Statistics and Probability Letters 81 (2011) 813–820

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Propensity score modelling in observational studies using dimension reduction methods

Debashis Ghosh

Departments of Statistics and Public Health Sciences, Penn State University, 514A Wartik Laboratory, University Park, PA, 16802, USA

ARTICLE INFO

Article history:
Available online 21 March 2011

Keywords:
Average causal effect
Matching
Model misspecification
Observational data
Potential outcomes

ABSTRACT

Conditional independence assumptions are very important in causal inference modelling as well as in dimension reduction methodologies. These are two very strikingly different statistical literatures, and we study links between the two in this article. The concept of covariate sufficiency plays an important role, and we provide theoretical justification when dimension reduction and partial least squares methods will allow for valid causal inference to be performed. The methods are illustrated with application to a medical study and to simulated data.

1. Introduction

In many medical and scientific studies, a major goal is understanding the relationship between a treatment and a response. Recently, there has been great interest in attempting to determine the causal effects associated with an intervention in an observational study setting. While the "gold-standard" approach would be to assess the intervention's effects using some type of randomized study design, in many situations this cannot be done because of logistic, economic, and/or ethical constraints.

In the literature on causal effect estimation, one of the major quantities that has played a central role is the propensity score (Rosenbaum and Rubin, 1983). This is the probability of receiving the treatment given a set of measured covariates. The causal effect estimation procedure normally proceeds in two steps. In the first step, the propensity score is modelled. Based on the estimated propensity score, the second step involves causal effect estimation. This can be done in a variety of ways, including matching, regression modelling, inverse probability weighted techniques and/or some combination thereof. A necessary condition for causal inference is termed the treatment ignorability assumption by Rosenbaum and Rubin (1983). It is described formally in Section 2.2, but the key observation is that the treatment ignorability assumption can be viewed as a conditional independence assumption involving the potential outcomes, the treatment and the covariates. Stone (1993) provides a nice discussion of what conditional independence assumptions are needed in order to make certain types of causal inferences.

There is another class of methods termed dimension reduction methods (Li, 1991) that involves conditional independence assumptions. Dimension reduction methodology has been a topic of intense research interest in the last twenty years, but for the purposes of exposition, we focus on sliced inverse regression (SIR) and partial least squares (PLS). The PLS method has been used primarily in the chemometrics literature; a comparison between partial least squares and dimension reduction methods was given by Naik and Tsai (2000). Much of this literature has focused on assessing the ability of the estimation procedures to capture the structure of the central subspace. These terms are more carefully defined in Section 2.2. However, in the causal inference framework, the central subspace is not our target estimand of interest. Rather, we use the output of the fitted model using PLS or SIR into a second regression model in order to estimate the average causal effect, defined in Section 2.1. We argue, mainly using simulation studies, that for the purposes of estimating causal

E-mail address: ghoshd@psu.edu.

0167-7152/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2011.03.002

effects, both PLS and SIR give fitted probabilities, or functionals thereof, that yield causal effect estimators with good finite-sample performance. Thus, misestimation of dimension reduction procedures, or equivalently, violation of distributional assumptions, appears to have very little effect on the average causal effect estimator.

A second goal of the paper is to use the ideas of conditional independence and in particular covariate sufficiency (Dawid, 1979) as a way to link dimension reduction methods to causal inference. This approach was also used by Nelson and Noorbaloochi (2009) in order to define what they term "dimension reduction summaries". In particular, some of the assumptions needed for validity of dimension reduction methods tie in nicely with a matching property used in causal inference termed *equal percent bias reduction* (EPBR) (Rubin and Thomas, 1992; Rubin and Stuart, 2006). The structure of this paper is as follows. In Section 2, we describe the observed data structures and review both the conditional independence assumptions needed for the causal inference, dimension reduction and partial least squares methods. In Section 3, we describe our proposed algorithm in conjunction with attendant theoretical properties of the proposed method. Application to a real dataset along with results from a limited simulation study are given in Section 4. We conclude with some discussion in Section 5. Proofs of the results are presented in a web Appendix.

2. Background and preliminaries

2.1. Data structures, causal estimands and conditional independence assumptions

Let the data be represented as $(Y_i, T_i, \mathbf{Z}_i),\ i = 1, \dots, n$, a random sample from the triple $(Y, T, \mathbf{Z})$, where Y denotes the response of interest, T denotes the treatment group, and Z is a p-dimensional vector of covariates. We assume that T takes the values $\{0, 1\}$. We adopt the causal inference framework that has been discussed by several other authors (Rubin, 1974; Holland, 1986). If we were given the counterfactuals $(Y(0), Y(1))$ for all n subjects, then we would be able to define causal effects, which are within-individual contrasts between the counterfactuals. In particular, given $(Y_i(0), Y_i(1)),\ i = 1, \dots, n$, we define the average causal effect:

$$\mathrm{ACE} = n^{-1}\sum_{i=1}^{n}\{Y_i(1) - Y_i(0)\}. \tag{1}$$

The standard assumption necessary for causal inference will be made:

$$T \perp \{Y(0), Y(1)\}|\mathbf{Z}, \tag{2}$$

i.e. treatment assignment is conditionally independent of the set of potential outcomes given covariates. This is the strong unconfounding or treatment ignorability assumption made by Rosenbaum and Rubin (1983). They then proposed the use of the propensity score for estimation of causal effects in observational studies. The propensity score is defined as

$$e(\mathbf{Z}) = P(T = 1|\mathbf{Z}) \tag{3}$$

and represents the probability of receiving treatment as a function of covariates. Use of the propensity score leads to balance in covariates between the groups with $T = 0$ and $T = 1$. Statistically, this corresponds to the conditional independence of T and $\mathbf{Z}$ conditional on $e(\mathbf{Z})$ and is summarized in Theorem 1 of Rosenbaum and Rubin (1983). Given the treatment ignorability assumption in (2), it also follows by Theorem 3 of Rosenbaum and Rubin (1983) that treatment is strongly ignorable given the propensity score, i.e.

$$\mathbf{Z} \perp \{Y(0), Y(1)\}|e(\mathbf{Z}).$$

Typically, the model fit for (3) involves a high-dimensional covariate vector. This has usually been done based on logistic regression. Logistic regression specifies the effects of covariates on the probability of treatment in a completely parametric manner. Given that the output from the model is the fitted values, the case can be made for adopting more flexible models of treatment. One such generalization is given in the next section.

2.2. Dimension reduction methods

Suppose we formulate a semiparametric model for the propensity score as

$$e(\mathbf{Z}) = g(\beta'\mathbf{Z}, u), \tag{4}$$

where β is a p-dimensional vector of unknown regression coefficients, u is an error term, and g is an unspecified link function. Because of the nonparametric nature of the link function, model (4) is semiparametric. Thus, (4) represents a flexible extension of the logistic regression model for propensity scores. Note that linear, logistic and log-linear regression models are special cases of (4). It is also an example of a single-index model in that the information about the covariate effects on the response is completely captured through the linear predictor, or equivalently, single-index $\beta'\mathbf{Z}$.

The starting point of dimension reduction methods is the conditional independence of T and $\mathbf{Z}$ given $e(\mathbf{Z})$. An implication of model (4) being true is that there exists a $p \times 1$ vector $\mathbf{B}$, where

$$T \perp \mathbf{Z}|\mathbf{B}'\mathbf{Z}. \tag{5}$$