Received: 25 December 2020 | Revised: 10 December 2021 | Accepted: 7 January 2022

DOI: 10.1111/sjos.12585

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

Multiply robust matching estimators of average and quantile treatment effects

Shu Yang | **Yunshu Zhang**

Department of Statistics, North Carolina State University, Raleigh, North Carolina, USA

Correspondence
Shu Yang, Raleigh, Department of Statistics, North Carolina State University, Raleigh, NC, USA.
Email: syang24@ncsu.edu

Funding information
National Institutes of Health, Grant/Award Numbers: P01 CA142538, 1R01AG066883, 1R01ES031651; National Science Foundation, Grant/Award Number: DMS 1811245

Abstract

Propensity score matching has been a long-standing tradition for handling confounding in causal inference, however, requiring stringent model assumptions. In this article, we propose novel double score matching (DSM) utilizing both the propensity score and prognostic score. To gain the protection of possible model misspecification, we posit multiple candidate models for each score. We show that the debiasing DSM estimator achieves the multiple robustness property in that it is consistent if any one of the score models is correctly specified. We characterize the asymptotic distribution for the DSM estimator requiring only one correct model specification based on the martingale representations of the matching estimators and theory for local normal experiments. We also provide a two-stage replication method for variance estimation and extend DSM for quantile estimation. Simulation demonstrates DSM outperforms single-score matching and prevailing multiply robust weighting estimators in the presence of extreme propensity scores.

KEYWORDS

Bahadur representation, causal effect on the treated, double robustness, quantile estimation, weighted bootstrap

1 | INTRODUCTION

Causal inference plays an important role in science, education, medicine, policy, and economics. If all confounders of the treatment–outcome relationship are observed, one can use standard techniques, such as regression adjustment, inverse probability of treatment weighting (IPW), augmented IPW (AIPW), and matching to adjust for confounding (Imbens & Rubin, 2015). Among them, the AIPW estimator is most popular because it achieves the so-called *double robustness* property by combining the use of models for the probability of treatment assignment, also known as the propensity score (Rosenbaum & Rubin, 1983b), and the outcome mean function. More specifically, it consistently estimates the treatment effect if either one of these functions is modeled correctly (e.g., Bang & Robins, 2005; Lunceford & Davidian, 2004). However, inevitably, weighting estimators can have a high variability by inverting the estimated propensity scores (e.g., Guo & Fraser, 2014; Kang & Schafer, 2007), especially if these probabilities are close to zero or one. Matching has multiple features that are desirable:

(a) Matching does not involve weighting by the inverse of the propensity score and therefore avoids the possibly large variability due to weighting (Frölich, 2004);
(b) Matching is transparent and intuitively appealing with the goal of replicating a randomized experiment from an observational study (Dehejia & Wahba, 1999, 2002; Heckman et al., 1997; Rosenbaum, 1989; Rubin, 2006; Stuart, 2010);
(c) Matching can be viewed as a hot deck imputation method that can provide valid estimators of general parameters depending on the entire distribution, such as quantiles (Ford, 1983).

Although matching has a substantial promise, it suffers from the issue of the curse of dimensionality. In the presence of many covariates, matching directly on high-dimensional covariates is incapable of removing all confounding biases. To overcome this challenge, researchers have proposed different dimension reduction techniques to facilitate matching. On the one hand, Rosenbaum and Rubin (1983b) demonstrated the central role of the propensity score as being a balancing score in the sense that the same propensity score distributions in different treatment groups lead to the same covariate distributions. Therefore, propensity score matching (PSM) can remove all confounding biases (e.g., Abadie & Imbens, 2016). However, PSM has recently received major backlash for emulating an unconditional randomized experiment, contrary to matching on covariates that mimics a block randomized experiment (King & Nielsen, 2019). On the other hand, Hansen (2008) proposed an alternative balancing score: the prognostic score, also called the disease risk score (i.e., a sufficient statistic for the potential outcomes given which the potential outcomes and covariates are independent). This score provides a balance of disease risks between the treatment groups, as distinct from the balance of treatment propensities provided by the propensity score. In economics, prognostic score matching (PGM) has been previously proposed in Imbens (2004) and Zhao (2004), where the prognostic score is a vector of linear predictors in treatment-specific outcome regressions. PGM is also similar to predictive mean matching (Yang & Kim, 2018; Yang & Kim, 2020) in the missing data literature to compensate for nonresponse. As analogous to AIPW, it is advantageous to combine the use of the propensity and prognostic score in matching (Hansen, 2008). Leacy and Stuart (2014) showed empirically that the joint use of two scores in matching (which we refer to as double score matching, DSM) improves the treatment effect estimation. Antonelli et al. (2018) later established the double robustness of matching jointly on propensity and prognostic scores in the sense that the matching estimator is consistent

for the ATE if either one of the score models is correctly specified. However, they only provided the rate of convergence but not the asymptotic distribution of the doubly robust matching estimator.

In this article, we propose new DSM estimators based on the propensity score and the prognostic score. Because each score creates a balance between the treated and control groups, the augmented score serves as a "double balancing score." To estimate the ATEs, existing DSM would require adjusting for the vector of the propensity score and possibly multiple treatment-specific prognostic scores (i.e., one for each treatment group). Instead of estimating the ATEs directly, we focus on estimating the average of the potential outcomes separately for each treatment level, which requires adjusting only for the propensity score and the prognostic score for that particular level of the treatment. This insight allows us to reduce the dimension of the double score further without giving up the double balancing property. This strategy also plays an important role in dimension reduction for constructing improved DSM estimators.

In practice, the double score is unknown and therefore requires modeling and estimation. The new DSM estimator is doubly robust, which includes one propensity score model and one prognostic score model. With an unknown data generating process, there is no guarantee that either of the two models is correctly specified. To gain additional protection against model misspecification, we posit multiple models for the propensity score and prognostic score. Doing so, however, may introduce bias due to matching discrepancy based on a moderately high-dimensional matching variable (Abadie & Imbens, 2011), although our strategy of estimating the average of potential outcomes separately helps dimension reduction. In this case, we propose the debiasing DSM estimator that corrects for the bias due to matching discrepancy. The current matching literature has focused primarily on estimating the ATEs; however, other aspects of the distribution such as quantiles may be more appropriate in certain applications. For example, a treatment strategy may not decrease average health cost but instead lowers the upper tail of the cost distribution, so focusing only on ATEs would not reveal the beneficial effect of the treatment strategy. In these cases, it is more informative to study quantile treatment effects (QTEs), which are defined as the differences in population quantiles of the potential outcome distributions. Taking the advantage of matching as a hot deck imputation method, we extend the multiply robust DSM framework to estimate QTEs.

We show that the DSM estimators have the *multiple robustness* property, which guarantees the estimation consistency if any one of the candidate models for the propensity score or prognostic score is correctly specified. This result is similar in essence to the multiply robust weighting (MRW) estimators in the missing data and survey literature (Chen & Haziza, 2017a, 2017b; Han, 2014; Han et al., 2019; Han & Wang, 2013; Li et al., 2020). Naik et al. (2016) and Wang (2019) proposed MRW estimators for estimating the ATEs in causal inference. The DSM estimator has an intrinsic connection with the doubly robust AIPW estimator in that they exhibit similar asymptotic expansions; see (6) and (7). However, AIPW is extremely unstable when some estimated propensity scores are close to zero or one due to the weight construction, and it is sensitive to slight model misspecification (Kang & Schafer, 2007). Instead of inverting the estimated propensity score, the DSM estimator uses the matching weights and hence is more robust to extreme propensity score values. MRW also mitigates the instability of (A)IPW by estimating the weights directly under balance constraints. However, it requires exact covariate balance and thus can still have large variability in finite samples compared to DSM, especially when the overlap of the covariate distribution is limited.

The rest of this paper proceeds as follows. Section 2 introduces notation, assumptions, and lemmas for various balancing scores. Section 3 provides the new perspective of using the double score as a dimension reduction tool and proposes the new DSM estimator of the

ATE with multiple candidate models for the double score. Section 4 establishes the multiple robustness of the DSM estimator and its limiting distribution, which allows quantifying the impact of the nuisance parameter estimation. Section 5 extends the DSM framework to the estimation of the QTE. Section 6 provides the extensions to the average treatment effect on the treated (ATT) and QTE on the treated (QTT). Section 7 uses simulation to evaluate the finite-sample properties of the DSM estimators. The simulation results demonstrate that matching estimators outperform weighting estimators. Section 8 applies the DSM estimators to an observational study from the job training program. Section 9 concludes, appendix and Data S1 contains the proofs and additional empirical results, and an R package *dsmatch* is available at https://github.com/Yunshu7/dsmatch.

2 | NOTATION, ASSUMPTIONS, AND BALANCING SCORES

Let X_i be a vector of pretreatment covariates, A_i the binary treatment, and Y_i the outcome for unit $i = 1, \dots, n$. We follow the potential outcomes framework. Let $Y_i(a)$ be the potential outcome had unit i been given treatment a ($a = 0, 1$). The observed outcome is $Y_i = Y_i(A_i) = A_iY_i(1) + (1 - A_i)Y_i(0)$. We assume that $\{X_i, A_i, Y_i(0), Y_i(1)\}$, $i = 1, \dots, n$, are independent and identically distributed. Thus, (X_i, A_i, Y_i), $i = 1, \dots, n$, are also independent and identically distributed. Various causal estimands are useful to provide a comprehensive assessment of treatment effects. The ATE is $\tau = \mathbb{E}\{Y(1) - Y(0)\}$. For $\xi \in (0, 1)$, the overall ξ-QTE is $\Delta_\xi = q_{1,\xi} - q_{0,\xi}$, where $q_{a,\xi} = \inf_q[\mathbb{P}\{Y(a) \le q\} \ge \xi]$, $a = 0, 1$. When the outcome data follow a skewed distribution, QTEs may be more informative measures of treatment effect. Similarly, the ATT is $\tau_{\text{ATT}} = \mathbb{E}\{Y(1) - Y(0)|A = 1\}$, and the QTT is $\Delta_{\text{QTT},\xi} = q_{1,\xi|A=1} - q_{0,\xi|A=1}$, where $q_{a,\xi|A=1} = \inf_q[\mathbb{P}\{Y(a) \le q\} \ge \xi|A = 1]$, $a = 0, 1$. We illustrate the methodology development for estimating τ and Δ_ξ and provide extensions to the ATT and QTT in Section 6. For simplicity of exposition, for a generic variable V, denote

$$\mu_a(V) = \mathbb{E}\{Y(a)|V\},\ \sigma_a^2(V) = \mathbb{V}\{Y(a)|V\},\ e(V) = \mathbb{P}(A = 1|V),$$

where $\mu_a(V)$ is an outcome mean function, $\sigma_a^2(V)$ is a variance function, and $e(V)$ is the propensity score. Hansen (2008) introduced the notion of the prognostic score $\Psi_a(X)$ as a sufficient statistic for $Y(a)$ in the sense that $Y(a) \perp\!\!\!\perp X \,|\, \Psi_a(X)$ for $a = 0, 1$. For example, if $Y(a)$ follows a location-shift family $f_a\{y - \mu_a(X)\}$, then $\Psi_a(X) = \mu_a(X)$ for $a = 0, 1$.

We focus on the setting where the standard positivity and treatment ignorability assumptions hold (Rosenbaum & Rubin, 1983b).

Assumption 1. (i) There exist constants c_1 and c_2 such that $0 < c_1 \le e(X) \le c_2 < 1$ almost surely; and (ii) $\{Y(0), Y(1)\} \perp\!\!\!\perp A|X$, where $\perp\!\!\!\perp$ means "independent of."

Assumption 1 (i) implies a sufficient overlap of the covariate distribution between the treatment groups. If this assumption is violated, a common approach is to trim the sample; see Yang and Ding (2018). Generally, Assumption 1 (ii) can be made plausible by collecting detailed information on characteristics of the units that are related to treatment assignment and outcome. As a result, the dimension of X may be high. Balancing scores have been proposed for dimension reduction.

Lemma 1 (Balancing score; Antonelli et al. (2018)). *Under Assumption 1, (i)* $\{Y(0), Y(1)\} \perp\!\!\!\perp A|\{e(X), h(X)\}$, *and ii)* $\{Y(0), Y(1)\} \perp\!\!\!\perp A \,|\, \{h(X), \Psi(X)\}$ *for any* $h(X)$.

In particular, if $h(X)$ is a null set, Lemma 1 (i) shows that the propensity score is a balancing score (Rosenbaum & Rubin, 1983b) and (ii) shows that the prognostic score is a balancing score (Hansen, 2008). Antonelli et al. (2018) proposed matching based on the estimated double score $\{\widehat{e}(X), \widehat{\Psi}(X)\}$. If either $\widehat{e}(X)$ or $\widehat{\Psi}(X)$ is consistent for the corresponding score, the matching estimator is consistent for the ATE by either (i) or (ii) in Lemma 1.

DSM is an attractive alternative to PSM; however, the dimension reduction property of Lemma 1 depends on the dimension of the prognostic score. The problem is that if the dimension of the matching variable is higher, the bias order of the matching estimator becomes larger, see Section 3.1, suggesting that the advantages of PSM do not carry over to DSM. To preserve the simplicity of matching (avoiding de-biasing), we show that further improvement of the dimension reduction property of the double score is possible without additional assumptions. Then, the advantage of PSM carries over to DSM, see Section 3.2. Moreover, because the double score is unknown in practice, one must posit models and estimate the double score from the observed data. To gain robustness to model misspecification, we posit multiple candidate models for the double score. We propose a multiply robust DSM procedure and show that the de-biasing matching estimator achieves multiple robustness, see Section 3.3.

3 | DSM ESTIMATORS OF THE ATE

3.1 | General matching estimators

To fix ideas, we consider matching with replacement with the number of matches fixed at M. Matching estimators hinge on imputing the missing potential outcome for each unit. In practice, the most common choice of M is 1, then the matching procedure becomes nearest neighbor imputation (Chen & Shao, 2000, 2001; Little & Rubin, 2002). To be precise, for unit i, the potential outcome under A_i is the observed outcome Y_i; the (counterfactual) potential outcome under $1 - A_i$ is not observed but can be imputed by the observed outcomes of the nearest M units with $1 - A_i$.

To illustrate the properties of the matching estimator, we first consider a generic variable V as the matching variable. Table 1 summarizes the choices of V. To stabilize the numerical performance, it is desirable to standardize V such that each component has mean zero and variance one. Without loss of generality, we use the Euclidean distance to determine neighbors; the discussion applies to other distances (Abadie & Imbens, 2006). We denote $\mathcal{J}_{V,i}$ as the index set for these matched units for unit i and $K_{V,i} = \sum_{l=1}^{n} \mathbf{1}(i \in \mathcal{J}_{V,l})$ as the number of times that unit i is used as a match, where the subscript "V" in $\mathcal{J}_{V,i}$ and $K_{V,i}$ indicates the name of the matching variable. Table 1 (I) illustrates the above matching scheme to impute the missing potential outcomes. For unit i with $A_i = 1$, the imputed potential outcomes are $\hat{Y}_i(1) = Y_i$ and $\hat{Y}_i(0) = M^{-1} \sum_{j \in \mathcal{J}_{V,i}} Y_j$. For unit i' with $A_{i'} = 0$, $\hat{Y}_{i'}(1) = M^{-1} \sum_{j \in \mathcal{J}_{V,i'}} Y_j$ and $\hat{Y}_{i'}(0) = Y_{i'}$. Once we approximate both potential outcomes for all units, a simple matching estimator of τ is

$$\hat{\tau}_{\text{mat}} = n^{-1} \sum_{i=1}^{n} \{\hat{Y}_i(1) - \hat{Y}_i(0)\} = n^{-1} \sum_{i=1}^{n} (2A_i - 1)(1 + M^{-1} K_{V,i}) Y_i.$$

To establish the asymptotic properties of $\hat{\tau}_{\text{mat}}$, Abadie and Imbens (2006) derived the following decomposition

$$n^{1/2}(\hat{\tau}_{\text{mat}} - \tau) = B_n + C_n,$$

where

$$B_n = n^{-1/2}\sum_{i=1}^{n}(2A_i-1)\left[M^{-1}\sum_{j\in\mathcal{J}_{V,i}}\left\{\mu_{1-A_i}(V_i)-\mu_{1-A_i}(V_j)\right\}\right],$$
$$C_n = n^{-1/2}\sum_{i=1}^{n}\left[\mu_1(V_i)-\mu_0(V_i)-\tau+(2A_i-1)\left(1+M^{-1}K_{V,i}\right)\left\{Y_i-\mu_{A_i}(V_i)\right\}\right]. \quad (1)$$

By Assumption 1, for V to be X, propensity score, prognostic score or double score, we have $\mathbb{E}\{\mu_1(V)-\mu_0(V)\}=\tau$, and therefore $\mathbb{E}(C_n)=0$. The difference $\mu_{A_i}(V_i)-\mu_{A_i}(V_j)$ in (1) accounts for the matching discrepancy, so B_n contributes to the asymptotic bias of the matching estimator. In general, if the matching variable is d_V-dimensional, we have $\mathbb{E}(B_n)=O(n^{1/2-2/d_V})$ (Abadie & Imbens, 2006, theorem 1). Table 2 demonstrates the relationship of the bias order and d_V. If $d_V\geq 4$, the bias is nonnegligible. If $d_V=3$, the bias shrinks to zero as n increases but the convergence rate $-1/6$ is slow. If $d_V=2$ and 1, the bias shrinks to zero at much faster rates $-1/2$ and $-3/2$, respectively. Therefore, in finite samples, matching based on a three-dimensional double score $\{e(X),\Psi(X)\}$ is likely to have a noticeable bias. Reducing d_V to 2 or 1 is worthwhile to make the bias achieve faster rates of converging to zero.

TABLE 1 Two matching schemes for imputing potential outcomes. $\mathcal{J}_{V,i}$ denotes the index set for the matched units for unit i, where the subscript "V" represents the name of the matching variable. In (I), the matching variable V is the same for imputing the missing values of $Y(0)$ and $Y(1)$. In (II), the matching variables V_0 and V_1 are different for imputing the missing values of $Y(0)$ and $Y(1)$

(I) Matching imputation					**(II) New matching imputation**				
Unit	$\boldsymbol{A}$	$\boldsymbol{Y}$	$\boldsymbol{\hat{Y}(0)}$	$\boldsymbol{\hat{Y}(1)}$	**Unit**	$\boldsymbol{A}$	$\boldsymbol{Y}$	$\boldsymbol{\hat{Y}(0)}$	$\boldsymbol{\hat{Y}(1)}$
i	0	Y_i	Y_i	$M^{-1}\sum_{l\in\mathcal{J}_{V,i}}Y_l$	i	0	Y_i	Y_i	$M^{-1}\sum_{l\in\mathcal{J}_{\mathbf{V_1},i}}Y_l$
i'	1	$Y_{i'}$	$M^{-1}\sum_{l\in\mathcal{J}_{V,i'}}Y_l$	$Y_{i'}$	i'	1	$Y_{i'}$	$M^{-1}\sum_{l\in\mathcal{J}_{\mathbf{V_0},i'}}Y_l$	$Y_{i'}$

	(I) Matching Variable		**(II) Matching Variable**		
	$\boldsymbol{V}$	$\boldsymbol{d_V}$	$\boldsymbol{V_0}$	$\boldsymbol{V_1}$	$\boldsymbol{d_V}$
M.X	X	$\dim(X)$	X	X	$\dim(X)$
PSM	$e(X)$	1	$e(X)$	$e(X)$	1
PGM	$\{\Psi_0(X),\Psi_1(X)\}$	2	$\Psi_0(X)$	$\Psi_1(X)$	1
DSM	$S=\{e(X),\Psi_0(X),\Psi_1(X)\}$	3	$S_0=\{e(X),\Psi_0(X)\}$	$S_1=\{e(X),\Psi_1(X)\}$	2
DSM	$S=\{e^j(X),\Psi_0^k(X),\Psi_1^k(X):$ $j=1,\ldots J;k=1,\ldots,K\}$	$J+2K$	$S_0=\{e^j(X),\Psi_0^k(X):$ $j=1,\ldots J;k=1,\ldots,K\}$	$S_1=\{e^j(X),\Psi_1^k(X):$ $j=1,\ldots J;k=1,\ldots,K\}$	$J+K$

Abbreviations: DSM, double score matching; PGM, prognostic score matching; PSM, propensity score matching.

TABLE 2 The order of bias of the matching variable in terms of the dimension of the matching variable

$\boldsymbol{d_V}$	**1**	**2**	**3**	**4**	**> 4**
$O(n^{1/2-2/d_V})$	$O(n^{-3/2})$	$O(n^{-1/2})$	$O(n^{-1/6})$	$O(1)$	$O(n^{1/10})$

3.2 | New simple DSM estimator

Lemma 2 is the key result (Antonelli et al., 2018).

Lemma 2. *Under Assumption 1,* $Y(a) \perp\!\!\!\perp A | \{h(X), \Psi_a(X)\}$, $Y(a) \perp\!\!\!\perp A | \{e(X), h(X)\}$ *for any* $h(X)$ *and* $a = 0, 1$.

Lemma 2 implies that $\mathbb{E}\{Y(a)\} = \mathbb{E}\left[\mathbb{E}\{Y|A = a, e(X), \Psi_a(X)\}\right]$, $(a = 0, 1)$. For its interpretation, it is useful to compare it to the result in Lemma 1. By Lemma 1, we create subpopulations where we can simultaneously compare the treated units and the control units. These subpopulations were defined by common values for $\{e(X), \Psi(X)\}$. By Lemma 2, we do not construct such populations. The key insight is that in order to estimate τ, it is not necessary to do so. Instead, we construct subpopulations where we can estimate the average value of the potential outcomes for $a = 0$ and 1 separately. For a given a, these subpopulations are defined by the value of $\{e(X), \Psi_a(X)\}$. This difference allows us to reduce the dimension of the double score from three to two, a small reduction of the dimension of the matching variable, a big reduction of the bias order of the matching estimator.

We focus on estimating $\mu_a = \mathbb{E}\{Y(a)\}$ separately for $a = 0, 1$. Let the matching variable be the double score $S_a(X) = \{e(X), \Psi_a(X)\}$ or S_a for shorthand. Table 1 (II) illustrates the new matching scheme to impute the missing potential outcomes. For unit i with $A_i = 1$, $\hat{Y}_i(1) = Y_i$ and $\hat{Y}_i(0) = M^{-1}\sum_{l\in\mathcal{J}_{S_0,i}} Y_l$. For unit i' with $A_{i'} = 0$, $\hat{Y}_{i'}(1) = M^{-1}\sum_{l\in\mathcal{J}_{S_1,i'}} Y_l$ and $\hat{Y}_{i'}(0) = Y_{i'}$. Importantly, the new matching scheme uses different matching variables, namely S_0 and S_1, to impute the missing values of $Y(0)$ and $Y(1)$. This is in contrast to matching scheme (I) that uses the same matching variable for imputing the missing values of $Y(0)$ and $Y(1)$. Once we approximate both potential outcomes for all units, a simple DSM estimator of τ is

$$\hat{\tau}^{(0)}_{\text{dsm}} = \hat{\mu}^{(0)}_{1,\text{dsm}} - \hat{\mu}^{(0)}_{0,\text{dsm}}, \tag{2}$$

where $\hat{\mu}^{(0)}_{a,\text{dsm}} = n^{-1}\sum_{i=1}^n \hat{Y}_i(a) = n^{-1}\sum_{i=1}^n \mathbf{1}(A_i = a)\left(1 + M^{-1}K_{S_a,i}\right) Y_i$, for $a = 0, 1$. Because $\dim(S_a) = 2$, $\hat{\tau}^{(0)}_{\text{dsm}}$ is asymptotically unbiased.

3.3 | Multiply robust DSM

In practice, S_0 and S_1 are unknown, requiring modeling and estimation from the observed data. Following the empirical literature, one can posit a logistic regression model for the propensity score and a generalized linear model for the prognostic score. To provide additional protection against model misspecification, we can posit multiple candidate models for both scores. The intuition is that if at least one of the candidate models is correctly specified, whether it is a propensity score model or a prognostic score model, balancing at least one score suffices to remove confounding biases. Therefore, the DSM estimator achieves the so-called multiple robustness.

Following Han and Wang (2013), we postulate multiple candidate models

- $\mathcal{M}(\alpha) = \{e^j(X; \alpha^j) : j = 1, \ldots, J\}$ for $e(X)$ with unknown parameters $\alpha = (\alpha^{1,\mathrm{T}}, \ldots, \alpha^{J,\mathrm{T}})^{\mathrm{T}}$;
- $\mathcal{M}_0(\beta_0) = \{\Psi_0^k(X; \beta_0^k) : k = 1, \ldots, K\}$ and $\mathcal{M}_1(\beta_1) = \{\Psi_1^k(X; \beta_1^k) : k = 1, \ldots, K\}$ for $\Psi_0(X)$ and $\Psi_1(X)$, respectively, with unknown parameters $\beta_0 = (\beta_0^{1,\mathrm{T}}, \ldots, \beta_0^{K,\mathrm{T}})^{\mathrm{T}}$ and $\beta_1 = (\beta_1^{1,\mathrm{T}}, \ldots, \beta_1^{K,\mathrm{T}})^{\mathrm{T}}$.

Let $\hat{\alpha}^j$, $\hat{\beta}_0^k$, and $\hat{\beta}_1^k$ be the maximum likelihood estimators or the method of moments estimators of α^j, β_0^k, and β_1^k under the corresponding working model, respectively.

For each treatment level $a \in \{0,1\}$, let $S_a(\theta_a) = \{\mathcal{M}(\alpha), \mathcal{M}_a(\beta_a)\}$, where $\theta_a^{\mathrm{T}} = (\alpha^{\mathrm{T}}, \beta_a^{\mathrm{T}})$, be the set of candidate models for the propensity score and the prognostic score for treatment a, for $a = 0, 1$. Under matching scheme (II), we use $S_a(\hat{\theta}_a)$ to impute the missing values of $Y(a)$, separately for $a = 0, 1$. The corresponding dimension of the matching variable is $J + K$. Let $S(\theta) = \{\mathcal{M}(\alpha), \mathcal{M}_0(\beta_0), \mathcal{M}_1(\beta_1)\}$, where $\theta = (\alpha^{\mathrm{T}}, \beta_0^{\mathrm{T}}, \beta_1^{\mathrm{T}})^{\mathrm{T}}$, be the set of candidate models for the propensity score and the prognostic score for both treatment groups. Under matching scheme (I), one would use $S(\hat{\theta})$ as the matching variable; the corresponding dimension is thus $J + 2K$. If the number of candidate models for the prognostic score is large, the dimension reduction of the double score under new matching scheme (II) can be much larger than under matching scheme (I).

The initial DSM estimator of τ is given by $\hat{\tau}_{\mathrm{dsm}}^{(0)}$ in (2) with S_a replaced by $S_a(\hat{\theta}_a)$ for $a = 0, 1$. We denote the initial estimator as $\hat{\tau}_{\mathrm{dsm}}^{(0)}(\hat{\theta})$ to reflect its dependence on $\hat{\theta}$. As discussed in Section 3.2, if $J = K = 1$, the dimension of $S_a(\hat{\theta})$ is two for $a = 0, 1$. In this case, the asymptotic bias of the matching estimator due to the matching discrepancy is negligible. We do not require further steps to correct the asymptotic bias of $\hat{\tau}_{\mathrm{dsm}}^{(0)}$. This preserves the simplicity of matching in practice. However, if $J, K \geq 2$, the dimension of each matching variable is larger than or equal to four. Consequently, as shown in Table 2, the bias of the matching estimator due to matching discrepancy is not asymptotic negligible. In this case, we propose the de-biasing matching estimator that corrects the asymptotic bias due to matching discrepancy.

Let $\hat{\mu}_a(S_a)$ be a nonparametric estimator of $\mu_a(S_a)$, for $a = 0, 1$, for example, using the method of sieves (Chen, 2007). The de-biasing DSM estimator of τ is

$$\hat{\tau}_{\mathrm{dsm}}(\hat{\theta}) = \hat{\tau}_{\mathrm{dsm}}^{(0)}(\hat{\theta}) - n^{-1/2}\hat{B}_n, \tag{3}$$

where $\hat{B}_n$ is an estimator of B_n by replacing $\mu_a(S_a)$ with $\hat{\mu}_a(S_a)$ for $a = 0, 1$.

Before delving into the discussion of the theoretical properties of $\hat{\tau}_{\mathrm{dsm}}(\hat{\theta})$, we summarize the DSM algorithm that contains nuts and bolts as follows.

Step 1. Posit multiple candidate parametric models $\mathcal{M}(\alpha)$, $\mathcal{M}_0(\beta_0)$, and $\mathcal{M}_1(\beta_1)$ for $e(X)$, $\Psi_0(X)$, and $\Psi_1(X)$, respectively. Obtain the parameter estimators $\hat{\alpha}$, $\hat{\beta}_0$, and $\hat{\beta}_1$. For each unit i, calculate $S_{a,i}(\hat{\theta}_a) = \{\mathcal{M}(\hat{\alpha}), \mathcal{M}_a(\hat{\beta}_a)\}$ for $a = 0, 1$. The propensity scores are probability estimates, ranging from zero to one. To stabilize the numerical performance, it is desirable to use a monotone mapping to transform each propensity score estimate $e^j(X_i; \hat{\alpha}_j) \in (0,1)$ in $\mathcal{M}(\hat{\alpha})$, for example, to $\mathrm{logit}\{e^j(X_i; \hat{\alpha}_j)\} \in \mathcal{R}$. We also recommend standardize $S_{a,i}(\hat{\theta}_a)$ such that each component has mean zero and variance one for $a = 0, 1$.

Step 2. For each unit i with treatment $A_i = a$, find M nearest neighbors from the treatment group $1 - a$ based on the matching variable $S_{1-a,i} = S_{1-a,i}(\hat{\theta})$. Obtain $\mathcal{J}_{S_{1-a}(\hat{\theta}),i}$ that contains the indexes of the matched units for unit i and calculate $K_{S_a(\hat{\theta}),i}$ that counts the number of time that unit i is matched to other units. Obtain the initial matching estimator $\hat{\tau}_{\mathrm{dsm}}^{(0)}(\hat{\theta})$ in (2) with S_a replaced by $S_a(\hat{\theta}_a)$.

If $J = K = 1$, let the DSM estimator be $\hat{\tau}_{\mathrm{dsm}}(\hat{\theta}) = \hat{\tau}_{\mathrm{dsm}}^{(0)}(\hat{\theta})$. If $J, K \geq 2$, we proceed to Steps 3 and 4 below. Even with $J = K = 1$, Steps 3 and 4 can help to reduce the matching discrepancy in finite samples.

Step 3. Obtain a nonparametric estimator of $\mu_a(S_a)$, denoted by $\hat{\mu}_a(S_a)$, for example, by the method of sieves based on $[\{Y_i, S_{a,i}(\hat{\theta}_a)\} : A_i = a]$, for $a = 0, 1$.

Step 4. The DSM estimator of τ is given by $\hat{\tau}_{\text{dsm}}(\hat{\theta})$ in (3) with $S_{a,i}$ replaced by $S_{a,i}(\hat{\theta}_a)$.

4 | MAIN RESULTS

In this section, we establish the asymptotic properties of $\hat{\tau}_{\text{dsm}}(\hat{\theta})$, which depends on the estimators of all nuisance parameters in the propensity score and prognostic score models. Without loss of generality, we consider the prognostic score $\Psi_a(X) = \mu_a(X; \beta_a)$ and multiple candidate models $\Psi_a^k(X; \beta_a^k) = \mu_a^k(X; \beta_a^k)$, for $k = 1, \ldots, K$ and $a = 0, 1$. Consider $\hat{\alpha}^j$, $\hat{\beta}_0^k$, and $\hat{\beta}_1^k$ that solve the estimating equation

$$n^{-1/2} \sum_{i=1}^{n} \begin{pmatrix} U_1^j(A_i, X_i; \alpha^j) \\ U_2^k\left(A_i, X_i, Y_i; \beta_0^k\right) \\ U_3^k\left(A_i, X_i, Y_i; \beta_1^k\right) \end{pmatrix} = 0, \tag{4}$$

where

$$\begin{aligned} U_1^j(A, X; \alpha^j) &= \frac{\partial e^j(X; \alpha^j)}{\partial \alpha^j} \frac{A - e^j(X; \alpha^j)}{e^j(X; \alpha^j)\{1 - e^j(X; \alpha^j)\}}, \\ U_2^k\left(A, X, Y; \beta_0^k\right) &= (1 - A) \frac{\partial \mu_0^k\left(X; \beta_0^k\right)}{\partial \beta_0^q} \{Y - \mu_0^k(X; \beta_0^k)\}, \\ U_3^k\left(A, X, Y; \beta_1^k\right) &= A \frac{\partial \mu_1^k\left(X; \beta_1^k\right)}{\partial \beta_1^k} \left\{Y - \mu_1^k\left(X; \beta_1^k\right)\right\}, \end{aligned}$$

for $j = 1, \ldots, J$ and $k = 1, \ldots, K$. Then, $\hat{\theta}$ solves the joint estimating equation $\mathcal{U}_n(\theta) = n^{-1/2} \sum_{i=1}^{n} U(A_i, X_i, Y_i; \theta) = 0$, where $U(\theta)$ stacks $U_1^j(A_i, X_i, \alpha^j)$ for $j = 1, \ldots, J$, $U_2^k\left(A_i, X_i, Y_i; \beta_0^k\right)$ and $U_3^k\left(A_i, X_i, Y_i; \beta_1^k\right)$ for $k = 1, \ldots, K$.

Let θ^* be the probability limit of $\hat{\theta}$. We divide our theoretical investigation into two steps: first, we establish the asymptotic results for the DSM estimator when θ^* is known, and second, building on the step-one results, we quantify the impact of the estimation of θ^* on the asymptotic distribution.

4.1 | Asymptotic results with known θ^*

We allow possible model misspecification, so θ^* may not be the true parameter values. If $e^j(X; \alpha^j)$ is a correctly specified model, we have $e^j(X; \alpha^{j*}) = e(X)$; if $\mu_a^k(X; \beta_a^k)$ is a correctly specified model, we have $\mu_a^k(X; \beta_a^{k*}) = \mu_a(X)$, for $a = 0, 1$. The key insight is that if any model of the propensity score or prognostic score is correctly specified, $S_a(\theta^*)$ remains as a balancing score in the sense that $Y(a) \perp\!\!\!\perp A | S_a(\theta^*)$ holds for $a = 0, 1$ (Lemma 2). In the following theorem, we establish the multiple robustness and asymptotic distribution of $\hat{\tau}_{\text{dsm}}(\theta^*)$.

Theorem 1. *Under Assumption 1 and regularity conditions in Assumption A1, if any model of the propensity score or prognostic score is correctly specified, we have* $n^{1/2}\{\hat{\tau}_{\text{dsm}}(\theta^*) - \tau\} \to \mathcal{N}(0, V_\tau)$, *in distribution, as* $n \to \infty$, *where*

$$V_\tau = \mathbb{E}\left[\{\mu_1(S_1) - \mu_0(S_0) - \tau\}^2\right] + \mathbb{E}\left(\sigma_1^2(S_1)\left[\frac{1}{e(S_1)} + \frac{1}{2M}\left\{\frac{1}{e(S_1)} - e(S_1)\right\}\right]\right) + \mathbb{E}\left(\sigma_0^2(S_0)\left[\frac{1}{1-e(S_0)} + \frac{1}{2M}\left\{\frac{1}{1-e(S_0)} - 1 + e(S_0)\right\}\right]\right). \quad (5)$$

Remark 1 (Connection with AIPW). The de-biased matching estimator has an intrinsic connection with the AIPW estimator. The AIPW estimator of τ is

$$\hat{\tau}_{\text{aipw}} = n^{-1}\sum_{i=1}^{n}\left[\mu_1(X_i) - \mu_0(X_i) + \frac{2A_i - 1}{\widehat{\mathbb{P}}(A_i \mid X_i)}\{Y_i - \mu_A(X_i)\}\right] + o_{\text{P}}\left(n^{-1/2}\right). \quad (6)$$

It is well-known that $\hat{\tau}_{\text{aipw}}$ is doubly robust and semiparametrically efficient when the outcome model and the propensity score model are correctly specified. In the proof of Theorem 1, we show that the DSM estimator $\hat{\tau}_{\text{dsm}}(\theta^*)$ exhibits a similar asymptotic expansion as the AIPW estimator

$$\hat{\tau}_{\text{dsm}}(\theta^*) = n^{-1}\sum_{i=1}^{n}\left[\mu_1(S_{1,i}) - \mu_0(S_{0,i}) + (2A_i - 1)\left(1 + M^{-1}K_{S_{A,i}}\right)\left\{Y_i - \mu_A(S_{A,i})\right\}\right] + o_{\text{P}}(n^{-1/2}). \quad (7)$$

When one of the prognostic score models is correctly specified, $\mu_a(S_{a,i}) = \mu_a(X_i)$ for $a = 0, 1$. From Theorem 1, the asymptotic variance of $\hat{\tau}_{\text{dsm}}(\theta^*)$ (5) does not achieve the semiparametric efficiency bound (Hahn, 1998) for a fixed M, but it becomes closer when M is larger. However, $\hat{\tau}_{\text{aipw}}$ is extremely unstable when the estimated propensity scores are close to zero or one due to the weight construction, and it is sensitive to slight model misspecification. Instead of inverting the estimated propensity score, the DSM estimator uses the matching weights, $1 + M^{-1}K_{S_A,i}$, for covariate balancing and hence is more robust in the scenarios with extreme propensity score values. This is confirmed in the simulation study.

Remark 2. From Theorem 1, the consistency of the DSM estimator is guaranteed if any model for the propensity score or prognostic score is correctly specified. Both the number of the posited models and their functional forms can affect the efficiency of the DSM estimator in a very complex way. In addition, with a finite sample size, the matching performance can be unstable if there are a large number of working models. In particular, the discrepancy of the matched units may be large when some of the models are poorly constructed. To reduce the chance of running into these issues, we suggest positing a few well-constructed working models instead of a large number of poorly built ones.

4.2 | Asymptotic results with estimated θ^*

To acknowledge the fact that θ^* is estimated prior to matching, we will establish the approximate distribution of $\hat{\tau}_{\text{dsm}}(\hat{\theta})$ and examine the impact of nuisance parameter estimation on the properties of the DSM estimator. As in Abadie and Imbens (2016), the typical Taylor expansion technique can not be used because of the nonsmooth nature of matching. Our derivation is based on the technique developed by Andreou and Werker (2012), which offers a general approach for deriving the limiting distribution of statistics that involve estimated nuisance parameters. This technique has been successfully used by Abadie and Imbens (2016) for the PSM estimators of the

ATE and ATT based on a correctly specified propensity score model. We extend their results to the DSM estimator requiring only one of the double score models to be correctly specified.

Theorem 2. *Under Assumption 1 and regularity conditions in Assumptions A1–A4, if any model of the propensity score or prognostic score is correctly specified, the approximate distribution of* $n^{1/2}\left\{\hat{\tau}_{\text{dsm}}(\hat{\theta})-\tau\right\}$ *is* $\mathcal{N}(0, V_{\tau,\text{adj}})$, *where*

$$V_{\tau,\text{adj}} = V_\tau - \gamma_1^{\mathrm{T}}\Sigma_U^{-1}\gamma_1 + \gamma_2^{\mathrm{T}}\Sigma_{\theta^*}\gamma_2, \tag{8}$$

where V_τ *is given in (5),* $\Sigma_U = \mathbb{E}\{U(A,X,Y;\theta^*)\ U(A,X,Y;\theta^*)^{\mathrm{T}}\}$, $\Sigma_{\theta^*} = \Gamma_{\theta^*}^{-1}\Sigma_U(\Gamma_{\theta^*}^{-1})^{\mathrm{T}}$, $\Gamma_{\theta^*} = \mathbb{E}\left\{\partial U(A,X,Y;\theta^*)/\partial\theta^{\mathrm{T}}\right\}$, γ_1 *and* γ_2 *are given in (A14) and (A10), respectively.*

We discuss the impact of estimating the nuisance parameters on the matching estimators. Abadie and Imbens (2016) showed that for τ, matching on the estimated propensity score always improves the estimation efficiency compared to matching on the true propensity score. This improvement is due to the correlation of the matching estimator and the score function for the parameters in the propensity score. In our context, comparing the asymptotic variances in Theorems 1 and 2, the difference between $V_{\tau,\text{adj}}$ and V_τ, $-\gamma_1^{\mathrm{T}}\Sigma_U^{-1}\gamma_1 + \gamma_2^{\mathrm{T}}\Sigma_{\theta^*}\gamma_2$, can be either positive, negative, or zero; that is, matching on the estimated double score can either increase, decrease, or maintain the estimation efficiency compared to matching on the true double score. To explain the difference, we note that the variance reduction term $-\gamma_1^{\mathrm{T}}\Sigma_U^{-1}\gamma_1$ is still due to the correlation of the matching estimator and the score function for the parameters in the double score, while the variance inflation term $\gamma_2^{\mathrm{T}}\Sigma_{\theta^*}\gamma_2$ is because if either the prognostic score model or the propensity score model is misspecified, τ may depend on the nuisance parameters through $\tau = \mathbb{E}\left[\mu_1\{S_1(\theta^*)\} - \mu_0\{S_0(\theta^*)\}\right]$, which contributes to the variance inflation term. On the other hand, Abadie and Imbens (2016) focused on the setting when the propensity score model is the only nuisance model and is correctly specified. In this case, τ does not depend on α^*, γ_2 is zero, and therefore the variance inflation term is zero.

4.3 | Variance estimation and inference

Theorem 2 provides guidance for variance estimation of the DSM estimators that can take all sources of variability into account. However, such variance estimators are complicated to construct. We consider variance estimation based on replication methods (Efron, 1979; Wolter, 2007). Lack of smoothness makes the standard replication methods invalid for the matching estimator. When the number of matches remains fixed, Abadie and Imbens (2008) demonstrated the failure of the bootstrap for matching estimators. This is because the nonparametric bootstrap cannot preserve the distribution of the number of times that each unit is used as a match. In this case, Otsu and Rai (2017) proposed a wild bootstrap procedure for the matching estimator when matching is directly based on the covariates. Yang and Kim (2020) proposed a replication-based procedure for predictive mean matching in survey data.

Given the two-stage estimation procedure for the DSM estimator, the variability of the matching estimator results from two sources: first, the estimation of the double score function, and second, matching. Following Yang and Kim (2020), we propose a two-stage replication variance estimation procedure, in parallel to the two-stage point estimation procedure. First, we construct replicates of the nuisance parameter estimators in the double score. Second, based on the asymptotic linear representations of the DSM estimator, we construct replicates of the DSM

estimator directly based on the linear terms with the replicated nuisance parameters. In this way, the distribution of the number of times that each unit is used as a match can be retained.

Specifically, the replication variance estimation algorithm proceeds as follows.

VE-Step 1. Obtain a bootstrap sample, or equivalently the bootstrap replication weights $\omega_i^* = n^{-1}m_i^*$ with $(m_1^*, \dots, m_n^*)$ is a multinomial random vector with n draws on n equal probability cells. Obtain a bootstrap replicate of $\hat{\theta}$, $\hat{\theta}^*$, by solving the estimating equation $n^{-1/2}\sum_{i=1}^n \{\omega_i^* U(A_i, X_i, Y_i; \theta)\} = 0$. For each unit i, calculate $S_{a,i}(\hat{\theta}^*)$ for $a = 0, 1$.

VE-Step 2. Obtain a bootstrap replicate of $\hat{\tau}_{\text{dsm}}(\hat{\theta})$ as

$$\hat{\tau}_{\text{dsm}}^*(\hat{\theta}^*) = n^{-1}\sum_{i=1}^n \omega_i^* \left[\hat{\mu}_1\{S_{1,i}(\hat{\theta}^*)\} - \hat{\mu}_0\{S_{0,i}(\hat{\theta}^*)\}\right] + n^{-1}\sum_{i=1}^n \omega_i^*(2A_i - 1)\left\{1 + M^{-1}K_{S_{A_i}(\hat{\theta}),i}\right\}\left[Y_i - \hat{\mu}_{A_i}\{S_{A_i,i}(\hat{\theta}^*)\}\right].$$

VE-Step 3. Repeat VE-Steps 1 and 2 a large number of times. Calculate the bootstrap variance estimator of $\hat{\tau}_{\text{dsm}}(\hat{\theta})$ as the empirical variance of $\hat{\tau}_{\text{dsm}}^*(\hat{\theta}^*)$ over a large number of bootstrap replicates.

5 | MULTIPLY ROBUST MATCHING ESTIMATOR OF THE QTE

Matching is attractive for general causal estimation because it can be viewed as a hot deck imputation method (Ford, 1983), where for each unit the donors for the missing potential outcome are actually observed values from the opposite treatment group. An advantage of hot deck imputation is that it preserves the distribution of the potential outcomes so that valid estimators for parameters depending on the entire distribution of the potential outcomes such as the mean and quantiles can be obtained based on the imputed dataset. In this section, we extend the DSM framework to estimate the QTE.

We focus on estimating $q_{a,\xi}$ separately for $a = 0, 1$. By Lemma 1, we have

$$q_{a,\xi} = \inf_q \left(\mathbb{E}[\mathbb{P}\{Y \le q \mid A = a, e(X), \Psi_a(X)\}] \ge \xi\right).$$

Based on the above equation, we propose the DSM estimator of $q_{a,\xi}$ as

$$\hat{q}_{a,\xi,\text{dsm}} = \inf_q \left\{\hat{F}_{a,\text{dsm}}(q) \ge \xi\right\}, \tag{9}$$

where

$$\hat{F}_{a,\text{dsm}}(q) = \hat{F}_{a,\text{dsm}}^{(0)}(q) - n^{-1/2}\hat{B}_{a,n}(q), \tag{10}$$

$$\hat{F}_{a,\text{dsm}}^{(0)}(q) = n^{-1}\sum_{i=1}^n \mathbf{1}(A_i = a)\left(1 + M^{-1}K_{S_a,i}\right)\mathbf{1}(Y_i \le q),$$

$$\hat{B}_{a,n}(q) = -n^{-1/2}\sum_{i=1}^n \mathbf{1}(A_i = 1 - a)M^{-1}\sum_{j\in\mathcal{J}_{S_{a,i}}}\left\{\hat{F}_a(q; S_{a,i}) - \hat{F}_a(q; S_{a,j})\right\}, \tag{11}$$

and $\hat{F}_a(q;S_a)$ is a semi/nonparametric estimator of $F_a(q;S_a)=\mathbb{P}\{Y(a)\le q\mid S_a\}$, for $a=0,1$. Note that $\hat{F}^{(0)}_{a,\mathrm{dsm}}(q)$ is an initial matching estimator of $F_a(q)=\mathbb{P}\{Y(a)\le q\}$, $\hat{B}_{a,n}(q)$ is the bias correction term. Then the DSM estimator of Δ_ξ is $\hat{\Delta}_{\xi,\mathrm{dsm}}=\hat{q}_{1,\xi,\mathrm{dsm}}-\hat{q}_{0,\xi,\mathrm{dsm}}$.

For estimating Δ_ξ, Steps 1 and 2 of DSM in Section 3.3 remain the same; Steps 3′ and 4′ proceed as follows:

Step 3′. Obtain a semiparametric estimator of $F_a(q;S_a)$ based on $[\{Y_i,S_{a,i}(\hat{\theta})\}:A_i=a]$, for $a=0,1$.

Step 4′. The DSM estimator of $q_{a,\xi}$ is given by (9) with S_a replaced by $S_a(\hat{\theta})$. We denote the final estimator of $q_{a,\xi}$ as $\hat{q}_{a,\xi\mathrm{dsm}}(\hat{\theta})$ to reflect its dependence on $\hat{\theta}$, for $a=0,1$. Then, the DSM estimator of Δ_ξ is $\hat{\Delta}_{\xi,\mathrm{dsm}}(\hat{\theta})=\hat{q}_{1,\xi,\mathrm{dsm}}(\hat{\theta})-\hat{q}_{0,\xi,\mathrm{dsm}}(\hat{\theta})$.

Remark 3. In Step 3′, many choices can be considered for estimating $F_a(q;S_a)$: for example, the method of sieves for the normal linear model after a Box-Cox transformation of Zhang et al. (2012), the single-index conditional distribution model of Chiang and Huang (2012), or the distribution regression models of Foresi and Peracchi (1995) and Chernozhukov et al. (2013). The data at hand and subject matter knowledge can be used to guide the choice of the models. For example, if the transformed outcome is believed to follow a normal distribution, Zhang et al. (2012)'s method is preferable; otherwise, other semiparametric approaches are desirable.

Under Assumption 1, regularity conditions in Assumptions A1 (i) and S1, if any model of the propensity score or prognostic score is correctly specified, similar to the proof in Section A.1, we have $\mathrm{d}\hat{F}_{a,\mathrm{dsm}}(q_{a,\xi})/\mathrm{d}q=f_a(q_{a,\xi})+o_{\mathrm{P}}(n^{-1/2})$, and then we express $\hat{q}_{a,\xi,\mathrm{dsm}}$ as

$$\hat{q}_{a,\xi,\mathrm{dsm}}-q_{a,\xi}=-\frac{\hat{F}_{a,\mathrm{dsm}}(q_{a,\xi})-F_a(q_{a,\xi})}{f_a(q_{a,\xi})}+o_P\left(n^{-1/2}\right), \tag{12}$$

$q_{a,\xi}$ lies in a closed interval $\mathcal{I}$. Expression (12) is called the Bahadur-type representation for $\hat{q}_{a,\xi,\mathrm{dsm}}$ (Francisco & Fuller, 1991). With the representation (12), we can then extend the multiple robustness and asymptotic distributions of the ATE estimation to $\hat{\Delta}_{\xi,\mathrm{dsm}}(\theta^*)$ and $\hat{\Delta}_{\xi,\mathrm{dsm}}(\hat{\theta})$.

Theorem 3. *Under Assumption 1, regularity conditions in Assumptions A1 (i) and S1, if any model of the propensity score or prognostic score is correctly specified, for all* $\xi\in\tilde{\mathcal{I}}=\{\xi: F_a(x)=\xi, x\in\mathcal{I}\}$, $n^{1/2}\{\hat{\Delta}_{\xi,\mathrm{dsm}}(\theta^*)-\Delta_\xi\}\to\mathcal{N}(0,V_\xi)$, *in distribution, as* $n\to\infty$, *where* V_ξ *is given in (S3).*

Theorem 4. *Under Assumption 1, regularity conditions in Assumptions A1 (i) and S1, if any model of the propensity score or prognostic score is correctly specified, the approximate distribution of* $n^{1/2}\left\{\hat{\Delta}_{\xi,\mathrm{dsm}}(\theta^*)-\Delta_\xi\right\}$ *is* $\mathcal{N}(0,V_{\xi,\mathrm{adj}})$, *where*

$$V_{\xi,\mathrm{adj}}=V_\xi-\gamma_3^{\mathrm{T}}\Sigma_U^{-1}\gamma_3+\gamma_4^{\mathrm{T}}\Sigma_{\theta^*}\gamma_4, \tag{13}$$

V_ξ *is given in (S3),* Σ_U *and* Σ_{θ^*} *are given in Theorem 3,* γ_3 *and* γ_4 *are given in (S4) and (S5), respectively.*

For variance estimation of $\hat{\Delta}_{\xi,\mathrm{dsm}}(\hat{\theta})$, VE-Step 1 in Section 4.3 remains the same; VE-Steps 2 and 3 proceed as follows:

VE-Step 2′. For $a = 0, 1$, obtain a bootstrap replicate of $\hat{q}_{a,\xi\mathrm{dsm}}(\hat{\theta})$, $\hat{q}^*_{a,\xi\mathrm{dsm}}(\hat{\theta}^*)$, by solving

$$\hat{F}^*_{a,\mathrm{dsm}}(q) = n^{-1}\sum_{i=1}^{n}\omega_i^*\hat{F}_a\{q; S_{a,i}(\hat{\theta}^*)\} + n^{-1}\sum_{i=1}^{n}\omega_i^*\mathbf{1}(A_i = a)\left\{1 + M^{-1}K_{S_a(\hat{\theta}),i}\right\}\left[\mathbf{1}(Y_i \le q) - \hat{F}_a\{q; S_{a,i}(\hat{\theta}^*)\}\right] = \xi,$$

for q. Then a bootstrap replicate of $\hat{\Delta}_{\xi,\mathrm{dsm}}(\hat{\theta})$ is $\hat{\Delta}^*_{\xi,\mathrm{dsm}}(\hat{\theta}^*) = \hat{q}^*_{1,\xi,\mathrm{dsm}}(\hat{\theta}^*) - \hat{q}^*_{0,\xi,\mathrm{dsm}}(\hat{\theta}^*)$.

VE-Step 3′. Repeat VE-Steps 1 and 2′ for a large number of times. Calculate the bootstrap variance estimator of $\hat{\Delta}_{\xi,\mathrm{dsm}}(\hat{\theta})$ as the empirical variance of $\hat{\Delta}^*_{\xi,\mathrm{dsm}}(\hat{\theta}^*)$ over a large number of bootstrap replicates.

6 | EXTENSIONS TO THE CAUSAL EFFECTS ON THE TREATED

In this extension, we estimate the average causal effect on the treated τ_{ATT} and the QTE on the treated $\Delta_{\mathrm{ATT},\xi} = q_{1,\xi|A=1} - q_{0,\xi|A=1}$, where $q_{a,\xi|A=1} = \inf_q[P\{Y(a) \le q\} \ge \xi \,|\, A = 1]$, $a = 0, 1$. Here, because $f\{Y(1) \,|\, A = 1\} = f(Y \,|\, A = 1)$, the outcome distribution for the treated is identifiable. Therefore, $\mathbb{E}\{Y(1) \,|\, A = 1\} = \mathbb{E}(Y \,|\, A = 1)$ and $q_{1,\xi|A=1} = \inf_q \{\mathbb{P}(Y \le q \,|\, A = 1) \ge \xi\}$.

To identify the outcome distribution for the control, Assumption 1 can be relaxed (Heckman et al., 1997).

Assumption 2. (i) $Y(0) \perp\!\!\!\perp A \,|\, X$; and (ii) there exists a constant c such that $e(X) \le c < 1$ almost surely.

For the causal effects on the treated, the prognostic score $\Psi_0(X)$ is a sufficient statistic for $Y(0)$ in the sense that $Y(0) \perp\!\!\!\perp X|\Psi_0(X)$ according to Hansen (2008). Then, under Assumption 2,

$$\begin{aligned}\tau_{\mathrm{ATT}} &= \mathbb{E}[\mathbb{E}(Y|A = 1) - \mathbb{E}\{Y \,|\, A = 0, e(X)\} \,|\, A = 1] \\ &= \mathbb{E}[\mathbb{E}(Y \,|\, A = 1) - \mathbb{E}\{Y \,|\, A = 0, \Psi_0(X)\} \,|\, A = 1],\end{aligned}$$

and

$$\begin{aligned}q_{0,\xi|A=1} &= \inf_q (\mathbb{E}[\mathbb{P}\{Y \le q \,|\, A = 0, e(X)\} \,|\, A = 1] \ge \xi) \\ &= \inf_q (\mathbb{E}[\mathbb{P}\{Y \le q \,|\, A = 0, \Psi_0(X)\} \,|\, A = 1] \ge \xi),\end{aligned}$$

encoding the double balancing properties of $S = \{e(X), \Psi_0(X)\}$.

The DSM estimators for τ_{ATT} and $\Delta_{\mathrm{ATT},\xi}$ follow similar steps as for τ and Δ_ξ. We describe the differences below.

In the matching step, for each unit i with treatment $A_i = 1$, find M nearest neighbors from the control group $A_i = 0$ based on the matching variable $S_i = S_i(\hat{\theta})$. Let these matched units for unit i be indexed by $\mathcal{J}_{S(\hat{\theta}),i}$.

The initial and de-biasing DSM estimators of τ_{ATT} are

$$\hat{\tau}^{(0)}_{\text{ATT,dsm}} = n_1^{-1}\sum_{i=1}^{n} A_i\left\{Y_i - \hat{Y}_i(0)\right\},\ \hat{Y}_i(0) = M^{-1}\sum_{j\in\mathcal{J}_{S,i}} Y_j,$$

$$\hat{\tau}_{\text{ATT,dsm}} = \hat{\tau}^{(0)}_{\text{ATT,dsm}} - n_1^{-1}\sum_{i=1}^{n} A_i\left\{\hat{\mu}_0(S_i) - M^{-1}\sum_{j\in\mathcal{J}_{S,i}}\hat{\mu}_0(S_j)\right\}.$$

Let the estimator of $F_1(q|A=1) = \mathbb{P}\{Y(1) < q \,|\, A=1\}$ be

$$\hat{F}_1(q\,|\,A=1) = n_1^{-1}\sum_{i=1}^{n} A_i\mathbf{1}(Y_i \le q).$$

Then, we estimate $q_{1,\xi|A=1}$ by

$$\hat{q}_{1,\xi|A=1} = \inf_q\{\hat{F}_1(q\,|\,A=1) \ge \xi\}.$$

The initial and de-biasing DSM estimators of $F_0(q\,|\,A=1) = \mathbb{P}\{Y(0) < q\,|\,A=1\}$ are

$$\hat{F}^{(0)}_{0,\text{dsm}}(q|A=1) = n_1^{-1}\sum_{i=1}^{n} A_i M^{-1}\sum_{j\in\mathcal{J}_{S,i}}\mathbf{1}(Y_j \le q) = n_1^{-1}\sum_{i=1}^{n}(1-A_i)M^{-1}K_{S,i}\mathbf{1}(Y_i \le q),$$

$$\hat{F}_{0,\text{dsm}}(q|A=1) = \hat{F}^{(0)}_{0,\text{dsm}}(q|A=1) - n_1^{-1/2}\hat{B}_{0,n}(q),$$

$$\hat{B}_{0,n}(q) = -n_1^{-1/2}\sum_{i=1}^{n} A_i M^{-1}\sum_{j\in\mathcal{J}_{S,i}}\left\{\hat{F}_0(q;S_i) - \hat{F}_0(q;S_j)\right\}.$$

Then, we estimate $q_{0,\xi|A=1}$ by

$$\hat{q}_{0,\xi|A=1,\text{dsm}} = \inf_q\{\hat{F}_{0,\text{dsm}}(q|A=1) \ge \xi\}.$$

Lastly, the DSM estimator of $\Delta_{\text{ATT},\xi}$ is $\hat{\Delta}_{\text{ATT},\xi,\text{dsm}} = \hat{q}_{1,\xi|A=1} - \hat{q}_{0,\xi|A=1,\text{dsm}}$.

For variance estimation, we replace the VE-Step 2 and VE-Step 2' for τ and Δ_ξ by the following steps:

ATT-VE-Step 2. Obtain a bootstrap replicate of $\hat{\tau}_{\text{ATT,dsm}}(\hat{\theta})$,

$$\hat{\tau}^*_{\text{ATT,dsm}}(\hat{\theta}^*) = n_1^{-1}\sum_{i=1}^{n}\omega_i^* A_i\left[\hat{\mu}_1\{S_i(\hat{\theta}^*)\} - \hat{\mu}_0\{S_i(\hat{\theta}^*)\}\right]$$
$$+ n_1^{-1}\sum_{i=1}^{n}\omega_i^*\left\{A_i - (1-A_i)M^{-1}K_{S(\hat{\theta}),i}\right\}\left[Y_i - \hat{\mu}_{A_i}\{S_i(\hat{\theta}^*)\}\right].$$

QTT-VE-Step 2′. For $a=1$, obtain a bootstrap replicate of $\hat{q}_{1,\xi|A=1}(\hat{\theta})$, $\hat{q}^*_{1,\xi|A=1}(\hat{\theta}^*)$, by solving

$$\hat{F}^*_1(q|A=1) = n_1^{-1}\sum_{i=1}^{n}\omega_i^* A_i\mathbf{1}(Y_i \le q) = \xi.$$

For $a = 0$, obtain a bootstrap replicate of $\hat{q}_{0,\xi|A=1,\mathrm{dsm}}(\hat{\theta})$, $\hat{q}^*_{0,\xi|A=1,\mathrm{dsm}}(\hat{\theta}^*)$, by solving

$$\hat{F}^*_{0,\mathrm{dsm}}(q|A=1) = n_1^{-1}\sum_{i=1}^{n}\omega_i^* A_i \hat{F}_0\{q; S_i(\hat{\theta}^*)\} + n_1^{-1}\sum_{i=1}^{n}\omega_i^* \mathbf{1}(A_i=0) M^{-1} K_{S(\hat{\theta}),i}\left[\mathbf{1}(Y_i \le q) - \hat{F}_0\{q; S_i(\hat{\theta}^*)\}\right] = \xi,$$

for q. Then a bootstrap replicate of $\hat{\Delta}_{\mathrm{ATT},\xi,\mathrm{dsm}}(\hat{\theta})$ is $\hat{\Delta}^*_{\mathrm{ATT},\xi,\mathrm{dsm}}(\hat{\theta}^*) = \hat{q}^*_{1,\xi|A=1}(\hat{\theta}^*) - \hat{q}^*_{0,\xi|A=1,\mathrm{dsm}}(\hat{\theta}^*)$.

7 | SIMULATION STUDY

We conduct a simulation study to investigate the finite-sample performance of the proposed DSM estimators relative to existing weighting and matching estimators. In the causal inference and missing data literature, previous simulations (e.g., Kang & Schafer, 2007) have found that weighting estimators can have high variability, especially if the probabilities are close to zero or one. Frölich (2004) found that the weighting estimator was inferior to matching estimators in terms of root mean squared error. It has been found that matching on high-dimensional covariates is not practical for commonly found sample sizes (e.g., Abadie & Imbens, 2006). In the comparative effectiveness research, PGM has been shown to be more advantageous than PSM when the propensity score distributions are strongly separated (Kumamaru et al., 2016; Wyss et al., 2015). Imbens (2004) noted that if the regression models are misspecified, PGM may be inconsistent. These results motivate us to compare the weighting and matching estimators in a setting with complex data generative models, and where the propensity scores may be extreme (i.e., close to zero or one) or nonextreme.

Let the sample size be $n = 1000$. Confounder $X \in \mathbb{R}^{10}$ is generated by $X_j \overset{\mathrm{iid}}{\sim}$ Uniform $[1-\sqrt{3}, 1+\sqrt{3}]$ for $j = 1, \ldots, 10$. To introduce nonlinear relationships between X and dependent variables, let $Z \in \mathbb{R}^{10}$ be a nonlinear transformation of X, where $Z_1 = \exp(X_1/2)$, $Z_2 = \exp(X_2/3)$, $Z_3 = \log\{(X_3+1)^2\}$, $Z_4 = \log\{(X_4+1)^2\}$, $Z_5 = \mathbf{1}(X_5 > 0.5)$, $Z_6 = \mathbf{1}(X_6 > 0.75)$, $Z_7 = \sin(X_7 - X_8)$, $Z_8 = \cos(X_7 + X_8)$, $Z_9 = \sin(X_9)$, and $Z_{10} = \cos(X_{10})$, which are further scaled and centered such that $\mathbb{E}(Z_j) = 1$ and $\mathbb{V}(Z_j) = 1$ for all j. The potential outcomes are $Y(0) = \beta_0^{\mathrm{T}} Z + \epsilon(0)$ and $Y(1) = Y(0) - \epsilon(0) + \epsilon(1)$, where $\beta_0^{\mathrm{T}} = (0.1, 1, 1, 1, 1, -1, -1, -1, -1, -1)$, $\epsilon(0) \sim \mathcal{N}(0, 2^2)$, and $\epsilon(1) \sim \mathcal{N}(0,1)$. Under the data generative model, the ATE τ is 0 and the 75th QTE is -0.45. An additional simulation with heterogeneous treatment effects and log-normal errors is presented in the supplementary material. The treatment indicator A follows Bernoulli$\{e(X)\}$, where $\mathrm{logit}\{e(X)\} = \alpha_0^{\mathrm{T}} Z$. We consider two scenarios for the propensity score distribution: in the first case, $\alpha_0^{\mathrm{T}} = (5, -5, 1, 1, 2, -2, -2, 1, -1, -1)$, resulting in some extreme values of $e(X)$ that are close to zero or one; and in the second case, the propensity score distribution is not extreme, where $\alpha_0^{\mathrm{T}} = (5, -5, 1, 1, 2, -2, -2, 1, -1, -1)/4$. Visualization of the propensity score distributions in the two cases is presented in Figure S2.

To assess the multiple robustness property of the DSM estimators, we consider two model specifications for the propensity score: (1) a correctly specified logistic regression model $\mathrm{logit}\{e^1(X; \alpha^1)\} = \alpha^{1,\mathrm{T}} Z$; and (2) a misspecified logistic regression model $\mathrm{logit}\{e^2(X; \alpha^2)\} = \alpha^{2,\mathrm{T}} X$; we also consider two model specifications for the prognostic score; (3) a correctly specified

regression model $\mu_a^1(X;\beta_a^1) = \beta_a^{1,\mathrm{T}} Z$ for $a = 0, 1$; and (4) a misspecified regression model $\mu_a^2\left(X;\beta_a^2\right) = \beta_a^{2,\mathrm{T}} X$ for $a = 0, 1$.

We compare the following estimators:

(a) naive, which is the simple difference of standard estimators from two treatment groups;
(b) the weighting estimators including the IPW, AIPW, and MRW estimators ("ipw," "aipw," and "mrw");
c) the matching estimators based on X ("m.x"; bias-corrected Abadie & Imbens, 2011), or propensity score ("psm") or prognostic score ("pgm") or double score ("dsm") with $M = 5$.

Each weighting and matching estimator is assigned a name in the form of "method-0000," where each digit of the four-digit number, from left to right, indicates if $e^1(X;\alpha^1)$, $e^2(X;\alpha^2)$, $\{\mu_a^1(X;\beta_a^1)\}_{a=0}^1$, or $\{\mu_a^2(X;\beta_a^2)\}_{a=0}^1$ is used in the construction of the method, with "1" meaning yes and "0" meaning no, respectively. For example, "ipw1000" is the IPW estimator with the propensity score model $e^1(X;\alpha^1)$ and "dsm1110" is the DSM estimator with two propensity score models $e^1(X;\alpha^1)$, $e^2(X;\alpha^2)$ and one prognostic score model $\{\mu_a^1(X;\beta_a^1)\}_{a=0}^1$. We implement standard IPW and AIPW estimators for the ATE estimation and the corresponding estimators of Zhang et al. (2012) for the QTE estimation. MRW is implemented by the R package "MultiRobust." For all matching estimators, the conditional outcome mean functions are approximated using power series, and the conditional distribution functions are approximated based on the power series for the normal linear model (Zhang et al., 2012).

Figure 1 shows the distributions of the estimation error (i.e., the estimator minus the true parameter value) based on 1000 repeated sampling. The naive estimator is biased for the 75th

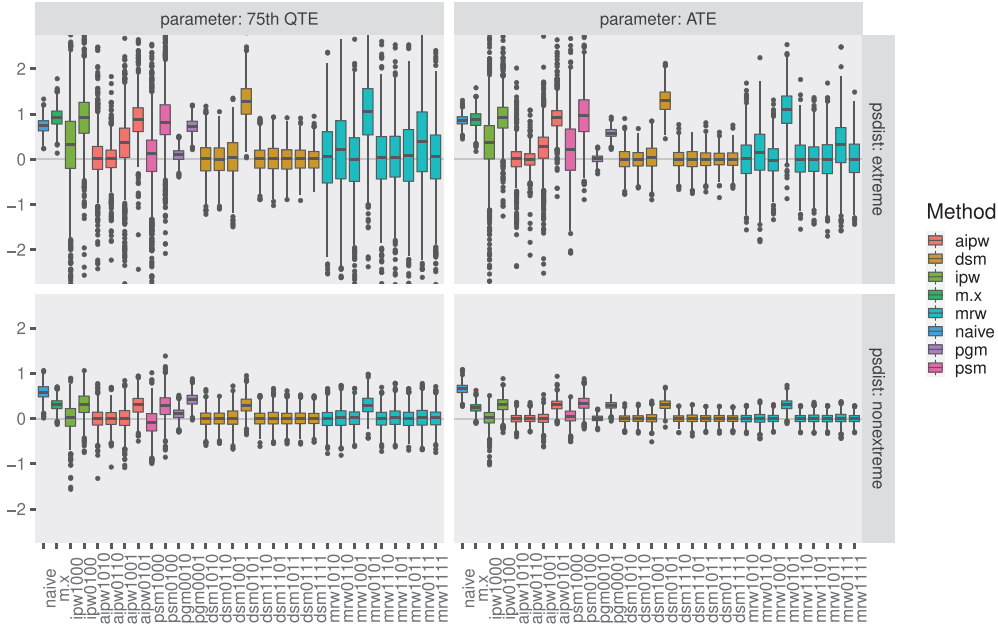

FIGURE 1 Simulation results of various weighting and matching estimators. There are four panels of results: the left for the 75th quantile treatment effect, the right for the ATE, the top for the extreme propensity score, and the bottom for the nonextreme propensity score. Each box plot shows the distribution of the estimator subtracting the true parameter value based on 1000 Monte Carlo simulated datasets

QTE and ATE. Matching directly based on 10-dimensional X (indicated by "m.x") is biased for the QTE and ATE even with bias correction. This suggests that matching on high-dimensional covariates is not practical and calls for dimension reduction. IPW is unstable and sensitive to the extreme values of the propensity score. Even when the propensity score model is correctly specified (indicated by "ipw1000"), IPW is biased of the QTE and ATE. By theory, AIPW is supposed to be doubly robust: it should have small biases if either the propensity score model or the prognostic model is correctly specified. However, as indicated by "aipw1001," AIPW is biased of the QTE and ATE when the propensity score is extreme even if its model is correctly specified. We examine the empirical distribution of the estimated propensity score weights and find that there are extremely large weights that dominate other weights. Therefore, weighting estimators by inverting the estimated propensity scores are sensitive to outliers of the propensity score estimates. To mitigate this issue, one can stabilize the weighting estimators by normalizing the weights (Hernán et al., 2001). However, this strategy is not effective in our setting. Although AIPW is constructed to be semiparametrically efficient, its performance can be poor when it involves large weights. By construction, matching does not invert the estimated propensity scores and therefore is more robust to outliers of the propensity score estimates. We now compare the performances of the score-based matching estimators. The single score matching estimators (indicated by "psm1000," "psm0100," "pgm0010," "pgm0001") are singly robust and rely on a correct specification of the underlying score model. DSM and MRW are multiply robust in that they have small biases for the QTE and the ATE if any model of the propensity score or prognostic score is correctly specified. Compared to MRW, DSM is more robust to extreme values of the propensity score estimates. Therefore, DSM is advantageous in practice compared to weighting.

Table 3 reports the coverage rates of the DSM estimators of the 75th QTE and the ATE using the proposed replication-based method. Under the multiple robustness condition (i.e., if any model of the propensity score or prognostic score is correctly specified), the coverage rates are all close to the nominal coverage except for "dsm0101."

TABLE 3 Simulation results based on 1000 Monte Carlo simulated datasets for the coverage properties of the double score matching estimators using the replication-based method: empirical coverage rate and (empirical coverage rate ± 1.96×Monte Carlo standard error)

	Nonextreme PS				**Extreme PS**			
	75th QTE		**ATE**		**75th QTE**		**ATE**	
"dsm1010"	94.8	(93.4,96.2)	95.2	(93.9,96.5)	96.0	(94.8, 97.2)	95.8	(94.6, 97.0)
"dsm0110"	94.4	(93.0,95.8)	96.0	(94.8,97.2)	95.7	(94.4, 97.0)	95.2	(93.9, 96.5)
"dsm1001"	95.4	(94.1,96.7)	95.9	(94.7,97.1)	95.9	(94.7, 97.1)	95.7	(94.4, 97.0)
"dsm0101"	72.9	(70.1,75.6)	29.2	(26.4,32.0)	19.4	(16.9, 21.9)	0.2	(-0.1, 0.5)
"dsm1111"	95.0	(93.6,96.4)	95.5	(94.2,96.8)	95.4	(94.1, 96.7)	95.8	(94.6, 97.0)
"dsm1110"	95.4	(94.1,96.7)	95.6	(95.5,97.7)	95.7	(94.4, 97.0)	96.2	(95.0, 97.4)
"dsm1101"	94.8	(93.4,96.2)	95.4	(94.1,96.7)	96.3	(95.1, 97.5)	95.4	(94.1, 96.7)
"dsm1011"	94.0	(92.5,95.5)	95.8	(94.6,97.0)	95.7	(94.4, 97.0)	95.8	(94.6, 97.0)
"dsm0111"	95.3	(94.0,96.6)	95.2	(93.9,96.5)	96.0	(94.8, 97.2)	95.4	(94.1, 96.6)

8 | REAL-DATA APPLICATION

In this section, we apply the proposed DSM method as well as other matching methods in Section 7 to the well-known National Supported Work (NSW) data (Firpo, 2007; LaLonde, 1986). This dataset documented the effect of a job training program for the unemployed on future earnings. Following Dehejia and Wahba (1999), we include the comparison group from Westat's Matched Current Population Survey-Social Security (CPS) Administration File. In our analysis, we include 185 treated units and 689 control units from the NSW, as well as 429 comparison units from the CPS-3, a subset of the CPS data (Firpo, 2007; LaLonde, 1986). Seven baseline confounding covariates are used for this application: age, education, race, Hispanic, married, having no college degree, and real earnings in 1975. The outcome of interest is the real earnings in 1978.

Because the outcome distributions are highly skewed (see Figure S3), the average treatment effect may not provide a comprehensive evaluation of the job training program. Therefore, we estimate the ATT and QTTs. To gain the robustness and reliability of the results, we posit two propensity score models and two prognostic score models. Following Dehejia and Wahba (1999), one propensity score model is a logistic regression model with all first-order terms of the covariates and second-order terms of numerical variables, and one prognostic score model is a linear regression of the earnings with the same predictors as in the propensity score models for the control group. Given the popularity of probit models, we consider the second propensity score model to be a probit regression model with the same predictors in the first propensity score model. Given the skewness of the outcome distribution, we consider the second prognostic score to be a linear regression model for the logarithm of the real earnings in 1978.

Matching admits a transparent assessment of covariate balance before and after matching. Table 4 presents the means of all covariates by treatment group and the standardized difference in means before and after DSM. The standardized difference is calculated as the difference of the group means divided by the overall standard error in the original sample. DSM makes standardized differences fall between −0.05 and 0.05 for all covariates, reducing the differences of the observed covariates in the treated and the control.

Table 5 shows the estimated ATTs and QTTs at the 0.1, 0.25, 0.3, 0.5, 0.75, and 0.9 quantiles, and 95% Wald confidence intervals from the four matching methods, as well as ATE and QTE estimated by the naive method. All four matching estimators show that the job training program does not have a significant effect on the average earning for the treated. Figure 2 shows the QTT plot estimated by DSM algorithm. A closer inspection of the QTT plot reveals that the effect is, in fact, significant around the percentile of 0.3, which suggests that the program is beneficial for the lower middle class.

TABLE 4 Covariate balance check before and after double score matching

		age	educ	black	hisp	married	nodegr	re75
Before Matching	Treatment group mean	24.63	10.38	0.80	0.09	0.17	0.73	3066
	Control group mean	26.25	10.21	0.50	0.13	0.34	0.70	2745
	Stand diff. in means	−0.19	0.08	0.61	−0.10	−0.37	0.06	0.07
After Matching	Treatment group mean	24.63	10.38	0.80	0.09	0.17	0.73	3066
	Control group mean	25.01	10.31	0.69	0.12	0.21	0.73	2876
	Stand diff. in means	−0.04	−0.03	0.22	−0.08	−0.10	0.00	0.04

TABLE 5 Estimated ATT and QTTs at the 0.1, 0.25, 0.3, 0.5, 0.75 and 0.9 quantiles, and 95% Wald confidence intervals

Estimand	m.x	psm	pgm	dsm	Naive (ATE and QTE)
ATT	372 (−746,1489)	918 (−222,2058)	−150 (−1215,914)	641 (−429,1711)	−65 (−957,827)
0.1-QTT	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
0.25-QTT	549 (−90,1189)	549 (−55,1154)	549 (−104,1202)	549 (−62,1160)	549 (−55,1153)
0.3-QTT	935 (57,1813)	1064 (268,1860)	1039 (96,1982)	1036 (211,1861)	604 (−641,1170)
0.5-QTT	524 (−1606,2654)	889 (−741,2519)	1296 (−252,2844)	680 (−766,2126)	69 (−1093,1230)
0.75-QTT	−391 (−2752,1970)	617 (−1451,2686)	737 (−2070,3544)	558 (−1297,2414)	−195 (−1441,1578)
0.9-QTT	−963 (−3482,1556)	897 (−1750,3543)	−936 (−4770,1795)	−718 (−3711,2275)	−2326 (−2021,2158)

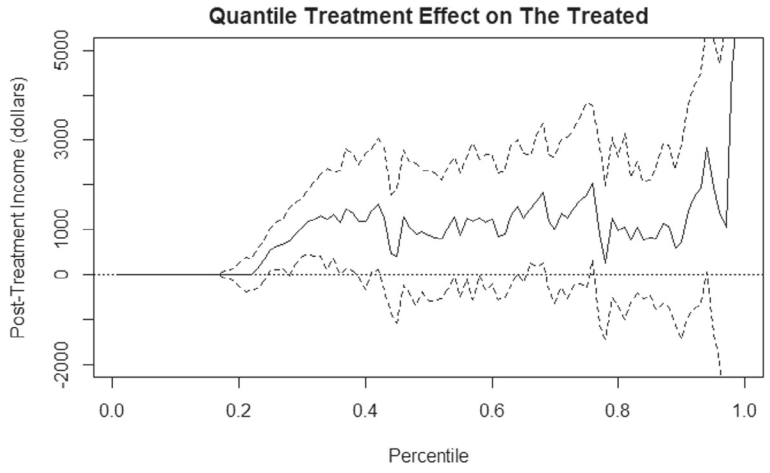

FIGURE 2 Quantile effect plot on the treated from the double score matching algorithm

9 | DISCUSSION

We have developed multiply robust matching estimators for general causal estimands. This framework offers a new "metric" to summarize the differential roles of different covariates and also serves as a powerful dimensional reduction tool in high-dimensional confounding. The improved robustness comes from multiple model specifications for the propensity score and prognostic score. The proposed DSM thus provides multiple protections to model misspecification and therefore is an attractive alternative to existing weighting estimators.

Several issues are worth discussing. As with PSM, although the matching variables are well balanced, individual covariates may not for a given application. In this case, if the researchers know important confounders based on substantive knowledge, they can augment the double score by adding those confounders to ensure balance for these confounders; however, adding too many variables will results in potential bias as demonstrated in our simulation. Alternatively, one can use regression adjustment for the matched sample Abadie and Spiess (2016), which can remove remaining confounding biases. We focus on a binary treatment. Yang et al. (2016) have developed the generalized PSM for estimating the treatment effects for more than two treatments. It is of interest to extend our DSM algorithm to more than two treatment comparisons. The current DSM framework focuses on continuous and binary outcomes, and it would be an important task to extend DSM to handle survival outcomes (Tang et al., 2019) and clustered data (Yang, 2018) and estimate the heterogeneous treatment effects Huang and Yang (2022). It is crucial to highlight that as for all existing matching methods, the DSM method cannot account for unmeasured confounding. Following Rosenbaum and Rubin (1983a) and Yang and Lok (2017), we will develop sensitivity analyses to no unmeasured confounding in the matching framework.

ACKNOWLEDGMENTS

We are grateful to Alberto Abadie for providing comments. Shu Yang is partially supported by the National Science Foundation grant DMS 1811245, National Cancer Institute grant P01 CA142538, National Institute on Aging grant 1R01AG066883, and National Institute of Environmental Health Science grant 1R01ES031651.

ORCID

Shu Yang https://orcid.org/0000-0001-7703-707X

REFERENCES

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

Abadie, A., & Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, *76*, 1537–1557.

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

Abadie, A., & Imbens, G. W. (2012). A martingale representation for matching estimators. *Journal of the American Statistical Association*, *107*, 833–843.

Abadie, A., & Imbens, G. W. (2016). Matching on the estimated propensity score. *Econometrica*, *84*, 781–807.

Abadie, A., & Spiess, J. (2016). *Robust post-matching inference* [Unpublished paper]. MIT and Harvard University. https://editorialexpress. com/cgi-bin/conference/down load

Andreou, E., & Werker, B. J. (2012). An alternative asymptotic analysis of residual-based statistics. *The Review of Economics and Statistics*, *94*, 88–99.

Antonelli, J., Cefalu, M., Palmer, N., & Agniel, D. (2018). Doubly robust matching estimators for high dimensional confounding adjustment. *Biometrics*, *74*, 1171–1179.

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, *61*, 962–973.

Bickel, P. J., Klaassen, C., Ritov, Y., & Wellner, J. (1993). *Efficient and adaptive inference in semiparametric models*. Johns Hopkins University Press.

Billingsley, P. (1995). *Probability and measure* (3rd ed.). New York, NY: Wiley.

Chen, J., & Shao, J. (2000). Nearest neighbor imputation for survey data. *The Journal of Official Statistics*, *16*, 113–131.

Chen, J., & Shao, J. (2001). Jackknife variance estimation for nearest-neighbor imputation. *Journal of the American Statistical Association*, *96*, 260–269.

Chen, S., & Haziza, D. (2017a). Multiply robust imputation procedures for the treatment of item nonresponse in surveys. *Biometrika*, *104*, 439–453.

Chen, S., & Haziza, D. (2017b). Multiply robust nonparametric multiple imputation for the treatment of missing data. *Statistica Sinica*, *29*, 2035–2053.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, *6*, 5549–5632.

Chernozhukov, V., Fernández-Val, I., Hahn, J., & Newey, W. (2013). Average and quantile effects in nonseparable panel models. *Econometrica*, *81*, 535–580.

Chiang, C.-T., & Huang, M.-Y. (2012). New estimation and inference procedures for a single-index conditional distribution model. *Journal of Multivariate Analysis*, *111*, 271–285.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Re-evaluating the evaluation of training programs. *Journal of the American Statistical Association*, *94*, 1053–1062.

Dehejia, R. H., & Wahba, S. (2002). Propensity score matching methods for non-experimental causal studies. *The Review of Economics and Statistics*, *84*, 151–161.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*, 1–26.

Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, *75*, 259–276.

Ford, B. L. (1983). An overview of hot-deck procedures. *Incomplete Data in Sample Surveys*, *2*(Part IV), 185–207.

Foresi, S., & Peracchi, F. (1995). The conditional distribution of excess returns: An empirical analysis. *Journal of the American Statistical Association*, *90*, 451–466.

Francisco, C. A., & Fuller, W. A. (1991). Quantile estimation with a complex survey design. *Annals of Statistics*, *19*, 454–469.

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

Guo, S., & Fraser, M. W. (2014). *Propensity score analysis: Statistical methods and applications* (Vol. *11*). SAGE.

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

Han, P. (2014). Multiply robust estimation in regression analysis with missing data. *Journal of the American Statistical Association*, *109*(507), 1159–1173.

Han, P., Kong, L., Zhao, J., & Zhou, X. (2019). A general framework for quantile estimation with incomplete data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *81*, 305–333.

Han, P., & Wang, L. (2013). Estimation with missing data: Beyond double robustness. *Biometrika*, *100*, 417–430.

Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika*, *95*, 481–488.

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, *64*, 605–654.

Hernán, M. A., Brumback, B., & Robins, J. M. (2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association*, *96*, 440–448.

Huang, M.-Y., & Yang, S. (2022). Robust inference of conditional average treatment effects using dimension reduction. *Statistica Sinica*, *32*, 547–567.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, *86*, 4–29.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

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

King, G., & Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political Analysis*, *27*, 435–454.

Kumamaru, H., Schneeweiss, S., Glynn, R. J., Setoguchi, S., & Gagne, J. J. (2016). Dimension reduction and shrinkage methods for high dimensional disease risk scores in historical data. *Emerging Themes in Epidemiology*, *13*, 5.

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

Leacy, F. P., & Stuart, E. A. (2014). On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: A simulation study. *Statistics in Medicine*, *33*, 3488–3508.

Li, W., Yang, S., & Han, P. (2020). Robust estimation for moment condition models with data missing not at random. *Journal of Statistical Planning and Inference*, *207*, 246–254.

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Wiley.

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

Naik, C., McCoy, E. J., & Graham, D. J. (2016). Multiply robust dose-response estimation for multivalued causal inference problems, *arXiv preprint arXiv:1611.02433*.

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

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, *84*, 1024–1032.

Rosenbaum, P. R., & Rubin, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *45*, 212–218.

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

Rubin, D. B. (2006). *Matched sampling for causal effects*. Cambridge University Press.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*, 1–21.

Tang, S., Yang, S., Wang, T., Cui, Z., Li, L., & Faries, D. E. (2019). Causal inference of hazard ratio based on propensity score matching, *arXiv preprint arXiv:1911.12430*.

van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge University Press.

Wang, L. (2019). Multiple robustness estimation in causal inference. *Communications in Statistics-Theory and Methods*, *48*, 5701–5718.

Wolter, K. (2007). *Introduction to variance estimation* (2nd ed.). Springer.

Wyss, R., Ellis, A. R., Brookhart, M. A., Jonsson Funk, M., Girman, C. J., Simpson, R. J., Jr., & Stürmer, T. (2015). Matching on the disease risk score in comparative effectiveness research of new treatments. *Pharmacoepidemiology and Drug Safety*, *24*, 951–961.

Yang, S. (2018). Propensity score weighting for causal inference with clustered data. *Journal of Causal Inference*, *6*, 20170027.

Yang, S., & Ding, P. (2018). Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika*, *105*, 487–493.

Yang, S., Imbens, G. W., Cui, Z., Faries, D. E., & Kadziola, Z. (2016). Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics*, *72*, 1055–1065.

Yang, S., & Kim, J. K. (2018). Nearest neighbor imputation for general parameter estimation in survey sampling. *Advances in Econometrics*, *39*, 211–236.

Yang, S., & Kim, J. K. (2020). Asymptotic theory and inference of predictive mean matching imputation using a superpopulation model framework. *The Scandinavian Journal of Statistics*, *47*, 839–861. https://doi.org/10.1111/sjos.12429

Yang, S., & Lok, J. J. (2017). Sensitivity analysis for unmeasured confounding in coarse structural nested mean models. *Statistica Sinica*, *28*, 1703–1723.

Zhang, Z., Chen, Z., Troendle, J. F., & Zhang, J. (2012). Causal inference on quantiles with an obstetric application. *Biometrics*, *68*, 697–706.

Zhao, Z. (2004). Using matching to estimate treatment effects: Data requirements, matching metrics, and Monte Carlo evidence. *The Review of Economics and Statistics*, *86*, 91–107.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Yang, S., & Zhang, Y. (2022). Multiply robust matching estimators of average and quantile treatment effects. *Scandinavian Journal of Statistics*, 1–31. https://doi.org/10.1111/sjos.12585

APPENDIX

Sections A.1 and A.2 present the proofs of Theorems 1 and 2 for ATE estimation. Parallel proofs of Theorems 3 and 4 for QTE estimation are presented in Data S1.

A.1 Proof of Theorem 1

Before presenting the asymptotic properties of $\hat{\tau}_{\text{dsm}}(\theta^*)$, we require technical conditions. For simplicity, let $S_a = S_a(\theta_a^*)$ and let $f_1(S_a)$ and $f_0(S_a)$ be the conditional density of S_a given $A = 1$ and $A = 0$, respectively.

Assumption A1. For $a = 0, 1$, (i) the matching variable S_a has a compact and convex support $\mathcal{S}$, with a continuous density bounded and bounded away from zero: there exist constants C_{1L} and C_{1U} such that $C_{1L} \le f_1(S_a)/f_0(S_a) \le C_{1U}$ for any $S_a \in \mathcal{S}$; (ii) $\mu_a(S_a)$ and $\sigma_a^2(S_a)$ satisfy Lipschitz continuity conditions: there exists a constant C_2 such that $|\mu_a(S_{a,i}) - \mu_a(S_{aj})| < C_2||S_{a,i} - S_{aj}||$ for any $S_{a,i}$ and S_{aj}, and similarly for $\sigma_a^2(S_a)$; (iii) there exists $\delta > 0$ such that $\mathbb{E}\left\{|Y(a)|^{2+\delta}|S_a\right\}$ is uniformly bounded for any $S_a \in \mathcal{S}$; and (iv) $\hat{B}_n = B_n + o_{\text{P}}(1)$.

Assumption A1 has been considered by Abadie and Imbens (2006) and Abadie and Imbens (2016) for matching estimators based on the covariates and the propensity score. Assumption A1 (i) is a convenient regularity condition. Assumption A1 (ii) imposes smoothness conditions for the outcome mean function $\mu_a(S_a)$ and the variance function $\sigma_a^2(S_a)$. Assumption A1 (iii) is a moment condition for establishing the central limit theorem. Assumption A1 (iv) requires regularity conditions on $\mu_a(S_a)$ $(a = 0, 1)$ and the nonparametric estimators; detailed discussions have appeared in Abadie and Imbens (2011, 2012) and Otsu and Rai (2017).

Under Assumption A1, similar to the proof of theorem 2 in Abadie and Imbens (2011), $\hat{\tau}_{\text{dsm}}(\theta^*)$ has the following asymptotic linear form:

$$n^{1/2}\{\hat{\tau}_{\text{dsm}}(\theta^*) - \tau\} = n^{-1/2}\sum_{i=1}^{n}\left\{\mu_1(S_{1,i}) - \mu_0(S_{0,i}) - \tau\right\} \tag{A1}$$

$$+ n^{-1/2}\sum_{i=1}^{n} A_i\left(1 + M^{-1}K_{S_1,i}\right)\left\{Y_i - \mu_1(S_{1,i})\right\} \tag{A2}$$

$$- n^{-1/2}\sum_{i=1}^{n}(1 - A_i)\left(1 + M^{-1}K_{S_0,i}\right)\left\{Y_i - \mu_0(S_{0,i})\right\} + o_{\text{P}}(1). \tag{A3}$$

If any model of the propensity score or prognostic score is correctly specified, by Lemma 2, we have $\mathbb{E}\{\mu_1(S_{1,i}) - \mu_0(S_{0,i})\} = \tau$ and therefore the asymptotic bias of $n^{1/2}\{\hat{\tau}_{\text{dsm}}(\theta^*) - \tau\}$ is zero.

Let the three terms in (A1) be

$$T_{1n} = n^{-1/2}\sum_{i=1}^{n}\left\{\mu_1(S_{1,i}) - \mu_0(S_{0,i}) - \tau\right\},$$

$$T_{2n} = n^{-1/2}\sum_{i=1}^{n} A_i\left(1 + M^{-1}K_{S_1,i}\right)\left\{Y_i - \mu_1(S_{1,i})\right\},$$

$$T_{3n} = -n^{-1/2}\sum_{i=1}^{n}(1 - A_i)\left(1 + M^{-1}K_{S_0,i}\right)\left\{Y_i - \mu_0(S_{0,i})\right\}.$$

We show the covariances of the three terms are zero:

$$\begin{aligned}
\text{cov}(T_{1n}, T_{2n}) &= n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\text{cov}\left[\mu_1(S_{1,i}) - \mu_0(S_{0,i}) - \tau, A_j\left(1 + M^{-1}K_{S_1,j}\right)\left\{Y_j - \mu_1(S_{1,j})\right\}\right] \\
&= n^{-1}\sum_{i=1}^{n}\text{cov}\left[\mu_1(S_{1,i}) - \mu_0(S_{0,i}) - \tau, A_i\left(1 + M^{-1}K_{S_1,i}\right)\left\{Y_i - \mu_1(S_{1,i})\right\}\right] \\
&= n^{-1}\sum_{i=1}^{n}\text{cov}\left(\mathbb{E}\left\{\mu_1(S_{1,i}) - \mu_0(S_{0,i}) - \tau \mid S_{1,i}, S_{0,i}\right\},\right. \\
&\quad \left.\mathbb{E}\left[A_i\left(1 + M^{-1}K_{S_1,i}\right)\left\{Y_i - \mu_1(S_{1,i})\right\} \mid S_{1,i}, S_{0,i}\right]\right) \\
&\quad + n^{-1}\sum_{i=1}^{n}\mathbb{E}\left(\text{cov}\left\{\mu_1(S_{1,i}) - \mu_0(S_{0,i}) - \tau \mid S_{1,i}, S_{0,i},\right.\right. \\
&\quad \left.\left.A_i\left(1 + M^{-1}K_{S_1,i}\right)\left\{Y_i - \mu_1(S_{1,i})\right\} \mid S_{1,i}, S_{0,i}\right\}\right) \\
&= 0,
\end{aligned}$$

similarly, $\mathrm{cov}(T_{1n}, T_{3n}) = 0$, and by construction, $\mathrm{cov}(T_{2n}, T_{3n}) = 0$. Thus, the asymptotic variance of $n^{1/2}\{\hat{\tau}_{\mathrm{dsm}}(\theta^*) - \tau\}$ is

$$\mathbb{V}\left[n^{-1/2}\sum_{i=1}^{n}\left\{\mu_1(S_{1,i}) - \mu_0(S_{0,i}) - \tau\right\}\right] + \mathbb{V}\left[n^{-1/2}\sum_{i=1}^{n}A_i\left(1 + M^{-1}K_{S_1,i}\right)\left\{Y_i - \mu_1(S_{1,i})\right\}\right] + \mathbb{V}\left[n^{-1/2}\sum_{i=1}^{n}(1-A_i)\left(1 + M^{-1}K_{S_0,i}\right)\left\{Y_i - \mu_0(S_{0,i})\right\}\right].$$

The first term becomes $\mathbb{E}\left[\{\mu_1(S_1) - \mu_0(S_0) - \tau\}^2\right]$. Following Abadie and Imbens (2006), the second and third term, as $n \to \infty$, becomes

$$\begin{aligned}
&\mathrm{plim}_{n\to\infty}\left[n^{-1}\sum_{i=1}^{n}A_i\left(1 + M^{-1}K_{S_1,i}\right)^2\mathbb{V}(Y_i \mid S_{1,i})\right] \\
&+ \mathrm{plim}_{n\to\infty}\left[n^{-1}\sum_{i=1}^{n}(1-A_i)\left(1 + M^{-1}K_{S_0,i}\right)^2\mathbb{V}(Y_i \mid S_{0,i})\right] \\
&= \mathbb{E}\left(\sigma_1^2(S_1)\left[\frac{1}{e(S_1)} + \frac{1}{2M}\left\{\frac{1}{e(S_1)} - e(S_1)\right\}\right]\right) \\
&\quad + \mathbb{E}\left(\sigma_0^2(S_0)\left[\frac{1}{1-e(S_0)} + \frac{1}{2M}\left\{\frac{1}{1-e(S_0)} - 1 + e(S_0)\right\}\right]\right).
\end{aligned}$$

This completes the proof of Theorem 1.

A.2 Proof of Theorem 2

We follow the technique in Andreou and Werker (2012) and Abadie and Imbens (2016). In Abadie and Imbens (2016), the PSM estimators rely on the nuisance parameter estimator under a correct specification of the propensity score model. In our setting, the nuisance parameters include both parameters in the propensity score model and the prognostic score model, and require only one of the models to be correctly specified. Without loss of generality, we assume one working model $e(X;\alpha)$ for the propensity score and one working model $\Psi(X;\beta) = \{\Psi_0(X;\beta_0), \Psi_1(X;\beta_1)\}$ for the prognostic score. The proof for the case with more than two working models for each score is similar at the expense of heavier notation. Let $\mathbb{P}$ be the distribution of $\{(A_i, X_i, Y_i) : i = 1, \ldots, n\}$. Consider $\mathbb{P} = \mathbb{P}^{\theta^*}$ to be indexed by $\theta^* = (\alpha^{*\mathrm{T}}, \beta_0^{*\mathrm{T}}, \beta_1^{*\mathrm{T}})^{\mathrm{T}}$, which satisfies

$$\mathbb{E}\{U(A,X,Y;\theta^*)\} = \mathbb{E}\left\{\begin{pmatrix} U_1(A,X,Y;\alpha^*) \\ U_2(A,X,Y;\beta_0^*) \\ U_3(A,X,Y;\beta_1^*)\end{pmatrix}\right\} = 0. \tag{A4}$$

We invoke standard regularity conditions on Z-estimation (van der Vaart, 2000) as follows.

Assumption A2. (i) Under $\mathbb{P}^{\theta^*}$, $\mathcal{U}_n(\theta^*) \to \mathcal{N}(0, \Sigma_U)$ in distribution, as $n \to \infty$, where $\Sigma_U = \mathbb{E}\{U(A,X,Y;\theta^*)\,U(A,X,Y;\theta^*)^{\mathrm{T}}\}$; (ii) $\Gamma_\theta = \mathbb{E}\{\partial U(A,X,Y;\theta)/\partial\theta^{\mathrm{T}}\}$ is nonsingular around θ^*; and (iii) for any vector of constant h, $\exp\{n^{1/2}h^{\mathrm{T}}\Gamma_{\theta^*}\Sigma_U^{-1}\mathcal{U}_n(\theta^*)\}$ is uniformly integrable.

Under Assumption A2,

$$n^{1/2}(\hat{\theta} - \theta^*) = -\Gamma_{\theta^*}^{-1}\mathcal{U}_n(\theta^*) + o_{\mathrm{P}}(1) \to \mathcal{N}(0, \Sigma_{\theta^*}), \tag{A5}$$

in distribution, as $n \to \infty$, where $\Sigma_{\theta^*} = \Gamma_{\theta^*}^{-1}\Sigma_U(\Gamma_{\theta^*}^{-1})^{\mathrm{T}}$.

To derive the large sample distribution of $\hat{\tau}_{\mathrm{dsm}}(\hat{\theta})$, following Abadie and Imbens (2016), we impose the following regularity conditions.

Assumption A3. There exists a neighborhood of θ^*, such that for any θ in this region, the following conditions hold: for $a = 0, 1$, (i) the matching variable $S_a(\theta)$ has a compact and convex support $\mathcal{S}$, with a continuous density bounded and bounded away from zero; (ii) $\mu_a\{S_a(\theta)\}$ and $\sigma_a^2\{S_a(\theta)\}$ satisfy the Lipschitz continuity condition; and (iii) there exists $\delta > 0$ such that $\mathbb{E}\left\{|Y(a)|^{2+\delta}|S_a(\theta)\right\}$ is uniformly bounded for any $S_a(\theta) \in \mathcal{S}$.

Following Andreou and Werker (2012), because we consider a semiparametric model for θ^*, to invoke the Le Cam's lemma, we specify an auxiliary parametric model $\mathbb{P}^{\theta_n}$ defined locally though θ^*, $\theta_n = \theta^* + n^{-1/2}h$, with a density

$$\frac{\exp\left\{n^{1/2}(\theta_n - \theta^*)^{\mathrm{T}}\Gamma_{\theta^*}\Sigma_U^{-1}\mathcal{U}_n(\theta^*) - 2^{-1}n(\theta_n - \theta^*)^{\mathrm{T}}\Sigma_{\theta^*}^{-1}(\theta_n - \theta^*)\right\}}{\mathbb{E}\left[\exp\left\{n^{1/2}(\theta_n - \theta^*)^{\mathrm{T}}\Gamma_{\theta^*}\Sigma_U^{-1}\mathcal{U}_n(\theta^*) - 2^{-1}n(\theta_n - \theta^*)^{\mathrm{T}}\Sigma_{\theta^*}^{-1}(\theta_n - \theta^*)\right\}\right]}. \tag{A6}$$

By Assumption A2 (iii), $\exp\{n^{1/2}(\theta_n - \theta^*)^{\mathrm{T}}\Gamma_{\theta^*}\Sigma_U^{-1}\mathcal{U}_n(\theta^*)\}$ is uniformly integrable, and thus model (A6) is uniformly locally asymptotically normal. Because under $\mathbb{P}^{\theta^*}$, $\mathcal{U}_n(\theta^*) \to \mathcal{N}(0, \Sigma_U)$ in distribution, the normalizing constant in the denominator converges to one as $n \to \infty$. The Fisher information under the parametric model (A6) is $n\Sigma_{\theta^*}^{-1}$. Therefore, $\hat{\theta}$ is efficient under model (A6).

Now consider (A_i, X_i, Y_i), for $i = 1, \ldots, n$, with the local shift $\mathbb{P}^{\theta_n}$ (Bickel et al., 1993). Under model (A6), the likelihood ratio under $\mathbb{P}^{\theta_n}$ is

$$\begin{aligned}\log(\mathrm{d}\mathbb{P}^{\theta^*}/\mathrm{d}\mathbb{P}^{\theta_n}) &= -h^{\mathrm{T}}\Gamma_{\theta^*}\Sigma_U^{-1}\mathcal{U}_n(\theta^*) + \frac{1}{2}h^{\mathrm{T}}\Sigma_{\theta^*}^{-1}h + o_{\mathrm{P}}(1)\\ &= -h^{\mathrm{T}}\Gamma_{\theta^*}\Sigma_U^{-1}\mathcal{U}_n(\theta_n) - \frac{1}{2}h^{\mathrm{T}}\Sigma_{\theta^*}^{-1}h + o_{\mathrm{P}}(1),\end{aligned} \tag{A7}$$

where the second equality follows by the Taylor expansion of $\mathcal{U}_n(\theta^*)$ at θ_n. Moreover, under $\mathbb{P}^{\theta_n}$: $\mathcal{U}_n(\theta_n) \to \mathcal{N}(0, \Sigma_U)$ in distribution, as $n \to \infty$, and

$$n^{1/2}(\hat{\theta} - \theta_n) = \Gamma_{\theta^*}^{-1}\mathcal{U}_n(\theta_n) + o_{\mathrm{P}}(1). \tag{A8}$$

We also assume the following regularity condition.

Assumption A4. For all bounded continuous functions $h(A, X, Y)$, the conditional expectation $\mathbb{E}_{\theta_n}\{h(A, X, Y)\}$ converges in distribution to $\mathbb{E}\{h(A, X, Y)\}$, where $\mathbb{E}_{\theta_n}(\cdot)$ is the expectation taken with respect to P^{θ_n}.

We derive the results in Theorem 2 in two steps.

In the first step, under $\mathbb{P}^{\theta_n}$, we write $\tau = \tau(\theta_n)$ to reflect its dependence on θ_n; to be specific, we have

$$\tau(\theta_n) = \mathbb{E}\left[\mu_1\{S_1(\theta_n)\} - \mu_0\{S_0(\theta_n)\}\right].$$

We derive that under $\mathbb{P}^{\theta_n}$,

$$\begin{pmatrix} n^{1/2}\{\hat{\tau}_{\mathrm{dsm}}(\theta_n) - \tau(\theta_n)\} \\ n^{1/2}(\hat{\theta} - \theta_n) \\ \log(\mathrm{d}\mathbb{P}^{\theta^*}/\mathrm{d}\mathbb{P}^{\theta_n}) \end{pmatrix} \to \mathcal{N}\left\{\begin{pmatrix} 0 \\ 0 \\ -\frac{1}{2}h^{\mathrm{T}}\Sigma_{\theta^*}^{-1}h \end{pmatrix}, \begin{pmatrix} V_\tau & \gamma_1^{\mathrm{T}}\Gamma_{\theta^*}^{-1} & -\gamma_1^{\mathrm{T}}\Sigma_U^{-1}\Gamma_{\theta^*}h \\ \Gamma_{\theta^*}^{-1}\gamma_1 & \Sigma_{\theta^*} & -h \\ -h^{\mathrm{T}}\Gamma_{\theta^*}\Sigma_U^{-1}\gamma_1 & -h^{\mathrm{T}} & h^{\mathrm{T}}\Sigma_{\theta^*}^{-1}h \end{pmatrix}\right\}, \tag{A9}$$

in distribution, as $n \to \infty$. We then express $\tau(\theta_n) = \tau(\theta^*) + \gamma_2^{\mathrm{T}}(n^{-1/2}h) + o(n^{-1/2})$, where

$$\gamma_2 = \left.\frac{\partial \tau(\theta)}{\partial \theta}\right|_{\theta=\theta^*} = \mathbb{E}\left[\left.\frac{\partial \mu_1\{S_1(\theta)\} - \mu_0\{S_0(\theta)\}}{\partial \theta}\right|_{\theta=\theta^*}\right]. \tag{A10}$$

By Le Cam's third lemma, under $\mathbb{P}^{\theta^*}$,

$$\begin{pmatrix} n^{1/2}\{\hat{\tau}_{\mathrm{dsm}}(\theta_n) - \tau\} \\ n^{1/2}(\hat{\theta} - \theta_n) \end{pmatrix} \to \mathcal{N}\left\{\begin{pmatrix} -\gamma_1^{\mathrm{T}}\Sigma_U^{-1}\Gamma_{\theta^*}h - \gamma_2^{\mathrm{T}}h \\ -h \end{pmatrix}, \begin{pmatrix} V_\tau & \gamma_1^{\mathrm{T}}\Gamma_{\theta^*}^{-1} \\ \Gamma_{\theta^*}^{-1}\gamma_1 & \Sigma_{\theta^*} \end{pmatrix}\right\},$$

in distribution, as $n \to \infty$. Replacing θ_n by $\theta^* + n^{-1/2}h$ yields that under $\mathbb{P}^{\theta^*}$,

$$\begin{pmatrix} n^{1/2}\{\hat{\tau}_{\mathrm{dsm}}(\theta^* + n^{-1/2}h) - \tau\} \\ n^{1/2}(\hat{\theta} - \theta^*) \end{pmatrix} \to \mathcal{N}\left\{\begin{pmatrix} -\gamma_1^{\mathrm{T}}\Sigma_U^{-1}\Gamma_{\theta^*}h - \gamma_2^{\mathrm{T}}h \\ 0 \end{pmatrix}, \begin{pmatrix} V_\tau & \gamma_1^{\mathrm{T}}\Gamma_{\theta^*}^{-1} \\ \Gamma_{\theta^*}^{-1}\gamma_1 & \Sigma_{\theta^*} \end{pmatrix}\right\}, \tag{A11}$$

in distribution, as $n \to \infty$.

In the second step, we provide a heuristic derivation for (A11) to obtain the approximate distribution (8). If the normal distribution were exact, then

$$n^{1/2}\{\hat{\tau}_{\mathrm{dsm}}(\theta^* + n^{-1/2}h) - \tau\} | n^{1/2}(\hat{\theta} - \theta^*) = h \sim \mathcal{N}\left(-\gamma_2^{\mathrm{T}}h, V_\tau - \gamma_1^{\mathrm{T}}\Sigma_U^{-1}\gamma_1\right). \tag{A12}$$

Given that $n^{1/2}(\hat{\theta} - \theta^*) = h$, we have $\theta^* + n^{-1/2}h = \hat{\theta}$, and hence $\hat{\tau}_{\mathrm{dsm}}(\theta^* + n^{-1/2}h) = \hat{\tau}_{\mathrm{dsm}}(\hat{\theta})$. Marginalizing (A12) over the asymptotic distribution of $n^{1/2}(\hat{\theta} - \theta^*)$, we derive (8). The formal technique to derive (8) can be find in Andreou and Werker (2012) and Abadie and Imbens (2016). To avoid repetition, we omit this step.

In the following, we provide the proof to (A9) in the first step of the proof. Asymptotic normality of $n^{1/2}\{\hat{\tau}_{\mathrm{dsm}}(\theta_n) - \tau(\theta_n)\}$ under $\mathbb{P}^{\theta_n}$ follows from Theorem 1 and the uniform local asymptotic normality of model (A6). Asymptotic joint normality of $\log(\mathrm{d}\mathbb{P}^{\theta^*}/\mathrm{d}\mathbb{P}^{\theta_n})$ and $n^{1/2}(\hat{\theta} - \theta_n)$ follows from (A7) and (A8). Also, $n^{1/2}\{\hat{\tau}_{\mathrm{dsm}}(\theta_n) - \tau(\theta_n)\} = D_n(\theta_n) + o_{\mathrm{P}}(1)$, where

$$\begin{aligned} D_n(\theta_n) = {} & n^{-1/2}\sum_{i=1}^{n}\left[\mu_1\{S_{1,i}(\theta_n)\} - \mu_0\{S_{0,i}(\theta_n)\} - \tau(\theta_n)\right] \\ & + n^{-1/2}\sum_{i=1}^{n} A_i\left\{1 + M^{-1}K_{S_1(\theta_n),i}\right\}\left[Y_i - \mu_1\{S_{1,i}(\theta_n)\}\right] \\ & - n^{-1/2}\sum_{i=1}^{n}(1 - A_i)\left\{1 + M^{-1}K_{S_0(\theta_n),i}\right\}\left[Y_i - \mu_0\{S_{0,i}(\theta_n)\}\right] + o_{\mathrm{P}}(1). \end{aligned}$$

Therefore, the remaining is to show that, under $\mathbb{P}^{\theta_n}$:

$$\begin{pmatrix} D_n(\theta_n) \\ \mathcal{U}_n(\theta_n) \end{pmatrix} \to \mathcal{N}\left\{\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} V_\tau & \gamma_1^{\mathrm{T}} \\ \gamma_1 & \Sigma_U \end{pmatrix}\right\}, \tag{A13}$$

in distribution, as $n \to \infty$. To prove (A13), consider the linear combination

$$
\begin{aligned}
T_n &= c_0 D_n(\theta_n) + c^{\mathrm{T}} \mathcal{U}_n(\theta_n) \\
&= c_0 n^{-1/2} \sum_{i=1}^{n} \left[\mu_1\{S_{1,i}(\theta_n)\} - \mu_0\{S_{0,i}(\theta_n)\} - \tau(\theta_n) \right] \\
&\quad + c_0 n^{-1/2} \sum_{i=1}^{n} (2A_i - 1) \left\{ 1 + M^{-1} K_{S_{A_i}(\theta_n),i} \right\} \left[Y_i - \mu_{A_i}\{S_{A_i,i}(\theta_n)\} \right] \\
&\quad + c_1^{\mathrm{T}} n^{-1/2} \sum_{i=1}^{n} \left[\frac{\partial e(X_i;\alpha_n)}{\partial \alpha} \frac{A_i - e(X_i;\alpha_n)}{e(X_i;\alpha_n)\{1 - e(X_i;\alpha_n)\}} \right] \\
&\quad + c_2^{\mathrm{T}} n^{-1/2} \sum_{i=1}^{n} \left[(1 - A_i) \frac{\partial \mu_0(X_i;\beta_{0,n})}{\partial \beta_0} \{ Y_i - \mu_0(X_i;\beta_{0,n}) \} \right] \\
&\quad + c_3^{\mathrm{T}} n^{-1/2} \sum_{i=1}^{n} \left[A_i \frac{\partial \mu_1(X_i;\beta_{1,n})}{\partial \beta_1} \{ Y_i - \mu_1(X_i;\beta_{1,n}) \} \right] + o_{\mathrm{P}}(1),
\end{aligned}
$$

where $c = (c_1^{\mathrm{T}}, c_2^{\mathrm{T}}, c_3^{\mathrm{T}})^{\mathrm{T}}$. We analyze T_n using the martingale theory. We rewrite $T_n = \sum_{k=1}^{2n} \xi_{n,k}$, where

$$
\xi_{n,k} = \begin{cases} \sum_{j=1}^{8} \xi_{n,k}^{(j)}, & 1 \le k \le n, \\ \sum_{j=9}^{11} \xi_{n,k}^{(j)}, & n+1 \le k \le 2n, \end{cases}
$$

$$
\begin{aligned}
\xi_{n,k}^{(1)} &= c_0 n^{-1/2} \left[\mu_1\{S_{1,k}(\theta_n)\} - \mu_0\{S_{0,k}(\theta_n)\} - \tau(\theta_n) \right], \\
\xi_{n,k}^{(2)} &= c_0 n^{-1/2} (2A_k - 1) \left\{ 1 + M^{-1} K_{S_{A_k}(\theta_n),k} \right\} \left[\mu_{A_k}(X_k) - \mu_{A_k}\{S_{A_k,k}(\theta_n)\} \right] \\
\xi_{n,k}^{(3)} &= c_1^{\mathrm{T}} n^{-1/2} \left[\frac{\partial e(X_k;\alpha_n)}{\partial \alpha} \frac{e(X_k) - e(X_k;\alpha_n)}{e(X_k;\alpha_n)\{1 - e(X_k;\alpha_n)\}} \right], \\
\xi_{n,k}^{(4)} &= c_2^{\mathrm{T}} n^{-1/2} \{1 - e(X_k)\} \frac{\partial \mu_0(X_k;\beta_{0,n})}{\partial \beta_0} \{ \mu_0(X_k) - \mu_0(X_k;\beta_{0,n}) \}, \\
\xi_{n,k}^{(5)} &= c_3^{\mathrm{T}} n^{-1/2} e(X_k) \frac{\partial \mu_1(X_k;\beta_{1,n})}{\partial \beta_1} \{ \mu_1(X_k) - \mu_1(X_k;\beta_{1,n}) \}, \\
\xi_{n,k}^{(6)} &= c_1^{\mathrm{T}} n^{-1/2} \left[\frac{\partial e(X_k;\alpha_n)}{\partial \alpha} \frac{A_k - e(X_k)}{e(X_k;\alpha_n)\{1 - e(X_k;\alpha_n)\}} \right], \\
\xi_{n,k}^{(7)} &= -c_2^{\mathrm{T}} n^{-1/2} \left[\{A_k - e(X_k)\} \frac{\partial \mu_0(X_k;\beta_{0,n})}{\partial \beta_0} \{ \mu_0(X_k) - \mu_0(X_k;\beta_{0,n}) \} \right], \\
\xi_{n,k}^{(8)} &= c_3^{\mathrm{T}} n^{-1/2} \left[\{A_k - e(X_k)\} \frac{\partial \mu_1(X_k;\beta_{1,n})}{\partial \beta_1} \{ \mu_1(X_k) - \mu_1(X_k;\beta_{1,n}) \} \right], \\
\xi_{n,k}^{(9)} &= c_0 n^{-1/2} (2A_{k-n} - 1) \left\{ 1 + M^{-1} K_{S_{A_{k-n}}(\theta_n),k-n} \right\} \left\{ Y_{k-n} - \mu_{A_{k-n}}(X_{k-n}) \right\} \\
\xi_{n,k}^{(10)} &= c_2^{\mathrm{T}} n^{-1/2} (1 - A_{k-n}) \frac{\partial \mu_0(X_{k-n};\beta_{0,n})}{\partial \beta_0} \{ Y_{k-n} - \mu_0(X_{k-n}) \}, \\
\xi_{n,k}^{(11)} &= c_3^{\mathrm{T}} n^{-1/2} A_{k-n} \frac{\partial \mu_1(X_{k-n};\beta_{1,n})}{\partial \beta_1} \{ Y_{k-n} - \mu_1(X_{k-n}) \}.
\end{aligned}
$$

Consider the σ-fields

$$\mathcal{F}_{n,k} = \begin{cases} \sigma(A_1, \ldots, A_k, X_1, \ldots, X_k), & 1 \le k \le n, \\ \sigma(A_1, \ldots, A_n, X_1, \ldots, X_n, Y_{k-1}, \ldots, Y_{k-n}), & 2n+1 \le k \le 3n. \end{cases}$$

Then, we have $\left\{ \sum_{k=1}^{i} \xi_{n,i}, \mathcal{F}_{n,i}, 1 \le i \le 2n \right\}$ is a martingale for each $n \ge 1$, which follows by the following reasons:

(i) because $S_{a,k}(\theta_n)$ is a double balancing score,

$$\mathbb{E}_{\theta_n}(\xi_{n,k}^{(1)} | \mathcal{F}_{n,k-1}) = \mathbb{E}\left(c_0 n^{-1/2}\left[\mu_1\{S_{1,k}(\theta_n)\} - \mu_0\{S_{0,k}(\theta_n)\} - \tau(\theta_n)\right] \mid \mathcal{F}_{n,k-1}\right) = 0;$$

(ii) let $\mathcal{F}_{n,k}^0 = \sigma\{A_1, \ldots, A_k, S_1(\theta_n), \ldots, S_k(\theta_n)\}$ for $1 \le k \le n$, then

$$\begin{aligned} \mathbb{E}_{\theta_n}(\xi_{n,k}^{(2)} | \mathcal{F}_{n,k-1}) &= \mathbb{E}_{\theta_n}\{\mathbb{E}_{\theta_n}(\xi_{n,k}^{(2)} | \mathcal{F}_{n,k-1}^0) \mid \mathcal{F}_{n,k-1}\} \\ &= c_0 n^{-1/2} \mathbb{E}_{\theta_n}\left((2A_k - 1)\left\{1 + M^{-1} K_{S_{A_k}(\theta_n),k}\right\}\right. \\ &\quad \left. \times \mathbb{E}_{\theta_n}\left[\mu_{A_k}(X_k) - \mu_{A_k}\{S_{A_k,k}(\theta_n)\} | \mathcal{F}_{n,k-1}^0\right] \mid \mathcal{F}_{n,k-1}\right) \\ &= c_0 n^{-1/2} \mathbb{E}_{\theta_n}\left[(2A_k - 1)\left\{1 + M^{-1} K_{S_{A_k}(\theta_n),k}\right\} \times 0 \mid \mathcal{F}_{n,k-1}\right] \\ &= 0; \end{aligned}$$

(iii) $\mathbb{E}_{\theta_n}\left(\xi_{n,k}^{(3)} | \mathcal{F}_{n,k-1}\right) = \mathbb{E}_{\theta_n}\left(\xi_{n,k}^{(4)} | \mathcal{F}_{n,k-1}\right) = \mathbb{E}_{\theta_n}\left(\xi_{n,k}^{(5)} | \mathcal{F}_{n,k-1}\right) = 0$ because $\mathbb{E}_{\theta_n}\{U(\theta_n)\} = 0$;

(iv) by the conditioning argument,

$$\mathbb{E}_{\theta_n}\left(\xi_{n,k}^{(6)} | \mathcal{F}_{n,k-1}\right) = \mathbb{E}_{\theta_n}\left[c_1^{\mathrm{T}} n^{-1/2} \frac{\partial e(X_k;\alpha_n)}{\partial \alpha} \frac{\mathbb{E}\left\{A_k - e(X_k) | \mathcal{F}_{n,k-1}, X_k\right\}}{e(X_k;\alpha_n)\{1 - e(X_k;\alpha_n)\}} | \mathcal{F}_{n,k-1}\right] = 0;$$

(v) $\mathbb{E}_{\theta_n}\left(\xi_{n,k}^{(7)} | \mathcal{F}_{n,k-1}\right) = 0$ and $\mathbb{E}_{\theta_n}\left(\xi_{n,k}^{(8)} | \mathcal{F}_{n,k-1}\right) = 0$ due to that fact that $A_k - e(X_k)$ is unbiased conditional on X_k;

(vi) $\mathbb{E}_{\theta_n}\left(\xi_{n,k}^{(9)} | \mathcal{F}_{n,k-1}\right) = 0$ because $(1 - A_{k-n})\{Y_{k-n} - \mu_0(X_{k-n})\}$ is unbiased given $\mathcal{F}_{n,k-1}$;

(vii) $\mathbb{E}_{\theta_n}\left(\xi_{n,k}^{(10)} | \mathcal{F}_{n,k-1}\right) = 0$ because $A_{k-n}\{Y_{k-n} - \mu_1(X_{k-n})\}$ is unbiased given $\mathcal{F}_{n,k-1}$.

Therefore, we can apply the martingale central limit theorem (Billingsley, 1995) to derive the limiting distribution of T_n. Under Assumption A3, we can verify the conditions for the martingale central limit theorem hold. It follows that under $\mathbb{P}^{\theta_n}$, $T_n \to \mathcal{N}(0, \sigma^2)$ in distribution, as $n \to \infty$, where $\sigma^2 = \operatorname{plim} \sum_{k=1}^{2n} \mathbb{E}_{\theta_n}(\xi_{n,k}^2 | \mathcal{F}_{n,k-1})$. Under Assumption A4, we thus derive the expression of σ^2 and specify the components in (A13) with

$$\gamma_1 = (\gamma_{11}^{\mathrm{T}}, \gamma_{12}^{\mathrm{T}}, \gamma_{13}^{\mathrm{T}})^{\mathrm{T}}, \tag{A14}$$

$$\gamma_{11} = \mathbb{E}\left(\left[\mu_1\{S_1(\theta^*)\} - \mu_0\{S_0(\theta^*)\} - \tau\right] \frac{\partial e(X;\alpha^*)}{\partial \alpha} \frac{A - e(X;\alpha^*)}{e(X;\alpha^*)\{1 - e(X;\alpha^*)\}}\right)$$

$$+\mathbb{E}\left(\left[\mu_1(X)-\mu_1\{S_1(\theta^*)\}\right]\frac{\partial e(X;\alpha^*)}{\partial\alpha}\frac{1-e(X;\alpha^*)}{e(X;\alpha^*)\{1-e(X;\alpha^*)\}}\right)$$
$$-\mathbb{E}\left(\left[\mu_0(X)-\mu_0\{S_0(\theta^*)\}\right]\frac{\partial e(X;\alpha^*)}{\partial\alpha}\frac{-e(X;\alpha^*)}{e(X;\alpha^*)\{1-e(X;\alpha^*)\}}\right),$$

$$\gamma_{12}=-\mathbb{E}\left(\left[\mu_1\{S_1(\theta^*)\}-\mu_0\{S_0(\theta^*)\}-\tau\right](1-A)\frac{\partial\mu_0(X;\beta_0^*)}{\partial\beta_0}\{\mu_0(X)-\mu_0(X;\beta_0^*)\}\right)$$
$$-\mathbb{E}\left(\left[\mu_0(X)-\mu_0\{S_0(\theta^*)\}\right]\frac{\partial\mu_0(X;\beta_0^*)}{\partial\beta_0}\{\mu_0(X)-\mu_0(X;\beta_0^*)\}\right)-\mathbb{E}\left\{\frac{\partial\mu_0(X;\beta_0^*)}{\partial\beta_0}\sigma_0^2(X)\right\},$$

and

$$\gamma_{13}=-\mathbb{E}\left(\left[\mu_1\{S_1(\theta^*)\}-\mu_0\{S_0(\theta^*)\}-\tau\right]A\frac{\partial\mu_1(X;\beta_1^*)}{\partial\beta_1}\{\mu_1(X)-\mu_1(X;\beta_1^*)\}\right)$$
$$+\mathbb{E}\left(\left[\mu_1(X)-\mu_1\{S_1(\theta^*)\}\right]\frac{\partial\mu_1(X;\beta_1^*)}{\partial\beta_1}\{\mu_1(X)-\mu_1(X;\beta_1^*)\}\right)-\mathbb{E}\left\{\frac{\partial\mu_1(X;\beta_1^*)}{\partial\beta_1}\sigma_1^2(X)\right\}.$$