1000 effective draws for V when $R^* > 1$ and for W when $R^* < 1$. The SD and SE samplers behave as expected — the SD sampler is slow for both V and W when $R^* > 1$ while the SD sampler is slow for both V and W when $R^* < 1$. The SD-SE GIS, Triple GIS and CIS algorithms appear to be the big winners here and are almost indistinguishable. All three algorithms are slightly slower for both V and W when R^* is near one, though for larger T, when R^* is near or below one all three are slow for W (plots available in Section 2.L). Compared to the state sampler, all three offer large gains over most of the parameter space. There appears to be no difference between a GIS algorithm and the corresponding alternating algorithm in terms of log time per 1000 effective draws, so the SD-SE Alt and Triple Alt algorithms are both just as efficient as the best interweaving algorithms. This may not always be the case though — the GIS version of an algorithm is computationally cheaper than the Alt version since it consists of three of the four same steps, and in the fourth step the Alt algorithm has to obtain a random draw while the GIS algorithm typically only has to make a transformation. The more expensive that draw is relative to the transformation, the faster GIS will be relative to Alt.

2.L Plots for all values of T

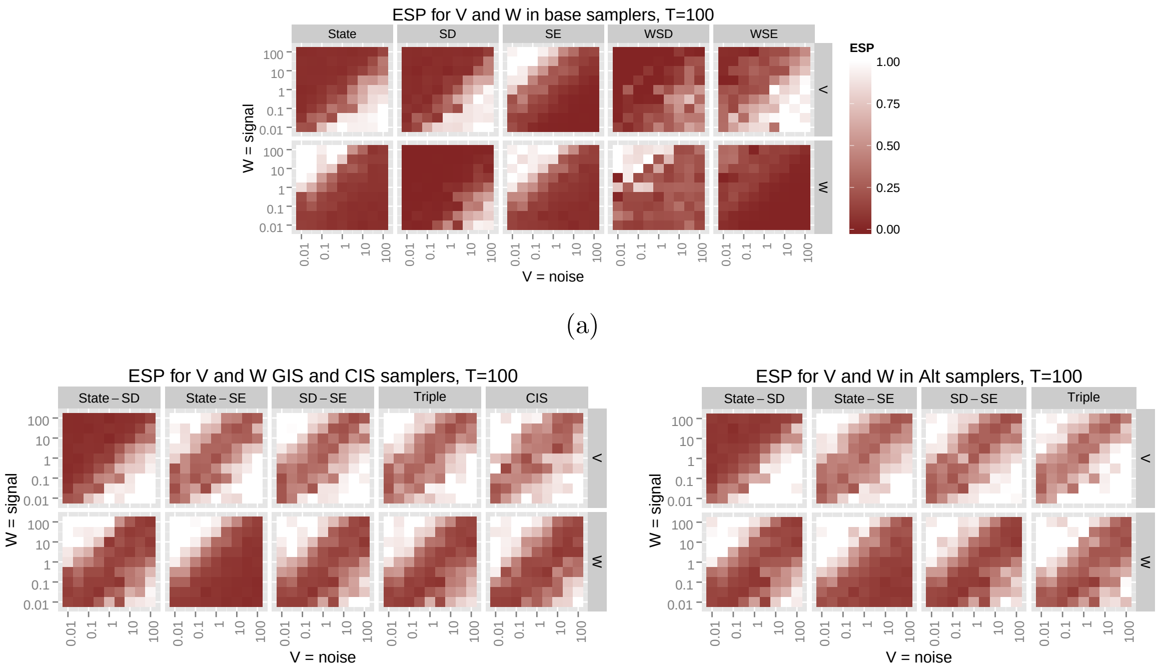


(a)

(b)

(c)

Figure 2.L.1: Effective sample proportion in the posterior sampler for a time series of length $T = 10$, for V and W in the each sampler. Figure 2.L.1a contains ESP for V and W for the base samplers, Figure 2.L.1b contains ESP in the GIS and CIS samplers, and Figure 2.L.1c contains ESP in the Alt samplers. X and Y axes indicate the true values of V and W respectively for the simulated data. Note that the signal-to-noise ratio is constant moving up any diagonal. In the upper left the signal is high, in the lower right the noise is high.

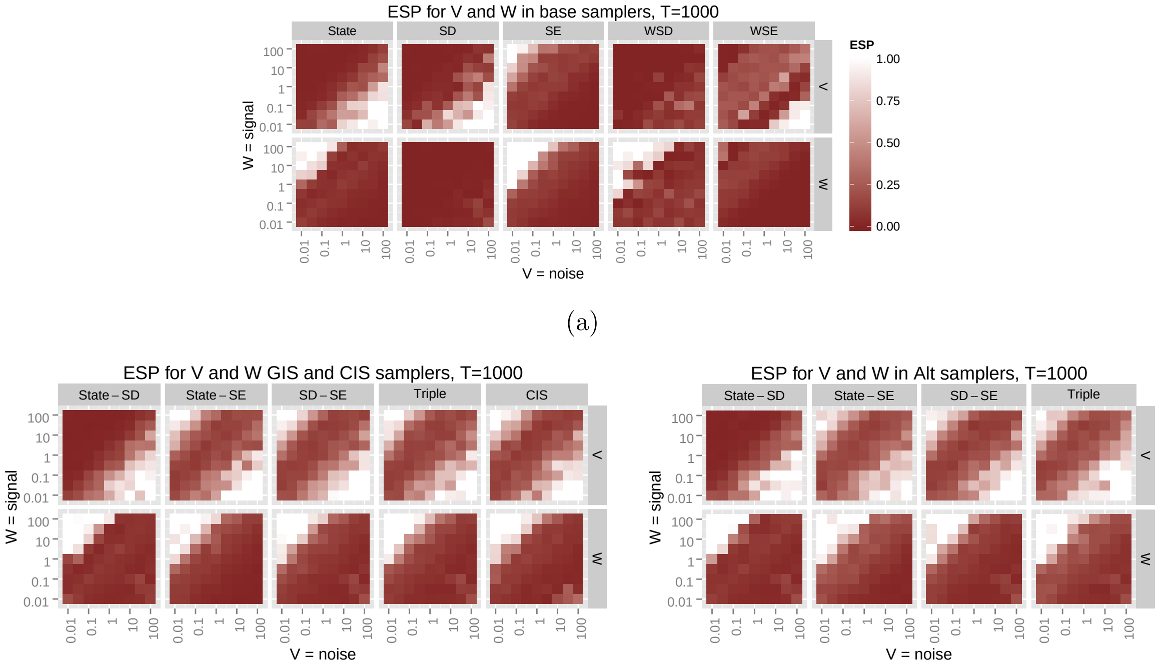


(a)

(b)

(c)

Figure 2.L.2: Effective sample proportion in the posterior sampler for a time series of length $T = 100$, for V and W in the each sampler. Figure 2.L.2a contains ESP for V and W for the base samplers, Figure 2.L.2b contains ESP in the GIS and CIS samplers, and Figure 2.L.2c contains ESP in the Alt samplers. X and Y axes indicate the true values of V and W respectively for the simulated data. Note that the signal-to-noise ratio is constant moving up any diagonal. In the upper left the signal is high, in the lower right the noise is high.

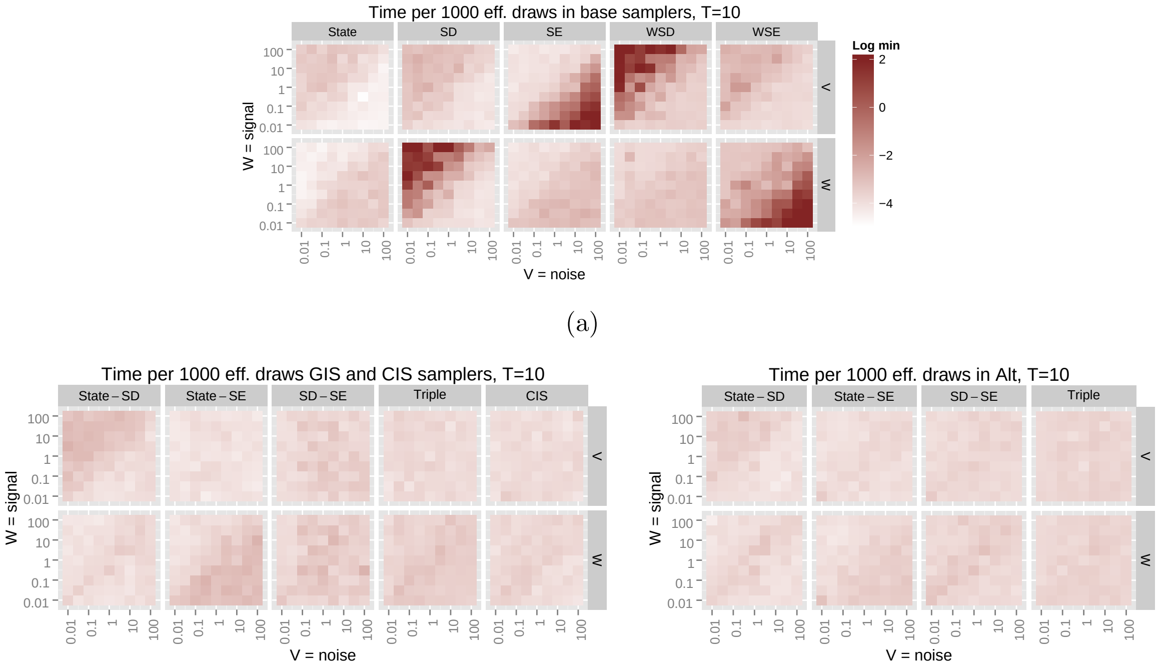


(a)

(b)

(c)

Figure 2.L.3: Effective sample proportion in the posterior sampler for a time series of length $T = 1000$, for V and W in the each sampler. Figure 2.L.3a contains ESP for V and W for the base samplers, Figure 2.L.3b contains ESP in the GIS and CIS samplers, and Figure 2.L.3c contains ESP in the Alt samplers. X and Y axes indicate the true values of V and W respectively for the simulated data. Note that the signal-to-noise ratio is constant moving up any diagonal. In the upper left the signal is high, in the lower right the noise is high.

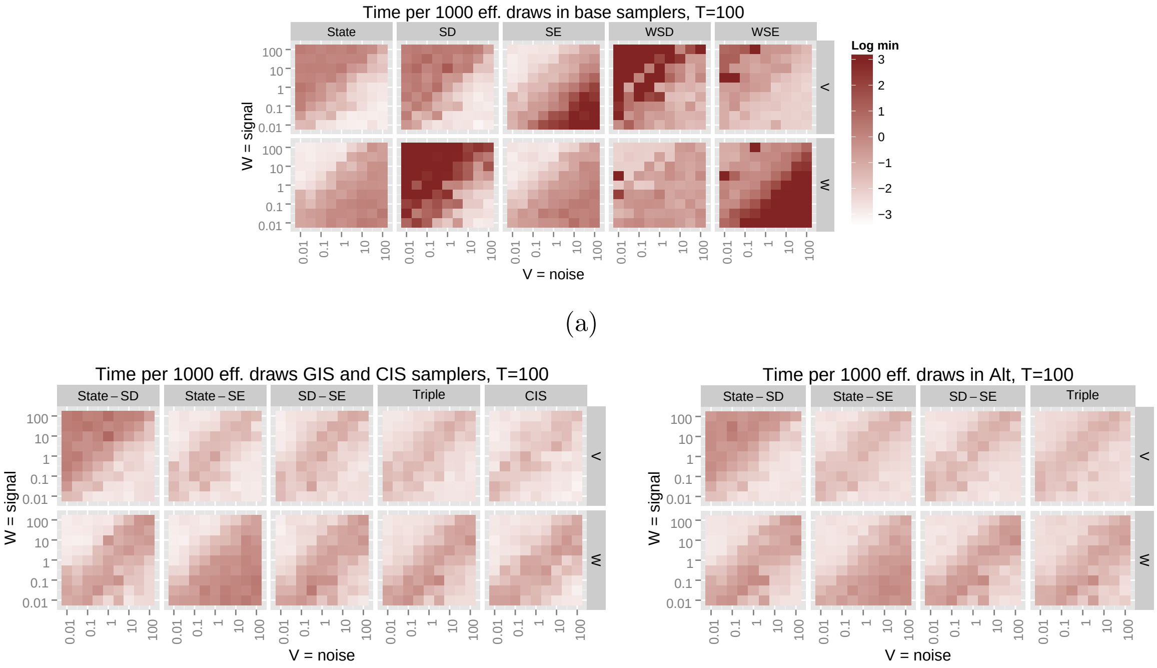


(a)

(b)

(c)

Figure 2.L.4: Log of the time in minutes per 1000 effective draws in the posterior sampler for V and W, for $T = 10$ in each sampler. Figure 2.L.4a contains the base samplers, Figure 2.L.4b contains the GIS and CIS samplers, while Figure 2.L.4c contains the Alt samplers. Log times larger than three log min are rounded down to three for plotting purposes.

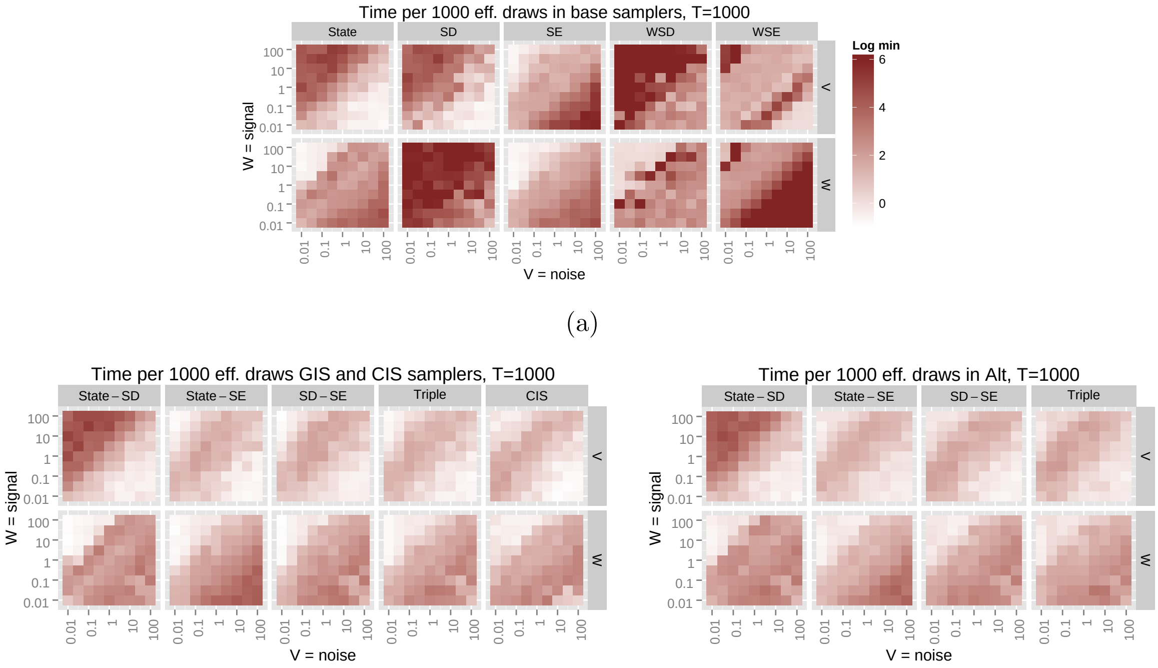


(a)

(b)

(c)

Figure 2.L.5: Log of the time in minutes per 1000 effective draws in the posterior sampler for V and W, for $T = 100$ in each sampler. Figure 2.L.5a contains the base samplers, Figure 2.L.5b contains the GIS and CIS samplers, while Figure 2.L.5c contains the Alt samplers. Log times larger than three log min are rounded down to three for plotting purposes.



(a)

(b)

(c)

Figure 2.L.6: Log of the time in minutes per 1000 effective draws in the posterior sampler for V and W, for $T = 1000$ in each sampler. Figure 2.L.6a contains the base samplers, Figure 2.L.6b contains the GIS and CIS samplers, while Figure 2.L.6c contains the Alt samplers. Log times larger than three log min are rounded down to three for plotting purposes.

CHAPTER 3. APPLICATION OF INTERWEAVING IN DLMS TO AN EXCHANGE AND SPECIALIZATION EXPERIMENT

A paper to appear in *Bayesian Statistics from Methods to Models and Applications*

Abstract

Markov chain Monte Carlo is often particularly challenging in dynamic models. In statespace models, the data augmentation algorithm (Tanner and Wong, 1987) is a commonly used approach, e.g. Frühwirth-Schnatter (1994) and Carter and Kohn (1994) in dynamic linear models. Using two data augmentations, Yu and Meng (2011) introduces a method of "interweaving" between the two augmentations in order to construct an improved algorithm. Picking up on this, Simpson et al. (2014) introduces several new augmentations for the dynamic linear model and builds interweaving algorithms based on these augmentations. In the context of a multivariate model using data from an economic experiment intended to study the disequilibrium dynamics of economic efficiency under a variety of conditions, we use these interweaving ideas and show how to implement them simply despite complications that arise because the model has latent states with a higher dimension than the data.

3.1 Introduction

Several innovations on the original data augmentation (DA) algorithm (Tanner and Wong, 1987) have been proposed in the literature, see e.g. Van Dyk and Meng (2001) for a thorough overview. One such innovation is the notion of interweaving two separate DAs together (Yu and Meng, 2011). This general idea has been picked up on in the dynamic setting by Kastner and Frühwirth-Schnatter (2014) in stochastic volatility models and Simpson et al. (2014) in dynamic linear models (DLMs). Previous literature exploring alternate DAs in statespace models includes Pitt and Shephard (1999) for the AR(1) plus noise model, Frühwirth-Schnatter (2004) for dynamic regression models, Strickland et al. (2008) for nonlinear models including the stochastic volatility model, Frühwirth-Schnatter and Sögner (2008) for the stochastic volatility model, and Frühwirth-Schnatter and Wagner (2010) in the context of model selection, though there are many more.

Much of this literature focuses on stochastic volatility and similar models (Shephard, 1996; Frühwirth-Schnatter and Sögner, 2003; Roberts et al., 2004; Bos and Shephard, 2006; Strickland et al., 2008; Frühwirth-Schnatter and Sögner, 2008; Kastner and Frühwirth-Schnatter, 2014), though Simpson et al. (2014) focuses on DLMs and develops several new data augmentations for a general class DLMs. Using these DAs, they construct several Markov chain Monte Carlo (MCMC) algorithms including interweaving algorithms based on Yu and Meng (2011), and compare these algorithms in a simulation study using the local level model. We seek to illustrate the interweaving methods introduced in Simpson et al. (2014) in the context of model that can be expressed either as a hierarchical DLM with equal state and data dimensions or simply a DLM with a state dimension larger than the data dimension. The latter representation in particular provides some difficulty in directly applying the methods discussed in Simpson et al. (2014), though we show how to easily overcome this.

Throughout this article we will use the notation $p(.|.)$ to denote the potentially conditional density of the enclosed random variables, $x_{1:T} = (x_1, \ldots, x_T)'$ when x_t is a scalar, and $x_{1:T} = (x'_1, \ldots, x'_T)'$ when x_t is a column vector so that $x_{1:T}$ is also a column vector in both cases. The rest of this paper is organized as follows: Section 3.2 will describe the data which arise from a series of economics experiments, and Section 3.3 will describe the model we wish to fit to these data. Section 4.6 will cover how to do MCMC in this model, including a fairly standard DA algorithm and an interweaving algorithm based on the ideas in Simpson et al. (2014) and Yu and Meng (2011). Finally, Section 3.5 will contain the results of fitting the model using both algorithms and Section 3.6 will briefly conclude.

3.2 Data

Economists are interested in determining the factors that affect the level of economic efficiency within an economy where economic efficiency can roughly be defined as the proportion of maximum possible dollar value of the total benefits to all actors in the economy, also known as Kaldor-Hicks efficiency and based on compensating variation (Kaldor, 1939; Mas-Colell et al., 1995). Studying this in the real world is messy and difficult in part because computing this proportion is nontrivial. In addition, most economic models only allow the analysis of equilibrium efficiency. To the extent that efficiency dynamics are studied, they are typically studied as equilibrium dynamics. Disequilibrium dynamics are difficult to study but potentially important. In order to avoid these difficulties while still learning something about the disequilibrium dynamics of efficiency, a series of laboratory experiments were designed and run by a group of experimental economists in order to explore what factors impact the disequilibrium dynamics of a small laboratory economy (Crockett et al., 2009; Kimbrough et al., 2010). What follows is a brief description of these experiments.[1]

In a single session of the experiment, 2, 4, or 8 subjects are recruited to participate, depending on the treatment. Each subject sits at a computer visually isolated from the rest of the subjects. On the computer, each subject controls an avatar in a virtual village where they can interact with the other subjects in the experiment. At any time during the experiment, subjects can communicate with each other by typing into a chat window. Each subject in a given session has control over a house and a field within the village and can view each other subject's house and field. The experiment runs for 40 periods, each lasting 100 seconds. Within a period, each

[1] For a more detailed description of the experimental design, see Crockett et al. (2009) especially, but also Kimbrough et al. (2010).

subject has to make a production decision and a consumption decision. Every seventh period is a 'rest' period where no production or consumption takes place, but the subjects can still communicate. This results in 35 periods of production and consumption.

There are two types of goods in this world, each produced in a subject's field: *red* and *blue*, and two types of subjects: *odd* and *even*. Half of the subjects are *odd* and half are *even*. Both *odd* and *even* subjects can produce both types of goods and earn money for consuming both types of goods, but they produce and consume in different ways. *Odd* subjects must *red* and *blue* in a fixed proportion of 1 *red* for every 3 *blue* to earn U.S. cents. *Even* subjects, on the other hand, must consume 2 *red* for every 1 *blue* to earn U.S. cents. However, *even* subjects are more effective at producing *blue* while *odd* subjects are more effective at producing *red*. Production occurs in the first 10 seconds of a period where each subject must decide how much of that time to devote to producing *red* and *blue* respectively using a slider on their screen. The last 90 seconds of the period is reserved for trading and consumption, though subjects have to discover that they may trade by noticing that they can use their mouse to drag and drop red and/or blue icons (representing one unit of *red* or *blue* respectively) onto another subject's house. The maximum level of village wide production takes place when each subject spends 100% of their time producing the good that they can produce the most efficiently, i.e. *odd* subjects produce only *red* and *even* subjects produce only *blue*. Maximum consumption and thus maximum profit occurs when under maximum production and the subjects trade extensively with each other. In every period, the efficiency level of the village is recorded.

A wide variety of treatments were applied to the various sessions of this experiment, including variations on group size and group formation, various levels of knowledge about the subject's own production function, allowing theft or not and if so, whether mechanisms for punishing theft are available. See Crockett et al. (2009) and Kimbrough et al. (2010) for a detailed description of these treatments. Each treatment consists of several replications — anywhere from four to six. The challenge, then, is to model a time series of proportions that takes into account the nested structure of the replications within the treatments. To deal with the proportions, we simply transform the efficiencies to the real line using the logit transformation, i.e. $\text{logit}(x) = \log(x/(1-x))$. In some replications of some treatments, efficiencies of

100% or 0% are obtained which causes a problem the logit and other plausible transformations. We only consider the Steal treatment of Kimbrough et al. (2010) in order to avoid this issue and simplify the model a bit. This allows for a useful illustration of Simpson et al. (2014) without too much additional complication. In short, the Steal treatment uses the Build8 structure from previous treatments that starts the subjects in four groups of two for several periods, then combines them into two groups of four for several more periods, then finally combines the groups into a single group of eight for the rest of the experiment. The only change from this structure is that Steal allows subjects to steal either of the goods from each other, which was not possible in previous treatments. Reference Kimbrough et al. (2010) has further details about this treatment and the various treatments it spawned in order to see what institutional arrangements help subjects prevent theft.

3.3 Model

Let $j = 1, 2, \ldots, J$ denote the replications of the treatment and $t = 1, 2, \ldots, T$ denote periods within these replications. Then let $y_{j,t}$ denote the observed logit efficiency of the j'th replication in the t'th period. Consider the following model

$$\begin{aligned} y_{j,t} =& \mu_t + \theta_{j,t} + v_{j,t} && \text{(observation equation)} \\ \theta_{j,t} =& \theta_{j,t-1} + w_{j,t} && \text{(replication level system equation)} \\ \mu_t =& \mu_{t-1} + u_t && \text{(treatment level system equation)} \end{aligned} \tag{3.1}$$

for $j = 1, 2, \ldots, J$ and $t = 1, 2, \ldots, T$, where $(v_{1:J,1:T}, w_{1:J,1:T}, u_{1:T})$ are mutually independent with $v_{j,t} \sim N(0, V_j)$, $w_{j,t} \sim N(0, W_j)$, and $u_t \sim N(0, U)$. The latent treatment level logit efficiency is represented by μ_t and evolves via a random walk. On the replication level, $\theta_{j,t}$ represents replication j's deviation from the the treatment logit efficiency in period t which also evolves over time via a random walk. Then $\mu_t + \theta_{j,t}$ is replication level latent logit efficiency. Finally $y_{j,t}$ represents the observed logit efficiency of replication j in period t. The amount replication j's path tends to differ from the treatment level path is controlled by the relative values of W_j and U — the larger W_j is relative to U, the less replication j's path is affected by the treatment level path. Finally, V_j represents how much of the change in logit efficiency is independent of previous changes. The relative size of V_j compared to W_j and U tells us how much logit efficiency changes over time due to independent sources of error relative to the replication and treatment level evolutionary processes. So in this sense, $\mu_t + \theta_{j,t}$ can be seen as the portion of replication j's logit efficiency that is carried on into the next period, or sustainable in a certain sense.

Another way to represent this model is by writing it in terms of the replication level latent logit efficiencies, $\phi_{j,t} = \mu_t + \theta_{j,t}$. Under this parameterization, the model is

$$\begin{aligned} y_{j,t} =& \phi_{j,t} + v_{j,t} \\ \phi_{j,t} =& \phi_{j,t-1} + w_{j,t} + u_t \end{aligned} \tag{3.2}$$

for $j = 1, 2, \ldots, J$ and $t = 1, 2, \ldots, T$ where we substitute u_t in for $\mu_t - \mu_{t-1}$. This representation shows us that the replication level latent logit efficiencies evolve according to a correlated random walk where U controls the degree of correlation between the replications.

Finally, if we let $\theta_t = (\mu_t, \theta'_{1:J,t})'$, $y_t = y_{1:J,t}$, $V = diag(V_1, \ldots, V_J)$, $W = diag(U, W_1, \ldots, W_J)$, and $F = [1_{J\times 1}\ \mathrm{I}_{J\times J}]$, we can write the model as a multivariate DLM:

$$\begin{aligned} y_t|\theta_{0:T} \sim& N_J(F\theta_t, V) \\ \theta_t|\theta_{0:(t-1)} \sim& N_{J+1}(\theta_{t-1}, W) \end{aligned} \tag{3.3}$$

for $t = 1, 2, \ldots, T$. This representation will be useful for constructing MCMC algorithms for the model. Using this representation, we need priors for the V_j's, W_j's, U, and θ_0 to complete the model. We will suppose that they are independent with $\theta_0 \sim N_{J+1}(m_0, C_0)$, $V_j \sim IG(a_{V_j}, b_{V_j})$, $W_j \sim IG(a_{W_j}, b_{W_j})$, and $U \sim IG(a_U, b_U)$. We will set $m_0 = 0_{J+1}$, $C_0 = diag(100)$, $a_{V_j} = a_{W_j} = a_u = 1.5$ and $b_{V_j} = b_{W_j} = b_U = 0.25$. This prior on the variance parameters has essentially zero mass below 0.02 and above 2, which allows for a fairly wide range of parameter estimates relative to the scale of the data. These priors are chosen for convenience in illustrating the MCMC method of Simpson et al. (2014) and for simplicity, but a simple way to use the inverse gamma priors without their well known inferential problems (Gelman, 2006) is to put gamma hyperpriors on the b parameters rather than fixing them. The marginal priors on the standard deviations will then be half-t and in the MCMC samplers we discuss a Gibbs step will have to be added for drawing the b's from a Gamma distribution. This prior is the hierarchical inverse Wishart prior of Huang and Wand (2013) in the scalar case.

3.4 Markov chain Monte Carlo

We construct two separate MCMC samplers for this model. One is a naive data augmentation algorithm and the other takes advantage of the interweaving technology of Yu and Meng (2011), particularly the developments of Simpson et al. (2014) for DLMs. We primarily use the DLM representation of the model given in (3.3).

3.4.1 Naive Data Augmentation

The standard DA algorithm characterizes the posterior of (V, W) by using a Gibbs sampler to draw from the posterior distribution of $(V, W, \theta_{0:T})$ (Tanner and Wong, 1987). In this particular case we are also interested in the posterior of $\theta_{0:T}$, which is common in dynamic models, but this does not change the MCMC strategy. The sampler is based on Frühwirth-Schnatter (1994) and Carter and Kohn (1994) and consists of two steps, a draw from $p(\theta_{0:T}|V, W, y_{1:T})$ and a draw from $p(V, W|\theta_{0:T}, y_{1:T})$. In order to construct this algorithm we need these two densities.

First, from the DLM representation of the model in (3.3), and the priors we can write the joint posterior density of V, W, and $\theta_{0:T}$ as

$$\begin{aligned} p(V, W,\theta_{0:T}|y_{1:T}) \propto\, & |V|^{-T/2} \exp\left[-\frac{1}{2}\sum_{t=1}^{T}(y_t - F\theta_t)'V^{-1}(y_t - F\theta_t)\right] \\ \times & |W|^{-T/2} \exp\left[-\frac{1}{2}\sum_{t=1}^{T}(\theta_t - \theta_{t-1})'W^{-1}(\theta_t - \theta_{t-1})'\right] \\ \times & \exp\left[-\frac{1}{2}(\theta_0 - m_0)'C_0^{-1}(\theta_0 - m_0)\right] U^{-a_U-1} \exp\left[-\frac{1}{U}b_U\right] \\ \times & \prod_{j=1}^{J} V_j^{-a_{V_j}-1} \exp\left[-\frac{1}{V_j}b_{V_j}\right] W_j^{-a_{W_j}-1} \exp\left[-\frac{1}{W_j}b_{W_j}\right]. \end{aligned} \tag{3.4}$$

From here we can derive the smoothing density, or conditional posterior density of $\theta_{0:T}$. We use the method of McCausland et al. (2011), based on Rue (2001), for drawing from this density, called the mixed Cholesky factor algorithm (MCFA) by Simpson et al. (2014). The following derivation closely follows Appendix C of Simpson et al. (2014). The full conditional density of $\theta_{0:T}$ can be written as

$$p(\theta_{0:T}|V,W,y_{1:T}) \propto \exp\left[-\frac{1}{2}g(\theta_{0:T})\right]$$

where

$$\begin{aligned} g(\theta_{0:T}) = &\sum_{t=1}^{T}(y_t - F\theta_t)'V^{-1}(y_t - F\theta_t) + \sum_{t=1}^{T}(\theta_t - \theta_{t-1})'W^{-1}(\theta_t - \theta_{t-1}) \\ &+ (\theta_0 - m_0)'C_0^{-1}(\theta_0 - m_0). \end{aligned}$$

Then g has the form $g(\theta_{0:T}) = \theta_{0:T}'\Omega\theta_{0:T} - 2\theta_{0:T}'\omega + K$ where K is some constant with respect to $\theta_{0:T}$, Ω is a square, symmetric matrix of dimension $(J+1)(T+1)$ and ω is a column vector of dimension $(J+1)(T+1)$. This gives $\theta_{0:T}|V,W,y_{1:T} \sim N_{(J+1)(T+1)}(\Omega^{-1}\omega, \Omega^{-1})$. Further, Ω is block tridiagonal since there are no cross product terms involving θ_t and θ_{t+k} where $|k| > 1$. Because of this, the Cholesky factor and thus inverse of Ω can be efficiently computed leading to the Cholesky factor algorithm (CFA) (Rue, 2001). Instead of computing the Cholesky factor of Ω all at once before drawing $\theta_{0:T}$ as in the CFA, the same technology can be used to draw θ_T, then $\theta_t|\theta_{(t+1):T}$ recursively in a backward sampling structure, resulting in the MCFA. In simulations, the MCFA has been found to be significantly cheaper than Kalman filter based methods and often cheaper than the CFA (McCausland et al., 2011).

In order to implement the algorithm, we need to first characterize the diagonal and off diagonal blocks of Ω and the blocks of ω:

$$\begin{aligned} \Omega_{0,0} &= C_0^{-1} + G_1'W^{-1}G_1 & & \\ \Omega_{t,t} &= F'V^{-1}F + 2W^{-1} & &\text{for } \ t = 1,2,\dots T-1 \\ \Omega_{T,T} &= F'V^{-1}F + W^{-1} & & \\ \Omega_{t,t-1} &= -W_t^{-1} = \Omega_{t-1,t} & &\text{for } \ t = 1,2,\dots T \\ w_0 &= C_0^{-1}m_0 & & \\ w_t &= F'V^{-1}y_t & &\text{for } \ t = 1,2,\dots T. \end{aligned}$$

Now let $\Sigma_0 = \Omega_{0,0}^{-1}$, $\Sigma_t = (\Omega_{t,t} - \Omega_{t,t-1}\Sigma_{t-1}\Omega_{t-1,t})^{-1}$ for $t = 1,2,\dots,T$, $h_0 = \Sigma_0 w_0$, and $h_t = \Sigma_t(w_t - \Omega_{t,t-1}h_{t-1})$ for $t = 1,2,\dots,T$. Then to complete the MCFA we perform the following draws recursively

$$\begin{aligned}\theta_T \sim & N(h_T, \Sigma_T)\\ \theta_t|\theta_{(t+1):T} \sim & N(h_t - \Sigma_t\Omega_{t,t+1}\theta_{t+1}, \Sigma_t) & \text{for}\ \ t = T-1, T-2, \dots, 0.\end{aligned}$$

The second step of the DA algorithm requires a draw from $p(V, W|\theta_{0:T}, y_{1:T})$. Recalling that $V = diag(V_1, \dots, V_J)$ and $W = diag(U, W_1, \dots, W_J)$, this density is

$$\begin{aligned}p(V, W|\theta_{0:T}, y_{1:T}) \propto & U^{-a_U - T/2-1}\exp\left[-\frac{1}{U}\left(b_U + \frac{1}{2}\sum_{t=1}^{T}(\mu_t - \mu_{t-1})^2\right)\right]\\ & \times\prod_{j=1}^{J}V_j^{-a_{V_j}-T/2-1}\exp\left[-\frac{1}{V_j}\left(b_{V_j} + \frac{1}{2}\sum_{t=1}^{T}(y_{j,t} - \mu_t - \theta_{j,t})^2\right)\right]\\ & \times\prod_{j=1}^{J}W_j^{-a_{W_j}-T/2-1}\exp\left[-\frac{1}{W_j}\left(b_{W_j} + \frac{1}{2}\sum_{t=1}^{T}(\theta_{j,t} - \theta_{j,t-1})^2\right)\right].\end{aligned}$$

This is the product of inverse gamma densities, so a draw from this density can easily be accomplished by

$$\begin{aligned}V_j &\sim IG(\tilde{a}_{V_j}, \tilde{b}_{V_j}) & \text{for } j = 1, 2, \dots, J\\ W_j &\sim IG(\tilde{a}_{W_j}, \tilde{b}_{W_j}) & \text{for } j = 1, 2, \dots, J\\ U &\sim IG(\tilde{a}_U, \tilde{b}_U)\end{aligned}$$

where $\tilde{a}_U = a_U + T/2$, $\tilde{b}_U = b_U + \sum_{t=1}^{T}(\mu_t - \mu_{t-1})^2/2$, and for $j = 1, 2, \dots, J$, $\tilde{a}_{V_j} = a_{V_j} + T/2$, $\tilde{b}_{V_j} = b_{V_j} + \sum_{t=1}^{T}(y_{j,t} - \mu_t - \theta_{j,t})^2/2$, $\tilde{a}_{W_j} = a_{W_j} + T/2$, and $\tilde{b}_{W_j} = b_{W_j} + \sum_{t=1}^{T}(\theta_{j,t} - \theta_{j,t-1})^2/2$. So we can write the naive DA algorithm as follows:

1. Draw $\theta_{0:T} \sim N(\Omega^{-1}\omega, \Omega^{-1})$ using the MCFA.

2. Draw $U \sim IG(\tilde{a}_U, \tilde{b}_U)$.

3. For $j = 1, 2, \dots, J$ draw $V_j \sim IG(\tilde{a}_{V_j}, \tilde{b}_{V_j})$ and $W_j \sim IG(\tilde{a}_{W_j}, \tilde{b}_{W_j})$.

Note that step 2 and the $2J$ sub steps of step 3 can be parallelized since the draws are all independent, though we do not explore this possibility.

3.4.2 Interweaving

The basic idea of interweaving is to use two separate DAs and "weave" them together (Yu and Meng, 2011). Suppose were have the DAs $\gamma_{0:T}$ and $\psi_{0:T}$. Then an alternating algorithm for our model consists of four steps:

$$[\gamma_{0:T}|V, W, y_{1:T}] \to [V, W|\gamma_{0:T}, y_{1:T}] \to [\psi_{0:T}|V, W, y_{1:T}] \to [V, W|\psi_{0:T}, y_{1:T}].$$

The first two steps are simply the two steps of the DA algorithm based on $\gamma_{0:T}$ while the last two steps are the two steps of the DA algorithm based on $\psi_{0:T}$. A global interweaving strategy (GIS) using these two augmentations is very similar:

$$[\gamma_{0:T}|V, W, y_{1:T}] \to [V, W|\gamma_{0:T}, y_{1:T}] \to [\psi_{0:T}|V, W, \gamma_{0:T}, y_{1:T}] \to [V, W|\psi_{0:T}, y_{1:T}].$$

The only difference is that in step 3, we condition on $\gamma_{0:T}$ as well as V, W, and $y_{1:T}$. Often, this is a transformation using the definition of $\gamma_{0:T}$ and $\psi_{0:T}$, and not a random draw. When step 3 is a transformation, this reduces the computational cost relative to the alternating algorithm. Depending on the properties of the data augmentations used, changing step 3 in this manner can also drastically improve the behavior of the Markov chain whether or not step 3 is a transformation (Yu and Meng, 2011).

Reference Simpson et al. (2014) defines several DAs for the DLM, including the following two — the scaled disturbances, defined by $\gamma_t = L_W^{-1}(\theta_t - \theta_{t-1})$, and the scaled errors, defined by $\psi_t = L_V^{-1}(y_t - F\theta_t)$ for $t = 1, 2, \ldots, T$ and $\psi_0 = \gamma_0 = \theta_0$ where L_X denotes the lower triangular Cholesky factor of the symmetric and positive definite matrix X. Since the dimension of y_t and θ_t are not the same, the scaled errors cannot be directly used without some additional augmentation. Another option is to use a representation of the model which removes the treatment level states, given in (3.2). Using this is unwieldy because the full conditional posterior of $(W_{1:J}, U)$ becomes complicated since the $\phi_{j,t}$'s are correlated across groups. Instead of either of those, we will take a particularly simple approach. Consider the hierarchical representation of the model given in (3.1). For $j = 1, 2, \ldots, J$ define the replication level scaled disturbances as $\gamma_{j,t} = (\theta_{j,t} - \theta_{j,t-1})/\sqrt{W_j}$ for $t = 1, 2, \ldots, T$ and $\gamma_{j,0} = \theta_{j,0}$ and the replication level scaled errors as $\psi_{j,t} = (y_{j,t} - \mu_t - \theta_{j,t})/\sqrt{V_j}$ for $t = 1, 2, \ldots, T$ and $\psi_{j,0} = \theta_{j,0}$. Now let $\gamma_t = (\mu_t, \gamma_{1:J,t}')'$

and $\psi_t = (\mu_t, \psi'_{1:J,t})'$ Then we can easily interweave between $\gamma_{0:T}$ and $\psi_{0:T}$ since these are one-to-one transformations of each other. Specifically the GIS algorithm we seek to construct is

1. Draw $\gamma_{0:T} \sim p(\gamma_{0:T}|V_{1:J}, W_{1:J}, U, y_{1:T})$.
2. Draw $(V_{1:J}, W_{1:J}, U) \sim p(V_{1:J}, W_{1:J}, U|\gamma_{0:T}, y_{1:T})$
3. Transform $\gamma_{0:T} \to \psi_{0:T}$ and draw $(V_{1:J}, W_{1:J}, U) \sim p(V_{1:J}, W_{1:J}, U|\psi_{0:T}, y_{1:T})$

In order to complete this algorithm, we need to characterize the relevant full conditionals. First, consider the transformation from $\theta_{j,0:T}$ to $\gamma_{j,0:T}$. The Jacobian is triangular with a one and T copies of $\sqrt{W_j}$ along the diagonal. So the joint posterior of $V_{1:T}, W_{1:J}, U$, and $\gamma_{0:T}$ is

$$\begin{aligned}
&p(V_{1:T}, W_{1:J}, U, \gamma_{0:T}|y_{1:T}) \propto U^{-a_U - T/2 - 1} \exp\left[-\frac{1}{U}\left(b_U + \frac{1}{2}\sum_{t=1}^{T}(\mu_t - \mu_{t-1})^2\right)\right] \\
&\times \exp\left[-\frac{1}{2}\sum_{j=1}^{J}\sum_{t=1}^{T}\gamma_{j,t}^2\right] \exp\left[-\frac{1}{2}(m_0 - \gamma_0)' C_0^{-1}(m_0 - \gamma_0)\right] \\
&\times \prod_{j=1}^{J} V_j^{-a_{V_j} - T/2 - 1} \exp\left[-\frac{1}{V_j}\left(b_{V_j} + \frac{1}{2}\sum_{t=1}^{T}\left(y_{j,t} - \mu_t - \gamma_{j,0} - \sqrt{W_j}\sum_{s=1}^{t}\gamma_{j,s}\right)^2\right)\right] \\
&\times \prod_{j=1}^{J} W_j^{-a_{W_j} - 1} \exp\left[-\frac{1}{W_j} b_{W_j}\right].
\end{aligned}$$

This allows us to write the model as

$$\begin{aligned}
y_{j,t} &= \mu_t + \sqrt{W_j}\sum_{s=1}^{t}\gamma_{j,s} + \gamma_{j,0} + v_{j,t} \\
\mu_t &= \mu_{t-1} + u_t \qquad (3.5)
\end{aligned}$$

where $(v_{1:J,1:T}, \gamma_{1:J,1:T}, u_{1:T})$ are mutually independent with $\gamma_{j,t} \sim N(0,1)$, $v_{j,t} \sim N(0, V_j)$, and $u_t \sim N(0, U)$ for $j = 1, 2, \ldots, J$ and $t = 1, 2, \ldots, T$. The full conditional of $\gamma_{0:T}$ is a bit more complicated than that of $\theta_{0:T}$, but we can just use the MCFA to draw from $\theta_{0:T}$'s full conditional and transform to $\gamma_{0:T}$. The full conditional of $(V_{1:J}, W_{1:J}, U)$ is

$$p(V_{1:T}, W_{1:J}, U|\gamma_{0:T}, y_{1:T}) = p(U|\gamma_{0:T}, y_{1:T})\prod_{j=1}^{J} p(V_j, W_j|\gamma_{0:T}, y_{1:T}).$$

Here $p(U|\gamma_{0:T}, y_{1:T}) = p(U|\theta_{0:T}, y_{1:T})$, i.e. the same inverse gamma distribution as when we conditioned on $\theta_{0:T}$. However, $p(V_j, W_j|\gamma_{0:T}, y_{1:T})$ is complicated and difficult to sample from efficiently. Instead of drawing V_j and W_j jointly, we draw from their full conditionals. It turns out that $V_j|W_j, \gamma_{0:T}, y_{1:T} \sim IG(\tilde{a}_{V_j}, \tilde{b}_{V_j})$, which is the same as when we conditioned on $\theta_{0:T}$. The full conditional density is of W_j is still rather complicated:

$$p(W_j|V_j, \gamma_{0:T}, y_{1:T}) \propto W_j^{-a_{W_j}-1} \exp\left[-b_{W_j}\frac{1}{W_j} + c_{W_j}\sqrt{W_j} - d_{W_j}W_j\right]$$

where

$$c_{W_j} = \frac{\sum_{t=1}^{T}(y_{j,t} - \mu_t - \gamma_{j,0})\sum_{s=1}^{t}\gamma_{j,s}}{V_j} \in \Re\,, \qquad d_{W_j} = \frac{\sum_{t=1}^{T}\left(\sum_{s=1}^{t}\gamma_{j,s}\right)^2}{2V_j} > 0.$$

The double summations in c_{W_j} and d_{W_j} are one consequence of the model no longer having the Markov property, which can easily be seen from (3.5). These summations can be expensive for large datasets, though in our experience this is typically not the most important computational bottleneck. In any case the summations can be attained much more efficiently via parallelization, especially using a GPU. In order to sample from this density, we follow Simpson et al. (2014) (Appendix E) and use an adaptive rejection sampling approach (Gilks and Wild, 1992) when it is log concave, and otherwise we use a Cauchy approximation in a rejection sampling scheme for the density of $\log(W_j)$.

Now we need to characterize the full conditionals given $\psi_{0:T}$. The Jacobian matrix of the transformation from $\theta_{j,0:T}$ to $\psi_{j,0:T}$ is diagonal with a one and T copies of $\sqrt{V_j}$ along the diagonal. So the joint posterior of $V_{1:T}, W_{1:J}, U$, and $\psi_{0:T}$ is

$$\begin{aligned}
p(V_{1:T}, W_{1:J}, U, \psi_{0:T}|y_{1:T}) \propto{}& U^{-a_U - T/2 - 1}\exp\left[-\frac{1}{U}\left(b_U + \frac{1}{2}\sum_{t=1}^{T}(\mu_t - \mu_{t-1})^2\right)\right]\\
&\times \exp\left[-\frac{1}{2}\sum_{j=1}^{J}\sum_{t=1}^{T}\psi_{j,t}^2\right]\exp\left[-\frac{1}{2}(m_0 - \psi_0)'C_0^{-1}(m_0 - \psi_0)\right]\\
&\times \prod_{j=1}^{J} W_j^{-a_{W_j} - T/2 - 1}\exp\left[-\frac{1}{W_j}\left(b_{W_j} + \frac{1}{2}\sum_{t=1}^{T}\left(\Delta y_{j,t} - \Delta\mu_t - \sqrt{V_j}\Delta\psi_{j,t}\right)^2\right)\right]\\
&\times \prod_{j=1}^{J} V_j^{-a_{V_j}-1}\exp\left[-\frac{1}{V_j}b_{V_j}\right]
\end{aligned}$$

where we define $\Delta x_{j,t} = x_{j,t} - x_{j,t-1}$ for $t = 2, 3, \ldots, T$ and $\Delta x_{j,1} = x_{j,1}$ for any variable $x_{j,t}$ except in the case of $x_{j,t} = y_{j,t}$ where we define $\Delta y_{j,1} = y_{j,1} - \psi_{j,0}$. This allows us to write the model as

$$y_{j,t} = y_{j,t-1} + \sqrt{V_j}\Delta\psi_{j,t} + u_t + w_{j,t} \tag{3.6}$$

where we define $y_{j,0} = (\sqrt{V_j}-1)\psi_{j,0}$ and where $(w_{1:J,1:T}, \psi_{1:J,1:T}, u_{1:T})$ are mutually independent with $\psi_{j,t} \sim N(0,1)$, $w_{j,t} \sim N(0, W_j)$, and $u_t \sim N(0, U)$ for $j = 1, 2, \ldots, J$ and $t = 1, 2, \ldots, T$. While the model is no longer a statespace model under this parameterization, it can be viewed as a statespace model for the $\Delta y_{j,t}$'s with latent states $\Delta\psi_{j,t}$'s and $u_t = \Delta\mu_t$ so long as care is taken in defining the initial values of the data and states. We did not explore this parameterization mainly because the scaled disturbances and scaled errors are natural opposites in the sense that they tend to yield efficient DA algorithms in opposite ends of the parameter space (Simpson et al., 2014), and as such are desirable candidates for interweaving.

Similar to the scaled disturbances case, we have

$$p(V_{1:T}, W_{1:J}, U|\psi_{0:T}, y_{1:T}) = p(U|\psi_{0:T}, y_{1:T}) \prod_{j=1}^{J} p(V_j, W_j|\psi_{0:T}, y_{1:T}).$$

Once again $p(U|\psi_{0:T}, y_{1:T}) = p(U|\theta_{0:T}, y_{1:T})$, which is the same inverse gamma draw. In fact, the parameters $\tilde{a}_U$ and $\tilde{b}_U$ do not change from the γ step to the ψ step, so the second draw of U is redundant and can be removed from the algorithm. The conditional density $p(V_j, W_j|\psi_{0:T}, y_{1:T})$ is once again complicated and has the same form as $p(W_j, V_j|\gamma_{0:T}, y_{1:T})$, i.e. it switches the positions of V_j and W_j. So again we draw V_j and W_j in separate Gibbs steps, and $W_j|V_j, \psi_{0:T}, y_{1:T}$ has the same inverse gamma density as $W_j|\theta_{0:T}, y_{1:T}$. The density of $V_j|W_j, \psi_{0:T}, y_{1:T}$ has the form

$$p(V_j|W_j\psi_{0:T}, y_{1:T}) \propto V_j^{-a_{V_j}-1} \exp\left[-b_{V_j}\frac{1}{V_j} + c_{V_j}\sqrt{V_j} - d_{V_j}V_j\right]$$

where

$$c_{V_j} = \frac{\sum_{t=1}^{T}\Delta\psi_{j,t}(\Delta y_{j,t} - \Delta\mu_t)}{W_j} \in \Re\ , \qquad\qquad d_{V_j} = \frac{\sum_{t=1}^{T}(\Delta\psi_{j,t})^2}{2W_j} > 0.$$

This density has the same form as $p(W_j|V_j, \gamma_{0:T}, y_{1:T})$ so the same rejection sampling strategy can be used to sample from it.

Finally we can write the GIS algorithm as follows:

1. Draw $\theta_{0:T} \sim N(\Omega^{-1}\omega, \Omega^{-1})$ using the MCFA.

2. Draw $U \sim IG(\tilde{a}_U, \tilde{b}_U)$.

3. For $j = 1, 2, \dots, J$:

   (a) Draw $V_j \sim IG(\tilde{a}_{V_j}, \tilde{b}_{V_j})$

   (b) Transform $\theta_{j,0:T} \to \gamma_{j,0:T}$ and draw $W_j \sim p(W_j|V_j, \gamma_{0:T}, y_{1:T})$.

   (c) Transform $\gamma_{j,0:T} \to \psi_{j,0:T}$ and draw $V_j \sim p(V_j|W_j, \psi_{0:T}, y_{1:T})$.

   (d) Draw $W_j \sim IG(\tilde{a}_{W_j}, \tilde{b}_{W_j})$.

Since $(U, V_1, \dots, V_J, W_1, \dots, W_J)$ are conditionally independent in the posterior no matter which of the DAs we use, Step 3 can be parallelized and step 2 can come before or after step 3, though we did not experiment with these possibilities. Steps 3.b and 3.c can both be accomplished using the rejection sampling method described from Appendix E of Simpson et al. (2014), briefly described above. Note that the transformation from $\gamma_{j,0:T} \to \psi_{j,0:T}$ is defined as $\psi_{j,t} = (y_{j,t} - \mu_t - \sqrt{W}_j \sum_{s=1}^{t} \gamma_{j,s} - \gamma_{j,0})/\sqrt{V_j}$ for $j = 1, 2, \dots, J$ and $t = 1, 2, \dots, T$.

In (3.5) and (3.6) it is apparent that using the scaled disturbances or the scaled errors, the model no longer has the Markov property. This is undesirable for computational reasons — it causes the double summations in the definitions of c_{W_i} and d_{W_i} and increases the computational cost associated with drawing the latent states — but the cost is worthwhile for convergence and mixing because the parameterizations are natural opposites in a particular sense. According to both theorem 1 and theorem 2 of Yu and Meng (2011), the convergence rate of an interweaving algorithm is faster when the convergence rate of the fastest underlying DA algorithm is faster, so in their words it is desirable to seek a “beauty and the beast” pair of DAs where when one DA algorithm is bad the other is good and vice-versa. Reference Simpson et al. (2014) showed in the local level model that the scaled disturbances and scaled errors yield DA algorithms which are efficient in opposite ends of the parameter space so that they exhibit precisely this “beauty and the beast” behavior.

It is also possible to transform the μ_t's in an interweaving approach. The problem becomes what two parameterizations to use. The scaled disturbances and the scaled errors make a natural pair because they work well in opposite ends of the parameter space which, in turn, seems to be driven by one being a data level reparameterization and the other a latent state level reparameterization. The scaled version of the μ_t's would still be a latent state level parameterization, and there is no clear data level reparameterization which corresponds to them. This is a consequence of the model having a higher dimensional latent state than data, though one method to overcome this issue that Simpson et al. (2014) mentions is via additional augmentation — that is define missing data on the data level so that the full data, consisting of the observed and missing data, has the same dimension as the latent state. We sidestep this issue by leaving the μ_t's untransformed through the algorithm, though there are potential gains to be made by experimenting with reparameterizing this component of the DA.

3.5 Results

We fit the model in R using both MCMC algorithms, running five chains for each algorithm at diverse starting points for $20,000$ iterations per chain. For both algorithms, convergence appeared to be attained for all parameters in all chains in the first $5,000$ iterations according to both trace plots and the Gelman-Rubin diagnostic (Brooks and Gelman, 1998), so we throw away those initial draws as burn in. The GIS algorithm appeared to converge slightly slower according to the Gelman-Rubin diagnostic for some of the parameters, though this difference was not apparent in trace plots.

Table 3.1: Effective sample size (n_{eff}) and time in seconds per $1,000$ effective draws (Time) for each MCMC algorithm computed after burn in for all chains. Actual sample size is $60,000$ for each algorithm.

		V_1	V_2	V_3	V_4	V_5	V_6
DA n_{eff}		24633	20656	20558	18883	21003	24897
GIS n_{eff}		44894	43659	35400	43843	23364	40913
DA Time		3.08	3.68	3.70	4.02	3.62	3.05
GIS Time		4.85	4.98	6.15	4.96	9.31	5.32
	U	W_1	W_2	W_3	W_4	W_5	W_6
DA n_{eff}	14583	15072	18713	15137	10609	13228	29458
GIS n_{eff}	19571	23706	23560	22768	15051	17753	29729
DA Time	5.21	5.04	4.06	5.02	7.16	5.74	2.58
GIS Time	11.12	9.18	9.24	9.56	14.46	12.26	7.32

There were, however, significant differences in mixing between the two algorithms. Table

3.1 contains the effective sample size, n_{eff} (Gelman et al., 2013), for each parameters as well as the time in seconds to achieve an effective sample size of $1,000$ for each parameter, computed for both MCMC algorithms using all $60,000$ post burn-in iterations. The GIS algorithm has higher n_{eff} for all parameters. For some parameters, e.g. V_5 and W_6, this difference is rather small. For others, such as V_1 and V_2, the GIS algorithm has an n_{eff} roughly twice as large as the DA algorithm. In time per $1,000$ effective draws, however, the GIS algorithm under-performs across the board. When evaluating these times, note that the algorithms were implemented in R where the code was interpreted, not compiled. Absolute times may differ dramatically from the times listed in Table 3.1 under different programming languages or based on whether the code was interpreted or compiled, though relative times should be roughly comparable at least for interpreted code from other languages. The steps to draw from $p(W_j|V_j, \gamma_{0:T}, y_{1:T})$ and $p(V_j|W_j, \psi_{0:T}, y_{1:T})$ are the main culprits — they are often very expensive. As the number of periods in the experiment increases, Simpson et al. (2014) found that in the local level model the GIS algorithm looks stronger relative to the DA algorithm since GIS is able to use adaptive rejection sampling more often and the relative advantage of the improved mixing becomes more important, and we expect this to hold in our model. Similarly, a judicious choice of priors which allows for easier full conditionals in the offending steps should result in a faster computational times for GIS relative to the DA algorithm.

Table 3.2: Parameter estimates, including the posterior mean, posterior median, and a 95% credible interval for each parameter.

	Mean	50%	2.5%	97.5%		Mean	50%	2.5%	97.5%
V_1	0.144	0.136	0.070	0.263	W_1	0.101	0.092	0.042	0.216
V_2	0.086	0.080	0.040	0.163	W_2	0.083	0.075	0.035	0.171
V_3	0.116	0.106	0.045	0.248	W_3	0.078	0.072	0.035	0.158
V_4	0.102	0.095	0.046	0.196	W_4	0.104	0.095	0.043	0.216
V_5	0.208	0.196	0.075	0.415	W_5	0.110	0.096	0.038	0.258
V_6	0.162	0.153	0.077	0.296	W_6	0.085	0.076	0.034	0.188
					U	0.044	0.041	0.023	0.079

Table 3.2 contains the parameter estimates for the model. The treatment level variance appears to be smaller than both the replication and observation level variances, suggesting that changes in logit efficiency over time are driven less by treatment level dynamics and

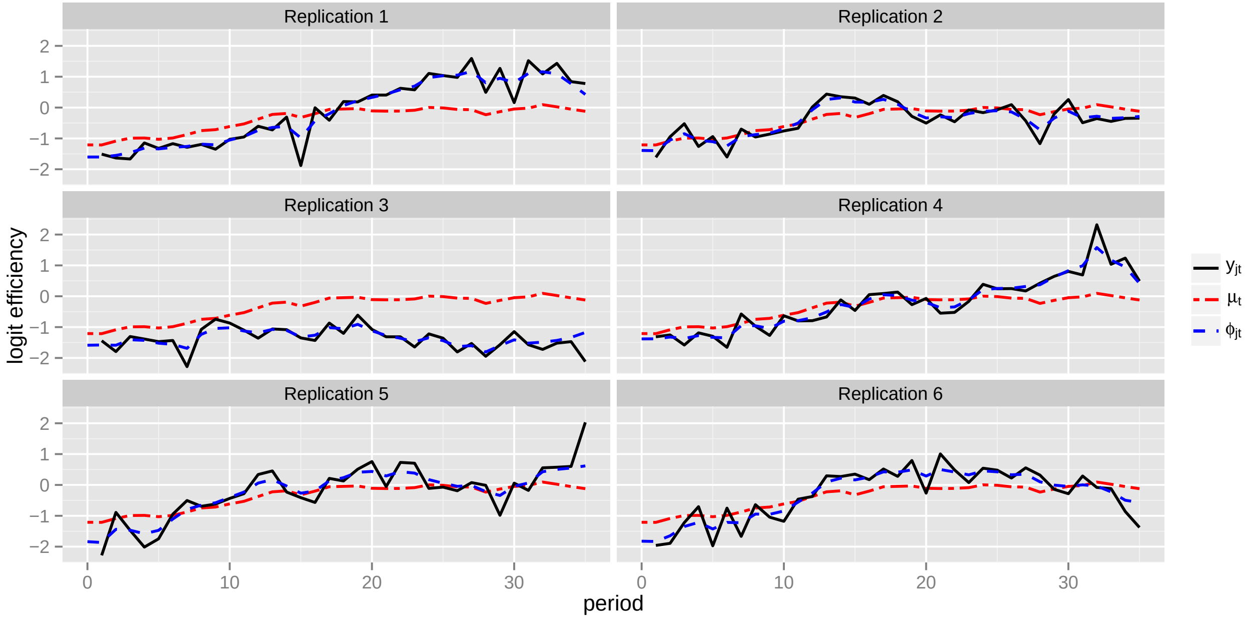


Figure 3.1: Plots by replication of the observed logit efficiency ($y_{j,t}$), posterior median latent replication logit efficiency ($\phi_{j,t}$), and posterior median latent treatment logit efficiency (μ_t).

more by random noise and replication level dynamics. Figure 3.1 also contains plots of each replication's observed logit efficiency trajectory, each replication's posterior median latent logit efficiency trajectory, and the treatment wide posterior median latent efficiency trajectory. The replication level latent logit efficiency follows the observed logit efficiency very closely in each case — it is essentially a smoothed version of the observed logit efficiency. The treatment latent logit efficiency follows the observed logit efficiencies of replications 2, 4, 5, and 6 fairly closely, but replication 3 consistently under performs the treatment average while replication 1 consistently over performs, at least in the latter half of periods.

3.6 Conclusion

Reference Simpson et al. (2014) explored the interweaving algorithms of Yu and Meng (2011) for DLMs, but only implemented them in the univariate local level model. We use their approach in a model that can be represented as independent local level models conditional on a univariate sequence of latent states, or as a slightly more complicated DLM with J

dimensional data and $J+1$ dimensional state. This poses some problems with directly applying the methods in Simpson et al. (2014), but we show that they are easily overcome. The resulting sampler has similar convergence and improved mixing properties compared to the standard data augmentation algorithm with this particular dataset. In terms of end user time required to adequately characterize the posterior, the DA algorithm is a bit faster for this particular problem despite worse mixing, but this is largely due to an inefficient rejection sampling step in the interweaving algorithm that likely can be improved (Simpson et al., 2014). This step also tends to become relatively more efficient in problems with more data as well as less important relative to improved mixing so that the interweaving algorithm will eventually, with enough data, outperform the DA algorithm (Simpson et al., 2014).

CHAPTER 4. A SLIDING SCALE OF PARTIAL IDENTIFICATION: WEAKENING PARTIAL IDENTIFICATION ASSUMPTIONS FOR CREDIBILITY

A working paper

Abstract

The National School Lunch Program (NSLP) provides free or reduced-price meals to children of households with low income. Evaluating the causal effectiveness of this program is difficult because of the missing counterfactual problem and misreporting of program participation. Following previous work on this program, we introduce several new methods to account for the missing counterfactual problem using a Bayesian treatment effects approach. We build two endogenous selection models with a plurality of unidentified parameters. To identify these parameters, we construct credible prior distributions that use dependence between identified and unidentified parameters to learn about the unidentified parameters; e.g. by choosing prior distributions that embody monotone treatment selection and similar assumptions. The analysis is extended to allow for post-stratification by fitting each model on a subgroup in a hierarchical structure with other subgroups.

4.1 Introduction

The National School Lunch program (NSLP) gave free or reduced price lunches to over 31 million U.S. children each school day in 2012 at a cost of about $11.6 billion for fiscal year 2012. Households which were under 130% of the poverty line received free school lunches for their children, while households between 130% and 185% of the poverty line paid a small price — 40 cents in 2001–2004, the period our data comes from. Presumably this would result in better health outcomes for children on the program, but the evidence is mixed. This is partially because identifying relevant treatment effect parameters is difficult. First, the standard missing counterfactual problem is present — we observe a household on the school lunch program or not, but never both. Second, there is evidence that the treatment status of the household (whether they are on the school lunch program or not) is underreported. Gundersen et al. (2012) attempt to deal with both of these issues through partial identification. We focus purely on the the missing counterfactual problem and construct several models in order to deal with this issue within a Bayesian framework.

Each model works by defining a model for the data we wish we had, identifying parameters which are unidentified given the data we do have, then constructing a reasonable prior that allows us to learn about unidentified parameters and in particular treatment effect parameters through what we learn about the identified parameters. These priors embody, e.g., the monotone treatment selection assumption (MTS), which bounds unidentified parameters with respect to identified parameters. Learning takes place through these bounds even in the Bayesian context, as in Manski (1999). There are well known problems with Bayesian inference in partial identification problems from a frequentist point of view. In particular Bayesian credible intervals will have incorrect frequentist coverage (Moon and Schorfheide, 2012), though a Bayesian approach can still be used to fit the reduced model using MCMC which then can be used with the bounds in order to construct correct frequentist confidence intervals (Kline and Tamer, 2013).

Our innovation within the Bayesian context is twofold. First, we construct priors which only force inequalities such as MTS to hold some proportion of the time. This proportion is unidentified and we have no intention to use the data to learn about it; rather, its purpose is to better represent our uncertainty about the problem. It is not often the case that we truly believe that a moment inequality holds with 100% certainty, so our prior should reflect that. By combining various reasonable moment inequalities for the problem with varying degrees of certainty in models with differing levels of detail, we are able to construct a continuum of possible methods for estimating treatment effect parameters. Our second innovation is to fit these models by conditioning on various sub-populations and shrinking the parameter estimates to a common mean. This allows us to compute treatment effects parameters by post-stratifying in order to take into account mismatch between sample and population. See Gelman and Little (1997), Gelman and Carlin (2001), and Park et al. (2004) for discussion and examples of post-stratification in the Bayesian context.

4.2 The Modeling Framework

Suppose we have some treatment $d \in \{0, 1\}$ and we are interested in the treatment's effect on some binary response y. Then $y_i(d) \in \{0, 1\}$ is the potential response for observation i under treatment condition d, $i = 1, 2, \cdots, N$. $y_i(.)$ can be thought of as a function that maps the treatment applied to an outcome – either successful or unsuccessful. Let d_i denote the actual treatment status for observation i. Note that $y_i(d_i)$ is observed, but $y_i(1 - d_i)$ is not. This is the essence of the problem – we do not observe the missing counterfactual, $y_i(1 - d_i)$ and observational units (e.g. households) select into their treatment status based on unobservables, so we cannot easily compare success rates among those who chose to go on the treatment and those who did not. Let y_i denote the observed response for observation i. Then

$$y_i = y_i(0) + d_i \left[y_i(1) - y_i(0)\right] = y_i(0) + d_i TE_i \tag{4.1}$$

where $TE_i = y_i(1) - y_i(0)$ is the i'th observation's treatment effect.

The most basic model we can write down in this context separates the potential outcomes into four categories based on which scenario we are considering and which scenario the observation actually entered into. This gives us five Bernoulli probabilities to estimate, $P(y_i(a) = 1|d_i = b)$ for $a, b \in \{0, 1\}$ and $P(d_i = 1)$, and is the starting point for Gundersen et al. (2012). Here we do not want to assume that $y_i(0)$ and $y_i(1)$ are independent conditional on d_i – it seems intuitive that household i's potential outcome while on the treatment is related to its potential outcome when not on the treatment, even conditional on the household's treatment choice. In order to take into account the dependence we need to decompose $P(y_i(a) = 1|d_i = 1 - a)$ into

$$\begin{aligned} &P(y_i(a) = 1|d_i = 1 - a) = \\ &\qquad P\left[y_i(a) = 1|y_i(1-a) = 0, d_i = 1 - a\right] P[y_i(1-a) = 0|d_i = 1 - a] \\ &\qquad +P\left[y_i(a) = 1|y_i(1-a) = 1, d_i = 1 - a\right] P[y_i(1-a) = 1|d_i = 1 - a]. \end{aligned}$$

This gives us seven Bernoulli probabilities to estimate, except this time the draws are independent within their categories:

$$\begin{aligned} p_d &= P(d_i = 1) \\ p_{a|a} &= P(y_i(a) = 1|d_i = a) \\ q_{a|b} &= P(y_i(a) = 1|y_i(1-a) = b, z_i = 1 - a) \end{aligned}$$

for $a, b \in \{0, 1\}$. In other words, the model is

$$\begin{aligned} y_i(1-a)|y_i(a) = b, d_i = a &\overset{iid}{\sim} \text{Ber}(q_{a|b}) \\ y_i(a)|d_i = a &\overset{iid}{\sim} \text{Ber}(p_{a|a}) \\ d_i &\overset{iid}{\sim} \text{Ber}(p_d). \end{aligned} \tag{4.2}$$

We can write this model in an equivalent form using the multinomial distribution. Let $x_i = y_i(0) + 2y_i(1)$ so that $x_i \in \{1, 2, 3, 4\}$. Then we have

$$\begin{aligned} x_i|d_i = d &\overset{iid}{\sim} \text{Multinomial}(\boldsymbol{p}_{|d}) \\ d_i &\overset{iid}{\sim} \text{Ber}(p_d) \end{aligned} \tag{4.3}$$

where $d \in \{0, 1\}$ and $\boldsymbol{p}_{|d} = (p_{00|d}, p_{01|d}, p_{10|d}, p_{11|d})$ with $p_{ab|d} = P(y_i(0) = a, y_i(1) = b|d_i = d)$ and $a, b, d \in \{0, 1\}$. Since $\boldsymbol{p}_{|d}$ lives in the simplex it only contains three unknown parameters, so that the model still contains a total of seven unknown parameters.

The version of the model in (4.3) is more convenient for some purposes including MCMC sampling, but the version in (4.2) is much more convenient for reasoning about identification. We do not observe $(d_i, y_i(0), y_i(1))$, but rather (d_i, y_i) with d_i defined in (4.1) and, as a result, the parameters in the first line of (4.2) are unidentified – $q_{a|b}$ for $a, b \in \{0, 1\}$. The rest of the parameters, p_d, $p_{0|0}$, and $p_{1|1}$, are identified. Using (4.3), p_d is again identified as well as $p_{0|0} = p_{00|0} + p_{01|0}$ & $p_{1|1} = p_{01|1} + p_{11|1}$ and $1 - p_{0|0} = p_{10|0} + p_{11|0}$ & $1 - p_{1|1} = p_{00|1} + p_{10|1}$. Note that the unidentified marginal probabilities can be written as

$$p_{0|1} = p_{1|1}q_{0|1} + (1 - p_{1|1})q_{0|0} = p_{10|1} + p_{11|1}$$

and

$$p_{1|0} = p_{0|0}q_{1|1} + (1 - p_{0|0})q_{1|0} = p_{01|0} + p_{11|0}.$$

4.3 Partial Identification of Mean Treatment Effects

In order to learn about the unidentified parameters in the model the model, we will use partial identification. Crucially we are interested in identifying treatment effects parameters and not necessary every parameter in the model. Note that

$$\begin{aligned}
ATE &= E[TE_{new}] = P(y_{new}(1) = 1) - P(y_{new}(0) = 1) \\
&= p_{1|1}p_d + p_{1|0}(1 - p_d) - p_{0|1}p_d - p_{0|0}(1 - p_d) \\
&= (p_{01|1} - p_{10|1})p_d + (p_{01|0} - p_{10|0})(1 - p_d) \\
ATT &= E[TE_{new}|d_{new} = 1] = P(y_{new}(1) = 1|d_{new} = 1) - P(y_{new}(0) = 1|d_{new} = 1) \\
&= p_{1|1} - p_{0|1} \\
&= p_{01|1} - p_{10|1}. \qquad (4.4)
\end{aligned}$$

So we can leave the full model unidentified so long as we can identify $p_{0|1}$ and $p_{1|0}$ – if we can find a way to learn about these two parameters, we can learn about the treatment effects parameters of interest.

While many assumptions will allow us to partially identify mean treatment effect, not all of them will identify the treatment effect *distribution*. Consider the full treatment effects distributions from the posterior predictive distribution. The ATE distribution is the distribution of a new household's treatment effect, $p(TE_{new})$, and is a discrete distribution on $\{-1,0,1\}$. Similarly the ATT distribution is the distribution of a new household's treatment effect conditional on the household selecting the treatment, $p(TE_{new}|d_{new}=1)$, and is also a discrete distribution on $\{-1,0,1\}$. Often we can make assumptions which partially identify the means of these distributions while not partially identifying the full distributions. The full distributions depend on the model parameters as follows:

$$\begin{aligned}
P(TE_{new}=-1) &= P(y_{new}(0)=1, y_{new}(1)=0)\\
&= (1-p_{1|1})q_{0|1}p_d + p_{0|0}(1-q_{1|0})(1-p_d)\\
&= p_{10|0}(1-p_d) + p_{10|1}p_d\\
P(TE_{new}=0) &= P(y_{new}(0)=0, y_{new}(1)=0) + P(y_{new}(0)=1, y_{new}(1)=1)\\
&= \left[p_{1|1}q_{0|1} + (1-p_{1|1})(1-q_{0|0})\right]p_d + \left[p_{0|0}q_{1|1} + (1-p_{0|0})(1-q_{1|0})\right](1-p_d)\\
&= [p_{00|0}+p_{11|0}](1-p_d) + [p_{00|1}+p_{11|1}]p_d\\
P(TE_{new}=1) &= P(y_{new}(0)=0, y_{new}(1)=1)\\
&= p_{1|1}(1-q_{0|1})p_d + (1-p_{0|0})q_{1|0}(1-p_d)\\
&= p_{01|0}(1-p_d) + p_{01|1}p_d
\end{aligned} \tag{4.5}$$

and

$$\begin{aligned}
P(TE_{new}=-1|d_{new}=1) &= P(y_{new}(0)=1, y_{new}(1)=0|d_{new}=1)\\
&= (1-p_{1|1})q_{0|1} = p_{10|1}\\
P(TE_{new}=0|d_{new}=1) &= P(y_{new}(0)=0, y_{new}(1)=0|d_{new}=1) + P(y_{new}(0)=1, y_{new}(1)=1|d_{new}=1)\\
&= p_{1|1}q_{0|1} + (1-p_{1|1})(1-q_{0|0}) = p_{00|1}+p_{11|1}\\
P(TE_{new}=1|d_{new}=1) &= P(y_{new}(0)=0, y_{new}(1)=1|d_{new}=1)\\
&= p_{1|1}(1-q_{0|1}) = p_{01|1}.
\end{aligned} \tag{4.6}$$

In order to identify these parameters it is sufficient to identify each of the $p_{ab|d}$'s, but it is not sufficient to identify each of the $p_{a|d}$'s. In this section, we will focus on identifying mean treatment effects parameters.

4.3.1 Monotone Treatment Selection

The monotone treatment selection (MTS) assumption partially identifies $p_{0|1}$ and $p_{1|0}$ by assuming that $p_{0|0} > p_{0|1}$ and $p_{1|1} < p_{1|0}$. In other words, MTS assumes that when both types of households are either on the treatment or off the treatment, households that chose to go on the treatment are on average worse off than households that chose not to go on the treatment. We can restate this assumption in terms of the $p_{ab|d}$'s as

$$p_{11|0} + p_{10|0} > p_{11|1} + p_{10|1}$$

and

$$p_{11|0} + p_{01|0} > p_{11|1} + p_{01|1}$$

which can be restated using the simplex constraint as

$$p_{00|0} + p_{01|0} < p_{00|1} + p_{01|1}$$

and

$$p_{00|0} + p_{10|0} > p_{00|1} + p_{10|1}$$

This is sufficient to partially identify $p_{0|1}$ & $p_{1|0}$ and thus ATE & ATT, but not the full ATE and ATT distributions – it does not allow us to learn about the $p_{ab|d}$'s without further assumptions.

In order to translate MTS into the Bayesian framework, we need to translate the bounds into a prior distribution on $(\boldsymbol{p}_{|0}, \boldsymbol{p}_{|1})$. We will keep things simple and for the moment and assume we are only interested in estimating mean ATE and mean ATT, so all we need is a prior on $(p_d, p_{0|0}, p_{0|1}, p_{1|0}, p_{1|1})$. The easy way to do this is to assume a uniform prior subject to the MTS constraints, but we could generalize slightly and assume that each parameter has a beta distribution subject to the MTS constraints. This approach is natural, but skipping forward a little bit it will cause problems for post-stratification when we try to construct a hierarchical version of the model. Suppose we have data $z_i \overset{iid}{\sim} \mathcal{B}(\alpha, \beta)$. Then the posterior is $p(\alpha,\beta|\boldsymbol{z}) \propto \frac{1}{B(\boldsymbol{\alpha},\boldsymbol{\beta})^n}\exp\left[-\alpha v_\alpha - \beta v_\beta\right]p(\alpha,\beta)$ where $p(\alpha,\beta)$ is the prior and v_α & v_β are functions of $\boldsymbol{z}$. The term involving the beta function prevents any convenient conditionally conjugate form from showing up and furthermore will sometimes be relatively expensive to evaluate every iteration of an MCMC algorithm. In addition, the Dirichlet distribution is known to be a relatively inflexible model for parameters that live in the simplex (Aitchison, 1986). These will be larger problems when we model the $p_{ab|c}$'s in Section 4.4, but they still pose some difficulty here since the beta is a special case of the Dirichlet.

In order to deal with this issue, we will use the approach of Gelman et al. (1996), also discussed in Gelman (1995), that uses normal distributions properly normalized to obtain a prior on the simplex. Specifically let $\theta = (p_d, p_{0|0}, p_{1|0}, p_{0|1}, p_{1|1})$ and $\lambda_k \overset{ind}{\sim} \mathcal{N}(\mu_k, \sigma_k^2)$ for $k = 1, 2, \ldots, 10$. This distribution is transformed to the simplex by

$$\theta_k = \frac{e^{\lambda_{2k}}}{e^{\lambda_{2k-1}} + e^{\lambda_{2k}}}$$

for $k = 1, 2, \ldots, 5$. We will call this prior the *normalized lognormal* or NLN prior, denoted by $p \sim \mathcal{NLN}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ with $\boldsymbol{\mu} = (\mu_1, \mu_2)$ and $\boldsymbol{\sigma}^2 = (\sigma_1^2, \sigma_2^2)$. The full set of λ_k's is not identifiable since we can add a constant to any pair without impacting any of the θ_k's, but as long we are only interested in the θ_k's this causes no inferential problems. The normalized lognormal prior causes each the distribution of each of the θ_k's to depend on four parameters, providing a greater deal of flexibility. The relative values of both μ_1 & μ_2 and σ_1^2 and σ_2^2 control the expected value of p, though in rather opaque ways, but Gelman (1995) shows how to use prior information about the moments of the θ_k's to choose the μ_k's and σ_k's to match. We can easily model each of those parameters hierarchically across groups as well using conditionally conjugate normal and inverse gamma distributions, which we will do in Section 4.5. The cost is that the full conditional distribution of each of the λ_k's is complicated. In Section 4.6 we show how to draw from these full conditionals using a random walk Metropolis step which works fairly well.

So the unconstrained prior on $\theta = (p_d, p_{0|0}, p_{1|0}, p_{0|1}, p_{1|1})$ is

$$\begin{aligned} p_{UN}(\theta) = {} & \mathcal{NLN}(p_d; \boldsymbol{\mu}_{1:2}, \boldsymbol{\sigma}^2_{1:2})\mathcal{NLN}(p_{0|0}; \boldsymbol{\mu}_{3:4}, \boldsymbol{\sigma}^2_{3:4})\mathcal{NLN}(p_{1|0}; \boldsymbol{\mu}_{5:6}, \boldsymbol{\sigma}^2_{5:6}) \\ & \times\mathcal{NLN}(p_{0|1}; \boldsymbol{\mu}_{7:8}, \boldsymbol{\sigma}^2_{7:8})\mathcal{NLN}(p_{1|1}; \boldsymbol{\mu}_{9:10}, \boldsymbol{\sigma}^2_{9:10}) \end{aligned} \tag{4.7}$$

where $\mathcal{NLN}(p;\boldsymbol{\mu},\boldsymbol{\sigma}^2)$ is the pdf of the normalized lognormal distribution. The MTS prior uses the same unconstrained prior except restricted to the space where MTS holds. To wit:

$$\begin{aligned}p_{MTS}(\theta) \propto\, & \mathcal{NLN}(p_d;\boldsymbol{\mu}_{1:2},\boldsymbol{\sigma}^2_{1:2})\mathcal{NLN}(p_{0|0};\boldsymbol{\mu}_{3:4},\boldsymbol{\sigma}^2_{3:4})\mathcal{NLN}(p_{1|0};\boldsymbol{\mu}_{5:6},\boldsymbol{\sigma}^2_{5:6}) \\ & \times\mathcal{NLN}(p_{0|1};\boldsymbol{\mu}_{7:8},\boldsymbol{\sigma}^2_{7:8})\mathcal{NLN}(p_{1|1};\boldsymbol{\mu}_{9:10},\boldsymbol{\sigma}^2_{9:10})\mathbb{1}\{p_{0|0}<p_{0|1}\}\mathbb{1}\{p_{1|0}<p_{1|1}\}.\end{aligned} \tag{4.8}$$

The full conditionals of the resulting posterior are complicated by the MTS constraints, but in Section 4.6 we show that the constraint on θ_k conditional on θ_{-k} is simply an interval function of θ_{-k}, (L,U) where $-\infty \leq L < U \leq \infty$.

4.3.2 Mean Monotone Treatment Response

Another popular identifying assumption, Monotone treatment response (MTR) says that $y_i(1) \geq y_i(0)$ for all i. This assumption is rather strong since it rules out the possibility that the treatment hurts the recipient. A weakening of MTR is possible that we might call mean monotone treatment response or MMTR. There are a couple of ways we could translate this statement into mathematics. The first way translates it literally and says $E[y_i(1)] > E[y_i(0)]$, i.e. $P(y_i(1)=1) > P(y_i(0)=1)$. The second way says that $E[y_i(1)|d_i=d] > E[y_i(0)|d_i=d]$, i.e. $P(y_i(1)=1|d_i=d) > P(y_i(0)=1|d_i=d)$ for all $d =\in \{0,1\}$. We will consider each version of this assumption in turn.

Version 1 of MMTR says $P(y_i(1)=1) > P(y_i(0)=1)$, which is equivalent to

$$\begin{aligned}&(p_{01|0}+p_{11|0})(1-p_d)+(p_{01|1}+p_{11|1})p_d > (p_{10|0}+p_{11|0})(1-p_d)+(p_{10|1}+p_{11|1})p_d \\ &\iff p_{01|0}(1-p_d)+p_{01|1}p_d > p_{10|0}(1-p_d)+p_{10|1}p_d.\end{aligned}$$

This assumption is a bit strange since it does *not* imply that, conditional on a household's treatment selection, the household is more likely to have a successful outcome while on the treatment than while off the treatment. So while it is a strict translation of "monotone mean treatment response" into mathematics, it does not faithfully capture the spirit of the idea.

Version 2 of MMTR says $P(y_i(1)=1|d_i=d) > P(y_i(0)=1|d_i=d)$ for $d\in\{0,1\}$, which is equivalent to $p_{1|0} > p_{0|0}$ and $p_{1|1} > p_{0|1}$, or equivalently $p_{01|0} > p_{10|0}$ and $p_{01|1} > p_{10|1}$. So version 2 of MMTR says that a given household is more likely to be successful under the treatment and unsuccessful off of it than the unsuccessful under the treatment and successful off of it. In particular, it implies version 1 of MMTR and furthermore, MTR $\implies$ MMTR v2 $\implies$ MMTR v1. Both versions of MMTR can be combined with MTS in order to buy additional identifying power, but we will use version 2.

To translate MMTR into a prior distribution, we will again use the unconstrained prior truncated to satisfy the MMTR constraints.

$$\begin{aligned} p_{MMTR}(\theta) \propto & \mathcal{NLN}(p_d; \boldsymbol{\mu}_{1:2}, \boldsymbol{\sigma}^2_{1:2})\mathcal{NLN}(p_{0|0}; \boldsymbol{\mu}_{3:4}, \boldsymbol{\sigma}^2_{3:4})\mathcal{NLN}(p_{1|0}; \boldsymbol{\mu}_{5:6}, \boldsymbol{\sigma}^2_{5:6}) \\ & \times\mathcal{NLN}(p_{0|1}; \boldsymbol{\mu}_{7:8}, \boldsymbol{\sigma}^2_{7:8})\mathcal{NLN}(p_{1|1}; \boldsymbol{\mu}_{9:10}, \boldsymbol{\sigma}^2_{9:10})\mathbb{1}\{p_{0|0} < p_{1|0}\}\mathbb{1}\{p_{0|1} < p_{1|1}\}. \end{aligned} \tag{4.9}$$

It turns out that the MTS and MMTR constraints are orthogonal, so we can impose both at the same time in order to further increase identification power. The combined inequalities state that $p_{0|0} < p_{1|0} < p_{1|1}$ and $p_{0|0} < p_{0|1} < p_{1|1}$, yielding the MTS+MMTR prior.

$$\begin{aligned} p_{MTS+MMTR}(\theta) \propto & \mathcal{NLN}(p_d; \boldsymbol{\mu}_{1:2}, \boldsymbol{\sigma}^2_{1:2})\mathcal{NLN}(p_{0|0}; \boldsymbol{\mu}_{3:4}, \boldsymbol{\sigma}^2_{3:4})\mathcal{NLN}(p_{1|0}; \boldsymbol{\mu}_{5:6}, \boldsymbol{\sigma}^2_{5:6}) \\ & \times \mathcal{NLN}(p_{0|1}; \boldsymbol{\mu}_{7:8}, \boldsymbol{\sigma}^2_{7:8})\mathcal{NLN}(p_{1|1}; \boldsymbol{\mu}_{9:10}, \boldsymbol{\sigma}^2_{9:10})\mathbb{1}\{p_{0|0} < p_{1|0} < p_{1|1}\}\mathbb{1}\{p_{0|0} < p_{0|1} < p_{1|1}\}. \end{aligned} \tag{4.10}$$

4.3.3 Probable MTS, MMTR, and MTS+MMTR

One issue with MTS, MMTR, and MTS+MMTR is that they assume the relevant bounds hold with 100% certainty, but we do not necessarily believe this – more likely they hold with a high probability. We use this idea to develop the probable version of each of the above assumptions – PMTS, PMMTR, and PMTS+PMMTR. Strict MTS assumes that $p_{0|0} > p_{0|1}$ and $p_{1|0} > p_{1|1}$, but instead we will assume that they hold with some probability η and that the parameters are unconstrained with probability $1 - \eta$. So the PMTS prior is a mixture of the MTS and UN priors, i.e.

$$p_{PMTS}(\theta, m) = (1 - \eta)^{1-m} p_{UN}(\theta) + \eta^m p_{MTS}(\theta)$$

where $m = 1$ indicates that MTS holds and $\eta = P(\text{MTS holds})$. Marginalizing out m yields

$$p_{PMTS}(\theta) = (1 - \eta)p_{UN}(\theta) + \eta p_{MTS}(\theta). \tag{4.11}$$

PMTS is a generalization of MTS since $\eta = 1$ yields MTS and the unconstrained case since $\eta = 0$. This prior provides a sliding scale of assumptions about the success rate of individuals who chose the treatment compared to the success rate of individuals who did not choose the treatment while on the other hand MTS dogmatically sets $\eta = 1$ – it says that we are 100% certain that individuals who chose not to go on the treatment have a higher success rate than individuals who chose to go on the treatment. When MTS is plausible but we have some misgivings, shrinking η away from one will allow us to obtain more credible estimates of $p_{0|1}$ and $p_{1|0}$ as well as mean ATE and mean ATT. It might seem suspicious that while using PMTS inference for the parameters of interest is highly dependent on η, a parameter which cannot be estimated from the data. This is true, of course, but MTS falls prey to the same criticism – inference will always be highly sensitive to these sort of identifying assumptions, and why is $\eta = 1$ so sacred, after all? The key is to represent our uncertainty faithfully and to be transparent about where our analysis is dependent on these sorts of choices.

Analogous to PMTS we can define PMMTR, but nothing is meaningfully different. So we will actually define PMTS+PMMTR first and see that PMTS and PMMTR are special cases. Let $\varepsilon = P(\text{MMTR holds})$ and assume that whether MTS holds and whether MMTR holds are independent. Then the PMTS+PMMTR prior is

$$\begin{aligned} p_{PMTS+PMMTR}(\theta) = {} & (1-\eta)(1-\varepsilon)p_{UN}(\theta) + \eta(1-\varepsilon)p_{MTS}(\theta) \\ & + (1-\eta)\varepsilon p_{MMTR}(\theta) + \eta\varepsilon p_{MTS+MMTR}(\theta). \end{aligned} \tag{4.12}$$

This gives us two sliding scales of partial identification that we can set independently according to the credibility of the corresponding assumptions for the problem at hand. Here $\eta = 0$ and $\varepsilon = 1$ yields the MMTR prior while $\eta = 1$ and $\varepsilon = 0$ yields the MTS prior. When $\eta = \varepsilon = 1$, we have the MTS+MMTR prior.

When we constructed the PMTS prior, we assumed that when MTS did not hold the prior support was the entire unconstrained space. A reasonable alternative is to restrict the prior support to the region of the space that contradicts MTS. This is the strategy that Bollinger and Hasselt (2009) used for constructing priors that partially identify measurement error models. Theoretically we could do this for both the MTS and MMTR assumptions, but it is not as simple in our context. For example the MTS assumption consists of two inequalities and when MTS is false, one or both inequalities is inverted giving three possibilities for not-MTS. With

MTS+MMTR the number of inequalities increases to four yielding seven possibilities for not-MTS+MMTR. The upshot is that computing, e.g., θ_k's conditional support given θ_{-k} under not-MTS+MMTR will be more complicated. The complication is easily surmountable, but it is not clear what we gain.

4.4 Partial Identification of the Treatment Effect Distributions

In this section we specify priors on $\boldsymbol{p}_{|0} = (p_{00|0}, p_{01|0}, p_{10|0}, p_{11|0})$ and $\boldsymbol{p}_{|1} = (p_{00|1}, p_{01|1}, p_{10|1}, p_{11|1})$ as well as p_d using a variety of the assumptions discussed above. This will allow us to learn more about the full treatment effects distributions – in particular more than just the mean. We will start with the NLN prior – strictly speaking this is a prior on the simplex, so when we write $p \sim \mathcal{NLN}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ for a scalar p, implicitly the distribution is on $(p, 1-p)$ with $1 - p = e^{\lambda_1}/(e^{\lambda_1} + e^{\lambda_2})$. When p is not a scalar, we will take it to satisfy the simplex constraints and require $\boldsymbol{\mu}$ & $\boldsymbol{\sigma}^2$ to have the same dimension as p. So for the unconstrained prior let $\phi = (p_d, \boldsymbol{p}_{|0}, \boldsymbol{p}_{|1})$ and assume

$$p_{UN}(\phi) = \mathcal{NLN}(p_d; \boldsymbol{\mu}_{1:2}, \boldsymbol{\sigma}^2_{1:2})\mathcal{NLN}(\boldsymbol{p}_{|0}; \boldsymbol{\mu}_{3:6}, \boldsymbol{\sigma}^2_{3:6})\mathcal{NLN}(\boldsymbol{p}_{|1}; \boldsymbol{\mu}_{7:10}, \boldsymbol{\sigma}^2_{7:10}). \tag{4.13}$$

It turns out that this prior implicitly imposes dependence between $p_{0|0}$ and $p_{1|0}$ as well as between $p_{0|1}$ and $p_{1|1}$. The dependence comes through the definition of the probabilities and can be more easily seen with independent Dirichlet priors on $\boldsymbol{p}_{|0}$ and $\boldsymbol{p}_{|1}$. On the one hand this dependence is disturbing – it gives us partial identification before we impose any constraints. But this dependence is actually very natural and based on the structure of the problem – if we learn about $p_{0|0}$ we should learn about $p_{1|0}$ because of their common component $p_{11|0}$. This relationship is what drives Manski (1999)'s worst case bounds. From the definitions of $p_{0|0}$, $p_{1|0}$, $p_{0|1}$, and $p_{1|1}$ we get the following set of inequalities:

$$\begin{array}{llll} p_{10|0} < p_{0|0} & p_{01|0} < 1 - p_{0|0} & p_{00|0} < 1 - p_{0|0} & p_{11|0} < p_{0|0} \\ p_{10|1} < 1 - p_{1|1} & p_{01|1} < p_{1|1} & p_{00|1} < 1 - p_{1|1} & p_{11|1} < p_{1|1}. \end{array}$$

In other words, the bounds are informative on each of the underlying probabilities of the basic model. This leads to the follow set of inequalities

$$-p_{0|0} < p_{01|0} - p_{10|0} < (1 - p_{0|0})$$
$$-(1 - p_{1|1}) < p_{01|1} - p_{10|1} < p_{1|1}$$

which then implies that

$$-(1 - p_{1|1})p_d - p_{0|0}(1 - p_d) < ATE < p_{1|1}p_d + (1 - p_{0|0})(1 - p_d)$$
$$-(1 - p_{1|1}) < ATT < p_{1|1}$$

which are just the worst case bounds from Manski (1999). We can see how the bounds impact the full distribution as well:

$$P(TE_{new} = -1) < p_{0|0}(1 - p_d) + (1 - p_{1|1})p_d, \quad P(TE_{new} = 1) < (1 - p_{0|0})(1 - p_d) + p_{1|1}p_d$$
$$P(TE_{new} = -1|d_{new} = 1) < 1 - p_{1|1}, \qquad P(TE_{new} = 1|d_{new} = 1) < p_{1|1}.$$

However, there are no binding bounds on $P(TE_{new} = 0)$ or $P(TE_{new} = 0|d = 1)$ under any circumstances – they are both always sharply bounded between 0 and 1. So the only way we can learn about the treatment effects distribution with the unconstrained prior is by chopping off the lower part of the distribution or the upper part which then pulls the mean up or down, but we never learn about the probability of no effect.

The MTS, MMTR, and MTS+MMTR versions of this prior are constructed analogously to the previous section:

$$p_{MTS}(\phi) \propto p_{UN}(\phi)\mathbb{1}\{p_{10|0} + p_{11|0} > p_{10|1} + p_{11|1}\}\mathbb{1}\{p_{01|0} + p_{11|0} > p_{01|1} + p_{11|1}\} \tag{4.14}$$
$$p_{MMTR}(\phi) \propto p_{UN}(\phi)\mathbb{1}\{p_{01|0} > p_{10|0}\}\mathbb{1}\{p_{01|1} > p_{10|1}\} \tag{4.15}$$
$$p_{MTS+MMTR}(\phi) \propto p_{MTS}(\phi)\mathbb{1}\{p_{01|0} > p_{10|0}\}\mathbb{1}\{p_{01|1} > p_{10|1}\}. \tag{4.16}$$

Each of these priors will allow us to learn more about the ATE and ATT distributions than their mean, though none of them identifies the full distributions.

4.4.1 Monotone Treatment Response

By specifying priors on the $p_{ab|c}$'s we are now able to apply MTR. Recall that MTR says that $y_i(1) \geq y_i(0)$ for all i, in other words households cannot be hurt by going on the treatment.

In terms of the model parameters this is equivalent to $p_{10|0} = p_{10|1} = 0$. This further implies that

$$p_{0|0} = p_{10|0} + p_{11|0} = p_{11|0}$$

and

$$p_{1|1} = p_{01|1} + p_{11|1} = 1 - p_{00|1} - p_{10|1} = 1 - p_{00|1}$$

which identifies $p_{11|0}$ and $p_{00|1}$. This assumption is not mutually exclusive with MTS either – they can be combined in order to buy even more identification power.

The next question is whether we can learn about the full distribution of treatment effect parameters using MTR. For both the ATT and ATE distributions in (4.5) and (4.6), $P(TE_{new} = -1) = 0$ by assumption using MTR. Now consider just the ATT distribution in (4.6). Since $p_{00|1} = 1 - p_{1|1}$, we are able to learn about $P(TE_{new} = 1|d_{new} = 1)$. This also allows us to learn about $P(TE_{new} = 1|d_{new} = 1)$ since MTR ensures that this distribution is a binary distribution and thus is completely determined by a single parameter. The logic is the same for the ATE distribution in (4.5) except we also need to use the fact that $p_{11|0} = p_{0|0}$ according to MTR.

The MTR prior is the unconstrained prior modified so that $p_{10|0} = p_{10|1} = 0$ or, equivalently it sets $\mu_4 = \mu_8 = -\infty$. So the density is

$$p_{MTR}(\phi) = \mathcal{NLN}(p_d; \boldsymbol{\mu}_{1:2}, \boldsymbol{\sigma}^2_{1:2})\mathcal{NLN}(\boldsymbol{p}_{|0}; \boldsymbol{\mu}_{3:6}, \boldsymbol{\sigma}^2_{3:6})\mathcal{NLN}(\boldsymbol{p}_{|1}; \boldsymbol{\mu}_{7:10}, \boldsymbol{\sigma}^2_{7:10}) \tag{4.17}$$

with $\mu_4 = \mu_8 = -\infty$. The MTS+MTR density is then

$$p_{MTS+MTR}(\phi) \propto p_{MTR}(\phi)\mathbb{1}\{p_{10|0} + p_{11|0} > p_{10|1} + p_{11|1}\}\mathbb{1}\{p_{01|0} + p_{11|0} > p_{01|1} + p_{11|1}\}. \tag{4.18}$$

MTR+MMTR is redundant with MTR since MTR implies MMTR, so we have exhausted all possible combinations of the three basic partial identification assumptions.

4.4.2 PMTS+PMMTR+PMTR

Now that we have set up the prior for each of the six possible assumptions we can make – unconstrained, MTS, MTR, MMTR, MTS + MMTR, and MTS + MTR – we can put sliding scales on each of these assumptions to create the PMTS+PMTR+PMMTR prior. As before, let $\eta = P(\text{MTS})$ and $\varepsilon = P(\text{MMTR})$. Further, define $\delta = P(\text{MTR}|\text{MMTR})$. Then

$$\begin{aligned} p_{PMTS+PMMTR+PMTR}(\phi) = &(1-\eta)(1-\varepsilon)p_{UN}(\phi) + (1-\eta)\varepsilon(1-\delta)p_{MMTR}(\phi) \\ &+ (1-\eta)\varepsilon\delta p_{MTR}(\phi) + \eta(1-\varepsilon)p_{MTS}(\phi) \\ &+ \eta\varepsilon(1-\delta)p_{MTS+MMTR}(\phi) + \eta\varepsilon\delta p_{MTS+MTR}(\phi). \end{aligned} \tag{4.19}$$

The main wrinkle with choosing $(\eta, \varepsilon, \delta)$ is that δ is a conditional probability. So this effectively gives us three sliding scales to adjust on the prior to determine how strongly we assert the identifying assumptions.

4.5 Post-stratification through Hierarchical Modeling

The previous analyses can be applied to an entire population, though often a representative sample from that population is not available. Instead we can fit each of the above models to separate sub-populations and then post-stratify by simulating new from the population distribution and then simulating those households' treatment effects from the predictive distribution of the model. Analogous to fixed effects, we can fit a model to each of these sub-populations completely separately but this throws away information. It seems likely that each sub-population's parameters will be somewhat close to each other, and a hierarchical model can allow us to capture this intuition while allowing information from one sub-population to spill over to another sub-population's parameters.

We will set up the hierarchical model using $\phi = (p_d, \boldsymbol{p}_{|0}, \boldsymbol{p}_{|1})$, though for the model with θ everything is analogous. Let $g = 1, 2, \ldots, G$ denote each sub-population or subgroup, and let ϕ_g denote that group's probabilities and ϕ_{kg} denote the k'th element of ϕ_g. Using the LLN prior on ϕ_g for $g = 1, 2, \ldots, G$ – potentially with constraints – we have

$$\phi_{1g} = \frac{e^{\lambda_{2g}}}{e^{\lambda_{1g}} + e^{\lambda_{2g}}}$$

and for $k = 3, 4, 5, 6$

$$\phi_{kg} = \frac{e^{\lambda_{kg}}}{\sum_{j=1}^{4} e^{\lambda_{jg}}} \qquad \& \qquad \phi_{k+4,g} = \frac{e^{\lambda_{k+4,g}}}{\sum_{j=1}^{4} e^{\lambda_{j+4,g}}}$$

with

$$\boldsymbol{\lambda}_g \overset{ind}{\sim} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\mathbb{1}\{\boldsymbol{\lambda}_g \in A\}$$

for $g = 1, 2, \ldots, G$ where we define $\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\sigma}^2)$. So now each $\boldsymbol{\lambda}_g$ comes from the same distribution which depends on the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$. Here $A \subseteq \Re^{10G}$ represents the subset of λ-space we are restricted to by whatever identifying assumptions we use, e.g. from MTS or MTS+MMTR. Each subgroup has the same set of constraints applied to its parameters, and this will make applying the probable version the priors straightforward. In that case each subgroup gets the same constraints obtained from the same set of sliding scales (η, ε, and δ). Each group could have its own set of sliding scales and therefore its own set of constraints, but this increases the number of sliding scales that need to be set a priori to $3G$. We will focus on the case where each group uses the same constraint.

Now we complete the model with the following prior:

$$\begin{aligned} \mu_k|\boldsymbol{\sigma}^2 &\overset{ind}{\sim} \mathcal{N}\left(\bar{\mu}_k, \frac{\sigma_k^2}{S_k\gamma_k}\right) \\ \sigma_k^2 &\overset{ind}{\sim} \mathcal{IG}\left(\frac{\nu_k}{2}, \frac{\nu_k}{2}S_k\right) \end{aligned}$$

for $k = 1, 2, \ldots, 10$. Under normal circumstances the inverse gamma prior on σ_k^2 is highly informative and can have an undue influence on inference (Gelman, 2006), especially on the variance in question. In this case the variances are part of an additional set of parameters in order to create a more flexible distribution on the simplex and in order to allow for higher level modeling. The inverse gamma prior reduces the flexibility a bit, but not much in practice. Furthermore, the constraining properties of the inverse gamma prior are actually desirable since we need these variances to be somewhat constrained near zero for two reasons. First, for large variances the MCMC algorithm will have problems for the μ's since they are unidentified and their prior is so diffuse – in the non-hierarchical case this means that we have to set σ_k^2 small or at least not much larger than one. Second, for a diffuse enough prior on σ^2 we find that no pooling happens at the level of the $p_{ab|c}$'s – for example with a Half-t prior on each of the σ_k's with a degrees of freedom near 10, the identified parameters, $p_{0|0,g}$, $p_{1|1,g}$, and $p_{d,g}$, are essentially estimated to be the empirically observed probabilities within those groups. This is not the case for large degrees of freedom (> 100) or for the inverse gamma prior. The inverse gamma prior combined with the normal prior on μ_k also has the benefit of being jointly conditionally conjugate to keep MCMC simple and fast.

In order to fully specify the prior we need to specify $\bar{\mu}_k$, γ_k, ν_k, and S_k for $k = 1, 2, \ldots, 10$, so 40 parameters. Shrinkage is controlled by S_k – the smaller S_k is, the more each group's probabilities are shrunk towards each other. Shrinkage towards a prior set of probabilities is controlled by γ_k, and the reason S_k appears in the prior for μ_k is to ensure that γ_k alone controls this sort of shrinkage rather than both γ_k and S_k. *Larger* values of γ_k cause shrinkage of the estimated probabilities to the prior set of probabilities. The prior set of probabilities is controlled in a complicated fashion by all of the hyperparameters, but primarily by the $\bar{\mu}_k$'s. These hyperparameters are only identified up to an additive constant even if we could directly observe the λ_{kg}'s, so only relative values matter. Finally ν_k can manipulate shrinkage to some extent as well. With $\gamma_k = 1$, $\nu_k = 1$, $\bar{\mu}_k = 0$, and $S_k = 1$ for all i and a modest amount of data in each subgroup – $n = 500$ to $n = 1000$ – each group's set of probabilities is essentially estimated independently of the prior and the other groups. As S_k decreases these probabilities are shrunk back towards each other and as γ_k increases they are shrunk towards a prior set of probabilities – a factor of 100 or so is enough to see meaningful changes in the estimates in both cases when the number of observations per group is around 1,000. In order to choose these hyperparameters more intelligently, the method of Gelman (1995) can also be applied.

4.6 MCMC

We will construct two MCMC algorithms here – one for the model using θ and one for the model using ϕ – then extend both algorithms to deal with the hierarchical version of both models. Both algorithms we construct assume a fixed set of constraints are used. In order to account for probable constraints the algorithm should be run for each set of fixed constraints allowed by the prior.

We will start with the model using $\theta = (p_d, p_{0|0}, p_{1|0}, p_{0|1}, p_{1|1})$. The posterior for this model in terms of the λ's used in the NLN prior is

$$p(\boldsymbol{\lambda}|\boldsymbol{d},\boldsymbol{y}) \propto \frac{e^{\lambda_1 F_d} e^{\lambda_2 T_d}}{(e^{\lambda_1}+e^{\lambda_2})^N} \frac{e^{\lambda_3 F_0} e^{\lambda_4 T_0}}{(e^{\lambda_3}+e^{\lambda_4})^{F_d}} \frac{e^{\lambda_9 F_1} e^{\lambda_{10} T_1}}{(e^{\lambda_9}+e^{\lambda_{10}})^{T_d}} \exp\left[-\frac{1}{2}\sum_{k=1}^{10}\frac{(\lambda_k-\mu_k)^2}{\sigma_k^2}\right] \mathbb{1}\{\boldsymbol{\lambda}\in A\} \tag{4.20}$$

where $T_d = \sum_i d_i$, $F_d = N - T_d$, $T_0 = \sum_i (1-d_i)y_i$, $F_0 = F_d - T_0$, $T_1 = \sum_i d_i y_i$, and $F_1 = T_d - T_1$ are the observed counts in each of the relevant categories. Here $A \subseteq \Re^{10}$ is the set $\boldsymbol{\lambda}$ is constrained to lie in by the prior, determined by which of the MTS and MMTR constraints hold. These constraints in terms of the λ's are

$$\begin{array}{llcl} \text{MTS:} & \lambda_7+\lambda_4 > \lambda_3+\lambda_8 & \& & \lambda_9+\lambda_6 > \lambda_5+\lambda_{10} \\ \text{MMTR:} & \lambda_3+\lambda_6 > \lambda_4+\lambda_5 & \& & \lambda_7+\lambda_{10} > \lambda_8+\lambda_9. \end{array}$$

The full conditional density of each λ_k has the form

$$p^*(\lambda_k|\boldsymbol{\lambda}_{-k},\boldsymbol{d},\boldsymbol{y}) \propto \frac{e^{\lambda_k T_k}}{(e^{\lambda_k}+C_k)^{N_k}} e^{-\frac{1}{2\sigma_k^2}(\lambda_k-\mu_k)^2} \mathbb{1}\{\lambda_k \in (a_k,b_k)\} \tag{4.21}$$

where T_k & N_k are functions of $\boldsymbol{y}$ and C_k, a_k, and b_k are functions of $\boldsymbol{\lambda}_{-k}$ with $-\infty \le a_k < b_k \le \infty$. The functions a_k and b_k come directly from the constraint we are assuming, e.g. MTS or MMTR or both or neither. When $T_k = N_k = 0$, e.g. for λ_4, this distribution is simply a scalar truncated normal. There are several well known algorithms for efficiently drawing from a truncated normal distribution, e.g. Geweke (1991). When T_k or N_k is nonzero the density is nonstandard and difficult to sample from. Instead we use a random walk Metropolis-Hastings step as follows:

1. Simulate $\lambda_k^{(*)} \sim \mathcal{N}_{(a_k,b_k)}(\lambda_k^{(t)}, u_k^2)$.

2. Compute the acceptance ratio

$$R = \frac{p^*(\lambda_k^{(*)}|\boldsymbol{\lambda}_{-k},\boldsymbol{y})}{p^*(\lambda_k^{(t)}|\boldsymbol{\lambda}_{-k},\boldsymbol{y})} \frac{\Phi\left(\frac{b_k-\lambda_k^{(t)}}{u_k}\right)-\Phi\left(\frac{a_k-\lambda_k^{(t)}}{u_k}\right)}{\Phi\left(\frac{b_k-\lambda_k^{(*)}}{u_k}\right)-\Phi\left(\frac{a_k-\lambda_k^{(*)}}{u_k}\right)}$$

   and set $\lambda_k^{(t+1)} = \lambda_k^{(*)}$ with probability $\min(1,R)$ and $\lambda_k^{(t+1)} = \lambda_k^{(t)}$ with probability $1 - \min(1,R)$.

Here $\Phi(.)$ is the standard normal cdf and u_k is a tuning parameter. In step 1 the proposal, $\lambda_k^{(*)}$, is simulated from a truncated normal distribution instead of an unconstrained normal, which is why this is not simply a random walk Metropolis step. While this is not strictly speaking a random walk Metropolis step we can still adaptively set the value of u_k during the burn in period using the ideas in Roberts and Rosenthal (2009) in order to achieve a target acceptance rate of about 0.44. The full algorithm for the single group model using θ is then a Gibbs sampler that consists of six Metropolis-Hastings steps for λ_1, λ_2, λ_3, λ_4, λ_9, and λ_{10} using the algorithm in (4.6), and four steps where we draw from a scalar truncated normal.

The MCMC algorithm for the single group model using $\phi = (p_d, \boldsymbol{p}_{|0}, \boldsymbol{p}_{|1})$ is similar but slightly more complicated. Here we use a data augmentation algorithm (Tanner and Wong, 1987) where we draw the missing counterfactual for each observation. So the full set of augmented data – both observed and missing data – is $(\boldsymbol{d}, \boldsymbol{y}(0), \boldsymbol{y}(1))$. The full augmented data posterior can be written as

$$p(\boldsymbol{\lambda}, \boldsymbol{y}(\mathbf{1}-\boldsymbol{d})|\boldsymbol{d}, \boldsymbol{y}(\boldsymbol{d})) \propto \frac{e^{\lambda_1 T_1} e^{\lambda_2 T_2}}{(e^{\lambda_1}+e^{\lambda_2})^{T_1+T_2}} \frac{e^{\lambda_3 T_3} e^{\lambda_4 T_4} e^{\lambda_5 T_5} e^{\lambda_6 T_6}}{(e^{\lambda_3}+e^{\lambda_4}+e^{\lambda_5}+e^{\lambda_6})^{T_3+T_4+T_5+T_6}}$$
$$\times \frac{e^{\lambda_7 T_7} e^{\lambda_8 T_8} e^{\lambda_9 T_9} e^{\lambda_{10} T_{10}}}{(e^{\lambda_7}+e^{\lambda_8}+e^{\lambda_9}+e^{\lambda_{10}})^{T_7+T_8+T_9+T_{10}}} \exp\left[-\frac{1}{2}\sum_{k=1}^{10}\frac{(\lambda_k-\mu_k)^2}{\sigma_k^2}\right] \mathbb{1}\{\boldsymbol{\lambda} \in A\} \tag{4.22}$$

where we define

$$\begin{aligned}
&T_1 = \#(d_i = 0), && T_2 = \#(d_i = 1),\\
&T_3 = \#(d_i = 0, y_i(0) = 0, y_i(1) = 0), && T_4 = \#(d_i = 0, y_i(0) = 1, y_i(1) = 0),\\
&T_5 = \#(d_i = 0, y_i(0) = 0, y_i(1) = 1), && T_6 = \#(d_i = 0, y_i(0) = 1, y_i(1) = 1),\\
&T_7 = \#(d_i = 1, y_i(0) = 0, y_i(1) = 0), && T_8 = \#(d_i = 1, y_i(0) = 1, y_i(1) = 0),\\
&T_9 = \#(d_i = 1, y_i(0) = 0, y_i(1) = 1), \quad \text{and} && T_{10} = \#(d_i = 1, y_i(0) = 1, y_i(1) = 1)
\end{aligned}$$

and recall

$$\phi_1 = \frac{e^{\lambda_2}}{e^{\lambda_1}+e^{\lambda_2}},$$

for $k = 2, 3, 4, 5$

$$\phi_k = \frac{e^{\lambda_{k+1}}}{\sum_{j=3}^{6} e^{\lambda_j}},$$

and for $k = 6, 7, 8, 9$

$$\phi_k = \frac{e^{\lambda_{k+1}}}{\sum_{j=7}^{10} e^{\lambda_j}}.$$

Here the MTS, MMTR, and MTR constraints in terms of the λ's are

$$\begin{aligned} \text{MTS:} \qquad & (e^{\lambda_5} + e^{\lambda_6})(e^{\lambda_7} + e^{\lambda_8}) > (e^{\lambda_9} + e^{\lambda_{10}})(e^{\lambda_3} + e^{\lambda_4}) \\ \& \qquad & (e^{\lambda_4} + e^{\lambda_6})(e^{\lambda_7} + e^{\lambda_9}) > (e^{\lambda_8} + e^{\lambda_{10}})(e^{\lambda_3} + e^{\lambda_5}) \\ \text{MMTR:} \qquad & \lambda_4 < \lambda_5 \;\; \& \;\; \lambda_8 < \lambda_9 \\ \text{MTR:} \qquad & \lambda_4 = -\infty \;\; \& \;\; \lambda_8 = -\infty. \end{aligned}$$

In the first step of the algorithm we need to draw $\boldsymbol{y}(\mathbf{1} - \boldsymbol{d})$ conditional on ϕ so we can form the T_k's. Their full conditional density is

$$p(\boldsymbol{y}(\mathbf{1} - \boldsymbol{d})|\phi, \boldsymbol{d}, \boldsymbol{y}(\boldsymbol{d})) = \prod_{i=1}^{n} p(y_i(1 - d_i)|\phi, d_i, y_i(d_i))$$

where

$$p(y_i(1)|\phi, d_i = 0, y_i(0) = y_i) \propto \left(\phi_2^{1-y_i}\phi_3^{y_i}\right)^{1-y_i(1)} \left(\phi_4^{1-y_i}\phi_5^{y_i}\right)^{y_i(1)}$$

and

$$p(y_i(0)|\phi, d_i = 1, y_i(1) = y_i) \propto \left(\phi_7^{1-y_i}\phi_9^{y_i}\right)^{1-y_i(0)} \left(\phi_8^{1-y_i}\phi_{10}^{y_i}\right)^{y_i(0)}.$$

So when $d_i = 0$, $y_i(0) = y_i$ and

$$y_i(1) \sim \text{Ber}\left(\frac{\phi_4^{1-y_i}\phi_5^{y_i}}{\phi_2^{1-y_i}\phi_3^{y_i} + \phi_4^{1-y_i}\phi_5^{y_i}}\right)$$

while when $d_i = 1$, $y_i(1) = y_i$ and

$$y_i(0) \sim \text{Ber}\left(\frac{\phi_8^{1-y_i}\phi_{10}^{y_i}}{\phi_7^{1-y_i}\phi_9^{y_i} + \phi_8^{1-y_i}\phi_{10}^{y_i}}\right).$$

From (4.22), the full conditional of each of the λ_k's has the form in (4.21), so we can use the algorithm in (4.6) for each λ_k step. Putting it all together, we first draw each $y_i(1 - d_i)$ from their full conditionals, independent across i, then we draw each λ_k conditional on the others using the Metropolis-Hastings algorithm above (4.6).

In the hierarchical version of both models we need to make the same two changes. First, each group g has its own $\boldsymbol{\lambda}_g$ which needs to be drawn in a separate set of Gibbs steps. Conditional on $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$, the $\boldsymbol{\lambda}_g$'s are independent across g so this step can be parallelized. Second, we need to draw $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ from their full conditional distribution. Since (μ_k, σ_k^2) is independent across k in the prior, they are also independent across k in their conditional posterior. Furthermore, the Normal-inverse Gamma prior on (μ_k, σ_k^2) is conditionally conjugate, so we can draw them jointly as follows (Bernardo and Smith, 2009):

1. Draw $\sigma_k^2 \sim \mathcal{IG}(\hat{a}_k, \hat{b}_k)$ where

$$\hat{a}_k = \frac{\nu_k + G}{2} \qquad \text{and} \qquad \hat{b}_k = \frac{1}{2}\left(\nu_k S_k + S_{\lambda_k}^2 + \frac{\gamma_k S_k G(\bar{\mu}_k - \bar{\lambda}_k)^2}{\gamma_k S_k + G}\right)$$

2. Draw $\mu_k \sim \mathcal{N}(\hat{\mu}_k, \hat{\sigma}_k^2)$ where

$$\hat{\mu}_k = \frac{G\bar{\lambda}_k + \gamma_k S_k \bar{\mu}_k}{G + \gamma_k S_k} \qquad \text{and} \qquad \hat{\sigma}_k^2 = \frac{\sigma_k^2}{G + \gamma_k S_k}.$$

Here $\bar{\lambda}_k = \sum_g \lambda_{kg}/G$ and $S_{\lambda_k}^2 = \sum_g (\lambda_{kg} - \bar{\lambda}_k)^2$. This algorithm will work well as long as G, the number of subgroups, is not too large. The number of Metropolis-Hastings steps required every iteration is $10G$ in both hierarchical models, so for a large number of groups this may be costly. For example with about 2000 groups a single iteration of the algorithm may take as long as a full second when programmed in `R`.

4.7 Analyzing the NSLP

Next we fit the models above to data from the national school lunch program. Our data comes from the 2001–2004 National Health and Nutrition Examination Survey (NHANES), conducted by the National Center for Health Statistics, Centers for Disease Control (NCHS/CDC). The NHANES uses surveys and physical examinations on a sample of about 5000 people per year, half of which are children. Vulnerable groups are over-sampled. Detailed measures on a variety of health related outcomes are included in the NHANES. For now, we restrict our attention to 2693 children in the sample who appear to be eligible to receive free or reduced price lunches through the NSLP. This includes children ages 6 to 17 who reside in households with income less than 185% of the federal poverty line and are reported to be attending schools with the NSLP. Parents also self report their participation in the school lunch program. We focus on one outcome: food security. Food security is a binary variable (1 = food secure, 0 = food insecure) that is measured with a series of 18 questions about food-related needs and resources in the household, e.g. "I worried whether our food would run out before we got money to buy more." The household is considered to be food insecure if the respondent answers affirmatively to three or more of these questions.

Table 4.1 contains summary statistics for each of the key variables. About 77% of eligible children in the sample are in households taking advantage of the NSLP. About 58% of recipients and 67% of non-recipients are food secure, so a naive comparison of households on the the NSLP to households not on the NSLP would suggest that the NSLP actually *decreases* food security. Note, however, that non-recipients appear to have higher income on average than recipients.

Table 4.1: Summary of key variables by National School Lunch Program participation. Note that these statistics do not take into account the sample weights.

	Income-eligible children	Recipients	Non-recipients
Age in years	11.88	11.41	13.47
	(3.33)	(3.21)	(3.21)
NSLP recipient	0.77	1	0
	(0.42)	(0)	(0)
Ratio of income to the poverty line	0.92	0.88	1.06
	(0.47)	(0.46)	(0.49)
Food secure household	0.6	0.58	0.67
	(0.49)	(0.49)	(0.47)

In order to fit the model we broke the dataset into six groups based on the ratio of their income to the poverty line as follows: $[0, 0.4]$, $(0.4, 0.7]$, $(0.7, 1]$, $(1, 1.3]$, $(1.3, 1.6]$, and $(1.6, 1.85]$. We fit each model using θ and each model using ϕ with each fixed set of constraints, including with no constraints. Once for θ and once for ϕ we fit a model with probabilistic constraints. In the probabilistic constraint model for θ we set $\eta = P(\text{MTS}) = 0.8$ and $\varepsilon = P(\text{MMTR}) = 0.5$. We used these same values for η and ε in the probabilistic model for ϕ and in addition we set $\delta = P(\text{MTR}|\text{MMTR}) = 0.25$. In both models we set the hyperparameters as follows: for $k = 1, 2, \ldots, 10$, $\bar{\mu}_k = 0$, $\gamma_k = 1$, $\nu_k = 1$, and $S_k = 1/10$. One Markov chain was obtained for each model fit with 330,000 iterations, 30,000 of which were used as the tuning period for the Metropolis-Hastings steps and were thrown away as burn in. Because so many of the parameters of interest are unidentified in the likelihood by design, it takes a very large sample size in order to adequately characterize the posterior. Essentially the chain mixes very poorly for the unidentified parameters, and some of those parameters are directly relevant to our scientific question.

Table 4.2 contains the post-stratified estimates of the mean of the ATE distribution and the mean of the ATT distribution using the models based on θ. In the unconstrained model the only source of identification is the worst case bounds and, as a result, the estimates of the mean treatment effects parameters are close to zero. Furthermore, the credible intervals do not contain zero. Both the MTS and MMTR assumptions strongly suggest that the mean ATE and mean ATT are larger than zero with more uncertainty about ATT, and combining these two assumptions only serves to reinforce this. With the probabilistic prior the estimates are still large but the intervals are much wider and still include zero, suggesting that we do not have strong evidence of a positive treatment effect.

Table 4.2: Estimates of post-stratified mean treatment effects parameters based on models using θ under a variety of prior assumptions. Under Prob we assume that $\eta = 0.8$ and $\varepsilon = 0.5$. Intervals are 95% posterior credible intervals for the mean of the post-stratified predictive distribution as a function of the model parameters.

	E[ATE]	2.5%	97.5%	E[ATT]	2.5%	97.5%
UN	0.10	-0.39	0.60	0.22	-0.35	0.81
MTS	0.45	0.11	0.64	0.53	0.08	0.83
MMTR	0.45	0.11	0.64	0.53	0.08	0.82
MTS+MMTR	0.46	0.13	0.64	0.54	0.10	0.83
Prob	0.42	-0.17	0.64	0.42	-0.17	0.64

The estimates for the model using ϕ are in Table 4.3. The unconstrained prior yields similar estimates for the mean of the treatment effects distributions in this model as it does in the model using θ, though it has a much smaller degree of uncertainty. The posterior credible interval for the mean is much narrower under the ϕ prior though still centered at the same value. Under the other priors that are shared across both models (MTS, MMTR, MTS+MMTR) the model using θ yields substantially higher estimates and intervals bounded significantly farther away from zero. This could be driven by the priors – while the MTS prior for both models uses the same hyperparameters, the meaning of these hyperparameters is different so it may be that in the θ model it puts more mass on a positive treatment effects or has a greater or lesser degree of shrinkage between the groups. This issue merits further investigation. As a result of the difference between the θ and ϕ models in the unconstrained prior, the probabilistic prior under ϕ yields a credible interval which only barely contains zero and more strongly suggests a positive treatment effect.

A big difference in the ϕ model is that we are able to do more than estimate and construct intervals for the mean of the treatment effects distribution – we can look at the full posterior predictive treatment effects distributions and integrate out the model parameters. In Table 4.3 we see posterior predictive probabilities of obtaining a negative, neutral, and positive treatment effect under a variety of prior assumptions. We can see that the constrained priors seem to differ compared to the unconstrained prior typically by moving mass from a negative effect to either a neutral or positive effect, though occasionally mass is moved from a neutral effect to a positive effect. By and large most of the estimates of the probability of a negative effect are at about 0.1, though keep in mind that the MTR assumption forces this probability to be zero. So it seems fairly likely that the NSLP is not hurting, though the posterior predictive probability of a positive effect is generally about 0.3 to 0.4.

4.8 Conclusions and Further Work

Typically in treatment effect problems such as program evaluation, we assume that certain unidentified parameters are bound by some function of identified parameters in order to partially identify the treatment effects. While these assumptions are useful to help identify the parameters we are interested in, they are not always credible. Rather than assume that they hold with certainty we develop an approach that allows us to assume that a given constraint holds with some probability. Using this approach and several commonly used constraints, we constructed two models and a prior for each that allows for a sliding scale of partial identification depending on how strongly we assert each of the constraints hold. Both models were extended to a hierarchical analysis of sub-populations which allows for post-stratification to

Table 4.3: Estimates of the post-stratified treatment effects distributions based on models using ϕ under a variety of prior assumptions. Under Prob we assume that $\eta = 0.8$, $\varepsilon = 0.5$, and $\delta = 0.25$. Intervals are 95% posterior credible intervals for the mean of the post-stratified predictive distribution as a function of the model parameters. Both the ATE and ATT distributions are discrete on $\{-1, 0, 1\}$ and the post-stratified posterior predictive probabilities of each of these possibilities is listed to the right of the estimate of the mean.

	E[ATE]	2.5%	97.5%	P(ATE=-1)	P(ATE=0)	P(ATE=1)
UN	0.10	-0.26	0.45	0.20	0.49	0.31
MTS	0.29	0.05	0.54	0.11	0.49	0.40
MMTR	0.31	0.05	0.57	0.08	0.53	0.39
MTR	0.30	0.02	0.59	0.00	0.70	0.30
MTS+MMTR	0.30	0.07	0.54	0.10	0.51	0.40
MTS+MTR	0.35	0.07	0.61	0.00	0.65	0.35
Prob	0.28	-0.04	0.56	0.10	0.52	0.38

	E[ATT]	2.5%	97.5%	P(ATT=-1)	P(ATT=0)	P(ATT=1)
UN	0.22	-0.20	0.70	0.14	0.49	0.37
MTS	0.35	0.03	0.71	0.13	0.39	0.48
MMTR	0.38	0.03	0.75	0.09	0.44	0.47
MTR	0.36	0.00	0.77	0.00	0.64	0.36
MTS+MMTR	0.36	0.05	0.71	0.12	0.41	0.47
MTS+MTR	0.41	0.06	0.79	0.00	0.59	0.41
Prob	0.28	-0.04	0.56	0.10	0.52	0.38

correct for sample-population mismatch and to potentially learn about the differences between various sub-populations.

We then applied the models to analyzing whether the NSLP increases food security among income-eligible children by breaking children in the data set into groups based on the ratio of household income to the poverty level. This analysis suggests that the program at least does no harm under a wide variety of assumptions but there is not strong evidence that the program helps. One key problem with the analysis, however, is the only covariate we took into account is income. This causes two problems. First it is likely that there is more information in the data about which sorts of children are more likely to be helped by the program just by taking into account race, parents' age, etc, and this is likely biasing our estimates. Second, the sample is likely biased relative to the population based on more than just income, so post-stratification should take this into account.

A key problem with expanding the number of groups is that MCMC slows down as the number of groups increases, and the number of groups increases quickly with the number of covariates we are interested in. In the above analysis we used six income groups. If we add five parents' age categories, four categories for race plus a Hispanic indicator, a sex indicator, two child's age categories, a married parents indicator, four household size categories, and four education level categories we suddenly have $6 \times 5 \times 4 \times 2 \times 2 \times 2 \times 2 \times 4 \times 4 = 30,720$ subgroups and thus $307,200$ group level parameters. Furthermore, it is unlikely that we have more than one observation is almost all of these groups and the vast majority of them we will have zero observations. By focusing on subgroups we are effectively forcing ourselves to consider all possible interactions, which is probably not necessary. So we might be able to get the group size down by considering some sort of linear model at the group level. For example instead of giving every group its own mean we could make each group's mean is a linear function of a certain set of covariates plus a group specific error. This does not completely remove the computational problems since we will still have a large number of groups, but something on the order of 100 is much more manageable.

As it stands MCMC for the models above will work well when the number of groups is relatively small – in the hundreds at most. For many more groups than that posterior computation may take something on the order of days to complete because we need such large sample sizes due to poor mixing of unidentified parameters.

BIBLIOGRAPHY

Aitchison, J. (1986). The statistical analysis of compositional data.

Alvarez-Castro, I., Simpson, M., and Niemi, J. (2014). Covariance matrix prior distributions for hierarchical linear models. In *Kansas State University Conference on Applied Statistics in Agriculture*.

Basu, D. (1955). On statistics independent of a complete sufficient statistic. *Sankhyā: The Indian Journal of Statistics*, 15(4):377–380.

Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M. (2003). Non-centered parameterisations for hierarchical models and data augmentation. In *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*, pages 307–326. Oxford University Press, London.

Bernardo, J. M. and Smith, A. F. (2009). *Bayesian theory*, volume 405. John Wiley & Sons.

Bollinger, C. and Hasselt, M. (2009). A Bayesian analysis of binary misclassification: Inference in partially identified models. *Manuscript, University of Kentucky and University of Western Ontario*.

Bos, C. S. and Shephard, N. (2006). Inference for adaptive time series models: Stochastic volatility and conditionally Gaussian state space form. *Econometric Reviews*, 25(2-3):219–244.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455.

Carter, C. K. and Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika*, 81(3):541–553.

Crockett, S., Smith, V. L., and Wilson, B. J. (2009). Exchange and specialisation as a discovery process. *The Economic Journal*, 119(539):1162–1188.

Dagpunar, J. (1989). An easily implemented generalised inverse Gaussian generator. *Communications in Statistics-Simulation and Computation*, 18(2):703–710.

De Jong, P. and Shephard, N. (1995). The simulation smoother for time series models. *Biometrika*, 82(2):339–350.

Dempster, A. P., Laird, N. M., Rubin, D. B., et al. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal statistical Society*, 39(1):1–38.

Devroye, L. (2012). Random variate generation for the generalized inverse Gaussian distribution. *Statistics and Computing*, 24(2):1–8.

Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of Time Series Analysis*, 15(2):183–202.

Frühwirth-Schnatter, S. (2004). Efficient Bayesian parameter estimation for state space models based on reparameterizations. In *State Space and Unobserved Component Models: Theory and Applications*, pages 123–151. Cambridge University Press, Cambridge, UK.

Frühwirth-Schnatter, S. and Sögner, L. (2003). Bayesian estimation of the Heston stochastic volatility model. In Harvey, A., Koopman, S. J., and Shephard, N., editors, *Operations Research Proceedings 2002*, pages 480–485. Springer.

Frühwirth-Schnatter, S. and Sögner, L. (2008). Bayesian estimation of the multi-factor Heston stochastic volatility model. *Communications in Dependability and Quality Management*, 11(4):5–25.

Frühwirth-Schnatter, S. and Tüchler, R. (2008). Bayesian parsimonious covariance estimation for hierarchical linear mixed models. *Statistics and Computing*, 18(1):1–13.

Frühwirth-Schnatter, S. and Wagner, H. (2006). Auxiliary mixture sampling for parameter-driven models of time series of counts with applications to state space modelling. *Biometrika*, 93(4):827–841.

Frühwirth-Schnatter, S. and Wagner, H. (2010). Stochastic model specification search for Gaussian and partial non-Gaussian state space models. *Journal of Econometrics*, 154(1):85–100.

Frühwirth-Schnatter, S. and Wagner, H. (2011). Bayesian variable selection for random intercept modeling of Gaussian and non-Gaussian data. In Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M., editors, *Bayesian Statistics 9*, pages 165–200. Oxford University Press, Oxford.

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995). Efficient parametrisations for normal linear mixed models. *Biometrika*, 82(3):479–488.

Gelman, A. (1995). Method of moments using monte carlo simulation. *Journal of Computational and Graphical Statistics*, 4(1):36–54.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515–534.

Gelman, A., Bois, F., and Jiang, J. (1996). Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *Journal of the American Statistical Association*, 91(436):1400–1412.

Gelman, A. and Carlin, J. B. (2001). Poststratification and weighting adjustments. In Groves, R. M., Dillman, D., Eltinge, J., and Little, R., editors, *Survey Nonresponse*, pages 289–302. Wiley.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis (3rd ed.).* CRC press, New York.

Gelman, A. and Little, T. C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology*, 23:127–135.

Geweke, J. (1991). Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities. In *Computing science and statistics: Proceedings of the 23rd symposium on the interface*, pages 571–578.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 41(2):337–348.

Gundersen, C., Kreider, B., and Pepper, J. (2012). The impact of the national school lunch program on child health: A nonparametric bounds analysis. *Journal of Econometrics*, 166(1):79–91.

Hobert, J. P. and Marchev, D. (2008). A theoretical comparison of the data augmentation, marginal augmentation and PX-DA algorithms. *The Annals of Statistics*, 36(2):532–554.

Huang, A. and Wand, M. (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis*, 8(2):439–452.

Jørgensen, B. (1982). *Statistical Properties of the Generalized Inverse Gaussian Distribution.* Springer, New York.

Kaldor, N. (1939). Welfare propositions of economics and interpersonal comparisons of utility. *The Economic Journal*, pages 549–552.

Kastner, G. and Frühwirth-Schnatter, S. (2014). Ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC estimation of stochastic volatility models. *Computational Statistics & Data Analysis*, 76:408–423.

Kimbrough, E. O., Smith, V. L., and Wilson, B. J. (2010). Exchange, theft, and the social formation of property. *Journal of Economic Behavior & Organization*, 74(3):206–229.

Kline, B. and Tamer, E. (2013). Default bayesian inference in a class of partially identified models. *manuscript, Northwestern University.*

Koopman, S. J. (1993). Disturbance smoother for state space models. *Biometrika*, 80(1):117–126.

Liu, J. S. and Wu, Y. N. (1999). Parameter expansion for data augmentation. *Journal of the American Statistical Association*, 94(448):1264–1274.

Manski, C. F. (1999). *Identification problems in the social sciences.* Harvard University Press.

Mas-Colell, A., Whinston, M. D., Green, J. R., et al. (1995). *Microeconomic Theory*, volume 1. Oxford university press New York.

McCausland, W. J., Miller, S., and Pelletier, D. (2011). Simulation smoothing for state–space models: A computational efficiency analysis. *Computational Statistics & Data Analysis*, 55(1):199–212.

Meng, X.-L. and Van Dyk, D. (1997). The EM algorithm–an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(3):511–567.

Meng, X.-L. and Van Dyk, D. (1998). Fast EM-type implementations for mixed effects models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3):559–578.

Meng, X.-L. and Van Dyk, D. (1999). Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika*, 86(2):301–320.

Moon, H. R. and Schorfheide, F. (2012). Bayesian and frequentist inference in partially identified models. *Econometrica*, 80(2):755–782.

Papaspiliopoulos, O. and Roberts, G. (2008). Stability of the Gibbs sampler for Bayesian hierarchical models. *The Annals of Statistics*, 36(1):95–117.

Papaspiliopoulos, O., Roberts, G. O., and Sköld, M. (2007). A general framework for the parametrization of hierarchical models. *Statistical Science*, 22(1):59–73.

Park, D. K., Gelman, A., and Bafumi, J. (2004). Bayesian multilevel estimation with post-stratification: state-level estimates from national polls. *Political Analysis*, 12(4):375–385.

Petris, G., Campagnoli, P., and Petrone, S. (2009). *Dynamic Linear Models with R*. Springer, New York.

Pitt, M. K. and Shephard, N. (1999). Analytic convergence rates and parameterization issues for the Gibbs sampler applied to state space models. *Journal of Time Series Analysis*, 20(1):63–85.

Prado, R. and West, M. (2010). *Time Series: Modeling, Computation, and Inference.* CRC Press, London.

Roberts, G. O., Papaspiliopoulos, O., and Dellaportas, P. (2004). Bayesian inference for non-Gaussian Ornstein–Uhlenbeck stochastic volatility processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2):369–393.

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive mcmc. *Journal of Computational and Graphical Statistics*, 18(2):349–367.

Roberts, G. O. and Sahu, S. K. (1997). Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2):291–317.

Rodriguez, P. P. (2009). *ars: Adaptive Rejection Sampling.* R package version 0.4, original C++ code from Arnost Komarek based on ars.f written by P. Wild and W. R. Gilks.

Rue, H. (2001). Fast sampling of Gaussian markov random fields. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):325–338.

Shephard, N. (1996). *Statistical Aspects of ARCH and Stochastic Volatility.* Springer, London.

Simpson, M. (2014). Application of interweaving in dlms to an exchange and specialization experiment. In Frühwirth-Schnatter, S., editor, *Bayesian Statistics from Methods to Models and Applications.* Springer.

Simpson, M., Niemi, J., and Roy, V. (2014). Interweaving Markov chain Monte Carlo strategies for efficient estimation of dynamic linear models. *Working Paper.*

Strickland, C. M., Martin, G. M., and Forbes, C. S. (2008). Parameterisation and efficient MCMC estimation of non-Gaussian state space models. *Computational Statistics & Data Analysis*, 52(6):2911–2930.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540.

Van Dyk, D. and Meng, X.-L. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50.

Van Dyk, D. and Meng, X.-L. (2010). Cross-fertilizing strategies for better EM mountain climbing and DA field exploration: A graphical guide book. *Statistical Science*, 25(4):429–449.

Van Dyk, D. A. and Tang, R. (2003). The one-step-late pxem algorithm. *Statistics and Computing*, 13(2):137–152.

West, M. and Harrison, J. (1999). *Bayesian Forecasting & Dynamic Models (2nd ed.)*. Springer, New York.

Yu, Y. and Meng, X.-L. (2011). To center or not to center: That is not the question - an ancillarity–sufficiency interweaving strategy (ASIS) for boosting MCMC efficiency. *Journal of Computational and Graphical Statistics*, 20(3):531–570.