

# Bayesian Synthetic Control Methods

Sungjin Kim, Clarence Lee, and Sachin Gupta 

## Abstract

The authors propose a new Bayesian synthetic control framework to overcome limitations of extant synthetic control methods (SCMs). The proposed Bayesian synthetic control methods (BSCMs) do not impose any restrictive constraints on the parameter space a priori. Moreover, they provide statistical inference in a straightforward manner as well as a natural mechanism to deal with the “large p, small n” and sparsity problems through Markov chain Monte Carlo procedures. Using simulations, the authors find that for a variety of data-generating processes, the proposed BSCMs almost always provide better predictive accuracy and parameter precision than extant SCMs. They demonstrate an application of the proposed BSCMs to a real-world context of a tax imposed on soda sales in Washington state in 2010. As in the simulations, the proposed models outperform extant models, as measured by predictive accuracy in the posttreatment periods. The authors find that the tax led to an increase of 5.7% in retail price and a decrease of 5.5%~5.8% in sales. They also find that retailers in Washington overshifted the tax to consumers, leading to a pass-through rate of approximately 121%.

## Keywords

Bayesian estimation, soda tax, synthetic control, treatment effect

Online supplement: <https://doi.org/10.1177/0022243720936230>

The questions that motivate most studies in marketing are not associational but causal in nature. Although the gold standard for causal inference is a randomized controlled experiment, experiments are often difficult or even impossible to implement for financial, political, and ethical reasons. Therefore, a substantial amount of empirical work in marketing relies on observational data and quasiexperimental methods. Similar to randomized controlled trials, quasiexperimental methods aim to identify the impact of a treatment by comparing treated units with control units. Common examples of quasiexperimental methods include propensity score matching (PSM; Rosenbaum and Rubin 1983), difference in differences (DID; Card and Krueger 1994), and, more recently, synthetic control (Abadie, Diamond, and Hainmueller 2010; Abadie and Gardeazabal 2003).

Because in observational data the assignment of treatment is typically not random, quasiexperimental methods rely on assumptions or unique data characteristics to estimate treatment effects free of confoundedness. Different quasiexperimental methods have varied data requirements and assumptions. For example, PSM uses observed characteristics to estimate the probability (the propensity score) that a unit will receive treatment and matches control units to treated units based on this probability (Rosenbaum and Rubin 1983). A

critical assumption (“conditional independence” assumption) of PSM is that assignment of units to treatment and control groups based on the propensity score is as good as random. Therefore, the efficacy of PSM depends crucially on how well the observed characteristics determine the propensity score. Accordingly, PSM requires extensive cross-sectional data on treated and control units’ observed characteristics. If there are unobserved characteristics that affect assignment to treatment and are not orthogonal to the outcome, then the estimated treatment effects can be biased.

Difference in differences (DID) requires data on outcomes of both treated and control units over time and compares the change in outcomes before and after treatment in the treated unit with the change in outcomes in equivalent time periods in the control unit. This method exploits the fact that we have information on both treated and control groups over time and

---

Sungjin Kim is Assistant Professor of Marketing, Shidler College of Business, University of Hawaii at Manoa, USA (email: [sungjin.kim@hawaii.edu](mailto:sungjin.kim@hawaii.edu)). Clarence Lee is Assistant Professor, Johnson Graduate School of Management, S.C. Johnson College of Business, Cornell University, USA (email: [clarence.lee@cornell.edu](mailto:clarence.lee@cornell.edu)). Sachin Gupta is Henrietta Johnson Louis Professor of Management, Johnson Graduate School of Management, S.C. Johnson College of Business, Cornell University, USA (email: [sachin.gupta@cornell.edu](mailto:sachin.gupta@cornell.edu)).

accounts for unobservable characteristics by taking the difference of within-group, over-time differences. Conventional DID methods, however, rely on the “parallel trends” assumption, which requires that, in the absence of treatment, potential outcomes of treated and control units would have followed parallel paths (e.g., Card and Krueger 1994). To fulfill this assumption, DID methods leave the choice of comparison units to the analyst, prompting ambiguity about how comparison units are chosen (Abadie, Diamond, and Hainmueller 2011).

The synthetic control method (SCM) proposed by Abadie and Gardeazabal (2003) and Abadie, Diamond, and Hainmueller (2010) generalizes the DID methods by providing an explicit data-driven control unit selection procedure. The SCM moves away from using a single control or a simple average of control units; instead, it uses a weighted average of the set of controls as a “synthetic control.” Because of its clarity and simplicity of control unit selection, Athey and Imbens (2017, p. 9) stated that synthetic control is “arguably the most important innovation in the policy evaluation literature in the last 15 years.” In the marketing literature, SCM has been applied to understand the impact of offline television advertising on multiple dimensions of online chatter (Tirunillai and Tellis 2017), to examine how consumers’ search and click activities changed when search engines banned foreign pharmacies from sponsored search (Chesnes, Dai, and Jin 2017), and to investigate the effect of a soda tax on firms’ and consumers’ behaviors in Berkeley, California (Bollinger and Sexton 2018; Rojas and Wang 2017).

The literature has argued that the standard SCM proposed by Abadie and Gardeazabal (2003) and Abadie, Diamond, and Hainmueller (2010) has at least three limitations: (1) restrictive constraints, (2) no inference theory, and (3) lack of an explicit mechanism for the “large  $p$ , small  $n$ ” problem. The first limitation is that the constraints in the standard SCM—namely, nonnegativity of weights, summing-to-one of weights, and no intercept—hinder accommodating various data patterns. Doudchenko and Imbens (2016) argued that none of the SCM constraints on the parameters is likely to hold in practice, and that there are many settings in which allowing negative weights and nonzero intercept would be desirable. Hsiao, Ching, and Wan (2012) dealt with this issue by proposing an alternative panel data method that relaxes all SCM constraints. Moreover, Ferman and Pinto (2016) and Carvalho, Masini, and Medeiros (2018) pointed out that the restriction to convex combinations of the control units biases the SCM estimator.

The second and more critical limitation is that there is no formal inference theory in SCM (Li 2019). Instead, Abadie, Diamond, and Hainmueller (2010) and most of the studies that followed (e.g., Abadie, Diamond, and Hainmueller 2015) adopted a form of permutation test, also known as a “placebo test,” for statistical inference. This test involves applying the SCM to control units instead of the treated unit to answer the question: how often would we get the “unusually” large estimated treatment effect by chance under the null hypothesis of no treatment effect? Hahn and Shi (2017) questioned the validity of this permutation test in the SCM context and concluded that the “permutation test in its current form is likely to fail and cannot

serve as a proper tool for inference with the synthetic control method” because the symmetry assumption, one of the crucial conditions for the validity of the permutation test, is violated. Li (2019) proposed an alternative statistical inference approach for SCM by deriving the asymptotic distributions based on subsampling methods. Firpo and Possebom (2018) proposed another method to calculate the confidence intervals of SCM estimates based on Fisher’s exact hypothesis testing procedure. Although these studies have proposed alternative approaches to inference for the SCM, so far there is no consensus on the best method.

The third limitation of the standard SCM is the lack of an explicit mechanism to deal with the “large  $p$ , small  $n$ ” and/or sparsity problems. The “large  $p$ , small  $n$ ” problem refers to a situation where the number of covariates exceeds the number of data points. In this case, there is typically no unique solution and therefore estimation and prediction are infeasible unless the parameter space is reduced through regularization or dimension-reduction methods. The sparsity problem, also known as a “large  $p$ ” problem, occurs when only a small number of covariates is relevant in a large available set of covariates. In this case, it is challenging to discover which covariates are important (i.e., are useful signals) for prediction. The constraints of the standard SCM, such as sum-to-one and nonnegativity, may help in regularizing the parameter space by forcing coefficients to be within prespecified ranges. However, if our objective is accurate prediction of the potential outcome of the treated unit under no treatment, it is not obvious why the synthetic control should be a convex combination of control units. Therefore, a more systematic way of regularizing parameters in SCM is needed.

In the SCM context, Doudchenko and Imbens (2016) and Carvalho, Masini, and Medeiros (2018) proposed the use of frequentist regularized regression methods—elastic net and lasso, respectively—to overcome this limitation.<sup>1</sup> However, a critical disadvantage of frequentist regularized regression methods is that they do not provide valid statistical inference (i.e., the second SCM limitation). Mallick and Yi (2013) pointed out that frequentist penalized regression is essentially an optimization problem that provides a point estimate of coefficients but not the confidence intervals of estimates. In principle, the confidence intervals can be obtained using bootstrap, but in practice this is difficult due to the adaptive nature of the estimation procedure (i.e., using cross-validation to estimate the penalty parameter) (Hastie, Tibshirani, and Wainwright 2015, chapter 6). Knight and Fu (2000) and Kyung et al. (2010) have shown that bootstrap estimates of the standard error of lasso estimates are inconsistent and unstable because bootstrap sampling introduces a bias that does not vanish asymptotically if the true parameters are zero or close to zero. Although some recent studies have made progress on inference theory for frequentist penalized regression methods (Chatterjee and Lahiri 2011,

<sup>1</sup> Carvalho, Masini, and Medeiros (2018) propose statistical inference only for the average treatment effect. However, Fonseca et al. (2018) point out that estimation of the covariance matrix is practically challenging for some choices of robust estimators.

**Table 1.** A Framework for Bayesian Synthetic Control Methods Incorporating Extant SCMs and Two New Methods.

Method	Reference	Analogous Frequentist Model	Limitations of SCM			
			Constraints on Parameter Space	Approach to Inference	“Large p, Small n” Mechanism	Bayesian Prior Specification
1. BSCM-Horseshoe	Current research	—	No	Posterior distribution through MCMC procedure is used to obtain valid uncertainty measure	Bayesian shrinkage priors	Horseshoe prior
2. BSCM spike and slab	Current research	—	No	Posterior distribution through MCMC procedure is used to obtain valid uncertainty measure	Bayesian shrinkage priors	Spike-and-slab prior
3. Standard SCM	Abadie, Diamond, and Hainmueller (2010)	Constrained regression	Yes	Placebo test, no formal theory	Constraints on parameter space	Uninformative prior with parameter constraints
4. Panel data method	Hsiao, Ching, and Wan (2012)	Linear regression	No	Asymptotic theory	No mechanism	Uninformative prior without parameter constraints
5. Ridge regression	Ben-Michael, Feller, and Rothstein (2018)	Ridge regression	No	Bootstrap with cross validation	Penalty term(s)	Normal prior
6. ArCo model	Carvalho, Masini, and Medeiros (2018)	Lasso regression	No	Bootstrap with cross validation	Penalty term(s)	Laplace prior
7. Modified SCM	Doudchenko and Imbens (2016)	Elastic net	No	Bootstrap with cross validation	Penalty term(s)	Scale mixture of normal priors

2013), there is no “default” approach that provides valid confidence intervals. In summary, although the literature has dealt with some of the individual limitations of SCMs, no single approach has resolved all of them concurrently.

In this research, we make two contributions to the literature. First, we propose two pure Bayesian synthetic control methods (BSCMs; see Table 1, rows 1 and 2) for which frequentist analogs do not exist. The BSCM horseshoe uses the horseshoe shrinkage prior (Carvalho, Polson, and Scott 2010; Piironen and Vehtari 2017), while the BSCM spike and slab uses Bayesian discrete mixtures (George and McCulloch 1993; Mitchell and Beauchamp 1988). As we describe in detail in the “Modeling” section, our proposed models do not suffer from the three limitations of the standard SCM. By relaxing all restrictive constraints on the parameter space, our proposed models accommodate various data patterns. Moreover, our Bayesian methods provide exact statistical inference (e.g., credible intervals) through a Markov chain Monte Carlo (MCMC) procedure even when the sample size is small, and have a natural mechanism to deal with the “large p, small n” problem through shrinkage priors. Second, our Bayesian framework incorporates extant SCM approaches by specifying particular prior distributions on weights. In rows 3–7 of Table 1, we show how the choice of Bayesian prior specifications leads to analogs of the following frequentist SCM methods: row 3: standard SCM, which is a constrained

regression model (Abadie, Diamond, and Hainmueller 2010); row 4: the panel data model of Hsiao, Ching, and Wan (2012), which is a linear regression; row 5: the ridge regression model that is explored by Ben-Michael, Feller, and Rothstein (2018); row 6: the ArCo model of Carvalho, Masini, and Medeiros (2018), which is a lasso regression; and row 7: the modified SCM, which is an elastic net model (Doudchenko and Imbens 2016). Importantly, the Bayesian analogs shown in rows 3–7 do not suffer from the second limitation of the standard SCM; namely, they do provide statistical inference.

We apply our proposed methods to understand the causal effect of the imposition of a soda tax in the state of Washington on prices and consumption of soda. Rojas and Wang (2017) appear to be the first to study a soda tax in Washington. Washington levied a 1/6 cent per ounce tax on both regular and diet soda, starting July 1, 2010. However, the tax was repealed on December 1 of the same year mainly due to lobbying efforts of the beverage industry. Using our proposed BSCMs, we find that the 1/6 cent per ounce tax on soda caused a 5.5%–5.8% decrease in sales and 5.7% increase in price. These percentages translate to retail price increases of more than .2 cents per ounce, which is greater than the tax of 1/6 cents per ounce. Accordingly, we estimate the price elasticity of soda from our two proposed models to be .96 and 1.05, and the retail tax pass-through rate to be 121% on average. We find

that although consumption of soda decreased significantly during the five months when the tax was in effect, it recovered quickly after the tax was removed. Our results are different from those of Rojas and Wang (2017), who used DID to analyze the effect, which yielded less precise estimates. They found that the Washington soda tax caused a statistically insignificant sales decrease and a retail price increase of .18 cents per ounce, which translated to a 105% pass-through rate.

The remainder of the article is organized as follows. In the next section, we discuss extant SCMs and our proposed BSCMs. We then evaluate the performance of the proposed methods as well as extant models in several simulation studies. In the empirical application section, we apply the proposed methods to the Washington tax event. Finally, we conclude with a summary of findings and discussion of limitations.

## Model

We first present the extant (frequentist) SCM approaches. We then introduce how our Bayesian framework can incorporate extant approaches through specifications of priors. Next, we introduce our two Bayesian synthetic control methods which have added advantages over other Bayesian models. Then, we explain how our framework can include covariates other than outcome values and why we included only outcome variables. Finally, we discuss three other Bayesian prior specifications to test the robustness of the results.

## Extant SCM Approaches

### *The Standard SCM and Panel Data Method of Hsiao, Ching, and Wan (2012)*

The standard SCM of Abadie and Gardeazabal (2003) and Abadie, Diamond, and Hainmueller (2010) considers a single treated unit (e.g., the state of Washington) with multiple control units (e.g., each of the other states in the United States). Without loss of generality, suppose we observe  $J + 1$  units over  $t = 1, \dots, T$  periods, with unit 0 being treated and units  $\{1, \dots, J\}$  being untreated. A “treatment” occurs at period  $T_0 + 1$  and remains in place subsequently, where  $T_0$  is the number of pretreatment periods, with  $1 \leq T_0 < T$ . Let  $Y_{jt}^I$  and  $Y_{jt}^N$  be the potential outcomes for unit  $j$  at time  $t$  with and without treatment, respectively.<sup>2</sup> To identify the treatment effect, Abadie, Diamond, and Hainmueller (2010) assume that the treatment has no effect on the outcome before treatment, and therefore  $Y_{jt}^I = Y_{jt}^N$  for all  $j$  at time  $t < T_0 + 1$ . Let  $\alpha_{jt} = Y_{jt}^I - Y_{jt}^N$  be the effect of the treatment for unit  $j$  and time  $t$ . The observed outcome for unit  $j$  at time  $t$  is

$$Y_{jt} = Y_{jt}^N + \alpha_{jt}D_{jt}, \quad (1)$$

where  $D_{jt}$  is an indicator such that  $D_{jt} = 1$  if unit  $j$  is under treatment at time  $t > T_0$ , and  $D_{jt} = 0$  otherwise. We want to estimate the effect of the treatment for treated unit 0 for  $t > T_0$  as follows:

$$\alpha_{0t} = \underbrace{Y_{0t}^I}_{\text{observed}} - \underbrace{Y_{0t}^N}_{\text{unobserved}}. \quad (2)$$

Note that we do not simultaneously observe  $Y_{0t}^I$  and  $Y_{0t}^N$ . Therefore, the challenge lies in estimating the counterfactual potential outcome of the treated unit under no treatment. Causal inference via SCM depends critically on the validity of the counterfactual prediction in the treated group, which is proxied by the out-of-sample predictive ability of the model.

Abadie and Gardeazabal (2003) and Abadie, Diamond, and Hainmueller (2010, 2015) define a synthetic control unit as a weighted average of control units, which is meant to approximate the counterfactual outcome for the treated unit under no treatment. Specifically, a synthetic control unit can be represented by a  $J \times 1$  vector of weights  $\beta_j = (\beta_1, \dots, \beta_J)'$  such that the weighted average of the control units is closest to the treated unit in terms of outcomes in the pretreatment periods. Let  $Y_{0t}$  be a  $T_0 \times 1$  vector that contains the values of the pretreatment outcomes of the treated unit, and let  $Y_{jt}$  be a  $T_0 \times J$  matrix that collects the outcome values of the control units. Then, one can select  $\hat{\beta}$  as the solution to the following constrained minimization problem:

$$\hat{\beta} = \arg \min_{\beta \in \Lambda} \sum_{t=1}^{T_0} \left( Y_{0t} - \beta_0 - \sum_{j=1}^J \beta_j Y_{jt} \right)^2, \quad (3)$$

where  $\beta_0$  is an intercept and  $\Lambda$  represents the constraints imposed on  $\beta$  such that

$$\Lambda = \left\{ \beta \in \mathbb{R}^{J+1} : \beta_0 = 0, \beta_j \geq 0 \text{ for } j = 1, \dots, J \text{ and } \sum_{j=1}^J \beta_j = 1 \right\}. \quad (4)$$

Having estimated  $\hat{\beta}$  using Equations 3 and 4, the treatment effect for period  $t > T_0$  is

$$\hat{\alpha}_{0t} = Y_{0t}^I - \sum_{j=1}^J \hat{\beta}_j Y_{jt}. \quad (5)$$

Hsiao, Ching, and Wan (2012) proposed a panel data method that is based on Equation 3 without the restrictions in Equation 4, as follows:

$$\hat{\beta}_{\text{panel}} = \arg \min_{\beta} \sum_{t=1}^{T_0} \left( Y_{0t} - \beta_0 - \sum_{j=1}^J \beta_j Y_{jt} \right)^2. \quad (6)$$

We refer the reader to the comparison between the standard SCM and Hsiao, Ching, and Wan's (2012) method discussed by Gardeazabal and Vega (2017). Although the panel data method provides statistical inference based on asymptotic

<sup>2</sup> To remain consistent with the notation in Abadie, Diamond, and Hainmueller (2010), we use superscript I for intervention or treatment and N for no treatment.

theory, it does not have a mechanism to deal with the “large  $p$ , small  $n$ ” problem. Instead, Hsiao, Ching, and Wan (2012) relied on institutional knowledge to regularize the parameters in their empirical application.

### Regularized Regression Methods

More recently, researchers have proposed frequentist regularized regression methods such as ridge, lasso, and elastic net for SCM (Ben-Michael, Feller, and Rothstein 2018; Carvalho, Masini, and Medeiros 2018; Doudchenko and Imbens 2016). Doudchenko and Imbens (2016) argue that none of the constraints on the parameters in Equation 4 is likely to hold in practice. In particular, they point out that the no-intercept constraint ( $\beta_0 = 0$ ) rules out the possibility of a systematic difference between the treated unit and a synthetic control unit. The sum-to-one restriction ( $\sum_{j=1}^J \beta_j = 1$ ) can help find a unique set of weights when there are relatively many control units. However, these two restrictions are implausible if a treated unit is on the extreme end of the distribution of units (i.e., it is the largest or smallest in terms of outcome values). Carvalho, Masini, and Medeiros (2018) argue that the nonnegativity constraint ( $\beta_j \geq 0$ ) is strong because it implicitly assumes positive correlation between outcomes of the treated unit and control units. Although the nonnegativity constraint may help regularize the number of significant weights, we do not want to assume a nonnegative relationship a priori but instead let the data inform these relationships.

Doudchenko and Imbens (2016) propose the use of the (frequentist) elastic net method that does not incorporate the restrictions in Equation 4 that the standard SCM imposes. Instead, the elastic net regularizes the estimators through penalty terms (Zou and Hastie 2005):

$$\hat{\beta} = \arg \min_{\beta} \left\{ \sum_{t=1}^{T_0} \left( Y_{0t} - \beta_0 - \sum_{j=1}^J \beta_j Y_{jt} \right)^2 + \lambda_1 \sum_{j=1}^J |\beta_j| + \lambda_2 \sum_{j=1}^J \beta_j^2 \right\}, \quad (7)$$

where nonnegative  $\lambda_1$  and  $\lambda_2$  are penalty parameters. The penalty terms discourage complex models primarily to avoid overfitting. Put differently, regularized regression methods improve predictive accuracy through penalty terms that allow some bias in the estimates (i.e., bias-variance trade-off). As Equation 7 indicates, the elastic net linearly combines the  $L_1$  penalty of the least absolute shrinkage and selection operator (lasso) and the  $L_2$  penalty of ridge. In other words, Equation 7 becomes a ridge regression by eliminating the first penalty term (i.e.,  $\lambda_1 \sum_{j=1}^J |\beta_j|$ ) which was explored by Ben-Michael, Feller, and Rothstein (2018), and it becomes a lasso regression by eliminating the second penalty term (i.e.,  $\lambda_2 \sum_{j=1}^J \beta_j^2$ ), which has been termed the ArCo model in the literature (Carvalho, Masini, and Medeiros 2018). Frequentist regularized regression methods only compute point estimates of the regression coefficients. Statistical inference on the estimates is difficult and often requires various kinds of asymptotic approximations.

This limitation is particularly crucial in the SCM context because we are interested in knowing whether the estimated treatment effect is statistically significant.

## Bayesian Formulations of Extant SCM

### The Standard SCM and Panel Data Methods

We specify and estimate Bayesian formulations of the standard SCM model shown in Equation 3, with constraints in Equation 4, using uninformative priors as follows:

$$\beta_j | \sigma \sim \text{uniform}(-\infty, \infty) \text{ for } j = 1, \dots, J, \quad (8a)$$

$$\sigma \sim \text{uniform}(0, \infty), \quad (8b)$$

$$\text{s.t. } \beta_0 = 0, \beta_j \geq 0, \text{ and } \sum_{j=1}^J \beta_j = 1. \quad (8c)$$

Similarly, we specify the panel data model of Hsiao, Ching, and Wan (2012) in a Bayesian fashion by replacing Equation 8c with  $\beta_0 \sim \text{uniform}(-\infty, \infty)$ . The equivalence between the frequentist linear model and its Bayesian counterpart is well-known (see, e.g., Gelman et al. 2013; McElreath 2015). With a uniform prior, ordinary least squares is equivalent to maximizing the posterior probability or maximizing the likelihood. Therefore, our Bayesian specifications of standard SCM and panel data methods are equivalent to their frequentist counterparts.<sup>3</sup>

### Bayesian Regularized Regression Methods

The mathematical and empirical equivalence between frequentist ridge and lasso and their Bayesian counterparts is also well known (Hsiang 1975; Park and Casella 2008). For example, Park and Casella (2008) have shown empirically that “results from the Bayesian lasso are strikingly similar to those from the ordinary lasso.” We have included sketch proofs of the equivalence of these methods in Web Appendix A.

From a Bayesian lens, ridge regression identifies the mode of the posterior distribution if normal priors with mean 0 and variance  $\sigma^2/\lambda$  are placed on each coefficient (Hsiang 1975):

$$\beta_j | \lambda, \sigma \sim \text{Normal}\left(0, \frac{\sigma^2}{\lambda}\right) \text{ for } j = 1, \dots, J,$$

$$\lambda \sim \text{Cauchy}^+(0, 10),$$

$$\sigma \sim \text{Cauchy}^+(0, 10),$$

$$\beta_0 \sim \text{Cauchy}(0, 10), \quad (9)$$

where  $\sigma^2$  is the variance of the error term. The “penalty” parameter  $\lambda$  determines the amount of shrinkage, with larger values resulting in more shrinkage of the coefficients toward zero. Note that we specify parameters  $\beta_0$  and  $\sigma$  to follow Cauchy and half-

<sup>3</sup> We tested other uninformative but proper priors such as  $N(0, 10,000)$  instead of the uniform distribution and found that our empirical results remained unchanged in simulations, as we discuss subsequently.



Cauchy distributions, respectively, which are known to be “conservative” weakly informative priors (Gelman et al. 2008).

Park and Casella (2008) propose a Bayesian counterpart of the lasso. The Bayesian lasso is obtained by specifying Laplace priors on the regression coefficients:

$$\begin{aligned}\beta_j | \lambda, \sigma &\sim \text{Laplace}\left(0, \frac{\sigma}{\lambda}\right) \text{ for } j = 1, \dots, J, \\ \lambda &\sim \text{Cauchy}^+(0, 10), \\ \sigma &\sim \text{Cauchy}^+(0, 10), \\ \beta_0 &\sim \text{Cauchy}(0, 10).\end{aligned}\quad (10)$$

As we have mentioned, this specification is the Bayesian counterpart of the ArCo model (Carvalho, Masini, and Medeiros 2018), which is an adaptation of SCM.

The equivalence between frequentist elastic net and its Bayesian counterpart needs further investigation, though they are mathematically equivalent (Li and Lin 2010). This is because elastic net has two penalty parameters, and the procedures to estimate these parameters are different between frequentist and Bayesian methods. Specifically, frequentist elastic net estimates the ridge-type penalty term first, followed by the lasso-type penalty term, which introduces extra bias. This problem has been termed “double shrinkage” in the literature (Zou and Hastie 2005). However, Bayesian elastic net estimates the two penalty parameters simultaneously, avoiding “double shrinkage.” Due to this difference, Li and Lin (2010) reported that Bayesian elastic net performs significantly better than its frequentist counterpart for more complicated models.<sup>4</sup> The Bayesian elastic net prior is proposed by Li and Lin (2010) as the following scale mixture of normal distributions:

$$\begin{aligned}\beta_j | \lambda_2, \tau_j, \sigma &\sim \text{Normal}\left[0, \left(\frac{\lambda_2}{\sigma^2} \tau_j - 1\right)^{-1}\right] \text{ for } j = 1, \dots, J, \\ \tau_j | \lambda_1, \lambda_2, \sigma &\sim \text{Truncated Gamma}\left[\frac{1}{2}, \frac{8\lambda_2\sigma^2}{\lambda_1^2}, (1, \infty)\right], \\ \lambda_1 &\sim \text{Cauchy}^+(0, 10), \\ \lambda_2 &\sim \text{Cauchy}^+(0, 10), \\ \sigma &\sim \text{Cauchy}^+(0, 10), \\ \beta_0 &\sim \text{Cauchy}(0, 10).\end{aligned}\quad (11)$$

<sup>4</sup> To confirm this discrepancy in the context of SCM, we generated 100 data sets in each of our simulations S1–S10 described in the “Simulation Study” section, and investigated the performance of the Bayesian elastic net and frequentist elastic net in terms of the equivalence of the two estimates, as well as the bias in parameter estimates (relative to true parameter values) for each. We found that frequentist point estimates and the Bayesian posterior modes are not equal due to the aforementioned “double shrinkage,” but they are close enough to be considered comparable in that 95% highest posterior densities include frequentist point estimates in most cases. Additional details of the analyses are available from the authors on request.

As we have mentioned, this specification is the Bayesian counterpart of the modified SCM (Doudchenko and Imbens 2016).

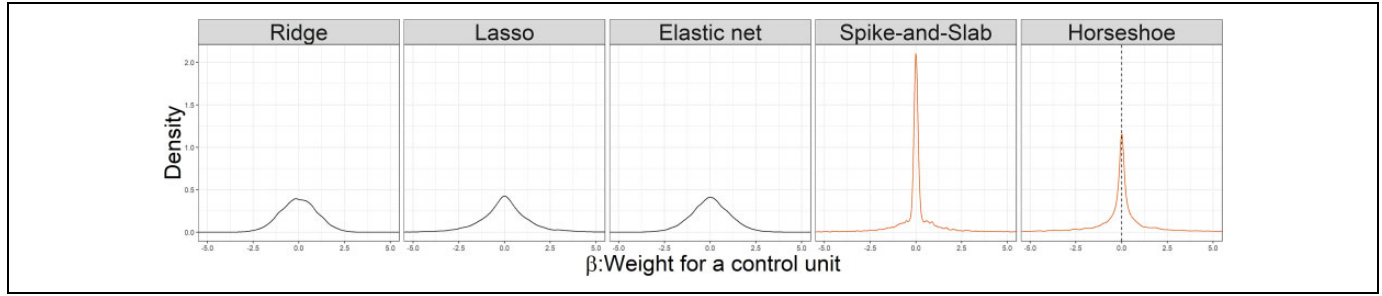
The rationale to use these shrinkage priors is in line with the use of penalty terms in the frequentist regularized regression methods. If a large number of control units is available and included in the model, a higher proportion of in-sample variability in the treated unit outcomes can be explained, but such a specification can suffer from overfitting and poor out-of-sample predictive accuracy. Shrinkage priors discourage overfitting in such high dimensional settings.

## Proposed Bayesian Synthetic Control Methods

Bayesian regularized regression methods discussed previously offer several advantages over frequentist methods. First and foremost, Bayesian methods enable exact statistical inference even when the sample size is small (Mallick and Yi 2013). That is, because we have samples from the posterior distribution as a byproduct of the MCMC procedure, we can summarize the posterior in any way that we choose to obtain, for instance, the posterior density, the posterior mean, and the posterior standard deviation. The second advantage is that Bayesian methods use shrinkage priors as a natural mechanism to deal with two issues: the sparsity problem, and the “large p, small n” problem. Third, Bayesian methods enable estimation of the penalty parameters (i.e.,  $\lambda$ ) jointly with the parameters of interest (i.e.,  $\beta_j$ ), so that model selection and parameter estimation are performed concurrently (Hastie, Tibshirani, and Wainwright 2015). As we have discussed, this advantage is especially relevant when there is more than one penalty parameter because Bayesian methods provide a natural way to avoid the double shrinkage problem by estimating all penalty parameters  $\lambda$  and coefficients  $\beta_j$  simultaneously (Li and Lin 2010).

In this article, we propose BSCM horseshoe and BSCM spike and slab, which are purely Bayesian ways of regularizing parameters; for these models, frequentist counterparts are not known. Our proposed models have the added advantage of better predictive accuracy in addition to the aforementioned advantages of all Bayesian regularized regression methods. As highlighted in Figure 1, the horseshoe and spike-and-slab densities are quite distinct from those of Bayesian ridge, lasso, and elastic net priors. Specifically, horseshoe and spike-and-slab densities show tall spikes around zero (i.e., noise) and flat tails for nonzero parameter values (i.e., signals), which provide accurate inferences about whether each  $\beta_j$  is signal or noise.<sup>5</sup> However, other priors, such as the Laplace (i.e., Equation 10), “compromise between shrinking noise and flagging signals” (Carvalho, Polson, and Scott 2010). The ability to clearly distinguish noise from signal is especially desirable in the SCM context where the “large p, small n” and/or sparsity problems are common concerns. Moreover, this property

<sup>5</sup> The priors of the BSCM horseshoe and BSCM spike-and-slab models are proper. Therefore, their posterior distributions are proper.



**Figure 1.** Densities of shrinkage priors.

leads to better predictive accuracy, as we further discuss in the “Simulation Study” section.

Carvalho, Polson, and Scott (2010) propose the following horseshoe prior that shares desirable properties of spike and slab, such as tall spikes and flat tails:

$$\begin{aligned} \beta_j | \lambda_j &\sim \text{Normal}\left(0, \lambda_j^2\right) \text{ for } j = 1, \dots, J, \\ \lambda_j | \tau &\sim \text{Cauchy}^+(0, \tau), \\ \tau | \sigma &\sim \text{Cauchy}^+(0, \sigma), \\ \sigma &\sim \text{Cauchy}^+(0, 10). \end{aligned} \quad (12)$$

$\tau$  is a global shrinkage parameter that provides severe shrinkage for all the parameters toward zero, while the local shrinkage parameter  $\lambda_j$  allows some  $\beta_j$ s to escape the severe shrinkage by providing heavy tails. Different levels of sparsity can be accommodated via the values of  $\tau$ : large values of  $\tau$  lead to diffuse priors with little shrinkage for all the coefficients, whereas values of  $\tau$  close to zero will shrink all the coefficients  $\beta_j$  to zero. A distinct characteristic of the horseshoe is clear separation between the global and local shrinkage effects that is not possible under the Bayesian ridge, lasso, and elastic net models. By contrast, the horseshoe prior yields estimates that are robust both to unknown sparsity as well as to large signals. Carvalho, Polson, and Scott (2010) suggest that the horseshoe should in general have faster computational times than spike and slab while showing similar predictive accuracy.

The spike-and-slab prior is a discrete mixture of a tall spike around zero (i.e., the spike) to determine whether the coefficient is close to zero, and flat tails for nonzero components (i.e., the slab) to correctly estimate signals (George and McCulloch 1993; Mitchell and Beauchamp 1988). The spike and slab is considered the “gold standard” for sparse Bayesian estimation both theoretically and empirically because of its clear mechanism that distinguishes signals from noise (Carvalho, Polson, and Scott 2010; Piironen and Vehtari 2017). Specifically, the spike-and-slab prior can be written as follows (George and McCulloch 1993):

$$\begin{aligned} \beta_j | \gamma_j, \tau_j, \phi_j &\sim (\gamma_j) \cdot \text{Normal}\left(0, \tau_j^2\right) + (1 - \gamma_j) \cdot \text{Normal}\left(0, \phi_j^2\right) \text{ for } j = 1, \dots, J, \\ \gamma_j &\sim \text{uniform}(0, 1), \end{aligned}$$

$$\tau_j^2 \sim \text{Inverse Gamma}\left(\frac{1}{2}, \frac{1}{2}\right), \quad (13)$$

where  $\gamma_j$  is the prior inclusion probability; therefore,  $\beta_j$  is assigned probability  $(1 - \gamma_j)$  of being zero a priori.  $\tau_j$  is given a noninformative prior so that the variance of the slab component is primarily estimated by the data, and  $\phi_j^2$  is fixed to a small number (.001) to generate the spike.<sup>6</sup> The main disadvantage of this model is that estimation can be computationally demanding when there is a large number of covariates due to rapid combinatorial growth of the model space (Carvalho, Polson, and Scott 2010).

## Inclusion of Covariates Other than Outcomes

Our framework can potentially include covariates other than outcome values in a straightforward manner. Suppose  $Y_{jt}$  is the observed outcome for unit  $j$  at time  $t$ , and  $X_{jt}$  are observed covariates other than outcome values. Then a simple extension to incorporate the observed covariates is

$$(\hat{\beta}, \hat{\theta}) = \arg \min_{\beta, \theta} \sum_{t=1}^{T_0} \left( Y_{0t} - \beta_0 - \sum_{j=1}^J \beta_j Y_{jt} - \sum_{j=1}^J \theta_j X_{jt} \right)^2, \quad (14)$$

where  $\theta_j$  are the weights of the observed covariates of unit  $j$ . The priors of  $\theta_j$  can be defined in a similar way to  $\beta_j$ .

Instead of using outcomes of each control unit as separate predictors as in Equation 3, Abadie and Gardeazabal (2003) and Abadie, Diamond, and Hainmueller (2010, 2015) included only some linear combinations of pretreatment outcome values (e.g., the means over pretreatment period outcomes of each control unit) and other characteristics as covariates. Although allowing covariates is possible, we prefer to include only outcomes of each control unit as separate predictors. Recent studies have advocated this idea due to the higher predictive power of outcomes of control units (Doudchenko and Imbens 2016), or due to concerns about overfitting (Powell 2018). Moreover, Kaul et al. (2015) have demonstrated both theoretically and empirically that

<sup>6</sup> Previous literature has suggested setting  $\phi_j^2$  “suitably small” (Ishwaran and Rao 2005) and proposed  $\phi_j^2 = .001$  as a default value (e.g., Van Erp, Oberski, and Mulder 2019), which we have adopted here.

using outcomes of each control unit as separate predictors renders all other covariates irrelevant. Therefore, following these studies, we also use only the outcomes of control units as covariates. However, a fuller investigation of the role of covariates in improving the predictive performance of SCMs is desirable.

## Robustness Checks

A common concern when using Bayesian estimation is the robustness of results across different specifications of priors and hyperpriors. In Equations 9–11, we used  $\text{Cauchy}(0, 10)$  for the prior of  $\beta_0$  and assigned  $\text{Cauchy}^+(0, 10)$  for the hyperprior of penalty terms (i.e.,  $\lambda$ ) and  $\sigma$  as “sufficiently weakly informative” priors (Gelman et al. 2008). Here, we discuss the sensitivity of our results to specifications of priors and hyperpriors. We do not change the specifications related to weights  $\beta_j$ , because they are part of the SCM model specification.

The extant literature provides some guidance in choice of priors for this examination. Park and Casella (2008) proposed using a gamma distribution for penalty parameter priors mainly due to conjugacy. Chung et al. (2013) proposed using  $\text{Gamma}(2, \alpha)$  with  $\alpha \rightarrow 0$  as a “default” weakly informative prior that provides some direction but still allows inference to be driven by the data. They used  $\text{Gamma}(2, .1)$  in their simulation studies.

To conduct robustness tests, we consider three cases: (1) more informative priors, (2) less informative priors, and (3) alternative weakly informative priors. We define more informative priors as follows:

$$\begin{aligned}\beta_0 &\sim \text{Cauchy}(\beta_{0\text{true}}, 1), \\ \lambda &\sim \text{Cauchy}^+(0, 10), \\ \sigma &\sim \text{Cauchy}^+(\sigma_{\text{true}}, 1),\end{aligned}\quad (15)$$

where  $\beta_{0\text{true}}$  and  $\sigma_{\text{true}}$  are the true values for  $\beta_0$  and  $\sigma$  that we will use in the data-generating process (DGP) in the simulation studies. In addition, we set the scale parameters to be smaller by a factor of 10 than the scale parameters that we previously used. The smaller scale parameter generates a taller and thinner prior distribution that enhances its informativeness. We retain the prior for  $\lambda$  to be conservative and follow  $\text{Cauchy}^+(0, 10)$  because we do not have any ex ante expectations about this hyperparameter.

We define less informative priors as follows:

$$\begin{aligned}\beta_0 &\sim \text{Cauchy}(0, 10^4), \\ \lambda &\sim \text{Cauchy}^+(0, 10^4), \\ \sigma &\sim \text{Cauchy}^+(0, 10^4).\end{aligned}\quad (16)$$

Here, we increase the variance from 10 in the proposed specification to  $10^4$ , thereby reducing the informativeness of priors.

Following Chung et al.’s (2013) suggestion, we consider the following alternative weakly informative priors:

$$\beta_0 \sim \text{Gamma}(2, 0.1),$$

$$\begin{aligned}\lambda &\sim \text{Gamma}(2, 0.1), \\ \sigma &\sim \text{Gamma}(2, 0.1).\end{aligned}\quad (17)$$

## Simulation Study

We describe Monte Carlo simulation experiments that are designed to assess performance of alternative models under various conditions: (1) whether the DGP satisfies the standard SCM constraints; (2) whether the “large p, small n” problem occurs; (3) whether “large p,” or sparsity, is a concern; and (4) whether there are many relevant control units. We generate simulated data sets that are similar in structure to the state-level weekly Nielsen retail sales data that we use in the “Empirical Application” section.

We estimate the model parameters using an open-source software, Stan, which implements the No U-Turn Sampling algorithm as a default sampler (Carpenter et al. 2017).<sup>7</sup> All estimation is done using four chains and 4,000 iterations; the first 2,000 are used for burn-in. We present leave-one-out cross-validation (LOO-CV) as the measure of predictive performance. Because we use the same data to estimate and validate each model, lower LOO-CV means better predictive accuracy (Veh-tari, Gelman, and Gabry 2017). We also calculated widely applicable information criterion and find that these results are qualitatively the same as for LOO-CV. Therefore, we omit these results for space considerations. We use mean estimation time of four chains on a PC desktop with 4.00 GHz central processing unit as a measure of computational efficiency.

## Simulation Design

We consider four factors in the design of the simulation study. The first factor is the number of pretreatment periods, which is at two levels:  $n = 100$  weeks or  $n = 40$  weeks. The second factor is the number of control units, which is also at two levels:  $p = 50$  or  $p = 5$ . The third factor is the number of relevant control units that “predict” the treated unit: only two control units are relevant for the prediction ( $k = 2$ ), or all control units are relevant for the prediction ( $k = 50$  or  $5$ ). Crossing design factors 1–3 leads to a total of  $2 \times 2 \times 2 = 8$  cases (described in Table 2). Of these eight, three cases are not informative, as we discuss subsequently, leading to five cases that we study for each level of factor 4. The fourth factor is whether or not the DGP satisfies the standard SCM constraints. When the constraints are satisfied, we denote the case “S,” and when the constraints are violated, we denote the case “V.” Simulations S1–S5 pertain to level “S” of factor 4, and simulations S6–S10 pertain to level “V” of factor 4. In all ten simulation cases, we assume that the number of posttreatment periods is equal to the number of pretreatment periods, which is 100 or 40 depending on the simulation scenarios.

The DGP and parameter values in simulation cases S1–S5, where the DGP satisfies the standard SCM constraints (“S”) are as follows:

<sup>7</sup> Estimation code is available in Web Appendix B.



**Table 2.** Summary of Simulation Design.

Factor 2: Total # of Control Units $p$	Factor 3: # of Relevant Weights (Control Units) $k$	Factor 1: Number of Pretreatment Periods ( $n$ )	
		100 weeks	40 weeks
50	2	S1 and S6: Sparsity (large $p$ )	S2 and S7: Large $p$ , small $n$ ( $p = 50 > n = 40$ )
50	50	S3 and S8: MRW	S4 and S9: Large $p$ , small $n$ and MRW
5	2	S5 and S10: No concern	Not informative
5	5	Not informative	Not informative

Notes: MRW = many relevant weights.

$$Y_{0t} = \sum_{j=1}^{p=5 \text{ or } 50} \beta_j Y_{jt} + \epsilon_{0t},$$

$$Y_{jt} \sim \text{Normal}(\mu, 10),$$

$$\epsilon_{0t} \sim \text{Normal}(0, 1), \quad (18)$$

where  $\beta_j$  and  $\mu$  are as follows:

$$\beta_j = \begin{bmatrix} \beta_1 = .2 & \beta_1 = .24 \\ \beta_2 = .8 & \beta_2 = .24 \\ \beta_3 = 0 & \beta_3 = (1 - 24/p)/(p - 2) \\ \beta_4 = 0 & \text{or } \beta_4 = (1 - 24/p)/(p - 2) \\ \beta_5 = 0 & \beta_5 = (1 - 24/p)/(p - 2) \\ \beta_6 = 0 \text{ or N.A.} & \beta_6 = (1 - 24/p)/(p - 2) \\ \vdots & \vdots \\ \underbrace{\beta_p = 0 \text{ or N.A.}}_{\{S1, S2, S5\}} & \underbrace{\beta_p = (1 - 24/p)/(p - 2)}_{\{S3, S4\}} \end{bmatrix}_{p \times 1}, \quad (19)$$

$$\mu = \begin{bmatrix} \mu_1 = 15 \\ \mu_2 = 35 \\ \mu_3 = 10 \\ \mu_4 = 20 \\ \mu_5 = 30 \\ \mu_{6 \sim 12} = 10 \text{ or N.A.} \\ \mu_{13 \sim 21} = 20 \text{ or N.A.} \\ \mu_{22 \sim 30} = 30 \text{ or N.A.} \\ \mu_{31 \sim 40} = 40 \text{ or N.A.} \\ \mu_{41 \sim 50} = 50 \text{ or N.A.} \end{bmatrix}_{p \times 1}.$$

Equation 18 shows that the DGP of the treated unit outcome  $Y_{0t}$  ( $n \times 1$  vector) is the sum of a weighted average of control unit outcomes  $Y_{jt}$  ( $n \times p$  matrix) and the error term  $\epsilon_{0t}$  ( $n \times 1$  vector).  $Y_{jt}$  is normally distributed with mean vector  $\mu$  and a variance of

10. In Equation 19, the first column of  $\beta_j$  shows the weights of the control units in simulation cases S1, S2, and S5; the second column indicates weights in simulation cases S3 and S4. In simulation studies S1, S2, and S5, there are only two relevant control units, whereas in S3 and S4, all control units are relevant. Note that this DGP satisfies all the standard SCM constraints: (1) non-negativity ( $\beta_1 \sim \beta_p \geq 0$ ), (2) sum to one ( $\sum_{j=1}^p \beta_j = 1$ ), and (3) no intercept. Depending on the simulation cases,  $\beta_6 \sim \beta_{50}$  and  $\mu_6 \sim \mu_{50}$  may or may not exist. For instance, when there are only five control units ( $p = 5$ ), these terms do not exist.

The DGP and parameter values in simulation cases S6–S10 when the DGP violates all the standard SCM constraints (“V”) are as follows:

$$Y_{0t} = \beta_0 + \sum_{j=1}^{p=5 \text{ or } 50} \beta_j Y_{jt} + \epsilon_{0t},$$

$$\beta_0 = 5,$$

$$Y_{jt} \sim \text{Normal}(\mu, 10),$$

$$\epsilon_{0t} \sim \text{Normal}(0, 1), \quad (20)$$

$$\beta_j = \begin{bmatrix} \beta_1 = -.5 & \beta_1 = -.5 \\ \beta_2 = 2 & \beta_2 = 2 \\ \beta_3 = 0 & \beta_3 = 1/p \\ \beta_4 = 0 & \text{or } \beta_4 = 1/p \\ \beta_5 = 0 & \beta_5 = 1/p \\ \beta_6 = 0 \text{ or N.A.} & \beta_6 = 1/p \\ \vdots & \vdots \\ \underbrace{\beta_p = 0 \text{ or N.A.}}_{\{S6, S7, S10\}} & \underbrace{\beta_p = 1/p}_{\{S8, S9\}} \end{bmatrix}_{p \times 1}, \quad (21)$$

$$\mu = \begin{bmatrix} \mu_1 = 15 \\ \mu_2 = 35 \\ \mu_3 = 10 \\ \mu_4 = 20 \\ \mu_5 = 30 \\ \mu_{6 \sim 12} = 10 \text{ or N.A.} \\ \mu_{13 \sim 21} = 20 \text{ or N.A.} \\ \mu_{22 \sim 30} = 30 \text{ or N.A.} \\ \mu_{31 \sim 40} = 40 \text{ or N.A.} \\ \mu_{41 \sim 50} = 50 \text{ or N.A.} \end{bmatrix}_{p \times 1}.$$

Note that, in this case, the DGP violates all the standard SCM constraints: (1) nonnegativity ( $\beta_1 < 0$ ), (2) sum to one ( $\sum_{j=1}^p \beta_j \neq 1$ ), and (3) zero intercept ( $\beta_0 = 5$ ).

The ten cases in this simulation can be summarized as follows. We describe S1–S5 first for the situation when the DGP satisfies (“S”) SCM constraints. We describe analogous cases S6–S10 subsequently. In S1 (S-Sparsity), we have  $n = 100$ ,  $p = 50$ , and  $k = 2$ . In this case, only 2 out of 50 parameters are nonzero, and thus “large  $p$ ” or sparsity arises. In S2 (S-Large  $p$ , small  $n$ ), we have  $n = 40$ ,  $p = 50$ ,  $k = 2$ ; thus, we have a “large  $p$ , small  $n$ ” problem. There is no unique solution in this case, and some form of regularization is necessary. The standard SCM handles this problem via its constraints, whereas other methods (except the panel data method) regularize parameters through shrinkage priors. In S3 (S-Many relevant weights), we have  $n = 100$  and  $p = 50$ , which is similar to S1, but all the control units are relevant to predict the treated unit (i.e.,  $k = 50$ ). In this case, our models as well as other Bayesian regularizing methods can potentially overly shrink the relevant (but small in magnitude) parameter estimates toward zero. We consider a realistic case where weights of some control units are more important (e.g.,  $\beta_1 = \beta_2 = .24$ ) than others (e.g.,  $\beta_{3 \sim p} = (1 - 24/p)/(p - 2)$ ). In S4 (S-Large  $p$ , small  $n$ , and Many relevant weights), we have  $n = 40$ ,  $p = 50$ , which is similar to S2, but now all the control units are relevant to predict the treated unit (i.e.,  $k = 50$ ). In this case, the expected performance of our models is ambiguous: on the one hand, our models are expected to perform well by handling “large  $p$ , small  $n$ ” problem better than other models. On the other hand, however, our models may perform worse than other models by overly shrinking relevant parameter estimates toward zero. In S5 (S-No concern), we have  $n = 100$ ,  $p = 5$ , and  $k = 2$ . In this case, we have neither the “large  $p$ , small  $n$ ” nor “overshrinkage” concerns. Moreover, sparsity is not an issue because only two of five parameters are relevant. Although we expect that in this case all models will perform well, we want to investigate whether shrinkage priors harm predictive accuracy when regularizing is not necessary. Note that we do not include three cases ( $n = 100$ ,  $p = 5$ ,  $k = 5$ ), ( $n = 40$ ,  $p = 5$ ,  $k = 5$ ), ( $n = 40$ ,  $p = 5$ ,  $k = 2$ ) because they are not informative relative to cases S1–S5.

Analogous to cases S1–S5, we define cases S6–S10 for the situation wherein the DGP violates (“V”) all the standard SCM constraints. S6 (V-Sparsity) uses  $n = 100$ ,  $p = 50$ ,  $k = 2$ ; S7 (V-Large  $p$ , small  $n$ ) uses  $n = 40$ ,  $p = 50$ ,  $k = 2$ ; S8 (V-Many relevant weights) uses  $n = 100$ ,  $p = 50$ ,  $k = 50$ ; S9 (V-Large  $p$ , small  $n$  & Many relevant weights) uses  $n = 40$ ,  $p = 50$ ,  $k = 50$ ; and S10 (V-No concern) uses  $n = 100$ ,  $p = 5$ ,  $k = 2$ .

## Simulation Results

In Table 3, we show the predictive accuracy measure, LOO-CV, and computational efficiency for each model for the ten simulation cases. In Figure 2, Panels A and B, we present the percentage difference in the actual outcome of the treated unit and predicted values of the outcome in the synthetic control unit in the posttreatment periods. The results of Bayesian ridge and elastic net are mostly worse than Bayesian lasso, and the results of spike-and-slab model are almost identical to the

results of horseshoe. Therefore, we only show results for standard SCM, panel data method, Bayesian lasso, and horseshoe model for ease of exposition.

The key findings from the simulation studies are as follows. In terms of predictive accuracy (LOO-CV), horseshoe and spike-and-slab models almost always outperform other models, including the standard SCM and other Bayesian regularized regression methods. When the DGP meets all the SCM constraints and when “large  $p$ , small  $n$ ” or “large  $p$ ” concerns exist, the horseshoe and spike-and-slab models have better holdout predictions. The standard SCM shows equivalent performance to the other models only if there is a sufficient number of pretreatment periods and sparsity is not a problem. The panel data method is unusable when “large  $p$ , small  $n$ ” problem exists. When the DGP violates the SCM constraints, the standard SCM is unusable, whereas the proposed models work well. The predictive performance of horseshoe and spike-and-slab specifications is similar, and it does not degrade in conditions where regularization is unnecessary. These results are robust across different prior specifications. We also found that the BSCMs tended to have longer computation times, but the maximum time was only 120.2 seconds (see Table 3a and 3b). In summary, BSCMs are broadly preferable to extant SCM models in that they perform better, or at least equivalently, and provide valid statistical inference in a variety of data situations. Next, we present the findings in greater depth.

### When the DGP Satisfies the Standard SCM Constraints

Recall that in simulation studies S1–S5 the DGP satisfies all the standard SCM constraints. In Table 3a, S1 (S-Sparsity) results show that the horseshoe and spike-and-slab models outperform the other models in terms of predictive accuracy. Note that even though the DGP meets all the standard SCM constraints, the standard SCM shows relatively poor predictive performance. Our models also outperform other Bayesian regularization methods in that the difference in predictive accuracy between BSCMs and the second best model, Bayesian lasso, is statistically significant.<sup>8</sup> This result verifies that the distinctive property of our models that distinguishes signals from noise leads to better predictive accuracy under sparsity.

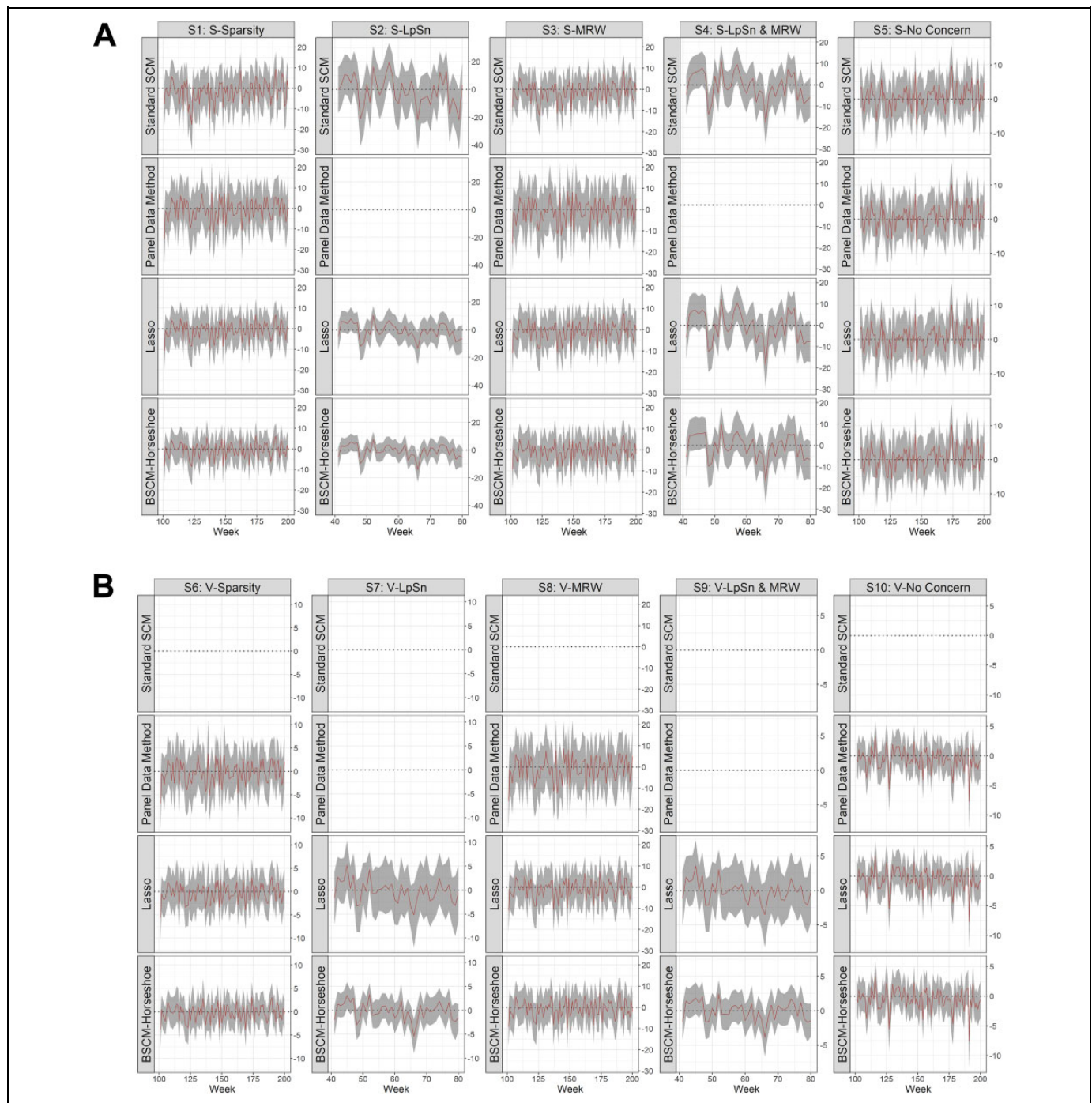
The results for S2 (S-Large  $p$ , small  $n$ ) are similar to those for S1; the horseshoe and spike-and-slab models outperform other models in predictive accuracy. In addition, Figure 2, Panel A, indicates that the predicted values of horseshoe and spike and slab have considerably narrower credible intervals compared to other models. This indicates that horseshoe and spike-and-slab specifications predict well while avoiding overly complex models. Once again, the standard SCM shows especially poor in-sample fit as well as out-sample predictions

<sup>8</sup> Statistical tests for model comparisons using LOO-CV and the widely applicable information criterion can be found in Vehtari, Gelman, and Gabry (2017).

**Table 3.** Predictive Accuracy and Computational Efficiency of Alternative Models in Simulation Studies.

<b>A: S1 ~ S5 (SCM constraints satisfied)</b>										
<b>S1: S-Sparsity</b>			<b>S2: S-Large p, Small n</b>		<b>S3: S-MRW</b>		<b>S4: S-Large p, Small n and MRW</b>		<b>S5: S-No Concern</b>	
LOO-CV	Estimation Time (Sec)		LOO-CV	Estimation Time (Sec)	LOO-CV	Estimation Time (Sec)	LOO-CV	Estimation Time (Sec)	LOO-CV	Estimation Time (Sec)
Standard SCM	362.4	6.8	187.4	3.6	312.8	11.7	126.6	6.8	284.7	.7
Panel data method	382.9	39.3	—	—	382.2	43.1	—	—	289.6	2.1
Ridge	370.9	30.9	160.9	24.0	343.0	28.5	129.7	11.3	289.5	2.5
Lasso	330.7	25.6	130.8	18.1	326.2	22.9	128.5	10.5	288.6	2.9
Elastic net	370.3	49.3	152.9	48.5	343.1	42.1	130.5	29.5	289.3	15.0
BSCM horseshoe	<b>303.9</b>	105.9	<b>114.3</b>	86.6	313.6	82.5	<b>122.5</b>	60.6	287.1	32.1
BSCM spike and slab	318.2	53.9	123.2	62.3	321.0	57.2	124.7	59.2	286.2	3.7
<b>B: S6 ~ S10 (SCM constraints violated)</b>										
<b>S6: V-Sparsity</b>			<b>S7: V-Large p, Small n</b>		<b>S8: V-MRW</b>		<b>S9: V-Large p, Small n and MRW</b>		<b>S10: V-No Concern</b>	
LOO-CV	Estimation Time (Sec)		LOO-CV	Estimation Time (Sec)	LOO-CV	Estimation Time (Sec)	LOO-CV	Estimation Time (Sec)	LOO-CV	Estimation Time (Sec)
Standard SCM	955.3	2.0	394.7	1.6	1,110.6	2.1	447.7	1.5	989.0	.3
Panel data method	382.3	43.5	—	—	384.9	46.9	—	—	271.0	2.2
Ridge	379.2	39.3	177.9	30.6	378.9	40.6	183.1	32.5	270.9	2.5
Lasso	343.6	31.1	136.7	26.9	349.4	32.1	138.8	29.0	270.9	2.5
Elastic net	378.0	65.3	159.1	72.5	376.5	76.9	165.7	64.2	270.5	28.1
BSCM horseshoe	<b>303.0</b>	115.7	<b>110.6</b>	82.4	<b>326.6</b>	120.2	<b>121.8</b>	76.8	270.5	32.8
BSCM spike and slab	317.7	59.5	120.2	52.1	326.6	67.1	126.9	68.3	<b>270.3</b>	3.8

Notes: MRW = many relevant weights. Estimation time is measured as mean estimation time of four chains in seconds. In S2, S4, S7, and S9, the panel data method is not identified and did not converge due to a lack of regularization mechanisms. The lowest LOO-CV values are highlighted in bold.



**Figure 2.** Predictive accuracy of alternative models in simulation studies.

**Notes:** The red line is the percentage difference between the outcomes in the actual treated unit and the predicted mean outcome values of the synthetic control unit in the posttreatment periods. The shaded areas are 95% posterior credible intervals. In S2, S4, S7, and S9, the panel data method did not converge due to a lack of regularization mechanisms. Therefore, we omit these results. In S6 through S10, the standard SCM results are not comparable to other methods due to very large prediction errors because the DGP violates the SCM constraints. We omit these results as well. MRW = many relevant weights; S = satisfied; V = violated; LpSn = large p small n.

compared with proposed models. The panel data method did not converge because it does not have a regularization mechanism to address the “large p, small n” situation.

In S3 (S-Many relevant weights), we consider a data situation where all weights are relevant for the prediction of the

treated unit. We feared that our model may perform worse than other models since our models may overly shrink some relevant weights. We find that the standard SCM performs the best across models in this case. However, confounding our fears, our model performed statistically as well as the standard SCM.



This suggests that our models work well even when there are many “weak signals.”

Results for S4 (S-Large  $p$ , small  $n$  and Many relevant weights) indicate that our models show slightly better predictive accuracy compared with the standard SCM in this challenging data situation. Once again, the panel data method did not converge, because the model cannot handle a “large  $p$ , small  $n$ ” situation.

In S5 (S-No concern), where there are neither “large  $p$ , small  $n$ ” nor sparsity concerns, as we expected, all models performed equivalently in terms of predictive accuracy. The standard SCM performed slightly better than other models, but the difference is not statistically significant. The results in this case provide evidence that the use of shrinkage priors does not hurt predictive accuracy when regularizing is not necessary.

Abadie, Diamond, and Hainmueller (2015) make it clear that the standard SCM is not suitable when the number of pretreatment periods is small. In addition, we have verified via simulation studies S1–S5 that the standard SCM performs poorly when either “large  $p$ , small  $n$ ” or sparsity are potential challenges, while the proposed Bayesian models perform well. Moreover, Figure 2, Panel A, shows that statistical inference in Bayesian methods is straightforward in that we can obtain credible intervals of potential outcomes under no treatment as byproducts of MCMC procedures. This obviates the need for placebo tests, which, as we have discussed, are controversial.

### *When the DGP Violates the Standard SCM Constraints*

We turn next to simulation studies S6–S10 wherein the DGP violates all the standard SCM constraints. Although the standard SCM models converge, their predictive accuracies are not comparable to other models, because they have very large prediction errors and credible intervals. Therefore, we omit the results of the standard SCM in Figure 2, Panel B. In Table 3, Panel B and Figure 2, Panel B, results of S6 through S9 show that the horseshoe and spike-and-slab models again outperform all the other models in terms of predictive accuracy and provide considerably narrower credible intervals. Furthermore, once again the panel data method did not converge, and therefore it is not usable when there is a “large  $p$ , small  $n$ ” problem. In S10, where “large  $p$ ” and/or “small  $n$ ” are not concerns, all models performed equivalently in terms of predictive accuracy except for the standard SCM. To summarize, studies S6–S10 demonstrate that the proposed Bayesian methods are usable even when the standard SCM and panel data method are infeasible. Moreover, as we expected, our models show better predictive accuracy due to their ability to clearly distinguish noise from signal.

### *Robustness Checks*

In Web Appendix C, Tables I.A through III.B show the LOO-CV and computational efficiency of models across different simulation scenarios under more informative, less informative,

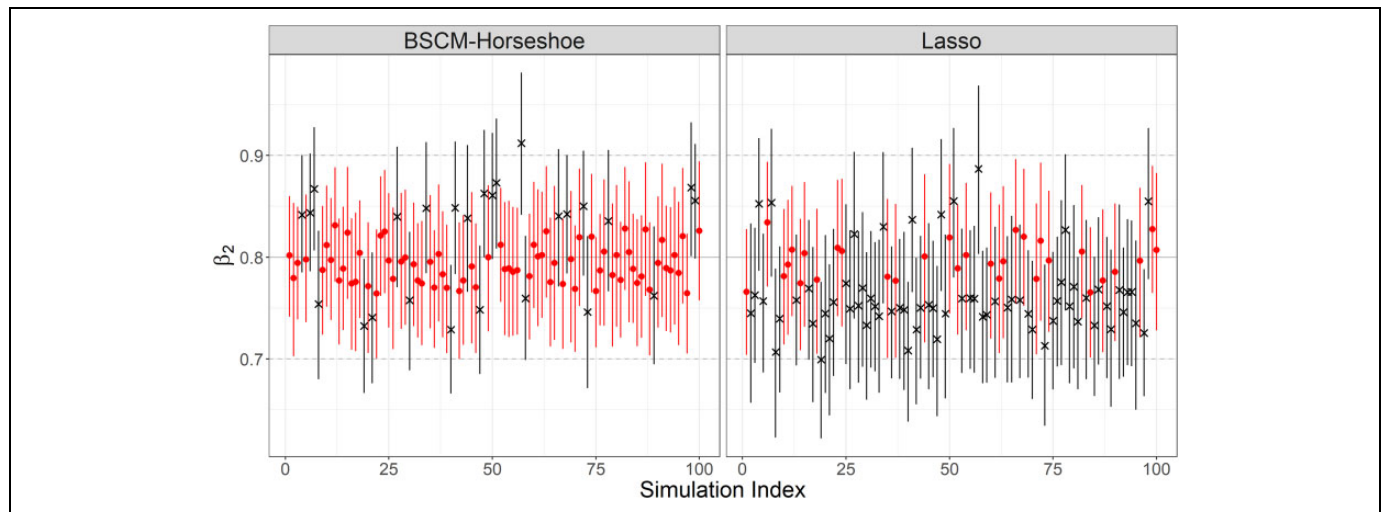
and alternative prior specifications. Not surprisingly, the predictive performance of most models is slightly better when we provide more rather than less informative priors, but the differences are very small. Moreover, the rank orders of models based on predictive performance rarely change across different prior specifications. Therefore, the conclusions based on our simulation studies are robust to specifications of priors.

We also conducted several additional robustness checks. For example, it is possible that outcomes of control units are correlated; thus, we investigate this case by including a correlation among control units. Moreover, unobserved random factors may play a larger role in reality; we examined this by increasing the magnitude of the variance of the error term. Our fundamental conclusions were unaltered in all these checks. Details of these additional analyses are available in Web Appendix C.

### **Precision of Parameter Estimates**

Previously, we have only been concerned with the predictive performance of the overall model because our primary goal is to predict the counterfactual outcomes via the synthetic control unit. Next, we turn our attention to the precision of estimates of individual parameters. In our simulation studies, the sample sizes are fixed and equal to the number of pretreatment periods. We focus, therefore, on the degree of precision in parameter estimates of each model given the fixed sample size. Maxwell, Kelley, and Rausch (2008) emphasize the close link between traditional power analysis and the accuracy in parameter estimation metric. They argue that “sample size planning should sometimes focus on obtaining a sample large enough to have an adequate probability to reject the null hypothesis, whereas other times the focus should be on an adequate probability of obtaining a sufficiently narrow confidence interval” (p. 542). In a frequentist setting, confidence intervals provide a useful framework for simultaneously considering the direction, the magnitude, and the accuracy of an effect (Maxwell, Kelley, and Rausch 2008).

In a Bayesian setting, we can use the posterior distribution to discern the credible values of the parameter (Kruschke 2014). If all the credible values (e.g., the credible intervals) are almost equal to the true value of the parameter, then we can conclude that there is sufficient precision in our model estimates. In particular, following Kruschke (2014), we subjectively define a region of practical equivalence (ROPE) which is a small range of parameter values that are considered to be practically equivalent to the true value. Once a ROPE is set, we can decide to “accept” a true value according to the following rule: “A parameter value is declared to be accepted for practical purposes if that value’s ROPE completely contains the 95% credible interval of the posterior of that parameter” (Kruschke 2014, p. 337). The ROPE limits, by definition, cannot be uniquely “correct” but instead are established by practical considerations, bearing in mind that wider ROPEs yield more decisions to accept the accuracy in parameter estimation. More



**Figure 3.** An example of precision of parameter estimate analysis.

Notes: The y-axis indicates the parameter estimates and the x-axis is the simulation index. Within the figure, the marker indicates the mean and the vertical line represent the 95% credible interval of the parameter estimate. Red (circle) indicates that the credible intervals of the parameter estimate is completely included within the ROPE (horizontal dotted lines at .7 and .9), whereas black (cross) represents the cases where the ROPE does not completely contain the credible intervals of the parameter estimates.

detailed explanations of this procedure of Bayesian power/precision analysis are available in Kruschke (2014).

In Figure 3, we show an example of how precision in parameter estimates can be calculated using the  $\beta_2$  parameter in simulation case S1. We generate 100 data sets in the S1 simulation case, estimate the parameters for each data set, for each model, and show the 100 mean estimates and their 95% credible intervals. Red (circle) lines indicate that the credible interval of the parameter estimate is completely included within the ROPE, whereas black (cross) lines represent data sets where the ROPE does not completely contain the credible intervals of the parameter estimates. For example, if we set the ROPE to be (.7, .9) for the true parameter value of .8, then we can “accept” a credible interval of (.73, .85) as an accurate estimate. However, we would consider a credible interval of (.67, .85) as an inaccurate estimate and “reject” it (even though it contains the true value) since the ROPE does not completely contain this credible interval. In the left panel, there are 74 out of 100 accurate estimates using BSCM horseshoe, and therefore the measure of precision/power is .74. Bayesian lasso in the right panel yields only 31 out of 100 precise estimates, and thus the precision/power is .31. Therefore, in this instance we conclude that our model has a higher degree of precision than Bayesian lasso, given a fixed sample size.

In Table 4, Panels A and B, we show the results of such analyses across all simulation cases and models. For each simulation case, we generate 100 data sets and estimate the parameters to calculate the accuracy in parameter estimation. We indicate the true parameter values (e.g.,  $\beta_1 = .2$ ,  $\beta_2 = .8$  in S1) and associated ROPEs (e.g., (.10, .30), (.70, .90) in S1). We find that the standard SCM shows higher precision than our models only when the DGP satisfies the SCM constraints and the “large p, small n” and sparsity conditions do not occur (i.e.,

S5: S-No concern). This finding is in line with the overall model performance results in Table 3, Panel A, that the standard SCM shows slightly better predictive accuracy than other models in this setting. Other than this specific case, our proposed models dominate all other models in all scenarios in terms of precision of parameter estimates. Put differently, other than computational time, our proposed models are broadly preferable to other models except when we have strong reasons to believe that the DGP conforms to the conditions of the standard SCM, which is unlikely to ever hold.

## Empirical Application

### Background

We apply our proposed methods to understand the causal effect of the imposition of a soda tax in the state of Washington on prices and consumption of soda (Rojas and Wang 2017). The motivation for the tax is the belief in a positive link between the intake of sugar-sweetened beverages (SSBs) and high obesity/diabetes rates (Roberto et al. 2019; Silver et al. 2017). Berkeley, California, passed a 1 cent per ounce tax on SSBs in March 2015, followed by Philadelphia, which levied a 1.5 cents per ounce tax. These events captured the interest of marketing researchers since a tax can cause changes in firms’ and consumers’ behaviors. Rojas and Wang (2017) and Bollinger and Sexton (2018) found that the effect of the SSB tax in Berkeley on prices and consumption was limited. Bollinger and Sexton (2018) further identified the reasons for this limited effect; the tax can be easily avoided by buying SSBs from retail stores outside the taxed jurisdiction; also, chain retailers tend to pursue uniform local pricing whereby all stores in large geographic regions charge a common price. Seiler, Tuchman, and Yao (2019) analyzed the effect of an SSB tax in

**Table 4.** Accuracy in Parameter Estimation of Alternative Models in Simulation Studies.

<b>A: <math>S1 \sim S5</math> (SCM constraints satisfied)</b>										
	<b>S1: S-Sparsity</b>		<b>S2: S-Large p, Small n</b>		<b>S3: S-MRW</b>		<b>S4: S-Large p, Small n and MRW</b>		<b>S5: S-No Concern</b>	
	$\beta_1 = .20$ (.10, .30)	$\beta_2 = .80$ (.70, .90)	$\beta_1 = .20$ (.05, .35)	$\beta_2 = .80$ (.65, .95)	$\beta_1 = .24$ (.12, .36)	$\beta_2 = .24$ (.12, .36)	$\beta_1 = .24$ (.06, .44)	$\beta_2 = .24$ (.06, .44)	$\beta_1 = .20$ (.10, .30)	$\beta_2 = .80$ (.70, .90)
Standard SCM	.00	.00	.00	.00	.38	.15	.00	.01	.98	.99
Panel data	.20	.16	—	—	.56	.52	—	—	.79	.54
method										
Ridge	.37	.05	.01	.00	.10	.11	.07	.33	.77	.52
Lasso	.51	.31	.04	.35	.69	.51	.16	.48	.76	.52
Elastic net	.39	.07	.03	.00	.19	.16	.18	.53	.79	.54
BSCM horseshoe	.73	.74	.11	.77	.93	.86	.41	.75	.74	.57
BSCM spike and slab	.67	.71	.08	.72	.92	.83	.44	.75	.71	.59
<b>B: <math>S6 \sim S10</math> (SCM constraints violated)</b>										
	<b>S6: V-Sparsity</b>		<b>S7: V-Large p, Small n</b>		<b>S8: V-MRW</b>		<b>S9: V-Large p, Small n and MRW</b>		<b>S10: V-No Concern</b>	
	$\beta_1 = -.50$ (-.60, -.40)	$\beta_2 = 2.00$ (1.90, 2.10)	$\beta_1 = -.50$ (-.70, -.30)	$\beta_2 = 2.00$ (1.80, 2.20)	$\beta_1 = -.50$ (-.60, -.40)	$\beta_2 = 2.00$ (1.90, 2.10)	$\beta_1 = -.50$ (-.70, -.30)	$\beta_2 = 2.00$ (1.80, 2.20)	$\beta_1 = -.50$ (-.60, -.40)	$\beta_2 = 2.00$ (1.90, 2.10)
Standard SCM	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
Panel data	.17	.16	—	—	.16	.19	—	—	.16	.19
method										
Ridge	.30	.12	.00	.00	.30	.12	.00	.00	.30	.12
Lasso	.57	.37	.28	.61	.54	.34	.28	.61	.54	.34
Elastic net	.34	.13	.00	.00	.34	.13	.00	.00	.34	.13
BSCM horseshoe	.80	.75	.83	.99	.72	.70	.64	.96	.72	.70
BSCM spike and slab	.76	.71	.82	.98	.75	.68	.73	.93	.75	.68

Notes: In each simulation case, we indicate the true parameter values (e.g.,  $\beta_1 = .20$ ,  $\beta_2 = .80$  in S1) and associated ROPEs (e.g., [.10, .30], [.70, .90] in S1). The highest precision values are highlighted in bold. Numbers shown are proportion of 100 samples in which estimated 95% credible interval is completely within the ROPE. MRW = many relevant weights.

Philadelphia and surrounding areas. They found that soda taxes lead to a 34% retail price increase and a drop of 46% in demand. Similar to Bollinger and Sexton (2018), Seiler, Tuchman, and Yao (2019) also found that consumers shifted to buying SSBs outside the city.

As previously mentioned, Washington imposed a 1/6 cent per ounce tax on both regular and diet soda effective July 1, 2010. The governor of Washington secretly progressed the soda tax legislation to avert almost certain opposition from the beverage industry. There were no committee hearings on the soda tax and it was the least publicized among all the taxes considered by the state legislature in 2010. Soon after the tax bill was publicized, the American Beverage Association spent more than \$1 million gathering signatures to get an antitax initiative on the ballot, and later spent over \$16 million on supporting campaigns against the soda tax. The beverage industry eventually got what it wanted through a ballot vote, and the tax was repealed effective December 1st, 2010. A more detailed history of the Washington soda tax legislation is available in Rojas and Wang (2017).

The Washington event is less known compared to the Berkeley and Philadelphia events and is unique in several ways. First, the Washington tax was levied on all soda products, not only SSBs.<sup>9</sup> Second, unlike the Berkeley and Philadelphia events, the Washington tax affected firms and consumers in the entire state. This implies that consumers were less likely to purchase at retailers outside taxed jurisdictions in response to the tax. Third, the size of the Washington soda tax (1/6 cent per oz) was much smaller compared with Berkeley's (1 cent per oz) and Philadelphia's (1.5 cent per oz).

Rojas and Wang (2017) used difference in differences with Oregon as a control unit to measure the effect of the Washington soda tax. They found that the reduction in sales volume due to the tax was statistically insignificant. They also found that the tax caused a retail price increase of .18 cents per ounce, which implies 105% pass-through on average.<sup>10</sup> We reexamine these results using SCMs, in particular our Bayesian framework, and estimate the effect of the tax on retail prices and sales volume of soda.

## Data

We use Nielsen scanner data provided by Kilts Center for Marketing<sup>11</sup> to investigate the effect of the tax. The data set

consists of store-level weekly average prices scanned at participating retailers as well as volume sales for each universal product code (UPC) of all soda products. We also observe anonymized store identification numbers, store location information such as state, county, and three-digit zip codes, and promotion information such as whether a UPC was promoted by a feature ad and/or in-store display. The data span 157 weeks: January 2009 to December 2011. Because the tax was in effect from July 1 to December 1 of 2010, we have 78 pretreatment weeks (i.e., from the beginning of January 2009 to the end of June 2010), 22 weeks during which the tax was in effect (i.e., from the beginning of July 2010 to the end of November 2010), and 57 posttax weeks (i.e., from the beginning of December 2010 to the end of December 2011).

Because Washington was the only state that adopted the soda tax in 2010, it is the only treated unit in our analysis. The other 48 states in the contiguous United States comprise the control units. To obtain volume-weighted average price for soda for each week at the state level, we summed volume sales at the UPC/store/week level and computed the volume-weighted average of UPC-/store-/week-level prices, using the UPC's volume in a store-week as the weight.

## Results

We estimate all the Bayesian models described in the model section and compare predictive performance and computational efficiencies across models. Note that we have a sufficient number of pretreatment periods to have a unique solution. However, because we have a large number of control units (i.e., 48 states), sparsity can be a potential challenge. This situation is similar to S1/S6 in the simulation studies.

In Table 5, we present the LOO-CV, which is a method for estimating out-of-sample predictive accuracy from fitted Bayesian models (Vehtari, Gelman, and Gabry 2017). The results show that the horseshoe and spike-and-slab models have better predictive accuracy compared to other models although Bayesian lasso has very similar performance. In the analysis of sales volume, Bayesian elastic net initially did not converge. The model converged after increasing the number of iterations to three times more (i.e., 12,000 iterations, 6,000 burn-in) than for other models. We conjecture that the inherent flexibility of the elastic net, which intuitively combines L1 and L2 regularization in the form of complicated Bayesian priors, is a double-edged sword, and it comes at the cost of model complexity which may hinder MCMC convergence. In terms of computational efficiency, the proposed models are generally slower than the other extant models, but their maximum estimation time is less than 200 seconds.

<sup>9</sup> SSBs are different from sodas in that they refer to any liquids that are sweetened with various forms of added sugars, such as regular sodas, fruit drinks, sports drinks, energy drinks, sweetened waters, and coffee/tea beverages. SSBs exclude sugar-free drinks such as diet sodas. In contrast, sodas are any carbonated drinks; these include sugar-free drinks but exclude some of the SSBs, such as fruit drinks and coffee/tea beverages.

<sup>10</sup> Pass-through is defined as the change in retail price per ounce expressed as a percentage of the tax measured in dollars per ounce.

<sup>11</sup> Researchers' own analyses were calculated (or derived) based in part on data from The Nielsen Company (US), LLC and marketing databases provided through the Nielsen Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions

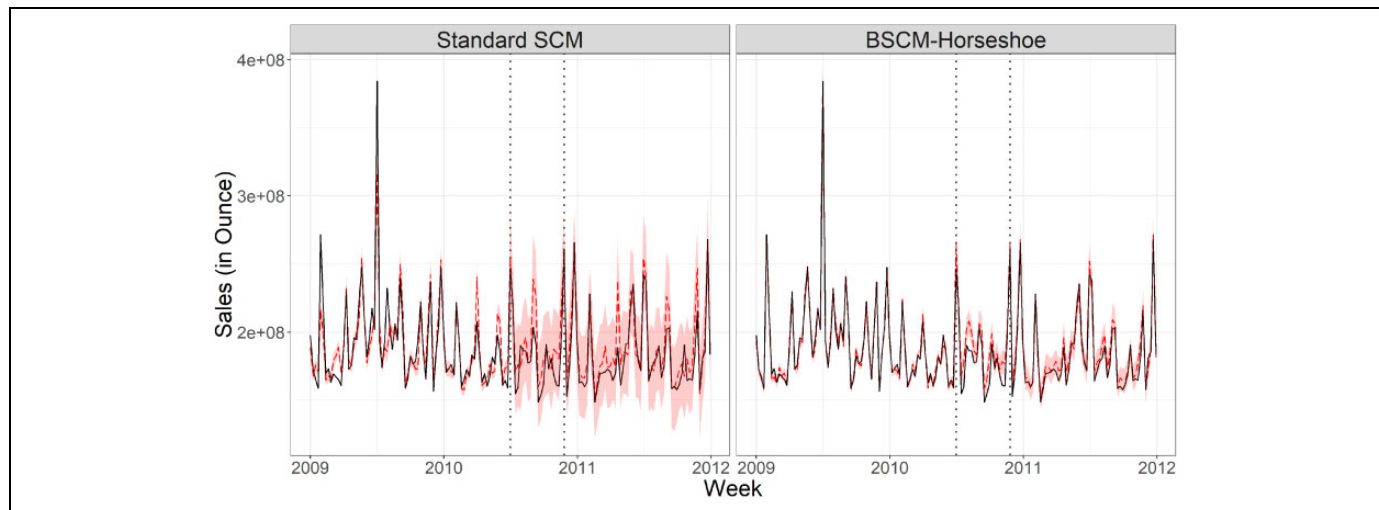
drawn from the Nielsen data are those of the researchers and do not reflect the views of Nielsen. Nielsen is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.



**Table 5.** Predictive Accuracy and Computational Efficiency of Alternative Models: Washington Soda Price and Sales.

	Price		Sales	
	LOO-CV	Estimation Time (Sec)	LOO-CV	Estimation Time (Sec)
Standard SCM	516.7	3.1	2,817.5	9.9
Panel data method	445.5	61.4	2,645.2	56.1
Ridge	417.2	62.7	2,640.3	43.6
Lasso	403.4	64.5	2,630.2	35.0
Elastic net	418.8	287.3	2,640.2	295.7
BSCM horseshoe	<b>400.9</b>	107.0	2,624.9	105.1
BSCM spike and slab	404.4	187.1	<b>2,612.6</b>	96.0

Notes: Estimation time is measured as mean estimation time of four chains in seconds. The lowest LOO-CV values are highlighted in boldface.

**Figure 4.** Total sales of soda in Washington and synthetic Washington.

Notes: The black (solid) line is the total sales of soda in Washington state while the red (dashed) line is the fitted and predicted sales of soda of synthetic Washington. The first dotted vertical line is the first week of July 2010, when the soda tax came into effect, while the second dotted vertical line indicates the last week of November 2010, when the tax was repealed. The shaded areas are 95% posterior credible intervals.

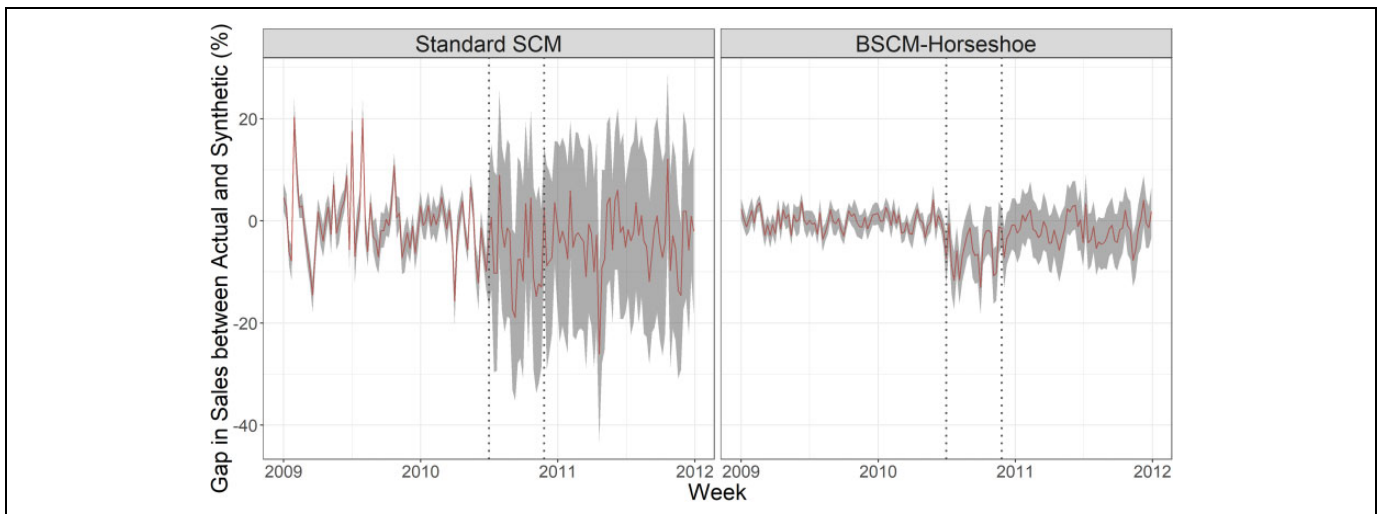
### Sales Analysis

Figure 4 shows total sales of soda in actual Washington and synthetic Washington. For ease of exposition, we only show results for standard SCM and the BSCM horseshoe.<sup>12</sup> The black (solid) line is the actual weekly soda sales in Washington while the red (dashed) line is the fitted and predicted weekly soda sales of synthetic Washington. In addition, the first vertical line represents the first week in July 2010 when the tax came into effect while the second vertical line indicates the last week of November 2010 when the tax was repealed. Compared with BSCMs, the standard SCM shows poor in-sample fit. The red line of standard SCM starts to diverge from the black line before the tax came into effect, which makes the divergence in the tax periods questionable. Moreover, the credible intervals of synthetic Washington include the actual Washington sales during the tax period. Using the standard SCM, the conclusion is that the tax did not affect the

sales of soda. On the other hand, the synthetic Washington based on the horseshoe model shows excellent in-sample fit in the pretreatment periods. During the tax period, the sales of synthetic Washington, which are the estimates of the treated unit outcome under no treatment, are higher than actual Washington sales. This implies that soda sales would have been significantly higher were it not for the tax, a conclusion in stark contrast with that from the standard SCM model.

In Figure 5, the red line shows the percentage gap between actual soda sales in Washington and its synthetic counterpart. Specifically, for each model we calculate  $(\text{Actual Sales} - \text{Synthetic Control Sales}) / \text{Actual Sales} \times 100$  using posterior samples. The standard SCM result shows no significant decrease in sales: the average treatment effect is estimated to be  $-6.17\%$ , but the 95% credible interval is very large and contains zero (95% credible interval:  $[-23.13, 10.81]$ ), which corresponds to Rojas and Wang (2017)'s findings. By contrast, the horseshoe and spike-and-slab models show that the decrease in sales due to the tax is quite substantial, although the credible interval occasionally includes zero.

<sup>12</sup> The results of BSCM spike and slab are very similar to BSCM horseshoe in both sales and price analyses. Therefore, we omit these figures.



**Figure 5.** Percentage sales gap between Washington and synthetic Washington.

Notes: The red line shows the percentage gap between actual total soda sales and its synthetic counterpart. The shaded areas are 95% posterior credible intervals.

More specifically, we estimate the average treatment effect on sales during the tax period to be  $-5.82\%$  (95% credible interval:  $[-10.87, -.82]$ ) using the horseshoe model and  $-5.45\%$  (95% credible interval:  $[-10.25, -.61]$ ) using the spike-and-slab model. Accordingly, we estimate the mean price elasticity to be 1.02 and .96 for the horseshoe and spike-and-slab models respectively. This shows that good predictive accuracy is very important in the SCM context because it can substantively change the conclusion of the analysis.

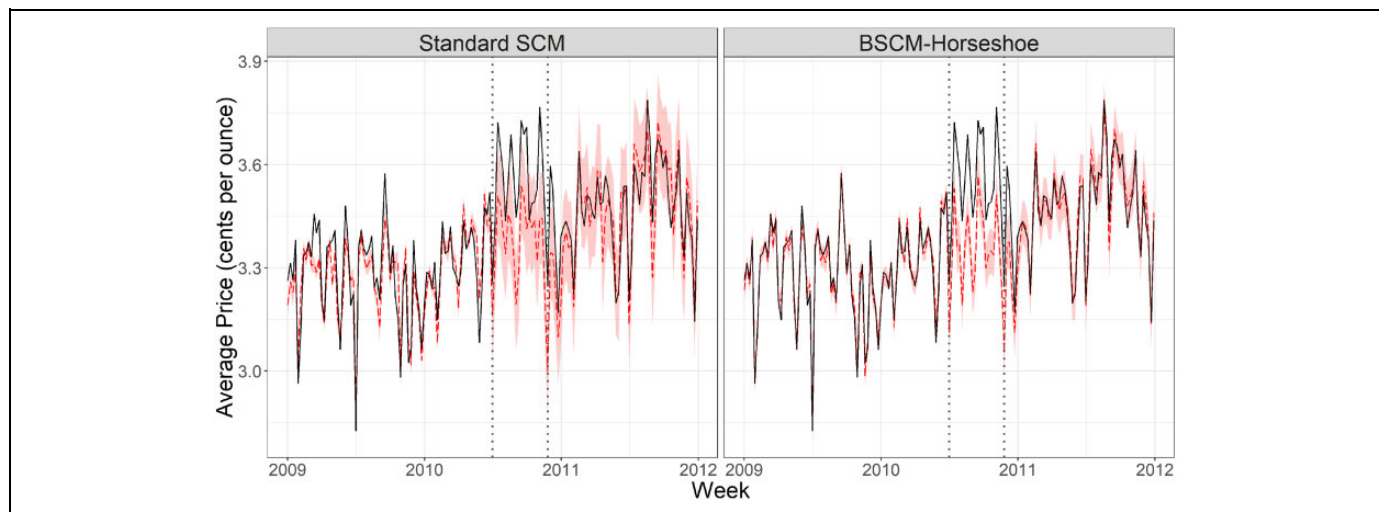
Turning to the estimated weights on control units, in the spike-and-slab estimates, the weights on Oregon, Montana, Maryland, and Virginia do not contain zero; the mean estimates are 1.33 (95% credible interval:  $[1.19, 1.47]$ ) for Oregon, 1.10 (95% credible interval:  $[.53, 1.80]$ ) for Montana, .27 (95% credible interval:  $[.08, .41]$ ) for Maryland, and  $-.10$  (95% credible interval:  $[-.20, -.01]$ ) for Virginia. Using the horseshoe model, only the credible intervals of weights for Oregon and Maryland do not contain zero; the mean estimates are 1.35 (95% credible interval:  $[1.19, 1.51]$ ) for Oregon, and .19 (95% credible interval:  $[.01, .35]$ ) for Maryland. In the case of the standard SCM, we can judge the statistical significance of an estimated parameter by whether the posterior mean is no less than twice the posterior standard deviation. This is because we have restricted all parameter estimates to be greater than zero via the nonnegativity constraint, and therefore the posterior distributions of all parameters do not contain zero. Using this criterion, only California's mean weight estimate is "significant": the mean estimate is .05 with .02 posterior standard deviation. The poor predictive accuracy of the standard SCM casts doubts on whether the model has correctly identified the sparse parameters.

### Price Analysis

In Figure 6, we show average price of soda in Washington and its synthetic control. The black (solid) line is the actual average

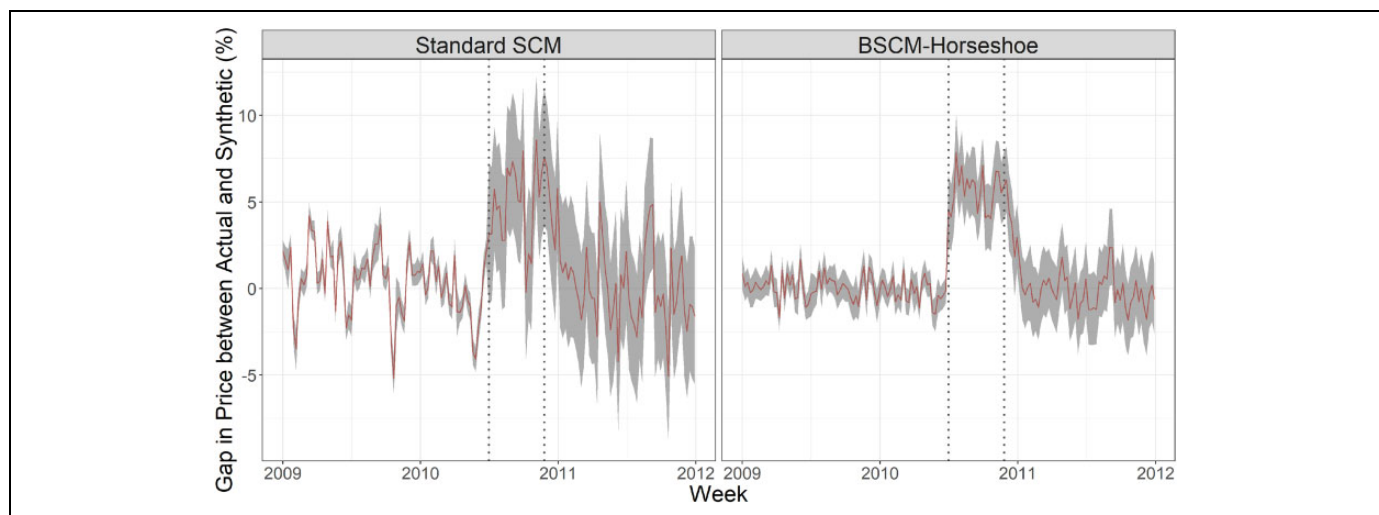
price of soda in Washington while the red (dashed) line is the fitted and predicted average price of soda of synthetic Washington. In the results for all model specifications, we note that the red line closely follows the black line in pretreatment periods, but the two lines start to sharply diverge when the tax is imposed. However, the two lines converge again when the tax is repealed. Moreover, the results of all models show that synthetic Washington's credible intervals (i.e., the shaded areas) do not include the actual price of soda (i.e., the black lines), implying that retailers in Washington changed the retail price of soda during the tax period. We also see that the standard SCM has wider credible intervals than the proposed methods. An additional finding is that the retailers did not adjust their prices immediately upon repeal of the tax on Dec 1, 2010; the posttax price remained high during December 2010.

In Figure 7, the red line is the percentage price gap between actual average price in Washington and its synthetic counterpart. The BSCM models show especially salient patterns; the gap between actual Washington and synthetic Washington is no different from zero in each pretreatment period, but jumps to an increase of more than 5% during the tax periods. The gap then returns to zero during the posttax periods. We estimate the average treatment effect during the tax period to be 5.69% (95% credible interval:  $[3.77, 7.59]$ ) using spike-and-slab model, and 5.66% (95% credible interval:  $[3.77, 7.59]$ ) using the horseshoe model. These percentages translate to increases of more than .2 cents per ounce, which is greater than the tax of 1/6 cents per ounce. Therefore, we estimate the pass-through rate to be 121%. Results of the standard SCM are noisier than the results of the horseshoe and spike-and-slab models. This reflects the poor in-sample fit and wider credible intervals shown in Figure 7. Using the standard SCM, the average treatment effect is estimated to be 4.99% (95% credible interval:  $[1.26, 8.75]$ ), which translates to retail price increases of .18 cents per ounce and 106% pass-through rate. The results



**Figure 6.** Average price of soda in Washington and synthetic Washington.

Notes: The black (solid) line is the actual average price of price in Washington state while the red (dashed) line is the fitted and predicted average price of soda of synthetic Washington. The first dotted vertical line in each graph is the first week of July 2010, when the soda tax came into effect, while the second dotted vertical line indicates the last week of November 2010, when the tax was repealed. The shaded areas are 95% posterior credible intervals.



**Figure 7.** Percentage price gap between Washington and Synthetic Washington.

Notes: The red line shows the percentage gap between actual price of soda in Washington and its synthetic counterpart. The shaded areas are 95% posterior credible intervals.

obtained by the standard SCM are again very similar to those of Rojas and Wang's (2017) findings.

In terms of weights on control units, we find that in the proposed models, only the weight on Oregon is different from zero (i.e., the credible interval does not include zero). The mean estimates for Oregon are .72 (95% credible interval: [.59, .85]) in the spike-and-slab model, and .72 (95% credible interval: [.60, .85]) in the horseshoe model. Using the standard SCM, Oklahoma and Wyoming have significant weights, while Oregon does not.

In summary, we find statistically significant decreases in sales of 5.5%–5.8% due to the tax, unlike the null effect finding of Rojas and Wang (2017). Moreover, we find slightly higher

retail price increase and pass-through rates (i.e., .20 cents per ounce and 121%) compared with difference-in-differences model results of Rojas and Wang (2017) (i.e., .18 cents per ounce and 105%).

## Conclusion

We propose a new Bayesian synthetic control framework that addresses three key limitations of extant SCMs and also has better predictive accuracy. The proposed Bayesian synthetic control methods do not impose any restrictive constraints on the parameter space a priori. Moreover, they provide a natural mechanism to deal with the “large p, small n” and sparsity

problems through shrinkage prior approaches. In simulation studies, we find that the proposed BSCMs almost always dominate other models in terms of predictive accuracy in a variety of data situations. Even when the DGP is consistent with all the standard SCM constraints, the proposed models perform better when “large  $p$ , small  $n$ ” or sparsity concerns exist. When the DGP violates the SCM constraints, the standard SCM is not identified whereas the proposed models show excellent predictive accuracy. Our Bayesian framework incorporates the extant models by specifying particular priors on weights. Unlike frequentist models in the literature, the proposed Bayesian framework allows for statistical inference in a straightforward manner.

We demonstrate how the proposed methods can be applied to real world data by estimating the impact of a tax on soda in the state of Washington in 2010. The proposed models show good in-sample fit in the pretreatment periods and better predictive accuracy compared to extant models. We find that the tax in Washington led to a 5.7% increase in retail price and 5.5%–5.8% decrease in sales of soda. We also find that retailers in Washington overshifted the tax to consumers, with a pass-through rate of about 121%. Our substantive findings stand in contrast with those of Rojas and Wang (2017), and appear to be more face valid.

The proposed method is not without limitations. The method implicitly assumes that we have a sufficient number of control units that can be used to form a synthetic control unit. However, it is possible that one has limited information on control units making creation of a valid synthetic control unit challenging. In such cases, we recommend using structural time-series models as in Brodersen et al. (2015) that utilize the treated unit’s pretreatment period information to predict the potential outcomes without treatment. Moreover, as in the standard synthetic control method, our methods focus on the case where only one unit is treated. However, this is done without loss of generality. In cases where multiple units are treated, our method can be applied to each affected unit separately. If one does not want to find synthetic control units for each treated unit one by one, Xu (2017) has proposed a method to obtain multiple treated counterfactuals in a single estimation run based on an interactive fixed effects model. Future research can extend our model to accommodate many treated units in a single run. Despite these limitations, we expect that the proposed methods will provide a useful alternative to the extant SCM approaches in many empirical problems.

### Acknowledgments

This paper is part of the PhD dissertation of the first author at Cornell University. The authors thank David Ruppert, attendees at the Marketing Dynamics Conference 2019 and seminar participants at the following schools for feedback: Cornell University, University of Western Ontario, Monash University, Deakin University, University of Washington, University of Texas at Dallas, San Diego State University, University of Hawaii at Manoa, Texas Christian University, and the Indian Institute of Management, Bangalore.

### Guest Editor

Robert Meyer

### Associate Editor

Eric Bradlow


### Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: at the time of acceptance Sachin Gupta was a coeditor of the *Journal of Marketing Research*.

### Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

### ORCID iD

Sachin Gupta  <https://orcid.org/0000-0001-9459-5233>

### References

- Abadie, Alberto and Javier Gardeazabal (2003, March), “The Economic Costs of Conflict: A Case Study of the Basque Country,” *American Economic Review*, 93 (1), 113–32.
- 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.
- Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2011), “Synth: An R Package for Synthetic Control Methods in Comparative Case Studies,” *Journal of Statistical Software*, 42 (13), 1–17.
- Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2015), “Comparative Politics and the Synthetic Control Method,” *American Journal of Political Science*, 59 (2), 495–510.
- Athey, Susan and Guido W. Imbens (2017), “The State of Applied Econometrics: Causality and Policy Evaluation,” *Journal of Economic Perspectives*, 31 (2), 3–32.
- Ben-Michael, Eli, Avi Feller, and Jesse Rothstein (2018), “The Augmented Synthetic Control Method,” *arXiv:1811.04170 [econ, stat]*. arXiv: 1811.04170.
- Bollinger, Bryan and Steven Sexton (2018), “Local Excise Taxes, Sticky Prices, and Spillovers: Evidence from Berkeley’s Soda Tax,” SSRN Scholarly Paper ID 3087966, Social Science Research Network.
- Brodersen, Kay H., Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L. Scott (2015), “Inferring Causal Impact Using Bayesian Structural Time-Series Models,” *Annals of Applied Statistics*, 9 (1), 247–74.
- Card, David and Alan B. Krueger (1994), “Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania,” *American Economic Review*, 84 (4), 772–93.
- Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, et al. (2017), “Stan: A Probabilistic Programming Language,” *Journal of Statistical Software*, 76 (1), 1–32.



- Carvalho, Carlos M., Ricardo Masini, and Marcelo C. Medeiros (2018, December), "ArCo: An Artificial Counterfactual Approach for High-dimensional Panel Time-series Data," *Journal of Econometrics*, 207 (2), 352–80.
- Carvalho, Carlos M., Nicholas G. Polson, and James G. Scott (2010), "The Horseshoe Estimator for Sparse Signals," *Biometrika*, 97 (2), 465–80.
- Chatterjee, A. and S.N. Lahiri (2011), "Bootstrapping Lasso Estimators," *Journal of the American Statistical Association*, 106 (494), 608–25.
- Chatterjee, A. and S. N. Lahiri (2013), "Rates of Convergence of the Adaptive LASSO Estimators to the Oracle Distribution and Higher Order Refinements by the Bootstrap," *Annals of Statistics*, 41 (3), 1232–59.
- Chesnes, Matthew, Weijia Daisy Dai, and G. Zhe Jin (2017), "Banning Foreign Pharmacies from Sponsored Search: The Online Consumer Response," *Marketing Science*, 36 (6), 879–907.
- Chung, Yeojin, Sophia Rabe-Hesketh, Vincent Dorie, Andrew Gelman, and Jingchen Liu (2013), "A Nondegenerate Penalized Likelihood Estimator for Variance Parameters in Multilevel Models," *Psychometrika*, 78 (4), 685–709.
- Doudchenko, Nick and Guido W. Imbens (2016), "Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis," Working Paper 22791, National Bureau of Economic Research, <http://www.nber.org/papers/w22791>.
- Ferman, Bruno and Cristine Pinto (2016), "Revisiting the Synthetic Control Estimator," working paper.
- Firpo, Sergio and Vitor Possebom (2018), "Synthetic Control Method: Inference, Sensitivity Analysis and Confidence Sets," *Journal of Causal Inference*, 6 (2), 1–26.
- Fonseca, Yuri R., Ricardo P. Masini, Marcelo C. Medeiros, and Gabriel F.R. Vasconcelos (2018), "ArCo: An R Package to Estimate Artificial Counterfactuals," *R Journal*, 10 (1), 91–108.
- Gardeazabal, Javier and Ainhoa B. Vega (2017), "An Empirical Comparison Between the Synthetic Control Method and HSIAO et al.'s Panel Data Approach to Program Evaluation," *Journal of Applied Econometrics*, 32 (5), 983–1002.
- Gelman, Andrew, Aleks Jakulin, Maria G. Pittau, and Yu-Sung Su (2008), "A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models," *Annals of Applied Statistics*, 2 (4), 1360–83.
- Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin (2013), *Bayesian Data Analysis*. 3rd ed. Boca Raton, FL: Chapman and Hall/CRC.
- George, Edward I. and Robert E. McCulloch (1993), "Variable Selection Via Gibbs Sampling," *Journal of the American Statistical Association*, 88 (423), 881–89.
- Hahn, Jinyong and Ruoyao Shi (2017), "Synthetic Control and Inference," *Econometrics*, 5 (4), 52.
- Hastie, Trevor, Robert Tibshirani, and Martin Wainwright (2015), *Statistical Learning with Sparsity: The Lasso and Generalizations*. 1st ed. Boca Raton: Chapman and Hall/CRC.
- Hsiang, T.C. (1975), "A Bayesian View on Ridge Regression," *Journal of the Royal Statistical Society. Series D (The Statistician)*, 24 (4), 267–68.
- 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.
- Ishwaran, Hemant and J. Sunil Rao (2005), "Spike and Slab Variable Selection: Frequentist and Bayesian Strategies," *Annals of Statistics*, 33 (2), 730–73.
- Kaul, Ashok, Stefan Klöyner, Gregor Pfeifer, and Manuel Schieler (2015), "Synthetic Control Methods: Never Use All Pre-Intervention Outcomes Together With Covariates," working paper.
- Knight, Keith and Wenjiang Fu (2000), "Asymptotics for Lasso-Type Estimators," *Annals of Statistics*, 28 (5), 1356–78.
- Kruschke, John K. (2014), *Doing Bayesian Data Analysis, Second Edition: A Tutorial with R, JAGS, and Stan*, 2nd ed. Boston: Academic Press.
- Kyung, Minjung, Jeff Gill, Malay Ghosh, and George Casella (2010), "Penalized Regression, Standard Errors, and Bayesian Lassos," *Bayesian Analysis*, 5 (2), 369–411.
- Li, Kathleen T. (2019), "Statistical Inference for Average Treatment Effects Estimated by Synthetic Control Methods," *Journal of the American Statistical Association* (published online December 9), DOI: 10.1080/01621459.2019.1686986.
- Li, Qing and Nan Lin (2010), "The Bayesian Elastic Net," *Bayesian Analysis*, 5 (1), 151–70.
- Mallick, Himel and Nengjun Yi (2013), "Bayesian Methods for High Dimensional Linear Models," *Journal of Biometrics & Biostatistics*, 1, 005.
- Maxwell, Scott E., Ken Kelley, and Joseph R. Rausch (2008), "Sample Size Planning for Statistical Power and Accuracy in Parameter Estimation," *Annual Review of Psychology*, 59, 537–63.
- McElreath, Richard (2015), *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Boca Raton, FL: Chapman and Hall/CRC.
- Mitchell, T.J. and J.J. Beauchamp (1988), "Bayesian Variable Selection in Linear Regression," *Journal of the American Statistical Association*, 83 (404), 1023–32.
- Park, Trevor and George Casella (2008), "The Bayesian Lasso," *Journal of the American Statistical Association*, 103 (482), 681–86.
- Piironen, Juho and Aki Vehtari (2017), "Sparsity Information and Regularization in the Horseshoe and Other Shrinkage Priors," *Electronic Journal of Statistics*, 11 (2), 5018–51.
- Powell, D. (2018), "Imperfect Synthetic Controls: Did the Massachusetts Health Care Reform Save Lives?" *SSRN Scholarly Paper ID 3192710*, *Social Science Research Network*.
- Roberto, Christina A., Hannah G. Lawman, Michael T. LeVasseur, Nandita Mitra, Ana Peterhans, Bradley Herring, et al. (2019), "Association of a Beverage Tax on Sugar-Sweetened and Artificially Sweetened Beverages With Changes in Beverage Prices and Sales at Chain Retailers in a Large Urban Setting," *JAMA*, 321 (18), 1799–810.
- Rojas, Christian and Emily Y. Wang (2017), "Do Taxes for Soda and Sugary Drinks Work? Scanner Data Evidence from Berkeley and Washington," *SSRN Scholarly Paper ID 3041989*, *Social Science Research Network*.

- Rosenbaum, Paul R. and Donald B. Rubin (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70 (1), 41–55.
- Seiler, Stephan, Anna Tuchman, and Song Yao (2019), "The Impact of Soda Taxes: Pass-Through, Tax Avoidance, and Nutritional Effects," SSRN Scholarly Paper ID 3302335, Social Science Research Network.
- Silver, Lynn D., Shu Wen Ng, Suzanne Ryan-Ibarra, Lindsey S. Taillie, Marta Induni, and Donna R. Miles (2017), "Changes in Prices, Sales, Consumer Spending, and Beverage Consumption One Year After a Tax on Sugar-Sweetened Beverages in Berkeley, California, US: A Before-and-After Study," *PLoS Medicine*, 14 (4), e1002283.
- Tirunillai, Seshadri and Gerard J. Tellis (2017), "Does Offline TV Advertising Affect Online Chatter? Quasi-Experimental Analysis Using Synthetic Control," *Marketing Science*, 36 (6), 862–78.
- Van Erp, Sara, Daniel L. Oberski, and Joris Mulder (2019), "Shrinkage Priors for Bayesian Penalized Regression," *Journal of Mathematical Psychology*, 89 (April), 31–50.
- Vehtari, Aki, Andrew Gelman, and Jonah Gabry (2017), "Practical Bayesian Model Evaluation Using Leave-One-Out Cross-Validation and WAIC," *Statistics and Computing*, 27 (5), 1413–32.
- Xu, Y. (2017), "Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models," *Political Analysis*, 25 (1), 57–76.
- Zou, Hui and Trevor Hastie (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67 (2), 301–20.