Covariate selection and propensity score specification in causal inference

Ingeborg Waernbaum

Doctoral Dissertation
Department of Statistics
Umeå University
SE-901 87 Umeå, Sweden

ISSN: 1100-8989
ISBN: 978-91-7264-564-6
Printed by Print & Media, Umeå 2008:2004588

Abstract

This thesis makes contributions to the statistical research field of causal inference in observational studies. The results obtained are directly applicable in many scientific fields where effects of treatments are investigated and yet controlled experiments are difficult or impossible to implement.

In the first paper we define a partially specified directed acyclic graph (DAG) describing the independence structure of the variables under study. Using the DAG we show that given that unconfoundedness holds we can use the observed data to select minimal sets of covariates to control for. General covariate selection algorithms are proposed to target the defined minimal subsets.

The results of the first paper are generalized in Paper II to include the presence of unobserved covariates. Morevoer, the identification assumptions from the first paper are relaxed.

To implement the covariate selection without parametric assumptions we propose in the third paper the use of a model-free variable selection method from the framework of sufficient dimension reduction. By simulation the performance of the proposed selection methods are investigated. Additionally, we study finite sample properties of treatment effect estimators based on the selected covariate sets.

In paper IV we investigate misspecifications of parametric models of a scalar summary of the covariates, the propensity score. Motivated by common model specification strategies we describe misspecifications of parametric models for which unbiased estimators of the treatment effect are available. Consequences of the misspecification for the efficiency of treatment effect estimators are also studied.

Keywords and phrases: Covariate selection, graphical models, matching, observational studies, treatment effects.

AMS 2000 subject classification: 62G05, 62B99, 62G35.

Preface

My first encounter with the science of statistics was when visiting my brother who was studying at the university. He brought me along to a lecture on a course in mathematical statistics. Not that I was enthusiastic at that time, I remember mostly understanding nothing at all when looking at all formulas on the blackboard. I also remember that they were talking all the time about something I had not heard about before. They called it estimation. Several years later when I had my own first experience with the subject it was love at second sight. I knew immediately that this was what I wanted to study. I wanted to learn how to learn about the world around us and I haven't stopped wanting to learn more since then. In that sense the years spent on the work with this thesis have been a privilege for me.

First of all I want to thank my supervisor, Professor Xavier de Luna. We have shared an amazing journey along the intriguing paths of causal inference. Thank you Xavier for all your invaluable help and support. You have made the years as a PhD student a very happy chapter of my life and for that I am truly grateful.

I want to thank my present and former colleagues at the Department of Statistics. Professor Göran Broström, my co-supervisor, for your help with the work on this thesis. I am very grateful to my first teacher Bengt Lundquist who inspired me and shared his view of statistics as a science for the science beyond formulas and technical procedures. I am indebted to Maria Karlsson for insightful and constructive comments on earlier drafts of several of the papers and the introduction to this thesis. To my friends at the department, thank you for listening and supporting me especially through tough times. Very special thanks to my dear friend Marie Lindkvist. Marie, you have really been a good friend and with you I have shared most of my problems and joys.

To my friends in other departments, History, Mathematical Statistics, Economics, Economic History, Odontology, Law, Public Health and Clinical Medicine, thank you for your friendship. Meeting different disciplines and perspectives but sharing the same reality in the academic life has been important for both my personal and scientific development.

Many thanks to my friends and family for your love and support. Most of all I want to thank my wife Maria and our son Ove. You are the supervisors in my life and I am very happy that we are sharing it together.

Umeå, May 2008
Ingeborg Waernbaum

Contents

List of papers

The thesis is based on the following papers:

I. de Luna, X., Waernbaum I. (2005). Covariate selection for non-parametric estimation of treatment effects. *IFAU Working Paper*, 2005:4, Institute for Labour Market Policy Evaluation, Uppsala

II. de Luna, X., Richardson T. S., Waernbaum I. (2007). Identification of minimal sets of covariates for the non-parametric estimation of an average treatment effect. Research Report 2007, Department of Statistics Umeå University, Umeå.

III. Waernbaum I. (2008). Model-free variable selection in causal inference. Research Report 2008, Department of Statistics Umeå University, Umeå.

IV. Waernbaum I. (2008). Propensity score model specification for estimation of average treatment effects. Research Report 2008, Department of Statistics Umeå University, Umeå.

1 Introduction

A typical goal of scientific research is to draw inferences about the effects of treatments. A treatment can be interpreted both in the conventional sense such as a medication, surgery or other medical therapy, or more broadly as an employment program, a specific education or any other action or intervention assigned to an individual. A gold standard in this context is the randomized experiment. Here, treatments are assigned to units randomly ensuring that the effect of a treatment does not depend on systematic differences between the treated and controls. However, often a randomized experiment is not possible or desirable to implement. In such cases there may be registers or databases available with observations of characteristics of individuals, treatment variables and outcomes that may be used instead. The main problem with such observations is that the possible effect of the treatment cannot be distinguished from other effects of variables whose distributions differ between the treated and controls. Therefore, a main approach when analyzing data from an observational study is to "control" for differences between the treatment groups.

In empirical research an implicit criterion for regarding an effect as being caused by a treatment is that the treatment has been randomly assigned to the units under study. Although many researchers intuitively comprehend the role of randomization we require a formal statistical formulation of what a causal effect is. In the following we adopt the definition of a causal effect from the potential outcome framework developed by Rubin (1974, 1977, 1978) also referred to as the Rubin causal model, see Holland (1986). Even though this research field dates back to the 70's, there has been a recent surge of new results due to an increasing interest in the evaluation of non-randomized treatments, see e.g. the review by Imbens (2004).

This thesis consists of four paper which contribute to the statis-

tical theory of causal inference in several ways. A main contribution concerns characterizing subsets of pre-treatment variables (covariates) that we control for in order to perform causal inference. Informally, it is often stated that the variables needed to control for are the variables that affect both the treatment and the outcome. This set of variables is often referred to as confounders since they are the variables confounding the true effect of the treatment in a non-randomized study. Three of the papers are committed to this central problem in observational studies. Here, we formally define minimal sets of covariates needed to be controlled for. We show that the defined covariate sets can be selected from the observed data and we describe properties of nonparametric estimators of treatment effects based on the defined covariate sets. A novel approach in this context is that for some of the results of the papers we describe assumptions through a graphical model to illustrate properties of the variables in the Rubin causal model.

The fourth and last paper is dedicated to the specification of a summarizing function of the covariates called the propensity score, the probability of being treated given the covariates. For a broad class of parametric models we show that some estimators of the treatment effect using the propensity score are unbiased even though the propensity score model is misspecified. The implications of a misspecification for the efficiency of treatment effect estimators are also studied.

This summary is organized as follows. Section 2 presents an overview of the Rubin causal model, i.e., the definition of a causal effect through potential outcomes and the assumptions required for its identification. Here, we also describe the propensity score and its role in the estimation of a treatment effect along with a general description of nonparametric estimators of average causal effects. Section 3 describes the graphical assumptions imposed on the variables in the Rubin causal model. In Section 4 the contents of Papers

I-IV are summarized, and in the last section conclusions and future research directions are presented.

2 Causal inference

2.1 Defining causal effects through potential outcomes

In the following we consider studies where the objective is to estimate the effect of a binary treatment, T, where $T = 1$ for an active treatment and $T = 0$ for a control treatment. Let Y_1 denote a response variable that would be observed under the active treatment and Y_0 denote a response variable under the control treatment. The two variables are called potential outcomes (Neyman 1923; Rubin, 1974, 1977). Assume that a random sample of N units, indexed $i = 1, \ldots, N$, is drawn from a large population. For a unit i a causal effect of the treatment is defined as the difference between the potential outcomes

$$Y_{1i} - Y_{0i}.$$

Since for each individual, Y_{1i} and Y_{0i} are never jointly observed we face what is called the "fundamental problem of causal inference" (Holland 1986, Holland and Rubin 1988). The fundametal problem of causal inference implies that such a unit level causal effect can never be calculated with the observed data. To assess a causal effect of the treatment we must instead rely on comparisons of the potential outcomes across units and also add some identifying assumptions. An important assumption in this framework, often not explicitly stated, is the *stable unit treatment value assumption* (Rubin 1980). The assumption assures that there is no interference between units leading to different outcomes depending on the treatment other units receive.

If we denote by $\mathbf{X}$ a vector of pre-treatment variables we can classify a subpopulation of units by their values of the covariates and define an average causal effect over such a subpopulation

$$\beta(\mathbf{X}) = E(Y_1 - Y_0 \mid \mathbf{X}).$$

The parameter of main interest is often the average causal effect over the population

$$\beta = E(Y_1 - Y_0),$$

or the average causal effect in the subpopulation of the treated

$$\gamma = E(Y_1 - Y_0 \mid T = 1).$$

Summarizing the variables in this framework we have that each unit under study is sampled from the joint distribution of $(Y, \mathbf{X}, T)$, where Y is defined as $Y = TY_1 + (1 - T)Y_0$.

2.2 Identifying assumptions

We can use observations of different units to gain knowledge on the average causal effect if the assignment to treatment and the potential outcomes are independent conditionally on the covariates $\mathbf{X}$:

$$A.1.\ \text{(Unconfoundedness)} \quad (Y_1, Y_0) \perp\!\!\!\perp T \mid \mathbf{X}.$$

This assumption together with the assumption of an overlapping distribution,

$$A.2.\ \text{(Overlap)} \quad 0 < P(T = 1 \mid \mathbf{X}) < 1,$$

implies that an average causal effect can be estimated with the data at hand by observing that

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

We now see that estimating a causal effect of a treatment requires a set of covariates satisfying A.1 and A.2. The covariate set, $\mathbf{X}$, should include all of the covariates affecting both the treatment and outcomes. Without further information unconfoundedness is an untestable assumption. As an illustration consider a distribution F of the potential outcome Y_0. Under uncondfoundedness it holds that

$$F(Y_0|T = 0, \mathbf{X}) = F(Y_0|\mathbf{X}), \quad (2)$$

but since Y_0 is not observed for all individuals we cannot use the observed data to estimate the right hand side of (2). This means that we must rely on subject matter knowledge to judge the credibility of the assumption. Augmenting the set $\mathbf{X}$ as well as reducing the set can lead to a violation of the assumption for a given set $\mathbf{X}$ that satisfies unconfoundedness. Adding a new variable can make a noncounfounding variable to a confounder. On the contrary, controlling for yet one more variable can make a variable that was previously a confounder to a nonconfounder implying that the latter is no longer needed in the conditioning set. A powerful way of studying conditional independence properties of the joint distribution of $(Y_j, \mathbf{X}, T)$, $j = 0, 1$ is provided by the use of a graphical representation of the Rubin model as decribed in Section 3.

2.3 The propensity score

Controlling for a high dimensional vector of covariates $\mathbf{X}$ can be an insurmountable task to accomplish. As an example conditioning on ten binary covariates yields a number of $2^{10} = 1024$ possible realizations of the covariates. To avoid the need to compare treated and

controls on the values of all covariates one can use a scalar function of the covariates, the propensity score, $e(\mathbf{X}) = P(T = 1 \mid \mathbf{X})$. The propensity score is a, so called, balancing score, $b(\mathbf{X})$. A balancing score is a function of the covariates such that the conditional distribution of $\mathbf{X}$ given $b(\mathbf{X})$ is the same for the treated and control units, i.e., $\mathbf{X} \perp\!\!\!\perp T \mid b(\mathbf{X})$. Rosenbaum and Rubin (1983) showed that the coarsest balancing score is the propensity score. Formally they showed that, $\mathbf{X} \perp\!\!\!\perp T \mid b(\mathbf{X})$, if and only if $e(\mathbf{X})$ is a function of $b(\mathbf{X})$. They also showed that under assumption A.1 and A.2

$$(Y_1, Y_0) \perp\!\!\!\perp T \mid b(\mathbf{X}),$$

which implies that adjustments for any balancing score, e.g., the propensity score, suffices for removing all biases associated with differences in the covariates. The propensity score is used in the identification of an average causal effect either by replacing $\mathbf{X}$ with $e(\mathbf{X})$ in (1) or by noting that

$$\beta = E(Y_1 - Y_0) = E\left(\frac{TY}{e(\mathbf{X})}\right) - E\left(\frac{(1-T)Y}{1-e(\mathbf{X})}\right), \tag{3}$$

which is estimable with the observed data.

2.4 Nonparametric estimators and their properties

Equations (1) and (3) form the basis of how a treatment can be estimated nonparametrically with the data at hand. A nonparametric estimator of the treatment effect is an estimator not imposing distributional or functional form assumptions. Estimators following equation (1) first estimates the average treatment effect for a subpopulation with covariates $\mathbf{X} = \mathbf{x}$ and then the average of these conditional effects are taken over the relevant distribution of $\mathbf{X}$.

As an example, consider a simple matching estimator, $\hat{\beta}_{\text{SM}}$

$$\hat{\beta}_{\text{SM}} = \frac{1}{N}\sum_{i=1}^{N} \hat{Y}_{1i} - \hat{Y}_{0i}, \tag{4}$$

where for $i = 1, \ldots, N$ we have that

$$\hat{Y}_{1i} = \begin{cases} Y_{1i}, & \text{if } T_i = 1 \\ \tilde{Y}_{1i}, & \text{if } T_i = 0 \end{cases} \text{ and } \hat{Y}_{0i} = \begin{cases} \tilde{Y}_{0i}, & \text{if } T_i = 1 \\ Y_{0i}, & \text{if } T_i = 0, \end{cases}$$

with $\tilde{Y}_{1i}$ and $\tilde{Y}_{0i}$ being the observed value of the potential outcome for a matched unit having the same or almost the same observed values of the covariates as unit i. A matching estimator that estimate the causal effect on the treated, γ, is obtained by solely averaging the differences

$$\hat{\gamma}_{\text{SM}} = \frac{1}{N_1}\sum_{i=1}^{N_1} Y_{1i} - \hat{Y}_{0i}, \tag{5}$$

for the N_1 individuals in the sample that received the treatment. Abadie and Imbens (2006) study the estimators $\hat{\beta}_{\text{SM}}$ and $\hat{\gamma}_{\text{SM}}$ when matching is performed on the covariate vector norm.

For a propensity score matching estimator the matching criterion is changed from matching on the covariates directly to matching on the estimated propensity score, $\hat{e}(\mathbf{X})$. Another estimator of an average causal effect is obtained by replacing $\hat{Y}_{1i}$ and $\hat{Y}_{0i}$ with some estimates of the conditional expectations, $E(Y_1 \mid \mathbf{X}_i) = \mu_1(\mathbf{X}_i)$ and $E(Y_0 \mid \mathbf{X}_i) = \mu_0(\mathbf{X}_i)$, obtained by nonparametric regression. Specific nonparametric methods for estimating the regression functions have been studied for their properties in the estimation of a causal effect, see e.g. Heckman, Ichimura, and Todd (1998) and Imbens, Newey, and Ridder (2005).

An estimator directly applying equation (3) weights the observed outcome by the inverse of the probability of receiving the treatment actually received,

$$\hat{\beta}_{\text{IPW}} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{T_i Y_i}{\hat{e}(\mathbf{X}_i)} - \frac{(1-T_i)\,Y_i}{(1-\hat{e}(\mathbf{X}_i))}\right), \tag{6}$$

where $\hat{e}(\mathbf{X}_i)$ is an estimated value of the propensity score for unit i. As with the matching estimator this estimator can be adjusted to target γ

$$\hat{\gamma}_{\text{IPW}} = \frac{1}{N}\sum_{i=1}^{N}\hat{e}(\mathbf{X}_i)\left(\frac{T_i Y_i}{\hat{e}(\mathbf{X}_i)} - \frac{(1-T_i)\,Y_i}{(1-\hat{e}(\mathbf{X}_i))}\right)\Big/\frac{1}{N}\sum_{i=1}^{N}\hat{e}(\mathbf{X}_i). \tag{7}$$

Properties of the inverse probability weighting estimator using nonparametric estimation of the propensity score have been studied by Hirano, Imbens, and Ridder (2003); see also Wooldridge (2002) for parametric models of the propensity score.

To make inferences concerning the population parameters β and γ we rely on asymptotic properties of the estimator used. There is an efficiency bound for nonparametric estimators of the average treatment effects. Formally, Hahn (1998) showed that for an estimator $\hat{\beta}$ such that

$$\sqrt{N}(\hat{\beta}-\beta) \xrightarrow{d} \mathcal{N}(0,V),$$

we have that

$$V \geq E\left(\frac{V(Y_1 \mid \mathbf{X})}{e(\mathbf{X})} + \frac{V(Y_0 \mid \mathbf{X})}{1-e(\mathbf{X})} + (\beta(\mathbf{X})-\beta)^2\right). \tag{8}$$

Similarly for an estimator $\hat{\gamma}$ such that

$$\sqrt{N}(\hat{\gamma} - \gamma) \xrightarrow{d} \mathcal{N}(0, V),$$

we have

$$V \geq E\left(\frac{e(\mathbf{X})V(Y_1 \mid \mathbf{X})}{E(e(\mathbf{X}))^2} + \frac{e(\mathbf{X})^2 V(Y_0 \mid \mathbf{X})}{E(e(\mathbf{X}))^2(1 - e(\mathbf{X}))} + \frac{(\beta(\mathbf{X}) - \gamma)^2 e(\mathbf{X})}{E(e(\mathbf{X}))^2}\right). \tag{9}$$

In the sequel we call estimators $\hat{\beta}$ and $\hat{\gamma}$ asymptotically efficient if they are $\sqrt{N}$-consistent, asymptotically normal and reach the asymptotic efficiency bounds (8) and (9) respectively. Among efficient estimators we have estimators of the form (4) but where the value of the missing potential oucome is an estimated conditional mean from a nonparametric regression, see e.g. the estimator proposed by Hahn (1998). Alternatively both the observed and the missing potential outcome can be imputed from a nonparametric regression. Such an estimator based on series estimation of the regression functions $\mu_1(\mathbf{X})$ and $\mu_0(\mathbf{X})$ is studied in Imbens, Newey, and Ridder (2005). The inverse probability weighting estimators $\hat{\beta}_{\text{IPW}}$ and $\hat{\gamma}_{\text{IPW}}$ are also asymptotically efficient when the propensity score is estimated with a sieve approach using a nonparametric series logit estimator, see the results by Hirano, Imbens, and Ridder (2003).

Matching estimators where the matching criterion is the covariate norm is not generally $\sqrt{N}$-consistent. Abadie and Imbens (2006) show that the estimator is asymptotically biased if the number of continuous covariates is larger than one. Because the propensity score is a scalar function of the covariates the results of Abadie and Imbens (2006) imply that the estimators (4) and (5), where matching is performed on the propensiy score, are $\sqrt{N}$-consistent and asymptotically normally distributed. However, even when the matching estimator is $\sqrt{N}$-consistent it is generally not asymptotically efficient for a fixed number of matches. Asymptotic comparisons between

covariate matching and propensity score matching is further studied by Frölich (2007). Lower bounds for the variance are compared if the matching estimators are $\sqrt{N}$-consistent. Here, it is shown that matching on the true propensity score is less efficient than matching on the covariates directly.

A class of estimators called doubly robust combines regression and weighting using parametric models for both the propensity score and the outcome regression. The double robustness refers to the estimators being consistent and asymptotically normally distributed when at most one of the models are misspecified (see e.g. Bang and Robins 2005 for a review of doubly robust estimators).

Even though asymptotic results provide a tool for the comparison of alternative estimators these results can not be directly translated to the finite sample performance of the estimators. In Frölich (2004) the finite sample properties of matching and weighting estimators were studied by simulations. The mean squared error was calculated for nonparametric regression estimators and then compared towards a benchmark matching estimator. In the simulations both optimal and datadriven bandwidth choices were applied. There were substantial efficiency gains from the regression estimators for a variety of simulation designs. In all of the designs the largest MSE were observed from the weighting estimator.

The finite sample behaviour of propensity score matching compared to covariate matching is studied by Zhao (2004) and Angrist and Hahn (2004). The latter use a panel framework to show that even if propensity score matching is asymptotically less efficient than covariate matching there may be gains in efficiency for propensity score matching in finite samples.

The asymptotic properties shown for the estimators discussed above do not offer complete guideance to which estimator should be preferred in practice. All of the estimators mentioned above involve the estimation of the two conditional expectations $\mu_1(\mathbf{X})$, $\mu_0(\mathbf{X})$

and/or the propensity score $e(\mathbf{X})$. To identify the average causal effect we often have a covariate vector of high dimension. This implies that the methods studied for the nonparametric regression of $\mu_1(\mathbf{X})$ and $\mu_0(\mathbf{X})$ and the propensity score $e(\mathbf{X})$ for asymptotically efficient estimators of β and γ are not feasible in practice. For an unknown propensity score, parametric models are often used and their apropriateness are evaluated by how well they achieve balance in the covariates between treated and controls, see for instance Rosenbaum and Rubin (1984) and Dehejia and Wahba (1999, 2002).

3 Causal inference and graphical models

A graph is a pair $\mathcal{G} = (V, E)$ where V is a set of vertices and E is a set of edges. A graphical model displays a multivariate statistical model in which a joint distribution satisfies independence statements depicted in the graph. In a graphical model the set V represents random variables, and the set E describes dependencies among the variables. For a theoretical overview of graphical modeling and their applications see, for instance, the textbooks by Lauritzen (1996), Cox and Wermuth (1996) and Edwards (2000).

The graphical models of concern in this thesis are directed acyclic graphs (DAGs) and ancestral graphs. A DAG is a graph where the edge set E consists of directed edges, drawn as arrows, such that there are no directed cycles. The absence of directed cylces is equivalent to the existence of a partial ordering of the vertices. The ordering is such that arrows point only from lower order vertices to higher order vertices. From an applied perspective the partial order corresponds to some ordering of the variables given by subject-matter knowledge. An ancestral graph is an extension of a graphical model, e.g. a DAG, allowing for unobserved variables (Richardson and Spirtes 2002). Ancestral graphs display the inde-

pendence structure that results from having unmeasured variables that are not explicitly included in the model. A graphical implication of accounting for unobserved variables is the introduction of bi-directed edges. The bi-directed edges represent the existence of unobserved variables without precise assumptions on their relationships to other unobserved and observed variables. Also, an ancestral graph may have an undirected edge for cases when an unobserved variable has been conditioned on, see Richardson and Spirtes (2003) for a description of underlying independence relations of unobserved variables represented in an ancestral graph.

An alternative formalization of causality for statistical inference is DAGs describing assumptions on causal structures between variables, see e.g. Spirtes, Glymour, and Scheines (1993), Pearl (1995) and Pearl (2000). This connection between a DAG and causality is the foundation of another framework of causal inference (Pearl 1995; Pearl 2000). In this framework a causal effect on a system of variables is defined by how these variables are affected by an intervention. For this purpose a calculus of intervention has been defined (Pearl 1995). The calculus of intervention described by, e.g., Lauritzen (2001) and Dawid (2002), is such that it differentiates conditioning by observation from conditioning by intervention. Using the definition of intervention conditioning one can characterize the class of graphs where a causal effect can be identified from observational data.

In this thesis a DAG is interpreted as a carrier of independence assumptions and a causal effect is solely defined through the potential outcomes, i.e., within the Rubin causal model. The assumptions made are expressed as two partially specified graphs, $\mathcal{G}_j^R$, of the variables $\mathbf{X}$, T and Y_j, $j = 0, 1$. Other graphical representations of the Rubin causal model have been made by, e.g., Pearl (1993) and Edwards (2000, Sec. 8.2.2). Here, we assume a partially specified graph of the variables of the Rubin model such that the arrows of

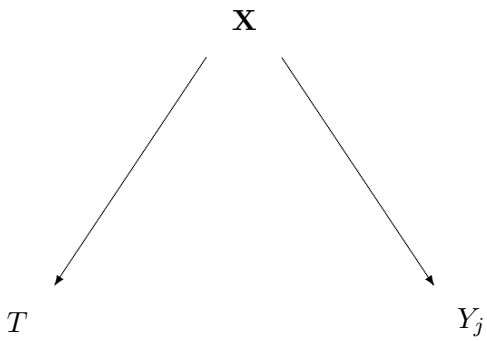

Figure 1: Graph, $\mathcal{G}_j^R$, $j = 0, 1$ of the Rubin causal model.

the graph are unspecified with exceptions: i) If there is an arrow between a vertex $R \in \mathbf{X}$ and T then $R \to T$, and ii) if there is an arrow between a vertex $R \in \mathbf{X}$ and Y_j, then $R \to Y_j$, for $j = 0, 1$. In the partially specified graphs, $\mathcal{G}_j^R$, $j = 0, 1$, the assumption that $\mathbf{X}$ are pre-treatment variables is expressed through the abscence of arrows pointing from T or Y_j to $\mathbf{X}$, see an example in Figure 3. The graph in Figure 3 is a DAG and/or an ancestral graph. Note, that it is a simplification of the graphical assumption described above since there may be arrows within $\mathbf{X}$ of any directions.

4 Summary of papers

This work consists of four papers in which the first three papers focus on reducing the set of covariates to condition on when estimating an average causal effect nonparametrically. In paper I and II we define covariate subsets necessary to condition on for identifying an average causal effect. A set of covariates is defined as minimal when the

treatment ceases to be unconfounded given any proper subset of the minimal set of covariates. We show that the covariate sets we define are minimal under certain assumptions and that they can be identified with the observed data. In paper I we assume that all relevant covariates are observed, i.e., there are no unmeasured confounders. We also assume that the variables in the Rubin causal model can be described by a partially specified DAG. In paper II we allow for the presence of unobserved variables within certain restrictions that are given through an assumption of a partially specified DAG. In paper III the results from paper I and II are implemented through a model-free variable selection method. Further, finite sample properties of nonparametric estimators utilizing the covariate sets defined in paper I and II are studied via simulations. In paper IV the general focus is changed from the selection of covariate sets to the functional form within which the covariates are conditioned on. Here, we study the consequences of misspecifications of parametric models of the propensity score for the estimation of an average causal effect.

4.1 Paper I: Covariate selection for non-parametric estimation of treatment effects

In this paper we assume that the variables in the Rubin causal model can be represented by a partially specified DAG, $\mathcal{G}_j^R$. The DAG is used to define minimal conditioning sets for independence of the treatment and the potential outcomes. The minimal conditioning sets have the important property that they can be selected by using subsets of the data within which all variables are observed. The results form the basis of covariate selection procedures proposed in two general algorithms. In the first algorithm, in a first step, the variables affecting the treatment are selected. Among the covariates selected from the first step we select the variables that affects the outcome for each treatment group respectively. In the second

algorithm, the two steps are taken in the reverse order.

A practical implementation of the algorithms are suggested. Here, we assume that the covariates affect the treatment and the potential outcomes respectively only through the mean function. The procedures proposed are then applied to select covariate sets in an application where the treatment is an employment program. The resulting covariate sets are then used to estimate the treatment effect of the program on the population of treated with a matching estimator. Each treated unit is matched to a control unit on the basis of the Mahalanobis' distance. Here, we observe a reduced mean distance for the matched pairs when using the minimal covariate sets selected with the proposed algorithms.

4.2 Paper II: Identification of minimal sets of covariates for the non-parametric estimation of an average treatment effect

In Paper II, assuming that the treatment is unconfounded given a set of covariates, we define subsets of the original covariate set such that unconfoundedness still holds. Earlier results exists on efficiency bounds for nonparametric estimators of β and γ (Hahn 1998). Also, so called exclusion restrictions, have been explored concerning sets of covariates affecting exclusively either the treatment or the outcome variables (Hahn 2004). Using the results of the exclusion restrictions we can compare asymptotic variance bounds of estimators when estimation is based on the covariate subsets that we define. Here, we show that using the defined covariate subsets may result in efficiency gains.

Under a graphical condition we can show that the defined covariate subsets are minimal. The condition is such that the variables $(\mathbf{U}, \mathbf{X}, T, Y_j)$, $j = 0, 1$, where $\mathbf{U}$ is a set of unobserved variables, can be described by a partially specified DAG. Moreover, we define a

covariate set as identified if it fulfills a condition that can be checked with the observed data. We show under general assumptions that the defined covariate subsets are identified.

4.3 Paper III: Model-free variable selection in causal inference

In the third paper we apply a model-free variable selection approach from the framework of sufficient dimension reduction in regression (Cook 1994; Cook 1996; Cook 1998). Here, we use marginal coordinate hypothesis tests, see Cook (2004), applied in each step of the general covariate selection algorithms proposed in paper I. The marginal coordinate tests are compared to standard model based procedures in a simulation study. In the simulations, the model based selection procedure fail in selecting covariates when the functional form of the model is misspecified. For those models a method using marginal coordinate hypothesis tests in a backward elimination procedure succeeds in selecting covariate sets upholding unconfoundedness.

Moreover, we study by simulations, the finite sample properties of three nonparametric estimators when estimation is based on the subsets of covariates defined and characterized in paper I and II. Estimation is performed by simple matching with replacement, regression imputation and inverse probability weighting. Coherent with the large sample results for the simple matching estimator, see Abadie and Imbens (2006), we observe a reduced bias when matching is performed on the covariate norm of the reduced subsets. For all three estimators the largest variance is obtained when following the common practice to condition on all covariates affecting the treatment.

The added uncertainty of using the covariate sets selected from the samples does not have a large impact on the estimated treatment

effect. The overall patterns from estimation based on the true covariate sets are preserved when the covariates are selected with the proposed marginal coordinate hypothesis tests.

4.4 Paper IV: Propensity score model specification for estimation of average treatment effects

Paper IV focuses on misspecifications of parametric models for the propensity score, the probability of being treated given the covariates, when estimating an average treatment effect. We assume that the propensity score can be described by a parametric model, represented by a strictly montonic transformation (link function) of a polynomial function of the covariates. We investigate misspecifications of parametric models of this form that are balancing scores, i.e., where the true propensity score is a function of the misspecification. When the misspecification is a balancing score, unbiased treatment effect estimators are available. It is shown that choosing the wrong link function does not introduce bias to the treatment effect estimator. Also, we describe misspecifications that are balancing scores when the order of the polynomial in the true model is higher than the order of the polynomial in the misspecification.

By decomposing the variance of a treatment effect estimator on a general form we show that the estimator is not necessarily less efficient if the propensity score model is misspecified.

A common model specification approach for propensity scores is to use the balancing property of the propensity score. Such a model specification procedure implies that the true propensity score cannot be distinguished from any other balancing score. In simulations we observe that the lack of distinction between the true propensity score and other balancing scores is not essential when using estimators that utilise the propensity score as a balancing score, e.g. estimators matching on the propensity score. On the other hand, when using

the propensity score as a probability, e.g., as in the inverse probability weighting estimator, we observe bias when the propensity score model is misspecified.

5 Conclusions and further research

Causal inference in observational studies have been an intense research area in statistics and econometrics during the last decades, see e.g. the review by Imbens (2004). To identify a causal effect in an observational study we rely on an assumption that the treatment is unconfounded conditional on a set of covariates. The assumption of an unconfounded treatment is an assumption not testable with the observed data but based on knowledge of the subject matter. In applications when we have access to large databases such as registers there is typically a large set of covariates available for the researcher. In such cases there is a demand for methods where subject matter knowledge can be combined with data driven procedures. The need to reduce the set of covariates is also emphasized by recent results concerning the bias and efficiency of nonparametric estimators of treatment effects (Abadie and Imbens 2006; Hahn 2004).

In this thesis it is shown that if the untestable assumption of unconfoundedness holds for a given set of covariates we can use the observed data to seek a reduction of the original covariate set. For this purpose we propose algorithms for covariate selection.

Using the model-free covariate selection methods from the framework of sufficient dimension reduction introduced also the possibility of reducing dimensions of the covariate vector not only by removing the covariates themselves but, in addition, to use the dimension reduction from the central subspace spanned by the remaining variables, see e.g. Cook (1996) for a definition of the central subspace. Studying a treatment effect estimator utilizing an estimated basis

for the central subspace is an interesting topic for future research. Such an estimator could also be implemented straightforwardly in combination with a covariate selection procedure using marginal coordinate hypothesis tests.

Another question that has been raised in connection to selection of minimal sets of covariates is the effect of the dimension of the covariate vector on treatment effect estimators utilizing the propensity score. Rosenbaum and Rubin (1983) showed that unconfoundedness is fulfilled when conditioning on a scalar function of the covariates, the propensity score. However, in simulations we have seen that the balancing property of the propensity score deteriorates if there are redundant covariates in the propensity score model. A research direction of interest would be to show formal results on the effect of redundant covariates on the finite and infinite sample behavior of propensity score based estimators.

When studying bias and efficiency of treatment effect estimators we have had much gain from using graphical diagnostic methods, e.g., plots of causal effects estimates conditional on covariates or propensity scores. Such plots have been very helpful in studying a range of topics from model misspecification to the detection of influential observations. Formalizing and studying tools for such graphical diagnostics is another future direction of research.

References

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

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

Bang, H. and J. M. Robins (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics 61*, 962–972.

Cook, R. D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association 89*, 177–189.

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

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

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

Edwards, D. (2000). *Introduction to graphical modelling*. New York: Springer-Verlag.

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

Frölich, M. (2007). On the inefficiency of propensity score matching. *Advances in Statistical Analysis 91*, 279–290.

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

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

Holland, P. and D. B. Rubin (1988). Causal inference in retrospective studies. *Evaluation Review 12*, 203–231.

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., W. Newey, and G. Ridder (2005). Mean-squared error calculations for average treatment effects. *IEPR Working Paper*, 05:34, Institute of Economic Policy Research, University of Southern California.

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*, London: Chapman and Hall, pp. 63-107.

Neyman, J. (1923). On the application of probability theory to agricultural experiments, essay on principles. Roczniki nauk Rolczych *X*, 1-51. In Polish, English translation by D.M.

Dabrowska and T. P. Speed in *Statstical Science*, *5*, 465-472, 1990.

Pearl, J. (1993). Aspects of graphical models connected with causality. In *Proceedings of 49th Session of the International Statistics Institute*, Florence, Italy, pp. 391–401.

Pearl, J. (1995). Casual diagrams for empirical research. *Biometrika 82*, 669–688.

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

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

Richardson, T. S. and P. Spirtes (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.

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

Rosenbaum, P. R. and D. B. Rubin (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association 79*, 516–524.

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

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

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomisation. *The Annals of Statistics 7*, 34–58.

Rubin, D. B. (1980). Comment on: Randomization analysis of experimental data: The Fisher randomization test. *Journal of the American Statistical Association 75*, 591–593.

Spirtes, P., C. Glymour, and R. Scheines (1993). *Causation prediction and search*. New York: Lecture Notes in Statstics, No. 81. Springer Verlag.

Wooldridge, J. (2002). Inverse probability weighted m-estimators for sample selection, attrition, and stratification. *Portugese Economic Journal 1*, 117–139.

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