Computational Statistics and Data Analysis 105 (2017) 280–292

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

# Data-driven algorithms for dimension reduction in causal inference

Emma Persson *, Jenny Häggström, Ingeborg Waernbaum, Xavier de Luna

*Department of Statistics, USBE, Umeå University, SE-90187 Umeå, Sweden*

ARTICLE INFO

*Article history:*
Received 12 October 2015
Received in revised form 19 August 2016
Accepted 19 August 2016
Available online 27 August 2016

*Keywords:*
Covariate selection
Marginal co-ordinate hypothesis test
Matching
Kernel smoothing
Type 1 diabetes mellitus

ABSTRACT

In observational studies, the causal effect of a treatment may be confounded with variables that are related to both the treatment and the outcome of interest. In order to identify a causal effect, such studies often rely on the unconfoundedness assumption, i.e., that all confounding variables are observed. The choice of covariates to control for, which is primarily based on subject matter knowledge, may result in a large covariate vector in the attempt to ensure that unconfoundedness holds. However, including redundant covariates can affect bias and efficiency of nonparametric causal effect estimators, e.g., due to the curse of dimensionality. Data-driven algorithms for the selection of sufficient covariate subsets are investigated. Under the assumption of unconfoundedness the algorithms search for minimal subsets of the covariate vector. Based, e.g., on the framework of sufficient dimension reduction or kernel smoothing, the algorithms perform a backward elimination procedure assessing the significance of each covariate. Their performance is evaluated in simulations and an application using data from the Swedish Childhood Diabetes Register is also presented.

## 1. Introduction

We consider observational studies where the goal is to investigate the causal effect of a treatment on an outcome of interest. In such studies the effect of the treatment may be confounded with other variables that are associated with both the treatment and the outcome of interest. The causal effect of the treatment can be identified if all confounders are observed, which is an assumption commonly referred to as unconfoundedness or no unmeasured confounding. The assumption of unconfoundedness is not testable in general, and thus it must be based on subject matter knowledge. In applications where there is a rich set of pretreatment variables, referred to as covariates in the sequel, the unconfoundedness assumption may be more credible. Such applications are nowadays common in the medical and social sciences due to the increasing possibilities to link administrative and health registers at the individual level.

This paper proposes and studies methods for data-driven selection of sufficient covariate sets, i.e. sets of covariates such that unconfoundedness holds. There may be several different sets of sufficient covariates in any application. Tests for sufficiency of a given subset were proposed by Robins (1997) and further described in a graphical model setting in Greenland et al. (1999). Although the tests are useful, they require that the empirical researcher defines the specific subset to be tested. A directed acyclic graph (DAG) model can also be applied as a basis to evaluate the sufficiency of a covariate set (Pearl, 2009). Using a DAG places great demands on the researcher's knowledge, since all the relations between the observed variables need to be specified. This complete specification is not necessary, and de Luna et al. (2011) have emphasized conditional

[*] Corresponding author. Fax: +46 907866614.
*E-mail address:* emma.persson@umu.se (E. Persson).

http://dx.doi.org/10.1016/j.csda.2016.08.012
0167-9473/© 2016 Elsevier B.V. All rights reserved.

independence properties of the variables involved as a guidance for selection of sets (see also VanderWeele and Shpitser, 2011). A large part of the literature is concerned with the balancing property (Rosenbaum and Rubin, 1983; Rubin, 1997). This corresponds to identifying covariates whose distribution differs between treated and untreated, e.g., selecting relevant covariates for a propensity score model. Brookhart et al. (2006), Kelcey (2011) and Vansteelandt et al. (2012) study covariate selection in a parametric setting, where the association of the covariates with both the treatment and the outcome is considered. In such a parametric setting a variance reduction of the estimator is obtained even when the covariates included are associated with outcome and not with the treatment assignment. For a semi-parametric estimator, an inverse probability weighted (IPW) estimator, Lunceford and Davidian (2004) describe the variance reduction when adding a covariate in the propensity score model that is solely related to outcome. In similar contexts, different methods for simultaneous covariate selection and model fitting have been proposed by van der Laan and Gruber (2010); Hill (2012) and McCaffrey et al. (2004).

In general, the covariate set will have an influence on both the large and small sample properties of an estimator. For estimators of the average causal effect under unconfoundedness, using a subset containing all covariates predicting the outcome has advantages when it comes to efficiency (de Luna et al., 2011; White and Lu, 2011). However, knowledge of the reduced subset discarding covariates related to the outcome and not to the treatment, is sometimes necessary to reach a lower efficiency bound (Hahn, 2004).

For nonparametric estimators, bias typically dominates variance, where the former dramatically depends on the dimension of the covariate set (e.g., Abadie and Imbens, 2006). Therefore, it is important to keep the cardinality of the covariate vector as low as possible. de Luna et al. (2011) developed a theory for selection of minimal subsets of covariates which are sufficient for unconfoundedness, and proposed two general algorithms for covariate selection. In this paper, we build on this theory by proposing data-driven algorithms for the selection of sufficient covariate sets. In the case of continuous covariates we study the use of marginal co-ordinate hypothesis tests (Cook, 2004; Li et al., 2005) based on the theory of sufficient dimension reduction in regression (Cook, 1994, 1996), to find sufficient subsets. When discrete covariates are present, as it is often the case in applications, we study the use of a kernel smoothing method (Hall et al., 2004, 2007; Li et al., 2009). Other model-free covariate selection methods could be used to implement the algorithms. The approach is here to select covariates without making strong model assumptions with the final aim to estimate a causal effect nonparametrically, e.g. using matching (Rosenbaum and Rubin, 1983; Abadie and Imbens, 2006). However, for comparison, we also implement the algorithms using parametric models for the outcome and treatment in combination with AIC and LASSO (Friedman et al., 2010) selection.

We study the finite sample properties of the algorithms, where continuous, discrete and mixed continuous–discrete sets of covariates are considered. In particular, the properties of matching estimators and an IPW estimator, are studied when sufficient covariate subsets are selected with the described algorithms. For instance, smaller mean squared errors (MSE) are achieved when using our algorithms, than when using all the covariates predicting the treatment assignment. In general, decreasing cardinality of the covariate set when possible yields better results. Covariate selection in the context of record linkage studies is illustrated in an application where we estimate the effect of low compulsory school grades on acute complications of Type 1 Diabetes Mellitus.

The paper is organized as follows. In the next section the theoretical framework and the covariate selection procedure are introduced. In Section 3, results from a simulation study are presented and an application concerning the effect of low compulsory school grades on acute complications of Type 1 Diabetes Mellitus can be found in Section 4. A discussion concludes the paper.

## 2. Covariate selection: context, theory and algorithms

### 2.1. Context

We consider a binary treatment, $T$, which will take on the value of 1 if treatment is received and 0 otherwise. For each unit we define two potential outcomes (Neyman, 1923; Rubin, 1974, 1977), $Y_1$ if the unit is treated and $Y_0$ if the unit is untreated. Only one of the potential outcomes can be observed for each unit, and we denote the observed response $Y$, where $Y = TY_1 + (1 - T)Y_0$. Let $X$ denote a set of covariates observed for all units. The parameter of interest considered is the average treatment effect,

$$\beta = \mathrm{E}(Y_1 - Y_0), \tag{1}$$

although the results presented below are useful for other summaries of the distribution of $Y_1 - Y_0$. In observational studies, where treatment assignment is not randomized, unconfoundedness, when it holds, allows us to identify causal parameters such as $\beta$. Consider the following assumptions,

A.1 [unconfoundedness] $Y_t \perp\!\!\!\perp T \mid X, t = 0, 1,$
A.2 [positivity] $\mathrm{P}(T = t \mid X) > 0, t = 0, 1.$

Under A.1 and A.2, the average treatment effect can be identified since $\beta = \mathrm{E}(Y_1 - Y_0) = \mathrm{E}[\mathrm{E}(Y_1 \mid T = 1, X) - \mathrm{E}(Y_0 \mid T = 0, X)]$. In this paper, we refer to assumption A.1 when discussing unconfoundedness, sometimes referred to as weak unconfoundedness (Imbens, 2000). In situations where A.1 and A.2 hold the parameter $\beta$ may be estimated nonparametrically by conditioning on the covariates $X$, e.g., using matching and/or IPW estimators; see, e.g., Imbens and Wooldridge (2009) for a review on estimation of causal effects under unconfoundedness.