Counterfactual Analysis with Time Varying Coefficients

A State Space Model Approach with Bayesian Shrinkage

Danny Klinenberg*

Last Updated: 2020-07-01

Abstract

Current synthetic control approaches model the relationship between the treated unit and the control unit as constant throughout the observation period. There are no tests available to determine the validity of the modeling choice. I propose explicitly modeling the relationship between control units and the treated unit as dynamic using a linear Gaussian state space framework. I decompose each coefficient into a time varying and a time invariant coefficient. To avoid overfitting, Bayesian shrinkage priors are used. Therefore, the model produces counterfactual estimates as accurate (measured by mean squared forecast error) 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 method is tested against the existing state space approach, *Causal Impact*, on simulated data and applied to a practical example (forthcoming).

*University of California, Santa Barbara

1 Introduction

A common approach to counterfactual construction is the synthetic control framework. The approach represents the data generating process as a linear factors model (Abadie, Diamond, and Hainmueller 2010). This allows for the creation of a counterfactual using a weighted average of untreated units over a pre-period window. The weights are assumed to be approximately constant (Abadie, Diamond, and Hainmueller 2010).

This assumption creates a tension when performing synthetic control analysis. The researcher wants a long pretreatment period that maintains a constant relationship between the treated unit and untreated units. If the relationship between the untreated units and treated unit is dynamic, the linear factors model used in synthetic control may not be an accurate representation of the true data generating process. The misspecification can lead to biases (Abadie 2019). There are no tests for constant relationships. Without a formal test, determining if the relationships are constant becomes a judgment call.

To address this tension, I develop a state space model for counterfactual analysis with time varying coefficients. An immediate concern is overfitting. This risk is reduced by decomposing each coefficient into a time invariant and a time varying coefficient. The decomposition, referred to as the non-centered parameterization of a state space model, allows for an automatic Bayesian process reducing time varying coefficients to time invariant. This process reduces the risk of overfitting. The model should perform similarly to existing methods (in regards to mean squared forecast error) when the coefficients are relatively constant and superior when the coefficients are dynamic.

This paper is most closely related to Brodersen et al. (2015). Their approach, *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 approach. My proposal builds off of *Causal Impact* but differs in two key ways. First, I incorporate the decomposition of time varying coefficients. This allows for the inclusion of time varying coefficients while limiting the concerns discussed by Brodersen et al. (2015). Second, I use a different set of priors to create the Bayesian LASSO[3].

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.

This paper contributes to the literature by applying a non-centered parameterization state space model to the synthetic control setting. The modeling choice allows for dynamic coefficients with reduced risk of overfitting. I compare this model to the existing state space model, *Casual Impact*, through out of sample mean squared forecast errors, coverage in the post treatment, and estimated treatment effect.

2 Setup

Following the Bayesian philosophy, the observed data, X, is assumed fixed while the model parameters and missing values are assumed to come from some unknown distribution. Let $y_{0,t}$ represent a treated unit observed over $t = 1, ..., T_0 - 1, T_0, T_0 + 1, ...T$. Suppose T_0 represents the year in which a policy change occurs. For example, let $y_{0,t}$ be the GDP of the United Kingdom in period t and T_0 represent the year the Brexit referendum passed (2016).

[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] In the future, this paper will generalize the model to many priors including the slab and spike.

Suppose there are also $j = 1, ..., J$ untreated aggregate units unaffected by the treatment (i.e. no spillover effects). 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 onward. The researcher also observes $X_{j,t} = x_{j,t}(0)$ 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 $\alpha_{0,t} = y_{0,t}(1) - y_{0,t}(0)$. 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 $t \geq T_0$. Since $y_{0,T_0+i}(0)$ for $i \in \{0, 1, 2, ..., T - T_0\}$ is missing, I assume $y_{0,T_0+i}(0) \sim f\left(y_{0,T_0+i}(0)\right)$, where f is some unknown distribution. Thus, $f\left(y_{0,T_0+i}(0)\right)$ must be estimated when $i \in \{0, 1, 2, ..., T - T_0\}$.

3 Model Specification

The model can be decomposed into two parts: linear Gaussian state space model and Bayesian shrinkage priors.

3.1 Linear Gaussian State Space Models

Define the linear Gaussian state space model as:

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

$$\beta_{j,t} = \beta_{j,t-1} + \eta_{j,t} \qquad \eta_{j,t} \sim N(0, \theta_j) \quad \forall j \tag{2}$$

$$\beta_{j,0} \sim N(\beta_j, \theta_j P_{jj}) \qquad \forall j \tag{3}$$

where $X_{J+1,t} = 1$ for all t (intercept). In the state space literature, equation (1) is known as the *observation equation* and equation (2) is referred to as the *state equations*. 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 researcher will use observations before T_0 to calculate $\beta_{j,t}$ and then predict $\hat{y}_{0,t}(0)$ for $t \geq T_0$. 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. The choice of random walk allows for the decomposition of the coefficients into time-varying and constant parts.

θ_j governs the amount of variability in the error term, $\eta_{j,t}$ which in turn controls the variability in $\beta_{j,t}$. As θ_j approaches 0, the distribution of $\eta_{j,t}$ collapses to 0 and $\beta_{j,t}$ simplifies to a time invariant coefficient. β_j represents the mean of $\beta_{j,t}$ at t=0. P_{jj} is a hyperparameter to be set.

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.

The state equation can be rewritten to decompose $\beta_{j,t}$ into a time varying and constant components.[4]

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

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. Thinking back 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). Substituting the reformulation back into the original equation yields the state space model:

$$Y_{0,t} = \sum_{j=1}^{J+1} \left(\beta_j + \tilde{\beta}_{j,t}\sqrt{\theta_j}\right) X_{j,t} + \epsilon_t \qquad \epsilon_t|\sigma^2 \sim N(0,\sigma^2) \tag{4}$$

$$\tilde{\beta}_{j,t} = \tilde{\beta}_{j,t-1} + \tilde{\eta}_{j,t} \qquad \tilde{\eta}_{j,t} \sim N(0,1) \tag{5}$$

$$\tilde{\beta}_{j,0} \sim N(0, P_{jj}) \tag{6}$$

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

[4]See the appendix for further explanation.

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).

3.2 Bayesian Shrinkage Priors

An increasingly common issue in economic analysis (especially counterfactual analysis) is more covariates than observations. For example, Abadie, Diamond, and Hainmueller (2010) considers a situation in which there are 29 covariates and 17 pretreatment periods. Past researchers have addressed this issue with machine learning techniques (e.g. Doudchenko and Imbens (2016), Athey et al. (2018)). A Bayesian solution to this problem is shrinkage priors. Shrinkage priors 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.

I set up the prior distribution for coefficients $\beta = [\beta_1, \beta_2, ..., \beta_{J+1}]$ with variances $\tau^2 = [\tau_1^2, \tau_2^2, ..., \tau_{J+1}^2]$ as a *global-local* shrinkage prior:

$$\begin{aligned} \beta|\tau^2, \lambda^2 &\sim \mathcal{N}_{J+1}(0_{J+1}, \lambda^2 diag[\tau_1^2, ...\tau_{J+1}^2]) \\ \tau_j^2 &\sim \pi(\tau_j^2) \\ \lambda^2 &\sim \pi(\lambda^2) \end{aligned}$$

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 τ_j^2 produces individual shrinkage. This formulation allows for strong shrinkage on small coefficients while leaving larger coefficients relatively unshrunk.

τ_j^2 is assigned an exponential distribution with rate 1. The hierarchical formulation of β and τ^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 $\tau^2 = [\tau_1^2, \tau_2^2, ..., \tau_{J+1}^2]$ are:

$$\beta|\tau^2, \lambda^2 \sim \mathcal{N}_{J+1}(0_{J+1}, \lambda^2 diag[\tau_1^2, ...\tau_{J+1}^2]) \tag{7}$$

$$\tau_j^2 \sim exp\,(1) \tag{8}$$

$$\lambda^2|\zeta_\beta \sim InverseGamma\left(\frac{1}{2}, \frac{1}{\zeta_\beta}\right) \tag{9}$$

$$\zeta_\beta \sim InverseGamma\left(\frac{1}{2}, 1\right) \tag{10}$$

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:

$$\sqrt{\theta}|\xi^2, \kappa^2 \sim \mathcal{N}_{J+1}(0_{J+1}, \kappa^2 diag[\xi_1^2, ...\xi_{J+1}^2]) \tag{11}$$

$$\xi_j^2 \sim exp\,(1) \tag{12}$$

$$\kappa^2|\zeta_{\sqrt{\theta}} \sim InverseGamma\left(\frac{1}{2}, \frac{1}{\zeta_{\sqrt{\theta}}}\right) \tag{13}$$

$$\zeta_{\sqrt{\theta}} \sim InverseGamma\left(\frac{1}{2}, 1\right) \tag{14}$$

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.

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

(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 (2002) for the state space model. First, rewrite equations (4), (5), and (6) as:

$$Y'_{0,t} = \sum_{j=1}^{J+1} \tilde{\beta}_{j,t} Z_{j,t} + \epsilon_t \qquad \epsilon_t|\sigma^2 \sim N(0, \sigma^2) \tag{15}$$

$$\tilde{\beta}_{j,t} = \tilde{\beta}_{j,t-1} + \tilde{\eta}_{j,t} \qquad \tilde{\eta}_{j,t} \sim N(0,1) \tag{16}$$

$$\tilde{\beta}_{j,0} \sim N(0, P_{jj}) \tag{17}$$

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

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 $\tilde{\beta}'_t = \tilde{\beta}^*_t - \tilde{\beta}^+_t + \tilde{\beta}^{*,+}_t$.

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 and $\sqrt{\theta}$

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

$$\mathcal{N}_{2(J+1)}\left((\tilde{X}^T\tilde{X}+\sigma^2V^{-1})^{-1}\tilde{X^T}Y_0, \sigma^2(\tilde{X}^T\tilde{Y}+\sigma^2V^{-1})^{-1}\right) \tag{18}$$

Where:

$$\tilde{X}=\begin{pmatrix} X_{1,1} & X_{2,1} & ... & X_{J+1,1} & \tilde{\beta}_{1,1}X_{1,1} & \tilde{\beta}_{2,1}X_{2,1} & ... & \tilde{\beta}_{J+1,1}X_{J+1,1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{1,T_0-1} & X_{2,T_0-1} & ... & X_{J+1,T_0-1} & \tilde{\beta}_{1,T_0-1}X_{1,T_0-1} & \tilde{\beta}_{2,T_0-1}X_{2,T_0-1} & ... & \tilde{\beta}_{J+1,T_0-1}X_{J+1,T_0-1} \end{pmatrix} \tag{19}$$

$$V = diag\left[\lambda^2\tau_1^2, \lambda^2\tau_2^2, ..., \lambda^2\tau_{J+1}^2, \kappa^2\xi_1^2, \kappa^2\xi_2^2, ..., \kappa^2\xi_{J+1}^2\right] \tag{20}$$

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:

$$IG\left(\sqrt{\frac{2\lambda^2}{\beta_j^2}}, 2\right) \text{ for j=1,...J+1} \tag{21}$$

4.2.3 Sample λ^2

Draw λ^2 from the conditional inverse gamma prior:

$$InverseGamma\left(shape = \frac{J+1}{2}, rate = \frac{1}{\zeta_\beta} + \frac{1}{2}\sum_{j=1}^{J+1}\frac{\beta_j^2}{\tau_j^2}\right) \tag{22}$$

4.2.4 Sample ζ_β

Draw ζ_β from the conditional inverse gamma prior:

$$InverseGamma\left(1, 1 + \frac{1}{\lambda^2}\right) \tag{23}$$

4.2.5 Sample ξ^2

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

$$IG\left(\sqrt{\frac{2\kappa^2}{\theta_j}}, 2\right) \text{ for j=1,...J+1} \tag{24}$$

4.2.6 Sample κ^2

Draw κ^2 from the conditional gamma prior:

$$InverseGamma\left(shape = \frac{J+1}{2}, rate = \frac{1}{\zeta_{\sqrt{\theta}}} + \frac{1}{2}\sum_{j=1}^{J+1}\frac{\sqrt{\theta}_j^{\,2}}{\xi_j^2}\right) \tag{25}$$

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

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

$$InverseGamma\left(1, 1 + \frac{1}{\kappa^2}\right) \tag{26}$$

4.2.8 Sample σ^2

Draw σ^2 from the posterior distribution:

$$InverseGamma\left(a_1 + \frac{T_0 - 1}{2}, a_2 + \frac{\sum_{t=1}^{T_0-1}\left(Y_{0,t} - \sum_{j=1}^{J+1}\left(\beta_j + \tilde{\beta}_{j,t}\sqrt{\theta_j}\right)X_{j,t}\right)^2}{2}\right) \quad (27)$$

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 $\sum_{j=1}^{J+1}\tau_j^2$ 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.[5] To account for this, **all covariates except the intercept are scaled to mean zero variance one** prior to analysis.

4.3 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:

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

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

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.

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 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. 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 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:

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

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.

5.1 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:

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

with initial conditions:

$$w_{1,0} = .2$$

$$w_{2,0} = .8$$

$$w_{j,0} = 0 \quad j \notin \{1, 2\}$$

5.2 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.

5.3 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.

6 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	$\alpha = 0$	$\alpha = 5$	$\alpha = 0$	$\alpha = 5$
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.

7 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.

8 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).

9 Additional Tables

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 Appendix

10.1 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:

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

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

10.2 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 $\eta_{j,t}$.

10.3 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. The matrix of covariates will still be referred to 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:

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

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:

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

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

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

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

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

10.3.1 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:

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

Combining exponents yields:

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

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

$$D = diag\left[\lambda^2\tau_1^2, ...\lambda^2\tau_{J+1}^2, \kappa^2\xi_1^1, ..., \kappa^2\xi_{J+1}^2\right]_{2(J+1),2(J+1)}$$

.

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:

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

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

10.3.3 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:

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

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 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.

10.3.4 Conditional Distribution of λ^2 and κ^2

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

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

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

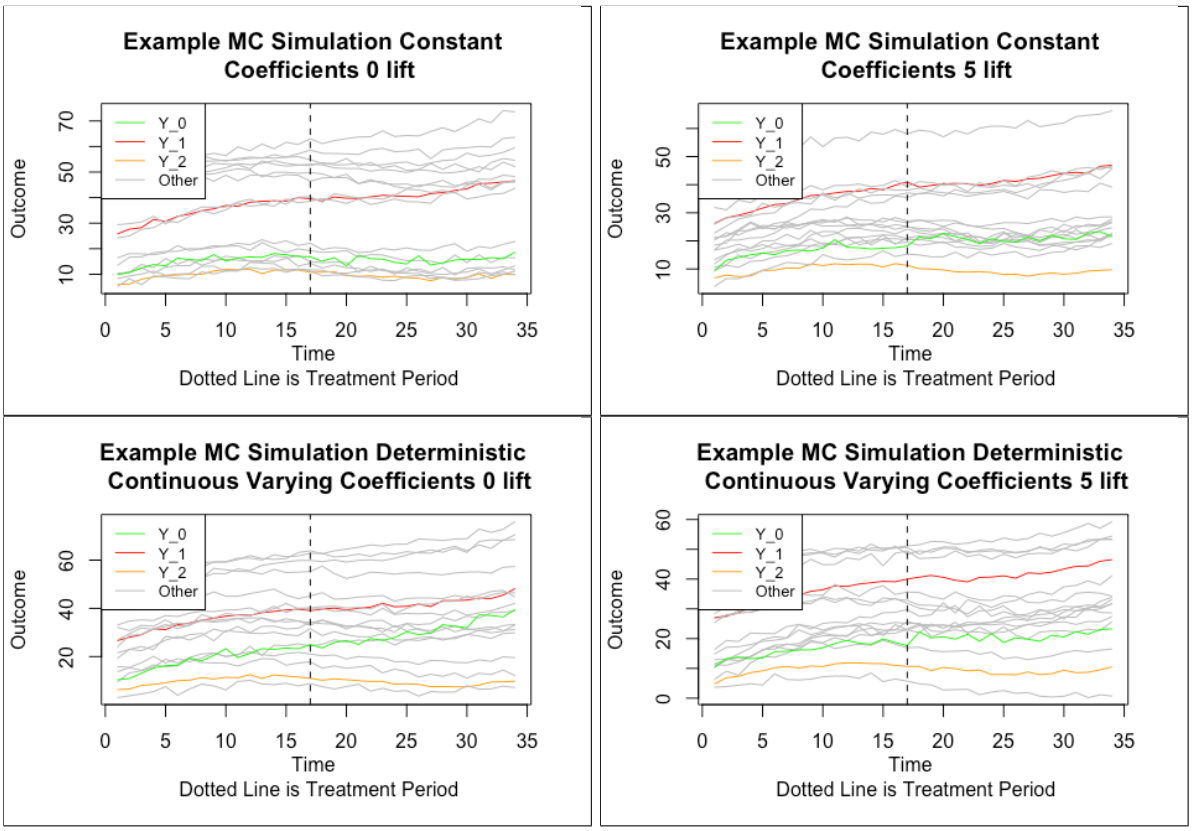
Similarly, κ^2 will follow an inverse gamma distribution with *shape* parameter $\frac{J+1}{2}$ and *rate* parameter $\frac{1}{\zeta_{\sqrt{\theta}}} + \frac{1}{2}\sum_{j=1}^{J+1}\frac{\sqrt{\theta_j}^2}{\xi_j^2}$.

10.3.5 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 $1 + \frac{1}{\kappa^2}$.

10.4 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

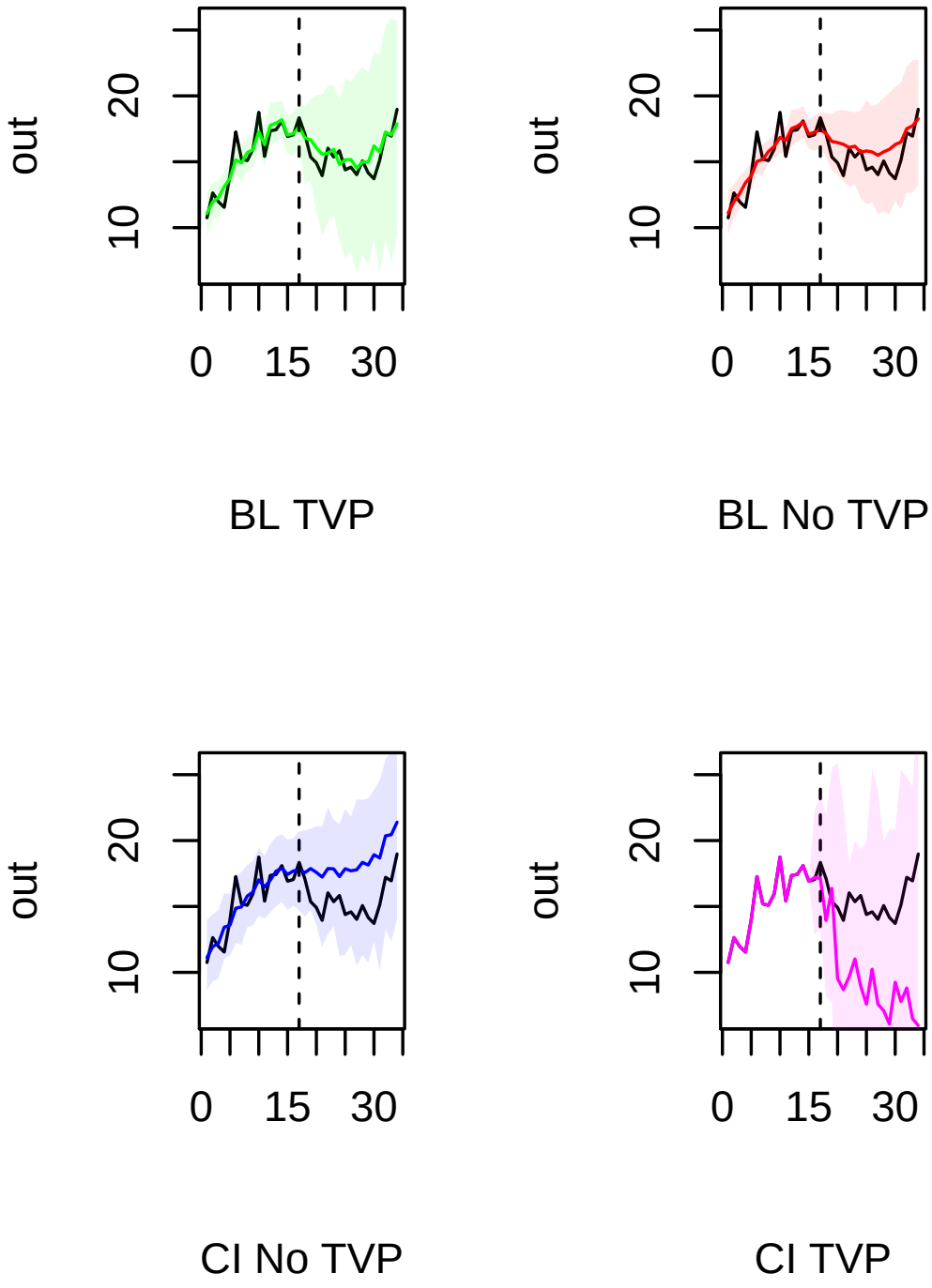


Figure 3: Example Plots Constant Coefficients 5 Treatment Effect

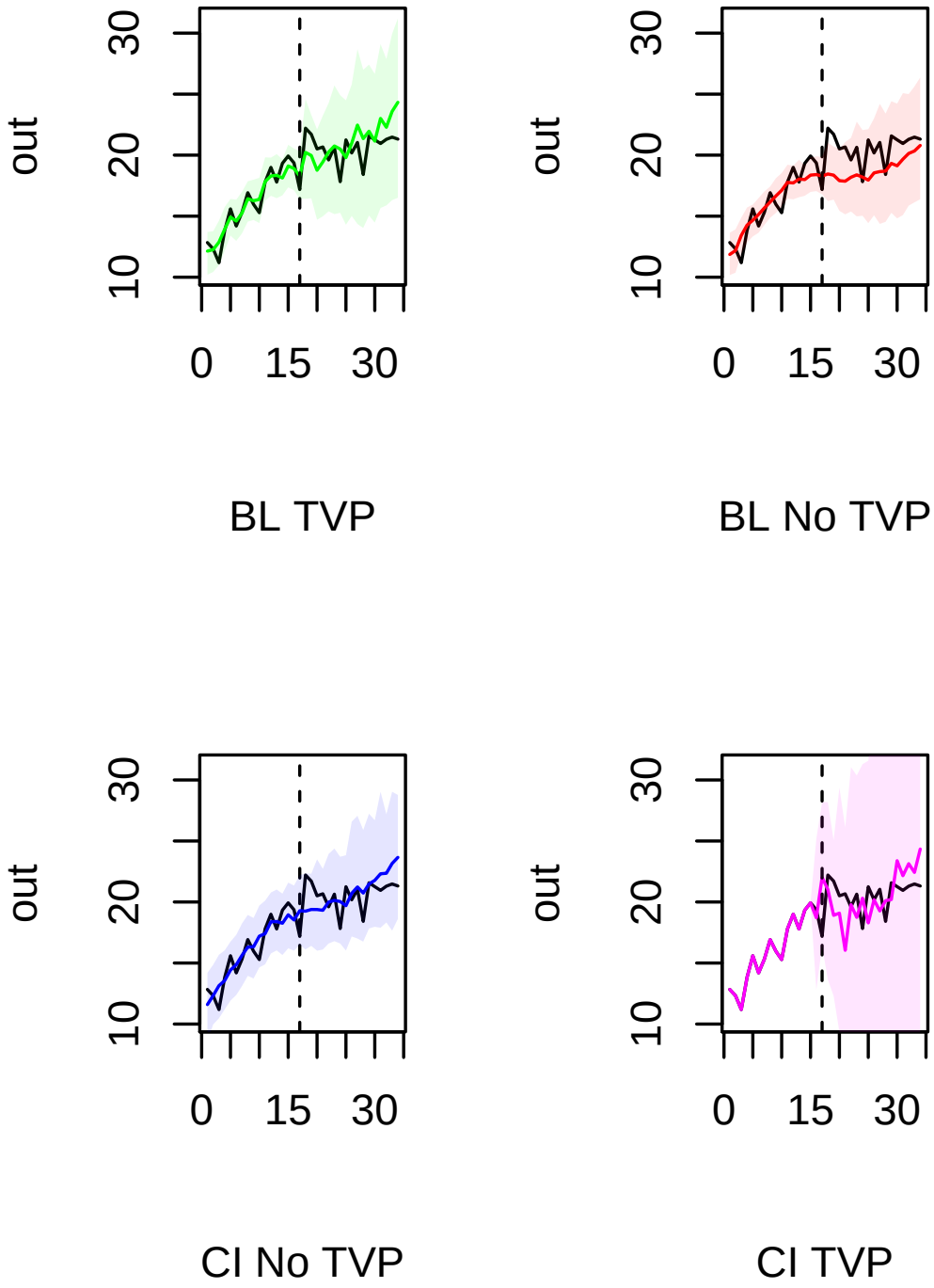


Figure 4: Example Plots Deterministic Continuous Varying Coefficients 0 Treatment Effect

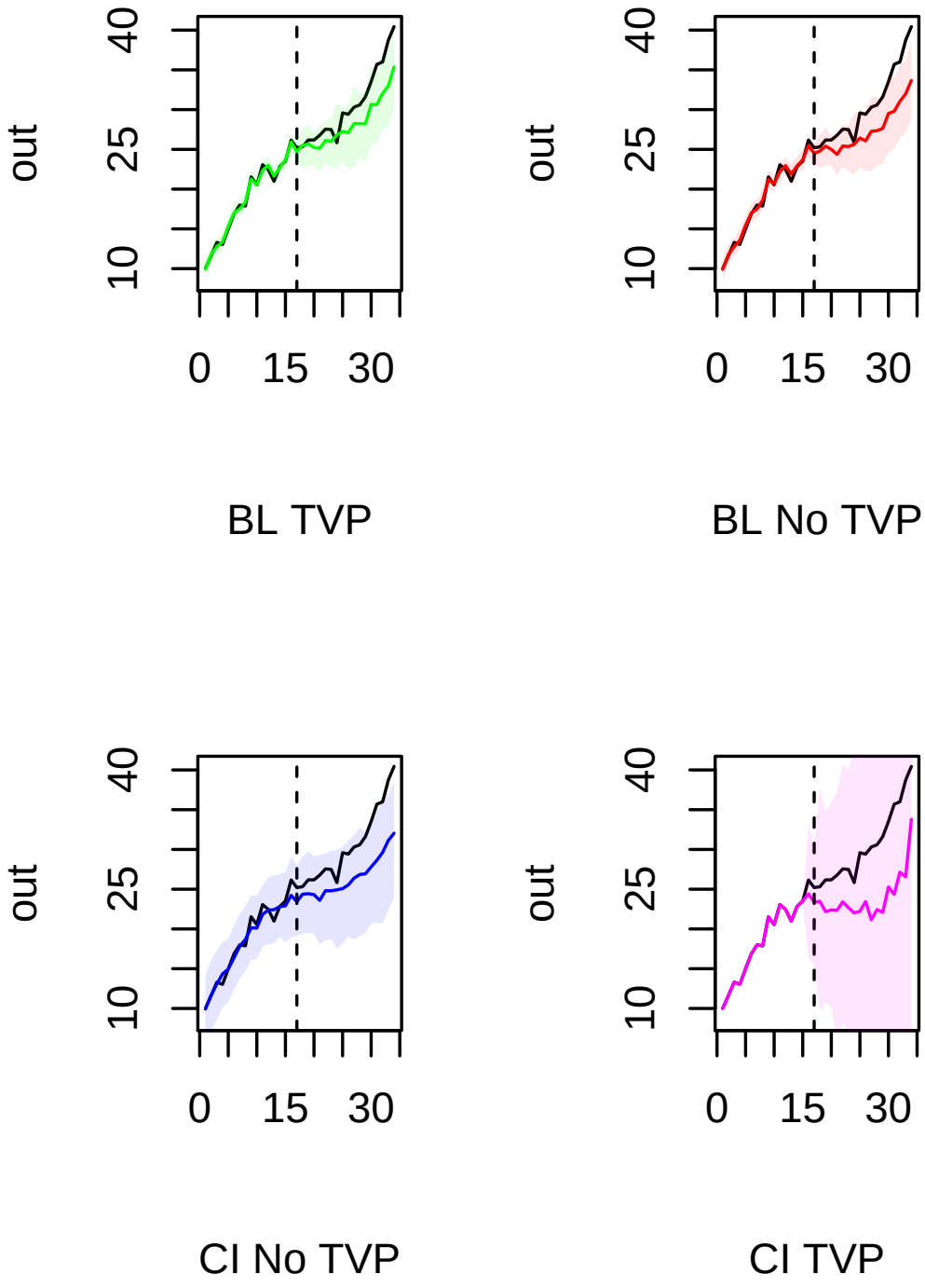
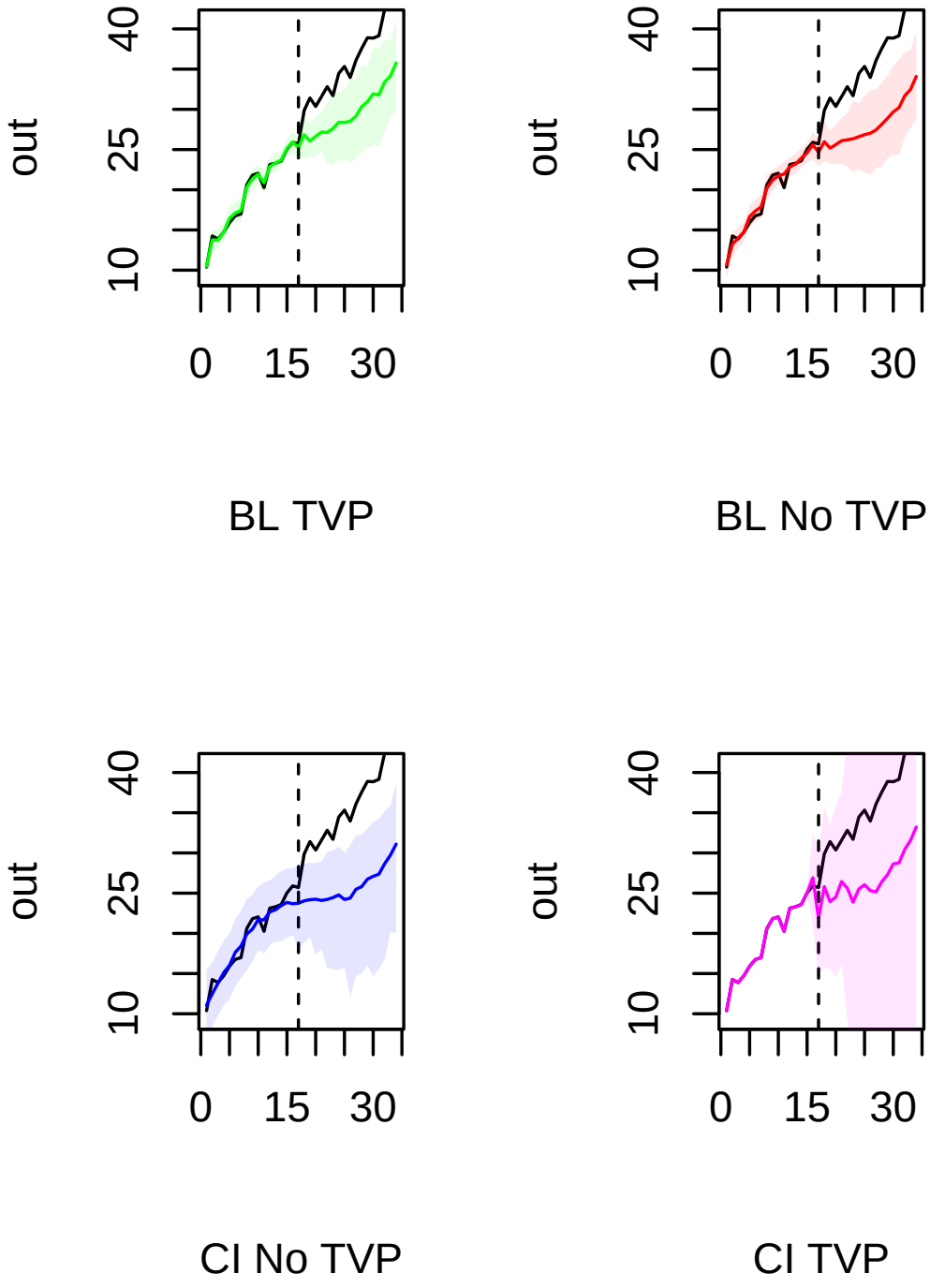


Figure 5: Example Plots Deterministic Continuous Varying Coefficients 5 Treatment Effect



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.

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.

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. 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.

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.

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.

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.

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.