Bayesian Shrinkage Among Time Varying Coefficients in Counterfactual Analysis

Danny Klinenberg

Last Updated: 2020-05-26

Abstract

Current synthetic control approaches model the coefficients as constant throughout the observation period. This method may lead to poor fits when shocks affect the treated or control units in the pretreatment period. I propose explicitely modeling the coefficients as dynamic using a Bayesian state space framework. I decompose each coefficient into a time varying portion and a constant portion. I then apply the Bayesian Lasso, an automatic process that reduces time-varying coefficients to static ones if the model is overfitting. Therefore, this model produces counterfactual estimates as accurate as a constant coefficient model when the data generating process includes only constant coefficients and better counterfactual estimates when the data generating process involves time varying coefficients. The main contribution of this paper is applying time varying coefficients to a synthetic control framework.

1 Introduction

Many important economic questions focus on policy implications at an aggregate level. Examples of these policy changes include countries entering or leaving trade unions, changes

to national healthcare, and foreign aid allocation. These major policy changes tend to happen to individual countries. One example is brexit in the United Kingdom. Researchers have proposed creating a counterfactual to analyze the effect of such policies. This paper suggests using a Bayesian structural state space model to analyze the effect of policies on aggregate levels. Unlike existing methods, this approach allows for time varying relationships between covariates and the outcome variable. Bayesian shrinkage methods are employed to set time varying coefficients as constant when the model is overfitting. Therefore, the model should perform similarly to existing methods when the relationship is relatively constant and superior when the relationship is dynamic.

Current research has focused on creating a counterfactual of the treated unit using untreated units. Suppose there exists a treated unit y_0 observed over $t = 1, ..., T_0 - 1, T_0, T_0 + 1, ...T$. Let $y_{0,t}$ represent unit y_0 in period t. Suppose T_0 represents the year in which a policy change occurs. For example, $y_{0,t}$ may represent GDP of the United Kingdom in period t and T_0 represent the year the brexit referendum passed (2016). Suppose there are also $j = 1, ..., J$ untreated aggregate units. These can represent GDPs of the United States, France, Canada, etc. The goal is to create a fake, or synthetic, treated unit in which the treatment did not occur. This would be a Great Britain in which the brexit referendum did not pass.

Let $W_{i,t}$ be an indicator that denotes the treatment status of unit i in period t. If $W_{i,t} = 1$, then unit i is treated in period t. Define $W_{i,t}$:

$$W_{i,t} = \begin{cases} 1 & i = 0 \ \& \ t \geq T_0 \\ 0 & otherwise \end{cases}$$

Let $y_{i,t}(W_{i,t})$ denote the outcome of unit i in period t with treatment status $W_{i,t}$. For example, Great Britain in period t in which brexit had not occurred would be $y_{0,t}(0)$. Great Britain in period t in which brexit had occurred would be $y_{0,t}(1)$.

Assume the researcher observes $Y_{0,t} = (1 - W_{0,t})y_{0,t}(0) + W_{0,t}y_{0,t}(1)$. This means the

researcher observes an untreated Great Britain before 2016 and a treated Great Britain 2016 onwards. The researcher also observes $Y_j = y_{j,t}(0)$ for all $j = 1, ..., J$ and $t = 1, ...T$, implying the control units are never treated.

The goal is to determine the treatment effect of the policy change. The treatment effect for the treated observation at period t is defined as $treat_{0,t} = y_{0,t}(1) - y_{0,t}(0)$. This would be comparing the GDP of Great Britain where brexit passed to the GDP of Great Britain where brexit did not pass in period t. The researchers observe $y_{0,t}(1)$ when $t \geq T_0$. Thus, they must estimate $y_{0,t}(0)$ when $t \geq T_0$. Athey et al. (2018) and Doudchenko and Imbens (2016) describe estimating the counterfactual as a missing data problem. The main body of literature assumes there exists $\beta^* = [\beta_1^*, ...\beta_J^*]$ such that:

$$y_{0,t}(0) = \sum_{j=1}^{J} \beta_j^* y_{j,t}(0) \tag{1}$$

Variants of this assumptions have also assumed matching on observed covariates (Abadie, Diamond, and Hainmueller 2010). This paper restricts the scope of analysis to outcomes on observables without considering covariates. A growing body of literature has supported synthetic control analysis without covariates. Athey and Imbens (2017) and Doudchenko and Imbens (2016) argue the outcomes tend to be far more important than covariates in terms of predictive power. They further argue that minimizing the difference between treated outcomes and control outcomes prior to treatment tend to be sufficient to construct a synthetic control. Kaul et al. (2018) showed covariates become redundant when all lagged outcomes are included in ADH approach. Botosaru and Ferman (2019) showed the counterfactual estimated by using only pre-treatment outcomes were very close to the original ADH. Brodersen et al. (2015) opt to omit covariates. Finally, both Kinn (2018) and Samartsidis et al. (2019) do not use covariates in their model comparisons.

The researcher can then estimate β^* using ordinary least squares (OLS) as:

$$\widehat{\beta}^{OLS} = argmin_b \frac{1}{T_0 - 1} \sum_{t=1}^{T_0 - 1} \left(Y_{0,t} - \sum_{j=1}^{J} b_j Y_{j,t} \right)^2 \tag{2}$$

and use $\widehat{\beta}^{OLS}$ to produce estimates of $y_{0,t}(0)$ for $t \geq T_0$:

$$\widehat{y}_{0,t}(0) = \sum_{j=1}^{J} \widehat{\beta}_j^{OLS} Y_{j,t} \tag{3}$$

Two main issues arise when using the OLS estimator. The first is identification. When the number of covariates (J) is larger than the number of observations ($T_0 - 1$), then the OLS coefficients are not identifiable. This is a common scenario in policy analysis. For example, Abadie, Diamond, and Hainmueller (2010) analyze the effect of changes on California tobacco policies using J=29 control states and $T_0 = 17$ pretreatment periods. The second concern is overfitting. While the best linear unbiased estimator, OLS has been shown to provide poor out of sample predictions. This is because OLS will fit the data too closely misrepresenting the data generating process. The misrepresentation leads to poor out of sample predictions. The most basic example of overfitting is when the number of observations equals the number covariates. In this case, the OLS will simply play "connect the dots". One remedy commonly used is the *penalized regression*. These methods aim to shrink coefficients towards their mean for better out of sample predictions. Examples of these estimators are the Least Absolute Shrinkage and Selection Operator (LASSO) regression, Ridge regression, and the elastic net. Doudchenko and Imbens (2016) employ the elastic net to estimate β^* for their predictions of $\widehat{y}_{0,t}^{EN}$:

$$(\hat{\beta}^{EN}, \hat{\mu}^{EN}) = argmin_{b,m} \frac{1}{T_0 - 1} \sum_{t=1}^{T_0 - 1} \left(Y_{0,t} - m - \sum_{j=1}^{J} b_j Y_{j,t} \right)^2 + \lambda \left(\frac{1-\alpha}{2} \sum_{j=1}^{J} b_j^2 + \alpha \sum_{j=1}^{J} |b_j| \right)$$

where $\hat{\mu}^{EN}$ is the estimated intercept, λ and α are hyperparameters set through a process known as *cross validation*. When $\alpha = 0$, the elastic net becomes a *ridge regression*. When $\alpha = 1$, the elastic net becomes a *LASSO regression*. The counterfactual is then estimated as:

$$\hat{y}_{0,t}^{EN}(0) = \hat{\mu}^{EN} + \sum_{j=1}^{J} \hat{\beta}_j^{EN} Y_{j,t} \tag{4}$$

An alternative approach is the *reduced ADH approach (rADH)*. This approach estimates β^* as:

$$\hat{\beta}^{rADH} = argmin_c \frac{1}{T_0 - 1} \sum_{t=1}^{T_0 - 1} \left(Y_{0,t} - \sum_{j=1}^{J} c_j Y_{j,t} \right)^2 \quad sbj. \sum_{j=1}^{J} c_j = 1$$
$$c_j \in [0, 1] \ \forall j$$

where the counterfactual is estimated as:

$$\hat{y}_{0,t}^{rADH}(0) = \sum_{j=1}^{J} \hat{\beta}_j^{rADH} Y_{j,t} \tag{5}$$

Doudchenko and Imbens (2016) compared this method to the elastic net and Abadie, Diamond, and Hainmueller (2010) full approach. All of these methods have modeled β^* as constant.

Fully parametric methods have been proposed to analyze counterfactuals. Scott and Varian

(2013) present Bayesian Structural Time Series for prediction. This method uses a Bayesian variable selection process, spike and slab, to identify relevant variables for prediction. Like synthetic controls, Scott and Varian (2013) performs variable selection by shrinking the coefficient on irrelevant covariates to 0. This was then extended by Brodersen et al. (2015) as an alternative for synthetic control.

Modeling the coefficients as constant throughout the pre-treatment and post-treatment may be problematic. On an intuitive level, the notion of constant weights means the relationship between the treated unit and controls is constant throughout the analysis period. Thinking back to Great Britain and brexit, each control unit cannot have a significant change throughout the analysis. This means no major governmental shifts, policy reforms, or economic booms/busts. If a researcher were to use only 10 years of pre-treatment, this would mean the pre-treatment analysis of brexit would go straight through the Great Recession. In order for the constant weights to hold (even approximately), each country must be affected relatively similar to the Great Recession, begin recovery around the same time and recover in a similar fashion. If this is not the case, synthetic control may produce poor fits. Abadie, Diamond, and Hainmueller (2010) recommend using other tools in such circumstances. Abadie (n.d.) followed up on this point stating the linear model is meant to be an approximation for a non-linear data generating process. If the process that determines the outcome is non-linear, then even a close fit by a synthetic control can lead to large biases in the post treatment periods. In conclusion, there exists a tension in synthetic controls: the researcher wants a long pre-treatment period that still maintains a linear relationship between the treated unit and covariates throughout the pre and post periods.

This proposal aims to tackle this tension explicitely. Rather than modeling (1) as a constant coefficient model, I assume:

$$y_{0,t}(0) = \sum_{j=1}^{J} \beta^*_{j,t} y_{j,t}(0) \tag{6}$$

There now exists (potentially) different β^* for each time period t. Without loss of generality, an intercept term may be added in by including $y_{j+1,t} = 1$ for all t. The time varying component will model the different effects the Great Recession had on the relationship between the United Kingdom and other countries. The different effects of the Great Recession may be some countries experienced stronger negative shocks than others or countries economic recoveries began at different years with different levels of intensity.

An immediate concern is identification. With time invariant coefficients, there are J coefficients being fit to $T_0 - 1$ equations. Now, there are J coefficients being fit to 1 equation $T_0 - 1$ times. In addition, a structure must be added to the time variation in order to project forward. Without a structure, there would be no way to determine the future values of the coefficients.

One approach to model the time varying component is through Bayesian state space modeling. A downside to this approach is model specification. The reduced form methods described above make minimal assumptions. The proposed method requires defining priors on all covariates, the structure of the time varying parameters, and hyperparameters. There are three main benefits to modeling using Bayesian state space modeling. First, the state space framework allows the researcher to incorporate additional features including seasonality and trend. Second, the Bayesian framework allows for posterior inference. Posterior credibility intervals can be calculated. Finally, the Bayesian framework allows for model uncertainty (Brodersen et al. 2015). The model uncertainty allows the researcher to take an agnostic stance on which variables should and shouldn't be included in the model similarly to the penalized regression. The Bayesian state space model has the additional benefit of automatically determining which variables are time varying and which are constant. The

cost of using a Bayesian state space model are structural assumptions. The benefits are posterior inference and the inclusion of time varying coefficients.

The purpose of this paper is to use Bayesian state space models for counterfactual estimates. This model incorporates Bayesian shrinkage to automatically classify coefficients as (i) time varying, (ii) constant, or (iii) irrelevant. This method allows for time varying coefficients to be included in the model without the risk of overfitting.

2 Model[1]

The Bayesian state space model proposed is inspired by Belmonte, Koop, and Korobilis (2014). First, estimate equation (6) as:

$$Y_{0,t} = \sum_{j=1}^{J} \beta_{j,t} Y_{j,t} + \epsilon_t \quad \epsilon_t | \sigma^2 \sim N(0, \sigma^2)$$

In the state space literature, this equation is known as the *observation equation*. As in the nonparametric examples, the researcher will use observations before T_0 to calculate $\beta_{j,t}$ and then predict $\widehat{y}_{0,t}(0)$ for $t \geq T_0$. However, structure on the evolution of $\beta_{j,t}$ must now be employed. The equations that govern the evolution of $\beta_{j,t}$ are referred to as *transition equations*. Dangl and Halling (2012) compared different transition equations for out of sample predictions in stock prices. The best out of sample predictor was the random walk. Belmonte, Koop, and Korobilis (2014) and Bitto and Frühwirth-Schnatter (2019) also use a random walk model for the transition equations. In addition to the positive results seen in stock predictions, the random walk keeps the model relatively parsimonious. I model the transition equations as:

[1] A brief introduction to Linear Gaussian State Space Models can be found in the appendix.

$$\beta_{j,t} = \beta_{j,t-1} + \eta_{j,t} \quad \eta_{j,t} \sim N(0, \theta_j) \quad \forall j$$
$$\beta_{j,0} \sim N(\beta_j, \theta_j P_{jj}) \quad \forall j$$

Therefore, the full Bayesian state space model is defined as:

$$Y_{0,t} = \sum_{j=1}^{J} \beta_{j,t} Y_{j,t} + \epsilon_t \quad \epsilon_t | \sigma^2 \sim N(0, \sigma^2)$$
$$\beta_{j,t} = \beta_{j,t-1} + \eta_{j,t} \qquad \eta_{j,t} \sim N(0, \theta_j) \quad \forall j$$
$$\beta_{j,0} \sim N(\beta_j, \theta_j P_{jj}) \quad \forall j$$

where P_{jj} is a hyperparameter. This specification of θ_j lends itself to a useful interpretation: θ_j governs the dynamics of $\beta_{j,t}$ (Bitto and Frühwirth-Schnatter 2019). In the special case where $\theta_j = 0$ for all j, the model collapses back to a constant coefficient model.

The errors are assumed to be independent of one another and independent of all leads and lags. The errors between coefficients are assumed to be independent (e.g. $cov(\eta_{j,t}, \eta_{i,t}) = 0$ for $i \neq j$). This assumption is to keep the model relatively parsimonious (Belmonte, Koop, and Korobilis (2014), Bitto and Frühwirth-Schnatter (2019)).

The transition equation can be rewritten to decompose β_j into a time varying and constant components[2] (Frühwirth-Schnatter and Wagner 2010).

$$\beta_{j,t} = \beta_j + \tilde{\beta}_{j,t} \sqrt{\theta_j}$$
$$\tilde{\beta}_{j,t} = \tilde{\beta}_{j,t-1} + \tilde{\eta}_{j,t} \quad \tilde{\eta}_{j,t} \sim N(0, 1)$$
$$\tilde{\beta}_{j,0} \sim N(0, P_{jj})$$

β_j can now be interpreted as the time invariant portion of $\beta_{j,t}$ and $\sqrt{\theta_j}\tilde{\beta}_{j,t}$ the time varying portion. $\sqrt{\theta_j}$ is the root of the variance of the time varying coefficient. The absolute value of $\sqrt{\theta_j}$ can be interperted as the standard deviation of time varying coefficient. Thinking back

[2]See the appendix for further explanation.

to Great Britain and brexit, β_j would be the average relationship between country j and Great Britain and $\sqrt{\theta_j}\tilde{\beta}_{j,t}$ would be the period specific effect (i.e. slow recession recovery). The advantage of this formulation is that shrinkage estimation can be performed on both aspects of the coefficient. Substituting the reformulation back into the original equation yields the state space model:

$$Y_{0,t} = \sum_{j=1}^{J} \left(\beta_j + \tilde{\beta}_{j,t}\sqrt{\theta_j}\right) Y_{j,t} + \epsilon_t \qquad \epsilon_t|\sigma^2 \sim N(0,\sigma^2) \tag{7}$$

$$\tilde{\beta}_{j,t} = \tilde{\beta}_{j,t-1} + \tilde{\eta}_{j,t} \qquad \tilde{\eta}_{j,t} \sim N(0,1) \tag{8}$$

$$\tilde{\beta}_{j,0} \sim N(0, P_{jj}) \tag{9}$$

Equations (7), (8), and (9) constitute the Bayesian state space model. Frühwirth-Schnatter and Wagner (2010) refer to this setup as the *non-centered parameterization of state space models*. This is extremely useful because the problem of *variance selection* has now been recast as one of *variable selection*. Variable selection problems are far better understood and applied. With this formulation there are 2J+1 parameters to be estimated: the J time invariant coefficients (i.e. β_j), the J time varying coefficients (i.e. $\sqrt{\theta_j}$) and the variance (σ^2).

3 The Priors

The choice of priors is left to the researcher in a Bayesian framework. This paper opts to follow a variant of Park and Casella (2008). Park and Casella (2008) show their choice of priors creates the Bayesian Lasso, which is analogous to the nonparametric Lasso. The Bayesian Lasso is especially appealing because of it's relationship to the nonparametric methods discussed previously. Doudchenko and Imbens (2016) use a variant of the Lasso

in their estimation. In addition, Kinn (2018) argued that the synthetic control proposed by Abadie, Diamond, and Hainmueller (2010) is a restricted version of the nonparametric Lasso. Therefore, using the Bayesian Lasso is staying within the "Lasso family" of shrinkage estimators.

Park and Casella (2008) define the Bayesian Lasso coefficients (i.e. the β_j's and $\sqrt{\theta_j}$'s) to have an independent identical laplace prior distributions. This choice results in the mode of the posterior distributions corresponding to the nonparametric LASSO estimators (Tibshirani 1996). The laplace prior can be represented heirarchically using a normal and exponential distribution. λ^2 is represented as a half-cauchy distribution with mean 0 and scale parameter 1. Polson and Scott (2011) provide arguments for the benefits of using the half-cauchy for the global shrinkage estimator. The main motivators are the flexibility and better behavior near 0 compared to alternatives (i.e. inverse gamma distribution).

Like the laplace distribution, the half-cauchy has a heirachrical representation. The heirarchy between λ^2 and ζ_β form the half-cauchy distribution. Therefore, the prior distribution for $\beta = [\beta_1, \beta_2, ..., \beta_j]$ with variances $\tau^2 = [\tau_1^2, \tau_2^2, ..., \tau_J^2]$ are:

$$\beta|\tau^2, \lambda^2 \sim \mathcal{N}_J(0_J, \lambda^2 diag[\tau_1^2, ...\tau_J^2]) \tag{10}$$

$$\tau_j^2 \sim exp\,(1) \tag{11}$$

$$\lambda^2|\zeta_\beta \sim InverseGamma\left(\frac{1}{2}, \frac{1}{\zeta_\beta}\right) \tag{12}$$

$$\zeta_\beta \sim InverseGamma\left(\frac{1}{2}, 1\right) \tag{13}$$

This formulation of the Bayesian Lasso create a *global-local shrinkage estimator*. λ^2 controls the overall model size while τ_j^2 controls the variable specific size.

Traditionally, variances have been defined by the inverse gamma distribution. However, the inverse gamma does not allow for effective shrinkage given it's support. Frühwirth-Schnatter

and Wagner (2010) provide an in depth argument for the use of the normal distribution as an alternative. Briefly, inverse gammas perform poorly in terms of shrinkage and the normal distribution does not. Similarly to β, assign the prior of $\sqrt{\theta} = \left[\sqrt{\theta_1}, \sqrt{\theta_2}, ..., \sqrt{\theta_J}\right]$ with variances $\xi^2 = [\xi_1^2, \xi_2^2, ...\xi_J^2]$ as:

$$\sqrt{\theta}|\xi^2, \kappa^2 \sim \mathcal{N}_J(0_J, \kappa^2 diag[\xi_1^2, ...\xi_J^2]) \tag{14}$$

$$\xi_j^2 \sim exp\,(1) \tag{15}$$

$$\kappa^2|\zeta_{\sqrt{\theta}} \sim InverseGamma\left(\frac{1}{2}, \frac{1}{\zeta_{\sqrt{\theta}}}\right) \tag{16}$$

$$\zeta_{\sqrt{\theta}} \sim InverseGamma\left(\frac{1}{2}, 1\right) \tag{17}$$

σ^2 is defined as $\frac{1}{\sigma^2} \sim Gamma(a_1, a_2)$ with *shape* hyperparameter a_1 and *scale* hyperparameter a_2. Notice that if $\sqrt{\theta_j} = 0$ for all j, the model collapses to a time invariant estimation with the Bayesian Lasso performing shrinkage. This would be the Bayesian version of the LASSO regression discussed earlier.

4 The Posterior Estimation (MCMC)

In order to draw predictions for the counterfactual, the posterior distribution must be calculated: $P(\tilde{\beta}, \beta, \tau^2, \lambda^2, \sqrt{\theta}, \xi^2, \kappa^2, \zeta_\beta, \zeta_{\sqrt{\theta}}, \sigma^2|Y_0)$ where $Y_0 = \{Y_{0,t}\}_{t=1}^{T_0-1}$. With values drawn from the posterior distribution, $\hat{y}_{0,t}(0)$ can then be estimated for $t \geq T_0$. A closed form does not exist for the posterior. Therefore, a Gibbs sampler must be used. The Gibbs sampler is a work-around in which the joint posterior is simulated by iteratively sampling through conditional posteriors. After a sufficiently large initial sample, or *burn in*, the draws from the conditional posterior will be simulations of the joint posterior.

Park and Casella (2008) suggest one formulation of the Gibbs sampler for the Bayesian Lasso

which Belmonte, Koop, and Korobilis (2014) then extended to a non-centered parameterization of state space models. I will build off of these two papers adding a synthetic control prediction step. A major benefit of the Bayesian Lasso formulation is that, conditional on τ^2 and ξ^2, the model follows a standard linear regression with normal priors.

The Posterior estimation can be broken into three main steps:

(i) Estimation of $\tilde{\beta}|\beta, \tau^2, \lambda^2, \sqrt{\theta}, \xi^2, \kappa^2, \zeta_\beta, \zeta_{\sqrt{\theta}}, \sigma^2, Y_0$.

(ii) Estimation of the parameters: $P(\beta, \tau^2, \lambda^2, \sqrt{\theta}, \xi^2, \kappa^2, \zeta_\beta, \zeta_{\sqrt{\theta}}, \sigma^2|Y_0)$.

(iii) Estimation of $\hat{y}_{0,t}(0)$ for $t \geq T_0$.

4.1 Estimation of $\tilde{\beta}|\beta, \tau^2, \lambda, \sqrt{\theta}, \xi^2, \kappa^2, \zeta_\beta, \zeta_{\sqrt{\theta}}, \sigma^2, Y_0$

Draw $\tilde{\beta}$ using Durbin and Koopman (2012) for the state space model. First, rewrite equations (7), (8), and (9) as:

$$Y'_{0,t} = \sum_{j=1}^{J} \tilde{\beta}_{j,t} Z_{j,t} + \epsilon_t \qquad \epsilon_t|\sigma^2 \sim N(0, \sigma^2) \tag{18}$$

$$\tilde{\beta}_{j,t} = \tilde{\beta}_{j,t-1} + \tilde{\eta}_{j,t} \qquad \tilde{\eta}_{j,t} \sim N(0, 1) \tag{19}$$

$$\tilde{\beta}_{j,0} \sim N(0, P_{jj}) \tag{20}$$

where $Z_{j,t} = \sqrt{\theta_j} Y_{j,t}$, $Y'_{0,t} = Y_{0,t} - \sum_{j=1}^{J} \beta_j Y_{j,t}$.

Many algorithms have been proposed to simulate latent variables in a state space framework. I use the method proposed by Durbin and Koopman (2012). I first run the Kalman filter and smoother given the data and parameters to produce $\hat{\tilde{\beta}}_t$. I then simulate new ${Y'}^{+}_{0,t}$ and $\tilde{\beta}^{+}_{j,t}$ for all j using equations (7), (8), and (9). I then run the Kalman filter and smoother on ${Y'}^{+}_{0,t}$ and $\tilde{\beta}^{+}_{j,t}$ for all j producing $\hat{\tilde{\beta}}^{+}_{t}$. My new simulated draw of $\tilde{\beta}_t$ is $\tilde{\tilde{\beta}}_t = \hat{\tilde{\beta}}_t - \tilde{\beta}^{+}_t + \hat{\tilde{\beta}}^{+}_t$.

4.2 Estimation of the parameters: $P(\beta, \tau^2, \lambda, \sqrt{\theta}, \xi^2, \kappa^2, \zeta_\beta, \zeta_{\sqrt{\theta}}, \sigma^2 | Y_0)$

Attempting to sample $P(\beta, \tau^2, \lambda^2, \sqrt{\theta}, \xi^2, \kappa^2, \zeta_\beta, \zeta_{\sqrt{\theta}}, \sigma^2 | Y_0)$ would lead to the same problem as before: no analytic posterior exists. Rather than sampling all parameters at once, I will sample the parameters as blocks. The sampling distributions are derived in the appendix.

4.2.1 Sample β

Block draw β from the normal conditional posterior:

$$\mathcal{N}_J \left((\tilde{Y^T}\tilde{Y} + \frac{\sigma^2}{\lambda^2} V_\beta^{-1})^{-1} \tilde{Y^T} \tilde{Y_0}, \sigma^2 (\tilde{Y^T}\tilde{Y} + \frac{\sigma^2}{\lambda^2} V_\beta^{-1})^{-1} \right) \tag{21}$$

Where:

$$\tilde{Y} = \begin{pmatrix} Y_{1,1} & Y_{2,1} & ... & Y_{J,1} \\ \vdots & \vdots & \vdots & \vdots \\ Y_{1,T_0-1} & Y_{2,T_0-1} & ... & Y_{J,T_0-1} \end{pmatrix}$$

$$V_\beta = diag[\tau_1^2, \tau_2^2, ..., \tau_J^2]$$

$$\tilde{Y_0} = \begin{pmatrix} Y_{0,1} - \sum_{j=1}^{J} \sqrt{\theta_j} \tilde{\beta}_{j,1} Y_{j,1} \\ Y_{0,2} - \sum_{j=1}^{J} \sqrt{\theta_j} \tilde{\beta}_{j,2} Y_{j,2} \\ \vdots \\ Y_{0,T_0-1} - \sum_{j=1}^{J} \sqrt{\theta_j} \tilde{\beta}_{j,T_0-1} Y_{j,T_0-1} \end{pmatrix}$$

Sampling from sparse matrices can lead preset matrix inversion techniques to fail. To avoid such failures, I implement the algorithm proposed by Bhattacharya, Chakraborty, and Mallick (2016).

4.2.2 Sample τ^2

Draw τ^2 using the fact $\frac{1}{\tau_j^2}$ each have independent inverse-Gaussian (IG) conditional priors:

$$IG\left(\sqrt{\frac{2\lambda^2}{\beta_j^2}}, 2\right) \text{ for j=1,...J} \tag{22}$$

4.2.3 Sample λ^2

Draw λ^2 from the conditional inverse gamma prior:

$$InverseGamma\left(shape = \frac{J+1}{2}, rate = \frac{1}{\zeta_\beta} + \frac{1}{2}\sum_{j=1}^{J}\frac{\beta_j^2}{\tau_j^2}\right) \tag{23}$$

4.2.4 Sample ζ_β

Draw ζ_β from the conditional inverse gamma prior:

$$InverseGamma\left(1, 1 + \frac{1}{\lambda^2}\right) \tag{24}$$

4.2.5 Sample $\sqrt{\theta}$

Block draw $\sqrt{\theta}$ from the normal conditional posterior:

$$\mathcal{N}_J\left((\bar{Y^T}\bar{Y} + \frac{\sigma^2}{\kappa^2}V_{\sqrt{\theta}}^{-1})^{-1}\bar{Y^T}\bar{Y}_0, \sigma^2(\bar{Y^T}\bar{Y} + \frac{\sigma^2}{\kappa^2}V_{\sqrt{\theta}}^{-1})^{-1}\right) \tag{25}$$

Where:

$$\bar{Y} = \begin{pmatrix} \tilde{\beta}_{1,1}Y_{1,1} & \tilde{\beta}_{2,1}Y_{2,1} & ... & \tilde{\beta}_{J,1}Y_{J,1} \\ \vdots & \vdots & \vdots & \vdots \\ \tilde{\beta}_{1,T_0-1}Y_{1,T_0-1} & \tilde{\beta}_{2,T_0-1}Y_{2,T_0-1} & ... & \tilde{\beta}_{J,T_0-1}Y_{J,T_0-1} \end{pmatrix}$$

$$V_{\sqrt{\theta}} = diag[\xi_1^2, \xi_2^2, ..., \xi_J^2]$$

$$\bar{Y_0} = \begin{pmatrix} Y_{0,1} - \sum_{j=1}^{J} \beta_{j,1}Y_{j,1} \\ Y_{0,2} - \sum_{j=1}^{J} \beta_{j,2}Y_{j,2} \\ \vdots \\ Y_{0,T_0-1} - \sum_{j=1}^{J} \beta_{j,T_0-1}Y_{j,T_0-1} \end{pmatrix}$$

I use Bhattacharya, Chakraborty, and Mallick (2016) algorithm here as well.

4.2.6 Sample ξ^2

Draw ξ^2 using the fact $\frac{1}{\xi_j^2}$ each have independent inverse-Gaussian (IG) conditional priors:

$$IG\left(\sqrt{\frac{2\kappa^2}{\theta_j}}, 2\right) \text{ for j=1,...J} \tag{26}$$

4.2.7 Sample κ^2

Draw κ^2 from the conditional gamma prior:

$$InverseGamma\left(shape = \frac{J+1}{2}, rate = \frac{1}{\zeta_{\sqrt{\theta}}} + \frac{1}{2}\sum_{j=1}^{J}\frac{\sqrt{\theta_j}^2}{\xi_j^2}\right) \tag{27}$$

4.2.8 Sample $\zeta_{\sqrt{\theta}}$

Draw $\zeta_{\sqrt{\theta}}$ from the conditional inverse gamma prior:

$$InverseGamma\left(1, 1 + \frac{1}{\kappa^2}\right) \tag{28}$$

4.2.9 Sample σ^2

Draw σ^2 from the posterior distribution:

$$InverseGamma\left(a_1 + \frac{T_0 - 1}{2}, a_2 + \frac{\left(Y_{0,t} - \sum_{j=1}^{J}\left(\beta_j + \tilde{\beta}_{j,t}\sqrt{\theta_j}\right)Y_{j,t}\right)^2}{2}\right) \tag{29}$$

Frühwirth-Schnatter and Wagner (2010) note an identification problem arises when using the non-centered paramaterization. There is no way to distinguish between $\sqrt{\theta_j}\tilde{\beta}_{j,t}$ and $(-\sqrt{\theta_j})(-\tilde{\beta}_{j,t})$. This problem is referred to as *label switching problem.* This issue is a common occurrence in Bayesian estimation when a distribution is multi-modal, as is the case with the square root of a variance. To solve this identification problem, Frühwirth-Schnatter and Wagner (2010) suggest a random sign change at the end of each iteration of the Gibbs Sampler. With 50% chance, the signs on $\tilde{\beta}$ and $\sqrt{\theta}$ are switched. Both Belmonte, Koop, and Korobilis (2014) and Bitto and Frühwirth-Schnatter (2019) employ this method.

A final note of interest is the formulation of λ^2 (and κ^2). Notice that the conditional distribution of λ^2 relies on $\sum_{j=1}^{J}\tau_j^2$ where each posterior τ_j^2 relies on β_j. This direct reliance on β_j in the conditional distributions can lead to scaling issues. Data that is bigger in magnitude can dominate the distribution of λ^2. The issue of scaling is common in nonparametric

shrinkage estimators.[3] To account for this, all covariates can be scaled to mean zero variance one prior to analysis. The coefficients can then be rescaled back into the proper units after the MCMC is complete for ease of interpertation.

4.3 Sample of $\widehat{y}_{0,t}(0)$ for $t \geq T_0$.

After a sufficiently large *burn in* period, use the proceeding draws to calculate $\widehat{y}_{0,t}(0)$ for $t \geq T_0$. Namely, perform the following steps:

(1) Simulate $\tilde{\beta}_{j,t} = \tilde{\beta}_{j,t-1} + \tilde{\eta}_{j,t}$ for all j for $t \geq T_0$. Use $\tilde{\tilde{\beta}}_{j,T_0-1}$ simulated in section 4.1 as an initial value. Notice that each iteration of the Gibbs sampler will mean a new $\tilde{\tilde{\beta}}_{j,T_0-1}$.

(2) Using the simulated $\tilde{\beta}_{j,t}$, predict $\widehat{y}_{0,t}(0)$ as:

$$\widehat{y}_{0,t}(0) = \mu + \sum_{j=1}^{J} \left(\beta_j + \tilde{\beta}_{j,t} \sqrt{\theta_j} \right) Y_{j,t} + \epsilon_t$$

drawing $\epsilon_t | \sigma^2 \sim N(0, \sigma^2)$. Section 4.2.1 provides the draws for β_j for all j. Section 4.2. provides the draws for $\sqrt{\theta_j}$ for all j. Section 4.2.7 provides the draw for σ^2 used for determining ϵ_t. Each iteration of the Gibbs sampler will produce new parameter and state values.

In conclusion, the Gibbs Sampling procedure is as follows:

5 Monte Carlo Simulation Data

The simulation is restricted to the outcomes of the observed units, without considering covariates. A growing body of literature has supported synthetic control analysis without

[3]For example, think back to the original LASSO: $\beta = argmin_b \sum_i (y_i - X_i b)^2 + \lambda \sum_i |b_i|$. If one covaraite is scaled 100 times larger than the others, then it will dominate $\lambda \sum_i |b_i|$. Rather than shrinking based on the relationship between the covariate and outcome, the shrinkage will be based on a combination of the relationship and magnitude of the covariate.

covariates. Athey and Imbens (2017) and Doudchenko and Imbens (2016) argue the outcomes tend to be far more important than covariates in terms of predictive power. They further argue that minimizing the difference between treated outcomes and control outcomes prior to treatment tend to be sufficient to construct a synthetic control. Kaul et al. (2018) showed covariates become redundant when all lagged outcomes are included in ADH approach. Botosaru and Ferman (2019) showed the counterfactual estimated by using only pre-treatment outcomes were very close to the original ADH. Brodersen et al. (2015) opt to omit covariates. Finally, both Kinn (2018) and Samartsidis et al. (2019) do not use covariates in their model comparisons.

For the purpose of this paper, the argument that covariates follow the same time varying weight structure as the outcome would be hard to rationalize theoretically or empirically. Because of this, the simulation opts to avoid covariates entirely.

The Monte Carlo simulation is based off of Kinn (2018) setup. Assume the following data generating process:

$$y_{j,t}(0) = \xi_{j,t} + \psi_{j,t} + \epsilon'_{j,t} \qquad \text{j=1,..,J}$$
$$y_{0,t}(0) = \sum_{j=1}^{J} w_{j,t}(\xi_{j,t} + \psi_{j,t}) + \epsilon'_{1.t}$$

for t=1,..,T where ξ_{jt} is the trend component, ψ_{jt} is the seasonality component, and $\epsilon'_{jt} \sim N(0, \sigma^2)$. Specifically, $\xi_{jt} = c_j t + z_j$ where $c_j,\ z_j \in \mathbb{R}$. This will allow for each observation to have a unit-specific time varying confounding factor and a time-invariant confounding factor. Seasonality will be represented as $\psi_{j,t} = \gamma_j sin\left(\frac{\pi t}{\rho_j}\right)$. Parallel trends are created when $c_j = c\ \forall\ j$ and $\gamma_j = 0\ \forall\ j$. The explicit data generating process is:

$$y_{j,t}(0) = c_j t + z_j + \gamma_j sin\left(\frac{\pi t}{\rho_j}\right) + \epsilon'_{j,t} \qquad \text{j=2,..,J}$$
$$y_{1,t}(0) = \sum_{j=2}^{J} w_{j,t}\left(c_j t + z_j + \gamma_j sin\left(\frac{\pi t}{\rho_j}\right)\right) + \epsilon'_{1.t}$$

Following Kinn (2018), a sparse set of controls will have nonzero weights. This means properly identifying the correct controls will be important for an accurate counterfactual. The treatment begins at period T_0.

This paper proposes one scenario to test continuous time varying weights.

5.1 Deterministic Continuous Varying Weights

To simulate continuous varying weights, $c_{2,t}$ and $c_{3,t}$ are defined .75 and .25 respectively. All other $c_{j,t}$ are randomly drawn from U[0,1]. In order to avoid $y_{2,t}$ and $y_{3,t}$ from crossing, set $z_2 = 25$ and $z_3 = 5$. In addition, set $\psi_{j,t} = 0$ for all j,t. Finally, define $w_{1,t} = .2 + .6\frac{t}{T}$ and $w_{2,t} = 1 - w_{1,t}$. In order to compare to ADH, the sum of the weights was sets to unity. This ensure the convex hull assumption is met.

To summarize, the parameters of this simulation are:

1) $c_{1,t} = .75$, $c_{2,t} = .25$, and $c_{j,t} \sim U[0,1]$ for all $j \notin \{1,2\}$

2) $z_1 = 25$, $z_2 = 5$ and z_j is sampled from $\{1,2,3,4,...,50\}$.

3) $\epsilon'_{j,t} \sim N(0,1)$.

4) T = 34, $T_0 = 17$.

5) J = 17.

6) $w_{1,t} = .2 + .6\frac{t}{T}$, $w_{2,t} = 1 - w_{2,t}$, and $w_{j,t} = 0$ for all else

7) $\gamma_j = 0 \ \forall j$.

The data generating process can be rewritten in recursive form:

$$
\begin{aligned}
y_{0,t}(0) &= \sum_{j=1}^{J} w_{j,t}\left(c_j t + z_j + \gamma_j sin\left(\frac{\pi t}{\rho_j}\right)\right) + \epsilon'_{1.t} \\
w_{1,t} &= w_{1,t-1} + \frac{.6}{T} \\
w_{2,t} &= w_{2,t-1} - \frac{.6}{T} \\
w_{j,t} &= w_{j,t-1} \qquad\qquad j \notin \{1,2\}
\end{aligned}
$$

with initial conditions:

$$
\begin{aligned}
w_{1,0} &= .2 \\
w_{2,0} &= .8 \\
w_{j,0} &= 0 \quad j \notin \{1,2\}
\end{aligned}
$$

5.2 Model Testing and Comparison

This simulation will test the accuracy of the estimates of the treatment effect and the accuracy of the inference (significant or not). The treatment effect sizes will be tested at 0%, 0.1%, 1%, 10%, and 100% similarly to Brodersen et al. (2015). These treatment effects will be calculated by defining $Y_{0,t}(1) = \rho Y_{0,t}(0)$ for $\rho \in \{1, 1.001, 1.01, 1.1, 2\}$. For inference, a causal effect is assumed only if 95% of the posteriod probability interval excludes 0.

In order to compare the TVPS state space model to ADH synthetic control, the median observation of the posterior distribution at each post treatment period will be used. This test will compare the recovered treatment effect size versus the actual. Zero percent, 0.1%, 1%, 10%, and 100% will be used for treatment effect sizes with $Y_{0,t}(1)$ defined as before.

6 Conclusion

This proposal adds shrinkage among time varying weights to counterfactual analysis. The addition of shrinkage among time varying weights will extend the scope of synthetic control to data previously restricted from analysis as well as add to the very limited existing literature of state space models in counterfactual analysis. Future research will include extending the model to multiple outcome variables (e.g. GDP and unemployment).

7 Appendix

7.1 Linear Gaussian State Space Models

This section presents an introduction to concepts in Linear Gaussian State Space Models following Durbin and Koopman (2012). All notation used in this section of the appendix is only meant for this section of the appendix.

Identifying time varying coefficients can be thought of as a latent variable estimation problem. State space modeling is a time series concept that allows for modeling latent variables explicitly. This means modeling unobserved components like time trends, seasonality, and time varying coefficients. A state space model is composed of an observation equation and transition equation. A general form of these equations follows:

$$\begin{aligned} y_t &= Z_t\alpha_t + \epsilon_t \qquad \text{observation equation} \\ \alpha_{t+1} &= T_t\alpha_t + R_t\eta_t \qquad \text{transition equation} \\ \alpha_0 &\sim \mathcal{N}(a_0, P_0) \end{aligned}$$

where $\epsilon_t \sim \mathcal{N}(0, \sigma_t^2)$ and $\eta_t \sim \mathcal{N}(0, Q_t)$ are independent of all unknown factors. y_t is the observed data and α_t is a combination of observed data (e.g. control variables) and unobserved components (e.g. trend and cycle). In the case of a scalar output, y_t, with m

variables and r time varying components, Z_t would be a 1 x m dimensional matrix, α_t a m x 1 matrix, and ϵ_t a scalar. α_{t+1} would also be a m x 1 matrix, T_t an m x m matrix, R_t a m x r matrix and Q_t an r x r matrix. Finally, a_0 is m x 1 and P_0 is m x m. Linear Gaussian state space models are structural models. The assumptions necessary for linear Gaussian state space models are:

1) $\epsilon_t \sim \mathcal{N}(0, \sigma_t^2)$ and $\eta_t \sim \mathcal{N}(0, Q_t)$. These errors are also assumed to be serially uncorrelated. This is because they are meant to be random disturbances within the model.

2) The errors must be normal.

3) the transition equations can be of lag order 1. Any additional lag orders can be rewritten as order 1 using the state space framework.

7.2 Transition Equation Derivation

To verify these representations of $\beta_{j,t}$ are equal, note:

$$\begin{aligned}
\beta_{j,t} - \beta_{j,t-1} &= (\beta_j + \sqrt{\theta}_j\tilde{\beta}_{j,t}) - (\beta_j + \sqrt{\theta_j}\tilde{\beta}_{j,t-1}) && \text{Plugging in} \\
&= \sqrt{\theta_j}(\tilde{\beta}_{j,t} - \tilde{\beta}_{j,t-1}) && \text{Regroup} \\
&= \sqrt{\theta_j}(\tilde{\beta}_{j,t-1} + \tilde{\eta}_{j,t} - \tilde{\beta}_{j,t-1}) && \text{Plug in} \\
&= \sqrt{\theta_j}\tilde{\eta}_{j,t} && \text{Simplify}
\end{aligned}$$

Notice that $\tilde{\eta}_{j,t} \sim N(0,1)$. Therefore $\sqrt{\theta_j}\tilde{\eta}_{j,t} \sim N(0, \theta_j)$ which is $\eta_{j,t}$.

7.3 Deriving Distributions for the Gibbs Sampler

The derivations are based off of Park and Casella (2008). Notable changes have been made for this specific application. Namely, the model is larger, β and $\sqrt{\theta}$ are not conditioned on σ^2, and the heirarchical structure is redifined to be a *global-local* shrinkage estimator. Park

and Casella (2008) use a heirarchical formulation where the local shrinkage is dependent on the global shrinkage. Park and Casella (2008) also use an inverse gamma distribution to represent the global shrinkage while this paper opts to use a half cauchy distribution.

For clarity, I will refer to the outcome variable, $Y_{0,t}$ as Y and the matrix of covariates as X. This is done simply for clarity in the derivations of the conditional probabilities. The slight change of notation only pertains to this section of the appendix.

Recall:

$$Y_t = \sum_{j=1}^{J} \left(\beta_j + \tilde{\beta}_{j,t}\sqrt{\theta_j}\right) X_{j,t} + \epsilon_t$$

The prior of Y is defined as $\mathcal{N}\left(X\beta_j + (X * \tilde{\beta}_j)\sqrt{\theta_j}), \sigma^2 I\right)$ where * denotes element wise multiplication. Conditional on τ_i^2 and ξ_i^2, the model follows a standard linear regression with normal priors. Textbook tools can be used to derive the distributions for the Gibbs sampler. The joint density is defined as:

$$f(Y|\beta, \sqrt{\theta}, \sigma^2)\pi(\sigma^2)\pi(\lambda^2)\pi(\kappa^2)\prod_{j=1}^{J}\pi(\beta_j|\tau_j^2)\pi(\tau_j^2)\pi(\sqrt{\theta_j}|\xi_j^2)\pi(\xi^2) =$$

$$\frac{1}{(2\pi\sigma^2)^{\frac{T_0-1}{2}}}\ exp\left\{\frac{-1}{2\sigma^2}\left(Y - X\beta - (X * \tilde{\beta})\sqrt{\theta}\right)^T\left(Y - X\beta - (X * \tilde{\beta})\sqrt{\theta}\right)\right\}$$

$$\frac{a_2^{a_1}}{\Gamma(a_1)}(\sigma^2)^{-a_1-1}exp\left\{-\frac{a_2}{\sigma^2}\right\}\frac{\frac{1}{\zeta_\beta}^{\frac{1}{2}}}{\Gamma\left(\frac{1}{2}\right)}(\lambda^2)^{-\frac{1}{2}-1}exp\left\{-\frac{\frac{1}{\zeta_\beta}}{\lambda^2}\right\}\frac{\frac{1}{\zeta_{\sqrt{\theta}}}^{1/2}}{\Gamma\left(\frac{1}{2}\right)}(\kappa^2)^{\frac{-1}{2}-1}exp\left\{-\frac{\frac{1}{\zeta_{\sqrt{\theta}}}}{\kappa^2}\right\}$$

$$\frac{1^{1/2}}{\Gamma(1/2)}\zeta_\beta^{-\frac{1}{2}-1}exp\left\{\frac{-1}{\zeta_\beta}\right\}\frac{1^{1/2}}{\Gamma(1/2)}\zeta_{\sqrt{\theta}}^{-\frac{1}{2}-1}exp\left\{\frac{-1}{\zeta_{\sqrt{\theta}}}\right\}$$

$$\prod_{j=1}^{J}\frac{1}{(2\pi\tau_j^2\lambda^2)^{\frac{1}{2}}}exp\left\{\frac{-1}{(2\tau_j^2\lambda^2)}\beta_j^2\right\}exp\left\{-\tau_j^2\right\}\frac{1}{(2\pi\xi_j^2\kappa^2)^{\frac{1}{2}}}exp\left\{\frac{-1}{(2\xi_j^2\kappa^2)}\sqrt{\theta}_j^2\right\}exp\left\{\xi_j^2\right\}$$

7.3.1 Conditional Distribution of β and $\sqrt{\theta}$

To solve for the conditional distribution of β, drop the terms that don't involve β. This only leaves 2 exponential terms:

$$exp\left\{\frac{-1}{2\sigma^2}\left(Y - X\beta - (X * \tilde{\beta})\sqrt{\theta}\right)^T\left(Y - X\beta - (X * \tilde{\beta})\sqrt{\theta}\right)\right\}\prod_{j=1}^{J} exp\left\{\frac{-1}{(2\tau_j^2\lambda^2)}\beta_j^2\right\}$$

Combining exponents yields:

$$exp\left\{\frac{-1}{2\sigma^2}\left(Y - X\beta - (X * \tilde{\beta})\sqrt{\theta}\right)^T\left(Y - X\beta - (X * \tilde{\beta})\sqrt{\theta}\right) + \sum_{j=1}^{J}\frac{-\sigma^2}{(2\tau_j^2\lambda^2)}\beta_j^2\right\}$$

Focusing solely on the exponential term and rearranging yields:

$$\frac{-1}{2\sigma^2}\left[\left(Y - X\beta - (X * \tilde{\beta})\sqrt{\theta}\right)^T\left(Y - X\beta - (X * \tilde{\beta})\sqrt{\theta}\right) + \beta^T\frac{\sigma^2}{\lambda^2}D^{-1}\beta\right]$$

where $D = diag(\tau_1^2, ...\tau_J^2)$. Multiplying out and rearranging yields:

$$\frac{-1}{2\sigma^2}\left[\left(Y - X * \tilde{\beta}\sqrt{\theta}\right)^T\left(Y - X * \tilde{\beta}\sqrt{\theta}\right) - 2\left(Y - X * \tilde{\beta}\sqrt{\theta}\right)X\beta + \beta^T(X^TX + \frac{\sigma^2}{\lambda^2}D^{-1})\beta\right]$$

Focus solely on the terms within the brackets for a moment. Setting $A = X^TX + \frac{\sigma^2}{\lambda^2}D^{-1}$ and completing the square yields:

$$\begin{aligned}&\left(\beta - A^{-1}X^T\left(Y - X * \tilde{\beta}\sqrt{\theta}\right)\right)^T A\left(\beta - A^{-1}X^T\left(Y - X * \tilde{\beta}\sqrt{\theta}\right)\right)\\ &+ \left(Y - X * \tilde{\beta}\sqrt{\theta}\right)^T(I - XA^{-1}X^T)\left(Y - X * \tilde{\beta}\sqrt{\theta}\right)\end{aligned}$$

Therefore, the part of the conditional distribution that relies on β can be written as:

$$\frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{\frac{-1}{2\sigma^2}\left(\beta - A^{-1}X^T\left(Y - X*\tilde{\beta}\sqrt{\theta}\right)\right)^T A\left(\beta - A^{-1}X^T\left(Y - X*\tilde{\beta}\sqrt{\theta}\right)\right)\right\}$$

which can be summarized as β conditionally distributed as:

$$\mathcal{N}\left(A^{-1}X^T\left(Y - X*\tilde{\beta}\sqrt{\theta}\right), \sigma^2 A^{-1}\right)$$

Notice that $\sqrt{\theta}$ appears identically in the joint likelihood function as β. Following the same steps would yield the conditional distribution of $\sqrt{\theta}$:

$$\mathcal{N}\left(B^{-1}X^T\left(Y - X\beta\right), \sigma^2 B^{-1}\right)$$

with $B = (x*\tilde{\beta})^T(x*\tilde{\beta}) + \frac{\sigma^2}{\kappa^2} diag[\xi_1^2, ...\xi_J^2]^{-1}$.

7.3.2 Conditional Distribution of σ^2

Now, I will derive the conditional distribution for σ^2. Returning to the joint probability, drop all terms that do not include σ^2:

$$\frac{1}{(\sigma^2)^{\frac{T_0-1}{2}}}\ exp\left\{\frac{-1}{2\sigma^2}\left(Y - X\beta - (X*\tilde{\beta})\sqrt{\theta}\right)^T\left(Y - X\beta - (X*\tilde{\beta})\sqrt{\theta}\right)\right\}$$
$$(\sigma^2)^{-a_1-1} exp\left\{-\frac{a_2}{\sigma^2}\right\}$$

Rearranging yields:

$$(\sigma^2)^{-\frac{T_0-1}{2}-a_1-1}\ exp\left\{\frac{-1}{2\sigma^2}\left(Y - X\beta - (X*\tilde{\beta})\sqrt{\theta}\right)^T\left(Y - X\beta - (X*\tilde{\beta})\sqrt{\theta}\right) - \frac{a_2}{\sigma^2}\right\}$$

which is an inverse gamma distribution without the scaling term. Therefore, σ^2 is conditionally inverse gamma with *shape* parameter $\frac{T_0-1}{2} + a_1$ and *scale* parameter

$\frac{1}{2}\left(Y - X\beta - (X * \tilde{\beta})\sqrt{\theta}\right)^t \left(Y - X\beta - (X * \tilde{\beta})\sqrt{\theta}\right) + a_2.$

7.3.3 Conditional Distribution of τ_j^2 and ξ_j^2

Focusing only on terms involving τ_j^2, the conditional distribution is:

$$\frac{1}{(\tau_j^2)^{1/2}} exp\left\{\frac{-1}{(2\tau_j^2\lambda^2)}\beta_j^2 - \tau_j^2\right\}$$

Park and Casella (2008) note that by setting $\frac{1}{\tau_j^2} = \zeta^2$, the density can be rewritten proportionally as an inverse Gaussian:

$$\begin{aligned}(\zeta^2)^{-3/2} exp\left\{-\left(\frac{\beta_j^2\zeta_j^2}{2\lambda^2} + \tau_j^2\right)\right\} &\propto (\zeta^2)^{-3/2} exp\left\{\frac{-\beta_j^2}{2\zeta^2\lambda^2}\left[\zeta^2 - \sqrt{\frac{2\lambda^2}{\beta_j^2}}\right]^2\right\} \\ &= (\zeta^2)^{-3/2} exp\left\{\frac{-2}{2\zeta^2\frac{2\lambda^2}{\beta_j^2}}\left[\zeta^2 - \sqrt{\frac{2\lambda^2}{\beta_j^2}}\right]^2\right\}\end{aligned}$$

This is one of many parameterizations of the Inverse Gaussian distribution. The Inverse Gaussian distribution can be written as:

$$f(x) = \sqrt{\frac{\lambda'}{2\pi}}\, x^{-3/2} exp\left\{-\frac{\lambda'(x-\mu')^2}{2(\mu')^2 x}\right\}$$

with mean parameter μ' and scale parameter λ.

Therefore, $\frac{1}{\tau_j^2}$ is conditionally distributed Invserse Gaussian with mean paramters $\frac{2\lambda^2}{\beta_j^2}$ and scale parameter $\lambda' = 2$. ξ_j^2 is derived following the same steps.

7.3.4 Conditional Distribution of λ^2 and κ^2

Focusing solely on λ^2 in the joint distribution yields:

$$(\lambda^2)^{-\frac{J+1}{2}-1} exp\left\{\left(-\frac{\sum_{j=1}^{J}\tau_j^2}{2}-\frac{1}{\zeta_\beta}\right)\frac{1}{\lambda^2}\right\}$$

which is proportional to an inverse gamma distribution with *shape* parameter $\frac{J+1}{2}$ and *rate* parameter $\frac{1}{\zeta_\beta}+\frac{1}{2}\sum_{j=1}^{J}\frac{\beta_j^2}{\tau_j^2}$.

Similarly, κ^2 will follow an inverse gamma distribution with *shape* parameter $\frac{J+1}{2}$ and *rate* parameter $\frac{1}{\zeta_{\sqrt{\theta}}}+\frac{1}{2}\sum_{j=1}^{J}\frac{\sqrt{\theta_j}^2}{\xi_j^2}$.

7.3.5 Sample $\zeta_{\sqrt{\theta}}$ and ζ_β

Finally, ζ_β will follow an inverse gamma with shape 1 and rate $1+\frac{1}{\lambda^2}$. Similalry, $\zeta_{\sqrt{\theta}}$ will follow an inverse gamma with shape 1 and rate $1+\frac{1}{\kappa^2}$.

Work Cited and References

Abadie, Alberto. n.d. "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects," 44.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association* 105 (490): 493–505. https://doi.org/10.1198/jasa.2009.ap08746.

Athey, Susan, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. 2018. "Matrix Completion Methods for Causal Panel Data Models." *arXiv:1710.10251 [Econ, Math, Stat]*, September. http://arxiv.org/abs/1710.10251.

Athey, Susan, and Guido W. Imbens. 2017. "The State of Applied Econometrics: Causality and Policy Evaluation." *Journal of Economic Perspectives* 31 (2): 3–32. https://doi.org/10.1257/jep.31.2.3.

Belmonte, Miguel A.G., Gary Koop, and Dimitris Korobilis. 2014. "Hierarchical Shrinkage in Time-Varying Parameter Models: Hierarchical Shrinkage in Time-Varying Parameter Models." *Journal of Forecasting* 33 (1): 80–94. https://doi.org/10.1002/for.2276.

Bhattacharya, Anirban, Antik Chakraborty, and Bani K. Mallick. 2016. "Fast Sampling with Gaussian Scale-Mixture Priors in High-Dimensional Regression." *arXiv:1506.04778 [Stat]*, June. http://arxiv.org/abs/1506.04778.

Bitto, Angela, and Sylvia Frühwirth-Schnatter. 2019. "Achieving Shrinkage in a Time-Varying Parameter Model Framework." *Journal of Econometrics* 210 (1): 75–97. https://doi.org/10.1016/j.jeconom.2018.11.006.

Botosaru, Irene, and Bruno Ferman. 2019. "On the Role of Covariates in the Synthetic Control Method." *The Econometrics Journal*, January, utz001. https://doi.org/10.1093/ectj/utz001.

Brodersen, Kay H., Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L. Scott. 2015. "Inferring Causal Impact Using Bayesian Structural Time-Series Models." *The Annals of Applied Statistics* 9 (1): 247–74. https://doi.org/10.1214/14-AOAS788.

Dangl, Thomas, and Michael Halling. 2012. "Predictive Regressions with Time-Varying Coefficients." *Journal of Financial Economics* 106 (1): 157–81. https://doi.org/10.1016/j.jfineco.2012.04.003.

Doudchenko, Nikolay, and Guido Imbens. 2016. "Balancing, Regression, Difference-in-Differences and Synthetic Control Methods: A Synthesis." w22791. Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w22791.

Durbin, J., and S. J. Koopman. 2012. *Time Series Analysis by State Space Methods*. 2nd ed. Oxford Statistical Science Series 38. Oxford: Oxford University Press.

Frühwirth-Schnatter, Sylvia, and Helga Wagner. 2010. "Stochastic Model Specification Search for Gaussian and Partial Non-Gaussian State Space Models." *Journal of Econometrics* 154 (1): 85–100. https://doi.org/10.1016/j.jeconom.2009.07.003.

Kaul, Ashok, Stefan Kloßner, Gregor Pfeifer, and Manuel Schieler. 2018. "Synthetic Control Methods: Never Use All Pre-Intervention Outcomes Together with Covariates," 24.

Kinn, Daniel. 2018. "Synthetic Control Methods and Big Data." *arXiv:1803.00096 [Econ]*, February. http://arxiv.org/abs/1803.00096.

Park, Trevor, and George Casella. 2008. "The Bayesian Lasso." *Journal of the American Statistical Association* 103 (482): 681–86. https://doi.org/10.1198/016214508000000337.

Polson, Nicholas G., and James G. Scott. 2011. "On the Half-Cauchy Prior for a Global Scale Parameter." *arXiv:1104.4937 [Stat]*, September. http://arxiv.org/abs/1104.4937.

Samartsidis, Pantelis, Shaun R. Seaman, Anne M. Presanis, Matthew Hickman, and Daniela De Angelis. 2019. "Assessing the Causal Effect of Binary Interventions from Observational Panel Data with Few Treated Units." *Statistical Science* 34 (3): 486–503. https://doi.org/10.1214/19-STS713.

Scott, Steven, and Hal Varian. 2013. "Bayesian Variable Selection for Nowcasting Economic Time Series." w19567. Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w19567.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* 58 (1): 267–88. http://www.jstor.org/stable/2346178.