A Bayesian Alternative to Synthetic Control for Comparative Case Studies

Xun Pang[1], Licheng Liu[2] and Yiqing Xu[3]

[1] Department of International Relations, Tsinghua University, Beijing, China. E-mail: xpang@tsinghua.edu.cn
[2] Department of Political Science, Massachusetts Institute of Technology, Cambridge, MA, USA. E-mail: liulch@mit.edu
[3] Department of Political Science, Stanford University, Stanford, CA, USA. E-mail: yiqingxu@stanford.edu

Abstract

This paper proposes a Bayesian alternative to the synthetic control method for comparative case studies with a single or multiple treated units. We adopt a Bayesian posterior predictive approach to Rubin's causal model, which allows researchers to make inferences about both individual and average treatment effects on treated observations based on the empirical posterior distributions of their counterfactuals. The prediction model we develop is a dynamic multilevel model with a latent factor term to correct biases induced by unit-specific time trends. It also considers heterogeneous and dynamic relationships between covariates and the outcome, thus improving precision of the causal estimates. To reduce model dependency, we adopt a Bayesian shrinkage method for model searching and factor selection. Monte Carlo exercises demonstrate that our method produces more precise causal estimates than existing approaches and achieves correct frequentist coverage rates even when sample sizes are small and rich heterogeneities are present in data. We illustrate the method with two empirical examples from political economy.

Keywords: synthetic control, comparative case studies, panel data, TSCS data, causal inference, Bayesian statistics, stochastic model search, latent factor model.

1 Introduction

With the introduction of the synthetic control method (SCM) (Abadie and Gardeazabal 2003; Abadie, Diamond, and Hainmueller 2010), comparative case studies using time-series cross-sectional (TSCS) data, or long panel data, are becoming increasingly popular in the social sciences. Compared with other quantitative social science research, comparative case studies have several unique features: (1) the sample includes a small number of aggregate entities; (2) a handful of units, or one, receive an intervention that is not randomly assigned; and (3) the treatment effect often takes time to present itself (Abadie 2020). As a result, comparative case studies face two main challenges given limited data: to "provide good predictions of the [counterfactual] trajectory of the outcome" of the treated unit(s) (Abadie, Diamond, and Hainmueller 2015, p. 499, henceforth, ADH 2015); and to make credible statistical inferences about the treatment effects.

The SCM uses a convex combination of control outcomes to predict the treated counterfactuals. Inspired by the SCM, a fast-growing literature proposes various new methods to improve the SCM's counterfactual predictive performance and robustness, or to extend the SCM to accommodate multiple treated units. These methods can be broadly put into three categories: (1) matching or reweighting methods, such as best subset (Hsiao, Ching, and Wan 2012), regularized weights (Doudchenko and Imbens 2017), and panel matching (Imai, Kim, and Wang 2019); (2) explicit outcome modeling approaches, such as Bayesian structural time-series models (Brodersen *et al.* 2014), latent factor models (LFMs) (e.g., Bai 2009; Gobillon and Magnac 2016; Xu 2017), and matrix completion methods (Athey *et al.* 2018); and (3) doubly robust methods, such as the

Political Analysis (2022)
vol. 30: 269–288
DOI: 10.1017/pan.2021.22

Published
13 July 2021

Corresponding author
Yiqing Xu

Edited by
Daniel Hopkins

augmented SCM (Ben-Michael, Feller, and Rothstein 2020) and synthetic difference-in-differences (DiD) (Arkhangelsky *et al.* 2020).

However, both inference and prediction challenges are not fully addressed by existing methods. The SCM uses a placebo test as an inferential tool, but users cannot interpret it as a permutation test since the treatment is not randomly assigned (Hahn and Shi 2017). Hence, researchers cannot quantify the uncertainty of their estimates in traditional ways. Other frequentist inferential methods require a repeated sampling interpretation,[1] which is often at odds with the fixed population of units at the heart of many comparative case studies. Additionally, from a prediction perspective, researchers can use multiple sources of information in TSCS data for counterfactual prediction, including (1) temporal relationships between the known "past" and the unknown "future" of each unit, (2) cross-sectional information reflecting the similarity between units based on observed covariates, and (3) time-series relationships among units based on their outcome trajectories (Beck and Katz 2007; Pang 2010; Pang 2014). While better predictive performance can translate to more precise causal estimates, existing model-based approaches make relatively rigid parametric assumptions and therefore do not take full advantage of the information in data.

The Bayesian approach is an appealing alternative to meet these challenges. First, Bayesian uncertainty measures are easy to interpret. Bayesian inference provides a solution to the inferential problem by making "probability statements conditional on observed data and an assumed model" (Gelman 2008, 467). Second, Bayesian multilevel modeling is a powerful tool to capture multiple sources of heterogeneity and dynamics in data (Gelman 2006). It can accommodate flexible functional forms and use shrinkage priors to select model features, which reduces model dependency and incorporates modeling uncertainties.

In this paper, we adopt the Bayesian causal inference framework (Rubin 1978; Imbens and Rubin 1997; Rubin *et al.* 2010; Ricciardi, Mattei, and Mealli 2020) to estimate treatment effects in comparative case studies. This framework views causal inference as a missing data problem and relies on the posterior predictive distribution of treated counterfactuals to draw inferences about the treatment effects on the treated. Missingness under this assumption falls in the category of "missing not at random" (MNAR) (Rubin 1976) because the assignment mechanism is allowed to be correlated to unobserved potential outcomes. The basic idea is to perform a low-rank approximation of the observed untreated outcome matrix so as to predict treated counterfactuals in the $(T \times N)$ rectangular outcome matrix. A key assumption we rely on is called *latent ignorability* (Ricciardi, Mattei, and Mealli 2020), which states that treatment assignment is ignorable conditional on exogenous covariates and an unobserved latent variable, which is learned from data.

Conceptually, the latent ignorability assumption is an extension of the strict exogeneity assumption. Existing causal inference methods using TSCS data rely on either of the two types of assumptions for identification: strict exogeneity, which is behind the conventional two-way fixed effect approach and implies "parallel trends" in DiD designs, and sequential ignorability, which has gained popularity recently (e.g., Blackwell 2013; Blackwell and Glynn 2018; Ding and Li 2018; Hazlett and Xu 2018; Strezhnev 2018; Imai, Kim, and Wang 2019). Strict exogeneity requires that treatment assignment is independent of the entire time series of potential outcomes conditional on a set of exogenous covariates and unobserved fixed effects. It rules out potential feedback effects from past outcomes on current and future treatment assignments (Imai and Kim 2019). Its main advantage is to allow researchers to adjust for unit-specific heterogeneity and to use contemporaneous information from a fixed set of control units to predict treated counterfactuals, a key insight of the SCM. Sequential ignorability, on the other hand, allows the probability of

[1] For example, Xu (2017) proposes a bootstrapping procedure for quantifying uncertainties of an LFM; Bai and Ng (2020) provide a limiting theory for individual treatment effects in a factor model setup when both the number of pre-treatment periods and the number of control units are large.

https://doi.org/10.1017/pan.2021.22 Published online by Cambridge University Press

treatment to be affected by past information including realized outcomes. In this paper, we adopt the strict exogeneity framework because it is consistent with the assumptions behind the SCM.

Specifically, we propose a dynamic multilevel latent factor model (henceforth DM-LFM) and develop an estimation strategy using Markov Chain Monte Carlo (MCMC). It incorporates a latent factor term to correct biases caused by the potential correlation between the timing of the treatment and the time-varying latent variables that can be represented by a factor structure, such as diverging trends across different units. It also allows covariate coefficients to vary by unit or over time. Because the model is parameter-rich, we use Bayesian shrinkage priors to conduct stochastic variable and factor selection, thus reducing model dependency. The MCMC algorithm we develop incorporates both model selection and parameter estimation in the same iterative sampling process.

After estimating a DM-LFM, Bayesian prediction generates the posterior distribution of each counterfactual outcome by integrating out all model parameters. We then compare the observed outcomes with the posterior distributions of their predicted counterfactuals to generate the posterior distribution of a causal effect of interest, conditional on observed data. We can use the posterior mean as the point estimate and form its uncertainty measure using the Bayesian 95% credibility interval defined by the 2.5% and 97.5% quantiles of the empirical posterior distribution. The uncertainty measure is easily interpretable: conditional on the data and assumed model, the causal effect takes values from the interval with an estimated probability of 0.95. It captures three sources of uncertainties: (1) the uncertainties from the data generating processes (DGPs), or fundamental uncertainties (King, Tomz, and Wittenberg 2000); (2) the uncertainties from parameter estimation; and (3) the uncertainties from choosing the most suitable model, while existing frequentist methods, such as LFMs, only take into account the first two sources.

Several studies have used Bayesian multilevel modeling for counterfactual prediction. Some are interested in the time-series aspect of data (Belmonte, Koop, and Korobilis 2014; de Vocht *et al.* 2017), in short panels (e.g., two periods) (Ricciardi, Mattei, and Mealli 2020), in large datasets with many treated and control units (Gutman, Intrator, and Lancaster 2018), or in latent factor models without covariates (Samartsidis 2020). A few others have used the Bayesian approach to improve inference for comparative case studies. For example, Amjad, Shah, and Shen (2018) adopt an empirical Bayesian approach to construct error bounds of the treatment effects, but nevertheless rely on frequentist optimization to obtain weights of the SCM. Kim, Lee, and Gupta (2020) propose a fully Bayesian version of the SCM, focusing on the single treated unit case and aiming at estimating a uni-dimensional set of weights on controls.

Our method can be applied to comparative case studies with one or more treated units. Like the SCM, it requires a large number of pre-treatment periods and more control units than the treated to accurately estimate the treatment effect. Our simulation study suggests that the number of pre-treatment periods for the treated units needs to be greater than 20 for the method to achieve satisfactory frequentist properties. Compared with the SCM or LFMs, our method is most suitable when one of the following is true: (1) the uncertainty measures bear important policy or theoretical implications; (2) researchers suspect the latent factor structure is complex, the number of factors is large, or some of the factors are relatively weak; (3) many potential pre-treatment covariates are available and their relationships with the outcome variable may vary across units or over time; or (4) researchers have limited knowledge of how to select covariates for counterfactual prediction. Compared to its frequentist alternatives, this Bayesian method is computationally intense. We therefore develop an `R` package `bpCausal`, whose core functions are written in `C++`, for researchers to efficiently implement this method.

https://doi.org/10.1017/pan.2021.22 Published online by Cambridge University Press

2 Bayesian Causal Inference: Posterior Predictive Distributions

We start by introducing the basic setting and fixing notations. We then define causal quantities of interest and develop posterior predictive distributions of counterfactuals based on several key assumptions. When doing so, we will use two empirical examples. The first one is on German reunification from ADH (2015). It is appealing to illustrate our method for several reasons. First, only one unit (West Germany) in the dataset was treated, and the treatment—a historical event—took place only once. Second, the control group consists of 16 Organization for Economic Co-operation and Development (OECD) member countries, and no single country in this group or their simple average can serve as an appropriate counterfactual for West Germany. Furthermore, the impact of reunification may emerge gradually over time. The second example concerns the effect of election day registration (EDR) on voter turnout in the United States. Xu (2017) uses this example to demonstrate a frequentist LFM (a.k.a. the generalized synthetic control method, or Gsynth). Compared to the first example, it represents a more general setting of comparative case studies in which there are multiple treated units and the treatment starts at different points in time.

2.1 Setup and Estimands

We denote $i = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$ as the unit and time for which and when the outcome of interest is observed. Although our method can accommodate imbalanced panels, we assume a balanced panel instead for notational convenience. We consider a binary treatment w_{it} that, once it takes a value of 1, cannot reverse back to 0 (staggered adoption). Following Athey and Imbens (2018), we define the *timing of adoption* for each unit *i* as a random variable a_i that takes its values in $\mathbb{A} = \{1, 2, \ldots, T, c\}$, in which $a_i = c > T$ means that unit *i* falls in the residual category and does not get treated in the observed time window. We call unit *i* a treated unit if it adopts the treatment at any of the observed time periods ($a_i = 1, 2, \ldots, T$); we call it a control unit if it never adopts the treatment by period T ($a_i = c$). The number of pre-treatment periods for a treated unit is $T_{0,i} = a_i - 1$. Suppose there are N_{co} control units and N_{tr} treated units; $N_{co} + N_{tr} = N$. In comparative case studies, $N_{tr} = 1$ or a small integer.

For instance, in the German reunification example, $i = 1, \ldots, 17$, $T = 1, 2, \ldots, 44$. West Germany's a_i is 31 (the calendar year 1990); and $a_i > 44$ for the other 16 OECD countries serving as controls. The EDR example uses the data of 47 states ($i = 1, \ldots, 47$) in 24 presidential election years from 1920 to 2012 ($t = 1, \ldots, 24$). Among the 47 states, 9 states adopted EDR before 2012, whose $a_i \in \{15, 20, 23, 24\}$; they are considered the treated units. The other 38 states did not adopt EDR by 2012 ($a_i > 24$) and are considered the control units.

Denote $\mathbf{w}_i = (w_{i1}, \ldots, w_{iT})'$ as the treatment assignment vector for unit *i*. Staggered adoption implies that the adoption time a_i uniquely determines vector $\mathbf{w}_i$: $\mathbf{w}_i(a_i)$, in which $w_{it} = 0$ if $t < a_i$ and $w_{it} = 1$ if $t \geqslant a_i$, for $t = 1, 2, \ldots, T$. We further define an $(N \times T)$ treatment assignment matrix, $\mathbf{W} = \{\mathbf{w}_1, \ldots, \mathbf{w}_N\}$; similarly, $\mathbf{W}$ is fully determined by $\mathcal{A} = \{a_1, \ldots, a_i, \ldots, a_N\}$, the adoption time vector. Following Athey and Imbens (2018), we make the following two assumptions to rule out cross-sectional spillover and the anticipation effect.

Assumption 1 (Cross-sectional stable unit treatment value assumption (SUTVA))**.** *Potential outcomes of unit i are only functions of the treatment status of unit i:* $\boldsymbol{y}_{it}(\mathbf{W}) = \boldsymbol{y}_{it}(\mathbf{w}_i), \forall i,\ t.$

This assumption rules out cross-sectional spillover effects and significantly reduces the number of potential outcome trajectories. For each unit *i*, because now there are only $(T + 1)$ possibilities for $\mathbf{w}_i$, there are $(T + 1)$ potential outcome trajectories, denoted by $y_{it}(\mathbf{w}_i),\ t = 1, 2, \ldots, T$. In the German reunification example, this assumption rules out the possibility that German reunification affects economic growth in the other 16 countries, which may, in fact, be a strong assumption. In the EDR example, this assumption implies that the adoption of EDR laws in State B does not affect

State A's turnout with or without EDR laws. Because $\mathbf{w}_i$ is fully determined by a_i, we simply write $y_{it}(\mathbf{w}_i(a_i))$ as $y_{it}(a_i)$.

Assumption 2 (No anticipation). *For all unit i, for all time periods before adoption $t < a_i$:*

$$y_{it}(a_i) = y_{it}(c), \text{ for } t < a_i, \forall i$$

in which $y_{it}(c)$ is the potential outcome under the "pure control" condition, that is, the treatment vector $\mathbf{w}_i$ includes all zeros. This assumption says that the current untreated potential outcome does not depend on whether the unit gets the treatment in the future. The assumption is violated when anticipating that a unit will adopt the treatment in the future affects its outcome today. For example, if people anticipated German reunification to take place in 1990 and West Germany's economy adjusted to that expectation before 1990, the assumption would be violated.

Estimands. Under Assumptions 1 and 2, for treated unit *i* whose adoption time $a_i \leqslant T$, we define its treatment effect at $t \geqslant a_i$ as

$$\delta_{it} = y_{it}(a_i) - y_{it}(c), \text{ for } a_i \leqslant t \leqslant T.$$

In other words, we focus on the difference between the observed post-treatment outcome of treated unit *i* and the counterfactual outcome of the same unit that had never received the treatment by period *T*. In the German reunification example, the causal effect of interest is the difference between the observed gross domestic product (GDP) per capita of West Germany in reunified Germany since 1990 and that of the counterfactual West Germany had it remained separated from East Germany.

Because $y_{it}(a_i)$ of treated unit *i* is fully observed for $t \geqslant a_i$, the Bayesian framework regards it as data. The *counterfactual outcome* $y_{it}(c)$ of treated unit *i* for $t \geqslant a_i$, on the other hand, is an unknown quantity; we regard it as a random variable. We also define the sample average treatment effect on the treated (ATT) for units that have been under the treatment for a duration of *p* periods: $\delta_p = \frac{1}{N_{tr,p}} \sum_{i:T-p+1 \leqslant a_i \leqslant T} \delta_{i,a_i+p-1}$, where $N_{tr,p}$ is the number of treated units that have been treated for *p* periods in the sample.

Given that $y_{it}(a_i)$ is observed in the post-treatment period, estimating δ_{it} is equivalent to constructing the counterfactual outcome $y_{it}(c)$. Under Assumptions 1 and 2, we can denote $\mathbf{Y}(\mathbf{0})$, a $(N \times T)$ matrix, as the potential outcome matrix under $\mathbf{W} = \mathbf{0}$ (i.e., $a_i = c, \forall i$). Given any realization of $\mathbf{W}$, we can partition the indices for $\mathbf{Y}(\mathbf{0})$ into two sets: $\mathcal{S}_0 \equiv \{(it)|w_{it} = 0\}$, with which $y_{it}(c)$ is observed; and $\mathcal{S}_1 \equiv \{(it)|w_{it} = 1\}$, with which $y_{it}(c)$ is missing. Additionally, $\mathcal{S} = \mathcal{S}_0 \cup \mathcal{S}_1$. We denote the observed and missing parts of $\mathbf{Y}(\mathbf{0})$ as $\mathbf{Y}(\mathbf{0})^{obs}$ and $\mathbf{Y}(\mathbf{0})^{mis}$, respectively, and $\mathbf{Y}_i(\mathbf{0})^{obs}$ and $\mathbf{Y}_i(\mathbf{0})^{mis}$ as the row vectors in $\mathbf{Y}(\mathbf{0})^{obs}$ and $\mathbf{Y}(\mathbf{0})^{mis}$ corresponding to unit *i*, respectively. $\mathbf{X}_{it}$ is a $(p_1 \times 1)$ vector of exogenous covariates. $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{iT})'$ is a $(T \times p_1)$ covariate matrix, and we define $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_N\}$.

2.2 The Assignment Mechanism

Rubin *et al.* (2010) lay down the fundamentals of Bayesian causal inference, which views "causal inference entirely as a missing data problem" (p. 685). In general, using the observed outcomes and covariates, as well as the assignment mechanism, we can stochastically impute counterfactuals from their posterior predictive distribution $\Pr(\mathbf{Y}(\mathbf{W})^{mis}|\mathbf{X}, \mathbf{Y}(\mathbf{W})^{obs}, \mathbf{W})$. Since in this research the main causal quantity of interest is the (average) treatment effect on the treated, our primary goal is to predict counterfactuals $\mathbf{Y}(\mathbf{0})^{mis}$, the untreated outcomes of the treated units. Because staggered adoption implies that $\mathbf{W}$ is fully determined by $\mathcal{A}$, the adoption time vector, we write

https://doi.org/10.1017/pan.2021.22 Published online by Cambridge University Press

the posterior predictive distribution of $\mathbf{Y}(\mathbf{0})^{mis}$ as

$$\begin{aligned}\Pr(\mathbf{Y}(\mathbf{0})^{mis}|\mathbf{X},\mathbf{Y}(\mathbf{0})^{obs},\mathcal{A}) &= \frac{\Pr(\mathbf{X},\mathbf{Y}(\mathbf{0})^{mis},\mathbf{Y}(\mathbf{0})^{obs})\Pr(\mathcal{A}|\mathbf{X},\mathbf{Y}(\mathbf{0})^{mis},\mathbf{Y}(\mathbf{0})^{obs})}{\Pr(\mathbf{X},\mathbf{Y}(\mathbf{0})^{obs},\mathcal{A})}\\ &\propto \Pr(\mathbf{X},\mathbf{Y}(\mathbf{0})^{mis},\mathbf{Y}(\mathbf{0})^{obs})\Pr(\mathcal{A}|\mathbf{X},\mathbf{Y}(\mathbf{0})^{mis},\mathbf{Y}(\mathbf{0})^{obs})\\ &\propto \Pr(\mathbf{X},\mathbf{Y}(\mathbf{0}))\Pr(\mathcal{A}|\mathbf{X},\mathbf{Y}(\mathbf{0})). \end{aligned} \tag{1}$$

The Bayes rule gives the equality in Equation (1), and we obtain the proportionality by dropping the denominator as a normalizing constant since it contains no missing data. Hence, two component probabilities, the underlying "science" $\Pr(\mathbf{X},\mathbf{Y}(\mathbf{0}))$ and the treatment assignment mechanism $\Pr(\mathcal{A}|\mathbf{X},\mathbf{Y}(\mathbf{0}))$, help predict the counterfactuals.

Assumption 3 (Individualistic assignment and positivity)**.** $\Pr(\mathcal{A}|\mathbf{X},\mathbf{Y}(\mathbf{0})) = \prod_{i=1}^{n}\Pr(a_i|\mathbf{X}_i,\mathbf{Y}_i(\mathbf{0}))$ *and* $0 < \Pr(a_i|\mathbf{X}_i,\mathbf{Y}_i(\mathbf{0})) < 1$ *for all unit i.*

We assume that the treatment assignment is "individualistic" (Imbens and Rubin 2015, 31), that is, the adoption time of unit i does not depend on the covariates or potential outcomes of other units or their time of adoption, given $\mathbf{X}_i$ and $\mathbf{Y}_i(\mathbf{0})$. We also require that each i has some nonzero chances of getting treated. The first part of this assumption is violated if policy diffusion takes place—for example, State A adopts EDR following State B's policy shift—and such an emulation effect cannot be captured by a unit's pre-treatment covariates and untreated potential outcomes. The positivity assumption means that all units in the sample have some probability of getting treated, thus justifying using control information to predict treated counterfactuals.

The fact that $\Pr(a_i|\mathbf{X}_i,\mathbf{Y}_i(\mathbf{0})) = \Pr(a_i|\mathbf{X}_i,\mathbf{Y}_i(\mathbf{0})^{obs},\mathbf{Y}_i(\mathbf{0})^{mis})$ implies that the treatment assignment mechanism may be correlated with $\mathbf{Y}_i(\mathbf{0})^{mis}$. To rule out potential confounding, one possibility is to impose further restrictions and make an ignorability assignment assumption $\Pr(a_i|\mathbf{X}_i,\mathbf{Y}_i(\mathbf{0})^{obs})$ (Rubin 1978). However, in the more general setting of MNAR, this restriction is unlikely to be true. For instance, the timing of a state's adopting an EDR law could be driven by legislators' concern about future voter turnout in the absence of such laws. Therefore, we rely on the latent ignorability assumption to break the link between treatment assignment and control outcomes.

Assumption 4 (Latent ignorability)**.** *Conditional on the observed pre-treatment covariates* $\mathbf{X}_i$ *and a vector of latent variables* $\mathbf{U}_i = (u_{i1}, u_{i2}, \cdots, u_{iT})$*, the assignment mechanism is free from dependence on any missing or observed untreated outcomes for each unit i, that is,*

$$\Pr(a_i|\mathbf{X}_i,\mathbf{Y}_i(\mathbf{0}),\mathbf{U}_i) = \Pr(a_i|\mathbf{X}_i,\mathbf{Y}_i(\mathbf{0})^{mis},\mathbf{Y}_i(\mathbf{0})^{obs},\mathbf{U}_i) = \Pr(a_i|\mathbf{X}_i,\mathbf{U}_i). \tag{2}$$

Note that $\mathbf{X}_i$ can include both time-varying and time-invariant pre-treatment covariates. $\mathbf{U}_i$ captures both unit-level heterogeneity, such as unit fixed effects, and unit-specific time trends (e.g., $u_{it} = \gamma_i \cdot g(t)$, in which $g(\cdot)$ is a function of time). We expect $\mathbf{U}_i$ to be correlated with $\mathbf{Y}_i(\mathbf{0})$; in fact, we will extract it from $\mathbf{Y}_i(\mathbf{0})^{obs}$. Therefore, once we condition on $\mathbf{X}_i$ and $\mathbf{U}_i$, the entire time series of $\mathbf{Y}_i(\mathbf{0})$ is assumed to be independent of a_i. Thus, we can understand Assumption 4 as an extension of the strict exogeneity assumption often assumed in fixed effects models. Like strict exogeneity, it rules out dynamic feedback from the past outcomes on current and future treatment assignments, conditional on $\mathbf{U}_i$.

The latent ignorability assumption is key to our approach. It is easy to see that the "parallel trends" assumption is implied by this assumption when $\mathbf{U}_i$ is a unit-specific constant $u_{i1} = u_{i2} = \cdots = u_{iT} = u_i$. What does this assumption mean substantively? In the EDR example, suppose there is an unmeasured downward trend in voter turnout in all states since the 1970s, driven by unknown socioeconomic forces. Its impact may be different across states due to differences in

https://doi.org/10.1017/pan.2021.22 Published online by Cambridge University Press

geography, demography, party organizations, ideological orientation, and so on. At the same time, the adoption of EDR in a state is correlated with how much such a trend would affect the state's turnout in the absence of any policy changes. In this case, the "parallel time trends" assumption is invalid because u_{it} is a time-varying confounder; however, we can correct for the biases if we can estimate, then condition on, the state-specific impact of this common trend. Hence, we further make a feasibility assumption.

Assumption 5 (Feasible data extraction)**.** *Assume that, for each unit i, there exists an unobserved covariate vector* $\mathbf{U}_i$ *for each unit i, such that the stacked* $(N \times T)$ *matrix* $\mathbf{U} = (\mathbf{U}_1, \ldots, \mathbf{U}_N)$ *can be approximated by two lower-rank matrices (*$r \ll \min\{N, T\}$*), that is,* $\mathbf{U} = \boldsymbol{\Gamma}'\mathbf{F}$ *in which* $\mathbf{F} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_T)$ *is a* $(r \times T)$ *matrix of factors and* $\boldsymbol{\Gamma} = (\gamma_1, \ldots, \gamma_N)$ *is a* $(r \times N)$ *matrix of the factor loadings.*

This assumption is explicitly or implicitly made with the factor-augmented approach in the existing literature (Xu 2017; Athey *et al.* 2018; Bai and Ng 2020). It says that we can decompose the unit-specific time trends into multiple common trends with heterogeneous impacts. Assumption 5 is violated when $\mathbf{U}_1, \ldots, \mathbf{U}_n$ have no common components; for example, when unit-specific time trends are idiosyncratic.

2.3 Posterior Predictive Inference

Under Assumption 4, we temporarily consider $\mathbf{U}$ as part of the covariates and write $\mathbf{X}$ and $\mathbf{U}$ together as $\mathbf{X}'$. Then we have the posterior predictive distribution of $\mathbf{Y}(\mathbf{0})^{mis}$ as

$$\begin{aligned}
\Pr(\mathbf{Y}(\mathbf{0})^{mis}|\mathbf{X}', \mathbf{Y}(\mathbf{0})^{obs}, \mathcal{A}) &\propto \Pr(\mathbf{X}', \mathbf{Y}(\mathbf{0})^{mis}, \mathbf{Y}(\mathbf{0})^{obs}) \Pr(\mathcal{A}|\mathbf{X}', \mathbf{Y}(\mathbf{0})^{mis}, \mathbf{Y}(\mathbf{0})^{obs}) \\
&\propto \Pr(\mathbf{X}', \mathbf{Y}(\mathbf{0})^{mis}, \mathbf{Y}(\mathbf{0})^{obs}) \Pr(\mathcal{A}|\mathbf{X}') \\
&\propto \Pr(\mathbf{X}', \mathbf{Y}(\mathbf{0})). \qquad (3)
\end{aligned}$$

The first line is simply to re-write Equation (1) by replacing $\mathbf{X}$ with $\mathbf{X}'$; we reach the second step using the latent ignorability assumption; the last step drops the treatment assignment mechanism $\Pr(\mathcal{A}|\mathbf{X}')$ as a normalizing constant since it does not contain $\mathbf{Y}(\mathbf{0})^{mis}$.

Equation (3) says that the latent ignorability assumption makes the treatment assignment mechanism ignorable in counterfactual prediction; as a result, we can ignore it as long as we condition on $\mathbf{X}'$. In other words, under these assumptions, the task of developing the posterior predictive distributions of counterfactuals is reduced to model $\Pr(\mathbf{X}', \mathbf{Y}(\mathbf{0}))$. We need this trick because in comparative case studies, the number of treated units is small; as a result, we lack sufficient variation of the timing of adoption in data to model $\Pr(\mathcal{A}|\mathbf{X}')$. To model the underlying "science," we further make the assumption of exchangeability:

Assumption 6 (Exchangeability)**.** *When* $\mathbf{U}$ *is known,* $\{(\mathbf{X}'_{it}, y_{it}(c))\}_{i=1,\ldots,N; t=1,\ldots,T}$ *is an exchangeable sequence of random variables; that is, the joint distribution of* $\{(\mathbf{X}'_{it}, y_{it}(c))\}$ *is invariant to permutations in the index* it*.*

By de Finetti's theorem (de Finetti 1963), $\{(\mathbf{X}'_{it}, y_{it}(c))\}$ can be written as *i.i.d*, given some parameters and their prior distributions. Note that $\Pr(\mathbf{X}', \mathbf{Y}(\mathbf{0}))$ is equivalent to $\Pr(\{(\mathbf{X}'_{it}, y_{it}(c))\})$, and we now can write the posterior predictive distribution of $\mathbf{Y}(\mathbf{0})^{mis}$ in Equation (3) as

$$\begin{aligned}
\Pr(\mathbf{Y}(\mathbf{0})^{mis}|\mathbf{X}', \mathbf{Y}(\mathbf{0})^{obs}, \mathcal{A}) &\propto \Pr(\{(\mathbf{X}'_{it}, y_{it}(c))\}) \\
&\propto \int \underbrace{\left(\prod_{it \in S_1} f(y_{it}(c)^{mis}|\mathbf{X}_{it}, \theta') \right)}_{\text{posterior predictive distribution}} \underbrace{\left(\prod_{it \in S_0} f(y_{it}(c)^{obs}|\mathbf{X}_{it}, \theta') \right)}_{\text{likelihood}} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}, \qquad (4)
\end{aligned}$$

https://doi.org/10.1017/pan.2021.22 Published online by Cambridge University Press

where $\boldsymbol{\theta}$ are the parameters that govern the DGP of $y_{it}(c)$ given $\mathbf{X}'_{it}$, and $\boldsymbol{\theta}' = (\boldsymbol{\theta}, \mathbf{U})$ when we regard the latent covariates $\mathbf{U}$ as parameters. Note that in Equation (4), the likelihood is based on observed outcomes while the posterior predictive distribution is for predicting the missing potential outcomes. The second proportionality is reached by de Finetti's theorem and under the assumption that the set of parameters that govern the DGP of the covariates $\mathbf{X}$ are independent of $\boldsymbol{\theta}$. See Supplementary Material for a formal mathematical development of the posterior predictive distribution.

Recall that our objective is to impute the untreated potential outcomes for treated observations. If we assume that the DGPs of the outcomes in $\mathcal{S}_0$ and $\mathcal{S}_1$ follow the same functional form $f(\cdot)$, we can build a parametric model and estimate parameters based on the likelihood and then predict $y_{it}(c)^{mis}$ for $i \leqslant N_{co}$ at $t \geqslant a_i$ using the posterior predictive distribution. If we can correctly estimate $\pi(\mathbf{U}|\mathbf{X}, \mathbf{Y}(\mathbf{0})^{obs})$, the posterior distributions of $\mathbf{U}$, using a factor analysis, we can draw samples of treated counterfactuals $y_{it}(c)^{mis}$ from its posterior predictive distribution as in Equation (1) by integrating out the model parameters.

3 Modeling and Implementation

In this section, we discuss the modeling strategy for the likelihood function and the posterior predictive distribution. We explain the proposed DM-LFM and discuss Bayesian shrinkage method for factor selection and model searching to reduce model dependency.

3.1 A Multilevel Model with Dynamic Factors

Assumption 7 (Functional form)**.** *The untreated potential outcomes for unit $i = 1, \ldots, N$ at $t = 1, \ldots, T$ are specified as follows:*

$$y_{it}(c) = \mathbf{X}'_{it}\boldsymbol{\beta}_{it} + \boldsymbol{\gamma}'_i \mathbf{f}_t + \epsilon_{it}, \tag{5}$$

$$\boldsymbol{\beta}_{it} = \boldsymbol{\beta} + \boldsymbol{\alpha}_i + \boldsymbol{\xi}_t, \tag{6}$$

$$\boldsymbol{\xi}_t = \Phi_{\xi}\boldsymbol{\xi}_{t-1} + \boldsymbol{e}_t, \; \mathbf{f}_t = \Phi_f \mathbf{f}_{t-1} + \boldsymbol{\nu}_t. \tag{7}$$

Equation (5) is the individual-level regression, in which $y_{it}(c)$ is explained by three components. The first one, $\mathbf{X}_{it}\boldsymbol{\beta}_{it}$, captures the relationships between observed covariates and the outcome. The double subscripts of $\boldsymbol{\beta}_{it}$ indicate that we allow these relationships to be heterogeneous across units and over time. Equation (6) decomposes $\boldsymbol{\beta}_{it}$ into three parts: $\boldsymbol{\beta}$ is the mean of $\boldsymbol{\beta}_{it}$ and shared by all observations, and $\boldsymbol{\alpha}_i$ and $\boldsymbol{\xi}_t$ are zero-mean unit- and time-specific "residuals" of $\boldsymbol{\beta}_{it}$, respectively. The second component, $\boldsymbol{\gamma}'_i\mathbf{f}_t$, is the multifactor term, in which $\mathbf{f}_t$ and $\boldsymbol{\gamma}_i$ are $(r \times 1)$ vectors of factors and factor loadings, respectively. Consistent with Assumption 5, we use $\mathbf{f}_t$ and $\boldsymbol{\gamma}_i$ to approximate $\mathbf{U}_i$. The last component, ϵ_{it}, represents $i.i.d.$ idiosyncratic errors. We further model the dynamics in $\boldsymbol{\xi}_t$ and $\mathbf{f}_t$ by specifying autoregressive processes as shown in Equation (7). We assume both transition matrices Φ_ξ and Φ_f are diagonal: $\Phi_\xi = \text{Diag}(\phi_{\xi_1}, \ldots, \phi_{\xi_{p_3}})$ and $\Phi_f = \text{Diag}(\phi_{f_1}, \ldots, \phi_{f_r})$.[2] Finally, we assume the individual- and group-level errors, ϵ_{it}, $\boldsymbol{e}_t$, and $\boldsymbol{\nu}_t$, to be *i.i.d.* normal.

The DM-LFM allows the slope coefficient of each covariate to vary by unit, time, both, or neither. To illustrate this flexibility, we rewrite the individual-level model in a reduced and matrix format as

[2] Depending on the values of the transition matrices (determined by data), time-varying parameters may take one of the following three processes: (1) a stationary autoregressive process with order one if Φ_ξ and Φ_f are diagonal matrices and each element on the diagonal is nonzero and falls in the open interval $(-1, 1)$; (2) a local smoothing model if Φ_ξ when Φ_f are identity matrices; or (3) a static multi-level structure if Φ_ξ and Φ_f are null matrices.

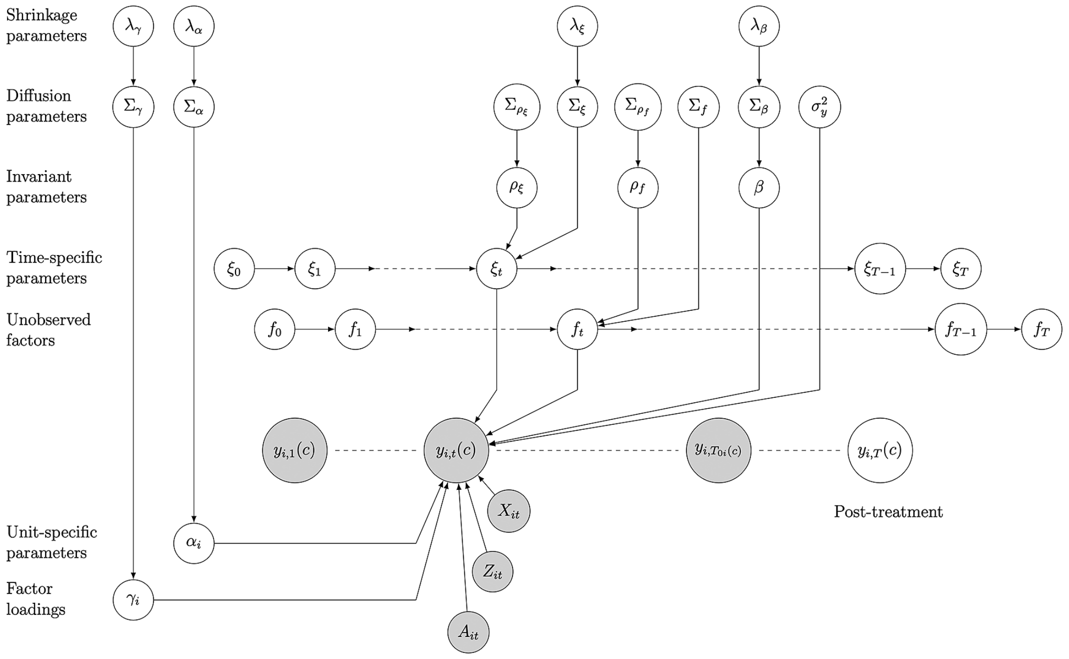


Figure 1. A graphic representation of dynamic multilevel latent factor model (DM-LFM). The shaded nodes represent observed data, including untreated outcomes and covariates; the unshaded nodes represent "missing" data (treated counterfactuals) and parameters. Only one (treated) unit i is shown. $T_{0i} = a_i - 1$ is the last period before the treatment starts to affect unit i. The focus of the graph is Period t. Covariates in periods other than t, as well as relationships between parameters and $y_{i1}(c)$, $y_{iT_{0i}(c)}$, and $y_{iT}(c)$, are omitted for simplicity.

$$\boldsymbol{y}_i(c) = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\alpha}_i + \boldsymbol{A}_i\boldsymbol{\xi} + \mathbf{F}\gamma_i + \epsilon_i, \tag{8}$$

in which $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_T)'$ is a $(T \times r)$ factor matrix. $\boldsymbol{Z}_i$ of dimension $(T \times p_2)$ are covariates that have unit-specific slopes $(\boldsymbol{\beta} + \boldsymbol{\alpha}_i)$. $\boldsymbol{A}_i$ of dimension $(T \times p_3)$ are covariates that have time-specific slopes $(\boldsymbol{\beta} + \boldsymbol{\xi}_t)$. Both $\boldsymbol{A}_i$ and $\boldsymbol{Z}_i$ are subsets of $\boldsymbol{X}_i$. When the kth covariate is in $\boldsymbol{A}_i \cap \boldsymbol{Z}_i$, it has a slope coefficient that varies across units and over time, that is, $\beta_{it}^{(k)} = \beta^{(k)} + \alpha_i^{(k)} + \xi_t^{(k)}$. Accordingly, $\boldsymbol{\beta}$ is a $(p_1 \times 1)$ coefficients vector, $\boldsymbol{\alpha}_i$ is a $(p_2 \times 1)$ vector, $\boldsymbol{\xi} = (\xi_1', \ldots, \xi_T')'$ is a $(p_3 \times 1)$ vector, and $p_1 \geqslant p_2, p_3$. Because $\boldsymbol{\alpha}_i, \boldsymbol{\xi}_t, \mathbf{f}_t$ are centered at zero, the systemic part of the model is $\mathbb{E}[\mathbf{y}_i(c)] = \mathbf{X}_i\boldsymbol{\beta}$. The rest of the components define the variance of the composite errors as $\boldsymbol{\Omega}_{\mathbf{y}_i(c) - \boldsymbol{X}_i\boldsymbol{\beta}} = (\boldsymbol{Z}_i\boldsymbol{\alpha}_i + \boldsymbol{A}_i\boldsymbol{\xi} + \mathbf{F}\gamma_i)'(\boldsymbol{Z}_i\boldsymbol{\alpha}_i + \boldsymbol{A}_i\boldsymbol{\xi} + \mathbf{F}\gamma_i) + \sigma_\epsilon^2\mathbf{I}$, which contains nonzero off-diagonal elements because of the components $\mathbf{Z}_i$ and $\mathbf{A}_i$. Note that we allow $\boldsymbol{Z}_i\boldsymbol{\alpha}_i$, $\boldsymbol{A}_i\boldsymbol{\xi}$, and $\mathbf{F}\gamma_i$ to be arbitrarily correlated.

Figure 1 presents the model graphically, and the shrinkage parameters λ's in the figure will be discussed later. Note that this outcome model governs $y_{it}(c)$ only. Without loss of generality, we can add $\delta_{it}w_{it}$ in the model in which δ_{it} is the causal effect for unit i at time t. This model is a dynamic and multilevel extension to several existing causal inference methods with TSCS data. For example, if we set $\mathbf{Z}_i = \mathbf{A}_i = (1, 1, \ldots, 1)'$, and $r = 0$, Equation (8) becomes a parametric linear DiD model with covariates and two-way fixed effects: $y_{it} = \delta_{it}w_{it} + \mathbf{X}_{it}'\beta + \alpha_i + \xi_t + \epsilon_{it}$. This is what Liu, Wang, and Xu (2020) call the fixed effects counterfactual model. When we put restrictions $\mathbf{Z}_{it} = \varnothing$ and $\mathbf{X}_i = \mathbf{A}_i$ is time-invariant, our model is reduced to a factor model that justifies the SCM (Abadie, Diamond, and Hainmueller 2010): $y_{it} = \delta_{it}w_{it} + \xi_t + \mathbf{X}_i'\beta_t + \gamma_i'\mathbf{f}_t + \epsilon_{it}$. Gsynth is also a special case of the model when we force the coefficients not to vary; that is, $\mathbf{Z}_{it} = \varnothing$ and $\mathbf{A}_{it} = \varnothing$ and $y_{it} = \delta_{it}w_{it} + \mathbf{X}_{it}'\beta + \gamma_i'\mathbf{f}_t + \epsilon_{it}$ (Xu 2017).

3.2 Bayesian Stochastic Model Specification Search

One advantage of the DM-LFM is that it is highly flexible. However, the large number of specification options poses a challenge to model selection. Bayesian stochastic model searching reduces the risks of model mis-specification and simultaneously incorporates model uncertainty. We use shrinkage priors to choose the number of latent factors and decide whether and how to include a covariate. Specifically, we adopt the Bayesian Lasso and Lasso-like hierarchical shrinkage methods based on recent research.[3] We apply the Bayesian Lasso shrinkage on $\boldsymbol{\beta}$ using the following hierarchical setting of mixture of a normal-exponential prior, $\beta_k|\tau^2_{\beta_k} \sim \mathcal{N}(0, \tau^2_{\beta_k})$, $\tau^2_{\beta_k}|\lambda_\beta \sim Exp(\frac{\lambda^2_\beta}{2})$, $\lambda^2_\beta \sim \mathcal{G}(a_1, a_2)$, $k = 1, \ldots, p_1$. The tuning parameter λ controls the sparsity and degree of shrinkage and can be understood as the Bayesian equivalent to the regulation penalty in a frequentist Lasso regression. Instead of fixing λ at a single value, we take advantage of Bayesian hierarchical modeling and give it a Gamma distribution with hyper-parameters a_1 and a_2.[4]

To select the other components of the model, we impose shrinkage on $\boldsymbol{\alpha}_i$, $\boldsymbol{\xi}_t$, or $\boldsymbol{\gamma}_i$ to determine whether to include a Z_j ($j = 1, 2, \ldots, p_2$), A_j ($j = 1, 2, \ldots, p_3$), or f_j ($j = 1, 2, \ldots, r$) in the model. We consider the Lasso-like hierarchical shrinkage approach with re-parameterization. Assume $\boldsymbol{\alpha}_i, \boldsymbol{\gamma}_i$, and $\boldsymbol{\xi}_t$ have diagonal variance–covariance matrices, $\mathbf{H}_0 = \text{Diag}(\omega^2_{\alpha_1}, \ldots, \omega^2_{\alpha_{p_2}})$, $\boldsymbol{\Gamma}_0 = \text{Diag}(\omega^2_{\gamma_1}, \ldots, \omega^2_{\gamma_r})$, $\Sigma_e = \text{Diag}(\omega^2_{\xi_1}, \ldots, \omega^2_{\xi_{p_3}})$, respectively. To have a shrinkage effect, we should allow the variance parameters ω^2 to have a positive probability to take the value zero. Therefore, we re-parameterize $\boldsymbol{\alpha}_i$, $\boldsymbol{\xi}_t$, and $\boldsymbol{\gamma}_i$ as $\boldsymbol{\alpha}_i = \boldsymbol{\omega}_\alpha \cdot \tilde{\boldsymbol{\alpha}}_i$, $\boldsymbol{\xi}_t = \boldsymbol{\omega}_\xi \cdot \tilde{\boldsymbol{\xi}}_t$, $\boldsymbol{\gamma}_i = \boldsymbol{\omega}_\gamma \cdot \tilde{\boldsymbol{\gamma}}_i$, where $\boldsymbol{\omega}_\alpha = (\omega_{\alpha_1}, \ldots, \omega_{\alpha_{p_2}})'$, $\boldsymbol{\omega}_\xi = (\omega_{\xi_1}, \ldots, \omega_{\xi_{p_3}})'$, and $\boldsymbol{\omega}_\gamma = (\omega_{\gamma_1}, \ldots, \omega_{\gamma_r})'$ are column vectors. After re-parameterization, the variances ω^2 appear in the model as coefficients $\boldsymbol{\omega}$ that can take values on the entire real line, and the new variance–covariance matrices become identity matrices.

Now we assign Lasso priors to each ω_α, ω_ξ, and ω_{γ_j} to shrink varying parameters grouped by unit or time. Together with the shrinkage on $\boldsymbol{\beta}$, the algorithm will decide *de facto* whether a certain covariate is included, whether its coefficient varies by time or across units, and how many latent factors are considered. Because the shrinkage priors do not have a point mass component at zero, parameters of less important covariates are not zeroed out completely. Instead, they stay in the model but with shrunk impacts and can be regarded as virtually excluded from the model. The posterior distribution of ω may be of different shapes. If it is clearly bimodal, it means that the associated parameter is included in the model; if it is close to unimodal and centered at zero, it means that the parameter is virtually excluded from the model; if, however, the posterior distribution of ω has three or more modes, it indicates that the data do not provide decisive information on whether the corresponding covariate or factor is sufficiently important—in some iterations, the parameter escapes the shrinkage while in others, it is trapped in a narrow neighborhood around zero. In each MCMC iteration, the algorithm samples a model consisting of the parameters that successfully escape the shrinkage, and posterior distributions of parameters generated by the stochastic search algorithm are based on a mixture of models in a continuous model space. In other words, this variable selection process is also a model-searching and model-averaging process.[5]

[3] See, for example, Park and Casella (2008), Kyung *et al.* (2010), Belmonte, Koop, and Korobilis (2014), and Bitto and Frühwirth-Schnatter (2019).

[4] Here, a_1 is the shape parameter and a_2 is the rate parameter, so the Gamma distribution has mean a_1/a_2 and variance a_1/a_2^2. We tune the model indirectly by choosing different values of the hyper-parameters. In general, when λ takes large values with higher probability, the shrinkage is more aggressive. The default values for the hyper-parameters are (0.001, 0.001) following the literature (Belmonte, Koop, and Korobilis 2014). In our simulated and empirical applications, such choices achieve sparsity while allowing important parameters to escape the shrinkage.

[5] The above re-parameterization makes $\boldsymbol{\omega}$'s and associated parameters unidentifiable, but this does not pose an issue for identifying the causal effects as long as the original parameters of $\boldsymbol{\gamma}_i$, $\boldsymbol{\xi}_t$, and $\boldsymbol{\alpha}_i$ are identified. We apply a permutation method to ensure the posterior distributions of $\boldsymbol{\omega}$ to be symmetric around zero.

https://doi.org/10.1017/pan.2021.22 Published online by Cambridge University Press

Table 1. Total number of parameters.

(a) Effective Parameters					
	Mean	Variance			
Parameters	$\boldsymbol{\beta}$	τ_α^2	τ_ξ^2	τ_γ^2	σ_ϵ^2
Number	p_1	p_2	p_3	r	1

Total: $p_1 + p_2 + p_3 + r + 1$.

(b) Parameters to be Integrated out								
	Parameter Expansion					Hyper-Parameters		
Parameters	$\boldsymbol{\alpha}_i$	$\boldsymbol{\xi}_t$	$\boldsymbol{\gamma}_i$	$\mathbf{f}_t$	$\boldsymbol{\omega}$	λ	$\boldsymbol{\phi}$	τ_β
Number	$p_2 N$	$p_3 T$	rN	rT	$(p_2 + p_3 + r)$	4	$(p_3 + r)$	p_1

Total: $(p_2 + r)N + (p_3 + r)T + (p_1 + p_2 + 2p_3 + 2r + 4)$.

The variance parameters for v, e, f_t do not appear in the table because we assume they have standard normal priors.

The DF-LFM has many parameters, but most of them, listed in Table 1(b), are included as part of Bayesian parameter expansion for computational convenience or as hyper-parameters to adjust for parameter uncertainties. They do not appear in the likelihood or can be integrated out of the likelihood given the effective parameters. The effective parameters, including $\boldsymbol{\beta}$ and the variance parameters, are reported in Table 1(a). Their number is usually much smaller than the number of observations.

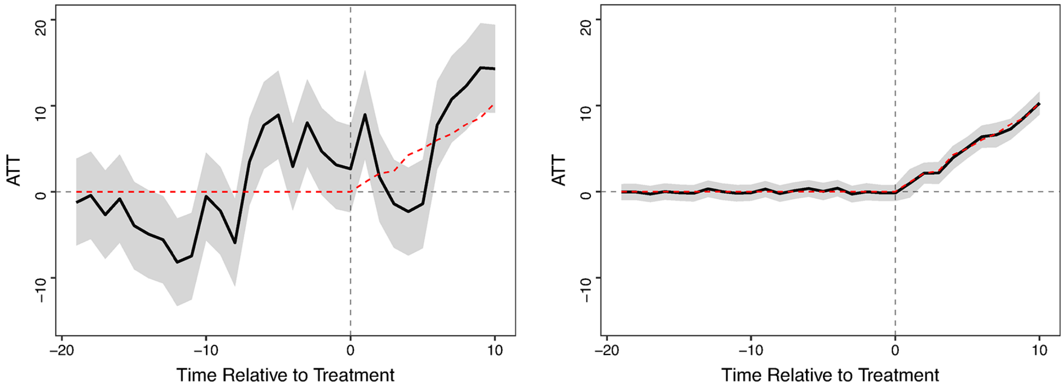
3.3 Implementing a DM-LFM

We develop an MCMC algorithm to estimate a DM-LFM. The core functions of the algorithm is written in C++. Due to space limitations, we present the details of choices of priors and the iterative steps of MCMC updating in Section A.2 in online Supplementary Material. Broadly speaking, implementing a DM-LFM takes the following three steps:

Step 1. Model searching and parameter estimation. We specify and estimate the DM-FLM model with Bayesian shrinkage to sample G draws (excluding draws in the burn-in stage) of the parameters from their posterior distributions, $\boldsymbol{\theta}_{it}^{(g)} \sim \pi(\boldsymbol{\theta}_{it}|\mathfrak{D})$, where $\mathfrak{D} = \{(\mathbf{X}_{it}, y_{it}(c)^{obs}) : it \in \mathcal{S}_0\}$ is the set of untreated observations. Because of Bayesian shrinkage, $\pi(\boldsymbol{\theta}_{it}|\mathfrak{D})$ is in effect a mixture of distributions.

Step 2. Prediction and integration. We conduct Bayesian prediction by generating draws of counterfactual $y_{it}(c)^{mis}$ for each treated unit at $a_i \leqslant t \leqslant T$ from its posterior predictive distribution: $f(y_{it}(c)^{mis}|\mathbf{X}, \mathbf{Y}(\mathbf{0})^{obs}) \propto \int f(y_{it}(c)^{mis}|\mathbf{X}_{it}, \boldsymbol{\theta}_{it})\pi(\boldsymbol{\theta}_{it}|\mathfrak{D})d\boldsymbol{\theta}_{it}$. Bayesian prediction is an empirical integration: a sample of the predicted counterfactual is generated by plugging each draw $\boldsymbol{\theta}_{it}^{(g)}$ from $\pi(\boldsymbol{\theta}_{it}|\mathfrak{D})$ into $f(y_{it}(c)^{mis}|\mathbf{X}_{it}, \boldsymbol{\theta}_{it})$ to obtain $y_{it}^{(g)}(c)$ for $g = 1, \ldots, G$. The sample of counterfactuals is drawn from $f(y_{it}(c)^{mis}|\mathbf{X}_{it}, \mathfrak{D}) = f(y_{it}(c)^{mis}|\mathbf{X}, \mathbf{Y}^{obs}(\mathbf{0}))$ without any unknown quantities.

Step 3. Inference and diagnostics. We make inference about the causal effect δ_{it} at $a_i \leqslant t \leqslant T$ for each treated unit i, by summarizing the empirical posterior distribution δ_{it} formed by $\delta_{it}^{(g)} = y_{it}(a_i) - y_{it}^{(g)}(c)$, $g = 1, \ldots G$. To summarize the results, we can obtain its posterior mean, variance, and the Bayesian 95% credibility interval. We can make inferences about other estimands, such as the ATT, by pooling the posterior draws of δ_{it} and summarizing their posterior distributions accordingly. We conduct Bayesian diagnostic tests on the convergence and mixing on main parameters' posterior distributions and find that the MCMC algorithm converges fast and mixes well in our simulation and empirical studies.



(a) DiD

(b) DM-LFM

Figure 2. Estimated average treatment effect on the treated (ATT): difference-in-differences (DiD) versus dynamic multilevel latent factor model (DM-LFM). The above figures show the ATT estimates and their 95% credibility intervals from the DiD and DM-LFM estimators, both implemented with Bayesian Markov Chain Monte Carlo (MCMC) algorithms. The red dashed lines represent the true ATT of the five treated units.

4 Simulation Studies

In this section, we first illustrate how the Bayesian DM-LFM works using a simulated example. We then study its properties by varying sample sizes and model specifications and compare its relative performance against existing methods, including SCM and Gsynth, in the case of a single treated unit. We report more findings from Monte Carlo exercises in Supplementary Material.

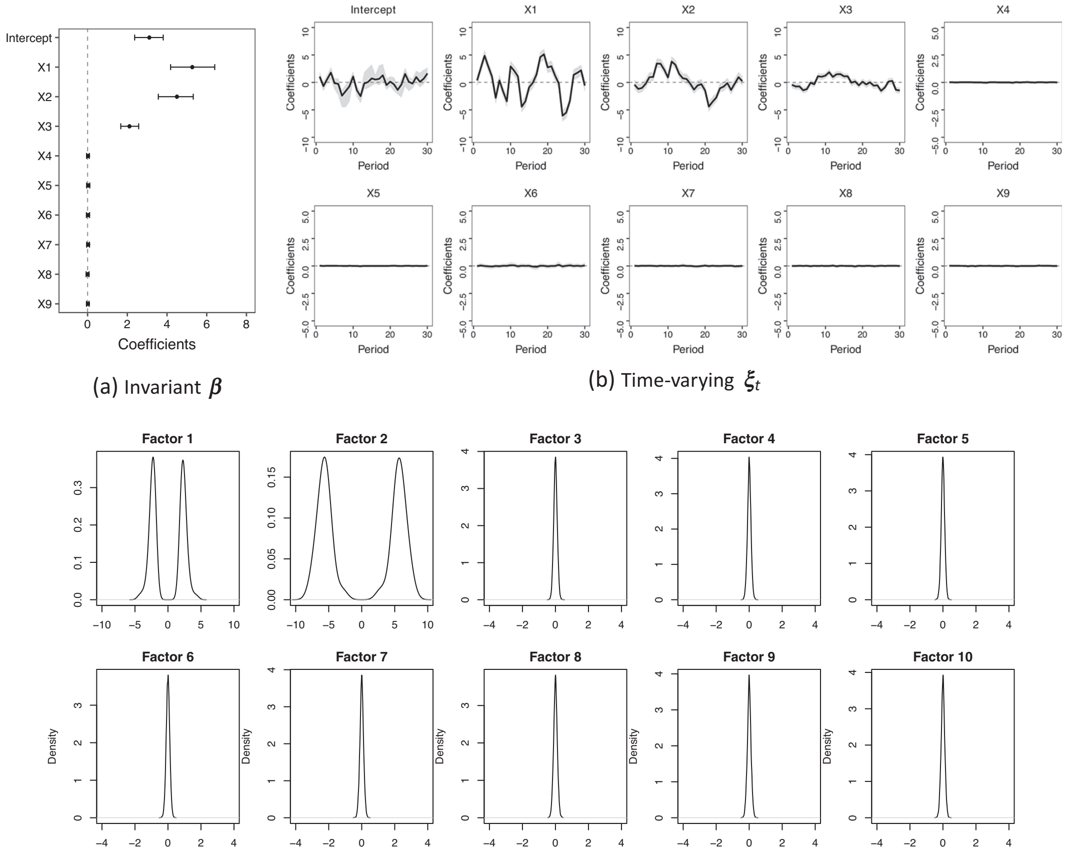
4.1 A Simulated Example

We simulate a panel dataset of 50 units and 30 time periods based on the following DGP:

$$y_{it} = \delta_{it} w_{it} + \mathbf{X}'_{it}\beta_{it} + \gamma'_i \mathbf{f}_t + \epsilon_{it} = \delta_{it} w_{it} + \mathbf{X}'_{it}(\beta + \alpha_i + \xi_t) + \gamma'_i \mathbf{f}_t + \epsilon_{it} \tag{9}$$

in which w_{it} is the treatment indicator and δ_{it} is the treatment effect. $\mathbf{X}_{it}$ is a vector of 10 covariates including an intercept and nine time-varying variables, but only the intercept and the first three covariates have nonzero, unit- and time-varying coefficients. We re-parameterize Equation (9) as $y_{it} = \delta_{it} w_{it} + \mathbf{X}'_{it}\beta + \mathbf{X}'_{it}(\omega_\alpha \cdot \tilde{\alpha}_i) + \mathbf{X}'_{it}(\omega_{\mathbf{x}} \cdot \tilde{\xi}_t) + (\omega_\gamma \cdot \tilde{\gamma}_i)'\mathbf{f}_t + \epsilon_{it}$ such that $\tilde{\alpha}_i$, $\tilde{\xi}_t$, and $\tilde{\gamma}_i$ all have univariances. Seven units receive the treatment starting from Period 21 and remain treated till Period 30. The remaining 23 units are never treated. The heterogeneous treatment effects are governed by $\delta_{it,t>20} = t - 20 + \tau_{it}$, in which $\tau_{it} \overset{i.i.d.}{\sim} N(0,1)$. This means the expected value of the treatment effect gradually increases as the treatment duration grows, for example, from 1 in Period 21 to 10 in Period 30. The factor vector $\mathbf{f}_t$ is two-dimensional and both factors follow an AR(1) process. The probability of getting treated is positively correlated with the sum of a unit's factor loadings $\tilde{\gamma}_i$, which are also *i.i.d.* $N(0,1)$. The selection on the factor loadings will cause biases in the causal estimates if a model does not account for the factor term or the covariates. We provide the details of the DGP in Supplementary Material.

Figure 2 shows the estimated ATT (posterior means) with their 95% credibility intervals using a Bayesian DiD model (a) and a DM-LFM model (b). The DiD model assumes $\omega_\alpha = \omega_\xi = \omega_\gamma = 0$; in other words, the coefficients for the covariates are assumed to be constant, and no factors are included in the model. Figure 2(a) shows that with the DiD estimator, multiple estimates in the pre-treatment periods are away from zero and significant biases exist for the ATT estimates in the post-treatment periods. On the contrary, the DM-LFM performs significantly better: we do not observe any pre-trend and the ATT estimates are close to the true values, which are covered by the corresponding 95% credibility intervals.



(a) Invariant β

(b) Time-varying ξ_t

(c) ω_γ for factor selection

Figure 3. Posterior distributions of coefficients. (a) and (b) show the posterior mean and 95% credibility intervals of the invariant component and time-varying component of β_{it}, respectively. (c) shows the posterior distribution of each ω_γ, which captures the extent to which a factor influences the outcome.

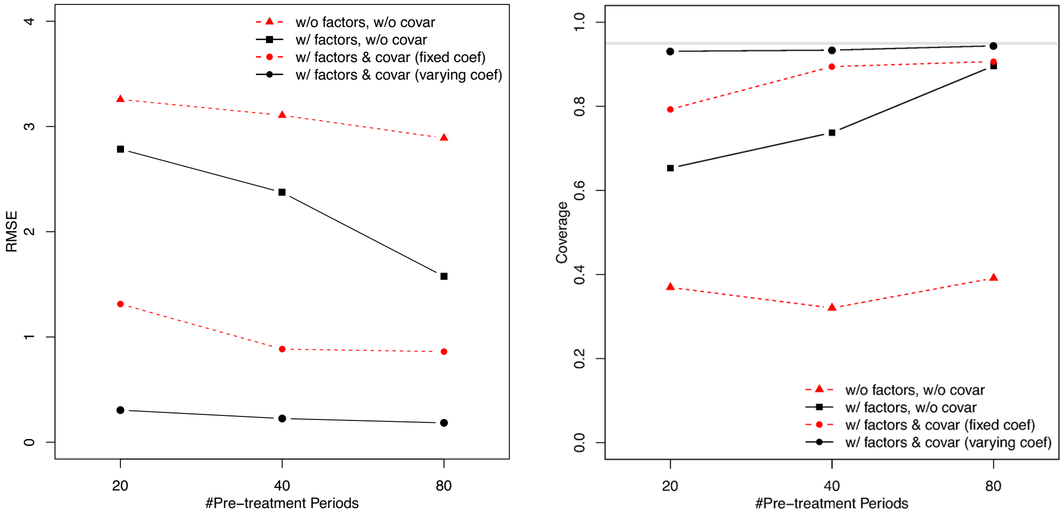
Figure 3(a) and Figure 3(b) show the invariant and time-varying components of the covariate coefficients, respectively. While β and ξ_t for the intercept and the first three covariates are clearly nonzero, the coefficients for the other six covariates are close to zero, which is consistent with the DGP. The algorithm also correctly selects the non-zero α_i, the unit-varying components of coefficients (reported in Supplementary Material). Figure 3c shows the posterior distributions of ω_γ, which measures the relative importance of each of the 10 factors subject to selection. It shows that two factors (Factors 1 and 2) have a ω with clear bimodal posteriors, and the others all have spike-shaped unimodal ones. This means that the model correctly identifies the number of factors to be two.

This simulated example demonstrates that the DM-LFM performs well even when the sample size is relatively small and when researchers have limited knowledge about which covariates to put in, whether the covariates' relationships with the outcome change over time or across units, or how many latent factors to include in the model.

4.2 Additional Monte Carlo Evidence

In addition, we conduct several sets of Monte Carlo exercises to study the properties of the proposed method and compare its relative performance against existing methods. Due to space limitations, we provide the details of these exercises in Supplementary Material and only briefly summarize two exercises and the major findings below.

In the first exercise, we study the role of each of the main components of a DM-LFM in estimation and investigate how the sample size affects the model performance. We simulate samples using



(a) Root Mean Square Errors

(b) Coverage

Figure 4. Comparison of model performance: root mean square errors (RMSE) and coverage. The above figures show the RMSE (a) and the coverage rate of 95% credibility intervals (b) of the average treatment effect on the treated (ATT) estimates using three sets of models: (1) models that do not include factors or covariates; (2) models that include 10 factors but no covariates; (3) models that include 10 factors and covariates with fixed coefficients; and (4) models that include 10 factors and covariates with varying coefficients.

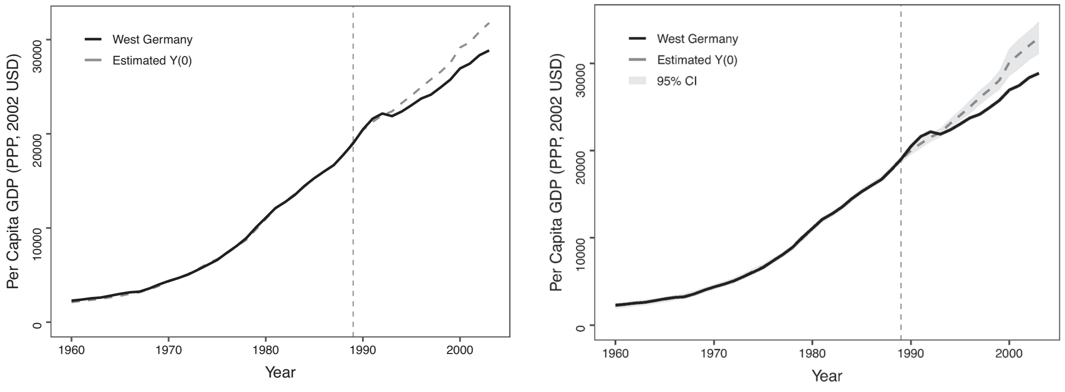
the DGP as in Equation (9) while varying the sample size (both the total number of units N and the number of pre-treatment periods T_0). We estimate and compare the full DM-LFM model with its three simpler variants: a model with covariates but without factors, analogous to a DiD model including covariates with fixed coefficients; a model without covariates but with 10 latent factors, analogous to Gsynth without time-varying covariates; and a model with factors and covariates with fixed coefficients. Figure 4 shows the comparison when the number of units is relatively small ($N = 40$), and the full results are reported in Supplementary Material. In general, we find that DM-LFM outperforms the other three models in terms of bias, standard deviation, root mean squared errors (RMSE), and coverage. This exercise demonstrates that each of the key components of the model contributes to improving performance in causal effect estimation. The factor term seems to have the most impact, but covariates with varying coefficients also notably improve precision and coverage.

The second exercise compares the performance of the DM-LFM with SCM and Gsynth in the case of a single treated unit. Tables A5–A7 in Supplementary Material show that, compared with the SCM, the DM-LFM has a much smaller RMSE. The DM-LFM outperforms Gsynth in the realistic scenario when the true number of factors is unknown, and when the number of factors is large and each of them produces relatively weak signals. Note that our Bayesian approach becomes significantly more computationally demanding as the sample size grows. For example, in one simulation in which $N = 50$ and $T_0 = 60$, it takes about 30 seconds to run a DM-LFM on a 2019 6-core MacBook Pro while Gsynth takes less than a second.

5 Empirical Applications

We apply the Bayesian DM-LFM to the two running examples introduced earlier. To gain confidence that the DM-LFM is appropriate for an application, we recommend users conduct the following diagnostic tests after running the model: (1) conduct Bayesian diagnostic tests on convergence and mixing of the MCMC output of key parameters, for example, by plotting of the traces of estimates; (2) examine whether the model fits the pre-treatment outcome trajectories of the treated units

https://doi.org/10.1017/pan.2021.22 Published online by Cambridge University Press



(a) SCM

(b) DM-LFM

Figure 5. Actual and estimated counterfactuals for West Germany. This figure shows the per capita gross domestic product (GDP) of actual West Germany and the posterior means (with 95% credibility intervals) of its untreated outcome from 1960 to 2003. The within-sample prediction is very precise and the interval is too narrow to be seen in the figure between 1960 and 1989.

reasonably well; and (3) conduct a placebo test using pre-treatment data of the treated units when there is a sufficient number of pre-treatment periods for the treated units.

5.1 Economic Impact of German Reunification

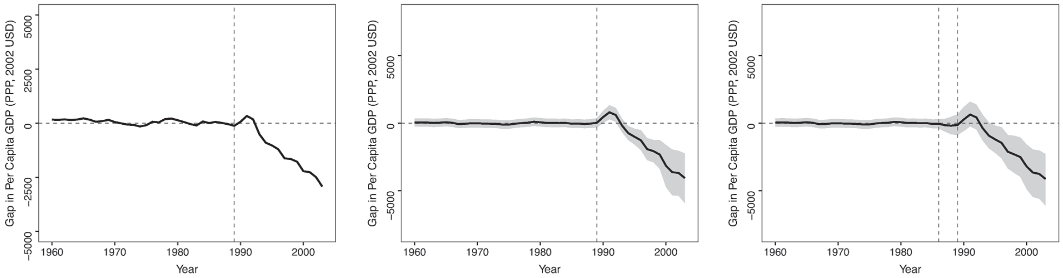
First, we build a DM-LFM incorporating all pre-treatment time-invariant covariates considered in ADH (2015), including pre-treatment averages of trade openness, inflation rate, industry share, schooling, and investment rate. The initial model we consider is $y_{it}(c) = \mathbf{X}_i(\boldsymbol{\beta} + \boldsymbol{\xi}_t) + \mathbf{f}_t\gamma_i + \epsilon_{it}$, in which $\mathbf{X}_i$ represents the time-invariant covariates. We do not include unit fixed effects or unit-varying coefficients to be consistent with the SCM, but time-varying coefficients are included. Figure A3 in Supplementary Material shows that the intercept has a strong time trend, but all the covariate coefficients are almost constant over time. Our result suggests four to six factors, with their ω's clearly having bimodal posteriors while several other factors exhibit mixed posteriors (Figure A4 in Supplementary Material).

We then produce an empirical posterior distribution of GDP per capita for counterfactual West Germany had German reunification not happened. To check the goodness-of-fit, we also calculate the model prediction of its GDP per capita in pre-treatment years. In Figure 5, we compare the counterfactual predictions of the SCM (a) with that of the DM-LFM (b), in which we shade the 95% Bayesian credibility intervals in gray. The dashed vertical line indicates the year 1989, 1 year before the adoption time. The two methods yield similar results: the GDP per capita of the counterfactual West Germany is higher than that of the actual West Germany during most of the post-reunification period except for the first few years after reunification.

Finally, we draw inferences about the treatment effects. Figure 6(a) and Figure 6(b) report the estimated effects of reunification on West Germany using the SCM and Bayesian DM-LFM, respectively. The corresponding 95% credibility intervals are added in (b). To lend further credibility to our causal estimates, we conduct a placebo test by setting 1987–1989, 3 years before reunification, as the placebo period. Figure 6(c) shows that the estimated effects at each time point during the placebo period are close to 0, buttressing our confidence in the identification assumptions.

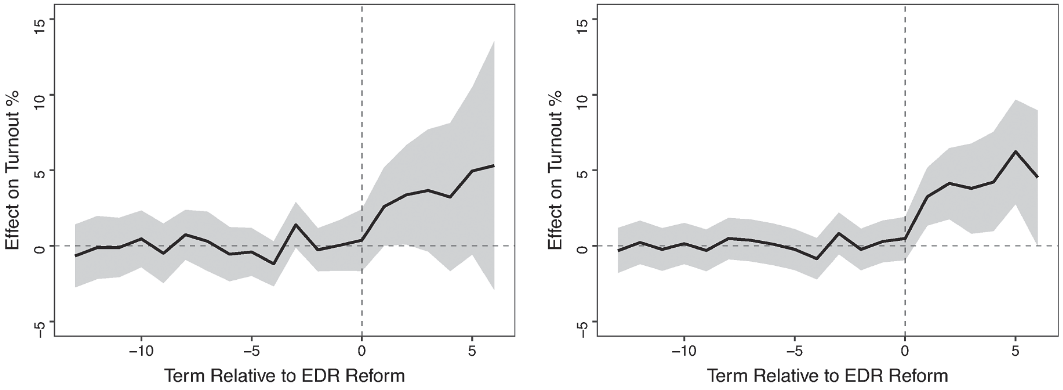
5.2 Election Day Registration and Voter Turnout

For this example, we specify a full DM-LFM model including the same time-varying covariates used in Xu (2017) (universal mail-in registration and motor voter registration) as well as 10 factors. Because there are only two covariates, we do not impose shrinkage on their β, but assign



(a) SCM (b) DM-LFM (c) DM-LFM (placebo)

Figure 6. Estimated effect of reunification on West Germany. This figure shows the estimated treatment effect of reunification on West Germany's per capita gross domestic product (GDP) using the synthetic control method (SCM) (a) and dynamic multilevel latent factor model (DM-LFM) (b), as well as the result from a placebo test using DM-LFM in which the 1987–1989 period (before reunification) is taken as "treated" (c).



(a) Gsynth (b) DM-LFM

Figure 7. The effect of election day registration (EDR) on voter turnout. This figure shows the estimated average treatment effect of EDR on voter turnout in the United States using Gsynth (Xu 2017) (a) and dynamic multilevel latent factor model (DM-LFM) (b), respectively. On the *x*-axis, the positive integers indicate the duration of treatment while the pre-treatment years are labeled as nonpositive integers. Period 1 is the first presidential election year in which a state implements EDR. Because the treated states adopted EDR at different points in time, the number of treated units decreases as *p* increases.

shrinkage priors to α_i, ξ_t, and γ_i. Our result suggests that at least six factors affect outcome prediction (Figure A9 in Supplementary Material). In contrast, Gsynth only includes two factors using a leave-one-out cross-validation procedure. As for α_i and ξ_t, the intercept varies in both time and space dimensions, but the varying parts of the slopes of the covariates are virtually shrunk to zero. We estimate the parameters using MCMC. Consistent with Xu (2017), we find that the two covariates do not explain much of the variation in turnout.

We then generate the posterior distributions of counterfactual outcomes for the nine treated states in their post-treatment years, based on which we estimate the effect of EDR on voter turnout. In Figure 7, we report the ATT for the same duration of adoption. To do so, we pool the posterior draws of the individual treatment effects of all treated states in p^{th} year after adoption for $p = 1, 2, \ldots, 6$. Using the posterior distributions of $\hat{\delta}_p$, we obtain their posterior means and Bayesian 95% credibility intervals.

Comparing the point and uncertainty estimates up to the sixth post-adoption presidential elections from Gsynth and the DM-LFM, the most notable difference between them is that the Bayesian 95% credibility intervals are considerably narrower than the 95% confidence intervals from Gsynth. We suspect that this is because our Bayesian approach has better predictive performance of individual counterfactuals than Gsynth as evidenced in Figure A11 in Supplementary Material, where we report the individual treatment effects on six treated states that have at least

Table 2. Comparing TSCS methods for comparative case studies.

	DiD	SCM	Gsynth	DM-LFM
Transparent designs	x	x		
Directly interpretable weights	x	x		
Addresses failure of “parallel trends”		x	x	x
Accommodates multiple treated units	x		x	x
Allows intercept shift	x		x	x
Accepts time-invariant covariates		x		x
Accepts time-varying covariates	x		x	x
Allows unit- and time-specific coefficients				x
Automated model or covariate selection		x	x	x
Model averaging				x
Inference on average treatment effects	x		x	x
Inference on individual treatment effects			x	x
Easily interpretable uncertainty measures				x
Low computational cost	x		x	

three post-treatment measures of the outcome using both Gsynth and the DM-LFM—our new method fits the trajectory of each treated unit in the pre-treatment period noticeably better than Gsynth.

6 Discussion

When is the DM-LFM applicable? And when is it more advantageous than existing methods? Table 2 summarizes its features in comparison to those of DiD, SCM, and Gsynth. The most important differences among the methods, we acknowledge, is that DiD and SCM adopt a clear design-based perspective and rely on more transparent identification assumptions—these assumptions are not necessarily weak, but they are widely understood and accepted by researchers. Moreover, the weights they impose on control units—uniform weights in DiD designs and non-negative weights with the SCM—are directly interpretable. Gsynth (a frequentist LFM) and Bayesian DM-LFM, on the other hand, use a model-based approach. Together with the SCM, they address potential failures of the “parallel trends” assumption by assuming a linear factor model. Unlike the SCM, they can easily accommodate comparative case studies with multiple treated units.

On modeling choices, DiD, Gsynth, and the DM-LFM allow intercept shift by assuming unit fixed effects, while the SCM does not (Doudchenko and Imbens 2017). The SCM can accept only time-invariant covariates while DiD and Gsynth can accept only time-varying ones. The DM-LFM, however, can accommodate both types of covariates and allow their coefficients to vary by unit and time. In terms of model selection, the SCM uses held-out pre-treatment periods to tweak the weighting matrix; Gsynth uses a cross-validation scheme to select the number of factors; and the DM-LFM conducts model selection scholastically in a larger model space and averages model predictions simultaneously.

All in all, the Bayesian DM-LFM is especially well suited for comparative case study, particularly when researchers suspect the conventional “parallel trends” assumption is unlikely to hold; when there are multiple treated units; when researchers would like to have easily interpretable uncertainty estimates on average or individual treatment effects; when many pre-treatment covariates are available and their relationships with the outcome are complex; or when time-varying confounders are complex and/or subtle enough that they need a relatively large number of factors to represent or some of the factors are relatively weak. The biggest disadvantage of the Bayesian approach is that it is computationally more expensive than frequentist methods such as Gsynth when the sample size is large. We also find that the frequentist properties of our method are unsatisfactory when there are too few control units or the number of pre-treatment periods is too small, for example, $T_0 < 20$ (see Table A4 in Supplementary Material).

Because of the strengths and weaknesses of different methods, we recommend researchers use multiple methods at the same time whenever possible to triangulate their findings. Future research should consider extending the method beyond the setting of staggered adoption, jointly modeling the treatment assignment mechanism and the response surface, and addressing potential SUTVA violations, such as policy diffusion and spillover effects.

Acknowledgements

We thank Dmitry Arkhangelsky, Neal Beck, Matt Blackwell, Sid Chib, Naoki Egami, Avi Feller, Matt Gabel, Jeff Gill, Kosuke Imai, Gary King, Dean Knox, John Londregan, Jacob Montgomery, Santiago Olivella, Jong Hee Park, Kevin Quinn, Marc Ratkovic, Guillermo Rosas, Betsy Sinclair, James F. Spriggs, Brandon Steward, Rocío Titiunik, Luwei Ying, Xiang Zhou, and participants at the Stanford Online Causal Inference Seminar, Annual Meeting of the Society for Political Methodology, the Harvard Applied Statistics Workshop, the Quantitative Social Science Colloquium at Princeton, and the workshop on "Frontiers of Bayesian Inference in Political Science" at Washington University in St. Louis. We are especially grateful to Associate Editor Dan Hopkins and three anonymous reviewers for their incredibly helpful comments. We thank Liz Crooks for excellent preproduction copy-editing. These authors contributed equally to this work and have decided to rotate the order of authorship in this and future collaborations.

Funding

This research is supported by the National Social Science Foundation of China (Grant No. 17ZDA110) as well as Stanford University.

Data Availability Agreement

Replication data and code for this article has been published at Harvard Dataverse at https://doi.org/10.7910/DVN/B6SWA1.

Supplementary Material

For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2021.22.

References

Abadie, A. 2020. "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects." *Journal of Economic Literature* 59(2):391–425.

Abadie, A., A. Diamond, and J. 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, A., A. Diamond, and J. Hainmueller. 2015. "Comparative Politics and the Synthetic Control Method." *American Journal of Political Science* 59(2):495–510.

Abadie, A., and J. Gardeazabal. 2003. "The Economic Costs of Conflict: A Case Study of the Basque Country." *The American Economic Review* 72(1):1–19.

Amjad, M., D. Shah, and D. Shen. 2018. "Robust Synthetic Control." *Journal of Machine Learning Research* 19:1–15.

Arkhangelsky, D., S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager. 2020. "Synthetic Difference in Differences." arXiv [stat.ME]. https://arxiv.org/abs/1812.09970.

Athey, S., M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi. 2018. "Matrix Completion Methods for Causal Panel Data Models." Technical report, National Bureau of Economic Research.

Athey, S., and G. W. Imbens. 2018. "Design-Based Analysis in Difference-in-Differences Settings with Staggered Adoption." Technical report, National Bureau of Economic Research.

Bai, J. 2009. “Panel Data Models with Interactive Fixed Effects.” *Econometrica* 77(4):1229–1279.

Bai, J., and S. Ng. 2020. “Matrix Completion, Counterfactuals, and Factor Analysis of Missing Data.” arXiv [econ.EM].

Beck, N., and J. N. Katz. 2007. “Random Coefficient Models for Time-Series-Cross-Section Data: Monte Carlo Experiments.” *Political Analysis* 15(2):182–195.

Belmonte, M. A., G. Koop, and D. Korobilis. 2014. “Hierarchical Shrinkage in Time-Varying Parameter Models.” *Journal of Forecasting* 33(1):80–94.

Ben-Michael, E., A. Feller, and J. Rothstein. 2020. “The Augmented Synthetic Control Method.” arXiv [stat.ME]. https://arxiv.org/abs/1811.04170.

Bitto, A., and S. Frühwirth-Schnatter. 2019. “Achieving Shrinkage in a Time-Varying Parameter Model Framework.” *Journal of Econometrics* 210(1):75–97.

Blackwell, M. 2013. “A Framework for Dynamic Causal Inference in Political Science.” *American Journal of Political Science* 57(2):504–519.

Blackwell, M., and A. Glynn. 2018. “How to Make Causal Inferences with Time-Series Cross-Sectional Data under Selection on Observables.” *American Political Science Review* 112(4):1067–1082.

Brodersen, K. H., F. Gallusser, J. Koehler, N. Remy, and S. L. Scott. 2014. “Inferring Causal Impact Using Bayesian Structural Time-Series Models.” *Annals of Applied Statistics* 9(1):247–274.

de Finetti, B. 1963. “Foresight: Its Logical Laws, Its Subjective Sources.” In *Studies in Subjective Probability*, edited by Jr. Henry E. Kyburg and H. E. Smokler. New York: John Wiley.

Ding, P., and F. Li. 2019. “A Bracketing Relationship between Difference-in-Differences and Lagged-Dependent-Variable Adjustment.” *Political Analysis* 27(4):605–615.

Doudchenko, N., and G. W. Imbens. 2017. “Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis.” arXiv [stat.AP]. https://arxiv.org/abs/1610.07748.

Gelman, A. 2006. “Analysis of Variance.” Technical Report, National Bureau of Economic Research.

Gelman, A. 2008. “Objections to Bayesian Statistics (Rejoint).” *Bayesian Analysis* 3(3):467–478.

Gobillon, L., and T. Magnac. 2016. “Regional Policy Evaluation: Interactive Fixed Effects and Synthetic Controls.” *Review of Economics and Statistics* 98(3):234–240.

Gutman, R., O. Intrator, and T. Lancaster. 2018. “A Bayesian Procedure for Estimating the Causal Effects of Nursing Home Bed-Hold Policy.” *Biostatistics* 19(4):444–460.

Hahn, J., and R. Shi. 2017. “Synthetic Control and Inference.” *Econometrics* 5(4):52–63.

Hazlett, C., and Y. Xu. 2018. Trajectory Balancing: A General Reweighting Approach to Causal Inference with Time-Series Cross-Sectional Data.” Social Science Research Network. https://papers.ssrn.com/abstract=3214231.

Hsiao, C., S. H. Ching, and S. K. 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):705–740.

Imai, K., and I. S. Kim. 2019. “When Should We Use Unit Fixed Effects Regression Model for Causal Inference with Longitudinal Data.” *American Journal of Political Science* 63(2):467–490.

Imai, K., I. S. Kim, and E. Wang. 2019. “Matching Methods for Causal Inference with Time-Series Cross-Sectional Data.” Mimeo, Harvard University. https://imai.fas.harvard.edu/research/tscs.html.

Imbens, G. W., and D. Rubin. 1997. “Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance.” *The Annuals of Statistics* 25(1):305–327.

Imbens, G. W., and D. B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press.

Kim, S., C. Lee, and S. Gupta. 2020. “Bayesian Synthetic Control Methods.” *Journal of Marketing Research* 57(5):831–852.

King, G., M. Tomz, and J. Wittenberg. 2000. “Making the Most of Statistical Analyses: Improving Interpretation and Presentation.” *American Journal of Political Science* 44(2):347–361.

Kyung, M., J. Gill, M. Ghosh, G. Casella, *et al.* 2010. “Penalized Tegression, Standard Errors, and Bayesian Lassos.” *Bayesian Analysis* 5(2):369–411.

Liu, L., Y. Wang, and Y. Xu. 2020. *A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data*. Stanford, CA: Mimeo, Stanford University.

Pang, X. 2010. “Modeling Heterogeneity and Serial Correlation in Binary Time-Series Cross-Sectional Data: A Bayesian Multilevel Model with AR(p) Errors.” *Political Analysis* 18(4):470–498.

Pang, X. 2014. “Varying Responses to Common Shocks and Complex Cross-Sectional Dependence: Dynamic Multilevel Modeling with Multifactor Error Structures for Time-Series Cross-Sectional Data.” *Political Analysis* 22(4):464–496.

Pang, X., L. Liu, and Y. Xu. 2021. “Replication Data for: A Bayesian Alternative to Synthetic Control for Comparative Case Studies.” https://doi.org/10.7910/DVN/B6SWA1, Harvard Dataverse, V1.

Park, T., and G. Casella. 2008. “The Bayesian Lasso.” *Journal of the American Statistical Association* 103(482):681–686.

Ricciardi, F., A. Mattei, and F. Mealli. 2020. “Bayesian Inference for Sequential Treatments Under Latent Sequential Ignorability.” *Journal of the American Statistical Association* 115(531):1498–1517.

https://doi.org/10.1017/pan.2021.22 Published online by Cambridge University Press

Rubin, D., X. Wang, L. Yin, and E. Zell. 2010. “Bayesian Causal Inference: Approaches to Estimating the Effect of Treating Hospital Type on Cancer Survival in Sweden Using Principal Stratification.” In *The Handbook of Applied Bayesian Analysis*, edited by A. O’Hagen and M. West. Oxford: Oxford University Press.

Rubin, D. B. 1976. “Inference and Missing Data.” *Biometrika* 63(3):581–592.

Rubin, D. B. 1978. “Bayesian Inference for Causal Effects: The Role of Randomization.” *The Annals of Statistics* 6(1):34–58.

Samartsidis, P. 2020. “A Bayesian Multivariate Factor Analysis Model for Evaluating an Intervention Using Observational Time-Series Data on Multiple Outcomes.” *Journal of the Royal Statistical Society. Series A: Statistics in Society* 183(4):1437–1459.

Strezhnev, A. 2018. “Semiparametric Weighting Estimators for Multi-Period Difference-in-Differences Designs.” *Working Paper*.

Vocht, F. de, K. Tilling, T. Pliakas, C. Angus, M. Egan, A. Brennan, R. Campbell, and M. Hickman. 2017. “The Intervention Effect of Local Alcohol Licensing Policies on Hospital Admission and Crime: A Natural Experiment Using a Novel Bayesian Synthetic Time-Series Method.” *Journal of Epidemiology & Community Health* 71(9):912–918.

Xu, Y. 2017. “Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models.” *Political Analysis* 25(1):57–76.

https://doi.org/10.1017/pan.2021.22 Published online by Cambridge University Press