Synthetic control methods

Yi Chen

December 2023

1 Standard synthetic control method

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.

Idea of synthetic control approach: use a combination of units as the control group.
Model:
Suppose we observe $J+1$ regions.
The first region is exposed to the intervention, and the remaining J regions are potential controls.
Time: $t = 1, ..., T$
T_0: the number of pre-intervention periods ($1 \leq T_0 < T$)
Y_{it}^N: the outcome that would be observed for region i at time t in the absence of the intervention.
Y_{it}^I: the outcome that would be observed for unit i at time t if unit i is exposed to the intervention in periods $T_0 + 1$ to T.
$\alpha_{it} = Y_{it}^I - Y_{it}^N$: the effect of the intervention for unit i at time t.
$D_{it} = \mathbb{1}[\text{unit } i \text{ is exposed to the intervention at time } t]$
Observed outcome: $Y_{it} = Y_{it}^N + \alpha_{it} D_{it}$ Since only the first region (region "one") is exposed to the intervention and only after period T_0, we have that

$$D_{it} = \begin{cases} 1 & \text{if } i = 1 \text{ and } t > T_0 \\ 0 & \text{otherwise} \end{cases}$$

We aim to estimate $(\alpha_{1T_0+1}, ..., \alpha_{1T})$, and for $t > T_0$,

$$\alpha_{1t} = Y_{1t}^I - Y_{1t}^N = Y_{1t} - Y_{1t}^N$$

Since Y_{1t} is observed. we just need to estimate Y_{1t}^N.
Suppose Y_{1t}^N is given by a factor model

$$Y_{it}^N = \delta_t + \boldsymbol{\theta}_t \mathbf{Z}_i + \boldsymbol{\lambda}_t \boldsymbol{\mu}_i + \epsilon_{it}$$

where

- δ_t: common time variation
- $\mathbf{Z}_i$: time-invariant observed covariates
- $\boldsymbol{\mu}_i$: unobserved factor loadings

Compared to difference-in-differences (fixed effects) model, model allows the effects of confounding unobserved characteristics to vary with time. $\beta = (\beta_2, ..., \beta_{J+1})'$: represents a potential synthetic control, i.e., a particular weighted average of control regions.
The value of the outcome variable for each synthetic control indexed by β is: $\sum_{j=2}^{J+1} = \delta_t + \boldsymbol{\theta}_t \sum_{j=2}^{J+1} \beta_j \mathbf{Z}_j + \boldsymbol{\lambda}_t \sum_{j=2}^{J+1} \beta_j \boldsymbol{\mu}_j + \sum_{j=2}^{J+1} \beta_j \epsilon_{jt}$ Suppose there are $(\beta_1^*, \beta_2^*, ..., \beta_{J+1}^*)$ such that

$$\sum_{j=2}^{J+1} \beta_j^* Y_{j1} = Y_{11}, ..., \sum_{j=2}^{J+1} \beta_j^* Y_{jT_0} = Y_{1T_0} \text{ and } \sum_{j=2}^{J+1} \beta_j^* \mathbf{Z}_j = \mathbf{Z}_1 \tag{1}$$

In practice, such $\beta*_i$'s usually do not exist. Instead, we find the β that is closest to the treated unit in terms of the outcomes in the pretreatment periods.

$$\hat{\beta} = \arg\min_{\beta \in \Lambda} \sum_{t=1}^{T_0} \left(Y_{1t} - \beta_1 - \sum_{j=2}^{J+1} \beta_j Y_{jt} \right)^2 \tag{2}$$

where w_1 is an intercept, and

$$\Lambda = \{\beta \in \mathbb{R}^J : \beta_1 = 0, \beta_j \geq 0 \text{ for } j = 2, ..., J+1 \text{ and } \sum_{j=2}^{J+1} \beta_j = 1\} \tag{3}$$

Thus we can use

$$\hat{\alpha}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} \hat{\beta}_j Y_{jt} \tag{4}$$

as an estimator of α_{1t} for $t \in \{T_0 + 1, ..., T\}$

Hsiao, Ching and Wan (2012) proposed a panel data method that is based on the equation (2) without restrictions (3).

Limitations:

1. restrictive constraints: non-negativity of weights, summing-to-one of weights, and no intercept
2. no inference theory
3. lack of an explicit mechanism for the "large p, small n" problem (a situation where the number of covariates exceeds the number of data points)

2 Regularized Regression methods

Doudchenko, Nick and Guido W. Imbens (2016), "Balancing, Regres- sion, Difference-In-Differences and Synthetic Control Methods: A Synthesis," Working Paper 22791, National Bureau of Economic Research, http://www.nber.org/papers/w22791.

Doudchenko and Imbens (2016) argue that none of the constraints on the parameters in Equation (3) is likely to hold in practice.

- The no-intercept constraint ($\beta_1 = 0$) rules out the possibility of a systematic difference between the treated unit and a synthetic control unit.
- The sum-to-one restriction ($\sum_{j=2}^{J+1} \beta_j = 1$) can help find a unique set of weights but is implausible if a treated unit is on the extreme end of the distribution of units. (i.e., the treated unit is the largest or smallest in terms of outcome values)
- The nonnegative constraint ($\beta_j \geq 0$) implicitly assumes positive correlation between outcomes of the treated unit and control units.

Thus they propose the use of the (frequentist) elastic net method that does not incorporate the restrictions in Equation (3):

$$\hat{\beta} = \arg\min_{\beta} \{ \sum_{t=1}^{T_0} (Y_{1t} - \beta_1 - \sum_{j=2}^{J+1} \beta_j Y_{jt})^2 + \lambda_1 \sum_{j=2}^{J+1} |\beta_j| + \lambda_2 \sum_{j=2}^{J+1} \beta_j^2 \} \tag{5}$$

where $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are penalty parameters. It combines the L_1 penalty of the lasso and the L_2 penalty of ridge.

3 Generalized Synthetic Control Method

Xu, Y. (2017). Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models. Political Analysis, 25(1), 57-76. doi:10.1017/pan.2016.2

DiD needs the "parallel trends" assumption, which is not directly testable. 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. However, parallel pretreatment trends are not supported by data in many cases.
One way to deal with this problem is the synthetic control method (Abadie et al., 2010) That is to match both pretreatment covariates and outcomes between a treated unit and a set of control units and uses pretreatment periods as criteria for good matches. Specifically, it constructs a "synthetic control unit" as the counterfactual for the treated unit by reweighting the control units. However, it only applies to the case of one treated unit and the uncertainty estimates it offers are not easily interpretable.
Another 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. However, such approach relies on a set of alternative identification assumptions that treatment assignment is ignorable conditional on both fixed effects and the imposed trends.
An alternative way is to model unobserved time-varying confounders semiparametrically. For example, use an interactive fixed effects (IFE) model which incorporates unit-specific intercepts interacted with time-varying coefficients (Bai, 2009).

- time-varying coefficients: (latent) factors
- unit-specific intercepts: factor loadings
- 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.

Therefore, 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.

1. Step 1: Estimate an IFE model using only the control group data, and obtain a fixed number of latent factors.
2. Step 2: Estimate factor loadings for each treated unit by linearly projecting pretreatment treated outcomes onto the space spanned by these factors.
3. Step 3: Impute treated counterfactuals based on the estimated factors and factor loadings.

3.1 Framework

3.1.1 Notation

- Y_{it}: outcome of unit i at time t
- N: total number of units; N_{tr}: number of treated units; N_{co}: number of control units; $N = N_{tr} + N_{co}$
- $\mathcal{T}$ is the set of units in treatment group. And $\mathcal{C}$ is the set of units in control group.
- T: time periods (from time 1 to time T)
- $T_{0,i}$: the number of pretreatment periods for unit i; thus the treated period is from $T_{0,i}+1$ to T. Total number of treated periods is $q_i = T - T_{0,i}$
  For simplicity for notation, we set all treated units to be exposed at the same time, i.e., $T_{0,i} = T_0$ and $q_i = q$, but this can be accommodated.

3.1.2 Assumptions

Assumption 3.1 *Functional form:* $Y_{it} = \delta_{it} D_{it} + x_{it}'\beta + \lambda_i' f_t + \epsilon_{it}$

- D_{it}: treatment indicator
- δ_{it}: heterogeneous treatment effect on unit i at time t
- x_{it}: $(k \times 1)$ vector of observed covariates
- β: $(k \times 1)$ vector of unknown parameters
- f_t: $(r \times 1)$ vector of unobserved common factors
- λ_i: $(r \times 1)$ vector of unknown factor loadings
- ϵ_{it}: unobserved idiosyncratic shocks for unit i at time t and has zero mean

The average treatment effect on the treated (ATT) at time t (when $t > T_0$) is:

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

Pooling all the time periods together, we can rewrite the DGP of each unit as:

$$Y_i = D_i \cdot \delta_i + X_i\beta + F\lambda_i + \epsilon_i, i \in 1, ..., N.$$

where $Y_i = [Y_{i1}, ..., Y_{iT}]'; D_i = [D_{i1}, ..., D_{iT}]'; \delta_i = [\delta_{i1}, ..., \delta_{iT}]'; \epsilon_i = epsilon_{i1}, ..., epsilon_{iT}]'$ are $(T \times 1)$ vectors; and $X_i = [x_{i1}, ..., x_{iT}]'$ is a $(T \times k)$ matrix; and $F = [f_1, ..., f_T]'$ is a $(T \times r)$ matrix. Specifically, the DGP of a control unit can be expressed as $Y_i = X_i\beta + F\lambda_i + \epsilon, i \in 1, ..., N_{co}$. Stacking all control units together, we have:

$$Y_{co} = X_{co}\beta + F\Lambda'_{co} + \epsilon_{co}$$

where $Y_{co} = [Y_1, ..., Y_{N_{co}}]$ and $\epsilon_{co} = [\epsilon_1, ..., \epsilon_{N_{co}}]$ are $(T \times N_{co}$ matrices; X_{co} is a three-dimentional $(T \times N_{co} \times p$ matrix; and $\Lambda_{co} = [\lambda_1, ..., \lambda_{N_{co}}]$ is a $(N_{co} \times r)$ matrix.
Other assumptions required for causal identification:

Assumption 3.2 *Strict exogeneity:*

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

Assumption 3.3 *Weak serial dependence of the error terms*

Assumption 3.4 *Regularity conditions*

Assumptions 3 and 4 are needed for the consistent estimation of β and the space spanned by F (or $F'F/T$)

Assumption 3.5 *The error terms are cross-sectionally independent and homoscedastic.*

Assumption 5 is needed for the valid inference based on a block bootstrap procedure. Heteroscedasticity across time is allowed.

3.2 Estimation Strategy

1. Step 1: estimate an IFE model using only the control group data and obtain $\hat{\beta}, \hat{F}, \hat{\Lambda_{co}}$:

$$(\hat{\beta}, \hat{F}, \hat{\Lambda}_{co}) = \arg \min_{\tilde{\beta}, \tilde{F}, \tilde{\Lambda}_{co}} \sum_{i \in \mathcal{C}} (Y_i - X_i\tilde{\beta} - \tilde{F}\tilde{\lambda}_i)'(Y_i - X_i\tilde{\beta} - \tilde{F}\tilde{\lambda}_i)$$

$$\text{s.t. } \tilde{F}'\tilde{F}/T = I_r \text{ and } \tilde{\Lambda}'_{co}\tilde{\Lambda}_{co}\text{=diagonal}$$

2. Step 2: Estimate factor loadings for each treated unit by minimizing the mean squared error of the predicated treated outcome in pretreatment periods:

$$\hat{\lambda}_i = \arg\min_{\tilde{\lambda}_i} (Y_i^0 - X_i^0\hat{\beta} - \hat{F}^0\tilde{\lambda}_i)'(Y_i^0 - X_i^0\hat{\beta} - \hat{F}^0\tilde{\lambda}_i) = (\hat{F}^{0\prime}\hat{F}^0)^{-1}\hat{F}^{0\prime}(Y_i^0 - X_i^0\hat{\beta}), i \in \mathcal{T}$$

   where "0" denotes the pretreatment periods.

3. Step 3: calculate treated counterfactuals based on $\hat{\beta}, \hat{F}, \hat{\lambda}_i$:

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

3.2.1 Model selection

In the previous sections, we assume the number of unobserved factors, r, is known. However, in practice, researchers may have limited knowledge of the exact number of factors. A cross-validation procedure to select models before estimating causal effect:

1. Start with a given number of factors r, estimate an IFE model using the control group data $\{Y_i, X_i\}_{i\in\mathcal{C}}$, obtaining $\hat{\beta}$ and $\hat{F}$;
2. Start a cross-validation loop that goes through all T_0 pretreatment periods:
   (a) In round $s \in \{1, ..., T_0\}$, 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:
   $$\hat{\lambda}_{i,-s} = (F^{0\prime}{}_{-s}F^0_{-s})^{-1}F^{0\prime}{}_{-s}(Y^0_{i,-s} - X^{0\prime}{}_{i,-s}\hat{\beta}), \forall u \in \mathcal{T},$$
   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 $i \in \mathcal{T}$.
3. Calculate the mean square prediction error (MSPE) given r,
   $$MSPE(r) = \sum_{s=1}^{T_0}\sum_{i\in\mathcal{T}} e^2_{is}/T_0$$
4. Repeat the above 3 steops with different r's and obtain corresponding MSPEs.
5. Choose r^* that minimizes the MSPE.

4 An Improved and Extended Bayesian Synthetic Control

Sean Pinkney. 2021. An Improved and Extended Bayesian Synthetic Control. In arXiv (arXiv.org). USA, 8 pages.

Drawback of Xu's approach (2017): the latent factors are fit only on control group data (for all time periods) and then the loadings are fit in a separate procedure, in only pre-treatment periods, potentially decreasing the efficiency of the estimates. The model proposed in this paper uses more of the data for estimation while simultaneously estimating the latent factors.

4.1 Framework

The outcome $y_j, j \in J$ follows:

$$y_j \sim \mathcal{N}(F\beta_j + X_j\gamma + \Delta + \kappa_j, \sigma)$$

- F is a $T \times L$ matrix;
- β_j is a $L \times 1$ vector;

- X is a $T \times k$ matrix of covariance with coefficients γ
- Δ is a $T \times 1$ vector that is the same across all J series; and κ is a individual fixed effect that does not vary across T.

To simplify the computation, assume that the loading matrix F is uncorrelated.
The priors for $F, \gamma, \Delta, kappa, \sigma$ are

$$F \sim \mathcal{N}(0,1)$$
$$\gamma \sim \mathcal{N}(0,1)$$
$$\Delta \sim \mathcal{N}(0,2)$$
$$\kappa \sim \mathcal{N}(0,1)$$
$$\sigma \sim \mathcal{N}(0,1)$$
$$(\beta_{l,j} \mid \lambda_l, \eta_j, \tau) \sim \mathcal{N}(0, \lambda_l)$$
$$(\lambda_l \mid \eta_j, \tau) \sim \mathcal{C}^+(0, \tau\eta_j)$$
$$\eta_j \sim \mathcal{C}^+(0,1)$$
$$\tau \sim \mathcal{C}^+(0,1)$$

5 A Bayesian Alternative Synthetic Control for Comparative Case Studies

Pang, X., Liu, L., & Xu, Y. (2022). A Bayesian Alternative to Synthetic Control for Comparative Case Studies. Political Analysis, 30(2), 269-288. doi:10.1017/pan.2021.22

5.1 Setup

- unit: $i = 1, 2, ..., N$. Suppose there are N_{co} control units and N_{tr} treated units; $N_{co} + N_{tr} = N$.
- time period: $t = 1, 2, ..., T$
- time of adoption for each i: $a_i \in \mathbb{A} = \{1, 2, ..., T, c\}$, $a_i = c > T$ means that unit i falls in the residual category and does not get treated in the observed time window. We call i a treated unit if $a_i = 1, ..., T$ and we call it a control unit if $a_i = c$.
- The number of pre-treatment periods for a treated unit: $T_{0,i} = a_i - 1$
- Denote $\mathcal{A} = \{a_1, ..., a_N\}$
- Denote $\mathbf{w_i} = (w_{i1}, ..., w_{iT})'$ as the treatment assignment vector for unit i. Define an $(N \times T)$ treatment assignment matrix: $\mathbf{W} = \{\mathbf{w}_1, ..., \mathbf{w}_N\}$

Assumption 5.1 *Cross-sectional stable unit treatment value assumption (SUTVA): Potential outcomes of unit i are only functions of the treatment status of unit i:* $\mathbf{y}_{it}(\mathbf{W}) = \mathbf{y}_{it}(\mathbf{w}_i), \forall i, t.$

This assumption rules out cross-sectional spillover effects.

Assumption 5.2 *No anticipation: For all unit i,for all time periods before adoption* $t < a_i$:

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

Estimand
For treated unit i whose adoption time $a_i \leq T$, we define its treatment effect at $t \geq a_i$ as

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

ATT for units that have been under the treatment for a duration of p periods:

$$\delta_p = \frac{1}{N_{tr,p}} \sum_{i:T-p+1 \leq a_i \leq T} \delta_{i,a_i+p-1},$$

where $N_{tr,p}$ is the number of treated units that have been treated for p periods in the sample.

5.2 Assignment Mechanism

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: $S_0 \equiv \{(it) \mid w_{it} = 0\}$, which $y_{it}(c)$ is observed; and $S_1 \equiv \{(it) \mid w_{it} = 1\}$, with which $y_{it}(c)$ is missing. $S = S_0 \cup S_1$. We denote the observed and missing parts of $\mathbf{Y}(\mathbf{0})$ as $\mathbf{Y}(0)^{obs}$ and $\mathbf{Y}(0)^{mis}$
$\mathbf{X}_{it}$ is a $(p_1 \times 1)$ vector of exogenous covariates. $\mathbf{X}_i = (\mathbf{X}_{i1}, ..., \mathbf{X}_{iT})'$ is a $(T \times p_1)$ covariate matrix and define $\mathbf{X} = \{\mathbf{X}_1, ..., \mathbf{X}_N\}$.

The posterior predictive distribution of $\mathbf{Y}(\mathbf{0})^{mis}$ can be written as

$$\begin{aligned}
\Pr(\mathbf{Y}(\mathbf{0})^{mis} \mid \mathbf{X}, \mathbf{Y}(\mathbf{0})^{obs}, \mathcal{A}) &= \frac{\Pr(\mathbf{X}, \mathbf{Y}(\mathbf{0})^{mis}, \mathbf{Y}(\mathbf{0})^{obs}) \Pr(\mathcal{A} \mid \mathbf{X}, \mathbf{Y}(\mathbf{0})^{mis}, \mathbf{Y}(\mathbf{0})^{obs})}{\Pr(\mathbf{X}, \mathbf{Y}(\mathbf{0})^{obs}, \mathcal{A})} \\
&\propto \Pr(\mathbf{X}, \mathbf{Y}(\mathbf{0})^{mis}, \mathbf{Y}(\mathbf{0})^{obs}) \Pr(\mathcal{A} \mid \mathbf{X}, \mathbf{Y}(\mathbf{0})^{mis}, \mathbf{Y}(\mathbf{0})^{obs}) \\
&\propto \Pr(\mathbf{X}, \mathbf{Y}(\mathbf{0})) \Pr(\mathcal{A} \mid \mathbf{X}, \mathbf{Y}(\mathbf{0})) \qquad (6)
\end{aligned}$$

Assumption 5.3 *Individualistic assignment and positivity:* $\Pr(\mathcal{A} \mid \mathbf{X}, \mathbf{Y}(\mathbf{0})) = \prod_{i=1}^{n} \Pr(a_i \mid \mathbf{X}_i, \mathbf{Y}_i(\mathbf{0}))$ *and* $0 < \Pr(a_i \mid \mathbf{X}_i, \mathbf{Y}_i(\mathbf{0})) < 1$ *for all unit i.*

This assumption ensures that the adoption time of unit i does not depend on other units and each i has some non-zero chances of getting treated.

Assumption 5.4 *Latent ignorability (exogeneity):*

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

Assumption 5.5 *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* $\mathbf{U} = (\mathbf{U}_1, ..., \mathbf{U}_N)$ *can be approximated by two lower-rank matrices* $(r << \min\{N, T\})$*, i.e.,* $\mathbf{U} = \mathbf{\Gamma}'\mathbf{F}$ *in which* $\mathbf{F} = (\mathbf{f}_1, ..., \mathbf{f}_T)$ *is a* $(r \times T)$ *matrix of factors and* $\mathbf{\Gamma} = (\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_N)$ *is a* $(r \times N)$ *matrix of the factor loadings.*

Assumption 5.6 *Exchangeability: When* $\mathbf{U}$ *is known,* $\{(\mathbf{X}'_{it}, y_{it}(c))\}_{i=1,...,N;t=1,...,T}$ *is an exchangeable sequence of random variables; i.e., the joint distribution of* $\{(\mathbf{X}'_{it}, y_{it}(c))\}$ *is invariant to permutations in the index it, i.e.,* $\{(\mathbf{X}'_{it}, y_{it}(c))\}$ *is i.i.d.*

We can 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

$$\begin{aligned}
\Pr(\mathbf{Y}(\mathbf{0})^{mis} \mid \mathbf{X}', \mathbf{Y}(\mathbf{0})^{obs}, \mathcal{A}) &= \frac{\Pr(\mathbf{X}', \mathbf{Y}(\mathbf{0})^{mis}, \mathbf{Y}(\mathbf{0})^{obs}) \Pr(\mathcal{A} \mid \mathbf{X}', \mathbf{Y}(\mathbf{0})^{mis}, \mathbf{Y}(\mathbf{0})^{obs})}{\Pr(\mathbf{X}', \mathbf{Y}(\mathbf{0})^{obs}, \mathcal{A})} \\
&\propto \Pr(\mathbf{X}', \mathbf{Y}(\mathbf{0})^{mis}, \mathbf{Y}(\mathbf{0})^{obs}) \Pr(\mathcal{A} \mid \mathbf{X}', \mathbf{Y}(\mathbf{0})^{mis}, \mathbf{Y}(\mathbf{0})^{obs}) \\
&\propto \Pr(\mathbf{X}', \mathbf{Y}(\mathbf{0})) \Pr(\mathcal{A} \mid \mathbf{X}', \mathbf{Y}(\mathbf{0})) \\
&\propto \Pr(\mathbf{X}', \mathbf{Y}(\mathbf{0})) \Pr(\mathcal{A} \mid \mathbf{X}') \text{ by latent ignorability} \\
&\propto \Pr(\mathbf{X}', \mathbf{Y}(\mathbf{0})) \\
&\propto \Pr(\{(\mathbf{X}'_{it}, y_{it}(c))\}) \text{ by exchangeability} \\
&\propto \int \underbrace{\left(\prod_{it \in S_1} f(y_{it}(c)^{mis} \mid \mathbf{X}_{it}, \boldsymbol{\theta}') \right)}_{\text{posterior predictive distribution}} \underbrace{\left(\prod_{it \in S_0} f(y_{it}(c)^{obs} \mid \mathbf{X}_{it}, \boldsymbol{\theta}') \right)}_{\text{likelihood}} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}
\end{aligned} \tag{7}$$

where $\boldsymbol{\theta}$ are the parameters that govern the data-generating process of $y_{it}(c)$ given $\mathbf{X}'_{it}$; and $\boldsymbol{\theta}' = (\boldsymbol{\theta}, \mathbf{U})$ when we regard the latent covaraites $\mathbf{U}$ as parameters.

5.3 Framework of a dynamic multilevel latent factor model (DM-LFM)

Assumption 5.7 *Functional form: The untreated potential outcomes for unit $i = 1, ..., N$ at $t = 1, ..., T$ are specified as follows:*

$$y_{it}(c) = \mathbf{X}'_{it}\boldsymbol{\beta}_{it} + \boldsymbol{\gamma}'_i \mathbf{f}_t + \epsilon_{it}, \tag{8}$$

$$\boldsymbol{\beta}_{it} = \boldsymbol{\beta} + \boldsymbol{\alpha}_i + \boldsymbol{\xi}_t, \tag{9}$$

$$\boldsymbol{\xi}_t = \Phi_\xi \boldsymbol{\xi}_{t-1} + \mathbf{e}_t, \mathbf{f}_t = \Phi_f \mathbf{f}_{t-1} + \boldsymbol{\nu}_t \tag{10}$$

Equation (10) models the dynamics in $\boldsymbol{\xi}_t$ and $\mathbf{f} + t$ by specifying autoregressive processes. Φ_ξ and Φ_f are assumed to be diagonal: $\Phi_\xi = \text{Diag}(\phi_{\xi_1}, ..., \phi_{\xi_{p_3}})$ and $\Phi_f = \text{Diag}(\phi_{f_1}, ..., \phi_{f_r})$. $\epsilon_{it}, \mathbf{e}_t, \boldsymbol{\nu}_t$ are i.i.d. normal.
The DM-LFM allows the slope coefficient of each covariate to vary by unit, time, both, or neither. We can rewrite the individual-level model in a reduced and matrix format as

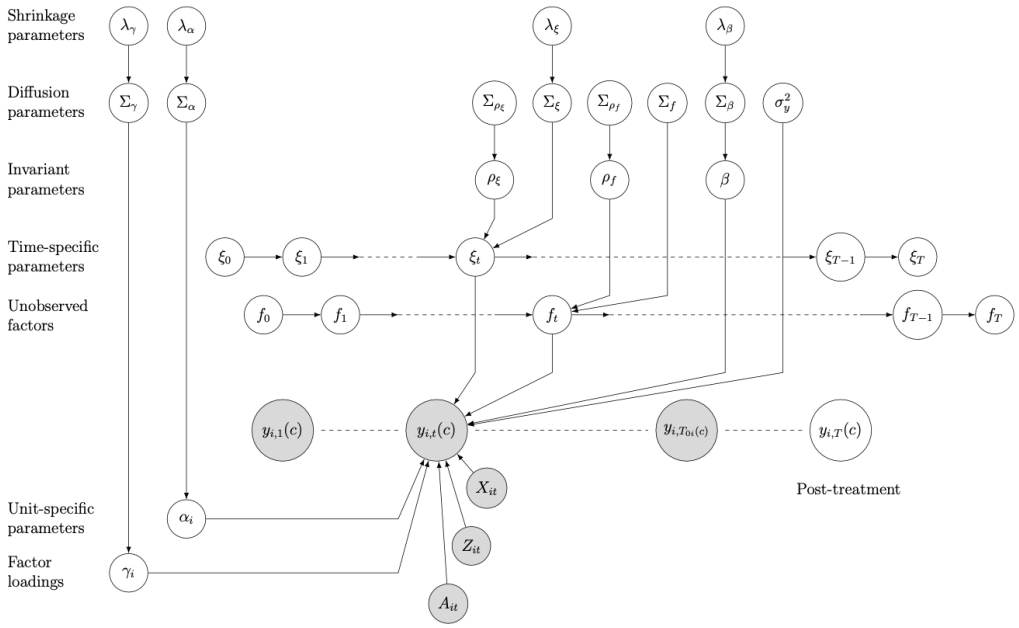
$$\mathbf{y}_i(c) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\alpha}_i + \mathbf{A}_i\boldsymbol{\xi} + \mathbf{F}\boldsymbol{\gamma}_i + \boldsymbol{\epsilon}_i \tag{11}$$

where

- $\mathbf{F} = (\mathbf{f}_1, ..., \mathbf{f}_T)'$ (a $T \times r$ matrix)
- $\mathbf{Z}_i$ (a $T \times p_2$ matrix) are covariates that have unit-specific slpes $\boldsymbol{\beta} + \boldsymbol{\alpha}_i$
- $\mathbf{A}_i$ (a $T \times p_3$ matrix) are covariates that have unit-specific slpes $\boldsymbol{\beta} + \boldsymbol{\xi}_i$. (Note: $\mathbf{A}_i$ and $\mathbf{Z}_i$ are subsets of $\mathbf{X}_i$.)

We can add $\delta_{it} w_{it}$ in the model, in which δ_{it} is the causal effect for unit i at time t. Then

$$y_{it} = \delta_{it} w_{it} + \mathbf{X}'_{it}\boldsymbol{\beta}_{it} + \boldsymbol{\gamma}'_i \mathbf{f}_t + \epsilon_{it}$$



Note: 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.

Figure 1: A graphic representation of DM-LFM.

5.4 Bayesian Stochastic Model Specification Search

This paper uses shrinkage priors to choose the number of latent factors and decide whether and how to include a covariate. Specifically, it applies the Bayesian Lasso shrinkage on β using the following hierarchical setting of mixture of a normal-exponential prior,

$$\beta_k \mid \tau^2_{\beta_k} \sim \mathcal{N}(0, \tau^2_{\beta_k}), \tau^2_{\beta_k} \mid \lambda_\beta \sim Exp(\frac{\lambda^2_\beta}{2}), \lambda^2_\beta \sim \mathcal{G}(a_1, a_2), k = 1, ..., p_1.$$

To select the other components of the model, it also imposes shrinkage on $\boldsymbol{\alpha}_i, \boldsymbol{\xi}_t$, or $\boldsymbol{\gamma}_i$ to determine whether to include a $Z_j (j = 1, 2, ..., p_2), A_j (j = 1, 2, ..., p_3), or f_j (j = 1, 2, ..., r)$ in the model. It uses Lasso-like hierarchical shrinkage approach with re-parameterization.
Assume $\boldsymbol{\alpha}_i, \boldsymbol{\gamma}_i$, and $\boldsymbol{\xi}_t$ have diagonal variance-covariance matrices, $\mathbf{H}_0 = \text{Diag}(\omega^2_{\alpha_1}, ..., \omega^2_{\alpha_{p_2}}), \boldsymbol{\Gamma}_0 = \text{Diag}(\omega^2_{\gamma_1}, ..., \omega^2_{\gamma_r}), \Sigma_e = \text{Diag}(\omega^2_{\xi_1}, ..., \omega^2_{\xi_{p_3}})$, respectively.
Then re-parameterize $\boldsymbol{\alpha}_i, \boldsymbol{\gamma}_i$, and $\boldsymbol{\xi}_t$ as $\boldsymbol{\alpha}_i = \boldsymbol{\omega}_\alpha \cdot \tilde{\boldsymbol{\alpha}}_i, \boldsymbol{\gamma}_i = \boldsymbol{\omega}_\gamma \cdot \tilde{\boldsymbol{\gamma}}_i, \boldsymbol{\xi}_t = \boldsymbol{\omega}_\alpha \cdot \tilde{\boldsymbol{\xi}}_t$, where $\boldsymbol{\omega}_\alpha = (\omega_{\alpha_1}, ..., \omega_{\alpha_{p_2}})', \boldsymbol{\omega}_\gamma = (\omega_{\gamma_1}, ..., \omega_{\gamma_r})', \boldsymbol{\omega}_\xi = (\omega_{\xi_1}, ..., \omega_{\xi_{p_3}})'$. Together with the shrinkage on β, the algorithm will decide in fact whether a certain covariate is included, whether its coefficient varies by time or across units, and how many latent factors are considered.

5.5 Implementing a DM-LFM

Implementing a DM-LFM takes the following three steps:

1. Step 1: Model searching and parameter estimation.
   We specify and estimate the DM-FLM model with Bayesian shrinkage to sample G draws of the parameters from their posterior distributions, $\boldsymbol{\theta}_{it}^{(g)} \sim \pi(\boldsymbol{\theta}_{it} \mid \mathcal{D})$, where $\mathcal{D} = \{(\mathbf{X}_{it}, y_{it}(c)^{obs}) : it \in S_0\}$ is the set of untreated observations.

2. Step 2: Prediction and integration.
   Conduct Bayesian prediction by generating draws of counterfactual $y_{it}(c)^{mis}$ for each treated unit at $a_i \leq t \leq T$ from its posterior predictive distribution. A sample of the predicted counterfactual is generated by plugging each draw $\boldsymbol{\theta}_{it}^{(g)}$ from $\pi(\boldsymbol{\theta}_{it} \mid \mathcal{D})$ into $f(y_{it}(c)^{mis} \mid \mathbf{X}_{it}, \boldsymbol{\theta}_{it})$ to obtain $y_{it}^{(g)}(c)$ for $g = 1, ..., G$.

3. Step 3: Inference and diagnostics.
   Summarize the results (δ_{it}^{g}): posterior mean, variance, and the Bayesian 95% credibility intereval.

The MCMC algorithm
Re-parameterize the reduced form model:

$$y(c)_{it} = \mathbf{X}_{it}'\boldsymbol{\beta} + \mathbf{Z}_{it}'(\boldsymbol{\omega}_\alpha \cdot \tilde{\boldsymbol{\alpha}}_i) + \mathbf{A}_{it}'(\boldsymbol{\omega}_{\boldsymbol{\xi}} \cdot \tilde{\boldsymbol{\xi}}_t) + (\boldsymbol{\omega}_{\boldsymbol{\gamma}} \cdot \tilde{\boldsymbol{\gamma}}_i)'\mathbf{f}_t + \epsilon_{it},$$

and $\tilde{\boldsymbol{\xi}}_t = \Phi_\xi \tilde{\boldsymbol{\xi}}_{t-1} + \tilde{\mathbf{e}}_t$ and $\tilde{\alpha}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{p_2}), \tilde{\boldsymbol{\gamma}}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_r), \tilde{\mathbf{e}}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{p_3})$. We assign Bayesian Lasso priors to the following parameters:

$$\begin{aligned}
&\beta_k \mid \tau_{\beta_k}^2 \sim \mathcal{N}(0, \tau_{\beta_k}^2), \tau_{\beta_k}^2 \mid \lambda_\beta \sim \text{Exp}(\frac{\lambda_\beta^2}{2}), \lambda_\beta^2 \sim \mathcal{G}(a_1, a_2), \forall 1 \leq k \leq p_1;\\
&\omega_{\alpha_j} \mid \tau_{\alpha_j}^2 \sim \mathcal{N}(0, \tau_{\alpha_j}^2), \tau_{\alpha_j}^2 \mid \lambda_\alpha \sim \text{Exp}(\frac{\lambda_\alpha^2}{2}), \lambda_\alpha^2 \sim \mathcal{G}(b_1, b_2), \forall 1 \leq j \leq p_2;\\
&\omega_{\xi_j} \mid \tau_{\xi_j}^2 \sim \mathcal{N}(0, \tau_{\xi_j}^2), \tau_{\xi_j}^2 \mid \lambda_\xi \sim \text{Exp}(\frac{\lambda_\xi^2}{2}), \lambda_\xi^2 \sim \mathcal{G}(c_1, c_2), \forall 1 \leq j \leq p_3;\\
&\omega_{\gamma_j} \mid \tau_{\gamma_j}^2 \sim \mathcal{N}(0, \omega_{\gamma_j}^2), \omega_{\gamma_j}^2 \mid \lambda_\gamma \sim \text{Exp}(\frac{\lambda_\gamma^2}{2}), \lambda_\gamma^2 \sim \mathcal{G}(k_1, k_2), \forall 1 \leq j \leq r. \qquad (12)
\end{aligned}$$

where $\mathbf{f}_t = \Phi_f \mathbf{f}_{t-1} + \boldsymbol{\nu}_t$ with $\boldsymbol{\nu}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_r)$. We assume $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$ and $\sigma_\epsilon^{-2} \sim \mathcal{G}(e_1, e_2)$.
Because the number of factors r is unknown, we presume a reasonably large positive integer for r (the initial number of factors) and let the algorithm determine its value, based on the posterior distributions of $\boldsymbol{\omega}_\gamma$
The MCMC algorithm for the DM-LFM (which simulates parameter posteriors and predicts $Y(\mathbf{0})^{miss}$ using data $\mathcal{D}$):

1. Start with the initial values of the parameters $\boldsymbol{\theta}^{(0)}$

2. in the gth iteration, sample from the following conditional distributions based on the most updated values $\boldsymbol{\theta}^{g-1}$

(a) update $\boldsymbol{\beta}$:

$$\begin{aligned}
&\boldsymbol{\beta} \sim \mathcal{N}(\bar{\boldsymbol{\beta}}, \mathbf{B}_1) \\
&\mathbf{B}_1 = \left(\sigma_\epsilon^{-2} \sum_{w_{it}=0} \mathbf{X}_{it}\mathbf{X}_{it}' + \mathbf{B}_0^{-1} \right)^{-1} \\
&\bar{\boldsymbol{\beta}} = \mathbf{B}_1(\sigma_\epsilon^{-2} \sum_{w_{it}=0} \mathbf{X}_{it}u_{it}) \\
&\mathbf{B}_0^{-1} = \mathrm{Diag}(\tau_{\beta_1}^{-2}, ..., \tau_{\beta_{p_1}}^{-2}) \\
&u_{it} = y_{it}(c)^{obs} - \mathbf{Z}_{it}'(\boldsymbol{\omega}_{\boldsymbol{\alpha}} \cdot \tilde{\boldsymbol{\alpha}}_i) - \mathbf{A}_{it}'(\boldsymbol{\omega}_{\boldsymbol{\xi}} \cdot \tilde{\boldsymbol{\xi}}_t) - (\boldsymbol{\omega}_{\boldsymbol{\gamma}} \cdot \tilde{\boldsymbol{\gamma}}_i)'\mathbf{f}_t; \qquad (13)
\end{aligned}$$

(b) update $(\tilde{\boldsymbol{\alpha}}_i', \tilde{\boldsymbol{\gamma}}_i')$:
Denote $\tilde{\mathbf{Z}}_{it} = (\mathbf{Z}_{it}' \cdot \boldsymbol{\omega}_\alpha', \boldsymbol{\omega}_\gamma' \cdot \mathbf{f}_t')'$

$$(\tilde{\boldsymbol{\alpha}}_i', \tilde{\boldsymbol{\gamma}}_i')' \sim \mathcal{N}(\bar{\boldsymbol{\alpha}}, \mathbf{H}_1)$$

$$\mathbf{H}_1 = \left(\sigma_\epsilon^{-2} \sum_{i, w_{it}=0} \tilde{\mathbf{Z}}_{it}\mathbf{Z}_{it}' + \mathbf{I}_{p_2+r} \right)^{-1}$$

$$\bar{\boldsymbol{\alpha}} = \mathbf{H}_1 \left(\sigma_\epsilon^{-2} \sum_{i, w_{it}=0} \mathbf{Z}_{it}u_{it} \right)$$

$$u_{it} = y_{it}(c)^{obs} - \mathbf{X}_{it}'\boldsymbol{\beta} - \mathbf{A}_{it}'(\boldsymbol{\omega}_{\boldsymbol{\xi}} \cdot \tilde{\boldsymbol{\xi}}_t)$$

(c) update $(\tilde{\boldsymbol{\xi}}_t', \mathbf{f}_t')'$:
Denote $\boldsymbol{\Psi}_t = (\tilde{\boldsymbol{\xi}}_t', \mathbf{f}_t')'$, $\tilde{\mathbf{A}}_{it} = (\mathbf{A}_{it}' \cdot \boldsymbol{\omega}_\xi', \boldsymbol{\omega}_\gamma \cdot \tilde{\boldsymbol{\gamma}}_i')'$ and $\Phi = \mathrm{diag}(\phi_{\xi_1}, ..., \phi_{\xi_{p_3}}, \phi_{f_1}, ..., \phi_{f_r})$.
Assume $\boldsymbol{\Psi}_0 = \mathbf{0}$ as initial state.

$$\boldsymbol{\Psi}_t \sim \mathcal{N}(\boldsymbol{\Omega}_t^{-1}\boldsymbol{\mu}_t, \boldsymbol{\Omega}_t^{-1})$$

$$\boldsymbol{\mu}_t = \begin{cases} \Phi\boldsymbol{\Psi}_{t-1} + \sigma_\epsilon^{-2} \sum_{t, w_{it}=0} \tilde{\mathbf{A}}_{it}u_{it} + \phi\Psi_{t+1}, 1 \le t \le T-1 \\ \Phi\boldsymbol{\Psi}_{t-1} + \sigma_\epsilon^{-2} \sum_{t, w_{it}=0} \tilde{\mathbf{A}}_{it}u_{it}, t = T \end{cases}$$

$$\boldsymbol{\Omega}_t = \begin{cases} \mathbf{I}_{p_3+r} + \sigma_\epsilon^{-2} \sum_{t, w_{it}=0} \tilde{\mathbf{A}}_{it}\tilde{\mathbf{A}}_{it}' + \Phi\Phi, 1 \le t \le T-1 \\ \mathbf{I}_{p_3+r} + \sigma_\epsilon^{-2} \sum_{t, w_{it}=0} \tilde{\mathbf{A}}_{it}\tilde{\mathbf{A}}_{it}', t = T \end{cases}$$

$$u_{it} = y_{it}(c)^{obs} - \mathbf{X}_{it}'\boldsymbol{\beta} - \mathbf{Z}_{it}'(\boldsymbol{\omega}_\alpha \cdot \tilde{\boldsymbol{\alpha}}_i);$$

(d) update $(\boldsymbol{\omega}_\alpha', \boldsymbol{\omega}_\xi', \boldsymbol{\omega}_\gamma')'$:
Denote $\boldsymbol{\omega} = (\boldsymbol{\omega}_\alpha', \boldsymbol{\omega}_\xi', \boldsymbol{\omega}_\gamma')'$ and $\tilde{\mathbf{Z}}_{it} = (\mathbf{Z}_{it}' \cdot \tilde{\boldsymbol{\alpha}}_i', \mathbf{A}_{it}' \cdot \tilde{\boldsymbol{\xi}}_t', \tilde{\boldsymbol{\gamma}}_i' \cdot \mathbf{f}_t')'$.

$$\boldsymbol{\omega} \sim \mathcal{N}(\bar{\boldsymbol{\omega}}, \boldsymbol{\Omega}_1)$$

$$\boldsymbol{\Omega}_1 = \left(\sigma_\epsilon^{-2} \sum_{w_{it}=0} \tilde{\mathbf{Z}}_{it} \tilde{\mathbf{Z}}'_{it} + \boldsymbol{\Omega}_0^{-1} \right)^{-1}$$

$$\bar{\boldsymbol{\omega}} = \boldsymbol{\Omega}_1 \left(\sigma_\epsilon^{-2} \sum_{w_{it}=0} \tilde{\mathbf{Z}}_{it} u_{it} \right)$$

$$\boldsymbol{\Omega}_0^{-1} = \text{Diag}(\tau_{\alpha_1}^{-2}, ..., \tau_{\alpha_{p_2}}^{-2}, \tau_{\xi_1}^{-2}, ..., \tau_{\xi_{p_3}}^{-2}, \tau_{f_1}^{-2}, ..., \tau_{f_r}^{-2}$$

$$u_{it} = y_{it}(c)^{obs} - \mathbf{X}'_{it}\boldsymbol{\beta}$$

(e) update autoregressive coefficients:

i. ϕ_{ξ_j} in Φ_ξ for $j = 1, ..., p_3$

$$\phi_{\xi_j} \sim \mathcal{N}(\bar{\phi}, \Phi_1)$$

$$\Phi_1 = \left(\sum_{t=1}^{T} \tilde{\xi}_{j,t-1}^2 + \sigma_\phi^{-2} \right)^{-1}$$

$$\bar{\phi} = \Phi_1 \left(\tilde{\xi}_{j,t} \tilde{\xi}_{j,t-1} \right)$$

ii. ϕ_{f_j} in Φ_f for $j = 1, ..., r$

$$\phi_{f_j} \sim \mathcal{N}(\bar{\phi}, \Phi_1)$$

$$\Phi_1 = \left(\sum_{t=1}^{T} f_{j,t-1}^2 + \sigma_\phi^{-2} \right)^{-1}$$

$$\bar{\phi} = \Phi_1 \left(f_{j,t} f_{j,t-1} \right)$$

(f) update $\tau_{\beta_j}^2$:

$$\tau_{\beta_j}^{-2} \sim IG(\sqrt{\frac{\lambda_\beta^2}{\beta_j^2}}, \lambda_\beta^2), \forall 1 \leq j \leq p_1;$$

(g) update $\tau_{\alpha_j}^2$:

$$\tau_{\alpha_j}^{-2} \sim IG(\sqrt{\frac{\lambda_\alpha^2}{\omega_{\alpha_j}^2}}, \lambda_\alpha^2), \forall 1 \leq j \leq p_2;$$

(h) update $\tau_{\xi_j}^2$:

$$\tau_{\xi_j}^{-2} \sim IG(\sqrt{\frac{\lambda_\xi^2}{\omega_{\xi_j}^2}}, \lambda_\xi^2), \forall 1 \leq j \leq p_3;$$

(i) update $\tau_{\gamma_j}^2$:

$$\tau_{\gamma_j}^{-2} \sim IG(\sqrt{\frac{\lambda_\gamma^2}{\omega_{\gamma_j}^2}}, \lambda_\gamma^2), \forall 1 \leq j \leq r;$$

(j) update λ_β^2:

$$\lambda_\beta^2 \sim \mathcal{G}(p_1 + a_1, \frac{1}{2} \sum_{j=1}^{p_1} \tau_{\beta_j}^2 + a_2);$$

(k) update λ_α^2:

$$\lambda_\alpha^2 \sim \mathcal{G}(p_2 + b_1, \frac{1}{2}\sum_{j=1}^{p_2} \tau_{\alpha_j}^2 + b_2);$$

(l) update λ_ξ^2:

$$\lambda_\xi^2 \sim \mathcal{G}(p_3 + c_1, \frac{1}{2}\sum_{j=1}^{p_3} \tau_{\xi_j}^2 + c_2);$$

(m) update λ_γ^2:

$$\lambda_\gamma^2 \sim \mathcal{G}(r + k_1, \frac{1}{2}\sum_{j=1}^{r} \tau_{\gamma_j}^2 + k_2);$$

(n) update σ_ϵ^2:

$$\sigma_\epsilon^{-2} \sim \mathcal{G}(N_{obs} + e_1, \frac{1}{2}\sum_{D_{it}=0} (y_{it} - u_{it})^2 + e_2)$$
$$N_{obs} = N \times T - N_{tr} \times (T - T_0)$$
$$u_{it} = y_{it}(c)^{obs} - \mathbf{X}_{it}'\boldsymbol{\beta} - \mathbf{Z}_{it}'(\boldsymbol{\omega_\alpha} \cdot \tilde{\boldsymbol{\alpha}}_i) - \mathbf{A}_{it}'(\boldsymbol{\omega_\xi} \cdot \tilde{\boldsymbol{\xi}}_t) - (\boldsymbol{\omega_\gamma} \cdot \tilde{\boldsymbol{\xi}}_i)'\mathbf{f}_t;$$

(o) update predicted $y_{it}(c)^{mis}$ for observations under treatment:

$$y_{it}(c)^{mis} \sim \mathcal{N}(\mu_{it}, \sigma_\epsilon^2), \text{ for } it \in S_1$$
$$\mu_{it} = \mathbf{X}_{it}'\boldsymbol{\beta} + \mathbf{Z}_{it}'(\boldsymbol{\omega_\alpha} \cdot \tilde{\boldsymbol{\alpha}}_i) + \mathbf{A}_{it}'(\boldsymbol{\omega_\xi} \cdot \tilde{\boldsymbol{\xi}}_t) + (\boldsymbol{\omega_\gamma} \cdot \tilde{\boldsymbol{\xi}}_i)'\mathbf{f}_t;$$

(p) obtain an estimate for δ_{it}: $\delta_{it} = y_{it} - y_{it}(c)^{mis}$, for $it \in S_1$

3. Repeat (a)-(p) until convergence and obtain G draws for each parameter, counterfactual and the individual causal effect.

6 Bayesian synthetic control methods

Kim, S., Lee, C., and Gupta, S. (2020). Bayesian Synthetic Control Methods. Journal of Marketing Research, 57(5), 831–852. https://doi.org/10.1177/0022243720936230

6.1 Bayesian Formulations of Extant SCM

6.1.1 The standard SCM of Panel Data Methods

Standard SCM:

$$\beta \mid \sigma \sim \text{uniform}(-\infty, \infty) \text{ for } j = 1...., J \tag{14}$$
$$\sigma \sim \text{uniform}(0, \infty) \tag{15}$$
$$\text{s.t. } \beta_1 = 0, \beta_j \geq 0, \text{ and } \sum_{j=2}^{J+1} \beta_j = 1 \tag{16}$$

Panel data methods:

$$\beta \mid \sigma \sim \text{uniform}(-\infty, \infty) \text{ for } j = 2...., J+1$$
$$\sigma \sim \text{uniform}(0, \infty)$$
$$\beta_1 \sim \text{uniform}(-\infty, \infty)$$

6.1.2 Bayesian regularized regression methods

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

$$\begin{aligned}
&\beta_j \mid \lambda, \sigma \sim \text{Normal } (0, \frac{\sigma^2}{\lambda}) \text{ for } j = 2, ..., J+1, \\
&\lambda \sim \text{Cauchy}^+(0, 10), \\
&\sigma \sim \text{Cauchy}^+(0, 10), \\
&\beta_1 \sim \text{Cauchy}^+(0, 10)
\end{aligned} \tag{17}$$

The Bayesian lasso is obtained by specifying Laplace priors on the regression coefficients (Park & Casella, 2008)

$$\begin{aligned}
&\beta_j \mid \lambda, \sigma \sim \text{Laplace } (0, \frac{\sigma^2}{\lambda}) \text{ for } j = 2, ..., J+1, \\
&\lambda \sim \text{Cauchy}^+(0, 10), \\
&\sigma \sim \text{Cauchy}^+(0, 10), \\
&\beta_1 \sim \text{Cauchy}^+(0, 10)
\end{aligned} \tag{18}$$

The Bayesian elastic net prior (Li and Lin, 2010):

$$\begin{aligned}
&\beta_j \mid \lambda_2, \tau_j, \sigma \sim \text{Normal } [0, \left(\frac{\lambda_2}{\sigma^2} \frac{\tau_j}{\tau_j - 1}\right)] \text{ for } j = 2, ..., J+1, \\
&\tau_j \mid \lambda_1, \lambda_2, \sigma \sim \text{Truncated Gamma}[\frac{1}{2}, \frac{8\lambda_2\sigma^2}{\lambda_1^2}, (1, \infty)], \\
&\lambda_1 \sim \text{Cauchy}^+(0, 10), \\
&\lambda_2 \sim \text{Cauchy}^+(0, 10), \\
&\sigma \sim \text{Cauchy}^+(0, 10), \\
&\beta_1 \sim \text{Cauchy}^+(0, 10)
\end{aligned} \tag{19}$$

Advantages of Bayesian regularized regression methods:

- Bayesian methods enable exact statistical inference even when the sample size is small.
- Bayesian methods use shrinkage priors as a natural mechanism to deal with the sparsity problem and the "large p, small n" problem.
- Bayesian methods enable estimation of the penalty parameters (i.e., λ) jointly with the parameters of interest (i.e., β_j) so that model selection and parameter estimation are performed concurrently.

6.2 Proposed Bayesian Synthetic Control Methods

The authors propose BSCM horseshoe and BSCM spike and slab. The two models have an additional advantage: better predictive accuracy.
As shown in figure **??**, horseshoe and spike-and-slab densities show tall spikes around zero (i.e.,

noise) and flat tails for nonzero parameter values (i.e., signals), which provide accurate inferences about whether each β_j is signal or noise. This ability to clearly distinguish noise from signal is especially desirable in the SCM context where the "large p, small n" and/or sparsity problems are common concerns.
Horseshoe prior (Carvalho et al., 2010):

$$\begin{aligned}
&\beta_j \mid \lambda_j \sim \text{Normal}(0, \lambda_j^2) \text{ for } j = 2, ..., J+1, \\
&\lambda_j \mid \tau \sim \text{Cauchy}^+(0, \tau), \\
&\tau \mid \sigma \sim \text{Cauchy}^+(0, \sigma), \\
&\sigma \sim \text{Cauchy}^+(0, 10)
\end{aligned} \tag{20}$$

where τ is a global shrinkage parameter that provides several shrinkage for all the parameters toward zero, while the local shrinkage parameter λ_j allows some β_js to escape the severe shrinkage by providing heavy tails.

7 On the Misspecification of Linear Assumptions in Synthetic Control

Nazaret, A., Shi, C., Blei, D. M. (2023). On the Misspecification of Linear Assumptions in Synthetic Control. arXiv preprint. Retrieved from https://arxiv.org/abs/2302.12777

In synthetic control methods, an implicit assumption is that the treated unit needs to be expressed as a linear combination of the control units. What if this requirement is not satisfied? This paper studies the practical situation where the synthetic control is misspecified.