AStA Advances in Statistical Analysis (2025) 109:483–508
https://doi.org/10.1007/s10182-025-00535-4

ORIGINAL PAPER

# Testing for causal effect for binary data when propensity scores are estimated through Bayesian Networks

**Paola Vicard[1] · Paola Maria Vittoria Rancoita[2] · Federica Cugnata[2] · Alberto Briganti[2] · Fulvia Mecatti[3] · Clelia Di Serio[2] · Pier Luigi Conti[4]**

Received: 29 July 2023 / Accepted: 12 July 2025 / Published online: 28 August 2025


**Abstract**
This paper proposes a new statistical approach for assessing treatment effect using Bayesian Networks (BNs). The goal is to draw causal inferences from observational data with a binary outcome and discrete covariates. The BNs are here used to estimate the propensity score, which enables flexible modeling and ensures maximum likelihood properties. When the propensity score is estimated by BNs, two point estimators are considered—Hájek and Horvitz–Thompson—based on inverse probability weighting, and their main distributional properties are derived for constructing confidence intervals and testing hypotheses about the absence of the treatment effect. Empirical evidence is presented to show the good behavior of the proposed methodology through a simulation study mimicking the characteristics of a real dataset of prostate cancer patients from Milan San Raffaele Hospital.

**Keywords** Bayesian Network · Propensity score · Covariate balance · Observational study · ATE estimation · Testing treatment effect

## 1 Introduction

Observational data in biomedical research are progressively gaining room, owing to increasing costs and regulatory hurdles associated with conducting clinical trials. However, its potential for external validity is still being debated. As a matter of fact, working with observational data has many challenges, such as non-randomized (uncontrolled) assignment of treatments and non-controlled study designs. Developing new rigorous statistical methods for overcoming these problems could potentially improve generalizability.

From a statistical perspective, the uncontrolled unbalanced design often associated to observational data could be approached by estimating the probability of receiving the treatments given the observed covariates (propensity score, PS). The PS can be then used for rebalancing the sample while statistically analyzing the data.

Extended author information available on the last page of the article

The actual statistical approaches for evaluating the treatment effects in this context are mainly based on: (1) matching (possibly based on PS); (2) stratification matching (possibly based on PS); and (3) inverse probability weighting (IPW) through estimated PS. In this article, we will focus on the latter approach, by defining a new test for the absence of treatment effect within the inverse probability weighting approach, as the other methods, even when using PS, have various disadvantages. For example, while stratification matching is useful for estimating the average treatment effect (ATE), as well as quantities such as average treatment effect on treated (ATT) or weighted average treatment effect (WATE), it is problematic to use for testing the presence of treatment effects and does not allow for straightforward estimation beyond the above mentioned quantities. In contrast, IPW avoids these limitations by easily extending to identifiable parameters other than ATE, ATT, WATE, and enabling the construction of tests for treatment effect (cfr. Conti and De Giovanni 2022; Donald and Hsu 2014 and references therein). Additionally, stratification matching requires a careful choice of the number of strata, which is subjective and must increase with the sample size to ensure consistency of estimates. Unfortunately, it is unclear how fast the number of strata should increase, which can lead to theoretically unsafe estimates. Furthermore, when stratification is based on covariate values, the number of resulting strata could become too large as the number of covariates and/or their levels increase, leading to insufficient data issues due to excessive granularity. Matching estimators for treatment effects are mainly studied under the assumption of continuous covariates with positive density; (cfr. the fundamental paper Abadie and Imbens 2016 and references therein). The extension to discrete covariates (and outcomes, as well), as in the present paper, requires non-trivial theoretical changes and is beyond the scope of this paper. Moreover, their use as a test-statistic is problematic, basically for two reasons. First of all, variance estimation is difficult. In the second place, results cannot be easily extended to quantities of interest different from ATE.

A common feature of the above mentioned methods (matching, stratification matching, and IPW) is that they typically require the estimation of PS. In this paper, we propose using Bayesian Networks (BNs) for PS estimation. BNs can capture the dependence structure between covariates and the treatment assignment in a data-driven perspective. By their nature, they have a clear interpretation in terms of both dependencies among covariates and influence of covariates on the probability of receiving treatment. Therefore, BNs are flexible; they also have the key advantage of providing consistent (and asymptotically efficient) estimates of propensity scores under the setting of the present paper (categorical covariates). This is a serious theoretical argument in favor of their use.

In conclusion, this paper has two main goals.

1. Proposal of a new Bayesian Networks-based estimator for propensity score, with sound statistical properties.
2. Moving beyond ATE point estimation to construct a test-statistic for the evaluation of the null hypothesis of the absence of treatment effect, by estimating potential outcome probabilities based on the estimated propensity score.

The paper is organized as follows. In Sect. 2, we introduce the basic concepts related to propensity score and Bayesian Networks. In Sect. 3, the properties of the Bayesian Networks-based propensity score estimators are explored. Additionally, two types of ATE estimators in the IPW class using the obtained propensity score are examined. It is shown that they are asymptotically equivalent, and their statistical efficiency is the same for large datasets. Then, a statistical test is developed to detect the possible presence of the treatment effect. In Sect. 4, it is shown that, under PS specification errors, the two types of ATE estimators behave differently, and have varying degrees of robustness with respect to misspecification. Finally, in Sect. 5, we present a motivating case study based on a large observational real data from prostate cancer patients. Section 6 describes a simulation study evaluating the performance of the considered ATE estimators and the corresponding tests. Proofs of theoretical results are in Appendix. Supplementary Material deals with the estimation of WATE and ATT.

## 2 Basic theory

### 2.1 Notation and basic concepts

Consider a random sample of $n$ independent subjects. Each of them could either receive a treatment $T$ or not; conventionally, $T = 1$ means a subject does receive the treatment, while $T = 0$ means the subject does not receive the treatment. Let $Y_{(1)}$ denote the potential outcome of a subject in the presence of the treatment (*i.e.* when $T = 1$), and let $Y_{(0)}$ denote the potential outcome of a subject in the absence of the treatment (*i.e.* when $T = 0$). The observed outcome for a subject is then:

$$Y = TY_{(1)} + (1 - T)Y_{(0)}. \tag{1}$$

Equation (1) can be more conveniently written by using the indicator function of the treatment. Let $I_{(T=1)}$ be the indicator of the event $T = 1$, namely

$$I_{(T=1)} = \begin{cases} 1 & \text{if } T = 1 \\ 0 & \text{if } T = 0, \end{cases}$$

thus $I_{(T=0)} = 1 - I_{(T=1)}$ is the indicator function of the event $T = 0$. The observed outcome $Y$ can be then written as

$$Y = Y_{(1)}I_{(T=1)} + Y_{(0)}I_{(T=0)}. \tag{2}$$

The treatment has no effect *in distribution* (or *on the average*, which is the same for dichotomous outcomes) when $Y_{(0)}$ and $Y_{(1)}$ have the same probability distribution. If $\stackrel{d}{=}$ denotes equality in distribution, then there is absence of treatment in distribution if and only if (iff) $Y_{(0)} \stackrel{d}{=} Y_{(1)}$. For the sake of brevity, from now on the term "absence of treatment effect" will be used in place of the more correct "absence of treatment effect in distribution" (or "on the average").

The assignment-to-treatment mechanism is not a "purely random" mechanism, as in an experimental framework. Due to the presence of confounding covariates, there could be considerable differences among subjects receiving different treatment levels. As usual in the literature (cfr. Imbens and Rubin 2015), we assume here that the assignment-to-treatment mechanism only depends on a set of $L$ observed, pre-treatment, covariates. From now on, the vector of relevant covariates is denoted by $\boldsymbol{X} = (X_1 \cdots X_L)$. Furthermore, the probability of receiving the treatment conditionally on $\boldsymbol{X} = \boldsymbol{x}$, namely, the *propensity score*, is denoted by

$$p_1(\boldsymbol{x}) = P(T = 1|\boldsymbol{X} = \boldsymbol{x}). \tag{3}$$

For the sake of completeness, in the sequel we will denote by

$$p_0(\boldsymbol{x}) = P(T = 0|\boldsymbol{X} = \boldsymbol{x}) \tag{4}$$

the probability of not receiving the treatment, given $\boldsymbol{X} = \boldsymbol{x}$. The relationship $p_0(\boldsymbol{x}) + p_1(\boldsymbol{x}) = 1$ holds.

In the present paper, we focus on the case where pre-treatment covariates are discrete, finite random variables (r.v.s), and the potential outcomes are dichotomous variables. The assumptions on which our analysis rests are listed below.

A1. *Discreteness*. Each covariate $X_l$ is discrete, finite. With no loss of generality, it may take nominal values $1, \ldots, r_l$ with positive probability. The potential outcomes $Y_{(k)}$ are dichotomic r.v.s. For the sake of simplicity, and with no loss of generality, in the sequel we assume that $Y_{(k)}$, $k = 0, 1$, may take values 0, 1 with (marginal) probability:

$$\theta_k = P(Y_{(k)} = 1); \;\; k = 0, 1. \tag{5}$$

A2. *Unconfoundedness*. $T \perp (Y_{(0)}, Y_{(1)})|\boldsymbol{X}$, where the symbol $\perp$ denotes stochastic independence.

A3. *Common support*. There exists a positive real $\delta$ for which $\delta \leq p_k(\boldsymbol{x}) \leq 1 - \delta$ for each $\boldsymbol{x}$ and $k = 0, 1$.

In the case under examination, the absence of treatment effect is equivalent to say that $\theta_0 = \theta_1$. Hence, testing for the absence of treatment effect reduces to the following hypothesis problem

$$\begin{cases} H_0 : \theta_1 = \theta_0 \\ H_1 : \theta_1 \neq \theta_0 \end{cases} \tag{6}$$

The hypothesis problem (6) can be also expressed in terms of the familiar notion of average treatment effect (ATE), defined as $ATE = E[Y_{(1)}] - E[Y_{(0)}]$. Since, from definition (5), $E[Y_{(1)}] = \theta_1$, $E[Y_{(0)}] = \theta_0$, the null hypothesis is equivalent to $ATE = 0$.

Observed data for $n$ subjects are formally defined as the triplets $(Y_i, T_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$. The r.v.s $(Y_i, T_i, \boldsymbol{X}_i)$ are assumed to be independent and identically distributed (*i.i.d.*).

As it appears from Imbens and Wooldridge (2009), testing whether the distribution of $Y_{(0)}$ is different from that of $Y_{(1)}$ is a problem of considerable interest, and relevance, as well. Bootstrap-based tests are proposed in Abadie (2002), and permutation tests for randomized experiments are studied in Ding (2017) and Wu and Ding (2018). Specifically devoted to the case of ordinal qualitative outcomes is the paper (Lu et al. 2019), where a test for an upper bound for $\gamma = P(Y_{(1)} > Y_{(0)}) - P(Y_{(1)} < Y_{(0)})$ is proposed; for further articles dealing with this subject, the reader is deferred to references in the above mentioned papers.

### 2.2 Bayesian Networks: basic aspects

The basic idea exploited in the present section is to estimate propensity scores $p_k(\boldsymbol{x})$, $k = 0, 1$, by using a Bayesian Network model for $(T, \boldsymbol{X})$; cfr. Cowell et al. (1999).

To simplify the notation, the r.v. $T$ is here denoted by $X_0$, so that $(T, \boldsymbol{X}) = (X_0, X_1, \ldots, X_L) = (X_0, \boldsymbol{X})$. A BN is a multivariate statistical model satisfying sets of conditional independence statements displayed in a directed acyclic graph (DAG), consisting in a set of *nodes* and a set of *directed arcs* connecting pairs of nodes. The nodes represent random variables and the arcs represent direct dependencies among the variables. Each node is associated with an index $l = 0, 1, \ldots, L$, which, in its turn, corresponds to the random variable $X_l$. A directed graph is acyclic, in the sense that it is forbidden to start from a node and, following arrows directions, go back to the starting node.

Next, let $pa(l)$ be the set of *parents* of node $l$, *i.e.* the set of all nodes with a directed arc pointing to node $l$. In equivalent terms, $\boldsymbol{X}_{pa(l)}$ is $ch(l)$ denotes the set of the set of all variables in $(X_0, \boldsymbol{X})$ "graphically" linked to $X_l$. Analogously, $ch(l)$ denotes the set of *children* of node $l$, *i.e.* the set of all nodes with a directed arc pointing to them from node $l$. Moreover, the Markov blanket of node $l$, $Mb(l)$, is the set of its parents, its children and the parents of its children. If $l$ is a childless node, then $Mb(l) = pa(l)$. Note that in the present paper, we focus on pre-treatment variables only, hence the treatment node is childless. Conditional independencies can be read off the DAG by means of the Markov properties. Specifically, the joint probability distribution over a DAG satisfies the local Markov property if each variable, say $X_l$, is independent of its non-descendants conditionally on its parents, where the set of non-descendants of $X_l$ is composed by all the variables for which there is no directed path from $X_l$ to them. For other Markov properties for DAGs (that have been shown to be equivalent), cfr. Lauritzen (1996). In a BN, each node of a DAG is associated with the distribution of the corresponding variable given its parents (if a node has no parents, it is associated with its marginal distribution). Formally speaking a BN is a pair DAG/joint probability distribution satisfying the Markov properties. The joint distribution can be factorized according to the DAG as the product of the conditional distributions associated to each node given its parents. In general, the chain rule for the distribution of $(X_0, X_1, \ldots, X_L)$ states that:

$$p(X_0 = x_0, \boldsymbol{X} = \boldsymbol{x}) = p(x_0, x_1, \ldots, x_L) = \prod_{l=0}^{L} p(x_l | \boldsymbol{x}_{pa(l)}).$$

When covariates $\mathbf{X}$ are discrete, the use of BNs to estimate propensity scores has several positive theoretical features and practical advantages. On the theoretical side, BNs allow to consider maximum likelihood estimators (MLEs) of propensity scores $p_k(\boldsymbol{x})$, that possess "usual" properties of MLEs, *i.e.* they are $\sqrt{n}$-consistent and asymptotically efficient. Computation of MLEs in BNs is carefully studied in Cowell et al. (1999). More explicitly, for each $l$ denote by $\mathbf{P}_l$ the Cartesian product

$$\mathbf{P}_l = \prod_{j \in pa(l)} \{1, \ldots, r_j\}.$$

i.e. the parent set configuration of node $l$, $l = 0, 1, \ldots, L$. The parameters of the BN are the conditional probabilities

$$\lambda_{x_l | \boldsymbol{x}_{pa(l)}} = p(x_l | \boldsymbol{x}_{pa(l)}), \quad x_l \in \{1, \ldots, r_l\}, \ \boldsymbol{x}_{pa(l)} \in \mathbf{P}_l, \ l = 0, 1, \ldots, L. \tag{7}$$

The constraints

$$\sum_{x_l=1}^{r_l} \lambda_{x_l | \boldsymbol{x}_{pa(l)}} = 1 \quad \forall \boldsymbol{x}_{pa(l)} \in \mathbf{P}_l, \ l = 0, 1, \ldots L$$

hold.

Let $\boldsymbol{\lambda}$ be the vector of parameters (7). The likelihood function for the BN under consideration essentially corresponds to a multinomial likelihood with parameters vector $\boldsymbol{\lambda}$. If $x_{il}$ denotes the value of the variable $X_l$ for the $i$th observation, and $\boldsymbol{x}_{ipa(l)}$ are the values of the parents of $X_l$ for the $i$th observation, the log-likelihood is equal to

$$\begin{aligned} l(\boldsymbol{\lambda}) &= \sum_{i=1}^{n} \sum_{l=0}^{L} \log \lambda_{x_{il} | \boldsymbol{x}_{ipa(l)}} \\ &= \sum_{l=0}^{L} \sum_{\boldsymbol{x}_{pa(l)} \in \mathbf{P}_l} \sum_{x_l=1}^{r_l} n(x_l, \boldsymbol{x}_{pa(l)}) \log \lambda_{x_l | \boldsymbol{x}_{pa(l)}} \end{aligned}$$

where $n(x_l, \boldsymbol{x}_{pa(l)})$ is the number of times the pair $(x_l, \boldsymbol{x}_{pa(l)})$ is observed in the sample. Hence, the MLE of $\lambda_{x_l | \boldsymbol{x}_{pa(l)}}$ is just the empirical conditional proportion

$$\widehat{\lambda}_{x_l | \boldsymbol{x}_{pa(l)}} = \frac{n(x_l, \boldsymbol{x}_{pa(l)})}{n(\boldsymbol{x}_{pa(l)})}$$

where

$$n(\boldsymbol{x}_{pa(l)}) = \sum_{x_l=1}^{r_l} n(x_l, \boldsymbol{x}_{pa(l)})$$

is the number of times $pa(x_l)$ is observed in the sample.

The MLE of $p(x_0, x_1, \ldots, x_L)$ is obtained through the invariance property:

$$\widehat{p}(x_0, x_1, \ldots, x_L) = \prod_{l=0}^{L} \widehat{\lambda}_{x_l|\boldsymbol{x}_{pa(l)}} = \prod_{l=0}^{L} \frac{n(x_l, \boldsymbol{x}_{pa(l)})}{n(\boldsymbol{x}_{pa(l)})}. \tag{8}$$

The association structure, *i.e.* the presence or absence of edges and their direction, between the variables and their conditional probability distributions, can be either known in advance by subject-matter knowledge or has to be learnt (estimated) from the data. In the second case, it is crucial to appropriately estimate the dependence model. To this aim there are three main classes of learning algorithms: constraint-based, score-based and hybrid algorithms (cfr. Drton and Maathuis 2017). In constraint-based learning, the DAG is learned according to a sequence of independence tests carried out on data, and to logic rules based on test results. In score-based learning, it is necessary to compute the maximum likelihood of each structure, penalize it in order to avoid overfitting (the penalized likelihood is also named score), and search for the structure with the best score. Penalization is usually performed based on AIC or BIC. When the number of nodes is larger than 5, the number of possible network structures is so large that it is necessary to resort to appropriate search algorithms such as the greedy search algorithm. Finally, hybrid algorithms combine the ideas and the advantages of the previous two types of algorithms. Compared to traditional statistical models, BNs allow a high degree of flexibility in discovering the dependence relationships among covariates and treatment levels, that is the presence of possible relationships of conditional independence (*i.e.* for the presence/absence of arcs) in the model, without requiring the identification, for each treatment level, of polynomial terms and interactions terms for the covariates in $\boldsymbol{x}$ to be included in the model.

In the sequel, theoretical results will be illustrated with a three-fold purpose. First of all, we consider estimates of propensity scores on the basis of an appropriate Bayesian Network (BN) model. As it will be seen, if the network (dependence) structure is (correctly) specified *a priori*, the obtained estimates possess highly desirable statistical properties. The effect of learning the network structure, leading to post-selection inference on propensity scores, will be also discussed. In the second place, two types of ATE estimators based on propensity score weighting are considered, and their statistical properties are studied. As it will be also seen, one of them is usually better under misspecification of the propensity score model.

Finally, estimated propensity scores are used to construct a test-statistic for the presence/absence of treatment effect.

## 3 Theoretical results

### 3.1 Estimation of propensity scores by Bayesian Network models

As already said, due to their flexibility in modeling the dependence structures among covariates, BNs naturally offer an excellent tool to estimate propensity scores. From

(8), and using again the invariance property, the MLE of the propensity score $p_k(\boldsymbol{x})$ is equal to

$$\widehat{p}_k(\boldsymbol{x}) = \frac{\widehat{p}(k, x_1, \ldots, x_L)}{\widehat{p}(1, x_1, \ldots, x_L) + \widehat{p}(0, x_1, \ldots, x_L)}, \quad k = 0, 1. \tag{9}$$

The main theoretical properties of (9) are consequences of general results on MLEs, and the same also holds for the marginal probability distribution of $(x_1, \ldots, x_L)$. In order to establish results in the proper way, in the sequel we will denote by $p_k^{TRUE}(\boldsymbol{x})$ the *true* propensity scores, and by $\mathcal{P} = \{(p_0(\boldsymbol{x}; \boldsymbol{\lambda})), p_1(\boldsymbol{x}; \boldsymbol{\lambda}); \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$ the *working* statistical model for propensity scores under BNs.

**Proposition 1** Suppose that assumptions *A*1-*A*3 are met. The following results hold.

(a) $p_k^{TRUE}(\boldsymbol{x}) \in \mathcal{P}$.
(b) The estimator (9) is consistent:

$$\widehat{p}_k(\boldsymbol{x}) \xrightarrow{p} p_k^{TRUE}(\boldsymbol{x}) \text{ as } n \to \infty, \quad k = 0, 1 \tag{10}$$

for every value of $\boldsymbol{x}$, the symbol $\xrightarrow{p}$ denoting convergence in probability. This statement holds either if the network structure is known in advance, or if it is learned from data through BIC.
(c) If the structure of the network is learned from data through BIC, then

$$\left|\widehat{p}_k(\boldsymbol{x}) - p_k^{TRUE}(\boldsymbol{x})\right| = O_p\left(\sqrt{\frac{\log n}{n}}\right). \tag{11}$$

In addition, if the network structure is known in advance, then the following two further statements hold.

(d) The estimator (9) is asymptotically Normally distributed:

$$\sqrt{n}(\widehat{p}_k(\boldsymbol{x}) - p_k^{TRUE}(\boldsymbol{x})) \xrightarrow{d} N(0, I_k^{-1}(\boldsymbol{x})) \text{ as } n \to \infty, \quad k = 0, 1 \tag{12}$$

where $I_k^{-1}(\boldsymbol{x})$ is the reciprocal of the Fisher information, and $\xrightarrow{d}$ denotes convergence in distribution.
(e) The estimator (9) is asymptotically efficient. If $\widetilde{p}_k(\boldsymbol{x})$ is another (sequence of) estimator(s) of the PS $p_k(\boldsymbol{x})$, consistent and asymptotically Normal with $\sqrt{n}(\widetilde{p}_k(\boldsymbol{x}) - p_k^{TRUE}(\boldsymbol{x})) \xrightarrow{d} N(0, V_k(\boldsymbol{x}))$ as $n \to \infty$, then:

$$V_k(\boldsymbol{x}) \geq I_k^{-1}(\boldsymbol{x}) \tag{13}$$

for all $\boldsymbol{x}$s and $k = 0, 1$.

Proposition 1 is "universal", in the sense that it holds true whatever the "true" value of the propensity score may be. In other terms, statement (*a*) above ensures

that the true PS is always a member of the working statistical model for BNs, thus avoiding the possibility of misspecification errors asymptotically (under conditions A1–A3, of course), even when their probabilistic structure is learned from data. To simplify the notation in the sequel, in case of non-ambiguity, the symbol $p_k(\boldsymbol{x})$ will be used in place of $p_k^{TRUE}(\boldsymbol{x})$.

Since the covariates are discrete, propensity scores could be simply estimated by using relative frequencies, and the resulting estimators would be consistent and asymptotically Normal. As also remarked by a referee, the main motivation for using Bayesian Networks is that they make it possible to exploit conditional independence relations to estimate propensity scores more efficiently than by simple frequencies. Proposition 1 offers the most relevant theoretical support in this direction. As a matter of fact, statements (*d*), (*e*) show that the estimator (9) obtained through BNs is not only consistent and asymptotically normally distributed, but it is also *asymptotically efficient*. Competitor PS estimators can be either as efficient as those obtained through BNs, or they are beated in terms of asymptotic Mean Squared Error (MSE), which is a reason in favor of using BNs.

One of the nicest features of BNs is the possibility to learn their probabilistic structure from data. Statement (*c*) shows that, in case of BIC penalization rule, $\widehat{p}_k(\boldsymbol{x})$ is still consistent when the network structure is learned from data, although at a lower speed of convergence; cfr. also Balov (2013), from which statement (*b*) can be deduced in the special case of absence of missing data, and Nandy et al. (2018), which is devoted to hybrid algorithms. Note that the same does not hold for AIC penalization, because of the lack of consistency; cfr. Shibata (1986).

We note *in passim* that, since Hahn (1998), the estimation of propensity score is usually considered as ancillary in estimating the parameter $\theta_1 - \theta_0$. Accordingly, a different approach to weighting consists in using moment method based on appropriate balancing with respect to covariates, or better in controlling the maximal imbalance; cfr. Zubizzarreta (2015), Chattopadhyay et al. (2020) and Ben-Michael et al. (2021). However, the speed of convergence of $\widehat{p}_k(\boldsymbol{x})$ has a non-trivial consequence in terms of imbalance. According to Ben-Michael et al. (2021), consider a bounded function $f(\boldsymbol{x})$, and let the imbalance with respect to $f$ be equal to

$$\left| \frac{1}{n} \sum_{i=1}^{n} I_{(T_i=k)} \widehat{p}_k(\boldsymbol{x}_i)^{-1} f(\boldsymbol{x}_i) - \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{x}_i) \right|. \tag{14}$$

The Law of Large Numbers and the consistency of $\widehat{p}_k$ guarantee that (14) tends in probability to 0 as $n$ increases. As a special case, the imbalance of single covariates (corresponding to $f(\boldsymbol{x}_i) = x_{ij}$, $j = 1, \ldots, L$), *i.e.*

$$\left| \frac{1}{n} \sum_{i=1}^{n} I_{(T_i=k)} \widehat{p}_k(\boldsymbol{x}_i)^{-1} x_{ij} - \frac{1}{n} \sum_{i=1}^{n} x_{ij} \right|, \; k = 0, 1, \; j = 1, \ldots, L$$

tends in probability to 0 as $n$ increases (cfr., for instance, Zubizzarreta 2015).

Furthermore, using Lemma 1 (see Appendix) it is easy to see that (14) tends to 0 in probability at the same speed at which $\sup_{\boldsymbol{x}} \left| \widehat{p}_k(\boldsymbol{x}_i)^{-1} - p_k(\boldsymbol{x})^{-1} \right|$ tends to 0. In our case, due to the assumptions made, the above supremum is $O_P(n^{-1/2})$, which

is the maximal speed. Interestingly enough, this result also holds when the structure of the Bayesian Network is learned from the data, *i.e.* in case of post-selection inference on propensity scores, although at a slightly reduced speed of convergence ($O_p(\sqrt{\log n/n})$ instead of $O_p(1/\sqrt{n})$).

### 3.2 Estimation of potential outcomes probabilities

In view of the estimator (9) of propensity score, an estimator of the marginal probability $\theta_k$ in Eq. (5) is needed. We consider here the following Hájek-type estimator (cfr., among the others, Hernán and Robins 2006), that is a slight modification of the commonly used inverse probability weighted estimator though with improved statistical properties:

$$\widehat{\theta}_k^H = \frac{1}{\sum_{i=1}^n I_{(T_i=k)}\widehat{p}_k(\boldsymbol{x}_i)^{-1}} \sum_{i=1}^n I_{(Y_i=1)}I_{(T_i=k)}\widehat{p}_k(\boldsymbol{x}_i)^{-1};\ k=0,\ 1. \tag{15}$$

As a natural competitor of (15), we also consider the inverse probability weighted estimator (*i.e.* the Horvitz–Thompson-type estimator) as in Lunceford and Davidian (2004):

$$\widehat{\theta}_k^{HT} = \frac{1}{n}\sum_{i=1}^n I_{(Y_i=1)}I_{(T_i=k)}\widehat{p}_k(\boldsymbol{x}_i)^{-1};\ k=0,\ 1. \tag{16}$$

Our first result is that both (15) and (16) are consistent estimators of $\theta_k$.

**Proposition 2** Under assumptions A1–A3, as $n \to \infty$:

$$\left|\widehat{\theta}_k^H - \theta_k\right| \xrightarrow{p} 0;\ k=0,\ 1; \tag{17}$$

$$\left|\widehat{\theta}_k^{HT} - \theta_k\right| \xrightarrow{p} 0;\ k=0,\ 1. \tag{18}$$

The main result of the present section concerns the asymptotic distribution of the estimators (15), (16), once properly normalized. Define first the vectors

$$\widehat{\boldsymbol{\theta}}^H = \begin{bmatrix}\widehat{\theta}_0^H \\ \widehat{\theta}_1^H\end{bmatrix},\ \widehat{\boldsymbol{\theta}}^{HT} = \begin{bmatrix}\widehat{\theta}_0^{HT} \\ \widehat{\theta}_1^{HT}\end{bmatrix},\ \boldsymbol{\theta} = \begin{bmatrix}\theta_0 \\ \theta_1\end{bmatrix},$$

and consider the bivariate r.v.s

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}^H - \boldsymbol{\theta}),\ \sqrt{n}(\widehat{\boldsymbol{\theta}}^{HT} - \boldsymbol{\theta}). \tag{19}$$

**Proposition 3** Suppose assumptions A1–A3 hold. Then, there exists a sequence of *i.i.d.* bivariate random vectors $\boldsymbol{h}^*(Y_i,\ T_i,\ \boldsymbol{X}_i)$, with

$$\boldsymbol{h}^*(Y_i,\, T_i,\, \boldsymbol{X}_i) = \begin{bmatrix} h_0^*(Y_i,\, T_i,\, \boldsymbol{X}_i) \\ h_1^*(Y_i,\, T_i,\, \boldsymbol{X}_i) \end{bmatrix}$$

and $E[h_k^*(Y_i,\, T_i,\, \boldsymbol{X}_i)] = 0$ having the following properties as $n \to \infty$:

1. $\sqrt{n}(\widehat{\boldsymbol{\theta}}^{HT} - \boldsymbol{\theta})$ and $n^{-1/2} \sum_{i=1}^n \boldsymbol{h}^*(Y_i,\, T_i,\, \boldsymbol{X}_i)$ possess the same limiting distribution;
2. $\sqrt{n}(\widehat{\boldsymbol{\theta}}^{H} - \boldsymbol{\theta})$ and $n^{-1/2} \sum_{i=1}^n \boldsymbol{h}^*(Y_i,\, T_i,\, \boldsymbol{X}_i)$ possess the same limiting distribution;
3. If $V[h_k^*(Y_i,\, T_i,\, \boldsymbol{X}_i)] > 0$, $k = 0,\, 1$, then $n^{-1/2} \sum_{i=1}^n \boldsymbol{h}^*(Y_i,\, T_i,\, \boldsymbol{X}_i) \xrightarrow{d} \boldsymbol{W}$

as $n$ goes to infinity, where $\boldsymbol{W}$ possesses a bivariate Normal $N_2(\boldsymbol{0},\, \boldsymbol{\Sigma})$ with null mean vector and covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{00} & \sigma_{01} \\ \sigma_{10} & \sigma_{11} \end{bmatrix} = E\left[\boldsymbol{h}^*(Y_i,\, T_i,\, \boldsymbol{X}_i)\, \boldsymbol{h}^*(Y_i,\, T_i,\, \boldsymbol{X}_i)^T\right].$$

As a consequence of Proposition 3, the estimators $\widehat{\boldsymbol{\theta}}^{H}$ and $\widehat{\boldsymbol{\theta}}^{HT}$ possess the same limiting distribution, and hence they are asymptotically equivalent. For this reason, we will mainly refer to $\widehat{\boldsymbol{\theta}}^{H}$.

Proposition 3 holds not only when the structure of the Bayesian Network used to estimate propensity score is known *a priori*, but also when it is learned from data (at least under BIC penalization). In this case, post-selection inference for propensity scores does not affect the limiting distribution of $\sqrt{n}(\widehat{\boldsymbol{\theta}}^{HT} - \boldsymbol{\theta})$ and of $\sqrt{n}(\widehat{\boldsymbol{\theta}}^{H} - \boldsymbol{\theta})$. The reason why this occurs is simple. Under post-selection inference for propensity scores based on BIC, we have $\widehat{p}_k(\boldsymbol{x}) = p_k(\boldsymbol{x}) + O_p(\sqrt{\log n/n})$, and this ensure the validity of Proposition 3. Using the arguments in Kim (2014), it is actually enough that $\widehat{p}_k(\boldsymbol{x}) = p_k(\boldsymbol{x}) + o_p(n^{-1/4})$. Incidentally, these results parallel a result established by Hahn (1998) for continuous covariates, establishing that the propensity score is ancillary for estimation of the average treatment effect.

Although the main interest of the present paper is in testing for $\Delta = \theta_1 - \theta_0 = 0$, we remark that, again as a consequence of the assumptions made, and using the same arguments as in Kim (2019), p. 9, the estimators $\widehat{\theta}_1^H - \widehat{\theta}_0^H$, $\widehat{\theta}_1^{HT} - \widehat{\theta}_0^{HT}$ are asymptotically efficient, in the sense that they attain the efficiency bound in Hahn (1998) and Kim (2019).

### 3.3 Testing for treatment effect

The primary goal of the present section is to construct a test for the hypothesis problem (6), *i.e.* a test for the absence of treatment effect. Define $\Delta = \theta_1 - \theta_0$. As already remarked, testing for the absence of treatment effect reduces to the following hypothesis problem

$$\begin{cases} H_0 : \Delta = 0 \\ H_1 : \Delta \neq 0 \end{cases} \tag{20}$$

A "natural" test-statistic for the above hypotheses problem is

$$D_n = \widehat{\theta}_1^H - \widehat{\theta}_0^H = \boldsymbol{a}^T \widehat{\boldsymbol{\theta}}^H$$

where $\boldsymbol{a}$ is the vector of components 1 and $-1$. The limiting distribution of $D_n$ is easily obtained from Proposition 4, which is an immediate consequence of the continuous mapping Theorem.

**Proposition 4** Suppose assumptions A1–A3 hold. The following two statements hold.

1. $\sqrt{n}(D_n - \Delta) \xrightarrow{d} N(0,\ \sigma^2)$ as $n \to \infty$, with $\sigma^2 = \boldsymbol{a}^T \boldsymbol{\Sigma} \boldsymbol{a}$.
2. Under $H_0$, $\sqrt{n} D_n \xrightarrow{d} N(0,\ \sigma_0^2)$ as $n \to \infty$.

The asymptotic variance $\sigma^2$ is unknown, and must be estimated from available data. A simple technique is jackknife, based on the systematic omission of a single observation from the sample data, on calculating the corresponding estimates of $\Delta$, $D_{n-1,(-i)}$, $i = 1, ..., n$, and on computing their average. Since for each sub-sample of size $n-1$ the construction of a BN is required, this technique may be computationally cumbersome, especially for large $n$. A different approach may be developed by exploiting (in a different context) ideas in Hirano et al. (2003). Define $\theta_k(\boldsymbol{x}) = P(Y_i = 1 | T_i = k, \boldsymbol{X}_i = \boldsymbol{x})$, $k = 0, 1$, and let $\widehat{\theta}_k(\boldsymbol{x})$ be the corresponding estimator obtained with the same approach used for estimating the propensity score (in our case by using a BN).

Define further

$$\widehat{h}_{i1} = \left( \frac{I_{(T_i=1)}}{\widehat{p}_1(\boldsymbol{x}_i)} I_{(Y_i=1)} - \widehat{\theta}_1 \right) - \frac{\widehat{\theta}_1(\boldsymbol{x}_i)}{\widehat{p}_1(\boldsymbol{x}_i)} (I_{(T_i=1)} - \widehat{p}_1(\boldsymbol{x}_i))$$

$$\widehat{h}_{i0} = \left( \frac{I_{(T_i=0)}}{\widehat{p}_0(\boldsymbol{x}_i)} I_{(Y_i=1)} - \widehat{\theta}_0 \right) - \frac{\widehat{\theta}_0(\boldsymbol{x}_i)}{\widehat{p}_0(\boldsymbol{x}_i)} (I_{(T_i=0)} - \widehat{p}_0(\boldsymbol{x}_i))$$

and

$$\widehat{d}_i = \widehat{h}_{i1} - \widehat{h}_{i0}, i = 1, ..., n.$$

Using the same approach as in Hirano et al. (2003), it is possible to see that

$$\widehat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (\widehat{d}_i)^2$$

is a consistent estimator of $\sigma^2$. As a consequence, in order to test for the presence of treatment effect, a simple procedure consists in constructing the following confidence interval at level $1 - \alpha$ for $\Delta$

$$\left[D_n - z_{\alpha/2}\frac{\widehat{\sigma}_n}{\sqrt{n}},\ D_n + z_{\alpha/2}\frac{\widehat{\sigma}_n}{\sqrt{n}}\right], \tag{21}$$

where $\widehat{z}_p$ is the $(1-p)$-quantile of the Standard Normal distribution, and in rejecting $H_0$ whenever the interval (21 ) does not contain 0.

Finally, as an approximated $p$-value for the above testing procedure, we may take $2(1-\Phi(\sqrt{n}D_n/\widehat{\sigma}_n))$, where $\Phi$ is the Standard Normal distribution function.

## 4 Estimation of treatment effects under misspecification of the propensity score model

As seen in the previous sections, BNs are an excellent tool for estimating propensity scores, in view of their "universal" consistency in Eq. (10). Asymptotically, BNs would ensure the removal of all bias in the potential outcomes probabilities estimates, since they are both universally consistent and parsimonious in terms of parameters used. However, under a limited sample size, or when structural learning is not used, or in case of omission of relevant covariates, there could be an uncorrect specification of propensity scores.

The goal of the present section is to study the behavior of estimators $\widehat{\theta}_k^{HT}$, $\widehat{\theta}_k^{H}$ when the model for propensity scores is incorrectly specified. More specifically, denote by $p_k^{TR}(\boldsymbol{x})$ the "true" propensity score, and assume that the working statistical model

$$p_k(\boldsymbol{x};\boldsymbol{\beta}) = P(T=k|,\boldsymbol{X}=\boldsymbol{x};\boldsymbol{\beta}),\ k=0,\ 1;\ \ \boldsymbol{\beta}\in\boldsymbol{\Upsilon} \tag{22}$$

is adopted, $\boldsymbol{\beta}$ being a multidimensional parameter and $\boldsymbol{\Upsilon}$ the corresponding parameter space. The working model is misspecified whenever $p_k(\boldsymbol{x};\boldsymbol{\beta}) \neq p_k^{TR}(\boldsymbol{x})$ for all $\boldsymbol{\beta}\in\boldsymbol{\Upsilon}$. Misspecification could consist in omission of relevant covariates (the working model (22) actually includes only some of the covariates in $\boldsymbol{X}$), or in omission of arcs in a Bayesian Network, or in a parametric misspecification of the model.

Let $\widehat{\boldsymbol{\beta}}_n$ be the MLE of $\boldsymbol{\beta}$. From White (1982), Th. 2.2, it is not difficult to see that, under mild regularity conditions on the model (22) (they are essentially Wald's conditions for consistency of MLEs), as $n$ increases $\widehat{\boldsymbol{\beta}}_n$ converges a.s. to

$$\boldsymbol{\beta}^* = \text{argmin}\left\{E\left[p_1(\boldsymbol{x})\log\frac{p_1(\boldsymbol{x})}{p_1(\boldsymbol{x};\boldsymbol{\beta})} + (1-p_1(\boldsymbol{x}))\log\frac{1-p_1(\boldsymbol{x})}{1-p_1(\boldsymbol{x};\boldsymbol{\beta})}\right]\right\}.$$

In other words, $p_1(\boldsymbol{x};\boldsymbol{\beta}^*)$ minimizes the (average) Kullback–Leibler divergence of the adopted probability distribution for the propensity score from the true one.

Consider now the Hájek and Horvitz–Thompson estimators of $\theta_k$ when the propensity score model is uncorrectly specified, namely

$$\widehat{\theta}_k^H = \frac{1}{\sum_{i=1}^n I_{(T_i=k)} p_k(\boldsymbol{x}_i; \widehat{\boldsymbol{\beta}}_n)^{-1}} \sum_{i=1}^n I_{(Y_i=1)} I_{(T_i=k)} p_k(\boldsymbol{x}_i; \widehat{\boldsymbol{\beta}}_n)^{-1}$$

$$\widehat{\theta}_k^{HT} = \frac{1}{n} \sum_{i=1}^n I_{(Y_i=1)} I_{(T_i=k)} p_k(\boldsymbol{x}_i; \widehat{\boldsymbol{\beta}}_n)^{-1}$$

Their limiting behavior, in terms of convergence in probability, is studied in Proposition 5.

**Proposition 5** Suppose that $p_k(\boldsymbol{x}; \boldsymbol{\beta})$ is a continuous function of $\boldsymbol{\beta}$ for each fixed $\boldsymbol{x}$, that assumptions $A1$-$A3$ in White (1982) and $A1$, $A2$ are satisfied for the true propensity score $p_k^{TR}(\boldsymbol{x})$, and that there exists $\delta > 0$ such that $\delta \leq p_k(\boldsymbol{x}; \boldsymbol{\beta}^*) \leq 1 - \delta$ for each $\boldsymbol{x}$. Then, the following two statements hold

$$\widehat{\theta}_k^{HT} - \theta_k \xrightarrow{p} \quad E\left[\theta_k(\boldsymbol{X})\left(\frac{p_k^{TR}(\boldsymbol{X})}{p_k(\boldsymbol{X}; \boldsymbol{\beta}^*)} - 1\right)\right] \quad \text{as } n \to \infty; \tag{23}$$

$$\widehat{\theta}_k^{H} - \theta_k \xrightarrow{p} \quad \frac{1}{E\left[\frac{p_k^{TR}(\boldsymbol{X})}{p_k(\boldsymbol{X}; \boldsymbol{\beta}^*)}\right]} E\left[(\theta_k(\boldsymbol{X}) - \theta_k)\left(\frac{p_k^{TR}(\boldsymbol{X})}{p_k(\boldsymbol{X}; \boldsymbol{\beta}^*)} - 1\right)\right] \quad \text{as } n \to \infty \tag{24}$$

for $k = 0, 1$.

The comparison of (23) and (24) makes it evident that $\widehat{\theta}_k^{HT}$ and $\widehat{\theta}_k^{H}$ are asymptotically equivalent only when the model for propensity score is correctly specified. In case of misspecification, their asymptotic behavior is different. From (23) it appears that the limit in probability of $\widehat{\theta}_k^{HT} - \theta_k$ is small *only* when $p_k(\boldsymbol{x}; \boldsymbol{\beta}^*)$ is close to $p_k^{TR}(\boldsymbol{x})$, *i.e.* when misspecification is negligible. However, (24) shows that the limit in probability of $\widehat{\theta}_k^{H} - \theta_k$ is small if *either* $p_k(\boldsymbol{x}; \boldsymbol{\beta}^*)$ is close to $p_k^{TR}(\boldsymbol{x})$ *or* $|\theta_k(\boldsymbol{x}) - \theta_k|$ is small, namely when $\theta_k(\boldsymbol{x})$ does not vary too much around $\theta_k$. In addition, from (24) and taking into account that $E[\theta_k(\boldsymbol{X})] = \theta_k$, the inequality

$$\left| \frac{1}{E\left[\frac{p_k^{TR}(\boldsymbol{X})}{p_k(\boldsymbol{X}; \boldsymbol{\beta}^*)}\right]} E\left[(\theta_k(\boldsymbol{X}) - \theta_k)\left(\frac{p_k^{TR}(\boldsymbol{X})}{p_k(\boldsymbol{X}; \boldsymbol{\beta}^*)} - 1\right)\right] \right| \leq \sup_{\boldsymbol{x}} |\theta_k(\boldsymbol{x}) - \theta_k|$$

is obtained. Hence, the limit in probability of $|\widehat{\theta}_k^{H} - \theta_k|$ is bounded. The same does not happen for the Horvitz–Thompson estimator $\widehat{\theta}_k^{HT}$, because the expectation

$$E\left[\theta_k(\boldsymbol{X})\left(\frac{p_k^{TR}(\boldsymbol{X})}{p_k(\boldsymbol{X}; \boldsymbol{\beta}^*)} - 1\right)\right]$$

is unbounded. This shows that $\widehat{\theta}_k^{H}$ is preferable, being less prone to misspecification of propensity scores model if compared to $\widehat{\theta}_k^{HT}$.

## 5 Application

The proposed approach is applied here to a clinical context. In particular, data from 6478 prostate cancer patients who underwent radical prostatectomy at San Raffaele Hospital (Milan, Italy) are considered. The goal is to evaluate whether receiving neoadjuvant hormonal therapy (NEOadjHT) before radical prostatectomy has an effect on the decision to perform lymphadenectomy during the surgery. The decisions on administered therapies, and in particular on neoadjuvant hormonal therapy, are based only on patient characteristics. In detail, the variables influencing decisions on neoadjuvant hormonal therapy are age, body mass index (BMI), Charlson Comorbidity Index (CCI), biopsy Gleason score (bxgg), clinical stage (clinstage), and total PSA (tpsa). Pre-treatment covariates also affect the potential outcome (lymphadenectomy), of course under the unconfoundedness assumption. Since categorical or categorized covariates are commonly used for decision-making in clinical practice, we have categorized all the quantitative covariates using clinical cut-offs in this study. This is actually a major reason to use BNs to estimate propensity scores: they should mimic clinical decisions, which are essentially based on discrete (or discretized) covariates.

The Bayesian networks (BNs) displayed in Fig. 1 were estimated on real prostate cancer data using the Tabu greedy search (TABU) algorithm with AIC and BIC score functions, respectively. Specific structural constraints (blacklists) were imposed to ensure that the relationships modeled by the BN are consistent with the temporal and clinical context of the variables. In particular, the treatment variable (NEOadjHT) has been constrained so that it cannot influence the other covariates, as they are pre-treatment characteristics. Therefore, NEOadjHT is a childless node and its Markov blanket coincides with its parent set. In addition, age was constrained to remain unaffected by other variables, and in its turn, BMI was constrained to remain unaffected by any variable other than age. In both cases, the estimated BN supports the presence of interaction terms among some covariates for the estimation of the propensity score. Therefore, in this real setting, the use of BN should provide an advantage for a good estimation of the propensity score and subsequently for the ATE estimation.

Figure 2 presents the ATE estimates, along with 95% confidence intervals, using the Hájek-type (15) and the Horvitz–Thompson-type (16) estimators. The propensity score was estimated by a Bayesian network based on either AIC or BIC score functions. All methods gave confidence intervals that did not include zero. Hence, there is sufficient evidence to reject the hypothesis that neoadjuvant hormonal therapy before radical prostatectomy did not have an effect on the decision to perform lymphadenectomy during the surgery. In this application, the differences in the ATE estimate among the different methodologies were negligible. To evaluate the goodness of the proposed approach and to delve deeper into the impact on the estimation of ATE of different PS estimations combined with the usage of the Hájek-type (15) and the Horvitz–Thompson-type (16) estimators, a simulation study has been designed in Sect. 6. It mimics the characteristics of this real dataset while varying the sample size.

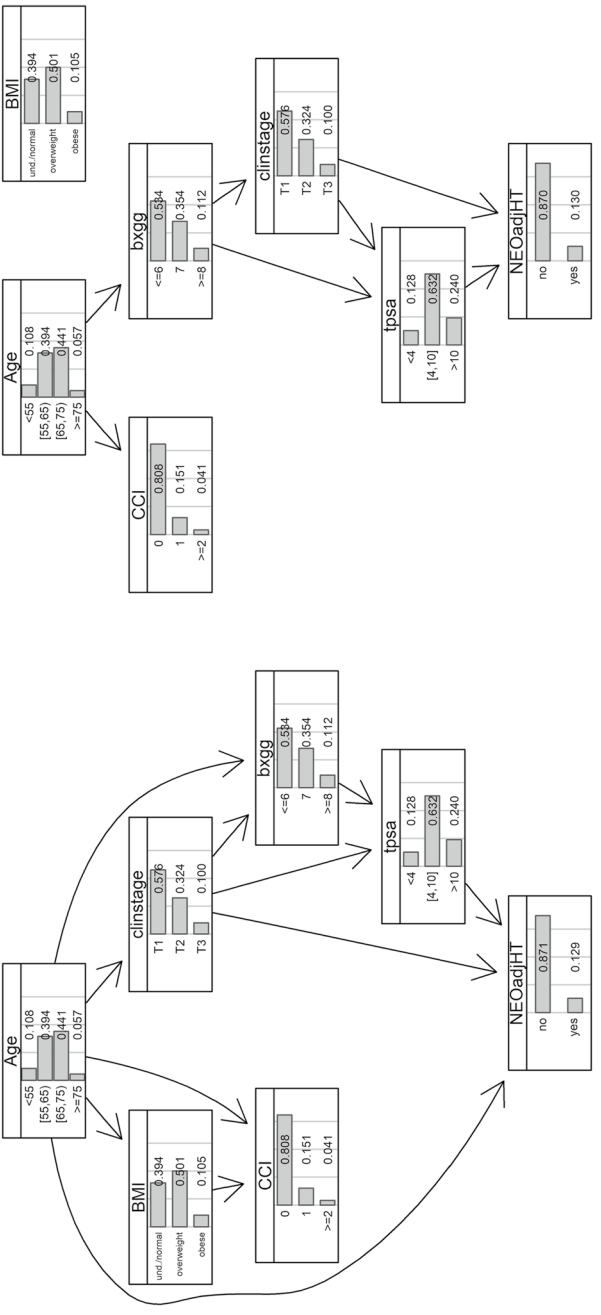


**Fig. 1** Bayesian networks obtained on the real prostate cancer data through the Tabu greedy search (TABU) algorithm with the AIC (on the left hand side) and BIC (on the right hand side) score functions. The variables in the data are: Body Mass Index (BMI), Charlson Comorbidity Index (CCI), biopsy Gleason score (bxgg), clinical stage (clinstage) and total PSA (tpsa)

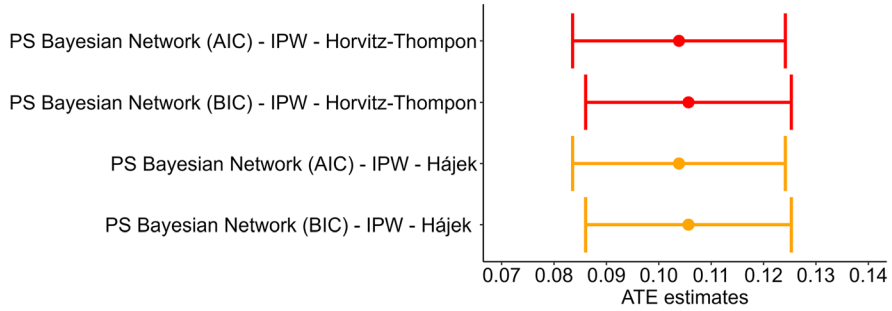


**Fig. 2** ATE estimates and 95% confidence intervals (bars)

## 6 Empirical evidence from data mimicking prostate cancer real data

In this section, we use artificial data that mimic the characteristics of a real dataset of prostate cancer patients, described in the previous section. Our main goal is to provide empirical evidence with finite sample sizes $n$ of the theoretical asymptotic results in Sects. 3 and 4.

### 6.1 Simulation study plan

The simulated data were generated in order to mimic the prostate cancer real data of Sect. 5. Treatment assignment and covariates have been generated based on the estimated BN by using the Tabu greedy search (TABU) algorithm with AIC score function (Fig. 1 left).

The binary potential outcomes $Y_{(0)}$ and $Y_{(1)}$ were then simulated through logistic models including the whole set of covariates. Again, values of the model coefficients were equal to the observed coefficients estimated on the prostate cancer data. In detail, if $B(p)$ denotes the Bernoulli distribution with parameter (success ptobability) $p$,

$$Y_{(k)}|X_c \sim B(P(Y_{(k)} = 1|\boldsymbol{X}_c)) \quad \text{for } k = 0, 1,$$

where $Y_{(k)} = 1$ if lymphadenectomy is performed and $Y_{(k)} = 0$ otherwise, for $k = 0, 1$, $\boldsymbol{X}_c$ denotes the vector of covariates, and

$$logit\big(P(Y_{(0)} = 1|\boldsymbol{X}_c)\big) = \alpha_0 + \boldsymbol{\beta}^T\boldsymbol{X}_c,$$
$$logit\big(P(Y_{(1)} = 1|\boldsymbol{X}_c)\big) = \alpha_0 + \alpha_1 + \boldsymbol{\beta}^T\boldsymbol{X}_c.$$

In order to compute the true ATE for the evaluation of the estimation methodologies, the probabilities $\theta_k = P(Y_{(k)} = 1)$, $k = 0, 1$, were obtained by marginalizing $P(Y_{(k)} = 1|\boldsymbol{X}_c)$ over $\boldsymbol{X}_c$. Fixing the parameters $\alpha_0$, $\alpha_1$ and $\boldsymbol{\beta}^T$ based on the real data, the corresponding ATE is equal to 0.094. The outcome $Y$ was generated according to Eq. (2), with treatment $T$ referring to the neoadjuvant hormonal therapy: $T = 1$ if the therapy was administered, $T = 0$ otherwise. In this set up mimicking real data, four

sample sizes ($n = 500, 1000, 2500, 5000$) we considered, with 1000 Monte Carlo (MC) runs for each sample size.

To estimate the propensity score we used Bayesian Networks learned with the Tabu greedy search algorithm with either AIC score function (BN AIC) or BIC score function (BN BIC). The Bayesian Networks were estimated by considering the same structural constraints (blacklists) defined in Sect. 5 to ensure consistency with the characteristics of the involved variables. As a comparison, the propensity score has been also estimated *via* logistic regression with BIC-based backward variable selection (BLR BIC), applied to the same set of candidate covariates. The estimated propensity score was then used to obtain the estimate of ATE and the related 95% confidence interval (CI) through two inverse probability weighting estimators for $\theta_k$: the Horvitz–Thompson estimator $\widehat{\theta}_k^{HT}$ in (16) and the Hájek estimator $\widehat{\theta}_k^{H}$ in (15).

### 6.2 Results and discussion

In Fig. 3, we present box-plots that summarize the differences in bias among the ATE estimators. The closer the median bias over simulations to zero, the higher the estimation accuracy. The lower the variability of the distribution, the higher the estimator efficiency. As far as the Horvitz–Thompson estimator is concerned, the ATE estimator's bias using the propensity score estimated by BN BIC is generally closer to zero than the bias obtained through BN AIC. In this latter case, we have less precision and a tendency to underestimate the ATE. In case of Hájek estimator, both BN AIC and BN BIC exhibit a good performance. Increasing the sample size, uniformly improves estimation.

Table 1 shows two additional evaluation metrics for increasing sample sizes: (1) the empirical coverage (EC) of 95% CIs, computed as the proportion of CIs in Eq. (21) that include the true ATE (0.094) within 1000 runs and (2) the empirical rejection rate (ERR) of the test for no treatment effect specified in Eq. (20) at the $\alpha = 0.05$ level, computed as the proportion of CIs in Eq. (21) not including zero. Clearly, the closer *EC* to the nominal level 0.95, the closer the significance level of the corresponding test to the nominal significance level 0.05. On the other hand, the higher the *ERR* term, the higher the power of the test, correspondingly to a real significance level $1 - EC$.

In case of Horvitz–Thompson estimator, BN BIC gives better results than BN AIC in terms of both empirical coverage and its closeness to the nominal level. With the Hájek estimator of ATE, the largest coverage is obtained by using the BN AIC propensity score, with a tendency to provide conservative CIs (i.e., coverage larger than 95%). The coverage obtained by using the BN BIC approach is the closest to the nominal 95%.

When propensity scores are estimated *via* logistic regression with BIC-based backward selection (BLR BIC), the real coverage probability is generally smaller than the nominal level 0.95. A comparison between BN BIC and BLR BIC shows that the first one gives considerably more accurate confidence intervals, especially for a sample size $n \leq 1000$. If ATE estimates are used to test hypotheses on treatment effect, this also means that the test based on BN BIC is conservative (its real

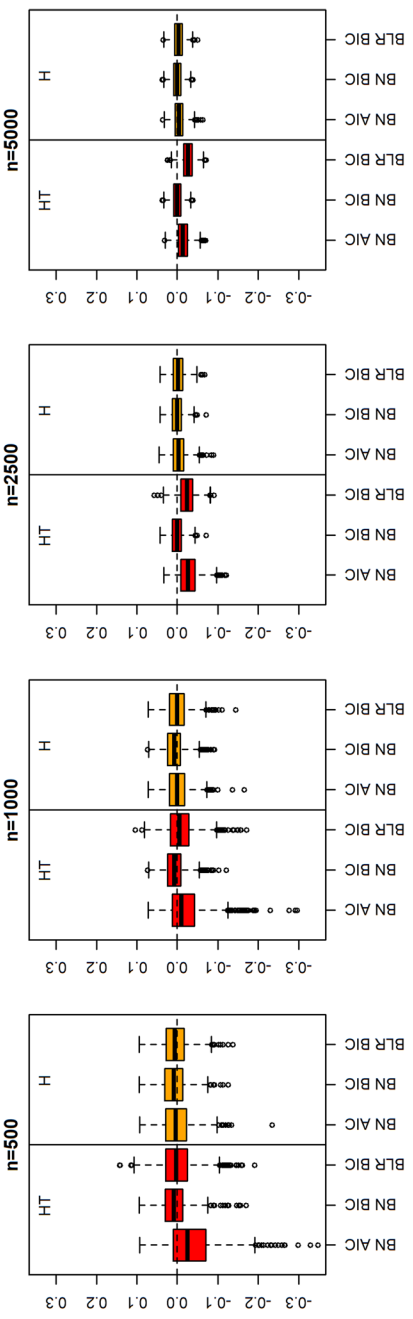


**Fig. 3** Bias of the ATE estimators obtained *via* BN AIC, BN BIS, BLR BIS for PS estimation on simulated data mimicking the real prostate cancer data. The bias is defined as the difference between the estimated and the true ATE. HT = Horvitz–Thompson estimator, H = Hájek estimator

**Table 1** Empirical coverage (EC) of CIs in Eq. (21) and empirical rejection rate (ERR) of no treatment effect test (20), for increasing sample size $n$, for the three PS estimation approaches (BN AIC, BN BIC, BLR BIC) and the two ATE estimators, by using the simulated data mimicking the real prostate cancer data

| n | Methods | Horvitz–Thompson | | Hájek | |
|---|---|---|---|---|---|
| | | EC | ERR | EC | ERR |
| 500 | BN AIC | 0.862 | 0.364 | 0.999 | 0.574 |
| 500 | BN BIC | 0.976 | 0.73 | 0.985 | 0.736 |
| 500 | BLR BIC | 0.836 | 0.705 | 0.879 | 0.738 |
| 1000 | BN AIC | 0.852 | 0.699 | 0.985 | 0.841 |
| 1000 | BN BIC | 0.972 | 0.932 | 0.975 | 0.933 |
| 1000 | BLR BIC | 0.842 | 0.819 | 0.902 | 0.877 |
| 2500 | BN AIC | 0.779 | 0.813 | 0.984 | 0.955 |
| 2500 | BN BIC | 0.973 | 0.999 | 0.975 | 0.999 |
| 2500 | BLR BIC | 0.686 | 0.94 | 0.952 | 0.995 |
| 5000 | BN AIC | 0.868 | 0.986 | 0.963 | 0.994 |
| 5000 | BN BIC | 0.946 | 1 | 0.945 | 1 |
| 5000 | BLR BIC | 0.428 | 0.995 | 0.934 | 1 |

significance level is smaller than the nominal 0.05), while the test based on BLR BIC is liberal (its real significance level is larger that the nominal 0.05). Furthermore, the *ERR* column shows that BN BIC offers a higher power than BLR BIC, even if the corresponding type I error probability is smaller.

With both types of ATE estimators, a larger proportion of rejection of the null hypothesis is achieved using the BN BIC rather than BN AIC propensity score. Moreover, this proportion increases to 1 as the sample size increases.

## 7 Conclusions

In this paper, a new method for estimating and testing the Average Treatment Effect (ATE) for binary outcomes using Bayesian Networks (BNs) for Propensity Score (PS) estimation is proposed. Both Horvitz–Thompson (HT) and Hájek (H) type estimators of ATE have been considered, the second being a modification of the first known to be more stable for small to moderate sample sizes. Asymptotic properties of the ATE estimators are derived by the statistical properties of BNs. The main conclusions of the present paper can be summarized as follows:

1. Estimating PS by BNs is particularly important when the dependence structure for treatments outcome and covariates is complex. BN structure can be learned directly from data (e.g., via structural learning) involving treatment assignment and covariates without necessarily imposing a priori functional relationships. Under the assumption of discrete covariates, BNs allow us to produce maximum likelihood (ML) estimators.
2. As a consequence of point 1, the use of BNs ensures the efficiency of both Horvitz–Thompson and Hájek estimators of ATE. It also allows us to define a test for the absence of treatment effect.

3. As theoretically explained in Sect. 4 and highlighted from our simulation study, the Hájek estimator seems recommendable over the more familiar Horvitz–Thompson estimator for inverse probability weighting to estimate the ATE, to produce interval estimates and for testing null treatment effect hypothesis.
4. As the sample size $n$ increases, BNs show a clear improvement, according to its "universal" consistency property given in Eq. (10) and discussed in Sect. 4
5. Since the estimation of propensity scores does not involve outcome data, the approach based on Bayesian Networks could be also used in case of continuous outcomes, provided that the pre-treatment covariates $\mathbf{X}$ are discrete. We do not go further into this direction, because it requires different, additional regularity conditions on outcomes distribution, that are far from the goals of the present paper.
6. As pointed out by a referee, including variables other than those influencing the treatment, and predictive for the treatment, generally improves the efficiency of ATE estimation; cfr. Rotnitzky and Smucler (2020) and Henckel et al. (2022). In the present paper, we have adopted a covariates selection criterion based on "control for all pre-treatment covariates", which may lead to a loss of efficiency in estimation. On the other hand, adopting the criterion in Rotnitzky and Smucler (2020) would require the construction of a full causal graph containing also the outcome variable, which could be a possible further development of the present paper. Furthermore, as already remarked, the estimator considered here is asymptotically efficient when (as in our case case) there are no additional variables other than pre-treatment covariates used for propensity score estimation.

## Appendix

**Proof of Proposition 1** Statement $(a)$ is obvious, in view of assumptions $A2$, $A3$. If the network structure is known *a priori*, statement $(b)$ is a consequence of well-known properties of MLEs; cfr. Davison (2003). If the network structure is learned from data through BIC, $(b)$ follows from the consistency of BIC criterion; cfr., in particular, Proposition 2.1 and Theorem 3.1 in Shibata (1986). Cfr. also Balov (2013), from which $(b)$ can be deduced as special case of no missing data. Statement $(c)$ is a consequence of Corollary 3.1 in Shibata (1986). Finally, statements $(d)$, $(e)$ follow from Davison (2003) and Le Cam (1953) and Le Cam (1960).

**Lemma 1** Under Assumptions A1, A3, the following results hold.

$$\sup_{\boldsymbol{x}} \left|\widehat{p}_k(\boldsymbol{x}) - p_k(\boldsymbol{x})\right| \xrightarrow{p} \quad 0 \ \text{ as } n \to \infty \ \ \forall\, k = 0,\ 1;$$

$$\sup_{\boldsymbol{x}} \left|\frac{1}{\widehat{p}_k(\boldsymbol{x})} - \frac{1}{p_k(\boldsymbol{x})}\right| \xrightarrow{p} \quad 0 \ \text{ as } n \to \infty \ \ \forall\, k = 0,\ 1.$$

**Proof** Immediate consequence of (10) and the finiteness of possible values of $\boldsymbol{x}$.

**Lemma 2** Under Assumptions A1–A3:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{I_{(T_i=k)}}{\widehat{p}_k(\boldsymbol{X}_i)}\xrightarrow{p}1 \quad \text{as } n\to\infty;\ \ k=0,\ 1. \tag{1}$$

**Proof** First of all, we may write

$$\frac{1}{n}\sum_{i=1}^{n}\frac{I_{(T_i=k)}}{\widehat{p}_k(\boldsymbol{X}_i)}=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\widehat{p}_k(\boldsymbol{X}_i)}-\frac{1}{p_k(\boldsymbol{X}_i)}\right)I_{(T_i=k)}+\frac{1}{n}\sum_{i=1}^{n}\frac{I_{(T_i=k)}}{p_k(\boldsymbol{X}_i)}. \tag{2}$$

In addition, since

$$E\left[\frac{I_{(T=k)}}{p_k(\boldsymbol{X})}\right]=E\left[\frac{1}{p_k(\boldsymbol{x})}E\left[I_{(T=k)}|\,\boldsymbol{X}=\boldsymbol{x}\right]\right]=1$$

from the Weak Law of Large Numbers (WLLN) it is seen that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{I_{(T_i=k)}}{p_k(\boldsymbol{X}_i)}\xrightarrow{p}1 \quad \text{as } n\to\infty.$$

Finally, observing that

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\widehat{p}_k(\boldsymbol{X}_i)}-\frac{1}{p_k(\boldsymbol{X}_i)}\right)I_{(T_i=k)}\right|\le\sup_{\boldsymbol{x}}\left|\frac{1}{\widehat{p}_k(\boldsymbol{x})}-\frac{1}{p_k(\boldsymbol{x})}\right|$$

conclusion (1) immediately follows from Lemma 1 and (2).

**Proof of Proposition 2** Define first the "pseudo-estimator" of $\theta_k$:

$$\widetilde{\theta}_k=\frac{1}{n}\sum_{i=1}^{n}\frac{I_{(T_i=k)}}{p_k(\boldsymbol{X}_i)}I_{(Y_i=1)},\quad k=0,\ 1.$$

From the WLLN, and using A2, it is immediate to see that

$$\begin{aligned}\widetilde{\theta}_k\xrightarrow{p}\ & E\left[\frac{I_{(T_i=k)}}{p_k(\boldsymbol{X}_i)}I_{(Y_i=1)}\right]=E\left[\frac{1}{p_k(\boldsymbol{x})}E\left[I_{(T=k)}|\,\boldsymbol{X}=\boldsymbol{x}\right]E\left[I_{(Y_{(k)}=1)}|\,\boldsymbol{X}=\boldsymbol{x}\right]\right]\\ =&\theta_k,\quad k=0,\ 1\end{aligned} \tag{3}$$

as $n\to\infty$.

The consistency of the Horvitz–Thompson estimator $\widehat{\theta}_k^{HT}(1)$ is then an immediate consequence of

$$\begin{aligned}\left|\widehat{\theta}_k^{HT}-\widetilde{\theta}_k\right|=&\left|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\widehat{p}_k(\boldsymbol{X}_i)}-\frac{1}{p_k(\boldsymbol{X}_i)}\right)I_{(T_i=k)}I_{(Y_i=1)}\right|\\ \le&\sup_{\boldsymbol{x}}\left|\frac{1}{\widehat{p}_k(\boldsymbol{x})}-\frac{1}{p_k(\boldsymbol{x})}\right|\end{aligned}$$

and Lemma 1.

As far as the Hájek-type estimator $\widehat{\theta}_k(1)$ is concerned, it can be proved as a consequence of the relationship

$$\widehat{\theta}_k - \widetilde{\theta}_k = \frac{1}{\frac{1}{n}\sum_{j=1}^{n} \frac{I_{(T_j=k)}}{\widehat{p}_k(\boldsymbol{X}_j)}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left( \frac{1}{\widehat{p}_k(\boldsymbol{X}_j)} - \frac{1}{p_k(\boldsymbol{X}_j)} \right) I_{(T_i=k)} I_{(Y_i=1)} \right\} + \left\{ \left( \frac{1}{n} \sum_{j=1}^{n} I_{(T_j=k)} \right)^{-1} - 1 \right\} \widetilde{\theta}_k.$$

and of Lemmas 1, 2 and (3).

**Proof of Proposition 3** Proof is in principle simple, apart a few complications in notation. Let us start with the HT-type estimator (16). Define

$$\theta_k(\boldsymbol{x}) = P(Y_i = 1 | T_i = k, \boldsymbol{X}_i = \boldsymbol{x}),\ k = 0,\ 1.$$

Taking into account that $\widehat{p}_k(\boldsymbol{x}) - p_k(\boldsymbol{x}) = o_p(n^{-1/2+\epsilon})$ for every positive $\epsilon$, and using the linearization technique in Hirano et al. (2003) and in Kim (2014) and Kim (2019), it is not difficult to see that the equality

$$\sqrt{n}(\widehat{\theta}_k^{HT} - \theta_k) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h_k^*(Y_1,\ T_i,\ \boldsymbol{X}_i) + o_p(1),\ k = 0,\ 1 \tag{4}$$

holds, where

$$h_k^*(Y_1,\ T_i,\ \boldsymbol{X}_i) = \left( \frac{1}{p_k(\boldsymbol{X}_i)} Y_i I_{(T_i=k)} - \theta_k \right) - \frac{\theta_k(\boldsymbol{X}_i)}{p_k(\boldsymbol{X}_i)} (I_{(T_i=k)} - p_k(\boldsymbol{X}_i)),\ k = 0,\ 1.$$

Next, observing that

$$E\left[h_k^*(Y_i,\ T_i,\ \boldsymbol{X}_i)\right] = E\left[ \frac{1}{p_k(\boldsymbol{X}_i)} y_i I_{(t_i=k)} \right] - \theta_k - E\left[ \frac{\theta_k(\boldsymbol{X}_i)}{p_k(\boldsymbol{X}_i)} (I_{(T_i=k)} - p_k(\boldsymbol{X}_i)) \right] = 0$$

and assuming that $V[h_k^*(Y_i,\ T_i,\ \boldsymbol{X}_i)] > 0$ for $k = 0,\ 1$, from (4) and the (bivariate) Central Limit Theorem, it follows that $\sqrt{n}(\widehat{\boldsymbol{\theta}}^{HT} - \boldsymbol{\theta})$ possesses Normal limiting distribution with zero expectation and variance-covariance matrix $\boldsymbol{\Sigma}$. This shows statements 1 and 3. Statement 2 can be proved by a technique similar to that of Proposition 2.

**Lemma 3** Suppose that $p_k(\boldsymbol{x}; \boldsymbol{\beta})$ is a continuous function of $\boldsymbol{\beta}$ for each fixed $\boldsymbol{x}$, that assumptions *A*1-*A*3 in White (1982) and *A*1, *A*2 are satisfied, and that there exists $\delta > 0$ such that $\delta \leq p_k(\boldsymbol{x}; \boldsymbol{\beta}^*) \leq 1 - \delta$ for each $\boldsymbol{x}$. The following two statements hold.

$$\sup_{\mathbf{x}} \left| p_k(\boldsymbol{x}; \widehat{\boldsymbol{\beta}}_n) - p_k(\boldsymbol{x}; \boldsymbol{\beta}^*) \right| \xrightarrow{p} 0 \text{ as } n \to \infty \ \forall k = 0, 1;$$

$$\sup_{x} \left| \frac{1}{p_k(\boldsymbol{x}; \widehat{\boldsymbol{\beta}}_n)} - \frac{1}{p_k(\boldsymbol{x}; \boldsymbol{\beta}^*)} \right| \xrightarrow{p} 0 \text{ as } n \to \infty \ \forall k = 0, 1.$$

**Proof** Similar to the proof of Lemma 1.

**Proof of Proposition 5** Similarly to Proposition 2 consider the pseudo-estimator

$$\widetilde{\theta}_k^{HT} = \frac{1}{n} \sum_{i=1}^{n} \frac{I_{(T_i=k)}}{p_k(\boldsymbol{X}_i; \boldsymbol{\beta}^*)} I_{(Y_i=1)}, \quad k = 0, 1.$$

By repeating *verbatim* the arguments of Proposition 2, it is not difficult to see that

$$\left| \widehat{\theta}_k^{HT} - \widetilde{\theta}_k^{HT} \right| \leq \sup_{x} \left| \frac{1}{p_k(\boldsymbol{x}; \widehat{\boldsymbol{\beta}}_n)} - \frac{1}{p_k(\boldsymbol{x}; \boldsymbol{\beta}^*)} \right| \xrightarrow{p} 0 \tag{5}$$

as $n \to \infty$, in view of Lemma 3. In the second place, from the Weak Law of Large Numbers, we also have

$$\begin{aligned} \widetilde{\theta}_k^{HT} &\xrightarrow{p} E\left[ \frac{1}{p_k(\boldsymbol{X}; \boldsymbol{\beta}^*)} I_{(Y=1)} I_{(T=k)} \right] \\ &= E\left[ \theta_k(1|\boldsymbol{X}) \frac{p_k(\boldsymbol{X})}{p_k(\boldsymbol{X}; \boldsymbol{\beta}^*)} \right] \end{aligned}$$

as $n$ increases, and hence, from (5),

$$\widehat{\theta}_k^{HT} \xrightarrow{p} E\left[ \theta_k(1|X) \frac{p_k(\boldsymbol{X})}{p_k(\boldsymbol{X}; \boldsymbol{\beta}^*)} \right] \text{ as } n \to \infty.$$

Taking now into account that $E[\theta_k(1|\boldsymbol{X})] = \theta_k$, relationship (23) is obtained.

As far as $\widehat{\theta}_k^H$ is concerned, using the above arguments it is immediate to see that

$$\frac{1}{n} \sum_{i=1}^{n} \frac{1}{p_k(\boldsymbol{X}_i; \boldsymbol{\beta}^*)} \xrightarrow{p} E\left[ \frac{p_k(\boldsymbol{X})}{p_k(\boldsymbol{X}; \boldsymbol{\beta}^*)} \right]$$

and hence, from (23),

$$\widehat{\theta}_k^H \xrightarrow{p} \frac{E\left[ \theta_k(1|X) \frac{p_k(\boldsymbol{X})}{p_k(\boldsymbol{X}; \boldsymbol{\beta}^*)} \right]}{E\left[ \frac{p_k(\boldsymbol{X})}{p_k(\boldsymbol{X}; \boldsymbol{\beta}^*)} \right]} \text{ as } n \to \infty$$

which is equivalent to (24).

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s10182-025-00535-4.

**Acknowledgements** Thanks are due to two anonymous referees, whose comments considerably improved an earlier version of the paper. Open access publishing permitted by Università degli Studi di Roma La Sapienza, as a part of the Springer - Sapienza agreement.



## References

Abadie, A.: Bootstrap tests for distributional treatment effects in instrumental variable models. J. Am. Stat. Assoc. **97**, 284–292 (2002)

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

Balov, N.: Consistent model selection of discrete Bayesian networks from incomplete data. Electron. J. Stat. **7**, 1935–7524 (2013)

Ben-Michael, E., Feller, A., Hirshberg, D.A., Zubizarreta, J.R.: The balancing act in causal inference (2021). arXiv:2110.14831

Chattopadhyay, A., Hase, C.H., Zubizarreta, J.R.: Balancing vs modeling approaches to weighting in practice. Stat. Med. **39**, 3227–3254 (2020)

Conti, P.L., De Giovanni, L.: Testing for the presence of treatment effect under selection on observables. AStA Adv. Stat. Anal. (2022). https://doi.org/10.1007/s10182-022-00454-8

Cowell, R.G., Dawid, A.P., Lauritzen, S.L., Spiegelhalter, D.J.: Probabilistic Networks and Expert Systems. Springer, New York (1999)

Davison, A.C.: Statistical Models. Cambridge University Press, Cambridge (2003)

Ding, P.: A paradox from randomization-based causal inference. Stat. Sci. **32**, 331–345 (2017)

Donald, S.G., Hsu, Y.C.: Estimation and inference for distribution functions and quantile functions in treatment effect models. J. Econom. **178**, 383–397 (2014)

Drton, M., Maathuis, M.H.: Structure learning in graphical modeling. Annu. Rev. Stat. Appl. **4**, 365–393 (2017)

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

Henckel, L., Perković, E., Maathuis, M.H.: Graphical criteria for efficient total effect estimation via adjustment in causal linear models. J. R. Stat. Soc. Ser. B Stat Methodol. **84**, 579–599 (2022)

Hernán, M.A., Robins, J.M.: Estimating causal effects from epidemiological data. J. Epidemiol. Community Health **60**, 578–586 (2006)

Hirano, K., Imbens, G.W., Ridder, G.: Efficient estimation of average treatment effects using the estimated propensity score. Econometrica **71**, 1161–1189 (2003)

Imbens, G.W., Rubin, D.B.: Causal Inference for Statistics, Social, and Biomedical Sciences. Cambridge University Press, Cambridge (2015)

Imbens, G.W., Wooldridge, J.M.: Recent developments in the econometrics of program evaluation. J. Econ. Lit. **47**, 5–86 (2009)

Kim, K.I.: An alternative efficient estimation of average treatment effects. J. Mark. Econ. **42**, 1–41 (2014)

Kim, K.I.: Efficiency of average treatment effect estimation when the true propensity is parametric. Econometrics **7**, 2–25 (2019). https://doi.org/10.3390/econometrics7020025

Lauritzen, S.L.: Graphical Models. Oxford University Press, Oxford (1996)
Le Cam, L.: On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. Univ. Calif. Publ. Stat. **1**, 277–330 (1953)
Le Cam, L.: Locally asymptotically normal families of distributions. Univ. Calif. Publ. Stat. **3**, 27–98 (1960)
Lu, J., Zhang, Y., Ding, P.: Sharp bounds on the relative treatment effect for ordinal outcomes. Biometrics (2019). https://doi.org/10.1111/biom.13148
Lunceford, J.K., Davidian, M.: Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Stat. Med. **23**, 2937–2960 (2004)
Nandy, P., Hauser, A., Maathuis, M.H.: High-dimensional consistency in score-based and hybrid structure learning. Ann. Stat. **46**, 3151–3183 (2018)
Rotnitzky, A., Smucler, E.: Efficient adjustment sets for population average causal treatment effect estimation in graphical models. J. Mach. Learn. Res. **21**, 1–86 (2020)
Shibata, R.: Consistency of model selection and parameter estimation. J. Appl. Probab. **23**, 127–141 (1986)
White, H.: Maximum likelihood estimation of misspecified models. Econometrica **50**, 1–25 (1982)
Wu, J., Ding, P.: Randomization tests for weak null hypotheses in randomized experiments (2018). arXiv:1809.07419 [stat.ME]
Zubizzarreta, J.R.: Stable weights that balance covariates for estimation with incomplete outcome data. J. Am. Stat. Assoc. **110**, 910–922 (2015)



## Authors and Affiliations

**Paola Vicard[1] · Paola Maria Vittoria Rancoita[2] · Federica Cugnata[2] · Alberto Briganti[2] · Fulvia Mecatti[3] · Clelia Di Serio[2] · Pier Luigi Conti[4]**

✉ Pier Luigi Conti
pierluigi.conti@uniroma1.it

Paola Vicard
paola.vicard@uniroma3.it

Paola Maria Vittoria Rancoita
rancoita.paolamaria@unisr.it

Federica Cugnata
cugnata.federica@unisr.it

Alberto Briganti
briganti.alberto@hsr.it

Fulvia Mecatti
fulvia.mecatti@unimib.it

Clelia Di Serio
clelia.diserio@unisr.it

[1] Roma Tre University, Rome, Italy

[2] Vita-Salute San Raffaele University, Milan, Italy

[3] University of Milano-Bicocca, Milan, Italy

[4] Sapienza University of Rome, Rome, Italy