Zulj, Valentin; Jin, Shaobo

**Working Paper**

# Can model averaging improve propensity score based estimation of average treatment effects?

Working Paper, No. 2024:1

**Provided in Cooperation with:**
IFAU - Institute for Evaluation of Labour Market and Education Policy, Uppsala

*Suggested Citation:* Zulj, Valentin; Jin, Shaobo (2024) : Can model averaging improve propensity score based estimation of average treatment effects?, Working Paper, No. 2024:1, Institute for Evaluation of Labour Market and Education Policy (IFAU), Uppsala

This Version is available at:
https://hdl.handle.net/10419/297024

WORKING PAPER 2024:1

# Can model averaging improve propensity score based estimation of average treatment effects?

Valentin Zulj
Shaobo Jin

IFAU
INSTITUTE FOR EVALUATION OF LABOUR MARKET AND EDUCATION POLICY

The Institute for Evaluation of Labour Market and Education Policy (IFAU) is a research institute under the Swedish Ministry of Employment, situated in Uppsala.

IFAU's objective is to promote, support and carry out scientific evaluations. The assignment includes: the effects of labour market and educational policies, studies of the functioning of the labour market and the labour market effects of social insurance policies. IFAU shall also disseminate its results so that they become accessible to different interested parties in Sweden and abroad.

Papers published in the Working Paper Series should, according to the IFAU policy, have been discussed at seminars held at IFAU and at least one other academic forum, and have been read by one external and one internal referee. They need not, however, have undergone the standard scrutiny for publication in a scientific journal. The purpose of the Working Paper Series is to provide a factual basis for public policy and the public policy discussion.

More information about IFAU and the institute's publications can be found on the website www.ifau.se

ISSN 1651-1166

# Can model averaging improve propensity score based estimation of average treatment effects?*

Valentin Zulj† and Shaobo Jin‡

**Abstract**

When drawing causal inferences from observational data, researchers often model the propensity score. To date, the literature on the estimation of propensity scores is vast, and includes covariate selection algorithms as well as super learners and model averaging procedures. The latter often tune the estimated scores to be *either* very accurate *or* to provide the best possible result in terms of covariate balance. This paper focuses on using inverse probability weighting to estimate average treatment effects, and makes the assertion that the context requires *both* accuracy *and* balance to yield suitable propensity scores. Using Monte Carlo simulation, the paper studies whether frequentist model averaging can be used to simultaneously account for both balance and accuracy in order to reduce the bias of estimated treatment effects. The candidate propensity scores are estimated using reproducing kernel Hilbert space regression, and the simulation results suggest that model averaging does not improve the performance of the individual estimators.

*We extend our gratitude to IFAU reviewers Ingeborg Waernbaum and Jenny Häggström, your ideas and comments have been very valuable and much appreciated. We also thank everyone who participated in our seminars given at IFAU and the Department of Statistics at Uppsala University.

†Department of Statistics, Uppsala University. valentin.zulj@statistik.uu.se.

‡Department of Mathematics and Department of Statistics, Uppsala University. shaobo.jin@math.uu.se

# Contents

# 1 Introduction

Estimation of causal effects is central to many fields of study. For example, it is common for researchers in economics, medicine, and epidemiology to face questions regarding the effect of exposure to some intervention or treatment (see e.g. Chang et al., 2017; Mörk et al., 2020; Yu et al., 2021, for examples). In many cases as such, these questions arise in observational studies, where the researchers do not have control over treatment assignment or background variables, meaning that confounding factors, if not dealt with properly, can severely skew results. As a way of accounting for confounders, estimation procedures based on the propensity score have been developed. Such procedures use the propensity score to even out differences between treatment groups, and are the main focus of the current study.

The propensity score, $e(\boldsymbol{x})$, is defined as the probability of being assigned to treatment, conditional on a vector of observed covariates $\boldsymbol{x}$. Rosenbaum and Rubin (1983) prove that the propensity score is a *balancing score*, meaning that, conditional on $e(\boldsymbol{x})$, the joint distribution of the pre-treatment variables is the same for both treated and untreated individuals. Thus, when the true propensity score is known, accounting for confounders is a simple task, and the average treatment effect can be estimated without bias. However, in observational studies the true propensity score is typically unknown, and thus needs to be estimated.

In a simulation study, Kang and Schafer (2007) show that a slightly misspecified propensity score model can result in large bias when estimating the average treatment effect (ATE). Since the true form of $e(\boldsymbol{x})$ is seldom known in observational studies, flexible methods that do not require model specification can be employed, attempting to find an estimator $\hat{e}(\boldsymbol{x})$ that predicts $e(\boldsymbol{x})$ as accurately as possible. Examples of such methods are random forests, boosting, and neural networks (see e.g., Lee et al., 2010; Cannas and Arpino, 2019). If $\hat{e}(\boldsymbol{x})$ predicts $e(\boldsymbol{x})$ perfectly, the ATE can be estimated without bias even though strict assumptions have not been made regarding the form of the propensity score. However, theory shows that accuracy is sufficient but not necessary, while balanced confounders is necessary, for unbiased estimation of the ATE. As such, unbiasedness is achievable even though the estimated propensity scores are not accurate, so long as care is taken to balance confounders. Thus, an alternative to the prediction based approaches is to use an estimator $\hat{e}(\boldsymbol{x})$ that balances treatment groups well empirically. To that end, Imai and Ratkovic (2014) and Ning et al. (2020) use moment restrictions when estimating the covariate balancing propensity score (CBPS), showing, as well, that the CBPS can contribute to reducing the bias of the estimated ATE. Zhao (2019) generalizes the CBPS to consider tailored loss functions.

In practice, perfect accuracy of $\hat{e}(\boldsymbol{x})$ is unlikely. Moreover, the set of true confounders is unknown in most cases, and confounding cannot be tested for, meaning that there is no guarantee that a balancing propensity score balances all confounders. Both of these issues are known to cause bias when estimating the ATE, and due to the uncertainties regarding accuracy and balancing ability, a third alternative is to compromise and use a propensity score estimator that takes both aspects into account. Here, the aim is for $\hat{e}(\boldsymbol{x})$ to have good enough balancing properties to properly balance the covariates used in estimation, and to mimic the true propensity score closely enough to moderate the bias generated by any unbalanced form of the confounders. One possible way of achieving this is to combine different estimators of the propensity score.

In the last decade, model compromise approaches, in the sequel referred to as *model averages*, have been used to tune prediction procedures for specific purposes. Following the seminal works of Hjort and Claeskens (2003) and Hansen (2007), frequentist model averaging (FMA) has been shown, both empirically and theoretically, to be a good alternative when tuning probability estimators for accuracy (see e.g., Zhao et al., 2019; Wan et al., 2014). In causal inference, this idea is put into practice by Pirracchio et al. (2015) and Wyss et al. (2018), who tune the combined propensity score estimator to achieve good accuracy. Further examples include Ju et al. (2019), who combine a model average with the high-dimensional propensity score variable selection algorithm, Li et al. (2016), who use a model average to estimate the effect of treatment on censored cost, and Autenrieth et al. (2021), who use stacked ensemble learning to estimate propensity scores. Working in the $\sqrt{n}$ local asymptotic framework, Lu (2015) develops a covariate selection criterion and suggests using it to find model averaging weights, while Kitagawa and Muris (2016) propose an averaging strategy that minimizes the estimated MSE of the estimated average treatment effect of the treated. Lastly, Pirracchio and Carone (2018) and Xie et al. (2019) deviate from prediction accuracy, and propose using model averaging to tune for good balancing properties.

In addition to providing an opportunity to tune the propensity score estimates, model averages to some extent allow for the circumvention of model selection uncertainty. As an example, consider Zhao (2019), who develops a covariate balancing scoring rule (CBSR), used in combination

with penalized reproducing kernel Hilbert space (RKHS) regression to compute propensity score estimates focused on covariate balancing. However, the degree of penalization, regulated by the tuning parameter $\lambda$, has to be specified, often without an obvious or motivated choice. Hence, using a weighted average of estimates $\hat{e}(\boldsymbol{x}; \lambda_1), \ldots, \hat{e}(\boldsymbol{x}; \lambda_M)$, is a suitable way of acknowledging the uncertainty regarding the best value of $\lambda$. Averaging tuning parameters has proved fruitful in both linear (Zhao et al., 2020; Schomaker, 2012) and generalized linear (Zulj and Jin, 2021) regression modeling.

The purpose of this paper is to study the use of model averages in propensity score estimation, and investigate whether averaging procedures can be used to make estimation of average treatment effects less biased. In particular, the paper proposes an average estimator given by $\sum_m \hat{w}_m \hat{e}(\boldsymbol{x}; \lambda_m)$, where the candidate propensity scores are estimated using the methodology developed by Zhao (2019), and the weights $w_m$ are estimated in an optimization procedure where the prediction error is minimized, subject to restrictions preventing the balancing abilities from being compromised. The proposal is evaluated in Monte Carlo simulation studies, where it is compared to the CBPS, random forests, and model averages that tune for either accuracy or balance, to see whether (i) averaging in general yields less bias than the single model estimate given by the CBPS, and (ii) weights estimated by jointly considering accuracy and balance produce less bias than weights estimated to consider one of the two aspects.

The rest of the paper is structured as follows: Section 2 sets the notation and gives some preliminaries regarding causal inference, balancing propensity scores and frequentist model averaging. Section 3 provides further motivation of why accuracy should be taken into account, and describes the proposed methodology. Lastly, Section 4 provides some Monte Carlo experiments and Section 5 summarizes and discusses the contribution of the study.

## 2 Preliminaries and notation

For each observational unit $i = 1, \ldots, n$, suppose that the observed data consists of a treatment indicator $T_i \in \{0, 1\}$, a $k$-dimensional vector of pre-treatment covariates $\boldsymbol{x}_i$, and an outcome $Y_i$. Here, $T_i = 1$ indicates the $i$:th subject has been assigned to treatment, $T_i = 0$ indicates the opposite, and the observed outcome is given by $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$, where $Y_i(1)$ and $Y_i(0)$ are the potential outcomes under treatment and control, respectively. Further, define the propensity score as the probability of receiving treatment given the observed covariates, $e(\boldsymbol{x}_i) = P(T_i = 1 | \boldsymbol{x}_i)$. For later use, let $\boldsymbol{T} = (T_1, \ldots, T_n)^T$, $\boldsymbol{e} = [e(\boldsymbol{x}_1), \ldots, e(\boldsymbol{x}_n)]^T$, and $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$.

The parameter of interest is the average treatment effect, defined as

$$\tau = \mathbb{E}\left[Y(1) - Y(0)\right]. \tag{1}$$

Rubin (1974; 2005) provides a framework in which $\tau$ can be identified and estimated using observational data, within which there are two assumptions relevant to the current study. The first assumption is that the propensity score has common, strictly positive, support for all individuals.

**Assumption 1** *For $i \in \{1, \ldots, n\}$, it holds that $0 < e(\boldsymbol{x}_i) < 1$.*

The second assumption is that the potential outcomes are independent of treatment allocation, given the pre-treatment covariates.

**Assumption 2** $T_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid \boldsymbol{x}_i$.

Rosenbaum and Rubin (1983) show that, given Assumption 2, it holds that

$$T_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid e(\boldsymbol{x}_i), \tag{2}$$

which is of theoretical convenience. Hence, adding the fact that $\boldsymbol{e}(\boldsymbol{x})$ is a balancing score, the propensity score simplifies the estimation of $\tau$.

As mentioned earlier, the propensity score is employed in several approaches to estimating the average treatment effect, such as propensity score matching and inverse probability weighting. However, since the true propensity score is seldom known, it needs to be estimated (see Sections 2.1-2.3 for further discussion). Let $\hat{e}(\boldsymbol{x}_i)$ denote an estimate of the propensity score of the $i$:th subject. Then, formally, $\hat{e}(\boldsymbol{x}_i)$ is perfectly accurate if $\hat{e}(\boldsymbol{x}_i) = e(\boldsymbol{x}_i)$, and it has perfect balancing properties if $\boldsymbol{x}_i \perp\!\!\!\perp T_i | \hat{e}(\boldsymbol{x}_i)$.

Based on the estimated propensity score, various estimators, $\hat{\tau}$, of the average treatment effect have been proposed in the literature. This study focuses on the standardized inverse probability weighting estimator

$$\hat{\tau} = \sum_{i=1}^{n} \left( \frac{w_1(\boldsymbol{x}_i)}{\sum_{i=1}^{n} w_1(\boldsymbol{x}_i)} - \frac{w_0(\boldsymbol{x}_i)}{\sum_{i=1}^{n} w_0(\boldsymbol{x}_i)} \right) Y_i, \tag{3}$$

where $w_1(\boldsymbol{x}_i) = T_i \cdot [\hat{e}(\boldsymbol{x}_i)]^{-1}$, and $w_0(\boldsymbol{x}_i) = (1 - T_i) \cdot [1 - \hat{e}(\boldsymbol{x}_i)]^{-1}$. This choice is motivated by the fact that $\hat{\tau}$ can be decomposed, when the true treatment effect is constant, in a way that clearly shows why both accuracy and balancing properties are desirable. Such a decomposition is given and discussed in Section 3.

## 2.1 Covariate balancing propensity score

In order to find a propensity score estimator that explicitly considers covariate balance as an objective, Imai and Ratkovic (2014) introduce an estimation procedure in which the propensity score is estimated subject to moment conditions. Assuming the true propensity score is generated by a logistic model, i.e.

$$e(\boldsymbol{x}_i) = \frac{\exp\{\boldsymbol{x}_i^T \boldsymbol{\beta}\}}{1 + \exp\{\boldsymbol{x}_i^T \boldsymbol{\beta}\}},$$

where $\boldsymbol{\beta}$ is an unknown parameter vector, the CBPS procedure estimates $\boldsymbol{\beta}$ by solving

$$\frac{1}{n} \sum_{i=1}^{n} \left[ \frac{T_i \boldsymbol{\phi}(\boldsymbol{x}_i)}{e(\boldsymbol{x}_i)} - \frac{(1 - T_i)\boldsymbol{\phi}(\boldsymbol{x}_i)}{1 - e(\boldsymbol{x}_i)} \right] = 0, \tag{4}$$

using the generalized method of moments (GMM) or empirical likelihood (EL). Here, $\boldsymbol{\phi}(\boldsymbol{x}_i) = (\phi_1(\boldsymbol{x}_i), \ldots, \phi_L(\boldsymbol{x}_i))^T$ is a vector valued function that determines what moments are considered by the balancing restrictions. For example, when $\boldsymbol{\phi}(\boldsymbol{x}_i) = \boldsymbol{x}_i$ the estimated propensity score balances the first order empirical moments of the covariates, and when $\boldsymbol{\phi}(\boldsymbol{x}_i) = (x_{i1}, \ldots, x_{ik}, x_{i1}^2, \ldots, x_{ik}^2)$ the CBPS balances both the first and second order empirical moments. In a similar way, $\boldsymbol{\phi}$ can easily be chosen to yield an estimate that balances higher order moments and interaction terms.

The form of $\boldsymbol{\phi}$ is selected by the researcher, but the choice can have serious consequences for $\hat{\tau}$. If $\boldsymbol{\phi}$ is chosen to match the correctly specified propensity score model, then (4) implies that Assumption 2 holds and $\tau$ can be estimated without bias using the resulting propensity score estimates. However, if $\boldsymbol{\phi}$ represents a misspecified propensity score model, unbalanced but relevant moments can bias the ATE estimator. The former case is theoretically ideal but unlikely in practice, meaning that $\hat{\tau}$ is likely to be biased. This paper examines whether the bias incurred might be mitigated by also taking the accuracy of the estimated propensity score into consideration.

## 2.2 Covariate balancing scoring rule

Zhao et al. (2020) introduces a more general framework for estimating propensity scores using tailored loss functions, as opposed to the likelihood function used by the standard logistic regression model. Assuming the propensity score model belongs to some pre-specified parametric family $\mathcal{P} = \{e_{\boldsymbol{\beta}}(\boldsymbol{x}) : \boldsymbol{\beta} \in B\}$, and given a *strictly proper* scoring rule $\mathcal{S}$ – see e.g. Section 2.1 in Zhao (2019) for reference – the parameter $\boldsymbol{\beta}$ is estimated as

$$\hat{\boldsymbol{\beta}}_n = \underset{\boldsymbol{\beta} \in B}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{S}\left(e_{\boldsymbol{\beta}}(\boldsymbol{x}_i), T_i\right). \tag{5}$$

When $\mathcal{P} = \mathcal{P}_{\text{GLM}}$ represents the family of finite-dimensional generalized linear models (GLM), that is $\mathcal{P}_{\text{GLM}} = \left\{e_{\boldsymbol{\beta}}(\boldsymbol{x}) : e_{\boldsymbol{\beta}}(\boldsymbol{x}) = l^{-1}(\boldsymbol{x}^T\boldsymbol{\beta}),\ \boldsymbol{x} \in \mathbb{R}^p\right\}$ for some link function $l$ and some $p < \infty$, Zhao (2019) identifies that the first order conditions required to solve (5) can be interpreted as covariate balancing constraints, giving rise to the covariate balancing scoring rule (CBSR) approach. The form of the balancing constraint depends on the scoring rule and the link function used to estimate the model. In the special case where $l$ is the logistic link and the scoring rule is

$$\mathcal{S}^*\left[e_{\boldsymbol{\beta}}(\boldsymbol{x}_i), T_i\right] = T_i\left\{\log\left[\frac{e_{\boldsymbol{\beta}}(\boldsymbol{x}_i)}{1-e_{\boldsymbol{\beta}}(\boldsymbol{x}_i)}\right] - \frac{1}{e_{\boldsymbol{\beta}}(\boldsymbol{x}_i)}\right\} + (1-T_i)\left\{\log\left[\frac{1-e_{\boldsymbol{\beta}}(\boldsymbol{x}_i)}{e_{\boldsymbol{\beta}}(\boldsymbol{x}_i)}\right] - \frac{1}{1-e_{\boldsymbol{\beta}}(\boldsymbol{x}_i)}\right\}, \tag{6}$$

the propensity score estimated using the CBSR is equivalent to the CBPS, meaning that the CBSR may be viewed as a generalization of the CBPS.

The CBSR approach requires specification of a family of models and a link function. Using $\mathcal{P}_{\text{GLM}}$ is suitable when there is a belief that the linear predictor of the true propensity score lies in span $\{\psi_1(\boldsymbol{x}), \ldots, \psi_L(\boldsymbol{x})\}$, for some functions $\psi_1, \ldots, \psi_L$. However, when little or nothing is known about the true form of $e(\boldsymbol{x})$, this assumption can be inhibitive since it restricts the assumed model form to belong to a fixed and relatively low-dimensional space. Zhao (2019) suggests working around this restriction by defining $\mathcal{P}_{\text{RKHS}} = l^{-1}(\mathcal{H})$, where $\mathcal{H}$ denotes the reproducing kernel Hilbert space (see Zhao (2019) for discussion of RKHS) of a given kernel function $K$. This model space is extensive and allows for more flexible estimation of $e(\boldsymbol{x})$.

Assuming the true form of $e(\boldsymbol{x})$ belongs to $\mathcal{P}_{\text{RKHS}}$, the estimated propensity score is given by

$$\hat{\boldsymbol{e}} = (\hat{e}(\boldsymbol{x}_1), \ldots, \hat{e}(\boldsymbol{x}_n))^T = \underset{p(\cdot)\in\mathcal{P}_{\text{RKHS}}}{\operatorname{argmax}} \left\{\frac{1}{n}\sum_{i=1}^{n} \mathcal{S}\left(p(\boldsymbol{x}_i), T_i\right) - \lambda\|p(\cdot)\|_{\mathcal{H}}\right\}, \tag{7}$$

where $\lambda$ is a regularization parameter, and the features required to form $\mathcal{P}_{\text{RKHS}}$ can be obtained by evaluation of the kernel $K$. Here $\mathcal{S}$ may be a balancing scoring rule, like $\mathcal{S}^*$, or some other form of strictly proper scoring rule. In view of Mercer's (1909) theorem, the approach can yield approximate balance of a possibly infinite range of moments, depending on the choice of kernel function. The moments to balance are determined by the choice of kernel function, meaning that specification of individual moments, as required when using the CBPS, is not part of using the CBSR.

## 2.3 Frequentist model average estimators

Suppose that there are $M$ propensity score models under consideration, and let $\hat{\boldsymbol{e}}_m = (\hat{e}_{1m}, \ldots, \hat{e}_{nm})^T$, $m = 1, \ldots, M$, collect the $m$:th candidate propensity score for all $n$ individuals. Then, the frequentist model average propensity score is given by

$$\bar{\boldsymbol{e}}(\boldsymbol{w}) = \sum_{m=1}^{M} w_m \hat{\boldsymbol{e}}_m. \tag{8}$$

Here, $w_m$ denotes the weight given to the $m$:th candidate, and forms part of the vector $\boldsymbol{w} = (w_1, \ldots, w_m)$ defined on the unit simplex

$$\mathcal{W} = \left\{\boldsymbol{w} : \boldsymbol{w} \in [0,1]^M, \sum_{m=1}^{M} w_m = 1\right\}. \tag{9}$$

The weights are not given a priori, meaning that their values need to be estimated.

In the early model averaging literature, much emphasis is placed on estimating weights that yield a final estimator, $\bar{\boldsymbol{e}}$, which is as accurate as possible. Here, accuracy may be measured in terms of, for example, mean squared error (Hansen, 2007; Hansen and Racine, 2012) or the Kullback-Leibler divergence (Zhang et al., 2015). However, since the covariate balancing properties of the estimator are also crucial, some adaptations of model averaging select weights that minimize disparities in covariate balance between the treatment and control groups. For example, Xie et al. (2019) use

$$\hat{w}e_{\text{RF}}(\boldsymbol{x}) + (1-\hat{w})e_{\text{LR}}(\boldsymbol{x}) \tag{10}$$

as the combined propensity score. Here $e_{\text{RF}}(\boldsymbol{x})$ is estimated from a random forest, $e_{\text{LR}}(\boldsymbol{x})$ is estimated using logistic regression, and $\hat{w}$ is selected to minimize some function of the absolute

standardized mean differences, such as their mean, median or maximum value. This approach is similar to the balance super learner of Pirracchio and Carone (2018), which combines different candidate algorithms based on a metric designed to quantify covariate imbalance. For each candidate algorithm, the metric is computed as

$$L = \frac{1}{k}\sum_{j=1}^{k} \omega_j \cdot \text{ASMD}_j, \tag{11}$$

where $\text{ASMD}_j$ is the absolute standardized mean difference of the $j$:th covariate and $\omega_j = \hat{\beta}_j \cdot \hat{\sigma}_j$. Here $\hat{\beta}_j$ is the coefficient given to the $j$:th covariate when regressing the outcomes $Y_i$ on the pre-treatment covariates, and $\hat{\sigma}_j$ is the sample standard deviation of covariate $j$. Thus, note that, unlike Xie et al. (2019) and the approach to be introduced below, the balance super learner of Pirracchio and Carone (2018) uses outcome data when producing the combined propensity score estimate.

# 3 Proposed methodology

As stated above, model average estimators of different forms have been discussed in the causal inference literature. However, there are no papers that suggest combining accuracy and covariate balancing conditions using model averaging. This paper studies whether the combination of propensity scores can be tailored to provide a weighted propensity score that is both accurate and that results in sufficient covariate balance.

To see why both aspects are relevant, suppose

$$Y(1) = \mu(\boldsymbol{x}) + \tau + \varepsilon, \tag{12}$$
$$Y(0) = \mu(\boldsymbol{x}) + \varepsilon, \tag{13}$$

where $\mu$ is some function of the pre-treatment covariates, $\tau$ denotes the homogeneous treatment effect, and $\varepsilon$ is a random error term. Then, assuming knowledge of the true forms of the propensity score and the function $\mu(\boldsymbol{x})$, the IPW estimator of $\tau$ can be decomposed as

$$\begin{aligned}
\hat{\tau} &= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\left[T_i\left(\mu\left(\boldsymbol{x}_i\right)+\tau+\varepsilon_i\right)\right]/e(\boldsymbol{x}_i)}{\sum_{i=1}^{n} T_i/e(\boldsymbol{x}_i)} - \frac{\left[T_i\left(\mu\left(\boldsymbol{x}_i\right)+\varepsilon_i\right)\right]/(1-e(\boldsymbol{x}_i))}{\sum_{i=1}^{n}(1-T_i)/(1-e(\boldsymbol{x}_i))}\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_i/e(\boldsymbol{x}_i)}{\sum_{i=1}^{n} T_i/e(\boldsymbol{x}_i)} - \frac{(1-T_i)/(1-e(\boldsymbol{x}_i))}{\sum_{i=1}^{n}(1-T_i)/(1-e(\boldsymbol{x}_i))}\right]\mu(\boldsymbol{x}_i) \qquad (14) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_i/e(\boldsymbol{x}_i)}{\sum_{i=1}^{n} T_i/e(\boldsymbol{x}_i)} - \frac{(1-T_i)/(1-e(\boldsymbol{x}_i))}{\sum_{i=1}^{n}(1-T_i)/(1-e(\boldsymbol{x}_i))}\right]\varepsilon_i \qquad (15) \\
&\quad + \tau,
\end{aligned}$$

Under the assumptions given in Section 2, by the law of large numbers, (14) and (15) tend to 0 as $n \to \infty$. However, in practice neither $e(\boldsymbol{x})$ nor $\mu(\boldsymbol{x})$ is known, and some estimated quantities have to be plugged in.

Now, suppose $\mu(\boldsymbol{x}) = \varphi(\boldsymbol{x}) + \rho(\boldsymbol{x})$, where $\varphi$ is correctly specified by the practitioner and $\rho$ can be thought of as some error containing terms that are not balanced by the estimated propensity score. Then, the term (14) expands to

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_i/e(\boldsymbol{x}_i)}{\sum_{i=1}^{n} T_i/e(\boldsymbol{x}_i)} - \frac{(1-T_i)/(1-e(\boldsymbol{x}_i))}{\sum_{i=1}^{n}(1-T_i)/(1-e(\boldsymbol{x}_i))}\right]\varphi(\boldsymbol{x}_i) \tag{16}$$
$$+\frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_i/e(\boldsymbol{x}_i)}{\sum_{i=1}^{n} T_i/e(\boldsymbol{x}_i)} - \frac{(1-T_i)/(1-e(\boldsymbol{x}_i))}{\sum_{i=1}^{n}(1-T_i)/(1-e(\boldsymbol{x}_i))}\right]\rho(\boldsymbol{x}_i). \tag{17}$$

Assuming that $\hat{e}(\boldsymbol{x})$ is estimated to balance the moments given in $\varphi$, plugging it into the above expression ensures (16) is approximately 0. However, this estimator does not necessarily balance the terms needed to compute $\rho$, and so an accurate propensity score estimator might balance $\rho$ to moderate the bias incurred from (17) being non-zero. This line of reasoning informs the construction of cross validation criteria given below.

### 3.1 Selection of weights by $V$-fold cross-validation

The literature on model averaging discusses a variety of approaches to estimating model average weights. For example, Hansen (2007) selects weights based on a Mallows criterion, while Zhang et al. (2015) use an approach based on the Kullback-Leibler divergence, and Hansen and Racine (2012) use a jackknife criterion. The studies mentioned above focus on model averaging in linear regression, while Zhao et al. (2019) employ $V$-fold cross-validation to estimate probabilities in discrete choice models. The latter approaches, based on cross-validation, have been widely applied, and their performance has been studied in detail. Moreover, other model average approaches developed for causal inference (for example Pirracchio et al., 2015; Pirracchio and Carone, 2018) are based on weights selected using cross-validation. As such, the weights used to compute the model averages in this study are selected using $V$-fold cross-validation based either on the covariate balancing constraint (4), on the mean squared prediction error $[\boldsymbol{T} - \bar{\boldsymbol{e}}(\boldsymbol{w})]^T[\boldsymbol{T} - \bar{\boldsymbol{e}}(\boldsymbol{w})]$, or a combination of the two criteria.

Suppose that the data $(\boldsymbol{T}, \boldsymbol{X})$ is split into $V$ independent groups, $v_1, \ldots, v_V$. Further, let $(\boldsymbol{T}, \boldsymbol{X})_v$ be the data on observations belonging to group $v$, and $(\boldsymbol{T}, \boldsymbol{X})_{-v}$ be the data for observations not in group $v$. Define $\tilde{\boldsymbol{e}}_m$ as the vector whose sub-vectors $\tilde{\boldsymbol{e}}_{m,v}$ collect the $m$:th candidate propensity scores of the subjects of group $v$, estimated using $(\boldsymbol{T}, \boldsymbol{X})_{-v}$, for $v \in \{v_1, \ldots, v_V\}$. In addition, let

$$\tilde{\boldsymbol{e}}(\boldsymbol{w}) = \sum_{m=1}^{M} w_m \tilde{\boldsymbol{e}}_m. \tag{18}$$

The marginal (in the sense that they consider *either* accuracy or balance) criteria for weight selection are given by

$$C_a(\boldsymbol{w}) = \left[\boldsymbol{T} - \tilde{\boldsymbol{e}}(\boldsymbol{w})\right]^T \left[\boldsymbol{T} - \tilde{\boldsymbol{e}}(\boldsymbol{w})\right], \tag{19}$$

and

$$C_b(\boldsymbol{w}) = \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{T_i \boldsymbol{\phi}(\boldsymbol{x}_i)}{\tilde{\boldsymbol{e}}(\boldsymbol{w})} - \frac{(1 - T_i)\boldsymbol{\phi}(\boldsymbol{x}_i)}{1 - \tilde{\boldsymbol{e}}(\boldsymbol{w})} \right] \right\}^2. \tag{20}$$

Here, the subscripts $a$ and $b$ indicate that the criteria focus either on the accuracy, $a$, or balancing properties, $b$, of $\bar{\boldsymbol{e}}$. Using these criteria, combination weights are estimated as

$$\hat{\boldsymbol{w}}_j = \underset{\boldsymbol{w} \in \mathcal{W}}{\operatorname{argmin}} \, C_j(\boldsymbol{w}), \; j \in \{a, b\}. \tag{21}$$

Plugging the estimates weights into (8) gives $\bar{\boldsymbol{e}}(\hat{\boldsymbol{w}})$, the $V$-fold cross-validation frequentist model average estimate of the propensity score. Furthermore, plugging $\bar{\boldsymbol{e}}(\hat{\boldsymbol{w}})$ into (3) produces the corresponding FMA estimate of the average treatment effect, $\hat{\tau}(\hat{\boldsymbol{w}})$.

As mentioned, $C_a(\boldsymbol{w})$ is a criterion that controls the mean squared prediction error of $\tilde{\boldsymbol{e}}(\boldsymbol{w})$ as a prediction of the vector of treatment indicators. Zhao et al. (2019) shows that, under certain conditions, when all candidate models are misspecified,

$$\frac{\|\bar{\boldsymbol{e}}(\hat{\boldsymbol{w}}_a) - \boldsymbol{e}\|^2}{\inf_{\boldsymbol{w} \in \mathcal{W}} \|\bar{\boldsymbol{e}}(\boldsymbol{w}) - \boldsymbol{e}\|^2} \xrightarrow{p} 1, \tag{22}$$

where $\|\boldsymbol{x} - \boldsymbol{y}\| = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^T(\boldsymbol{x} - \boldsymbol{y})}$. This means that asymptotically, $\bar{\boldsymbol{e}}(\hat{\boldsymbol{w}}_a)$ estimates $\boldsymbol{e}$ as accurately as the infeasible best model average estimator based on the current $M$ candidates. However, as discussed above, joint consideration of both accuracy and balancing properties may improve the estimator $\hat{\tau}$. Thus, the weights

$$\hat{\boldsymbol{w}}_{ab} = \underset{\boldsymbol{w} \in \mathcal{W}}{\operatorname{argmin}} \left\{ C_a(\boldsymbol{w}) \right\} \text{ s.t. } C_b(\boldsymbol{w}) \leq \min_{m \in 1, \ldots, M} C_b\left(\boldsymbol{w}_m^0\right) \tag{23}$$

are estimated to be as accurate as possible under some balancing restriction. Here, $\boldsymbol{w}_m^0$ is the vector that gives weight 1 to the $m$:th candidate model, and weight 0 to all other candidates, meaning that $\hat{\boldsymbol{w}}_{ab}$ is estimated to make $\bar{\boldsymbol{e}}(\boldsymbol{w})$ as accurate as possible while balancing the covariates at least as well as the best balancing candidate.

### 3.2 Formulation of candidate models

As discussed in Section 2.3, it is common to form a set of candidate models by letting the elements represent competing estimation methods. In Xie et al. (2019), for example, the number of candidates is $M = 2$, and the candidate propensity scores are computed using logistic regression and a random forest. In the current paper, however, the same framework of estimation is employed for all candidates, but the models themselves are distinguished by their value of the regularization parameter $\lambda$.

The parameter $\lambda$ regulates the degree to which the estimate of the penalized regression parameter is shrunk towards 0, and thus moderates the contribution of the regressors associated with the penalized coefficients to the final estimate. Since each possible $\lambda$ produces a unique candidate propensity, the specification of $\lambda$ can be interpreted as model selection. Essentially, the estimated propensity score is a function of $\lambda$, and the model average is given by

$$\bar{\boldsymbol{e}}_j(\hat{\boldsymbol{w}}) = \sum_{M=1}^{M} \hat{w}_m \hat{\boldsymbol{e}}(\lambda_m). \tag{24}$$

Since $\lambda$ is a theoretical construct, its value has to be determined, under some uncertainty, by the practitioner. Thus, averaging over a sequence of $\lambda$ values constitutes a way of incorporating model selection uncertainty into the estimation process.

This way of using $\lambda$ as a means of forming candidate models is employed similarly in Schomaker (2012), Zhao et al. (2020), and Zulj and Jin (2021). Zhao et al. (2020) prove that, for regularized linear regression models, the model average estimator attains the oracle property (22). Zulj and Jin (2021) prove a similar result for generalized linear models.

## 4 Monte Carlo experiments

Following the example set by Li and Li (2021), the sequence $\{\lambda_1, \ldots, \lambda_M\}$ is made up of 20 values between 0 and 1, equally spaced on the $\log_{10}$ scale. As such, each FMA propensity score incorporates the information of 20 separate candidate models, often producing markedly different estimates.

Two sets of candidate models are estimated and used to form competing model averages. The candidate sets are given by

- $\mathcal{M}_{\text{CBSR}}$: a set of 20 candidates estimated as in (7), using the CBSR given in (6). These candidates are estimated to provide good covariate balance.
- $\mathcal{M}_{\text{GLM}}$: a set of 20 candidates estimated as in (7), but using the Bernoulli log-likelihood instead of a scoring rule $\mathcal{S}$. These candidates can be viewed as augmented logistic regression models, and do not target covariate balance specifically.

Both estimation procedures use the logistic link function, meaning that

$$\mathcal{P}_{\text{RKHS}} = \left\{ p(\boldsymbol{x}, \boldsymbol{k}) : p(\boldsymbol{x}, \boldsymbol{k}) = \text{logit}^{-1}\left(\boldsymbol{x}^T\boldsymbol{\beta} + \boldsymbol{k}^T\boldsymbol{\alpha}\right) \right\}, \tag{25}$$

where $\boldsymbol{x}$ is the $i$:th row of the data matrix $\boldsymbol{X}$, and $\boldsymbol{k}$ is the $i$:th row of the kernel matrix $\boldsymbol{K}$, computed using the standard Gaussian kernel function. Moreover, the penalty function $\lambda\|p(\cdot)\|_{\mathcal{H}} = \lambda\boldsymbol{\alpha}^T\boldsymbol{K}\boldsymbol{\alpha}$, meaning that $\boldsymbol{\alpha}$ is penalized but $\boldsymbol{\beta}$ is not. Thus, note that, as $\lambda \to \infty$, a CBSR candidate using the covariate matrix $\boldsymbol{X}$ tends to the CBPS using the same candidates.

The model averages CBSR A and GLM A are computed with weights estimated using the accuracy criterion, $C_a(\boldsymbol{w})$, given in (19). Similarly, CBSR B and GLM B are formed with weights estimated using the balancing criterion $C_b(\boldsymbol{w})$, while the averages CBSR AB and GLM AB use weights estimated by solving the optimization problem posed in (23). The AB averages consider both accuracy and balance, and CBSR AB represents the main methodological proposal of the current paper. Here, weighting the CBSR models using the accuracy criterion can also be thought of as combining accuracy and balance simultaneously. This is because the candidate propensity scores themselves are estimated under balancing constraints, although tuning the average weights solely for accuracy may influence balancing properties of the combined estimates.

The average estimators are compared to the best model from each set of candidates, evaluated in terms of balancing ability, i.e. for each set $\mathcal{M}_{\text{CBSR}}$ and $\mathcal{M}_{\text{GLM}}$, the candidate model which

yields the smallest value of the balancing criterion (20) is selected for comparison. Henceforth, these models are referred to as CBSR $\lambda_{\min}$ and GLM $\lambda_{\min}$. Moreover, the standard CBPS (made to balance first and second order moments as well as the corresponding interaction terms) and a random forest are estimated and used for comparison.

Throughout the simulations that follow, the sample size is varied as $n \in \{300, 1000\}$, the number of folds in cross-validation is set to $V = 5$, and 2000 replications are made for each design. In order to estimate the weights, the `R` function `kernlab::ipop` (Karatzoglou et al., 2004) is used to solve quadratic programming problems, and the function `Rsolnp::solnp` (Ghalanos and Theussl, 2015) is used to solve nonlinear programming problems. Full access to simulation code is given here.

## 4.1 Data generation

The Monte Carlo experiments are performed using three different simulation designs, each aiming to investigate the performance of the suggested approach under different circumstances.

The first simulation design mimics one of the settings used by Lunceford and Davidian (2004). This is a well-known paper and the design has been replicated frequently. In this setting, all six covariates are used to generate the potential outcomes, but only three of them are used to generate the propensity scores. In other words, this scenario studies the case where the estimated propensity scores do not balance all covariates that influence the potential outcomes, while it does have the possibility to balance all confounders.

The second design copies the simulated example given by Kang and Schafer (2007). In this case, each potential outcome and the propensity score is generated using a four-dimensional covariate vector $\boldsymbol{z}$, but only a function of the generating data, $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{z})$, can be used to specify the propensity score model. As such, this design studies the way in which the estimation approaches work under a certain degree of model misspecification.

Finally, the third design studies the performance of the different approaches when the treatment effect is heterogeneous, i.e. different for every individual. Here, the potential outcome under treatment, $Y(1)$ is generated using a Gaussian process based on some covariates, while the potential outcome of individuals that have not been exposed to treatment is generated as a linear function of the same covariates. The designs are described in greater detail in the following three subsections.

### 4.1.1 First simulation design

The first simulation design is taken from Lunceford and Davidian (2004), and the parameter values are taken to give strong associations between the covariates and $Y$ and $T$ (i.e. $\boldsymbol{\xi}^{\text{str}}$ and $\boldsymbol{\beta}^{\text{str}}$ in the cited paper). The design includes six different covariates, three of which, $x_1$, $x_2$ and $x_3$, are associated with both $Y$ and $T$, while a further three, $v_1$, $v_2$ and $v_3$, are associated with the outcome only. $x_3$ is generated from Bernoulli(0.2). Then, conditional on $x_3$, $v_3$ is drawn from the Bernoulli distribution with probability given by $0.75 \cdot x_3 + 0.25 \cdot (1 - x_3)$, and, the remaining four covariates $(x_1, v_1, x_2, v_2)$ are generated from $\mathcal{N}_4(\boldsymbol{\gamma}_{x_3}, \boldsymbol{\Sigma})$. Here, $\boldsymbol{\gamma}_1 = (1, 1, -1, -1)^T$, $\boldsymbol{\gamma}_0 = (-1, -1, 1, 1)^T$, and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.5 & -0.5 & -0.5 \\ 0.5 & 1 & -0.5 & -0.5 \\ -0.5 & -0.5 & 1 & 0.5 \\ -0.5 & -0.5 & 0.5 & 1 \end{pmatrix}.$$

The propensity score is generated as $e(\boldsymbol{x}) = \{1 + \exp[-(0.6x_1 - 0.6x_2 + 0.6x_3)]\}^{-1}$, using which the treatment assignment is sampled from Bernoulli$(e(\boldsymbol{x}))$. Lastly, the outcome variable is simulated as

$$Y = -x_1 + x_2 - x_3 + 2T - v_1 + v_2 + v_3 + \varepsilon,$$

where $\varepsilon \sim \mathcal{N}(0, 1)$. This gives a true average treatment effect of $\tau = 2$. The candidate procedures are allowed consider all six covariates when estimating the propensity scores.

### 4.1.2 Second simulation design

The second Monte Carlo experiment follows the design first given by Kang and Schafer (2007). Here, observations $(z_1, z_2, z_3, z_4)$ are drawn independently from $\mathcal{N}_4(\boldsymbol{0}, \boldsymbol{I})$ for each individual $i = 1, \ldots, n$. The true propensity score is simulated as

$$e(\boldsymbol{z}) = \{1 + \exp[-(z_1 - 0.5z_2 + 0.25z_3 + 0.1z_4)]\}^{-1},$$

and used to sample a treatment indicator as before. The outcome is generated as

$$Y = 210 + 24.7z_1 + 13.7z_2 + 13.7z_3 + 13.7z_4 + \varepsilon,$$

with $\varepsilon \sim \mathcal{N}(0, 1)$. This gives a true average treatment effect of $\tau = 0$.

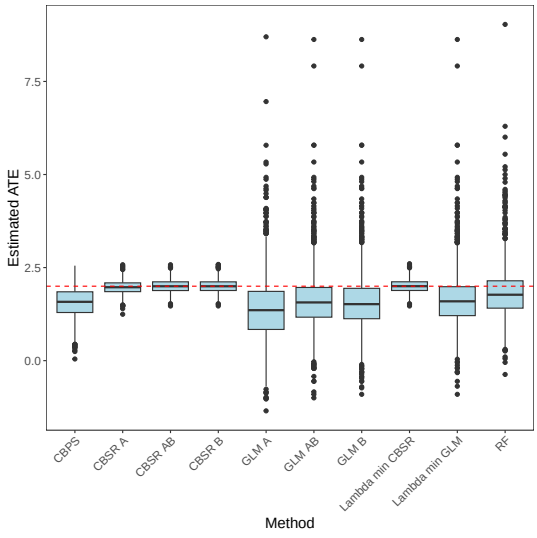
However, for the purpose of modeling the propensity score, only

$$x_1 = \exp\{z_1/2\},\ x_2 = \frac{z_2}{1 + \exp\{z_1\}} + 10,\ x_3 = \left(\frac{z_1 \cdot z_3}{25} + 0.6\right)^3,\ x_4 = (z_2 + z_4 + 20)^2$$

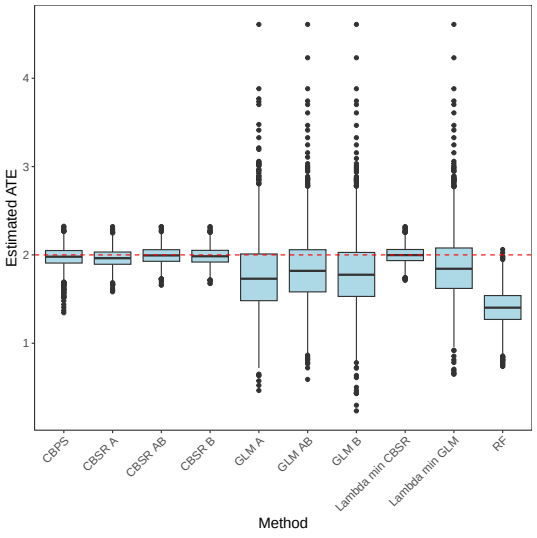
are given. The relationship between $(z_1, z_2, z_3, z_4)$ and $(x_1, x_2, x_3, x_4)$ are taken as unknown, necessarily giving rise to misspecification of the propensity score model. Here, the CBPS is expected to perform badly because it is tasked with balancing the moments of $\boldsymbol{\phi}(x_1, x_2, x_3, x_4)$ rather than $\boldsymbol{\phi}(z_1, z_2, z_3, z_4)$ which are the moments of importance.

### 4.1.3 Third simulation design

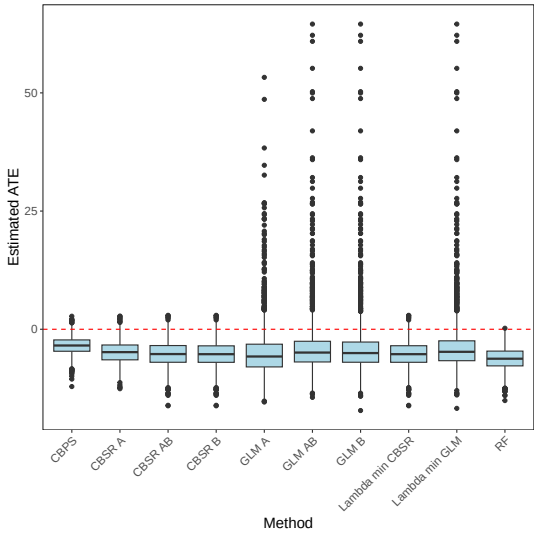
The third simulation setting considers a setting where the treatment effect is non-homogeneous. Let $\boldsymbol{y}_{\text{out}}$ and $\boldsymbol{y}_{\text{ps}}$ denote two realizations of a mean zero Gaussian process, where the standard Gaussian covariance function, and a sequence of 500 four dimensional Halton numbers have been used to form the kernel matrix. Further, let $f_{\text{out}}$ and $f_{\text{ps}}$ be approximations of $\boldsymbol{y}_{\text{out}}$ and $\boldsymbol{y}_{\text{ps}}$, estimated using Gaussian process regressions.



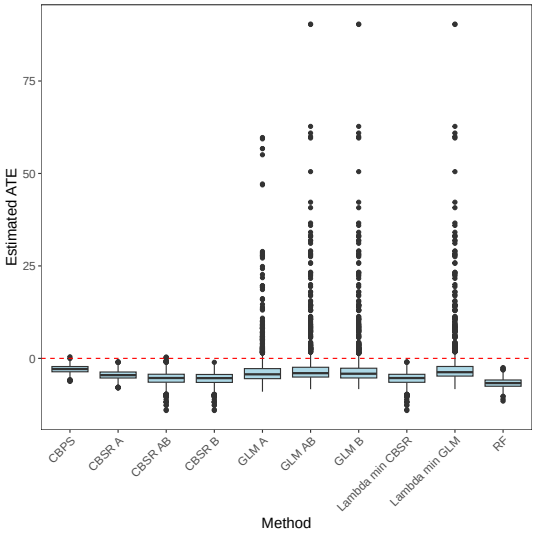

(a) First simulation design, $n = 300$



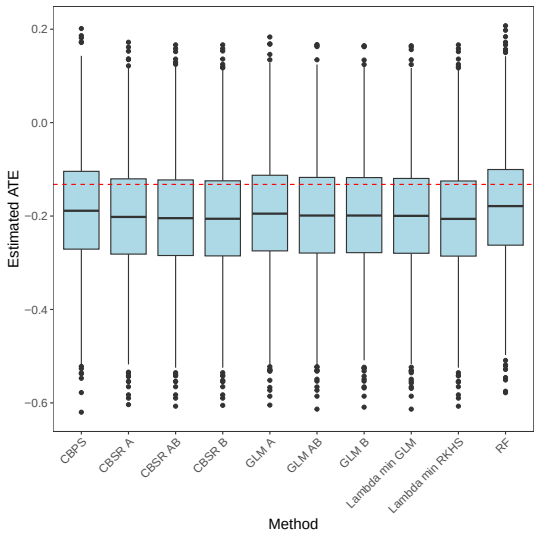

(b) First simulation design, $n = 1000$



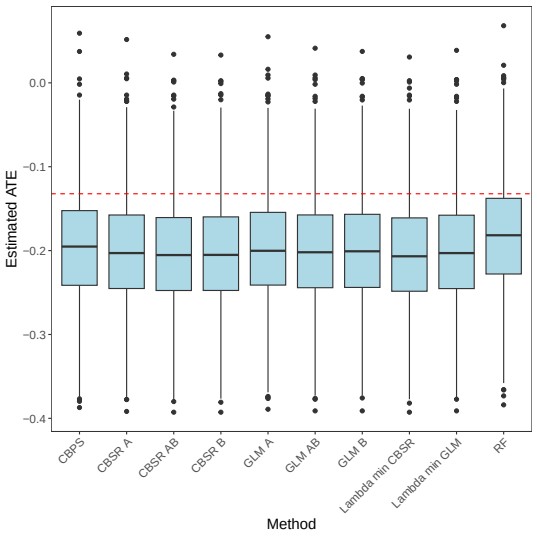

(c) Second simulation design, $n = 300$




(d) Second simulation design, $n = 1000$




(e) Third simulation design, $n = 300$




(f) Second simulation design, $n = 1000$

Figure 1: Box plots showing estimates of $\hat{\tau}$ for different designs (rows) and sample sizes (columns). In all plots, the red, dotted line represents the true average treatment effect.

In order to generate treatment indicators and outcome values, the pre-treatment covariate vectors $\boldsymbol{x}$ are generated from $\mathcal{N}_4(\mathbf{0}, \boldsymbol{I})$. The propensity scores are generated as

$$e(\boldsymbol{x}) = \{1 + \exp[-f_{\text{ps}}(\boldsymbol{x})]\}^{-1},$$

and used to generate the treatment indicator $T$ from Bernoulli$(e(\boldsymbol{x}))$. Depending on $T$, the outcome is simulated as

$$Y = \begin{cases} f_{\text{out}}(\boldsymbol{x}) + \nu & \text{if } T = 1, \text{ and} \\ 1.5c + c\boldsymbol{x}^T\boldsymbol{\beta} + \varepsilon, & \text{if } T = 0. \end{cases} \tag{26}$$

Here, $\boldsymbol{\beta} = (1, -1, 2, 2)^T$, $c = 0.4126469$ $\nu \sim \mathcal{N}(0, 0.35^2)$, and $\varepsilon \sim \mathcal{N}(0, 1)$, where $c$ and the variance of $\nu$ are chosen such that the signal-to-noise ratio is approximately 0.6 in both outcome settings. Using one million replicates of $Y(1)$ and $Y(0)$, the true treatment effect is approximated by the mean difference as $\tau = -0.132237$

The candidate models are given the raw covariates $\boldsymbol{x}$ to consider, and no knowledge of the generating process is assumed in the modeling stage. In the simulation, the candidate propensity scores and the weights are estimated according to the procedure outlined in the paper. That is, no special care is taken due to the fact that the treatment effect is heterogeneous.

## 4.2 Results

The treatment effect estimates rendered by the simulations are illustrated using box plots in Figure 1, where each row of plots represents one simulation design, and the left column holds simulations using $n = 300$ while the right columns holds simulations using $n = 1000$.

### 4.2.1 First simulation design

The results from the first simulation design are given in Figures 1a and 1b. As can be seen, the results suggest that the approaches based on CBSR estimate the average treatment effect well. The median bias is small compared to the other methods. However, there is no suggestion that the model averages outperform the single model approach CBSR $\lambda_{\text{min}}$, and there is no evidence that considering both balance and accuracy in the model average (CBSR A and CBSR AB) improves the result obtained when only considering balance (CBSR B).

When it comes to the regularized logistic regression, neither model average seems to better the selection of a single candidate model. The model selected using cross-validation and the balancing criterion performs similarly in terms of both bias and variance. The figures suggests that the combined approach GLM AB and the balance only approach produce bias of similar magnitude, as does the $\lambda_{\text{min}}$ approach. Moreover, the results suggest the random forest is preferable to any GLM based estimator, and that the CBPS is preferred to GLM A.

### 4.2.2 Second simulation design

The results from the second simulation design are given in Figures 1c and 1d, and reflect the results from the first design by indicating that there is little to gain from averaging candidates. Again, the CBSR approaches perform similarly and have relatively small spread, although here their bias is more similar to that of the competing modeling strategies. The GLM averages perform on par with their CBSR counterparts in terms of estimated median bias, but the variability in the estimates of $\tau$ is much larger when using GLM candidates. There is no clear pattern as to when averaging could be useful, and nothing suggests that combining balance and accuracy gives better treatment effect estimates in general.

In this setting, however, the random forest approach and the CBPS are comparable to the CBSR regressions, both in terms of bias and accuracy.

### 4.2.3 Third simulation design

The results from the third simulation design are given in Figures 1e and 1f. The results are different from those generated by the previous simulation designs in that all strategies used to model the propensity score and estimate the treatment effect perform similarly. This is the case both in terms of median bias and in terms of the spread of $\hat{\tau}$, although the CBPS and the random forest are slightly better than other approaches when it comes to bias. The results of the third

simulation do not support the claim that the combination of propensity scores improves estimation of treatment effects in general. Further, they do not indicate that combining accuracy and balance when estimating combination weights reduces the bias of the final ATE estimator.

# 5 Discussion and conclusions

Compromise estimators have been applied in many fields of statistics, largely as a means of weighting together predictions to increase prediction accuracy. In causal inference, such estimators have been implemented in order to estimate more accurate propensity scores, but also to specifically target the balancing properties of the estimated propensity scores. In practice, there is no way of evaluating the accuracy of estimated propensity scores, and there is no way of knowing whether they balance all confounders. Since good accuracy and balancing properties are known to mitigate bias of the ATE estimator, the current paper proposes a compromise estimator where the weights are estimated to target both accuracy and balance simultaneously. The proposal is implemented mainly through averaging of candidate models estimated using the covariate balancing scoring rule, a modern and flexible approach to estimating balancing propensity scores. Another implementation is given through averaging of RKHS logistic regression models.

The results given in Section 4 do not support the hypothesis that model averaging necessarily leads to less biased estimation of causal effects. Further, they provide no evidence that considering accuracy in combination with balance yields a better ATE estimator than considering balance or accuracy alone. The support for this claim is consistent over the three simulation designs presented in this paper, as well as several other simulations that, for brevity, have been left out of the final report.

In view of the discussion given in connection to (14)-(17), the simulation results are not as expected. The general literature on model averaging suggests that model averages suitably reduce the mean squared prediction error, as compared to the individual predictions. In particular, when used for estimation of probabilities, several studies show that averaging can increase accuracy, as discussed in connection to (22). Although the oracle property has not been proved for RKHS regression models, it is reasonable to assume that it may benefit the accuracy of such models as well. Furthermore, the results of Pirracchio and Carone (2018) indicate that averaging for balance can reduce the bias of the estimated treatment effects. It is conceivable that averaging for accuracy negatively impacts the balancing properties, and vice versa, so that the gains of averaging one aspect are canceled by the concession made in the other aspect. However, this trade-off should not be relevant when using the combined weights $\hat{\boldsymbol{w}}_{ab}$, since the accuracy is assumed to improve while the ability to balance the covariates is not compromised.

Considering the performance of the CBSR $\lambda_{\text{min}}$ approach, one explanation may be that the CBSR approach itself balances the covariates as well as possible. In that case, further tuning the accuracy may yield no or negligible improvement. This could also explain why, at least in the first simulation design, CBSR $\lambda_{\text{min}}$ performs better than the GLM $\lambda_{\text{min}}$, as the latter is estimated without taking balance into account in any way. However, it provides no clarity as to why tuning the RKHS GLM approaches for accuracy or balance offers no improvement.

In the general literature, model averaging forms part of a one-step estimation procedure, where the weight selection criterion is used to generate the model average that best fits the final outcome, with normal linear regression being the typical example. In the current situation, model averaging is used in the first (propensity score estimation) of two steps, and there is no known way of directly tuning the average to reduce or minimize bias in the second step (estimation of $\tau$). That is, while the propensity scores estimated may find the ultimate trade-off between accuracy and balance, in terms of the weight selection criterion, there is no guarantee that the improvement is expressed in terms of any notable improvement in the second step. Direct consideration of the second step is possible in the $\sqrt{n}$ local asymptotic framework, as detailed by Kitagawa and Muris (2016), but is difficult in standard asymptotic framework.

# References


Autenrieth, M., Levine, R. A., Fan, J., and Guarcello, M. A. (2021). Stacked ensemble learning for propensity score methods in observational studies. *Journal of Educational Data Mining*, 13(1):24–189.

Cannas, M. and Arpino, B. (2019). A comparison of machine learning algorithms and covariate balance measures for propensity score matching and weighting. *Biometrical Journal*, 61(4):1049–1072.

Chang, S.-H., Chou, I.-J., Yeh, Y.-H., Chiou, M.-J., Wen, M.-S., Kuo, C.-T., See, L.-C., and Kuo, C.-F. (2017). Association between use of non–vitamin k oral anticoagulants with and without concurrent medications and risk of major bleeding in nonvalvular atrial fibrillation. *JAMA : the Journal of the American Medical Association*, 318(13):1250–1259.

Ghalanos, A. and Theussl, S. (2015). *Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method*. R package version 1.16.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica*, 75(4):1175–1189.

Hansen, B. E. and Racine, J. S. (2012). Jackknife model averaging. *Journal of Econometrics*, 167(1):38–46.

Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, 98(464):879–899.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 76(1):243–263.

Ju, C., Combs, M., Lendle, S. D., Franklin, J. M., Wyss, R., Schneeweiss, S., and van der Laan, M. J. (2019). Propensity score prediction for electronic healthcare databases using super learner and high-dimensional propensity score methods. *Journal of Applied Statistics*, 46(12):2216–2236.

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

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20.

Kitagawa, T. and Muris, C. (2016). Model averaging in semiparametric estimation of treatment effects. *Journal of Econometrics*, 193(1):271–289.

Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29(3):337–346.

Li, J., Handorf, E., Bekelman, J., and Mitra, N. (2016). Propensity score and doubly robust methods for estimating the effect of treatment on censored cost. *Statistics in medicine*, 35(12):1985–1999.

Li, Y. and Li, L. (2021). Propensity score analysis methods with balancing constraints: A Monte Carlo study. *Statistical Methods in Medical Research*, 30(4):1119–1142.

Lu, X. (2015). A covariate selection criterion for estimation of treatment effects. *Journal of Business and Economic Statistics*, 33(4):506–522.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, 23(19):2937–2960.

Mercer, J. (1909). Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London. Series A, Containing papers of a mathematical or physical character*, 209:415–446.

Mörk, E., Sjögren, A., and Svaleryd, H. (2020). Consequences of parental job loss on the family environment and on human capital formation-evidence from workplace closures. *Labour Economics*, 67:101911–.

Ning, Y., Sida, P., and Imai, K. (2020). Robust estimation of causal effects via a high-dimensional covariate balancing propensity score. *Biometrika*, 107(3):533–554.

Pirracchio, R. and Carone, M. (2018). The balance super learner: A robust adaptation of the super learner to improve estimation of the average treatment effect in the treated based on propensity score matching. *Statistical Methods in Medical Research*, 27(8):2504–2518.

Pirracchio, R., Petersen, M. L., and van der Laan, M. (2015). Improving propensity score estimators'robustness to model misspecification using super learner. *American Journal of Epidemiology*, 181(2):108–119.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

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

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.

Schomaker, M. (2012). Shrinkage averaging estimation. *Statistical Papers (Berlin, Germany)*, 53(4):1015–1034.

Wan, A. T., Zhang, X., and Wang, S. (2014). Frequentist model averaging for multinomial and ordered logit models. *International journal of forecasting*, 30(1):118–128.

Wyss, R., Schneeweiss, S., van der Laan, M., Lendle, S. D., Ju, C., and Franklin, J. M. (2018). Using super learner prediction modeling to improve high-dimensional propensity score estimation. *Epidemiology*, 29(1):96–106.

Xie, Y., Zhu, Y., Cotton, C. A., and Wu, P. (2019). A model averaging approach for estimating propensity scores by optimizing balance. *Statistical Methods in Medical Research*, 28(1):84–101.

Yu, J., Green, M. D., Li, S., Sun, Y., Journey, S. N., Choi, J. E., Rizvi, S. M., Qin, A., Waninger, J. J., Lang, X., Chopra, Z., El Naqa, I., Zhou, J., Bian, Y., Jiang, L., Tezel, A., Skvarce, J., Achar, R. K., Sitto, M., Rosen, B. S., Su, F., Narayanan, S. P., Cao, X., Wei, S., Szeliga, W., Vatan, L., Mayo, C., Morgan, M. A., Schonewolf, C. A., Cuneo, K., Kryczek, I., Ma, V. T., Lao, C. D., Lawrence, T. S., Ramnath, N., Wen, F., Chinnaiyan, A. M., Cieslik, M., Alva, A., and Zou, W. (2021). Liver metastasis restrains immunotherapy efficacy via macrophage-mediated t cell elimination. *Nature Medicine*, 27(1):152–164.

Zhang, X., Zou, G., and J. Carroll, R. (2015). Model averaging based on kullback-leibler distance. *Statistica Sinica*, 25:1583–1598.

Zhao, Q. (2019). Covariate balancing propensity score by tailored loss functions. *The Annals of Statistics*, 47(2):965–993.

Zhao, S., Liao, J., and Yu, D. (2020). Model averaging estimator in ridge regression and its large sample properties. *Statistical Papers (Berlin, Germany)*, 61(4):1719–1739.

Zhao, S., Zhou, J., and Yang, G. (2019). Averaging estimators for discrete choice by m-fold cross-validation. *Economics Letters*, 174:65–69.

Zulj, V. and Jin, S. (2021). Frequentist model averaging with penalization in generalized linear models. Unpublished manuscript. Department of Statistics, Uppsala University.