A Quasi Synthetic Control Method for Nonlinear Models With High-Dimensional Covariates

Ying Fang

School of Economics, Xiamen University

Present to School of Economics, Nankai University, Aug. 19, 2024

Co-Authors: Zongwu Cai, Ming Lin and Zixuan Wu

1. Review of the Synthetic Control Method
2. A Quasi Synthetic Control Method for Nonlinear Models
   - Model Setup
   - Identification and Estimation Procedures
   - Asymptotic Property
   - Bootstrap Inference
3. Variable Selection
   - QSCM With a Diverging Number of Covariates
   - QSCM With Ultra-high Dimensional Covariates
4. Monte Carlo Simulations
   - Evaluating QSCM and the Bootstrap Procedure
   - Evaluating QSCM with Variable Selection
5. Empirical Example
6. Conclusion Remarks

Review of the Synthetic Control Method

Review of the Synthetic Control Method

The Synthetic Control Method

- The synthetic control method (SCM), proposed by Abadie and Gardeazabal (2003, AER), is a powerful tool for estimating average treatment effects (ATE), and gains increasing popularity in fields such as statistics, economics, political science, and marketing.

"The synthetic control approach ... is arguably the most important innovation in the policy evaluation literature in the last 15 years."

—Athey and Imbens (2017, JEP)

Setting

- We code the treatment status of unit i using the binary variable D_i, so $D_i = 1$ if i is treated and $D_i = 0$ otherwise.
- We adopt the potential outcomes framework proposed by Rubin (1974, JEP). Let Y_{1i} and Y_{0i} be random variables representing potential outcomes under treatment and without treatment, respectively, for unit i, and the realized outcome is defined as $Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}$.
- Let X_i be a $(d \times 1)$ vector of pretreatment predictors.
- Then, we observe $(Y_i, X_i) = (Y_{1i}, X_i)$ for n_1 treated units and $(Y_i, X_i) = (Y_{0i}, X_i)$ for n_0 control units. Combining these observables, we obtain the pooled dataset, $\{(Y_i, D_i, X_i)\}_{i=1}^{n}$, with $n = n_0 + n_1$. For simplicity, we reorder the observations so that the n_0 control units come first.

The Synthetic Control Method

- The quantity of interest is **the treatment effect on the treated units**, $\Delta_i = Y_{1i} - Y_{0i}$, for $i = n_0 + 1, \dots, n$, and the average treatment effect is given by

$$\Delta = \frac{1}{n_1} \sum_{i=n_0+1}^{n} (Y_{1i} - Y_{0i}).$$

- The difficulty in estimating Δ is that $\boxed{\{Y_{0i}\}_{i=n_0+1}^{n}}$ are not observable, which has been a key issue for researchers since the paper by Rubin (1974).

- Now, the SCM solves this problem by assuming that a combination of control units may approximate the characteristics of the treated unit well, and this combination can be used to estimate $\{Y_{0i}\}_{i=n_0+1}^{n}$.

The Synthetic Control Method

Concretely, for each treated unit $i = n_0 + 1, \ldots, n$, we can construct a synthetic control, which is a combination of control units represented by a $n_0 \times 1$ vector of weights $W_i^* = (W_{i,1}^*, \ldots, W_{i,n_0}^*)'$. Given a set of weights, W_i^*, the synthetic control estimator of Y_{0i} and Δ can be written as

$$\hat{Y}_{0i} = \sum_{j=1}^{n_0} W_{i,j}^* Y_j \tag{1}$$

and

$$\hat{\Delta} = \frac{1}{n_1} \sum_{i=n_0+1}^{n} \left(Y_i - \sum_{j=1}^{n_0} W_{i,j}^* Y_j \right) = \frac{1}{n_1} \sum_{i=n_0+1}^{n} Y_i - \boxed{\frac{1}{n_0} \sum_{j=1}^{n_0} a_j^* Y_j},$$

where $a_j^* = n_0 \sum_{i=n_0+1}^{n} W_{i,j}^* / n_1$.

Question: How to choose the weights $\{W_{i,j}^*\}$, $n_1 \times n_0$ parameters?

The Synthetic Control Method

The SCM proposes to choosing $W_{i,j}^*$ such that the synthetic control resembles the corresponding treated unit i in terms of the values of the predictors of the outcome variable. Mathematically speaking, the SCM seeks the solution to the following question:

$$\min_{W_i \in \mathbb{R}^{n_0}} \left(X_i - \sum_{j=1}^{n_0} W_{i,j} X_j \right)^\top V \left(X_i - \sum_{j=1}^{n_0} W_{i,j} X_j \right)$$

$$\text{s.t. } W_{i,1} \geq 0, \ldots, W_{i,n_0} \geq 0, \quad \text{and} \quad \sum_{j=1}^{n_0} W_{i,j} = 1, \qquad (2)$$

where V is a $d \times d$ matrix with the elements on the diagonal being all positive and reflecting the relative importance for each predictor.

Remarks

- The SCM has been widely applied in empirical research in economics and other disciplines. The paper by Abadie (2021, JEL) presents a thorough discussion on the advantages and the feasibility of the SCM.
- In the SCM, the weights are restricted to be non-negative and sum to one, which is called as the convex hull constraint. This constraint might not be needed nor necessarily satisfied in many cases. Several modifications have been proposed to relaxing this constraint (see, e.g., Doudchenko and Imbens (2016, WP), Li (2020, JASA), Kellogg et al. (2021, JASA)).

Remarks

- For more econometric/statistical theories and inferences on the SCM and its variants, the reader is referred to the paper by Li (2020) and the special section in *Journal of The American Statistical Association* in the last issue of 2021 on synthetic control methods edited by Abadie and Cattaneo (2021, JASA), which covers some new research directions on synthetic control estimation and inference, including the following four aspects:
  1. factor models and matrix completion methods proposed by Agarwal et al. (2021), Athey et al. (2021) and Bai and Ng (2021),
  2. time series analysis approach studied by Ferman (2021) and Masini and Medeiros (2021),
  3. extensions, modifications and generalizations investigated by Abadie and L'Hour (2021), Ben-Michael, Feller and Rothstein (2021) and Kellogg et al. (2021), and
  4. uncertainty quantification and inference explored by Cattaneo, Feng and Titiunik (2021), Chernozhukov, Wüthrich and Zhu (2021), and Shaikh and Toulis (2021).

Remarks

- It is easy to see from (1) that the SCM assumes implicitly that the prediction function of Y_{0i} given X_i is a linear or close to linear function of X_i, which might not be satisfied in real applications.
- Also, as pointed out by Abadie (2021), the optimization problem in (2) might not have a unique solution. Indeed, there are an infinite number of solutions.
- Furthermore, it is important to note that for any particular data set there are not ex ante guarantees on the size of the differences $X_i - \sum_{j=1}^{n_0} W_{i,j} X_j$ in (2). When these differences are large, the papers by Abadie, Diamond and Hainmueller (2010, JASA) and Abadie (2021) recommend against the use of synthetic controls because of the potential for substantial biases.

Remarks

- When n_0 is large, the computing burden to find the "optimal" weights in (2) is troublesome. To see this issue, in our empirical study, we will report the computing time based on our computing facility.

- In addition to the above computing issue, sparsities might exist among $\{W_{i,j}\}_{j=1}^{n_0}$. To address these challenges, the paper by Abadie and L'Hour (2021) propose a synthetic control estimator, termed as penalized synthetic control method (Pen-SCM), that penalizes the pairwise discrepancies between the characteristics of the treated units and of the corresponding synthetic control units. That is to add the following penalty term into (2)

$$\lambda \sum_{j=1}^{n_0} W_{i,j} ||X_i - X_j||^2,$$

  which is different from the conventional LASSO type methods imposing the penalty on parameters.

QSCM

Quasi Synthetic Control Method

Model Setup

Assume we observe n units, some of which are exposed to the treatment or intervention of our interest. For each unit $i = 1, \ldots, n$, denote

- $D_i = \{0, 1\}$ as the binary treatment variable
- Y_{1i} and Y_{0i} as the potential outcomes under treatment and no treatment, respectively
- $X_i \in \mathbb{R}^d$ as the $d \times 1$ vector of pre-treatment predictors of Y_{0i}[1]

Under the potential outcomes framework, the observed outcome Y_i satisfies $Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}$. Therefore, we obtain a pooled data set $\{Y_i, D_i, X_i\}_{i=1}^n$.

[1] d might be very large.

Model Setup

Denote n_1 and n_0 as the number of the treated observations and the untreated observations, respectively. For simplicity, we reorder the data so that the n_0 untreated observations come first.

The quantity of our interest is the average treatment effect on the treated (ATT):

$$\Delta = E(\Delta_i) = E(Y_{1i} - Y_{0i}), \; i = n_0 + 1, \ldots, n. \tag{3}$$

Still, the difficulty in estimating Δ_i and Δ is that Y_{0i} is not observable for $i = n_0 + 1, \ldots, n$.

Model Setup

- To estimate the unobservables $\{Y_{0i}\}_{i=n_0+1}^{n}$, we assume that the prediction function based on the conditional expectation of Y_{0i} given X_i, denoted by $m(x) = E(Y_{0i}|X_i = x)$, is in an index form as $m(x) = m(\beta_0^\top x) = m(z)$, where $z = \beta_0^\top x \in \mathbb{R}$.[2]

- Then, for $i = n_0 + 1, \ldots, n$,

$$E(Y_{0i}) = E\left[E(Y_{0i}|X_i)\right] = E\left[E(Y_{0i}|Z_i)\right]$$

where $Z_i = \beta_0^\top X_i$ for a given β_0, so that the estimation of $m(z)$ is one-dimensional, and the so-called *curse of dimensionality* in a nonparametric smoothing can be avoided.

[2] This covers a linear model as an special case. Of course, when d is small, one can estimate directly $m(x)$ by using a nonparametric method. Therefore, this case is much easier.

Identification

- From the above discussion, our method needs to identify both the unknown index vector β_0 and the function $m(z)$. In fact, it is a two-step procedure.

- Clearly, given $z_0 = \beta^\top x$, the function $m(z)$ can be identified nonparametrically under certain assumptions.

- To identify β_0, we introduce the following assumption.

Identification of the First Step

Denote $m_c(x) = E\left[Y_{0j} \,|\, X_j = x\right]$ for $j = 1, \ldots, n_0$ and $m_t(x) = E\left[Y_{0j} \,|\, X_j = x\right]$ for $i = n_0 + 1, \ldots, n$.

Assumption 1

Assume that $m_c(x) = m_t(x) = m(z)$, where $z = \beta_0^\top x$ and $\beta_0 \in \mathbb{B}$, where $\mathbb{B} = \{\beta \in \mathbb{R}^d : \beta_1 > 0, ||\beta||^2 = \sum_{k=1}^d \beta_k^2 = 1\}$. Furthermore, assume that the second order derivative of $m(z)$ is continuous.

By Assumption 1, we can identify β_0 using data $\{Y_j, X_j\}_{j=1}^{n_0}$.

Estimation of the First Step: A Brief Review

As introduced before, $E(Y_{0i}|X_i = x) = m(\beta^\top x)$ is identical to the well-known single index model (SIM), which assumes $Y_{0i} = m(\beta^\top X_i) + \varepsilon_i$, where $E(\varepsilon_i|X_i) = 0$ and β is called the parametric index vector.

- Estimation of β is very attractive both in theory and practice.
- The papers by Powell et al. (1989, ECTA) and Hädle and Stoker (1989, JASA) propose the average derivative estimation (ADE) method, which involves estimating a high-dimensional density function and its derivative.
- The paper by Ichimura (1993, JoE) proposes the semiparametric least squares (SLS) estimation. But the minimization is very difficult to implement.
- The paper by Xia et al. (2002, JRSSB) proposes the minimum average variance estimation (MAVE) method for the dimension reduction problem, which can be applied to the SIM directly.

Estimation of the First Step: the MAVE Method

Under the least squares loss,

$$\beta_0 = \arg\min_{\tilde{\beta}\in\mathbb{R}^d} E[Y - E(Y|\tilde{\beta}^\top X)]^2. \tag{4}$$

In our setting, we have data $\{Y_j, X_j\}_{j=1}^{n_0}$. Motivated by the local linear smoothing technique, the sample analogue of (4) can be written as

$$\begin{aligned} \hat{\beta}_{\mathsf{MAVE}} &= \arg\min_{\tilde{\beta}\in\mathbb{R}^d} \sum_{j=1}^{n_0} \{\min_{a_j,b_j} \sum_{i=1}^{n_0} [Y_i - a_j - b_j\tilde{\beta}^\top (X_i - X_j)]^2 w_{ij}\} \\ &= \arg\min_{\substack{\tilde{\beta}\in\mathbb{R}^d \\ a_j,b_j}} \sum_{j=1}^{n_0}\sum_{i=1}^{n_0} [Y_i - a_j - b_j\tilde{\beta}^\top (X_i - X_j)]^2 w_{ij} \end{aligned} \tag{5}$$

where $a_j = m(\tilde{\beta}^\top X_j)$, $b_j = \partial m(u)/\partial u|_{u=\tilde{\beta}^\top X_j}$, $w_{ij} = K_h(\tilde{\beta}^\top (X_i - X_j))$ with $K_h(v) = K(v/h)/h$ and $K(\cdot)$ being a kernel function as well as h being the bandwidth.

Estimation of the First Step: the MAVE Method

- The MAVE method solves (5) iteratively. First, given $\tilde{\beta}$, optimize (5) with respect to a_j and b_j, and then, given a_j and b_j, optimize (5) with respect to $\tilde{\beta}$.
- During the iteration, the weights w_{ij} are updated simultaneously accroding to the latest value of $\tilde{\beta}$.
- The paper by Xia (2006, ET) derives the asymptotic distribution of the estimator of β_0 based on the MAVE, and shows that it can achieve the information lower bound in the semiparametric sense.

The Second Step Estimation

Under Assumption 1, for any z, we can also derive the Nadaraya-Watson estimator of $m(z)$:

$$\hat{m}(z) = \sum_{j=1}^{n_0} \hat{m}(z) = \sum_{j=1}^{n_0} c_{j,h}(z) Y_j, \tag{6}$$

where $c_{j,h}(z) = K_h(Z_j - z)/\sum_{l=1}^{n_0} K_h(Z_l - z)$, $K_h(u) = K(u/h)/h$, and $K(u)$ is a kernel function, and h is the bandwidth.

Consequently, we can derive an infeasible[3] estimator of Y_{0i}:

$$\tilde{Y}_{0i} = \hat{m}(Z_i) = \sum_{j=1}^{n_0} c_{j,h}(Z_i) Y_j, \; i = n_0 + 1, \ldots, n. \tag{7}$$

[3] This estimator is infeasible because it is based on the unknown quantities $\{Z_j\}_{j=1}^{n_0}$

The Second Step Estimation

Then, the infeasible estimator of Δ, $\tilde{\Delta}$ is given by

$$\tilde{\Delta} = \frac{1}{n_1}\sum_{i=n_0+1}^{n}\left[Y_i - \sum_{j=1}^{n_0} c_{j,h}(Z_i)Y_j\right] = \frac{1}{n_1}\sum_{i=n_0+1}^{n} Y_i - \boxed{\frac{1}{n_0}\sum_{j=1}^{n_0} a_{j,h}Y_j}, \quad (8)$$

where $a_{j,h} = a_h(Z_j)$ and

$$a_h(z) = \frac{1}{n_1}\sum_{i=n_0+1}^{n} K_h(Z_i - z)\left[\frac{1}{n_0}\sum_{l=1}^{n_0} K_h(Z_i - Z_l)\right]^{-1}.$$

Clearly, (8) is similar to (1) and $a_{j,h}$ in (8) is similar to a_j^* in (1). Therefore, our method is called quasi synthetic control method (QSCM). Note that the key difference between SCM and QSCM is that the SCM is only valid for linear models but the QSCM can be accommodate nonlinear models.

Summary of the Estimation Procedure

We summarize our estimation procedure based on above discussion.

- Step 1. Using data $\{Y_j, X_j\}_{j=1}^{n_0}$, estimate the index vector β_0 by the MAVE method, and denote the estimator as $\hat{\beta}$.

- Step 2. Set $\hat{Z}_j = \hat{\beta}^\top X_j$ for $j = 1, \ldots, n_0$ and $\hat{Z}_i = \hat{\beta}^\top X_i$ for $i = n_0 + 1, \ldots, n$.

- Step 3. Plug $\{\hat{Z}_j\}_{j=1}^{n_0}$ and $\{\hat{Z}_i\}_{i=n_0+1}^{n}$ into (8), and compute the feasible estimator of Δ as

$$\hat{\Delta} = \frac{1}{n_1} \sum_{i=n_0+1}^{n} \left[Y_{1i} - \sum_{j=1}^{n_0} \hat{c}_{j,h}(\hat{Z}_i) Y_j \right] = \frac{1}{n_1} \sum_{i=n_0+1}^{n} Y_{1i} - \frac{1}{n_0} \sum_{j=1}^{n_0} \hat{a}_{j,h} Y_j, \quad (9)$$

where $\hat{a}_{j,h} = \hat{a}_h(\hat{Z}_j) = \frac{1}{n_1} \sum_{i=n_0+1}^{n} K_h(\hat{Z}_i - \hat{Z}_j) [\frac{1}{n_0} \sum_{l=1}^{n_0} K_h(\hat{Z}_i - \hat{Z}_l)]^{-1}$.

Notations

To derive the asymptotic property of the proposed estimator in (9), some assumptions are needed. Before presenting these assumptions, we first introduce some notations.

- Let $f_c(z)$ be the density of Z_j for $j = 1, \ldots, n_0$ and $f_t(z)$ be the density of Z_i for $i = n_0 + 1, \ldots, n$.
- Define $\mathcal{C}_1$ to be the support of Z_j for $j = 1, \ldots, n_0$ and $\mathcal{C}_2$ to be the support of Z_i for $i = n_0 + 1, \ldots, n$.

Assumptions

Assumption 2

$\{Y_{0j}, Y_{1j}, X_j\}_{j=1}^{n_0}$ for the control group and $\{Y_{0i}, Y_{1i}, X_i\}_{i=n_0+1}^{n}$ for the treated group are independent and identically distributed, respectively. Assume that $E(|Y_{di}|^s) < \infty$ for $d = 0, 1$ and some $s > 2$. We also assume that $\mathcal{C}_2 \subseteq \mathcal{C}_1$ and $f_c(z) > m_1 > 0$ for $z \in \mathcal{C}_2$.

Assumption 3

Assume that the second order of derivative of $r(z)$ is bounded, where $r(z) = f_t(z)/f_c(z)$, the ratio function to characterize the distributional changes of the single index between the treated and control units.[a]

[a] Indeed, $r(z)$ is interpreted as “acceptance probability" in rejection sampling instead of “importance re-weighting", or covariate shift in the machine learning literature, especially in marketing science.

Assumptions

Assumption 4

The kernel function $K(\cdot)$ is symmetric, bounded and positive. Further assume that the first derivative of $K(\cdot)$ is continuous.

Assumption 5

Assume that $n_0\, h^2 \to \infty$, $n_0\, h^4 \to 0$, and $n_1/n_0 \to \eta$ as $n_0 \to \infty$, where $0 < \eta < \infty$.

Assumption 6

Assume that for any estimate of β_0, $\hat{\beta}$ admits the following expression

$$\sqrt{n_0}\left(\hat{\beta} - \beta_0\right) = \frac{1}{\sqrt{n_0}} \sum_{j=1}^{n_0} \phi(X_j, Y_j) + o_p(1) \xrightarrow{d} N(0, \Sigma_{\beta_0}) \qquad (10)$$

for some function $\phi(\cdot)$ with variance $\Sigma_{\beta_0} = Var(\phi(X_j, Y_j))$ for $j = 1, \ldots, n_0$.

Asymptotic Property

Let $\varepsilon_j = Y_{0j} - E(Y_{0j} \mid X_j)$ for $j = 1, \ldots, n_0$. Define $\sigma_1^2 = \mathrm{Var}[Y_{1i} - m(Z_i)]$ for $i = n_0 + 1, \ldots, n$, $\sigma_2^2 = \mathrm{Var}[r(Z_j)\varepsilon_j]$ for $j = 1, \ldots, n_0$, and $\sigma_3^2 = \delta_a^\top \Sigma_{\beta_0} \delta_a$ with $\delta_a = E\left[m'(Z_i)X_i^\top\right]$ for $i = n_0 + 1, \ldots, n$, where $m'(z)$ is the first order derivative of $m(z)$, and Σ_{β_0} is given in Assumption 6. Define $\Sigma_{23} =$Cov$(\phi(X_j, Y_j), r(Z_j)\varepsilon_j)$.

Theorem 1

Under Assumptions 1 - 6, we have

$$\sqrt{n_1}\left(\hat{\Delta} - \Delta\right) \xrightarrow{d} N(0, \sigma_\Delta^2),$$

where $\boxed{\sigma_\Delta^2 = \sigma_1^2 + \eta\ \left[\sigma_2^2 + \sigma_3^2 + 2\,\delta_a^\top \Sigma_{23}\right]}$.

A Remark on Asymptotic Property

It follows from Theorem 1 that the asymptotic variance consists of four terms.

- The first term in σ^2_Δ stands for the variance of $Y_{1i} - m(Z_i)$.
- The second term is for charactering the variation for estimating Y_{0i}.
- The third term σ_3^2 is the variation carried over from the estimation of β.
- The last term depicts the correlation between the first step and the second step.

This is typical for a two-stage procedure as addressed in Cai, Das, Wu and Xiong (2006, JoE). Also, one can see that obtaining a consistent estimate of σ^2_Δ is not a straightforward task due to its complicated form of involving several terms. However, a Bootstrap procedure can overcome possibly this difficulty.

Bootstrap Procedure

To facilitate an easy inference, we propose the following (hybrid) Bootstrap procedure to estimate σ_Δ^2.

- Step 1. Given $\{Y_j, X_j\}_{j=1}^{n_0}$ and $\{Y_i, X_i\}_{i=n_0+1}^{n}$, estimate the treatment effect as $\hat{\Delta}$.
- Step 2. Generate the wild Bootstrap sample $\{(X_j, Y_j^*)\}_{j=1}^{n_0}$ of the control group, where $Y_j^* = \hat{m}(\hat{\beta}^\top X_j) + \varepsilon_j^*$ with $\hat{m}(\hat{\beta}^\top X_j) = \sum_{l=1}^{n_0} K_h(\hat{\beta}^\top X_j - \hat{\beta}^\top X_l) Y_l / \sum_{l=1}^{n_0} K_h(\hat{\beta}^\top X_j - \hat{\beta}^\top X_l)$, $\varepsilon_j^* = [Y_j - \hat{m}(\hat{\beta}^\top X_j)]\xi_j$, and $\{\xi_j\}_{j=1}^{n_0}$ being i.i.d. random disturbances with mean zero and unit variance.
- Step 3. Generate the nonparametric Bootstrap sample $\{(X_i^*, Y_i^*)\}_{i=n_0+1}^{n}$ of the treated group by drawing with replacement from the original dataset $\{(X_i, Y_i)\}_{i=n_0+1}^{n}$.

Bootstrap Procedure

- Step 4. Using the wild Bootstrap sample $\{(X_j, Y_j^*)\}_{j=1}^{n_0}$ to re-estimate the index parameter as $\hat{\beta}^*$. Set $\hat{Z}_j^* = X_j^\top \hat{\beta}^*$ for $j = 1, \ldots, n_0$ and $\hat{Z}_i^* = (X_i^*)^\top \hat{\beta}^*$ for $i = n_0 + 1, \ldots, n$. Then, obtain the quasi synthetic control estimator $\hat{\Delta}^*$ as

$$\hat{\Delta}^* = \frac{1}{n_1} \sum_{i=n_0+1}^{n} [Y_i^* - \sum_{j=1}^{n_0} \hat{c}_{j,h}^*(\hat{Z}_i^*) Y_j^*] = \frac{1}{n_1} \sum_{i=n_0+1}^{n} Y_i^* - \frac{1}{n_0} \sum_{j=1}^{n_0} \hat{a}_{j,h}^* Y_j^*,$$

where

$$\hat{a}_{j,h}^* = \hat{a}_h^*(\hat{Z}_j^*) = \frac{1}{n_1} \sum_{i=n_0+1}^{n} K_h(\hat{Z}_i^* - \hat{Z}_j^*) \left[\frac{1}{n_0} \sum_{l=1}^{n_0} K_h(\hat{Z}_i^* - \hat{Z}_l^*) \right]^{-1}.$$

- Step 5. Repeat steps 2 to 4 a large number of times, say, B times to obtain $\{\hat{\Delta}^{*(b)}\}_{b=1}^{B}$. Then σ_Δ^2 can be estimated as

$$\hat{\sigma}_\Delta^2 = n_1 \sum_{b=1}^{B} (\hat{\Delta}^{*(b)} - \hat{\Delta})^2 / (B - 1).$$

Bootstrap Theory

Theorem 2

Under the conditions imposed in Theorem 1, conditional on the original sample $\{X_j, Y_j\}_{j=1}^{n_0}$ *and* $\{X_i, Y_i\}_{i=n_0+1}^{n}$ *and in probability, one has*

$$\sqrt{n_1}(\hat{\Delta}^* - \hat{\Delta}) \xrightarrow{d} \mathrm{N}(0, \sigma_\Delta^2),$$

where σ_Δ^2 *is defined in Theorem 1.*

Penalized Approach

When the number of predictor variables is large, it is common that sparsity exists so that it is necessary to discriminate relevant variables from irrelevant variables, since the inclusion of irrelevant variables may harm estimation accuracy and model interpretability.

Generally, now we consider a $d_{n_0} \times 1$ vector of covariates X, which means that the dimension of the covariates changes with the sample size of the control group n_0. That is $d_0 = d_0(n_0) = O(n_0^{\gamma})$ for some $0 < \gamma < 1$, see Assumption 10 later on assumption on d_0 which depends on n_0.

For the ultra-dimensional case that $d_0 >> n$, say, $d_0 = O(\exp(n^{\xi}))$ for some $\xi > 0$, one need to use some screening approach first, such as the sure independence screening (SIR) method in Fan and Lv (2008, JRSSB), and then, use a penalized method.

Penalized Approach

Assume that the dimension of the covariates diverges with the sample size of the control group and denote it as d_{n_0}. Without loss of generality, we assume that the first s components of β_0 are non-zeros, *i.e.*, β_0 is partitioned to $\beta_{0,\mathcal{A}} = (\beta_{0,1}, \ldots, \beta_{0,s})^\top$ and $\beta_{0,\mathcal{A}^c} = (0, \ldots, 0)^\top$ with $d_{n_0} - s$ components, where $\mathcal{A} = \{1, \cdots, s\}$ and $\mathcal{A}^C = \{s+1, \cdots, d_{n_0}\}$.

To select the relevant covariates, we can add a penalty term to the least-squares-form loss function as

$$\sum_{j=1}^{n_0} [Y_j - \hat{m}(\beta^\top X_j)]^2 + n_0 \sum_{k=1}^{d_{n_0}} p_{\lambda_{n_0}}(|\beta_k|), \tag{11}$$

where $\beta = (\beta_1, \cdots, \beta_{d_{n_0}})^\top$, $\hat{m}(\cdot)$ is an estimate of the link function $m(\cdot)$, $p_{\lambda_{n_0}}(\cdot)$ denotes a penalty function and λ_{n_0} is the penalty parameter.

Penalized Approach

- For a given β, we can obtain $\hat{m}(\beta^\top X_j)$ using the local linear smoothing method. Specifically, we let

$$(\hat{a}_j,\,\hat{b}_j) = \arg\min_{a_j,\,b_j}\left\{\sum_{l=1}^{n_0}[Y_l - a_j - b_j(\beta^\top X_l - \beta^\top X_j)]^2 K_{h_1}(\beta^\top X_l - \beta^\top X_j)\right\}, \tag{12}$$

where $K_{h_1}(v) = K(v/h_1)/h_1$, $K(\cdot)$ is a kernel function and h_1 is the bandwidth. Then we have $\hat{m}(\beta^\top X_j) = \hat{a}_j$.

- For the penalty function, we choose the SCAD penalty and modify the objective function in (10) as

$$\hat{\beta}_{\text{SCAD}} = \arg\min_{\beta\in\mathbb{B}}\left\{\sum_{j=1}^{n_0}\left[Y_j - \hat{m}(\beta^\top X_j)\right]^2 + n_0\sum_{k=1}^{d_{n_0}} p_{\lambda_{n_0}}^{\text{SCAD}}(|\beta_k|)\right\}. \tag{13}$$

SCAD Algorithm

- Step 1. Given data $\{Y_j, X_j\}_{j=1}^{n_0}$, calculate the initial estimator $\hat{\beta}^{(0)}$ by the MAVE method. Set $t = 1$.

- Step 2. For $t \geq 1$, given $\hat{\beta}^{(t-1)}$, calculate

$$(\hat{a}_j^{(t-1)}, \hat{b}_j^{(t-1)}) = \arg\min_{a_j, b_j} \Big\{ \sum_{l=1}^{n_0} \big[Y_l - a_j - b_j (\hat{\beta}^{(t-1)})^\top (X_l - X_j) \big]^2 \cdot K_{h_1}\big((\hat{\beta}^{(t-1)})^\top (X_l - X_j) \big) \Big\}.$$

- Step 3. Given $\hat{a}_j^{(t-1)}$ and $\hat{b}_j^{(t-1)}$, update the estimate of β_0 by letting

$$\hat{\beta}^{(t)} = \arg\min_{\beta \in \mathbb{B}^{d_{n_0}}} \left\{ \sum_{j=1}^{n_0} \left[Y_j - \hat{a}_j^{(t-1)} - \hat{b}_j^{(t-1)} (\beta - \hat{\beta}^{(t-1)})^\top X_j \right]^2 + n_0 \sum_{k=1}^{d_{n_0}} p_{\lambda_{n_0}}^{\text{SCAD}}(|\beta_k|) \right\}$$

SCAD Algorithm

- Step 4. Let $\hat{\beta}^{(t)} = \text{sgn}(\hat{\beta}_1^{(t)})\hat{\beta}^{(t)}/\|\hat{\beta}^{(t)}\|$ and $t = t + 1$. Repeat Steps 2 and 3 until convergence reaches. Finally, let $\hat{\beta}_{\text{SCAD}} = \hat{\beta}^{(t)}$.

In summary, we can first use (13) to select relevant covariates and obtain $\hat{\beta}_{\text{SCAD}}$, then, set $\hat{Z}_i = \hat{\beta}_{\text{SCAD}}^\top X_i$ for the control group and the treated group, respectively. Finally, we can estimate the treatment effect using (9), denoted by $\hat{\Delta}_{\text{SCAD}}$.

Asymptotic Property

To derive the asymptotic property of $\hat{\Delta}_{\mathsf{SCAD}}$, we make following assumptions.

Assumption 7

For $j = 1, \ldots, n_0$, $Y_{0j} = m(\beta_0^\top X_j) + \varepsilon_j$, *where* $E(\varepsilon_j | X_j) = 0$ *and* $E(\varepsilon_j^4 | X_j) < M$ *for some* $M > 0$.

Assumption 8

Denote $\beta_{0,-1} = (\beta_{0,2}, \ldots, \beta_{0,d_{n_0}})^\top$ *and define a* $d_{n_0} \times (d_{n_0} - 1)$ *matrix as* $J_{\beta_0} = \begin{pmatrix} -\beta_{0,-1}^\top / \sqrt{1 - \|\beta_{0,-1}\|^2} \\ \mathbf{I}_{d_{n_0}-1} \end{pmatrix}$, *where* $\mathbf{I}_{d_{n_0}-1}$ *is the order* $d_{n_0} - 1$ *identity matrix. Assume that the smallest eigenvalue of* $J_{\beta_0}^\top \Sigma J_{\beta_0}$ *is larger than a positive constant* c, *where*

$$\Sigma = E\left\{[m'(Z_j)]^2 [E(X_j|Z_j) - X_j][E(X_j|Z_j) - X_j]^\top\right\}.$$

Variable Selection Theory

Assumption 9

For $j = 1, \ldots, n_0$, the marginal density of $\beta^\top X_j$ is positive and uniformly continuous in a neighborhood of β_0.

Assumption 10

$d_{n_0}/n_0 h_1^3 \to 0$ and $n_0 h_1^4 \to 0$ as n_0 goes to infinity.

Denote
$W_{\mathrm{SCAD}} = E\Big\{m'(\beta_0^\top X_j)^2 J^\top_{\beta_{0,\mathcal{A}}}[E(X_{j,\mathcal{A}}|\beta^\top_{0,\mathcal{A}}X_{j,\mathcal{A}}) - X_{j,\mathcal{A}}][E(X_{j,\mathcal{A}}|\beta^\top_{0,\mathcal{A}}X_{j,\mathcal{A}})$
$-X_{j,\mathcal{A}}]^\top J_{\beta_{0,\mathcal{A}}}\Big\}$, where $X_{j,\mathcal{A}} = (X_{j,1}, \cdots, X_{j,s})^\top$ and J_{β_0} denotes the
$s \times (s-1)$ matrix $\begin{pmatrix} -\beta^\top_{0,\mathcal{A},-1}/\sqrt{1-||\beta_{0,\mathcal{A},-1}||^2} \\ \mathbf{I}_{s-1} \end{pmatrix}$ with $\beta_{0,\mathcal{A},-1} = (\beta_{0,2}, \ldots, \beta_{0,s})^\top$.

Variable Selection Theory

Theorem 3

Under Assumptions 4 and 7 - 10, if the tuning parameter λ_{n_0} satisfies $\lambda_{n_0} \to 0$ and $\sqrt{n_0/d_{n_0}}\,\lambda_{n_0} \to \infty$, then, with probability approaching 1, we have:

(a) *Sparsity:* $\hat{\beta}_{SCAD,\mathcal{A}^c} = 0$.

(b) *Asymptotic representation:*

$$
\begin{aligned}
\hat{\beta}_{SCAD,\mathcal{A}} - \beta_{0,\mathcal{A}} =& \frac{1}{n_0}\sum_{j=1}^{n_0} J_{\beta_{0,\mathcal{A}}} W_{SCAD}^{-1} J_{\beta_{0,\mathcal{A}}}^{\top} m'(\beta_0^{\top} X_j)\{X_{j,\mathcal{A}} - E[X_{j,\mathcal{A}} | \beta_{0,\mathcal{A}}^{\top} X_{j,\mathcal{A}}]\}\varepsilon_j \\
&+ o_p(n_0^{-1/2}) \\
:=& \frac{1}{n_0}\sum_{j=1}^{n_0} \phi_{\mathcal{A}}(X_j, Y_j) + o_p(n_0^{-1/2}).
\end{aligned}
$$

Variable Selection Theory

From Part (b) of Theorem 3, it follows that $\sqrt{n_0}(\hat{\beta}_{\mathrm{SCAD},\mathcal{A}} - \beta_{0,\mathcal{A}}) \xrightarrow{d} \mathrm{N}(0, \Sigma_{\beta_0,\mathcal{A}})$, where $\Sigma_{\beta_0,\mathcal{A}} = \mathrm{Var}(\phi_{\mathcal{A}}(X_j, Y_j))$ for $j = 1, \ldots, n_0$. It also indicates that $\hat{\beta}_{\mathrm{SCAD}}$ satisfies Assumption 6. Hence, according to Theorem 1, we have the following corollary.

Corollary 1

Under the conditions imposed in Theorem 1 and Assumptions 7 - 10, one has

$$\sqrt{n_1}\left(\hat{\Delta}_{SCAD} - \Delta\right) \xrightarrow{d} \mathrm{N}\left(0, \sigma^2_{\Delta,SCAD}\right),$$

where $\sigma^2_{\Delta,SCAD} = \sigma_1^2 + \lambda\left(\sigma_2^2 + \sigma^2_{3,\mathcal{A}} + 2\delta_{a,\mathcal{A}}\Sigma_{23,\mathcal{A}}\right)$, σ_1^2 *and* σ_2^2 *defined in Theorem 1,* $\sigma^2_{3,\mathcal{A}} = \delta_{a,\mathcal{A}}\Sigma_{\beta,\mathcal{A}}\delta^\top_{a,\mathcal{A}}$, $\Sigma_{\beta,\mathcal{A}} = \mathrm{Var}(\phi_{\mathcal{A}}(X_j, Y_j))$, *and* $\Sigma_{23,\mathcal{A}} = \mathrm{Cov}(r(Z_j)\varepsilon_j, \phi_{\mathcal{A}}(X_j, Y_j))$ *for* $j = 1, \ldots, n_0$.

Screening Methods

In some real applications, the dimension of the covariates may be much larger than the sample size, which is termed as ultra-high dimensional covariates in the literature.

- For linear models with Gaussian predictors and responses, Fan and Lv (2008, JRSSB) proposed the sure independence screening (SIS) method.
- Fan, Feng, and Song (2011, JASA) developed a nonparametric independence screening method for sparse ultra-high dimensional additive models.
- Li, Zhong and Zhu (2012, JASA) proposed a sure independence screening procedure based on the distance correlation (DC-SIS).
- Zhong et al. (2016, Stat. Sin.) developed a robust DC-SIS procedure (DCRoSIS) that can be applied to the single index models.

DC-RoSIS-SCAD Method

When the dimension of covariates is ultra-high, we propose to first apply the DC-RoSIS procedure to reduce the dimensionality of the covariates, then, use (13) to estimate β. We denote the ultimate estimator for β_0 as $\hat{\beta}_{\text{DC-RoSIS-SCAD}}$ and the corresponding estimator for Δ as $\hat{\Delta}_{\text{DC-RoSIS-SCAD}}$.

Now, we let $F_{Y,0}(y)$ be the CDF of Y_j for the control group, and define $\hat{F}_{Y,0}(y) = \frac{1}{n_0}\sum_{j=1}^{n_0} I(Y_j \leq y)$. Denote $X_j = \left(X_{j,1}, \cdots, X_{j,d_{n_0}}\right)^\top$. The implementation of the corresponding DC-RoSIS procedure is summarized as follows.

DC-RoSIS Procedure

- Step 1. For $k = 1, \cdots, d_{n_0}$, we calculate the sample distance covariances $\widehat{\mathrm{dcov}}^2\{\hat F_{Y,0}(Y_j), \hat F_{Y,0}(Y_j)\}$, $\widehat{\mathrm{dcov}}^2\{X_{j,k}, X_{j,k}\}$ and $\widehat{\mathrm{dcov}}^2\{X_{j,k}, \hat F_{Y,0}(Y_j)\}$ for the control group. Here the sample distance covariance of two random variables U_j and V_j is defined as $\widehat{\mathrm{dcov}}^2\{U_j, V_j\} = \hat S_1 + \hat S_2 - 2\hat S_3$, where

$$\hat S_1 = \frac{1}{n_0^2}\sum_{j=1}^{n_0}\sum_{l=1}^{n_0} |U_j - U_l||V_j - V_l|,$$

$$\hat S_2 = \frac{1}{n_0^2}\sum_{j=1}^{n_0}\sum_{l=1}^{n_0} |U_j - U_l| \frac{1}{n_0^2}\sum_{j=1}^{n_0}\sum_{l=1}^{n_0} |V_j - V_l|,$$

and

$$\hat S_3 = \frac{1}{n_0^3}\sum_{j=1}^{n_0}\sum_{l=1}^{n_0}\sum_{q=1}^{n_0} |U_j - U_q||V_l - V_q|.$$

DC-RoSIS Procedure

- Step 2. For $k = 1, \cdots, d_{n_0}$, calculate the sample distance correlation

$$\hat{\omega}_k := \widehat{\text{dcorr}}\{X_{j,k}, \hat{F}_{Y,0}(Y_j)\} = \frac{\widehat{\text{dcov}}\{X_{j,k}, \hat{F}_{Y,0}(Y_j)\}}{\sqrt{\widehat{\text{dcov}}\{X_{j,k}, X_{j,k}\}\widehat{\text{dcov}}\{\hat{F}_{Y,0}(Y_j), \hat{F}_{Y,0}(Y_j)\}}}.$$

- Step 3. Keep covariates $X_{j,k}$ with $k \in \hat{\mathcal{A}} := \{k : \hat{\omega}_k \geq c n_0^{-\kappa}, k = 1, \ldots, d_{n_0}\}$, where $c > 0$ and $0 \leq \kappa < 1/2$ are pre-specified constants.

Using the DC-RoSIS, the number of covariates is reduced from d_{n_0} to $|\hat{\mathcal{A}}|$. Zhong et al. (2016, Stat. Sin.) demonstrated that the DC-RoSIS has the sure screening property; that is, $\Pr(\mathcal{A} \subseteq \hat{\mathcal{A}}) \to 1$ as $n_0 \to \infty$. [4]

[4] For the ultra-high dimensional case, the asymptotic property for the proposed ATE estimator, similar to that in Corollary 1, should be investigated, which is very challenging and warranted as a future research topic.

Monte Carlo Simulations

Monte Carlo Simulations

Simulation Settings

- We consider several different data generating processes (DGP).
- We set the bandwidth $h = 1 * n_0^{-1/3}$ and use the Gaussian kernel $K(v) = \frac{1}{\sqrt{2\pi}} \exp(-v^2/2)$.
- For each setting, the simulation is repeated 500 times.
- We use the mean of 500 absolute deviation errors (MADE) and root mean square error (RMSE) as the main evaluation metrics for different estimators.

Example 1: For each DGP, we vary the dimension of the covariates d and the true index vector β as following two cases.

- Case I: $d = 5$ and $\beta_0 = (1,\ 0.7,\ -0.5,\ 0.25,\ 0.8)^\top$.
- Case II: $d = 10$ with $\beta_0 = (1, 0.7, -0.5, 0.5, -0.75, 0.8, -0.4, 1, -0.2, 0.2)^\top$.

Simulation Settings

We consider the following linear and nonlinear model for the potential outcomes:

$$Y(0) = m(\beta^T X) + \varepsilon \quad \text{and} \quad Y(1) = Y(0) + 2,$$

where for $k = 1, \ldots, d$, $X_k \sim \mathrm{U}(-\sqrt{2}, \sqrt{2})$ for the treated units and $X_k \sim \mathrm{N}(0, 1)$ for the untreated units, and $\varepsilon \sim \mathrm{N}(0, 1)$. In this example, we consider two cases: $m(u) = u$ and $m(u) = 4 * \sqrt{|u+1|} + u$ respectively.

Clearly, the true treatment effect is $\Delta = 2$.

Example 1: Simulation Results

Table 1: Performance of SCM and QSCM under Example 1.

$m(u) = u$							
(n_0, n_1)		(200,100)		(400,200)		(800,400)	
	method	RMSE	MADE	RMSE	MADE	RMSE	MADE
$d = 5$	SCM	[illegible]	0.1287	0.1202	0.0969	0.0950	0.0740
	QSCM	[illegible]	0.1023	0.0886	0.0710	0.0618	0.0497
$d = 10$	SCM	0.1592	0.1237	0.1186	0.0957	0.0771	0.0619
	QSCM	0.1282	0.1015	0.0891	0.0713	0.0620	0.0498

$m(u) = 4 \cdot \sqrt{\lvert u + 1\rvert} + u$							
(n_0, n_1)		(200,100)		(400,200)		(800,400)	
	method	RMSE	MADE	RMSE	MADE	RMSE	MADE
$d = 5$	SCM	0.7781	0.7393	0.8075	0.7865	0.8729	0.8593
	QSCM	0.1280	0.0999	0.0870	0.0694	0.0618	0.0491
$d = 10$	SCM	0.7192	0.6721	0.7864	0.7657	0.8701	0.8594
	QSCM	0.1333	0.1046	0.0886	0.0709	0.0624	0.0503

Example 1: Simulation Results

- From the top panel of Table 1, we can see that both methods perform well with the linear potential outcome model, and our method is comparable to the SCM.
- From the bottom panel of Table 1, where the potential outcome model is nonlinear, we can see that the SCM is invalid and our method performs much better.
- The finite sample performance of the proposed estimator is well-behaved in the sense that both the MADE and RMSE are generally small.
- The RMSE decreases as the sample size n_1 increases, and the convergence rate is in line with our expectation.

Example 1: Simulation Results

Table 2: Coverage rates of the proposed Bootstrap procedure

$m(u) = u$						
$(n_0,\ n_1)$	(200,100)		(400,200)		(800,400)	
NCP	d=5	d=10	d=5	d=10	d=5	d=10
0.9	0.893	0.884	0.899	0.900	0.892	0.882
0.95	0.944	0.934	0.955	0.956	0.942	0.934
0.99	0.981	0.982	0.994	0.993	0.982	0.986
$m(u) = 4 + \sqrt{\lvert u+1\rvert} + u$						
$(n_0,\ n_1)$	(200,100)		(400,200)		(800,400)	
NCP	d=5	d=10	d=5	d=10	d=5	d=10
0.9	0.903	0.896	0.891	0.918	0.897	0.886
0.95	0.949	0.942	0.939	0.962	0.944	0.943
0.99	0.990	0.987	0.985	0.991	0.982	0.989

Example 2: Simulation Settings

Example 2: For simplicity, we illustrate the performance for high-dimensional variates, with the same setting as in Example 1 except that the number of covariates is set as $d_{n_0} = \lfloor 60 * n_0^{1/6} \rfloor$. And the true index vector is set as $\beta_0 = (1, 0.7, -0.5, 0.25, 0.8, \ 0, \dots, 0)^\top$.

- We set the bandwidth $h = 1 * n_0^{-1/3}$ and $h_1 = 1 * n_0^{-4/15}$, and use the Gaussian kernel.
- We use BIC to choose the penalty parameter λ_{n_0}.
- For each setting, the simulation is repeated 500 times.
- We still use MADE and RMSE as the main evaluation metrics for two different estimators (QSCM and pen-QSCM).
- We evaluate the performance of variable selection by the mean of true positive rate (TPR) and false positive rate (FPR) based on 500 replications.

Example 2: Simulation Results

Table 3: Performance of QSCM with variable selection

$m(u) = u$						
	QSCM		pen-QSCM		Variable Selection	
(n_0, n_1)	RMSE	MADE	RMSE	MADE	TPR	FPR
(200, 100)	0.2461	0.1943	0.1303	0.1026	0.9176	0.0260
(400, 200)	0.1198	0.0955	0.0865	0.0687	0.9724	0.0030
(800, 400)	0.0704	0.0561	0.0606	0.0483	0.9996	0.0018

$m(u) = 4 + \sqrt{\lvert u+1 \rvert} + u$						
	QSCM		pen-QSCM		Variable Selection	
(n_0, n_1)	RMSE	MADE	RMSE	MADE	TPR	FPR
(200, 100)	0.5958	0.4863	0.1691	0.1191	0.9996	0.0196
(400, 200)	0.1822	0.1424	0.0915	0.0725	1.0000	0.0005
(800, 400)	0.0753	0.0614	0.0633	0.0510	1.0000	0.0001

Example 2: Simulation Results

- Under both settings, the true positive rate is close to 1 and the false positive rate is relatively small and tends to 0 as the sample size n_0 increases.
- Compared with the QSCM estimator without variable selection, the penalized QSCM estimator behaves better with smaller RMSE and MADE.

Example 3: Simulation Settings

Example 3: For simplicity, we illustrate the performance for ultra-high dimensional variates, with the same setting as in Example 1 except that the number of covariates is set as $d_{n_0} = 5 * n_0$. And the true index vector is set as $\beta_0 = (1, 0.7, -0.5, 0.25, 0.8, 0, \ldots, 0)^\top$.

- In the DC-RoSIS procedure, we choose $c = 1$ and $\kappa = 1/3$.
- For each setting, the simulation is repeated 500 times.
- We still use MADE and RMSE as the main evaluation metrics for $\hat{\Delta}_{\text{DC-RoSIS-SCAD}}$.
- We evaluate the performance of variable selection by the mean of true positive rate (TPR) and false positive rate (FPR) based on 500 replications.

Example 3: Simulation Results

Table 4: Performance of QSCM with feature screening and variable selection

$m(u) = u$				
	DC-RoSIS-SCAD		Variable Selection	
(n_0, n_1)	RMSE	MADE	TPR	FPR
(200, 100)	0.1312	0.1033	0.8464	0.0056
(400, 200)	0.0890	0.0710	0.8968	0.0014
(800, 400)	0.0609	0.0489	0.9476	0.0005
$m(u) = 4 + \sqrt{\lvert u + 1\rvert} + u$				
	DC-RoSIS-SCAD		Variable Selection	
(n_0, n_1)	RMSE	MADE	TPR	FPR
(200, 100)	0.1443	0.1149	0.8724	0.0006
(400, 200)	0.0997	0.0784	0.9116	0.0000
(800, 400)	0.0645	0.0512	0.9560	0.0000

Example 3: Simulation Results

- Under both settings, the true positive rate is close to 1 and the false positive rate is relatively small and tends to 0 as the sample size n_0 increases.
- The RMSE and MADE values are generally small and approximately decrease at a rate of $1/\sqrt{n_1}$, as desired.

Empirical Example

Empirical Example

Empirical Example

- We apply our quasi synthetic control method to evaluate the effect of a labor market training program in the National Supported Work (NSW) Demonstration. It was originally analyzed by Lalonde (1986, AER), and subsequently, by researchers like Dehejia and Wahba (1999, JASA), Smith and Todd (2005, JoE), and Abadie and Imbens (2011, JBES).

- The NSW program was aimed at improving employment opportunities for individuals at the margins of the labor market by providing them with temporary subsidized jobs. It targeted individuals with low levels of education, individuals with criminal records, former drug addicts, and mothers who received welfare benefits for several years.

Empirical Example

- In the original experiment, individuals from the targeted population were randomly split between a treatment arm and a control arm, and the quantity of interest is the impact of the participation in the NSW program on 1978 yearly earnings in dollars for this specific population.

- Here, we use the version of the data in Dehejia and Wahba (1999) as experimental data.[5] Based on this experimental data, the ATE estimate is $\$1794$, which serves as an experimental benchmark in the literature. For details, see Dehejia and Wahba (1999).

- To estimate the effect of NSW program based on observational data, scholars propose to replace individuals in the control group of the experimental dataset with observations from the Panel Study of Income Dynamics (PSID).

[5] This data are available from Dehejia's website.

Empirical Example

We use the experimental participants and the non-experimental comparison group from the PSID:

- $D_i = \{0, 1\}$: an indicator for the participation of NSW program.
- Y_i: 1978 yearly earnings in dollars
- X_i: an 10×1 vector of covariates (age, education, black, hispanic, married, no degree, earnings in 1974, earnings in 1975, no earnings in 1974, and no earnings in 1975).
- There are $n_1 = 185$ treated units and $n_0 = 2490$ control units.

Empirical Example

Table 5: Summary statistics of 10 covariates.

	Experimental data				Non-experimental data	
	Treated ($n_1 = 185$)		Control ($n_0 = 260$)		PSID ($n_0 = 2490$)	
	Mean	Std	Mean	Std	Mean	Std
Covariates						
Age	25.82	7.16	25.05	7.06	34.85	10.44
Education	10.35	2.01	10.09	1.61	12.12	3.08
Black	0.84	0.36	0.83	0.38	0.25	0.43
Hispanic	0.06	0.24	0.11	0.31	0.03	0.18
Married	0.19	0.39	0.15	0.36	0.87	0.34
No degree	0.71	0.46	0.83	0.37	0.31	0.46
Earnings in 1974	2095.57	4886.62	2107.03	5687.91	19428.75	13406.88
Earnings in 1975	1532.06	3219.25	1266.91	3102.98	19063.34	13596.95
Unemployment in 1974	0.71	0.46	0.75	0.43	0.09	0.28
Unemployment in 1975	0.6	0.49	0.68	0.47	0.1	0.3
Outcome variable						
Earnings in 1978	6349.14	7867.4	4554.8	5483.84	21553.92	15555.35

Empirical Example

First, we would like to see if there exists a nonlinear relationship between the outcome and the index.

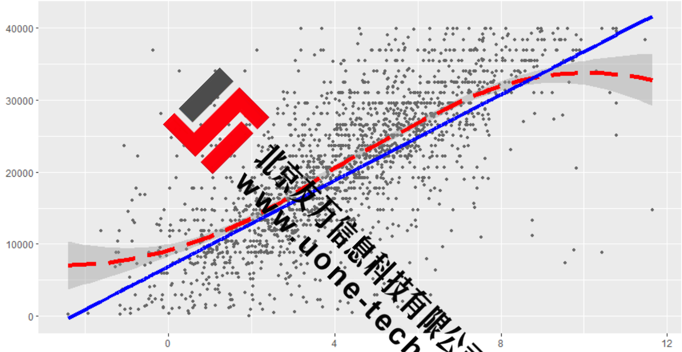

Figure 1: Scatterplot of Y_0 versus Z in PSID group, together with the *lowess* estimate of the unknown function $m(\cdot)$ in the dashed red line with its pointwise 95% confidence interval presented by the shaded area and a least-squares fitting of $m(\cdot)$ in the solid blue line.

Empirical Example

- As in Monte Carlo simulations, we use the Gaussian kernel, and the bandwidth is selected by cross-validation to minimize the mean squared error (MSE) of estimating Y_{0j} for the control units.
- We compare our quasi synthetic control estimator (QSCM) with a series of existing estimators:
  - the conventional synthetic control estimator (SCM)
  - the penalized synthetic control estimator which minimizes the bias (Pen. SCM) as in Abadie and L'Hour (2021)
  - the one-match nearest neighbor matching estimator (1-Matching)

Empirical Example

Table 6: Non-experimental estimates for the NSW data for various methods

Method	Benchmark	QSCM	SCM	Pen-SCM	1-Matching
Treatment effect	1794.34	1801.22	2118.61	1881.40	2236.87

Notes: The result for pen-SCM comes from Abadie and LHour (2021), and the result for 1-Matching is computed via the R package *Matching* by Sekhon and Saarinen (2023).

- From Table 6, we can see clearly that our QSCM estimator is **1801.22** is closest to the benchmark.
- The conventional SCM estimator is **2118.61**, which is substantially biased, as well as the one-match nearest neighbor matching estimator.
- We also compute the standard error of the QSCM estimator using the hybrid Bootstrap method and the standard error of $\hat{\Delta}_{QSCM}$ is **883.50**, which is much smaller than **1725.38**, the corresponding standard error for the 1-Matching estimate as in Abadie and Imbens (2006, ECTA).

Empirical Example

Finally, I need to mention about the computing time issue as mentioned earlier.

- In the conventional SCM, we need to calculate a 2490×1 vector of weights for each treated unit, so that this is computationally expensive.
- Indeed, our computing is carried out on a IBM X3550M4 dual processors server equipped with Twenty-Four Core Intel Xeon E5-2620 v2 @ 2.10GHz CPU, 64 GB RAM running Windows Server 2019. Using parallel computing in R language, it takes 1.69 hours to compute the conventional SCM estimate. Whereas, given a selected bandwidth, the computation time for our QSCM estimate is 13.6 seconds without parallel computation. Besides, as pointed out by Abadie and L'Hour (2021), the minimizer of (2) may not be unique with many treated units and/or many control units. Therefore, to search for the minimizer of (2), the computing is heavy.

Conclusion Remarks

Conclusion Remarks

Conclusion

- To overcome the shortcomings of the conventional synthetic control method, we propose a quasi synthetic control method, which can accommodate nonlinearity and feature fast computing.

- To address sparsity and variable selection, we propose to use the SCAD method to deal with a diverging number of covariates. And when the number of covariates is greater than the sample size, we suggest using a robust sure independence screening procedure based on the distance correlation to reduce the dimensionality first.

- We provide the inference theory for the QSC method, and derive the asymptotic distribution of the QSC ATE estimators with and without a penalty term.

- We also propose a carefully designed and easy-to-implement Bootstrap method and establish the validity of the subsampling method for inference.

THANKS

THANK YOU AGAIN for the INVITATION!

THANK YOU for YOUR ATTENTION!