Statistics and Its Interface Volume 15 (2022) 283–301

Empirical likelihood-based estimation and inference in randomized controlled trials with high-dimensional covariates

Wei Liang and Ying Yan*

In this paper, we propose a data-adaptive empirical likelihood-based approach for treatment effect estimation and inference, which overcomes the obstacle of the traditional empirical likelihood-based approaches in the high-dimensional setting by adopting penalized regression and machine learning methods to model the covariate-outcome relationship. In particular, we show that our procedure successfully recovers the true variance of Zhang's treatment effect estimator [30] by utilizing a data-splitting technique. Our proposed estimator is proved to be asymptotically normal and semiparametric efficient under mild regularity conditions. Simulation studies indicate that our estimator is more efficient than the estimator proposed by Wager et al. [26] when random forest is employed to model the covariate-outcome relationship. Moreover, when multiple machine learning models are imposed, our estimator is at least as efficient as any regular estimator with a single machine learning model. We compare our method to existing ones using the ACTG175 data and the GSE118657 data, and confirm the outstanding performance of our approach.

Keywords and phrases: Average treatment effect, Data-splitting, Machine learning, Multiple robustness, Semiparametric efficiency bound.

1. INTRODUCTION

Randomized controlled trials (RCTs) are recognized as the standard clinical design to eliminate sources of confounding bias. When the outcome of interest is a continuous variable, the difference of mean responses in the treatment and the control groups is an unbiased and consistent estimator for the average treatment effect (ATE), a commonly used estimand to evaluate the effect of a treatment or policy. When the baseline information is collected before assigning the treatment, such as age, sex, and other characteristics, adjusting for the pre-treatment covariates helps to improve the efficiency of the ATE estimator.

The key of covariate adjustment is to explore the relationship between the auxiliary covariates and response. Analysis of covariance (ANCOVA) is a classical regression method for covariate adjustment where a linear regression model for $\mathrm{E}[Y|X,D]$ is postulated, i.e.,

$$\mathrm{E}[Y|X,D] = \beta_0 + \beta_x^\top X + \beta_d D. \tag{1}$$

Here, Y is the outcome variable, X is the vector of covariates and D is the binary treatment indicator variable. The parameter of interest, the unconditional population-level treatment effect, is β_d. We can then make inference about ATE based on the asymptotic normality of the least square estimator, $\widehat{\beta}_d^{\mathrm{ols}}$, of β_d in (1) [12]. It follows from Leon, Tsiatis and Davidian [15] and Tsiatis et al. [25] that $\widehat{\beta}_d^{\mathrm{ols}}$ belongs to the class of all regular and asymptotically linear estimators, and more efficient estimators in this class can be obtained by positing two separate working regression models for $\eta^{(1)}(x) = \mathrm{E}[Y|X=x, D=1]$ and $\eta^{(0)}(x) = \mathrm{E}[Y|X=x, D=0]$, respectively.

Empirical likelihood (EL) is an alternative way to carry out covariate adjustment. EL was introduced by Owen [17, 18, 19] and primarily used to construct confidence intervals for the mean or to infer the parameters in the general estimating functions [20]. It has also been adopted as a tool to efficiently incorporate information of auxiliary covariates in causal inference problems. Huang, Qin and Follmann [11] used EL to derive efficient estimators in the pretest–posttest study with missing data. Qin and Zhang [21] and Zhang [29] studied EL-based estimation in the observational study. In particular, when multiple parametric regression models are imposed into constraints, the EL estimator has good performance as long as one of multiple models correctly specifies the covariate-outcome relationship without requiring the knowledge of which model is correct. This is known as the multiple robustness property [10]. Recently, Zhang [30] and Tan et al. [22] extended EL for statistical estimation and inference of ATE in RCTs. Their EL estimators enjoy some prominent advantages. For example, Zhang [30] proved that his EL estimator was at least as efficient as the existing regular estimators when the parametric models for the covariate-outcome relationship were mis-specified, and asymptotically as efficient as the semiparametric estimator of Tsiatis et al. [25] when the parametric models for the covariate-outcome relationship were correctly specified. As shown in the simulation studies of Zhang [30] and Tan et al. [22], both EL

*Corresponding author.

estimators were considerably more efficient than the estimator of Tsiatis et al. [25] when the imposed parametric regression models were mis-specified. In addition, Tan et al. [22] proved that the multiple robustness property of the EL estimator could be maintained in RCTs.

In practice, the true covariate-outcome model is unknown, which can be much more complicated than a simple linear combination of several variables as displayed in equation (1). Furthermore, in the big data era, the number of features may be high-dimensional, where ANCOVA and other traditional methods are no longer directly applicable. It inspires us to model the highly complex covariate-outcome relationship by modern machine learning (ML) methods, such as Lasso [23], SCAD [6], and random forests [2]. A general semiparametric framework for statistical inference of treatment effects under which infinite-dimensional nuisance parameters are modelled with ML methods is given by Chernozhukov et al. [4] and Belloni et al. [1], where two crucial points were presented:

1. They used Neyman orthogonal scores to remove the bias brought by regularization.
2. They split data to avoid overfitting.

Specifically, the Neyman orthogonal scores technique adjusts for the effect of covariates to reduce sensitivity with respect to the nuisance parameters, and thus promotes the efficiency of treatment effect estimation. It is straightforward to show that the score function developed by Tsiatis et al. [25] is Neyman orthogonal in RCTs. With an additional data-splitting procedure, Wager et al. [26] generalized the results of Tsiatis et al. [25] to the high-dimensional setting and adopted ML methods to model the covariate-outcome relationship. Under mild regularity conditions, they proposed valid inference of ATE by integrating the data-splitting procedure.

EL and Neyman orthogonal scores play similar roles in RCTs as they both achieve the goal of efficiency improvement of treatment effect estimation by incorporating information of auxiliary covariates. However, the estimator proposed by Wager and his colleagues does not enjoy some unique properties of EL, e.g., multiple robustness. When the single ML algorithm adopted by Wager et al. [26] does not successfully capture the covariate-outcome relationship, their ATE estimation may incur efficiency loss. Motivated by embedding the benefits of EL-based approaches with low-dimensional covariates in the high-dimensional setting, we propose a Machine Learning and Data-splitting-based Empirical Likelihood (MDEL) approach to estimate ATE, where we apply multiple ML algorithms to model the covariate-outcome relationship. Compared with the high-dimensional regression adjustment approach of Wager et al. [26], our proposed EL approach has the following advantages:

1. When the single ML estimator of nuisance parameters does not perform well, our proposed EL estimator is more efficient, as indicated by our simulation studies.
2. Different estimators of the nuisance parameters can be imposed simultaneously into constraints to enhance the performance of our estimator. Our simulation studies indicate that our EL estimator with multiple models tends to perform as good as that with the correct model without requiring the knowledge of which model is correct.

Our paper is organized as follows. In Section 2, we give a brief introduction to the model setup and notations. In addition, we review the semiparametric methods by Tsiatis et al. [25] and Wager et al. [26]. In Section 3, we introduce our proposed empirical likelihood approach. Then we discuss the practical implementation of our EL approach in Section 4. In Section 5, we compare our proposed EL approach to the existing ones in simulation studies, the ACTG175 data set, and the GSE118657 data set.

2. NOTATIONS AND REVIEWS

2.1 The model setup

We introduce our model setup under the potential outcome framework of Imbens and Rubin [12]. Suppose we have n observations $\{W_i = (Y_i, X_i, D_i), i = 1, \cdots, n\}$ from a binary experiment with the treatment indicator variable $D_i \in \{0, 1\}$. D_i takes the value 1 if the i-th unit is assigned to the treatment group and 0 if the i-th unit is assigned to the control group. Assume $Y_i(d)$ is the potential outcome under $D_i = d$ for $d = 0, 1$. The observed outcome of the i-th unit, Y_i, satisfies $Y_i = D_iY_i(1) + (1 - D_i)Y_i(0)$. The variable X_i is the covariates of the i-th unit and offers us pre-treatment information. We are interested in estimation and inference of the population level ATE, defined by $\theta = \mathrm{E}[Y_i(1) - Y_i(0)]$, under the assumption that $\{W_i, i = 1, \cdots, n\}$ are independent and identically distributed random samples from a random vector $W = (Y, X, D)$. In this paper, we focus on randomized controlled trials, where D_i is randomly assigned to either 0 or 1 and is independent of all pre-treatment variables and the potential outcomes, i.e.,

$$D_i \perp\!\!\!\perp \{Y_i(1), Y_i(0), X_i\} \quad \text{for} \quad i = 1, \cdots, n.$$

Let $\delta = \mathrm{P}(D = 1)$ be the probability of a unit being assigned to the treatment group. We further assume that $0 < \delta < 1$. Randomness of treatment assignments leads to that $\theta = \mathrm{E}[Y|D = 1] - \mathrm{E}[Y|D = 0]$, which leads to a natural and commonly used consistent estimator of ATE, the difference in the means, defined by

$$\widehat{\theta}_{\mathrm{dim}} = \bar{Y}^{(1)} - \bar{Y}^{(0)} = \sum_{i=1}^{n} \frac{D_i Y_i}{n_1} - \sum_{i=1}^{n} \frac{(1 - D_i)Y_i}{n_0},$$

where n_1 is the size of the treatment group and $n_0 = n - n_1$ is the size of the control group. $\widehat{\theta}_{\mathrm{dim}}$ ignores information of covariates and thus loses efficiency. A popular regression adjustment approach Tsiatis et al. [25] was proposed to exploit information of covariates and enhance the efficiency, and we review it in the following section.

2.2 A regression adjustment approach

Various regression adjustment methods for improving the efficiency of the treatment effect estimation in RCTs with low-dimensional covariates have emerged. Here, we only review the method of Tsiatis et al. [25], which can be reformulated as a method based on the efficient score. Wager et al. [26] generalized it to the high-dimensional case. For other methods, we refer to Zhang [30] and Tan et al. [22]. Different from ANCOVA, the approach of Tsiatis et al. [25] focuses on separately modelling the covariate-outcome relationships $\eta^{(d)}(x) = \mathrm{E}[Y|D = d, X = x], d = 0, 1$. By fitting $\eta^{(1)}(x)$ and $\eta^{(0)}(x)$ with two different parametric models $f_1(x, \alpha_1)$ and $f_0(x, \alpha_0)$, specified by two finite-dimensional parameters, α_1 and α_0, respectively, Tsiatis et al. [25] proposed to estimate θ with

$$\begin{aligned}\widehat{\theta}_{\mathrm{tdzl}} =& \bar{Y}^{(1)} - \bar{Y}^{(0)} - \sum_{i=1}^{n}\left(D_i - \frac{n_1}{n}\right)\Big\{n_0^{-1}f_0(X_i,\widehat{\alpha}_0) \\ &+ n_1^{-1}f_1(X_i,\widehat{\alpha}_1)\Big\},\end{aligned}$$

where $\widehat{\alpha}_0$ and $\widehat{\alpha}_1$ are estimators of α_0 and α_1, respectively. For example, $\hat{\alpha}_d$ can be the least square estimators or the stepwise regression estimators. Write $\widehat{f}_d(\cdot) = f(\cdot, \widehat{\alpha}_d), d = 0, 1$. The semiparametric theory [24] indicates that the efficient score of θ is given by

$$\begin{aligned}&\varphi(W, \theta, \delta, \eta^{(1)}, \eta^{(0)}) \\ =&\frac{D}{\delta}\left(Y - \eta^{(1)}(X)\right) - \frac{1-D}{1-\delta}\left(Y - \eta^{(0)}(X)\right) \\ &+ \eta^{(1)}(X) - \eta^{(0)}(X) - \theta.\end{aligned}$$

Here, $\eta^{(1)}$ and $\eta^{(0)}$ are treated as nuisance parameters, and θ is the parameter of interest. $\widehat{\theta}_{\mathrm{tdzl}}$ can be reformulated as the solution of

$$\frac{1}{n}\sum_{i=1}^{n}\varphi(W_i, \theta, \widehat{\delta} = \frac{n_1}{n}, \widehat{f}_1, \widehat{f}_0) = 0.$$

The semiparametric estimator $\widehat{\theta}_{\mathrm{tdzl}}$ reaches the semiparametric efficiency bound when both f_1 and f_0 correctly specify the true covariate-outcome relationships.

2.3 High-dimensional regression adjustment

Wager et al. [26] extended the approach of Tsiatis et al. [25] to the high-dimensional case. The nuisance parameters were proposed to be estimated using ML methods and a data-splitting procedure was adopted for valid inference with high-dimensional covariates. Let $\mathbb{I} = \{1, \cdots, n\}$ be the sample index set, $\mathbb{I}^{(1)}$ and $\mathbb{I}^{(0)}$ the index set of the treatment group and control group, respectively. We use the notation $|\mathbb{A}|$ as the size of a set $\mathbb{A}$. Suppose we randomly partition $\mathbb{I}^{(d)}$ into K subsets with equal size, denoted by $(\mathbb{I}_k^{(d)})_{k=1}^K$, for $d = 0, 1$. Let $\mathbb{I}_k = \mathbb{I}_k^{(1)} \cup \mathbb{I}_k^{(0)}$, ${\mathbb{I}_k^{(d)}}^c = \mathbb{I}^{(d)}\backslash\mathbb{I}_k^{(d)}$ and $\mathbb{I}_k^c = \mathbb{I}\backslash\mathbb{I}_k$. Generally, we set $\frac{|\mathbb{I}_k^{(d)}|}{|\mathbb{I}_k|} = \frac{n_d}{n}$ and $\frac{|\mathbb{I}_k^{(d)}|}{|\mathbb{I}^{(d)}|} = \frac{1}{K}$.

After data-splitting, Wager et al. [26] proposed to estimate θ with $\widehat{\theta}_{\mathrm{wdtt}} = \frac{1}{K}\sum_{k=1}^{K}\widehat{\theta}_{\mathrm{wdtt}}^k$, where the k-th sub-estimator, $\widehat{\theta}_{\mathrm{wdtt}}^k$, is the solution of

$$\frac{1}{|\mathbb{I}_k|}\sum_{i\in\mathbb{I}_k}\varphi(W_i, \theta, \widehat{\delta} = \frac{|\mathbb{I}_k^{(1)}|}{|\mathbb{I}_k|}, \widehat{g}_k^{(1)}, \widehat{g}_k^{(0)}) = 0.$$

Here, for fixed k and d, $\widehat{g}_k^{(d)}$ is an ML estimator of $\eta^{(d)}$ obtained via the sub-sample $(W_i)_{i\in{\mathbb{I}_k^{(d)}}^c} = \left\{W_i | i \in {\mathbb{I}_k^{(d)}}^c\right\}$. It follows immediately that conditional on the sample $(W_i)_{i\in{\mathbb{I}_k^{(d)}}^c}$, $\widehat{g}_k^{(d)}(x)$ is a non-random function of x. Therefore, the variance of $\widehat{\theta}_{\mathrm{wdtt}}$ can be directly estimated by

$$\widehat{\mathrm{Var}}(\widehat{\theta}_{\mathrm{wdtt}}) = \sum_{k=1}^{K}\frac{|\mathbb{I}_k|^2}{n^2}\widehat{\mathrm{Var}}(\widehat{\theta}_{\mathrm{wdtt}}^k),$$

where for a fixed k, $\widehat{\mathrm{Var}}(\widehat{\theta}_{\mathrm{wdtt}}^k)$ is a moment-based plug-in estimator for the conditional variance of $\widehat{\theta}_{\mathrm{wdtt}}^k$,

$$\begin{aligned}\sum_{d=0,1}\frac{1}{|\mathbb{I}_k^{(d)}|}\mathrm{Var}\Bigg[Y - \frac{|\mathbb{I}_k^{(0)}|}{|\mathbb{I}_k|}\widehat{g}_k^{(1)}(X) - \frac{|\mathbb{I}_k^{(1)}|}{|\mathbb{I}_k|}\widehat{g}_k^{(0)}(X)\Bigg| \\ \widehat{g}_k^{(1)}, \widehat{g}_k^{(0)}, D = d\Bigg].\end{aligned}$$

Wager et al. [26] demonstrated that $\frac{(\widehat{\theta}_{\mathrm{wdtt}} - \theta)}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\theta}_{\mathrm{wdtt}})}}$ is asymptotically standard normal under certain regularity conditions. Therefore, for statistical inference, the corresponding $1 - \alpha$ confidence interval for θ is given by

$$\left(\widehat{\theta}_{\mathrm{wdtt}} - z_{\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{Var}}(\widehat{\theta}_{\mathrm{wdtt}})}, \widehat{\theta}_{\mathrm{wdtt}} + z_{\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{Var}}(\widehat{\theta}_{\mathrm{wdtt}})}\right),$$

where $z_{\frac{\alpha}{2}}$ is the upper $\frac{\alpha}{2}$ quantile of the standard normal distribution.

3. EMPIRICAL LIKELIHOOD-BASED APPROACHES IN RCTS

3.1 Traditional EL based approaches in RCTs

Let $f(x) = (f_1(x), f_0(x))^\top$ be a vector function of x and $\xi = \mathrm{E}[f(X)]$. Based on two unbiased estimating functions

$$h_1(D, Y, \theta, \delta) = \frac{DY}{\delta} - \frac{(1-D)Y}{1-\delta} - \theta$$

and

$$h_2(D,X,\delta,f,\xi)=\frac{D-\delta}{\delta(1-\delta)}\{f(X)-\xi\},$$

Zhang [30] proposed to estimate θ by maximizing the nonparametric likelihood $L_F=\prod_{i=1}^n p_i$ subject to constraints $\sum_{i=1}^n p_i=1$, $p_i\geq 0$ and $\sum_{i=1}^n p_i\left(h_1(D_i,Y_i,\theta,\hat\delta),h_2^\top(D_i,X_i,\hat\delta,\hat f,\hat\xi)\right)^\top=0$, where $\hat\delta=\frac{n_1}{n}$, $\hat\xi=\frac{1}{n}\sum_{i=1}^n\hat f(X_i)$, $\hat f=(\hat f_1,\hat f_0)^\top$, and $\hat f_1$ and $\hat f_0$ are estimated working regression models for $\eta^{(1)}$ and $\eta^{(0)}$, respectively. The variance of $\widehat\theta_{\text{Zhang}}$ was estimated by a sandwich variance estimator. When the covariates are of low-dimension and f_d is a correctly specified parametric regression model for $\eta^{(d)}$, $d=1,0$, $\widehat\theta_{\text{Zhang}}$ is semiparametric efficient. However, when the covariates are of high-dimension and $\hat f_d$ is an ML estimator for model selection, many spurious variables which have high correlations with the response but do not belong to the true feature set will be selected and thus result in serious underestimation of the variance [5]. We conduct simulations to illustrate this point in Section 3.3.

Tan et al. [22] extended the two-sample EL approach of Wu and Yan [28] and proposed to estimate θ based on the property $\mathrm{E}[f(X)-\xi|D=d]=0$. Their estimator is $\widehat\theta_{\text{Tan}}=\sum_{i\in\mathbb{I}^{(1)}}\hat p_iY_i-\sum_{i\in\mathbb{I}^{(0)}}\hat p_iY_i$, where for $d=1,0$, the $\hat p_i$'s are obtained by maximizing the nonparametric likelihood $L_F=\prod_{i\in\mathbb{I}^{(d)}}p_i$ subject to constraints $\sum_{i\in\mathbb{I}^{(d)}}p_i=1,p_i\geq 0$ and $\sum_{i\in\mathbb{I}^{(d)}}p_i\hat f_d(X_i)=\frac{1}{n}\sum_{j=1}^n\hat f_d(X_j)$. Here $\hat f_d(x)$ is a guess of $\mathrm{E}[Y|X=x,D=d]$. Multiple guesses are allowed in Tan's method. Estimation for the variance of $\widehat\theta_{\text{Tan}}$ is given by the bootstrap method. Tan's approach is simple and easy to explain. Asymptotic theory and simulation studies of Tan et al. [22] verify its multiple robustness property, which means that the estimator achieves the semiparametric efficiency bound as long as one model of f_d is correctly specified. However, when $\hat f_d$ involves ML estimators, their proposed bootstrap re-sampling procedure is no longer applicable as Donsker conditions are inappropriate when the space of $\hat f_d$ is highly complicated.

Both EL approaches by Zhang [30] and Tan et al. [22] have desirable properties in the low-dimensional setting but fail to make valid inference in the high-dimensional setting. To maintain multiple robustness and other nice properties of EL estimators, as well as to overcome the invalid inference problem of traditional EL approaches, we are motivated to extend the approach of Tan et al. [22], which is very simple to implement, to RCTs with high-dimensional covariates by means of machine learning and data-splitting.

3.2 The proposed EL-based estimator with high-dimensional covariates

In our proposed approach, the nuisance parameters are allowed to be estimated using multiple ML methods. For $d=0,1$, assume we already have an r-dimensional vector of estimators of $\eta^{(d)}$, denoted by $\widehat g_k^{(d)}=\left(\widehat g_{k,1}^{(d)},\cdots,\widehat g_{k,r}^{(d)}\right)^\top$, where each component of $\widehat g_k^{(d)}$ is an ML estimator such as the random forests estimator or Lasso estimator of $\eta^{(d)}$ based on the sub-sample $(W_i)_{i\in\mathbb{I}_k^{(d)c}}$. Let $\widehat\xi^{(d)}=\frac{1}{n}\sum_{k=1}^K\sum_{i\in\mathbb{I}_k}\widehat g_k^{(d)}(X_i)$ and $\widehat\xi_k^{(d)}=\frac{1}{|\mathbb{I}_k|}\sum_{i\in\mathbb{I}_k}\widehat g_k^{(d)}(X_i)$ for $k=1,\cdots,K$ and $d=0,1$. It is easy to check that $\widehat\xi^{(d)}=\frac{1}{K}\sum_{k=1}^K\widehat\xi_k^{(d)}$. Let $k_d(y,x)$ be the conditional joint density of (Y,X) given $D=d$ and $p_i=k_d(Y_i,X_i)$, $i\in\mathbb{I}^{(d)}$ for $d=0,1$. Due to randomization, conditional on the sub-sample $(W_i)_{i\in\mathbb{I}_k^{(d)c}}$, we have

$$\mathrm{E}\left[\widehat g_k^{(d)}(X)\,\middle|\,D=d\right]=\mathrm{E}\left[\widehat g_k^{(d)}(X)\right]$$

for $k=1,\cdots,K$ and $d=0,1$. It leads to

$$\sum_{k=1}^K\sum_{i\in\mathbb{I}_k^{(d)}}p_i\widehat g_k^{(d)}(X_i)=\frac{1}{K}\sum_{k=1}^K\widehat\xi_k^{(d)}\tag{2}$$

for $d=0,1$. Based on (2) and $\iint k_d(y,x)dydx=1$, we impose the following constraints on p_i:

$$\begin{aligned}&\sum_{i\in\mathbb{I}^{(d)}}p_i=1,\quad d=0,1,\quad p_i\geq 0,\quad\forall i\in\mathbb{I},\\&\sum_{k=1}^K\sum_{i\in\mathbb{I}_k^{(d)}}p_i\widehat g_k^{(d)}(X_i)=\frac{1}{K}\sum_{k=1}^K\widehat\xi_k^{(d)},\quad d=0,1.\end{aligned}\tag{3}$$

We propose to estimate p_i by maximizing the nonparametric likelihood $L_k=\prod_{i\in\mathbb{I}^{(1)}}p_i\prod_{i\in\mathbb{I}^{(0)}}p_i$ subject to the constraints (3), which is equivalent to solving two separated minimization problems:

$$\begin{aligned}&\min_{\{p_i\geq 0,i\in\mathbb{I}^{(d)}\}}-\sum_{i\in\mathbb{I}^{(d)}}\log(p_i)\\&\quad\text{s.t.}\quad\sum_{i\in\mathbb{I}^{(d)}}p_i=1,\\&\qquad\quad\sum_{k=1}^K\sum_{i\in\mathbb{I}_k^{(d)}}p_i\left(\widehat g_k^{(d)}(X_i)-\widehat\xi^{(d)}\right)=0,\end{aligned}\tag{4}$$

for $d=0,1$. Let $\widehat{\mathrm{G}}\left(x,\widehat g_k^{(d)},\widehat\xi^{(d)}\right)=\widehat g_k^{(d)}(x)-\widehat\xi^{(d)}$. The Lagrange multiplier method shows that the dual problem of (4) is

$$\begin{aligned}\max_{\lambda_d\in\mathcal{A}_n^d}\ell(\lambda_d)=&\sum_{k=1}^K\sum_{i\in\mathbb{I}_k^{(d)}}\log\left\{1+\lambda_d^\top\widehat{\mathrm{G}}\left(X_i,\widehat g_k^{(d)},\widehat\xi^{(d)}\right)\right\}\\&+n_d\log n_d\end{aligned}\tag{5}$$

where

$$\mathcal{A}_n^d = \left\{ x \in \mathbb{R}^r : 1 + x^\top \widehat{\mathrm{G}}\left(X_i, \widehat{g}_k^{(d)}, \widehat{\xi}^{(d)}\right) > 0, i \in \mathbb{I}^{(d)} \right\}$$

and $\hat{p}_i$ is given by

$$\widehat{p}_i = \frac{1}{n_d}\left\{1 + \widehat{\lambda}_d^\top \widehat{\mathrm{G}}\left(X_i, \widehat{g}_k^{(d)}, \widehat{\xi}^{(d)}\right)\right\}^{-1}$$

for $i \in \mathbb{I}_k^{(d)}$, $k = 1, \cdots, K$, where $\widehat{\lambda}_d$ is the solution of (5). The existence and uniqueness of the solution of (5) could be guaranteed by two conditions [9, 3]: (b1) The strict concavity of $\ell(\lambda_d)$ over the domain $\mathcal{A}_n^d$; (b2) The zero vector is an interior point of the convex hull of $\left\{\widehat{\mathrm{G}}\left(X_i, \widehat{g}_k^{(d)}, \widehat{\xi}^{(d)}\right), i \in \mathbb{I}^{(d)}\right\}$. Below we show in propositions that (b1) and (b2) could be ensured with high probability as n goes to infinity. Simple calculation also reveals that, under (b1) and (b2), $\widehat{\lambda}_d$ is determined by

$$\frac{1}{n_d}\sum_{k=1}^K \sum_{i \in \mathbb{I}_k^{(d)}} \frac{\widehat{\mathrm{G}}\left(X_i, \widehat{g}_k^{(d)}, \widehat{\xi}^{(d)}\right)}{1 + \widehat{\lambda}_d^\top \widehat{\mathrm{G}}\left(X_i, \widehat{g}_k^{(d)}, \widehat{\xi}^{(d)}\right)} = 0. \tag{6}$$

Our proposed Machine learning and Data-splitting based Empirical Likelihood (MDEL) estimator for θ is

$$\widehat{\theta}_{\mathrm{mdel}} = \widehat{\theta}_{\mathrm{mdel}}^{(1)} - \widehat{\theta}_{\mathrm{mdel}}^{(0)} = \sum_{i=1}^n D_i \widehat{p}_i Y_i - \sum_{i=1}^n (1 - D_i)\widehat{p}_i Y_i.$$

To solve the dual problem (5) for $d = 0, 1$ and obtain the MDEL estimator, we carry out a modified Newton-Raphson algorithm such that the search for the solution of (6) always falls in $\mathcal{A}_n^d$. The details of the algorithm have been extensively discussed in Chen, Sitter and Wu [3], Wu [27], and Tan et al. [22]. The following regularity condition ensures that (b2) holds with high probability.

(**A1**) With probability tending to 1 as $n \to \infty$, $\mathrm{Var}\left\{\left[\widehat{g}_k^{(d)}(X)\right]_j\right\} > 0$ and $\mathrm{E}\left[\left[\widehat{g}_k^{(d)}(X)\right]_j^2\right] < \infty$ conditionally on $(W_i)_{i \in \mathbb{I}_k^{(d)c}}$ for $j = 1, \cdots, r$, $k = 1, \cdots, K$ and $d = 0, 1$, where $[x]_i$ denotes the i-th element of a vector x.

Assumption (**A1**) is mild and does not give any restriction on the dimension of covariates. The solution of (6) exists and is unique with high probability when n is large enough, even if the dimension of covariates grows with the sample size, which is indicated by the following proposition.

Proposition 1. *Under Assumption (**A1**), the vector* 0 *is in the convex hull of* $\left\{\widehat{\mathrm{G}}\left(X_i, \widehat{g}_k^{(d)}, \widehat{\xi}^{(d)}\right), i \in \mathbb{I}^{(d)}\right\}$ *with probability tending to 1 as* $n \to \infty$ *for* $d = 0, 1$.

For $d = 0, 1$, define

$$\begin{aligned}
\ddot{\xi}^{(d)} &= \frac{1}{K}\sum_{k=1}^K \mathrm{E}\left[\widehat{g}_k^{(d)}(X) \middle| (W_i)_{i \in \mathbb{I}_k^{(d)c}}\right], \\
\widehat{\mathrm{G}}\left(x, \widehat{g}_k^{(d)}, \ddot{\xi}^{(d)}\right) &= \widehat{g}_k^{(d)}(x) - \ddot{\xi}^{(d)}, \\
\widehat{V}_n^{(d)} &= \frac{1}{n}\sum_{k=1}^K \sum_{i \in \mathbb{I}_k^{(d)}} \frac{\widehat{\mathrm{G}}\left(X_i, \widehat{g}_k^{(d)}, \widehat{\xi}^{(d)}\right)^{\otimes 2}}{(2d-1)\delta + 1 - d}, \\
\ddot{J}_n^{(d)} &= \frac{1}{n}\sum_{k=1}^K \sum_{i \in \mathbb{I}_k^{(d)}} \frac{(Y_i - \theta_d)\widehat{\mathrm{G}}\left(X_i, \widehat{g}_k^{(d)}, \ddot{\xi}^{(d)}\right)}{(2d-1)\delta + 1 - d}, \\
\ddot{S}_n^{(d)} &= \frac{1}{n}\sum_{k=1}^K \sum_{i \in \mathbb{I}_k^{(d)}} \frac{\widehat{\mathrm{G}}\left(X_i, \widehat{g}_k^{(d)}, \ddot{\xi}^{(d)}\right)^{\otimes 2}}{(2d-1)\delta + 1 - d},
\end{aligned}$$

where $H^{\otimes 2} = HH^\top$ for any vector or matrix H. The following regularity condition ensures (b1) and helps to study the rate of $\hat{\lambda}_d$.

(**A2**) With probability tending to 1 as $n \to \infty$, $V_1 < \widehat{V}_n^{(d)} < V_2$, where V_1 and V_2 are two finite and positive definite matrices.

Assumption (**A2**) is mild and reasonable as we do not require that $\widehat{V}_n^{(d)}$ converges to any finite and positive definite matrix, but is bounded by two finite and positive definite matrices with probability tending to 1. The full rank requirement of $\widehat{V}_n^{(d)}$ suggests that the r ML models we impose should be different as much as possible. It is easy to see that the second derivative of $\ell(\lambda_d)$ is negative definite under (**A2**) and thus (b1) is satisfied. (**A2**) together with (**A1**) are sufficient for us to derive the order of $\widehat{\lambda}_d$, which is given by the following proposition:

Proposition 2. *Under Assumptions (**A1**)–(**A2**), we have* $||\widehat{\lambda}_d|| = O_p(\frac{1}{\sqrt{n}})$ *for* $d = 0, 1$, *where* $||\cdot||$ *denotes the Euclidean norm.*

The difference between $\ddot{S}_n^{(d)}$ and $\widehat{V}_n^{(d)}$ is only a term that vanishes in probability. Therefore, under (**A2**), $\ddot{S}_n^{(d)}$ is invertible at least when n is sufficiently large.

(**A3**) $\mathrm{E}[Y^2|D = d] < \infty$, for $d = 0, 1$.

Assumption (**A3**) is common in the empirical likelihood theory. Now, we derive the decomposition and asymptotic properties of the MDEL estimator.

Proposition 3. *Under Assumptions (**A1**)–(**A3**), we have*

$$\sqrt{n}\left(\widehat{\theta}_{mdel}-\theta\right) = \frac{1}{\sqrt{n}}\sum_{k=1}^{K}\sum_{i\in\mathbb{I}_k}\left[\frac{D_i}{\delta}\left(Y_i-\theta_1\right) - \frac{D_i-\delta}{\delta}\ddot{J}_n^{(1)^\top}\ddot{S}_n^{(1)^{-1}}\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(1)},\ddot{\xi}^{(1)}\right) - \frac{1-D_i}{1-\delta}\left(Y_i-\theta_0\right) + \frac{D_i-\delta}{1-\delta}\ddot{J}_n^{(0)^\top}\ddot{S}_n^{(0)^{-1}}\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(0)},\ddot{\xi}^{(0)}\right)\right] + o_p(1),$$

where $\theta_1 = \mathrm{E}[Y(1)]$ *and* $\theta_0 = \mathrm{E}[Y(0)]$.

It is easy to know from Proposition 3 that each term in the decomposition of $\sqrt{n}\left(\widehat{\theta}_{\text{mdel}}-\theta\right)$ is correlated with all the observations. When $r=1$ and the ML estimators for the nuisance parameters are consistent conditionally, however, these terms become asymptotically independent. In this case, the semiparametric theory implies that proposed MDEL estimator is asymptotically normal and semiparametric efficient.

Theorem 4. *Under Assumptions (**A1**)–(**A3**), if* $r=1$ *and*

$$\mathrm{E}\left[\left(\widehat{g}_k^{(d)}(X)-\eta^{(d)}(X)\right)^2\middle|(W_i)_{i\in\mathbb{I}_k^{(d)c}}\right]\to 0$$

in probability as $n\to\infty$ *for* $k=1,\cdots,K$ *and* $d=0,1$, $\sqrt{n}\left(\widehat{\theta}_{mdel}-\theta\right)$ *is asymptotically normal with mean zero and variance* $\mathrm{Var}[\varphi(W,\theta,\delta,\eta^{(1)},\eta^{(0)})]$, *where*

$$\varphi(W,\theta,\delta,\eta^{(1)},\eta^{(0)}) = \frac{D}{\delta}\left(Y-\eta^{(1)}(X)\right) - \frac{1-D}{1-\delta}\left(Y-\eta^{(0)}(X)\right) + \eta^{(1)}(X)-\eta^{(0)}(X)-\theta.$$

Therefore, $\widehat{\theta}_{mdel}$ *achieves the semiparametric efficiency bound.*

The "conditional consistency" assumption $\mathrm{E}\left[\left(\widehat{g}_k^{(d)}(X)-\eta^{(d)}(X)\right)^2\middle|(W_i)_{i\in\mathbb{I}_k^{(d)c}}\right]\to 0$ in probability as $n\to\infty$ is called "risk consistency" in Wager et al. [26]. It is mild for penalized regression methods when the regression model is correctly specified and sufficient sparsity is satisfied. The "risk consistency" assumption is also mild for many ML methods such as the random forests and neural networks when sufficient sparsity is satisfied. We refer to Chernozhukov et al. [4] for more explanations. When $r>1$, i.e., multiple ML methods are imposed to estimate the nuisance parameters, it is our expectation that Theorem 4 could hold when any one of estimators for the nuisance parameter satisfies "risk consistency" assumption. Moreover, we expect that the convergence rate of the estimator with multiple models is identical to that with the oracle model. The asymptotic theory under this situation, however, needs further complicated assumptions about the structure of covariates such that the weak law of large numbers can be applied to dependent terms, and we reserve it as future work. Instead, we use simulation studies in Section 5 to show that our proposed estimator attains multiple robustness property and is approximately normally distributed with reasonable coverage rates.

3.3 Variance recovery for valid inference

Based on the decomposition of $\widehat{\theta}_{\text{mdel}}$ in Proposition 3, we propose to estimate the variance of $\widehat{\theta}_{\text{mdel}}$ with

(7)

$$\widehat{\sigma}^2_{\text{mdel}} = \frac{1}{n}\sum_{d=0,1}\sum_{k=1}^{K}\sum_{i\in\mathbb{I}_k^{(d)}}\frac{n_d}{n}\widehat{p}_i\left\{\frac{n}{n_1}D_i(Y_i-\widehat{\theta}^{(1)}_{\text{mdel}}) - \frac{n}{n_1}(D_i-\frac{n_1}{n})\widehat{J}_n^{(1)^\top}\widehat{S}_n^{(1)^{-1}}\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(1)},\widehat{\xi}^{(1)}\right) - \frac{n}{n_0}(1-D_i)(Y_i-\widehat{\theta}^{(0)}_{\text{mdel}}) + \frac{n}{n_0}(D_i-\frac{n_1}{n})\widehat{J}_n^{(0)^\top}\widehat{S}_n^{(0)^{-1}}\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(0)},\widehat{\xi}^{(0)}\right)\right\}^2,$$

where

$$\widehat{J}_n^{(d)} = \sum_{k=1}^{K}\sum_{i\in\mathbb{I}_k^{(d)}}\widehat{p}_iY_i\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(d)},\widehat{\xi}^{(d)}\right),$$

$$\widehat{S}_n^{(d)} = \sum_{v\in\{0,1\}}\sum_{k=1}^{K}\sum_{i\in\mathbb{I}_k^{(v)}}\frac{n_v}{n}\widehat{p}_i\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(d)},\widehat{\xi}^{(d)}\right)\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(d)},\widehat{\xi}^{(d)}\right)^\top,$$

for $d=0,1$. The following theorem indicates that this variance estimator converges to the true variance in probability. That is, our variance estimator successfully recovers the true asymptotic variance of the proposed MDEL estimator.

Theorem 5. *Under the assumption and regularity conditions of Theorem 4, we have*

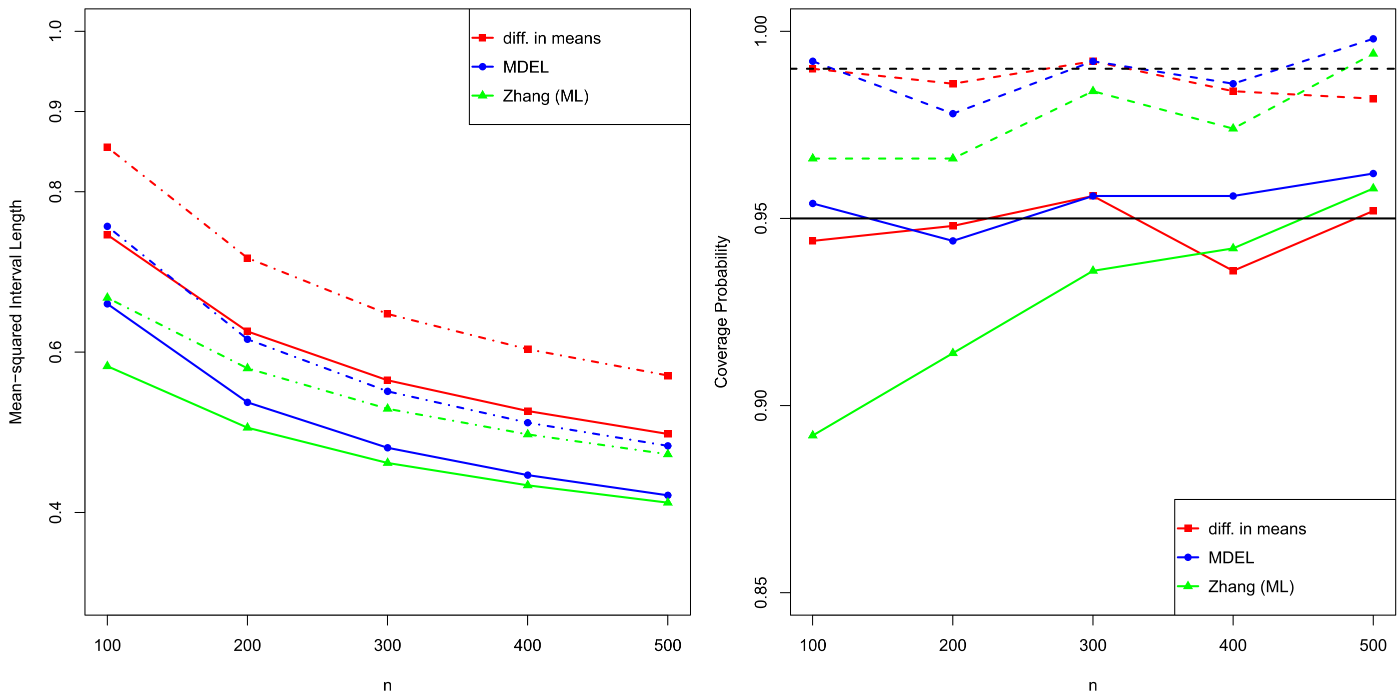
$$\widehat{\sigma}^2_{mdel}\to\frac{\mathrm{Var}\left[\varphi(W,\theta,\delta,\eta^{(1)},\eta^{(0)})\right]}{n}$$

in probability as $n\to\infty$.

Corollary 6. *Under the assumption and regularity conditions of Theorem 4,* $(\widehat{\sigma}_{mdel})^{-1}\left(\widehat{\theta}_{mdel}-\theta\right)$ *is asymptotically standard normal.*

Corollary 6 leads to a $100(1-\alpha)\%$ Wald confidence interval of θ:

$$\mathrm{CI} := \left(\widehat{\theta}_{\text{mdel}}-z_{\alpha/2}\widehat{\sigma}_{\text{mdel}},\widehat{\theta}_{\text{mdel}}+z_{\alpha/2}\widehat{\sigma}_{\text{mdel}}\right).$$



(a) Mean-squared length of confidence intervals.

(b) Coverage probability of confidence intervals.

Figure 1. Simulation results based on 500 Monte Carlo replications with $\beta^{(1)} = \beta^{(0)} = (1, 0, \cdots, 0)$, $p = 500$, $\rho = 0$, $\delta = 0.5$ *and sample size* n *ranging from* 100 *to* 500 *under a simple setting described in section 5.1. In the left panel, solid lines depict the mean-squared lengths of* 95% *Wald confidence intervals and dashed-dotted lines depict the mean squared lengths of* 99% *Wald confidence intervals. In the right panel, solid lines depict coverage proportions of* 95% *Wald confidence intervals that cover the true* θ *and dashed lines depict coverage proportions of* 99% *Wald confidence intervals that cover the true* θ.

As we mentioned in Section 3.1, Zhang's approach with the covariate-outcome relationship estimated by an ML method seriously underestimates the variance. In contrast, our proposed EL approach recovers the true variance. To illustrate this point, we conduct a simulation study following the setting of Wager et al. [26]. The setting is a special case of the simulation studies in Section 5.1 with coefficients equal to $(1, 0, \cdots, 0)$ or a permutation of $(1, \frac{1}{2}, \cdots, \frac{1}{p})$. We compare the difference-in-means estimator $\widehat{\theta}_{\text{dim}}$, Zhang's estimator, and our proposed MDEL estimator. In both Zhang's and our proposed MDEL approaches, the covariate-outcome relationship is modelled by Lasso.

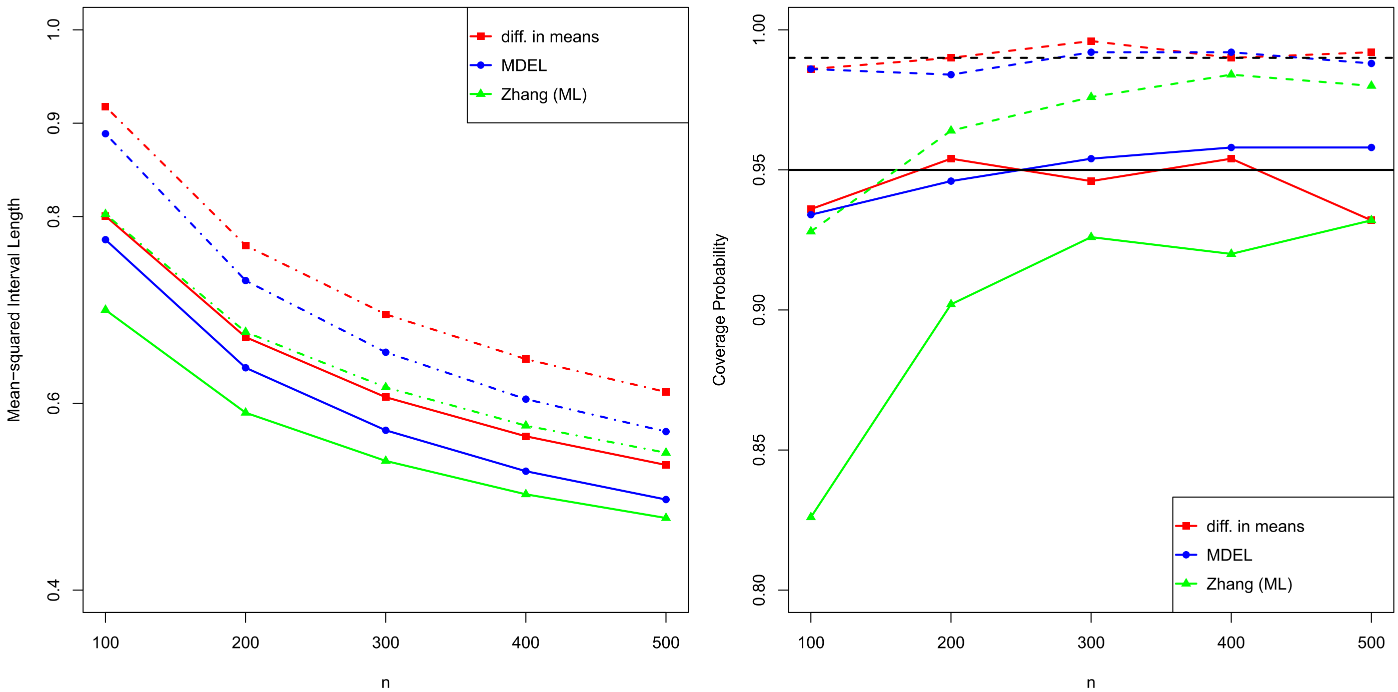
In Figure 1, where the true signal is very sparse, the 95% confidence intervals and 99% confidence intervals of the proposed MDEL approach are shorter than the corresponding confidence intervals based on $\widehat{\theta}_{\text{dim}}$. The coverage probability of the 95% or 99% confidence intervals of Zhang's EL approach is substantially lower than the true level for $n = 100, 200$. In contrast, the coverage probability of our approach resembles the nominal level for any sample size.

In Figure 2, where the true signal is geometric, the 95% confidence intervals and 99% confidence intervals of the proposed MDEL approach are still shorter than the corresponding ones based on $\widehat{\theta}_{\text{dim}}$. The coverage probability of the 95% or 99% confidence intervals of Zhang's approach is significantly below the nominal level for n ranging from 100 to 500. In contrast, the coverage probability of our approach are very close to the nominal level for any sample size.

In summary, when the true signal is either $(1, 0, \cdots, 0)$ or a permutation of $(1, \frac{1}{2}, \cdots, \frac{1}{p})$, Zhang's approach with machine learning to model the covariate-outcome relationship is not desirable as the corresponding confidence intervals fail to cover θ in reasonable proportions, whereas our approach recovers true variance and the coverage probabilities are close to the nominal levels.

4. PRACTICAL IMPLEMENTATIONS

In our simulation studies and real data analysis, we utilize three popular ML methods to estimate $\eta^{(d)}$. The first two methods are Lasso [23] and SCAD [6]. Lasso and SCAD are both penalized regression methods. Generally, they both lead to sparse solutions and thus work well for variable selection purpose. Compared with Lasso, however, large coefficients would not be shrunken by SCAD and some small coefficients cannot survive after punishment. Therefore, SCAD works better for models with strong and sparse signals. The third method we use is random forests [2], which is increas-



(a) Mean-squared length of confidence intervals. (b) Coverage probability of confidence intervals.

Figure 2. Simulation results based on 500 Monte Carlo replications with $\beta^{(d)}$ equal to a random permutation of $(\frac{1}{1}, \frac{1}{2}, \cdots, \frac{1}{p})$ for $d = 0, 1$, $p = 500$, $\rho = 0.5$, $\delta = 0.5$ and sample size n ranging from 100 to 500 under the setting described in Section 5.1. In the left panel, solid lines depict the mean-squared length of 95% Wald confidence intervals and dashed-dotted lines depict the mean-squared length of 99% Wald confidence intervals. In the right panel, solid lines depict coverage proportions of 95% Wald confidence intervals that cover the true θ and dashed lines depict coverage proportions of 99% Wald confidence intervals that cover the true θ.

ingly popular in recent years because of its flexibility and outstanding prediction ability for real complex data.

For Lasso, the penalty parameter λ is determined by 10-fold cross-validation criterion using `cv.glmnet` in R-package `glmnet` in this paper. For SCAD, the first tuning parameter a is chosen to be default 3.7 and penalty parameter λ is determined by 10-fold cross-validation criterion using `cv.ncvreg` in R-package `ncvreg`. For random forests, we build 500 regression trees using `ranger` in R-package `ranger`, a fast implementation of random forests for high dimensional data in C++ and R, with parameters set to be default.

5. SIMULATION STUDIES AND REAL DATA ANALYSIS

5.1 Simulation studies

We consider linear models for $\eta^{(d)}$ with the dimension of covariates p larger than the sample size n. The universal settings of our simulations are as follows. The covariates $X_i, i = 1, \cdots, n$ are independent and identically generated from multivariate Gaussian $\mathcal{N}(1_p, \Sigma)$, where $1_p = (1, \cdots, 1)^\top$ is a p-dimensional vector. The assignment probability is fixed to be $\delta = 0.5$ and $D_i \overset{i.i.d}{\sim} \text{Bernoulli}(\delta)$. The outcome Y_i of the i-th unit under treatment $D_i = d_i$ are generated from $\mathcal{N}(X_i^\top \beta^{(d_i)} + 5I(d_i = 1), 1)$, $i = 1, \cdots, n$. We consider four different size scales

$$(n, p) = \{(80, 200), (160, 200), (200, 1000), (800, 1000)\}.$$

Define $0^0 = 1$, signals and the covariance matrix of the covariates, $\mathbf{\Sigma}$, are different as follows.

Simulation 1 (Sufficient sparsity). $\beta_i^{(1)} = 3 \cdot 1(i \leq 3), \beta_j^{(0)} = 2 \cdot 1(j \leq 3)$ and $\mathbf{\Sigma}_{ij} = \rho^{1(i \neq j)}$.

Simulation 2 (Fan, Guo and Hao [5]).

$$(\beta_i^{(d)})_{i=1,2,3,5,7,11,13,17,19,23} = (1.01, -0.06, 0.72, 1.55, 2.32, -0.36, 3.75, -2.04, -0.13, 0.61)^\top, d = 0, 1$$

and $\mathbf{\Sigma}_{ij} = \rho^{|i-j|}$.

Simulation 3 (Dense geometry [26]). $\beta_i^{(1)} = 11^{-10i/p}, \beta_j^{(0)} = 10^{-10j/p}$ and $\mathbf{\Sigma}_{ij} = \rho^{|i-j|}$.

Table 1. Results of Simulation 1 based on 5000 Monte Carlo replications

	$\rho = 0$						$\rho = 0.5$					
Estimator	Bias	SD	SE	RMSE	Cov95	Cov99	Bias	SD	SE	RMSE	Cov95	Cov99
$(n, p) = (80, 200)$												
$\widehat{\theta}_{\rm dim}$	-0.004	1.002	1.015	1.002	0.950	0.988	-0.012	1.389	1.417	1.389	0.953	0.991
$\widehat{\theta}_{\rm wdtt}$(LASSO)	0.000	0.386	0.383	0.386	0.943	0.987	0.000	0.407	0.409	0.407	0.951	0.991
$\widehat{\theta}_{\rm wdtt}$(SCAD)	0.001	0.318	0.315	0.318	0.948	0.989	0.002	0.381	0.378	0.381	0.947	0.988
$\widehat{\theta}_{\rm wdtt}$(RF)	-0.005	0.959	0.970	0.959	0.947	0.988	-0.010	0.820	0.841	0.820	0.955	0.990
$\widehat{\theta}_{\rm mdel}$(LASSO)	0.001	0.349	0.349	0.349	0.949	0.987	0.000	0.399	0.406	0.399	0.952	0.991
$\widehat{\theta}_{\rm mdel}$(SCAD)	0.001	0.318	0.316	0.318	0.945	0.990	0.002	0.382	0.381	0.382	0.948	0.988
$\widehat{\theta}_{\rm mdel}$(RF)	-0.008	0.945	0.948	0.945	0.948	0.988	-0.008	0.743	0.772	0.743	0.956	0.992
$\widehat{\theta}_{\rm mdel}$(MULTI)	0.003	0.321	0.341	0.321	0.959	0.992	0.002	0.374	0.397	0.374	0.957	0.993
$(n, p) = (160, 200)$												
$\widehat{\theta}_{\rm dim}$	-0.004	0.713	0.718	0.713	0.950	0.990	-0.016	0.999	1.003	0.999	0.951	0.991
$\widehat{\theta}_{\rm wdtt}$(LASSO)	-0.001	0.225	0.228	0.225	0.953	0.989	-0.002	0.261	0.263	0.261	0.946	0.992
$\widehat{\theta}_{\rm wdtt}$(SCAD)	-0.001	0.210	0.212	0.210	0.948	0.989	-0.002	0.250	0.251	0.250	0.949	0.990
$\widehat{\theta}_{\rm wdtt}$(RF)	-0.004	0.653	0.658	0.653	0.951	0.990	-0.005	0.550	0.556	0.550	0.951	0.990
$\widehat{\theta}_{\rm mdel}$(LASSO)	-0.001	0.215	0.218	0.215	0.950	0.990	-0.001	0.259	0.262	0.259	0.950	0.992
$\widehat{\theta}_{\rm mdel}$(SCAD)	-0.001	0.210	0.212	0.210	0.950	0.989	-0.002	0.250	0.252	0.250	0.952	0.991
$\widehat{\theta}_{\rm mdel}$(RF)	-0.004	0.536	0.560	0.536	0.959	0.992	-0.001	0.510	0.519	0.510	0.952	0.991
$\widehat{\theta}_{\rm mdel}$(MULTI)	-0.001	0.212	0.222	0.212	0.956	0.991	-0.002	0.251	0.261	0.251	0.959	0.991
$(n, p) = (200, 1000)$												
$\widehat{\theta}_{\rm dim}$	-0.008	0.642	0.642	0.642	0.946	0.989	-0.019	0.895	0.896	0.895	0.947	0.990
$\widehat{\theta}_{\rm wdtt}$(LASSO)	0.002	0.206	0.207	0.206	0.950	0.988	0.002	0.237	0.238	0.237	0.950	0.989
$\widehat{\theta}_{\rm wdtt}$(SCAD)	0.003	0.190	0.189	0.190	0.948	0.990	0.003	0.225	0.225	0.225	0.951	0.989
$\widehat{\theta}_{\rm wdtt}$(RF)	-0.007	0.618	0.618	0.618	0.946	0.990	-0.009	0.510	0.508	0.510	0.947	0.988
$\widehat{\theta}_{\rm mdel}$(LASSO)	0.004	0.194	0.194	0.194	0.949	0.990	0.003	0.235	0.236	0.235	0.951	0.989
$\widehat{\theta}_{\rm mdel}$(SCAD)	0.003	0.190	0.190	0.190	0.950	0.990	0.003	0.225	0.225	0.225	0.952	0.990
$\widehat{\theta}_{\rm mdel}$(RF)	-0.005	0.565	0.576	0.565	0.953	0.991	-0.006	0.483	0.482	0.483	0.950	0.987
$\widehat{\theta}_{\rm mdel}$(MULTI)	0.003	0.191	0.196	0.191	0.955	0.992	0.003	0.226	0.231	0.226	0.954	0.991
$(n, p) = (800, 1000)$												
$\widehat{\theta}_{\rm dim}$	-0.005	0.321	0.320	0.321	0.949	0.990	-0.004	0.450	0.447	0.450	0.951	0.991
$\widehat{\theta}_{\rm wdtt}$(LASSO)	-0.001	0.096	0.096	0.096	0.949	0.991	-0.002	0.114	0.113	0.114	0.949	0.989
$\widehat{\theta}_{\rm wdtt}$(SCAD)	-0.001	0.093	0.094	0.093	0.950	0.991	-0.002	0.112	0.112	0.112	0.949	0.989
$\widehat{\theta}_{\rm wdtt}$(RF)	-0.004	0.294	0.293	0.294	0.951	0.991	-0.005	0.238	0.237	0.238	0.949	0.989
$\widehat{\theta}_{\rm mdel}$(LASSO)	-0.001	0.094	0.094	0.094	0.950	0.991	-0.002	0.113	0.113	0.113	0.949	0.989
$\widehat{\theta}_{\rm mdel}$(SCAD)	-0.001	0.093	0.094	0.093	0.951	0.991	-0.002	0.112	0.112	0.112	0.950	0.989
$\widehat{\theta}_{\rm mdel}$(RF)	-0.002	0.189	0.194	0.189	0.957	0.991	-0.005	0.230	0.230	0.230	0.948	0.989
$\widehat{\theta}_{\rm mdel}$(MULTI)	-0.001	0.093	0.095	0.093	0.952	0.992	-0.002	0.112	0.113	0.112	0.949	0.989

Bias = average bias of 5000 Monte Carlo estimators, SD = sample standard deviation of estimators, SE = average of model-based standard error, RMSE = empirical root mean square error, Cov95 = proportion of 95% Wald confidence intervals covering the true θ, Cov99 = proportion of 99% Wald confidence intervals covering the true θ.

Simulation 1 has sparse and strong signals. Simulation 2 has sparse signals with more challenging cofficients. Simulation 3 is identical to the geometric case of Wager et al. [26]. Results of simulations are all based on 5000 Monte Carlo data sets and given in Table 1, 2 and 3. First, we summarize the results for Simulations 1 and 2 given in Table 1 and 2 (sparse case):

(a) Compared with the simple approach of difference in means, the EL estimators with any outcome model have significantly smaller SDs and RMSEs.

(b) Among the EL estimators with one outcome model, the estimators using SCAD perform relatively better than other estimators, and estimators using random forests perform worst in sense of RMSE. As expected, the EL estimators with multiple models perform closest to those with SCAD, and better than all other estimators when $\rho = 0.5$ and $(n, p) = (80, 200)$ and when $(n, p) = (800, 1000)$.

(c) Using SCAD to model the covariate-outcome relationship, the EL estimators perform similarly to Wager's

Table 2. Results of Simulation 2 based on 5000 Monte Carlo replications

	$\rho = 0$						$\rho = 0.5$					
Estimator	Bias	SD	SE	RMSE	Cov95	Cov99	Bias	SD	SE	RMSE	Cov95	Cov99
$(n,p) = (80, 200)$												
$\widehat{\theta}_{\text{dim}}$	0.028	1.196	1.210	1.196	0.950	0.988	0.017	1.231	1.243	1.231	0.946	0.988
$\widehat{\theta}_{\text{wdtt}}$(LASSO)	0.010	0.584	0.586	0.584	0.951	0.987	0.006	0.536	0.535	0.536	0.952	0.990
$\widehat{\theta}_{\text{wdtt}}$(SCAD)	0.007	0.451	0.447	0.451	0.949	0.989	0.007	0.448	0.442	0.448	0.946	0.989
$\widehat{\theta}_{\text{wdtt}}$(RF)	0.026	1.141	1.155	1.141	0.950	0.989	0.017	1.153	1.165	1.153	0.946	0.988
$\widehat{\theta}_{\text{mdel}}$(LASSO)	0.008	0.539	0.541	0.539	0.953	0.987	0.006	0.494	0.489	0.494	0.951	0.989
$\widehat{\theta}_{\text{mdel}}$(SCAD)	0.008	0.457	0.453	0.457	0.950	0.988	0.007	0.457	0.448	0.457	0.948	0.989
$\widehat{\theta}_{\text{mdel}}$(RF)	0.023	1.130	1.129	1.131	0.948	0.987	0.013	1.105	1.112	1.105	0.951	0.987
$\widehat{\theta}_{\text{mdel}}$(MULTI)	0.007	0.457	0.466	0.457	0.957	0.990	0.006	0.440	0.448	0.440	0.956	0.991
$(n,p) = (160, 200)$												
$\widehat{\theta}_{\text{dim}}$	-0.018	0.859	0.853	0.859	0.947	0.988	-0.010	0.891	0.878	0.891	0.947	0.988
$\widehat{\theta}_{\text{wdtt}}$(LASSO)	0.000	0.226	0.223	0.226	0.947	0.988	-0.001	0.217	0.215	0.217	0.949	0.988
$\widehat{\theta}_{\text{wdtt}}$(SCAD)	0.002	0.182	0.182	0.182	0.945	0.988	0.001	0.183	0.183	0.183	0.953	0.989
$\widehat{\theta}_{\text{wdtt}}$(RF)	-0.016	0.788	0.784	0.789	0.948	0.989	-0.008	0.794	0.782	0.794	0.947	0.989
$\widehat{\theta}_{\text{mdel}}$(LASSO)	0.001	0.208	0.207	0.208	0.950	0.989	0.000	0.201	0.201	0.201	0.952	0.987
$\widehat{\theta}_{\text{mdel}}$(SCAD)	0.001	0.183	0.183	0.183	0.948	0.989	0.001	0.183	0.184	0.183	0.953	0.990
$\widehat{\theta}_{\text{mdel}}$(RF)	-0.011	0.655	0.675	0.655	0.958	0.992	-0.004	0.622	0.638	0.622	0.955	0.992
$\widehat{\theta}_{\text{mdel}}$(MULTI)	0.001	0.184	0.200	0.184	0.960	0.994	0.001	0.184	0.201	0.184	0.966	0.993
$(n,p) = (200, 1000)$												
$\widehat{\theta}_{\text{dim}}$	-0.003	0.773	0.763	0.773	0.945	0.988	-0.002	0.796	0.785	0.796	0.949	0.988
$\widehat{\theta}_{\text{wdtt}}$(LASSO)	0.001	0.226	0.222	0.226	0.945	0.991	0.001	0.212	0.209	0.212	0.945	0.990
$\widehat{\theta}_{\text{wdtt}}$(SCAD)	0.002	0.164	0.163	0.164	0.950	0.990	0.003	0.166	0.164	0.166	0.948	0.989
$\widehat{\theta}_{\text{wdtt}}$(RF)	-0.003	0.746	0.736	0.746	0.946	0.988	-0.002	0.758	0.748	0.758	0.949	0.987
$\widehat{\theta}_{\text{mdel}}$(LASSO)	0.001	0.199	0.198	0.199	0.949	0.992	0.002	0.185	0.185	0.185	0.951	0.991
$\widehat{\theta}_{\text{mdel}}$(SCAD)	0.002	0.164	0.164	0.164	0.952	0.990	0.003	0.167	0.165	0.167	0.948	0.990
$\widehat{\theta}_{\text{mdel}}$(RF)	-0.003	0.695	0.696	0.695	0.949	0.989	-0.006	0.671	0.677	0.671	0.949	0.991
$\widehat{\theta}_{\text{mdel}}$(MULTI)	0.002	0.165	0.175	0.165	0.965	0.995	0.003	0.168	0.177	0.168	0.962	0.993
$(n,p) = (800, 1000)$												
$\widehat{\theta}_{\text{dim}}$	0.010	0.385	0.381	0.385	0.949	0.990	0.009	0.396	0.392	0.397	0.949	0.989
$\widehat{\theta}_{\text{wdtt}}$(LASSO)	0.000	0.076	0.077	0.076	0.952	0.991	0.000	0.076	0.076	0.076	0.947	0.991
$\widehat{\theta}_{\text{wdtt}}$(SCAD)	-0.001	0.072	0.073	0.072	0.953	0.992	-0.001	0.072	0.073	0.072	0.955	0.991
$\widehat{\theta}_{\text{wdtt}}$(RF)	0.009	0.354	0.350	0.354	0.947	0.990	0.009	0.351	0.348	0.352	0.946	0.990
$\widehat{\theta}_{\text{mdel}}$(LASSO)	0.000	0.073	0.074	0.073	0.954	0.991	0.000	0.073	0.074	0.073	0.954	0.992
$\widehat{\theta}_{\text{mdel}}$(SCAD)	-0.001	0.072	0.073	0.072	0.954	0.992	-0.001	0.072	0.073	0.072	0.956	0.991
$\widehat{\theta}_{\text{mdel}}$(RF)	0.003	0.236	0.237	0.236	0.955	0.991	0.005	0.224	0.226	0.224	0.953	0.989
$\widehat{\theta}_{\text{mdel}}$(MULTI)	-0.001	0.072	0.075	0.072	0.958	0.992	-0.001	0.072	0.075	0.072	0.958	0.992

Bias = average bias of 5000 Monte Carlo estimators, SD = sample standard deviation of estimators, SE = average of model-based standard error, RMSE = empirical root mean square error, Cov95 = proportion of 95% Wald confidence intervals covering the true θ, Cov99 = proportion of 99% Wald confidence intervals covering the true θ.

estimators in terms of SD and RMSE. However, when Lasso or random forests model are adopted, the EL estimators outperform Wager's estimators. In Simulation 1, compared with Wager's estimators with random forests, the EL estimators with random forests have an average of 15.9% reduction in RMSE for $\rho = 0$ and 6.3% reduction in RMSE for $\rho = 0.5$. In simulation 2, compared with Wager's estimators with random forests, the EL estimators with random forests have an average of 14.5% reduction in RMSE for $\rho = 0$ and 18.4% reduction in RMSE for $\rho = 0.5$. The reduction is more obvious when n is larger.

(d) The SEs of the EL estimators with one outcome model are very close to their corresponding SDs, and the coverage probabilities of the EL estimators with one model are close to the nominal levels. However, the variances of the EL estimators with multiple models are slightly overestimated, but in a reasonable range.

Results in Table 3 are summarized as follows (dense case):

Table 3. Results of Simulation 3 based on 5000 Monte Carlo replications

	$\rho = 0$						$\rho = 0.5$					
Estimator	Bias	SD	SE	RMSE	Cov95	Cov99	Bias	SD	SE	RMSE	Cov95	Cov99
$(n, p) = (80, 200)$												
$\widehat{\theta}_{\text{dim}}$	0.003	0.489	0.490	0.489	0.949	0.990	0.002	0.732	0.737	0.732	0.949	0.990
$\widehat{\theta}_{\text{wdtt}}$(LASSO)	0.002	0.453	0.451	0.453	0.946	0.987	0.000	0.435	0.432	0.435	0.945	0.985
$\widehat{\theta}_{\text{wdtt}}$(SCAD)	0.004	0.458	0.460	0.458	0.946	0.988	0.002	0.497	0.497	0.497	0.945	0.984
$\widehat{\theta}_{\text{wdtt}}$(RF)	0.002	0.478	0.478	0.478	0.947	0.991	0.000	0.667	0.671	0.667	0.948	0.990
$\widehat{\theta}_{\text{mdel}}$(LASSO)	0.003	0.464	0.452	0.464	0.943	0.987	0.000	0.425	0.422	0.425	0.945	0.988
$\widehat{\theta}_{\text{mdel}}$(SCAD)	0.003	0.469	0.459	0.469	0.940	0.987	0.003	0.510	0.507	0.510	0.946	0.984
$\widehat{\theta}_{\text{mdel}}$(RF)	0.003	0.489	0.478	0.489	0.940	0.987	-0.002	0.583	0.599	0.582	0.952	0.992
$\widehat{\theta}_{\text{mdel}}$(MULTI)	0.005	0.470	0.446	0.470	0.932	0.983	0.001	0.433	0.425	0.433	0.943	0.986
$(n, p) = (160, 200)$												
$\widehat{\theta}_{\text{dim}}$	-0.002	0.350	0.346	0.350	0.947	0.989	-0.005	0.527	0.521	0.527	0.948	0.989
$\widehat{\theta}_{\text{wdtt}}$(LASSO)	-0.001	0.259	0.255	0.259	0.950	0.989	-0.002	0.220	0.219	0.220	0.948	0.990
$\widehat{\theta}_{\text{wdtt}}$(SCAD)	0.000	0.260	0.256	0.260	0.949	0.988	-0.001	0.257	0.254	0.257	0.949	0.990
$\widehat{\theta}_{\text{wdtt}}$(RF)	-0.002	0.336	0.332	0.336	0.946	0.990	-0.003	0.447	0.442	0.446	0.948	0.989
$\widehat{\theta}_{\text{mdel}}$(LASSO)	-0.001	0.260	0.255	0.260	0.946	0.989	-0.002	0.211	0.211	0.211	0.952	0.989
$\widehat{\theta}_{\text{mdel}}$(SCAD)	0.000	0.266	0.261	0.266	0.947	0.988	-0.001	0.262	0.260	0.262	0.951	0.991
$\widehat{\theta}_{\text{mdel}}$(RF)	-0.003	0.326	0.322	0.326	0.945	0.988	0.000	0.304	0.317	0.304	0.959	0.992
$\widehat{\theta}_{\text{mdel}}$(MULTI)	0.000	0.258	0.250	0.258	0.942	0.988	-0.002	0.212	0.215	0.212	0.952	0.991
$(n, p) = (200, 1000)$												
$\widehat{\theta}_{\text{dim}}$	0.001	0.663	0.660	0.663	0.945	0.988	-0.001	1.113	1.109	1.113	0.947	0.988
$\widehat{\theta}_{\text{wdtt}}$(LASSO)	0.001	0.640	0.641	0.640	0.950	0.989	0.002	0.753	0.758	0.753	0.950	0.990
$\widehat{\theta}_{\text{wdtt}}$(SCAD)	0.001	0.645	0.646	0.645	0.949	0.989	0.003	0.864	0.867	0.864	0.950	0.991
$\widehat{\theta}_{\text{wdtt}}$(RF)	0.002	0.655	0.653	0.655	0.946	0.988	0.000	1.064	1.062	1.064	0.948	0.988
$\widehat{\theta}_{\text{mdel}}$(LASSO)	0.002	0.651	0.643	0.651	0.947	0.987	-0.001	0.735	0.740	0.735	0.953	0.990
$\widehat{\theta}_{\text{mdel}}$(SCAD)	0.002	0.654	0.646	0.654	0.946	0.987	0.002	0.858	0.861	0.858	0.949	0.991
$\widehat{\theta}_{\text{mdel}}$(RF)	0.002	0.661	0.653	0.661	0.946	0.987	0.001	0.959	0.978	0.958	0.952	0.991
$\widehat{\theta}_{\text{mdel}}$(MULTI)	0.001	0.653	0.640	0.653	0.945	0.985	-0.002	0.734	0.737	0.734	0.952	0.989
$(n, p) = (800, 1000)$												
$\widehat{\theta}_{\text{dim}}$	0.004	0.330	0.330	0.330	0.951	0.991	0.008	0.553	0.555	0.553	0.950	0.991
$\widehat{\theta}_{\text{wdtt}}$(LASSO)	0.001	0.155	0.155	0.155	0.954	0.990	0.001	0.111	0.111	0.111	0.953	0.990
$\widehat{\theta}_{\text{wdtt}}$(SCAD)	0.000	0.138	0.139	0.138	0.954	0.992	-0.001	0.165	0.167	0.165	0.950	0.993
$\widehat{\theta}_{\text{wdtt}}$(RF)	0.004	0.320	0.321	0.320	0.951	0.991	0.008	0.495	0.495	0.495	0.949	0.991
$\widehat{\theta}_{\text{mdel}}$(LASSO)	0.001	0.146	0.146	0.146	0.953	0.989	0.001	0.104	0.104	0.104	0.953	0.989
$\widehat{\theta}_{\text{mdel}}$(SCAD)	0.000	0.139	0.140	0.139	0.954	0.991	0.000	0.168	0.169	0.167	0.950	0.992
$\widehat{\theta}_{\text{mdel}}$(RF)	0.003	0.294	0.299	0.294	0.953	0.991	0.007	0.301	0.306	0.301	0.958	0.989
$\widehat{\theta}_{\text{mdel}}$(MULTI)	0.000	0.134	0.135	0.134	0.955	0.992	0.001	0.104	0.105	0.104	0.957	0.989

Bias = average bias of 5000 Monte Carlo estimators, SD = sample standard deviation of estimators, SE = average of model-based standard error, RMSE = empirical root mean square error, Cov95 = proportion of 95% Wald confidence intervals covering the true θ, Cov99 = proportion of 99% Wald confidence intervals covering the true θ.

(a) When $\rho = 0$ and n is small compared to the dimension of covariates ($(n, p) = (80, 200), (200, 1000)$), compared with the simple approach of difference in means, there is no significant reduction in RMSE for the EL estimators. Otheriwse, compared with the difference in means estimators, the EL estimators with any outcome model have significantly smaller SDs and RMSEs.

(b) When $\rho = 0$ and n is small compared to the dimension of covariates, there is no significant difference among different estimators. In other cases, among the EL estimators using one outcome model, the estimators with Lasso generally perform best and the estimators with random forests perform worst in the sense of RMSE. As expected, the EL estimators with multiple ML models perform closest to the ones with the best model.

(c) Under Lasso or SCAD model, the EL estimators perform similarly to Wager's estimators in terms of RMSE. However, under random forests model, the EL estimators significantly outperform Wager's estimators.

Table 4. Point and interval estimates of θ for ACTG 175 data

Estimator	Estimate	SE	Relative Efficiency	95% Confidence Interval	99% Confidence Interval
		5-Fold Cross Validation			
$\widehat{\theta}_{\text{dim}}$	46.811	6.760	1.000	(33.56, 60.06)	(29.40, 64.22)
$\widehat{\theta}_{\text{tdzl}}$(Forward-1)	49.896	5.139	1.738	(39.82, 59.97)	(36.66, 63.13)
$\widehat{\theta}_{\text{tdzl}}$(Forward-2)	51.589	5.070	1.797	(41.65, 61.53)	(38.53, 64.65)
$\widehat{\theta}_{\text{Zhang}}$(Forward-1)	49.872	5.128	1.738	(39.82, 59.92)	(36.66, 63.08)
$\widehat{\theta}_{\text{Zhang}}$(Forward-2)	51.395	5.028	1.808	(41.54, 61.25)	(38.44, 64.34)
$\widehat{\theta}_{\text{wdtt}}$(LASSO)	49.785	5.233	1.669	(39.53, 60.04)	(36.31, 63.26)
$\widehat{\theta}_{\text{wdtt}}$(SCAD)	49.508	5.216	1.680	(39.29, 59.73)	(36.07, 62.94)
$\widehat{\theta}_{\text{wdtt}}$(RF)	53.442	5.253	1.656	(43.15, 63.74)	(39.91, 66.97)
$\widehat{\theta}_{\text{mdel}}$(LASSO)	49.938	5.200	1.690	(39.75, 60.13)	(36.54, 63.33)
$\widehat{\theta}_{\text{mdel}}$(SCAD)	49.483	5.197	1.692	(39.30, 59.67)	(36.10, 62.87)
$\widehat{\theta}_{\text{mdel}}$(RF)	53.160	5.216	1.680	(42.94, 63.38)	(39.72, 66.60)
$\widehat{\theta}_{\text{mdel}}$(MULTI)	50.396	5.150	1.723	(40.30, 60.49)	(37.13, 63.66)
		10-Fold Cross Validation			
$\widehat{\theta}_{\text{dim}}$	46.811	6.760	1.000	(33.56, 60.06)	(29.40, 64.22)
$\widehat{\theta}_{\text{tdzl}}$(Forward-1)	49.896	5.139	1.738	(39.82, 59.97)	(36.66, 63.13)
$\widehat{\theta}_{\text{tdzl}}$(Forward-2)	51.589	5.070	1.797	(41.65, 61.53)	(38.53, 64.65)
$\widehat{\theta}_{\text{Zhang}}$(Forward-1)	49.872	5.128	1.738	(39.82, 59.92)	(36.66, 63.08)
$\widehat{\theta}_{\text{Zhang}}$(Forward-2)	51.395	5.028	1.808	(41.54, 61.25)	(38.44, 64.34)
$\widehat{\theta}_{\text{wdtt}}$(LASSO)	49.854	5.224	1.675	(39.62, 60.09)	(36.40, 63.31)
$\widehat{\theta}_{\text{wdtt}}$(SCAD)	49.991	5.210	1.684	(39.78, 60.20)	(36.57, 63.41)
$\widehat{\theta}_{\text{wdtt}}$(RF)	53.885	5.262	1.651	(43.57, 64.20)	(40.33, 67.44)
$\widehat{\theta}_{\text{mdel}}$(LASSO)	50.059	5.178	1.705	(39.91, 60.21)	(36.72, 63.40)
$\widehat{\theta}_{\text{mdel}}$(SCAD)	49.979	5.177	1.705	(39.83, 60.13)	(36.64, 63.32)
$\widehat{\theta}_{\text{mdel}}$(RF)	53.542	5.212	1.682	(43.33, 63.76)	(40.12, 66.97)
$\widehat{\theta}_{\text{mdel}}$(MULTI)	50.665	5.140	1.730	(40.59, 60.74)	(37.42, 63.90)

SE = standard error, Relative Efficiency = (SE^2 of corresponding estimator)/(SE^2 of $\widehat{\theta}_{\text{dim}}$).

For example, when $(n, p) = (160, 200)$ and $\rho = 0.5$, the RMSE of $\widehat{\theta}_{\text{mdel}}$(RF) is 0.304 while the RMSE of $\widehat{\theta}_{\text{wdtt}}$(RF) is 0.446.

(d) When $\rho = 0$ and $(n, p) = (80, 200)$, the variances of EL estimators are underestimated and the coverage rates of the EL estimators are smaller than the nominal levels, but still in a reasonable range.

5.2 Analysis of ACTG 175 data set

In this section, we apply our proposed MDEL method to data from 2139 HIV-infected patients enrolled in AIDS Clinical Trials Group Protocol 175 (ACTG175) Study [8]. It is a double-blinded randomized experiment which was designed to study the treatment of patients receiving three different drugs and their combinations. Patients whose CD4 cell counts from 200 to 500 per cubic millimeter were randomly assigned to different antiretroviral regimens: zidovudine (ZDV) monotherapy, ZDV + didanosine (ddI), ZDV + zalcitabine, and ddI monotherapy. We follow the work of Tsiatis et al. [25], Huang, Qin and Follmann [11], and Zhang [30], where two treatment groups are considered: patients who received ZDV monotherapy alone, with $n_0 = 532$ and patients who received either ZDV + ddI, or ZDV + zalcitabine, or ddI alone, with $n_1 = 1607$. Pre-treatment baseline covariates are five continuous variables: cd40 = CD4 count (cells/mm^3), cd80 = CD8 count (cells/mm^3), age = age (years), wtkg = weight (kg), karnof = Karnofsky score (scale of 0–100), and seven binary variables: hemo = hemophilia, homo = homosexual activity, drug = history of intravenous drug use, race = race (0=white, 1=non-white), gender=gender (0=female), anti=antiretroviral history (0=naive, 1=experienced), and symp=symptomatic status (0=asymptomatic).

In the previous work of Tsiatis et al. [25], Zhang [30], and Tan et al. [22], Forward-1, a forward stepwise regression model allowing for linear terms of covariates, and Forward-2, a forward stepwise regression model allowing for linear, quadratic and interaction terms of baseline variables, are adopted. Our proposed MDEL approach enables us to consider a much richer feature set. Therefore, we take linear and quadratic terms of continuous variables, linear and interaction terms of binary variables, and interaction terms of above two sets of coordinates as our final feature set, i.e.,

$$\mathbb{F}_{\text{ACTG175}} = \Big\{(\text{cd40+cd80+age+wtkg+karnof+1})^2 \times \Big((\text{hemo+homo+drug+race+gender +anti+symp+1})^2 - \text{hemo}^2 - \text{homo}^2 - \text{drug}^2 - \text{race}^2 - \text{gender}^2 - \text{anti}^2 - \text{symp}^2\Big)\Big\}.$$

This leads to 608 explanatory variables (excluding the intercept) and we adopt Lasso, SCAD, and random forests to estimate the variable-outcome relationship using this feature set. Table 4 displays the estimates, standard errors, and confidence intervals of our proposed approach and some existing approaches described in Section 2 and 3.

Here, $\widehat{\theta}_{\text{mdel}}(\text{ML})$ denotes the point estimator of our MDEL approach, where the choices of ML include LASSO, SCAD, RF, and MULTI. Here, RF indicates the random forests method, and MULTI means we make use of multiple ML methods (LASSO, SCAD, and RF) in our MDEL method.

For inference on θ, both 95% and 99% Wald confidence intervals are provided. The results of Table 4 give us strong evidence to reject the null hypothesis that there is no difference in treatment effect between two groups with different therapies. It is worth to note that, despite a much richer feature set with $p = 608$ variables is considered, our proposed approach does not improve the estimation efficiency. This indicates that the original explanatory variables are adequate for modeling $\eta^{(d)}(\cdot), d = 0, 1$. However, our data analysis result of ACTG 175 data set is still meaningful because we provide further reliability to use the original set of explanatory variables.

5.3 Analysis of GSE118657 data set

Gene Expression Omnibus dataset (GSE118657, available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE118657 and R-package `GEOquery`) is a Phase II/III randomized controlled trial examining the use of lactoferrin to prevent nosocomial infections in critically ill patients undergoing mechanical ventilation [16, 14]. This data set consists of longitudinal measurements of 61 patients, among which 32 patients were randomized to receive lactoferrin and the remaining ones were assigned to the placebo group. We are interested in studying the effect of lactoferrin on the length of stay in ICU. For covariate-adjustment, we consider four important variables of patients before receiving the treatment-age, sex, SOFA score, and APACHE II score, denoted by X_b, and the first-day gene expression data of patients, denoted by X_g. In the following data analysis, approaches of Zhang [30] and Tsiatis et al. [25] are based on modelling $\mathrm{E}[Y|X_b, D = d], d = 0, 1$ with Forward-1 or Forward-2 model. To make use of information of the gene expression data, we model $\mathrm{E}[Y|X_b, X_g, D = d], d = 0, 1$ by ML methods and subsequently apply the approach of Wager et al. [26] and our proposed MDEL approach. Since the dimension of $X = (X_b, X_g)$, $p \approx 50000$, is too high, we use sure independent screening (SIS) method [7] to filter out variables that are relatively weak-correlated with the response, and reduce the dimension of X to a low level, say $d_X = O(n)$, before modelling $\mathrm{E}[Y|X, D = 1]$ and $\mathrm{E}[Y|X, D = 0]$.

Results given in Table 5 indicate that there is no improvement about the length of stay in ICU for patients after the use of lactoferrin. Our approach with multiple ML methods, $\widehat{\theta}_{\text{mdel}}(\text{MULTI})$, is more efficient than other estimators with the shortest confidence intervals.

6. CONCLUSIONS AND FURTHER DISCUSSIONS

In this paper, we propose a machine learning and data-splitting based EL approach to make statistical inference on the average treatment effect in randomized controlled trials. Our approach not only maintains the advantages of the traditional EL approaches, but also overcomes the disadvantage that the traditional EL approaches usually make invalid inference in high-dimensional settings. Compared with the regression adjustment approach proposed by Wager et al. [26], our proposed approach has two attractive characteristics, which are illustrated by our simulation studies: (i). Our MDEL estimator performs better when we use random forests to estimate the nuisance parameters; (ii). Our MDEL estimator with multiple ML models is likely to perform as good as the oracle model, which is known as the multiple robustness property.

For future work, we plan to (i). study the asymptotic theory of the proposed EL estimator with multiple models; (ii). generalize our proposed approach to high-dimensional observational studies by modelling propensity scores and imposing additional constraints about the propensity scores.

APPENDIX: LEMMAS AND PROOFS

Lemmas

Lemma 1 (Chernozhukov et al. [4]; Conditional convergence implies unconditional)**.** *Let $\{X_m\}$ and $\{Y_m\}$ be random vectors. (a) If for $\varepsilon_m \to 0$, $\mathbb{P}(||X_m|| > \varepsilon_m | Y_m) \to 0$ as $m \to \infty$, then $\mathbb{P}(||X_m|| > \varepsilon_m) \to 0$ as $m \to \infty$. (b) Let $\{A_m\}$ be a sequence of positive constants. If $||X_m|| = O_p(A_m)$ conditional on Y_m, then $||X_m|| = O_p(A_m)$ holds unconditionally.*

Lemma 2. $\widehat{\mathrm{G}}\left(X_i, \widehat{g}_k^{(d)}, \widehat{\xi}^{(d)}\right) = \widehat{\mathrm{G}}\left(X_i, \widehat{g}_k^{(d)}, \ddot{\xi}^{(d)}\right) + O_p(\frac{1}{\sqrt{n}})$, *$i \in \mathbb{I}_k^{(d)}$ for $k = 1, \cdots, K$ and $d = 0, 1$.*

Proof. Simple algebra gives

$$\widehat{\mathrm{G}}\left(X_i, \widehat{g}_k^{(d)}, \ddot{\xi}^{(d)}\right) - \widehat{\mathrm{G}}\left(X_i, \widehat{g}_k^{(d)}, \widehat{\xi}^{(d)}\right) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{|\mathbb{I}_k|}\sum_{i\in\mathbb{I}_k}\left(\widehat{g}_k^{(d)}(X_i) - \mathrm{E}\left[\widehat{g}_k^{(d)}(X_i)\Big|\,(W_j)_{j\in\mathbb{I}_k^{(d)c}}\right]\right).$$

Table 5. *Point and interval estimates of θ for GSE118657 data with 5-fold cross validation*

Estimator	Estimate	SE	Relative Efficiency	95% Confidence Interval	99% Confidence Interval
			$d_X = 100$		
$\widehat{\theta}_{\text{dim}}$	-8.489	13.737	1.000	(-35.41, 18.44)	(-43.87, 26.90)
$\widehat{\theta}_{\text{tdzl}}$(Forward-1)	-7.769	13.701	1.005	(-34.62, 19.08)	(-43.06, 27.52)
$\widehat{\theta}_{\text{tdzl}}$(Forward-2)	-10.993	13.965	0.968	(-38.36, 16.38)	(-46.96, 24.98)
$\widehat{\theta}_{\text{Zhang}}$(Forward-1)	-7.933	13.204	1.082	(-33.81, 17.95)	(-41.94, 26.08)
$\widehat{\theta}_{\text{Zhang}}$(Forward-2)	-10.083	14.413	0.908	(-38.33, 18.17)	(-47.21, 27.04)
$\widehat{\theta}_{\text{wdtt}}$(LASSO)	-9.310	13.789	0.993	(-36.34, 17.72)	(-44.83, 26.21)
$\widehat{\theta}_{\text{wdtt}}$(SCAD)	-10.493	14.165	0.941	(-38.26, 17.27)	(-46.98, 25.99)
$\widehat{\theta}_{\text{wdtt}}$(RF)	-13.732	14.065	0.954	(-41.30, 13.83)	(-49.96, 22.50)
$\widehat{\theta}_{\text{mdel}}$(LASSO)	-8.540	12.970	1.122	(-33.96, 16.88)	(-41.95, 24.87)
$\widehat{\theta}_{\text{mdel}}$(SCAD)	-8.661	12.181	1.272	(-32.53, 15.21)	(-40.04, 22.71)
$\widehat{\theta}_{\text{mdel}}$(RF)	-8.647	13.457	1.042	(-35.02, 17.73)	(-43.31, 26.02)
$\widehat{\theta}_{\text{mdel}}$(MULTI)	-7.237	10.755	1.632	(-28.32, 13.84)	(-34.94, 20.46)
			$d_X = 500$		
$\widehat{\theta}_{\text{dim}}$	-8.489	13.737	1.000	(-35.41, 18.44)	(-43.87, 26.90)
$\widehat{\theta}_{\text{tdzl}}$(Forward-1)	-7.769	13.701	1.005	(-34.62, 19.08)	(-43.06, 27.52)
$\widehat{\theta}_{\text{tdzl}}$(Forward-2)	-10.993	13.965	0.968	(-38.36, 16.38)	(-46.96, 24.98)
$\widehat{\theta}_{\text{Zhang}}$(Forward-1)	-7.933	13.204	1.082	(-33.81, 17.95)	(-41.94, 26.08)
$\widehat{\theta}_{\text{Zhang}}$(Forward-2)	-10.083	14.413	0.908	(-38.33, 18.17)	(-47.21, 27.04)
$\widehat{\theta}_{\text{wdtt}}$(LASSO)	-10.720	14.094	0.950	(-38.34, 16.90)	(-47.02, 25.58)
$\widehat{\theta}_{\text{wdtt}}$(SCAD)	-11.145	14.072	0.953	(-38.73, 16.44)	(-47.39, 25.10)
$\widehat{\theta}_{\text{wdtt}}$(RF)	-15.856	14.167	0.940	(-43.62, 11.91)	(-52.35, 20.64)
$\widehat{\theta}_{\text{mdel}}$(LASSO)	-8.193	12.881	1.137	(-33.44, 17.05)	(-41.37, 24.99)
$\widehat{\theta}_{\text{mdel}}$(SCAD)	-6.061	12.325	1.242	(-30.22, 18.10)	(-37.81, 25.69)
$\widehat{\theta}_{\text{mdel}}$(RF)	-8.648	13.420	1.048	(-34.95, 17.66)	(-43.22, 25.92)
$\widehat{\theta}_{\text{mdel}}$(MULTI)	-7.202	11.924	1.327	(-30.57, 16.17)	(-37.92, 23.51)

SE = standard error, Relative Efficiency = (SE^2 of corresponding estimator)/(SE^2 of $\widehat{\theta}_{\text{dim}}$).

The proof is completed by the Central Limit Theorem and lemma 1. □

Proofs

Proof of Proposition 1

Proof. Let $n_d^* = \frac{n_d}{K}$. n_d^* is an integer as $\mathbb{I}_k^{(d)}$, $k = 1, \cdots, K$ are of equal size for $d = 0, 1$. Write $(W_i)_{i \in \mathbb{I}_k^{(d)}} = \{W_{k1}^d, W_{k2}^d, \cdots, W_{kn_d^*}^d\}$ for $k = 1, \cdots, K$ and $d = 0, 1$ with random orders. Let $T_{udj} = \frac{1}{K}\sum_{k=1}^{K}\left[\widehat{\mathrm{G}}\left(X_{kj}^d, \widehat{g}_k^{(d)}, \widehat{\xi}^{(d)}\right)\right]_u$ for $u = 1, \cdots, r$, $j = 1, \cdots, n_d^*$ and $d = 0, 1$. It suffices to prove that 0 is contained in the convex hull of $\{T_{ud1}, \cdots, T_{udn_d^*}\}$ with probability tending to 1 as $n \to \infty$ for $u = 1, \cdots, r$ and $d = 0, 1$. To prove it, it suffices to show that for any given u and d, $\mathrm{P}(\max_{1\le j\le n_d^*} T_{udj} \le 0) \to 0$ and $\mathrm{P}(\min_{1\le j\le n_d^*} T_{udj} \ge 0) \to 0$. Now, we only prove that $\mathrm{P}(\max_{1\le j\le n_d^*} T_{udj} \le 0) \to 0$ and the proof of $\mathrm{P}(\min_{1\le j\le n_d^*} T_{udj} \ge 0) \to 0$ will be similar. Let $T_{udj}^q = \left[\widehat{g}_q^{(d)}(X_{qj}^d)\right]_u - \left[\widehat{\xi}_q^{(d)}\right]_u$ for $q = 1, \cdots, K$. It is easy to check that $T_{udj} = \frac{1}{K}\sum_{q=1}^{K} T_{udj}^q$. Simple calculation gives

$$\mathrm{P}(\max_{1\le j\le n_d^*} T_{udj} \le 0) = \mathrm{P}(\max_{1\le j\le n_d^*} \frac{1}{K}\sum_{q=1}^{K} T_{udj}^q \le 0)$$
$$\le \mathrm{P}(\min_{1\le q\le K}\max_{1\le j\le n_d^*} T_{udj}^q \le 0).$$

Therefore, it suffices to prove that $\mathrm{P}(\max_{1\le j\le n_d^*} T_{udj}^q \le 0) \to 0$ for $q = 1, \cdots, K$. The proof below follows the technique of Jing, Yuan and Zhou [13]. For a given q, let $\nu_{udj} = \psi(T_{udj}^q)$, where $\psi(x)$ is a nondecreasing, twice differentiable function with bounded first and second derivatives such that

$$\psi(x) = \begin{cases} 0, & \text{if } x \le 0 \\ a(x), & \text{if } 0 < x < \epsilon \\ 1, & \text{if } x \ge \epsilon \end{cases} \tag{8}$$

with $0 < a(x) < 1$ for $0 < x < \epsilon$. Simple algebra gives that, conditional on $(W_i)_{i \in \mathbb{I}_q^{(d)c}}$,

$$\mathrm{P}\left(\max_{1\le j\le n_d^*} T_{udj}^q \le 0\right)$$

$$
\begin{aligned}
&= \mathrm{P}\left(\nu_{ud1}=0,\cdots,\nu_{udn_d^*}=0\right) \\
&= \mathrm{P}\left(\sum_{j=1}^{n_d^*}\nu_{udj}=0\right) \\
&\le \mathrm{P}\left(\left|\sum_{j=1}^{n_d^*}(\nu_{udj}-\mathrm{E}\nu_{udj})\right| \ge n_d^*\mathrm{E}\nu_{udj}\right) \\
&\le \frac{\mathrm{E}\left[\left(\sum_{j=1}^{n_d^*}(\nu_{udj}-\mathrm{E}\nu_{udj})\right)^2\right]}{n^2(\mathrm{E}\nu_{udj})^2} \\
&\le \frac{n_d^*\mathrm{Var}(\nu_{ud1})+n_d^*(n_d^*-1)\mathrm{Cov}(\nu_{ud1},\nu_{ud2})}{n_d^{*2}(\mathrm{E}\nu_{ud1})^2}
\end{aligned}
$$

It suffices to show that, conditional on $(W_i)_{i\in\mathbb{I}_q^{(d)c}}$,

(a) $\mathrm{Var}(\nu_{ud1}) \le 1$.
(b) $\lim_{n\to\infty} \mathrm{E}\nu_{ud1} \ge c > 0$.
(c) $\mathrm{Cov}(\nu_{ud1},\nu_{ud2}) \to 0$.

(a) holds because $0 \le \nu_{ud1} \le 1$. Simple algebra indicates that

$$
\begin{aligned}
T_{udj}^q &= \left[\widehat{g}_q^{(d)}(X_{qj}^d)\right]_u - \mathrm{E}\left[\left[\widehat{g}_q^{(d)}(X_{qj}^d)\right]_u \middle| (W_i)_{i\in\mathbb{I}_q^{(d)c}}\right] \\
&\quad - \left[\widehat{\xi}_q^{(d)}\right]_u + \mathrm{E}\left[\left[\widehat{g}_q^{(d)}(X_{qj}^d)\right]_u \middle| (W_i)_{i\in\mathbb{I}_q^{(d)c}}\right] \\
&= Q_u(X_{qj}^d,\widehat{g}_q^{(d)}) + \mathrm{Rem}_{ud}^q.
\end{aligned}
$$

Under condition (**A1**), The Central Limit Theory (CLT) and Lemma 1 gives $\mathrm{Rem}_{ud}^q = O_p(\frac{1}{\sqrt{n}})$. A Taylor expansion yields

$$
\nu_{udj} = \psi(T_{udj}^q) = \psi(Q_u(X_{qj}^d,\widehat{g}_q^{(d)})) + C_n\mathrm{Rem}_{ud}^q,
$$

where $C_n < C$ for some constant C. Therefore, it is easy to verify that $\mathrm{E}\nu_{udj} \to \mathrm{E}\left[\psi(Q_u(X_{qj}^d,\widehat{g}_q^{(d)}))\right]$. Note that conditional on $(W_i)_{i\in\mathbb{I}_q^{(d)c}}$, $\mathrm{E}\left[Q_u(X_{qj}^d,\widehat{g}_q^{(d)})\right] = 0$ and $\mathrm{Var}\left(Q_u(X_{qj}^d,\widehat{g}_q^{(d)})\right) > 0$ (By condition (**A1**)). Therefore, we have $\mathrm{P}(Q_u(X_{qj}^d,\widehat{g}_q^{(d)}) > 0) > 0$, which implies that $\mathrm{E}\left[\psi(Q_u(X_{qj}^d,\widehat{g}_q^{(d)}))\right] > 0$ conditional on $(W_i)_{i\in\mathbb{I}_q^{(d)c}}$. This completes the proof of (b). (c) is obvious because conditional on $(W_i)_{i\in\mathbb{I}_q^{(d)c}}$, $\psi(Q_u(X_{q1}^d,\widehat{g}_q^{(d)}))$ and $\psi(Q_u(X_{q2}^d,\widehat{g}_q^{(d)}))$ are independent and $C_n\mathrm{Rem}_{ud}^q \to 0$ with probability tending to 1. This completes the proof. □

Proof of Proposition 2

Proof. For fixed d, from $\frac{1}{n_d}\sum_{k=1}^{K}\sum_{i\in\mathbb{I}_k^{(d)}}\frac{\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(d)},\widehat{\xi}^{(d)}\right)}{1+\widehat{\lambda}_d^\top\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(d)},\widehat{\xi}^{(d)}\right)} = 0$, simple algebra gives

$$
\begin{aligned}
0 &= \frac{1}{n_d}\sum_{k=1}^{K}\sum_{i\in\mathbb{I}_k^{(d)}}\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(d)},\widehat{\xi}^{(d)}\right)\left[1 - \frac{1}{1+\widehat{\lambda}_d^\top\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(d)},\widehat{\xi}^{(d)}\right)}\right] \\
&= \frac{1}{n_d}\sum_{k=1}^{K}\sum_{i\in\mathbb{I}_k^{(d)}}\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(d)},\widehat{\xi}^{(d)}\right) \\
&\quad - \frac{1}{n_d}\sum_{k=1}^{K}\sum_{i\in\mathbb{I}_k^{(d)}}\frac{\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(d)},\widehat{\xi}^{(d)}\right)^{\otimes 2}\widehat{\lambda}_d}{1+\widehat{\lambda}_d^\top\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(d)},\widehat{\xi}^{(d)}\right)}.
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
&\frac{1}{n_d}\sum_{k=1}^{K}\sum_{i\in\mathbb{I}_k^{(d)}}\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(d)},\widehat{\xi}^{(d)}\right) \\
&= \frac{1}{n_d}\sum_{k=1}^{K}\sum_{i\in\mathbb{I}_k^{(d)}}\frac{\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(d)},\widehat{\xi}^{(d)}\right)^{\otimes 2}\widehat{\lambda}_d}{1+\widehat{\lambda}_d^\top\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(d)},\widehat{\xi}^{(d)}\right)}.
\end{aligned}
\tag{9}
$$

For fixed k, conditional on $(W_i)_{i\in\mathbb{I}_k^{(d)c}}$, the Central Limit Theorem indicates that

$$
\begin{aligned}
&\frac{1}{|\mathbb{I}_k^{(d)}|}\sum_{i\in\mathbb{I}_k^{(d)}}\left(\widehat{g}_k^{(d)}(X_i)-\widehat{\xi}_k^{(d)}\right) \\
&\quad - \mathrm{E}\left[\widehat{g}_k^{(d)}(X)-\widehat{\xi}_k^{(d)} \middle| (W_i)_{i\in\mathbb{I}_k^{(d)c}}\right] = O_p(\frac{1}{\sqrt{n}}).
\end{aligned}
$$

Then, lemma 1 gives $\frac{1}{|\mathbb{I}_k^{(d)}|}\sum_{i\in\mathbb{I}_k^{(d)}}\left(\widehat{g}_k^{(d)}(X_i)-\widehat{\xi}_k^{(d)}\right) = O_p(\frac{1}{\sqrt{n}})$ unconditionally. Hence, the left term of (9) is

$$
\begin{aligned}
&\frac{1}{n_d}\sum_{k=1}^{K}\sum_{i\in\mathbb{I}_k^{(d)}}\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(d)},\widehat{\xi}^{(d)}\right) \\
&= \frac{1}{K}\sum_{k=1}^{K}\frac{1}{|\mathbb{I}_k^{(d)}|}\sum_{i\in\mathbb{I}_k^{(d)}}\left(\widehat{g}_k^{(d)}(X_i)-\widehat{\xi}_k^{(d)}\right) \\
&= O_p(\frac{1}{\sqrt{n}}).
\end{aligned}
$$

Turn to the right term of (9), and let $\nu_d = \frac{\widehat{\lambda}_d}{||\widehat{\lambda}_d||}$, where $||\cdot||$ is the Euclidean norm. We have

$$
\begin{aligned}
&1+\widehat{\lambda}_d^\top\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(d)},\widehat{\xi}^{(d)}\right) \\
&\le 1+||\widehat{\lambda}_d||\nu_d^\top\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(d)},\widehat{\xi}^{(d)}\right) \\
&\le 1+2||\widehat{\lambda}_d||\sqrt{r}\max_{k\in\{1,\cdots,K\}}\max_{j=1,\cdots,r}\max_{i\in\mathbb{I}_k}\left|[\widehat{g}_k^{(d)}(X_i)]_j\right|.
\end{aligned}
$$

Condition (**A1**), lemma 11.2 in Owen [19], and lemma 1 indicate that

$$\max_{k\in\{1,\cdots,K\}}\max_{j=1,\cdots,r}\max_{i\in\mathbb{I}_k}\left|[\widehat{g}_k^{(d)}(X_i)]_j\right| = o_p(n^{1/2})$$

Multiplying $\nu_d^\top$ on both sides of (9), we have

$$\begin{aligned}&||\widehat{\lambda}_d||\frac{1}{n_d}\sum_{k=1}^K\sum_{i\in\mathbb{I}_k^{(d)}}\nu_d^\top\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(d)},\widehat{\xi}^{(d)}\right)^{\otimes 2}\nu_d\\ \leq&\frac{\nu_d^\top}{n_d}\sum_{k=1}^K\sum_{i\in\mathbb{I}_k^{(d)}}\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(d)},\widehat{\xi}^{(d)}\right)\Bigg(1+2||\widehat{\lambda}_d||\sqrt{r}\max_{k\in\{1,\cdots,K\}}\\ &\max_{j=1,\cdots,r}\max_{i\in\mathbb{I}_k}\left|[\widehat{g}_k^{(d)}(X_i)]_j\right|\Bigg).\end{aligned}$$

Condition (**A2**) implies

$$\frac{1}{n_d}\sum_{k=1}^K\sum_{i\in\mathbb{I}_k^{(d)}}\nu_d^\top\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(d)},\widehat{\xi}^{(d)}\right)^{\otimes 2}\nu_d\asymp 1.$$

It follows from all above results that

$$(10)\qquad ||\widehat{\lambda}_d||\leq O_p(\frac{1}{\sqrt{n}})(1+2||\widehat{\lambda}_d||o_p(n^{1/2})).$$

Equation (10) indicates that $||\hat{\lambda}_d||=O_p(\frac{1}{\sqrt{n}})$. This completes the proof. □

Proof of Proposition 3

Proof. First, we consider the case $d=1$ and the case $d=0$ will be similar. Taylor expansion, Proposition 1, and Lemma 2 lead to

(11)

$$\begin{aligned}0=&\sqrt{n}\frac{1}{n_1}\sum_{k=1}^K\sum_{i\in\mathbb{I}_k^{(1)}}\frac{\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(1)},\widehat{\xi}^{(1)}\right)}{1+\widehat{\lambda}_1^\top\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(1)},\widehat{\xi}^{(1)}\right)}\\ =&\frac{1}{\sqrt{n}}\sum_{k=1}^K\sum_{i\in\mathbb{I}_k}\frac{D_i}{\delta}\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(1)},\widehat{\xi}^{(1)}\right)\\ &-\sqrt{n}\frac{1}{K}\sum_{k=1}^K\frac{1}{|\mathbb{I}_k|}\sum_{i\in\mathbb{I}_k}\frac{D_i}{\delta}\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(1)},\widehat{\xi}^{(1)}\right)^{\otimes 2}\widehat{\lambda}_1+o_p(1)\\ =&\frac{1}{\sqrt{n}}\sum_{k=1}^K\sum_{i\in\mathbb{I}_k}\frac{D_i-\delta}{\delta}\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(1)},\ddot{\xi}^{(1)}\right)\\ &-\sqrt{n}\frac{1}{K}\sum_{k=1}^K\frac{1}{|\mathbb{I}_k|}\sum_{i\in\mathbb{I}_k}\frac{D_i}{\delta}\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(1)},\ddot{\xi}^{(1)}\right)^{\otimes 2}\widehat{\lambda}_1+o_p(1).\end{aligned}$$

Therefore, we have

(12)

$$\sqrt{n}\widehat{\lambda}_1=\ddot{S}_n^{(1)^{-1}}\frac{1}{\sqrt{n}}\sum_{k=1}^K\sum_{i\in\mathbb{I}_k}\frac{D_i-\delta}{\delta}\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(1)},\ddot{\xi}^{(1)}\right)+o_p(1).$$

Under condition (**A3**), by Taylor expansion, Proposition 1, Lemma 2, and (12), we have

(13)

$$\begin{aligned}&\sqrt{n}\left(\widehat{\theta}_{\mathrm{mdel}}^{(1)}-\theta_1\right)\\ =&\sqrt{n}\sum_{k=1}^K\sum_{i\in\mathbb{I}_k}D_i\widehat{p}_i\left(Y_i-\theta_1\right)\\ =&\frac{1}{\sqrt{n}}\sum_{k=1}^K\sum_{i\in\mathbb{I}_k}\frac{D_i}{\delta}\frac{Y_i-\theta_1}{1+\widehat{\lambda}_1^\top\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(1)},\widehat{\xi}^{(1)}\right)}+o_p(1)\\ =&\frac{1}{\sqrt{n}}\sum_{k=1}^K\sum_{i\in\mathbb{I}_k}\frac{D_i}{\delta}\left(Y_i-\theta_1\right)\\ &-\frac{1}{\sqrt{n}}\sum_{k=1}^K\sum_{i\in\mathbb{I}_k}\frac{D_i}{\delta}\left(Y_i-\theta_1\right)\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(1)},\ddot{\xi}^{(1)}\right)\widehat{\lambda}_1+o_p(1)\\ =&\frac{1}{\sqrt{n}}\sum_{k=1}^K\sum_{i\in\mathbb{I}_k}\frac{D_i}{\delta}\left(Y_i-\theta_1\right)\\ &-\ddot{J}_n^{(1)^\top}\ddot{S}_n^{(1)^{-1}}\frac{1}{\sqrt{n}}\sum_{k=1}^K\sum_{i\in\mathbb{I}_k}\frac{D_i-\delta}{\delta}\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(1)},\ddot{\xi}^{(1)}\right)\\ &+o_p(1)\\ =&\frac{1}{\sqrt{n}}\sum_{k=1}^K\sum_{i\in\mathbb{I}_k}\Bigg[\frac{D_i}{\delta}\left(Y_i-\theta_1\right)\\ &-\frac{D_i-\delta}{\delta}\ddot{J}_n^{(1)^\top}\ddot{S}_n^{(1)^{-1}}\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(1)},\ddot{\xi}^{(1)}\right)\Bigg]+o_p(1)\end{aligned}$$

It is easy to give the form of $\sqrt{n}\left(\widehat{\theta}_{\mathrm{mdel}}^{(0)}-\theta_0\right)$ in a similar way:

$$\begin{aligned}\sqrt{n}\left(\widehat{\theta}_{\mathrm{mdel}}^{(0)}-\theta_0\right)=&\frac{1}{\sqrt{n}}\sum_{k=1}^K\sum_{i\in\mathbb{I}_k}\Bigg[\frac{1-D_i}{1-\delta}\left(Y_i-\theta_0\right)\\ &-\frac{D_i-\delta}{1-\delta}\ddot{J}_n^{(0)^\top}\ddot{S}_n^{(0)^{-1}}\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(0)},\ddot{\xi}^{(0)}\right)\Bigg]\\ &+o_p(1).\end{aligned}$$

Above results give that

$$\begin{aligned}\sqrt{n}\left(\widehat{\theta}_{\mathrm{mdel}}-\theta\right)=&\frac{1}{\sqrt{n}}\sum_{k=1}^K\sum_{i\in\mathbb{I}_k}\Bigg[\frac{D_i}{\delta}\left(Y_i-\theta_1\right)\\ &-\frac{D_i-\delta}{\delta}\ddot{J}_n^{(1)^\top}\ddot{S}_n^{(1)^{-1}}\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(1)},\ddot{\xi}^{(1)}\right)\\ &-\frac{1-D_i}{1-\delta}\left(Y_i-\theta_0\right)\\ &+\frac{D_i-\delta}{1-\delta}\ddot{J}_n^{(0)^\top}\ddot{S}_n^{(0)^{-1}}\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(0)},\ddot{\xi}^{(0)}\right)\Bigg]\\ &+o_p(1).\end{aligned}$$

This completes the proof. □

Proof of Theorem 4

Proof. When $r=1$, conditional on $(W_i)_{i\in\mathbb{I}_k^{(d)c}}$, the Holder Inequality gives that

$$\begin{aligned}&\mathrm{E}\left[\left|\widehat{g}_k^{(d)}(X)-\eta^{(d)}(X)\right|\quad\left|(W_i)_{i\in\mathbb{I}_k^{(d)c}}\right.\right]\\ \leq&\sqrt{\mathrm{E}\left[\left(\widehat{g}_k^{(d)}(X)-\eta^{(d)}(X)\right)^2\middle|(W_i)_{i\in\mathbb{I}_k^{(d)c}}\right]}.\end{aligned}$$

Therefore, we have $\mathrm{E}\left[\left|\widehat{g}_k^{(d)}(X)-\eta^{(d)}(X)\right|\quad\left|(W_i)_{i\in\mathbb{I}_k^{(d)c}}\right.\right]\to 0$ in probability as $n\to\infty$ for $k=1,\cdots,K$. Let $\mathrm{G}(X_i,\eta^{(d)},\theta_d)=\eta^{(d)}(X_i)-\theta_d$. For simplicity, write $\varsigma_k^{(d)}(X_i)=\widehat{g}_k^{(d)}(X_i)-\eta^{(d)}(X_i)$. It is straightforward to show that $\widehat{\mathrm{G}}\left(X_i,\widehat{g}_k^{(d)},\ddot{\xi}^{(d)}\right)-\mathrm{G}(X_i,\eta^{(d)},\theta_d)=\varsigma_k^{(d)}(X_i)+o_p(1)$ by lemma 1. Following (11), it is easy to verify that

(14)

$$\begin{aligned}0=&\frac{1}{\sqrt{n}}\sum_{i\in I}\frac{D_i-\delta}{\delta}\mathrm{G}(X_i,\eta^{(1)},\theta_1)\\ &-\sqrt{n}\frac{1}{n}\sum_{i\in I}\frac{D_i}{\delta}\mathrm{G}(X_i,\eta^{(1)},\theta_1)^2\widehat{\lambda}_1+o_p(1)\\ &+\frac{1}{\sqrt{n}}\sum_{k=1}^K\sum_{i\in\mathbb{I}_k}\frac{D_i-\delta}{\delta}\varsigma_k^{(1)}(X_i)+\frac{1}{\sqrt{n}}\sum_{k=1}^K\sum_{i\in\mathbb{I}_k}\varsigma_k^{(1)}(X_i)^2\widehat{\lambda}_1\\ &-\frac{2}{\sqrt{n}}\sum_{k=1}^K\sum_{i\in\mathbb{I}_k}\frac{D_i}{\delta}\varsigma_k^{(1)}(X_i)\mathrm{G}(X_i,\eta^{(1)},\theta_1)\widehat{\lambda}_1.\end{aligned}$$

Now, we bound

$$A=\frac{1}{\sqrt{n}}\sum_{k=1}^K\sum_{i\in\mathbb{I}_k}\frac{D_i-\delta}{\delta}\varsigma_k^{(1)}(X_i),$$

$$B=\frac{1}{\sqrt{n}}\sum_{k=1}^K\sum_{i\in\mathbb{I}_k}\varsigma_k^{(1)}(X_i)^2\widehat{\lambda}_1$$

and

$$C=\frac{1}{\sqrt{n}}\sum_{k=1}^K\sum_{i\in\mathbb{I}_k}\frac{D_i}{\delta}\varsigma_k^{(1)}(X_i)\mathrm{G}(X_i,\eta^{(1)},\theta_1)\widehat{\lambda}_1,$$

respectively. Conditional on $(W_i)_{i\in\mathbb{I}_k^{(1)c}}$, the mean of $\frac{1}{\sqrt{|\mathbb{I}_k|}}\sum_{i\in\mathbb{I}_k}\frac{D_i-\delta}{\delta}\varsigma_k^{(1)}(X_i)$ is zero and the variance is given by

$$\mathrm{E}[(D-\delta)^2]\cdot\mathrm{E}\left[\varsigma_k^{(1)}(X)^2\middle|(W_i)_{i\in\mathbb{I}_k^{(1)c}}\right],$$

which converges to zero in probability as $n\to\infty$. Then $A=o_p(1)$ by the Chebyshev's Inequality and lemma 1. B vanishes in probability because $\sqrt{n}\widehat{\lambda}_1=O_p(1)$. For C, the Cauchy-Schwarz Inequality gives that

$$\begin{aligned}C\leq&\sqrt{n}\widehat{\lambda}_1\frac{1}{K}\sum_{k=1}^K\sqrt{\frac{1}{|\mathbb{I}_k|}\sum_{i\in\mathbb{I}_k}\varsigma_k^{(1)}(X_i)^2}\\ &\times\sqrt{\frac{1}{|\mathbb{I}_k|}\sum_{i\in\mathbb{I}_k}\left(\frac{D_i}{\delta}\mathrm{G}(X_i,\eta^{(1)},\theta_1)\right)^2}.\end{aligned}$$

Conditional on $(W_i)_{i\in\mathbb{I}_k^{(d)c}}$, the right term of above inequality converges to 0 in probability as $n\to\infty$; therefore $C=o_p(1)$ by lemma 1. Above results give that

$$\begin{aligned}&\sqrt{n}\widehat{\lambda}_1\\ =&\mathrm{E}\left[\mathrm{G}(X_i,\eta^{(1)},\theta_1)^2\right]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n\frac{D_i-\delta}{\delta}\mathrm{G}(X_i,\eta^{(1)},\theta_1)\\ &+o_p(1).\end{aligned}$$

Similarly, it is easy to check that

$$\begin{aligned}&\sqrt{n}\widehat{\lambda}_0\\ =&\mathrm{E}\left[\mathrm{G}(X_i,\eta^{(0)},\theta_0)^2\right]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n\frac{D_i-\delta}{1-\delta}\mathrm{G}(X_i,\eta^{(0)},\theta_0)\\ &+o_p(1).\end{aligned}$$

Using the above results, Taylor expansion indicates that

$$\begin{aligned}&\sqrt{n}\left(\widehat{\theta}_{\mathrm{mdel}}^{(1)}-\theta_1\right)\\ =&\frac{1}{\sqrt{n}}\sum_{i=1}^n\frac{D_i}{\delta}\left(Y_i-\theta_1\right)\\ &-\frac{1}{\sqrt{n}}\sum_{i=1}^n\frac{D_i}{\delta}\left(Y_i-\theta_1\right)\mathrm{G}(X_i,\eta^{(1)},\theta_1)\widehat{\lambda}_1+o_p(1)\\ =&\frac{1}{\sqrt{n}}\sum_{i=1}^n\left\{\frac{D_i}{\delta}\left(Y_i-\theta_1\right)\right.\\ &-\frac{D_i-\delta}{\delta}\mathrm{E}\left[\frac{D}{\delta}(Y-\theta_1)\mathrm{G}(X,\eta^{(1)},\theta_1)\right]\\ &\left.\times\mathrm{E}\left[\mathrm{G}(X,\eta^{(1)},\theta_1)^2\right]^{-1}\mathrm{G}(X,\eta^{(1)},\theta_1)\right\}+o_p(1).\end{aligned}$$

Following from

$$\begin{aligned}&\mathrm{E}\left[\frac{D}{\delta}(Y-\theta_1)\mathrm{G}(X,\eta^{(1)},\theta_1)\right]\\ =&\frac{\mathbb{P}(D=1)}{\delta}\mathrm{E}\left[(Y-\theta_1)\mathrm{G}(X,\eta^{(1)},\theta_1)|D=1\right]\\ =&\mathrm{E}\left[\mathrm{E}\left[(Y-\theta_1)|X,D=1\right]\mathrm{G}(X,\eta^{(1)},\theta_1)\right]\\ =&\mathrm{E}\left[\mathrm{G}(X,\eta^{(1)},\theta_1)^2\right],\end{aligned}$$

we have

$$\sqrt{n}\left(\widehat{\theta}^{(1)}_{\rm mdel}-\theta_1\right)$$
$$=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{D_i}{\delta}\left(Y_i-\theta_1\right)-\frac{D_i-\delta}{\delta}\left(\eta^{(1)}(X_i)-\theta_1\right)\right\}$$
$$+o_p(1)$$
$$=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{D_i}{\delta}\left(Y_i-\eta^{(1)}(X_i)\right)+\left(\eta^{(1)}(X_i)-\theta_1\right)\right\}$$
$$+o_p(1).$$

Similarly, when $d=0$, it is easy to verify that

$$\sqrt{n}\left(\widehat{\theta}^{(0)}_{\rm mdel}-\theta_0\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{1-D_i}{1-\delta}\left(Y_i-\eta^{(0)}(X_i)\right)+\left(\eta^{(0)}(X_i)-\theta_0\right)\right\}+o_p(1).$$

Therefore, we have

$$\sqrt{n}\left(\widehat{\theta}_{\rm mdel}-\theta\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{D_i}{\delta}\left(Y_i-\eta^{(1)}(X_i)\right)-\frac{1-D_i}{1-\delta}\left(Y_i-\eta^{(0)}(X_i)\right)+\eta^{(1)}(X_i)-\eta^{(0)}(X_i)-\theta\right\}+o_p(1).$$

This completes the proof. □

Proof of Theorem 5

Proof. Following the proof of Theorem 4, it is easy to verify that

$$\widehat{J}^{(d)}_n=\frac{1}{n_d}\sum_{k=1}^{K}\sum_{i\in\mathbb{I}_k^{(d)}}Y_i\widehat{\mathrm{G}}\left(X_i,\eta^{(1)},\theta_1\right)+o_p(1)$$

and

$$\widehat{S}^{(d)}_n=\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathbb{I}_k}\widehat{\mathrm{G}}\left(X_i,\eta^{(1)},\theta_1\right)^2+o_p(1).$$

Then, some algebra gives

(15)
$$\widehat{\sigma}^2_{\rm mdel}=\frac{1}{n}\sum_{d=0,1}\sum_{k=1}^{K}\sum_{i\in\mathbb{I}_k^{(d)}}\frac{n_d}{n}\widehat{p}_i\left\{\frac{n}{n_1}D_i(Y_i-\eta^{(1)}(X_i)+\theta_1-\widehat{\theta}^{(1)}_{\rm mdel})+\eta^{(1)}(X_i)-\theta_1-\frac{n}{n_0}(1-D_i)(Y_i-\eta^{(0)}(X_i)+\theta_0-\widehat{\theta}^{(0)}_{\rm mdel})-\eta^{(0)}(X_i)+\theta_0\right\}^2+o_p(1)$$
$$=\frac{1}{n^2}\sum_{i=1}^{n}\left\{\frac{D_i}{\delta}\left(Y_i-\eta^{(1)}(X_i)\right)-\frac{1-D_i}{1-\delta}\left(Y_i-\eta^{(0)}(X_i)\right)+\eta^{(1)}(X_i)-\eta^{(0)}(X_i)-\theta\right\}^2+o_p(1).$$

This completes the proof. □

SUPPLEMENTARY FILES

The R-code of the modified Newton-Raphson algorithm for the empirical likelihood optimization problem is available at http://intlpress.com/site/pub/files/_supp/sii/2022/0015/0003/SII-2022-0015-0003-s001.zip.

ACKNOWLEDGEMENTS

We are very grateful to two reviewers for their helpful and constructive comments that have greatly improved our original submission. Ying Yan's research is supported by the National Natural Science Foundation of China (NSFC) (Grant No. 11901599).

Received 5 October 2020

REFERENCES

[1] Belloni, A., Chernozhukov, V., Fernández-Val, I. and Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica* **85** 233–298. MR3611771

[2] Breiman, L. (2001). Random forests. *Machine learning* **45** 5–32. MR3874153

[3] Chen, J., Sitter, R. and Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika* **89** 230–237. MR1888365

[4] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* **21** C1–C68. MR3769544

[5] Fan, J., Guo, S. and Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74** 37–65. MR2885839

[6] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association* **96** 1348–1360. MR1946581

[7] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70** 849–911. MR2530322

[8] Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M. et al. (1996). A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine* **335** 1081–1090.

[9] Han, P. (2014). Multiply robust estimation in regression analysis with missing data. *Journal of the American Statistical Association* **109** 1159–1173. MR3265688

[10] Han, P. and Wang, L. (2013). Estimation with missing data: beyond double robustness. *Biometrika* **100** 417–430. MR3068443

[11] Huang, C.-Y., Qin, J. and Follmann, D. A. (2008). Empirical likelihood-based estimation of the treatment effect in a pretest–posttest study. *Journal of the American Statistical Association* **103** 1270–1280. MR2462898

[12] Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences: an introduction*. Cambridge University Press. MR3309951

[13] Jing, B.-Y., Yuan, J. and Zhou, W. (2009). Jackknife empirical likelihood. *Journal of the American Statistical Association* **104** 1224–1232. MR2562010

[14] Lee, W.-C. and Lin, J.-H. (2019). A test for treatment effects in randomized controlled trials, harnessing the power of ultrahigh dimensional big data. *Medicine* **98**.

[15] Leon, S., Tsiatis, A. A. and Davidian, M. (2003). Semiparametric estimation of treatment effect in a pretest–posttest study. *Biometrics* **59** 1046–1055. MR2025129

[16] Muscedere, J., Maslove, D. M., Boyd, J. G., O'Callaghan, N., Sibley, S., Reynolds, S., Albert, M., Hall, R., Jiang, X., Day, A. G. et al. (2018). Prevention of nosocomial infections in critically ill patients with lactoferrin: a randomized, double-blind, placebo-controlled study. *Critical Care Medicine* **46** 1450–1456.

[17] Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75** 237–249. MR0946049

[18] Owen, A. B. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics* 90–120. MR1041387

[19] Owen, A. B. (2001). *Empirical likelihood*. CRC press.

[20] Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *the Annals of Statistics* 300–325. MR1272085

[21] Qin, J. and Zhang, B. (2007). Empirical-likelihood-based inference in missing response problems and its application in observational studies. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69** 101–122. MR2301502

[22] Tan, Y., Wen, X., Liang, W. and Yan, Y. (2020). Empirical Likelihood Weighted Estimation of Average Treatment Effects. *arXiv preprint arXiv:2008.12989*.

[23] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58** 267–288. MR1379242

[24] Tsiatis, A. (2007). *Semiparametric theory and missing data*. Springer Science & Business Media. MR2233926

[25] Tsiatis, A. A., Davidian, M., Zhang, M. and Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Statistics in medicine* **27** 4658–4677. MR2528575

[26] Wager, S., Du, W., Taylor, J. and Tibshirani, R. J. (2016). High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences* **113** 12673–12678. MR3576188

[27] Wu, C. (2004). Some algorithmic aspects of the empirical likelihood method in survey sampling. *Statistica Sinica* 1057–1067. MR2126339

[28] Wu, C. and Yan, Y. (2012). Empirical likelihood inference for two-sample problems. *Statistics and its Interface* **5** 345–354. MR2979214

[29] Zhang, B. (2016). Empirical likelihood in causal inference. *Econometric Reviews* **35** 201–231. MR3427150

[30] Zhang, B. (2018). Empirical likelihood inference in randomized clinical trials. *Statistical Methods in Medical Research* **27** 3770–3784. MR3878656

Wei Liang
School of Mathematics
Sun Yat-sen University
No. 135, Xingang Xi Road
Guangzhou 510275
China
E-mail address: liangw53@mail2.sysu.edu.cn

Ying Yan
School of Mathematics
Sun Yat-sen University
No. 135, Xingang Xi Road
Guangzhou 510275
China
E-mail address: yanying7@mail.sysu.edu.cn