Covariate selection for the non-parametric estimation of an average treatment effect

Running headline: Covariate selection for non-parametric estimation of a treatment effect

Xavier de Luna[a], Thomas S. Richardson[b] and Ingeborg Waernbaum[c]

[a]Department of Statistics, Umeå University, SE-90187 Umeå, Sweden

and
The Institute of Labour Market Policy Evaluation,
Box 513, SE-75120 Uppsala, Sweden
xavier.deluna@stat.umu.se

[b]Department of Statistics, University of Washington,
Box 354322, WA 98195-4322 Seattle, USA
thomasr@u.washington.edu

[c]Department of Statistics, Umeå University, SE-90187 Umeå, Sweden
ingeborg.waernbaum@stat.umu.se

June 19, 2008

Abstract

Observational studies where the effect of a non-randomized treatment on an outcome of interest is estimated are common, for instance, in labor economics and epidemiology. Such studies often rely on an assumption of unconfounded treatment when controlling for a given set of observed pre-treatment covariates. Which covariates must be controlled for in order to guarantee unconfoundedness should primarily be based on subject matter grounds. However, theory gives typically only partial guidance. It is tempting to include many covariates in the controlling set to try to make realistic the assumption of an unconfounded treatment. In this paper we show that including unnecessary covariates is suboptimal when the effect of a binary treatment is estimated non-parametrically. For instance, when using $\sqrt{n}$-consistent estimator a loss of efficiency may result from using covariates that are irrelevant for the unconfoundedness assumption. Moreover, bias may dominate the variance when many covariates are used. Embracing the Neyman-Rubin model typically used in conjunction with non-parametric estimators of treatment effects, we characterize subsets from the original reservoir of covariates, which are sufficient to condition on for the treatment assignment to be unconfounded. We define a set of covariates as minimal when the treatment ceases to be unconfounded given any proper subset of the minimal set of covariates. We provide a general sufficient condition to show minimality of certain subsets of covariates that are defined in the paper. These subsets of covariates are, moreover, shown to be identified under mild assumptions. Finally, we study with simulations the finite sample properties of a matching and an inverse probability weighting estimator based on data-driven selected sets covariates.

Key words: Graphical models, minimal sets of covariates, marginal coordinate hypothesis, Neyman-Rubin model, unconfoundedness.

1 Introduction

The context of this paper is an observational study of a treatment effect on an outcome of interest. Numerous such studies are performed in economics and epidemiology. In observational studies a set of characteristics (covariates in the sequel) not affected by the treatment assignment is observed for the study participants. We call the treatment assignment unconfounded ("exogenous", "ignorable" or "selected on observables" are also used in the literature) if it is independent of the potential outcomes resulting from assignment to treatment or no treatment when conditioning on a set of observed covariates. The concept of unconfoundedness/exogeneity has been defined in different ways in the literature, see, e.g., de Luna and Johansson (2006) and the references therein. We consider throughout a binary treatment assignment which is unconfounded, and estimators of average treatment effects which are non-parametric in the sense that they do not make distributional and functional form assumptions; see, e.g., Rubin (1974), Rosenbaum and Rubin (1983), Rosenbaum (2002) and the review by Imbens (2004). Unconfoundedness of the treatment assignment may be considered to be realistic when a rich set of covariates is available. On the other hand, non-parametric estimators based on a high dimensional covariate vector may have poor performance (e.g., Hahn, 2004, Abadie and Imbens, 2006). Thus, the reduction of the initial covariate set dimension is an important practical issue. Despite the popularity of non-parametric estimators of average treatment effects, there is little theoretical work on dimension reduction of the covariate set to be used with non-parametric estimators. This is to be contrasted with an extremely large statistical literature on the identification and selection of optimal covariates sets to be used in parametric regression models. The articles by Abadie and Imbens (2006), Hahn (2004) and Heckman, Ichimura, and Todd (1998) are notable exceptions. Abadie and Imbens showed that the asymptotic behavior of a specific type of estimator based on matching depends on the number of continuous covariates utilized. In particular, an asymptotically non-ignorable bias appears when using two or more such covariates. When non-parametric estimators are $\sqrt{n}$-consistent, Hahn (2004) showed that the number of covariates utilized has consequences on the semi-parametric efficiency bound. More precisely, including covariates which affect exclusively either the outcome or the treatment may yield (depending on the parameter of interest) a larger asymptotic variance, see also Heckman, Ichimura, and Todd (1998).

In this paper, we provide new theoretical results on the important issue of covariate dimension reduction. We first provide results on which dimension reduction can be achieved in general situations. Embracing the Neyman-Rubin model (Neyman, 1923, Rubin, 1974, Holland, 1986) typically used in conjunction with non-parametric estimators of treatment effects, we characterize subsets from the original reservoir of covariates, which are sufficient to condition on for the treatment assignment to be unconfounded. We show that dimension reduction is essential for the optimality of nonparametric estimators by using the results obtained by Abadie and Imbens (2006) and Hahn (1998, 2004). We further define a set of covariates as *minimal* when the treatment ceases to be unconfounded given any proper subset of the minimal set of covariates. We provide a general sufficient condition for subsets of covariates to be minimal, and show that, under mild assumptions, the minimal subsets of covariates defined are identified. These theoretical results are completed with a set of simulation experiments. We use a model-free procedure proposed by Cook (2004) to select minimal sets covariates. Based on such a data-driven covariate selection we estimate the average treatment effect with matching and inverse probability weighting. The finite sample properties of the combined procedure (model-free covariate selection and non-parametric estimation of treatment effect) are thus investigated. We observe that estimators based on selected sets of covariates often outperform (in terms of mean squared error) estimators using larger sets. In particular, our results suggest that dimension reduction is indeed an essential issue for the performance of the estimators studied, and that using the covariates sets that affect treatment assignment (the common practice) is not optimal.

Section 2 briefly introduces the Neyman-Rubin model and estimators of average treatment effects. Section 3 and 4 presents our theoretical results on dimension reduction. The simulation experiments are described and commented in Section 5. Section 6 concludes.

2 Assumptions and estimators

2.1 The Neyman-Rubin model

The potential outcome framework was introduced by Neyman (1923) in the context of randomized trials. It was then adapted by Rubin (1974) to deal with observational studies, where the treatment is not randomized. We consider studies where the purpose is to estimate the effect of a binary treatment T ($T = 0$ when not treated, and $T = 1$ when treated) on an outcome of interest. Each individual/unit in the study has an observed treatment assignment T. Moreover, two (potential) outcome variables are defined: the outcome when not treated, Y_0, and the outcome when treated, Y_1. It is assumed that only one of them is observed for each unit, i.e. each unit is either treated or not treated. The estimand of interest is in practice often an expected/average value of the difference between the potential outcomes: the average treatment effect, $E(Y_1 - Y_0)$ and/or the average treatment effect on the treated $E(Y_1 - Y_0 \mid T = 1)$ being often used in applications.

Within this framework, when the treatment assignment is randomized, we have $\Pr(Y_0, Y_1|T = 1) = \Pr(Y_0, Y_1)$; see Holland (1986) for a detailed presentation and discussion. That is the treatment assignment is independent of the potential outcomes, denoted herein

$$Y_1, Y_0 \perp\!\!\!\perp T. \tag{1}$$

Thus, because (1) implies $E(Y_1 - Y_0) = E(Y_1 \mid T = 1) - E(Y_0 \mid T = 0)$, an unbiased estimate of the average treatment effect can be obtained by taking the difference between the sample average of the outcomes for the treated and the sample average of the outcomes for the controls (untreated).

In observational studies where treatment is not randomized, (1) typically does not hold. Instead, further modelling assumptions are made. Let $\mathbf{X}$ be a set of pre-treatment characteristics, called covariates in the sequel, and assume that these characteristics are observed for all the units in the study. Let us also consider a distribution $P(Y_0, Y_1, T|\mathbf{X})$ with well defined densities as needed below for the conditional independence statements and the expectation operators.

Alternatively, one may only consider the two marginals $P(Y_0, T|\mathbf{X})$ and $P(Y_1, T|\mathbf{X})$. Then, we use the following assumptions

(A.1) $Y_0 \perp\!\!\!\perp T|\mathbf{X}$, (A.2) $\Pr(T = 1 \mid \mathbf{X}) < 1$,
(A.3) $Y_1 \perp\!\!\!\perp T|\mathbf{X}$, (A.4) $\Pr(T = 0 \mid \mathbf{X}) < 1$.

Here, $Y_j \perp\!\!\!\perp T|\mathbf{X}$, $j = 0, 1$ denotes that the treatment assignment T and the potential outcome Y_j are independent given $\mathbf{X}$; see Dawid (1979) for a general reference on conditional independence. Instead of (A.1) and (A.3), it is sometimes necessary to assume the stronger assumption

(A.[1&3]) $(Y_0, Y_1) \perp\!\!\!\perp T|\mathbf{X}$,

see Section 2.2. The estimand of interest may be estimated by noting that, if (A.1-A.2) hold then $E(Y_0 \mid \mathbf{X}) = E(Y_0 \mid \mathbf{X}, T = 0)$. Hence, we have for the average treatment effect on the treated

$$\gamma = E(Y_1 - Y_0|T = 1) = E(Y_1 \mid T = 1) - E(E(Y_0 \mid \mathbf{X}, T = 0) \mid T = 1). \qquad (2)$$

Moreover, if (A.3-A.4) hold then $E(Y_1 \mid \mathbf{X}) = E(Y_1 \mid \mathbf{X}, T = 1)$, and, hence, assuming (A.1-A.4) we have for the average treatment effect

$$\beta = E(Y_1 - Y_0) = E(E(Y_1 \mid \mathbf{X}, T = 1) - E(Y_0 \mid \mathbf{X}, T = 0)). \qquad (3)$$

Equations (2) and (3) show that unbiased estimation of β and γ is possible with the data at hand, because Y_j is observed given $T = j$ for $j = 0, 1$.

2.2 Non-parametric estimators and their properties

Many estimators of average treatment effects have been proposed in the literature that make no distributional or functional form assumptions, see, e.g., Rubin (1991) and the recent review by Imbens (2004). In the next section we will use asymptotic results to highlight the importance of reducing the dimension of the covariate vector. Thus, it may be relevant here to consider a taxonomy of non-parametric estimators based on their asymptotic properties. For the sake of conciseness we refrain from explicitly defining these estimators but refer the reader to the precise references provided. We distinguish estimators which are $\sqrt{n}$-consistent, where n denotes the sample size, from those which are not. Also, among the $\sqrt{n}$-consistent estimators, we also distinguish those that are asymptotically efficient. Many estimators proposed and used in the literature are $\sqrt{n}$-consistent, including the double robust estimator of Robins and

Rotnitzky (1995, Thm 1), and the kernel-based matching estimator of Heckman, Ichimura, and Todd (1998, Thm 2). However, matching estimators may not yield $\sqrt{n}$-consistency as was shown in Abadie and Imbens (2006, Thm 1). Estimators shown to be asymptotically efficient include the imputation estimator of Hahn (1998, Thm 6), the inverse probability weighting estimator of Hirano, Imbens, and Ridder (2003, Thm 1), and the imputation estimator of Imbens, Newey, and Ridder (2003, Thm 3.1). The last theorem also shows that a non-parametric version of the double robust estimator and the two imputation estimators mentioned above are asymptotically equivalent. Estimators based on matching on covariates are not asymptotically efficient in general, see Abadie and Imbens (2006, Thm 5).

A final remark here is that in order to obtain semiparametric efficiency bounds for non-parametric estimators (Hahn, 1998), (A.[1&3]) is assumed. Therefore, the theory developed in the sequel will consider both cases, assuming only (A.1) and (A.3) and the situation where the stronger condition (A.[1&3]) holds.

3 Reducing the dimension of the set of covariates

In this section we discuss and build upon some important theoretical results by Abadie and Imbens (2006) and Hahn (1998, 2004). However, we start by giving some preliminary definitions and results on the possibility to reduce the dimension of $\mathbf{X}$.

3.1 Preliminary results

We are first reducing the covariate set by removing variables conditionally independent of T given the rest, and then further removing variables conditionally independent of the potential outcomes (Y_0 and/or Y_1) given the remainder.

Let us define $\mathbf{X}_T \subseteq \mathbf{X}$ of minimum cardinality such that $p(T|\mathbf{X}) = p(T|\mathbf{X}_T)$, or equivalently $T \perp\!\!\!\perp \mathbf{X} \setminus \mathbf{X}_T | \mathbf{X}_T$. Then, under assumptions (A.2) and (A.4), $\mathbf{X}_T$ is uniquely defined (Pearl, 1988, p. 119). Moreover, we have the following result.

Proposition 1 *Under assumptions (A.1-A.4), we have* $Y_0 \perp\!\!\!\perp T|\mathbf{X}_T$ *and* $Y_1 \perp\!\!\!\perp T|\mathbf{X}_T$. *Moreover, assuming also (A.[1&3]), then* $(Y_0, Y_1) \perp\!\!\!\perp T|\mathbf{X}_T$.

Proof. Rewrite (A.1) as $Y_0 \perp\!\!\!\perp T|\mathbf{X} \setminus \mathbf{X}_T, \mathbf{X}_T$. Also, by definition, $T \perp\!\!\!\perp \mathbf{X} \setminus \mathbf{X}_T|\mathbf{X}_T$.

Hence, by Lemma 4.3 in Dawid (1979), $(Y_0, \mathbf{X} \setminus \mathbf{X}_T) \perp\!\!\!\perp T | \mathbf{X}_T$, implying $Y_0 \perp\!\!\!\perp T | \mathbf{X}_T$. ■

We further define the sets $\mathbf{Q}_0, \mathbf{Q}_1 \subseteq \mathbf{X}_T$ of minimum cardinality such that $p(Y_j|\mathbf{X}_T) = p(Y_j|\mathbf{Q}_j)$, for $j = 0, 1$. Similarly we define $\mathbf{Q} \subseteq \mathbf{X}_T$ of minimum cardinality such that $p(Y_0, Y_1|\mathbf{X}_T) = p(Y_0, Y_1|\mathbf{Q})$. Then, under assumption

(A.5) the density function $p(Y_0, Y_1|\mathbf{X})$ is positive: $p(Y_0, Y_1|\mathbf{X}) > 0$,

$\mathbf{Q}_j$, $j = 0, 1$, and $\mathbf{Q}$ are uniquely defined (Pearl, 1988, p. 119). The following result also holds.

Proposition 2 *Under assumptions (A.1-A.5), we have* $Y_0 \perp\!\!\!\perp T|\mathbf{Q}_0$ *and* $Y_1 \perp\!\!\!\perp T|\mathbf{Q}_1$. *Moreover, assuming also (A.[1&3]), then* $(Y_0, Y_1) \perp\!\!\!\perp T|\mathbf{Q}$.

Proof. Use Lemma 4.3 in Dawid (1979) as in Proposition 1. ■

Example 3 *Consider the situation with* $\mathbf{Q}_0 = \mathbf{Q}_1 = \mathbf{Q}$, $\mathbf{Q} \subset \mathbf{X}_T \subset \mathbf{X}$ *and*

$$Y_0 = g_0(\mathbf{Q}) + \varepsilon_0, \qquad Y_1 = g_1(\mathbf{Q}) + \varepsilon_1, \qquad T = g_T(\mathbf{X}_T) + \varepsilon_T,$$

where $\varepsilon_0, \varepsilon_1$ *and* ε_T *are independent of each other and* $\mathbf{X}$. *Denoting by* $Y = (1 - T)Y_0 + TY_1$, *the observed outcome, we have*

$$Y = g_0(\mathbf{Q}) + T(g_1(\mathbf{Q}) - g_0(\mathbf{Q})) + \varepsilon_Y,$$

where $\varepsilon_Y = (1 - T)\varepsilon_0 + T\varepsilon_1$. *Thus, we have that* $\mathbf{Q}$ *and* $\mathbf{X}_T \setminus \mathbf{Q}$ *affect* T, *and* $\mathbf{Q}$ *and* T *affect* Y.

We should often have situations similar to Example 3, i.e. where, for $j = 0, 1$, $\mathbf{Q}_j \subset \mathbf{X}_T \subset \mathbf{X}$. In such cases the dimension of the conditioning/matching set may be reduced by using $\mathbf{Q}$ instead of $\mathbf{X}_T$.

Alternatively we may reverse the order by first reducing the covariate set by removing variables independent of potential outcomes given the rest, and then further removing variables conditionally independent of T.

We define $\mathbf{X}_0$, $\mathbf{X}_1$ and $\mathbf{X}_Y$ of minimum cardinality such that $p(Y_0|\mathbf{X}) = p(Y_0|\mathbf{X}_0)$, $p(Y_1|\mathbf{X}) = p(Y_1|\mathbf{X}_1)$ and $p(Y_0, Y_1|\mathbf{X}) = p(Y_0, Y_1|\mathbf{X}_Y)$ respectively. These covariates sets are uniquely defined under assumption (A.5), see again Pearl (1988, p. 119).

Proposition 4 *Under assumptions (A.1-A.5), we have* $Y_0 \perp\!\!\!\perp T|\mathbf{X}_0$ *and* $Y_1 \perp\!\!\!\perp T|\mathbf{X}_1$. *Moreover, assuming also (A.[1&3]), then* $(Y_0, Y_1) \perp\!\!\!\perp T|\mathbf{X}_Y$.

Proof. Use Lemma 4.3 in Dawid (1979) as in Proposition 1. ■

Further, define $\mathbf{Z}_j \subseteq \mathbf{X}_j$, for $j = 0, 1$, and $\mathbf{Z} \subseteq \mathbf{X}_Y$ of minimum cardinality such that $p(T|\mathbf{X}_j) = p(T|\mathbf{Z}_j)$, and $p(T|\mathbf{X}_Y) = p(T|\mathbf{Z})$ respectively. These covariates sets are uniquely defined under assumption (A.2) and (A.4).

Proposition 5 *Under assumptions (A.1-A.5), we have* $Y_0 \perp\!\!\!\perp T|\mathbf{Z}_0$ *and* $Y_1 \perp\!\!\!\perp T|\mathbf{Z}_1$. *Moreover, assuming also (A.[1&3]), then* $(Y_0, Y_1) \perp\!\!\!\perp T|\mathbf{Z}$.

Proof. Use Lemma 4.3 in Dawid (1979) as in Proposition 1. ■

In Example 3 we have $\mathbf{X}_0 = \mathbf{X}_1 = \mathbf{X}_Y = \mathbf{Z}_0 = \mathbf{Z}_1 = \mathbf{Q}$. Another, different, particular case is given in Example 6.

Example 6 *Consider the situation with* $\mathbf{Z}_0 = \mathbf{Z}_1 = \mathbf{Z}$, $\mathbf{Z} \subset \mathbf{X}_Y \subset \mathbf{X}$ *and*

$$Y_0 = g_0(\mathbf{X}_Y) + \varepsilon_0, \qquad Y_1 = g_1(\mathbf{X}_Y) + \varepsilon_1, \qquad T = g_T(\mathbf{Z}) + \varepsilon_T,$$

where $\varepsilon_0, \varepsilon_1$ *and* ε_T *are independent of each other and* $\mathbf{X}$. *For the observed outcome* Y, *we have*

$$Y = g_0(\mathbf{X}_Y) + T(g_1(\mathbf{X}_Y) - g_0(\mathbf{X}_Y)) + \varepsilon_Y,$$

where $\varepsilon_Y = (1 - T)\varepsilon_0 + T\varepsilon_1$. *Thus, we have that* $\mathbf{Z}$ *affects* T, *and* $\mathbf{X}_Y$, $\mathbf{X}_Y \setminus \mathbf{Z}$ *and* T *affect* Y.

In general, and in contrast with the two examples given above, we can also expect to have situations such that $\mathbf{Z} \subset \mathbf{X}_Y$,$\mathbf{Q} \subset \mathbf{X}_T$, with $\mathbf{Z} \not\subset \mathbf{Q}$ and/or $\mathbf{Q} \not\subset \mathbf{Z}$, see the examples of Figure 1.

3.2 Asymptotic bias

The large sample properties of a simple matching estimator were studied in detail in Abadie and Imbens (2006). In particular, they show that, for a simple matching estimator $\hat{\beta}$ (defined in Abadie and Imbens, 2006, Equation (3)), $\sqrt{N}(\hat{\beta} - \beta) = A_N + B_N$, where $A_N = O_p(1)$ with a normal limiting distribution, and $B_N = O_p(N^{1/2-1/k})$ (a conditional bias term), where k is the number of continuous valued covariates and N is the sample size (Abadie and Imbens, 2006, Thm 1). When estimating the average treatment effect on the treated, γ, with a simple matching estimator, the same conditional bias is of order $O_p(N_1^{1/2-r/k})$, where N_1 is the number of treated individuals in the sample, and $r \leq 1$, where it is assumed that $N_1^r/(N - N_1) \to M$, $0 < M < \infty$ (Abadie and Imbens, 2006, Thm 2).

These results demonstrate that the bias term cannot be ignored asymptotically unless $k = 1$ for the simple matching estimator of β, and unless $r > k/2$ when estimating γ. This problem is an example of what is often called the "curse of the dimensionality", see Bellman (1961). As noted by Abadie and Imbens (2006, Footnote 1) discrete covariates can be matched exactly if there are a few of them, otherwise the impossibility of finding exact matches will also introduce non-ignorable bias.

3.3 Asymptotic efficiency

When non-parametric estimators are $\sqrt{n}$-consistent (e.g., Hahn, 1998, Hirano, Imbens, and Ridder, 2003), the asymptotic lower bounds that can be achieved for their variance depend on the dimension of the set of covariates utilized, as can be shown by using the results in Hahn (1998, 2004).

Proposition 7 *Assume (A.2), (A.4), (A.5) and (A.[1&3]), and consider* $\mathbf{X}_T$ *and* $\mathbf{Q}$ *as defined above with the exclusion restriction:*

(B.1) $\mathbf{Q} \subset \mathbf{X}_T$.

Then the semi-parametric efficiency bounds for the estimation of β and γ using the knowledge that (B.1) holds are lower than the bounds not using this knowledge.

Proof. Under case (C.1) $p(Y_0, Y_1|\mathbf{X}_T) = p(Y_0, Y_1|\mathbf{Q})$ and the results stated in Hahn (2004, Sec. III) can be directly applied to prove the proposition. ■

Note that estimators achieving the asymptotic efficiency bound have been proposed by Hahn (1998) and Hirano, Imbens, and Ridder (2003). Such estimators are efficient when using $\mathbf{Q}$. Note also that Heckman, Ichimura, and Todd (1998, Sec. 7) show that $\mathbf{Q}$ yields the lowest achievable variance when using their kernel-based matching estimator.

Proposition 8 *Assume (A.2), (A.4), (A.5) and (A.[1&3]), and consider* $\mathbf{X}_Y$ *and* $\mathbf{Z}$ *as defined above with the exclusion restriction:*

(B.2) $\mathbf{Z} \subset \mathbf{X}_Y$.

Then the semi-parametric efficiency bound for the estimation of γ *using the knowledge that (B.2) holds is lower than the bound not using this knowledge. For the estimation of* β*, the bounds with and without the knowledge on (B.2) are equal.*

Proof. Under case (C.3) $p(T|\mathbf{X}_Y) = p(T|\mathbf{Z})$ and, hence, the proposition is a direct consequence of Theorems 1 and 2 in Hahn (2004). ■

Under (C.3) and when estimating γ, the estimators of Hahn (1998) and Hirano, Imbens, and Ridder (2003) can be modified in order for them to achieve the semi-parametric asymptotic efficiency bound, see Hahn (2004, Prop. 1).

4 Identification of minimal sets of covariates

A consequence of the results presented in the previous section is that, when the relevant covariate sets are known, the performance of non-parametric estimators of β and γ can be optimized by using as few covariates as possible: the bias of simple matching estimators dominates their asymptotic distribution when matching for too many covariates, and $\sqrt{n}$-consistent estimators gain in efficiency when not using unnecessary covariates. In this section we present results on the identifiability of minimal sets of covariates. We start by defining a concept of minimality.

Definition 9 *Let three disjoint subsets of random variables* $\mathbf{A}$*,* $\mathbf{B}$ *and* $\mathbf{C}$*, the latter possibly empty, be such that* $\mathbf{A} \perp\!\!\!\perp \mathbf{B}|\mathbf{C}$*, then* $\mathbf{C}$ *is said to be a minimal conditioning set for the independence of* $\mathbf{A}$ *and* $\mathbf{B}$*, if no proper subset* $\mathbf{C}^- \subset \mathbf{C}$ *is such that* $\mathbf{A} \perp\!\!\!\perp \mathbf{B}|\mathbf{C}^-$.

Without further assumptions on the joint distribution $P(Y_1, Y_0, T, \mathbf{X})$, the sets $\mathbf{Q}_j$, $j = 0, 1$, and $\mathbf{Q}$ as defined in Section 3.1 are not necessarily minimal for the conditional independence of Y_0 and T, Y_1 and T, and (Y_0, Y_1) and T, respectively. The same is true for $\mathbf{Z}_j$, $j = 0, 1$, and $\mathbf{Z}$. This is illustrated with an example in the next section (Figure 1). In order to obtain conditions guaranteeing minimality of the set of covariates defined in this paper, we need to impose some more structure on the joint distribution of the random variables involved. This is conveniently done by using graphical models (Lauritzen, 1996), and we thus start with a brief introduction to their related theory.

4.1 Graphical Markov properties and independence models

A directed graph, $\mathcal{G} = (\mathbf{V}, E)$, is a set of vertices (variables) $\mathbf{V}$ and directed edges ($\rightarrow$) E. In the graph $A \rightarrow B$, A is said to be a *parent* of B. The set of parents of a vertex B in a graph $\mathcal{G}$ is denoted $pa_{\mathcal{G}}(B)$.

A *path* in a graph is a sequence of consecutive edges of any directionality. A *directed path* from A to B is a path of the form $A \rightarrow \cdots \rightarrow B$ on which every arrowhead is pointing towards B. A vertex A is an *ancestor* of B, if and only if either there is a directed path from A to B or $A = B$. The set of ancestors of a vertex B in a graph $\mathcal{G}$ is denoted $an_{\mathcal{G}}(B)$. If E is such that there is no directed cycle (directed path from one vertex to itself) then we have a directed acyclic graph (DAG).

A non-endpoint vertex Z on a path is a *collider* on the path if the edges preceding and succeeding Z have an arrowhead at Z, $\rightarrow Z \leftarrow$. If the non-endpoint vertex is not a collider it is a *noncollider*. A path from A to B in a DAG $\mathcal{G}$ is said to be *d-connecting* given a set of nodes $\mathbf{C}$ possibly empty ($A, B \notin \mathbf{C}$) if and only if:

i) every noncollider on the path is not in $\mathbf{C}$, and

ii) every collider on the path is an ancestor of $\mathbf{C}$.

If there is no path d-connecting any vertex in $\mathbf{A}$ to any other vertex in $\mathbf{B}$ given $\mathbf{C}$ then $\mathbf{A}$ and $\mathbf{B}$ are said to be *d-separated* given $\mathbf{C}$ (or $\mathbf{C}$ d-separates $\mathbf{A}$ and $\mathbf{B}$). For instance, in the DAG represented in Figure 1(a), X_1 is a collider on the path $Y_0 \leftarrow X_3 \rightarrow X_1 \leftarrow T$ so this path does not d-connect given $\emptyset$. However, Y_0 and T are d-connected given $\emptyset$ by the path $Y_0 \leftarrow X_2 \rightarrow T$. Similarly, Y_0 and T are d-separated given X_2. A DAG, $\mathcal{G}$, together with the d-separation criterion define an *independence model* $\mathcal{I}_d(\mathcal{G})$, the set of triples $\langle \mathbf{A}, \mathbf{B} \mid \mathbf{C} \rangle$, where $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$ are three disjoint subsets of $\mathbf{V}$, $\mathbf{C}$ possibly empty, such that $\mathbf{C}$ d-separates $\mathbf{A}$ and $\mathbf{B}$.

A distribution is said to be *compatible* with a DAG, $\mathcal{G}$, (or is said to obey the Markov property for $\mathcal{G}$) if it can be factored into a product of conditional densities for each variable in $\mathbf{V}$ given its parents. For instance, a distribution P compatible

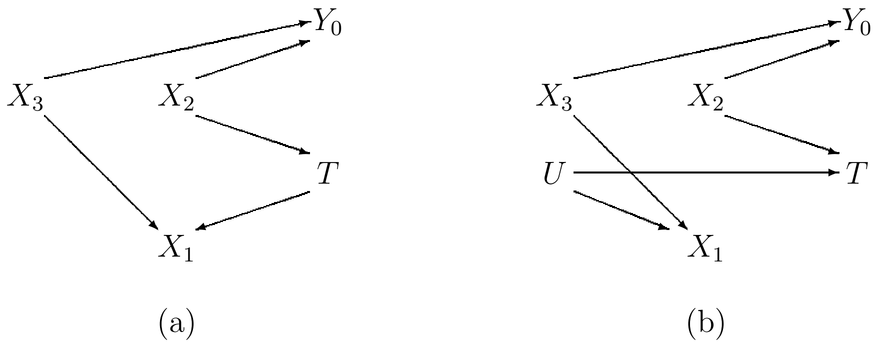

Figure 1: Two examples where $\mathbf{Q}_0 = (X_2, X_3)$ is not minimal.

with the DAG of Figure 1(a) can be factored as follows:

$$p(y_0, t, x_1, x_2, x_3) = p(y_0|x_2, x_3)p(t|x_2)p(x_1|t, x_3)p(x_2)p(x_3).$$

Two different DAGs may define the same set of compatible distributions, e.g. the three graphs $X \to Y \to Z$, $X \leftarrow Y \to Z$, and $X \leftarrow Y \leftarrow Z$, define the same set of compatible distributions.

We have that $\langle \mathbf{A}, \mathbf{B} \mid \mathbf{C} \rangle \in \mathcal{I}_d(\mathcal{G})$ implies $\mathbf{A} \perp\!\!\!\perp \mathbf{B}|\mathbf{C}[P]$ for any P compatible with $\mathcal{G}$. Finally, P (compatible with $\mathcal{G}$) is said to be *faithful* to $\mathcal{G}$ if also $\mathbf{A} \perp\!\!\!\perp \mathbf{B}|\mathbf{C}[P]$ implies $\langle \mathbf{A}, \mathbf{B} \mid \mathbf{C} \rangle \in \mathcal{I}_d(\mathcal{G})$. In particular if a distribution P is faithful to a DAG then the latter summarizes all the (conditional) independence statements holding under P. It may be shown that "almost all" distributions that are compatible with a given DAG are faithful (for smooth measures over the set of distributions); see Spirtes, Glymour, and Scheines (2000, Sec. 3.5) and Meek (1995). DAGs are hence a powerful graphical tool to both discuss and impose conditions on the conditional independence statements holding for a given set of distributions.

As illustrative examples consider the two DAGs given in Figure 1, together with a faithful distribution for the variables (vertices) involved. Then, for panel (a), we have, for $\mathbf{X} = (X_1, X_2, X_3)$, that $\mathbf{X}$ d-separates Y_0 and T, or equivalently $Y_0 \perp\!\!\!\perp T|\mathbf{X}$. Moreover, also using the d-separation criterion we see that $\mathbf{X}_T = \mathbf{X}$ and $\mathbf{Q}_0 = (X_2, X_3)$ for the sets defined in Section 3.1. Note that here $Y_0 \perp\!\!\!\perp T|X_2$, that is $\mathbf{Q}_0$ is not a minimal conditioning set for the independence of Y_0 and T. It may be argued against this example on the grounds that there is an arrow pointing from T to X_1, which contradicts the usual assumptions in observational studies that $\mathbf{X}$ should be restricted to pre-treatment variables. This does not happen in panel (b) of Figure

1. Consider here $\mathbf{X} = (X_1, X_2, X_3)$, that is U is a latent (unobserved) variable. Then, $Y_0 \perp\!\!\!\perp T|\mathbf{X}$. Here, as in the previous case, $\mathbf{X}_T = \mathbf{X}$ and $\mathbf{Q}_0 = (X_2, X_3)$. On the other hand, $Y_0 \perp\!\!\!\perp T|X_2$ and $\mathbf{Q}_0$ is not a minimal conditioning set for the independence of Y_0 and T, thereby illustrating again that $\mathbf{Q}_0$ as defined in Section 3.1 does not need to be minimal. Note here that the graphs drawn cannot have a causal interpretation because of the presence of a potential outcome Y_0, although one may wish to give causal interpretation to subgraphs, for instance, where Y_0 and its adjacent edges are deleted. The causal interpretation of DAGs (e.g., Pearl, 2000, Lauritzen, 2001, Richardson and Spirtes, 2003, Dawid, 2002) is, however, out of the scope of this paper.

4.2 Results on minimality

We show in the sequel that the following assumption is sufficient for the sets $\mathbf{Q}_j$ and $\mathbf{Z}_j$, $j = 0, 1$ to be minimal conditioning sets for the independence of Y_j and T.

(A.6) For $j = 0, 1$, the distribution P_j over the variables $(Y_j, T, \mathbf{X}, \mathbf{U})$, $\mathbf{U}$ unobserved, is faithful to a DAG $\mathcal{G}_j$. This DAG $\mathcal{G}_j$ is such that:

i) $\forall R \in \mathbf{X}$, if R and T are d-connected given every subset of $\mathbf{X}$ then R is an ancestor of T in $\mathcal{G}_j$;

ii) $\forall R \in \mathbf{X}$, if R and Y_j are d-connected given every subset of $\mathbf{X}$ then R is an ancestor of Y_j in $\mathcal{G}_j$.

Similarly we show below that the following condition is sufficient for the minimality of $\mathbf{Q}$ and $\mathbf{Z}$ for the independence of (Y_0, Y_1) and T. For that purpose, we consider only one graph where (Y_0, Y_1) is represented as a single node.

(A.7) The distribution P over the variables $(Y_0, Y_1, T, \mathbf{X}, \mathbf{U})$, $\mathbf{U}$ unobserved, is faithful to a DAG $\mathcal{G}$, where (Y_0, Y_1) is a single node in the graph. This DAG $\mathcal{G}$ is such that:

i) $\forall R \in \mathbf{X}$, if R and T are d-connected given every subset of $\mathbf{X}$ then R is an ancestor of T in $\mathcal{G}$;

ii) $\forall R \in \mathbf{X}$, if R and (Y_0, Y_1) are d-connected given every subset of $\mathbf{X}$ then R is an ancestor of (Y_0, Y_1) in $\mathcal{G}$.

Figure 1 displays the two basic structures prohibited by these assumptions. The first structure (Panel (a)) is a situation where a covariate is affected by the treatment assignment. The second prohibited structure (Panel (b)) occurs when a covariate (X_1) is not directly related with the treatment assignment, while the effect of X_1 on T is confounded by an unobserved variable (U). While the first structure can, generally, be avoided in practice by considering only covariates that are measured before treatment assignment, it is more difficult to have control on the second.

A primary motivation for assuming faithfulness is a presupposition that there is an underlying structural DAG $\mathcal{G}$ relating the variables. Since, as stated earlier, almost all distributions compatible with a given DAG are faithful, it may be argued to be reasonable to assume faithfulness. However, if the variables are deterministically related then faithfulness will likely be violated.

Proposition 10 *Assume (A.1-A.6). Then,* $\mathbf{Q}_j$ *is a minimal conditioning set for the independence of* T *and* Y_j, $j = 0, 1$. *Moreover, when replacing (A.6) with (A.7) then* $\mathbf{Q}$ *is a minimal conditioning set for the independence of* T *and* (Y_0, Y_1).

The proof is given in the appendix.

Proposition 11 *Assume (A.1-A.6). Then,* $\mathbf{Z}_j$*, is a minimal conditioning set for the independence of* T *and* Y_j, $j = 0, 1$. *Moreover, when replacing (A.6) with (A.7) then* $\mathbf{Z}$ *is a minimal conditioning set for the independence of* T *and* (Y_0, Y_1).

The proof is given in the appendix.

4.3 Identification

We call a set of variables $\mathbf{C}$ identified if it fulfills a condition that can be checked by using data. This motivates the following definition of identifiability, because tests of conditional independence statements (Su and White, 2003) may be used to retrieve an identified set of observed variables.

Definition 12 *Let a random variable A and two subsets of variables $\mathbf{B}$ and $\mathbf{C}$ be such that $A \notin \mathbf{B}$ and $\mathbf{C} \subseteq \mathbf{B}$, then we say that $\mathbf{C}$ is identified (in $\mathbf{B}$ for A) if it is the unique minimal conditioning set for the independence of A and $\mathbf{B} \setminus \mathbf{C}$, and the random variables involved are observed for all the individuals.*

Thus, by definition $\mathbf{X}_T$ is identified, since it is the unique minimal set such that $T \perp\!\!\!\perp \mathbf{X} \setminus \mathbf{X}_T | \mathbf{X}_T$ with all involved random variables observed on all units. On the other hand, e.g., $\mathbf{Q}_0$ was defined as the minimal set such that $Y_0 \perp\!\!\!\perp \mathbf{X}_T \setminus \mathbf{Q}_0 | \mathbf{Q}_0$, but Y_0 is not observed for all units. The identification of such sets of covariates can be retrieved under weak assumptions.

Proposition 13 (Identification of $\mathbf{Q}_j$) *Assume (A.1), (A.3) and $P(Y_1, Y_0, T, \mathbf{X}) > 0$, then, for $j = 0, 1$, $\mathbf{Q}_j$ is the unique minimal conditioning set for the independence of Y_j and $\mathbf{X}_T \setminus \mathbf{Q}_j$, given that $T = j$.*

Proof. By Propositions 1 and 2 we can write $P(Y_j|\mathbf{X}_T, T = j) = P(Y_j|\mathbf{X}_T) = P(Y_j|\mathbf{Q}_j) = P(Y_j|\mathbf{Q}_j, T = j)$ which gives the desired property $Y_j \perp\!\!\!\perp \mathbf{X}_T \setminus \mathbf{Q}_j | \mathbf{Q}_j, T = j$. Now it remains to show that $\mathbf{Q}_j$ is unique minimal for the latter conditional independence statement. We start by looking at the minimality property and suppose for a contradiction that there exists $\mathbf{Q}_j^- \subset \mathbf{Q}_j$ such that $Y_j \perp\!\!\!\perp \mathbf{X}_T \setminus \mathbf{Q}_j^- | \mathbf{Q}_j^-, T = j$ holds. We have that $Y_j \perp\!\!\!\perp \mathbf{X}_T \setminus \mathbf{Q}_j^- | \mathbf{Q}_j^-, T = j$ and $Y_j \perp\!\!\!\perp T | \mathbf{X}_T \setminus \mathbf{Q}_j^-, \mathbf{Q}_j^-$ (holds by Proposition 1) implies that $Y_j \perp\!\!\!\perp (T, \mathbf{X}_T \setminus \mathbf{Q}_j^-) | \mathbf{Q}_j^-$ (by (3.6e) in Pearl, 1988, and positivity). Hence, $Y_j \perp\!\!\!\perp \mathbf{X}_T \setminus \mathbf{Q}_j^- | \mathbf{Q}_j^-$ (by Lemma 4.2 in Dawid, 1979). Given that $\mathbf{Q}_j$ is of minimum cardinality such that $P(Y_j|\mathbf{X}_T) = P(Y_j|\mathbf{Q}_j)$, we have a contradiction and therefore $\mathbf{Q}_j$ is minimal. For uniqueness, suppose for a contradiction that there exists a set $\mathbf{Q}_j^* \neq \mathbf{Q}_j$ and of same cardinality than $\mathbf{Q}_j$. Then, as for $\mathbf{Q}_j^-$, we have that $P(Y_j|\mathbf{X}_T) = P(Y_j|\mathbf{Q}_j^*)$ which is in contradiction with the fact that $\mathbf{Q}_j$ was uniquely defined as the set of minimum cardinality such that $P(Y_j|\mathbf{X}_T) = P(Y_j|\mathbf{Q}_j)$ (Pearl, 1988, p. 119); see Section 3.1. ■

The identification of $\mathbf{Z}_j$ is a direct consequence of the following result.

Proposition 14 (Identification of $\mathbf{Z}_j$) *Assume (A.1), (A.3) and $P(Y_1, Y_0, T, \mathbf{X}) > 0$, then, for $j = 0, 1$, $\mathbf{X}_j$ is the unique minimal conditioning set for the independence of Y_j and $\mathbf{X} \setminus \mathbf{X}_j$, given $T = j$.*

The proof is similar to the one given for Proposition 13.

The identification of the minimal sets of covariates obtained in Section 4.2 is a direct consequence of the latter results. Under the assumptions of Propositions 10 and 13, $\mathbf{Q}_j$, $j = 0, 1$, is an identifiable minimal conditioning set for the independence of T and Y_j, $j = 0, 1$ respectively. Under the assumptions of Propositions 11 and 14, $\mathbf{Z}_j$, $j = 0, 1$, is an identifiable minimal conditioning set for the independence of T and Y_j.

Given the above results, asymptotic identification of the sets $\mathbf{Q}_j$ and $\mathbf{Z}_j$ may be obtained by using a consistent test of conditional independence. Consistent tests typically exist under the faithfulness assumption, see Robins, Scheines, Spirtes, and Wasserman (2003, Thm 1-2). In the latter paper it is also shown that no uniform consistent test exists without further assumptions. Uniform consistency is of interest because it yields a worst case error bound for a finite sample size. Zhang and Spirtes (2003) proposed a stronger version of the faithfulness assumption which is sufficient (and almost necessary) for the existence of uniformly consistent tests of conditional independence statements.

5 Simulation study

We perform simulations to study the finite sample performances that can be achieved by reducing the dimension of covariate set when estimating average treatment effects non-parametrically.

5.1 Data generating mechanisms

We generate values for ten covariates $\mathbf{X} = (X_1, X_2, \ldots, X_{10})$ from a multivariate standard normal distribution. All covariates are independent except for a partial correlation of 0.5 between X_7 and X_8. Further, a treatment assignment mechanism is specified as

$$\Pr(T = 1|\mathbf{X}) = (1 + \exp(-0.5X_1 - 0.5X_2 - 0.5X_3 - 0.5X_4 - 0.5X_7))^{-1}.$$

This model yields an expected number of treated equal to half the sample size.

We study three different data generating mechanism for the response. Two linear models, of which one with constant treatment effect,

$$Y|\mathbf{X} \sim N(2 + 2T + 2X_1 + 2X_2 + 2X_5 + 2X_6 + 2X_8, 1), \tag{4}$$

and one whose treatment effect is function of X_8,

$$Y|\mathbf{X} \sim N(2 + 2T + 8TX_8 + 2X_1 + 2X_2 + 2X_5 + 2X_6 + 2X_7 + 2X_8, 1), \tag{5}$$

and a heteroskedastic model with constant treatment effect,

$$Y|\mathbf{X} \sim N(2 + 2T + 2X_1 + 2X_5 + 2X_6 + 2X_8, \{2.5(2 + X_2)\}^2). \tag{6}$$

By design we have for model (4) and (6) that $\mathbf{X}_T = (X_1, X_2, X_3, X_4, X_7)$, $\mathbf{X}_Y = \mathbf{X}_0 = \mathbf{X}_1 = (X_1, X_2, X_5, X_6, X_8)$, $\mathbf{Q} = \mathbf{Q}_0 = \mathbf{Q}_1 = (X_1, X_2, X_7)$, $\mathbf{Z} = \mathbf{Z}_0 = \mathbf{Z}_1 = (X_1, X_2, X_8)$, while for model (5) we can write that $\mathbf{X}_T = (X_1, X_2, X_3, X_4, X_7)$, $\mathbf{X}_Y = \mathbf{X}_0 = \mathbf{X}_1 = (X_1, X_2, X_5, X_6, X_7, X_8)$, $\mathbf{Q} = \mathbf{Q}_0 = \mathbf{Q}_1 = (X_1, X_2, X_7)$, $\mathbf{Z} = \mathbf{Z}_0 = \mathbf{Z}_1 = (X_1, X_2, X_7)$.

Further, model (5) is such that the asymptotic semi-parametric efficiency bound for γ is stricty lower when using the knowledge that $\mathbf{Z} \subset \mathbf{X}_Y$ than when not using it. This can be shown by deriving the bounds of Theorems 1 and 2 in Hahn (2004) for this model.

For each of the models (4), (5) and (6) we generate 1000 random samples of size $n = 500$ and 1000. All data generation and computations are performed with the software R (R Development Core Team, 2007).

5.2 Model-free covariate selection

We start by studying the covariate selection stage. Estimation of the average treatment effects is left to the next subsection. Table 1 describes two two-steps algorithms inspired by our theoretical results. Algorithm A yields $\hat{\mathbf{X}}_T$ in its first step, and $\hat{\mathbf{Q}}_j$, $j = 0, 1$, in its second step, while Algorithm B yields $\hat{\mathbf{X}}_j$, $j = 0, 1$, in its first step, and $\hat{\mathbf{Z}}_j$, $j = 0, 1$, in its second step.

Because the aim is to estimate an average treatment effect without making distributional or functional form assumptions, we use a model-free covariate selection method proposed by Cook (1998, 2004) in all stages of these algortithms. Backward

Table 1: Identification of covariate sets

	Algorithm A: Identification of $\mathbf{Q}_0$ and $\mathbf{Q}_1$
Step 1.	Identify $\mathbf{X}_T$ such that $(T \perp\!\!\!\perp \mathbf{X} \setminus \mathbf{X}_T \| \mathbf{X}_T)$ holds.
Step 2.	For $j = 0, 1$: Identify $\mathbf{Q}_j \subseteq \mathbf{X}_T$ such that $(Y_j \perp\!\!\!\perp \mathbf{X}_T \setminus \mathbf{Q}_j \| \mathbf{Q}_j, T = j)$ holds.
	Algorithm B: Identification of $\mathbf{Z}_0$ and $\mathbf{Z}_1$
For $j = 0, 1$:	
Step 1.	Identify $\mathbf{X}_j$ such that $(Y_j \perp\!\!\!\perp \mathbf{X} \setminus \mathbf{X}_j \| \mathbf{X}_j, T = j)$ holds.
Step 2.	Identify $\mathbf{Z}_j \subseteq \mathbf{X}_j$ such that $(T \perp\!\!\!\perp \mathbf{X}_j \setminus \mathbf{Z}_j \| \mathbf{Z}_j)$ holds.

elimination covariate selection is performed starting with the whole set of the covariates, and testing hypothesis of the type $T \perp\!\!\!\perp B|\mathbf{C}$ or $Y \perp\!\!\!\perp B|\mathbf{C}, T = j$, $j = 0, 1$, $B \in \mathbf{X}, \mathbf{C} \subset \mathbf{X}, B \notin \mathbf{C}$. For instance, to infer about the conditional independence statement $T \perp\!\!\!\perp B|\mathbf{C}$, Cook (2004) proposes to test the marginal coordinate hypothesis

$$P_{\mathcal{H}}\mathcal{S}_{T|\mathbf{C}} = \mathcal{O}_p \text{ versus } P_{\mathcal{H}}\mathcal{S}_{T|\mathbf{C}} \neq \mathcal{O}_p,$$

where $P_{(\cdot)}$ is the projection operator with respect to the standard inner product, $\mathcal{H}$ is the subspace of R^p corresponding to the coordinate B, $p = dim(\mathbf{C})$, $\mathcal{S}_{T|\mathbf{C}}$ is the central subspace, i.e. the intersection of all subspaces $\mathcal{S}$ of the covariate space, satisfying $T \perp\!\!\!\perp \mathbf{C}|P_{\mathcal{S}}\mathbf{C}$, and $\mathcal{O}_p$ indicates the origin in $\mathbf{R}^p$. This central subspace exists and is unique under mild conditions (Cook, 1996). A test statistic and its corresponding asymptotic distribution was deduced by Cook (2004, Eq.10), where the central subspace is estimated by sliced inverse regression (Li, 1991). A detailed description of the test used is given in Li, Cook, and Nachtsheim (2005, Sec. 2.2). We use the implementation `dr.step` and `drop1.dr` from the software R (R Development Core Team, 2007), library `dr`, and select the level of significance to 10%.

For comparison, we use also two widely used model-based covariate selection procedure, the Akaike information criterion (AIC, Akaike, 1973) and the Bayesian information criterion (BIC, Schwarz, 1978), fitting a linear logistic regression model for T and a linear regression model for $Y_j \mid T = j$. Thus, the model fitted are misspecified except for linear model (4). The backward elimination implementation `stepAIC`, library `MASS`, of AIC and BIC from the softare R (R Development Core Team, 2007) is used.

Table 2: Model (5) − The percentage of times the selected covariate sets are such that the unconfoundedness assumption holds given the selected sets.

		Algorithm A		Algorithm B	
n	Method	$\mathbf{X}_T \subseteq \hat{\mathbf{X}}_T$	$Y_0 \perp\!\!\!\perp T \mid \hat{\mathbf{Q}}_0$	$\mathbf{X}_0 \subseteq \hat{\mathbf{X}}_0$	$Y_0 \perp\!\!\!\perp T \mid \hat{\mathbf{Z}}_0$
500	AIC	99.7	99.9	100.0	99.8
	BIC	95.8	98.1	100.0	96.8
	MCH	99.6	99.8	100.0	99.7
1000	AIC	100.0	100.0	100.0	100.0
	BIC	100.0	100.0	100.0	100.0
	MCH	100.0	100.0	100.0	100.0

Table 3: Model (6) − The percentage of times the selected covariate sets are such that the unconfoundedness assumption holds given the selected sets.

		Algorithm A		Algorithm B	
n	Method	$\mathbf{X}_T \subseteq \hat{\mathbf{X}}_T$	$Y_0 \perp\!\!\!\perp T \mid \hat{\mathbf{Q}}_0$	$\mathbf{X}_0 \subseteq \hat{\mathbf{X}}_0$	$Y_0 \perp\!\!\!\perp T \mid \hat{\mathbf{Z}}_0$
500	AIC	99.7	19.1	25.5	22.2
	BIC	95.7	2.8	5.4	2.3
	MCH	99.3	79.5	98.5	80.5
1000	AIC	100.0	20.2	26.1	25.4
	BIC	100.0	2.4	4.3	3.7
	MCH	100.0	90.4	100.0	96.7

Results are summarized for $j = 0$ and models (5) and (6) in Tables 2 and 3, by giving the percentage of times the selected covariate set (in both steps of the algorithms) are d-separators, i.e. the unconfoundedness assumption holds by conditioning on the sets. Results for (4) are similar to those for (5) and are therefore omitted. As in Li, Cook, and Nachtsheim (2005) the model-free marginal coordinate hypothesis (MCH) procedure outperforms AIC and BIC when the latter use misspecified models, here in the case of model (6). Similar results were obtained for $j = 1$.

5.3 Covariate selection and estimation

In this section we study the consequences of covariate selection for the estimation of average treatment effects. We use MCH as described in the previous section to obtain selected covariates subsets. For estimation of the average treatment effects, we control/match for the sets $\mathbf{X}_T$, $\mathbf{Q}$, $\mathbf{X}_Y$, and $\mathbf{Z}$, as well as their selected version, $\hat{\mathbf{X}}_T$, $\hat{\mathbf{Q}} = \hat{\mathbf{Q}}_0 \cup \hat{\mathbf{Q}}_1$, $\hat{\mathbf{X}}_Y = \hat{\mathbf{X}}_0 \cup \hat{\mathbf{X}}_1$, and $\hat{\mathbf{Z}} = \hat{\mathbf{Z}}_0 \cup \hat{\mathbf{Z}}_1$.

Two estimators of average treatment effects are considered, a matching and an inverse probability weighting estimator. Treatment effects are estimated by one-to-one matching with replacement, i.e. where each unit has a single match. The matching estimator we use is

$$\begin{aligned} \hat{\beta}_{\text{SM}} &= \frac{1}{n}\left\{\sum_{i=1}^{n_1}\left(Y_{1i} - \hat{Y}_{0i}\right) + \sum_{i=1}^{n_0}\left(\hat{Y}_{1i} - Y_{0i}\right)\right\}, \\ \hat{\gamma}_{\text{SM}} &= \frac{1}{n_1}\sum_{i=1}^{n_1}\left(Y_{1i} - \hat{Y}_{0i}\right), \end{aligned} \tag{7}$$

where n_1 and n_0 are the number of treated and control units respectively, $n = n_1 + n_0$, and $\hat{Y}_{1i}$ and $\hat{Y}_{0i}$ are the outcome of the closest match to unit i among the treated and control respectively. Closest matches are chosen to minimize the Euclidian vector norm

$$\|\mathbf{C}_i - \mathbf{C}_m\| = \{(\mathbf{C}_i - \mathbf{C}_m)'(\mathbf{C}_i - \mathbf{C}_m)\}^{1/2},$$

where $\mathbf{C}_i$ and $\mathbf{C}_m$ denotes the covariate vector to be matched for, for the observed unit i and its potential match m respectively.

The second estimator we study is an inverse probability weighting estimator (Hirano, Imbens, and Ridder, 2003)

$$\hat{\beta}_{\text{IPW}} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\left(\frac{T_iY_i}{\hat{e}(\mathbf{C}_i)} - \frac{(1-T_i)Y_i}{(1-\hat{e}(\mathbf{C}_i))}\right), \quad \hat{\gamma}_{\text{IPW}} = \frac{\frac{1}{n}\sum_{i=1}^{n}\hat{e}(\mathbf{C}_i)\left(\frac{T_iY_i}{\hat{e}(\mathbf{C}_i)} - \frac{(1-T_i)Y_i}{(1-\hat{e}(\mathbf{C}_i))}\right)}{\frac{1}{n}\sum_{i}^{n}\hat{e}(\mathbf{C}_i)},$$

where $e(\mathbf{C}_i) = \Pr(T = 1|\mathbf{C}_i)$, called propensity score, is fitted by using a linear (in the covariates) logistic regression yielding $\hat{e}(\mathbf{C}_i)$, and $\mathbf{C}_i$ is the covariate vector to be controlled for. However, when the latter is $\mathbf{Z}$ or $\hat{\mathbf{Z}}$, i.e. covariates affecting the outcome solely have been discarded, then we follow Hahn (2004) and replace $\hat{\gamma}_{\text{IPW}}$ by

$$\hat{\gamma}_{\text{IPW}} = \frac{\frac{1}{n}\sum_{i=1}^{n}\hat{e}(\mathbf{D}_i)\left(\frac{T_iY_i}{\hat{e}(\mathbf{C}_i)} - \frac{(1-T_i)Y_i}{(1-\hat{e}(\mathbf{C}_i))}\right)}{\frac{1}{n}\sum_{i}^{n}\hat{e}(\mathbf{D}_i)},$$

where $\mathbf{C} = \mathbf{X}_Y$ and $\mathbf{D} = \mathbf{Z}$, or $\mathbf{C} = \hat{\mathbf{X}}_Y$ and $\mathbf{D} = \hat{\mathbf{Z}}$

The results (bias, standard deviation and mean squared error, MSE) for both estimators are found in Tables 4-9 (Appendix B) and summarized in Figures 2-4.

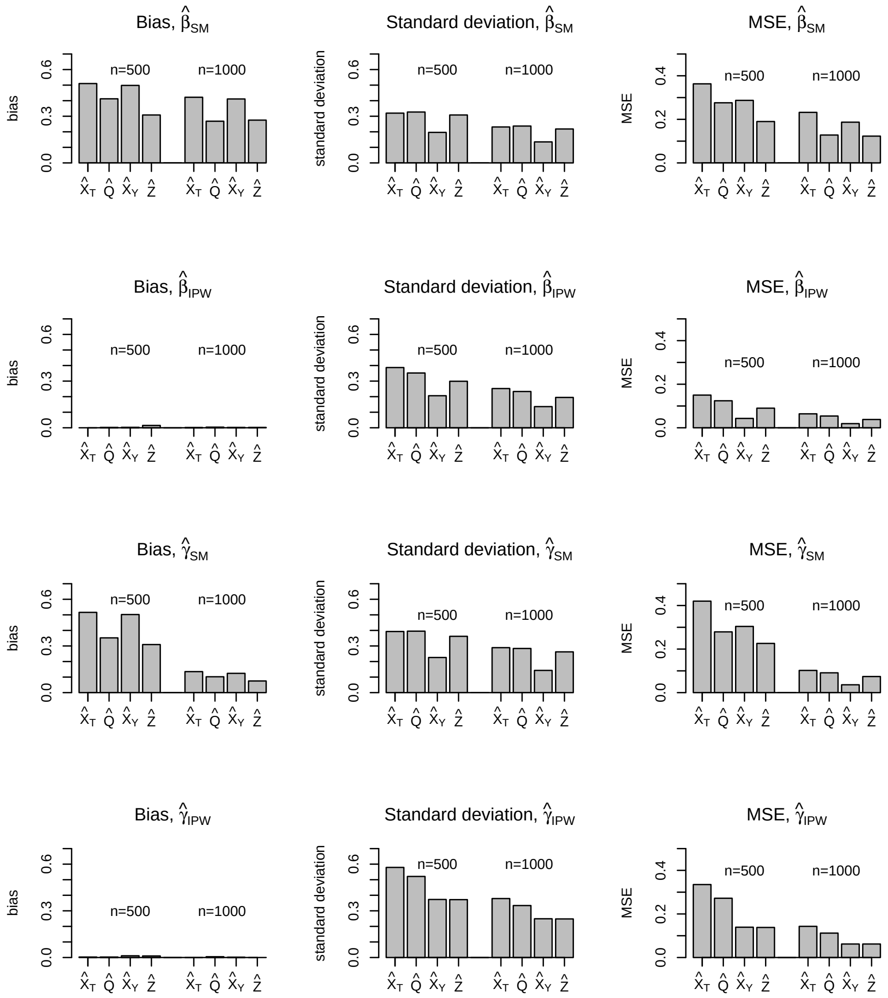

Figure 2: Model (4) – Bias, standard devation and MSE of $\hat{\beta}_{\text{SM}}$, $\hat{\beta}_{\text{IPW}}$, $\hat{\gamma}_{\text{SM}}$ and $\hat{\gamma}_{\text{IPW}}$ when estimators utilize the sets $\hat{\mathbf{X}}_T$, $\hat{\mathbf{Q}}$, $\hat{\mathbf{X}}_Y$ and $\hat{\mathbf{Z}}$.

For all our experiments we see that estimators based on the selected set $\hat{\mathbf{Q}}$ yield lower MSE than $\hat{\mathbf{X}}_T$. Moreover, $\hat{\mathbf{Q}}$ itself yields at best only as low MSE than $\hat{\mathbf{X}}_Y$ and/or $\hat{\mathbf{Z}}$. Thus, this study indicates that Algorithm B in Table 1 is to be preferred to Algorithm A. This is in line with Kuroki and Miyakawa (2003), where by restrict-

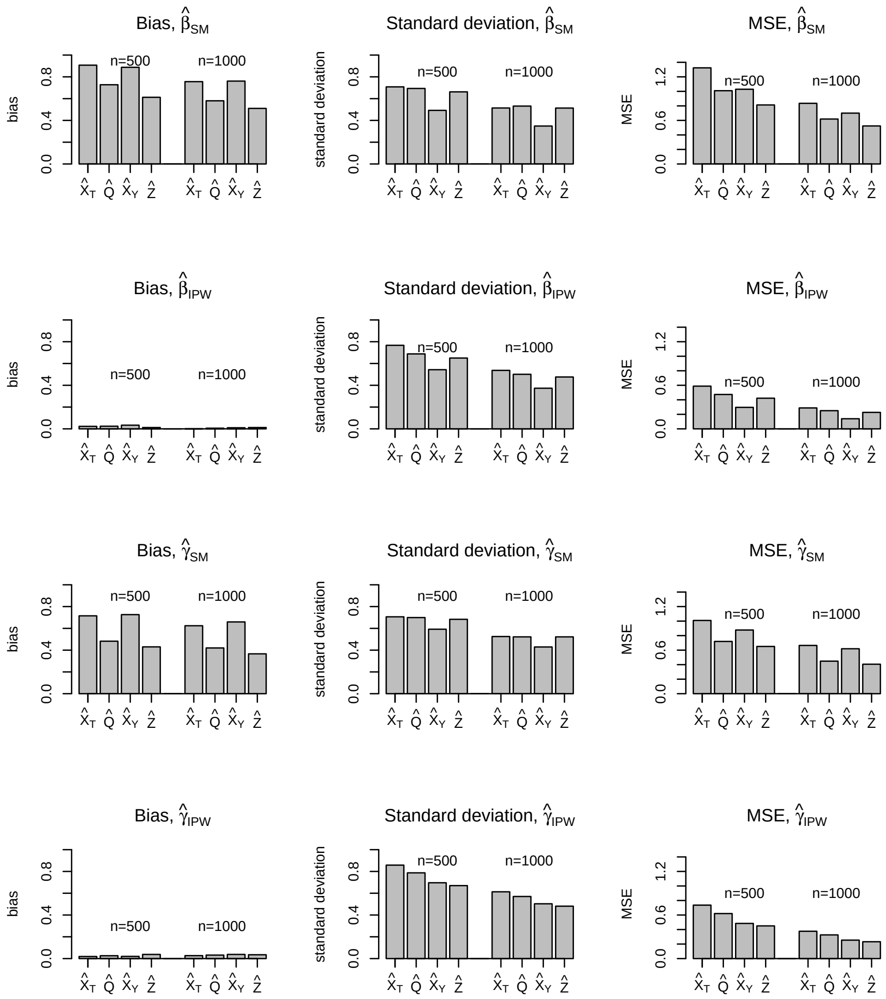

Figure 3: Model (5) – Bias, standard devation and MSE of $\hat{\beta}_{\text{SM}}$, $\hat{\beta}_{\text{IPW}}$, $\hat{\gamma}_{\text{SM}}$ and $\hat{\gamma}_{\text{IPW}}$ when estimators utilize the sets $\hat{\mathbf{X}}_T$, $\hat{\mathbf{Q}}$, $\hat{\mathbf{X}}_Y$ and $\hat{\mathbf{Z}}$.

ing attention to linear structural equations, i.e. a parametric situation, it is shown that it is more efficient to estimate the treatment effect by conditioning on the covariates that affects directly the outcome rather than conditioning on variables that are related to the outcome only through other variables. Finally, we note that our results show that both $\hat{\mathbf{X}}_Y$ and $\hat{\mathbf{Z}}$ may yield lowest MSE. This typically depends on whether bias dominates the variance, hence on the data generating mechanism and the estimator used. Our suggestion would be to use $\hat{\mathbf{Z}}$ instead of $\hat{\mathbf{X}}_Y$ in situations

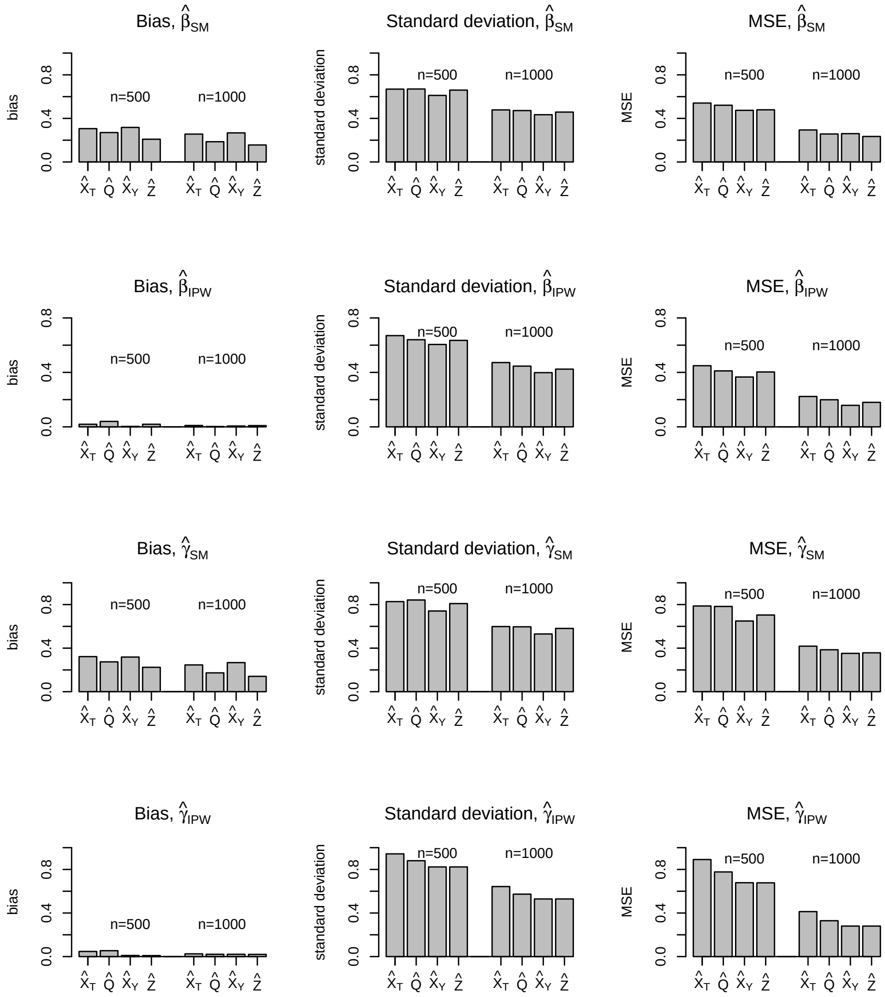

Figure 4: Model (6) − Bias, standard devation and MSE of $\hat{\beta}_{\text{SM}}$, $\hat{\beta}_{\text{IPW}}$, $\hat{\gamma}_{\text{SM}}$ and $\hat{\gamma}_{\text{IPW}}$ when estimators utilize the sets $\hat{\mathbf{X}}_T$, $\hat{\mathbf{Q}}$, $\hat{\mathbf{X}}_Y$ and $\hat{\mathbf{Z}}$.

where the cardinality of the former is much lower than that of the former, since then one may expect bias to be more of a problem. Finally, we should note here that the bias obtained with the inverse probability weighting estimator is probably unrealistically low. This is due to the fact that this estimator is using the correctly specified parametric propensity score. Using a non-parametric estimate of the latter should increase bias and decrease variance.

6 Discussion

In this paper we have argued that non-parametric estimation of an average treatment effect can be improved by reducing the dimension of the set of covariates utilized. We have defined reduced sets of covariates which guarantee the unconfoundedness of the treatment, and we have provided a general sufficient condition under which the defined subsets are minimal in the presence of unobserved variables. We have also shown that the defined subsets of covariates are identified through the conditional independence statements given in Propositions 13 and 14. With simulations we study the finite sample properties of non-parametric estimators based on covariate selected with the data at hand. Estimators based on selected sets of covariates are often shown to outperform estimators using larger sets in terms of mean squared error. In particular, our results suggest that using the covariates sets that affect treatment assignment (common practice) is not optimal, and that the focus should lie on covariates affecting the outcome instead.

There is some controversy in the literature concerning the relevance of asymptotic properties of non-parametric estimators of treatment effects when dealing with finite sample sizes, see Robins and Ritov (1997) and Angrist and Hahn (2004). In particular, $\sqrt{n}$-consistency of the estimators of an average treatment effect is typically obtained by averaging over non-parametric estimates of regression functions over the covariates. The latter estimates have slower rates of convergence (the more covariates the slower) and this may be translated into poor small sample performance even for the resulting $\sqrt{n}$-consistent averages. This issue merely highlights the need for the reduction of the dimension of the covariate set discussed herein.

The minimal sets of covariates that we characterize are unique within the subsets of covariates in which they are defined. In general, other minimal sets may be available, possibly of lower cardinality, although we conjecture that the minimal sets defined in this paper are the only identified ones if no further assumptions are made.

Acknowledgments

The Swedish Research Council (Linneaus grant: Ageing and Living Condition program), the Swedish Council for Working Life and Social Research and the Swedish

Foundation for International Cooperation in Research and Higher Education are acknowledged for their financial support.

References

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

Akaike, H. (1973). Information theory and extension of the maximum likelihood principle. In B. Petrov and F. Csàki (Eds.), *Proc. 2nd Int. Symp. Information Theory*, Budapest, pp. 267–281. Akadémiai Kiadó.

Angrist, J. and Hahn, J. (2004). When to control for covariates? Panel asymptotics for estimates of treatment effects. *The Review of Economics and Statistics*, **86**, 58–72.

Bellman, R. E. (1961). *Adaptive Control Processes.* Princeton: Princeton University Press.

Cook, R. D. (1996). Graphics for regressions with a binary response. *Journal of the American Statistical Association*, **91**, 983–992.

Cook, R. D. (1998). *Regression Graphics.* New York: Wiley.

Cook, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. *The Annals of Statistics*, **32**, 1062–1092.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B*, **41**, 1–31.

Dawid, A. P. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review*, **70**, 161–189.

de Luna, X. and Johansson, P. (2006). Exogeneity in structural equation models. *Journal of Econometrics*, **132**, 527–543.

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

Hahn, J. (2004). Functional restriction and efficiency in causal inference. *The Review of Economics and Statistics*, **86**, 73–76.

Heckman, J., Ichimura, H., and Todd, P. (1998). Matching as an econometric evaluation estimator. *Review of Economic Studies*, **65**, 261–294.

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

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

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., Newey, W., and Ridder, G. (2003). Mean-squared-error calculations for average treatment effects. Department of Economics, UC Berkeley, unpublished manuscript.

Kuroki, M. and Miyakawa, M. (2003). Covariate selection for estimating the causal effect of control plans by using causal diagrams. *Journal of the Royal Statistical Society: Series B*, **65**, 209–222.

Lauritzen, S. (1996). *Graphical Models*. Oxford: Oxford University Press.

Lauritzen, S. (2001). Causal inference from graphical models. In Barndorff-Nielsen, O.E., Cox, D.R. and Klüppelberg, C. (Eds.), *Complex Stochastic Systems*, pp. 63-107, London: Chapman and Hall.

Li, K. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, **86**, 316–327.

Li, L., Cook, R. D., and Nachtsheim, C. J. (2005). Model-free variable selection. *Journal of the Royal Statistical Society Series B*, **67**, 285–299.

Meek, C. (1995). Strong completeness and faithfulness in Bayesian networks. In P. Besnard and S. Hanks (Eds.), *Uncertainty in Artificial Intelligence: Proceedings of the* 11$^{\text{th}}$ *Conference*, San Francisco, pp. 411–418. Morgan Kaufmann.

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, **X**, 1–51. In Polish, English translation by D. Dabrowska and T. Speed in *Statistical Science*, **5**, 465–472, 1990.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kauffman.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Richardson, T. S. and Spirtes, P. (2002). Ancestral graph Markov models. *The Annals of Statistics*, **30**, 962–1030.

Richardson, T. S. and Spirtes, P. (2003). Causal inference via ancestral graph models. In Green, P. J., Hjort, N.L., and Richardson, S. (Eds.), *Highly Structured Stochastic Systems*, Oxford: Oxford University Press, pp. 83-108.

Robins, J. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine*, **16**, 285–319.

Robins, J., Scheines, R., Spirtes, P., and Wasserman, L. (2003). Uniform consistency in causal inference. *Biometrika*, **90**, 491–515.

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, **90**, 122–129.

Rosenbaum, P. R. (2002). *Observational Studies (2nd edition)*. New York: Springer.

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

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

Rubin, D. B. (1991). Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics*, **47**, 1213–1234.

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

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction and Search, (2nd edition)*. Cambridge, MA: The MIT Press.

Su, L. and White, H. (2003). Testing conditional independence via empirical likelihood. *Working Paper*, 2003-14, Department of Economics, University of California, San Diego.

Zhang, J. and Spirtes, P. (2003). Strong faithfulness and uniform consistency in causal inference. In *Proceedings of the 19th Annual Conference on Uncertainty in Artificial Intelligence (UAI-03)*, San Francisco, CA, pp. 632–63. Morgan Kaufmann.

A Marginalizing a DAG

A.1 Preliminaries

An independence model $\mathcal{I}$ over a vertex set $\mathbf{V}$ is a set of triples $\langle \mathbf{A}, \mathbf{B} | \mathbf{C} \rangle$, where $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are disjoint subsets of $\mathbf{V}$ and $\mathbf{C}$ is possibly empty. A separation criterion Γ associates an independence model $\mathcal{I}_\Gamma(\mathcal{G})$ with a graph $\mathcal{G}$: $\langle \mathbf{A}, \mathbf{B} | \mathbf{C} \rangle \in \mathcal{I}_\Gamma(\mathcal{G})$ if and only if $\mathbf{A}$ and $\mathbf{B}$ are separated given $\mathbf{C}$ in $\mathcal{G}$ under criterion Γ. For instance, the independence model resulting from applying the d-separation criterion was denoted $\mathcal{I}_d(\mathcal{G})$ in Section 4.1. An independence model $\mathcal{I}_\Gamma(\mathcal{G})[_\mathbf{U}$, after *marginalizing out a subset* $\mathbf{U}$ is the subset of triples that do not involve any vertices in $\mathbf{U}$:

$$\mathcal{I}_\Gamma(\mathcal{G})[_\mathbf{U} \equiv \{\langle \mathbf{A}, \mathbf{B} | \mathbf{C} \rangle | \langle \mathbf{A}, \mathbf{B} | \mathbf{C} \rangle \in \mathcal{I}_d(\mathcal{G}); (\mathbf{A} \cup \mathbf{B} \cup \mathbf{C}) \cap \mathbf{U} = \emptyset\},$$

where $\emptyset$ denotes the empty set. We now define a graphical transformation for a DAG $\mathcal{G}$, $\mathcal{G} \mapsto \mathcal{G}[_\mathbf{U}$, in such a way that the independence model corresponding to the transformed graph, $\mathcal{G}[_\mathbf{U}$, will be the independence model obtained when marginalizing over the unobserved variables in the original DAG. Note that the transformed graph is not, in general, a DAG.

Definition 15 *Graphical transformation for a DAG $\mathcal{G}$, $\mathcal{G} \mapsto \mathcal{G}[_\mathbf{U}$*

The graph $\mathcal{G}[_\mathbf{U}$ has vertex set $\mathbf{V} \setminus \mathbf{U}$ and edges specified as follows: If A and B are such that $\forall \mathbf{Z}$, with $\mathbf{Z} \subseteq \mathbf{V} \setminus \mathbf{U} \cup \{A, B\}$, $\langle A, B | \mathbf{Z} \rangle \notin \mathcal{I}_d(\mathcal{G})$,

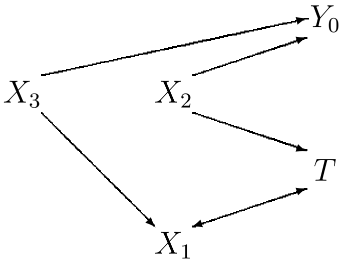

Figure 5: An ancestral graph, marginalizing Figure 1(b) over U.

$$\textit{and} \left\{ \begin{array}{c} A \in an(B);\, B \notin an(A) \\ A \notin an(B);\, B \in an(A) \\ A \notin an(B);\, B \notin an(A) \end{array} \right\},\ \textit{then} \left\{ \begin{array}{c} A \rightarrow B \\ A \leftarrow B \\ A \leftrightarrow B \end{array} \right\},\ \textit{in}\ \mathcal{G}[_{\mathbf{U}}.$$

In particular, $\mathcal{G}[_{\mathbf{U}}$ is a graph where all vertices in $\mathbf{U}$ have been deleted. Two vertices A and B are adjacent in $\mathcal{G}[_{\mathbf{U}}$ if for every subset $\mathbf{Z}$ of $\mathbf{V} \setminus \mathbf{U}$, A is not d-separated from B given $\mathbf{Z}$ in $\mathcal{G}$. The graph $\mathcal{G}[_{\mathbf{U}}$ is called an *ancestral graph* by Richardson and Spirtes (2002). For example, the ancestral graph corresponding to the DAG of Figure 1(b) where U is marginalized is drawn in Figure 2.

For an ancestral graph with directed and bi-directed edges the d-separation criterion extends to all colliders $\rightarrow Z \leftarrow$, $\leftrightarrow Z \leftrightarrow$, $\leftrightarrow Z \leftarrow$, $\rightarrow Z \leftrightarrow$, and is then called the *m-separation criterion*, see Richardson and Spirtes (2002). The graphical independence model resulting from applying the m-separation criterion to a graph $\mathcal{G}$ is denoted $\mathcal{I}_m(\mathcal{G})$. Note that for a DAG $\mathcal{G}$ we have $\mathcal{I}_d(\mathcal{G})[_U = \mathcal{I}_m(\mathcal{G}[_U)$, see Richardson and Spirtes (2002, Thm 4.18).

By construction we have that if $\langle \mathbf{A}, \mathbf{B} | \mathbf{C} \rangle \in \mathcal{I}_m(\mathcal{G}[_U)$ then $\mathbf{A} \perp\!\!\!\perp \mathbf{B}|\mathbf{C}[P]$. Conversely, if $(\mathbf{A} \cup \mathbf{B} \cup \mathbf{C}) \cap \mathbf{U} = \emptyset$ and $\mathbf{A} \perp\!\!\!\perp \mathbf{B}|\mathbf{C}[P]$, then $\langle \mathbf{A}, \mathbf{B} | \mathbf{C} \rangle \in \mathcal{I}_d(\mathcal{G})$ by the faithfulness assumption and, hence, $\langle \mathbf{A}, \mathbf{B} | \mathbf{C} \rangle \in \mathcal{I}_m(\mathcal{G}[_U)$. In the sequel we use $\mathbf{A} \perp\!\!\!\perp \mathbf{B}|\mathbf{C}$ to denote conditional independence under P, d-separation in $\mathcal{G}$ and m-separation in $\mathcal{G}[_U$.

A.2 Proof of Propositions 10 and 11

Let P be a faithful distribution to a DAG $\mathcal{G}$ which is such that assumption (A.7) holds. To obtain results under (A.6), replace in the proofs P with P_j and $\mathcal{G}$ with

$\mathcal{G}_j$, $j = 0, 1$. We denote in the sequel the transformed graph $\mathcal{G}^* = \mathcal{G}[_{\mathbf{U}}$. We start by showing two lemmas.

Lemma 16 *Assume (A.1-A.4) and (A.6) or (A.7) then* $T \perp\!\!\!\perp \mathbf{X} \setminus pa_{\mathcal{G}^*}(T) | pa_{\mathcal{G}^*}(T)$.

Proof. Suppose for a contradiction that there is a path π m-connecting T and $X^* \in \mathbf{X} \setminus pa_{\mathcal{G}^*(T)}(T)$. The first vertex R in π is adjacent to T and, hence, by (A.6) $R \in pa_{\mathcal{G}^*(T)}(T)$. Consequently R is a non-collider on the path π, and the latter cannot m-connect T and X^*. This contradiction establishes the result. ■

Lemma 17 *Assume (A.1-A.4) and (A.6) or (A.7) then* $pa_{\mathcal{G}^*(T)}(T) = \mathbf{X}_T$.

Proof. Let $R \in pa_{\mathcal{G}^*(T)}(T)$, then no subset of $\mathbf{X}$ m-separates R from T. Hence, $R \in \mathbf{X}_T$ showing that $pa_{\mathcal{G}^*(T)}(T) \subseteq \mathbf{X}_T$. Moreover, by Lemma 1 $\mathbf{X}_T \subseteq pa_{\mathcal{G}^*(T)}(T)$, thereby completing the proof. ■

Proof of Proposition 10. Let $\mathbf{Q}_j^- \subset \mathbf{Q}_j$ and assume for a contradiction that $T \perp\!\!\!\perp Y_j | \mathbf{Q}_j^-$. By the definition of $\mathbf{Q}_j$, we have that $Y_j \not\perp\!\!\!\perp \mathbf{X}_T \setminus \mathbf{Q}_j^- | \mathbf{Q}_j^-$, hence by our assumptions Y_j is m-connected to $X_T^- \in \mathbf{X}_T \setminus \mathbf{Q}_j^-$ in $\mathcal{G}^*$. Hence there is an m-connecting path π from Y_j to X_T^- given $\mathbf{Q}_j^-$. By Lemma 17 we have that $X_T^- \in pa_{\mathcal{G}^*}(T)$ so that π may be extended to a path m-connecting Y_j and T given $\mathbf{Q}_j^-$, which is a contradiction. ■

The proof of Proposition 11 is obtained as above by replacing $\mathbf{Q}_j$ by $\mathbf{Q}$ and Y_j by (Y_0, Y_1).

B Simulation results

We display here details of the simulation results described and discussed in Section 5.

Table 4: Model (4) $-$ Bias, standard deviation and MSE for $\hat{\beta}_{\text{SM}}$ and $\hat{\gamma}_{\text{SM}}$.

		$\hat{\beta}$			$\hat{\gamma}$		
n	matched covariates	bias	standard deviation	MSE	bias	standard deviation	MSE
500	$\mathbf{X}_T$	0.475	0.382	0.372	0.481	0.445	0.428
	$\mathbf{Q}$	0.341	0.377	0.258	0.211	0.432	0.231
	$\hat{\mathbf{X}}_T$	0.510	0.320	0.363	0.516	0.393	0.420
	$\hat{\mathbf{Q}}$	0.412	0.327	0.276	0.352	0.395	0.279
	$\mathbf{X}_Y$	0.422	0.167	0.206	0.424	0.200	0.220
	$\mathbf{Z}$	0.200	0.330	0.149	0.203	0.380	0.185
	$\hat{\mathbf{X}}_Y$	0.498	0.196	0.287	0.502	0.226	0.304
	$\hat{\mathbf{Z}}$	0.308	0.308	0.190	0.309	0.362	0.226
1000	$\mathbf{X}_T$	0.386	0.273	0.223	0.130	0.321	0.120
	$\mathbf{Q}$	0.270	0.272	0.147	0.062	0.306	0.097
	$\hat{\mathbf{X}}_T$	0.422	0.231	0.232	0.135	0.289	0.102
	$\hat{\mathbf{Q}}$	0.268	0.237	0.128	0.102	0.284	0.091
	$\mathbf{X}_Y$	0.344	0.103	0.129	0.103	0.128	0.027
	$\mathbf{Z}$	0.138	0.224	0.069	0.041	0.292	0.087
	$\hat{\mathbf{X}}_Y$	0.411	0.135	0.187	0.124	0.143	0.036
	$\hat{\mathbf{Z}}$	0.275	0.218	0.123	0.075	0.262	0.074

Table 5: Model (5) $-$ Bias, standard deviation and MSE for $\hat{\beta}_{\text{SM}}$ and $\hat{\gamma}_{\text{SM}}$.

		$\hat{\beta}$			$\hat{\gamma}$		
n	matched covariates	bias	standard deviation	MSE	bias	standard deviation	MSE
500	$\mathbf{X}_T$	0.845	0.811	1.372	0.660	0.767	1.023
	$\mathbf{Q}$	0.658	0.806	1.082	0.286	0.744	0.635
	$\hat{\mathbf{X}}_T$	0.907	0.708	1.324	0.715	0.706	1.009
	$\hat{\mathbf{Q}}$	0.727	0.693	1.009	0.482	0.699	0.720
	$\mathbf{X}_Y$	0.718	0.450	0.717	0.642	0.577	0.745
	$\mathbf{Z}$	0.658	0.806	1.082	0.286	0.744	0.635
	$\hat{\mathbf{X}}_Y$	0.887	0.492	1.029	0.726	0.592	0.877
	$\hat{\mathbf{Z}}$	0.612	0.662	0.813	0.430	0.683	0.651
1000	$\mathbf{X}_T$	0.709	0.572	0.829	0.582	0.558	0.650
	$\mathbf{Q}$	0.548	0.569	0.623	0.261	0.540	0.359
	$\hat{\mathbf{X}}_T$	0.756	0.514	0.835	0.624	0.525	0.665
	$\hat{\mathbf{Q}}$	0.580	0.532	0.619	0.420	0.522	0.448
	$\mathbf{X}_Y$	0.609	0.319	0.472	0.584	0.416	0.513
	$\mathbf{Z}$	0.548	0.569	0.623	0.261	0.540	0.359
	$\hat{\mathbf{X}}_Y$	0.761	0.348	0.700	0.659	0.429	0.619
	$\hat{\mathbf{Z}}$	0.510	0.513	0.524	0.366	0.522	0.406

Table 6: Model (6) − Bias, standard deviation and MSE for $\hat{\beta}_{\text{SM}}$ and $\hat{\gamma}_{\text{SM}}$.

		$\hat{\beta}$			$\hat{\gamma}$		
	matched		standard			standard	
n	covariates	bias	deviation	MSE	bias	deviation	MSE
500	$\mathbf{X}_T$	0.283	0.688	0.553	0.292	0.859	0.822
	$\mathbf{Q}$	0.226	0.679	0.513	0.178	0.856	0.764
	$\hat{\mathbf{X}}_T$	0.306	0.669	0.541	0.322	0.827	0.787
	$\hat{\mathbf{Q}}$	0.270	0.670	0.521	0.274	0.842	0.783
	$\mathbf{X}_Y$	0.269	0.587	0.417	0.278	0.721	0.597
	$\mathbf{Z}$	0.120	0.651	0.438	0.130	0.807	0.668
	$\hat{\mathbf{X}}_Y$	0.317	0.611	0.474	0.318	0.741	0.649
	$\hat{\mathbf{Z}}$	0.209	0.660	0.479	0.224	0.809	0.704
1000	$\mathbf{X}_T$	0.227	0.499	0.300	0.216	0.615	0.424
	$\mathbf{Q}$	0.152	0.491	0.264	0.065	0.599	0.362
	$\hat{\mathbf{X}}_T$	0.256	0.478	0.294	0.246	0.598	0.418
	$\hat{\mathbf{Q}}$	0.186	0.472	0.257	0.173	0.596	0.385
	$\mathbf{X}_Y$	0.204	0.416	0.215	0.209	0.514	0.308
	$\mathbf{Z}$	0.074	0.459	0.215	0.062	0.570	0.328
	$\hat{\mathbf{X}}_Y$	0.267	0.434	0.260	0.267	0.530	0.352
	$\hat{\mathbf{Z}}$	0.156	0.458	0.234	0.141	0.581	0.357

Table 7: Model (4) $-$ Bias, standard deviation and MSE for $\hat{\beta}_{\text{IPW}}$ and $\hat{\gamma}_{\text{IPW}}$.

		$\hat{\beta}_{\text{IPW}}$			$\hat{\gamma}_{\text{IPW}}$		
n	covariates	bias	sd	MSE	bias	sd	MSE
500	$\mathbf{X}_T$	0.011	0.440	0.194	0.020	0.612	0.374
	$\mathbf{Q}$	0.014	0.355	0.126	0.024	0.445	0.199
	$\hat{\mathbf{X}}_T$	0.000	0.387	0.150	-0.003	0.579	0.335
	$\hat{\mathbf{Q}}$	-0.002	0.352	0.124	-0.003	0.521	0.272
	$\mathbf{X}_Y$	0.004	0.151	0.023	0.006	0.279	0.078
	$\mathbf{Z}$	0.002	0.305	0.093	0.005	0.277	0.077
	$\hat{\mathbf{X}}_Y$	0.003	0.206	0.043	0.011	0.373	0.139
	$\hat{\mathbf{Z}}$	0.015	0.299	0.090	0.010	0.372	0.138
1000	$\mathbf{X}_T$	-0.013	0.300	0.090	-0.011	0.413	0.170
	$\mathbf{Q}$	-0.006	0.247	0.061	-0.004	0.302	0.091
	$\hat{\mathbf{X}}_T$	-0.001	0.252	0.064	0.000	0.379	0.143
	$\hat{\mathbf{Q}}$	0.004	0.233	0.054	0.005	0.334	0.112
	$\mathbf{X}_Y$	0.006	0.099	0.010	0.006	0.176	0.031
	$\mathbf{Z}$	-0.002	0.202	0.041	0.006	0.175	0.031
	$\hat{\mathbf{X}}_Y$	0.002	0.136	0.019	0.001	0.249	0.062
	$\hat{\mathbf{Z}}$	0.002	0.195	0.038	0.000	0.248	0.062

Table 8: Model (5) $-$ Bias, standard deviation amd MSE for $\hat{\beta}_{\text{IPW}}$ and $\hat{\gamma}_{\text{IPW}}$.

		$\hat{\beta}$			$\hat{\gamma}$		
n	matched covariates	bias	standard deviation	MSE	bias	standard deviation	MSE
500	$\mathbf{X}_T$	0.008	0.848	0.719	0.011	0.908	0.823
	$\mathbf{Q}$	-0.004	0.714	0.509	0.003	0.761	0.578
	$\hat{\mathbf{X}}_T$	0.023	0.767	0.588	0.019	0.858	0.736
	$\hat{\mathbf{Q}}$	0.024	0.688	0.473	0.026	0.787	0.620
	$\mathbf{X}_Y$	0.022	0.470	0.221	0.011	0.636	0.405
	$\mathbf{Z}$	-0.004	0.714	0.509	0.030	0.561	0.315
	$\hat{\mathbf{X}}_Y$	0.035	0.543	0.296	0.020	0.696	0.484
	$\hat{\mathbf{Z}}$	0.012	0.650	0.422	0.038	0.670	0.450
1000	$\mathbf{X}_T$	0.000	0.588	0.345	0.030	0.643	0.414
	$\mathbf{Q}$	0.011	0.521	0.272	0.041	0.551	0.305
	$\hat{\mathbf{X}}_T$	0.001	0.537	0.288	0.027	0.613	0.376
	$\hat{\mathbf{Q}}$	0.006	0.501	0.250	0.032	0.570	0.326
	$\mathbf{X}_Y$	0.008	0.337	0.114	0.039	0.453	0.207
	$\mathbf{Z}$	0.011	0.521	0.272	0.040	0.402	0.163
	$\hat{\mathbf{X}}_Y$	0.010	0.373	0.139	0.038	0.503	0.254
	$\hat{\mathbf{Z}}$	0.012	0.476	0.227	0.035	0.481	0.232

Table 9: Model (6) − Bias, standard deviation and MSE for $\hat{\beta}_{\text{IPW}}$ and $\hat{\gamma}_{\text{IPW}}$.

		$\hat{\beta}$			$\hat{\gamma}$		
n	matched covariates	bias	standard deviation	MSE	bias	standard deviation	MSE
500	$\mathbf{X}_T$	0.0116	0.694	0.481	0.039	0.954	0.911
	$\mathbf{Q}$	0.009	0.631	0.397	0.020	0.788	0.621
	$\hat{\mathbf{X}}_T$	0.019	0.670	0.449	0.047	0.943	0.891
	$\hat{\mathbf{Q}}$	0.040	0.640	0.411	0.054	0.880	0.777
	$\mathbf{X}_Y$	-0.003	0.554	0.307	0.001	0.717	0.514
	$\mathbf{Z}$	0.001	0.602	0.362	0.000	0.715	0.510
	$\hat{\mathbf{X}}_Y$	0.003	0.605	0.366	0.010	0.823	0.678
	$\hat{\mathbf{Z}}$	0.019	0.635	0.403	0.009	0.823	0.677
1000	$\mathbf{X}_T$	-0.008	0.409	0.167	-0.029	0.522	0.273
	$\mathbf{Q}$	-0.015	0.455	0.207	-0.034	0.558	0.312
	$\hat{\mathbf{X}}_T$	-0.010	0.472	0.223	-0.025	0.643	0.413
	$\hat{\mathbf{Q}}$	-0.002	0.446	0.199	-0.021	0.573	0.329
	$\mathbf{X}_Y$	-0.003	0.378	0.143	-0.016	0.471	0.222
	$\mathbf{Z}$	-0.010	0.425	0.181	-0.014	0.471	0.222
	$\hat{\mathbf{X}}_Y$	-0.006	0.398	0.158	-0.021	0.529	0.280
	$\hat{\mathbf{Z}}$	-0.009	0.424	0.180	-0.020	0.529	0.280