Econometric Theory, **41**, 2025, 1468–1505.
doi:10.1017/S0266466625100029

MODEL AVERAGING FOR TREATMENT EFFECT ESTIMATION WITH HETEROGENEITY AND HETEROSKEDASTICITY

YUTING WEI
Nanjing University of Information Science and Technology
and
University of Science and Technology of China

GUANGREN YANG
Jinan University

ZHANSHOU CHEN
Qinghai Normal University

XINYU ZHANG
University of Science and Technology of China
and
Academy of Mathematics and Systems Science

The primary focus of this article is to capture heterogeneous treatment effects measured by the conditional average treatment effect. A model averaging estimation scheme is proposed with multiple candidate linear regression models under heteroskedastic errors, and the properties of this scheme are explored analytically. First, it is shown that our proposal is asymptotically optimal in the sense of achieving the lowest possible squared error. Second, the convergence of the weights determined by our proposal is provided when at least one of the candidate models is correctly specified. Simulation results in comparison with several related existing methods favor our proposed method. The method is applied to a dataset from a labor skills training program.

We thank the Editor (Professor Peter Phillips), the Co-Editor (Professor Michael Jansson), and three anonymous referees for constructive comments which led to great improvements of the article. This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFA1008704. Wei acknowledges support from the Startup Foundation for Introducing Talent of NUIST (Grant No. 2024r013) and the Natural Science Foundation of Jiangsu Province (Grant No. BK20240690). Yang's research was supported by the National Social Science Foundation of China (Grant Nos. 24BTJ069 and 24BTJ070) and the National Statistical Scientific Research Center Projects 2024LY009. Chen acknowledges support from National Natural Science Foundation of China (NNSFC) (Grant No. 12161072). Zhang acknowledges support from NNSFC (Grant Nos. 72525001 and 72495124) and Beijing Natural Science Foundation (Z240004). Address correspondence to Xinyu Zhang, SKLMS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China, e-mail: xinyu@amss.ac.cn.

© The Author(s), 2025. Published by Cambridge University Press.

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

1. INTRODUCTION

Evaluation of the causal effect of a treatment on an outcome variable is a fundamental concern across various disciplines, particularly in econometrics. For example, in economic studies, researchers may be interested in the effect that an advertising strategy has on consumer purchase intent or the effect that a job-training program has on employment income. In many such studies, treatment effects often exhibit heterogeneity. For example, the advertising strategy may be more persuasive to teenagers, and the job-training program may be more helpful to a highly educated workforce. In this case, accurately capturing the heterogeneous treatment effects is crucial since it enables researchers to make more effective decisions and thus increase the profit of the program of interest. A commonly used measure of heterogeneous treatment effects is the conditional average treatment effect (CATE) which has received considerable recent attention. In this article, we restrict attention to CATE estimation.

In some studies, the CATE was estimated either by assuming a parametric model for the conditional mean of the potential outcome given the baseline covariates for each treatment group or by fitting a parametric model for the relationship between the observed outcome and the baseline covariates as well as the treatment assignment (see, e.g., Rubin, 1977; Crump et al., 2008; Cai et al., 2011; Bhattacharya and Dupas, 2012; Imai and Ratkovic, 2013; Tian et al., 2014). However, rather than relying on a single assumed parametric model, practitioners often have many candidate models that are postulated based on different interpretations of the observed data. Thus, Rolling and Yang (2014) developed a model selection method, named the treatment effect cross-validation (TECV) method, for choosing the most suitable model for estimating CATE. However, the model selection method may impose some risk in the subsequent analysis since we are "putting all our inferential eggs in one unevenly woven basket" (Longford, 2005).

Model averaging is a smoothed extension of model selection and can substantially reduce risk relative to selection (Hansen, 2014). For further evidence of the gains from model averaging, we refer to Yuan and Yang (2005) and Zhang et al. (2020). Due to these superiorities, model averaging has attracted considerable attention over the past decades (see, e.g., Buckland, Burnham, and Augustin, 1997; Yang, 2001; Hjort and Claeskens, 2003; Hansen, 2007; Wan, Zhang, and Zou, 2010; Liang et al., 2011; Hansen and Racine, 2012; Liu and Okui, 2013; Ullah and Wang, 2013; Cheng and Hansen, 2015; Liu, 2015; Liu, Okui, and Yoshimura, 2016; De Luca, Magnus, and Peracchi, 2018; Fang et al., 2019a; Fang, Li, and Wang, 2019b; Yan et al., 2021). In this article, our goal is to develop a model averaging method to combine CATE estimators in the presence of multiple available candidate models. To the best of our knowledge, there are only two methods available for this issue. The first method, proposed by Rolling, Yang, and Velez (2019), is referred to as the treatment effect estimation by mixing (TEEM) method. The second is the model averaging method of Zhao et al. (2024) developed based on the idea of Mallows model averaging (MMA) as introduced

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

by Hansen (2007). It should be noted that both the TEEM method of Rolling et al. (2019) and the method of Zhao et al. (2024) theoretically rely on pairwise matching of treated units to control units using a covariate probability density support partition to match. When the number of covariates is large or even moderate, it may be hard to find a treatment-control pair in each cell after partitioning.

In this article, under heteroskedastic errors, a model averaging estimation for CATE with multiple candidate linear regression models is developed by borrowing the idea of the jackknife criterion of Hansen and Racine (2012). Our proposed method is called the MATEE-HH method where MA, TEE, the first H and the second H stand for model averaging, treatment effect estimation, heterogeneity and heteroskedasticity, respectively. The theoretical properties of our proposed method are provided. We first present the asymptotic optimality of our proposed method. We next present the convergence of the model weights determined by our proposal when at least one of the candidate models is correctly specified. Notably, this second theoretical property is lacking in the theoretical analyses of Zhao et al. (2024).

The rest of this article is organized as follows. In Section 2, we first describe the model framework and then introduce our proposed model averaging procedure. The theoretical properties of our proposed method are established in Section 3. A Monte Carlo study is conducted in Section 4 to investigate the finite-sample performance of our proposal as well as its comparison with some related approaches. An empirical example is presented in Section 5. Some conclusions are made in Section 6. All the technical details are given in the Appendix.

2. MODEL FRAMEWORK AND ESTIMATION PROCEDURE

2.1. Model Framework

$\{Y_i, \mathrm{X}_i, T_i\}_{i=1}^n$ is a random sample from (Y, X, T), where Y is the scalar response variable, X is the vector of baseline covariates and T is the treatment indicator variable, taking value 1 if the individual is in the treated group, and 0 otherwise. Following the Rubin causal model (Holland, 1986), we postulate the existence of a pair of potential outcomes Y_t and Y_c corresponding to the Y that is observed when the individual's choice is treatment and control. Then, we have

$$Y_i = T_i Y_{t,i} + (1 - T_i) Y_{c,i}, \tag{1}$$

and $Y_{t,i} - Y_{c,i}$ represents the treatment effect for the ith individual, where $Y_{t,i}$ and $Y_{c,i}$ are the realized values of Y_t and Y_c for sampled unit i, respectively. Obviously, $Y_{t,i} - Y_{c,i}$ is infeasible since only one of the potential outcomes $Y_{t,i}$ or $Y_{c,i}$ can be observed depending on the individual's choice of treatment or control. Holland (1986) called this dilemma the "fundamental problem of causal inference." Assume that

$$Y_{t,i} - Y_{c,i} = \mu_i + e_i, \quad E(e_i|\mathrm{X}_i) = 0, \quad Var(e_i|\mathrm{X}_i) = \sigma_i^2. \tag{2}$$

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

Then, μ_i is the CATE for the ith individual, and its estimation is the focus of this article. In other words, we construct a suitable estimator for $\mu = (\mu_1, \mu_2, \ldots, \mu_n)^\top$.

We consider a sequence of linear regression models. Specifically, under the mth model, we have

$$\mu_i = \sum_{j=1}^{k_m} X_{(m),ij}\theta_{(m),j}, \tag{3}$$

where $X_{(m),ij}$ is the jth element of $\mathrm{X}_{(m),i}$, $\mathrm{X}_{(m)}$ is a row vector comprised of k_m different elements of X, and $\theta_{(m),j}$ is the corresponding regression coefficient. For T, the treatment assignment, we assume that

$$P(T = 1|Y_t, Y_c, \mathrm{X}) = P(T = 1|\mathrm{X}) := \pi(\mathrm{X}), \tag{4}$$

where $\pi(\mathrm{X})$ is referred to as the propensity score function in the spirit of Hahn (1998). This assumption, commonly adopted in the causal effect literature, is known as unconfounded assignment. The goal of this article is to find an asymptotically optimal model average estimator for the conditional average treatment effects μ.

2.2. Estimation Procedure

We first consider the case where the propensity score function $\pi(\mathrm{X})$ is known and postpone the discussion of the estimation of $\pi(\mathrm{X})$ at the end of this section. Assume that there exists a positive constant c_π such that

$$0 < c_\pi \le \pi(\mathrm{X}) \le 1 - c_\pi < 1, \tag{5}$$

with probability one. Then, we know that $\pi(\mathrm{X}_i)$ and $1-\pi(\mathrm{X}_i)$ are invertible with probability one; thus, we define $Z_{\pi,i} = \pi(\mathrm{X}_i)^{-1}T_iY_i - \{1 - \pi(\mathrm{X}_i)\}^{-1}(1 - T_i)Y_i$. Then, according to (1), (2), and (4), we have

$$\begin{aligned} E(Z_{\pi,i}|\mathrm{X}_i) &= E\left[\pi(\mathrm{X}_i)^{-1}T_iY_i - \{1-\pi(\mathrm{X}_i)\}^{-1}(1-T_i)Y_i|\mathrm{X}_i\right] \\ &= \pi(\mathrm{X}_i)^{-1}E(T_i|\mathrm{X}_i)E(Y_{t,i}|\mathrm{X}_i) - \{1-\pi(\mathrm{X}_i)\}^{-1}E(1-T_i|\mathrm{X}_i)E(Y_{c,i}|\mathrm{X}_i) \\ &= \mu_i. \end{aligned}$$

Thus, letting $e_{\pi,i} = Z_{\pi,i} - \mu_i$, we can obtain that

$$Z_{\pi,i} = \mu_i + e_{\pi,i}, \quad E(e_{\pi,i}|\mathrm{X}_i) = 0 \quad \text{and} \quad Var(e_{\pi,i}|\mathrm{X}_i) = \sigma^2_{\pi,i}. \tag{6}$$

As a result, under the mth model, we have

$$Z_{\pi,i} = \sum_{j=1}^{k_m} \mathrm{X}_{(m),ij}\theta_{(m),j} + e_{\pi,i}. \tag{7}$$

Note that $\{Z_{\pi,i}, \mathrm{X}_i\}_{i=1}^n$ is fully observed when $\pi(\mathrm{X})$ is known. This together with the heteroskedasticity in (6) and the linearity in (7) indicates that μ can be estimated by applying the jackknife model averaging (JMA) method of Hansen and Racine

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

(2012) to $\{Z_{\pi,i}, \mathrm{X}_i\}_{i=1}^n$ since the JMA approach is a model averaging method developed for linear regression models with heteroskedastic errors where data are complete.

In what follows, we present the details of applying the JMA method to obtain a model average estimator of μ. Let $\mathrm{Z}_\pi = (Z_{\pi,1}, Z_{\pi,2}, \ldots, Z_{\pi,n})^\top$, $\mathbf{X}_{(m)}$ be an $n \times k_m$ matrix with the ijth element being $X_{(m),ij}$ and $\mathrm{e}_\pi = (e_{\pi,1}, e_{\pi,2}, \ldots, e_{\pi,n})^\top$. Then, the matrix form of model (7) is

$$\mathrm{Z}_\pi = \mathbf{X}_{(m)}\theta_{(m)} + \mathrm{e}_\pi,$$

where $\theta_{(m)}$ is the vector of regression coefficients. Assume that $\mathbf{X}_{(m)}$ is of full column rank. Then, based on model (7), the least squares estimator of $\theta_{(m)}$ is $\hat{\theta}_{(m),\pi} = (\mathbf{X}_{(m)}^\top \mathbf{X}_{(m)})^{-1}\mathbf{X}_{(m)}^\top \mathrm{Z}_\pi$. The corresponding estimator of μ is

$$\hat{\mu}_{(m),\pi} = \mathbf{X}_{(m)}\hat{\theta}_{(m),\pi} = \mathbf{P}_{(m)}\mathrm{Z}_\pi, \tag{8}$$

where $\mathbf{P}_{(m)} = \mathbf{X}_{(m)}(\mathbf{X}_{(m)}^\top \mathbf{X}_{(m)})^{-1}\mathbf{X}_{(m)}^\top$. Then, the corresponding model average estimator of μ is

$$\hat{\mu}_\pi(\omega) = \sum_{m=1}^{M_n} \omega_m \hat{\mu}_{(m),\pi} = \sum_{m=1}^{M_n} \omega_m \mathbf{P}_{(m)}\mathrm{Z}_\pi = \mathbf{P}(\omega)\mathrm{Z}_\pi, \tag{9}$$

where M_n is the number of candidate models, $\mathbf{P}(\omega) = \sum_{m=1}^{M_n} \omega_m \mathbf{P}_{(m)}$ and $\omega = (\omega_1, \omega_2, \ldots, \omega_{M_n})^\top$ is a weight vector in $\mathscr{H}_n = \left\{\omega \in [0,1]^{M_n} : \sum_{m=1}^{M_n} \omega_m = 1\right\}$. We consider the jackknife criterion for choosing ω. Specifically, let $\tilde{\mu}_{(m),\pi}$ be the estimator of μ based on (7) when leave-one-out cross-validation is applied. Then, careful calculation shows that $\tilde{\mu}_{(m),\pi} = \tilde{\mathbf{P}}_{(m)}\mathrm{Z}_\pi$, where $\tilde{\mathbf{P}}_{(m)} = \mathbf{D}_{(m)}(\mathbf{P}_{(m)} - \mathbf{I}_n) + \mathbf{I}_n$, $\mathbf{I}_n$ is an $n \times n$ identity matrix, $\mathbf{D}_{(m)}$ is an $n \times n$ diagonal matrix with the ith diagonal element $(1 - h_{m,ii})^{-1}$, and $h_{m,ii}$ is the ith diagonal element of $\mathbf{P}_{(m)}$. This leads to the following weight choice criterion

$$\mathrm{CV}_\pi(\omega) = \|\mathrm{Z}_\pi - \tilde{\mu}_\pi(\omega)\|^2, \tag{10}$$

where

$$\tilde{\mu}_\pi(\omega) = \sum_{m=1}^{M_n} \omega_m \tilde{\mu}_{(m),\pi} = \tilde{\mathbf{P}}(\omega)\mathrm{Z}_\pi \quad \text{and} \quad \tilde{\mathbf{P}}(\omega) = \sum_{m=1}^{M_n} \omega_m \tilde{\mathbf{P}}_{(m)}. \tag{11}$$

Finally, we discuss the estimation of the propensity score function $\pi(\mathrm{X})$. We assume a parametric model $\tilde{\pi}(\mathrm{X};\alpha)$ for $\pi(\mathrm{X})$ with α being the unknown model parameter vector. Denote $\hat{\alpha}_n$ as an estimator of α chosen by the user. In what follows, a Greek letter subscripted by $\hat{\pi}$ indicates that it is obtained by replacing $\{\pi(\mathrm{X}_i)\}_{i=1}^n$ in its corresponding estimator with $\{\tilde{\pi}(\mathrm{X}_i;\hat{\alpha}_n)\}_{i=1}^n$, respectively. For example, $\mathrm{Z}_{\hat{\pi}}$ is obtained by replacing $\{\pi(\mathrm{X}_i)\}_{i=1}^n$ in Z_π with $\{\tilde{\pi}(\mathrm{X}_i;\hat{\alpha}_n)\}_{i=1}^n$. With $\{\tilde{\pi}(\mathrm{X}_i;\hat{\alpha}_n)\}_{i=1}^n$, $\mathrm{CV}_\pi(\omega)$ becomes

$$\mathrm{CV}_{\hat{\pi}}(\omega) = \|\mathrm{Z}_{\hat{\pi}} - \tilde{\mu}_{\hat{\pi}}(\omega)\|^2. \tag{12}$$

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

Let $\hat{\omega}_{cv}$ be the minimizer of $\mathrm{CV}_{\hat{\pi}}(\omega)$ among $\mathscr{H}_n$; that is,

$$\hat{\omega}_{cv} = \arg\min_{\omega \in \mathscr{H}_n} \mathrm{CV}_{\hat{\pi}}(\omega). \tag{13}$$

Then, the corresponding model average estimator of μ is $\hat{\mu}_{\hat{\pi}}(\hat{\omega}_{cv})$. Let $\mathbf{B}_{\hat{\pi}} = (\mathrm{Z}_{\hat{\pi}} - \tilde{\mathbf{P}}_{(1)}\mathrm{Z}_{\hat{\pi}}, \ldots, \mathrm{Z}_{\hat{\pi}} - \tilde{\mathbf{P}}_{(M_n)}\mathrm{Z}_{\hat{\pi}})$; then, we have $\mathrm{CV}_{\hat{\pi}}(\omega) = \omega^\top \mathbf{B}_{\hat{\pi}}^\top \mathbf{B}_{\hat{\pi}} \omega$, a quadratic function of ω. Thus, $\hat{\omega}_{cv}$ can be efficiently computed by using numerous software packages (e.g., the quadprog package in R and the quadprog function in MATLAB).

3. THEORETICAL PROPERTIES

In this section, we present the theoretical results of this article. Specifically, we first demonstrate the asymptotic optimality of $\hat{\mu}_{\hat{\pi}}(\hat{\omega}_{cv})$. We then present the convergence of the model weights $\hat{\omega}_{cv}$ when at least one of the candidate models is correctly specified. It is important to note that, in this article, the mth candidate model (3) is considered correctly specified, or a correct model, if there exists $\theta^*_{(m)}$ such that $\mu = \mathbf{X}_{(m)}\theta^*_{(m)}$; otherwise, model (3) is considered misspecified, or an incorrect model.

3.1. Asymptotic Optimality

In this subsection, we demonstrate the asymptotic optimality of $\hat{\mu}_{\hat{\pi}}(\hat{\omega}_{cv})$. For ease of expression, we first introduce some notations. Let $\mu_{t,i} = E(Y_{t,i}|\mathrm{X}_i)$ and $e_{t,i} = Y_{t,i} - \mu_{t,i}$. Assume that, conditional on $\{\mathrm{X}_i\}_{i=1}^n$, the random errors $\{e_{t,i}\}_{i=1}^n$ are heteroskedastic, and let $\sigma^2_{t,i} = E(e^2_{t,i}|\mathrm{X}_i)$. Define the loss function and the risk function of $\hat{\mu}_\pi(\omega)$ in (9) as

$$L_\pi(\omega) = \|\mu - \hat{\mu}_\pi(\omega)\|^2 \quad \text{and} \quad R_\pi(\omega) = E\{L_\pi(\omega)|\mathbf{X}\}, \tag{14}$$

respectively, where $\|\cdot\|$ is the Euclidean norm and $\mathbf{X} = (\mathrm{X}_1^\top, \mathrm{X}_2^\top, \ldots, \mathrm{X}_n^\top)^\top$. Let $\xi_\pi = \inf_{\omega \in \mathscr{H}_n} R_\pi(\omega)$, $\bar{h} = \max_{1 \le m \le M_n} \max_{1 \le i \le n} h_{m,ii}$ and $\bar{k} = \max_{1 \le m \le M_n} k_m$. The asymptotic optimality of $\hat{\mu}_{\hat{\pi}}(\hat{\omega}_{cv})$ requires the following conditions where all the limiting processes discussed here and throughout the article are with respect to $n \to \infty$.

(C1) Assume that the propensity score function $E(T|\mathrm{X})$ is specified as the parametric model $\tilde{\pi}(\mathrm{X};\alpha)$. $\sqrt{n}\|\hat{\alpha}_n - \alpha_0\| = O_p(1)$ where α_0 is the true value of α. In addition, the first derivative of $\tilde{\pi}(\mathrm{X};\alpha)$ with respect to α is continuous and bounded.

(C2) $\max_{1 \le i \le n} \max\left\{E(e_i^4 \mid \mathrm{X}_i), E(e_{t,i}^4 \mid \mathrm{X}_i), |\mu_i|, |\mu_{t,i}|\right\}$ is bounded above by a constant $\bar{C}$ almost surely. Besides, $\min_{1 \le i \le n} \sigma^2_{t,i} \ge c_t$ almost surely where c_t is a positive constant.

(C3) $M_n^2 \xi_\pi^{-1} \to 0$ almost surely.

(C4) $\bar{h} \to 0$ and $\bar{h}\bar{k}\xi_\pi^{-1} \to 0$ almost surely.

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

THEOREM 1. *If Conditions (C1)-(C4) are satisfied, then*

$$\frac{L_{\hat{\pi}}(\hat{\omega}_{cv})}{\inf_{\omega\in\mathscr{H}_n} L_{\hat{\pi}}(\omega)} \xrightarrow{p} 1. \tag{15}$$

This theorem states that selected weight vector $\hat{\omega}_{cv}$ yields a squared error that is asymptotically identical to that of the infeasible optimal weight vector. This implies the asymptotic optimality of $\hat{\mu}_{\hat{\pi}}(\hat{\omega}_{cv})$.

The specification of an appropriate form for the parametric model $\tilde{\pi}(\mathrm{X};\alpha)$ requires much scientific contextual knowledge. In this article, we consider the logistic regression model for $\tilde{\pi}(\mathrm{X};\alpha)$, since it is the most widely used model for relating probabilities to the covariates (Claeskens and Hjort, 2008). Instead of the logistic regression model, one may consider a probit model for $\tilde{\pi}(\mathrm{X};\alpha)$ since the probit model is also a classical model for modeling binomial probabilities (Hastie, Tibshirani, and Friedman, 2009). The probit model takes the form $\tilde{\pi}(\mathrm{X};\alpha) = \Phi(\mathrm{X}^\top\alpha)$ where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. The second part of Condition (C1) requires $\hat{\alpha}_n$ to be $\sqrt{n}-$consistent. This is a reasonable restriction since, according to White (1982), the commonly used maximum likelihood estimator (MLE) satisfies this restriction under certain regularity conditions. Instead of MLE, when the number of covariates included in $\tilde{\pi}(\mathrm{X};\alpha)$ increases with the sample size n, the penalized likelihood estimator (PLE) of α, whose properties are investigated in Xie, Yan, and Tang (2021), may be preferred. From Theorem S1.1 in the supplementary material of Xie et al. (2021), it follows that this PLE also satisfies the first part of (C1) under certain conditions. The last part of Condition (C1) imposes some restrictions on $\partial\tilde{\pi}(\mathrm{X};\alpha)/\partial\alpha$ and is satisfied if $\tilde{\pi}(\mathrm{X};\alpha)$ is continuous and the support sets of (X,α) are bounded and closed. In addition, this condition is implied by Assumption 1 of Xie et al. (2021).

Condition (C2) places a bound on the conditional moments $\{E(e_i^4 \mid \mathrm{X}_i), E(e_{t,i}^4 \mid \mathrm{X}_i)\}_{i=1}^n$. Such a placement is common in model averaging literature such as Liu and Okui (2013); Ando and Li (2014) and Fang et al. (2019a). In addition, it also imposes a restriction on the conditional means $\{\mu_i, \mu_{t,i}\}_{i=1}^n$, and this restriction is satisfied if both $E(Y_t|\mathrm{X})$ and $E(Y_c|\mathrm{X})$ are bounded. The last part of Condition (C2) is similar to Liu and Okui (2013, Assumption 2.1) which places a lower bound on $\{\sigma_{t,i}^2\}_{i=1}^n$.

Apparently, under Condition (C3), the number of candidate models M_n is allowed to be fixed or go to infinity. It is clear that a necessary condition for (C3) is

$$\xi_\pi \to \infty, \quad \text{a.s.} \tag{16}$$

As noted by Hansen (2007), (16) requires that there is no finite dimensional correct model in the class of candidate models. To clarify this, assume that the mth model

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

is a finite dimensional correct model. Then, by the definition of ξ_π and (14), we have that

$$\xi_\pi \le E\left\{\|\mu - \hat{\mu}_{(m),\pi}\|^2 | \mathbf{X}\right\} = E\left\{\|\mu - \mathbf{P}_{(m)}\mu - \mathbf{P}_{(m)}e_\pi\|^2 | \mathbf{X}\right\} = \mathrm{tr}\{\mathbf{P}_{(m)}\mathbf{\Omega}_\pi \mathbf{P}_{(m)}^\top\}, \tag{17}$$

where the last equality is due to the definition of the correct model presented at the beginning of Section 3 and $\mathbf{\Omega}_\pi$ is an $n \times n$ diagonal matrix with ith diagonal element $\sigma^2_{\pi,i}$. For any matrix $\mathbf{A}$, denote $\lambda_{\max}\{\mathbf{A}\}$ as the largest singular value of $\mathbf{A}$. Then, by Lemma 3 and Lemma 6 presented in Appendix A.1, we have

$$\mathrm{tr}\{\mathbf{P}_{(m)}\mathbf{\Omega}_\pi \mathbf{P}_{(m)}^\top\} \le \lambda_{\max}\{\mathbf{\Omega}_\pi\}\mathrm{tr}\{\mathbf{P}_{(m)}\} \le \max_{1\le i\le n} \sigma^2_{\pi,i} k_m \le c^2_{\pi,2} k_m < \infty,$$

almost surely where $c_{\pi,2}$ is a positive constant given in Lemma 3. This together with (17) implies that Condition (C3) excludes the case where there is a finite dimensional correct model in the class of candidate models. As a result, in the next subsection, we analyze the theoretical property of the model weights $\hat{\omega}_{cv}$ when correct models are present in the class of candidate models. What is worth mentioning is that a related condition, used in Zhang, Wan, and Zou (2013) for proving the asymptotic optimality of the JMA estimator, is

$$M_n \xi_\pi^{-2G} \sum_{m=1}^{M_n} R_\pi^G(\omega_m^0) \to 0, \quad \text{a.s.}, \tag{18}$$

for some fixed positive integer G, where ω_m^0 is an $M_n \times 1$ vector whose mth element is one while the others are zeros. Our Condition (C2) corresponds to (18) with $G = 1$. When $G = 1$, (18) is clearly stronger than Condition (C3). In addition, Condition (C3) is simpler and more easily interpretable than (18).

The first part of Condition (C4) is the same as Hansen and Racine (2012, Condition (A.9)). As noted by Hansen and Racine (2012), this condition actually requires that the diagonal element $h_{m,ii}$ is asymptotically negligible and excludes extremely unbalanced design matrices as candidate linear regression models. A commonly used restriction imposed on $h_{m,ii}$ typical for the studies of cross-validation is that for any $m \in \{1, \ldots, M_n\}$,

$$h_{m,ii} \le \Lambda k_m n^{-1}, \quad a.s., \tag{19}$$

for some positive constant Λ. See, for example, Li (1987); Andrews (1991) and Fang et al. (2019a). Clearly, neither the first part of (C4) nor (19) is stronger than the other. It is clear that under (19), the first part of Condition (C4) is satisfied if $\bar{k}n^{-1} = o(1)$. In addition, we know that under (19) and Condition (C3), the second part of Condition (C4) is satisfied if

$$\bar{k}^2 n^{-1} M_n^{-2} = O(1). \tag{20}$$

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

It is worth mentioning that (20) is weaker than some related conditions used in the literature on model averaging. See, for example, Condition (12) in Wan et al. (2010) and Assumption 9 of Liu et al. (2016).

Recently, Wei and Wang (2021) have developed a model average method for linear models with response missing at random. Our proposal is similar to that of Wei and Wang (2021): both construct a fully observed dataset with the use of the propensity score function and then apply the JMA approach to this dataset. However, compared with Wei and Wang (2021), this article differs in the following three aspects. First, this article is concerned with causal inference while Wei and Wang (2021) focused on missing data. The "fundamental problem of causal inference" gives rise to new challenges especially in the construction of Z_π and the establishment of the theoretical properties. Second, the conditions required for the asymptotic optimality of our proposal are different from that required in Wei and Wang (2021) since we adopt different proof techniques. A key condition related to ξ_π required in Wei and Wang (2021) is

$$\xi_\pi^{-2}\sum_{m=1}^{M_n} R_\pi(\omega_m^0) \to 0, \quad \text{a.s.} \tag{21}$$

Note that $\xi_\pi^{-2}\sum_{m=1}^{M_n} R_\pi(\omega_m^0) \geq \xi_\pi^{-1} M_n$. Therefore, when M_n is fixed, our Condition (C3) is weaker than (21). In addition, Condition (C3) is simpler and more easily interpretable. Another key condition required in Wei and Wang (2021) is (19) which has already been explained and compared in the previous paragraph. Third, this article investigates the convergence of the model weights (see the next subsection) while Wei and Wang (2021) did not. In general, this article is by no means a straightforward or a routine extension of Wei and Wang (2021)'s work.

3.2. Weight Convergence

In this subsection, we consider the case where at least one of the candidate models is correctly specified and present the convergence of the model weights $\hat{\omega}_{cv}$ in this scenario. We first show that the sum of the weights obtained by our proposal for the correct candidate models converges to one in probability as the sample size goes to infinity. To demonstrate such a theoretical result, we first introduce some notations. Denote m_0 as the number of candidate models that are correctly specified. Without loss of generality, assume that these correct candidate models are the first m_0 candidate models. Let $\hat{s}_{cv} = \sum_{m=1}^{m_0} \hat{\omega}_{cv,m}$, where $\hat{\omega}_{cv,m}$ is the mth element of $\hat{\omega}_{cv}$. Then, $\hat{s}_{cv}$ is the sum of the weights obtained by our proposal for the correct candidate models. Let $\mathscr{H}_F = \{\omega \in [0,1]^{M_n} : \sum_{m=m_0+1}^{M_n} \omega_m = 1\}$ and $\xi_{\pi,F} = \inf_{\omega\in\mathscr{H}_F} R_\pi(\omega)$. Assume the following conditions.

(C5) $\xi_{\pi,F}^{-1}\max\left\{\sqrt{n},\ (M_n - m_0)^2,\ m_0\bar{k}\right\} \to 0$ almost surely.
(C6) $\bar{h} = O(n^{-1/2})$ almost surely.

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

THEOREM 2. *If Conditions (C1)-(C2) and (C5)-(C6) are satisfied, then* $\hat{s}_{cv} \to 1$ *in probability.*

Theorem 2 indicates that our proposed model average estimator $\hat{\mu}_{\hat{\pi}}(\hat{\omega}_{cv})$ will automatically exclude the incorrect models and be determined only by the correct models when the sample size is sufficiently large. This, to some extent, can be interpreted as a form of model selection consistency. $\xi_{\pi,F}$ is the optimal risk when the weights are assigned to all the incorrect models. Condition (C5) requires that this optimal risk diverges to infinity at a faster rate than that of $\sqrt{n}$, $(M_n - m_0)^2$ and $m_0\bar{k}$. If $\sqrt{n}$ is larger than $(M_n - m_0)^2$ and $m_0\bar{k}$, then (C5) is almost surely equivalent to $\xi_{\pi,F}^{-1}\sqrt{n} \to 0$, and thus is similar to Fang, Li, and Xia (2022, Condition (C6)) and implied by Ando and Li (2014, Conditions (7) and (8)). If $(M_n - m_0)^2$ is larger than $\sqrt{n}$ and $m_0\bar{k}$, then (C5) is almost surely equivalent to $\xi_{\pi,F}^{-1}(M_n - m_0)^2 \to 0$, and thus is similar to Condition (C3). Notably, Condition (C5) allows both the number of correct models m_0 and the number of incorrect models $M_n - m_0$ to be fixed or go to infinity. Condition (C6) is the same as Liu and Okui (2013, Assump. 2.4). In addition, it can be verified that, under (19), (C6) is satisfied if $\bar{k}^2 n^{-1} = O(1)$. As discussed by Liu and Okui (2013), such a condition excludes peculiar models, such as a model that contains a dummy variable on some single observation.

If a correct candidate model has the smallest dimension among all the correct candidate models, then it is referred to as the smallest correct model. Otherwise, it is called over-fitted. If a candidate model is misspecified, then it is called under-fitted. Theorem 2 implies that the weight selected by (13) for an under-fitted model converges to zero in probability as the sample size goes to infinity. It may not be true for over-fitted models (see, for example, Zhang and Liu, 2019, Thm. 4, with linear regression). However, in the special case where the candidate models are nested, by modifying (12) with a proper penalty term, we can show that the resulting model averaging estimator asymptotically assigns zero weight to both under-fitted and over-fitted models. Concretely, we modify (12) to

$$\mathrm{CV}_{\hat{\pi},\phi_n}(\omega) = \|\mathrm{Z}_{\hat{\pi}} - \tilde{\mu}_{\hat{\pi}}(\omega)\|^2 + \phi_n \omega^\top \mathrm{K}, \tag{22}$$

where ϕ_n is a tuning parameter and $\mathrm{K} = (k_1, k_2, \ldots, k_{M_n})^\top$. Let $\breve{\omega}_{\phi_n}$ be the minimizer of $\mathrm{CV}_{\hat{\pi},\phi_n}(\omega)$ among $\mathscr{H}_n$; that is,

$$\breve{\omega}_{\phi_n} = \arg\min_{\omega \in \mathscr{H}_n} \mathrm{CV}_{\hat{\pi},\phi_n}(\omega). \tag{23}$$

Suppose we have a sequence of nested candidate models which is commonly seen in the literature of model averaging. See, for example, Hansen (2007); Hansen (2014); Zhang and Liu (2019). Assume that the number of candidate models M_n is fixed. Denote m_o as the label of the smallest correct model. Then we have the mth model is under-fitted if $m < m_o$ and over-fitted if $m > m_o$. Assume the following conditions.

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

(C7) $n^{-1}\mathbf{X}^\top\mathbf{X} \to E(\mathrm{X}_i^\top \mathrm{X}_i)$ in probability and $E(\mathrm{X}_i^\top \mathrm{X}_i)$ is a positive definite matrix.
(C8) $\|\mathbf{X}^\top \mathrm{e}_\pi\| = O_p(\sqrt{n})$.
(C9) $\bar{h} = O(n^{-1})$.

THEOREM 3. *If Conditions (C1)–(C2) and (C7)–(C9) are satisfied, then*

(i) *for any* $m \in \{1, 2, \ldots, m_o - 1\}$, $\breve{\omega}_{\phi_n, m} = O_p(\phi_n n^{-1})$;
(ii) *for any* $m \in \{m_o + 1, m_o + 2, \ldots, M_n\}$, $\breve{\omega}_{\phi_n, m} = O_p(\phi_n^{-1})$,

where $\breve{\omega}_{\phi_n, m}$ *is the* m*th element of* $\breve{\omega}_{\phi_n}$.

Theorem 3 shows that, under some conditions, the modified model averaging estimator asymptotically assigns zero weight to both under-fitted and over-fitted models when both $n/\phi_n \to \infty$ and $\phi_n \to \infty$. One can take $\phi_n = \log(n)$ which corresponds to the penalty term in the Bayesian information criterion (BIC; Schwarz, 1978). If X_i is one-dimensional and $E(\mathrm{X}_i^2) < \infty$, then Condition (C7) follows from the weak law of large numbers. According to Liu (2015), for multidimensional covariate vector X_i, Condition (C7) holds if $Y_{t,i} - Y_{c,i}$ is a stationary and ergodic martingale difference sequence with finite fourth moments. It can be verified that, under Condition (C2), a sufficient condition for Condition (C8) is $E(\mathrm{X}_i \mathrm{X}_i^\top) < \infty$. Clearly, such a sufficient condition holds for a multivariate-normal vector X_i. Condition (C9) is quite mild and is equivalent to (19) in the model framework of this article.

4. A MONTE CARLO STUDY

In this section, a Monte Carlo study is conducted to investigate the finite-sample performance of our proposed method. Recall that whether Condition (C3) holds is related to whether there exists a correct model in the class of candidate models, and the convergence of $\hat{\omega}_{cv}$ is established when at least one of the candidate models is correctly specified. We consider two designs in our Monte Carlo study. Specifically, in the first design, the candidate models are all misspecified, while in the second design, correct models exist in the class of candidate models. For a better analysis, we also consider several related existing methods as our proposal's competitors. Specifically, the first competitor is a simple rule that assigns equal weights to each candidate model, referred to as the equal weighting (EW). The second competitor is the model average approach proposed by Zhao et al. (2024), referred to as the Zhao et al. (2024)'s model average (ZMA) method. The rest are Akaike information criterion (AIC; Akaike, 1974), BIC, the smoothed AIC (SAIC) and smoothed BIC (SBIC) of Buckland et al. (1997), TECV of Rolling and Yang (2014) and the TEEM algorithm of Rolling et al. (2019). The details and results of the Monte Carlo study are given below.

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

Design 1 (without correct candidate models) The data-generating process is

$$Y_{t,i} = -1.2\sum_{j=1}^{J} j^{-1}X_{ij} + 0.5X_{i2}^2 + 0.5X_{i3}^2 + e_{t,i} \quad \text{and}$$

$$Y_{c,i} = -\sum_{j=1}^{J} j^{-1}X_{ij} + 0.5X_{i2}^2 + e_{c,i},$$

where $X_{i1} = 1$, $\{X_{ij}\}_{j=2}^{J}$ are independent standard normal random variables, $\{e_{t,i}, e_{c,i}\}$ are independent random errors sampled from $N(0, \sigma^2 X_{i2}^2)$ and $J = 10^3$. This leads to

$$\mu_i = -0.2\sum_{j=1}^{J} j^{-1}X_{ij} + 0.5X_{i3}^2. \tag{24}$$

The parameter σ^2 is determined by the population R^2, which is defined as $R^2 = var(\mu_i)/var(Y_{t,i} - Y_{c,i})$. σ^2 is chosen to let R^2 vary on a grid between 0.1 and 0.9. The propensity score function is taken as

$$\pi(\mathrm{X}) = \frac{\exp(0.5X_{i2} + 0.5X_{i3})}{1 + \exp(0.5X_{i2} + 0.5X_{i3})}.$$

Based on $\pi(\mathrm{X})$, we can obtain $\{T_i\}_{i=1}^{n}$ and thus $\{Y_i\}_{i=1}^{n}$. The candidate models are strictly nested with the mth linear regression model using the first m covariates in $\{X_{ij}\}_{j=1}^{J}$. By (24) and the definition of correct models given at the beginning of Section 3, one can know that the considered candidate models are all misspecified. M_n is set to be the nearest integer to $1.5n^{1/3}$. The sample size is considered to be $n = 100,\ 200$, and 400, so that $M_n = 6,\ 8$, and 11, respectively. The model $\tilde{\pi}(\mathrm{X};\alpha)$ used for our proposal is taken to be $\tilde{\pi}(\mathrm{X};\alpha) = \exp(\sum_{j=1}^{\bar{k}} \alpha_j X_j)/(1 + \sum_{j=1}^{\bar{k}} \alpha_j X_j)$, where X_j and α_j are the jth elements of X and α, respectively. Noting that, in the current design, $\bar{k}$ is equal to M_n and thus increases with n, we consider the PLE of α. Following Hansen (2007), we use the following MSE as the measure of performance of each method, which is given by

$$\frac{1}{nD}\sum_{d=1}^{D} \|\{\hat{\mu}(\omega)\}^{(d)} - \mu^{(d)}\|^2, \tag{25}$$

where $\mu^{(d)}$ is the CATEs in the dth trial, $\{\hat{\mu}(\omega)\}^{(d)}$ is the model average estimator of $\mu^{(d)}$ and D is the number of simulation trials. We report the ratios of the MSEs computed with the MSE of the EW as the denominator. The number of simulation trials is 1,000.

Figure 1 plots the MSE ratio curves of the ten considered estimators of the CATEs μ. In this figure, it is clear that our proposed MATEE-HH method achieves a lower MSE ratio than its competitors in the majority of the combinations of

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

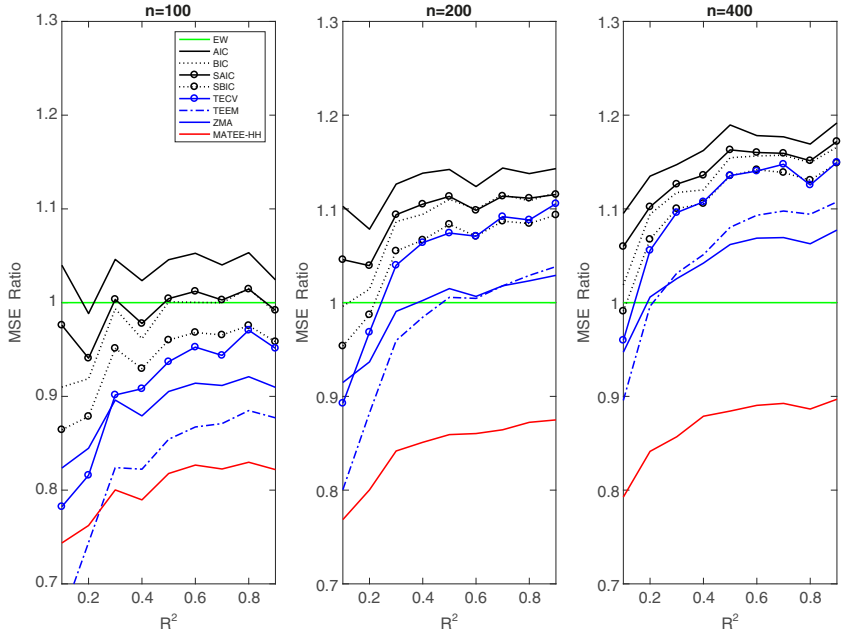


FIGURE 1. The MSE ratio curves of different methods under Design 1.

sample sizes and the population R^2 considered. This, to some extent, demonstrates the superiority of our proposal over its competitors.

Finally, we assess the asymptotic optimality of $\hat{\mu}_{\hat{\pi}}(\hat{\omega}_{cv})$ via numerical inspection. Specifically, we calculate the means of $\mathrm{LR} = L_{\hat{\pi}}(\hat{\omega}_{cv}) / \inf_{\omega \in \mathscr{H}_n} L_{\hat{\pi}}(\omega)$ based on 1,000 replications under different sample sizes with $R^2 = 0.5$. The mean curves are displayed in Figure 2.

Figure 2 shows that the mean LR decreases and approaches one as the sample size n increases. This confirms the asymptotic optimality of our proposal presented in Theorem 1.

Design 2 (with correct candidate models) The data-generating process is

$$Y_{t,i} = X_{i2} - 0.6X_{i4} - 0.5X_{i5} + 0.8X_{i2}^2 + e_{t,i} \quad \text{and}$$
$$Y_{c,i} = X_{i2} - 0.6X_{i3} - 0.5X_{i5} + 0.8X_{i2}^2 + e_{c,i},$$

where $\{X_{ij}\}_{j=2}^5$ are independent standard normal random variables, $\{e_{t,i}, e_{c,i}\}$ are independent random errors sampled from $N(0, \sigma^2 X_{i4}^2)$. This leads to

$$\mu_i = 0.6X_{i3} - 0.6X_{i4}. \tag{26}$$

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

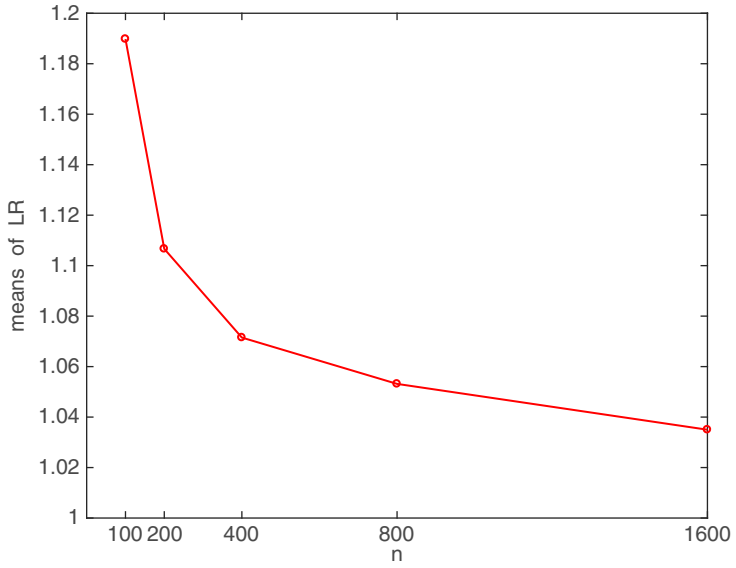


FIGURE 2. Assessing the asymptotic optimality of $\hat{\mu}_{\hat{\pi}}(\hat{\omega}_{cv})$ under Design 1.

σ^2 is also chosen to let R^2 vary on a grid between 0.1 and 0.9, where R^2 is defined in the same way as that in Design 1. The propensity score function is taken as

$$\pi(\mathrm{X}) = \frac{\exp(0.75X_{i2})}{1+\exp(0.75X_{i2})}.$$

Based on $\pi(\mathrm{X})$, we can obtain $\{T_i\}_{i=1}^n$ and thus $\{Y_i\}_{i=1}^n$. All the candidate models include the intercept term and are constructed by varying combinations of $\{X_2, X_3, X_4, X_5\}$. This leads to $M_n = 2^4$. Based on (26) and the definition of correct models given at the beginning of Section 3, it follows that the considered candidate models including X_3 and X_4 are correctly specified. The sample size is also considered to be $n = 100$, 200, and 400. The model $\tilde{\pi}(\mathrm{X};\alpha)$ is taken to be $\tilde{\pi}(\mathrm{X};\alpha) = \exp(\sum_{j=1}^5 \alpha_j X_j)/(1+\sum_{j=1}^5 \alpha_j X_j)$ where $X_1 = 1$. Noting that the number of covariates included in $\tilde{\pi}(\mathrm{X};\alpha)$ is fixed, we consider the MLE of α. The number of simulation trials is 1,000. The results are displayed in Figure 3.

It is observed that our proposal achieves a lower MSE ratio than its competitors except for the TEEM algorithm and the ZMA method for almost all the combinations of sample sizes and the population R^2 considered. Additionally, although our proposal performs slightly worse than the TEEM algorithm and the ZMA method when $n = 100$, it performs better than these two methods when $n = 200$ or 400. Moreover, the superiority of our proposal over these two methods becomes more prominent as the sample size increases.

Finally, we investigate the convergence of the model weights $\hat{\omega}_{cv}$ via numerical inspection. We report the means of $\hat{s}_{cv}$, the sum of the weights corresponding to

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

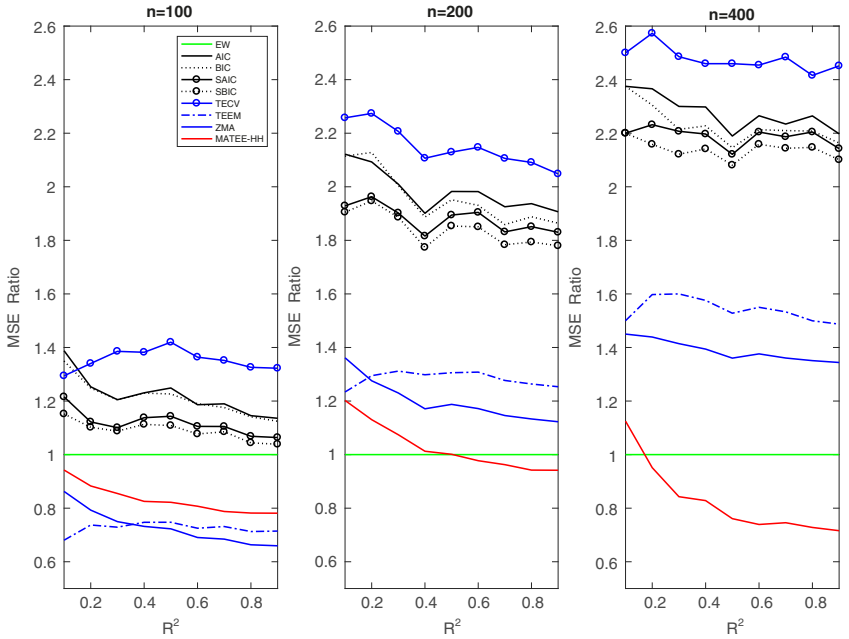


FIGURE 3. The MSE ratio curves of different methods under Design 2.

correct candidate models, based on 1,000 replications under different sample sizes with $R^2 = 0.5$. The mean curves are displayed in Figure 4.

Figure 4 shows that the mean of $\hat{s}_{cv}$ increases and approaches one as the sample size n increases. This confirms the convergence of $\hat{\omega}_{cv}$ presented in Theorem 2.

5. AN EMPIRICAL EXAMPLE

In this section, we apply our proposal to the well-known National Supported Work (NSW) Demonstration dataset which was collected by Lalonde (1986) and analyzed in Rolling et al. (2019). This dataset is available at http://users.nber.org/~rdehejia/nswdata2.html. In this example, we are interested in estimating the heterogeneity of the treatment effect. The outcome variable Y is taken as the earnings in 1978 divided by 10^3. The baseline covariates are age, years of education, indicator for black, indicator for Hispanic, marital status, indicator for no degree and earnings in 1975 divided by 10^3. This dataset contains $n = 722$ sample points, with 297 participants in the training program and 425 non-participants. We take the group of individuals who participated in the training program as the treated group and those who did not participate as the control group.

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

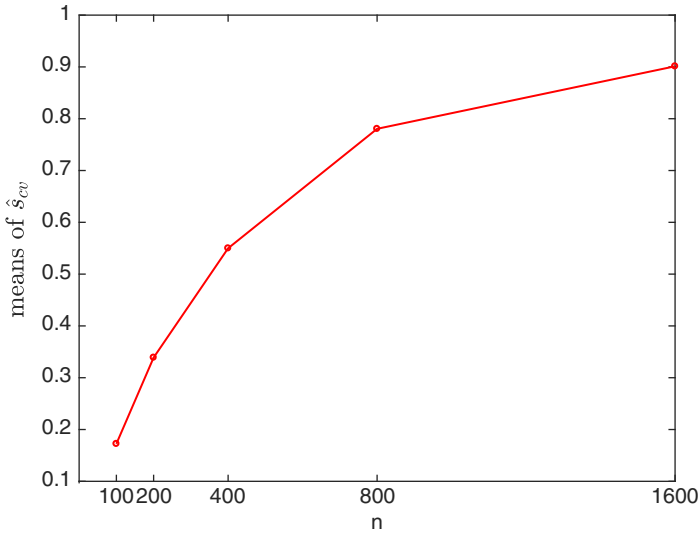


FIGURE 4. Assessing the convergence of the model weights $\hat{\omega}_{cv}$ under Design 2.

TABLE 1. The simple correlation coefficient of each baseline covariates with $\hat{\mu}_{\text{pair}}$

Symbol	Description	Correlation with $\hat{\mu}_{\text{pair}}$
X_1	Earnings in 1975 divided by 10^3	−0.1767
X_2	Indicator for Hispanic (1 if Hispanic, 0 otherwise)	−0.1286
X_3	Indicator for black (1 if black, 0 otherwise)	0.1170
X_4	Years of education	0.0666
X_5	Marital status (1 if married, 0 otherwise)	0.0242
X_6	Indicator for no degree (1 if no degree, 0 otherwise)	−0.0132
X_7	Age	0.0125

We calculate the simple correlation coefficient of each baseline covariate with $\hat{\mu}_{\text{pair}}$, where $\hat{\mu}_{\text{pair}}$ is a nonparametric estimator of the CATE μ based on nearest neighbor pairing used in Rolling et al. (2019). Subsequently, we label the baseline covariates as $X_1, X_2, \ldots, X_7$ in descending order of the absolute values of the correlation coefficients. See Table 1 for details.

Although our proposal does not restrict the candidate models to be nested, for simplicity, we assume that the mth candidate model for μ uses the first m covariates in $\{X_1, X_2, \ldots, X_7\}$ with an intercept term. To implement our proposal, we specify a logistic model with all the covariates for the propensity score function $\pi(\mathrm{X})$.

We conduct a "guided simulation experiment" to evaluate the performance of the methods considered in the above section. Specifically, we first apply a model selection approach to the dataset to select an appropriate model. Let $\hat{m}$ be the index

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

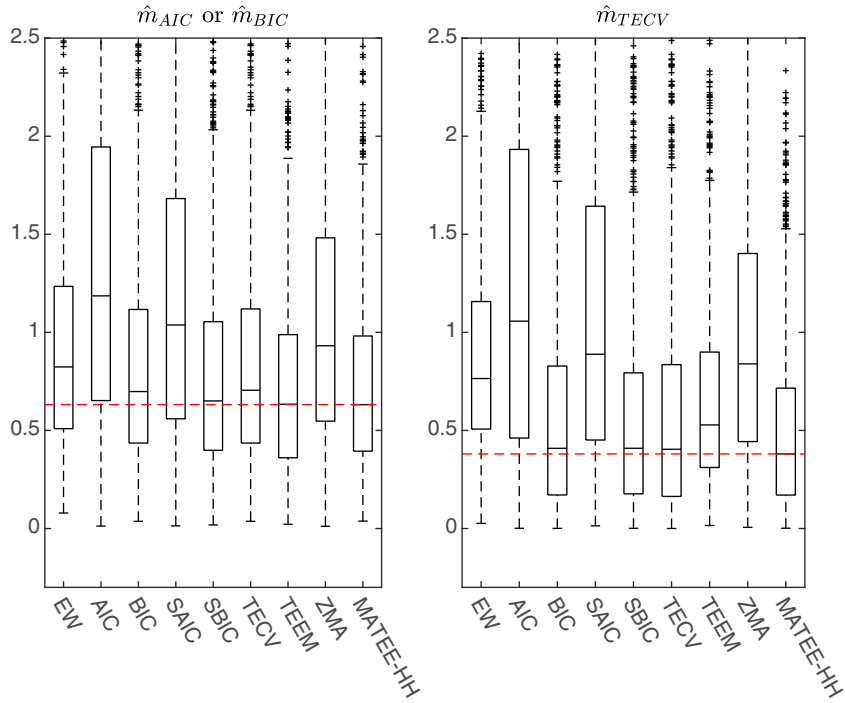


FIGURE 5. Box-plots of MSEs by different methods for the NSW Demonstration dataset under different selected models.

of the order of this selected model in the class of candidate models. Then, based on the $\hat{m}$th model, we generate $\{e_{(\hat{m}),t,i}\}_{i=1}^n$ and $\{e_{(\hat{m}),c,i}\}_{i=1}^n$ from $N(0,\hat{\sigma}^2_{(\hat{m}),t})$ and $N(0,\hat{\sigma}^2_{(\hat{m}),c})$, respectively, where $\hat{\sigma}^2_{(\hat{m}),t}$ and $\hat{\sigma}^2_{(\hat{m}),c}$ are the error variance estimates derived based on the selected model with the original dataset $\{Y_i, \mathrm{X}_i, T_i\}_{i=1}^n$. This leads us to a simulation dataset $\{Y_{(\hat{m}),i}, \mathrm{X}_i, T_i\}_{i=1}^n$ where

$$Y_{(\hat{m}),i} = T_i Y_{(\hat{m}),t,i} + (1-T_i) Y_{(\hat{m}),c,i}, \quad Y_{(\hat{m}),t,i} = \mathrm{X}_{(\hat{m}),i}\hat{\beta}_{(\hat{m})} + e_{(\hat{m}),t,i},$$
$$Y_{(\hat{m}),c,i} = \mathrm{X}_{(\hat{m}),i}\hat{\eta}_{(\hat{m})} + e_{(\hat{m}),c,i},$$

and $\hat{\beta}_{(\hat{m})}$ and $\hat{\eta}_{(\hat{m})}$ are the regression coefficient estimates obtained based on the selected model with the original dataset $\{Y_i, \mathrm{X}_i, T_i\}_{i=1}^n$. For this simulation dataset, the "true" μ_i is known and is equal to $\mathrm{X}_{(\hat{m}),i}\hat{\beta}_{(\hat{m})} - \mathrm{X}_{(\hat{m}),i}\hat{\eta}_{(\hat{m})}$. Thus, we adopt the MSE $\|\{\hat{\mu}(\omega)\} - \mu\|^2/n$ as our performance metric, where $\hat{\mu}(\omega)$ is now obtained with the simulation dataset $\{Y_{(\hat{m}),i}, \mathrm{X}_i, T_i\}_{i=1}^n$. We repeat this experiment 1,000 times and report the results in Figure 5. In Figure 5, $\hat{m}_{AIC}, \hat{m}_{BIC}$ and $\hat{m}_{TECV}$

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

correspond to the model chosen by applying AIC, BIC, and the TECV method on the original dataset $\{Y_i, \mathrm{X}_i, T_i\}_{i=1}^n$, respectively. It should be noted that $\hat{m}_{AIC}$ and $\hat{m}_{BIC}$ are the same.

Figure 5 shows that our proposal performs better than its competitors except for the TEEM algorithm for all situations. Additionally, compared to the TEEM algorithm, our proposed method shows a comparable performance under $\hat{m}_{AIC}$ or $\hat{m}_{BIC}$ and clearly better performance under $\hat{m}_{TECV}$.

6. DISCUSSION

In this article, we develop a model averaging method, called MATEE-HH, for estimating CATE with multiple candidate linear regression models. Using Z_π is crucial to our proposal and is also a big difference between our proposal and some related methods. It helps us to circumvent the "fundamental problem of causal inference" and thus makes the use of the JMA approach possible. We first establish the asymptotic optimality of our proposed method and then demonstrate the convergence of the model weights determined by our proposal when at least one of the candidate models is correctly specified. The simulation and empirical example results favor our proposal over its main competitors.

Our proposal requires a correct specification of the parametric model for $\pi(\mathrm{X})$. Lee, Okui, and Whang (2017) developed a nonparametric CATE estimator with parametric estimations of $\pi(\mathrm{X}), E(Y|X, T=1)$ and $E(Y|X, T=0)$. They showed that their proposed estimator is doubly robust in the sense that it is consistent if either the parametric model for $\pi(\mathrm{X})$ is correctly specified or the parametric models for $E(Y|X, T=1)$ and $E(Y|X, T=0)$ are correctly specified. Inspired by this, we think one possible way to alleviate the sensitivity of our proposal to the specification of the parametric model for $\pi(\mathrm{X})$ is to combine the idea of Lee et al. (2017) with our proposal. However, such a combination is nontrivial and needs further investigation. Instead of using a parametric estimator of $\pi(\mathrm{X})$, one may choose to estimate $\pi(\mathrm{X})$ nonparametrically in our proposed method. According to Graham (2011) and Abrevaya, Hsu, and Lieli (2015), this choice presumably increases the efficiency of our proposal. However, a theoretical validation of the efficiency gain from this choice is challenging due to the difficulties in deriving the asymptotic distribution of the final model average estimators, and is beyond the scope of this article. We leave it as a future study. Additionally, in practice, researchers may also have many candidate models available for $\pi(\mathrm{X})$. In this case, adapting our proposed MATEE-HH method to combine the information contained in these candidate models postulated for $\pi(\mathrm{X})$ is undoubtedly an interesting topic and warrants future studies.

In this article, we propose a model average estimator for $\mu = (\mu_1, \mu_2, \ldots, \mu_n)^\top$ where μ_i is the CATE for the ith individual, and the corresponding weights are shown to possess desirable properties. Here, we point out that the CATE for a new observation i^* can be estimated by

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

$$\hat{\mu}_{i^*} = \sum_{m=1}^{M_n} \hat{\omega}_{cv,m} \mathrm{X}_{(m),i^*} (\mathbf{X}_{(m)}^\top \mathbf{X}_{(m)})^{-1} \mathbf{X}_{(m)}^\top \mathrm{Z}_{\hat{\pi}},$$

where $\mathrm{X}_{(m),i^*}$ is the covariate vector corresponding to the i^*th observation under the mth model. In addition to CATE, one might be interested in the average treatment effect which is defined at the population level and is given by $\theta_{ATE} = E(Y_t - Y_c)$. Furthermore, one might be interested in the treatment effect at a specific subgroup level. For example, one may be interested in $\theta_{sub} = E(Y_t - Y_c | X_1 = 1)$ where X_1 is a binary variable, taking value 1 if the individual is female, and 0 otherwise. Based on $\hat{\mu}_{\hat{\pi}}(\hat{\omega}_{cv})$, we can obtain model average estimates for these quantities of interest. Since $\theta_{ATE} = E\left\{E(Y_{t,i} - Y_{c,i} | \mathrm{X}_i)\right\} = E(\mu_i)$, a model average estimator for θ_{ATE} can be computed as $\hat{\theta}_{ATE}(\hat{\omega}_{cv}) = n^{-1} \sum_{i=1}^n \hat{\mu}_{\hat{\pi},i}(\hat{\omega}_{cv})$ where $\hat{\mu}_{\hat{\pi},i}(\hat{\omega}_{cv})$ is the ith element of $\hat{\mu}_{\hat{\pi}}(\hat{\omega}_{cv})$. Note that

$$\begin{aligned}\theta_{sub} &= E(Y_t - Y_c | X_1 = 1) = E\left\{E(Y_{t,i} - Y_{c,i} | \mathrm{X}_i) | \mathrm{X}_{i1} = 1\right\} \\ &= E(\mu_i | \mathrm{X}_{i1} = 1) = \frac{E\{\mu_i I(\mathrm{X}_{i1} = 1)\}}{P(\mathrm{X}_{i1} = 1)},\end{aligned}$$

where X_{i1} is the first element of X_i and $I(\mathscr{S})$ is the indicator of set $\mathscr{S}$. Then, a model average estimator for θ_{sub} can be obtained by $\hat{\theta}_{sub}(\hat{\omega}_{cv}) = \left\{\sum_{i=1}^n I(\mathrm{X}_{i1} = 1)\right\}^{-1} \sum_{i=1}^n \hat{\mu}_{\hat{\pi},i}(\hat{\omega}_{cv}) I(\mathrm{X}_{i1} = 1)$. The theoretical properties as well as the finite sample performances of $\hat{\theta}_{ATE}(\hat{\omega}_{cv})$ and $\hat{\theta}_{sub}(\hat{\omega}_{cv})$ are not analyzed since they are outside the scope of this article. However, they are worthy of investigation and we leave them for future studies.

Last, like many papers in the field of model averaging, the asymptotic distribution of the final model averaging estimator $\hat{\mu}_{\hat{\pi}}(\hat{\omega}_{cv})$ is lacking in this article. Finding the asymptotic distribution of $\hat{\mu}_{\hat{\pi}}(\hat{\omega}_{cv})$ is important since it may help us to obtain confidence intervals for μ. However, this is not an easy task due to the random weights and possible model misspecification. Greater efforts must be dedicated to this endeavor.

A. APPENDIX

A.1. Lemmas and its Proofs

Before presenting the proof of Theorem 1, let us first present some lemmas that are used in the course of the proof. In the following context, we use c to denote a generic positive constant that could take different values in different occasions.

LEMMA 1. *Let* $\tilde{\mathrm{e}}_\pi = (\tilde{e}_{\pi,1}, \tilde{e}_{\pi,2}, \ldots, \tilde{e}_{\pi,n})^\top$ *where* $\tilde{e}_{\pi,i} = e_{\pi,i}/\sigma_{\pi,i}$. *Let* $\mathbf{D}_\tau$ *be an* $n \times n$ *diagonal matrix with ith diagonal element* $\tau_i = E(\tilde{e}_{\pi,i}^4 | \mathrm{X}_i) - 3$. *Then, for any two* $n \times n$ *matrices* $\mathbf{O}_1$ *and* $\mathbf{O}_2$, *we have*

$$E(\tilde{\mathrm{e}}_\pi^\top \mathbf{O}_1 \tilde{\mathrm{e}}_\pi \tilde{\mathrm{e}}_\pi^\top \mathbf{O}_2 \tilde{\mathrm{e}}_\pi | \mathbf{X}) = tr\{\mathbf{O}_1\} tr\{\mathbf{O}_2\} + tr\{\mathbf{O}_1 \mathbf{O}_2 + \mathbf{O}_1^\top \mathbf{O}_2\} + tr\{\mathbf{O}_1 * (\mathbf{D}_\tau \mathbf{O}_2)\},$$

where $*$ *denotes the Hadamard product.*

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

Proof of Lemma 1. By the definition of $\tilde{e}_{\pi,i}$ and (6), after some careful calculations, we have

$$E(\tilde{e}_{\pi,i}\tilde{e}_{\pi,j}\tilde{\mathrm{e}}_\pi\tilde{\mathrm{e}}_\pi^\top|\mathbf{X}) = \begin{cases} \mathbf{E}_{ij}+\mathbf{E}_{ji} & i\neq j, \\ \mathbf{I}_n+2\mathbf{E}_{ii}+\tau_i\mathbf{E}_{ii} & i=j, \end{cases}$$

where $\mathbf{E}_{ij}$ is an $n\times n$ matrix whose ijth element is one while the other elements are zeros. Thus, we have

$$\begin{aligned} E\left\{(\tilde{\mathrm{e}}_\pi\tilde{\mathrm{e}}_\pi^\top)\otimes(\tilde{\mathrm{e}}_\pi\tilde{\mathrm{e}}_\pi^\top)|\mathbf{X}\right\} &= E\left[\left\{\sum_{i=1}^p\sum_{j=1}^p(\tilde{e}_{\pi,i}\tilde{e}_{\pi,j}\mathbf{E}_{ij})\right\}\otimes(\tilde{\mathrm{e}}_\pi\tilde{\mathrm{e}}_\pi^\top)\middle|\mathbf{X}\right] \\ &= \sum_{i=1}^p\sum_{j=1}^p\left\{\mathbf{E}_{ij}\otimes E(\tilde{e}_{\pi,i}\tilde{e}_{\pi,j}\tilde{\mathrm{e}}_\pi\tilde{\mathrm{e}}_\pi^\top|\mathbf{X})\right\} \\ &= \sum_{i=1}^p\sum_{j=1}^p\left\{\mathbf{E}_{ij}\otimes(\mathbf{E}_{ij}+\mathbf{E}_{ji})\right\}+\sum_{i=1}^p\{\mathbf{E}_{ii}\otimes(\mathbf{I}_n+\tau_i\mathbf{E}_{ii})\}, \end{aligned}$$

where $\otimes$ denotes the Kronecker product. This implies that

$$\begin{aligned} E\left(\tilde{\mathrm{e}}_\pi^\top\mathbf{O}_1\tilde{\mathrm{e}}_\pi\tilde{\mathrm{e}}_\pi^\top\mathbf{O}_2\tilde{\mathrm{e}}_\pi|\mathbf{X}\right) &= E\left\{(\tilde{\mathrm{e}}_\pi^\top\otimes\tilde{\mathrm{e}}_\pi^\top)(\mathbf{O}_1\otimes\mathbf{O}_2)(\tilde{\mathrm{e}}_\pi\otimes\tilde{\mathrm{e}}_\pi)|\mathbf{X}\right\} \\ &= tr\left[(\mathbf{O}_1\otimes\mathbf{O}_2)E\left\{(\tilde{\mathrm{e}}_\pi\tilde{\mathrm{e}}_\pi^\top)\otimes(\tilde{\mathrm{e}}_\pi\tilde{\mathrm{e}}_\pi^\top)|\mathbf{X}\right\}\right] \\ &= \sum_{i=1}^p\sum_{j=1}^p\left[tr\left\{\mathbf{O}_1\mathbf{E}_{ij}\right\}tr\left\{\mathbf{O}_2\mathbf{E}_{ij}+\mathbf{O}_2\mathbf{E}_{ji}\right\}\right] \\ &\quad+\sum_{i=1}^p[tr\{\mathbf{O}_1\mathbf{E}_{ii}\}tr\{\mathbf{O}_2\mathbf{I}_n+\tau_i\mathbf{O}_2\mathbf{E}_{ii}\}] \\ &= tr\{\mathbf{O}_1\}tr\{\mathbf{O}_2\}+tr\{\mathbf{O}_1\mathbf{O}_2+\mathbf{O}_1^\top\mathbf{O}_2\}+tr\{\mathbf{O}_1*(\mathbf{D}_\tau\mathbf{O}_2)\}, \end{aligned}$$

which completes the proof of Lemma 1. □

The following lemma is a key to the proof of Theorem 1. It shows that the distance between Z_π and its estimator $\mathrm{Z}_{\hat{\pi}}$ converges at an appropriate rate.

LEMMA 2. *Provided that Conditions (C1) and (C2) hold, we have* $\|\mathrm{Z}_{\hat{\pi}}-\mathrm{Z}_\pi\|^2 = O_p(1)$.

Proof of Lemma 2. According to the monotonicity of L_r-norm presented in Lin and Bai (2011), we have

$$\begin{aligned} \max\left\{\max_{1\le i\le n}\sigma_i^2,\ \max_{1\le i\le n}\sigma_{t,i}^2\right\} &= \max\left\{\max_{1\le i\le n}E(e_i^2|\mathrm{X}_i),\ \max_{1\le i\le n}E(e_{t,i}^2|\mathrm{X}_i)\right\} \\ &\le \max\left[\max_{1\le i\le n}\{E(e_i^4|\mathrm{X}_i)\}^{1/2},\ \max_{1\le i\le n}\{E(e_{t,i}^4|\mathrm{X}_i)\}^{1/2}\right] \\ &\le \bar{C}, \end{aligned} \tag{A.1}$$

almost surely where the last inequality is due to (C2). Thus, we have, for every $\epsilon>0$, there exists an integer N_ϵ such that $P(\max_{1\le i\le n}\sigma_i^2>\bar{C})\le\epsilon/2$, for all $n\ge N_\epsilon$. Let $M_\epsilon = 2\bar{C}/\epsilon$.

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

Then, by Markov's inequality, we have

$$
\begin{aligned}
&\sup_{n\geq 1} P\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2 > M_\epsilon\right) \\
&\quad = \sup_{n\geq 1}\left\{P\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2 > M_\epsilon,\ \max_{1\leq i\leq n}\sigma_i^2 \leq \bar{C}\right) + P\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2 > M_\epsilon,\ \max_{1\leq i\leq n}\sigma_i^2 > \bar{C}\right)\right\} \\
&\quad \leq \sup_{n\geq 1} E\left\{I\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2 > M_\epsilon\right) I\left(\max_{1\leq i\leq n}\sigma_i^2 \leq \bar{C}\right)\right\} + \sup_{n\geq N_\epsilon} P\left(\max_{1\leq i\leq n}\sigma_i^2 > \bar{C}\right) \\
&\quad \leq \sup_{n\geq 1} E\left[E\left\{I\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2 > M_\epsilon\right)\Big|\mathbf{X}\right\} I\left(\max_{1\leq i\leq n}\sigma_i^2 \leq \bar{C}\right)\right] + \frac{\epsilon}{2} \\
&\quad = \sup_{n\geq 1} E\left\{P\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2 > M_\epsilon\Big|\mathbf{X}\right) I\left(\max_{1\leq i\leq n}\sigma_i^2 \leq \bar{C}\right)\right\} + \frac{\epsilon}{2} \\
&\quad \leq \sup_{n\geq 1} E\left\{M_\epsilon^{-1} E\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2\Big|\mathbf{X}\right) I\left(\max_{1\leq i\leq n}\sigma_i^2 \leq \bar{C}\right)\right\} + \frac{\epsilon}{2} \\
&\quad = \sup_{n\geq 1} E\left\{M_\epsilon^{-1}\frac{1}{n}\sum_{i=1}^{n}\sigma_i^2 I\left(\max_{1\leq i\leq n}\sigma_i^2 \leq \bar{C}\right)\right\} + \frac{\epsilon}{2} \\
&\quad \leq M_\epsilon^{-1}\bar{C} + \frac{\epsilon}{2} \\
&\quad = \epsilon.
\end{aligned}
$$

Thus, we have

$$\frac{1}{n}\sum_{i=1}^{n} e_i^2 = O_p(1). \tag{A.2}$$

Similarly, it can be shown that

$$\frac{1}{n}\sum_{i=1}^{n} e_{t,i}^2 = O_p(1). \tag{A.3}$$

By Cauchy–Schwarz inequality, we have

$$
\begin{aligned}
&\|Z_{\hat{\pi}} - Z_\pi\|^2 \\
&\quad = \sum_{i=1}^{n}\left\{\frac{T_i}{\tilde{\pi}(\mathrm{X}_i;\hat{\alpha}_n)} Y_{t,i} - \frac{1-T_i}{1-\tilde{\pi}(\mathrm{X}_i;\hat{\alpha}_n)} Y_{c,i} - \left(\frac{T_i}{\pi(\mathrm{X}_i)} Y_{t,i} - \frac{1-T_i}{1-\pi(\mathrm{X}_i)} Y_{c,i}\right)\right\}^2 \\
&\quad = \sum_{i=1}^{n}\left[\left\{\frac{T_i - \tilde{\pi}(\mathrm{X}_i;\hat{\alpha}_n)}{\tilde{\pi}(\mathrm{X}_i;\hat{\alpha}_n)(1-\tilde{\pi}(\mathrm{X}_i;\hat{\alpha}_n))} - \frac{T_i - \pi(\mathrm{X}_i)}{\pi(\mathrm{X}_i)(1-\pi(\mathrm{X}_i))}\right\} Y_{t,i}\right.
\end{aligned}
$$

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

$$
\begin{aligned}
&+\left\{\frac{1-T_i}{1-\tilde{\pi}(\mathrm{X}_i;\hat{\alpha}_n)}-\frac{1-T_i}{1-\pi(\mathrm{X}_i)}\right\}(Y_{t,i}-Y_{c,i})\Bigg]^2\\
&\le c\Bigg[\left\{\sqrt{n}\max_{1\le i\le n}\left|\frac{T_i-\tilde{\pi}(\mathrm{X}_i;\hat{\alpha}_n)}{\tilde{\pi}(\mathrm{X}_i;\hat{\alpha}_n)(1-\tilde{\pi}(\mathrm{X}_i;\hat{\alpha}_n))}-\frac{T_i-\pi(\mathrm{X}_i)}{\pi(\mathrm{X}_i)(1-\pi(\mathrm{X}_i))}\right|\right\}^2\frac{1}{n}\sum_{i=1}^n Y_{t,i}^2\\
&\quad+\left\{\sqrt{n}\max_{1\le i\le n}\left|\frac{1-T_i}{1-\tilde{\pi}(\mathrm{X}_i;\hat{\alpha}_n)}-\frac{1-T_i}{1-\pi(\mathrm{X}_i)}\right|\right\}^2\frac{1}{n}\sum_{i=1}^n(Y_{t,i}-Y_{c,i})^2\Bigg]\\
&\le c\Bigg[\left\{\sqrt{n}\max_{1\le i\le n}\left|\frac{T_i-\tilde{\pi}(\mathrm{X}_i;\hat{\alpha}_n)}{\tilde{\pi}(\mathrm{X}_i;\hat{\alpha}_n)(1-\tilde{\pi}(\mathrm{X}_i;\hat{\alpha}_n))}-\frac{T_i-\pi(\mathrm{X}_i)}{\pi(\mathrm{X}_i)(1-\pi(\mathrm{X}_i))}\right|\right\}^2\\
&\quad\times\left(\frac{1}{n}\sum_{i=1}^n\mu_{t,i}^2+\frac{1}{n}\sum_{i=1}^n e_{t,i}^2\right)\\
&\quad+\left\{\sqrt{n}\max_{1\le i\le n}\left|\frac{1-T_i}{1-\tilde{\pi}(\mathrm{X}_i;\hat{\alpha}_n)}-\frac{1-T_i}{1-\pi(\mathrm{X}_i)}\right|\right\}^2\left(\frac{1}{n}\sum_{i=1}^n\mu_i^2+\frac{1}{n}\sum_{i=1}^n e_i^2\right)\Bigg]\\
&=\left\{\sqrt{n}\max_{1\le i\le n}\left|\frac{T_i-\tilde{\pi}(\mathrm{X}_i;\hat{\alpha}_n)}{\tilde{\pi}(\mathrm{X}_i;\hat{\alpha}_n)(1-\tilde{\pi}(\mathrm{X}_i;\hat{\alpha}_n))}-\frac{T_i-\pi(\mathrm{X}_i)}{\pi(\mathrm{X}_i)(1-\pi(\mathrm{X}_i))}\right|\right\}^2 O_p(1)\\
&\quad+\left\{\sqrt{n}\max_{1\le i\le n}\left|\frac{1-T_i}{1-\tilde{\pi}(\mathrm{X}_i;\hat{\alpha}_n)}-\frac{1-T_i}{1-\pi(\mathrm{X}_i)}\right|\right\}^2 O_p(1),
\end{aligned}
$$

where the last equality is due to (C2), (A.2), and (A.3). Therefore, to prove Lemma 2, it suffices to prove

$$
\sqrt{n}\max_{1\le i\le n}\left|\frac{T_i-\tilde{\pi}(\mathrm{X}_i;\hat{\alpha}_n)}{\tilde{\pi}(\mathrm{X}_i;\hat{\alpha}_n)(1-\tilde{\pi}(\mathrm{X}_i;\hat{\alpha}_n))}-\frac{T_i-\pi(\mathrm{X}_i)}{\pi(\mathrm{X}_i)(1-\pi(\mathrm{X}_i))}\right|=O_p(1), \tag{A.4}
$$

$$
\sqrt{n}\max_{1\le i\le n}\left|\frac{1-T_i}{1-\tilde{\pi}(\mathrm{X}_i;\hat{\alpha}_n)}-\frac{1-T_i}{1-\pi(\mathrm{X}_i)}\right|=O_p(1). \tag{A.5}
$$

Applying Taylor expansion to $\{\tilde{\pi}(\mathrm{X}_i;\hat{\alpha}_n)(1-\tilde{\pi}(\mathrm{X}_i;\hat{\alpha}_n))\}^{-1}\{T_i-\tilde{\pi}(\mathrm{X}_i;\hat{\alpha}_n)\}$ around α_0, we obtain

$$
\begin{aligned}
&\sqrt{n}\max_{1\le i\le n}\left|\frac{T_i-\tilde{\pi}(\mathrm{X}_i;\hat{\alpha}_n)}{\tilde{\pi}(\mathrm{X}_i;\hat{\alpha}_n)(1-\tilde{\pi}(\mathrm{X}_i;\hat{\alpha}_n))}-\frac{T_i-\pi(\mathrm{X}_i)}{\pi(\mathrm{X}_i)(1-\pi(\mathrm{X}_i))}\right|\\
&=\max_{1\le i\le n}\left|\left\{\frac{T_i-\tilde{\pi}(\mathrm{X}_i;\alpha_i^*)}{\tilde{\pi}(\mathrm{X}_i;\alpha_i^*)(1-\tilde{\pi}(\mathrm{X}_i;\alpha_i^*))}\right\}^2\cdot\frac{\partial\tilde{\pi}(\mathrm{X}_i;\alpha)}{\partial\alpha^\top}\Big|_{\alpha=\alpha_i^*}\cdot\sqrt{n}(\hat{\alpha}_n-\alpha_0)\right|\\
&\le c\left\{\min_{1\le i\le n}\tilde{\pi}(\mathrm{X}_i;\alpha_i^*)\right\}^{-2}\left\{1-\max_{1\le i\le n}\tilde{\pi}(\mathrm{X}_i;\alpha_i^*)\right\}^{-2}\\
&\quad\times\max_{1\le i\le n}\left\|\frac{\partial\tilde{\pi}(\mathrm{X}_i;\alpha)}{\partial\alpha^\top}\Big|_{\alpha=\alpha_i^*}\right\|\sqrt{n}\|\hat{\alpha}_n-\alpha_0\|,
\end{aligned}
\tag{A.6}
$$

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

where α_i^* is a vector between $\hat{\alpha}_n$ and α_0. Expanding $\tilde{\pi}(\mathrm{X}_i;\alpha_i^*)$ in a Taylor series around α_0 and considering (5), we have

$$\begin{aligned}\min_{1\le i\le n}\tilde{\pi}(\mathrm{X}_i;\alpha_i^*) &= \min_{1\le i\le n}\left\{\pi(\mathrm{X}_i)+\frac{\partial\tilde{\pi}(\mathrm{X}_i;\alpha)}{\partial\alpha^\top}\Big|_{\alpha=\alpha_i^{**}}\cdot(\alpha_i^*-\alpha_0)\right\}\\ &\ge c_\pi-\max_{1\le i\le n}\left\|\frac{\partial\tilde{\pi}(\mathrm{X}_i;\alpha)}{\partial\alpha^\top}\Big|_{\alpha=\alpha_i^{**}}\right\|\cdot\|\hat{\alpha}_n-\alpha_0\|,\\ 1-\max_{1\le i\le n}\tilde{\pi}(\mathrm{X}_i;\alpha_i^*) &= 1-\max_{1\le i\le n}\left\{\pi(\mathrm{X}_i)+\frac{\partial\tilde{\pi}(\mathrm{X}_i;\alpha)}{\partial\alpha^\top}\Big|_{\alpha=\alpha_i^{**}}\cdot(\alpha_i^*-\alpha_0)\right\}\\ &\ge c_\pi-\max_{1\le i\le n}\left\|\frac{\partial\tilde{\pi}(\mathrm{X}_i;\alpha)}{\partial\alpha^\top}\Big|_{\alpha=\alpha_i^{**}}\right\|\cdot\|\hat{\alpha}_n-\alpha_0\|,\end{aligned}$$

where α_i^{**} is a vector between α_i^* and α_0. This together with (A.6) and Condition (C1) implies that (A.4) is valid. Similarly, it can be shown that (A.5) is valid. This completes the proof of Lemma 2. □

LEMMA 3. *Under Condition (C2), there exists positive constants $c_{\pi,1}, c_{\pi,2}$ and τ_π such that*

$$0<c_{\pi,1}^2\le\min_{1\le i\le n}\sigma_{\pi,i}^2\le\max_{1\le i\le n}\sigma_{\pi,i}^2\le c_{\pi,2}^2<\infty,\quad a.s., \tag{A.7}$$

$$\max_{1\le i\le n}\tau_i\le\tau_\pi,\quad a.s. \tag{A.8}$$

Proof of Lemma 3. By (4), (5), (6), (A.1), (C2) and Cauchy-Schwarz inequality, we have

$$\begin{aligned}\sigma_{\pi,i}^2 &= E\left\{\left(\frac{T_i}{\pi(\mathrm{X}_i)}Y_{t,i}-\frac{1-T_i}{1-\pi(\mathrm{X}_i)}Y_{c,i}-\mu_i\right)^2\Big|\mathrm{X}_i\right\}\\ &= E\left[\left\{\frac{T_i-\pi(\mathrm{X}_i)}{\pi(\mathrm{X}_i)(1-\pi(\mathrm{X}_i))}(\mu_{t,i}+e_{t,i})+\frac{\pi(\mathrm{X}_i)-T_i}{1-\pi(\mathrm{X}_i)}\mu_i+\frac{1-T_i}{1-\pi(\mathrm{X}_i)}e_i\right\}^2\Big|\mathrm{X}_i\right]\\ &= E\left[\left\{\frac{T_i-\pi(\mathrm{X}_i)}{\pi(\mathrm{X}_i)(1-\pi(\mathrm{X}_i))}\mu_{t,i}+\frac{\pi(\mathrm{X}_i)-T_i}{1-\pi(\mathrm{X}_i)}\mu_i\right\}^2\Big|\mathrm{X}_i\right]\\ &\quad+E\left[\left\{\frac{T_i-\pi(\mathrm{X}_i)}{\pi(\mathrm{X}_i)(1-\pi(\mathrm{X}_i))}e_{t,i}+\frac{1-T_i}{1-\pi(\mathrm{X}_i)}e_i\right\}^2\Big|\mathrm{X}_i\right]\\ &\le c\left(E\left[\left\{\frac{T_i-\pi(\mathrm{X}_i)}{\pi(\mathrm{X}_i)(1-\pi(\mathrm{X}_i))}\mu_{t,i}\right\}^2\Big|\mathrm{X}_i\right]+E\left[\left\{\frac{\pi(\mathrm{X}_i)-T_i}{1-\pi(\mathrm{X}_i)}\mu_i\right\}^2\Big|\mathrm{X}_i\right]\right.\\ &\quad\left.+E\left[\left\{\frac{T_i-\pi(\mathrm{X}_i)}{\pi(\mathrm{X}_i)(1-\pi(\mathrm{X}_i))}e_{t,i}\right\}^2\Big|\mathrm{X}_i\right]+E\left[\left\{\frac{1-T_i}{1-\pi(\mathrm{X}_i)}e_i\right\}^2\Big|\mathrm{X}_i\right]\right)\\ &\le c\left(c_\pi^{-2}\bar{C}^2+c_\pi^{-1}\bar{C}^2+c_\pi^{-2}\sigma_{t,i}^2+c_\pi^{-2}\sigma_i^2\right)\\ &\le c\left(c_\pi^{-2}\bar{C}^2+c_\pi^{-1}\bar{C}^2+2c_\pi^{-2}\bar{C}\right),\end{aligned}$$

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

$$\begin{aligned}\sigma_{\pi,i}^2 &\geq E\left[\left\{\frac{T_i-\pi(\mathrm{X}_i)}{\pi(\mathrm{X}_i)(1-\pi(\mathrm{X}_i))}e_{t,i}+\frac{1-T_i}{1-\pi(\mathrm{X}_i)}e_i\right\}^2\Big|\mathrm{X}_i\right]\\ &= E\left[\left\{\frac{1}{\pi(\mathrm{X}_i)}-1\right\}\frac{1}{1-\pi(\mathrm{X}_i)}e_{t,i}^2+\frac{1}{1-\pi(\mathrm{X}_i)}(e_{t,i}-e_i)^2\Big|\mathrm{X}_i\right]\\ &\geq (1-c_\pi)^{-1}\left\{(1-c_\pi)^{-1}-1\right\}\sigma_{t,i}^2\\ &\geq (1-c_\pi)^{-1}\left\{(1-c_\pi)^{-1}-1\right\}c_t,\end{aligned}$$

almost surely. Thus, it is readily seen that (A.7) is valid. By (5), (6), (C2), and Minkowski inequality, we have

$$\begin{aligned}E(e_{\pi,i}^4|\mathrm{X}_i) &= E\left[\left\{\frac{T_i-\pi(\mathrm{X}_i)}{\pi(\mathrm{X}_i)(1-\pi(\mathrm{X}_i))}(\mu_{t,i}+e_{t,i})+\frac{\pi(\mathrm{X}_i)-T_i}{1-\pi(\mathrm{X}_i)}\mu_i+\frac{1-T_i}{1-\pi(\mathrm{X}_i)}e_i\right\}^4\Big|\mathrm{X}_i\right]\\ &\leq \left(E\left[\left\{\frac{T_i-\pi(\mathrm{X}_i)}{\pi(\mathrm{X}_i)(1-\pi(\mathrm{X}_i))}\mu_{t,i}\right\}^4\Big|\mathrm{X}_i\right]^{\frac{1}{4}}+E\left[\left\{\frac{\pi(\mathrm{X}_i)-T_i}{1-\pi(\mathrm{X}_i)}\mu_i\right\}^4\Big|\mathrm{X}_i\right]^{\frac{1}{4}}\right.\\ &\quad\left.+E\left[\left\{\frac{T_i-\pi(\mathrm{X}_i)}{\pi(\mathrm{X}_i)(1-\pi(\mathrm{X}_i))}e_{t,i}\right\}^4\Big|\mathrm{X}_i\right]^{\frac{1}{4}}+E\left[\left\{\frac{1-T_i}{1-\pi(\mathrm{X}_i)}e_i\right\}^4\Big|\mathrm{X}_i\right]^{\frac{1}{4}}\right)^4\\ &\leq \left\{2c_\pi^{-1}(c_\pi^{-1}+1)(\bar{C}+\bar{C}^{\frac{1}{4}})\right\}^4,\end{aligned}$$

almost surely. This together with the definition of τ_i given in Lemma 1 implies that (A.8) is valid. The proof is complete. □

LEMMA 4. *Rao (1973) For a $p\times p$ positive-definite and symmetric matrix $\mathbf{B}$ and a p-dimensional vector* v,

$$\sup_{\mathrm{z}}\frac{(\mathrm{z}^\top\mathrm{v})^2}{\mathrm{z}^\top\mathbf{B}\mathrm{z}}=\mathrm{v}^\top\mathbf{B}\mathrm{v}.$$

LEMMA 5. *Zhang (2021) Let $\widetilde{\omega}=\operatorname{argmin}_{\omega\in\mathcal{W}}\{L_n(\omega)+a_n(\omega)+b_n\}$, where $a_n(\omega)$ is a term related to ω and b_n is a term unrelated to ω. Let $R_n(\omega)=E\{L_n(\omega)|\mathbf{X}\}$. If*

$$\sup_{\omega\in\mathcal{W}}\frac{|a_n(\omega)|}{R_n(\omega)}=o_p(1),\quad \sup_{w\in\omega}\frac{|R_n(\omega)-L_n(\omega)|}{R_n(\omega)}=o_p(1),$$

and there exists a constant c and a positive integer N^* so that when $n\geq N^*$, $\inf_{\omega\in\mathcal{W}}R_n(\omega)\geq c>0$ almost surely, then $L_n(\widetilde{\omega})/\inf_{\omega\in\mathcal{W}}L_n(\omega)\to 1$ in probability.

LEMMA 6. *Zhang (2021) For any $n_1\times n_2$ matrices $\mathbf{B}_1$ and $\mathbf{B}_2$,*

$$\lambda_{\max}\{\mathbf{B}_1\mathbf{B}_2\}\leq\lambda_{\max}\{\mathbf{B}_1\}\lambda_{\max}\{\mathbf{B}_2\},\quad \lambda_{\max}\{\mathbf{B}_1+\mathbf{B}_2\}\leq\lambda_{\max}\{\mathbf{B}_1\}+\lambda_{\max}\{\mathbf{B}_2\}.$$

A.2. Proof of Theorem 1

Proof. By the definition of $R_\pi(\omega)$ in (14), it can be shown that $R_\pi(\omega)=\|\mathbf{A}(\omega)\mu\|^2+\operatorname{tr}\{\mathbf{P}(\omega)\mathbf{\Omega}_\pi\mathbf{P}(\omega)^\top\}$ where $\mathbf{A}(\omega)=\mathbf{I}_n-\mathbf{P}(\omega)$. Define the loss function and the risk function

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

of $\tilde{\mu}_\pi(\omega)$ in (11) as $\tilde{L}_\pi(\omega) = \|\mu - \tilde{\mu}_\pi(\omega)\|^2, \tilde{R}_\pi(\omega) = E\{\tilde{L}_\pi(\omega)|\mathbf{X}\}$, respectively. Then, similarly, we can obtain $\tilde{R}_\pi(\omega) = \|\tilde{\mathbf{A}}(\omega)\mu\|^2 + \mathrm{tr}\{\tilde{\mathbf{P}}(\omega)\mathbf{\Omega}_\pi\tilde{\mathbf{P}}(\omega)^\top\}$ where $\tilde{\mathbf{A}}(\omega) = \mathbf{I}_n - \tilde{\mathbf{P}}(\omega)$. Define

$$V_\pi(\omega) = \|\mathbf{A}(\omega)\mu\|^2 + \mathrm{tr}\{\mathbf{P}(\omega)\mathbf{\Omega}_\pi\mathbf{P}(\omega)\} \quad \text{and}$$
$$\tilde{V}_\pi(\omega) = \|\tilde{\mathbf{A}}(\omega)\mu\|^2 + \mathrm{tr}[\tilde{\mathbf{P}}(\omega)\mathbf{\Omega}_\pi\{\tilde{\mathbf{P}}(\omega)\}^\top].$$

Then, we have

$$\begin{aligned}
&\frac{L_{\hat{\pi}}(\hat{\omega}_{cv})}{\inf_{\omega\in\mathscr{H}_n} L_{\hat{\pi}}(\omega)} - 1 \\
&= \sup_{\omega\in\mathscr{H}_n}\left\{\frac{L_{\hat{\pi}}(\hat{\omega}_{cv})}{L_{\hat{\pi}}(\omega)} - 1\right\} \\
&= \sup_{\omega\in\mathscr{H}_n}\left\{\frac{L_{\hat{\pi}}(\hat{\omega}_{cv})}{V_\pi(\hat{\omega}_{cv})}\frac{V_\pi(\hat{\omega}_{cv})}{\tilde{V}_\pi(\hat{\omega}_{cv})}\frac{\tilde{V}_\pi(\hat{\omega}_{cv})}{\tilde{L}_\pi(\hat{\omega}_{cv})}\frac{\tilde{L}_\pi(\hat{\omega}_{cv})}{\tilde{L}_\pi(\omega)}\frac{\tilde{L}_\pi(\omega)}{\tilde{R}_\pi(\omega)}\frac{\tilde{R}_\pi(\omega)}{R_\pi(\omega)}\frac{R_\pi(\omega)}{L_{\hat{\pi}}(\omega)} - 1\right\} \\
&\le \sup_{\omega\in\mathscr{H}_n}\left(\frac{L_{\hat{\pi}}(\omega)}{R_\pi(\omega)}\right)\sup_{\omega\in\mathscr{H}_n}\left(\frac{R_\pi(\omega)}{\tilde{R}_\pi(\omega)}\right)\sup_{\omega\in\mathscr{H}_n}\left(\frac{\tilde{R}_\pi(\omega)}{\tilde{L}_\pi(\omega)}\right)\sup_{\omega\in\mathscr{H}_n}\left(\frac{\tilde{L}_\pi(\omega)}{\tilde{R}_\pi(\omega)}\right) \\
&\quad\times \sup_{\omega\in\mathscr{H}_n}\left(\frac{\tilde{R}_\pi(\omega)}{R_\pi(\omega)}\right)\sup_{\omega\in\mathscr{H}_n}\left(\frac{R_\pi(\omega)}{L_{\hat{\pi}}(\omega)}\right)\frac{\tilde{L}_\pi(\hat{\omega}_{cv})}{\inf_{\omega\in\mathscr{H}_n}\tilde{L}_\pi(\omega)} - 1.
\end{aligned}$$

Therefore, to prove Theorem 1, it suffices to show that

$$\sup_{\omega\in\mathscr{H}_n}\left|\frac{\tilde{R}_\pi(\omega)}{R_\pi(\omega)} - 1\right| = o_p(1), \tag{A.9}$$

$$\sup_{\omega\in\mathscr{H}_n}\left|\frac{\tilde{L}_\pi(\omega)}{\tilde{R}_\pi(\omega)} - 1\right| = o_p(1), \tag{A.10}$$

$$\sup_{\omega\in\mathscr{H}_n}\left|\frac{L_{\hat{\pi}}(\omega)}{R_\pi(\omega)} - 1\right| = o_p(1), \tag{A.11}$$

$$\frac{\tilde{L}_\pi(\hat{\omega}_{cv})}{\inf_{\omega\in\mathscr{H}_n}\tilde{L}_\pi(\omega)} - 1 = o_p(1). \tag{A.12}$$

Under (C2)–(C4), it can be verified that the assumptions of Zhang (2021, Thm. 2) are satisfied for the dataset $\{\mathrm{Z}_{\pi,i}, \mathrm{X}_i\}_{i=1}^n$. Thus, by Zhang (2021), it can be shown that (A.9) and (A.10) are valid.

Proof of (A.11): By Cauchy–Schwarz inequality, it can be shown that

$$\left|\frac{L_{\hat{\pi}}(\omega)}{R_\pi(\omega)} - 1\right| \le \left|\frac{L_\pi(\omega)}{R_\pi(\omega)} - 1\right| + \frac{2\{L_\pi(\omega)\}^{1/2}\cdot\|\hat{\mu}_\pi(\omega) - \hat{\mu}_{\hat{\pi}}(\omega)\|}{R_\pi(\omega)} + \frac{\|\hat{\mu}_\pi(\omega) - \hat{\mu}_{\hat{\pi}}(\omega)\|^2}{R_\pi(\omega)}.$$

Thus, to prove (A.11), it suffices to show that

$$\sup_{\omega\in\mathscr{H}_n}\left|\frac{L_\pi(\omega)}{R_\pi(\omega)} - 1\right| = o_p(1), \tag{A.13}$$

$$\sup_{\omega\in\mathscr{H}_n}\frac{\|\hat{\mu}_\pi(\omega) - \hat{\mu}_{\hat{\pi}}(\omega)\|^2}{R_\pi(\omega)} = o_p(1). \tag{A.14}$$

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

Similar to (A.9), by Zhang (2021), we know (A.13) is valid. By Lemma 2, Lemma 6, and (C3), we have

$$\begin{aligned}\sup_{\omega\in\mathscr{H}_n}\frac{\|\hat{\mu}_\pi(\omega)-\hat{\mu}_{\hat{\pi}}(\omega)\|^2}{R_\pi(\omega)} &= \sup_{\omega\in\mathscr{H}_n}\frac{(\mathrm{Z}_{\hat{\pi}}-\mathrm{Z}_\pi)^\top\mathbf{P}(\omega)^\top\mathbf{P}(\omega)(\mathrm{Z}_{\hat{\pi}}-\mathrm{Z}_\pi)}{R_\pi(\omega)}\\ &\le \xi_\pi^{-1}\sup_{\omega\in\mathscr{H}_n}\lambda_{\max}\left\{\mathbf{P}(\omega)^\top\mathbf{P}(\omega)\right\}\|\mathrm{Z}_{\hat{\pi}}-\mathrm{Z}_\pi\|^2\\ &\le \xi_\pi^{-1}\|\mathrm{Z}_{\hat{\pi}}-\mathrm{Z}_\pi\|^2\\ &= o_p(1).\end{aligned}$$

As a result, (A.14) is valid and hence (A.11) is valid.

Proof of (A.12): In fact, by Zhang (2021), we can obtain $\sup_{\omega\in\mathscr{H}_n}\left|\tilde{R}_\pi(\omega)/R_\pi(\omega)-1\right| \xrightarrow{\text{a.s.}} 0$, a stronger result than (A.9). This together with (C3) leads us to

$$\tilde{\xi}_\pi^{-1}M_n^2=\xi_\pi^{-1}M_n^2\xi_\pi\tilde{\xi}_\pi^{-1}\le\xi_\pi^{-1}M_n^2\left\{\sup_{\omega\in\mathscr{H}_n}\left|\frac{R_\pi(\omega)}{\tilde{R}_\pi(\omega)}-1\right|+1\right\}\xrightarrow{\text{a.s.}}0, \tag{A.15}$$

where $\tilde{\xi}_\pi=\inf_{\omega\in\mathscr{H}_n}\tilde{R}_\pi(\omega)$. According to (12), after some careful calculations, we have $\mathrm{CV}_{\hat{\pi}}(\omega)=\tilde{L}_\pi(\omega)+\tilde{a}_n(\omega)+\|\mathrm{Z}_{\hat{\pi}}-\mu\|^2$ where

$$\begin{aligned}\tilde{a}_n(\omega)=&\|\tilde{\mu}_\pi(\omega)-\tilde{\mu}_{\hat{\pi}}(\omega)\|^2+2\left(\mathrm{Z}_{\hat{\pi}}-\mathrm{Z}_\pi\right)^\top\{\mu-\tilde{\mu}_\pi(\omega)\}+2\mathrm{e}_\pi^\top\tilde{\mathbf{A}}(\omega)\mu-2\mathrm{e}_\pi^\top\tilde{\mathbf{P}}(\omega)\mathrm{e}_\pi\\ &+2\mathrm{e}_\pi^\top\tilde{\mathbf{P}}(\omega)\left(\mathrm{Z}_\pi-\mathrm{Z}_{\hat{\pi}}\right)+2\{\mu-\tilde{\mu}_\pi(\omega)\}^\top\left\{\tilde{\mu}_\pi(\omega)-\tilde{\mu}_{\hat{\pi}}(\omega)\right\}.\end{aligned}$$

Therefore, by (13), Lemma 5, (A.10), and (A.15), we know that (A.12) is valid if the following holds:

$$\sup_{\omega\in\mathscr{H}_n}\frac{|\tilde{a}_n(\omega)|}{\tilde{R}_\pi(\omega)}=o_p(1). \tag{A.16}$$

Note that

$$\tilde{\mathbf{P}}_{(m)}=\mathbf{D}_{(m)}(\mathbf{P}_{(m)}-\mathbf{I}_n)+\mathbf{I}_n=\mathbf{P}_{(m)}-\mathbf{Q}_{(m)}\mathbf{A}_{(m)}, \tag{A.17}$$

where $\mathbf{Q}_{(m)}=\mathbf{D}_{(m)}-\mathbf{I}_n$ and $\mathbf{A}_{(m)}=\mathbf{I}_n-\mathbf{P}_{(m)}$. This together with Lemma 6 and the idempotent symmetric property of $\mathbf{P}_{(m)}$ leads us to

$$\begin{aligned}\lambda_{\max}\{\tilde{\mathbf{P}}(\omega)\}&\le\sum_{m=1}^{M_n}\omega_m\lambda_{\max}\left\{\tilde{\mathbf{P}}_{(m)}\right\}\\ &\le\sum_{m=1}^{M_n}\omega_m\left[\lambda_{\max}\left\{\mathbf{P}_{(m)}\right\}+\lambda_{\max}\left\{-\mathbf{Q}_{(m)}\mathbf{A}_{(m)}\right\}\right]\\ &\le\sum_{m=1}^{M_n}\omega_m\left[\lambda_{\max}\left\{\mathbf{P}_{(m)}\right\}+\lambda_{\max}\left\{-\mathbf{Q}_{(m)}\right\}\lambda_{\max}\left\{\mathbf{A}_{(m)}\right\}\right]\\ &\le\sum_{m=1}^{M_n}\omega_m\left[1+\max_{1\le i\le n}\frac{h_{m,ii}}{1-h_{m,ii}}\right]\\ &\le(1-\bar{h})^{-1}.\end{aligned}$$

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

Therefore, by Cauchy–Schwarz inequality and Lemma 6, we know that

$$|\tilde{a}_n(\omega)| \le \|\tilde{\mu}_\pi(\omega) - \tilde{\mu}_{\hat{\pi}}(\omega)\|^2 + 2\|Z_{\hat{\pi}} - Z_\pi\| \left\{\tilde{L}_\pi(\omega)\right\}^{\frac{1}{2}} + 2\left|e_\pi^\top \tilde{\mathbf{A}}(\omega)\mu\right| + 2\left|e_\pi^\top \tilde{\mathbf{P}}(\omega)e_\pi\right|$$
$$+ 2\|\tilde{\mathbf{P}}(\omega)^\top e_\pi\| \left\|Z_{\hat{\pi}} - Z_\pi\right\| + 2\left\{\tilde{L}_\pi(\omega)\right\}^{\frac{1}{2}} \|\tilde{\mu}_\pi(\omega) - \tilde{\mu}_{\hat{\pi}}(\omega)\|,$$
$$\|\tilde{\mu}_\pi(\omega) - \tilde{\mu}_{\hat{\pi}}(\omega)\|^2 = \|\tilde{\mathbf{P}}(\omega)\left(Z_{\hat{\pi}} - Z_\pi\right)\|^2 \le \left[\lambda_{\max}\left\{\tilde{\mathbf{P}}(\omega)\right\}\right]^2 \left\|Z_{\hat{\pi}} - Z_\pi\right\|^2 \tag{A.18}$$
$$\le \{1 - \bar{h}\}^{-2} \left\|Z_{\hat{\pi}} - Z_\pi\right\|^2.$$

This together with (A.10), (C4), Lemma 2, and (A.15) implies that, to prove (A.16), we only need to prove

$$\sup_{\omega \in \mathscr{H}_n} \frac{\left|e_\pi^\top \tilde{\mathbf{A}}(\omega)\mu\right|}{\tilde{R}_\pi(\omega)} = o_p(1), \tag{A.19}$$

$$\sup_{\omega \in \mathscr{H}_n} \frac{\left|e_\pi^\top \tilde{\mathbf{P}}(\omega)e_\pi\right|}{\tilde{R}_\pi(\omega)} = o_p(1), \tag{A.20}$$

$$\sup_{\omega \in \mathscr{H}_n} \frac{\|\tilde{\mathbf{P}}(\omega)^\top e_\pi\|}{\tilde{R}_\pi(\omega)} = o_p(1). \tag{A.21}$$

By the proofs of (A.46) and (A.48) in Zhang (2021), it can be shown that (A.19) and (A.20) are valid. Since

$$\|\tilde{\mathbf{P}}(\omega)^\top e_\pi\|^2 \le \left|e_\pi^\top \tilde{\mathbf{P}}(\omega)\tilde{\mathbf{P}}(\omega)^\top e_\pi - \mathrm{tr}\{\tilde{\mathbf{P}}(\omega)^\top \Omega_\pi \tilde{\mathbf{P}}(\omega)\}\right| + \mathrm{tr}\{\tilde{\mathbf{P}}(\omega)^\top \mathbf{\Omega}_\pi \tilde{\mathbf{P}}(\omega)\},$$

to prove (A.21), we only need to prove

$$\sup_{\omega \in \mathscr{H}_n} \frac{\left|e_\pi^\top \tilde{\mathbf{P}}(\omega)\tilde{\mathbf{P}}(\omega)^\top e_\pi - \mathrm{tr}\{\tilde{\mathbf{P}}(\omega)^\top \mathbf{\Omega}_\pi \tilde{\mathbf{P}}(\omega)\}\right|}{\left\{\tilde{R}_\pi(\omega)\right\}^2} = o_p(1), \tag{A.22}$$

$$\sup_{\omega \in \mathscr{H}_n} \frac{\mathrm{tr}\{\tilde{\mathbf{P}}(\omega)^\top \mathbf{\Omega}_\pi \tilde{\mathbf{P}}(\omega)\}}{\left\{\tilde{R}_\pi(\omega)\right\}^2} = o_p(1). \tag{A.23}$$

By the proof of (A.49) in Zhang (2021), it can be shown that (A.22) is valid. Similar to (A.15), we have

$$\bar{h}\bar{k}\tilde{\xi}_\pi^{-1} \xrightarrow{\text{a.s.}} 0. \tag{A.24}$$

Due to (A.17), Lemma 6 and the fact that the diagonal elements of $\tilde{\mathbf{P}}_{(m)}$ are all zeros, we have

$$\mathrm{tr}\{\tilde{\mathbf{P}}_{(m)}^\top \tilde{\mathbf{P}}_{(m)}\} = \mathrm{tr}\{(\mathbf{P}_{(m)} - \mathbf{I}_n)\mathbf{D}_{(m)}\tilde{\mathbf{P}}_{(m)}\}$$
$$= \mathrm{tr}\{\mathbf{P}_{(m)}\mathbf{D}_{(m)}\tilde{\mathbf{P}}_{(m)}\}$$
$$= \mathrm{tr}\{\mathbf{P}_{(m)}\mathbf{D}_{(m)}\mathbf{D}_{(m)}(\mathbf{P}_{(m)} - \mathbf{I}_n)\} + \mathrm{tr}\{\mathbf{P}_{(m)}\mathbf{D}_{(m)}\}$$

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

$$\begin{aligned}
&\le \mathrm{tr}\{\mathbf{P}_{(m)}\mathbf{D}_{(m)}\mathbf{D}_{(m)}\mathbf{P}_{(m)}\} + \mathrm{tr}\{\mathbf{P}_{(m)}\mathbf{D}_{(m)}\} \\
&\le (1-\bar{h})^{-2}(2-\bar{h})\mathrm{tr}\{\mathbf{P}_{(m)}\} \\
&= (1-\bar{h})^{-2}(2-\bar{h})k_m. \qquad \textbf{(A.25)}
\end{aligned}$$

Similar to the proof of (A.51) in Zhang (2021), we can show that

$$\tilde{R}_\pi(\omega) \ge \omega^\top \left(\tilde{\mathbf{\Phi}} + \min_{1\le i\le n} \sigma_{\pi,i}^2 \mathbf{\Psi}_o\right)\omega - 8 \max_{1\le i\le n} \sigma_{\pi,i}^2 \bar{h}(1-\bar{h})^{-1}\bar{k}, \qquad \textbf{(A.26)}$$

where $\tilde{\mathbf{\Phi}} = \tilde{\mathbf{G}}^\top\tilde{\mathbf{G}}$, $\tilde{\mathbf{G}} = (\tilde{\mathbf{A}}_{(1)}\mu, \tilde{\mathbf{A}}_{(2)}\mu, \ldots, \tilde{\mathbf{A}}_{(M_n)}\mu)$, $\tilde{\mathbf{A}}_{(m)} = \mathbf{I}_n - \tilde{\mathbf{P}}_{(m)}$ and $\mathbf{\Psi}_o$ is an $M_n \times M_n$ diagonal matrix with mth diagonal element k_m. Let $\mathbf{T}_{B,\pi}$ be an $M_n \times M_n$ matrix with msth element $\mathrm{tr}\{\tilde{\mathbf{P}}_{(m)}^\top \mathbf{\Omega}_\pi \tilde{\mathbf{P}}_{(s)}\}$. Obviously, $\mathbf{T}_{B,\pi}$ is symmetric and thus there exists an $M_n \times M_n$ matrix $\mathbf{B}_\pi$ such that $\mathbf{T}_{B,\pi} = \mathbf{B}_\pi\mathbf{B}_\pi^\top = \sum_{m=1}^{M_n} \mathrm{b}_{\pi,(m)}\mathrm{b}_{\pi,(m)}^\top$ where $\mathrm{b}_{\pi,(m)}$ is the mth column of $\mathbf{B}_\pi$. By Lemma 3, Lemma 4, (C4), (A.15), (A.25), and (A.26), we have

$$\begin{aligned}
&\sup_{\omega\in\mathscr{H}_n} \frac{\mathrm{tr}\{\tilde{\mathbf{P}}(\omega)^\top\mathbf{\Omega}_\pi\tilde{\mathbf{P}}(\omega)\}}{\left\{\tilde{R}_\pi(\omega)\right\}^2} \\
&\le \tilde{\xi}_\pi^{-1} \sup_{\omega\in\mathscr{H}_n} \frac{\omega^\top\mathbf{T}_{B,\pi}\omega}{\tilde{R}_\pi(\omega)} \\
&\le \tilde{\xi}_\pi^{-1} \sum_{m=1}^{M_n} \sup_{\omega\in\mathscr{H}_n} \frac{\left(\omega^\top \mathrm{b}_{\pi,(m)}\right)^2}{\tilde{R}_\pi(\omega)} \\
&\le \tilde{\xi}_\pi^{-1} \sum_{m=1}^{M_n} \sup_{\omega\in\mathscr{H}_n} \frac{\left(\omega^\top \mathrm{b}_{\pi,(m)}\right)^2}{\tilde{R}_\pi(\omega) + 8\max_{1\le i\le n}\sigma_{\pi,i}^2\bar{h}(1-\bar{h})^{-1}\bar{k}} \\
&\qquad \cdot \frac{\tilde{R}_\pi(\omega) + 8\max_{1\le i\le n}\sigma_{\pi,i}^2\bar{h}(1-\bar{h})^{-1}\bar{k}}{\tilde{R}_\pi(\omega)} \\
&\le \tilde{\xi}_\pi^{-1} \sum_{m=1}^{M_n} \sup_{\omega\in\mathscr{H}_n} \frac{\left(\omega^\top \mathrm{b}_{\pi,(m)}\right)^2}{\omega^\top\left(\min_{1\le i\le n}\sigma_{\pi,i}^2\mathbf{\Psi}_o\right)\omega} \cdot \left(\sup_{\omega\in\mathscr{H}_n} \frac{8\max_{1\le i\le n}\sigma_{\pi,i}^2\bar{h}(1-\bar{h})^{-1}\bar{k}}{\tilde{R}_\pi(\omega)} + 1\right) \\
&\le \tilde{\xi}_\pi^{-1} \sum_{m=1}^{M_n} \mathrm{b}_{\pi,(m)}^\top \left(\min_{1\le i\le n}\sigma_{\pi,i}^2\mathbf{\Psi}_o\right)^{-1} \mathrm{b}_{\pi,(m)} \cdot \left(8\tilde{\xi}_\pi^{-1}\max_{1\le i\le n}\sigma_{\pi,i}^2\bar{h}(1-\bar{h})^{-1}\bar{k} + 1\right) \\
&= \tilde{\xi}_\pi^{-1}\mathrm{tr}\left\{\left(\min_{1\le i\le n}\sigma_{\pi,i}^2\mathbf{\Psi}_o\right)^{-1} \sum_{m=1}^{M_n} \mathrm{b}_{\pi,(m)}\mathrm{b}_{\pi,(m)}^\top\right\}(1+o_p(1)) \\
&= \tilde{\xi}_\pi^{-1}\left(\min_{1\le i\le n}\sigma_{\pi,i}^2\right)^{-1} \sum_{m=1}^{M_n} k_m^{-1}\mathrm{tr}\{\tilde{\mathbf{P}}_{(m)}^\top\mathbf{\Omega}_\pi\tilde{\mathbf{P}}_{(m)}\}(1+o_p(1)) \\
&\le \tilde{\xi}_\pi^{-1}\left(\min_{1\le i\le n}\sigma_{\pi,i}^2\right)^{-1} \max_{1\le i\le n}\sigma_{\pi,i}^2 \sum_{m=1}^{M_n} k_m^{-1}\mathrm{tr}\{\tilde{\mathbf{P}}_{(m)}^\top\tilde{\mathbf{P}}_{(m)}\}(1+o_p(1))
\end{aligned}$$

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

$$\le \tilde{\xi}_\pi^{-1}\left(\min_{1\le i\le n}\sigma_{\pi,i}^2\right)^{-1}\max_{1\le i\le n}\sigma_{\pi,i}^2(1-\bar{h})^{-2}(2-\bar{h})M_n(1+o_p(1))$$
$$= o_p(1).$$

Thus (A.23) is valid. This completes the proof of (A.12) and hence the proof of Theorem 1. □

A.3. Proof of Theorem 2

Proof. Let $\psi_n(\omega) = Z_{\hat{\pi}} - Z_\pi + \hat{\mu}_\pi(\omega) - \tilde{\mu}_\pi(\omega) + \tilde{\mu}_\pi(\omega) - \tilde{\mu}_{\hat{\pi}}(\omega)$. Then, according to (12), we know

$$\begin{aligned}
\mathrm{CV}_{\hat{\pi}}(\hat{\omega}_{cv}) &= \|Z_{\hat{\pi}} - Z_\pi + Z_\pi - \hat{\mu}_\pi(\hat{\omega}_{cv}) + \hat{\mu}_\pi(\hat{\omega}_{cv}) - \tilde{\mu}_\pi(\hat{\omega}_{cv}) + \tilde{\mu}_\pi(\hat{\omega}_{cv}) - \tilde{\mu}_{\hat{\pi}}(\hat{\omega}_{cv})\|^2 \\
&= \|\mu - \hat{\mu}_\pi(\hat{\omega}_{cv}) + e_\pi + \psi_n(\hat{\omega}_{cv})\|^2 \\
&= \Bigg\| \hat{s}_{cv}\sum_{m=1}^{m_0}\frac{\hat{\omega}_{cv,m}}{\hat{s}_{cv}}\{\mu - \hat{\mu}_{(\pi,m)}\} + (1-\hat{s}_{cv}) \\
&\quad \times \sum_{m=m_0+1}^{M_n}\frac{\hat{\omega}_{cv,m}}{1-\hat{s}_{cv}}\{\mu - \hat{\mu}_{(\pi,m)}\} + e_\pi + \psi_n(\hat{\omega}_{cv})\Bigg\|^2 \\
&= \|\hat{s}_{cv}\{\mu - \hat{\mu}_\pi(\hat{\omega}_C)\} + (1-\hat{s}_{cv})\{\mu - \hat{\mu}_\pi(\hat{\omega}_F)\} + e_\pi + \psi_n(\hat{\omega}_{cv})\|^2, \qquad \textbf{(A.27)}
\end{aligned}$$

where $\hat{\omega}_C = (\hat{\omega}_{cv,1},\ldots,\hat{\omega}_{cv,m_0},0,\ldots,0)/\hat{s}_{cv} \in \mathscr{H}_n$ and $\hat{\omega}_F = (0,\ldots,0,\hat{\omega}_{cv,m_0+1},\ldots,\hat{\omega}_{cv,M_n})/(1-\hat{s}_{cv}) \in \mathscr{H}_n$. Similarly, we have

$$\mathrm{CV}_{\hat{\pi}}(\hat{\omega}_C) = \|\mu - \hat{\mu}_\pi(\hat{\omega}_C) + e_\pi + \psi_n(\hat{\omega}_C)\|^2. \qquad \textbf{(A.28)}$$

According to (13), one can see that $\mathrm{CV}_{\hat{\pi}}(\hat{\omega}_{cv}) \le \mathrm{CV}_{\hat{\pi}}(\hat{\omega}_C)$. This together with (A.27), (A.28), and Cauchy–Schwarz inequality leads us to

$$\begin{aligned}
&(1-\hat{s}_{cv})^2 \\
&\le \Big[(1-\hat{s}_{cv}^2)\|\mu - \hat{\mu}_\pi(\hat{\omega}_C)\|^2 + \|\psi_n(\hat{\omega}_C)\|^2 + 2e_\pi^\top\left[\{\mu - \hat{\mu}_\pi(\hat{\omega}_C)\} + \psi_n(\hat{\omega}_C)\right] \\
&\quad + 2\{\mu - \hat{\mu}_\pi(\hat{\omega}_C)\}^\top\psi_n(\hat{\omega}_C) + 2\left[\hat{s}_{cv}\{\mu - \hat{\mu}_\pi(\hat{\omega}_C)\} + \psi_n(\hat{\omega}_{cv})\right]^\top\{\mu - \hat{\mu}_\pi(\hat{\omega}_F)\} \\
&\quad + \|\psi_n(\hat{\omega}_{cv})\|^2 + 2e_\pi^\top\{\mu - \hat{\mu}_\pi(\hat{\omega}_F)\} + 2e_\pi^\top\left[\hat{s}_{cv}\{\mu - \hat{\mu}_\pi(\hat{\omega}_C)\} + \psi_n(\hat{\omega}_{cv})\right] \\
&\quad + 2\hat{s}_{cv}\{\mu - \hat{\mu}_\pi(\hat{\omega}_C)\}^\top\psi_n(\hat{\omega}_{cv})\Big]/\|\mu - \hat{\mu}_\pi(\hat{\omega}_F)\|^2 \\
&\le \Big[2\|\mu - \hat{\mu}_\pi(\hat{\omega}_C)\|^2 + 2\sup_{\omega\in\mathscr{H}_n}\|\psi_n(\omega)\|^2 \\
&\quad + 4\left|e_\pi^\top\left\{\mu - \hat{\mu}_\pi(\hat{\omega}_C)\right\}\right| + 4\|e_\pi\|\sup_{\omega\in\mathscr{H}_n}\|\psi_n(\omega)\| \\
&\quad + 4\|\mu - \hat{\mu}_\pi(\hat{\omega}_C)\|\sup_{\omega\in\mathscr{H}_n}\|\psi_n(\omega)\|
\end{aligned}$$

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

$$
\begin{aligned}
&+2\Big\{\|\mu-\hat{\mu}_\pi(\hat{\omega}_C)\|+\sup_{\omega\in\mathscr{H}_n}\|\psi_n(\omega)\|\Big\}\|\mu-\hat{\mu}_\pi(\hat{\omega}_F)\|\\
&+2\mathrm{e}_\pi^\top\mathbf{A}(\hat{\omega}_F)\mu-2\mathrm{e}_\pi^\top\mathbf{P}(\hat{\omega}_F)\mathrm{e}_\pi\Big]\frac{1}{R_\pi(\hat{\omega}_F)}\frac{R_\pi(\hat{\omega}_F)}{L_\pi(\hat{\omega}_F)}\\
\le&\Bigg[\Big\{2\|\mu-\hat{\mu}_\pi(\hat{\omega}_C)\|^2+2\sup_{\omega\in\mathscr{H}_n}\|\psi_n(\omega)\|^2+4\Big|\mathrm{e}_\pi^\top\{\mu-\hat{\mu}_\pi(\hat{\omega}_C)\}\Big|\\
&+4\|\mathrm{e}_\pi\|\sup_{\omega\in\mathscr{H}_n}\|\psi_n(\omega)\|+4\|\mu-\hat{\mu}_\pi(\hat{\omega}_C)\|\sup_{\omega\in\mathscr{H}_n}\|\psi_n(\omega)\|\Big\}\xi_{\pi,F}^{-1}\\
&+2\frac{|\mathrm{e}_\pi^\top\mathbf{A}(\hat{\omega}_F)\mu|}{R_\pi(\hat{\omega}_F)}+2\frac{\mathrm{e}_\pi^\top\mathbf{P}(\hat{\omega}_F)\mathrm{e}_\pi}{R_\pi(\hat{\omega}_F)}\\
&+2\Big\{\|\mu-\hat{\mu}_\pi(\hat{\omega}_C)\|+\sup_{\omega\in\mathscr{H}_n}\|\psi_n(\omega)\|\Big\}\\
&\times\Big\{\frac{L_\pi(\hat{\omega}_F)}{R_\pi(\hat{\omega}_F)}\Big\}^{1/2}\xi_{\pi,F}^{-1/2}\Bigg]\sup_{\omega\in\mathscr{H}_F}\Bigg[\Big|\frac{R_\pi(\omega)}{L_\pi(\omega)}-1\Big|+1\Bigg]\\
\le&\Bigg[\Big\{2\|\mu-\hat{\mu}_\pi(\hat{\omega}_C)\|^2+2\sup_{\omega\in\mathscr{H}_n}\|\psi_n(\omega)\|^2\\
&+4\Big|\mathrm{e}_\pi^\top\{\mu-\hat{\mu}_\pi(\hat{\omega}_C)\}\Big|+4\|\mathrm{e}_\pi\|\sup_{\omega\in\mathscr{H}_n}\|\psi_n(\omega)\|\\
&+4\|\mu-\hat{\mu}_\pi(\hat{\omega}_C)\|\sup_{\omega\in\mathscr{H}_n}\|\psi_n(\omega)\|\Big\}\xi_{\pi,F}^{-1}+2\sup_{\omega\in\mathscr{H}_F}\frac{|\mathrm{e}_\pi^\top\mathbf{A}(\omega)\mu|}{R_\pi(\omega)}+2\sup_{\omega\in\mathscr{H}_F}\frac{\mathrm{e}_\pi^\top\mathbf{P}(\omega)\mathrm{e}_\pi}{R_\pi(\omega)}\\
&+2\xi_{\pi,F}^{-1/2}\Big\{\|\mu-\hat{\mu}_\pi(\hat{\omega}_C)\|+\sup_{\omega\in\mathscr{H}_n}\|\psi_n(\omega)\|\Big\}\Big\{\sup_{\omega\in\mathscr{H}_F}\Big|\frac{L_\pi(\omega)}{R_\pi(\omega)}-1\Big|+1\Big\}^{1/2}\Bigg]\\
&\times\sup_{\omega\in\mathscr{H}_F}\Big\{\Big|\frac{R_\pi(\omega)}{L_\pi(\omega)}-1\Big|+1\Big\}.
\end{aligned}
$$

This together with Condition (C5) indicates that, to prove Theorem 2, it suffices to prove

$$\xi_{\pi,F}^{-1}\|\mu-\hat{\mu}_\pi(\hat{\omega}_C)\|^2=o_p(1), \tag{A.29}$$

$$\xi_{\pi,F}^{-1}\Big|\mathrm{e}_\pi^\top\{\mu-\hat{\mu}_\pi(\hat{\omega}_C)\}\Big|=o_p(1), \tag{A.30}$$

$$\sup_{\omega\in\mathscr{H}_n}\|\psi_n(\omega)\|^2=O_p(1), \tag{A.31}$$

$$\xi_{\pi,F}^{-2}\|\mathrm{e}_\pi\|^2=o_p(1), \tag{A.32}$$

$$\sup_{\omega\in\mathscr{H}_F}\Big|\frac{L_\pi(\omega)}{R_\pi(\omega)}-1\Big|=o_p(1), \tag{A.33}$$

$$\sup_{\omega\in\mathscr{H}_F}\frac{|\mathrm{e}_\pi^\top\mathbf{A}(\omega)\mu|}{R_\pi(\omega)}=o_p(1), \tag{A.34}$$

$$\sup_{\omega\in\mathscr{H}_F}\frac{\mathrm{e}_\pi^\top\mathbf{P}(\omega)\mathrm{e}_\pi}{R_\pi(\omega)}=o_p(1). \tag{A.35}$$

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

Proofs of (A.29) and (A.30): By the definition of the correct model presented at the beginning of Section 3, we know that $\mathbf{P}_{(m)}\mu = \mathbf{P}_{(m)}\mathbf{X}_{(m)}\theta^*_{(m)} = \mathbf{X}_{(m)}\theta^*_{(m)} = \mu$ for $m = 1, 2, \ldots, m_0$. This indicates that

$$\begin{aligned}
\|\mu - \hat{\mu}_\pi(\hat{\omega}_C)\|^2 &= \left\| \sum_{m=1}^{m_0} \frac{\hat{\omega}_{cv,m}}{\hat{s}_{cv}} \{\mu - \mathbf{P}_{(m)} Z_\pi\} \right\|^2 \\
&= \left\| \sum_{m=1}^{m_0} \frac{\hat{\omega}_{cv,m}}{\hat{s}_{cv}} \mathbf{P}_{(m)} \mathrm{e}_\pi \right\|^2 \\
&= \sum_{m=1}^{m_0} \frac{\hat{\omega}_{cv,m}}{\hat{s}_{cv}} \sum_{t=1}^{m_0} \frac{\hat{\omega}_{cv,t}}{\hat{s}_{cv}} \mathrm{e}_\pi^\top \mathbf{P}_{(m)} \mathbf{P}_{(t)} \mathrm{e}_\pi \\
&\le \frac{1}{2} \sum_{m=1}^{m_0} \frac{\hat{\omega}_{cv,m}}{\hat{s}_{cv}} \sum_{t=1}^{m_0} \frac{\hat{\omega}_{cv,t}}{\hat{s}_{cv}} \left(\mathrm{e}_\pi^\top \mathbf{P}_{(m)} \mathbf{P}_{(m)} \mathrm{e}_\pi + \mathrm{e}_\pi^\top \mathbf{P}_{(t)} \mathbf{P}_{(t)} \mathrm{e}_\pi \right) \\
&= \sum_{m=1}^{m_0} \frac{\hat{\omega}_{cv,m}}{\hat{s}_{cv}} \mathrm{e}_\pi^\top \mathbf{P}_{(m)} \mathrm{e}_\pi \\
&\le \max_{1 \le m \le m_0} \mathrm{e}_\pi^\top \mathbf{P}_{(m)} \mathrm{e}_\pi .
\end{aligned}$$

Therefore, to prove (A.29), we only need to prove

$$\xi_{\pi,F}^{-1} \max_{1 \le m \le m_0} \mathrm{e}_\pi^\top \mathbf{P}_{(m)} \mathrm{e}_\pi = o_p(1). \tag{A.36}$$

From Lemma 2, we know that, for every $\epsilon > 0$, there exists an integer N_ϵ such that $P(\max_{1 \le i \le n} \sigma^2_{\pi,i} > c^2_{\pi,2}) \le \epsilon/2$, for all $n \ge N_\epsilon$. Let $M_\epsilon = 2c^2_{\pi,2}/\epsilon$. Then, by Markov's inequality, we have

$$\begin{aligned}
&\sup_{n \ge 1} P\left(\frac{1}{m_0 \bar{k}} \max_{1 \le m \le m_0} \mathrm{e}_\pi^\top \mathbf{P}_{(m)} \mathrm{e}_\pi > M_\epsilon \right) \\
&= \sup_{n \ge 1} \left\{ P\left(\frac{1}{m_0 \bar{k}} \max_{1 \le m \le m_0} \mathrm{e}_\pi^\top \mathbf{P}_{(m)} \mathrm{e}_\pi > M_\epsilon, \max_{1 \le i \le n} \sigma^2_{\pi,i} \le c^2_{\pi,2} \right) \right. \\
&\quad \left. + P\left(\frac{1}{m_0 \bar{k}} \max_{1 \le m \le m_0} \mathrm{e}_\pi^\top \mathbf{P}_{(m)} \mathrm{e}_\pi > M_\epsilon, \max_{1 \le i \le n} \sigma^2_{\pi,i} > c^2_{\pi,2} \right) \right\} \\
&\le \sup_{n \ge 1} \sum_{m=1}^{m_0} P\left(\frac{1}{m_0 \bar{k}} \mathrm{e}_\pi^\top \mathbf{P}_{(m)} \mathrm{e}_\pi > M_\epsilon, \max_{1 \le i \le n} \sigma^2_{\pi,i} \le c^2_{\pi,2} \right) + \sup_{n \ge 1} P\left(\max_{1 \le i \le n} \sigma^2_{\pi,i} > c^2_{\pi,2} \right) \\
&\le \sup_{n \ge 1} \sum_{m=1}^{m_0} E\left\{ I\left(\frac{1}{m_0 \bar{k}} \mathrm{e}_\pi^\top \mathbf{P}_{(m)} \mathrm{e}_\pi > M_\epsilon \right) I\left(\max_{1 \le i \le n} \sigma^2_{\pi,i} \le c^2_{\pi,2} \right) \right\} \\
&\quad + \sup_{n \ge N_\epsilon} P\left(\max_{1 \le i \le n} \sigma^2_{\pi,i} > c^2_{\pi,2} \right) \\
&\le \sup_{n \ge 1} \sum_{m=1}^{m_0} E\left[E\left\{ I\left(\frac{1}{m_0 \bar{k}} \mathrm{e}_\pi^\top \mathbf{P}_{(m)} \mathrm{e}_\pi > M_\epsilon \right) \middle| \mathbf{X} \right\} I\left(\max_{1 \le i \le n} \sigma^2_{\pi,i} \le c^2_{\pi,2} \right) \right] + \frac{\epsilon}{2}
\end{aligned}$$

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

$$
\begin{aligned}
&\le \sup_{n\ge 1}\sum_{m=1}^{m_0} E\left\{P\left(\frac{1}{m_0\bar{k}}\mathrm{e}_\pi^\top \mathbf{P}_{(m)}\mathrm{e}_\pi > M_\epsilon \middle| \mathbf{X}\right) I\left(\max_{1\le i\le n}\sigma_{\pi,i}^2 \le c_{\pi,2}^2\right)\right\} + \frac{\epsilon}{2}\\
&\le \sup_{n\ge 1}\sum_{m=1}^{m_0} E\left\{\frac{1}{m_0\bar{k}} M_\epsilon^{-1} E\left(\mathrm{e}_\pi^\top \mathbf{P}_{(m)}\mathrm{e}_\pi \middle| \mathbf{X}\right) I\left(\max_{1\le i\le n}\sigma_{\pi,i}^2 \le c_{\pi,2}^2\right)\right\} + \frac{\epsilon}{2}\\
&\le \sup_{n\ge 1}\sum_{m=1}^{m_0} E\left\{\frac{1}{m_0\bar{k}} M_\epsilon^{-1}\mathrm{tr}\{\mathbf{P}_{(m)}\mathbf{\Omega}_\pi\} I\left(\max_{1\le i\le n}\sigma_{\pi,i}^2 \le c_{\pi,2}^2\right)\right\} + \frac{\epsilon}{2}\\
&\le \sup_{n\ge 1} E\left\{\frac{1}{m_0\bar{k}} m_0 M_\epsilon^{-1} k_m \max_{1\le i\le n}\sigma_{\pi,i}^2 I\left(\max_{1\le i\le n}\sigma_{\pi,i}^2 \le c_{\pi,2}^2\right)\right\} + \frac{\epsilon}{2}\\
&\le M_\epsilon^{-1} c_{\pi,2}^2 + \frac{\epsilon}{2}\\
&= \epsilon.
\end{aligned}
$$

Thus, we have $(m_0\bar{k})^{-1}\max_{1\le m\le m_0}\mathrm{e}_\pi^\top \mathbf{P}_{(m)}\mathrm{e}_\pi = O_p(1)$. This together with Condition (C5) indicates (A.36) which completes the proof of (A.29). Note that

$$
\left|\mathrm{e}_\pi^\top\{\mu - \hat{\mu}_\pi(\hat{\omega}_C)\}\right| = \left|\sum_{m=1}^{m_0}\frac{\hat{\omega}_{cv,m}}{\hat{s}_{cv}}\mathrm{e}_\pi^\top \mathbf{P}_{(m)}\mathrm{e}_\pi\right| \le \max_{1\le m\le m_0}\mathrm{e}_\pi^\top \mathbf{P}_{(m)}\mathrm{e}_\pi .
$$

Therefore, from the above proof of (A.29), we know that (A.30) is valid.

Proof of (A.31): By Cauchy–Schwarz inequality, Lemma 2, (A.18), and (C6), we know that

$$
\begin{aligned}
\sup_{\omega\in\mathscr{H}_n}\|\psi_n(\omega)\|^2 &= \sup_{\omega\in\mathscr{H}_n}\|\mathrm{Z}_{\hat{\pi}} - \mathrm{Z}_\pi + \hat{\mu}_\pi(\omega) - \tilde{\mu}_\pi(\omega) + \tilde{\mu}_\pi(\omega) - \tilde{\mu}_{\hat{\pi}}(\omega)\|^2\\
&\le 2\sup_{\omega\in\mathscr{H}_n}\left\{\|\mathrm{Z}_{\hat{\pi}} - \mathrm{Z}_\pi\|^2 + \|\hat{\mu}_\pi(\omega) - \tilde{\mu}_\pi(\omega)\|^2 + \|\tilde{\mu}_\pi(\omega) - \tilde{\mu}_{\hat{\pi}}(\omega)\|^2\right\}\\
&\le 2\{1 + (1-\bar{h})^{-2}\}\|\mathrm{Z}_{\hat{\pi}} - \mathrm{Z}_\pi\|^2 + 2\sup_{\omega\in\mathscr{H}_n}\|\hat{\mu}_\pi(\omega) - \tilde{\mu}_\pi(\omega)\|^2\\
&= O_p(1) + 2\sup_{\omega\in\mathscr{H}_n}\|\hat{\mu}_\pi(\omega) - \tilde{\mu}_\pi(\omega)\|^2. \qquad \textbf{(A.37)}
\end{aligned}
$$

By (A.17), Lemma 6, and Cauchy–Schwarz inequality, one can obtain that

$$
\begin{aligned}
&\sup_{\omega\in\mathscr{H}_n}\|\hat{\mu}_\pi(\omega) - \tilde{\mu}_\pi(\omega)\|^2\\
&\quad= \sup_{\omega\in\mathscr{H}_n}\|\mathbf{P}(\omega)\mathrm{Z}_\pi - \tilde{\mathbf{P}}(\omega)\mathrm{Z}_\pi\|^2\\
&\quad= \sup_{\omega\in\mathscr{H}_n}\left\|\sum_{m=1}^{M_n}\omega_m \mathbf{Q}_{(m)}\mathbf{A}_{(m)}\mathrm{Z}_\pi\right\|^2\\
&\quad= \sup_{\omega\in\mathscr{H}_n}\sum_{m=1}^{M_n}\omega_m\sum_{s=1}^{M_n}\omega_s \mathrm{Z}_\pi^\top \mathbf{A}_{(m)}\mathbf{Q}_{(m)}\mathbf{Q}_{(s)}\mathbf{A}_{(s)}\mathrm{Z}_\pi
\end{aligned}
$$

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

$$\le \frac{1}{2}\sup_{\omega\in\mathscr{H}_n}\sum_{m=1}^{M_n}\omega_m\sum_{s=1}^{M_n}\omega_s\left(\mathrm{Z}_\pi^\top\mathbf{A}_{(m)}\mathbf{Q}_{(m)}\mathbf{Q}_{(m)}\mathbf{A}_{(m)}\mathrm{Z}_\pi+\mathrm{Z}_\pi^\top\mathbf{A}_{(s)}\mathbf{Q}_{(s)}\mathbf{Q}_{(s)}\mathbf{A}_{(s)}\mathrm{Z}_\pi\right)$$
$$=\sup_{\omega\in\mathscr{H}_n}\sum_{m=1}^{M_n}\omega_m\mathrm{Z}_\pi^\top\mathbf{A}_{(m)}\mathbf{Q}_{(m)}\mathbf{Q}_{(m)}\mathbf{A}_{(m)}\mathrm{Z}_\pi$$
$$\le\sup_{\omega\in\mathscr{H}_n}\sum_{m=1}^{M_n}\omega_m\lambda_{\max}\{\mathbf{A}_{(m)}\mathbf{Q}_{(m)}\mathbf{Q}_{(m)}\mathbf{A}_{(m)}\}\mathrm{Z}_\pi^\top\mathrm{Z}_\pi$$
$$\le 2(1-\bar{h})^{-2}\bar{h}^2 n\left(\frac{1}{n}\mu^\top\mu+\frac{1}{n}\mathrm{e}_\pi^\top\mathrm{e}_\pi\right).$$

This together with (C2), (C6), and (A.37) indicates that, to prove (A.31), we only need to prove

$$\frac{1}{n}\mathrm{e}_\pi^\top\mathrm{e}_\pi=O_p(1). \tag{A.38}$$

Similar to the proof of (A.2), one can show that (A.38) is valid and hence (A.31) is valid. By (A.38) and (C5), it is easy to see that (A.32) is valid.

Proofs of (A.33)–(A.35): Under Conditions (C5) and (C6), it is easy to see that

$$(M_n-m_0)^2\xi_{\pi,F}^{-1}\to 0,\quad \bar{h}\to 0,\quad \bar{h}\bar{k}\xi_{\pi,F}^{-1}\to 0,$$

almost surely. This implies that Conditions (C3) and (C4) are satisfied for $\{\mathrm{Z}_{\pi,i},\mathrm{X}_i\}_{i=1}^n$ with $\omega\in\mathscr{H}_F$. Hence, similar to (A.13), by Zhang (2021), we know (A.33)–(A.35) are valid. □

A.4. Proof of Theorem 3

Proof. We first introduce some notations with their properties as a preparation for the proofs of (i) and (ii) of Theorem 3. Let $\hat{a}_m=\mathrm{Z}_{\hat{\pi}}^\top(\mathbf{I}_n-\mathbf{P}_{(m)})\mathrm{Z}_{\hat{\pi}}$. For any $m>s$, we have $\mathbf{P}_{(m)}\mathbf{P}_{(s)}=\mathbf{P}_{(s)}\mathbf{P}_{(m)}=\mathbf{P}_{(m)}$. Thus, we have

$$\hat{a}_m-\hat{a}_s=\mathrm{Z}_{\hat{\pi}}^\top(\mathbf{P}_{(s)}-\mathbf{P}_{(m)})\mathrm{Z}_{\hat{\pi}}=\mathrm{Z}_{\hat{\pi}}^\top\mathbf{P}_{(m)}(\mathbf{P}_{(s)}-\mathbf{I}_n)\mathbf{P}_{(m)}\mathrm{Z}_{\hat{\pi}}\le 0,\quad m>s. \tag{A.39}$$

By Equation (A.8) in Zhang and Liu (2019), for any $m\in\{1,2,\ldots,m_o-1\}$, we know that $\mathrm{Z}_\pi^\top(\mathbf{P}_{(M_n)}-\mathbf{P}_{(m)})\mathrm{Z}_\pi=O_p(n)$ under (C7) and (C8). This together with Cauchy–Schwarz inequality, Lemmas 2 and 6 indicates that, for any $m\in\{1,2,\ldots,m_o-1\}$,

$$\begin{aligned}\hat{a}_m-\hat{a}_{M_n}&=\mathrm{Z}_{\hat{\pi}}^\top(\mathbf{P}_{(M_n)}-\mathbf{P}_{(m)})\mathrm{Z}_{\hat{\pi}}\\&=(\mathrm{Z}_{\hat{\pi}}-\mathrm{Z}_\pi+\mathrm{Z}_\pi)^\top(\mathbf{P}_{(M_n)}-\mathbf{P}_{(m)})(\mathrm{Z}_{\hat{\pi}}-\mathrm{Z}_\pi+\mathrm{Z}_\pi)\\&\le 2\left\{(\mathrm{Z}_{\hat{\pi}}-\mathrm{Z}_\pi)^\top(\mathbf{P}_{(M_n)}-\mathbf{P}_{(m)})(\mathrm{Z}_{\hat{\pi}}-\mathrm{Z}_\pi)+\mathrm{Z}_\pi^\top(\mathbf{P}_{(M_n)}-\mathbf{P}_{(m)})\mathrm{Z}_\pi\right\}\\&\le 2\lambda_{\max}\{\mathbf{P}_{(M_n)}-\mathbf{P}_{(m)}\}\|\mathrm{Z}_{\hat{\pi}}-\mathrm{Z}_\pi\|^2+O_p(n)\\&=O_p(n).\end{aligned} \tag{A.40}$$

By Equation (A.5) in Zhang and Liu (2019), for any $m\in\{1,2,\ldots,M_n\}$, we know that $\mathrm{e}_\pi^\top\mathbf{P}_{(m)}\mathrm{e}_\pi=O_p(1)$ under (C7) and (C8). This together with Cauchy–Schwarz inequality,

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

Lemmas 2 and 6 implies that, for any $m, s \in \{m_o, m_o+1, \ldots, M_n\}$,

$$\begin{aligned}\hat{a}_m - \hat{a}_s &= \mathrm{Z}_{\hat{\pi}}^\top (\mathbf{P}_{(s)} - \mathbf{P}_{(m)}) \mathrm{Z}_{\hat{\pi}} \\ &\le (\mathrm{Z}_{\hat{\pi}} - \mathrm{Z}_\pi + \mathrm{e}_\pi)^\top \mathbf{P}_{(s)} (\mathrm{Z}_{\hat{\pi}} - \mathrm{Z}_\pi + \mathrm{e}_\pi)^\top \\ &\le 2 \left\{ \mathrm{e}_\pi^\top \mathbf{P}_{(s)} \mathrm{e}_\pi + \|\mathrm{Z}_{\hat{\pi}} - \mathrm{Z}_\pi\|^2 \right\} \\ &= O_p(1). \end{aligned} \tag{A.41}$$

Let $\mathbf{X}_{(m)^c}$ be the matrix obtained by eliminating the columns of $\mathbf{X}_{(m)}$ in $\mathbf{X}$. Denote $\theta_{(m)^c}$ as the vector of regression coefficients corresponding to $\mathbf{X}_{(m)^c}$. If the mth model is correct, then $\theta_{(m)^c} = 0$. Let $\hat{\mathbf{\Upsilon}}, \mathbf{\Psi}$ and $\hat{\mathbf{\Gamma}}$ be $M_n \times M_n$ matrixes whose msth elements are $\hat{\mathbf{\Upsilon}}_{ms} = \hat{a}_{\max\{m,s\}}$,

$$\begin{aligned}\mathbf{\Psi}_{ms} &= \left(\mathrm{e}_\pi + \mathbf{X}_{(m)^c} \theta_{(m)^c}\right)^\top (\mathbf{I}_n - \mathbf{P}_{(m)})(\mathbf{Q}_{(m)} + \mathbf{Q}_{(s)} + \mathbf{Q}_{(m)}\mathbf{Q}_{(s)})(\mathbf{I}_n - \mathbf{P}_{(s)}) \\ &\quad \times \left(\mathrm{e}_\pi + \mathbf{X}_{(s)^c} \theta_{(s)^c}\right), \\ \hat{\mathbf{\Gamma}}_{ms} &= (\mathrm{Z}_{\hat{\pi}} - \mathrm{Z}_\pi)^\top (\mathbf{I}_n - \mathbf{P}_{(m)})(\mathbf{Q}_{(m)} + \mathbf{Q}_{(s)} + \mathbf{Q}_{(m)}\mathbf{Q}_{(s)})(\mathbf{I}_n - \mathbf{P}_{(s)}) \left(\mathrm{e}_\pi + \mathbf{X}_{(s)^c} \theta_{(s)^c}\right) \\ &\quad + \left(\mathrm{e}_\pi + \mathbf{X}_{(m)^c} \theta_{(m)^c}\right)^\top (\mathbf{I}_n - \mathbf{P}_{(m)})(\mathbf{Q}_{(m)} + \mathbf{Q}_{(s)} + \mathbf{Q}_{(m)}\mathbf{Q}_{(s)})(\mathbf{I}_n - \mathbf{P}_{(s)})(\mathrm{Z}_{\hat{\pi}} - \mathrm{Z}_\pi) \\ &\quad + (\mathrm{Z}_{\hat{\pi}} - \mathrm{Z}_\pi)^\top (\mathbf{I}_n - \mathbf{P}_{(m)})(\mathbf{Q}_{(m)} + \mathbf{Q}_{(s)} + \mathbf{Q}_{(m)}\mathbf{Q}_{(s)})(\mathbf{I}_n - \mathbf{P}_{(s)})(\mathrm{Z}_{\hat{\pi}} - \mathrm{Z}_\pi), \end{aligned}$$

respectively. $\mathbf{Q}_{(m)}$ is given right below (A.17). By Equation (A.14) in Zhang and Liu (2019), we know that, under (C7)–(C9),

$$\mathbf{\Psi}_{ms} = O_p(1), \quad m, s \in \{1, 2, \ldots, M_n\}. \tag{A.42}$$

By Cauchy–Schwarz inequality, (A.38), Lemmas 2 and 6, (C7), and (C9), we have

$$\begin{aligned}&\left|(\mathrm{Z}_{\hat{\pi}} - \mathrm{Z}_\pi)^\top (\mathbf{I}_n - \mathbf{P}_{(m)})(\mathbf{Q}_{(m)} + \mathbf{Q}_{(s)} + \mathbf{Q}_{(m)}\mathbf{Q}_{(s)})(\mathbf{I}_n - \mathbf{P}_{(s)}) \left(\mathrm{e}_\pi + \mathbf{X}_{(s)^c} \theta_{(s)^c}\right)\right| \\ &\le \|\mathrm{Z}_{\hat{\pi}} - \mathrm{Z}_\pi\| \lambda_{\max}\{(\mathbf{I}_n - \mathbf{P}_{(m)})(\mathbf{Q}_{(m)} + \mathbf{Q}_{(s)} + \mathbf{Q}_{(m)}\mathbf{Q}_{(s)})(\mathbf{I}_n - \mathbf{P}_{(s)})\} \|\mathrm{e}_\pi + \mathbf{X}_{(s)^c} \theta_{(s)^c}\| \\ &\le \|\mathrm{Z}_{\hat{\pi}} - \mathrm{Z}_\pi\| \{2\bar{h}(1-\bar{h})^{-1} + \bar{h}^2(1-\bar{h})^{-2}\} \sqrt{2(\|\mathrm{e}_\pi\|^2 + \|\mathbf{X}_{(s)^c} \theta_{(s)^c}\|^2)} \\ &= O_p(1) \cdot O_p(n^{-1}) \cdot O_p(n^{1/2}) \\ &= o_p(1). \end{aligned} \tag{A.43}$$

Similarly, one can obtain that

$$\begin{aligned}&\left|\left(\mathrm{e}_\pi + \mathbf{X}_{(m)^c} \theta_{(m)^c}\right)^\top (\mathbf{I}_n - \mathbf{P}_{(m)})(\mathbf{Q}_{(m)} + \mathbf{Q}_{(s)} + \mathbf{Q}_{(m)}\mathbf{Q}_{(s)})(\mathbf{I}_n - \mathbf{P}_{(s)})(\mathrm{Z}_{\hat{\pi}} - \mathrm{Z}_\pi)\right| = o_p(1), \\ &(\mathrm{Z}_{\hat{\pi}} - \mathrm{Z}_\pi)^\top (\mathbf{I}_n - \mathbf{P}_{(m)})(\mathbf{Q}_{(m)} + \mathbf{Q}_{(s)} + \mathbf{Q}_{(m)}\mathbf{Q}_{(s)})(\mathbf{I}_n - \mathbf{P}_{(s)})(\mathrm{Z}_{\hat{\pi}} - \mathrm{Z}_\pi) = o_P(1). \end{aligned}$$

This together with (A.43) leads us to

$$\hat{\mathbf{\Gamma}}_{ms} = o_p(1), \quad m, s \in \{1, 2, \ldots, M_n\}. \tag{A.44}$$

Let $\mathbf{1}_n$ be an $M_n \times 1$ vector whose elements are all one. Then, $\mathbf{1}_n^\top \omega = 1$. Note that $\mathbf{D}_{(m)} = \mathbf{I}_n + \mathbf{Q}_{(m)}$. Then, by (22), we have

$$\begin{aligned}\mathrm{CV}_{\hat{\pi}, \phi_n}(\omega) &= \|\mathrm{Z}_{\hat{\pi}} - \tilde{\mu}_{\hat{\pi}}(\omega)\|^2 + \phi_n \omega^\top \mathrm{K} \\ &= \sum_{m=1}^{M_n} \omega_m \sum_{s=1}^{M_n} \omega_s \mathrm{Z}_{\hat{\pi}}^\top (\mathbf{I}_n - \tilde{\mathbf{P}}_{(m)})^\top (\mathbf{I}_n - \tilde{\mathbf{P}}_{(s)}) \mathrm{Z}_{\hat{\pi}} + \phi_n \omega^\top \mathrm{K} \mathbf{1}_n^\top \omega \end{aligned}$$

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

$$
\begin{aligned}
&= \sum_{m=1}^{M_n} \omega_m \sum_{s=1}^{M_n} \omega_s Z_{\hat{\pi}}^\top (\mathbf{I}_n - \mathbf{P}_{(m)}) \mathbf{D}_{(m)} \mathbf{D}_{(s)} (\mathbf{I}_n - \mathbf{P}_{(s)}) Z_{\hat{\pi}} + \phi_n \omega^\top \mathrm{K} \mathbf{1}_n^\top \omega \\
&= \sum_{m=1}^{M_n} \omega_m \sum_{s=1}^{M_n} \omega_s Z_{\hat{\pi}}^\top (\mathbf{I}_n - \mathbf{P}_{(m)}) \left(\mathbf{I}_n + \mathbf{Q}_{(m)} + \mathbf{Q}_{(s)} + \mathbf{Q}_{(m)} \mathbf{Q}_{(s)}\right) \\
&\quad \times (\mathbf{I}_n - \mathbf{P}_{(s)}) Z_{\hat{\pi}} + \phi_n \omega^\top \mathrm{K} \mathbf{1}_n^\top \omega \\
&= \sum_{m=1}^{M_n} \omega_m \sum_{s=1}^{M_n} \omega_s \Big\{ Z_{\hat{\pi}}^\top (\mathbf{I}_n - \mathbf{P}_{(m)}) (\mathbf{I}_n - \mathbf{P}_{(s)}) Z_{\hat{\pi}} \\
&\quad + \left(\mathrm{e}_\pi + \mathbf{X}_{(m)^c} \theta_{(m)^c} + Z_{\hat{\pi}} - Z_\pi\right)^\top (\mathbf{I}_n - \mathbf{P}_{(m)}) \\
&\quad \times \left(\mathbf{Q}_{(m)} + \mathbf{Q}_{(s)} + \mathbf{Q}_{(m)} \mathbf{Q}_{(s)}\right) (\mathbf{I}_n - \mathbf{P}_{(s)}) \left(\mathrm{e}_\pi + \mathbf{X}_{(s)^c} \theta_{(s)^c} + Z_{\hat{\pi}} - Z_\pi\right) \Big\} \\
&\quad + \phi_n \omega^\top \mathrm{K} \mathbf{1}_n^\top \omega \\
&= \omega^\top \hat{\mathbf{\Upsilon}} \omega + \omega^\top \mathbf{\Psi} \omega + \omega^\top \hat{\mathbf{\Gamma}} \omega + \phi_n \omega^\top \mathrm{K} \mathbf{1}_n^\top \omega. \qquad \textbf{(A.45)}
\end{aligned}
$$

Proof of (i) of Theorem 3: For any $m_F \in \{1, 2, \ldots, m_o - 1\}$, let $\breve{\omega}_{m_F}$ be an $M_n \times 1$ vector whose m_Fth element is $-\breve{\omega}_{\phi_n, m_F}$, M_nth element is $\breve{\omega}_{\phi_n, m_F}$ and all other elements are zeros. Then, by (A.39), (A.42), (A.44), (A.45), and (23), we have

$$
\begin{aligned}
&0 \le \mathrm{CV}_{\hat{\pi}, \phi_n}(\breve{\omega}_{m_F} + \breve{\omega}_{\phi_n}) - \mathrm{CV}_{\hat{\pi}, \phi_n}(\breve{\omega}_{\phi_n}) \\
&= (\breve{\omega}_{m_F} + \breve{\omega}_{\phi_n})^\top (\hat{\mathbf{\Upsilon}} + \mathbf{\Psi} + \hat{\mathbf{\Gamma}} + \phi_n \mathrm{K} \mathbf{1}_n^\top)(\breve{\omega}_{m_F} + \breve{\omega}_{\phi_n}) - \breve{\omega}_{\phi_n}^\top (\hat{\mathbf{\Upsilon}} + \mathbf{\Psi} + \hat{\mathbf{\Gamma}} + \phi_n \mathrm{K} \mathbf{1}_n^\top) \breve{\omega}_{\phi_n} \\
&= \breve{\omega}_{\phi_n, m_F}^2 (\hat{a}_{m_F} - \hat{a}_{M_n}) + 2 \breve{\omega}_{\phi_n, m_F} \sum_{s=1}^{M_n} \breve{\omega}_{\phi_n, s} \left(\hat{a}_{M_n} - \hat{a}_{\max\{s, m_F\}}\right) + \phi_n \breve{\omega}_{\phi_n, m_F} (k_{M_n} - k_{m_F}) \\
&\quad + \breve{\omega}_{\phi_n, m_F}^2 (\mathbf{\Psi}_{m_F m_F} - \mathbf{\Psi}_{M_n m_F} + \mathbf{\Psi}_{M_n M_n} - \mathbf{\Psi}_{m_F M_n}) + 2 \breve{\omega}_{\phi_n, m_F} \sum_{s=1}^{M_n} \breve{\omega}_{\phi_n, s} (\mathbf{\Psi}_{s M_n} - \mathbf{\Psi}_{s m_F}) \\
&\quad + \breve{\omega}_{\phi_n, m_F}^2 (\hat{\mathbf{\Gamma}}_{m_F m_F} - \hat{\mathbf{\Gamma}}_{M_n m_F} + \hat{\mathbf{\Gamma}}_{M_n M_n} - \hat{\mathbf{\Gamma}}_{m_F M_n}) + 2 \breve{\omega}_{\phi_n, m_F} \sum_{s=1}^{M_n} \breve{\omega}_{\phi_n, s} (\hat{\mathbf{\Gamma}}_{s M_n} - \hat{\mathbf{\Gamma}}_{s m_F}) \\
&\le -\breve{\omega}_{\phi_n, m_F}^2 (\hat{a}_{m_F} - \hat{a}_{M_n}) + \breve{\omega}_{\phi_n, m_F} \left\{ O_p(1) + \phi_n O_p(1) \right\}.
\end{aligned}
$$

This together with (A.40) leads us to

$$
\breve{\omega}_{\phi_n, m_F} \le (\hat{a}_{m_F} - \hat{a}_{M_n})^{-1} \left\{ O_p(1) + \phi_n O_p(1) \right\} = O_p(\phi_n n^{-1}).
$$

Hence, the proof of (i) of Theorem 3 is completed.

Proof of (ii) of Theorem 3: For any $m_c \in \{m_o + 1, m_o + 2, \ldots, M_n\}$, let $\breve{\omega}_{m_c}$ be an $M_n \times 1$ vector whose m_oth element is $\breve{\omega}_{\phi_n, m_c}$, m_cth element is $-\breve{\omega}_{\phi_n, m_c}$ and all other elements are zeros. By (A.41), (A.42), (A.44), and (23), we have

$$
\begin{aligned}
&0 \le \mathrm{CV}_{\hat{\pi}, \phi_n}(\breve{\omega}_{m_c} + \breve{\omega}_{\phi_n}) - \mathrm{CV}_{\hat{\pi}, \phi_n}(\breve{\omega}_{\phi_n}) \\
&= \breve{\omega}_{\phi_n, m_c}^2 (\hat{\mathbf{\Upsilon}}_{m_o m_o} - \hat{\mathbf{\Upsilon}}_{m_c m_o} - \hat{\mathbf{\Upsilon}}_{m_o m_c} + \hat{\mathbf{\Upsilon}}_{m_c m_c}) + 2 \breve{\omega}_{\phi_n, m_c} \sum_{s=1}^{M_n} \breve{\omega}_{\phi_n, s} (\hat{\mathbf{\Upsilon}}_{s m_o} - \hat{\mathbf{\Upsilon}}_{s m_c})
\end{aligned}
$$

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

$$+\check{\omega}^2_{\phi_n,m_c}(\boldsymbol{\Psi}_{m_om_o}-\boldsymbol{\Psi}_{m_cm_o}-\boldsymbol{\Psi}_{m_om_c}+\boldsymbol{\Psi}_{m_cm_c})+2\check{\omega}_{\phi_n,m_c}\sum_{s=1}^{M_n}\check{\omega}_{\phi_n,s}(\boldsymbol{\Psi}_{sm_o}-\boldsymbol{\Psi}_{sm_c})$$

$$+\check{\omega}^2_{\phi_n,m_c}(\hat{\boldsymbol{\Gamma}}_{m_om_o}-\hat{\boldsymbol{\Gamma}}_{m_cm_o}-\hat{\boldsymbol{\Gamma}}_{m_om_c}+\hat{\boldsymbol{\Gamma}}_{m_cm_c})+2\check{\omega}_{\phi_n,m_c}\sum_{s=1}^{M_n}\check{\omega}_{\phi_n,s}(\hat{\boldsymbol{\Gamma}}_{sm_o}-\hat{\boldsymbol{\Gamma}}_{sm_c})$$

$$+\phi_n\check{\omega}_{\phi_n,m_c}(k_{m_o}-k_{m_c})$$

$$=O_p(1)+\phi_n\check{\omega}_{\phi_n,m_c}(k_{m_o}-k_{m_c}).$$

Thus, we have $\check{\omega}_{\phi_n,m_c}=O_p(\phi_n^{-1})$. This completes the proof of (ii) of Theorem 3 and thus the proof of Theorem 3. □

REFERENCES

Abrevaya, J., Hsu, Y.-C., & Lieli, R. P. (2015). Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, 33, 485–505.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723.

Ando, T., & Li, K.-C. (2014). A model-averaging approach for high-dimensional regression. *Journal of the American Statistical Association*, 109, 254–265.

Andrews, D. W. (1991). Asymptotic optimality of generalized C_L, cross-validation, and generalized cross-validation in regression with heteroskedastic errors. *Journal of Econometrics*, 47, 359–377.

Bhattacharya, D., & Dupas, P. (2012). Inferring welfare maximizing treatment assignment under budget constraints. *Journal of Econometrics*, 167, 168–196.

Buckland, S. T., Burnham, K. P., & Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics*, 53, 603–618.

Cai, T., Tian, L., Wong, P. H., & Wei, L. J. (2011). Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics*, 12, 270–282.

Cheng, X., & Hansen, B. E. (2015). Forecasting with factor-augmented regression: A frequentist model averaging approach. *Journal of Econometrics*, 186, 280–293.

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge University Press.

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2008). Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics*, 90, 389–405.

De Luca, G., Magnus, J. R., & Peracchi, F. (2018). Weighted-average least squares estimation of generalized linear models. *Journal of Econometrics*, 204, 1–17.

Fang, F., Lan, W., Tong, J., & Shao, J. (2019a). Model averaging for prediction with fragmentary data. *Journal of Business & Economic Statistics*, 37, 517–527.

Fang, F., Li, J., & Wang, J. (2019b). Optimal model averaging estimation for correlation structure in generalized estimating equations. *Communications in Statistics - Simulation and Computation*, 48, 1574–1593.

Fang, F., Li, J., & Xia, X. (2022). Semiparametric model averaging prediction for dichotomous response. *Journal of Econometrics*, 229, 219–245.

Graham, B. S. (2011). Efficiency bounds for missing data models with semiparametric restrictions. *Econometrica*, 79, 437–452.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66, 315–331.

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

Hansen, B. E. (2014). Model averaging, asymptotic risk, and regressor groups. *Quantitative Economics*, 5, 495–530.

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

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

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

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

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81, 945–960.

Imai, K., & Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7, 443–470.

Lalonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76, 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, 1207–1225.

Li, K.-C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *The Annals of Statistics*, 15, 958–975.

Liang, H., Zou, G., Wan, A. T. K., & Zhang, X. (2011). Optimal weight choice for frequentist model average estimators. *Journal of the American Statistical Association*, 106, 1053–1066.

Lin, Z., & Bai, Z. (2011). *Probability inequalities*. Springer Science & Business Media.

Liu, C.-A. (2015). Distribution theory of the least squares averaging estimator. *Journal of Econometrics*, 186, 142–159.

Liu, Q., & Okui, R. (2013). Heteroscedasticity-robust C_p model averaging. *The Econometrics Journal*, 16, 463–472.

Liu, Q., Okui, R., & Yoshimura, A. (2016). Generalized least squares model averaging. *Econometric Reviews*, 35, 1692–1752.

Longford, N. T. (2005). Editorial: Model selection and efficiency—Is "which model…?" the right question?. *Journal of the Royal Statistical Society (Series A)*, 168, 469–472.

Rao, C. R. (1973). *Linear statistical inference and its applications*. (2nd ed.) John Wiley & Sons.

Rolling, C. A., & Yang, Y. (2014). Model selection for estimating treatment effects. *Journal of the Royal Statistical Society (Series B)*, 76, 749–769.

Rolling, C. A., Yang, Y., & Velez, D. (2019). Combining estimates of conditional treatment effects. *Econometric Theory*, 35, 1089–1110.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, 2, 1–26.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461–464.

Tian, L., Alizadeh, A. A., Gentles, A. J., & Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109, 1517–1532.

Ullah, A., & Wang, H. (2013). Parametric and nonparametric frequentist model selection and model averaging. *Econometrics*, 1, 157–179.

Wan, A. T., Zhang, X., & Zou, G. (2010). Least squares model averaging by Mallows criterion. *Journal of Econometrics*, 156, 277–283.

Wei, Y., & Wang, Q. (2021). Cross-validation-based model averaging in linear models with response missing at random. *Statistics & Probability Letters*, 171, 108990.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1–25.

Xie, J., Yan, X., & Tang, N. (2021). A model-averaging method for high-dimensional regression with missing responses at random. *Statistica Sinica*, 31, 1005–1026.

Yan, X., Wang, H., Wang, W., Xie, J., Ren, Y., & Wang, X. (2021). Optimal model averaging forecasting in high-dimensional survival analysis. *International Journal of Forecasting*, 37, 1147–1155.

Yang, Y. (2001). Adaptive regression by mixing. *Journal of the American Statistical Association*, 96, 574–588.

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press

Yuan, Z., & Yang, Y. (2005). Combining linear regression models: When and how?. *Journal of the American Statistical Association*, 100, 1202–1214.

Zhang, X. (2021). A new study on asymptotic optimality of least squares model averaging. *Econometric Theory*, 37, 388–407.

Zhang, X., & Liu, C.-A. (2019). Inference after model averaging in linear regression models. *Econometric Theory*, 35, 816–841.

Zhang, X., Wan, A. T., & Zou, G. (2013). Model averaging by jackknife criterion in models with dependent data. *Journal of Econometrics*, 174, 82–94.

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, 972–984.

Zhao, Z., Zhang, X., Zou, G., Wan, A. T., & Tso, G. K. (2024). Model averaging for estimating treatment effects. *Annals of the Institute of Statistical Mathematics*, 76, 73–92.

https://doi.org/10.1017/S0266466625100029 Published online by Cambridge University Press