2011

Propensity score modelling in observational studies using dimension reduction methods

Debashis Ghosh, *Penn State University*

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, U.S.A.

ghoshd@psu.edu

Summary

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.

Keywords: Causal effect; counterfactual; link-free inference; matching; model misspecification; observational data; potential outcomes.

1 Introduction

In many medical and scientific studies, a major goal is understanding the relationship between a treatment with 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. Lunceford and Davidian (2004) provide a nice review of various modelling approaches for causal inference based on the propensity score. In this article, we focus on the use of matching techniques.

For causal inference, the most commonly used model is the so-called potential outcomes model (Rubin, 1974; Holland, 1986). We will assume that our treatment is binary and that for each subject i, one can formulate the counterfactual variables $\{Y_i(0), Y_i(1)\}$, $i = 1, \ldots, n$. We observe $Y_i = Y_i(T_i) \equiv T_i Y_i(1) + (1 - T_i)Y_i(0)$ $(i = 1, \ldots, n)$, where $T_i \in \{0, 1\}$ denotes the observed treatment for the ith individual. Note that only one of the potential outcomes is observed for each individual. The causal effect is based on the individual-level contrasts $Y_i(1) - Y_i(0)$, $i = 1, \ldots, n$. 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. The sliced inversion regression (SIR) algorithm proposed by Li (1991) used a combination of stratification of the response variable in conjunction with calculation of a slice-weighted correlation matrix of the predictors and its singular value decomposition. 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 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). More recently, Li et al. (2007) combined PLS with SIR and developed a general algorithm that they term partial inverse regression, and Cook et al. (2007) developed a general sufficient dimension reduction method for so-called multi-index models based on combining PLS with dimension reduction methods. 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 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) developed by Rubin and colleagues (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 causal inference, dimension reduction and partial least squares methods. In Section 3, we describe our proposed algorithm in conjunction and present some 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.

2 Background and Preliminaries

2.1 Data Structures, Causal Estimands and Conditional Independence Assumptions

Let the data be represented as $(Y_i, T_i, \mathbf{Z}_i)$, $i = 1, \ldots, n$, 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). The issue of existence of counterfactual variables is a controversial one in certain contexts; see Holland (1986) and Dawid (2000) for more discussion on this. Throughout this manuscript, we will be assuming the existence of counterfactual variables.

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 $(Y_i(0), Y_i(1))$, $i = 1, \ldots, n$, we define the average causal effect:

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

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) in their seminal paper and which allows for the estimation of causal effects.

Rosenbaum and Rubin (1983) 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 Theorexm 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

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

Another way of stating (5) is that the projection $\mathbf{B}'\mathbf{Z}$ provides a sufficient data reduction and contains the essential information about the relationship between T and $\mathbf{Z}$. More generally, we can define a projection operator $\mathbf{P}_B$ to be the projection operator onto the subspace spanned by the columns of $\mathbf{B}$. Then (5) can be reexpressed as

$$T \perp \mathbf{Z} | \mathbf{P}_B\mathbf{Z}. \tag{6}$$

If (6) holds, then it also holds for any subspace $\mathbf{C}$ such that the span of $\mathbf{B}$ is the same as the span of $\mathbf{C}$. Let $S(\mathbf{B})$ be the subspace generated by the columns of $\mathbf{B}$. Let S_* denote the intersection of all possible subspaces; if S_* is also a subspace, i.e. it satisfies (6), then we will refer to S_* as the central subspace (Li, 1991; Cook, 1998). We will assume throughout that the central subspace exists (Cook, 1998; Yin, Li and Cook, 2008). We will now discuss methods of estimating this quantity.

2.3 Sliced Inverse Regression

We assume, without loss of generality, that $\mathbf{Z}$ has mean zero vector and covariance matrix equal to the identity matrix. One key assumption is necessary for implementation of SIR: (A1) the distribution of $\mathbf{Z}$, conditional on $\mathbf{P}_B\mathbf{Z}$, satisfies a conditional linearity in the mean, i. e.

$$E(\mathbf{Z}|\mathbf{P}_B\mathbf{Z}) = \mathbf{P}_B\mathbf{Z}.$$

Assumptions (A1) pertains to the marginal distribution of $\mathbf{Z}$ and means that all the information about $\mathbf{Z}$ is contained in its projection onto the subspace spanned by B. One class of distributions that satisfy assumptions (A1) is the family of elliptically symmetric distributions. This includes distributions such as the multivariate normal distributions and scale mixtures of multivariate normal distributions. The SIR method of Li (1991) is implemented as follows:

1. Standardize the predictor observations as

$$\tilde{\mathbf{Z}}_i = \hat{\Sigma}^{-1/2}(\mathbf{Z}_i - \hat{\mu}), (i = 1, \ldots, n),$$

   where $\hat{\mu}$ and $\hat{\Sigma}$ are the sample mean and covariance matrices of $Z_1, \ldots, Z_n$.

2. Calculate sample mean estimates within dose groups:

$$\bar{\mathbf{Z}}_d = \frac{1}{n_d} \sum_{i=1}^{n} I(T_i = d)\tilde{\mathbf{Z}}_i,$$

   where $n_d = \sum_{i=1}^{n} I(T_i = d)$, $d = 0, 1$.

3. Estimate the population covariance matrix of $\mathbf{Z}$ given T by

$$\hat{\Theta} = \sum_{d=0}^{1} \frac{n_d}{n} \bar{\mathbf{Z}}_d \bar{\mathbf{Z}}_d'.$$

4. Calculate the eigenvalues of $\hat{\Theta}$. These are the estimates of the basis vectors for the minimal propensity score subspace. Take $\hat{t}_1$ to be the first eigenvector of the matrix.

One would then compute $\hat{\beta} = \hat{\Sigma}^{-1/2}\hat{t}_1$ as the estimate of the regression coefficient vector from (3). This algorithm is termed "inverse regression" because effectively, information on the "backwards regression" $E(\mathbf{Z}|T)$ is being estimated here rather than the "forward regression" $E(T|\mathbf{Z})$. Li (1991) argues that this approach circumvents the usual issue of the curse of dimensionality. Other advantages of the sliced inverse regression algorithm is that it avoids multivariate nonparametric smoothing and is quite easy to fit. Some more intuition about the result from this algorithm is that it can be viewed as the solution to maximization of the correlation between the transformed response and the linear predictor (Chen and Li, 1998). To be specific, from Theorem 3.2. of Chen and Li (1998), we have that the estimated direction $\hat{\beta}_1$ is the solution to the following maximization problem:

$$\max_{b,H} Corr(H(T), b'\mathbf{Z}),$$

where H is taken over all transformations of the response. Assumption (A1) guarantees the identifiability of β, up to a scale parameter. Thus, another interpretation of sliced inverse regression is that it achieves the goal of maximizing correlations between T and $\mathbf{Z}$ subject to transforming the response variable.

2.4 Partial least squares

A popular regression method in chemometrics is partial least squares. The partial least squares algorithms attempt to simultaneously find linear combinations of the predictors whose correlation is maximized with the response and which are uncorrelated over the training sample. There are several algorithms available for numerical estimation using partial least squares; a nice overview of these methods can be found in Denham (1994).

Recently, Naik and Tsai (2000) proposed the use of partial least squares for fitting single-index models of the form (3). The partial least squares they use is given by the following formula:

$$\hat{\beta}_{PLS} = \hat{R}(\hat{R}'\hat{\Sigma}\hat{R})^{-1}\hat{R}'\mathbf{w}, \tag{7}$$

where $\mathbf{w} = (n-1)^{-1}\sum_{i=1}^{n}(\mathbf{Z}_i - \bar{\mathbf{Z}})(T_i - \bar{T})$, $\bar{T} = n^{-1}\sum_{i=1}^{n} T_i$, and $\hat{R} \equiv (\mathbf{w}, \hat{\Sigma}\mathbf{u}, \ldots, \hat{\Sigma}^{q-1}\mathbf{w})$ is called a Kryvlov sequence. Under a multivariate normality condition on $\mathbf{Z}$ and some mild moment conditions, Naik and Tsai (2000) show that the partial least squares can be used for estimation of β in model (3). In the simulation studies, they compare the PLS estimator with the SIR estimator in the usual regression setting and finds that the latter method yields better performance.

3 Proposed Methodology

3.1 Propensity Score/Matching Algorithm

Suppose we have that the conditions needed for SIR or PLS to estimate the central subspace. This yields the following proposition:

Proposition 1: Suppose that (2) and (5) hold and $0 < P(T = 1|\mathbf{Z}) < 1$ for all $\mathbf{Z}$, then

$$(Y(0), Y(1)) \perp T|\mathbf{B}'\mathbf{Z}.$$

Proof: By Theorem 3 in Rosenbaum and Rubin (1983, p. 45), (2) and $0 < P(T = 1|\mathbf{Z}) < 1$ imply that $(Y(0), Y(1)) \perp T|e(\mathbf{Z})$. This in conjunction with (5) gives the desired result.

The implication of Proposition 1 is that estimation of the basis of the population version central subspace will provide a sufficient reduction of the data while maintaining balance in the distribution of the observed covariates. Since both SIR and PLS can estimate this

under certain assumptions, either algorithm should produce balancing scores for these cases. Given the estimated propensity score using either SIR or PLS, we are now in a position to assess the causal effect of T on Y. As discussed in Lunceford and Davidian (2004), several procedures are available for estimation. Here, we employ an approach based on matching. The idea of matching is to find for each subject with $T = 1$ the subject(s) with $T = 0$ whose propensity score is closest. Matching individuals from the treatment and control groups with very similar propensity scores has the effect of balancing the two populations based on the observed covariates $\mathbf{Z}$ as described for the population version of the propensity score in Theorem 1 of Rosenbaum and Rubin (1983).

An alternative approach would involve fitting a regression model of the outcome variable on T incorporating the estimated propensity score either as a covariate and/or a weight. Regression estimators are constructed using weighted estimation of the differences between the propensity scores for $T = 1$ and $T = 0$ and typically involve some degree of extrapolation. By contrast, matching estimators do not involve any such extrapolation. The extrapolation can be problematic if the $T = 1$ and $T = 0$ have radically different propensity score distributions. This is why we pursue the matching approach here.

One issue that arises is how many subjects with $T = 0$ should be matched to each individual in the $T = 1$ group. In classical epidemiology, typically 1:1 matching (i.e., one person from $T = 1$ to one person with $T = 0$) has been advocated. More recently, arguments have been made for using more complex matching structures, such as full matching (Rosenbaum 2002, §10.4). With the recent availability of these types of methods in the R programming language (Hansen, 2004), we will use a full matching algorithm in the numerical examples in Section 4. To summarize, our algorithm for estimating causal effects is as follows:

1. Estimate β in (3) using either SIR or PLS, and denote the resulting estimator by $\hat{\beta}$.
2. Based on $\hat{\beta}'\mathbf{Z}$, perform a full matching of subjects from $T = 1$ and $T = 0$.
3. Estimate the average causal effects using the matched sample. Following the recommendation of Ho et al. (2007), we ignore the matching at this stage.

The issue of how to perform inference after matching is currently an open one and will not

be dealt with in the paper.

3.2 Theoretical properties

We now describe some theoretical aspects of our procedure. The first has to do with the conditional independence assumptions inherent in making causal inference statements. Before stating our proposition, we describe three types of causal null hypotheses that can be formulated in observational studies (Stone, 1993): letting $\mathbf{U}$ denote unobserved covariates,

1. No causation: $Y \perp T|(\mathbf{Z}, \mathbf{U})$: equivalent to $y = y(t, \mathbf{z}, \mathbf{u})$ does not depend on t for any $(\mathbf{z}, \mathbf{u})$.

2. No distribution effect: this means that $p\{y(t, \mathbf{Z}, \mathbf{U})|\mathbf{Z} = \mathbf{z}\}$ does not depend on t.

3. No mean effect: this means that $E\{y(t, \mathbf{Z}, \mathbf{U})|\mathbf{Z} = \mathbf{z}\}$ does not depend on t.

Proposition 2: *Suppose that (2) hold and that conditions hold for the validity of SIR or PLS for estimating the central subspace. Then the following results hold:*

(a) No association $\Leftrightarrow$ no causation.

(b) The no unmeasured confounders assumption (van der Laan and Robins, 2002) holds.

(c) The strongly ignorable treatment assumption (Rubin, 1974) holds.

Proof: The assumptions imply that the central subspace is sufficient, in the sense of Dawid (1979), for the conditional independence of the potential outcomes with T. Under this assumption, (a)-(c) follow directly from the discussion in Stone (1993).

As a simple consequence of Proposition 2, no association is equivalent to no distribution effect, and no mean difference is the same as no mean effect. While a very simple proposition to prove, Proposition 2 underlies the crucial role that covariate sufficiency (Dawid, 1979) plays.

We now describe another theoretical property for our procedure that relates to the matching stage. It deals with a concept termed equal percent bias reduction. Equal percent bias reduction means that matching will reduce bias in all dimensions of $\mathbf{Z}$ by the same amount. Matching in this setting also reduces bias in any function of the X's, including the outcome

of interest (Rubin and Thomas, 1992). Let $\mathbf{z}_0$ and $\mathbf{z}_1$ denote the $n_0 \times p$ and $n_1 \times p$ matrices of covariates for the subjects in the control and treatment populations. A matching method generically selects a subset of rows from the two data matrices. Let $\mathcal{M}_i$ denote the set of indices of the selected rows from $\mathbf{z}_i$, $i = 0, 1$. An *affinely invariant* matching method is one for which $\mathcal{M}_0$ and $\mathcal{M}_1$ do not change when we apply an affine transformation to $\mathbf{z}_0$ and $\mathbf{z}_1$. Let $\alpha'\mathbf{Z}$ denote a linear combination of $\mathbf{Z}$, where α is a unit-length vector, i.e., $\alpha'\alpha = 1$. One can then decompose the linear combination into components along and orthogonal to the best linear discriminant $U = (\mu_1 - \mu_0)'\Sigma_0\mathbf{Z}$. A matching method is said to be equal percent bias reducing (Rubin and Thomas, 1992, Corollary 3.2) if

$$\frac{E(\bar{Y}_{m1} - \bar{Y}_{m0})}{E(\bar{Y}_{r1} - \bar{Y}_{r0})} = \frac{E(\bar{U}_{m1} - \bar{U}_{m0})}{E(\bar{U}_{r1} - \bar{U}_{r0})}, \tag{8}$$

with the following notation:

$$\bar{Y}_{m1} = \frac{1}{m_1}\sum_{j\in\mathcal{M}_1} Y_j; \bar{Y}_{m0} = \frac{1}{m_0}\sum_{j\in\mathcal{M}_0} Y_j; \bar{U}_{m1} = \frac{1}{m_1}\sum_{j\in\mathcal{M}_1} U_j; \bar{U}_{m0} = \frac{1}{m_0}\sum_{j\in\mathcal{M}_0} U_j;$$

$$\bar{Y}_{r1} = \frac{1}{m_1^*}\sum_{j\in\mathcal{M}_1^*} Y_j; \bar{Y}_{r0} = \frac{1}{m_0^*}\sum_{j\in\mathcal{M}_0^*} Y_j; \bar{U}_{r1} = \frac{1}{m_1^*}\sum_{j\in\mathcal{M}_1^*} U_j; \bar{U}_{r0} = \frac{1}{m_0^*}\sum_{j\in\mathcal{M}_0^*} U_j;$$

where m_0 and m_1 are the cardinality of $\mathcal{M}_0$ and $\mathcal{M}_1$, respectively. In addition, M_1^* and M_0^* are the indices of m_1^* and m_0^* randomly chosen individuals with $T = 1$ and $T = 0$, respectively. The cardinalities have to match for the two populations, and the expectations in (8) are taken with respect to randomly selecting subjects with $T = 1$ and $T = 0$. Note that because $E(\bar{U}_{m1} - \bar{U}_{m0})/E(\bar{U}_{r1} - \bar{U}_{r0})$ takes the same value for all Y, (8) implies that the percent bias reduction is the same for any linear combination of $\mathbf{Z}$. We now have the following result:

Theorem 1: *(a) Suppose that* $\mathbf{Z}$ *has an elliptically symmetric distribution. Then an affinely invariant matching procedure based on* $\beta'\mathbf{Z}$*, where* β *is the in-limit probability of* $\hat{\beta}_{SIR}$ *or* $\hat{\beta}_{PLS}$*, is equal percent bias reducing.*

(b) Suppose that $\mathbf{Z}|T$ *has an elliptically symmetric distribution. Then an affinely invariant matching procedure based on* $\beta'\mathbf{Z}$*, where* β *is the in-limit probability of* $\hat{\beta}_{SIR}$ *or* $\hat{\beta}_{PLS}$*, is equal percent bias reducing.*

Proof: For (a), the assumption on $\mathbf{Z}$ implies that condition (A1) holds so that either estimate of β (SIR or PLS) will be estimating the same direction and will converge in probability to β. The condition on $\mathbf{Z}$ implies that Theorem 3.1. and Corollaries 3.1. and 3.2 of Rubin and Thomas (1992) apply so that any affinely invariant matching method will be equal percent bias reducing. For (b), the assumption on the conditional distribution of $\mathbf{Z}$ given T implies that the marginal distribution of $\mathbf{Z}$ will be a mixture of elliptically symmetric distributions. Again, condition (A1) holds so that either estimate of β (SIR or PLS) will be estimating the same direction and will converge in probability to β. The condition on $\mathbf{Z}|T$ implies that Theorem 3.1. of Rubin and Thomas (1992) and Corollaries 3.1. and 3.2 of Rubin and Stuart (2006) apply so that any affinely invariant matching method will be equal percent bias reducing.

4 Numerical examples

While we have shown that in theory validity of the PLS and SIR estimators for the single-index model will yield matching procedures that are EPBR, in practice the distributional assumptions embodied in assumption (A1) could be violated on real datasets. We illustrate the effects of the modelling strategy on real and simulated data that violate assumption (A1) to illustrate the robustness of the methodology.

4.1 SUPPORT study

In this example, we apply the proposed methodology to data from the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT). This is a multicenter observational trial that was designed to study the outcomes for patients who were hospitalized with serious conditions. The question we consider here is whether or not the treatment, right heart catherization (RHC), has an effect on 30-day survival (dead/alive at 30 days). Further information about the study can be found in Connors et al. (1996). The dataset contains information on 5735 patients, 2184 of whom received RHC. Table 1 lists the covariates we use in our analysis. This is a small subset of what is available and is meant for illustratory purposes. A simple analysis of death rates with the RHC treatment revealed

a strong association between RHC and death. Of the 2184 who received RHC treatment, 830 died, while among the 3571 subjects who did not receive RHC, 1088 died (risk difference = -0.074, p-value = $1e-08$). However, because RHC treatment was not randomized, there are in fact systematic differences in baseline covariates between the treated and nontreated groups. Some summary statistics are provided in Table 1. Based on the table, we find that patients who underwent RHC tended to be younger, male, more highly educated, have higher levels of serum bilirubin, be more likely to have respiratory infection and less likely to have immune disease and cardiovascular disease relative to the non-RHC subjects. Note also that four of the variables are binary, which is a violation of the assumptions typically needed for dimension reduction procedures to hold.

Based on the sliced inverse regression algorithm, we estimated the propensity score as proportional to

$$0.0049\text{age} - 0.17\text{sex} - 0.25\text{immunhx} - 0.31\log(\text{bili1}) + 0.41\text{resp} - 0.80\text{card} - 0.031\text{edu}.$$

For the partial least squares estimator, the first partial least squares component is

$$-0.012\text{age} + .018\text{sex} + 0.16\text{immunhx} + 0.046\log(\text{bili1}) - 0.052\text{resp} + 0.058\text{card} + .018\text{edu}.$$

Discrepancies from the two estimators are a result of the data distribution not being multivariate normal and/or elliptically symmetric. After performing a full matching based on the fitted values from the sliced inverse regression, we find that the risk difference associated with RHC changes sign (risk difference = 0.065, p-value = $3e-06$). Similar results are obtained after performing a full matching based on the fitted values from partial least squares (risk difference = 0.065, p-value = $1.5e-05$). These effects and associated p-values were obtained from a linear regression of 30 day survival on RHC treatment with adjustment for the matched sets from the full matching as a factored variable. Interestingly, even though neither PLS or SIR is expected to capture the directions underlying subspace structure correctly, both yield highly similar causal effect estimates.

4.2 Simulation studies

Next, we conducted limited simulation studies based on the same setup that was used by Lee et al. (2010), which is similar to the SUPPORT data in that the covariates are a mix of continuous and discrete variables. We now describe the simulation setup. We generated a covariate vector $\mathbf{W} \equiv (W_1, \ldots, W_{10})$, a 10-dimensional multivariate normal vector with mean zero and correlation matrix Σ^W with elements σ^W_{ij}, $1 \leq i, j \leq 10$. The off-diagonal elements are all zero except for the following: $\sigma^W_{15} = 0.2$, $\sigma^W_{26} = 0.9$, $\sigma^W_{38} = 0.2$, $\sigma^W_{49} = 0.9$. Then $(W_1, W_3, W_5, W_6, W_8, W_9)$ were dichotomized at their means; this yields the final covariate vector $\mathbf{Z}$, which is a mixture of discrete $(Z_1, Z_3, Z_5, Z_6, Z_8, Z_9)$ and continuous variables (Z_2, Z_4, Z_7, Z_{10}). We considered the following scenarios for the true propensity score model: letting $\beta^* = (0.8, -0.25, 0.6, -0.4, -0.8, -0.5, 0.7, 0, 0, 0)$,

1. Scenario A:

$$\text{logit}P(T = 1|\mathbf{Z}) = \mathbf{Z}'\beta^*$$

2. Scenario B:

$$\text{logit}P(T = 1|\mathbf{Z}) = \mathbf{Z}'\beta^* - 0.25Z_2^2$$

3. Scenario C:

$$\text{logit}P(T = 1|\mathbf{Z}) = \mathbf{Z}'\beta^* - 0.25Z_2^2 - 0.4Z_4^2 - 0.7Z_7$$

4. Scenario D:

$$\text{logit}P(T = 1|\mathbf{Z}) = \mathbf{Z}'\beta^* + 0.4Z_1Z_3 - 0.175Z_2Z_4 - 0.2Z_4Z_5 - 0.4Z_5Z_6$$

5. Scenario E:

$$\text{logit}P(T = 1|\mathbf{Z}) = \mathbf{Z}'\beta^* - 0.25Z_2^2 + 0.4Z_1Z_3 - 0.175Z_2Z_4 - 0.2Z_4Z_5 - 0.4Z_5Z_6$$

6. Scenario F:

$$\begin{aligned}\text{logit}P(T = 1|\mathbf{Z}) = \mathbf{Z}'\beta^* &+ 0.4Z_1Z_3 - 0.175Z_2Z_4 + 0.3Z_1Z_3 - 0.28Z_4Z_6 - 0.4Z_5Z_7 + 0.4Z_5Z_6 \\ &- 0.175Z_2Z_3 + 0.3Z_3Z_4 - 0.2Z_4Z_5 - 0.4Z_5Z_6\end{aligned}$$

These correspond exactly to simulation models A.) - F.) in Lee et al. (2010, p. 339) and represent differing levels of nonlinearity and interactions. For the outcome model, we assumed that the true effect of the treatment on the average response was $-.4$. We compared the properties of the resulting average causal effect estimators, where full matching, as described in Section 3.1, was used to calculate the average causal effect estimates. The propensity scores were estimated using partial least squares and sliced inverse regression. We generated 1000 simulation samples for each scenario, each with 1000 observations per simulated dataset. We evaluated the estimators based on the following properties: (1) average standardized absolute mean distance, a measure of covariate balance; (2) bias of treatment effect: the relative distance of the treatment effect; (3) standard error of the estimate of the treatment effect and (4) coverage probabilities of the 95% CIs.

The results for the average standardized absolute mean distance and standard errors of the treatment effect are given in Figures 1 – 2. Based on the pictures, we find that while SIR leads to better covariate balance relative to PLS, the standard errors of the treatment effect from PLS tend to be smaller. However, the bias of the estimated treatment effect was around 20% for all situations with PLS, while it was 1% for the SIR method. Neither method performed well in terms of coverage probabilities (data not shown); the method using PLS tended to have lower coverage relative to the nominal coverage, while using SIR led to higher than nominal coverage. It should be noted that the simulation model considered is not consistent with the assumptions for either SIR or PLS to estimate β consistently in (3). The simulations suggest that SIR still has utility as a means of reducing covariate imbalance and adjusting for confounding.

5 Discussion

In this article, we have used conditional independence as an approach to unify dimension reduction methods with propensity score modelling approaches to causal inference. Intuitively, we have assumed that $\mathbf{Z}$ constitutes "sufficient covariates" in the sense of Dawid (1979). This assumption, in conjunction with distributional assumption (A), can be used to justify the validity of dimension reduction procedures. Equivalently, the assumptions (2) and (5) are

sufficient for identifying $\mathbf{Z}$ as being sufficient covariates for performing causal inference in the sense of Stone (1993). However, the simulation studies show that even distributional assumptions are violated, the PLS and SIR algorithms estimate propensity scores that yield very similar average causal effect estimators, as in the SUPPORT study example. The simulation studies show that the SIR estimator performs better than the PLS estimator, so we would recommend the former over the latter in practice. Because the target of estimation is the average causal effect, it appears that misspecification of the dimension reduction has small impact.

The issue of propensity score model specification has attracted some attention in the statistical literature (Drake, 1993). While more recent work has focused on how to select variables for the propensity score model (e.g., Hirano and Imbens, 2001), we have focused here on fitting a particular semiparametric model. One related issue to what was discussed here is that of link misspecification in a model such as (3). Under condition (A1), fitting a linear regression of T on $\mathbf{Z}$ will identify β in (3) up to a scalar. However, this will not have any effect on propensity-score based matching estimators for the average causal effect, as the "order-preserving" property of propensity scores is maintained (Zhao, 2008).

Finally, we note that an important contribution was in thinking about the central role that conditional independence plays in both the causal inference and dimension reduction literatures. This link is meant to develop further cross-fertilization between these two methodologies.

Acknowledgments

The author would like to thank Brian Lee for making his simulation code publicly available, Yeying Zhu for useful discussions and the Penn State Causal Inference Working Group for helpful discussions on the topic. He also would like to thank the reviewer whose comments substantially improved this manuscript. This work was supported by the National Institute on Drug Abuse grant P50 DA010075. The content of this manuscript is solely the responsibility of the author(s) and does not necessarily represent the official views of the National Institute on Drug Abuse or the National Institutes of Health.

References

Chen, C. H., Li, K. C. 1998. Can SIR ever be as popular as multiple regression? *Statistica Sinica* 8, 298 – 316.

A. F. Connors et al. 2001. The effectiveness of right heart catheterization in the initial care of critically ill patients. *Journal of the American Medical Association* 276, 889 – 897.

Cook, R. D. 1998. Regression Graphics. Wiley, New York.

Cook, R. D., Li, B. and Chiaromonte, F. 2007. Dimension reduction in regression without matrix inversion. *Biometrika* , 569 – 584

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

David, A. P. 2000. Causal thinking without counterfactuals. Journal of the American Statistical Association 95, 407-424.

Denham, M. C. 1994. Implementing partial least squares. Statistics and Computing 5, 191 – 202.

Drake, C. 1993. Effects of misspecification of the propensity score on estimators of treatment effect. Biometrics 49, 1231 – 1236.

Hirano, K., Imbens, G. 2001. Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. Health Services and Outcomes Research Methodology 2, 259 – 278.

Hansen, B. 2004. Full matching in an observational study of coaching for the SAT. Journal of the American Statistical Association 99, 609 – 618.

Ho, D.E., Imai, K., King, G., and Stuart, E.A. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Political Analysis 15(3): 199-236.

Holland, P. 1986. Statistics and causal inference (with discussion). Journal of the American Statistical Association 81, 945 – 970.

Lee, B. K., Lessler, J., Stuart E. A. 2010. Improving propensity score weighting using machine learning. Statistics in Medicine 29, 337 – 346.

Li, K. C. 1991. Sliced inverse regression for dimension reduction (with discussion). Journal of the American Statistical Association 86, 316 – 342.

Li, L., Cook, R. D. and Tsai, C. L. (2007). Partial inverse regression. Biometrika, 94, 615-625.

Lunceford, J. K., Davidian, M. 2004. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in Medicine 23, 2937 – 2960.

Naik, P. A., Tsai, C. L. 2000. Partial least squares estimator for single-index models. Journal of the Royal Statistical Society Series B 62, 763 – 771.

Nelson, D., Noorbaloochi, S. 2009. Dimension reduction summaries for balanced contrasts. Journal of Statistical Planning and Inference 139, 617 – 628.

Rosenbaum, P. R. 2002. Observational Studies, 2nd ed. Springer-Verlag, New York.

Rosenbaum, P. R., 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., Stuart, E.A. 2006 Affinely invariant matching methods with discriminant mixtures of proportional ellipsoidally symmetric distributions. The Annals of Statistics 34, 1814-1826.

Rubin, D. B., Thomas, N. 1992. Affinely invariant matching methods with ellipsoidal distributions. The Annals of Statistics 20, 1079-93.

Rubin, D. B., Thomas, N. 1996. Matching using estimated propensity scores: relating theory to practice. Biometrics 52, 249 – 264.

Stone, R. 1993. The assumptions on which causal inferences rest. Journal of the Royal Statistical Society, Series B 55, 455-466.

van der Laan, M. J., Robins, J. M. 2002. Unified Approach to Censored Data and Causality. Springer-Verlag, New York.

X. Yin, B. Li and R.D. Cook (2008). Successive dimension extraction for estimating the central subspace in multiple-index regression. Journal of Multivariate Analysis, 99, 1733–1757.

Zhao, Z. 2008. Sensitivity of propensity score methods to the specifications. Economics Letters 98, 309 – 319.

Table 1. *Summary variables for analysis of RHC data*

Variable	Description	Average value among those with RHC	Average value among those without RHC	P-value
Age	age at baseline (years)	60.75	61.76	0.03
Gender	0 = male; 1 = female	0.41	0.46	7×10^{-4}
Education	years education	11.85	11.57	8×10^{-4}
Immunehx	0 = no immune disease history 1 = immune disease history	0.71	0.74	3×10^{-3}
Bilirubin	serum bilirubin	0.30	0.09	2×10^{-16}
Resp	0 = no diagnosis of respiratory infection 1 = diagnosis of respiratory infection	0.71	0.58	2×10^{-16}
Card	0 = no diagnosis of respiratory infection 1 = diagnosis of respiratory infection	0.58	0.71	2×10^{-16}

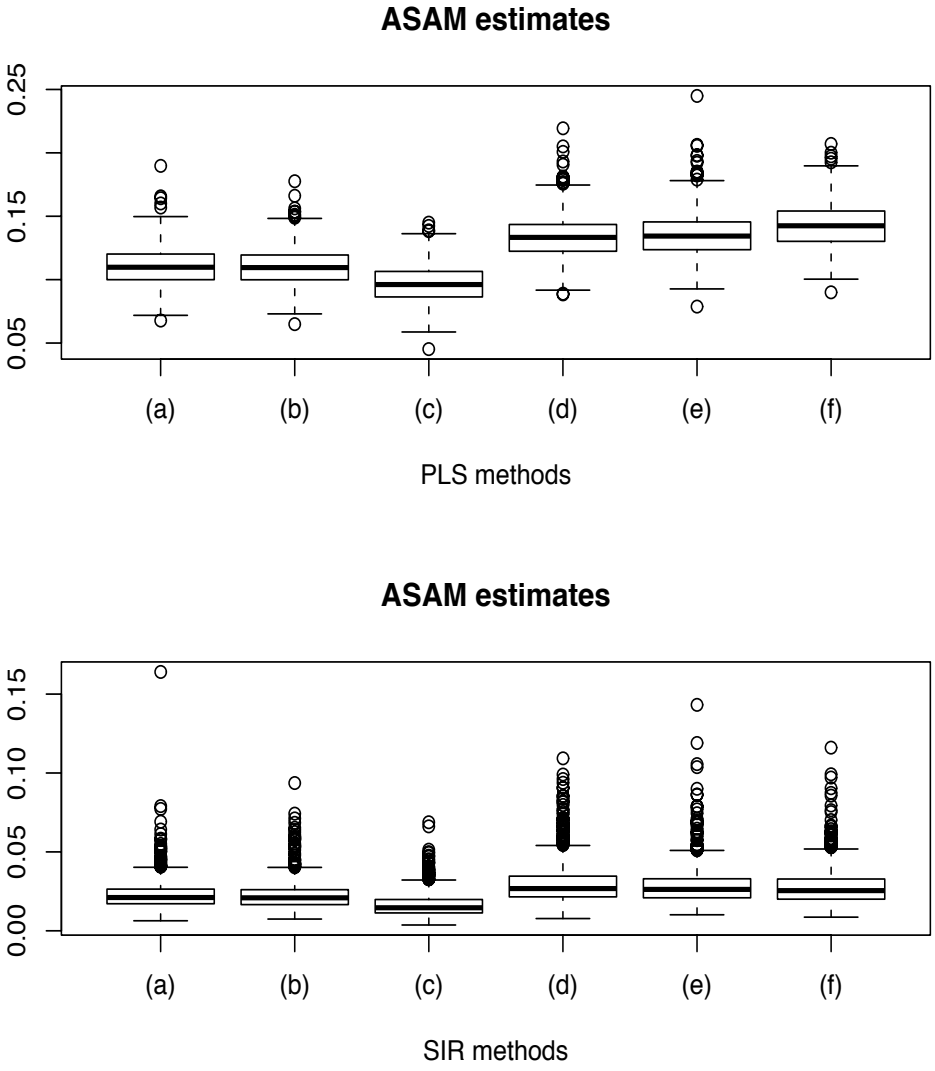

Figure 1: Average standard absolute mean difference in average causal effect estimate across simulation studies. The symbols (a) - (f) correspond to simulation models A - F described in Section 4.2. The upper plot denotes using partial least squares, while the lower plot denotes the output from sliced inverse regression.

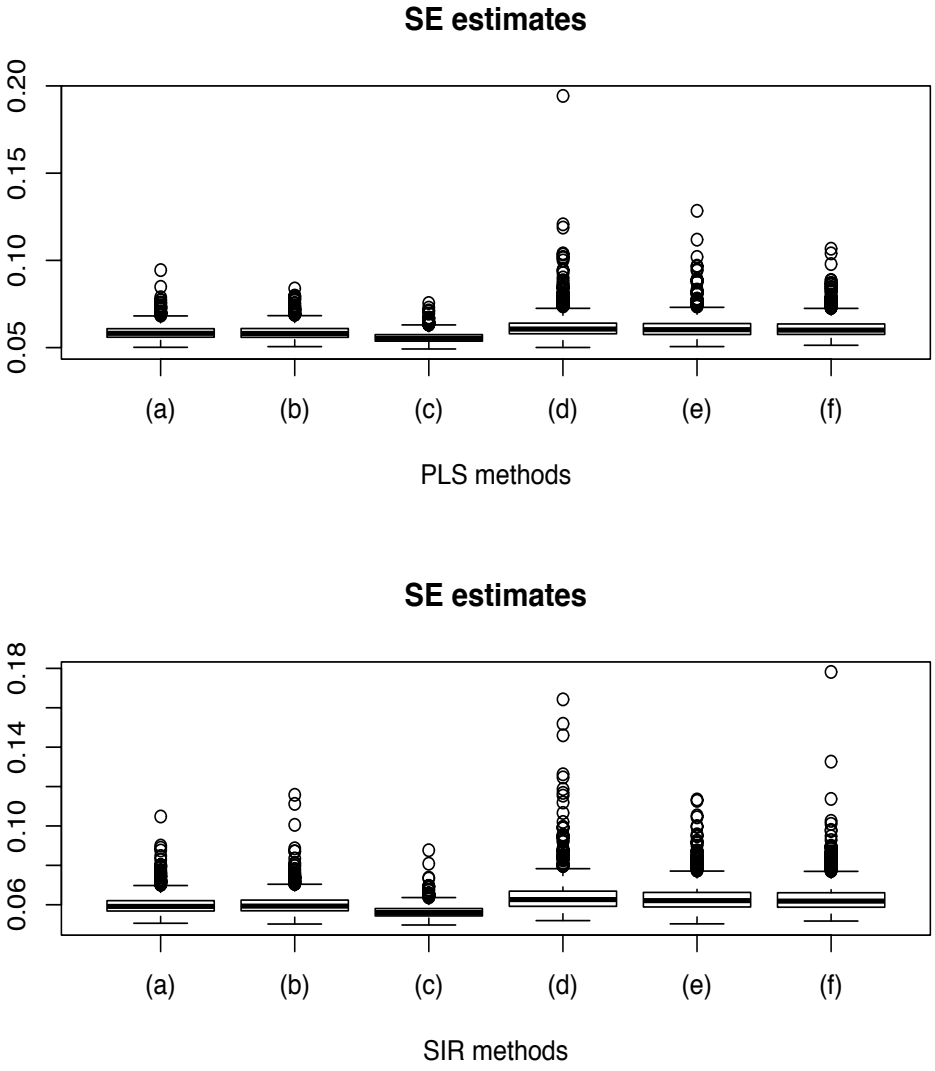

Figure 2: Average standard errors of average causal effect estimate across simulation studies. The symbols (a) - (f) correspond to simulation models A - F described in Section 4.2. The upper plot denotes using partial least squares, while the lower plot denotes the output from sliced inverse regression.