Annals of the Institute of Statistical Mathematics (2024) 76:73–92
https://doi.org/10.1007/s10463-023-00876-4

# Model averaging for estimating treatment effects

**Zhihao Zhao[1] · Xinyu Zhang[2] · Guohua Zou[3] · Alan T. K. Wan[4] · Geoffrey K. F. Tso[4]**

Received: 16 August 2022 / Revised: 17 March 2023 / Accepted: 9 May 2023 /
Published online: 30 June 2023


## Abstract

The estimation of treatment effects on the response variable is often a primary goal in empirical investigations in disciplines such as medicine, economics and marketing. Typically, the investigator would select one model from a multitude of models and estimate the treatment effects based on this single winning model. In this paper, we consider an alternative model averaging approach, where estimates of treatment effects are obtained from not one single model but a weighted ensemble of models. We develop a weight choice method based on a minimisation of the approximate risk under squared error loss of the model average estimator of the conditional treatment effects. We prove that the model average estimator resulting from this criterion has an optimal asymptotic property. The results of a simulation study show that the proposed approach is superior to various existing model selection and averaging methods in a large region of the parameter space in finite samples. The proposed method is applied to a data set on HIV treatment.

**Keywords** Model average · Treatment effects · Causal inference · Asymptotic optimality

Zhao's work was supported by Capital University of Economics and Business: The Fundamental Research Funds for Beijing Universities (Grant No. XRZ2021042) and Youth Academic Innovation Team Construction project of Capital University of Economics and Business (Grant No. QNTD202303). Zhang's work was partially supported by the National Natural Science Foundation of China (Grant Nos. (71925007, 72091212 and 12288201) and the CAS Project for Young Scientists in Basic Research (YSBR-008). Zou's work was partially supported by the National Natural Science Foundation of China (Grant Nos. 11971323, 12031016). Wan's work was supported by the Hong Kong Research Grants Council (Grant No. 11500419) and the National Natural Science Foundation of China (Grant No. 71973116).

Extended author information available on the last page of the article

## 1 Introduction

Inferring the causal effects of a treatment on a response is often a primary goal of a statistical analysis. The term 'treatment effect' refers to the causal effect of a binary (0, 1) variable on an outcome variable of interest. Treatment effects abound in economics, marketing, medicine and many other fields. Examples include the effects of training programmes, college degrees, drug use, etc. One major problem associated with treatment effect estimation is the selection bias that results from the differences between treated and non-treated observations due to reasons other than the treatment status *per se*. There is now a considerable body of literature on treatment effect estimation. Some well-known examples include (Ashenfelter 1978, Ashenfelter and Card 1984, LaLonde 1986, Abadie and Imbens 2011), among others.

One common method of estimating treatment effects is by regressing the response against a treatment indicator variable and some baseline covariates. An important problem is how to deal with the uncertainty of which covariates should be used. Belloni and Hansen (2014) proposed robust methods for inference about the effect of a treatment variable on a scalar outcome in the presence of many covariates in a model with possibly non-Gaussian and heteroscedastic disturbances. Rolling and Yang (2014) proposed a treatment effect cross-validation (TECV) criterion that selects the model with the smallest treatment effect estimation errors. This is different from traditional cross-validation criteria, which typically emphasise the model's prediction errors of the response variable. Not surprisingly, as the objective has changed, a model that performs well for prediction is not necessarily good for estimating treatment effect. Lee et al. (2017) and Belloni et al. (2017) also provided some methods for handling model uncertainty when estimating treatment effects.

In addition to the above methods, another approach to dealing with model uncertainty that is enjoying rising popularity is model averaging, where estimates are obtained based on not one single model but a weighted ensemble of models. A large part of the model averaging literature is about finding optimal weights for computing the model average oriented towards some form of optimality. The weight choice methods that have been developed include weighting based on information scores (Buckland 1997), adaptive regression by mixing (ARM) (Yang 2001; Yuan and Yang 2005), the Mallows criterion (e.g. (Hansen 2007; Zhu et al. 2019)), minimisation of estimator's MSE (Liang et al. 2011), cross-validation (CV) (Hansen and Racine 2012), Kullback–Leibler-type measures (Zhang et al. 2015) and parsimonious model averaging (PMA) (Zhang et al. 2020). Some studies have also examined inference after model averaging (e.g. Hjort and Claeskens (2003); Liu (2015); Zhang and Liu (2019)). For the alternative Bayesian approach, readers may consult Hoeting et al. (1999).

The literature on model averaging typically emphasises the models' predictive accuracy; few studies have considered the estimation of treatment effects. Kuersteiner and Okui (2010) and Lee and Shin (2021) applied the model averaging approach to the first stage of the two-stage least squares estimator. Seng and Li

(2022) investigated a model averaging approach to estimate the treatment effects in observational studies in order to incorporate the instruments into the two-stage least square estimation procedure. Kitagawa and Muris (2016) developed a model averaging method that minimises the approximated MSE of a semiparametric estimator related to treatment effects. They proved that this model averaging procedure is Bayes optimal with respect to a given prior. Rolling et al. (2019) proposed a model averaging method known as treatment effect estimation by mixing (TEEM) based on the idea of ARM but focused on the conditional average treatment effects. In this paper, we propose a model averaging method based on a minimisation of an unbiased estimator of the expected squared error loss of the model average estimator of the conditional average treatment effects. This is in the spirits of Liang et al. (2011), who focused on predictive accuracy. Although our approach bears some similarity to that of Kitagawa and Muris (2016), there are important differences. The objective function of our analysis is the sum of squared errors for conditional average treatment effects, which differs from the mean squared error criterion of the average treatment effect on which Kitagawa and Muris (2016) focused. In particular, the work of Kitagawa and Muris (2016) was developed under a local neighbourhood asymptotic framework (Hjort and Claeskens 2003). In this paper, we consider a fixed parameter asymptotic framework. Our method is also entirely different from TEEM, which is an adaptive, but also a computationally intensive procedure that requires a large number of random permutations of the order of observations. We prove that our proposed weight choice criterion has an optimal asymptotic property.

The remainder of this paper is organised as follows. Section 2 describes the model framework. In Sect. 3, we develop our weight choice criterion, and demonstrate its asymptotic optimality. Section 4 is devoted to an investigation of the finite sample properties of the proposed approach. Section 5 applies the proposed method to a data set on HIV treatment. Technical proofs are given in supplementary material.

## 2 Model framework

Our model framework follows closely that of Rolling and Yang (2014). Let $Y_i$ be the response variable, and $T_i \in \{t, c\}$ be a binary variable such that the individual $i$ is under either treatment ($T_i = t$) or control ($T_i = c$) (i.e. no treatment). We write

$$Y_i = \left[f_t(\boldsymbol{u}_i) + \zeta_i\right] I(T_i = t) + \left[f_c(\boldsymbol{u}_i) + v_i\right] I(T_i = c), \qquad 1 \le i \le n, \tag{1}$$

where $\boldsymbol{u}_i$ represents a set of $p$ pre-treatment covariates and $\zeta_i$ and $v_i$ are the random errors under treatment and control, respectively. We assume that

$$0 < P(T_i = t | \boldsymbol{u}_i = \boldsymbol{u}) < 1, \tag{2}$$

$$E(\zeta_i | \boldsymbol{u}_i) = E(v_i | \boldsymbol{u}_i) = 0, \ \zeta_i \perp v_j \text{ for } i \neq j, \tag{3}$$

where '$\perp$' denotes independence, and

$$E(\zeta_i^2|\boldsymbol{u}_i) = \sigma_t^2,\ E(v_i^2|\boldsymbol{u}_i) = \sigma_c^2, \quad 1 \le i,j \le n. \tag{4}$$

Following Imbens and Wooldridge (2009) and Rolling and Yang (2014), we conceptualise the treatment effect on $Y$ in terms of 'potential outcomes' as in the Rubin causal model (RCM) (Holland 1986). Let $Y_{it}$ and $Y_{ic}$ denote the potential outcomes when $T_i = t$ and $T_i = c$ , respectively. The effect of $T_i$ on $Y_i$ is thus equal to $Y_{it} - Y_{ic}$, which is unobserved because for any given $i$, only one of $Y_{it}$ and $Y_{ic}$ can be observed. Holland (1986) suggested inferring on $Y_{it} - Y_{ic}$ based on the average of $Y_{it} - Y_{ic}$ over all $i$'s. Typically, treatment effects are heterogeneous with respect to the covariates. Allowing for this heterogeneity, we define the conditional average treatment effect (CATE) as

$$\text{CATE}(\boldsymbol{u}) := E[(Y_{it} - Y_{ic})|\boldsymbol{u_i} = \boldsymbol{u}]. \tag{5}$$

To connect (1) with (5), we need a condition that requires the covariates to contain all potential confounding information of the relationship between the treatment and potential outcomes, i.e. $\{Y_{it}, Y_{ic}\} \perp T_i|\boldsymbol{u}_i$. We label this condition as Assumption 1 here. It always holds in randomised experiments. Rolling and Yang (2014) also assumed the same condition and they referred to it as the condition of unconfounded assignment. As Rolling and Yang (2014) remarked, this assumption is always satisfied under randomised experiments but not necessarily under observational studies. For the latter, increasing the number of covariates can enhance the plausibility of the assumption. Under Assumption 1, we can write

$$\begin{aligned}
\text{CATE}(\boldsymbol{u}) =& E[(Y_{it} - Y_{ic})|\boldsymbol{u_i} = \boldsymbol{u}] \\
=& E(Y_{it}|\boldsymbol{u_i} = \boldsymbol{u}) - E(Y_{ic}|\boldsymbol{u_i} = \boldsymbol{u}) \\
=& E(Y_{it}|T_i = t, \boldsymbol{u_i} = \boldsymbol{u}) - E(Y_{ic}|T_i = c, \boldsymbol{u_i} = \boldsymbol{u}) \\
=& E(Y_i|T_i = t, \boldsymbol{u_i} = \boldsymbol{u}) - E(Y_i|T_i = c, \boldsymbol{u_i} = \boldsymbol{u}) \\
=& f_t(\boldsymbol{u}) - f_c(\boldsymbol{u}) \\
:=& \Delta(\boldsymbol{u}).
\end{aligned} \tag{6}$$

Our objective here is to estimate $\Delta(\boldsymbol{u})$, the difference between the regression functions under treatment and control, which may be interpreted as the conditional average treatment effect. Hereafter, we will simply refer to $\Delta(\boldsymbol{u})$ as the treatment effect.

Clearly, different subsets of $u_i$ will give rise to different models of $Y_i$ and using all components of $u_i$ may lead to an inefficient estimator of treatment effect. Let there be $K$ candidate (or approximating) models of $Y_i$. Denote the $i$th vector of observations of covariates under the $k$th approximating model as $\boldsymbol{u}_i^{(k)} = (u_{1i}^{(k)}, \ldots, u_{l_k i}^{(k)})'$, where $l_k$ is the number of covariates, $i = 1, \cdots, n$. The $k$th approximating model of $Y_i$ is thus given by

$$\begin{aligned}
Y_i = \left[\boldsymbol{u}_i^{(k)\prime}\boldsymbol{\beta}^{(k)} + \zeta_i\right] I(T_i = t) + \left[\boldsymbol{u}_i^{(k)\prime}\boldsymbol{\gamma}^{(k)} + v_i\right] I(T_i = c), \\
i = 1, \ldots, n, k = 1, \ldots, K,
\end{aligned} \tag{7}$$

where $\boldsymbol{\beta}^{(k)}$ and $\boldsymbol{\gamma}^{(k)}$ are unknown coefficient vectors to be estimated. We call the subsamples of $Y_i$ comprising the observations of $Y_i = t$ and $Y_i = c$ the treatment and control samples, respectively.

Let $\boldsymbol{Y}_{ta}$ $(n_t \times 1)$ and $\boldsymbol{Y}_{ca}$ $(n_c \times 1)$ be the vectors of observations of $Y_i$ in the treatment and control samples, respectively, with the corresponding number of observations given by $n_t$ and $n_c$ , respectively. For the $k$th approximating model, let $\boldsymbol{U}_{ta}^{(k)}$ $(n_t \times l_k)$ and $\boldsymbol{U}_{ca}^{(k)}$ $(n_c \times l_k)$ contain the covariate observations corresponding to the treatment and control samples, respectively. It is assumed that $\boldsymbol{U}_{ta}^{(k)}$ and $\boldsymbol{U}_{ca}^{(k)}$ are of full column rank. Under this framework, the least squares estimators of $\boldsymbol{\beta}^{(k)}$ and $\boldsymbol{\gamma}^{(k)}$ are $\hat{\boldsymbol{\beta}}^{(k)} = (\boldsymbol{U}_{ta}^{(k)\prime}\boldsymbol{U}_{ta}^{(k)})^{-1}\boldsymbol{U}_{ta}^{(k)\prime}\boldsymbol{Y}_{ta}$ and $\hat{\boldsymbol{\gamma}}^{(k)} = (\boldsymbol{U}_{ca}^{(k)\prime}\boldsymbol{U}_{ca}^{(k)})^{-1}\boldsymbol{U}_{ca}^{(k)\prime}\boldsymbol{Y}_{ca}$ , respectively. The estimator of the treatment effect $\Delta(\boldsymbol{u}_i)$ under the $k$th model is thus equal to

$$\hat{\Delta}_i^{(k)} = \boldsymbol{u}_i^{(k)\prime}\hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{u}_i^{(k)\prime}\hat{\boldsymbol{\gamma}}^{(k)}. \tag{8}$$

As discussed in Sect. 1, Rolling and Yang (2014) proposed the TECV model selection criterion targeted for treatment effect estimation. Here, instead of drawing inference based on one model, we derive the estimator of the treatment effect based on a weighted ensemble of the candidate models. Now, let $\boldsymbol{w} = (w_1, ..., w_K)'$ be a weight vector belonging to the set

$$H_n = \left\{ \boldsymbol{w} \in [0,1]^K \,:\, \sum\nolimits_{k=1}^{K} w_k = 1 \right\}.$$

The model average estimator of $\Delta(\boldsymbol{u}_i)$ is given by

$$\hat{\Delta}_i(\boldsymbol{w}) = \sum\nolimits_{k=1}^{K} w_k \hat{\Delta}_i^{(k)}, \tag{9}$$

which combines the estimators from different models. The question is how to determine $\boldsymbol{w}$ in order to obtain desirable properties of $\hat{\Delta}_i(\boldsymbol{w})$. We address this question in the next section.

## 3 Weight choice and asymptotic optimality

### 3.1 Partition and match

Our estimation of $\hat{\Delta}_i(\boldsymbol{w})$ and method for choosing $\boldsymbol{w}$ are based on partitioning the feature space. Write $n_1 = \min(n_t, n_c)$. Let $\mathscr{U} \subset \mathscr{R}^p$ be the support of the probability density of $\boldsymbol{u}$, and denote the lower bound of the covariate density by $\underline{c} > 0$. The goal is to partition $\mathscr{U}$ into cells of appropriate sizes. Without loss of generality, we let $\mathscr{U} = [0,1]^p$, $h$ be the side length of cells, and $n_2$ be the number of cells.

From each cell, we randomly select a treatment–control pair $(i, i^*)$ such that $T_i = t$ and $T_{i^*} = c$. Following Rolling and Yang (2014), after choosing a suitable $h$, a treatment–control pair can be found with high probability. Let $\boldsymbol{u}_m^t$ and $\boldsymbol{u}_m^c$ be the vectors of covariates corresponding to the treatment and control observations of the $m$th treatment–control pair, respectively, and $Y_m^t$, $Y_m^c$, $\zeta_m^t$ and $v_m^c$ be the corresponding

responses and errors. Typically, as $h$ goes to zero, $\boldsymbol{u}_m^t$ approaches $\boldsymbol{u}_m^c$. Hence, the treatment effect $\Delta(\boldsymbol{u}_m^t)$ may be approximated by $Y_m^t - Y_m^c$. Let $b_m = f_c(\boldsymbol{u}_m^t) - f_c(\boldsymbol{u}_m^c)$ and $\eta_m = \zeta_m^t - v_m^c$. It is readily seen that

$$\begin{aligned} Y_m^t - Y_m^c &= f_t(\boldsymbol{u}_m^t) + \zeta_m^t - \left(f_c(\boldsymbol{u}_m^c) + v_m^c\right) \\ &= \left[f_t(\boldsymbol{u}_m^t) - f_c(\boldsymbol{u}_m^t)\right] + \left[f_c(\boldsymbol{u}_m^t) - f_c(\boldsymbol{u}_m^c)\right] + \left[\zeta_m^t - v_m^c\right] \\ &= \Delta(\boldsymbol{u}_m^t) + b_m + \eta_m, \end{aligned} \tag{10}$$

where $\eta_m$ may be interpreted as an error term, and $b_m$ can be thought of as the bias arising from the approximation of $\Delta(\boldsymbol{u}_m^t)$ by $Y_m^t - Y_m^c$.

After partitioning the feature using $h$ as in Rolling and Yang (2014), the bias terms have the uniform bound

$$\sup_{1\le m\le n_2} |b_m| \le C\left\{ \left\lfloor \left( \frac{cn_1}{2\log n_1} \right)^{1/p} \right\rfloor \right\}^{-1} = O\left( p\left( \frac{\log n}{n} \right)^{1/p} \right), \tag{11}$$

where $C$ is a constant. See Lemma 1 in supplementary material of Rolling and Yang (2014) for details.

### 3.2 Weight choice criterion

Unless otherwise stated, all limiting processes discussed in this and subsequent sections are with respect to $n \to \infty$. Now, let $\hat{\boldsymbol{\Delta}}^{(k)} = (\hat{\Delta}_1^{(k)}, \ldots, \hat{\Delta}_{n_2}^{(k)})'$, $\hat{\boldsymbol{\Delta}}(\boldsymbol{w}) = (\hat{\Delta}_1(\boldsymbol{w}), \ldots, \hat{\Delta}_{n_2}(\boldsymbol{w}))'$, $\boldsymbol{\Delta} = (\Delta(\boldsymbol{u}_1^t), \ldots, \Delta(\boldsymbol{u}_{n_2}^t))'$, and $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)'$. Define the squared error and the corresponding conditional risk of $\boldsymbol{\Delta}(\boldsymbol{w})$ as $L_n(\boldsymbol{w}) = \|\hat{\boldsymbol{\Delta}}(\boldsymbol{w}) - \boldsymbol{\Delta}\|^2$ and $R_n(\boldsymbol{w}) = E\left[L_n(\boldsymbol{w})\middle|\boldsymbol{U}\right]$, respectively. Our method of choosing $\boldsymbol{w}$ is based on an approximation of $R_n(\boldsymbol{w})$.

Let $\boldsymbol{U}_t^{(k)}$ and $\boldsymbol{U}_c^{(k)}$, both of dimension $n_2 \times l_k$, be matrices containing the covariate observations within the $n_2$ pairs corresponding to the treatment and control, respectively. Let

$$\begin{aligned} &\tilde{Y}_m = Y_m^t - Y_m^c, && \tilde{\boldsymbol{Y}} = (\tilde{Y}_1, \ldots, \tilde{Y}_{n_2})', \\ &\boldsymbol{P}_t^{(k)} = \boldsymbol{U}_t^{(k)}(\boldsymbol{U}_{ta}^{(k)\prime}\boldsymbol{U}_{ta}^{(k)})^{-1}\boldsymbol{U}_{ta}^{(k)\prime}, && \boldsymbol{P}_{\tilde{c}}^{(k)} = \boldsymbol{U}_t^{(k)}(\boldsymbol{U}_{ca}^{(k)\prime}\boldsymbol{U}_{ca}^{(k)})^{-1}\boldsymbol{U}_{ca}^{(k)\prime}, \\ &\boldsymbol{P}_t(\boldsymbol{w}) = \sum_{k=1}^K w_k \boldsymbol{P}_t^{(k)}, && \boldsymbol{P}_{t0}^{(k)} = \left(\boldsymbol{P}_t^{(k)\prime},\ \boldsymbol{0}_{n_t\times(n_t-n_2)}\right)', \\ &\boldsymbol{P}_{\tilde{c}}(\boldsymbol{w}) = \sum_{k=1}^K w_k \boldsymbol{P}_{\tilde{c}}^{(k)}, && \boldsymbol{P}_{\tilde{c}0}^{(k)} = \left(\boldsymbol{P}_{\tilde{c}}^{(k)\prime},\ \boldsymbol{0}_{n_c\times(n_c-n_2)}\right)', \\ &\boldsymbol{P}_{t0}(\boldsymbol{w}) = \sum_{k=1}^K w_k \boldsymbol{P}_{t0}^{(k)}, && \text{and} \quad \boldsymbol{P}_{\tilde{c}0}(\boldsymbol{w}) = \sum_{k=1}^K w_k \boldsymbol{P}_{\tilde{c}0}^{(k)}. \end{aligned} \tag{12}$$

Now, write $\xi_n = \inf_{\boldsymbol{w}\in H_n} R_n(\boldsymbol{w})$. Theorem 1 shows that

$$C_n(\boldsymbol{w}) = \|\hat{\boldsymbol{\Delta}}(\boldsymbol{w}) - \tilde{\boldsymbol{Y}}\|^2 + 2\sigma_t^2 tr\left(\boldsymbol{P}_{t0}(\boldsymbol{w})\right) + 2\sigma_c^2 tr\left(\boldsymbol{P}_{\tilde{c}0}(\boldsymbol{w})\right) \tag{13}$$

is approximately an unbiased estimator of the expected squared error of $\hat{\Delta}(\boldsymbol{w})$ save for a constant.

**Theorem 1** *Provided that*

$$p^2\left(\frac{n}{\log n}\right)^{1-(2/p)}\xi_n^{-1} = o(1), \quad a.s. \tag{14}$$

*then*

$$E\big(C_n(\boldsymbol{w})|\boldsymbol{U}\big) = R_n(\boldsymbol{w})\big(1+o(1)\big) + n_2(\sigma_t^2+\sigma_c^2). \quad a.s. \tag{15}$$

***Proof*** See Appendix A.1 in supplementary material. □

***Remark 1*** Condition (14) places a constraint on the growth rates of $\xi_n$, $n$ and $p$. Li (1987) used a similar condition. To satisfy Condition (14), it is necessary that $\xi_n \to \infty$. As Hansen (2007) remarked, this simply means there is no finite approximating model for which the bias is zero. We provide an example in Appendix A.4 under which Condition (14) holds.

Our optimal weight vector, $\hat{\boldsymbol{w}}$, is obtained by minimising $C_n(\boldsymbol{w})$, i.e.

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}\in H_n}{\operatorname{argmin}}\, C_n(\boldsymbol{w}). \tag{16}$$

Alternatively, one can express $C_n(\boldsymbol{w})$ as a quadratic function of $\boldsymbol{w}$ by rewriting (13) as $C_n(\boldsymbol{w}) = \boldsymbol{w}'\boldsymbol{\Upsilon}'\boldsymbol{\Upsilon}\boldsymbol{w} + \boldsymbol{w}'\boldsymbol{g}$, where $\boldsymbol{\Upsilon} = (\hat{\boldsymbol{\Delta}}^{(1)} - \tilde{\boldsymbol{Y}}, \ldots, \hat{\boldsymbol{\Delta}}^{(K)} - \tilde{\boldsymbol{Y}})'$ and $\boldsymbol{g} = \Big(2\sigma_t^2 tr\big(\boldsymbol{P}_{t0}^{(1)}\big) + 2\sigma_c^2 tr\big(\boldsymbol{P}_{t0}^{(1)}\big), \ldots, 2\sigma_t^2 tr\big(\boldsymbol{P}_{t0}^{(K)}\big) + 2\sigma_c^2 tr\big(\boldsymbol{P}_{t0}^{(K)}\big)\Big)'$. This is a convenient feature that facilitates the computation of $\hat{\boldsymbol{w}}$ by software packages.

### 3.3 Asymptotic optimality

In this section, we show that our proposed method has an asymptotic optimal property. Now, let $\boldsymbol{f}_t = (f_t(\boldsymbol{u}_1^t), \ldots, f_t(\boldsymbol{u}_{n_2}^t))'$, $\boldsymbol{f}_c = (f_c(\boldsymbol{u}_1^c), \ldots, f_c(\boldsymbol{u}_{n_2}^c))'$, $\hat{\boldsymbol{f}}_t(\boldsymbol{w}) = \boldsymbol{P}_t(\boldsymbol{w})\boldsymbol{Y}_{ta}$, $\hat{\boldsymbol{f}}_c(\boldsymbol{w}) = \boldsymbol{P}_{\tilde{c}}(\boldsymbol{w})\boldsymbol{Y}_{ca}$, $L_{n_t}(\boldsymbol{w}) = \|\hat{\boldsymbol{f}}_t(\boldsymbol{w}) - \boldsymbol{f}_t\|^2$, $R_{n_t}(\boldsymbol{w}) = E\big[L_{n_t}(\boldsymbol{w})\big|\boldsymbol{U}\big]$, $L_{n_{\tilde{c}}}(\boldsymbol{w}) = \|\hat{\boldsymbol{f}}_c(\boldsymbol{w}) - \boldsymbol{f}_c\|^2$, and $R_{n_{\tilde{c}}}(\boldsymbol{w}) = E\big[L_{n_{\tilde{c}}}(\boldsymbol{w})\big|\boldsymbol{U}\big]$.

**Theorem 2** *Assume that* (14) *holds, and for some fixed integer* $1 \le G < \infty$,

$$E(\eta_m^{4G}|\boldsymbol{U}) \le \kappa < \infty, \quad m = 1, \ldots, n_2, \quad a.s. \tag{17}$$

$$K\xi_n^{-2G}\sum_{k=1}^{K}\left[\big(R_{n_t}(\boldsymbol{w}_k^0)\big)^G + \big(R_{n_{\tilde{c}}}(\boldsymbol{w}_k^0)\big)^G\right] \to 0, \quad a.s. \tag{18}$$

$$0 < C_0 \leq \underline{\lambda}\left\{\frac{\boldsymbol{U}_{ca}^{(k)'}\boldsymbol{U}_{ca}^{(k)}}{n_c}\right\}, \quad k = 1, \ldots, K, \quad a.s. \tag{19}$$

*and*

$$\frac{p}{\log n}\left(\frac{\log n}{n}\right)^{2/p} = o(1). \tag{20}$$

*Then*,

$$\frac{L_n(\hat{\boldsymbol{w}})}{\inf\limits_{\boldsymbol{w}\in H_n} L_n(\boldsymbol{w})} \to_p 1, \tag{21}$$

*where* $\underline{\lambda}(\boldsymbol{A})$ *denotes the minimum singular value of* $\boldsymbol{A}$, $C_0$ *and* $\kappa$ *are constants and* $\boldsymbol{w}_0^k$ *is a* $K \times 1$ *vector with the kth element taking on the value of unity and all other elements zeros*.

***Proof*** See Appendix A.2 in supplementary material. □

Theorem 2 shows that the squared error of $\hat{\boldsymbol{\Delta}}$ resulting from using $\hat{\boldsymbol{w}}$ is asymptotically equivalent to that from using the infeasible optimal weight vector, meaning that our proposed method has an optimal asymptotic property.

***Remark 2*** Condition (17) places a bound on the conditional moments. This condition is similar to Condition (7) of Wan et al. (2010). Condition (18) places a restriction on the rates of increase of the minimum risk $\xi_n$ of the model average estimator, the number of candidate models $K$ and $\sum_{k=1}^{K}\left[\left(R_{n_t}(\boldsymbol{w}_k^0)\right)^G + \left(R_{n_{\bar{c}}}(\boldsymbol{w}_k^0)\right)^G\right]$. If we define $\eta_n$ as the maximum of $R_{n_t}(\boldsymbol{w}_k^0)$ and $R_{n_{\bar{c}}}(\boldsymbol{w}_k^0)$ for $k = 1, \ldots, K$, it can be seen that $K^2(\eta_n\xi_n^{-2})^G \to 0$ is a sufficient condition for Condition (18). Now, if $\xi_n \to \infty$ holds, then $K^2(\eta_n\xi_n^{-2})^G \to 0$ also holds as long as $K^2\eta_n^G$ tends to infinity at a rate slower than that of $\xi_n^{2G}$ to infinity. Under the nested model framework proposed by Hansen (2007), Condition (18) is typically implied by the usual condition $K = O(n^v)$ (where $v > 0$ is a small constant), as demonstrated in Wan et al. (2010). Specifically, this holds automatically when $K$ is bounded. Subsection 3.1 of Zhang (2021) provides an extensive discussion of similar conditions. Condition (19) is similar to Condition C.4 of Zhang et al. (2016) and Condition 1 of Lv and Liu (2014). Condition (20) places a constraint on the growth rates of $n$ and $p$, where $p$ is allowed to be of $O((\log n)^{\beta_0})$ for some $0 < \beta_0 < 1$.

Now, we consider the case where the unknown $\sigma_t^2$ and $\sigma_c^2$ are replaced by estimates. Denote the $K^*$th approximating model as the model such that $l_{K^*} = \max\{l_1, l_2, \ldots, l_K\}$. Write

$$\boldsymbol{P}_{ta}^{(k)} = \boldsymbol{U}_{ta}^{(k)}(\boldsymbol{U}_{ta}^{(k)'}\boldsymbol{U}_{ta}^{(k)})^{-1}\boldsymbol{U}_{ta}^{(k)'}, \qquad \boldsymbol{P}_{ca}^{(k)} = \boldsymbol{U}_{ca}^{(k)}(\boldsymbol{U}_{ca}^{(k)'}\boldsymbol{U}_{ca}^{(k)})^{-1}\boldsymbol{U}_{ca}^{(k)'},$$
$$\hat{\boldsymbol{f}}_{ta}^{(k)} = \boldsymbol{P}_{ta}^{(k)}\boldsymbol{Y}_{ta} \qquad \text{and} \quad \hat{\boldsymbol{f}}_{ca}^{(k)} = \boldsymbol{P}_{ca}^{(k)}\boldsymbol{Y}_{ca}.$$

In addition, let

$$\hat{\sigma}_{K_t^*}^2 = (n_t - l_{K^*})^{-1}\|\boldsymbol{Y}_{ta} - \hat{\boldsymbol{f}}_{ta}^{(K^*)}\|^2 \tag{22}$$

and

$$\hat{\sigma}_{K_c^*}^2 = (n_c - l_{K^*})^{-1}\|\boldsymbol{Y}_{ca} - \hat{\boldsymbol{f}}_{ca}^{(K^*)}\|^2, \tag{23}$$

be the estimators of $\sigma_t^2$ and $\sigma_c^2$ , respectively.

To facilitate analysis, we introduce notations analogous to those in Subsect. 3.3. Let $\boldsymbol{f}_{ta} = E(\boldsymbol{Y}_{ta}|\boldsymbol{U})$, $\boldsymbol{f}_{ca} = E(\boldsymbol{Y}_{ca}|\boldsymbol{U})$, $\boldsymbol{P}_{ca}(\boldsymbol{w}) = \sum_{k=1}^{K} w_k \boldsymbol{P}_{ca}^{(k)}$, $\boldsymbol{P}_{ta}(\boldsymbol{w}) = \sum_{k=1}^{K} w_k \boldsymbol{P}_{ta}^{(k)}$, $\hat{\boldsymbol{f}}_{ta}(\boldsymbol{w}) = \boldsymbol{P}_{ta}(\boldsymbol{w})\boldsymbol{Y}_{ta}$, $\hat{\boldsymbol{f}}_{ca}(\boldsymbol{w}) = \boldsymbol{P}_{ca}(\boldsymbol{w})\boldsymbol{Y}_{ca}$, $L_{n_{ca}}(\boldsymbol{w}) = \|\hat{\boldsymbol{f}}_{ca}(\boldsymbol{w}) - \boldsymbol{f}_{ca}\|^2$, $L_{n_{ta}}(\boldsymbol{w}) = \|\hat{\boldsymbol{f}}_{ca}(\boldsymbol{w}) - \boldsymbol{f}_{ta}\|^2$, $R_{n_{ca}}(\boldsymbol{w}) = E[L_{n_{ca}}(\boldsymbol{w})|\boldsymbol{U}]$ and $R_{n_{ta}}(\boldsymbol{w}) = E[L_{n_{ta}}(\boldsymbol{w})|\boldsymbol{U}]$. The following theorem shows that under some mild conditions, Theorem 2 remains valid when $\sigma_t^2$ and $\sigma_c^2$ are replaced by $\hat{\sigma}_{K_t^*}^2$ and $\hat{\sigma}_{K_c^*}^2$ , respectively.

**Theorem 3** *When $\sigma_t^2$ and $\sigma_c^2$ are replaced by $\hat{\sigma}_{K_t^*}^2$ and $\hat{\sigma}_{K_c^*}^2$ , respectively, Theorem 2 remains valid, provided that*

$$\boldsymbol{f}_{ta}'\boldsymbol{f}_{ta}/n_t = O(1), \quad \boldsymbol{f}_{ca}'\boldsymbol{f}_{ca}/n_c = O(1), \quad a.s. \tag{24}$$

$$K\xi_n^{-2G}\sum_{k=1}^{K}\left[\left(R_{n_{ca}}(\boldsymbol{w}_k^0)\right)^G + \left(R_{n_{ta}}(\boldsymbol{w}_k^0)\right)^G\right] \to 0, \quad a.s. \tag{25}$$

*and*

$$l_{K^*}^2/n_1 \le \varphi < \infty, \tag{26}$$

*where $\varphi$ is a constant.*

***Proof*** See Appendix A.3 in supplementary material. □

***Remark 3*** Condition (24) is a mild condition on the averages of $f_{it}$ and $f_{ic}$. Shao (1997) and Wan et al. (2010) assumed similar conditions. Condition (25) is analogous to Condition (18). Condition (26) places a constraint on the number of regressors in the largest approximating model. Similar assumptions have been made by Shibata (1980) and Newey (1997).

## 4 A simulation study

We label our proposed model averaging strategy as the OPT method to indicate that it possesses an optimal asymptotic property. In this section, we compare the finite sample performance of OPT with several other estimators, including the AIC and BIC model selection estimators and smoothed model averaging counterparts, denoted as the smoothed-AIC (SAIC) and smoothed-BIC (SBIC) estimators. Let $\hat{\sigma}_k^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \hat{f}_{T_i}^{(k)})^2$ be the estimator of $\sigma^2$ in the $k$th model. The AIC and BIC scores for the $k$th model are $\mathrm{AIC}_k = n_2 \log \hat{\sigma}_k^2 + 2l_k$ and $\mathrm{BIC}_k = n_2 \log \hat{\sigma}_k^2 + (\log n_2) l_k$ , respectively. The AIC and BIC model selection estimators select the model with the smallest AIC and BIC scores, respectively. The SAIC estimator introduced by Buckland (1997) is a model average estimator that assigns the weight

$$w_k = \exp\left(-\frac{1}{2}\mathrm{AIC}_k\right) / \sum_{k=1}^K \exp\left(-\frac{1}{2}\mathrm{AIC}_k\right)$$

to the $k$th model. The SBIC estimator is defined analogously. Our analysis also includes a comparison with the TECV model selection and the TEEM model averaging methods. For a detailed discussion of these methods, readers may refer to Rolling and Yang (2014) and Rolling et al. (2019).

Our simulation study is based on the model setup in (1). We set $n = 100, 400, 800, p = 2$, $P(T_i = t) = 0.5$ and $(u_{1i}, \ldots, u_{pi})' \sim N_p(\mathbf{0}, \mathbf{\Sigma})$, where the $kj$th element of $\mathbf{\Sigma}$ is set to $\rho^{|k-j|}$; $k, j \in \{1, \ldots, p\}$, $\rho = 0, 0.5, 0.8$. The larger the value of $\rho$, the stronger the correlation among the covariates. Define $R^2 = \mathrm{var}\left(I(T_i = t) \times f_t(\boldsymbol{u}_i) + I(T_i = c) \times f_c(\boldsymbol{u}_i)\right) / \mathrm{var}(Y_i)$. Consider $T_i \in \{t, c\}$ and denote $h_1(\boldsymbol{u}_i) = f_t(\boldsymbol{u}_i)/c_r$, $h_0(\boldsymbol{u}_i) = f_c(\boldsymbol{u}_i)/c_r$ and $e_i = I(T_i = t) \times \zeta_i + I(T_i = c) \times v_i$. Then, we can write $R^2 = \mathrm{var}\left(c_r \times h_{T_i}(\boldsymbol{u}_i)\right) / \left(\mathrm{var}\left(c_r \times h_{T_i}(\boldsymbol{u}_i) + \mathrm{var}\,(e_i)\right)\right)$. In our simulations, we vary $c_r$ to match the values of $R^2 = 0.2, 0.3, \ldots, 0.8$. A small (large) $R^2$ indicates that the true model has a high (small) noise level. The simulation setting of Rolling et al. (2019) corresponds to our setting with $R^2$ being close to 0.5. We assume the following data generating process of $Y_i$:

$$Y_i = [0.5u_{i1}^2 + 0.5u_{i2} + 0.5u_{i1} + 0.5u_{i2}^2 + \zeta_i] I(T_i = t) + [0.5u_{i1}^2 + 0.5u_{i2} + v_i] I(T_i = c),$$

where the errors are generated from one of the following two designs:

**Table 1** Candidate models for Designs 1-2

| Number | Candidate Model |
|---|---|
| 1 | $Y_i = [\beta_0 + \zeta_i] I(T_i = t) + [\gamma_0 + v_i] I(T_i = c)$ |
| 2 | $Y_i = [\beta_0 + \beta_1 u_{i1} + \zeta_i] I(T_i = t) + [\gamma_0 + \gamma_1 u_{i1} + v_i] I(T_i = c)$ |
| 3 | $Y_i = [\beta_0 + \beta_2 u_{i2} + \zeta_i] I(T_i = t) + [\gamma_0 + \gamma_2 u_{i2} + v_i] I(T_i = c)$ |
| 4 | $Y_i = [\beta_0 + \beta_1 u_{i1} + \beta_2 u_{i2} + \zeta_i] I(T_i = t) + [\gamma_0 + \gamma_1 u_{i1} + \gamma_2 u_{i2} + v_i] I(T_i = c)$ |

**Design 1:** $v_i \sim N(0,1)$ and $\zeta_i \sim N(0,1)$;

**Design 2:** $v_i \sim N(0,1)$ and $\zeta_i$'s are generated from a double-exponential (Laplace) distribution with variance equal to one.

The purpose of Design 2 is to examine the robustness of our method when the treatment and control groups have different error distributions.

We consider four candidates models given in Table 1. We gauge the estimators' performance by the average squared errors $\text{ASE}_j = \frac{1}{N_{new}} \sum_{i=1}^{N_{new}} (\hat{\Delta}_n(\boldsymbol{u}_i) - \Delta(\boldsymbol{u}_i))^2$, where $N_{new} = 10^6$ is the size of the evaluation sample and $\hat{\Delta}_n(\boldsymbol{u}_i)$ is an estimator of $\Delta(\boldsymbol{u}_i)$ based on a training sample with $n$ observations. We set the evaluation sample size to $10^6$ following the study of Rolling et al. (2019). Since the evaluation set here is a random sample of simulated predictor variables, setting $N_{new} = 10^6$ is easily achievable. Notably, the reliability of the results increases with the size of the evaluation sample. Each part of our simulation experiment is based on $J = 1000$ replications. We define the risk of the estimator as

$$\text{Risk} = \frac{1}{J} \sum_{j=1}^{J} \text{ASE}_j.$$

When reporting the results, we normalise the risk by dividing it by the risk of the infeasible (or best fitting) optimal least squares estimator in each replication.

***Remark 4*** When conducting our simulations, we originally adopted the partitioning approach described in Subsect. 3.1 with $h = \left\lfloor \left( \frac{n_1}{2 \log n_1} \right)^{1/p} \right\rfloor^{-1}$. The results show that the OPT method generally outperforms other methods. However, in order to draw a fair comparison with the method of Rolling et al. (2019), we present the simulation results based on a nearest-neighbour pairing approach that has the advantages of simplicity and producing a larger number of pairs than the partitioning and matching approach, especially when the sample size is small or $p$ is not very small. In the context of estimating the average treatment effect, Abadie and Imbens (2006) showed that matching with replacement via nearest neighbour produces better matching quality than matching without replacement via partitioning. Rolling et al. (2019) took a similar view. Remark 5 provides a brief description of the nearest-neighbour pairing approach. As the two approaches are similar, we do not undertake theoretical analysis based on the nearest-neighbour pairing approach here.

***Remark 5*** With nearest-neighbour pairing, the covariates are scaled so that each covariate has a common mean and variance. Let us denote the vector of standardised covariates as $\tilde{\boldsymbol{u}}$. For a given observation in a treatment group, the goal of nearest-neighbour pairing is to find the observation's 'nearest neighbour' from the other group. Here, the distances between any two neighbours are calculated by standardised covariates. Specifically, for each $i$ in a treatment group, the method seeks $i^*$ from the other group such that

$$i^* = \text{argmin}_{1 \le i' \le n} d(\tilde{\boldsymbol{u}}_i, \tilde{\boldsymbol{u}}_{i'}) \text{ subject to } T_i \neq T_{i'}$$

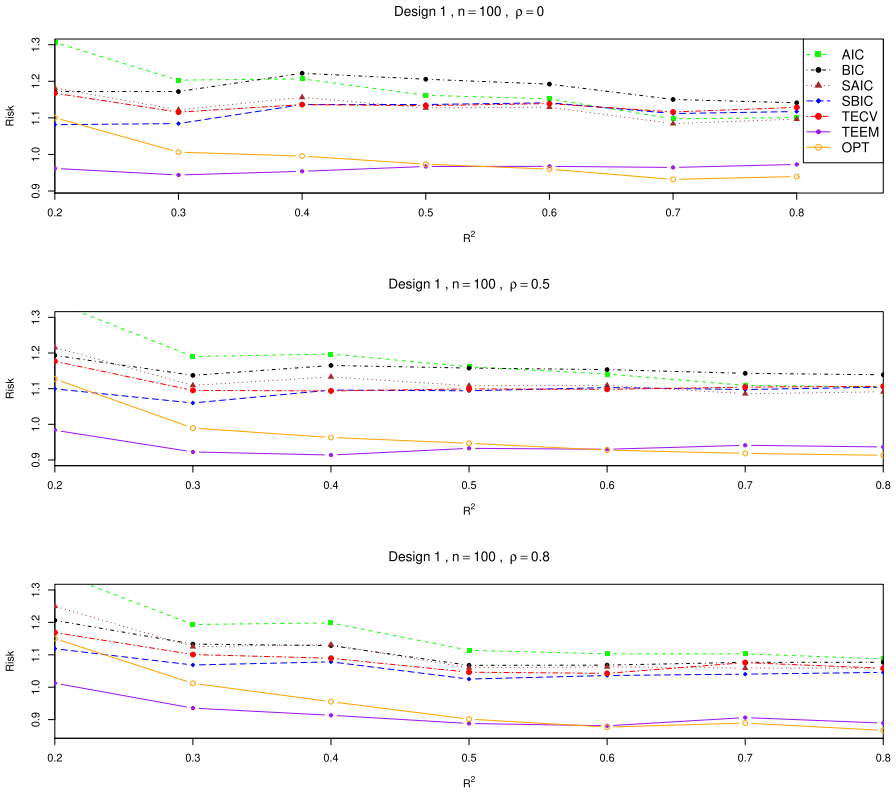


**Fig. 1** Results of simulations under Design 1 with $n = 100$

and hence $\tilde{Y}_i = Y_i^t - Y_{i^*}^c$, where $d(\cdot)$ represents the Euclidean distance.

The results under Design 1 are reported in Figs. 1, 2 and 3. The results suggest that OPT is the strategy that most frequently achieves the best risk outcomes. Without exception, the OPT estimator has the smallest risk when $R^2$ is moderate to large. In general, the bigger the sample size, the larger the range of $R^2$ values for which the OPT estimator yields the smallest risk, *ceteris paribus*. When $n$ is small, the value of $\rho$ makes little difference to the performance of the OPT estimator. However, when $n$ is large, a decrease in $\rho$ commonly has the effect of enlarging the range of values of $R^2$ where the OPT estimator delivers the best outcomes. The TEEM method also enjoys good risk properties. It frequently produces the next best estimates when the OPT yields the smallest risk, and can sometimes outperform the OPT estimator when $R^2$ is small. Interestingly, while an increase in sample size generally benefits the relative performance of the OPT estimator, it frequently worsens the performance of the TEEM estimator. The SAIC, SBIC, AIC, BIC and TECV estimators never result in the best risk outcomes. When both

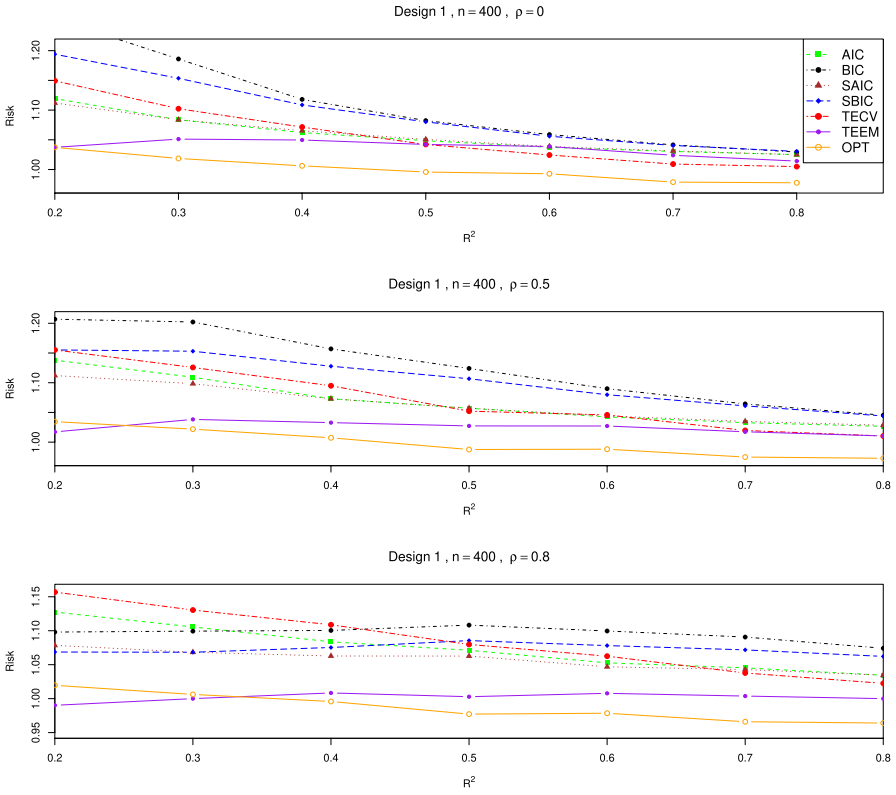


**Fig. 2** Results of simulations under Design 1 with $n = 400$

$n$ and $R^2$ are large, the TECV estimator can yield better estimates than the TEEM estimator but it cannot outperform the OPT estimator. The AIC estimator nearly always delivers the worst estimates when $n$ and $R^2$ are small, but its performance, as well as that of the BIC estimator, often improves as the sample size increases. In general, the SAIC and SBIC estimators enjoy better sampling properties than their model selection counterparts; exceptions occur, for example, at small values of $R^2$, where the AIC and BIC estimators can sometimes be the preferred strategies. The results under Design 2 are reported in Figs. 6–8 in Appendix A.5 of supplementary material. Generally speaking, the comments made above under Design 1 also apply to Design 2, where the error distributions are different for the treatment and control groups. In particular, the OPT estimator remains the best estimator most of the time.

We also evaluate the performance of our method when the probability $\mathrm{P}(T_i = t)$ is related to the covariates. We consider the following design:

**Design 3:** $n = 500$, $p = 10$, $\rho = 0.7$, and

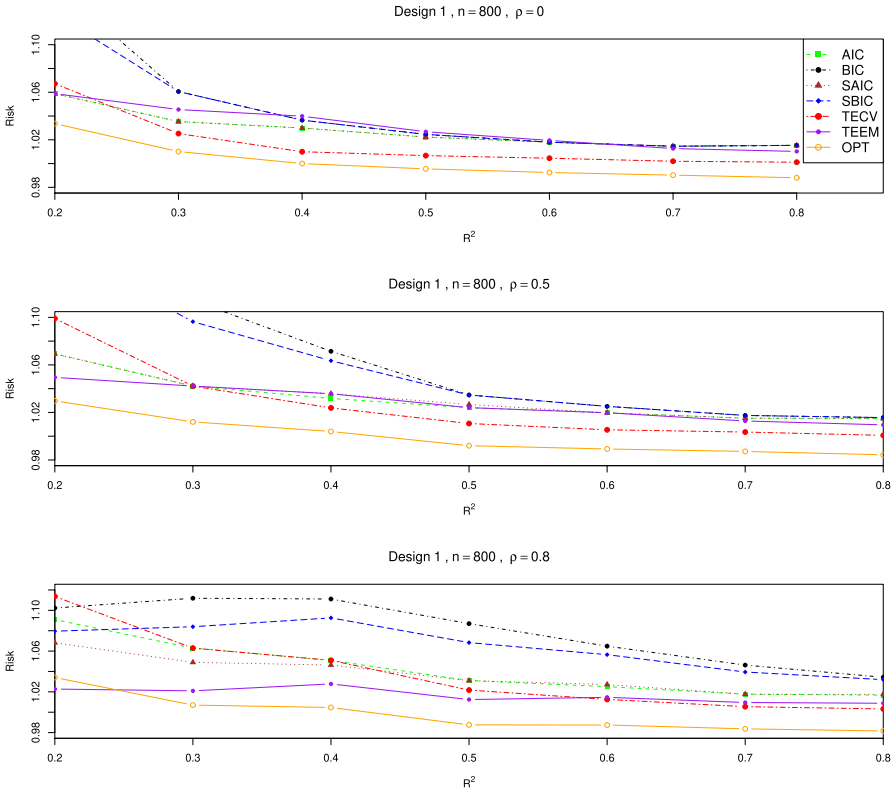


**Fig. 3** Results of simulations under Design 1 with $n = 800$

**Table 2** Candidate models for Design 3

| Number | Candidate model |
|---|---|
| 1 | $Y_i = [\beta_0 + \zeta_i]I(T_i = t) + [\gamma_0 + v_i]I(T_i = c)$ |
| 2 | $Y_i = [\beta_0 + \beta_1 u_{i1} + \zeta_i]I(T_i = t) + [\gamma_0 + \gamma_1 u_{i1} + v_i]I(T_i = c)$ |
| 3 | $Y_i = [\beta_0 + \sum_{j=1}^{2} \beta_j u_{ij} + \zeta_i]I(T_i = t) + [\gamma_0 + \sum_{j=1}^{2} \gamma_j u_{ij} + v_i]I(T_i = c)$ |
| 4 | $Y_i = [\beta_0 + \sum_{j=1}^{3} \beta_j u_{ij} + \zeta_i]I(T_i = t) + [\gamma_0 + \sum_{j=1}^{3} \gamma_j u_{ij} + v_i]I(T_i = c)$ |
| ⋮ | ⋮ |
| 11 | $Y_i = [\beta_0 + \sum_{j=1}^{10} \beta_j u_{ij} + \zeta_i]I(T_i = t) + [\gamma_0 + \sum_{j=1}^{10} \gamma_j u_{ij} + v_i]I(T_i = c)$ |

$$\mathrm{P}(T_i = t) = \frac{\exp \frac{1}{10}\sum_{j=1}^{10} u_{ij}}{1 + \exp \frac{1}{10}\sum_{j=1}^{10} u_{ij}},$$

under the following data generating process of $Y_i$:

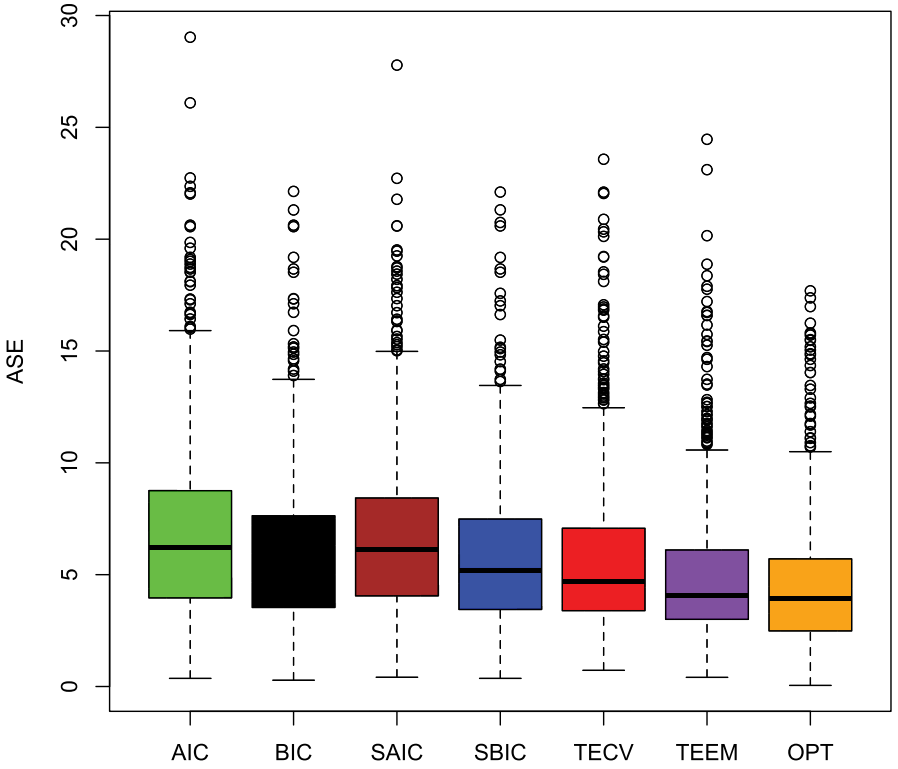


**Fig. 4** Boxplots of average squared errors under Design 3

$$Y_i = [3\sum_{j=1}^{3} u_{ij} + 2\sum_{j=4}^{6} u_{ij} + \zeta_i]I(T_i = t) + [2\sum_{j=1}^{6} u_{ij} + v_i]I(T_i = c),$$

where $v_i \sim N(0, 10)$ and $\zeta_i \sim N(0, 10)$. Under this design, we consider eleven nested candidates models given in Table 2. We use boxplots to show the results of ASEs based on $J = 1000$ replications for Design 3. The results under Design 3 are reported in Fig. 4. The results suggest that when the probability $\mathrm{P}(T_i = t)$ depends on the covariates, the OPT estimator is still the superior strategy.

## 5 An empirical example

This application is based on the same data example considered by Rolling and Yang (2014). The data are obtained from the Community Programs for Clinical Research on AIDS that contain results of a clinical trial known as FIRST that evaluated treatment strategies for patients who were HIV positive. These patients were then assigned to one of the two treatment strategies described in MacArthur et al. (2006) and Rolling and Yang (2014). The purpose of the study is to compare the difference in the change in CD4-cell counts under the two strategies.

We use the same $n = 1191$ observations as in Rolling and Yang (2014), of which 799 observations correspond to patients assigned to the less potent of the two strategies and the remaining 392 were assigned to the more potent strategy. Our response and baseline covariates are also the same as in Rolling and Yang (2014); namely, we use the difference between the square root of the patient's average CD4-cell count over all measurements taken at or after 32 months from enrolment and the square root of the patient's CD4-cell count at the baseline enrolment date as the response variable, and the square root of the baseline CD4-cell count ($CD4_0$), the logarithm of the baseline HIV ribonucleic acid concentration ($RNA_0$), the patient's age (Age), and the treatment variable $T$ as covariates. We assign $T = t$ to the more potent strategy and $T = c$ to the less potent strategy. For any given model, we consider it mandatory to include the intercept and $CD4_0$, and treat all other covariates as optional. In addition, when a covariate is included, the pairwise interaction of that covariate with every of the other covariates is also included in the same model. This results in 22 candidate models, each containing a distinct set of covariates. Table 3 shows the covariates of these models.

**Table 3** Candidate models in empirical example

| Model Index | Covariates |
|---|---|
| 1 | $CD4_0$ |
| 2 | Age, $CD4_0$ |
| 3 | $RNA_0$, $CD4_0$ |
| 4 | Age, $RNA_0$, $CD4_0$ |
| 5 | $T$, $CD4_0$ |
| 6 | $T$, $CD4_0$, T*$CD4_0$ |
| 7 | $T$, $CD4_0$, Age |
| 8 | $T$, $CD4_0$, $RNA_0$ |
| 9 | $T$, $CD4_0$, Age, T*$CD4_0$ |
| 10 | $T$, $CD4_0$, Age, $RNA_0$ |
| 11 | $T$, $CD4_0$, T*$CD4_0$, $RNA_0$ |
| 12 | $T$, $CD4_0$, Age, T*Age |
| 13 | $T$, $CD4_0$, $RNA_0$, T*$RNA_0$ |
| 14 | $T$, $CD4_0$, Age, T*Age, T*$CD4_0$ |
| 15 | $T$, $CD4_0$, $RNA_0$, T*$RNA_0$, T*$CD4_0$ |
| 16 | $T$, $CD4_0$, $RNA_0$, Age, T*$CD4_0$ |
| 17 | $T$, $CD4_0$, $RNA_0$, Age, T*Age |
| 18 | $T$, $CD4_0$, $RNA_0$, Age, T*$RNA_0$ |
| 19 | $T$, $CD4_0$, Age, T*Age, T*$CD4_0$, $RNA_0$ |
| 20 | $T$, $CD4_0$, Age, T*$RNA_0$, T*$CD4_0$, $RNA_0$ |
| 21 | $T$, $CD4_0$, Age, T*$RNA_0$, T*Age, $RNA_0$ |
| 22 | $T$, $CD4_0$, Age, T*$RNA_0$, T*Age, T*$CD4_0$, $RNA_0$ |

[a] All models contain an intercept term

Following Rolling et al. (2019), we evaluate the various methods by a 'guided simulation' experiment. We assume three 'true' processes based on three estimated regression functions, indexed as Models 14, 2 and 19 in Table 3. Models 2 and 19 are the models with the smallest AIC and BIC, respectively, while Model 14 is selected by a focused information criterion (see Rolling and Yang (2014) for details). For each regression function, we generate 1191 observations of $Y$ by augmenting the estimated regression with an i.i.d. noise variable obtained from a zero-mean Gaussian distribution with variance equal to the estimated error variance estimate of that regression, taking into account the sample values of $T_i$ and $\boldsymbol{u}_i$. We apply the same seven model selection and averaging methods considered in Sect. 4. Each of these methods produces, for each process and $i$, an estimate of $\Delta(\boldsymbol{u}_i)$. For a given method, the performance of the method is evaluated by the average of the squared errors $(\Delta(\boldsymbol{u}_i) - \hat{\Delta}(\boldsymbol{u}_i))^2$ across the $n = 1191$ observations. Note that in each case the 'true' $\Delta(\boldsymbol{u}_i)$ can be obtained from the assumed true process. We repeat this experiment 100 times, and different realisations of the noises are generated for each trial. As we consider three different processes of $Y$, this results in 300 average squared errors for each method.

Figure 5, which displays the boxplots of the average squared errors, shows that the OPT method generally results in the best estimates, followed by the TEEM and SAIC methods that usually outperform all other methods including the SBIC method, whose performance is on par with that of its model selection counterpart and the TECV method but generally worse than that of the AIC selection method.

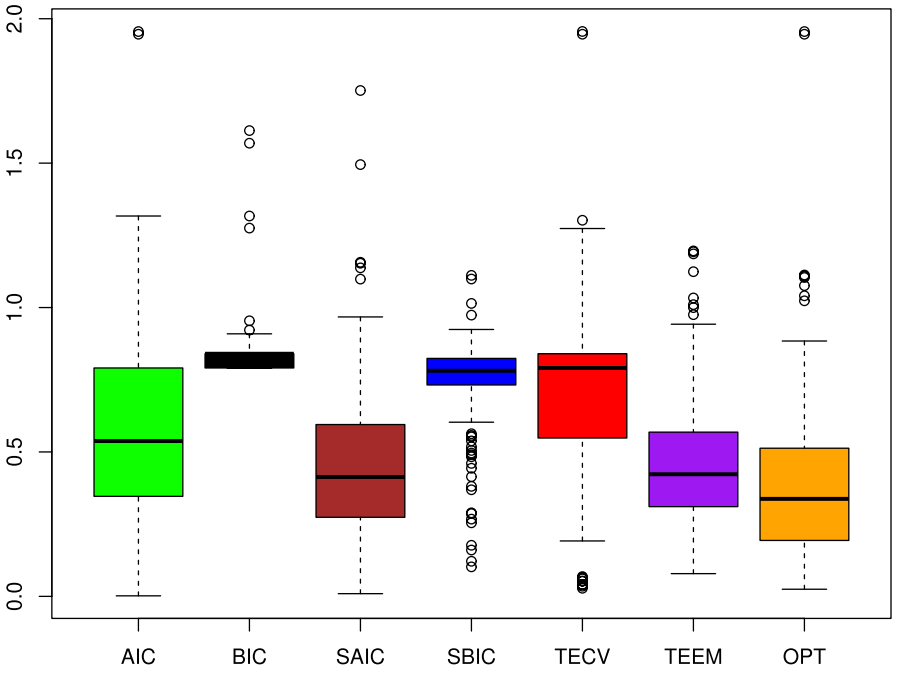


**Fig. 5** Boxplots of average squared errors in empirical example

## 6 Discussion

In this paper, we have developed a model averaging strategy for the estimation of treatment effects based on a minimisation of the approximate risk of the model averaging estimator under a squared error loss function. It is shown that the proposed strategy has an optimal asymptotic property, and fares well compared with other model selection and averaging strategies in finite samples. The current work can be extended in various ways. First, Condition (20) imposes a stringent constraint on the number of model parameters. It would be worthwhile to consider weakening this constraint in the theoretical analysis. Also, our analysis is limited to the case of binary, or yes-or-no, treatment. There is a need to consider the more complex type of multiple treatments, which commonly arise in health and medical research. See Cattaneo (2010), Feng et al. (2012), Linden et al. (2016), among others. An investigation of the distributional properties of the proposed model average estimators is another fruitful avenue for further research. The results would enable us to examine the implications of model averaging on interval, rather than just point, estimation. Recent works by Liu (2015) and Zhang and Liu (2019) may serve as a useful guide in this regard. This paper maintains homoskedasticity within each group. A natural question is how to extend our procedure to allow heteroskedasticity. Finally, we have consider parametric candidate models; model averaging criteria that incorporate nonparametric models for treatment effect estimation are yet to be developed.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s10463-023-00876-4.

**Acknowledgements** The authors are grateful to Drs. Craig Rolling and Yuhong Yang for providing their codes for the computation of the TECV estimates and to Dr. Yuhong Yang for several helpful discussions. This paper has benefitted from the suggestions and comments of two referees. All remaining errors are ours.

**Data availability statement** The data that support the findings of this study are available from the corresponding author upon reasonable request.

## References

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

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

Ashenfelter, O. (1978). Estimating the effect of training programs on earnings. *The Review of Economics and Statistics, 60*(1), 47–57.

Ashenfelter, O., Card, D. (1984). Using the longitudinal structure of earnings to estimate the effect of training programs. *Review of Economics and Statistics, 67*(4), 648–660.

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

Belloni, A., Hansen, C. (2014). Inference on treatment effects after selection amongst high-dimensional controls. *The Review of Economic Studies, 81*(2), 608–650.

Buckland, S. T. (1997). Model selection: An integral part of inference. *Biometrics, 53*(2), 603–618.

Cattaneo, M. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics, 155*(2), 138–154.

Feng, P., Zhou, X., Zou, Q., Fan, M., Li, X. (2012). Generalized propensity score for estimating the average treatment effect of multiple treatments. *Statistics in Medicine, 31*(7), 681–697.
Hansen, B. E. (2007). Least squares model averaging. *Econometrica, 75*(4), 1175–1189.
Hansen, B. E., Racine, J. S. (2012). Jackknife model averaging. *Journal of Econometrics, 167*(1), 38–46.
Hjort, N. L., Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association, 98*(1), 879–899.
Hoeting, J. A., Madigan, D., Raftery, A. E., Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science, 14*(4), 382–417.
Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association, 81*(396), 945–960.
Imbens, G. W., Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature, 47*(1), 5–86.
Kitagawa, T., Muris, C. (2016). Model averaging in semiparametric estimation of treatment effects. *Journal of Econometrics, 105*(4), 358–368.
Kuersteiner, G., Okui, R. (2010). Constructing optimal instruments by first-stage prediction averaging. *Econometrica, 78*(2), 697–718.
LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review, 76*(4), 604–620.
Lee, S., Okui, R., Whang, Y. J. (2017). Doubly robust uniform confidence band for the conditional average treatment effect function. *Journal of Applied Econometrics, 32*(7), 1207–1225.
Lee, S., Shin, Y. (2021). Complete subset averaging with many instruments. *The Econometrics Journal, 24*(2), 290–314.
Li, K.-C. (1987). Asymptotic optimality for $C_p$, $C_l$, cross-validation and generalized cross-validation: Discrete index set. *Annals of Statistics, 15*(3), 958–975.
Liang, H., Zou, G., Wan, A. T., Zhang, X. (2011). Optimal weight choice for frequentist model average estimators. *Journal of the American Statistical Association, 106*(495), 1053–1066.
Linden, A., Uysal, S., Ryan, A., Adams, J. (2016). Estimating causal effects for multivalued treatments: A comparison of approaches. *Statistics in Medicine, 35*(4), 534–552.
Liu, C.-A. (2015). Distribution theory of the least squares averaging estimator. *Journal of Econometrics, 186*(1), 142–159.
Lv, J., Liu, J. S. (2014). Model selection principles in misspecified models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76*(1), 141–167.
MacArthur, R. D., Novak, R. M., Peng, G., Chen, L., Xiang, Y., Hullsiek, K. H. (2006). A comparison of three highly active antiretroviral treatment strategies consisting of non-nucleoside reverse transcriptase inhibitors, protease inhibitors, or both in the presence of nucleoside reverse transcriptase inhibitors as initial therapy (CPCRA 058 FIRST Study): A long-term randomised trial. *Lancet, 368*(9553), 2125–2135.
Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics, 79*(1), 147–168.
Rolling, C. A., Yang, Y. (2014). Model selection for estimating treatment effects. *Journal of the Royal Statistical Society (Series B), 76*(4), 749–769.
Rolling, C. A., Yang, Y., Velez, D. (2019). Combining estimates of conditional treatment effects. *Econometric Theory, 35*(6), 1089–1110.
Seng, L., Li, J. (2022). Structural equation model averaging: Methodology and application. *Journal of Business and Economic Statistics, 40*(2), 815–828.
Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica, 7*(2), 221–242.
Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Annals of Statistics, 8*(1), 147–164.
Wan, A. T., Zhang, X., Zou, G. (2010). Least squares model averaging by Mallows criterion. *Journal of Econometrics, 156*(2), 277–283.
Yang, Y. (2001). Adaptive regression by mixing. *Journal of the American Statistical Association, 96*(454), 574–588.
Yuan, Z., Yang, Y. (2005). Combining linear regression models: When and how? *Journal of the American Statistical Association, 100*(472), 1202–1214.
Zhang, X. (2021). A new study on asymptotic optimality of least squares model averaging. *Econometric Theory, 37*(2), 388–407.
Zhang, X., Liu, C.-A. (2019). Inference after model averaging in linear regression models. *Econometric Theory, 35*(4), 816–841.

Zhang, X., Zou, G., Carroll, R. J. (2015). Model averaging based on Kullback-Leibler distance. *Statistica Sinica, 25*(4), 1583–1598.

Zhang, X., Yu, D., Zou, G., Liang, H. (2016). Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models. *Journal of the American Statistical Association, 111*(516), 1775–1790.

Zhang, X., Zou, G., Liang, H., Carroll, R. J. (2020). Parsimonious model averaging with a diverging number of parameters. *Journal of the American Statistical Association, 115*(530), 972–984.

Zhu, R., Wan, A. T., Zhang, X., Zou, G. (2019). A Mallows-type model averaging estimator for the varying-coefficient partially linear model. *Journal of the American Statistical Association, 114*(526), 882–892.



## Authors and Affiliations

**Zhihao Zhao[1] · Xinyu Zhang[2] · Guohua Zou[3] · Alan T. K. Wan[4] · Geoffrey K. F. Tso[4]**

✉ Xinyu Zhang
xinyu@amss.ac.cn

[1] School of Statistics, Capital University of Economics and Business, 121 Zhangjialukou, Huaxiang Fengtai District, Beijing 100070, China

[2] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, 55 Zhongguancun East Road, Haidian District, Beijing 100190, China

[3] School of Mathematical Sciences, Capital Normal University, 105 West Third Ring Road North, Haidian District, Beijing 100048, China

[4] Department of Management Sciences, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon Tong, Kowloon, Hong Kong, China