Assessing the causal effect of binary interventions from observational panel data with few treated units

Pantelis Samartsidis
MRC Biostatistics Unit, University of Cambridge

17th Armitage Workshop
Cambridge, 7 Nov. 2019
Motivation

**Problem:** causal inference in a setting where

1. Intervention is **non-randomised** and **binary**
2. Data at **multiple time points**, before and after intervention
3. Only a **few treated** units
Motivation

**Problem:** causal inference in a setting where

1. Intervention is **non-randomised** and **binary**
2. Data at **multiple time points**, before and after intervention
3. Only a **few treated** units

**Objectives:**

- Highlight the **importance/challenges** of the problem
- Review **state of the art** methodologies/ongoing work
- Outline directions for **future research**
Example 1: alcohol licensing (de Vocht et al., 2017)

- **Outcome:** alcohol-related hospitalisations per 10,000
- **Units:** local councils in UK
- **Intervention:** stricter licensing policies
- **Study period:** mid-2009 to 2013 (quarterly)
Example 2: new treatment against HCV

- Outcome: Hepatitis C virus prevalence
- Units: regions in Scotland
- Intervention: treatment scale-up among people who inject drugs
- Study period: 2011-? (ongoing)

<table>
<thead>
<tr>
<th>Unit</th>
<th>2008/9</th>
<th>2010</th>
<th>2011/12</th>
<th>2013/14</th>
<th>2015/16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tayside (treated)</td>
<td>30.2</td>
<td>40.2</td>
<td>38.5</td>
<td>46.7</td>
<td>43.6</td>
</tr>
<tr>
<td>Glasgow</td>
<td>66.1</td>
<td>63.6</td>
<td>60.1</td>
<td>65.9</td>
<td>60.8</td>
</tr>
<tr>
<td>Rest of Scotland</td>
<td>43.9</td>
<td>45.3</td>
<td>43.8</td>
<td>45.0</td>
<td>48.0</td>
</tr>
</tbody>
</table>
Example 3: California tobacco (Abadie et al., 2010)

- Outcome: per-capita cigarette sales in California
- Units: USA states
- Intervention: Proposition 99
- Study period: 1970-2003 (annually)
Challenges for causal inference

1. Adjust for observed confounders (covariates)

2. Potential of unobserved confounding

3. Account for temporal trends

4. Propensity score methodologies (Robins et al., 2000) not suitable
Notation & causal framework

Notation:
- $n$ units (indexed by $i$): $n_1$ controls and $n_2$ treated
- $T$ times (indexed by $t$): $T_1$ pre- and $T_2$ post-intervention
- Outcome $y_{it}$ and covariates $x_{it}$

Rubin causal model (Holland, 1986):
- Treatment free outcomes $y_{i0}(t)$
- Outcomes under intervention $y_{i1}(t)$
- Hence, observed data are: $y_{it} = \{y_{i1}(t), i > n_1 \text{ and } t > T_1\}
  \{y_{i0}(t), \text{otherwise}\}$
- Causal effects $\theta_{it} = y_{i1}(t) - y_{i0}(t)$
- Some untestable assumptions are required to estimate counterfactuals $\hat{y}_{i0}(t)$ from observed data
Notation & causal framework

Notation:

- $n$ units (indexed by $i$): $n_1$ controls and $n_2$ treated
- $T$ times (indexed by $t$): $T_1$ pre- and $T_2$ post-intervention
- Outcome $y_{it}$ and covariates $x_{it}$

Rubin causal model (Holland, 1986):

- Treatment free outcomes $y_{it}^{(0)}$ (all $i$ and $t$)
- Outcomes under intervention $y_{it}^{(1)}$ ($i > n_1$ and $t > T_1$)
Notation & causal framework

Notation:
- $n$ units (indexed by $i$): $n_1$ controls and $n_2$ treated
- $T$ times (indexed by $t$): $T_1$ pre- and $T_2$ post-intervention
- Outcome $y_{it}$ and covariates $x_{it}$

Rubin causal model (Holland, 1986):
- Treatment free outcomes $y_{it}^{(0)}$ (all $i$ and $t$)
- Outcomes under intervention $y_{it}^{(1)}$ ($i > n_1$ and $t > T_1$)
- Hence, observed data are: $y_{it} = \begin{cases} y_{it}^{(1)}, & i > n_1 \text{ and } t > T_1 \\ y_{it}^{(0)}, & \text{otherwise} \end{cases}$
Notation & causal framework

Notation:
- $n$ units (indexed by $i$): $n_1$ controls and $n_2$ treated
- $T$ times (indexed by $t$): $T_1$ pre- and $T_2$ post-intervention
- Outcome $y_{it}$ and covariates $x_{it}$

Rubin causal model (Holland, 1986):
- Treatment free outcomes $y_{it}^{(0)}$ (all $i$ and $t$)
- Outcomes under intervention $y_{it}^{(1)}$ ($i > n_1$ and $t > T_1$)
- Hence, observed data are: $y_{it} = \begin{cases} y_{it}^{(1)}, & i > n_1 \text{ and } t > T_1 \\ y_{it}^{(0)}, & \text{otherwise} \end{cases}$
- Causal effects $\theta_{it} = y_{it}^{(1)} - y_{it}^{(0)}$ ($i > n_1$ and $t > T_1$)
- Some untestable assumptions are required to estimate counterfactuals $\hat{y}_{it}^{(0)}$ ($i > n_1$ and $t > T_1$) from observed data
Causal framework: illustration

\[ y_{1t}^{(0)} \]

\[ y_{n1+1,t}^{(0)} \]

\[ y_{1t}^{(1)} \]

Control
- Treated (no intervention)
- Treated (intervention)
Causal framework: illustration
Factor analysis

The **factor analysis** (FA) model assumes that for all $i$ and $t$

\[
y_{it}^{(0)} = x_{it}^{\top} \hat{\beta} + \lambda_{i}^{\top} \hat{f}_{t} + \epsilon_{it}
\]

**Coefficients** $\beta \sim N(d(0,10^{3}I))$

**Loadings** $\lambda_{i} \sim N(p(0,I))$ for all units $i$ (latent)

**Factors** $f_{t} \sim N(p(0,I))$ for all times $t$ (latent)

**Errors** $\epsilon_{it} \sim N(0,\psi_{i}^{2})$, where $\psi_{i}^{2} \sim IG(0.01,0.01)$

**Comments:**

- If $\lambda_{i}$ is associated with intervention assignment then it is an unobserved confounder
- $f_{t}$ can represent an underlying (e.g. environmental) shock
- For $t>T_{1}$ and $i>n_{1}$, estimate $\hat{y}_{it}^{(0)} = x_{it}^{\top} \hat{\beta} + \hat{\lambda}_{i}^{\top} \hat{f}_{t}$
Factor analysis

The factor analysis (FA) model assumes that for all $i$ and $t$

$$y_{it}^{(0)} = \mathbf{x}_{it}^\top \beta$$

- Coefficients $\beta \sim N_d (\mathbf{0}, 10^3 \mathbf{I})$
Factor analysis

The factor analysis (FA) model assumes that for all $i$ and $t$

$$y_{it}^{(0)} = x_{it}^\top \beta + \lambda_i^\top f_t$$

- Coefficients $\beta \sim N_d(0, 10^3 I)$
- **Loadings** $\lambda_i \sim N_p(0, I)$ for all units $i$ (latent)
- **Factors** $f_t \sim N_p(0, I)$ for all times $t$ (latent)
Factor analysis

The **factor analysis** (FA) model assumes that for all \( i \) and \( t \)

\[
y_{it}^{(0)} = x_{it}^\top \beta + \lambda_i^\top f_t + \epsilon_{it}
\]

- Coefficients \( \beta \sim \mathcal{N}_d (0, 10^3 I) \)
- **Loadings** \( \lambda_i \sim \mathcal{N}_p (0, I) \) for all units \( i \) (latent)
- **Factors** \( f_t \sim \mathcal{N}_p (0, I) \) for all times \( t \) (latent)
- Errors \( \epsilon_{it} \sim \mathcal{N} (0, \psi_i^2) \), where \( \psi_i^2 \sim \mathcal{IG} (0.01, 0.01) \)

Comments:
- If \( \lambda_i \) is associated with intervention assignment then it is an unobserved confounder
- \( f_t \) can represent an underlying (e.g. environmental) shock
- For \( t>T_1 \) and \( i>n_1 \), estimate \( \hat{y}_{it}^{(0)} = x_{it}^\top \hat{\beta} + \lambda_i^\top \hat{f}_t + \hat{\epsilon}_{it} \)
Factor analysis

The factor analysis (FA) model assumes that for all \( i \) and \( t \)

\[
y^{(0)}_{it} = \mathbf{x}_{it}^\top \beta + \mathbf{\lambda}_{i}^\top \mathbf{f}_t + \varepsilon_{it}
\]

- Coefficients \( \beta \sim N_d (0, 10^3 \mathbf{I}) \)
- **Loadings** \( \mathbf{\lambda}_i \sim N_p (0, \mathbf{I}) \) for all units \( i \) (latent)
- **Factors** \( \mathbf{f}_t \sim N_p (0, \mathbf{I}) \) for all times \( t \) (latent)
- Errors \( \varepsilon_{it} \sim N (0, \psi_i^2) \), where \( \psi_i^2 \sim IG (0.01, 0.01) \)

Comments:

- If \( \mathbf{\lambda}_i \) is associated with intervention assignment then it is an unobserved confounder
- \( \mathbf{f}_t \) can represent an underlying (e.g. environmental) shock
- For \( t > T_1 \) and \( i > n_1 \), estimate \( \hat{y}_{it}^{(0)} = \mathbf{x}_{it} \hat{\beta} + \hat{\mathbf{\lambda}}_i^\top \hat{\mathbf{f}}_t \)
Difference-in-differences (DID)

- When
  1. $p = 2$
  2. $\lambda_i = (\kappa_i, 1)^\top$ for all $i$
  3. $f_t = (1, \mu_t)^\top$ for all $t$

- Then the FA model reduces to the difference-in-differences model (Angrist and Pischke, 2009)

$$y_{it}^{(0)} = x_{it}^\top \beta + \kappa_i + \mu_t + \varepsilon_{it},$$

- Stricter assumptions (parallel trends) compared to the FA model
  1. Effect of $\kappa_i$ constant over time
  2. Shocks $\mu_t$ affect all units in the same way
Estimation:

- Several methods proposed (Chan and Kwok, 2016; Gobillon and Magnac, 2016; Athey et al., 2017; Xu, 2017)
- Asymptotically \((T_1 \rightarrow \infty, n_1 \rightarrow \infty)\) unbiased estimates of \(\theta_{it}\)
- Factors/loadings generally not identifiable

Limitations:

- Choosing \(p\) is hard
- Does not account for temporal dependence in the time-series
- Can be applied to a single outcome
FA continued

Estimation:

- Several methods proposed (Chan and Kwok, 2016; Gobillon and Magnac, 2016; Athey et al., 2017; Xu, 2017)
- Asymptotically ($T_1 \to \infty, n_1 \to \infty$) unbiased estimates of $\theta_{it}$
- Factors/loadings generally **not identifiable**

Limitations:

- Choosing $p$ is hard
- Does not account for temporal dependence in the time-series
- Can be applied to a single outcome
The MVFA+AR model

- The multivariate FA (De Vito et al., 2018, MVFA) assumes that:

\[ y_{itk}^{(0)} = x_{it}^\top \beta_k + \lambda_{ik} f_{tk} + \gamma_i^\top s_{tk} + \varepsilon_{itk}, \]

where \( k \) indexes an outcome

- For each \( j = 1, \ldots, p \), we assume factors are AR(1) i.e.

\[ f_{tj} = \rho_j f_{t-1,j} + \eta_{tj}, \]

where \( \rho_j \sim \text{Uniform}(-1, 1) \) and \( \eta_{tj} \sim \mathcal{N}(0, 1) \). Analogous prior for \( s_{tk} \)

- Multiplicative Gamma process shrinkage prior on \( \lambda_{ik} \) and \( \gamma_i \) to account for uncertainty in \( p \) (Bhattacharya and Dunson, 2011)
MVFA+AR: simulation studies

- effect magnitude ($X$) against power to detect ($Y$)

**Setup VII**
- $T_1 = 10$
- $n_1 = 30$

**Setup III**
- $T_1 = 40$
- $n_1 = 5$
Synthetic controls: intuition

- Recall: California tobacco dataset
Synthetic controls: intuition

Controls shown: Connecticut, Montana, Nevada and New Mexico
Synthetic controls: intuition

\[ CA_t = 0.18 \cdot CT_t + 0.22 \cdot MT_t + 0.21 \cdot NV_t + 0.39 \cdot NM_t \]
Synthetic controls: intuition

\[ CA_t = 0.18 \cdot CT_t + 0.22 \cdot MT_t + 0.21 \cdot NV_t + 0.39 \cdot NM_t \]
Synthetic controls: intuition

\[ CA_t = 0.18 \cdot CT_t + 0.22 \cdot MT_t + 0.21 \cdot NV_t + 0.39 \cdot NM_t \]
The synthetic control method (SCM) (Abadie et al., 2010)

- For $i = 1, \ldots, n_1 + 1$ (single treated), let $y_i = (y_{i1}, \ldots, y_{iT_1})^\top$

- The SCM finds the $w_1, \ldots, w_{n_1}$ that minimise

$$
\left( y_{n_1 + 1} - \sum_{i=1}^{n_1} w_i y_i \right)^\top V \left( y_{n_1 + 1} - \sum_{i=1}^{n_1} w_i y_i \right),
$$

subject to $\sum_{i=1}^{n_1} w_1 = 1$ and $w_i \geq 0$

- For $t > T_1$, the predicted counterfactual is

$$\hat{y}_{n_1 + 1, t}^{(0)} = \sum_{i=1}^{n_1} w_i y_{it}$$
SCM continued

- “...arguably the most important innovation in the policy evaluation literature in the last 15 years...” (Athey and Imbens, 2017)

- Links to the matching literature

- No assumptions regarding the data generating mechanism

- Asymptotically \((T_1 \to \infty)\) unbiased estimates of \(\theta_{it}\) when the data are generated from the FA model

- Has led to a flexible class of estimators (Hsiao et al., 2012; Brodersen et al., 2015; Doudchenko and Imbens, 2016)

\[
\hat{y}_{n_1+1,t}^{(0)} = \hat{\beta}_0 + \sum_{i=1}^{n_1} \hat{\beta}_i y_{it}
\]
SCM extensions

- **Inference**: Chernozhukov *et al.* (2017); Amjad *et al.* (2018)

- **Multiple treated units**: Acemoglu *et al.* (2016); Kreif *et al.* (2016)

- **Matching high-order characteristics**: Hazlett and Xu (2018)

- **Sparsity of weights**: Brodersen *et al.* (2015); Doudchenko and Imbens (2016); Li and Bell (2017)

- **Multivariate outcomes**: Robbins *et al.* (2017)
The outcome of the treated unit is sum of 3 components: time-series, regression and error

For example, for $t = 1, \ldots, T$, assume that

\[
\begin{align*}
    y_{n_1+1,t}^{(0)} &= \beta_0 t + \sum_{i=1}^{n_1} \beta_i y_{it} + \epsilon_t \\
    \beta_{0,t_1+1} &= \beta_{0,t} + \delta_t + \eta_t \\
    \delta_{t+1} &= \delta_t + \zeta_t,
\end{align*}
\]

where $\epsilon_t \sim N(0, \sigma^2_\epsilon)$, $\eta_t \sim N(0, \sigma^2_\eta)$ and $\zeta_t \sim N(0, \sigma^2_\zeta)$

Spike-and-slab prior on $\beta_i$ for sparsity

More flexible time-series models can be adopted, e.g. seasonal
CIM simulation study

Interested in potential of CIM for use in epidemiological studies:

- Small $T_1$
- $n_1 > T_1$
- Outcomes are not known exactly (measurement error)

Simulation study questions:

1. Sufficient power to detect an effect?
2. Can spike-and-slab prior select the controls?
3. Measurement error: implications?
CIM simulation study results

- Data generated based on HCV dataset
- % prevalence decrease ($X$) against power ($Y$)

(A) Sensitivity to # controls
(B) Sensitivity to # participants

<table>
<thead>
<tr>
<th>Cohort recruited</th>
<th>Small</th>
<th>Moderate</th>
<th>Infinite</th>
</tr>
</thead>
<tbody>
<tr>
<td># additional controls</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Errors-in-variables CIM

- We account for the measurement error with a hierarchical errors-in-variables synthetic control model

- In the first (latent) level we assume

\[ y_{n_1+1,t}^{(0)} = \beta_0 + \sum_{i=1}^{n_1} \beta_i y_{it} + \varepsilon_t, \]

where \( \varepsilon_t \sim N(0, \sigma^2) \)

- In the second (observation) level we assume

\[ k_{it}^{(0)} \sim \text{Binomial} \left( N_{it}, \frac{\exp \left( y_{it}^{(0)} \right)}{1 + \exp \left( y_{it}^{(0)} \right)} \right), \]

where \( N_{it} \) is the sample size and \( k_{it}^{(0)} \) is the number of infected individuals
Connections

- In FA, we marginally have that for all $t$:

$$\text{Cov} \left( y_{t}^{(0)} \right) = \Lambda \Lambda^\top + \text{diag} \{ \psi_1^2, \ldots, \psi_n^2 \},$$

where $\Lambda$ is the $n \times p$ matrix with rows $\lambda_i$. 

Connections

- In FA, we marginally have that for all $t$:

$$\text{Cov} \left( y_t^{(0)} \right) = \Lambda \Lambda^\top + \text{diag} \left\{ \psi_1^2, \ldots, \psi_n^2 \right\},$$

where $\Lambda$ is the $n \times p$ matrix with rows $\lambda_i$.

- In synthetic control-type approaches, for $i > n_1$:

$$y_{it}^{(0)} = \beta_0 + \sum_{i=1}^{n_1} \beta_i y_{it} + \varepsilon_{it}$$

- Hence, both classes of methods exploit linear relationships between treated and control units to predict counterfactuals.

- These relationships remain constant over time.
Recommendations (1/2)

- Generally, none of the methods is superior
- Gobillon and Magnac (2016); O’Neill et al. (2016); Kinn (2018) perform extensive simulation studies
- Choice of method should be driven by data characteristics

Factor analysis

- Requires at least moderate $n_1$ and $T_1$
- Useful when covariates are highly predictive of outcome

Synthetic-control type approaches

- Best suited when $T_1$ is large
- When $n_1 \approx T_1$, some regularization is required
Recommendations (2/2)

Diagnostics

- Goodness-of-fit in the pre-intervention period (and post-intervention for controls)
- Overlap of support (both covariates and outcomes) between treated and controls to avoid extrapolation biases
- Identify controls with similar loadings to the treated units (FA) or large weights (SCM)

Sensitivity analyses

- Implement several methods to check for conflicting results
- Implement one method with various specifications (e.g. by removing some of the controls)
Open problems

1. Account for **geographical locations**
   - Units with spatial proximity tend to be correlated
   - Can improve efficiency of the causal estimates

2. Consider **non-linear** relationships
   - Machine learning techniques, e.g. Gaussian process regression
   - Theoretical properties

3. Relax the assumption that the intervention does not affect control units
Acknowledgements

For more details:

▶ *Statistical Science* 34(3), 486-503, 2019

Collaborators:

▶ BSU: D de Angelis, SR Seaman, AM Presanis
▶ External: M Hickman (Bristol), S Montagna (Torino), N Martin (UCSD), A Charlett (PHE)

Funding:

▶ NIHR Health Protection Unit on Evaluation of Interventions
Acknowledgements

For more details:

▶ *Statistical Science* 34(3), 486-503, 2019

Collaborators:

▶ BSU: D de Angelis, SR Seaman, AM Presanis
▶ External: M Hickman (Bristol), S Montagna (Torino), N Martin (UCSD), A Charlett (PHE)

Funding:

▶ NIHR Health Protection Unit on Evaluation of Interventions

THANK YOU 😄


Choosing the number of factors

- Multiplicative Gamma shrinkage process prior on $\lambda_i$
- Proposed by Bhattacharya and Dunson (2011)
- Let $\lambda_i = (\lambda_{i1}, \lambda_{i2}, \ldots)^\top$ be of dimension $\infty$ for each $i$
Choosing the number of factors

- **Multiplicative Gamma shrinkage process** prior on $\lambda_i$
- Proposed by Bhattacharya and Dunson (2011)
- Let $\lambda_i = (\lambda_{i1}, \lambda_{i2}, \ldots)^\top$ be of dimension $\infty$ for each $i$
- The MGPS assumes that

$$\lambda_{ij} \sim N \left(0, \frac{1}{\phi_{ij}\tau_j}\right),$$

where $\tau_j = \prod_{\ell=1}^j \delta_j$

- Hyperpriors: $\phi_{ij} \sim \text{Gamma} \left(\frac{3}{2}, \frac{3}{2}\right)$, $\delta_1 \sim \text{Gamma} \left(2.1, 1\right)$ and $\delta_j \sim \text{Gamma} \left(3.1, 1\right) (j > 1)$
- Elements of $\lambda_i$ will progressively shrink to zero
Alcohol licensing: evidence for common factors

- Factor index $j$ ($X$) against mean posterior of $\sum_{i=1}^{n} |\lambda_{ij}|$ ($Y$)
- Left/right: outcome-specific/shared loadings