Synthetic Control with Time Varying Coefficients

A State Space Approach with Bayesian Shrinkage

Danny Klinenberg*

Last Updated: 2020-07-16

Abstract

Synthetic control methods are a popular tool used to measure the effects of policy interventions on a treated unit. In practice, researchers must create a linear combination of untreated units that closely mimics the treated unit before the policy intervention. Recent literature has focused on situations in which a close fit is infeasible. I review recent advances in the synthetic control framework with a focus on estimation and poor pre-treatment fit. I then propose a new approach to estimate the synthetic control counterfactual dynamically relying on a state space framework and Bayesian shrinkage. The dynamics allow for a closer pre-treatment fit extending the methodology to new scenarios. I compare the proposed model to three existing synthetic control models in classic synthetic control settings.

*University of California, Santa Barbara

1 Introduction

In this paper, I consider the problem of estimating a causal effect of an intervention on an outcome of interest when:

1) there is one treated unit,

2) the treatment is binary,

3) the treatment is at an aggregate level (i.e. county, country, market) measured on an aggregate outcome (i.e. GDP per capita, mortality rates, fertility rates),

4) the relationship between the treated unit and control units is non-constant.

A common approach to this problem is the synthetic control framework. The goal is to construct a counterfactual for the treated unit as a linear combination of untreated units. This approach has employed by practitioners in all aspects of economics including (but not limited to) the effects of terrorism (Abadie and Gardeazabal (2003), Bilgel and Karahasan (2017), Dorsett (2013)), trade policies (Aytuğ et al. (2017), Billmeier and Nannicini (2013),), natural disasters (Cavallo et al. (n.d.), Fujiki and Hsiao (2015)) and social issues (Ben-Michael, Feller, and Rothstein (2019), Cunningham and Shah (n.d.), Cunningham, DeAngelo, and Smith (2020), Grossman and Slusky (2019), Powell (2018)).

To fix ideas, suppose...

Abadie, Diamond, and Hainmueller (2010) show if there exists some linear combination of untreated units such that a perfect pretreatment estimate of the treated unit exists, then the asymptotic bias of the treatment effect is zero. However, they explicitly warn against the uses of synthetic control with a poor pretreatment fit. The poor pretreatment fit is a sign that the weighted average of untreated units is not properly representing the unobserved confounders. The authors then suggest testing the assumption by comparing the synthetic control estimate to the true data with a placebo cutoff.

Ferman and Pinto (2019) investigate the validity of synthetic control when the pretreatment fit is imperfect. They show the synthetic control estimator is biased if the treatment assignment is correlated with unobserved confounders regardless of pretreatment size. The authors then suggest modified versions of the estimator that greatly reduce the bias. Add a sentence about curse of dimensionality from their paper.

This paper proposes using dynamic coefficients to achieve a perfect pre-treatment fit. Dynamic coefficients have gained prominence in macroeconomic forecasting (Dangl and Halling (2012), Bitto and Frühwirth-Schnatter (2019), Belmonte, Koop, and Korobilis (2014)). The counterfactual will be modeled using a state space non-centered parameterization. The non-centered parameterization decomposes time varying coefficients into dynamic and static components allowing for individual shrinkage to occur. Dynamic coefficients will be biased towards static coefficients and static towards 0 in the event of overfitting. This reduction allows the proposed model to perform as well as a static-coefficient model when the true data generating process involves only static coefficients and superior otherwise.

I compare the proposed model's performance to Brodersen et al. (2015) *Causal Impact*, Abadie, Diamond, and Hainmueller (2010) synthetic control, and Xu (2017) linear factor model using the classic German Reunification (Abadie, Diamond, and Hainmueller 2015) case study, California Tobacco tax (Abadie, Diamond, and Hainmueller 2010), and the effect of the joining the Eurasian Economic Union on Belarus's GDP per capita. To better understand the benefits of the non-centered parameterization, I perform simulation studies comparing *Causal Impact* with and without time varying parameters to the proposed model with and without time varying parameters.

1.1 Related Work

Developments to synthetic control can be summarized into three main categories: multiple treatments/outcomes, inference, and counterfactual estimation. This paper is focused on counterfactual estimation. Major contributions to the other two categories are briefly

discussed. For a full review of synthetic controls, the reader is directed to Abadie (2019). For an in depth comparison of multiple synthetic control approaches, the reader is directed to Samartsidis et al. (2019) and Kinn (2018).

The synthetic control literature has developed to accommodate multiple treated units. Xu (2017) extended the synthetic control method to multiple outcomes by explicitly modeling the linear factor model. His approach involves explicitly estimating the latent factors in a three step process. Another approach was suggested by Athey et al. (2020) utilizing matrix completion methods from the computer science literature. This approach views the synthetic control problem as one of a missing data problem. The problem then becomes one of imputation with the major contribution coming in the use of the nuclear norm. Additional approaches include L'Hour (2019) and Powell (2018).

Initial inference was based on a permutation test assuming the treatment was randomly assigned. Arkhangelsky et al. (2019) suggest using a jackknife approach to inference which was later utilized by Ben-Michael, Feller, and Rothstein (2019). Li (2019) derived asymptotic results in "long panel" settings. Cattaneo, Feng, and Titiunik (2019) propose prediction intervals that leverage two forms of randomness: the misspecification of the weights and unobserved stochastic error in the post-treatment period. Given the small number of pre/post treatments and small number of controls in typical synthetic control analysis, advances have also been made in small sample inference. Chernozhukov, Wuthrich, and Zhu (2019) proposed a model free inference procedure for synthetic control valid in finite samples.

Initially, Abadie, Diamond, and Hainmueller (2010) derives the synthetic control estimator with no intercept where the coefficients must sum to 1 and be non-negative. This was suggested to avoid extrapolation bias. Doudchenko and Imbens (2016) then investigated machine learning methods to estimate the counterfactual focusing heavily on the elastic net. The authors dropped the restrictions on the coefficients and added an intercept. Hsiao, Steve Ching, and Ki Wan (2012) propose using ordinary least squares to estimate the counterfactual. Powell (2018) uses a two step approach which first predicts values of the outcome

and uses the predicted outcomes to calculate counterfactuals. The benefit of this model is the addition of transitory shocks to the outcome variable and better pre-treatment fit. Ben-Michael, Feller, and Rothstein (2019) suggest using a bias correction to obtain a better pre-treatment fit. They derive assymptotic properties using an augmented ridge regression with an assumed linear factors model.

Bayesian methods have also been used to estimate the counterfactual. Brodersen et al. (2015)'s proposal, *Causal Impact*, models the counterfactual using a combination of spike and slab priors and linear Gaussian state space modeling. The spike and slab priors are used to perform automatic variable selection. The authors allow for the coefficients to be constant or dynamic. They warn of the dangers of overfitting and implausibly large probability intervals with dynamic coefficients (Brodersen et al. 2015). Although popularly cited in synthetic control literature[1] and included in simulation studies[2], there have been few developments to this specific approach.

Pang, Liu, and Xu (n.d.)[3] develop a Bayesian approach based off of the Xu (2017) model utilizing Bitto and Frühwirth-Schnatter (2019) non-centered parameterization. Their model can be viewed as a generalized version of the proposed model in this paper. The key differences are that: i) this model does not aim to measure the latent factor loadings, ii) this model focuses on the case of one treated unit and no controls, and iii) the Bayesian shrinkage used in this model follows a local global shrinkage approach inspired by Polson and Scott (2011b). Finally, Gutman, Intrator, and Lancaster (2018) proposed a Bayesian approach to synthetic control with multiple treated units. A major contribution of this paper is the clear identification of the underlying assumptions for Bayesian synthetic control approaches.

Samartsidis et al. (2020) developed a Bayesian model to identify the treatment effects of multiple units and multiple outcomes. Their approach builds off the linear factors model. This paper differs from Samartsidis et al. (2020) in scope and shrinkage method. I focus on

[1] See Athey et al. (2018), Abadie (2019) Doudchenko and Imbens (2016), and Xu (2017).
[2] See Kinn (2018) and Samartsidis et al. (2019).
[3] This paper is a working draft presented at a luncheon.

one treated unit, one outcome and utilize the global-local priors.

The decomposition and shrinkage of time varying coefficients in a state space framework is popular in macroeconomic forecasting. Frühwirth-Schnatter and Wagner (2010) first proposed this idea in addition to Bayesian shrinkage priors. This process shrinks time varying coefficients to time invariant when overfitting occurs. Belmonte, Koop, and Korobilis (2014) then extended this idea to a different set of Bayesian shrinkage. Bitto and Frühwirth-Schnatter (2019) generalized the set of Bayesian shrinkage priors used in Belmonte, Koop, and Korobilis (2014) to obtain finer predictions.

2 Setup

2.1 Potential Outcomes and Parameter of Interest

Let $\big(y_{j,t}(0), y_{j,t}(1)\big)$ represent potential outcomes in the presence and absence of a treatment with $t = 1, ..., T_0-1, T_0, T_0+1, ...T$ and $j = 0, 1, ..., J$. I assume the potential outcomes are fixed values that follow a linear factors model:

Abadie 2010 makes the point that synthetic control is meant for aggregate data. He then makes the argument that aggregate data is measured with little to no error, so arguing for a super-population may not make as much sense. In future works where we look at disaggregated data, such as multiple states being treated, assuming fixed potential outcomes is not warranted. The choice of fixed potential outcomes comes from the scope of the question: we are interested in a case study where only one aggregate unit has been treated and we are looking at an aggregate outcome.

Assumption 2.1. The potential outcomes are given as:

$$y_{j,t}(D_{j,t}) = \begin{cases} \delta_t + \lambda_t \mu_j + \epsilon_{j,t} & D_{j,t} = 0 \\ \tau_{j,t} + \delta_t + \lambda_t \mu_j + \epsilon_{j,t} & D_{j,t} = 1 \end{cases}$$

where δ_t is an unknown common factor with constant factor loadings across all j, λ_t is a (1xF) vector of common factors, μ_j is a (Fx1) vector of unknown factor loadings and $\epsilon_{j,t}$ are unobserved idiosyncratic shocks. I focus this analysis on the case where there are no 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.

Suppose only one unit, $j = 0$, is treated beginning in period $t = T_0$ and remains treated for all $t \geq T_0$. Suppose the other J units are unaffected by the treatment and no anticipation effects (SUTVA holds, Rubin (1990))[4]. Define $Y_{j,t} = (1 - D_{j,t})y_{j,t}(0) + D_{j,t}y_{j,t}(1)$ where $D_{j,t} = I(j = 0, t \geq T_0)$. The researcher observes the following:

$$Y_{j,t} = \begin{cases} y_{j,t}(1) & j = 0 \ \& \ t \geq T_0 \\ y_{j,t}(0) & else \end{cases}$$

WHY DOES FERMAN 2019 SAY THIS: We treat $\tau_{1,t}$ as given once the sample is drawn, as did Abadie et al. (2010) and Xu (2017).

The parameter of interest is the individual treatment effect of the treated unit at each $t \geq T_0$ denoted $\tau_{0,t} = y_{0,t}(1) - y_{0,t}(0) = Y_{0,t} - y_{0,t}(0)$. Estimating the treatment effect is an exercise in estimating the missing potential outcome, $y_{0,t}(0)$.

2.2 Identification

Assumption 2.2. Conditional Independence on Past Outcomes

[4]This is a common assumption in the synthetic control literature. Recently, Grossi et al. (2020) have introduced spillover effects in analyzing new light rail transit.

$$y_{j,T_0+i}(0) \perp\!\!\!\perp D_{j,T_0+i} | y_{j,1}(0), \ldots, y_{j,T_0}(0) \tag{1}$$

for $i \in \{1, \ldots, T - T_0\}$.

Intuition: The conditional independence assumption uses the full set of pretreatment outcomes to proxy for the unobserved confounders. A better fit on the pretreatment outcomes implies a better representation of the unobserved confounders and better prediction of the unobserved potential outcome. Botosaru and Ferman (2019) show if there exists a perfect estimate of pre-treatment outcomes for the treated unit, then the bias is bounded. More so, the estimate is asymptotically unbiased.

2.3 General Model

In order to estimate $y_{0,t}(0)$, define:

$$y_{0,t}(0) = f_v(\mathbf{Y}) + \epsilon_t \tag{2}$$

where v represents some parameters, $\mathbf{Y}$ represents the matrix of all untreated units in the pre-treatment period, and $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$. I propose using a time-varying coefficient model. Namely:

$$f_v(\mathbf{Y}) = \sum_{j=1}^{J+1} \beta_{j,t} y_{j,t}(0) \tag{3}$$

where $y_{J+1,t}(0) = 1$ for all t (intercept). With a time varying structure, a perfect fit can be made for each $y_{0,t}(0)$. More so, there exists an infinite combination of perfect matches. In addition, the model would have no out of sample predictive ability. The perfect fit would

have no differentiation between signal and noise. To account for this, I add structure to the time varying coefficients through state space modeling. The coefficients are modeled as random walks. Incorporating (3) into (2) and adding the random walk component yields:

$$y_{0,t}(0) = \sum_{j=1}^{J+1} \beta_{j,t} y_{j,t}(0) + \epsilon_t \qquad \epsilon_t \sim \mathcal{N}(0, \sigma^2) \tag{4}$$

$$\beta_{j,t} = \beta_{j,t-1} + \eta_{j,t} \qquad \eta_{j,t} \sim \mathcal{N}(0, \theta_j) \quad \forall j \tag{5}$$

$$\beta_{j,0} \sim \mathcal{N}(\beta_j, \theta_j P_{jj}) \qquad \forall j \tag{6}$$

Dangl and Halling (2012) compared different state 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 state equations. This model is **not** assuming the data follows a random walk. Rather, this model is assuming the coefficients follow a random walk. Finally, the choice of random walk allows for a useful decomposition of the coefficients into time-varying and constant parts used in following sections.

The parameters of the model are $\mathbf{v} = \{\sigma^2, \beta_1, ...\beta_{J+1}, \theta_1, ...\theta_{J+1}\}$ with ϵ_t and $\eta_{j,t}$ assumed independent of all other unknowns. P_{jj} is a hyperparameter set to ensure the initial distribution is disperse. In practice, P_{jj} is set to a very large number value (Durbin and Koopman 2012).

In general state space form, the model is written as:

$$y_{0,t}(0) = y_{.,t}(0)\Xi_t + \epsilon_t \qquad \epsilon_t \sim \mathcal{N}(0, \sigma^2) \quad \text{observation equation} \tag{7}$$

$$\Xi_{t+1} = \mathbf{T_t}\Xi_t + \mathbf{R_t}\eta_t \qquad \eta_t \sim \mathcal{N}(0, Q) \quad \text{state equation} \tag{8}$$

$$\Xi_0 \sim \mathcal{N}(a_0, P_0) \tag{9}$$

where

$$\mathbf{T_t} = I, \quad \mathbf{R_t} = I,$$

$$\mathbf{P_0} = diag\left[\theta_1 P_{11}, \dots, \theta_{J+1} P_{J+1,J+1}\right],$$

$$\mathbf{Q} = diag\left[\theta_1, \dots, \theta_{J+1}\right],$$

$$\Xi_t = \begin{bmatrix} \beta_{1,t} \\ \vdots \\ \beta_{J+1,t} \end{bmatrix}, \quad \Xi_0 = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_{J+1} \end{bmatrix}$$

where $diag(*)$ represents a diagonal matrix with the specified elements on the diagonal.

Notice that the evolution of $\beta_{j,t}$ can be changed by changing $\mathbf{T_t}$. For example, setting $\mathbf{T_t} = diag[\rho_1, \dots, \rho_{J+1}]$ creates an AR(1) process. In addition, adding off diagonal terms add interdependence between the evolution of the coefficients. Setting $\beta_{j,t} = \beta_j$ and allowing for a local linear time trend will yield the original estimator in Brodersen et al. (2015). The reader is referred to Durbin and Koopman (2012) for an advanced treatment of state space modeling.

3 Estimation of Parameters and Counterfactual

rework with paragraph doesn't bleong here

An increasingly common issue in economic analysis (especially counterfactual analysis) is more untreated units than pre-treatment periods. For example, Abadie, Diamond, and Hainmueller (2010) considers a situation in which there are 29 untreated units and 17 pre-treatment periods. Past researchers have addressed this issue with machine learning techniques (e.g. Doudchenko and Imbens (2016), Athey et al. (2018)). Bayesian solutions have also been suggested. Pang (2010) recommended a model comparison algorithm using Bayes factors while Pang, Liu, and Xu (n.d.) and Brodersen et al. (2015) incorporate Bayesian shrinkage priors, which place a majority of mass of the prior distribution at zero. This forces coefficients biased towards zero which allows for the usage of more covariates than

observations and better out of sample predictions while avoiding overfitting.

To perform regularization, I incorporate a variant of the Bayesian Lasso (Park and Casella 2008). I model the shrinkage using the global-local shrinkage framework (Polson and Scott 2011b). This framework is designed to apply stronger amounts of shrinkage for smaller parameter estimates allowing larger parameter estimates to "escape" leading to better in-sample and out-of-sample fits. Let $\pi(\mathbf{v})$ define the prior distribution on the parameters $\mathbf{v}$.

end of thing that doesn't belong

There are three sources of information to estimate $y_{0,T_0+i}(0)$:

i) Untreated pre-treatment units: $\mathbf{Y}$.

ii) Untreated post-treatment units.

iii) Treated pre-treatment units: $\mathbf{Y_0} = \Big(y_{0,1}(0), \dots, y_{0,T_0-1}(0)\Big)$.

The researcher will use all the observations prior to treatment and the control observations posttreatment to estimate $y_{0,T_0+i}(0)$. Namely, $y_{0,T_0+i}(0)$ will be estimated using the posterior predictive density:

$$f(y_{0,T_0+i}(0)|\mathbf{Y_0}, \mathbf{Y}) = \int f(y_{0,T_0+i}(0)|\mathbf{Y_0}, \mathbf{Y}, \mathbf{v}) Pr(\mathbf{v}|\mathbf{Y_0}, \mathbf{Y}) dv \tag{10}$$

Define $\hat{f}(y_{0,T_0+i}(0)|\mathbf{Y_0}, \mathbf{Y})$ as the estimate of $f(y_{0,T_0+i}(0)|\mathbf{Y_0}, \mathbf{Y})$ and $\hat{y}_{0,t}(0)$ is a random variable drawn from $\hat{f}(y_{0,T_0+i}(0)|\mathbf{Y_0}, \mathbf{Y})$.

The treatment effect is then estimated as:

$$\hat{\tau}_{0,t} = y_{0,t}(1) - \hat{y}_{0,t}(0) \tag{11}$$

$$= y_{0,t}(1) - y_{0,t}(0) + y_{0,t}(0) - \hat{y}_{0,t}(0) \tag{12}$$

$$= \tau_{0,t} + y_{0,t}(0) - \hat{y}_{0,t}(0) \tag{13}$$

A major benefit of the Bayesian approach is finite sample credibility intervals. Bayesian analysis captures uncertainty through the priors. Inference can be conducted from credibility intervals regardless of sample size. However, many researchers are also interested in the large sample properties of estimators. In addition to small sample inference, this model provides an asymptotically unbiased estimate of the treatment effect.

3.1 Asymptotic Unbiasedness

Prior to proving asymptotic unbiasedness of $\mathbb{E}\left[\alpha_{0,t}\right]$, some mechanisms must be introduced.

Definition 3.1. The Kullback-Leibler Divergence of the proposed distribution $Pr(y_{0,T_0+i}(0)|Y_0)$ with respect to the true distribution $f\left(y_{0,T_0+i}(0)\right)$ is defined as:

$$KL_f(\tilde{v}) = \mathbb{E}\left(log\left(\frac{f\left(y_{0,T_0+i}(0)\right)}{Pr(y_{0,T_0+i}(0)|Y_0)}\right)\right)$$

where $\tilde{v}$ denotes the parameters and f the true distribution. The $KL_f(v) \geq 0$ with equality when the true data generating process is $Pr(y_{0,T_0+i}(0)|Y_0)$ (Gelman 2014).

Define v^0 as the unique set of parameters such that:

$$v^0 = argmin_{\tilde{v}}\mathbb{E}\left(log\left(\frac{f\left(y_{0,T_0+i}(0)\right)}{Pr(y_{0,T_0+i}(0)|Y_0)}\right)\right)$$

If the true data generating process is included in the model specification (i.e. $f\left(y_{0,T_0+i}(0)\right) =$

$Pr(y_{0,T_0+i}(0)|Y_0))$, then $KL_f(v^0) = 0$. Otherwise, v^0 is the value of the parameters that minimizes the distance between the model and the true data generating process with respect to the Kullback-Leibler Divergence.

The treatment effect can be written as $\mathbb{E}[\widehat{\alpha}_{0,t}] = \mathbb{E}\left[y_{0,t}(1) - \widehat{y}_{0,t}(0)\right] = \mathbb{E}\left[\alpha_{0,t} + y_{0,t}(0) - \widehat{y}_{0,t}(0)\right]$. In order for $\mathbb{E}[\widehat{\alpha}_{0,t}] \xrightarrow{p} \mathbb{E}[\alpha_{0,t}]$ as $T_0 - 1 \to \infty$, $\mathbb{E}\left[y_{0,t}(0) - \widehat{y}_{0,t}(0)\right] \xrightarrow{p} 0$. It is sufficient to show $Pr(\widehat{y}_{0,T_0+i}(0)|Y_0) \xrightarrow{p} Pr(y_{0,T_0+i}(0)|Y_0)$ as $T_0 - 1 \to \infty$. In order to show this, it is sufficient to show that the posterior of the parameters converge to the true values (i.e. $Pr(v^0|Y_0) \xrightarrow{p} 1$).

Theorem 3.1 (Asymptotic Unbiasedness)**.** Assume that:

1) the model specification is the true data generating process.

2) v is defined on a finite parameter space V and $Pr(v = v^0) > 0$.

3) $\{y_{0,1}(0), ...y_{0,T}(0)\}$ is a stationary ergodic process.

4) The poster distribution $Pr(v|Y_0)$ is unimodal.

If all the above assumptions are met, then $\mathbb{E}[\widehat{\alpha}_{0,t}] \xrightarrow{p} \mathbb{E}[\alpha_{0,t}]$ as $T_0 - 1 \xrightarrow{p} \infty$.

I will prove this result in three steps. the first step is based off of Gelman (2014) (appendix B).

Proof. Step 1: $Pr(v^0|Y_0) \xrightarrow{p} 1$

Define $v' \neq v^0$. Consider the following:

$$log\left(\frac{Pr(v=v'|Y_0)}{Pr(v=v^0|Y_0)}\right) = log\left(\frac{\pi(v=v')}{\pi(v=v^0)}\right) + log\left(\frac{Pr(Y_0|v)}{Pr(Y_0|v^0)}\right) \quad \text{Bayes Rule} \tag{14}$$

$$= log\left(\frac{\pi(v=v')}{\pi(v=v^0)}\right) + log\left(\frac{\Pi_{t=1}^{T_0-1} Pr(y_{0,t}(0)|v=v')}{\Pi_{t=1}^{T_0-1} Pr(y_{0,t}(0)|v=v^0)}\right) \quad \text{Model Specification} \tag{15}$$

$$= log\left(\frac{\pi(v=v')}{\pi(v=v^0)}\right) + \sum_{t=1}^{T_0-1} log\left(\frac{Pr(y_{0,t}(0)|v=v')}{Pr(y_{0,t}(0)|v=v^0)}\right) \quad \text{log rules} \tag{16}$$

$$= log\left(\frac{\pi(v=v')}{\pi(v=v^0)}\right) + \frac{T_0-1}{T_0-1}\sum_{t=1}^{T_0-1} log\left(\frac{Pr(y_{0,t}(0)|v=v')}{Pr(y_{0,t}(0)|v=v^0)}\right) \quad \text{multiply by 1} \tag{17}$$

Focus for a moment on the second term. Notice that as $T_0 - 1 \to \infty$:

$$\frac{1}{T_0-1}\sum_{t=1}^{T_0-1} log\left(\frac{Pr(y_{0,t}(0)|v=v')}{Pr(y_{0,t}(0)|v=v^0)}\right) \to \mathbb{E}\left(log\left(\frac{Pr(y_{0,t}(0)|v=v')}{Pr(y_{0,t}(0)|v=v^0)}\right)\right) \quad \text{ergodic stationary} \tag{18}$$

$$= \mathbb{E}\left(log\left(\frac{Pr(y_{0,t}(0)|v=v')f(y_{0,t}(0))}{Pr(y_{0,t}(0)|v=v^0)f(y_{0,t}(0))}\right)\right) \quad \text{multiply by 1} \tag{19}$$

$$= KL_f(v^0) - KL_f(v') \tag{20}$$

$$< 0 \tag{21}$$

Combining this result and (10) yields:

$$log\left(\frac{Pr(v=v'|Y_0)}{Pr(v=v^0|Y_0)}\right) \to -\infty$$

as $T_0 - 1 \to \infty$. Rearranging concludes $Pr(v = v'|Y_0) \to 0$ for all $v \neq v^0$. Therefore, $Pr(v = v^0|Y_0) \to 1$. Equivalently, $v \to v^0$ as $T_0 - 1 \to \infty$.

Step 2: $Pr(\hat{y}_{0,T_0+i}(0)|Y_0) \xrightarrow{p} Pr(y_{0,T_0+i}(0)|Y_0)$

Recall:

$$Pr(\hat{y}_{0,T_0+i}(0)|Y_0) = \sum_{v' \in V} Pr(y_{0,T_0+i}(0)|Y_0, v = v')Pr(v = v'|Y_0)$$

I need help here. I'm not sure if I can pull the limit into the integral. I feel like I can't, but I also know this should be an asymptotic unbiased estimator from other papers.

step 3: $\mathbb{E}[\hat{\alpha}_{0,t}] \xrightarrow{p} \mathbb{E}[\alpha_{0,t}]$ as $T_0 - 1 \to \infty$

In order for $\mathbb{E}[\hat{\alpha}_{0,t}] \xrightarrow{p} \mathbb{E}[\alpha_{0,t}]$ as $T_0 - 1 \to \infty$, $\mathbb{E}\left[y_{0,T_0+i}(0) - \hat{y}_{0,T-0+i}(0)\right] \xrightarrow{p} 0$

—Note for me—

Once I get step 2, I think I can say if $Pr(\hat{y}_{0,T_0+i}(0)|Y_0) \xrightarrow{p} Pr(y_{0,T_0+i}(0)|Y_0)$, then $\mathbb{E}\left[y_{0,t}(0) - \hat{y}_{0,t}(0)|Y_0\right] \xrightarrow{p} 0$ which then leads to $\mathbb{E}\left[\mathbb{E}\left[y_{0,t}(0) - \hat{y}_{0,t}(0)|Y_0\right]\right] \xrightarrow{p} 0$.

—

□

Intuition As the pre-treatment period gets larger and larger, the choice of priors will become less important. The data will converge to a point mass on the true parameter value, v^0. In turn, the mean of the poster predictive distribution will converge to the true value $y_{0,T_0+i}(0)$ leading to an asymptotic unbiased estimator of the treatment effect.

Remark. So long as the choice of priors assigns non-zero probability to the true parameter values v^0, then the choice of priors is irrelevant in the limit. However, this is not true for finite sample estimation. In application, the choice of priors can have massive effects on inference.

Remark. This proof was shown using a finite parameter space. However, the assumption can be relaxed by assuming a compact set on the parameter space. Should I do it for continuous parameter space? I don't think it adds anything

Remark. This result provides assurance of unbiasedness in the limit. However, inference is not drawn from asymptotic results.

#

4 Reparameterization

Equation (5) can be rewritten to decompose $\beta_{j,t}$ into a time varying and constant components:[5]

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

β_j can now be interpreted as the time invariant component of $\beta_{j,t}$ and $\sqrt{\theta_j}\tilde{\beta}_{j,t}$ the time varying component. $\sqrt{\theta_j}$ is defined as the root of θ_j and allowed to take both positive and negative values. Defining $\sqrt{\theta_j}$ in this manner allows 0 to be an interior point in the prior distribution. This is a desirable feature when performing Bayesian shrinkage (Bitto and Frühwirth-Schnatter 2019). The absolute value of $\sqrt{\theta_j}$ is the standard deviation of time varying coefficient. Substituting the reformulation back into the original equation yields the state space model:

$$Y_{0,t} = \sum_{j=1}^{J+1} \left(\beta_j + \tilde{\beta}_{j,t}\sqrt{\theta_j}\right) Y_{j,t} + \epsilon_t \qquad \epsilon_t|\sigma^2 \sim N(0,\sigma^2) \tag{22}$$

$$\tilde{\beta}_{j,t} = \tilde{\beta}_{j,t-1} + \tilde{\eta}_{j,t} \qquad \tilde{\eta}_{j,t} \sim N(0,1) \tag{23}$$

$$\tilde{\beta}_{j,0} \sim N(0, P_{jj}) \tag{24}$$

Equations (4), (5), and (6) constitute the model. This setup is commonly known as the

[5] See the appendix for further explanation.

non-centered parameterization of state space models. This formulation allows estimation of the time varying and time invariant component of the coefficients individually which allows for 4 types of coefficients: (i) time varying non-zero, (ii) time invariant, (iii) time varying centered at zero, and (iv) time invariant zero coefficients (irrelevant).

With this formulation there are 2(J+1)+1 parameters to be estimated: the J+1 time invariant coefficients (i.e. β_j's), the J+1 time varying coefficients (i.e. $\sqrt{\theta_j}$'s) and the variance (σ^2).

4.1 Bayesian Shrinkage Priors

I set up the prior distribution for coefficients $\beta = [\beta_1, \beta_2, ..., \beta_{J+1}]$ with variances $\alpha^2 = [\alpha_1^2, \alpha_2^2, ..., \alpha_{J+1}^2]$ as a *global-local* shrinkage prior:

$$\begin{aligned} \beta|\alpha^2, \lambda^2 &\sim \mathcal{N}_{J+1}(0_{J+1}, \lambda^2 diag[\alpha_1^2, ...\alpha_{J+1}^2]) \\ \alpha_j^2 &\sim \pi(\alpha_j^2) \\ \lambda^2 &\sim \pi(\lambda^2) \end{aligned}$$

This prior formulation has gained popularity in the Bayesian framework due to it's attractive shrinkage properties (Makalic and Schmidt (2016), Polson and Scott (2011b)). λ^2 controls the overall complexity of the model while α_j^2 produces individual shrinkage. This formulation allows for strong shrinkage on small coefficients while leaving larger coefficients relatively unshrunk.

α_j^2 is assigned an exponential distribution with rate 1. The hierarchical formulation of β and α^2 is identical to independent Laplace priors. Such a prior forms the Bayesian LASSO proposed by Park and Casella (2008). Park and Casella (2008) showed this choice of priors leads to posterior performance similar to the frequentist machine learning approach LASSO (Tibshirani 1996). The authors derived the joint posterior distributions as well as formulated a Gibbs sampling technique.

λ^2 is represented as a half-Cauchy distribution with mean 0 and scale parameter 1. The

half-Cauchy is used for the global shrinkage prior because of the flexibility and better behavior near 0 compared to alternatives (Polson and Scott 2011a). In addition, the half-Cauchy has significant amounts of mass at the point 0 leading to better shrinkage properties.

Like the Laplace distribution, the half-Cauchy has a hierarchical representation where $\lambda^2|\zeta_\beta$ follows an inverse gamma with shape parameter 1/2 and rate $1/\zeta_\beta$. The hierarchical parameter, ζ_β, follows an inverse gamma with shape parameter 1/2 and rate parameter 1. Therefore, the prior distribution for $\beta = [\beta_1, \beta_2, ..., \beta_{J+1}]$ with variances $\alpha^2 = [\alpha_1^2, \alpha_2^2, ..., \alpha_{J+1}^2]$ are:

$$\beta|\alpha^2, \lambda^2 \sim \mathcal{N}_{J+1}(0_{J+1}, \lambda^2 diag[\alpha_1^2, ...\alpha_{J+1}^2]) \tag{25}$$

$$\alpha_j^2 \sim exp\,(1) \tag{26}$$

$$\lambda^2|\zeta_\beta \sim InverseGamma\left(\frac{1}{2}, \frac{1}{\zeta_\beta}\right) \tag{27}$$

$$\zeta_\beta \sim InverseGamma\left(\frac{1}{2}, 1\right) \tag{28}$$

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, the inverse gamma prior performs poorly in terms of shrinkage due to 0 being an extreme value in the distribution. This limits the amount of mass that can be placed at 0 in turn limiting the amount of shrinkage. The normal distribution allows for mass at 0 avoiding this problem. Similarly to β, assign the prior of $\sqrt{\theta} = \left[\sqrt{\theta_1}, \sqrt{\theta_2}, ..., \sqrt{\theta_{J+1}}\right]$ with variances $\xi^2 = [\xi_1^2, \xi_2^2, ...\xi_{J+1}^2]$ as:

$$\sqrt{\theta}|\xi^2, \kappa^2 \sim \mathcal{N}_{J+1}(0_{J+1}, \kappa^2 diag[\xi_1^2, ...\xi_{J+1}^2]) \tag{29}$$

$$\xi_j^2 \sim exp\,(1) \tag{30}$$

$$\kappa^2|\zeta_{\sqrt{\theta}} \sim InverseGamma\left(\frac{1}{2}, \frac{1}{\zeta_{\sqrt{\theta}}}\right) \tag{31}$$

$$\zeta_{\sqrt{\theta}} \sim InverseGamma\left(\frac{1}{2}, 1\right) \tag{32}$$

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.

5 The Posterior Estimation (MCMC)

In order to draw predictions for the counterfactual, the posterior distribution must be calculated: $P(\tilde{\beta}, \beta, \alpha^2, \lambda^2, \sqrt{\theta}, \xi^2, \kappa^2, \zeta_\beta, \zeta_{\sqrt{\theta}}, \sigma^2|Y_0)$. With values drawn from the posterior distribution, $\widehat{y}_{0,t}(0)$ can then be estimated for $t \geq T_0$. A closed form does not exist for the posterior. Therefore, I implement the Gibbs sampler. 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.

The posterior estimation can be broken into three main steps:

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

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

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

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

Draw $\tilde{\beta}$ using Durbin (2002) for the state space model. First, rewrite equations (4), (5), and (6) as:

$$Y'_{0,t} = \sum_{j=1}^{J+1} \tilde{\beta}_{j,t} Z_{j,t} + \epsilon_t \qquad \epsilon_t|\sigma^2 \sim N(0, \sigma^2) \tag{33}$$

$$\tilde{\beta}_{j,t} = \tilde{\beta}_{j,t-1} + \tilde{\eta}_{j,t} \qquad \tilde{\eta}_{j,t} \sim N(0,1) \tag{34}$$

$$\tilde{\beta}_{j,0} \sim N(0, P_{jj}) \tag{35}$$

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

Many algorithms have been proposed to simulate latent variables in a state space framework. I use the method proposed by Durbin (2002). I first run the Kalman filter and smoother given the data and parameters to produce $\tilde{\beta}^*_t$. I then simulate new $\tilde{\beta}^+_{j,t}$ and ${Y'}^+_{0,t}$ for all j using equations (15), (16), and (17). I then run the Kalman filter and smoother on ${Y'}^+_{0,t}$ and $\tilde{\beta}^+_{j,t}$ for all j producing $\tilde{\beta}^{*,+}_t$. My new simulated draw of $\tilde{\beta}_t$ (denoted $\tilde{\beta}'_t$) is $\tilde{\beta}'_t = \tilde{\beta}^*_t - \tilde{\beta}^+_t + \tilde{\beta}^{*,+}_t$.

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

Attempting to sample $P(\beta, \alpha^2, \lambda^2, \sqrt{\theta}, \xi^2, \kappa^2, \zeta_\beta, \zeta_{\sqrt{\theta}}, \sigma^2|Y_0)$ 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.

5.2.1 Sample β and $\sqrt{\theta}$

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

$$\mathcal{N}_{2(J+1)}\left((\tilde{Y}^T\tilde{Y}+\sigma^2V^{-1})^{-1}\tilde{Y^T}Y_0, \sigma^2(\tilde{Y}^T\tilde{Y}+\sigma^2V^{-1})^{-1}\right) \tag{36}$$

Where:

$$\tilde{Y} = \begin{pmatrix} Y_{1,1} & Y_{2,1} & ... & Y_{J+1,1} & \tilde{\beta}_{1,1}Y_{1,1} & \tilde{\beta}_{2,1}Y_{2,1} & ... & \tilde{\beta}_{J+1,1}Y_{J+1,1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ Y_{1,T_0-1} & Y_{2,T_0-1} & ... & Y_{J+1,T_0-1} & \tilde{\beta}_{1,T_0-1}Y_{1,T_0-1} & \tilde{\beta}_{2,T_0-1}Y_{2,T_0-1} & ... & \tilde{\beta}_{J+1,T_0-1}Y_{J+1,T_0-1} \end{pmatrix} \tag{37}$$

$$V = diag\left[\lambda^2\alpha_1^2, \lambda^2\alpha_2^2, ..., \lambda^2\alpha_{J+1}^2, \kappa^2\xi_1^2, \kappa^2\xi_2^2, ..., \kappa^2\xi_{J+1}^2\right] \tag{38}$$

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).

5.2.2 Sample α^2

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

$$IG\left(\sqrt{\frac{2\lambda^2}{\beta_j^2}}, 2\right) \text{ for j=1,...J+1} \tag{39}$$

5.2.3 Sample λ^2

Draw λ^2 from the conditional inverse gamma prior:

$$InverseGamma\left(shape = \frac{J+1}{2}, rate = \frac{1}{\zeta_\beta} + \frac{1}{2}\sum_{j=1}^{J+1}\frac{\beta_j^2}{\alpha_j^2}\right) \tag{40}$$

5.2.4 Sample ζ_β

Draw ζ_β from the conditional inverse gamma prior:

$$InverseGamma\left(1, 1 + \frac{1}{\lambda^2}\right) \tag{41}$$

5.2.5 Sample ξ^2

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

$$IG\left(\sqrt{\frac{2\kappa^2}{\theta_j}}, 2\right) \text{ for j=1,...J+1} \tag{42}$$

5.2.6 Sample κ^2

Draw κ^2 from the conditional gamma prior:

$$InverseGamma\left(shape = \frac{J+1}{2}, rate = \frac{1}{\zeta_{\sqrt{\theta}}} + \frac{1}{2}\sum_{j=1}^{J+1}\frac{\sqrt{\theta_j}^2}{\xi_j^2}\right) \tag{43}$$

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

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

$$InverseGamma\left(1, 1 + \frac{1}{\kappa^2}\right) \tag{44}$$

5.2.8 Sample σ^2

Draw σ^2 from the posterior distribution:

$$InverseGamma\left(a_1 + \frac{T_0 - 1}{2}, a_2 + \frac{\sum_{t=1}^{T_0-1}\left(Y_{0,t} - \sum_{j=1}^{J+1}\left(\beta_j + \tilde{\beta}_{j,t}\sqrt{\theta_j}\right)Y_{j,t}\right)^2}{2}\right) \quad (45)$$

Frühwirth-Schnatter and Wagner (2010) note an identification problem arises when using the non-centered parameterization. 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 $\sum_{j=1}^{J+1}\alpha_j^2$ where each posterior α_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.[6] To account for this, **all covariates except the intercept are scaled to mean zero variance one** prior to analysis.

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

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

[6]For example, the LASSO is defined as: $\beta = argmin_b \sum_i (y_i - Y_i b)^2 + \lambda \sum_i |b_i|$. If one covariate 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.

(1) Simulate $\tilde{\beta}_{j,t} = \tilde{\beta}_{j,t-1} + \tilde{\eta}_{j,t}$ for all j for $t \geq T_0$. Use $\tilde{\beta}'_{j,T_0-1}$ simulated in section 4.1 as an initial value. Notice that each iteration of the Gibbs sampler will create a new $\tilde{\beta}'_{j,T_0-1}$.

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

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

drawing $\epsilon_t|\sigma^2 \sim N(0, \sigma^2)$. Section 4.2.1 provides the draws for β_j for all j. Section 4.2.5 provides the draws for $\sqrt{\theta_j}$ for all j. Section 4.2.9 provides the draw for σ^2 used for determining ϵ_t. Each iteration of the Gibbs sampler will produce new parameter and state values.

It is often reasonable to assume the model estimated in the pre-treatment continues to represent the data generating process in the post-period. Reasons this assumption may not be applicable include the controls becoming treated or additional events affecting the treatment. The post-treatment period should be chosen such that the controls remain untreated and there are no other events affecting the treatment.

6 Monte Carlo Simulation Data

For the purpose of this paper, the argument that covariates follow the same time varying 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) data generating processes. 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:

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

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

The treatment begins at period T_0.

This paper proposes testing four scenarios: (i) deterministic continuous varying coefficients with no treatment effect, (ii) deterministic continuous varying coefficients with a 5 unit treatment effect, (iii) constant coefficients with no treatment effect, (iv) constant coefficients with a 5 unit treatment effect. Scenario (i) and (ii) will provide insight on the point prediction accuracy (via mean squared forecast error) and the probability interval size. Scenarios (iii) and (iv) will provide insight on the ability of the models to identify treatment effects.

6.1 Deterministic Continuous Varying Coefficients

To simulate continuous varying coefficients, $c_{1,t}$ and $c_{2,t}$ are defined .75 and .25 respectively. All other $c_{j,t}$ are randomly drawn from U[0,1]. In order to avoid $y_{1,t}$ and $y_{2,t}$ from crossing, set $z_1 = 25$ and $z_2 = 5$. 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 the time varying case.

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 (Time Varying)

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

The data generating process for the time varying coefficient case can be rewritten in recursive form:

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

with initial conditions:

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

6.2 Constant Coefficients

The setup for constant coefficients is identical to deterministic continuous varying coefficients except point (6) is replaced by (6'):

(6') $w_{1,t} = .2$, $w_{2,t} = 1 - w_{2,t}$, and $w_{j,t} = 0$ for all else (Time Invariant).

See the appendix for example plots of the four scenarios.

6.3 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). These treatment effects will be calculated by defining $Y_{0,t}(1) = \alpha + Y_{0,t}(0)$ for $\alpha \in \{0, 5\}$. Given this data generating process, 5 represents about a 20% treatment effect. Each specification is run twice: once with time varying coefficients and once without time varying coefficients

I will compare the mean squared forecast error (MSFE), post treatment coverage of the 95% probability interval (95% PI), and the estimated treatment effect (TE) in the post period. Each measurement will be defined as:

$$\text{MSFE} \equiv \frac{1}{T - T_0} \sum_{t=T_0}^{T} \left(Y_{0,t} - \hat{y}_{0,t}\right)^2$$

$$95\%\ \text{PI} \equiv \frac{1}{T - T_0} \sum_{t=T_0}^{T} I\left(Y_{0,t} \in \left[\hat{y}_{0,t}^{.025}, y_{0,t}^{.975}\right]\right)$$

$$\text{TE} \equiv \frac{1}{T - T_0} \sum_{t=T_0}^{T} \left(Y_{0,t} - \hat{y}_{0,t}\right)$$

where $\hat{y}_{0,t}$ is the median of the posterior predictive density created by each model specification, $\hat{y}_{0,t}^{.025}$ and $\hat{y}_{0,t}^{.975}$ are the 2.5^{th} and 97.5^{th} quantiles of the posterior estimations.

7 Preliminary Results

Initial simulations are run using the package *Causal Impact* to create estimates for **Causal Impact No TVP** and **Causal Impact TVP**. I then compare the results to the proposed model. In the results, I call the proposed model **Bayesian LASSO Time Varying Parameter (Bayesian LASSO TVP)**. I also compare the proposed model to the **Bayesian LASSO without time varying Parameter (Bayesian LASSO No TVP)**. Bayesian LASSO No TVP is the proposed model where $\sqrt{\theta} = 0$. The comparison between **Bayesian**

LASSO TVP and **Bayesian LASSO No TVP** showcases the benefits of time varying coefficients. The appendix has example plots of the four scenarios.

Table 1: Monte Carlo Simulation: Mean Squared Forecast Error

	Constant		Deterministic Continuous Varying	
Model	$\tau = 0$	$\tau = 5$	$\tau = 0$	$\tau = 5$
Causal Impact No TVP	6.550	5.340	18.941	61.657
Causal Impact TVP	8.683	19.460	16.344	51.702
Bayesian Lasso No TVP	2.303	11.760	22.799	69.871
Bayesian Lasso TVP	3.440	11.997	12.529	48.819

* Median results of 100 monte carlo simulations.
† Each simulation of Bayesian Lasso TVP is run 3000 times with a 1500 burn-in.
‡ All other models are run according to presets.
§ The preset Causal Impact model was used as described in Brodersen et al. 2015.

Initial simulations suggest the Bayesian Lasso TVP has a lower mean squared forecast error compared to *Causal Impact No TVP* and *Causal Impact TVP* in the constant coefficient case as well as in the deterministic continuous varying coefficient case. However, this could be due to either the choice of priors or the time varying coefficient decomposition. Both *Bayesian Lasso No TVP* and *Bayesian Lasso TVP* suggest lower mean squared forecast errors in the constant coefficient model.

The benefit of the Bayesian Lasso TVP is showcased in the deterministic continuous varying coefficient case. Bayesian Lasso TVP showcases lower mean squared forecast error compared to all three models. However, this is one simulation study run 100 times. These results are **suggestive** of potential benefits.

8 Conclusion

This proposal adds shrinkage among time varying coefficients to counterfactual analysis. The setup of the model automatically shrinks time varying coefficients towards static coefficients if the model is overfitting. Therefore, the formulation allows for the use of time

varying coefficients with reduced risk of overfitting. Initial simulations suggest this formulation performs better than the pre-existing state space models used in counterfactual analysis.

9 Things I Still Need

In order of importance:

1) A proof

I am investigating oracle inequality proofs. Unfortunately, this is requiring far more machine learning theory than I expected.

Samartsidis et al. (2019) provide a brief "proof" of asymptotic unbiasedness for *Casual Impact*. However, the proof seems a bit lacking.

2) A real life example

My initial real life example, Brexit, has received serious concern from the macroeconomic reading group. I am looking for an example in the performance based aid literature. This may be a stronger example because countries who receive aid are by definition changing faster.

3) The Gibbs Sampler

I can reorganize my block draw to speed up the process. Over summer, I plan to rewrite the code drawing β and $\sqrt{\theta}$ together. I am also investigating ways to speed up the draw of $\tilde{\beta}$. One option is using All Without a Loop (AWOL) proposed in Bitto and Frühwirth-Schnatter (2019).

Table 2: Monte Carlo Simulation: Constant Coefficients, Treatment Effect=0

Model	Pretreatment MSE	Pretreatment Coverage	Post Treat MSFE	95% CI	CI Spread	Estimated Treatment Effect
Causal Impact No TVP	0.644	1.000	6.550	1.000	6.246	-2.230
Causal Impact TVP	0.000	1.000	8.683	1.000	11.536	-0.358
Bayesian Lasso No TVP	0.559	0.882	2.303	0.941	6.202	-0.824
Bayesian Lasso TVP	0.243	1.000	3.440	0.941	8.152	-0.833

Median results of 100 monte carlo simulations. Each simulation is run 3000 times with a 1500 burn-in.
95% confidence intervals are calculated using the 97.5th percentile and 2.5th percentile.
Coverage refers to the average inclusion rate of the 95% credibility interval.
CI Spread is the 97.5% percentile less the 2.5% percentile

Table 3: Monte Carlo Simulation: Deterministic Continuous Varying Coefficients, Treatment Effect=0

Model	Pretreatment MSE	Pretreatment Coverage	Post Treat MSFE	95% CI	CI Spread	Estimated Treatment Effect
Causal Impact No TVP	0.712	1.000	18.941	0.824	9.951	3.824
Causal Impact TVP	0.000	1.000	16.344	1.000	16.804	3.088
Bayesian Lasso No TVP	0.439	0.941	22.799	0.647	8.804	4.216
Bayesian Lasso TVP	0.212	1.000	12.529	0.882	9.147	2.901

Median results of 100 monte carlo simulations. Each simulation is run 3000 times with a 1500 burn-in.
95% confidence intervals are calculated using the 97.5th percentile and 2.5th percentile.
Coverage refers to the average inclusion rate of the 95% credibility interval.
CI Spread is the 97.5% percentile less the 2.5% percentile

10 Additional Tables

11 Appendix

11.1 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 state equation. A general form of these equations follows:

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

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

11.2 State 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 $\eta_{j,t}$.

11.3 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 hierarchical structure is redefined to be a *global-local* shrinkage estimator. Park and Casella (2008) use a hierarchical 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. 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:

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

The conditional prior of $\mathbf{Y_0}$ is defined as $\mathcal{N}\left(\mathbf{Y}\beta_j + (\mathbf{Y} * \tilde{\beta}_j)\sqrt{\theta_j}), \sigma^2 I\right)$ where * denotes element wise multiplication. Conditional on α_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:

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

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

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

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

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

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

To solve for the conditional distribution of β and $\sqrt{\theta}$, drop the terms that don't involve β and $\sqrt{\theta}$. This only leaves 3 exponential terms:

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

Combining exponents yields:

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

Define:

$$\tilde{Y}=\left[X,X*\tilde{\beta}\right]_{T_0-1,2(J+1)}$$

,

$$\Theta = \left[\beta, \sqrt{\theta}\right]_{2(J+1),2(J+1)}$$

and

$$D = diag\left[\lambda^2\alpha_1^2, ...\lambda^2\alpha_{J+1}^2, \kappa^2\xi_1^1, ..., \kappa^2\xi_{J+1}^2\right]_{2(J+1),2(J+1)}$$

.

Focusing solely on the exponential term and rearranging yields:

$$\frac{-1}{2\sigma^2}\left[\left(Y - \tilde{Y}\Theta\right)^T\left(Y - \tilde{Y}\Theta\right) + \Theta^T\sigma^2V^{-1}\Theta\right]$$

Multiplying out and rearranging yields:

$$\frac{-1}{2\sigma^2}\left[Y^TY - 2Y\tilde{Y}\Theta + \Theta^T(\tilde{Y}^TY + \sigma^2V^{-1})\Theta\right]$$

Focus solely on the terms within the brackets including Θ for a moment. Setting $A = \tilde{Y}^T\tilde{Y} + \sigma^2V^{-1}$ and completing the square yields:

$$\left(\Theta - A^{-1}\tilde{Y}^TY\right)^T A\left(\Theta - A^{-1}\tilde{Y}^TY\right)$$
$$+ Y^T\left(I - \tilde{Y}A^{-1}\tilde{Y}^T\right)Y$$

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

$$\frac{1}{\sqrt{2\pi\sigma^2}}exp\left\{\frac{-1}{2\sigma^2}\left(\Theta - A^{-1}\tilde{Y}^TY\right)^T A\left(\Theta - A^{-1}\tilde{Y}^TY\right)\right\}$$

which can be summarized as Θ conditionally distributed as:

$$\mathcal{N}\left(A^{-1}\tilde{Y}^TY, \sigma^2A^{-1}\right)$$

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

11.3.3 Conditional Distribution of α_j^2 and ξ_j^2

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

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

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

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

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

11.3.4 Conditional Distribution of λ^2 and κ^2

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

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

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

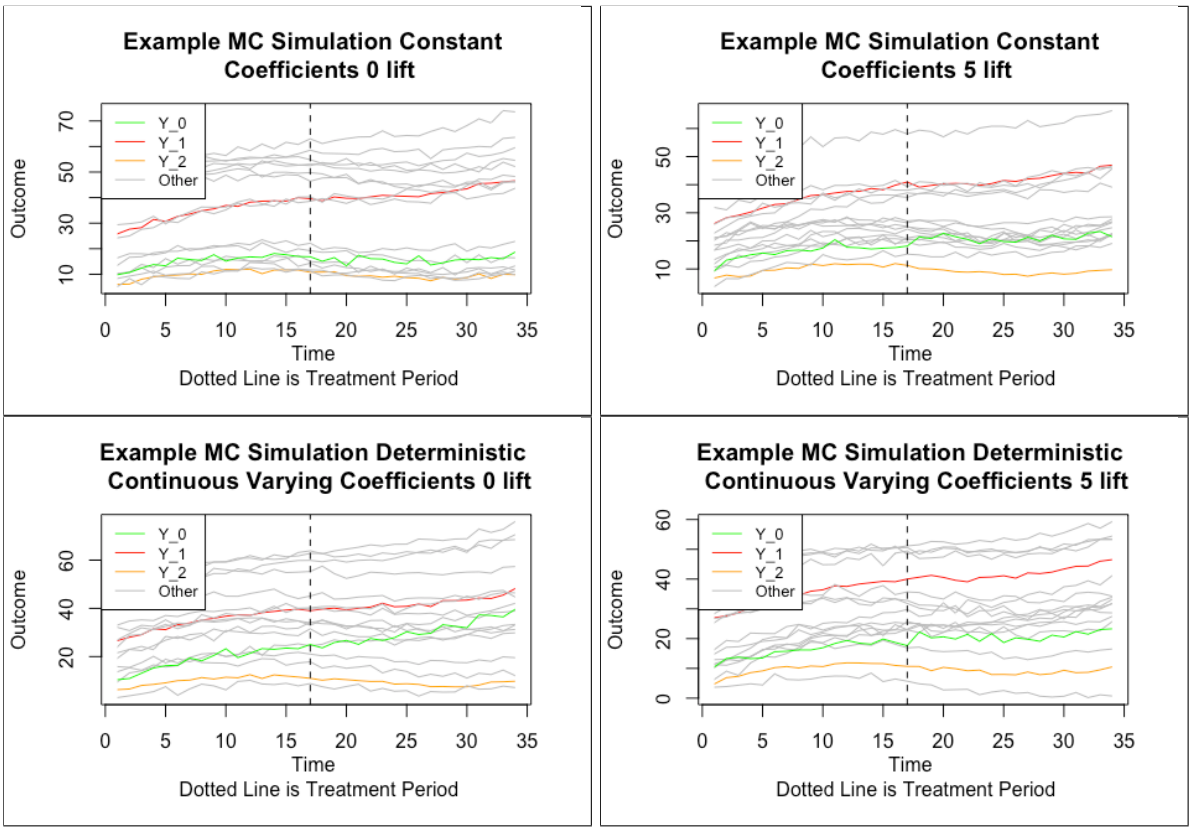
Similarly, κ^2 will follow an inverse gamma distribution with *shape* parameter $\frac{J+1}{2}$ and *rate* parameter $\frac{1}{\zeta_{\sqrt{\theta}}} + \frac{1}{2}\sum_{j=1}^{J+1}\frac{\sqrt{\theta_j}^2}{\xi_j^2}$.

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

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

11.4 Example Plots

Figure 1: Example MCMC Draws Graphed



The four models are abbreviated in the following plots:

BL TVP: Bayesian Lasso Time Varying Parameter (green)

BL No TVP: Bayesian Lasso No Time Varying Parameter (red)

CI No TVP: Causal Impact No Time Varying Parameter (blue)

CI TVP: Causal Impact Time Varying Parameter (pink)

Figure 2: Example Plots Constant Coefficients 0 Treatment Effect

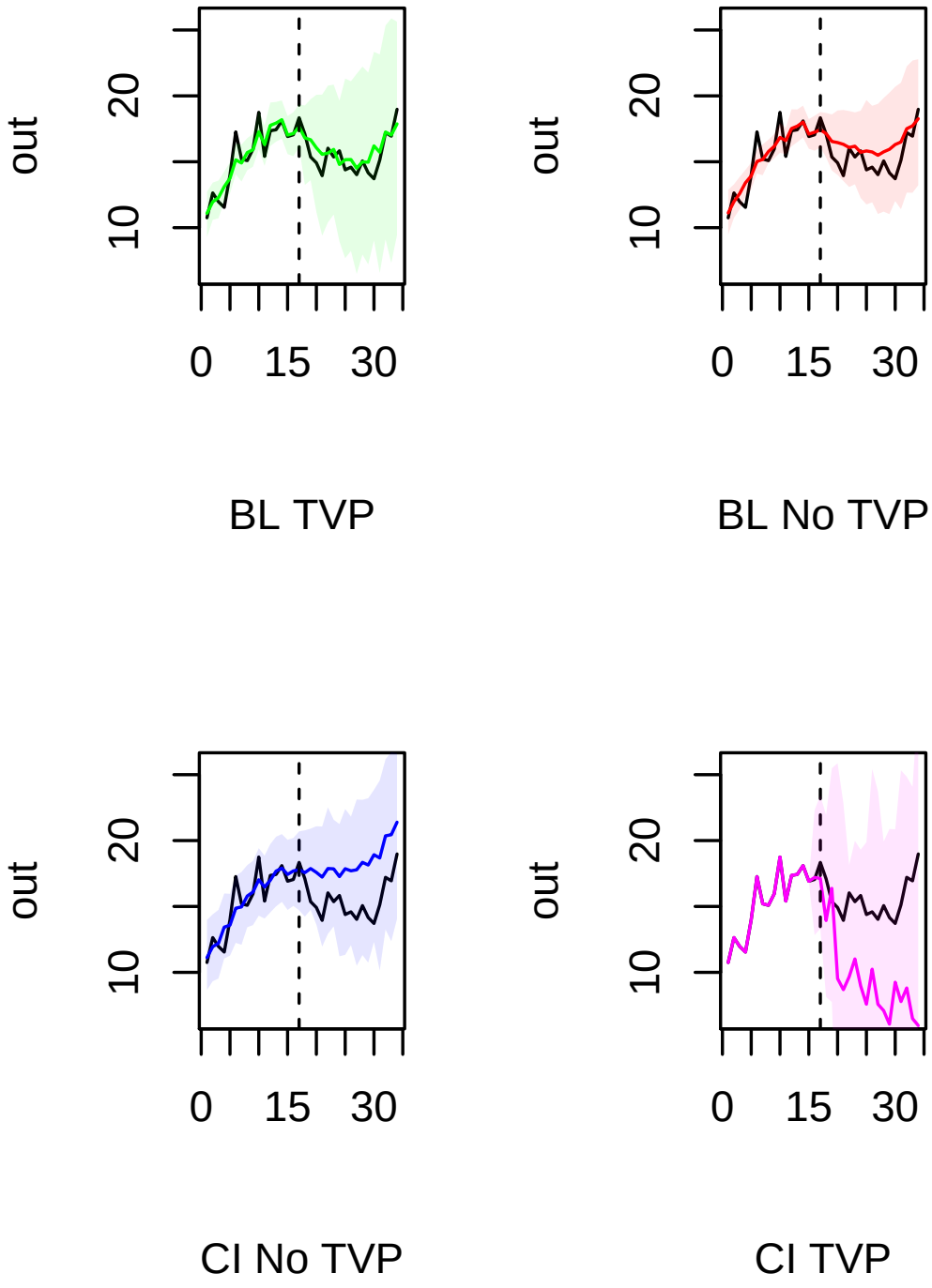


Figure 3: Example Plots Constant Coefficients 5 Treatment Effect

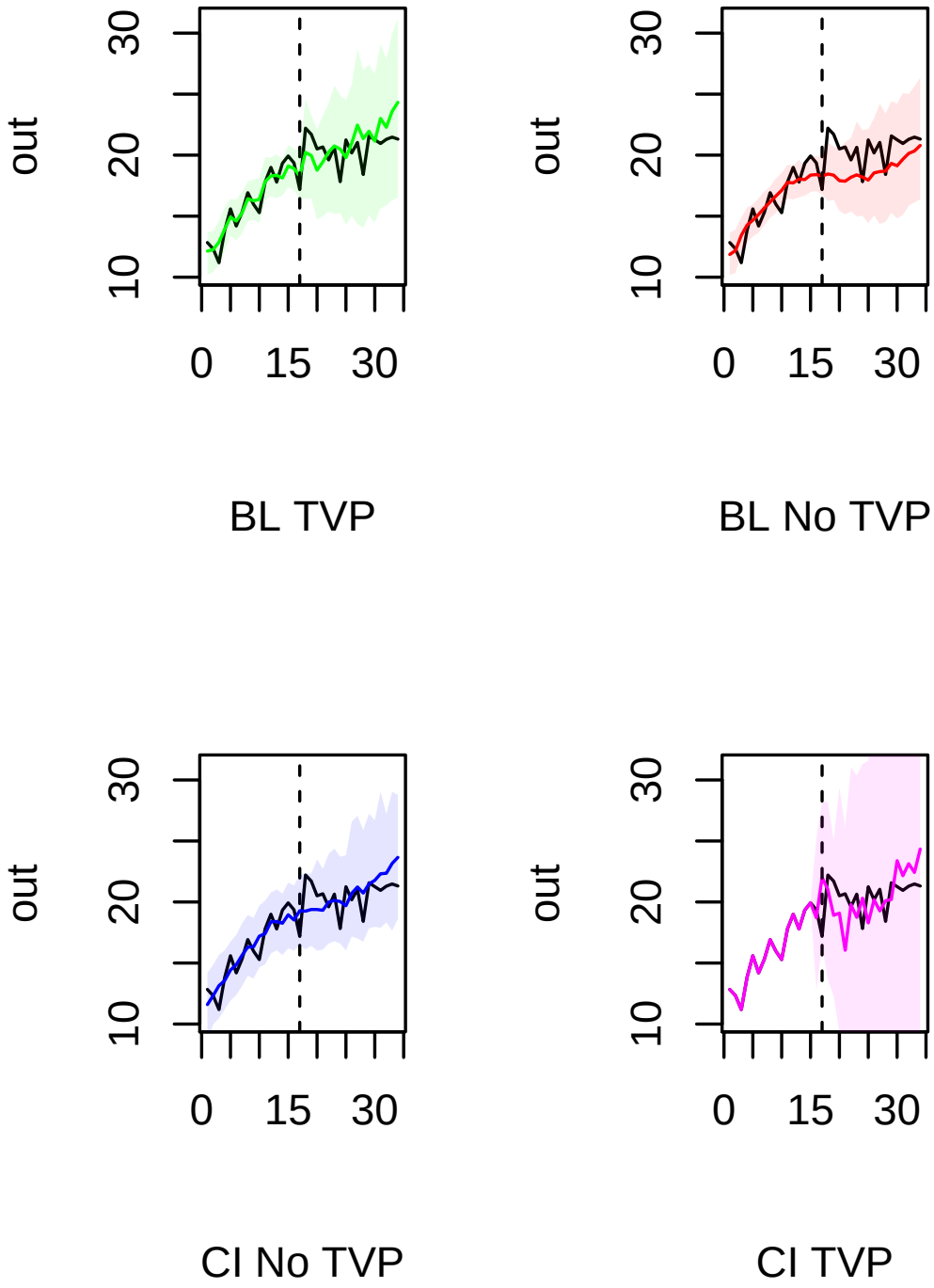


Figure 4: Example Plots Deterministic Continuous Varying Coefficients 0 Treatment Effect

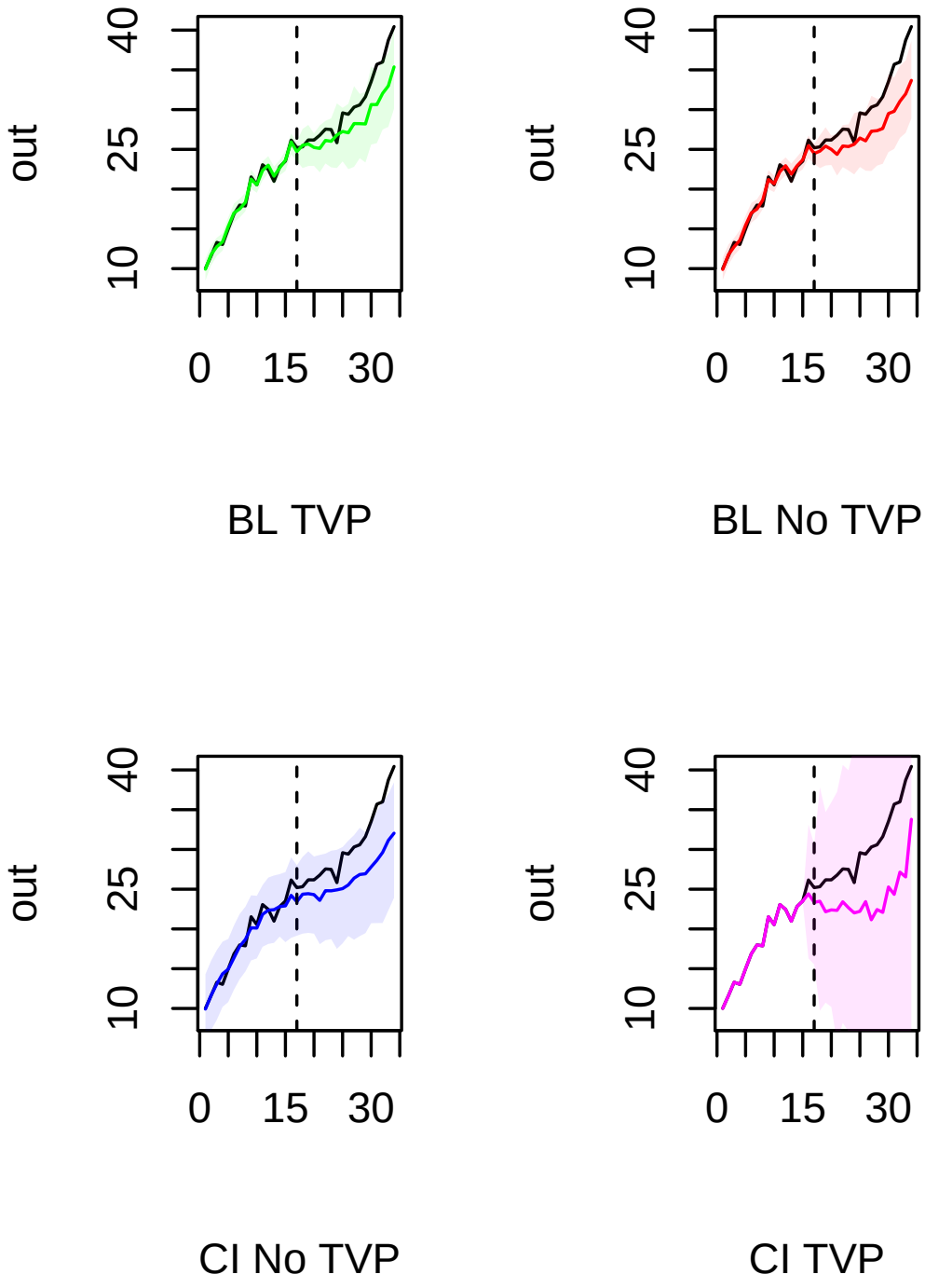
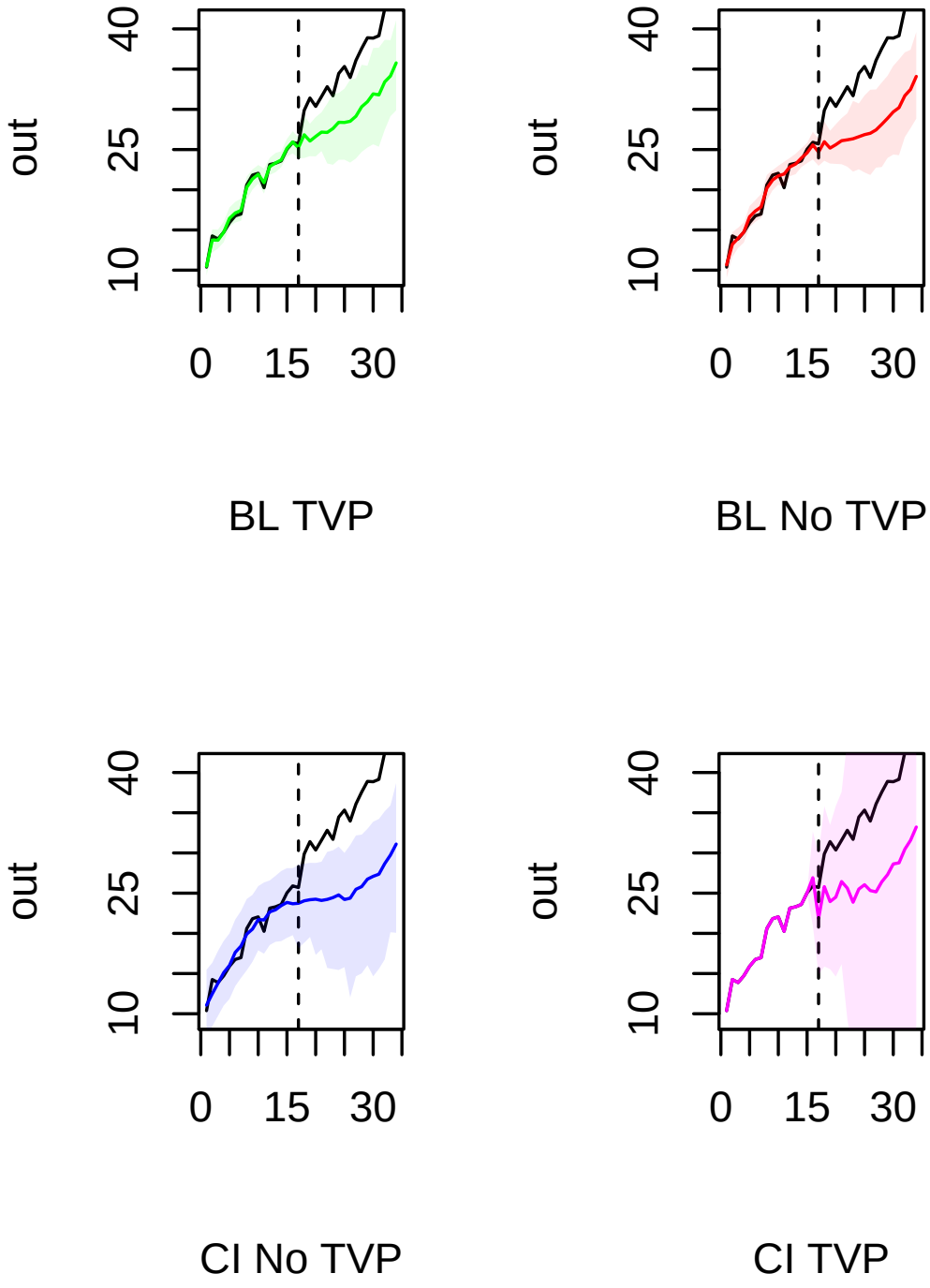


Figure 5: Example Plots Deterministic Continuous Varying Coefficients 5 Treatment Effect



Work Cited

Abadie, Alberto. 2019. "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.

———. 2015. "Comparative Politics and the Synthetic Control Method: COMPARATIVE POLITICS AND THE SYNTHETIC CONTROL METHOD." *American Journal of Political Science* 59 (2): 495–510. https://doi.org/10.1111/ajps.12116.

Abadie, Alberto, and Javier Gardeazabal. 2003. "The Economic Costs of Conflict: A Case Study of the Basque Country." *American Economic Review* 93 (1): 113–32. https://doi.org/10.1257/000282803321455188.

Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager. 2019. "Synthetic Difference in Differences." *arXiv:1812.09970 [Stat]*, January. http://arxiv.org/abs/1812.09970.

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.

———. 2020. "Matrix Completion Methods for Causal Panel Data Models." *arXiv:1710.10251 [Econ, Math, Stat]*, June. 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.

Aytuğ, Hüseyin, Merve Mavuş Kütük, Arif Oduncu, and Sübidey Togan. 2017. "Twenty Years of the EU-Turkey Customs Union: A Synthetic Control Method Analysis: Twenty Years of the EU-Turkey Customs Union: Effects of EU Integration." *JCMS: Journal of*

Common Market Studies 55 (3): 419–31. https://doi.org/10.1111/jcms.12490.

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.

Ben-Michael, Eli, Avi Feller, and Jesse Rothstein. 2019. "The Augmented Synthetic Control Method." *arXiv:1811.04170 [Econ, Stat]*, November. http://arxiv.org/abs/1811.04170.

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.

Bilgel, Fırat, and Burhan Can Karahasan. 2017. "The Economic Costs of Separatist Terrorism in Turkey." *Journal of Conflict Resolution* 61 (2): 457–79. https://doi.org/10.1177/0022002715576572.

Billmeier, Andreas, and Tommaso Nannicini. 2013. "Assessing Economic Liberalization Episodes: A Synthetic Control Approach." *Review of Economics and Statistics* 95 (3): 983–1001. https://doi.org/10.1162/REST_a_00324.

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.

Cattaneo, Matias D., Yingjie Feng, and Rocio Titiunik. 2019. "Prediction Intervals for Synthetic Control Methods." *arXiv:1912.07120 [Econ, Stat]*, December. http://arxiv.org/

abs/1912.07120.

Cavallo, Eduardo, Sebastian Galiani, Ilan Noy, and Juan Pantano. n.d. "CATASTROPHIC NATURAL DISASTERS AND ECONOMIC GROWTH." *THE REVIEW OF ECONOMICS AND STATISTICS*, 13.

Chernozhukov, Victor, Kaspar Wuthrich, and Yinchu Zhu. 2019. "An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls." *arXiv:1712.09089 [Econ, Stat]*, November. http://arxiv.org/abs/1712.09089.

Cunningham, Scott, Gregory DeAngelo, and Brock Smith. 2020. "Fracking and Risky Sexual Activity." *Journal of Health Economics* 72 (July): 102322. https://doi.org/10.1016/j.jhealeco.2020.102322.

Cunningham, Scott, and Manisha Shah. n.d. "Decriminalizing Indoor Prostitution: Implications for Sexual Violence and Public Health," 55.

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.

Dorsett, Richard. 2013. "The Effect of the Troubles on GDP in Northern Ireland." *European Journal of Political Economy* 29 (March): 119–33. https://doi.org/10.1016/j.ejpoleco.2012.10.003.

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. 2002. "A Simple and Efficient Simulation Smoother for State Space Time Series Analysis." *Biometrika* 89 (3): 603–16. https://doi.org/10.1093/biomet/89.3.603.

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.

Ferman, Bruno, and Cristine Pinto. 2019. "Synthetic Controls with Imperfect Pre-Treatment Fit." *arXiv:1911.08521 [Econ]*, November. http://arxiv.org/abs/1911.08521.

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.

Fujiki, Hiroshi, and Cheng Hsiao. 2015. "Disentangling the Effects of Multiple Treatments—Measuring the Net Economic Impact of the 1995 Great Hanshin-Awaji Earthquake." *Journal of Econometrics* 186 (1): 66–73. https://doi.org/10.1016/j.jeconom.2014.10.010.

Gelman, Andrew. 2014. "Bayesian Data Analysis." In *Bayesian Data Analysis*, Third edition, 585–88. Chapman & Hall/CRC Texts in Statistical Science. Boca Raton: CRC Press.

Grossi, Giulio, Patrizia Lattarulo, Marco Mariani, Alessandra Mattei, and Özge Öner. 2020. "Synthetic Control Group Methods in the Presence of Interference: The Direct and Spillover Effects of Light Rail on Neighborhood Retail Activity." *arXiv:2004.05027 [Econ, Stat]*, June. http://arxiv.org/abs/2004.05027.

Grossman, Daniel S., and David J. G. Slusky. 2019. "The Impact of the Flint Water Crisis on Fertility." *Demography* 56 (6): 2005–31. https://doi.org/10.1007/s13524-019-00831-0.

Gutman, Roee, Orna Intrator, and Tony Lancaster. 2018. "A Bayesian Procedure for Estimating the Causal Effects of Nursing Home Bed-Hold Policy." *Biostatistics* 19 (4): 444–60. https://doi.org/10.1093/biostatistics/kxx049.

Hsiao, Cheng, H. Steve Ching, and Shui Ki Wan. 2012. "A PANEL DATA APPROACH FOR PROGRAM EVALUATION: MEASURING THE BENEFITS OF POLITICAL AND ECONOMIC INTEGRATION OF HONG KONG WITH MAINLAND CHINA." *Journal of Applied Econometrics* 27 (5): 705–40. https://doi.org/10.1002/jae.1230.

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.

L'Hour, Alberto Abadie Jeremy. 2019. "A Penalized Synthetic Control Estimator for Disaggregated Data," 53.

Li, Kathleen T. 2019. "Statistical Inference for Average Treatment Effects Estimated by Synthetic Control Methods." *Journal of the American Statistical Association*, December, 1–16. https://doi.org/10.1080/01621459.2019.1686986.

Makalic, Enes, and Daniel F. Schmidt. 2016. "High-Dimensional Bayesian Regularised Regression with the BayesReg Package." *arXiv:1611.06649 [Stat]*, December. http://arxiv.org/abs/1611.06649.

Pang, Xun. 2010. "Modeling Heterogeneity and Serial Correlation in Binary Time-Series Cross-Sectional Data: A Bayesian Multilevel Model with AR(p) Errors." *Political Analysis* 18 (4): 470–98. https://doi.org/10.1093/pan/mpq019.

Pang, Xun, Licheng Liu, and Yiqing Xu. n.d. "Bayesian Predictive Synthesis for Causal Inference with TSCS Data: A Multilevel State-Space Factor Model with Hierarchical Shrinkage Priors," 54.

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

———. 2011b. "Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction*." In *Bayesian Statistics 9*, edited by José M. Bernardo, M. J. Bayarri, James O. Berger, A. P. Dawid, David Heckerman, Adrian F. M. Smith, and Mike West, 501–38. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199694587.003.0017.

Powell, David. 2018. "Imperfect Synthetic Controls:" 55.

Rubin, Donald B. 1990. "Formal Mode of Statistical Inference for Causal Effects." *Journal of Statistical Planning and Inference* 25 (3): 279–92. https://doi.org/10.1016/0378-3758(90)90077-8.

Samartsidis, Pantelis, Shaun R. Seaman, Silvia Montagna, André Charlett, Matthew

Hickman, and Daniela De Angelis. 2020. "A Bayesian Multivariate Factor Analysis Model for Evaluating an Intervention by Using Observational Time Series Data on Multiple Outcomes." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, May, rssa.12569. https://doi.org/10.1111/rssa.12569.

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.

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.

Xu, Yiqing. 2017. "Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models." *Political Analysis* 25 (1): 57–76. https://doi.org/10.1017/pan.2016.2.