Statistica Sinica Preprint No: SS-2023-0271	
Title	A Quasi Synthetic Control Method for Nonlinear Models with High-Dimensional Covariates
Manuscript ID	SS-2023-0271
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0271
Complete List of Authors	Zongwu Cai, Ying Fang, Ming Lin and Zixuan Wu
Corresponding Authors	Ming Lin
E-mails	linming50@xmu.edu.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

Statistica Sinica

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

Zongwu Cai[a], Ying Fang[b,c], Ming Lin[b,c,*], Zixuan Wu[c]

[a] *Department of Economics, University of Kansas, Lawrence, KS 66045, USA.*

[b] *Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen, Fujian 361005, China.*

[c] *Department of Statistics and Data Science, School of Economics, Xiamen University, Xiamen, Fujian 361005, China.*

Abstract: To make the conventional synthetic control method more flexible to estimate the average treatment effect (ATE), this article proposes a quasi synthetic control method for nonlinear models under the index model framework with possible high-dimensional covariates, together with a suggestion of using the minimum average variance estimation (MAVE) method to estimate parameters and the LASSO-type procedure to choose high-dimensional covariates. We derive the asymptotic distribution of the proposed ATE estimators for both finite and diverging dimensions of covariates. A properly designed Bootstrap method is proposed to obtain confidence intervals and its theoretical justification is provided. When the dimension of covariates is greater than the sample size, we suggest using the robust version of sure independence screening procedure based on the distance correlation to first reduce the dimensionality and then apply the MAVE approach to estimate parameters. Finally, Monte Carlo simulation studies are conducted to examine the finite sample performance of our proposed estimators and Bootstrap procedure. In ad-

*Corresponding author: Ming Lin (linming50@xmu.edu.cn).

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

dition, an empirical application to reanalyzing data from the National Supported Work Demonstration demonstrates the practical usefulness of our proposed method.

Key words and phrases: Average treatment effect; Bootstrap inference; Index model; Semiparametric estimation; Synthetic control method.

1. Introduction

When evaluating the impact of policy interventions, one of the main challenges lies in estimating unknown counterfactual outcomes. With observable covariates, a natural idea is to construct an outcome regression model. In practice, the classic linear regression model is usually inadequate or even incorrect. To fully capture the relationship between the covariates and the outcomes, researchers suggest using the nonparametric model which can avoid the risk of model misspecification. However, the nonparametric model is challenged by the so-called *curse of dimensionality*. Therefore, as a combination of the parametric and nonparametric models, the semiparametric model has been conceived to overcome the aforementioned limitations.

There is a vast literature concerning applying semiparametric techniques to estimate the treatment effect and the existing research can be divided into two categories: estimating the counterfactual outcomes directly and indirectly. For the former, various semiparametric approaches have been used to estimate the conditional mean function or the conditional quantile function. For example, Heckman, Ichimura and Todd (1998) proposed a kernel-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

matching-based estimator for the average treatment effect (ATE) and presented a rigorous distributional theory, while Chiburis (2010) discussed the semiparametric bounds on the average treatment effect of a binary treatment on a binary outcome. Under the framework of the latent factor model to vary cross-section, Hsiao, Ching and Wan (2012) initiated an approach, termed as the panel data approach (PDA), which assumes that the conditional mean of the outcomes of the treated units is a linear function of the outcomes of the control units. Furthermore, Li and Bell (2017) relaxed this linear conditional mean function assumption to allow for the conditional mean function to have any unknown functional form and they used a linear project argument to show that the PDA remains valid, although it may be less efficient than estimating the conditional mean function nonparametrically when the sample size is sufficiently large. As for the latter one, under the ignorability assumption, many scholars have proposed to first estimate the propensity score and then estimate the treatment effect by either re-weighting or matching technique. For details, see, for example, the papers by Abadie and Imbens (2006), Galvao and Wang (2015)and the references therein.

One of the most important semiparametric models is the single index model. On the one hand, the single index model projects the multi-dimensional covariates into a one-dimensional single index variable by a linear transformation. On the other hand, it assumes an unknown nonlinear link function for the single index variable, which is greatly flexible. In policy evaluation,

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

the single index model is usually used to estimate the propensity score. Song (2014) considered semiparametric models with single-index nuisance parameters. The single index component is allowed to be estimable only at the cube-root rate, and the corresponding single-index matching estimator can be identified under a weaker conditional median independence assumption. Sun, Yan and Li (2021) also considered a single index model for the propensity score, and developed the asymptotic theory for the two-step semiparametric ATE estimator. Park et al. (2021) further considered a constrained single index model for the interaction between a multi-valued treatment variable and covariates. The propensity score is sensitive to the model specification. As shown by Frölich (2004) and Kang and Schafer (2007), the misspecification of the propensity score can lead to misleading treatment effect estimates. Hence, using the single index model to flexibly characterize this relationship is advantageous.

In this article, our focus is on the statistical inference for the average treatment effect under the framework of the single index model. Our estimation procedure consists of two steps. First, parameters in the single index model are estimated by the minimum average variance estimation (MAVE) method proposed by Xia et al. (2002). In the second step, a nonparametric kernel smooth technique can be applied to estimate the weights for estimating the counterfactual outcomes. To address sparsity and variable selection, we propose to use the smoothly clipped absolute deviation (SCAD), a LASSO-type

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

method, proposed by Fan and Li (2001), to deal with a diverging number of covariates. 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, proposed by Zhong et al. (2016), and then using the MAVE approach to estimate parameters. Therefore, we make several contributions to the literature. First, our method is the first attempt to conduct a formal statistical inference for estimating the ATE for single index models. We derive the asymptotic inference theory for the corresponding ATE estimators for the high-dimensional covariate cases. Second, we propose a properly designed (hybrid) Bootstrap method by combining the wild Bootstrap and the classical nonparametric Bootstrap and show that the carefully designed Bootstrap method provides valid inferences theoretically and empirically. Third, we propose combining our method with the penalized and feature screening methods to address the ultra-high dimensional covariate cases. Finally, the proposed ATE estimator is fast to compute, and we demonstrate through simulations and an empirical example that the proposed method, which is robust to nonlinear model situations, can greatly enhance the applicability of estimating the ATE. Therefore, our work complements the existing inference work in the literature on treatment effects.

The rest of the paper is organized as follows. Section 2 first presents the model setup for our method, and the estimation procedure is described in detail. Additionally, this section provides the asymptotic theory for the

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

proposed estimator and presents a carefully designed Bootstrap method with a theoretical justification for valid inferences. Section 3 deals with choosing covariates and addressing sparsity. The LASSO-type method and the feature screening procedure are developed in this section. A simulation study is conducted in Section 4 to illustrate the finite sample performance. Section 5 reports the empirical analysis using our quasi-synthetic control method to analyze data from the National Supported Work (NSW) Demonstration. Finally, Section 6 concludes the paper. All detailed technical proofs are provided in the Supplementary Material.

2. Quasi Synthetic Control Method

2.1 Setup

Assume we observe n units and some of the units are exposed to the treatment or intervention of our interest. The treatment status of unit i is indicated by a binary variable D_i, where $D_i = 1$ if unit i is treated and $D_i = 0$ otherwise. To define treatment effects, we adopt the potential outcomes framework in Rubin (1974). Formally speaking, for each unit i, let Y_{1i} and Y_{0i} be the potential outcomes under treatment and without treatment, respectively. Then, the observed outcome Y_i can be written as $Y_i = D_i Y_{1i} + (1 - D_i)Y_{0i}$; that is to say, we can only observe Y_{1i} for the treated unit and Y_{0i} for the control unit. For each unit i, we can also observe a $d \times 1$ vector of covariates, denoted by X_i. Assume there are n_1 units being treated and the remaining $n_0 = n - n_1$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

units are not exposed to the treatment. For simplicity, we reorder these units so that the control units come first in the data set. Then, the observed data set can be written as $(Y_i, X_i)_{i=1}^n$ with $i = 1, \ldots, n_0$ being the control units and $i = n_0 + 1, \ldots, n$ being the treated units. Notice that while the synthetic control method is mostly used to deal with a panel data model, as pointed by Abadie and L¡¯Hour (2021), as a matching estimator essentially, the synthetic control method can also be used to analyze cross sectional data.

We are interested in estimating the average treated effect for the treated units, which is defined as

$$\Delta = E(Y_{1i} - Y_{0i}) \tag{2.1}$$

for $i = n_0 + 1, \ldots, n$. The difficulty of estimating Δ lies in the fact that $(Y_{0i})_{i=n_0+1}^n$ are not observable.

2.2 Estimation Procedure

We consider the prediction function based on the conditional expectation of Y_{0i} given X_i, denoted by $m(x) = E(Y_{0i}|X_i = x)$, in an index form as $m(x) = m(\beta_0^\top x) = m(z)$, where $m(\cdot)$ is an unknown function and $z = \beta_0^\top x \in \mathbb{R}$, which covers the linear model as a special case. Denote $\beta_0 = (\beta_{0,1}, \cdots, \beta_{0,d})^T$. For the identification purpose, it is commonly assumed, in what follows, that $\beta_{0,1} > 0$ and $||\beta_0||^2 = \sum_{k=1}^d \beta_{0,k}^2 = 1$.

When β_0 is given, the estimation of $m(z)$ is one-dimensional and the so-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

called *curse of dimensionality* in a nonparametric smoothing can be avoided.Define $Z_i = \beta_0^\top X_i$. The kernel type (Nadaraya-Watson) estimate of $m(z)$, based on the data $(Y_j, X_j)_{j=1}^{n_0}$ from the control group, is given by

$$\tilde{m}(z) = \sum_{j=1}^{n_0} c_{j,h}(z) Y_j, \tag{2.2}$$

where $c_{j,h}(z) = K_h(z - Z_j)/\sum_{l=1}^{n_0} K_h(z - Z_l)$, $K_h(u) = K(u/h)/h$, $K(u)$ is a kernel function, and h is the bandwidth. Now, an infeasible prediction of Y_{0i} is denoted by $\tilde{Y}_{0i}$, where

$$\tilde{Y}_{0i} = \tilde{m}(Z_i) = \sum_{j=1}^{n_0} c_{j,h}(Z_i) Y_j \tag{2.3}$$

for $i = n_0 + 1, \ldots, n$. Note that (2.3) is infeasible since it is based on the unknown parameter β_0. Accordingly, an infeasible estimate of Δ is

$$\tilde{\Delta} = \frac{1}{n_1} \sum_{i=n_0+1}^{n} \left[Y_{1i} - \sum_{j=1}^{n_0} c_{j,h}(Z_i) Y_j \right] = \frac{1}{n_1} \sum_{i=n_0+1}^{n} Y_i - \frac{1}{n_0} \sum_{j=1}^{n_0} a_{j,h} Y_j, \tag{2.4}$$

where $a_{j,h} = \frac{1}{n_1} \sum_{i=n_0+1}^{n} K_h(Z_i - Z_j) \left[\frac{1}{n_0} \sum_{l=1}^{n_0} K_h(Z_i - Z_l) \right]^{-1}$.

Clearly, under this setting, we first need to find β_0 such that $\beta_0^\top X_i$ can be the best to predict Y_{0i} for $i = 1, \ldots, n_0$. To do so, we suggest using the index model and its estimation approach is described in Section 2.4.

Interestingly, our method shares some similarities and differences with the synthetic control method (SCM) proposed by Abadie and Gardeazabal (2003). Although the SCM is originally designed to deal with the panel data

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

setting, Abadie and L¡¯Hour (2021) presented a penalized version of the SCM for disaggregated data. Apparently, $c_{j,h}(Z_i)$ in (2.3) also serves as an inidividual weight. However, different from Abadie and L¡¯Hour (2021), our weights $\{c_{j,h}(Z_i)\}$ take care of both the best prediction to resemble the characteristics of the potential outcome without the intervention and nonlinearity of prediction function since our model is in a semiparametric nature. Therefore, our method is termed as the quasi synthetic control method (QSCM).

From the above discussions, the QSCM for estimating Δ consists of the following two steps. First, use (2.9) given in Section 2.4 to obtain $\hat{\beta}$, and then, set $\hat{Z}_i = \hat{\beta}^\top X_i$ for $i = 1, \ldots, n$. Second, compute the feasible estimate of Δ based on (2.4), which is

$$\hat{\Delta} = \frac{1}{n_1} \sum_{i=n_0+1}^{n} \left[Y_{1i} - \sum_{j=1}^{n_0} \hat{c}_{j,h}(\hat{Z}_i) Y_j \right] = \frac{1}{n_1} \sum_{i=n_0+1}^{n} Y_i - \frac{1}{n_0} \sum_{j=1}^{n_0} \hat{a}_{j,h} Y_j, \quad (2.5)$$

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

2.3 Asymptotic Theory

To describe the asymptotic properties of $\hat{\Delta}$, some notations are introduced. 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 $i = n_0 + 1, \ldots, n$. 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 $i = n_0 + 1, \ldots, n$. Denote $m_c(x) = E\big[Y_{0j} \,|\, X_j = x\big]$ for $j = 1, \ldots, n_0$ and $m_t(x) = E\big[Y_{0j} \,|\, X_j = x\big]$ for

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

$i = n_0 + 1, \ldots, n$. Now we make the following assumptions.

A1. Assume that $m_c(x) = m_t(x) = m(z)$, where $z = \beta_0^\top x$ and $\beta_0 \in \mathbb{B} := \{\beta \in \mathbb{R}^d : \beta_1 > 0, ||\beta||^2 = \sum_{k=1}^d \beta_k^2 = 1\}$. Furthermore, assume that the second order derivative of $m(z)$ is continuous.

A2. $\{Y_{0j}, Y_{1j}, X_j\}_{j=1}^{n_0}$ 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 $f_c(z) \geq M_1 > 0$ for $z \in \mathcal{C}_2$.

A3. 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.

A4. The kernel function $K(\cdot)$ is symmetric, bounded and positive. Further assume that the first derivative of $K(\cdot)$ is continuous.

A5. 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 $0 < \eta < \infty$.

A6. Assume that for any estimate of β_0, $\hat{\beta}$ admits the following expression

$$\sqrt{n_0}\,\left(\hat{\beta} - \beta_0\right) = \frac{1}{\sqrt{n_0}} \sum_{j=1}^{n_0} \phi(X_j, Y_j) + o_p(1) \xrightarrow{d} N(0, \Sigma_{\beta_0}) \tag{2.6}$$

for some function $\phi(\cdot)$ with variance Σ_{β_0} =Var$(\phi(X_j, Y_j))$ for $j = 1, \ldots, n_0$.

Assumptions listed above are standard. Assumption A1 assumes that the conditional expectations of the outcomes for the treated and control units are the same in the absence of treatment, following a single-index model. The ra-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

tio function $r(z)$ in Assumption A3 is interpreted as *acceptance probability* in rejection sampling instead of *importance re-weighting*, or *covariate shift*, in the machine learning literature; see, for example, Wu, Ren and Mu (2016) and the references therein. Assumption A5 is under-smoothed in a nonparametric kernel smoothing estimation, which makes the asymptotic bias negligible and leads the practical choice of h in application to be not difficult. The assumption in A6 is common in the index model literature; see, for example, Cai, Juhl and Yang (2015) and the references therein. Indeed, under some regularity conditions, $\sqrt{n_0}(\hat{\beta} - \beta_0)$ can be expressed as in (2.6); see Section 2.4 for details.

Let $\varepsilon_j = Y_{0j} - E(Y_{0j} \,|\, X_j)$ 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 A6. Also, define $\Sigma_{23} = \mathrm{Cov}(\phi(X_j, Y_j), r(Z_j)\varepsilon_j)$. The asymptotic normality of $\hat{\Delta}$ is stated in the following theorem with its proof provided in the Supplementary Material.

Theorem 1: Under Assumptions A1 - A6, one has

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

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

It follows from Theorem 1 that the asymptotic variance consists of four terms. In particular, the first term in σ_Δ^2 stands for the variance of $Y_{1i} - m(Z_i)$, the second term characterizes the variation for estimating $m(Z_i)$ given β_0, the

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

third term σ_3^2 is the variation carried over from the estimation of β_0, and 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, Xiong and Wu (2006).

2.4 MAVE Method for Estimating β_0

We now discuss how to estimate β_0. To do so, assume that Y_{0i} follows a single index model

$$Y_{0i} = m(\beta_0^\top X_i) + \varepsilon_i = m(Z_i) + \varepsilon_i, \tag{2.7}$$

where $E(\varepsilon_i|X_i) = 0$, $m(\cdot)$ is an unknown link function, and $\beta_0 = (\beta_{0,1}, \ldots, \beta_{0,d})^\top$ is the $d \times 1$ index vector. The estimation of the index vector β_0 has attracted extensive attentions. For example, Ichimura (1993) proposed the semiparametric least squares estimation of the single index model based on the leave-one-out technique. Since the single index model shares a close connection with the central mean subspace in the sufficient dimension reduction, Xia et al. (2002) proposed the (conditional) minimum average variance estimation method for the dimension reduction problem, and later, Xia (2006) showed that this method can be applied to the single index model. We employ the MAVE method to estimate β_0, described as follows.

Notice that for the single index model (2.7),

$$\beta_0 = \arg\min_{\beta \in \mathbb{B}} E\left[Y_{0i} - E(Y_{0i}|\beta^\top X_i)\right]^2. \tag{2.8}$$

In our setting, the index is estimated by the observed data for the control

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

units, $\{Y_j, X_j\}_{j=1}^{n_0}$. Motivated by the local linear smoothing technique, the sample analogue of (2.8) is given by

$$\hat{\beta}_{\text{MAVE}} = \arg\min_{\substack{\beta\in\mathbb{B}\\ a_j,b_j}} \sum_{j=1}^{n_0}\sum_{l=1}^{n_0} \left[Y_l - a_j - b_j\beta^\top(X_l - X_j)\right]^2 w_{lj}, \tag{2.9}$$

where $w_{lj} = K_{h_1}(\beta^\top(X_l - X_j))$, $K_{h_1}(v) = K(v/h_1)/h_1$, $K(\cdot)$ is a kernel function and h_1 is the bandwidth. Xia (2006) proposed an easy-to-implement algorithm to estimate β_0, which is presented in the Supplementary Material.

Xia (2006) derived the asymptotic normality for $\hat{\beta}_{\text{MAVE}}$ and shows that the asymptotic covariance matrix of $\hat{\beta}_{\text{MAVE}}$ can achieve the information lower bound in the semiparametric sense. From Xia (2006), one can see that under some regularity conditions, $\hat{\beta}_{\text{MAVE}}$ satisfies Assumption A6 with $\phi(X_j, Y_j) = W_m^+ m'(\beta_0^\top X_j) v_{\beta_0}(X_j)\varepsilon_j$, where $m'(z)$ is the first derivative of $m(z)$, $v_{\beta_0}(X_j) = E(X_j|\beta_0^\top X_j) - X_j$, $W_m = E\{m'(\beta_0^\top X_j)^2 v_{\beta_0}(X_j) v_{\beta_0}^\top(X_j)\}$, and W_m^+ denotes the Moore-Penrose inverse of W_m. Therefore, Assumption A6 is reasonable.

2.5 A Bootstrap Inference

Clearly, Theorem 1 provides the asymptotic distribution for $\hat{\Delta}$, so that an inference can be made if σ_Δ^2 can be estimated consistently. But, one can see from Theorem 1 that the form of σ_Δ^2 is complicated so that it is not easy to get a consistent estimate. To facilitate an easy inference, we propose the following hybrid Bootstrap procedure by combining the (conditional) wild Bootstrap similar to that in Zhang, Huang and Liu (2020) for single index

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

models and the nonparametric Bootstrap, 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 by (2.5) as $\hat{\Delta}$.

Step 2. Generate the nonparametric Bootstrap sample $\{(X_i^*, Y_i^*)\}_{i=n_0+1}^{n}$ by drawing with replacement from the original treated group $\{(X_i, Y_i)\}_{i=n_0+1}^{n}$.

Step 3. 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. Using $\{(X_j, Y_j^*)\}_{j=1}^{n_0}$ to re-estimate the index parameter as $\hat{\beta}^*$.

Step 4. 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

$$\hat{\Delta}^* = \frac{1}{n_1} \sum_{i=n_0+1}^{n} Y_i^* - \frac{1}{n_0} \sum_{j=1}^{n_0} \hat{a}_{j,h}^* Y_j^*,$$

where $\hat{a}_{j,h}^* = \frac{1}{n_1} \sum_{i=n_0+1}^{n} K_h(\hat{Z}_i^* - \hat{Z}_j^*) \left[\frac{1}{n_0} \sum_{l=1}^{n_0} K_h(\hat{Z}_i^* - \hat{Z}_l^*) \right]^{-1}$, which is the Bootstrap version of $\hat{a}_{j,h}$ in (2.5).

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

A (1-α)100% Bootstrap confidence interval for Δ can be constructed as $(\hat{\Delta} - z_{\alpha/2}\hat{\sigma}_\Delta/\sqrt{n_1}, \hat{\Delta} + z_{\alpha/2}\hat{\sigma}_\Delta/\sqrt{n_1})$ based on the asymptotic normality of $\hat{\Delta}$ in Theorem 1, where $z_{\alpha/2}$ is the $(1-\alpha/2)$th quantile of the standard normal

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

distribution. The theoretical validity of this procedure can be confirmed by the following theorem with its proof in the Supplementary Material.

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

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

where σ_Δ^2 is defined in Theorem 1.

Our method shares a deep connection with the matching methods. Abadie and Imbens (2011) demonstrated that the standard Bootstrap method fails to conduct inference for matching estimators. To overcome this problem, Otsu and Rai (2017) proposed asymptotically valid inference methods for matching estimators based on the weighted Bootstrap. However, their method only deals with the case of a fixed number of matches. Our method matches each treated unit with all control units, which means that the number of matches increases with the size of the control group and is definitely not fixed.

3. Quasi Synthetic Control Method With Many Covariates

Based on the above discussion, we assume a single index model, as in (2.7), for Y_{0i}. When the number of predictor variables is large, it is necessary to discriminate relevant variables from irrelevant variables, since the inclusion of irrelevant variables may harm estimation accuracy and model interpretability. This negative effect may be amplified in our quasi synthetic control method since the method is intrinsically a two-step procedure.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

Variable selection methods for single index models have been widely discussed in the literature, for example, Kong and Xia (2007), Wang and Yin (2008), Zeng, He and Zhu (2012), Wang, Xu and Zhu (2013), and the references therein. However, existing literature mainly focuses on the case of finite-dimensional covariates, while in many cases, the dimension of covariates might grow with the sample size. In the following sections, we propose penalized QSCM estimation procedures with a diverging number of covariates and with ultra-high dimensional covariates.

3.1 QSCM With a Diverging Number of Covariates

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

$$\sum_{j=1}^{n_0} [Y_j - \hat{m}(\beta^\top X_j)]^2 + n_0 \sum_{k=1}^{d} p_{\lambda_{n_0}}(|\beta_k|), \tag{3.1}$$

where $\beta = (\beta_1, \cdots, \beta_{d_{n_0}})^T$, $\hat{m}(\cdot)$ 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. For a given β, we can estimate $\hat{m}(\beta^\top X_j)$ using the local linear smoothing method. Specifically, we let

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

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

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, different choices of $p_{\lambda_{n_0}}(\cdot)$ lead to different variable selection methods. One choice is to set $p_{\lambda_{n_0}}(|\beta_k|) = \lambda_{n_0}|\beta_k|$, which corresponds to the least absolute shrinkage and selection operator (LASSO) proposed by Tibshirani (1996). However, the LASSO estimator is biased. Alternatively, Fan and Li (2001) proposed the SCAD penalty, which is defined by $\left(p^{\mathrm{SCAD}}_{\lambda_{n_0}}(|\theta|)\right)' = \lambda_{n_0}\{I(|\theta| \le \lambda_{n_0}) + \frac{(a\lambda_{n_0} - |\theta|)_+}{(a-1)\lambda_{n_0}} I(|\theta| > \lambda)\}$ for some $a > 2$. We choose the SCAD penalty and modify the objective function in (3.1) as

$$\hat{\beta}_{\mathrm{SCAD}} = \arg\min_{\beta \in \mathbb{B}} \left\{ \sum_{j=1}^{n_0} \left[Y_j - \hat{m}(\beta^\top X_j)\right]^2 + n_0 \sum_{k=1}^{d_{n_0}} p^{\mathrm{SCAD}}_{\lambda_{n_0}}(|\beta_k|) \right\}. \qquad (3.2)$$

The algorithm to solve the optimization problem in (3.2) is summarized as follows:

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

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

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

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

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

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

Step 4. Let $\hat{\beta}^{(t)} = \mathrm{sgn}(\hat{\beta}_1^{(t)})\hat{\beta}^{(t)}/\|\hat{\beta}^{(t)}\|$ and $t = t + 1$, where $\hat{\beta}_1^{(t)}$ denotes the first component of $\hat{\beta}^{(t)}$. Repeat Steps 2 and 3 until convergence reaches. Finally, let $\hat{\beta}_{\mathrm{SCAD}} = \hat{\beta}^{(t)}$.

Based on the above discussion, we can first use (3.2) to select relevant covariates and obtain $\hat{\beta}_{\mathrm{SCAD}}$, then, set $\hat{Z}_i = \hat{\beta}_{\mathrm{SCAD}}^\top X_i$ for the control and treated groups, respectively. Finally, we can estimate the treatment effect using (2.5), denoted by $\hat{\Delta}_{\mathrm{SCAD}}$. To derive the asymptotic property of $\hat{\Delta}_{\mathrm{SCAD}}$, we make the following assumptions.

B1. 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 $M > 0$.

B2. 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 $\Sigma = E\left\{[m'(Z_j)]^2 [E(X_j|Z_j) - X_j][E(X_j|Z_j) - X_j]^\top\right\}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

B3. 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.

B4. $d_{n_0}/n_0\, h_1^3 \to 0$ and $n_0\, h_1^4 \to 0$ as n_0 goes to infinity.

One can see that the above assumptions are indeed regularity assumptions, also listed in Peng and Huang (2011), which discuss the penalized least squares estimator for single index models with finite number of covariates. Denote $W_{\text{SCAD}} = E\Big\{m'(\beta_0^\top X_j)^2 J_{\beta_{0,\mathcal{A}}}^\top [E(X_{j,\mathcal{A}}|\beta_{0,\mathcal{A}}^\top X_{j,\mathcal{A}}) - X_{j,\mathcal{A}}][E(X_{j,\mathcal{A}}|\beta_{0,\mathcal{A}}^\top X_{j,\mathcal{A}}) - X_{j,\mathcal{A}}]^\top J_{\beta_{0,\mathcal{A}}}\Big\}$, where $X_{j,\mathcal{A}} = (X_{j,1}, \cdots, X_{j,s})^\top$, and J_{β_0} denotes the $s \times (s-1)$ matrix $\begin{pmatrix} -\beta_{0,\mathcal{A},-1}^\top / \sqrt{1-||\beta_{0,\mathcal{A},-1}||^2} \\ \mathbf{I}_{s-1} \end{pmatrix}$ with $\beta_{0,\mathcal{A},-1} = (\beta_{0,2}, \ldots, \beta_{0,s})^\top$. Then, we have the following asymptotic result with its detailed proof given in the Supplementary Material.

Theorem 3: Under Assumptions A4 and B1 - B4, 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: $\hat{\beta}_{\text{SCAD},\mathcal{A}^C} = 0$.

(b) Asymptotic representation:

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

Evidently, from Part (b) of Theorem 3, it follows that $\sqrt{n_0}(\hat{\beta}_{\text{SCAD},\mathcal{A}} - \beta_{0,\mathcal{A}}) \xrightarrow{d} \text{N}(0, \Sigma_{\beta_{0,\mathcal{A}}})$, where $\Sigma_{\beta_{0,\mathcal{A}}} = \text{Var}(\phi_{\mathcal{A}}(X_j, Y_j))$ for $j = 1, \ldots, n_0$.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

It also indicates that $\hat{\beta}_{\text{SCAD}}$ satisfies Assumption A6. Hence, according to Theorem 1, we have the following corollary.

Corollary 1: Under the conditions imposed in Theorems 1 and 3, one has

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

where $\sigma^2_{\Delta,\text{SCAD}} = \sigma_1^2 + \eta\left(\sigma_2^2 + \sigma^2_{3,\mathcal{A}} + 2\delta_{a,\mathcal{A}}\Sigma_{23,\mathcal{A}}\right)$, σ_1^2 and σ_2^2 are defined in Theorem 1, $\sigma^2_{3,\mathcal{A}} = \delta_{a,\mathcal{A}}\Sigma_{\beta_0,\mathcal{A}}\delta^\top_{a,\mathcal{A}}$ with $\delta_{a,\mathcal{A}} = E\left[m'(Z_i)X^\top_{i,\mathcal{A}}\right]$ for $i = n_0 + 1, \cdots, n$, and $\Sigma_{23,\mathcal{A}} = \text{Cov}(r(Z_j)\varepsilon_j, \phi_{\mathcal{A}}(X_j, Y_j))$ for $j = 1, \ldots, n_0$.

To make inference of $\hat{\Delta}_{\text{SCAD}}$ in practice, we suggest using a Bootstrap method similar to the one introduced in Section 2.5. Specifically, we apply the wild Bootstrap method to the control group and the nonparametric Bootstrap method to the treated group to obtain a Bootstrap sample. Then, we can estimate $\hat{\Delta}_{\text{SCAD}}$ based on the Bootstrap sample. By repeating the above steps many times, we can obtain an estimate of $\sigma^2_{\Delta,\text{SCAD}}$. The theoretical validation of such a Bootstrap procedure should be one of our future research topics.

3.2 QSCM With Ultra-high Dimensional Covariates

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. As pointed out by Fan, Samworth and Wu (2009), for such cases, traditional regularization methods may not perform well. To

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

deal with ultra-high dimensional covariate cases, several feature screening procedures have been proposed. For linear models with Gaussian predictors and responses, Fan and Lv (2008) proposed the sure independence screening (SIS) method to reduce dimensionality from ultra-high to below the sample size. Later, Fan, Feng, and Song (2011) developed a nonparametric independence screening method for sparse ultra-high dimensional additive models. For more general model settings, Li, Zhong and Zhu (2012) proposed a sure independence screening procedure based on the distance correlation (DC-SIS). Furthermore, Zhong et al. (2016) developed a robust DC-SIS procedure (DC-RoSIS) that can be applied to the single index models.

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 (3.2) 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 $\hat{F}_{Y,0}(y) = \frac{1}{n_0}\sum_{j=1}^{n_0} I(Y_j \le y)$. Denote $X_j = \left(X_{j,1}, \cdots, X_{j,d_{n_0}}\right)^T$. The implementation of the corresponding DC-RoSIS procedure is summarized as follows.

Step 1. For $k = 1, \cdots, d_{n_0}$, we calculate the sample distance covariances $\widehat{\text{dcov}}^2\{\hat{F}_{Y,0}(Y_j), \hat{F}_{Y,0}(Y_j)\}$, $\widehat{\text{dcov}}^2\{X_{j,k}, X_{j,k}\}$ and $\widehat{\text{dcov}}^2\{X_{j,k}, \hat{F}_{Y,0}(Y_j)\}$ for the control group. Here the sample distance covariance of two random variables U_j and V_j is defined as $\widehat{\text{dcov}}^2\{U_j, V_j\} = \hat{S}_1 + \hat{S}_2 - 2\hat{S}_3$, where

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

$$\hat{S}_1 = \frac{1}{n_0^2}\sum_{j=1}^{n_0}\sum_{l=1}^{n_0}|U_j - U_l||V_j - V_l|, \quad \hat{S}_2 = \frac{1}{n_0^2}\sum_{j=1}^{n_0}\sum_{l=1}^{n_0}|U_j - U_l| \cdot \frac{1}{n_0^2}\sum_{j=1}^{n_0}\sum_{l=1}^{n_0}|V_j - V_l|,$$

and

$$\hat{S}_3 = \frac{1}{n_0^3}\sum_{j=1}^{n_0}\sum_{l=1}^{n_0}\sum_{q=1}^{n_0}|U_j - U_q||V_l - V_q|.$$

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. We keep covariates $X_{j,k}$ with $k \in \hat{\mathcal{A}} := \{k : \hat{\omega}_k \geq cn_0^{-\kappa}, k = 1, \ldots, d_{n_0}\}$, 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) demonstrated that under regularity conditions, the DC-RoSIS has the sure screening property; that is, $\Pr(\mathcal{A} \subseteq \hat{\mathcal{A}}) \to 1$ as $n_0 \to \infty$. 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.

4. Monte Carlo Simulation Studies

In the following simulation studies, we investigate the finite sample performance of our proposed estimators, methods for selecting covariates, and Bootstrap procedure.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

4.1 Evaluating QSCM and the Bootstrap Procudure

We first evaluate the performance of our proposed estimator $\hat{\Delta}$ in (2.5). Let $Y(0) = m\left(\beta_0^\top X\right) + \varepsilon$ and $Y(1) = Y(0) + 2$, where $X = (X_1, \cdots, X_d)^T$ with X_k's being i.i.d. following the $N(0,1)$ distribution and the $\mathrm{U}(-\sqrt{2}, \sqrt{2})$ distribution for the control units and the treated units, respectively. The independent noise $\varepsilon \sim N(0,1)$. Hence, the true ATE is $\Delta = 2$. To illustrate the universality of our method, we use both linear and nonlinear models for the potential outcomes. Specifically, we consider three cases: linear model as $m(u) = u$ and nonlinear models as $m(u) = 4 * \sqrt{|u+1|} + u$ and $m(u) = 2 * u + 10 * \exp(-u^2/5)$, respectively. Such nonlinear models are used in Zeng, He and Zhu (2012). We set the dimension of covariates as $d = 5$ and $d = 10$. The true index vector is $\beta_0 = (1,\, 0.7,\, -0.5,\, 0.25,\, 0.8)^\top$ for $d = 5$, and $\beta_0 = (1,\, 0.7,\, -0.5,\, 0.5,\, -0.75,\, 0.8,\, -0.4,\, 1,\, -0.2,\, 0.2)^\top$ for $d = 10$. The sample sizes are $(n_0,\, n_1) = (200, 100)$, $(400, 200)$, and $(800, 400)$.

In the simulation studies, the Gaussian kernel $K(v) = \frac{1}{\sqrt{2\pi}} \exp(-v^2/2)$ is used. We tried several different choices of the bandwidths and obtained similar results. For simplicity, we only present the results for bandwidths $h = 1 * n_0^{-1/3}$. The bandwidth h_1 used in the MAVE for estimating β_0 is chosen to minimize the mean integrated squared error as suggested in Xia et al. (2002). For each setting, we repeat the experiment 500 times independently. Both the root mean square error (RMSE) defined as $\mathrm{RMSE} = \left[\sum_{k=1}^{500} \left(\hat{\Delta}_k - \Delta\right)^2 / 500\right]^{1/2}$, and the mean of the 500 absolute deviation errors (MADE) are reported in

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

Table 1, compared with the results of the SCM in Abadie and L¡¯Hour (2021).

Table 1: Performance of SCM and QSCM

$m(u) = u$							
(n_0, n_1)		(200,100)		(400,200)		(800,400)	
	method	RMSE	MADE	RMSE	MADE	RMSE	MADE
$d = 5$	SCM	0.1599	0.1287	0.1202	0.0969	0.0950	0.0740
	QSCM	0.1297	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
$m(u) = 4 * \sqrt{\lvert u + 1 \rvert} + u$							
(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
$m(u) = 2 * u + 10 * \exp(-u^2/5)$							
(n_0, n_1)		(200,100)		(400,200)		(800,400)	
	method	RMSE	MADE	RMSE	MADE	RMSE	MADE
$d = 5$	SCM	1.7319	1.7037	1.8124	1.7967	1.9057	1.8963
	QSCM	0.1279	0.1012	0.0861	0.0678	0.0624	0.0495
$d = 10$	SCM	1.3674	1.3124	1.5454	1.5189	1.6481	1.6355
	QSCM	0.1319	0.1058	0.0897	0.0713	0.0632	0.0506

Based on the results in Table 1, one can see clearly that when the potential outcome model is linear, both methods perform well and our method is comparable to the SCM. However, when the potential outcome model is nonlinear, it is obvious that the SCM is invalid and our method performs much better. Furthermore, the finite sample performance of the QSCM is well-behaved in the sense that both the RMSE and MADE values are gen-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

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
$m(u) = 2 * u + 10 * \exp(-u^2/5)$						
(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	0.889	0.866	0.876	0.873	0.891	0.881
0.95	0.942	0.923	0.927	0.920	0.941	0.936
0.99	0.983	0.981	0.987	0.978	0.98	0.985

erally small. The RMSE and MADE values decrease as the sample size n_1 increases, and the convergence rate is in line with our expectation in the sense that, as in Theorem 1, the proposed estimator is $\sqrt{n_1}$-consistent.

Next, we consider the performance of the Bootstrap procedure proposed in Section 2.5. The number of Bootstrap replications is set as $B = 500$. For each setting, we repeat the experiment 1000 times independently and compare the sample coverage rates to the nominal coverage probabilities (NCP). The results are reported in Table 2, from which, we can see that, the proposed Bootstrap procedure has reasonably good estimated coverage probabilities.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

4.2 Evaluating QSCM With Variable Selection

Now, we conduct simulations to evaluate the effectiveness combining our QSCM estimator with the variable selection methods proposed in Section 3. We use the same models as in Section 4.1 except that the number of covariates is set as $d_{n_0} = \lfloor 60 * n_0^{1/6} \rfloor$ with the true index vector $\beta_0 = (1,\ 0.7,\ -0.5,\ 0.25,\ 0.8,\ 0, \ldots,\ 0)^\top$.

Following Bai, Rao and Wu (1999), we use BIC to choose the penalty parameter λ. We compare the RMSEs and MADEs of the estimators $\hat{\Delta}$ and $\hat{\Delta}_{\text{SCAD}}$ (pen-QSCM) based on 500 replications. We also evaluate the performance of the variable selection procedure by the mean of the true positive rates (TPR) and false positive rates (FPR), which are $\text{TPR} = \#\{1 \le j \le 5 : \hat{\beta}_{\text{SCAD},j} \neq 0\}/5$ and $\text{FPR} = \#\{6 \le j \le d_{n_0} : \hat{\beta}_{\text{SCAD},j} \neq 0\}\ /(d_{n_0} - 5)$ in our setting, respectively.

The results are presented in Table 3. We can see that $\hat{\Delta}_{\text{SCAD}}$ performs better than $\hat{\Delta}$ under all circumstances, indicating that the penalized method can effectively improve the performance of the QSCM. It is also observed that the results of variable selection are good. The true positive rates approach 1 and the false positive rates approach 0 as the sample sizes increase.

Finally, we consider an example with ultra-high dimensional covariates. We use the same same models as in Section 4.1, but the dimension of covariates is set as $d_{n_0} = 5 * n_0$ with the true index vector $\beta_0 = (1,\ 0.7,\ -0.5,\ 0.25,\ 0.8,\ 0, \ldots,\ 0)^\top$. In this case, we use the estimator $\hat{\Delta}_{\text{DC-RoSIS-SCAD}}$ proposed in Sec-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

Table 3: Performance of QSCM for $d_{n_0} = \lfloor 60 * n_0^{1/6} \rfloor$ 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
$m(u) = 2 * u + 10 * \exp(-u^2/5)$						
	QSCM		pen-QSCM		Variable Selection	
(n_0, n_1)	RMSE	MADE	RMSE	MADE	TPR	FPR
(200, 100)	1.7379	1.5918	0.7153	0.4071	0.9792	0.1833
(400, 200)	0.3672	0.2910	0.0912	0.0738	1.0000	0.0003
(800, 400)	0.0731	0.0582	0.0634	0.0506	1.0000	0.0003

tion 3.2. In the DC-RoSIS procedure, we choose $c = 1$ and $\kappa = 1/3$. Table 4 reports the RMSEs and MADEs of $\hat{\Delta}_{\text{DC-RoSIS-SCAD}}$ (DC-RoSIS-SCAD) based on 500 replications, along with the mean values of the TPR and FPR. We observe that the RMSE and MADE values are generally small and approximately decrease at a rate of $1/\sqrt{n_1}$, as desired.

From the above simulation, we can see clearly that if the true outcome model is within the class of linear model, both QSCM and SCM perform comparable. However, if the true model is from the class of nonlinear index model, the QSCM performs very well but the SCM fails. Finally, if the true

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

Table 4: Performance of QSCM for $d_{n_0} = 5 * n_0$ 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
$m(u) = 2 * u + 10 * \exp(-u^2/5)$				
	DC-RoSIS-SCAD		Variable Selection	
(n_0, n_1)	RMSE	MADE	TPR	FPR
(200, 100)	0.2066	0.1529	0.7988	0.0031
(400, 200)	0.1110	0.0870	0.8488	0.0003
(800, 400)	0.0728	0.0559	0.8896	0.0000

model is not from the class of an index model, in other words, the true model is mis-specified for both the QSCM and SCM, we conduct a simulation study to see how both methods perform. As a result, the simulation results conclude that the QSCM can still perform much better than that for the SCM, although both are inconsistent. The detailed model setting and simulation results are omitted here due to the space limitation and available upon a request.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

5. Revisit of the NSW Data

In this section, we study an empirical application by using our quasi synthetic control method to analyze the data from the National Supported Work Demonstration. The NSW program was a labor market program for underprivileged workers operated during the mid-1970s in the United States. By providing these workers with subsidized job for 9 to 18 months, the NSW program aimed to strengthen their job skills and enhance their employment opportunities. The NSW program randomly assigned the qualified applicants to the treated and control groups, making the program a randomized controlled trial, which is universally recognized as the golden standard to learn the treatment effect. This appealing feature of the NSW program motivates numerous researches.

LaLonde (1986) first analyzed the male sub-sample of the NSW program. In the Lalonde sample, the outcome of interest is the annual earnings in 1978. Additionally, the Lalonde sample also collects several individual characteristics: age, education, black, hispanic, married, no degree, and annual earnings in 1975. Dehejia and Wahba (1999) reorganized the Lalonde sample and collected the annual earnings in 1974. Excluding the individuals with the annual earnings in 1974 missed, the Dehejia-Wahba sample consists of $n_1 = 185$ treated units and $n_0 = 260$ control units, and the ATE estimate based on the Dehejia-Wahba sample is \$1794, termed as the experimental benchmark

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

value. Notice that the NSW program can be regarded as a randomized controlled trial. Consequently, the mean difference of the outcomes of the treated and control groups can serve as true value of the average treatment effect. For details on how to compute this benchmark value, one can refer to the papers by Dehejia and Wahba (1999) and Abadie and L¡¯Hour (2021).

Note that the Dehejia-Wahba sample has been widely used in many empirical studies. For example, Dehejia and Wahba (2002) applied the propensity score matching method to this dataset by using the Dehejia-Wahba sample. However, as pointed out by Smith and Todd (2005), estimates of the impact of NSW based on propensity score matching are highly sensitive to the set of variables included in the propensity score model, while Abadie and Imbens (2011) evaluated the performance of various matching estimators by analyzing the NSW data. For more literature on analyzing this dataset, the reader is referred to the paper by Abadie and L¡¯Hour (2021) and the references therein.

The Dehejia-Wahha sample is based on experimental data and provides us with an unbiased estimate of the ATE. To evaluate different estimators for treatment effects, it is recommended to use a non-experimental control group and estimate the treatment effect based on the experimental treated and non-experimental control groups. LaLonde (1986) constructed six non-experimental control groups from the Panel Study of Income Dynamics (PSID) and the Current Population Survey, as well as further subsets subtracted from

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

these two basic control groups. Referring to the existing literature, we use the experimental treated group from the Dehejia-Wahha sample ($n_1 = 185$) and the control group from the PSID ($n_0 = 2490$) to illustrate our quasi synthetic control method. The outcome variable is the annual earnings in 1978 and 10 covariates are considered. We present the summary statistics of the data used in our analysis in the Supplementary Material.

First, we would like to see if there exists a nonlinear relationship between the outcome and the index. To do so in a visual way, using data from PSID group, we plot the outcome Y_0 (y-axis) versus the estimated single index Z (x-axis) in Figure 1, together with a nonparametric estimate (*lowess* in R, locally-weighted polynomial regression technique) of the unknown function $m(\cdot)$ in the dashed line (with its pointwise 95% confidence interval presented by the shaded area), and a least-squares fitting of $m(\cdot)$ in the solid line. From Figure 1, it is clear that there does exist a nonlinear relationship between Y_{0i} and Z_i and this supports strongly that our nonlinear model is appropriate for this real data.

Now, to compute the QSCM estimator $\hat{\Delta}$, as in Monte Carlo simulations, we use the Gaussian kernel and the bandwidth $h = 0.23$, which is chosen through cross-validation to minimize the mean squared error (MSE) of estimating Y_{0j} for the control units. We compare our quasi synthetic control estimator with a series of existing estimators: the conventional synthetic control estimator (SCM), the penalized synthetic control estimator which minimizes

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

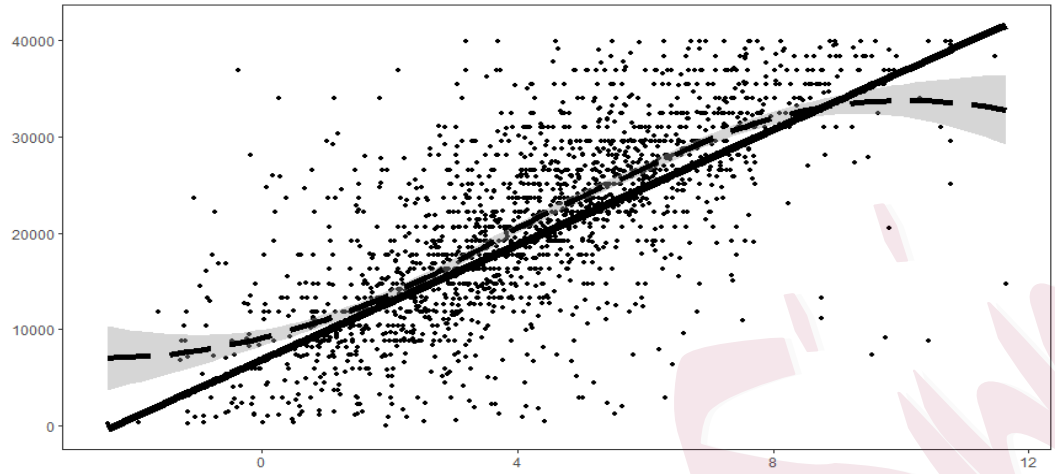

Figure 1: Scatterplot of Y_0 versus Z in PSID group, together with the estimate of the unknown function $m(\cdot)$ in the dashed line with its pointwise 95% confidence interval presented by the shaded area and a least-squares fitting of $m(\cdot)$ in the solid line.

the bias (pen-SCM) as in Abadie and L¡¯Hour (2021), and the one-match nearest neighbor matching estimator (1-Matching). Table 5 reports the empirical results. These four estimators yield treatment effects ranging from

Table 5: Estimated ATTs for the NSW data.

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

Note: The QSCM estimate is computed based on the bandwidth $h = 0.23$, which is chosen through cross-validation to minimize the MSE of estimating Y_{0j} for the control units. The result for pen-SCM come from Abadie and L¡¯Hour (2021), and the result for 1-Matching is computed via the R package *Matching* by Sekhon and Saarinen (2023).

\$1801.22 to \$2138.8. Given the experimental benchmark Δ = \$1794.34, our QSCM estimator is the best in the sense that it has the smallest bias from the benchmark value. This result indicates that our method effectively captures the unknown features of the NSW data with a possible nonlinear relationship

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

between Y_{0i} and Z_i. We also compute the standard error of the QSCM estimator using the hybrid Bootstrap method proposed in Section 2.5 and the standard error of $\hat{\Delta}_{\text{QSCM}}$ 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). Since the statistical inference for the SCM and the pen-QSCM has not been discussed yet, we do not compute the standard error of the SCM estimate and the pen-SCM estimate.

Finally, it is also interesting to note that in this empirical example, the conventional SCM needs to optimize a 2490×1 vector of weights for each of the total 185 treated units, which is computationally expensive in practice. Our computations were carried out on an 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 took 1.69 hours to compute the conventional SCM estimate. In contrast, the computation time for our QSCM estimate is 13.6 seconds without parallel computation. To assess this phenomenon, indeed, as pointed out by Abadie and L¡¯Hour (2021), the best synthetic control may not be unique with many control units. Therefore, searching for the best synthetic control involves heavy computing, and our method can significantly reduce the computation time.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

6. Conclusion

The SC method is a popular and powerful approach to estimating ATE, as addressed by Athey and Imbens (2017). However, as pointed out in the literature, the SC methods have some shortcomings. To overcome these difficulties, this paper proposes a QSC method, which can accommodate nonlinearity and feature fast computing. In particular, this article provides the inference theory for the QSC method, and we derive the asymptotic distribution of the QSC ATE estimators with and without a penalty term. Also, due to the complex structure of the asymptotic variances of the proposed estimators, we resolve this difficulty by proposing a carefully designed and easy-to-implement Bootstrap method and establish the validity of the subsampling method for inference. Our work complements the conventional SC method and its variants. In addition, our simulations show that the QSC method performs well in practice. Finally, we apply the QSC method to estimate ATE for the NSW data. The empirical application demonstrates that when the conventional SC method fits the data poorly, the QSC method can fit the data well and provide reasonable ATE estimation results.

Finally, in addition to the aforementioned future research topics, it is worth to note that under the current framework, one might extend easily the proposed methodology to estimate the quantile (distributional) treatment effects as investigated in Cai et al. (2022), which is warranted as a future

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

research topic.

Supplementary Material

The online Supplementary Material contains proofs of Theorems 1-3, the algorithm for the MAVE method, and summary statistics of the empirical data.

Acknowledgments

We thank the Co-Editor (Professor Huixia Judy Wang), the Associate Editor, and three anonymous referees for their constructive and helpful comments and suggestions that improved significantly the quality of the paper. Also, the authors gratefully acknowledge the financial supports, in part, from the National Science Fund of China (NSFC) key project grants #72033008 and #72133002, Basic Science Center Program of NSFC with grant #71988101.

Disclosure Statement

The authors claim that there are no relevant financial or non-financial competing interests to report for this article. Also, the authors declare that they do not use any generative AI and AI-assisted technologies in the writing process.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

References

Abadie, A. and Gardeazabal, J. (2003). The economic costs of conflict: A case study of the Basque country. *American Economic Review*, 93(1):113-132.

Abadie, A. and Imbens, G.W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235-267.

Abadie, A. and Imbens, G.W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1):1-11.

Abadie, A. and L'Hour, J. (2021). A penalized synthetic control estimator for disaggregated data. *Journal of the American Statistical Association*, 116(536):1817-1834.

Athey, S. and Imbens, G.W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(1):3-32.

Bai, Z.D., Rao, C. R., & Wu, Y. (1999). Model selection with data-oriented penalty. *Journal of Statistical Planning and Inference*, 77(1):103-117.

Cai, Z., Das, M., Xiong, H. and Wu, X. (2006). Functional coefficient instrumental variables models. *Journal of Econometrics* 133(1):207-241.

Cai, Z., Juhl, T. and Yang, B. (2015). Functional index coefficient models with variable selection. *Journal of Econometrics*, 189(2):272-284.

Cai, Z., Fang, Y., Lin, M. and Zhan, M. (2022). Estimating quantile treatment effects for panel data. *Working Paper*, Department of Economics, University of Kansas.

Chiburis, R. C. (2010). Semiparametric bounds on treatment effects. *Journal of Econometrics*, 159(2):267-275.

Dehejia, R.H., and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448):1053-1062.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

Dehejia, R.H. and Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, 84(1):151-161.

Fan, J., Feng, Y., & Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494):544-557.

Fan, J. and Gijbels, I. (1996). Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability. CRC Press.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348-1360.

Fan, J., Samworth, R., & Wu, Y. (2009). Ultrahigh dimensional feature selection: Beyond the linear model. *Journal of Machine Learning Research*, 10:2013-2038.

Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(5):849-911.

Fan, J., Samworth, R., & Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *The Journal of Machine Learning Research*, 10, 2013-2038.

Frölich, M. (2004). Finite-sample properties of propensity-score matching and weighting estimators. *Review of Economics and Statistics*, 86(1):77-90.

Galvao, A.F. and Wang, L. (2015). Uniformly semiparametric efficient estimation of treatment effects with a continuous treatment. *Journal of the American Statistical Association*, 110(512):1528-1542.

Heckman, J.J., Ichimura, H. and Todd, P. (1998). Matching as an econometric evaluation estimator. *Review of Economic Studies*, 65(2):261-294.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

Hsiao, C., Ching, S. and Wan, S.K. (2012). A panel data approach for program evaluation: Measuring the benefits of political and economic integration of Hong kong with mainland China. *Journal of Applied Econometrics*, 27(5):705-740.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, 58:71-120.

Kang, J.D.Y and Schafer, J.L. (2007). Demystifying double robustness: A Comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523-539.

Kong, E. and Xia, Y. (2007). Variable selection for the single-index model. *Biometrika*, 94(1):217-229.

LaLonde, R.J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76(4):604-620.

Li, K.T. and Bell, D.R. (2017). Estimation of average treatment effects with panel data: Asymptotic theory and implementation. *Journal of Econometrics*, 197(1):65-75.

Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129-1139.

Otsu, T. and Rai, Y. (2017). Bootstrap inference of matching estimators for average treatment effects. *Journal of the American Statistical Association*, 112(520):1720-1732.

Park, H., Petkova, E., Tarpey, T. and Ogden, R.T. (2021). A constrained single index regression for estimating interactions between a treatment and covariates. *Biometrics*, 77(2):506-518.

Peng, H. and Huang, T. (2011). Penalized least squares for single index models. *Journal of Statistical Planning and Inference*, 141(4):1362-1379.

Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688-701.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

Sekhon, J.S. and Saarinen T. (2023). Matching: Multivariate and Propensity Score Matching with Balance Optimization. *R package version 4.10-14*.

Smith, J.A. and Todd, P.E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators?. *Journal of Econometrics*, 125:305-353.

Song, K. (2014). Semiparametric models with single-index nuisance parameters. *Journal of Econometrics*, 178:471-483.

Sun, Y., Yan, K.X. and Li, Q. (2021). Estimation of average treatment effect based on a semiparametric propensity score. *Econometric Reviews*, 40(9):852-866.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B*, 58(1):267-288.

Wang, T., Xu, P. and Zhu, L. (2013). Penalized minimum average variance estimation. *Statistica Sinica*, 23(2):543-569.

Wang, Q. and Yin, X. (2008). A nonlinear multi-dimensional variable selection method for high dimensional data: Sparse MAVE. *Computational Statistics & Data Analysis*, 52(9):4512-4520.

Wu, Y., Ren, T. and Mu, L. (2016). Importance reweighting using adversarial-collaborative training. *Neural Information Processing Systems, 2016 Workshop on Adversarial Training*, 1-6.

Xia, Y. (2006). Asymptotic distributions for two estimators of the single-index model. *Econometric Theory*, 22(6):1112-1137.

Xia, Y., Tong, H., Li, W.K. and Zhu, L. (2002). An adaptive estimation of dimension reduction space (with discussions). *Journal of the Royal Statistical Society, Series B*, 64(2):363-410.

Zeng, P., He, T. and Zhu, Y. (2012). A LASSO-type approach for estimation and variable selection in single index models. *Journal of Computational and Graphical Statistics*, 21(1):92-109.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0271

Zhang, H., Huang, L. and Liu, L. L. (2020). On Bootstrap consistency of MAVE for single index models. *Computational Statistics & Data Analysis*, 141:28-39.

Zhong, W., Zhu, L., Li, R. and Cui, H. (2016). Regularized quantile regression and robust feature screening for single index models. *Statistica Sinica*, 26(1): 69-95.