Bayesian Matrix Completion Approach to Causal Inference with Panel Data

Masahiro Tanaka*

November 4, 2019

Abstract

This study proposes a new Bayesian approach to infer average treatment effect. The approach treats counterfactual untreated outcomes as missing observations and infers them by completing a matrix composed of realized and potential untreated outcomes using a data augmentation technique. We also develop a tailored prior that helps in the identification of parameters and induces the matrix of the untreated outcomes to be approximately low rank. While the proposed approach is similar to synthetic control methods and other relevant methods, it has several notable advantages. Unlike synthetic control methods, the proposed approach does not require stringent assumptions. Whereas synthetic control methods do not have a statistically grounded method to quantify uncertainty about inference, the proposed approach can estimate credible sets in a straightforward and consistent manner. Our proposal approach has a better finite sample performance than the existing Bayesian and non-Bayesian approaches, as we show through a series of simulation studies.

Keywords: average treatment effect, causal inference, low-rank matrix completion, panel data, shrinkage prior, synthetic control method.

1 Introduction

Program/policy evaluation and comparative case study using observational data are pervasive in social and natural sciences as well as in practice in government and business. In particular, causal inference is an integral part of social sciences, where randomized experiments are usually infeasible. For instance, although [Abadie et al., 2015] analyzed the economic cost of the German reunification in 1990, we cannot repeat such a political event many times in a controlled fashion.

The primary interest of this study is inference of average treatment effect (ATE). Suppose we have a panel data with $J$ units and $T$ time periods. An outcome of unit $j$ at period $t$ is denoted by $y_{j,t}(s)$, where $s = 1$ when it is exposed to treatment and $s = 0$ otherwise. Let $S_1$ and $S_0$ be sets of indices for treated and untreated observations, respectively. Then, ATE is represented as

$$ATE = \frac{1}{|S_1|} \sum_{(j,t) \in S_1} (y_{j,t}(1) - y_{j,t}(0)),$$

*Graduate School of Economics, Waseda University; Director-General for Policy Planning (Statistical Standards), Ministry of Internal Affairs and Communications. Address: 1-6-1, Nishi-Waseda, Shinjuku-ku, Tokyo 169-8050 Japan, Email: gspddlnit45@toki.waseda.jp.
where \(|\mathcal{A}|\) denotes the cardinality of a set \(\mathcal{A}\). Estimation of ATE amounts to estimation of counterfactual untreated outcomes \(y_{j,t}(0), (j,t) \in S_1\), or “potential outcome” in terms of Neyman–Rubin’s causal model \(\text{Holland} \ 1986\). Inference of ATE poses a serious challenge to statisticians, and numerous approaches have been proposed: the difference-in-difference estimator, regression discontinuity design, matching-based methods, etc.

In the study, we propose a new Bayesian approach for inferring ATE with panel data. We treat counterfactual untreated outcomes as missing observations and infer them via data augmentation \(\text{Tanner and Wong} \ 1987\). In other words, we transform a statistical inference of ATE into a matrix completion problem, which is an extensively studied issue in the machine learning community (e.g., \text{Srebro et al.} \ 2005; \text{Keshavan et al.} \ 2010). To facilitate the purpose, we develop a tailored prior that ensures identification of parameters and induces the model to be lower rank, adapting a cumulative shrinkage process prior \(\text{Legramanti et al.} \ 2019\).

Our Bayesian approach has two notable advantages. First, it can provide credible intervals in a consistent manner, while the existing non-Bayesian approaches do not have a statistically grounded method to quantify uncertainty about inference. As hypothesis testing is an essential part of scientific research, this advantage would be a strong reason to resort to Bayesian methods. Second, it has better finite sample performance than the existing approaches. Through a series of simulation studies, we show that our proposal consistently outperforms the existing approaches in terms of the precision of the prediction of potential outcomes.

Two strands of literature are particularly relevant to the study. First, the proposed approach is related to a class of synthetic control methods (SCMs) (e.g., \text{Abadie and Gardeazabal} \ 2003; \text{Abadie et al.} \ 2010; \text{Xu} \ 2017; \text{Doudchenko and Imbens} \ 2017). This class of methods is aimed at obtaining “synthetic” observations of untreated outcomes as weighted sums of the outcomes of the control units. Despite its increasing popularity, the original SCM (\text{Abadie et al.} \ 2010) has two notable shortcomings. First, it imposes strong assumption that the weights of synthetic observations are non-negative and sum-to-one. This implies that the treated unit falls in the convex hull of the control units and that synthetic observations are positively correlated to the control units, which is not plausible in many real situations. Second, it does not have a an effective method for assessing uncertainty on obtained estimates. Indeed, \text{Abadie et al.} \ (2010) propose to conduct a series of placebo studies, but it is merely a sort of robustness check. \text{Doudchenko and Imbens} \ (2017) propose a variant of SCM that is free from the three assumptions above, while \text{Kim et al.} \ (2019) develop a Bayesian version of their approach that can deliver credible sets.

Second, an approach developed by \text{Athey et al.} \ (2018) is particularly related to our proposal. They also treat counterfactual untreated outcomes as missing data and estimate them via matrix completion with the nuclear norm penalty \(\text{Mazumder et al.} \ 2010\). \text{Athey et al.} \ (2018) non-Bayesian approach does not have an estimator of confidence intervals.

The remainder of the study is structured as follows. In Section 2, we introduce a new Bayesian approach to causal inference with panel data and compare it with the existing alternatives. In Section 3, we illustrate the proposed approach by applying it to simulated and real data. We conduct a simulation study and show that our proposal is competitive with the existing approaches in terms of the precision of the predictions of counterfactual untreated outcomes. Then, the proposal is applied to the evaluation of the tobacco control program implemented in California 1988. The last section concludes the study.

---

1See, e.g., \text{Imbens and Rubin} \ (2015); \text{Athey and Imbens} \ (2017); \text{Abadie and Cattaneo} \ (2018).

2See \text{Abadie (forthcoming)} for a recent overview of the literature on SCMs.
2 Proposed Approach

2.1 Framework

An individual outcome is modeled as follows: for \( j = 1, \ldots, J; t = 1, \ldots, T \),

\[
y_{j,t} (s_{j,t}) = \theta_{j,t} + x_{j,t}^\top \beta + u_{j,t},
\]

where \( \theta_{j,t} \) is a unit- and time-specific intercept, \( x_{j,t} \) is an \( L \)-dimensional vector of covariates which may contain unit- and/or time-specific effects, \( \beta \) is the corresponding coefficient vector, and \( u_{j,t} \) is an error term. We focus on cases having a single type of treatment. Let \( S_1 \) and \( S_0 \) be sets of indices for treated and untreated observations, respectively. It is assumed that there is no interference between units. The covariates \( x_{j,t} \) are completely observed for all the units and periods. Let \( Y \) be a \( J \times T \) matrix composed of (actually) untreated outcomes, \( y_{j,t} (0), (j, t) \in S_0 \), and counterfactual untreated outcomes that are actually treated, \( y_{j,t} (0), (j, t) \in S_1 \).

We treat the latter elements as missing observations and infer them via matrix completion. Let \( \Xi = (\xi_{j,t}) \) with \( \xi_{j,t} = x_{j,t}^\top \beta \). Then the model can be posed in a matrix representation as

\[
Y = \Theta + \Xi + U,
\]

where \( U = (u_{j,t}) \).

A matrix of untreated outcomes \( Y \) can be structured flexibly. For instance, when only the \( J \)th unit is affected by the treatment for the last \( T_1 \) periods as in the standard SCM, \( Y \) is specified as

\[
Y = \begin{pmatrix}
y_{1,1} (0) & \cdots & y_{1,T_0} (0) & y_{1,T_0+1} (0) & \cdots & y_{1,T} (0) \\
\vdots & & \vdots & \vdots & & \vdots \\
y_{J,1} (0) & \cdots & y_{J,T_0} (0) & y_{J,T_0+1} (0) & \cdots & y_{J,T} (0)
\end{pmatrix},
\]

where \( \sqrt{\cdot} \) denotes a missing entry. It is possible to allow more than one treated unit:

\[
Y = \begin{pmatrix}
y_{1,1} (0) & \cdots & y_{1,T_0} (0) & y_{1,T_0+1} (0) & \cdots & y_{1,T_0+1} (0) \\
\vdots & & \vdots & \vdots & & \vdots \\
y_{J,1} (0) & \cdots & y_{J,T_0} (0) & y_{J,T_0+1} (0) & \cdots & y_{J,T_0+1} (0) \\
y_{J_0,1} (0) & \cdots & y_{J_0,T_0} (0) & y_{J_0,T_0+1} (0) & \cdots & y_{J_0,T_0+1} (0) \\
\vdots & & \vdots & \vdots & & \vdots \\
y_{J,T_1} (0) & \cdots & y_{J,T_0} (0) & \sqrt{\cdot} & \cdots & \sqrt{\cdot}
\end{pmatrix},
\]

Furthermore, it is possible to handle a more complex structure:

\[
Y = \begin{pmatrix}
\sqrt{\cdot} & \cdots & \sqrt{\cdot} & y_{1,e',0} (0) & \cdots & \cdots & \cdots & y_{1,T} (0) \\
\vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \vdots \\
y_{j,1} (0) & \sqrt{\cdot} & y_{j,e',+1} (0) & \sqrt{\cdot} & y_{j,e'+1} (0) & \cdots & \cdots & y_{j,T} (0) \\
\vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \vdots \\
y_{j',1} (0) & \cdots & y_{j',e'} (0) & \sqrt{\cdot} & \sqrt{\cdot} & y_{j',e'+3} (0) & \cdots & y_{j',T} (0) \\
\vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \vdots \\
y_{J,1} (0) & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & y_{J,T} (0)
\end{pmatrix}.
\]
To allow serial correlations in the error terms, their distribution is specified as
\[ u_j = (u_{j,1}, \ldots, u_{j,T})^\top \sim \mathcal{N}(0_r, \mathbf{Y}), \]
\[ \mathbf{Y} = \sigma^2 \mathbf{R}, \quad \mathbf{R} = (r_{t,t'}), \quad \text{with} \quad r_{t,t'} = \rho^{|t-t'|}, \]
where \( \sigma^2 \) is a variance parameter and \( \mathbf{R} \) is a correlation matrix whose generic element \( r_{t,t'} \) is specified as a function of an autocorrelation parameter \( \rho \in (-1, 1) \).

We define sets of observed and unobserved untreated outcomes as \( \mathcal{Y}^{\text{obs}} = \{y_{j,t} (0), (j, t) \in \mathcal{S}_0\} \) and \( \mathcal{Y}^{\text{miss}} = \{y_{j,t} (0), (j, t) \in \mathcal{S}_1\} \), respectively. The elements in the latter set are treated as unknown parameters. Then, the likelihood is represented as
\[
p(Y^{\text{obs}}|Y^{\text{miss}}, \Theta, \beta, \rho, \sigma^2; \mathbf{X}) = (2\pi)^{-\frac{HT}{2}} |\mathbf{Y}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \text{tr} \left( \mathbf{U}^\top \mathbf{U} \mathbf{Y}^{-1} \right) \right\}
\]
\[ \mathbf{U} = \mathbf{Y} - \Theta - \Xi. \]

### 2.2 Bayesian inference

We use a prior that induces \( \Theta \) to be low rank. We decompose \( \Theta \) into two parts as
\[ \Theta = \Phi \Psi^\top, \quad \text{with} \quad \Phi \in \mathbb{R}^{J \times H}, \quad \Psi \in \mathbb{R}^{T \times H}. \]

Setting \( H < \min (J, T) \) induces \( \Theta \) to be rank deficient a priori, although one can use an arbitrarily large \( H \). As the decomposition is not unique, some restriction is necessary for identification. In the study, we restrict \( \Psi \) to be unitary, i.e., \( \Psi^\top \Psi = I_H \), and assign a uniform Haar prior to \( \Psi \), \( p(\Psi) \propto \mathbb{1}(\Psi \in \mathcal{M}_{T \times H}) \), where \( \mathcal{M}_{T \times H} \) denotes a Stiefel manifold with dimensions of \( T \times H \) and \( \mathbb{1}(\cdot) \) denotes the indicator function. Let the singular value decomposition of \( \Theta \) be represented as \( \Theta = ED \Psi^\top \). Then, \( \Psi \) is interpreted as the right orthonormal matrix, while \( \Phi = ED \) is the product of the left orthonormal matrix \( E \) and the diagonal matrix having the eigenvalues in its principal diagonal \( D \).

For \( \Phi \), we adapt a cumulative shrinkage process prior (Legramanti et al., 2019) to our context. A prior of \( \Phi \) is specified by the following hierarchy:
\[
\phi_{j,h} | \lambda_h \sim \mathcal{N}(0, \lambda_h^2) \), \quad j = 1, \ldots, J; \quad h = 1, \ldots, H, \\
\lambda_h | \pi_h \sim \pi_h \delta_{\lambda_{\infty}} + (1 - \pi_h) \mathcal{IG}(\kappa_1, \kappa_2) \), \quad h = 1, \ldots, H, \\
\pi_h = \sum_{l=1}^{H} \omega_l, \quad \text{with} \quad \omega_l = \zeta_l \prod_{m=1}^{l-1} (1 - \zeta_m), \quad h = 1, \ldots, H, \\
\zeta_h \sim \mathcal{B}(1, \eta), \quad h = 1, \ldots, H - 1, \\
\zeta_H = 1,
\]
where \( \mathcal{IG}(a, b) \) is an inverse gamma distribution with shape \( a \) and rate \( b \), and \( \mathcal{B}(a, b) \) is a beta distribution (of the first kind) with scale parameters \( a \) and \( b \). The prior of \( \phi_{j,h} \) is a scale mixture of normal distributions. The prior distribution of the variances \( \lambda_h \) belongs to a class of spike-and-slab priors (e.g., Ishwaran et al., 2005), in that the prior consists of spike \( \delta_{\lambda_{\infty}} \) and slab \( \mathcal{IG}(\kappa_1, \kappa_2) \) parts. Although \( \delta_{\lambda_{\infty}} \) can be zero, we set it to a small non-zero value for the ease of posterior simulation (Ishwaran et al., 2005; Legramanti et al., 2019). The prior distribution of the weights \( \pi_h \) exploits the stick-breaking construction of the Dirichlet process (Ishwaran and...
As $h$ grows, the distribution of $\lambda_h$ concentrates around $\delta_{\lambda_h}$ since $\lim_{h \to \infty} \pi_h = 1$ almost surely.

Although we do not consider it in the study, [Bhattacharya and Dunson (2011)] consider a prior similar to the cumulative shrinkage process prior, called the multiplicative gamma process prior. This prior cannot control the rate of shrinkage and the prior for the active elements simultaneously. Thus, it readily overshrinks the model. See also [Durante (2017) and Legramanti et al. (2019)] for further discussion.

There are many fully Bayesian methods for estimating low-rank matrices (e.g., [Salakhutdinov and Mnih, 2008; Paisley and Carin, 2009; Ding et al., 2011; Babacan et al., 2012; Valera and Ghahramani, 2014; Fazayeli et al., 2014; Tang et al., 2019]). Compared with them, the above-mentioned prior construction method has two distinct features. First, the prior we use imposes the unitary restriction on $\Psi$, while the existing Bayesian approaches do not. The only exception is [Tang et al. (2019)] who decompose a matrix into three parts as in a standard singular value decomposition, that is, a diagonal matrix and two unitary matrices. Second, while the other Bayesian approaches to low-rank matrix estimation do not care about ordering of the eigenvalues, our prior specification progressively shrinks the prior variances of $\phi_{j,h}$ as $h$ increases. Although it does not impose an exactly monotone restriction, it can aid in the parameter identification.

In turn, for the remaining parameters, we employ standard priors. For $\beta$, we use an independent normal prior with mean zero and precision $\gamma$, $\beta \sim \mathcal{N}(0_L, \gamma^{-1}I_L)$. $\rho$ is assumed to be distributed according to a uniform distribution with support of $(-1, 1)$, $\rho \sim \mathcal{U}(-1, 1)$. A prior distribution of $\sigma^2$ is specified by an inverse gamma distribution, $\sigma^2 \sim \mathcal{IG}(\nu_1, \nu_2)$.

For posterior simulation, we develop a Markov chain Monte Carlo sampler that is a hybrid of three algorithms. For $\rho$, we use a random walk Metropolis–Hastings algorithm, while for $\Psi$ we employ the geodesic Monte Carlo on embedded manifolds (Byrne and Girolami, 2013). As the conditionals of the remaining parameters are standard, they are updated via Gibbs steps. See the Appendix for the computational details.

## 2.3 Comparison to existing approaches

We compare our proposal to existing approaches. For simplicity, in this section, we assume that there is no covariate and that only the $J$th unit is exposed to treatment during the last $T_1$ periods. Let $T_0$ denote the length of the pre-treatment periods (thus $T = T_0 + T_1$). Then, the set of indices of the treated observations is

$$ S_1 = \{(J, T_0 + 1), (J, T_0 + 2), \ldots, (J, T)\}.$$

Of the statistical methods for estimating ATE, a class of SCMs ([Abadie and Gardeazabal, 2003; Abadie et al., 2010]) is particularly related to the proposed approach. In SCMs, “synthetic” untreated outcomes are estimated as weighted sums of the untreated units. The original SCM (Abadie et al., 2010) specifies an estimation problem as

$$ \min_{\alpha_1, \ldots, \alpha_{J-1}} \sum_{t=1}^{T_0} \left( y_{J,t} (0) - \sum_{j=1}^{J-1} \alpha_j y_{j,t} (0) \right)^2, $$

subject to $\alpha_j \geq 0$, $\forall j$, $\sum_{j=1}^{J-1} \alpha_j = 1.$
This approach imposes three strong assumptions: no-intercept, non-negativity of the weights, and sum-to-one constraints on the weights. However, none of these assumptions seem to be plausible in many real cases. The proposed approach is free from such restrictions.

Doudchenko and Imbens (2017) propose an approach that does not impose any of these restrictions on the weights to be estimated. Instead, they propose estimating synthetic observations by solving the following optimization problem, which corresponds to the elastic net estimator (Zou and Hastie, 2005):

\[
\min_{\alpha_0, \alpha_1, \ldots, \alpha_{J-1}} \sum_{t=1}^{T} \left( y_{J,t} - \alpha_0 - \sum_{j=1}^{J-1} \alpha_j y_{j,t} \right)^2 + \varphi \left( \frac{1 - \vartheta}{2} \sum_{j=1}^{J-1} \alpha_j^2 + \vartheta \sum_{j=1}^{J-1} |\alpha_j| \right), \\
\text{(3)}
\]

where \( \varphi, \vartheta \geq 0 \) are tuning parameters. All the existing non-Bayesian approaches, including Abadie et al. (2010) and Doudchenko and Imbens (2017), share the same caveat: there is no statistically grounded method to estimate confidence intervals. Abadie et al. (2010) conduct a series of placebo studies: the original SCM is iteratively applied to every other state in replacing with California. Indeed, such an exercise can provide some notion of uncertainty about estimation, but it is a way of a robustness check, not a serious statistical method for hypothesis testing. This caveat is critical because hypothesis testing is an essential part of statistical inference in scientific research.

Kim et al. (2019) develop a Bayesian version of Doudchenko and Imbens’s (2017) approach. Instead of the elastic net penalty, they propose to use shrinkage priors. For instance, when the horseshoe prior (Carvalho et al., 2010) is employed, the probabilistic representation of a model considered in Kim et al. (2019) is specified as follows:

\[
y_{J,t} | \alpha_0, \alpha_1, \ldots, \alpha_{J-1}, \sigma^2 \sim N \left( \alpha_0 + \sum_{j=1}^{J-1} \alpha_j y_{j,t}, \sigma^2 \right), \quad t = 1, \ldots, T, \\
\text{(4)}
\]

\[
\alpha_j | \varphi, \vartheta_j \sim N \left( 0, \sigma^2 \varphi^2 \vartheta_j^2 \right), \quad j = 1, \ldots, J - 1, \\
\varphi \sim C^+ (0, 1), \\
\vartheta_j \sim C^+ (0, 1), \quad j = 1, \ldots, J - 1, \\
\alpha_0 \sim f_{\alpha_0}, \quad \sigma^2 \sim f_{\sigma^2}, \\
\text{(5)-(7)}
\]

where \( f_{\alpha_0} \) and \( f_{\sigma^2} \) are prior distributions of \( \alpha_0 \) and \( \sigma^2 \), respectively and \( C^+ (0, 1) \) denotes a standard half-Cauchy distribution with a probability density function (PDF).

\[
p(x) = \frac{2}{\pi (1 + x^2)}, \quad x > 0.
\]

As with the Bayesian inference, their approach can obtain credible intervals consistently. Our fully Bayesian approach also enjoys the same advantage.

Athey et al. (2018) develop a novel approach to causal inference with panel data. In a sense, our proposal can be seen as a Bayesian counterpart of their approach. As in Section 2.2, let \( Y \) denote a matrix composed of untreated outcomes, and observations exposed to treatment are handled as missing. Then, an estimation problem can be posed as a kind of matrix completion problem:\(^3\)

\[
\min_{\Theta} \frac{1}{2} \| P_{S_1} (Y - \Theta) \|_F + \varphi \| \Theta \|_k, \\
\text{(9)}
\]

\(^3\)See, e.g., Candes and Plan (2010) and Shi et al. (2017) for a survey.
where $\varphi \geq 0$ is a tuning parameter and $\|\cdot\|_F$ denotes the Frobenius norm. $\|A\|_*$ denotes the nuclear norm of a matrix $A$ defined as

$$\|A\|_* = \sum_i \sigma_i (A),$$

where $\sigma_i (A)$ is the $i$th eigenvalue of $A$. When $O$ denotes a set of indices corresponding to observed entries, the operator $P_O (\cdot)$ is defined for a matrix $A = (a_{n,m})$ as

$$P_O (A) = (\tilde{a}_{n,m}), \quad \tilde{a}_{n,m} = \begin{cases} a_{n,m} & (n, m) \in O \\ 0 & (n, m) \notin O \end{cases}.$$ 

Athey et al. (2018) call this estimator a matrix completion with a nuclear norm minimization estimator (MC-NNM). The penalty term induces $\Theta$ to be lower rank as the nuclear norm is a convex relaxation to the rank constraint (Fazel et al., 2001). The prior of $\Theta$ used in the proposed approach plays a similar role to the nuclear norm penalty. This family of approaches involving matrix completion has two notable advantages over a family of SCMs. First, a treatment is allowed to occur arbitrarily, not consecutively. Second, while SCMs only use the pre-treatment observations for estimation, this family exploits all the observations including the treated periods (except treated outcomes). Therefore, this class is likely to be statistically more efficient than SCMs, as shown in the simulation study below. In common with other non-Bayesian approaches, Athey et al.’s (2018) approach only provides a point estimation, while our proposal readily estimates credible intervals.

Lastly, Brodersen et al. (2015) also develop a Bayesian approach to estimate ATE that uses structural time series models, more specifically, state-space models. Their approach relies on the fit of a state-space model for the single time series of a treated unit. Their and our approaches are similar in that they tend to obtain counterfactual outcomes using Bayesian methods. On the other hand, they are targeting different types of data: Brodersen et al.’s (2015) approach is better suited for relatively long time series, while the proposed approach is designed for typical panel data with rather short time periods.

### 3 Application

#### 3.1 Simulated data

We conduct a simulation study to demonstrate the proposed approach. Only the $J$th unit is treated, and it is exposed to the treatment during the last $T_1$ periods of the total time periods $T$. Let $T_0$ denote the number of untreated periods; thus, $T = T_0 + T_1$. The realized treated outcomes are specified by the sums of hypothetical untreated outcomes $y_{J,t}(0)$ and the average treatment effect $ATE$:

$$y_{J,t}(1) = y_{J,t}(0) + ATE, \quad t = T_0 + 1, \ldots, T.$$ 

$y_{J,t}(0)$ is generated from a factor model: for $j = 1, \ldots, J; t = 1, \ldots, T$,

$$(y_{1,t}(0), \ldots, y_{J,t}(0))^\top = y_t(0) = \Phi \psi_t + u_t,$$

$$\Psi = (\psi_1, \ldots, \psi_T)^\top,$$

$$u_t = (u_{1,t}, \ldots, u_{J,t})^\top \sim N(0_J, I_J).$$
We do not include any covariate and assume no autocorrelation in the error term $u_{j,t}$, fixing $\rho$ to zero (not estimated). We consider two types of data-generating processes (DGPs), named DGP-independent and DGP-dependent. They only differ in how $\Psi$ is generated. In the DGP-independent case, the “latent factors” are independently distributed according to a uniform distribution specified as

$$
\psi_t \sim \mathcal{N}(0_3, I_3), \quad t = 1, \ldots, T.
$$

In the DGP-dependent case, the row of motion of $\psi_t$ is specified by a vector autoregressive process:

$$
\begin{align*}
\psi_1 &= \epsilon_1, \\
\psi_t &= \begin{pmatrix} 0.6 & 0.2 & 0.0 \\ 0.2 & 0.4 & 0.1 \\ 0.0 & 0.1 & 0.2 \end{pmatrix} \psi_{t-1} + \epsilon_t, \quad t = 2, 3, \ldots, T, \\
\epsilon_t &\sim \mathcal{N}(0_3, I_3), \quad t = 1, \ldots, T.
\end{align*}
$$

Entries in the “factor loading” $\Phi$ are generated independently from a standard normal distribution:

$$
\Phi = (\phi_{j,t}), \quad \text{with } \phi_{j,t} \sim \mathcal{N}(0, 1).
$$

We compare the proposed approach to four alternatives. The first is the original synthetic control method described in Abadie et al. (2010) (SCM-Plain), where an estimation problem is specified as (1)–(2). The second is a method proposed by Doudchenko and Imbens (2017) (SCM-EN) (3), and the third is a matrix completion with a nuclear norm minimization estimator (Athey et al. 2018) (MC-NNM) (9). In SCM-EN and MC-NNM, the tuning parameters $\varphi$ and $\vartheta$ are chosen using five-fold cross-validation. The fourth is a Bayesian SCM developed by Kim et al. (2019). According to their simulation study, specifications with the horseshoe (Carvalho et al., 2010) and spike-and-slab (Ishwaran et al., 2005) priors outperform other alternatives. While the performances of these two priors are fairly comparable, posterior simulation using the horseshoe prior is faster. For these reasons, we consider the horseshoe prior for Kim et al.’s (2019) approach and refer to this specific approach as BSCM-HS in what follows. The probabilistic representation of BSCM-HS is given by (4)–(8). While Kim et al. (2019) use non-informative priors for $\alpha_0$ and $\sigma^2$, we employ vague, proper priors. For $\alpha_0$, we use a normal prior with large variance, $\alpha_0 \sim \mathcal{N}(0, 10^4)$. We assign the same prior to $\sigma^2$ as the proposed approach, $\sigma^2 \sim IG(\nu_1, \nu_2)$. We choose $\nu_1 = \nu_2 = 0.001$, inducing the prior of $\sigma^2$ to be fairly non-informative. We sample $\alpha_0, \alpha_1, \ldots, \alpha_{J-1}$ using the elliptical slice sampler of Hahn et al. (2019) and the remaining parameters using Gibbs sampler, as in Makalic and Schmidt (2016). We simulate 10,000 warmup draws and estimate the posterior densities using the subsequent 40,000 draws.

For the proposed approach, following Legramanti et al. (2019), the prefixed hyperparameters for the cumulative shrinkage process prior are chosen as $\eta = 5$ and $\kappa_1 = \kappa_2 = 2$. While Legramanti et al. (2019) use $\delta_\infty = 0.05$, we set it to a smaller value, $\delta_\infty = 0.0001$. For a fair comparison to BSCM-HS, we use the same hyperparameters for the prior of $\sigma^2$, $\nu_1 = \nu_2 = 0.001$. The maximum rank of $\Theta$ is set to $H = \min(J, T)$, where $\lceil \rceil$ denotes the ceiling function. We obtain 40,000 draws after discarding the initial 10,000 draws.

We consider four types of sample size, namely, combinations of $J \in \{5, 20\}$ and $T_0 \in \{10, 50\}$, while the length of the treated periods is fixed to $T_1 = 20$. A total of 500 experiments are conducted for each case. As noted earlier, an estimation of ATE amounts to an estimation of
potential outcomes. Therefore, we evaluate the alternatives based on the precision of estimates of \( y_{j,t} (0), t = T_0 + 1, \ldots, T \), measured by the mean squared errors (MSE) and the absolute errors (MAE). For the Bayesian approaches, we compute posterior means of predicted potential outcomes. We also report computation times measured in seconds (Time). For each experiment, MSE and MAE are normalized by the corresponding values for SCM-Plain, (i.e., the smaller, the better than SCM-Plain, and vice versa). Medians of the performance measures over the experiments are calculated.

Table 1 summarizes the results of the simulation study for DGP-independent. In terms of MSE and MAE, irrespective of combinations of \((J, T_0)\), the proposed approach consistently outperforms the others as well as SCM-Plain. The recently proposed alternatives are comparable with SCM-Plain, and are not always better than the original. In terms of computational time in seconds, as expected, Bayesian approaches are slower than non-Bayesian options. Indeed BMC-CSP is computationally heavy, but the computational cost is not prohibitive. In our simulation study, SCM-EN is rather slow to converge, possibly due to non-smooth objective functions.

The simulation results for DGP-dependent are reported in Table 2. In terms of MSE and MAE, SCM-EN, MC-NNM, and BMC-SCP perform better than SCM-Plain, while BSCM-HS tends to be inferior to the others. For this DGP, approaches using matrix completion are likely to be more precise than a family of SCMs. BMC-CSP is slightly better than MC-NNM, being the absolute winner.

### 3.2 Real data

As an illustration, we apply the proposed approach to the evaluation of California’s tobacco control program implemented in 1988. We replicate Abadie et al.’s (2010) study using the same data, annual state-level panel spans from 1970 to 2000. The first 19 years are pre-treatment periods. Only California is treated, while the other 38 states are used as control units. We include seven time-invariant covariates: log of gross domestic product per capita, percentage share of 15–24-year-old people in the population, retail price, beer consumption per capita, and cigarette sales per capita in 1980 and 1975. See Abadie et al. (2010) for further details. We choose the hyperparameters as in the simulation study. We draw 50,000 posterior samples and use the last 40,000 samples for posterior analysis.

Figure 1 compares the realized per-capita cigarette sales in California (solid black line), the potential per-capita cigarette sales in “synthetic California” obtained using the original SCM (Abadie et al., 2010) (dashed black line), and the posterior mean estimates of the corresponding counterfactual outcomes obtained by the proposed method (solid red line). The estimates using the proposed method are in line with the estimates using the original SCM. Posterior estimates of 90% and 70% credible sets are also reported (shaded areas). As the credible sets do not include the realized California, we see that the program has statistically significant effects on tobacco consumption in California, confirming the conclusion in the original paper.

### 4 Concluding Remarks

This study develops a novel Bayesian approach to causal analysis using panel data. We treat a problem of inferring average treatment effect as a matrix completion problem: counterfac-

---

4 We wrote all the programs in Matlab R2019a (64 bit) and executed them on an Ubuntu Desktop 18.04 LTS (64 bit), running on Intel Xeon E5-2607 v3 processors (2.6 GHz).

5 The data and the Matlab program were downloaded from Jens Hainmueller’s personal website. (https://web.stanford.edu/~jhain/synthpage.html)
tual untreated outcomes are inferred using a data augmentation technique. We also propose a prior structured to ensure identification and to obtain a low-rank approximation of the panel data. While existing non-Bayesian methods cannot deliver confidence intervals, the proposed Bayesian approach can estimate credible intervals straightforwardly. Through a series of simulation studies, we show that the proposed approach outperforms the existing ones in terms of the prediction of hypothetical untreated outcomes, that is, the accuracy of estimation of ATE.

Appendix: Computational Details

This appendix describes the computational details of the posterior simulation of the proposed approach. Each sampling block is specified in what follows.

Sampling $\Phi$

For $j = 1, ..., J$,

$$
\phi_{(j)} | - \sim N \left( m_{\phi_{(j)}}, P_{\phi_{(j)}}^{-1} \right),
$$

$$
m_{\phi_{(j)}} = P_{\phi_{(j)}}^{-1} \Psi^\top Y^{-1} (y_{(j)} - \xi_{(j)}),
$$

$$
P_{\phi_{(j)}} = \text{diag} \left( \lambda_{1}^{-1}, ..., \lambda_{H}^{-1} \right) + \Psi^\top Y^{-1} \Psi,
$$

$$
Y = (y_{(1)}, ..., y_{(J)})^\top, \quad \Xi = (\xi_{(1)}, ..., \xi_{(J)})^\top.
$$

Sampling the shrinkage parameters

$$
p \left( z_{h} = l | - \right) \propto \omega_{l} N \left( \phi_{h} | 0_{J}, \lambda_{\infty} I_{J} \right) \quad l = 1, ..., h,
$$

$$
p \left( z_{l} > l | - \right) \propto \omega_{l} t_{c} \left( \phi_{h} | 0_{J}, \frac{\kappa_{2}}{\kappa_{1}} I_{J} \right) \quad l = h + 1, ..., H,
$$

where $N \left( x | a, B \right)$ is the PDF of a multivariate normal distribution with mean $a$ and covariance $B$ evaluated at $x$, and $t_{c} \left( x | a, B, c \right)$ is the PDF of a multivariate t distribution with location $a$, scale $B$, and $c$ degrees of freedom.

$$
\zeta_{l | -} \sim B \left( 1 + \sum_{h=1}^{H} \mathbb{I} \left( z_{h} = l \right), \quad \eta + \sum_{h=1}^{H} \mathbb{I} \left( z_{l} > l \right) \right), \quad l = 1, ..., H - 1,
$$

$$
\lambda_{h} \sim \mathbb{I} \left( z_{h} \leq h \right) \delta_{\theta_{h}} + \left( 1 - \mathbb{I} \left( z_{h} \leq h \right) \right) IG \left( \kappa_{1} + \frac{J}{2}, \quad \kappa_{2} + \frac{1}{2} \sum_{j=1}^{H} \phi_{j,h}^{2} \right),
$$

where $\mathbb{I} \left( \cdot \right)$ denotes the indicator function.

Sampling $\Psi$

To sample $\Psi$, we employ the geodesic Monte Carlo on embedded manifolds developed by Byrne and Girolami (2013). The algorithm for sampling $\Psi$ is summarized in Algorithm 1. Let $\pi \left( \Psi \right)$ be the posterior density of $\Psi$ conditional on the other parameters. Then the gradient with respect to $\Psi$ is derived as

$$
\nabla_{\Psi} \log \pi \left( \Psi \right) = - \left[ - \Psi^{-1} (Y - \Xi)^\top \Phi + \Psi^{-1} \Phi \Psi^\top \Phi \right].
$$
The step size $\varepsilon$ is adaptively tuned so as to keep the average acceptance rate being around a target value $a^*$. At the $i$th iteration, $\varepsilon$ is updated as follows:

$$
\log(\varepsilon) \leftarrow \log(\varepsilon) + i^{-1/\zeta} (a^* - \bar{a}_i),
$$

where $\bar{a}_i$ is the average acceptance rate at the $i$th iteration, and $\zeta \in (0.5, 1)$ is a tuning parameter. We choose $a^* = 0.6$ and $\zeta = 0.6$.

### Sampling $\beta$

$$
|\beta| \sim \mathcal{N}(m_\beta, P_\beta^{-1}),
$$

$$
m_\beta = P_\beta^{-1} X^\top (Y^{-1} \otimes I_J) \text{vec}(Y - \Theta),
$$

$$
P_\beta = \gamma I_L + X^\top (Y^{-1} \otimes I_J) X,
$$

$$
X = \left( \begin{array}{c} X_1 \\ \vdots \\ X_J \end{array} \right), \quad \text{with } X_j = \left( \begin{array}{c} x_{j,1}^\top \\ \vdots \\ x_{j,T}^\top \end{array} \right),
$$

where $\text{vec}(\cdot)$ denotes the column-wise vectorization operator and $\otimes$ denotes the Kronecker product.

### Sampling $\sigma^2$

$$
\sigma^2 \sim IG\left(\nu_1 + \frac{JT}{2}, \nu_2 + \frac{1}{2} \text{vec}(\bar{U})^\top (R^{-1} \otimes I_J) \text{vec}(\bar{U})\right),
$$

$$
\bar{U} = Y - \Theta - \Gamma - \Delta - \Xi.
$$

### Sampling $\rho$

As the conditional posterior of $\rho$ does not belong to a class of canonical distributions, we update $\rho$ using a random walk Metropolis–Hastings algorithm with adaptation (Haario et al., 2001). At the $i$th iteration, given a current state $\rho_{(i-1)}$, a proposal $\rho'$ is generated from a normal distribution $\rho' \sim \mathcal{N}(\rho_{(i-1)}, v_\rho)$. The proposal variance $v_\rho$ is adaptively tuned as $\varepsilon$. The target acceptance rate is set to 0.25.

### Sampling $Y_{\text{miss}}$

Without loss of generality, we focus on a case where the $J$th unit is treated during the last $T_1 = T - T_0$ periods, $Y_{\text{miss}} = \{y_{J,T_0+1}, ..., y_{J,T}\}$. Sampling procedures for other cases are analogously constructed. Given all the parameters, the distribution of the $J$th row of $Y$ is represented as

$$
y_{(J)} = (y_{J,1}, ..., y_{J,T})^\top \sim \mathcal{N}(\mu, \Upsilon)
$$

$$
\mu = \theta_{(J)} + \xi_{(J)},
$$

where $\theta_{(J)}$ and $\xi_{(J)}$ are the $J$th rows of $\Theta$ and $\Xi$, respectively. We define partitionings as

$$
y_{\text{obs}}_{(J)} = (y_{J,1}, ..., y_{J,T_0})^\top, \quad y_{\text{miss}}_{(J)} = (y_{J,T_0+1}, ..., y_{J,T})^\top,
$$

$$
\mu = \left( \begin{array}{c} \mu_1 \\ \mu_2 \end{array} \right), \quad \Upsilon = \left( \begin{array}{cc} \Upsilon_{1,1} & \Upsilon_{1,2} \\ \Upsilon_{2,1} & \Upsilon_{2,2} \end{array} \right),
$$

11
where $\mu_1 \in \mathbb{R}^{T_0}$, $\mu_2 \in \mathbb{R}^{T_1}$, $\Upsilon_{1,1} \in \mathbb{R}^{T_0 \times T_0}$, $\Upsilon_{1,2} \in \mathbb{R}^{T_0 \times T_1}$, and $\Upsilon_{2,2} \in \mathbb{R}^{T_1 \times T_1}$. Then, the conditional posterior distribution of the missing observations is

$$y_{(j)}^{\text{miss}} | - \sim \mathcal{N} \left( m_{y_{(j)}^{\text{miss}}}, V_{y_{(j)}^{\text{miss}}} \right),$$

$$m_{y_{(j)}^{\text{miss}}} = \mu_2 + \Upsilon_{1,2}^{-1} \Upsilon_{1,1}^{-1} \left( y_{(j)}^{\text{obs}} - \mu_1 \right),$$

$$V_{y_{(j)}^{\text{miss}}} = \Upsilon_{2,2} - \Upsilon_{1,2} \Upsilon_{1,1}^{-1} \Upsilon_{1,2}.$$
Ding, X., L. He, and L. Carin (2011) “Bayesian Robust Principal Component Analysis,” *IEEE Transactions on Image Processing*, Vol. 20, No. 12, 3419–3430.

Doudchenko, N. and G. W. Imbens (2017) “Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis,” Technical report, arXiv:1610.07748.

Durante, D. (2017) “A Note on the Multiplicative Gamma Process,” *Statistics and Probability Letters*, Vol. 122, 198–204.

Fazayeli, F., A. Banerjee, J. Kattge, F. Schrodt, and P. B. Reich (2014) “Uncertainty Quantified Matrix Completion Using Bayesian Hierarchical Matrix Factorization,” in *2014 13th International Conference on Machine Learning and Applications*, 312–317, IEEE.

Fazel, M., H. Hindi, and S. P. Boyd (2001) “A Rank Minimization Heuristic with Application to Minimum Order System Approximation,” in *Proceedings of the American Control Conference*, Vol. 6, 4734–4739, Citeseer.

Haario, H., E. Saksman, and J. Tamminen (2001) “An Adaptive Metropolis Algorithm,” *Bernoulli*, Vol. 7, No. 2, 223–242.

Hahn, P. R., J. He, and H. F. Lopes (2019) “Efficient Sampling for Gaussian Linear Regression with Arbitrary Priors,” *Journal of Computational and Graphical Statistics*, Vol. 28, No. 1, 142–154.

Holland, P. W. (1986) “Statistics and Causal Inference,” *Journal of the American Statistical Association*, Vol. 81, No. 396, 945–960.

Imbens, G. W. and D. B. Rubin (2015) *Causal Inference in Statistics, Social, and Biomedical Sciences*: Cambridge University Press.

Ishwaran, H. and L. F. James (2001) “Gibbs Sampling Methods for Stick-breaking Priors,” *Journal of the American Statistical Association*, Vol. 96, No. 453, 161–173.

Ishwaran, H., J. S. Rao et al. (2005) “Spike and Slab Variable Selection: Frequentist and Bayesian Strategies,” *Annals of Statistics*, Vol. 33, No. 2, 730–773.

Keshavan, R. H., A. Montanari, and S. Oh (2010) “Matrix Completion from Noisy Entries,” *Journal of Machine Learning Research*, Vol. 11, No. Jul, 2057–2078.

Kim, S., C. Lee, and S. Gupta (2019) “Bayesian Synthetic Control Methods,” Technical report, Cornell University.

Legramanti, S., D. Durante, and D. B. Dunson (2019) “Bayesian Cumulative Shrinkage for Infinite Factorizations,” Technical report, arXiv:1902.04349.

Makalic, E. and D. F. Schmidt (2016) “A Simple Sampler for the Horseshoe Estimator,” *IEEE Signal Processing Letters*, Vol. 23, No. 1, 179–182.

Mazumder, R., T. Hastie, and R. Tibshirani (2010) “Spectral Regularization Algorithms for Learning Large Incomplete Matrices,” *Journal of Machine Learning Research*, Vol. 11, No. Aug, 2287–2322.
Paisley, J. and L. Carin (2009) “Nonparametric Factor Analysis with Beta Process Priors,” in Proceedings of the 26th Annual International Conference on Machine Learning, 777–784, ACM.

Salakhutdinov, R. and A. Mnih (2008) “Bayesian Probabilistic Matrix Factorization Using Markov Chain Monte Carlo,” in Proceedings of the 25th international Conference on Machine learning, 880–887, ACM.

Shi, J., X. Zheng, and W. Yang (2017) “Survey on Probabilistic Models of Low-rank Matrix Factorizations,” Entropy, Vol. 19, No. 8, p. 424.

Srebro, N., J. D. M. Rennie, and T. S. Jaakkola (2005) “Maximum-margin Matrix Factorization,” in Advances in Neural Information Processing Systems, 1329–1336.

Tang, K., Z. Su, J. Zhang, L. Cui, W. Jiang, X. Luo, and X. Sun (2019) “Bayesian Rank Penalization,” Neural Networks, Vol. 116, 246–256.

Tanner, M. A. and W. H. Wong (1987) “The Calculation of Posterior Distributions by Data Augmentation,” Journal of the American Statistical Association, Vol. 82, No. 398, 528–540.

Valera, I. and Z. Ghahramani (2014) “General Table Completion Using a Bayesian Nonparametric Model,” in Advances in Neural Information Processing Systems, 981–989.

Xu, Y. (2017) “Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models,” Political Analysis, Vol. 25, No. 1, 57–76.

Zou, H. and T. Hastie (2005) “Regularization and Variable Selection Via the Elastic Net,” Journal of the Royal Statistical Society: Series B, Vol. 67, No. 2, 301–320.
Algorithm 1 Geodesic Monte Carlo

Input: $\Phi_0$ (current state), $\pi(\Phi)$ (target kernel), $\epsilon$ (step size), $N_{\text{step}}$ (number of steps)

vec($V$) $\sim$ $\mathcal{N}(0_{JH}, I_{JH})$

2: $V \leftarrow V - \frac{1}{2} \Phi_0 (\Phi_0^T V + V^T \Phi_0)$

3: $H_0 \leftarrow \log \pi(\Phi_0) - \frac{1}{2} \text{vec}(V)^T \text{vec}(V)$

4: $\Phi_1 \leftarrow \Phi_0$

5: for $i = 1, \ldots, N_{\text{step}}$ do

6: $V \leftarrow V + \frac{\epsilon}{2} \nabla_{\Phi_1} \log \pi(\Phi_1)$

7: $V \leftarrow V - \frac{1}{2} \Phi_1 (\Phi_1^T V + V^T \Phi_1)$

8: $A \leftarrow \Phi_1^T V$, $S \leftarrow V^T V$

9: $\begin{pmatrix} \Phi_1 & V \end{pmatrix} \leftarrow \begin{pmatrix} \Phi_1 & V \end{pmatrix} \exp \left( \epsilon \begin{pmatrix} A & -S \\ I_H & A \end{pmatrix} \right) \left( \exp \left( -\epsilon A \right) O_{H \times H} \right)$

10: $V \leftarrow V + \frac{\epsilon}{2} \nabla_{\Phi_1} \log \pi(\Phi_1)$

11: $V \leftarrow V - \frac{1}{2} \Phi_1 (\Phi_1^T V + V^T \Phi_1)$

end for

12: $H_1 \leftarrow \log \pi(\Phi_1) - \frac{1}{2} \text{vec}(V)^T \text{vec}(V)$

13: $w \sim \mathcal{U}(0, 1)$

14: if $w < \exp(H_1 - H_0)$ then

15: $\Phi \leftarrow \Phi_1$

16: end if

17: return $\Phi$

Note: $\exp(A)$ denotes the matrix exponential operator.
| $(J, T_0)$ | Approach   | MSE  | MAE  | Time |
|-----------|------------|------|------|------|
|           | SCM-Plain  | 1.00 | 1.00 | >0.1 |
|           | SCM-EN     | 1.14 | 1.06 | 5.8  |
| (5, 10)   | BSCM-HS    | 1.31 | 1.13 | 11.0 |
|           | MC-NNM     | 0.95 | 0.97 | 0.9  |
|           | BMC-CSP    | 0.89 | 0.95 | 123.4|
|           | SCM-Plain  | 1.00 | 1.00 | >0.1 |
|           | SCM-EN     | 1.18 | 1.07 | 20.5 |
| (5, 50)   | BSCM-HS    | 1.51 | 1.23 | 16.0 |
|           | MC-NNM     | 1.01 | 1.02 | 1.9  |
|           | BMC-CSP    | 0.88 | 0.94 | 637.0|
|           | SCM-Plain  | 1.00 | 1.00 | >0.1 |
|           | SCM-EN     | 0.99 | 0.99 | 4.8  |
| (20, 10)  | BSCM-HS    | 0.98 | 0.99 | 14.5 |
|           | MC-NNM     | 0.98 | 0.98 | 0.9  |
|           | BMC-CSP    | 0.91 | 0.96 | 148.4|
|           | SCM-Plain  | 1.00 | 1.00 | >0.1 |
|           | SCM-EN     | 1.04 | 1.02 | 22.7 |
| (20, 50)  | BSCM-HS    | 1.55 | 1.23 | 46.3 |
|           | MC-NNM     | 1.11 | 1.04 | 2.6  |
|           | BMC-CSP    | 0.94 | 0.96 | 848.8|

Note: MSE denotes the sum of the mean squared errors. MAE denotes the sum of the absolute errors. Reported values of MSE and MAE are normalized by the corresponding values for SCM-Plain. Time denotes computational time in seconds.
Table 2: Results of simulation study (2): DGP-dependent

| $(J, T_0)$ | Approach  | MSE | MAE | Time |
|------------|-----------|-----|-----|------|
| (5, 10)    | SCM-Plain | 1.00| 1.00| >0.1 |
|            | SCM-EN    | 0.85| 0.92| 5.9  |
|            | BSCM-HS   | 1.22| 1.10| 10.5 |
|            | MC-NNM    | 0.73| 0.85| 0.7  |
|            | BMC-CSP   | 0.70| 0.83| 124.1|
| (5, 50)    | SCM-Plain | 1.00| 1.00| >0.1 |
|            | SCM-EN    | 0.91| 0.96| 22.6 |
|            | BSCM-HS   | 1.55| 1.23| 16.9 |
|            | MC-NNM    | 0.86| 0.93| 1.4  |
|            | BMC-CSP   | 0.79| 0.89| 639.5|
| (20, 10)   | SCM-Plain | 1.00| 1.00| >0.1 |
|            | SCM-EN    | 0.77| 0.87| 5.4  |
|            | BSCM-HS   | 0.84| 0.91| 14.8 |
|            | MC-NNM    | 0.75| 0.85| 0.6  |
|            | BMC-CSP   | 0.73| 0.84| 150.7|
| (20, 50)   | SCM-Plain | 1.00| 1.00| >0.1 |
|            | SCM-EN    | 0.92| 0.95| 22.0 |
|            | BSCM-HS   | 1.60| 1.28| 43.4 |
|            | MC-NNM    | 0.89| 0.95| 1.9  |
|            | BMC-CSP   | 0.87| 0.93| 849.8|

Note: MSE denotes the sum of the mean squared errors. MAE denotes the sum of the absolute errors. Reported values of MSE and MAE are normalized by the corresponding values for SCM-Plain. Time denotes computational time in seconds.
Note: The solid black line traces the realized per-capita cigarette sales in California. The dashed black line and the solid red line trace the estimated potential per-capita cigarette sales using SCM-Plain and BMC-CSP, respectively. The light and dark shaded areas indicate 90% and 70% credible sets, respectively.