PA

Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models

Yiqing Xu

Department of Political Science, University of California, San Diego. 9500 Gilman Drive #0521, La Jolla, CA 92093, USA.
Email: yiqingxu@ucsd.edu

Abstract

Difference-in-differences (DID) is commonly used for causal inference in time-series cross-sectional data. It requires the assumption that the average outcomes of treated and control units would have followed parallel paths in the absence of treatment. In this paper, we propose a method that not only relaxes this often-violated assumption, but also unifies the synthetic control method (Abadie, Diamond, and Hainmueller 2010) with linear fixed effects models under a simple framework, of which DID is a special case. It imputes counterfactuals for each treated unit using control group information based on a linear interactive fixed effects model that incorporates unit-specific intercepts interacted with time-varying coefficients. This method has several advantages. First, it allows the treatment to be correlated with unobserved unit and time heterogeneities under reasonable modeling assumptions. Second, it generalizes the synthetic control method to the case of multiple treated units and variable treatment periods, and improves efficiency and interpretability. Third, with a built-in cross-validation procedure, it avoids specification searches and thus is easy to implement. An empirical example of Election Day Registration and voter turnout in the United States is provided.

1 Introduction

Difference-in-differences (DID) is one of the most commonly used empirical designs in today's social sciences. The identifying assumptions for DID include the "parallel trends" assumption, which states that in the absence of the treatment the average outcomes of treated and control units would have followed parallel paths. This assumption is not directly testable, but researchers have more confidence in its validity when they find that the average outcomes of the treated and control units follow parallel paths in pretreatment periods. In many cases, however, parallel pretreatment trends are not supported by data, a clear sign that the "parallel trends" assumption is likely to fail in the posttreatment period as well. This paper attempts to deal with this problem systematically. It proposes a method that estimates the average treatment effect on the treated using time-series cross-sectional (TSCS) data when the "parallel trends" assumption is not likely to hold.

The presence of unobserved time-varying confounders causes the failure of this assumption. There are broadly two approaches in the literature to deal with this problem. The first one is to condition on pretreatment observables using matching methods, which may help balance the influence of potential time-varying confounders between treatment and control groups. For example, Abadie (2005) proposes matching before DID estimations. Although this method is easy

Political Analysis (2017)
vol. 25:57–76
DOI: 10.1017/pan.2016.2

Published
21 February 2017

Corresponding author
Yiqing Xu

Edited by
R. Michael Alvarez

Author's note: The author is indebted to Matt Blackwell, Devin Caughey, Justin Grimmer, Jens Hainmueller, Danny Hidalgo, Simon Jackman, Jonathan Katz, Luke Keele, Eric Min, Molly Roberts, Jim Snyder, Brandon Stewart, Teppei Yamamoto, as well as seminar participants at the 2015 MPSA Annual Meeting and 2015 APSA Annual Meeting for helpful comments and suggestions. The author thanks the editor, Mike Alvarez, and two anonymous reviewers for their extremely helpful suggestions. He also thanks Jushan Bai for generously sharing the Matlab codes used in Bai (2009) and Melanie Springer for kindly providing the state-level voter turnout data (1920–2000). The source code and data used in the paper can be downloaded from the *Political Analysis* Dataverse at dx.doi.org/10.7910/DVN/8AKACJ (Xu 2016) as well as the author's website.

to implement, it does not guarantee parallel pretreatment trends. The synthetic control method proposed by Abadie, Diamond, and Hainmueller (2010, 2015) goes one step further. It matches both pretreatment covariates and outcomes between a treated unit and a set of control units and uses pretreatment periods as criteria for good matches.[1] Specifically, it constructs a "synthetic control unit" as the counterfactual for the treated unit by reweighting the control units. It provides explicit weights for the control units, thus making the comparison between the treated and synthetic control units transparent. However, it only applies to the case of one treated unit and the uncertainty estimates it offers are not easily interpretable.[2]

The second approach is to model the unobserved time-varying heterogeneities explicitly. A widely used strategy is to add in unit-specific linear or quadratic time trends to conventional two-way fixed effects models. By doing so, researchers essentially rely upon a set of alternative identification assumptions that treatment assignment is ignorable conditional on both the fixed effects and the imposed trends (Mora and Reggio 2012). Controlling for these trends, however, often consumes a large number of degrees of freedom and may not necessarily solve the problem if the underlying confounders are not in forms of the specified trends.

An alternative way is to model unobserved time-varying confounders semiparametrically. For example, Bai (2009) proposes an interactive fixed effects (IFE) model, which incorporates unit-specific intercepts interacted with time-varying coefficients. The time-varying coefficients are also referred to as (latent) *factors* while the unit-specific intercepts are labeled as *factor loadings*. This approach builds upon an earlier literature on factor models in quantitative finance.[3] The model is estimated by iteratively conducting a factor analysis of the residuals from a linear model and estimating the linear model that takes into account the influences of a fixed number of most influential factors. Pang (2010, 2014) explores nonlinear IFE models with exogenous covariates in a Bayesian multi-level framework. Stewart (2014) provides a general framework of estimating IFE models based on a Bayesian variational inference algorithm. Gobillon and Magnac (2016) show that IFE models outperform the synthetic control method in DID settings when factor loadings of the treatment and control groups do not share common support.[4]

This paper proposes a *generalized synthetic control (GSC)* method that links the two approaches and unifies the synthetic control method with linear fixed effects models under a simple framework, of which DID is a special case. It first estimates an IFE model using only the control group data, obtaining a fixed number of latent factors. It then estimates factor loadings for each treated unit by linearly projecting pretreatment treated outcomes onto the space spanned by these factors. Finally, it imputes treated counterfactuals based on the estimated factors and factor loadings. The main contribution of this paper, hence, is to employ a latent factor approach to address a causal inference problem and provide valid, simulation-based uncertainty estimates under reasonable assumptions.

This method is in the spirit of the synthetic control method in the sense that by essence it is a reweighting scheme that takes pretreatment treated outcomes as benchmarks when choosing weights for control units and uses cross-sectional correlations between treated and control units to predict treated counterfactuals. Unlike the synthetic matching method, however, it conducts dimension reduction prior to reweighting such that vectors to be reweighted on are smoothed across control units. The method can also be understood as a bias correction procedure for IFE

1 See Hsiao, Ching, and Wan (2012) and Angrist, Jord, and Kuersteiner (2013) for alternative matching methods along this line of thought.

2 To gauge the uncertainty of the estimated treatment effect, the synthetic control method compares the estimated treatment effect with the "effects" estimated from placebo tests in which the treatment is randomly assigned to a control unit.

3 See Campbell, Lo, and MacKinlay (1997) for applications of factor models in finance.

4 For more empirical applications of the IFE estimator, see Kim and Oka (2014) and Gaibulloev, Sandler, and Sul (2014).

models when the treatment effect is heterogeneous across units.[5] It treats counterfactuals of treated units as missing data and makes out-of-sample predictions for posttreatment treated outcomes based on an IFE model.

This method has several advantages. First, it generalizes the synthetic control method to cases of multiple treated units and/or variable treatment periods. Since the IFE model is estimated only once, treated counterfactuals are obtained in a single run. Users therefore no longer need to find matches of control units for each treated unit one by one.[6] This makes the algorithm fast and less sensitive to the idiosyncrasies of a small number of observations.

Second, the GSC method produces frequentist uncertainty estimates, such as standard errors and confidence intervals, and improves efficiency under correct model specifications. A parametric bootstrap procedure based on simulated data can provide valid inference under reasonable assumptions. Since no observations are discarded from the control group, this method uses more information from the control group and thus is more efficient than the synthetic matching method when the model is correctly specified.

Third, it embeds a cross-validation scheme that selects the number of factors of the IFE model automatically, and thus is easy to implement. One advantage of the DID data structure is that treated observations in pretreatment periods can naturally serve as a validation dataset for model selection. We show that with sufficient data, the cross-validation procedure can pick up the correct number of factors with high probability, therefore reducing the risks of overfitting.

The GSC method has two main limitations. First, it requires more pretreatment data than fixed effects estimators. When the number of pretreatment periods is small, "incidental parameters" can lead to biased estimates of the treatment effects. Second, and perhaps more importantly, modeling assumptions play a heavier role with the GSC method than the original synthetic matching method. For example, if the treated and control units do not share common support in factor loadings, the synthetic matching method may simply fail to construct a synthetic control unit. Since such a problem is obvious to users, the chances that users misuse the method are small. The GSC method, however, will still impute treated counterfactuals based on model extrapolation, which may lead to erroneous conclusions. To safeguard against this risk, it is crucial to conduct various diagnostic checks, such as plotting the raw data, fitted values, and predicted counterfactuals.

The rest of the paper is organized as follows. Section 2 sets up the model and defines the quantities of interest. Section 3 introduces the GSC estimator, describes how it is implemented, and discuss the parametric bootstrap procedure. Section 4 reports simulation results that explores the finite sample properties of the GSC estimator and compares it with several existing methods. Section 5 illustrates the method with an empirical example that investigates the effect of Election Day Registration (EDR) laws on voter turnout in the United States. The last section concludes.

2 Framework

Suppose Y_{it} is the outcome of interest of unit i at time t. Let $\mathcal{T}$ and $\mathcal{C}$ denote the sets of units in treatment and control groups, respectively. The total number of units is $N = N_{tr} + N_{co}$, where N_{tr} and N_{co} are the numbers of treated and control units, respectively. All units are observed for T periods (from time 1 to time T). Let $T_{0,i}$ be the number of pretreatment periods for unit i, which

5 When the treatment effect is heterogeneous (as it is almost always the case), an IFE model that imposes a constant treatment effect assumption gives biased estimates of the average treatment effect because the estimation of the factor space is affected by the heterogeneity in the treatment effect.

6 For example, Acemoglu *et al.* (2016), who estimate the effect of Tim Geithner connections on stock market returns, conduct the synthetic control method repeatedly for each connected (treated) firm; Dube and Zipperer (2015) estimate the effect of minimum wage policies on wage and employment by conducting the method for each of the 29 policy changes. The latter also extend Abadie, Diamond, and Hainmueller (2010)'s original inferential method to the case of multiple treated units using the mean percentile ranks of the estimated effects.

is first exposed to the treatment at time $(T_{0,i} + 1)$ and subsequently observed for $q_i = T - T_{0,i}$ periods. Units in the control group are never exposed to the treatment in the observed time span. For notational convenience, we assume that all treated units are first exposed to the treatment at the same time, i.e., $T_{0,i} = T_0$ and $q_i = q$; variable treatment periods can be easily accommodated. First, we assume that Y_{it} is given by a linear factor model.

ASSUMPTION 1. Functional form:

$$Y_{it} = \delta_{it} D_{it} + x_{it}'\beta + \lambda_i' f_t + \varepsilon_{it},$$

where the treatment indicator D_{it} equals 1 if unit i has been exposed to the treatment prior to time t and equals 0 otherwise (i.e., $D_{it} = 1$ when $i \in \mathcal{T}$ and $t > T_0$ and $D_{it} = 0$ otherwise).[7] δ_{it} is the heterogeneous treatment effect on unit i at time t; x_{it} is a $(k \times 1)$ vector of observed covariates, $\beta = [\beta_1, \ldots, \beta_k]'$ is a $(k \times 1)$ vector of unknown parameters,[8] $f_t = [f_{1t}, \ldots, f_{rt}]'$ is an $(r \times 1)$ vector of unobserved common factors, $\lambda_i = [\lambda_{i1}, \ldots, \lambda_{ir}]'$ is an $(r \times 1)$ vector of unknown factor loadings, and ε_{it} represents unobserved idiosyncratic shocks for unit i at time t and has zero mean. Assumption 1 requires that the treated and control units are affected by the same set of factors and the number of factors is fixed during the observed time periods, i.e., no structural breaks are allowed.

The factor component of the model, $\lambda_i' f_t = \lambda_{i1} f_{1t} + \lambda_{i2} f_{2t} + \cdots + \lambda_{ir} f_{rt}$, takes a linear, additive form by assumption. In spite of the seemingly restrictive form, it covers a wide range of unobserved heterogeneities. First and foremost, conventional additive unit and time fixed effects are special cases. To see this, if we set $f_{1t} = 1$ and $\lambda_{i2} = 1$ and rewrite $\lambda_{i1} = \alpha_i$ and $f_{2t} = \xi_t$, then $\lambda_{i1} f_{1t} + \lambda_{i2} f_{2t} = \alpha_i + \xi_t$.[9] Moreover, the term also incorporates cases ranging from unit-specific linear or quadratic time trends to autoregressive components that researchers often control for when analyzing TSCS data.[10] In general, as long as an unobserved random variable can be decomposed into a multiplicative form, i.e., $U_{it} = a_i \times b_t$, it can be absorbed by $\lambda_i' f_t$ while it cannot capture unobserved confounders that are independent across units.

To formalize the notion of causality, we also use the notation from the potential outcomes framework for causal inference (Neyman 1923; Rubin 1974; Holland 1986). Let $Y_{it}(1)$ and $Y_{it}(0)$ be the potential outcomes for individual i at time t when $D_{it} = 1$ or $D_{it} = 0$, respectively. We thus have $Y_{it}(0) = x_{it}'\beta + \lambda_i' f_t + \varepsilon_{it}$ and $Y_{it}(1) = \delta_{it} + x_{it}'\beta + \lambda_i' f_t + \varepsilon_{it}$. The individual treatment effect on treated unit i at time t is therefore $\delta_{it} = Y_{it}(1) - Y_{it}(0)$ for any $i \in \mathcal{T}$, $t > T_0$.

We can rewrite the DGP of each unit as:

$$Y_i = D_i \circ \delta_i + X_i\beta + F\lambda_i + \varepsilon_i, \quad i \in 1, 2, \ldots N_{co}, N_{co} + 1, \ldots, N,$$

where $Y_i = [Y_{i1}, Y_{i2}, \ldots, Y_{iT}]'$; $D_i = [D_{i1}, D_{i2}, \ldots, D_{iT}]'$ and $\delta_i = [\delta_{i1}, \delta_{i2}, \ldots, \delta_{iT}]'$ (symbol "$\circ$" stands for point-wise product); $\varepsilon_i = [\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iT}]'$ are $(T \times 1)$ vectors; $X_i = [x_{i1}, x_{i2}, \ldots, x_{iT}]'$ is a $(T \times k)$ matrix; and $F = [f_1, f_2, \ldots, f_T]'$ is a $(T \times r)$ matrix.

7 Cases in which the treatment switches on and off (or "multiple-treatment-time") can be easily incorporated in this framework as long as we impose assumptions on how the treatment affects current and future outcomes. For example, one can assume that the treatment only affect the current outcome but not future outcomes (no carryover effect), as fixed effects models often do. In this paper, we do not impose such assumptions. See Imai and Kim (2016) for a thorough discussion.

8 β is assumed to be constant across space and time mainly for the purpose of fast computation in the frequentist framework. It is a limitation compared with more flexible and increasingly popular random coefficient models in Bayesian multi-level analysis.

9 For this reason, additive unit and time fixed effects are not explicitly assumed in the model. An extended model that directly imposes additive two-way fixed effects is discussed in the next section.

10 In the former case, we can set $f_{1t} = t$ and $f_{2t} = t^2$; in the latter case, for example, we can rewrite $Y_{it} = \rho Y_{i,t-1} + x_{it}'\beta + \varepsilon_{it}$ as $Y_{it} = Y_{i0} \cdot \rho^t + x_{it}'\beta + v_{it}$, in which v_{it} is an AR(1) process and ρ^t and Y_{i0} are the unknown factor and factor loadings, respectively. See Gobillon and Magnac (2016) for more examples.

The control and treated units are subscripted from 1 to N_{co} and from $N_{co}+1$ to N, respectively. The DGP of a control unit can be expressed as: $Y_i = X_i\beta + F\lambda_i + \varepsilon_i, i \in 1, 2, \ldots N_{co}$. Stacking all control units together, we have:

$$Y_{co} = X_{co}\beta + F\Lambda_{co}' + \varepsilon_{co}, \tag{1}$$

in which $Y_{co} = [Y_1, Y_2, \ldots, Y_{N_{co}}]$ and $\varepsilon_{co} = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{N_{co}}]$ are $(T \times N_{co})$ matrices; X_{co} is a three-dimensional $(T \times N_{co} \times p)$ matrix; and $\Lambda_{co} = [\lambda_1, \lambda_2, \ldots, \lambda_{N_{co}}]'$ is a $(N_{co} \times r)$ matrix, hence, the products $X_{co}\beta$ and $F\Lambda_{co}'$ are also $(T \times N_{co})$ matrices. To identify β, F and Λ_{co} in Equation (1), more constraints are needed. Following Bai (2003, 2009), I add two sets of constraints on the factors and factor loadings: (1) all factor are normalized, and (2) they are orthogonal to each other, i.e.: $F'F/T = I_r$ and $\Lambda_{co}'\Lambda_{co}$ = diagonal.[11] For the moment, the number of factors r is assumed to be known. In the next section, we propose a cross-validation procedure that automates the choice of r.

The main quantity of interest of this paper is the average treatment effect on the treated (ATT) at time t (when $t > T_0$):

$$ATT_{t,t>T_0} = \frac{1}{N_{tr}} \sum_{i \in \mathcal{T}} [Y_{it}(1) - Y_{it}(0)] = \frac{1}{N_{tr}} \sum_{i \in \mathcal{T}} \delta_{it}.^{12}$$

Note that in this paper, as in Abadie, Diamond, and Hainmueller (2010), we treat the treatment effects δ_{it} as given once the sample is drawn.[13] Because $Y_{it}(1)$ is observed for treated units in posttreatment periods, the main objective of this paper is to construct counterfactuals for each treated unit in posttreatment periods, i.e., $Y_{it}(0)$ for $i \in \mathcal{T}$ and $t > T_0$. The problem of causal inference indeed turns into a problem of forecasting missing data.[14]

2.1 Assumptions for causal identification

In addition to the functional form assumption (Assumption 1), three assumptions are required for the identification of the quantities of interest. Among them, the assumption of strict exogeneity is the most important.

ASSUMPTION 2. Strict exogeneity.

$$\varepsilon_{it} \perp\!\!\!\perp D_{js}, x_{js}, \lambda_j, f_s \quad \forall i, j, t, s.$$

Assumption 2 means that the error term of any unit at any time period is independent of treatment assignment, observed covariates, and unobserved cross-sectional and temporal heterogeneities

11 These constraints do not lead to loss of generality because for an arbitrary pair of matrices F and Λ_{co}, we can find an $(r \times r)$ invertible matrix A such that $(FA)'(FA)/T = I_r$ and $(A^{-1}\Lambda_{co})'A^{-1}\Lambda_{co}$ is a diagonal matrix. To see this, we can then rewrite $\lambda_i'F$ as $\tilde{\lambda}_i'\tilde{F}$, in which $\tilde{F} = FA$ and $\tilde{\lambda}_i = A^{-1}\lambda_i$ for units in both the treatment and control groups such that $\tilde{F}$ and $\tilde{\Lambda}_{co}$ satisfy the above constraints. The total number of constraints is r^2, the dimension of the matrix space where A belongs. It is worth noting that although the original factors F may not be identifiable, the space spanned by F, a r-dimensional subspace of in the T-dimensional space, is identified under the above constraints because for any vector in the subspace spanned by $\tilde{F}$, it is also in the subspace spanned by the original factors F.

12 For a clear and detailed explanation of quantities of interest in TSCS analysis, see Blackwell and Glynn (2015). Using their terminology, this paper intends to estimate the Average Treatment History Effect on the Treated given two specific treatment histories: $\mathbb{E}[Y_{it}(\underline{a}_t^1) - Y_{it}(\underline{a}_t^0)|\underline{D}_{i,t-1} = \underline{a}_{t-1}^1]$ in which $\underline{a}_t^0 = (0, \ldots, 0)$, $\underline{a}_t^1 = (0, \ldots, 0, 1, \ldots, 1)$ with T_0 zeros and $(t - T_0)$ ones indicate the histories of treatment statuses. We keep the current notation for simplicity.

13 We attempt to make inference about the ATT in the sample we draw, not the ATT of the population. In other words, we do not incorporate uncertainty of the treatment effects δ_{it}.

14 The idea of predicting treated counterfactuals in a DID setup is also explored by Brodersen *et al.* (2014) using a structural Bayesian time-series approach.

of all units (including itself) at all periods. We call it a *strict* exogeneity assumption, which implies conditional mean independence, i.e., $\mathbb{E}[\varepsilon_{it}|D_{it}, x_{it}, \lambda_i, f_t] = \mathbb{E}[\varepsilon_{it}|x_{it}, \lambda_i, f_t] = 0$.[15]

Assumption 2 is arguably weaker than the strict exogeneity assumption required by fixed effects models when decomposable time-varying confounders are at present. These confounders are decomposable if they can take forms of heterogeneous impacts of a common trend or a series of common shocks. For instance, suppose a law is passed in a state because the public opinion in that state becomes more liberal. Because changing ideologies are often cross-sectionally correlated across states, a latent factor may be able to capture shifting ideology at the national level; the national shifts may have a larger impact on a state that has a tradition of mass liberalism or has a higher proportion of manufacturing workers than a state that is historically conservative. Controlling for this unobserved confounder, therefore, can alleviate the concern that the passage of the law is endogenous to changing ideology of a state's constituents to a great extent.

When such a confounder exists, with two-ways fixed effects models we need to assume that $(\varepsilon_{it} + \lambda_i f_t) \perp\!\!\!\perp D_{js}, x_{js}, \alpha_j, \xi_s, \forall i, j, t, s$ (with $\lambda_i f_t$, α_j and ξ_s representing the time-varying confounder for unit i at time t, fixed effect for unit j, and fixed effect for time s, respectively) for the identification of the constant treatment effect. This is implausible because $\lambda_i f_t$ is likely to be correlated with D_{it}, x_{it}, and α_i, not to mention other terms. In contrast, Assumption 2 allows the treatment indicator to be correlated with both x_{js} and $\lambda_j' f_s$ for any unit j at any time periods s (including i and t themselves).

Identifying the treatment effects also requires the following assumptions:

ASSUMPTION 3. Weak serial dependence of the error terms.

ASSUMPTION 4. Regularity conditions.

Assumptions 3 and 4 (see the Online Appendix in Supplementary Materials for details) are needed for the consistent estimation of β and the space spanned by F (or $F'F/T$). Similar, though slightly weaker, assumptions are made in Bai (2009) and Moon and Weidner (2015). Assumption 3 allows weak serial correlations but rules out strong serial dependence, such as unit root processes; errors of different units are uncorrelated. A sufficient condition for Assumption 3 to hold is that the error terms are not only independent of covariates, factors, and factor loadings, but also independent both across units and over time, which is assumed in Abadie, Diamond, and Hainmueller (2010). Assumption 4 specifies moment conditions that ensure the convergence of the estimator.

For valid inference based on a block bootstrap procedure discussed in the next section, we also need Assumption 5 (see Online Appendix for details). Heteroscedasticity across time, however, is allowed.

ASSUMPTION 5. The error terms are cross-sectionally independent and homoscedastic.

REMARK 1. Assumptions 3 and 5 suggest that the error terms ε_{it} can be serially correlated. Assumption 2 rules out dynamic models with lagged dependent variables; however, this is mainly for the purpose of simplifying proofs (Bai 2009, p. 1243). The proposed method can accommodate dynamic models as long as the error terms are not serially correlated.

3 Estimation Strategy

In this section, we first propose a GSC estimator for treatment effect of each treated unit. It is essentially an out-of-sample prediction method based on Bai (2009)'s factor augmented model.

15 Note that because ε_{it} is independent of D_{is} and x_{is} for all (t, s), Assumption 2 rules out the possibility that past outcomes may affect future treatments, which is allowed by the so called "sequential exogeneity" assumption. A directed acyclic graph (DAG) representation is provided in the Online Appendix. See Blackwell and Glynn (2015) and Imai and Kim (2016) for discussions on the difference between the strict ignorability and sequential ignorability assumptions. What is unique here is that we conditional on unobserved factors and factor loadings.

The GSC estimator for the treatment effect on treated unit i at time t is given by the difference between the actual outcome and its estimated counterfactual: $\hat{\delta}_{it} = Y_{it}(1) - \hat{Y}_{it}(0)$, in which $\hat{Y}_{it}(0)$ is imputed with three steps. In the first step, we estimate an IFE model using only the control group data and obtain $\hat{\beta}, \hat{F}, \hat{\Lambda}_{co}$:

$$\textbf{Step 1.} \qquad (\hat{\beta}, \hat{F}, \hat{\Lambda}_{co}) = \underset{\tilde{\beta}, \tilde{F}, \tilde{\Lambda}_{co}}{\operatorname{argmin}} \sum_{i \in C} (Y_i - X_i\tilde{\beta} - \tilde{F}\tilde{\lambda}_i)'(Y_i - X_i\tilde{\beta} - \tilde{F}\tilde{\lambda}_i)$$
$$s.t. \qquad \tilde{F}'\tilde{F}/T = I_r \quad \text{and} \quad \tilde{\Lambda}_{co}'\tilde{\Lambda}_{co} = \text{diagonal}.$$

We explain in detail how to estimate this model in the Online Appendix. The second step estimates factor loadings for each treated unit by minimizing the mean squared error of the predicted treated outcome in pretreatment periods:

$$\textbf{Step 2.} \qquad \hat{\lambda}_i = \underset{\tilde{\lambda}_i}{\operatorname{argmin}} (Y_i^0 - X_i^0\hat{\beta} - \hat{F}^0\tilde{\lambda}_i)'(Y_i^0 - X_i^0\hat{\beta} - \hat{F}^0\tilde{\lambda}_i)$$
$$= (\hat{F}^{0\prime}\hat{F}^0)^{-1}\hat{F}^{0\prime}(Y_i^0 - X_i^0\hat{\beta}), \quad i \in \mathcal{T},$$

in which $\hat{\beta}$ and $\hat{F}^0$ are from the first-step estimation and the superscripts "0"s denote the pretreatment periods. In the third step, we calculate treated counterfactuals based on $\hat{\beta}, \hat{F}$, and $\hat{\lambda}_i$:

$$\textbf{Step 3.} \qquad \hat{Y}_{it}(0) = x_{it}'\hat{\beta} + \hat{\lambda}_i'\hat{f}_t \quad i \in \mathcal{T},\ t > T_0.$$

An estimator for ATT_t therefore is: $\widehat{ATT}_t = (1/N_{tr}) \sum_{i \in \mathcal{T}} [Y_{it}(1) - \hat{Y}_{it}(0)]$ for $t > T_0$.

REMARK 2. In the Online Appendix, we show that, under Assumptions 1–4, the bias of the GSC shrinks to zero as the sample size grows, i.e., $\mathbb{E}_\varepsilon(\widehat{ATT}_t | D, X, \Lambda, F) \to ATT_t$ as $N_{co}, T_0 \to 0$ (N_{tr} is taken as given), in which $D = [D_1, D_2, \ldots, D_N]$ is a $(T \times N)$ matrix, X is a three-dimensional $(T \times N \times p)$ matrix; and $\Lambda = [\lambda_1, \lambda_2, \ldots, \lambda_N]'$ is a $(N \times r)$ matrix. Intuitively, both large N_{co} and large T_0 are necessary for the convergences of $\hat{\beta}$ and the estimated factor space. When T_0 is small, imprecise estimation of the factor loadings, or the "incidental parameters" problem, will lead to bias in the estimated treatment effects. This is a crucial difference from the conventional linear fixed effects models.

3.1 Model selection

In practice, researchers may have limited knowledge of the exact number of factors to be included in the model. Therefore, we develop a cross-validation procedure to select models before estimating the causal effect. It relies on the control group information as well as information from the treatment group in pretreatment periods. Algorithm 1 describes the details of this procedure.

ALGORITHM 1 (Cross-validating the number of factors). A leave-one-out-cross-validation procedure that selects the number of factors takes the following steps:

Step 1. Start with a given number of factors r, estimate an IFE model using the control group data $\{Y_i, X_i\}_{i \in C}$, obtaining $\hat{\beta}$ and $\hat{F}$;

Step 2. **Start** a cross-validation loop that goes through all T_0 pretreatment periods:

(a) In round $s \in \{1, \ldots, T_0\}$, hold back data of all treated units at time s. Run an OLS regression using the rest of the pretreatment data, obtaining factor loadings for each treated unit i:

$$\hat{\lambda}_{i,-s} = (F_{-s}^{0\prime}F_{-s}^0)^{-1}F_{-s}^{0\prime}(Y_{i,-s}^0 - X_{i,-s}^{0\prime}\hat{\beta}), \quad \forall i \in \mathcal{T},$$

in which the subscript "-s" stands for all pretreatment periods except for s.

(b) Predict the treated outcomes at time s using $\hat{Y}(0)_{is} = x'_{is}\hat{\beta} + \hat{\lambda}'_{i,-s}\hat{f}_s$ and save the prediction error $e_{is} = Y_{is}(0) - \hat{Y}_{is}(0)$ for all $i \in \mathcal{T}$.

End of the cross-validation loop;

Step 3. Calculate the mean square prediction error (MSPE) given r,

$$MSPE(r) = \sum_{s=1}^{T_0} \sum_{i \in \mathcal{T}} e_{is}^2 / T_0.$$

Step 4. Repeat Steps 1–3 with different r's and obtain corresponding MSPEs.

Step 5. Choose r^* that minimizes the MSPE.

The basic idea of the above procedure is to hold back a small amount of data (e.g., one pretreatment period of the treatment group) and use the rest of data to predict the held-back information. The algorithm then chooses the model that on average makes the most accurate predictions. A TSCS dataset with a DID data structure allows us to do so because (1) there exists a set of control units that are never exposed to the treatment and therefore can serve as the basis for estimating time-varying factors and (2) the pretreatment periods of treated units constitute a natural validation set for candidate models. This procedure is computationally inexpensive because with each r, the IFE model is estimated only once (Step 1). Other steps involves merely simple calculations. In the Online Appendix, we conduct Monte Carlo exercises and show that the above procedure performs well in term of choosing the correct number of factors even with relatively small datasets.

REMARK 3. Our framework can also accommodate DGPs that directly incorporate additive fixed effects, known time trends, and exogenous time-invariant covariates, such as:

$$Y_{it} = \delta_{it} D_{it} + x'_{it}\beta + \gamma'_i I_t + z'_i \theta_t + \lambda'_i f_t + \alpha_i + \xi_t + \varepsilon_{it}, \tag{2}$$

in which I_t is a $(q \times 1)$ vector of known time trends that may affect each unit differently; γ_i is $(q \times 1)$ unit-specific unknown parameters; z_i is a $(m \times 1)$ vector of observed time-invariant covariates; θ_t is a $(m \times 1)$ vector of unknown parameters; α_i and ξ_t are additive individual and time fixed effects, respectively. We describe the estimation procedure of this extended model in the Online Appendix.

3.2 Inference

We rely on a parametric bootstrap procedure to obtain the uncertainty estimates of the GSC estimator (deriving the analytical asymptotic distribution of the GSC estimator is a necessary step for future research). When the sample size is large, when N_{tr} is large in particular, a simple nonparametric bootstrap procedure can provide valid uncertainty estimates. When the sample size is small, especially when N_{tr} is small, we are unable to approximate the DGP of the treatment group by resampling the data nonparametrically. In this case, we simply lack the information of the joint distribution of $(X_i, \lambda_i, \delta_i)$ for the treatment group. However, we can obtain uncertainty estimates conditional on observed covariates and unobserved factors and factor loadings using a parametric bootstrap procedure via resampling the residuals. By resampling entire time series of residuals, we preserve the serial correlation within the units, thus avoiding underestimating the standard errors due to serial correlations (Beck and Katz 1995). Our goal is to estimate the conditional variance of ATT estimator, i.e., $\text{Var}_\varepsilon(\widehat{ATT}_t | D, X, \Lambda, F)$. Notice that the only random variable that is not being conditioned on is ε_i, which are assumed to be independent of treatment

assignment, observed covariates, factors and factor loadings (Assumption 2). We can interpret ε_i as measurement errors or variations in the outcome that we cannot explain but are unrelated to treatment assignment.[16]

In the parametric bootstrap procedure, we simulate treated counterfactuals and control units based on the following resampling scheme:

$$\tilde{Y}_i(0) = X_i\hat{\beta} + \hat{F}\hat{\lambda}_i + \tilde{\varepsilon}_i, \quad \forall i \in \mathcal{C};$$
$$\tilde{Y}_i(0) = X_i\hat{\beta} + \hat{F}\hat{\lambda}_i + \tilde{\varepsilon}_i^p, \quad \forall i \in \mathcal{T},$$

in which $\tilde{Y}_i(0)$ is a vector of simulated outcomes in the absence of the treatment; $X_i\hat{\beta} + \hat{F}\hat{\lambda}_i$ is the estimated conditional mean; and $\tilde{\varepsilon}_i$ and $\tilde{\varepsilon}_i^p$ are resampled residuals for unit i, depending on whether it belongs to the treatment or control group. Because $\hat{\beta}$ are $\hat{F}$ are estimated using only the control group information, $X_i\hat{\beta} + \hat{F}\hat{\lambda}_i$ fits $X_i\beta + F\lambda_i$ better for a control unit than for a treated unit (as a result, the variance of $\tilde{\varepsilon}_i^p$ is usually bigger than that of $\tilde{\varepsilon}_i$). Hence, $\tilde{\varepsilon}_i$ and $\tilde{\varepsilon}_i^p$ are drawn from different empirical distributions: $\tilde{\varepsilon}_i$ is the in-sample error of the IFE model fitted to the control group data, and therefore is drawn from the empirical distribution of the residuals of the IFE model, while $\tilde{\varepsilon}_i^p$ can be seen as the prediction error of the IFE model for treated counterfactuals.[17]

Although we cannot observe treated counterfactuals, $Y_{it}(0)$ is observed for all control units. With the assumptions that treated and control units follow the same factor model (Assumption 1) and the error terms are independent and homoscedastic across space (Assumption 5), we can use a cross-validation method to simulate ε_i^p based on the control group data (Efron 2012). Specifically, each time we leave one control unit out (to be taken as a "fake" treat unit) and use the rest of the control units to predict the outcome of left-out unit. The difference between the predicted and observed outcomes is a prediction error of the IFE model. ε_i^p is drawn from the empirical distributions of the prediction errors. Under Assumptions 1–5, this procedure provides valid uncertainty estimates for the proposed method without making particular distributional assumptions of the error terms. Algorithm 2 describes the entire procedure in detail.

ALGORITHM 2 (Inference). A parametric bootstrap procedure that gives the uncertainty estimates of the ATT is described as follows:

Step 1. **Start** a loop that runs B_1 times:

(a) In round $m \in \{1, \ldots, B_1\}$, randomly select one control unit i as if it was treated when $t > T_0$.

(b) Resample the rest of the control group with replacement of size N_{co} and form a new sample with one "treated" unit and N_{co} resampled control units.

(c) Apply the GSC method to the new sample, obtaining a vector of prediction error, or residuals; $\hat{\varepsilon}_{(m)}^p = Y_i - \hat{Y}_i(0)$.

End of the loop, collecting $\hat{\mathbf{e}}^{\mathbf{p}} = \{\hat{\varepsilon}_{(1)}^p, \hat{\varepsilon}_{(2)}^p \cdots, \hat{\varepsilon}_{(B_1)}^p\}$.

Step 2. Apply the GSC method to the original data, obtaining: (1) $\widehat{ATT}_t$ for all $t > T_0$, (2) estimated coefficients: $\hat{\beta}, \hat{F}, \hat{\lambda}_{co}$, and $\hat{\lambda}_{j,j\in\mathcal{T}}$, and (3) the fitted values and residuals of the control units: $\hat{\mathbf{Y}}_{\mathbf{co}} = \{\hat{Y}_1(0), \hat{Y}_2(0), \ldots, \hat{Y}_{N_{co}}(0)\}$ and $\hat{\mathbf{e}} = \{\hat{\varepsilon}_1, \hat{\varepsilon}_2, \ldots, \hat{\varepsilon}_{N_{co}}\}$.

16 ε_{it} may be correlated with $\hat{\lambda}_i$ when the errors are serially correlated because $\hat{\lambda}_i$ is estimated using the pretreatment data.

17 The treated outcome for unit i, thus can be drawn from $\tilde{Y}_i(1) = \tilde{Y}_i(0) + \delta_i$. We do not directly observe δ_i, but since it is taken as given, its presence will not affect the uncertainty estimates of $\widehat{ATT}_t$. Hence, in the bootstrap procedure, we use $\tilde{Y}_i(0)$ for both the treatment and control groups to form bootstrapped samples (set $\delta_i = \mathbf{0}$, for all $i \in \mathcal{T}$). We will add back $\widehat{ATT}_t$ when constructing confidence intervals.

Step 3. **Start** a bootstrap loop that runs B_2 times:

(a) In round $k \in \{1, \ldots, B_2\}$, construct a bootstrapped sample $S^{(k)}$ by:

$$\tilde{Y}_i^{(k)}(0) = \hat{Y}_i(0) + \tilde{\varepsilon}_i, \quad i \in \mathcal{C}$$
$$\tilde{Y}_i^{(k)}(0) = \hat{Y}_i(0) + \tilde{\varepsilon}_i^p, \quad j \in \mathcal{T}$$

in which each vector of $\tilde{\varepsilon}_i$ and $\tilde{\varepsilon}_j^p$ are randomly selected from sets $\mathbf{e}$ and $\mathbf{e}^p$, respectively, and $\hat{Y}_i(0) = X_i\hat{\beta} + \hat{F}\hat{\lambda}_i$. Note that the simulated treated counterfactuals do not contain the treatment effect.

(b) Apply the GSC method to $S^{(k)}$ and obtain a new ATT estimate; add $\widehat{ATT}_{t,t>T_0}$ to it, obtaining the bootstrapped estimate $\widehat{ATT}^{(k)}_{t,t>T_0}$.

End of the bootstrap loop.

Step 4. Compute the variance of $\widehat{ATT}_{t,t>T_0}$ using

$$\text{Var}(\widehat{ATT}_t|D, X, \Lambda, F) = \frac{1}{B}\sum_{k=1}^{B}\left(\widehat{ATT}_t^{(k)} - \frac{1}{B}\sum_{j=1}^{B}\widehat{ATT}_t^{(j)}\right)^2$$

and its confidence interval using the conventional percentile method (Efron and Tibshirani 1993).

4 Monte Carlo Evidence

In this section, we conduct Monte Carlo exercises to explore the finite sample properties of the GSC estimator and compare it with several existing methods, including the DID estimator, the IFE estimator, and the original synthetic matching method. We also investigate the extent to which the proposed cross-validation scheme can choose the number of factors correctly in relatively small samples.

We start with the following data generating process (DGP) that includes two observed time-varying covariates, two unobserved factors, and additive two-way fixed effects:

$$Y_{it} = \delta_{it}D_{it} + x_{it,1}\cdot 1 + x_{it,2}\cdot 3 + \lambda_i'f_t + \alpha_i + \xi_t + 5 + \varepsilon_{it} \tag{3}$$

where $f_t = (f_{1t}, f_{2t})'$ and $\lambda_i = (\lambda_{i1}, \lambda_{i2})'$ are time-varying factors and unit-specific factor loadings. The covariates are (positively) correlated with both the factors and factor loadings: $x_{it,k} = 1 + \lambda_i'f_t + \lambda_{i1} + \lambda_{i2} + f_{1t} + f_{2t} + \eta_{it,k}$, $k = 1, 2$. The error term ε_{it} and disturbances in covariates $\eta_{it,1}$ and $\eta_{it,2}$ are i.i.d. $N(0, 1)$. Factors f_{1t} and f_{2t}, as well as time fixed effects ξ_t, are also i.i.d. $N(0, 1)$. The treatment and control groups consist of N_{tr} and N_{co} units. The treatment starts to affect the treated units at time $T_0 + 1$ and since then 10 periods are observed ($q = 10$). The treatment indicator is defined as in Section 2, i.e., $D_{it} = 1$ when $i \in \mathcal{T}$ and $t > T_0$ and $D_{it} = 0$ otherwise. The heterogeneous treatment effect is generated by $\delta_{it,t>T_0} = \bar{\delta}_t + e_{it}$, in which e_{it} is i.i.d. N(0,1). $\bar{\delta}_t$ is given by: $[\bar{\delta}_{T_0+1}, \bar{\delta}_{T_0+1}, \ldots, \bar{\delta}_{T_0+10}] = [1, 2, \ldots, 10]$.

Factor loadings λ_{i1} and λ_{i2}, as well as unit fixed effects α_i, are drawn from uniform distributions $U[-\sqrt{3}, \sqrt{3}]$ for control units and $U[\sqrt{3} - 2w\sqrt{3}, 3\sqrt{3} - 2w\sqrt{3}]$ for treated units ($w \in [0, 1]$). This means that when $0 \leq w < 1$, (1) the random variables have variance 1; (2) the supports of factor loadings of treated and control units are not perfectly overlapped; and (3) the treatment indicator and factor loadings are positively correlated.[18]

18 The DGP specified here is modified based on Bai (2009) and Gobillon and Magnac (2016).

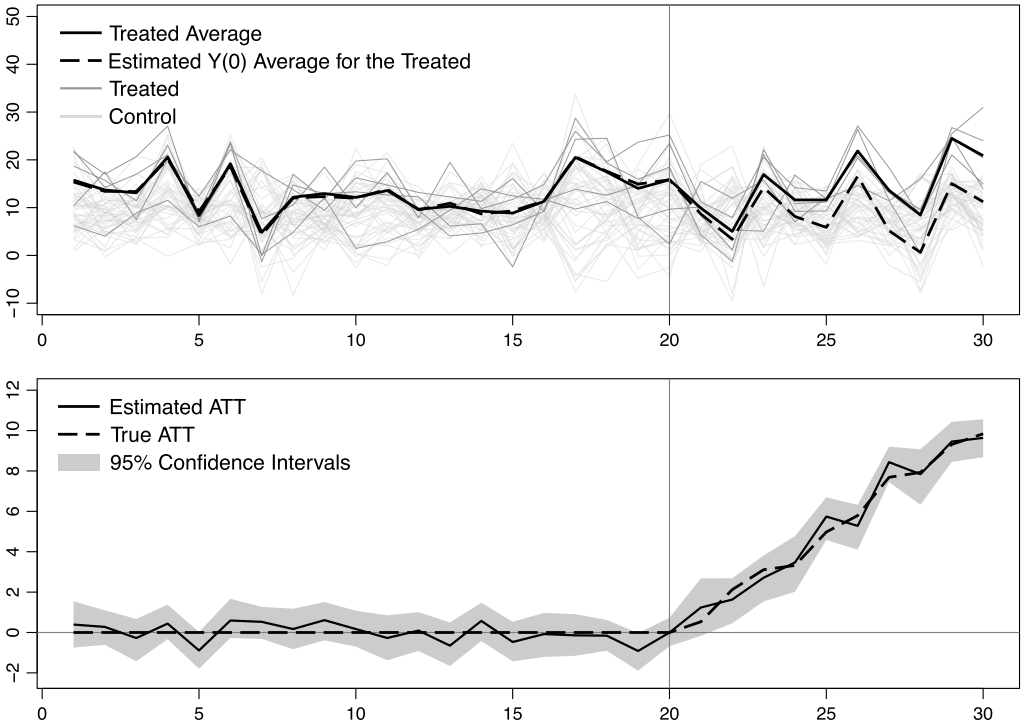


Figure 1. Estimated ATT for a simulated sample $N_{tr} = 5$, $N_{co} = 45$, $T = 30$, $T_0 = 10$.

4.1 A simulated example

We first illustrate the proposed method, as well as the DGP described above, with a simulated sample of $N_{tr} = 5$, $N_{co} = 45$, and $T_0 = 20$ (hence, $N = 50$, $T = 30$). w is set to be 0.8, which means that the treated units are more likely to have larger factor loadings than the control units. Figure 1 visualizes the raw data and estimation results. In the upper panel, the dark and light gray lines are time series of the treated and control units, respectively. The bold solid line is the average outcome of the five treated units while the bold dashed line is the average predicted outcome of the five units in the absence of the treatment. The latter is imputed using the proposed method.

The lower panel of Figure 1 shows the estimated ATT (solid line) and the true ATT (dashed line). The 95 percent confidence intervals for the ATT are based on bootstraps of 2,000 times. It shows that the estimated average treated outcome fits the data well in pretreatment periods and the estimated ATT is very close to the actual ATT. The estimated factors and factor loadings, as well as imputed counterfactual and individual treatment effect for each treat unit, are shown in the Online Appendix.

4.2 Finite sample properties

We present the Monte Carlo evidence on the finite sample properties of the GSC estimator in Table 1 (additional results are shown in the Online Appendix). As in the previous example, the treatment group is set to have five units. The estimand is the ATT at time $T_0 + 5$, whose expected value equals 5. Observables, factors, and factor loadings are drawn only once while the error term is drawn repeatedly; w is set to be 0.8 such that treatment assignment is positively correlated with factor loadings. Table 1 reports the bias, standard deviation (SD), and root mean squared error (RMSE) of $\widehat{ATT}_{T_0+5}$ from 5,000 simulations for each pair of T_0 and N_{co}.[19] It shows that

19 Standard deviation is defined as: $SD(\widehat{ATT}_t) = \sqrt{\mathbb{E}[\widehat{ATT}_t^{(k)} - \mathbb{E}(\widehat{ATT}_t^{(k)})]^2}$, while root mean squared error is defined as: $RMSE(\widehat{ATT}_t) = \sqrt{\mathbb{E}(\widehat{ATT}_t^{(k)} - ATT_t^{(k)})^2}$. The superscript (k) denotes the kth sample. We see that they are very close because the bias of the GSC estimator shrinks to zero as the sample size grows.

Table 1. Finite sample properties and coverage rates

N_{tr}	N_{co}	T_0	*Bias*	*SD*	*RMSE*	*Coverage*
5	40	15	0.053	0.589	0.591	0.947
5	80	15	0.017	0.535	0.536	0.949
5	120	15	0.010	0.524	0.524	0.949
5	200	15	0.011	0.518	0.518	0.949
5	40	30	0.046	0.538	0.540	0.946
5	80	30	0.021	0.504	0.505	0.948
5	120	30	0.024	0.494	0.495	0.949
5	200	30	0.008	0.487	0.487	0.949
5	40	50	0.031	0.519	0.520	0.947
5	80	50	0.016	0.497	0.498	0.948
5	120	50	0.003	0.475	0.475	0.949
5	200	50	0.016	0.468	0.469	0.949

the GSC estimator has limited bias even when T_0 and N_{co} are relatively small and the bias goes away as T_0 and N_{co} grow. As expected, both the SD and RMSE shrink when T_0 and N_{co} become larger. Table 1 also reports the coverage probabilities of 95 percent confidence intervals for $\widehat{ATT}_{i,T_0+5}$ constructed by the parametric bootstrap procedure (Algorithm 2). For each pair of T_0 and N_{co}, the coverage probability is calculated based on 5,000 simulated samples, each of which is bootstrapped for 1,000 times. These numbers show that the proposed procedure can achieve the correct coverage rate even when the sample size is relatively small (e.g., $T_0 = 15$, $N_{tr} = 5$, $N_{co} = 80$).

In the Online Appendix, we run additional simulations and compare the proposed method with several existing methods, including the DID estimator, the IFE estimator, and the synthetic matching method. We find that (1) the GSC estimator has less bias than the DID estimator in the presence of unobserved, decomposable time-varying confounders; (2) it has less bias than the IFE estimator when the treatment effect is heterogeneous; and (3) it is usually more efficient than the original synthetic matching estimator. It is worth emphasizing that these results are under the premise of correct model specifications. To address the concern that the GSC method relies on correct model specifications, we conduct additional tests and show that the cross-validation scheme described in Algorithm 1 is able to choose the number of factors correctly most of the time when the sample is large enough.

5 Empirical Example

In this section, we illustrate the GSC method with an empirical example that investigates the effect of EDR laws on voter turnout in the United States. Voting in the United States usually takes two steps. Except in North Dakota, where no registration is needed, eligible voters throughout the country must register prior to casting their ballots. Registration, which often requires a separate trip from voting, is widely regarded as a substantial cost of voting and a culprit of low turnout rates before the 1993 National Voter Registration Act (NVRA) was enacted (e.g., Highton 2004). Against this backdrop, EDR is a reform that allows eligible voters to register on Election Day when they arrive at polling stations. In the mid-1970s, Maine, Minnesota, and Wisconsin were the first adopters of this reform in the hopes of increasing voter turnout; while Idaho, New Hampshire, and Wyoming established EDR in the 1990s as a strategy to opt out the NVRA (Hanmer 2009). Before

the 2012 presidential election, three other states, Montana, Iowa, and Connecticut, passed laws to enact EDR, adding the number of states having EDR laws to nine.[20]

Most existing studies based on individual-level cross-sectional data, such as the Current Population Surveys and the National Election Surveys, suggest that EDR laws increase turnout (the estimated effect varies from 5 to 14 percentage points).[21] These studies do not provide compelling evidence of a causal effect of EDR laws because the research designs they use are insufficient to address the problem that states self-select their systems of registration laws. "Registration requirements did not descend from the skies," as Dean Burnham puts it (1980, p. 69). A few studies employ time-series or TSCS analysis to address the identification problem.[22] However, Keele and Minozzi (2013) cast doubts on these studies and suggest that the "parallel trends" assumption may not hold, as we will also demonstrate below.

In the following analysis, we use state-level voter turnout data for presidential elections from 1920 to 2012.[23] The turnout rates are calculated with total ballots counted in a presidential election in a state as the numerator and the state's voting-age population (VAP) as the denominator.[24] Alaska and Hawaii are not included in the sample since they were not states until 1959. North Dakota is also dropped since no registration is required. As mentioned above, up to the 2012 presidential election, nine states had adopted EDR laws (hereafter referred to as *treated*) and the rest thirty-eight states had not (referred to as *controls*). The raw turnout data for all forty-seven states are shown in the Online Appendix.[25]

First, we use a standard two-way fixed effects mode, which is often referred to as a DID model in the literature. The results are shown in Table 2 columns (1) and (2). Standard errors are produced by nonparametric bootstraps (blocked at the state level) of 2,000 times. In column (1), only the EDR indicator is included, while in column (2), we additionally control for indicators of universal mail-in registration and motor voter registration. The estimated coefficients of EDR laws are 0.87 and 0.78 percent using the two specifications, respectively, with standard errors around 3 percent.

The two-way fixed effects model presented in Table 2 assumes a constant treatment effect both across states and over time. Next we relax this assumption by literally employing a DID approach. In other words, we estimate the effect of EDR laws on voter turnout in the posttreatment period by subtracting the time intercepts estimated from the control group and the unit intercepts based on the pretreatment data. The predict turnout for state i in year t, therefore, is the summation of unit intercept i and time intercept t, plus the impact of the time-varying covariates. The result is visualized in the upper panel of Figure 2. Figure 2a shows the average actual turnout (solid line) and average predicted turnout in the absence of EDR laws (dashed line); both averages are taken based on the number of terms since (or before) EDR laws first took effect. Figure 2b shows the gap between the two lines, or the estimated ATT. The confidence intervals are produced by block bootstraps of 2,000 times. It is clear from both figures that the "parallel trends" assumption is not likely to hold since the average predicted turnout deviates from the average actual turnout in the pretreatment periods.

20 In the Online Appendix, we list the years during which EDR laws were enacted and first took effect in presidential elections.

21 See Wolfinger and Rosenstone (1980), Mitchell and Wlezien (1995), Rhine (1992), Highton (1997), Timpone (1998), Timpone (2002), Huang and Shields (2000), Alvarez, Ansolabehere, and Wilson (2002), Brians and Grofman (2001), Hanmer (2009), Burden *et al.* (2009), Cain, Donovan, and Tolbert (2011), Teixeira (2011) for examples. The results are especially consistent for the three early adopters, Maine, Minnesota, and Wisconsin.

22 See, for example, Fenster (1994), King and Wambeam (1995), Knack and White (2000), Knack (2001), Neiheisel and Burden (2012), Springer (2014).

23 The data from 1920 to 2000 are from Springer (2014). The data from 2004 to 2012 are from The United States Election Project, http://www.electproject.org/. Indicators of other registration laws, including universal mail-in registration and motor voter registration, also come from Springer (2014), with a few supplements. Replication files can be found in Xu (2016).

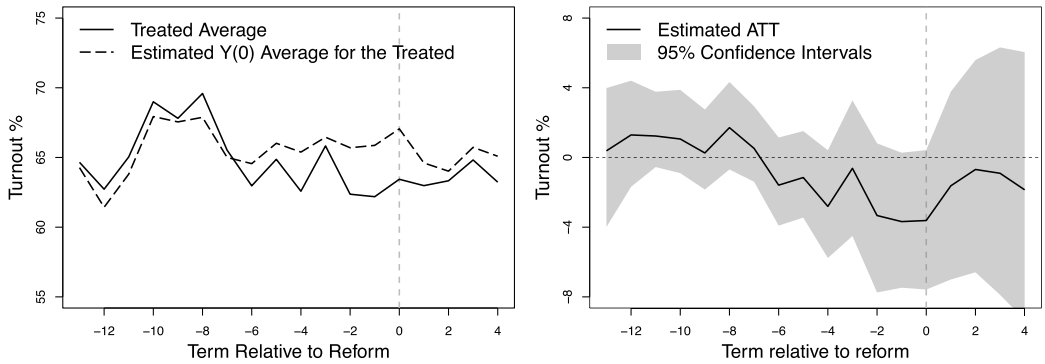
24 We do not use the voting-eligible population (VEP) as the denominator because they are not available in early years.

25 As is shown in the figure and has been pointed out by many, turnout rates are in general higher in states that have EDR laws than states that have not, but this does not necessarily imply a causal relationship between EDR laws and voter turnout.

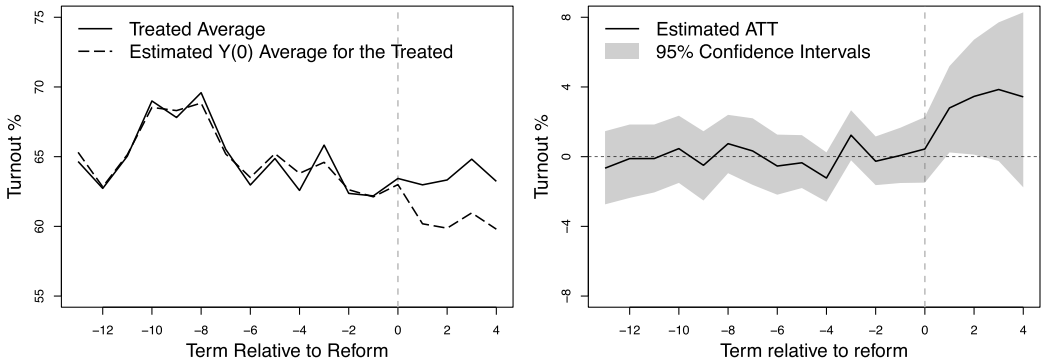
Table 2. The effect of EDR on voter turnout

	Voter turnout %			
	FE		*GSC*	
Outcome variable	***(1)***	***(2)***	***(3)***	***(4)***
Election Day Registration	0.87	0.78	5.13	4.90
	(3.01)	(3.31)	(2.27)	(2.27)
Universal mail-in registration		−0.94		0.15
		(1.80)		(0.80)
Motor voter registration		−0.21		−1.05
		(1.45)		(0.79)
State fixed effects	x	x	x	x
Year fixed effects	x	x	x	x
Unobserved factors	N/A	N/A	2	2
Observations	1,128	1,128	1,128	1,128
Treated states	9	9	9	9
Control states	38	38	38	38

Note: Standard errors in columns (1) and (2) are based on nonparametric bootstraps (blocked at the state level) of 2,000 times. Standard errors in columns (3) and (4) are based on parametric bootstraps (blocked at the state level) of 2,000 times.

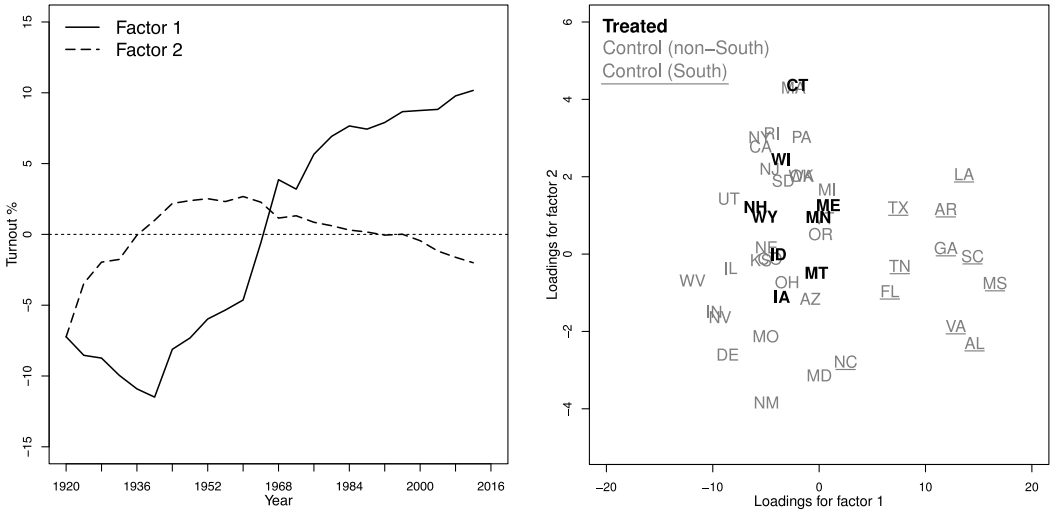


(a) Difference-in-differences



(b) Generalized synthetic control

Figure 2. The effect of EDR on turnout: Main results.



(a) Factors (b) Loadings

Figure 3. The Effect of EDR on turnout: Factors and loadings.

Next, we apply the GSC method to the same dataset. Table 2 columns (3) and (4) summarize the result.[26] Again, both specifications impose additive state and year fixed effects. In column (3), no covariates are included, while in column (4), mail-in and motor voter registration are controlled for (assuming that they have constant effects on turnout). With both specifications, the cross-validation scheme finds two unobserved factors to be important and after conditioning on both the factors and additive fixed effects, the estimated ATT based on the GSC method is around 5 percent with a standard error of 2.3 percent.[27] This means that EDR laws are associated with a statistically significant increase in voter turnout, consistent with previous OLS results based on individual-level data. The lower panel of Figure 2 shows the dynamics of the estimated ATT. Again, in the left figure, averages are taken after the actual and predicted turnout rates are realigned to the timing of the reform. With the GSC method, the average actual turnout and average predicted turnout match well in pretreatment periods and diverge after EDR laws took effect. The right figure shows that the gaps between the two lines are virtually flat in pretreatment periods and the effect takes off right after the adoption of EDR.[28]

Figure 3 presents the estimated factors and factor loadings produced by the GSC method.[29] Figure 3a depicts the two estimated factors. The x-axis is year and the y-axis is the magnitude of factors (rescaled by the square root of their corresponding eigenvalues to demonstrate their relative importance). Figure 3b shows the estimated factors loadings for each treated (black, bold) and control (gray) units, with x- and y-axes indicating the magnitude of the loadings for the first and second factors, respectively. Bearing in mind the caveat that estimated factors may not be directly interpretable because they are, at best, linear transformations of the true factors, we find that the estimated factors shown in this figure are meaningful. The first factor captures the sharp increase in turnout in the southern states because of the 1965 Voting Rights Act that removed Jim Crow laws, such as poll taxes or literacy tests, that suppressed turnout. As shown in the

26 Note that although the estimated ATT of EDR on voter turnout is presented in the same row as the coefficient of EDR using the FE model, the GSC method does not assume the treatment effect to be constant. In fact, it allows the treatment effect to be different both across states and over time. Predicted counterfactuals and individual treatment effect for each of the nine treated states are shown in the Online Appendix.

27 The results are similar if additive state and year fixed effects are not directly imposed, though not surprisingly, the algorithm includes an additional factor.

28 Although it is not guaranteed, this is not surprising since the GSC method uses information of all past outcomes and minimizes gaps between actual and predicted turnout rates in pretreatment periods.

29 The results are essentially the same with or without controlling for the other two registration reforms.

Table 3. The effect of EDR on voter turnout: Three waves

	Voter turnout %		
	1st wave	*2nd wave*	*3rd wave*
Outcome variable	*(1)*	*(2)*	*(3)*
Election Day Registration	7.27	2.17	−1.14
	(3.33)	(2.82)	(3.00)
Mail-in and motor voter registration	x	x	x
State fixed effects	x	x	x
Year fixed effects	x	x	x
Unobserved factors	2	2	2
Observations	1,128	1,128	1,128
Treated states	3	3	3
	(ME, MN, WI)	(ID, NH, WY)	(MT, IA, CT)
Control states	38	38	38

Note: Standard errors are based on parametric bootstraps (blocked at the state level) of 1,000 times.

right figure, the top eleven states that have the largest loadings on the first factor are exactly the eleven southern states (which were previously in the confederacy).[30] The labels of these states are underlined in Figure 3b. The second factor, which is set to be orthogonal to the first one, is less interpretable. However, its nonnegligible magnitude indicates a strong downward trend in voter turnout in many states in recent years. Another reassuring finding shown by Figure 3b is that the estimated factor loadings of the nine treated units mostly lie in the convex hull of those of the control units, which indicates that the treated counterfactuals are produced mostly by more reliable interpolations instead of extrapolations.

Finally, we investigate the heterogeneous treatment effects of EDR laws. Previous studies have suggested that the motivations behind enacting these laws are vastly different between the early adopters and later ones. For example, Maine, Minnesota, and Wisconsin, which established the EDR in mid-1970s, did so because officials in these states sincerely wanted the turnout rates to be higher, while the "reluctant adopters," including Idaho, New Hampshire, and Wyoming, introduced the EDR as a means to avoid the NVRA because officials viewed the NVRA as "a more costly and potentially chaotic system" (Hanmer 2009). Because of the different motivations and other reasons, we may expect the treatment effect of EDR laws to be different in states that adopted them in different times.

The estimation of heterogeneous treatment effects is embedded in the GSC method since it gives individual treatment effects for all treated units in a single run. Table 3 summarizes the ATTs of EDR on voter turnout among the three waves of EDR adopters. Again, additive state and year fixed effects, as well as indicators of two other registration systems, are controlled for. Table 3 shows that EDR laws have a large and positive effect on the early adopters (the estimate is about 7 percent with a standard error of 3 percent) while EDR laws were found to have no statistically significant impact on the other six states.[31] Such differential outcomes can be due to two reasons. First, the NVRA of 1993 substantially reduced the cost of registration: since almost everyone who

30 Although we can control for indicators of Jim Crow laws in the model, such indicators may not be able to capture the heterogeneous impacts of these laws on voter turnout in each state.

31 In the Online Appendix, we show that the treatment effects are positive (and relatively large) for all three early adopting states, Maine, Minnesota, and Wisconsin. Using a fuzzy regression discontinuity design, Keele and Minozzi (2013) show that EDR has almost no effect on the turnout in Wisconsin. The discrepancy with this paper could be mainly due to the difference in the estimands. Two biggest cities in Wisconsin, Milwaukee and Madison constitute a major part of Wisconsin's constituency but have neglectable influence to their local estimates. One advantage of Keele and Minozzi (2013)'s approach over ours is the use of fine-grained municipal level data.

has some intention to vote is a registrant after the NVRA was enacted, "there is now little room for enhancing turnout further by making registration easier" (Highton 2004). Second, because states having a strong "participatory culture" is more likely to be selected into an EDR system in earlier years, costly registration, as a binding constraint in these states, may not be a first-order issue in a state where many eligible voters have low incentives to vote in the first place. It is also possible that voters in early adopting states formed a habit to vote in the days when the demand for participation was high (Hanmer 2009).

In short, using the GSC method, we find that EDR laws increased turnout in early adopting states, including Maine, Minnesota, and Wisconsin, but not in states that introduced EDR as a strategy to opt out the NVRA or enacted EDR laws in recent years. These results are broadly consistent with evidence provided by a large literature based on individual-level cross-sectional data (see, for example, Leighley and Nagler 2013 for a summary). They are also more credible than results from conventional fixed effects models when the "parallel trends" assumption appears to fail.[32]

6 Conclusion

In this paper, we propose the GSC method for causal inference with TSCS data. It attempts to address the challenge that the "parallel trends" assumption often fails when researchers apply fixed effects models to estimate the causal effect of a certain treatment. The GSC method estimates the individual treatment effect on each treated unit semiparametrically. Specifically, it imputes treated counterfactuals based on a linear interactive fixed effects model that incorporates time-varying coefficients (factors) interacted with unit-specific intercepts (factor loadings). A built-in cross-validation scheme automatically selects the model, reducing the risks of overfitting.

This method is in spirit of the original synthetic control method in that it uses data from pretreatment periods as benchmarks to customize a reweighting scheme of control units in order to make the best possible predictions for treated counterfactuals. It generalizes the synthetic control method in two aspects. First, it allows multiple treated units and differential treatment timing. Second, it offers uncertainty estimates, such as standard errors and confidence intervals, that are easy to interpret.

Monte Carlo exercises suggest that the proposed method performs well even with relatively small T_0 and N_{co} and show that it has advantages over several existing methods: (1) it has less bias than the two-way fixed effects or DID estimators in the presence of decomposable time-varying confounders, (2) it corrects bias of the IFE estimator when the treatment effect is heterogeneous across units; and (3) it is more efficient than the synthetic control method. To illustrate the applicability of this method in political science, we estimate the effect of EDR laws on voter turnout in the United States. We show that EDR laws increased turnout in early adopting states but not in states that introduced them more recently.

Two caveats are worth emphasizing. First, insufficient data (with either a small T_0 or a small N_{co}) cause bias in the estimated treatment effect. In general, users should be cautious when $T_0 < 10$ or $N_{co} < 40$. Second, excessive extrapolations based on imprecisely estimated factors and factor loading can lead to erroneous results. To avoid this problem, we recommend the following diagnostics upon using this method: (1) plot raw data of treated and control outcomes as well as imputed counterfactuals and check whether the imputed values are within reasonable intervals; (2) plot estimated factor loadings of both treated and control units and check the overlap (as in Fig. 3). We provide software routines `gsynth` in `R` to implement the estimation procedure as well as these diagnostic tests. When excessive extrapolations appear to happen,

32 Glynn and Quinn (2011) argue that traditional cross-sectional methods in general overestimate the effect of EDR laws on voter turnout and suggest that EDR laws are likely to have minimum effect on turnout in non-EDR states (the ATC). In this paper, we focus on the effect of EDR in EDR states (the ATT) instead.

we recommend users to include a smaller number of factors or switch back to the conventional DID framework. We also recommend users to benchmark the results with estimates from the IFE model (Bai 2009) as well as Bayesian multi-level factor models (e.g., Pang 2014) whenever it is possible.

Another limitation of the proposed method is that it cannot accommodate complex DGPs that often appear in TSCS data (when T is much bigger than panel data), such as (1) dynamic relationships between the treatment, covariates, and outcome (e.g., Pang 2010, 2014, Blackwell and Glynn 2015), (2) structural breaks (e.g., Park 2010, 2012), and (3) multiple times of treatment and variable treatment intensity. Nor does it allow random coefficients for the observed time-varying covariates, as such modeling setups become increasing popular with Bayesian multi-level analysis. Future research is needed to accommodate these scenarios.

Supplementary material

For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2016.2.

References

Abadie, Alberto. 2005. Semiparametric difference-in-differences estimators. *The Review of Economic Studies* 72(1):1–19.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association* 105(490):493–505.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2015. Comparative politics and the synthetic control method. *American Journal of Political Science* 59(2):495–510.

Acemoglu, Daron, Simon Johnson, Amir Kermani, James Kwak, and Todd Mitton. 2016. The value of connections in turbulent times: Evidence from the United States. *Journal of Financial Economics* 121(2):368–391.

Alvarez, R. Michael, Stephen Ansolabehere, and Catherine H. Wilson. 2002. Election day voter registration in the United States: How one-step voting can change the composition of the American electorate. Working Paper, Caltech/MIT Voting Technology Project.

Angrist, Joshua D., Scar Jord, and Guido Kuersteiner. 2013. Semiparametric estimates of monetary policy effects: String theory revisited. NBER Working Paper No. 19355.

Bai, Jushan. 2003. Theory for factor models of large dimensions. *Econometrica* 71(1):135–137.

Bai, Jushan. 2009. Panel data models with interactive fixed effects. *Econometrica* 77:1229–1279.

Beck, Nathaniel, and Jonathan N. Katz. 1995. What to do (and not to do) with time-series cross-section data. *American Political Science Review* 89(3):634–647.

Blackwell, Matthew, and Adam Glynn. 2015. How to make causal inferences with time-series cross-sectional data. Harvard University. Mimeo.

Brians, Craig Leonard, and Bernard Grofman. 2001. Election day registration's effect on US voter turnout. *Social Science Quarterly* 82(1):170–183.

Brodersen, Kay H., Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L. Scott. 2014. Inferring causal impact using Bayesian structural time-series models. *Annals of Applied Statistics* 9(1):247–274.

Burden, Barry C., David T. Canon, Kenneth R. Mayer, and Donald P. Moynihan. 2009. The effects and costs of early voting, election day registration, and same day registration in the 2008 elections. University of Wisconsin-Madison. Mimeo.

Burnham, Walter Dean. 1980. The appearance and disappearance of the American voter. In *Electoral participation: A comparative analysis*, ed. Richard Rose. Beverly Hills, CA: Sage Publications.

Cain, Bruce E., Todd Donovan, and Caroline J. Tolbert. 2011. *Democracy in the states: Experiments in election reform*. Washington, DC: Brookings Institution Press.

Campbell, John Y., Andrew W. Lo, and A. Craig MacKinlay. 1997. *The econometrics of financial markets*. Princeton, NJ: Princeton University Press.

Dube, Arindrajit, and Ben Zipperer. 2015. Pooling multiple case studies using synthetic controls: An application to minimum wage policies. IZA Discussion Paper No. 8944.

Efron, Brad. 2012. The estimation of prediction error. *Journal of the American Statistical Association* 99(467):619–632.

Efron, Brad, and Rob Tibshirani. 1993. *An introduction to the bootstrap*. New York: Chapman & Hall.

Fenster, Mark J. 1994. The impact of allowing day of registration voting on turnout in US elections from 1960 to 1992 A research note. *American Politics Research* 22(1):74–87.
Gaibulloev, Khusrav, Todd Sandler, and Donggyu Sul. 2014. Dynamic panel analysis under cross-sectional dependence. *Political Analysis* 22(2):258–273.
Glynn, Adam N. Glynn, and Kevin M. Quinn. 2011. Why process matters for causal inference. *Political Analysis* 19:273–286.
Gobillon, Laurent, and Thierry Magnac. 2016. Regional policy evaluation: Interactive fixed effects and synthetic controls. *The Review of Economics and Statistics* 98(3):535–551.
Hanmer, Michael J. 2009. *Discount voting: Voter registration reforms and their effects*. New York: Cambridge University Press.
Highton, Benjamin. 1997. Easy registration and voter turnout. *The Journal of Politics* 59(2):565–575.
Highton, Benjamin. 2004. Voter registration and turnout in the United States. *Perspectives on Politics* 2(3):507–515.
Holland, Paul W. 1986. Statistics and causal inference. *Journal of American Statistical Association* 81(8):945–960.
Hsiao, Cheng, Steve H. Ching, and Shui Ki Wan. 2012. A panel data approach for program evaluation: Measuring the benefits of political and economic integration of Hong Kong with Mainland China. *Journal of Applied Econometrics* 27(5):705–740.
Huang, Chi, and Todd G. Shields. 2000. Interpretation of interaction effects in logit and probit analyses reconsidering the relationship between registration laws, education, and voter turnout. *American Politics Research* 28(1):80–95.
Imai, Kosuke, and In Song Kim. 2016. When should we use linear fixed effects regression models for causal inference with panel data. Princeton University. Mimeo.
Keele, L., and W. Minozzi. 2013. How much is Minnesota like Wisconsin? Assumptions and counterfactuals in causal inference with observational data. *Political Analysis* 21(2):193–216.
Kim, Dukpa, and Tatsushi Oka. 2014. Divorce law reforms and divorce rates in the USA: An interactive fixed-effects approach. *Journal of Applied Econometrics* 29(2):231–245.
King, James D., and Rodney A. Wambeam. 1995. Impact of election day registration on voter turnout: A quasi-experimental analysis. *Policy Studies Review* 14(3):263–278.
Knack, Stephen. 2001. Election-day registration The second wave. *American Politics Research* 29(1):65–78.
Knack, Stephen, and James White. 2000. Election-day registration and turnout inequality. *Political Behavior* 22(1):29–44.
Leighley, Jan E., and Jonathan Nagler. 2013. *Who votes now? Demographics, issues, inequality, and turnout in the United States*. Princeton, NJ: Princeton University Press.
Mitchell, Glenn E., and Christopher Wlezien. 1995. The impact of legal constraints on voter registration, turnout, and the composition of the American electorate. *Political Behavior* 17(2):179–202.
Moon, Hyungsik Roger, and Martin Weidner. 2015. Dynamic linear panel regression models with interactive fixed effects. *Econometric Theory* 33(1):158–195.
Mora, Ricardo, and Ilina Reggio. 2012. Treatment effect identification using alternative parallel assumptions. Universidad Carlos III de Madrid. Mimeo.
Neiheisel, J. R., and B. C. Burden. 2012. The impact of election day registration on voter turnout and election outcomes. *American Politics Research* 40(4):636–664.
Neyman, Jerzy. 1923. On the application of probability theory to agricultural experiments: Essay on principles. *Statistical Science* 5:465–480, Section 9 (translated in 1990).
Pang, Xun. 2010. Modeling heterogeneity and serial correlation in binary time-series cross-sectional data: A Bayesian multilevel model with AR(p) errors. *Political Analysis* 18(4):470–498.
Pang, Xun. 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.
Park, Jong Hee. 2010. Structural change in US presidents' use of force. *American Journal of Political Science* 54(3):766–782.
Park, Jong Hee. 2012. A unified method for dynamic and cross-sectional heterogeneity: Introducing hidden Markov panel models. *American Journal of Political Science* 56(4):1040–1054.
Rhine, Staci L. 1992. An analysis of the impact of registration factors on turnout in 1992. *Political Behavior* 18(2):171–185.
Rubin, Donald B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 5(66):688–701.
Springer, Melanie Jean. 2014. *How the states shaped the nation: American electoral institutions and voter turnout, 1920–2000*. Chicago, IL: University of Chicago Press.
Stewart, Brandon. 2014. Latent factor regressions for the social sciences. Princeton University. Mimeo.

Teixeira, Ruy A. 2011. *The disappearing American voter*. Washington, DC: Brookings Institution Press.
Timpone, Richard J. 1998. Structure, behavior, and voter turnout in the United States. *The American Political Science Review* 92(1):145.
Timpone, Richard J. 2002. Estimating aggregate policy reform effects: New baselines for registration, participation, and representation. *Political Analysis* 10(2):154–177.
Wolfinger, Raymond E., and Steven J. Rosenstone. 1980. *Who votes?* New Haven, CT: Yale University Press.
Xu, Yiqing. 2016. Replication data for: Generalized synthetic control method: Causal inference with interactive fixed effects models. doi:10.7910/DVN/8AKACJ, Harvard Dataverse.