Sufficient covariates and linear propensity analysis

Hui Guo **A. Philip Dawid**
Statistical Laboratory, University of Cambridge

Abstract

Working within the decision-theoretic framework for causal inference, we study the properties of "sufficient covariates", which support causal inference from observational data, and possibilities for their reduction. In particular we illustrate the rôle of a *propensity variable* by means of a simple model, and explain why such a reduction typically does not increase (and may reduce) estimation efficiency.

Keywords: Average causal effect; Propensity variable; Linear discriminant; Quadratic discriminant; Sufficient covariate.

1 Introduction: Decision-theoretic causality

Our concern is to understand the *causal effect* of a binary treatment variable T on a real-valued outcome variable Y, and to consider when and how it might be estimated from observational data. In particular, we shall be concerned with defining, and making appropriate adjustment for, confounding variables.

In contrast to the prevalent "potential outcomes" interpretation of statistical causality (Rubin 1974; Rubin 1978), we shall operate within the decision-theoretic framework for causal inference (Dawid 2002). This aims to identify appropriate assumptions allowing transfer of distributional information between various regimes, comprising an observational regime, whose properties can be identified from data, and interventional regimes, that arise when the treatment is assigned by external manipulation. We introduce a non-stochastic regime indicator variable F_T, with values $\emptyset, 0, 1$: $F_T = \emptyset$ labels the observational regime, while, for $t = 0, 1$, $F_T = t$ labels the interventional regime in which T is set to t. There will be a joint distribution P_f of all relevant variables associated with each regime f. Notations such as $\mathrm{P}_f(A)$ and $\mathrm{P}(A \,|\, F_T = f)$ will be used interchangeably.

Causal assumptions, relating the different regimes, can be conveniently expressed using the notation and calculus of conditional independence, extended to allow some of the variables (here, the regime indicator F_T) to be non-random (Dawid 1979a; Dawid 1980; Dawid 2002). For example, the "ignorable treatment assignment" assumption, which states that the distribution of Y given $T = t$ in the observational regime is the same as in the regime that intervenes to set $T = t$, can be expressed as

$$Y \perp\!\!\!\perp F_T \,|\, T. \tag{1}$$

This is however a strong condition that will rarely be appropriate in the absence of randomisation.

2 Sufficient covariate

For simplicity we confine attention to the *average causal effect* ACE of T on Y, defined by:

$$\mathrm{ACE} := \mathrm{E}_1(Y) - \mathrm{E}_0(Y). \tag{2}$$

Because it is defined in terms of interventional regimes, ACE has a direct causal interpretation. Our prime task is to try and identify ACE from data collected under the observational regime, $F_T = \emptyset$. The natural observational counterpart of ACE is the "face-value average causal effect", FACE $:= \mathrm{E}_\emptyset(Y \,|\, T = 1) - \mathrm{E}_\emptyset(Y \,|\, T = 0)$. We typically will not have FACE = ACE unless we can assume ignorable treatment assignment, which will often be unreasonable. However, external considerations may make it relatively easy to argue that a certain variable X is a *sufficient covariate*, defined as follows.

Definition 1 A (possibly multivariate) variable X is a *covariate (with respect to treatment T)* if:

Property 1: $X \perp\!\!\!\perp F_T$. □

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9.

Property 1 requires that the distribution of X be the same in all regimes, whether observational or interventional. This will typically be appropriate when X is an attribute of the unit to which treatment is applied, or of its environment, determined prior to the treatment decision.

Definition 2 X is a *sufficient covariate (for the effect of treatment T on outcome Y)* if, in addition to Property 1, we have

Property 2: $Y \perp\!\!\!\perp F_T \,|\, (X, T)$. □

Property 2 states, informally, that the conditional distribution of Y, given X and T, is the same in all regimes. Property 2 can also be described as "ignorable treatment assignment, given X" (Rosenbaum and Rubin 1983). In any given problem there may be several distinct sufficient covariates, or none at all. In contrast to the case for statistical (Fisher) sufficiency, there need not exist a minimal sufficient covariate.

A rigorous statement (Dawid 1979a; Dawid 1979b) of Property 2 is as follows. Let Z be a function of Y — which we henceforth notate as $Z \preceq Y$ — whose expectation exists in each regime — which we henceforth denote by "Z is integrable". Then there exists a random variable $W \preceq (X, T)$ such that, for each regime $f = 0, 1, \emptyset$, W serves as a version of the conditional expectation $\mathrm{E}_f(Z \,|\, X, T)$ under the distribution P_f associated with regime f. (Because we focus on ACE, we will only need this property for the case $Z \equiv Y$ — assumed integrable.)

Properties 1 and 2 can be represented graphically by means of the DAG (influence diagram) of Figure 1.[1]

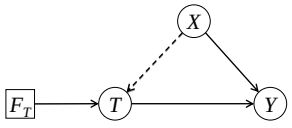

Figure 1: Sufficient covariate

For many purposes we will also require the following reasonable positivity condition, requiring that, for each possible level of X, both treatments are used in the observational regime:

Definition 3 A variable X is a *strongly sufficient covariate* if it is a sufficient covariate and, in addition:

Property 3: *For $t = 0$ or 1, $\mathrm{P}_\emptyset(T = t \,|\, X) > 0$ with probabilility 1.* □

[1]The dotted arrow indicates a link that disappears under an interventional regime: when $F_T = t$, T will have the 1-point distribution at t, independently of X.

Lemma 1 *Suppose X is a strongly sufficient covariate. Then, as distributions for (Y, X, T), $\mathrm{P}_t \ll \mathrm{P}_\emptyset \quad (t = 0, 1)$.*

Proof. Let A be an event for (Y, X, T). Property 2, expressed equivalently as $(Y, X, T) \perp\!\!\!\perp F_T \,|\, (X, T)$, asserts that there exists a function $w(X, T)$ such that $\mathrm{P}_f(A \,|\, X, T) = w(X, T)$, a.s. for $f = 0, 1, \emptyset$. If now $\mathrm{P}_\emptyset(A) = 0$, then, a.s. $[\mathrm{P}_\emptyset]$, $\sum_{t=0}^{1} \mathrm{P}_\emptyset(T = t \,|\, X)\, w(X, t) = \sum_{t=0}^{1} \mathrm{P}_\emptyset(T = t \,|\, X)\, \mathrm{P}_\emptyset(A \,|\, X, T = t) = \mathrm{P}_\emptyset(A \,|\, X) = 0$. By Property 3, for $t = 0, 1$, $w(X, t) = 0$ a.s. $[\mathrm{P}_\emptyset]$ and hence, by Property 1, a.s. $[\mathrm{P}_t]$. So $\mathrm{P}_t(A) = \mathrm{E}_t\{w(X, t)\} = 0$. □

Theorem 1 *Let X be a strongly sufficient covariate. Then for any integrable $Z \preceq (Y, X, T)$, and any versions of the conditional expectations,*

$$\mathrm{E}_t(Z \,|\, X) = \mathrm{E}_\emptyset(Z \,|\, X, T = t) \quad (t = 0, 1) \qquad (3)$$

almost surely in any regime. We can take $\mathrm{E}_\emptyset(Z \,|\, X, T)$ or $\mathrm{E}_T(Z \,|\, X)$ as a version of $\mathrm{E}(Z \,|\, X, T)$ in all regimes.

Proof. By Property 2 there exists a function $w(X, T)$ which is a version of $\mathrm{E}_f(Z \,|\, X, T)$ for $f = 0, 1, \emptyset$. In particular, $\mathrm{E}_\emptyset(Z \,|\, X, T) = w(X, T)$ a.s. $[\mathrm{P}_\emptyset]$ and thus, by Lemma 1, a.s. $[\mathrm{P}_t]$ $(t = 0, 1)$ — thus we can take $\mathrm{E}_\emptyset(Z \,|\, X, T)$ as the common version of $\mathrm{E}_f(Z \,|\, X, T)$ for $f = 0, 1, \emptyset$. In particular, a.s. $[\mathrm{P}_t]$, $\mathrm{E}_t(Z \,|\, X) = \mathrm{E}_t(Z \,|\, X, T = t) = \mathrm{E}_\emptyset(Z \,|\, X, T = t)$. So (3) holds a.s. $[\mathrm{P}_t]$ and thus, by Property 1, a.s. in each regime. □

Theorem 1 expresses rigorously what we mean by saying that the observational conditional distribution of Y, given a strongly sufficient covariate X, for those happening to receive treatment t, is the same as the interventional conditional distribution of Y given X, for those given treatment t.

2.1 Back-door formula

Let X be a covariate.

Definition 4 The *specific causal effect (of T on Y, relative to X)* is the random variable

$$\mathrm{SCE} := \mathrm{E}_1(Y \,|\, X) - \mathrm{E}_0(Y \,|\, X). \qquad \square$$

Then SCE is a function of X, defined almost surely (under any regime). Where we need to indicate its construction from the specific covariate X, we annotate SCE as SCE_X; we also write $\mathrm{SCE}(X)$ or $\mathrm{SCE}_X(X)$ to express SCE as a function of X. Because it is defined in terms of interventional regimes, SCE has a direct causal interpretation: $\mathrm{SCE}(x)$ is the average causal effect in the subpopulation having $X = x$.

Theorem 2 *For any covariate X, $\mathrm{ACE} = \mathrm{E}(\mathrm{SCE}_X)$ (where the expectation may be taken under any regime).*

Proof. By Property 1, $\mathrm{E}_\emptyset\{\mathrm{E}_t(Y \,|\, X)\} = \mathrm{E}_t\{\mathrm{E}_t(Y \,|\, X)\} = \mathrm{E}_t(Y)$. By subtraction, $\mathrm{ACE} = \mathrm{E}_f(\mathrm{SCE}_X)$ for $f = \emptyset$ and thus, again by Property 1, also for $f = 0, 1$. □

SCE_X is typically not identifiable from purely observational data. However, if X is strongly sufficient, by (3) we can also express SCE_X as $\mathrm{E}_\emptyset(Y \,|\, X, T = 1) - \mathrm{E}_\emptyset(Y \,|\, X, T = 0)$, which means that it is then estimable from data collected in the observational regime. Since we have

$$\mathrm{ACE} = \mathrm{E}_\emptyset\left(\mathrm{SCE}_X\right), \tag{4}$$

it follows that ACE is then expressible purely in terms of properties of the observational joint distribution of (T, X, Y), where X is any strongly sufficient covariate. Formula (4) is Pearl's "back-door formula" (Pearl 1993).

We can similarly define "the effect of treatment on the treated", as $\mathrm{ETT} = \mathrm{E}_\emptyset\left(\mathrm{SCE}_X \,|\, T = 1\right)$ — again the same for all choices of X (Geneletti and Dawid 2010).

3 Reduction of strongly sufficient covariate

Suppose X is a strongly sufficient covariate. It might simplify the application of formula (4) if we could replace X by a variable $V \preceq X$ which is itself a strongly sufficient covariate. When can we do this? Since, when $V \preceq X$, Properties 1 and 3 for V are automatically inherited from the same properties for X, we need only establish Property 2 for V, *viz.*:

$$Y \perp\!\!\!\perp F_T \,|\, (V, T). \tag{5}$$

The following theorem gives two alternative sufficient conditions for this to hold.[2]

Theorem 3 *Suppose X is a strongly sufficient covariate and $V \preceq X$. Then V will be a strongly sufficient covariate if either of the following conditions is satisfied:*

(a). **Response-sufficient reduction:**[3] *For $t = 0, 1$,*

$$Y \perp\!\!\!\perp X \,|\, (V, F_T = t). \tag{6}$$

That is, for each applied treatment, once V is known, any further information about X is of no value for predicting Y.

(b). **Treatment-sufficient reduction:**

$$T \perp\!\!\!\perp X \,|\, (V, F_T = \emptyset). \tag{7}$$

That is, in the observational regime, the choice of treatment depends on X only through V.

Proof. Assume first Condition (a). By (6), for any integrable $Z \preceq Y$ there exists a version, $w(X, t)$ say, of $\mathrm{E}_t(Z \,|\, X)$ that is a function of V $(t = 0, 1)$. By Theorem 1, $W := w(X, T)$ is a common version of $\mathrm{E}_f(Z \,|\, X, T)$ in every regime f; moreover $W \preceq (V, T)$. So the common conditional distribution for Y given (X, T) depends on X through V alone:

$$Y \perp\!\!\!\perp (X, F_T) \,|\, (V, T) \tag{8}$$

and (5) follows.

As for Condition (b), since we trivially have $X \perp\!\!\!\perp T \,|\, (V, F_T = t)$, (7) is equivalent to

$$X \perp\!\!\!\perp T \,|\, (V, F_T). \tag{9}$$

Now Property 1, with $V \preceq X$, implies $X \perp\!\!\!\perp F_T \,|\, V$, which together with (9) gives $X \perp\!\!\!\perp (T, F_T) \,|\, V$, whence $X \perp\!\!\!\perp F_T \,|\, (V, T)$. This now combines with Property 2 (equivalently expressed, since $V \preceq X$, as $Y \perp\!\!\!\perp F_T \,|\, (X, V, T)$), to yield $(Y, X) \perp\!\!\!\perp F_T \,|\, (V, T)$, and we once again deduce (5). □

We note that the proof of (b) does not require Property 3 and, once we have established (9), uses only the basic calculus of conditional independence (Dawid 1979a; Dawid 1980). In particular, it continues to hold when the qualifier "strongly" in the statement of Theorem 3 is omitted.

As an alternative proof of (b), we can first represent the recursive collection of conditional independence properties (i) Property 1: $X \perp\!\!\!\perp F_T$ (ii) $V \perp\!\!\!\perp F_T \,|\, X$ (a trivial consequence of (i) and $V \preceq X$), (iii) eqn. (9): $T \perp\!\!\!\perp X \,|\, (V, F_T)$[4] and (iv) $Y \perp\!\!\!\perp (V, F_T) \,|\, (X, T)$ (which follows from Property 2 and $V \preceq X$), by means of the DAG of Figure 2. We can then use the semantics of d-separation (Pearl 1986; Verma and Pearl 1990) or its moralisation equivalent (Lauritzen *et al.* 1990) to read the desired conclusion (5) directly off the graph.

A parallel graphical approach to (a) does not work since, without Property 3, (8) need not follow. However, while not supplying a proof, Figure 3 does conveniently embody the collection of conditional independencies Property 1, Property 2 and (8), as well as the

[2] However, (5) can hold even when neither of these conditions does.

[3] The *prognostic score* (Hansen 2008) corresponds to (6) confined to $t = 0$.

[4] Note that the dotted arrow in Figure 2 implies the equivalence of (7) and (9).

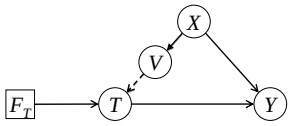

Figure 2: Treatment-sufficient reduction

trivial property $V \perp\!\!\!\perp T \,|\, (X, F_T)$; and also the conclusion (5).

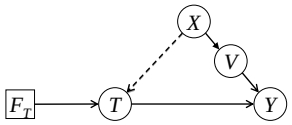

Figure 3: Response-sufficient reduction

3.1 Propensity variable

We observe that the defining property (7) for a treatment-sufficient reduction, expressed as

$$X \perp\!\!\!\perp T \,|\, (V, F_T = \emptyset), \tag{10}$$

states that the observational conditional distribution of X given V is the same for both treatments (*i.e.*, further conditioned on either $T = 0$ or $T = 1$): that is to say, V is a *balancing score* for X (Rosenbaum and Rubin 1983).[5]

The property (10) can also be fruitfully interpreted as follows. Consider the family $\mathcal{Q} = \{Q_0, Q_1\}$ consisting of the pair of observational conditional distibutions for X, given respectively $T = 0$ and $T = 1$: these are well-defined since, by Property 3, each of these conditioning events has positive probability. Then (10) asserts that V is a *sufficient statistic* (in the usual Fisherian sense) for this family. In particular, a *minimal* treatment-sufficient reduction[6] is obtained as a minimal sufficient statistic for $\mathcal{Q}$: *viz.*, any variable almost surely equal to a $(1, 1)$-function of the likelihood ratio statistic $\Lambda := q_1(X)/q_0(X)$ (where $q_i(\cdot)$ is a version of the density of Q_i with respect to some dominating measure). We term such a minimal treatment-sufficient covariate a *propensity variable*, since one form for it is

$$\Pi := \mathrm{P}_\emptyset(T = 1 \,|\, X) = \pi\,\Lambda/(1 - \pi + \pi\Lambda) \tag{11}$$

(where $\pi := \mathrm{P}_\emptyset(T = 1) \in (0, 1)$, by Property 3), which is known as the *propensity score* (Rosenbaum and Rubin 1983).

Note that it is entirely possible that we will obtain distinct propensity variables if we start from distinct strongly sufficient covariates.

4 Normal linear model

We illustrate and develop the above theory in the context of a simple but instructive example.

We have a univariate response Y, and a $(p \times 1)$ vector sufficient covariate X. The common conditional distribution for Y given (X, T) is specified as:

$$Y \,|\, (X, T, F_T) \sim \mathcal{N}(d + \delta T + b'X, \phi), \tag{12}$$

with parameters d and δ (scalar), b $(p \times 1)$, and ϕ (scalar). It is readily seen that the specific causal effect is constant: $\mathrm{SCE}_X \equiv \delta$. In particular, from (4), our main quantity of interest, the average causal effect ACE, is just δ.

From (12) we see that the *linear predictor* $\mathrm{LP} := b'X$ satisfies Condition (a) of Theorem 3, and thus LP is a response-sufficient reduction of X. Trivially $\mathrm{E}(Y \,|\, \mathrm{LP}, T) = d + \delta T + \mathrm{LP}$, with the desired coefficient δ of T.

Our model for the observational joint distribution of (T, X) is most easily described in terms of its marginal for T, and conditional for X given T, as follows:

$$\mathrm{P}_\emptyset(T = 1) = \pi \tag{13}$$

$$X \,|\, (T, F_T = \emptyset) \sim \mathcal{N}(\mu_T, \Sigma) \tag{14}$$

with parameters $\pi \in (0, 1)$, μ_0, μ_1 $(p \times 1)$, and Σ $(p \times p$ positive definite).

The observational marginal distribution of X will thus be a multivariate normal mixture:

$$X \,|\, F_T = \emptyset \sim (1 - \pi)\,\mathcal{N}(\mu_0, \Sigma) + \pi\,\mathcal{N}(\mu_1, \Sigma). \tag{15}$$

Because we are requiring Property 1 to hold, the marginal distribution for X in each interventional regime, $F_T = 0$ or 1, is also given by (15).

The observational conditional distribution of T given X is given by (11), with

$$\log \Lambda = c + \mathrm{LD} \tag{16}$$

where $-2c = \mu_1'\Sigma^{-1}\mu_1 - \mu_0'\Sigma^{-1}\mu_0$, and

$$\mathrm{LD} := \gamma' X, \tag{17}$$

with

$$\gamma := \Sigma^{-1}(\mu_1 - \mu_0), \tag{18}$$

[5] *Caution*: This balancing property need not imply that V is a sufficient covariate if the variable X we start from is not itself sufficient.

[6] *Caution*: Although any non-trivial reduction of such a variable can not be treatment-sufficient, that does not imply that such a reduction can not be a sufficient covariate. For example, if $Y \perp\!\!\!\perp X \,|\, (T, F_T)$ then the trivial reduction of X to a constant will be a sufficient covariate.

is Fisher's *linear discriminant*, best separating the pair of multivariate normal observational distributions for X, given $T = 0$ and $T = 1$. It is clear that Property 3 holds, so that X is strongly sufficient.

Lemma 2 *Suppose V is a linear sufficient covariate (*i.e.*, a linear function of X that is itself a sufficient covariate). Then the coefficient of T in the observational linear regression of Y on T and V is δ.*

Proof. Under the distributional assumptions made, linearity of V implies

$$\mathrm{E}_t(Y \mid V) = \tilde{d} + \tilde{\delta} t + \tilde{b}' V \tag{19}$$

for suitable parameter values. Thus $\mathrm{SCE}_V \equiv \tilde{\delta}$, so that $\tilde{\delta} = \mathrm{E}(\mathrm{SCE}_V) = \mathrm{ACE} = \delta$. If now V is sufficient (hence strongly sufficient, since X is), then (19) is also the observational conditional expectation $\mathrm{E}_\emptyset(Y \mid V, T = t)$. □

From (16) we see that LD is a propensity variable. As it is thus a linear sufficient covariate, from Lemma 2 we deduce:

Lemma 3 *Under the given distributional assumptions, the coefficient of T in the observational regression of Y on T and* LD *is δ.*

4.1 Further implications

A significant observation is the following. Since Lemma 3 refers solely to properties of the observational regime, it must still hold when the causal assumptions, relating that regime to interventional regimes, are dropped. Furthermore, because this result involves only first and second order moments, it must apply in still greater generality. Thus suppose (Y, T, X), with T binary, have an arbitrary joint distribution with finite second moments and T non-degenerate. Let $\mu_t := \mathrm{E}(X \mid T = t)$, let Σ be the dispersion matrix of $X - \mu_T$, and let $\mathrm{LD} := \gamma' X$, with $\gamma := \Sigma^{-1}(\mu_1 - \mu_0)$, be the linear discriminant function, based on X, for distinguishing between $T = 0$ and $T = 1$. Then we have

Theorem 4 *The coefficient of T in the linear regression of Y on (T, LD) is the same as that in the linear regression of Y on (T, X).*

Corollary 1 *Suppose we have data on (Y, T, X) for a sample of individuals. Let* LD^* *be the sample linear discriminant for T based on X. Then the coefficient of T in the sample linear regression of Y on T and* LD^* *is the same as that in the sample linear regression of Y on T and X.*

Proof. Apply Theorem 4 to the empirical distribution of (Y, T, X) formed from the sample. □

Rosenbaum and Rubin (1983) (§ 3.4) give this result with a brief non-causal argument.

The result of Corollary 1 is paradoxical: when we think our estimation procedure is adjusting for the treatment assignment process, by regressing on the estimated propensity variable LD^*, what we actually end up with is an adjustment for the full sufficient covariate X—which renders the treatment assignment process entirely irrelevant.

4.2 Propensity analysis does not increase precision

It might naïvely be thought that it would increase the precision of our estimator of ACE if we were to adjust for just one covariate, the sample-based propensity variable LD^*, rather than for all p variables X. This could be regarded as supported by the fact that, if we adjust X for some univariate linear function Z of X, the (sample-based) choice $Z = \mathrm{LD}^*$ will tend to maximise sample balance across the two treatment groups. However, Corollary 1 shows that there can be no increase in efficiency: indeed, adjusting for LD^* rather than for all p predictors makes *absolutely no difference* to our estimate — and, consequently, to its precision. Similar theoretical conclusions hold in other scenarios (Hahn 1998; Senn *et al.* 2007), and there is substantial empirical evidence that this is a general effect (Winkelmayer and Kurth 2004). Intuitively, the overfitting error introduced by selecting that variable best separating the two treatment groups in the data compensates for any increased accuracy that could be obtained by regressing on just one variable, rather than on p. The excellent sample balance achieved is likewise a feature of overfitting, and merely reflects the equivalence of (7) and (10), applied, with $V = \mathrm{LD}^*$, to the empirical distribution of the data.

There is a close analogy with the case of response-sufficiency: the sample analogue of the response-sufficient covariate LP is the estimated best linear predictor for the observed data, LP^*. But it would clearly not change the estimated coefficient of T if we were first to compute LP^* by regressing on all p variables, and finally regress Y on T and the single variable LP^*.

There is much theoretical and empirical evidence (Robins *et al.* 1992; Hirano *et al.* 2003) to support the claim that estimating ACE by adjusting for the *true* propensity variable (when this is known) yields *worse* precision than adjusting for an *estimated* propensity variable. In our setting the two approaches correspond to running regressions of Y on (T, LD) and on

(T, LD^*), respectively — each of which will yield an unbiased estimator of δ, the ACE. The above claim clearly can not be universally valid, since in the special case that LD happens to be the same as LP, by Corollary 1 the regression on LD^* is equivalent to adjusting for the estimated linear predictor LP^*, which by the Gauss-Markov theorem will be less accurate than adjusting for the true $\mathrm{LP} = \mathrm{LD}$ (though the discrepancy will be small in large samples). The claim is however likely to apply in the typical case that LD is not highly correlated with LP, for then regressing on LD is adjusting for a variable which is a less precise predictor of outcome than the true linear predictor LP; whereas, by Corollary 1, adjusting for LD^* will yield the same estimate as adjusting for LP^* which (in large enough samples) will approximate adjustment for LP, which is optimal.

For illustration, suppose we have the following true values for the parameters in (12), (13) and (14): $p = 2$, $d = 0$, $\delta = 0.5$, $b = (1, 0)'$, $\phi = 1$, $\pi = 0.5$, $\mu_1 = (0, 1)'$, $\mu_0 = (0, 0)'$, $\Sigma = I_2$. Then the population linear predictor is $\mathrm{LP} = X_1$, with $Y \mid X, T \sim \mathcal{N}(\frac{1}{2}T + X_1, 1)$, while the population linear discriminant is X_2 — which is however not even weakly predictive of Y, since $Y \mid X_2, T \sim \mathcal{N}(\frac{1}{2}T, 2)$.

The full regression model (model M_0 say) is as in (12), with all parameters taken as unknown. Fitting the true linear predictor $\mathrm{LP} = X_1$ is equivalent to setting $b_2 = 0$ (yielding model M_1), while fitting the true linear discriminant $\mathrm{LD} = X_2$ is equivalent to setting $b_1 = 0$ (model M_2). Note that all these models are "true"; for M_2 the true value of b_2 is 0, and the true residual variance ϕ is 2, as against $\phi = 1$ for M_0 and M_1. Finally, for any data-set we can construct the estimated propensity variable LD^*, and then fit the model $Y \sim \mathcal{N}(d + \delta T + b^* \mathrm{LD}^*, \phi)$ (model M_3).

For each model M_k the associated least-squares estimator $\widehat{\delta}_k$ of δ is unbiased, having expectation $\delta = 0.5$. By the Gauss-Markov theorem, conditionally on all regressors and hence also unconditionally, $\mathrm{var}(\widehat{\delta}_0)$ $(= \mathrm{var}(\widehat{\delta}_3)$ by Corollary 1) $\geq \mathrm{var}(\widehat{\delta}_1)$. In fact $\mathrm{var}(\widehat{\delta}_0) \sim 5/n$, $\mathrm{var}(\widehat{\delta}_1) \sim 4/n$. However, $\mathrm{var}(\widehat{\delta}_2) \sim 10/n$, so in this case it is indeed asymptotically less precise to adjust for the true propensity variable than for its estimate.

To investigate finite-sample behaviour, we generated 25 simulated datasets, each of size $n = 20$, from the above model. Figure 4 shows the empirical distribution of $\widehat{\delta}_k$ for each of the models. As expected, we see that adjusting for the true linear predictor LP is most precise; next is adjusting for the estimated propensity variable LD^* or — entirely equivalently — for all covariates (X_1, X_2); and last comes adjustment for the true propensity variable $\mathrm{LD} = X_2$.

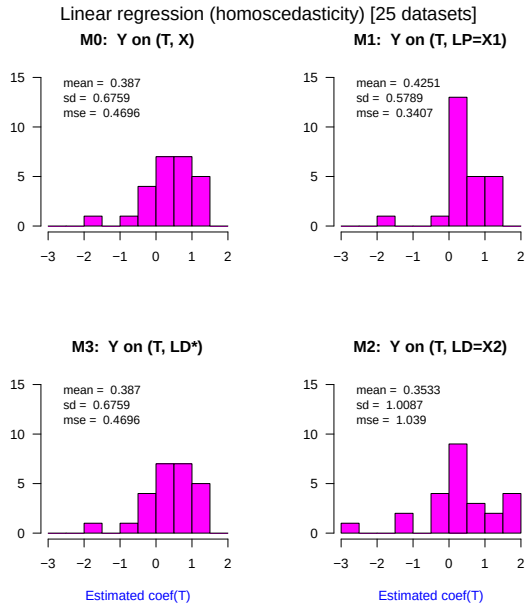

Figure 4: (Clockwise from top left.) Estimates of ACE = 0.5 from regressing Y on T and: (X_1, X_2); population linear predictor $\mathrm{LP} = X_1$; population linear discriminant (propensity variable) $\mathrm{LD} = X_2$; estimated linear discriminant (propensity variable) LD^*.

5 Heteroscedasticity

Suppose now that, keeping all other distributional assumptions of § 4 unchanged, we change (14) to:

$$X \mid (T, F_T = \emptyset) \sim \mathcal{N}(\mu_T, \Sigma_T) \tag{20}$$

with different within-group dispersion matrices Σ_0, Σ_1. The marginal distribution of X (in all regimes) is thus

$$X \sim (1 - \pi)\,\mathcal{N}(\mu_0, \Sigma_0) + \pi\,\mathcal{N}(\mu_1, \Sigma_1), \tag{21}$$

while expression (16) is replaced by:

$$\log \Lambda = c + \mathrm{QD} \tag{22}$$

where $-2c = \log\det\Sigma_1 - \log\det\Sigma_0 + \mu_1'\Sigma_1^{-1}\mu_1 - \mu_o'\Sigma_0^{-1}\mu_0$, and

$$\mathrm{QD} := \left(\Sigma_1^{-1}\mu_1 - \Sigma_0^{-1}\mu_0\right)' X - \frac{1}{2} X' \left(\Sigma_1^{-1} - \Sigma_0^{-1}\right) X \tag{23}$$

is the *quadratic discriminant* for distinguishing the observational distributions of X given $T = 0$ and $T = 1$. In particular, QD is a propensity variable — though not, now, a linear covariate.

It now follows that $\mathrm{ACE} = \mathrm{E}(\mathrm{SCE}_{\mathrm{QD}})$, with $\mathrm{SCE}_{\mathrm{QD}} = \mathrm{E}_\emptyset(Y \mid T = 1, \mathrm{QD}) - \mathrm{E}_\emptyset(Y \mid T = 0, \mathrm{QD})$. However,

since QD is quadratic in X, computation of the conditional and the unconditional expectations in this formula is difficult. As an approximation, we might replace the exact non-linear form of $\mathrm{E}_\emptyset(Y \mid T, \mathrm{QD})$ by the observational *linear* regression of Y on (T, QD), and extract the coefficient of T — although this will not be exactly ACE. A version of this using sample-based estimates of the required population parameters would then (it might be hoped) provide a reasonable estimate of ACE. Alternatively, $\mathrm{SCE}_{\mathrm{QD}}$ might be estimated nonparametrically, *e.g.* by subclassification (Rosenbaum and Rubin 1984) on QD, and averaged to estimate ACE.

A different approach can be based on the theory of § 4.1. The linear discriminant is

$$\mathrm{LD} = (\mu_1 - \mu_0)'\Sigma^{-1}X, \tag{24}$$

with $\Sigma = \pi_0\Sigma_0 + \pi_1\Sigma_1$. Even though LD is *not* now a sufficient covariate, Theorem 4 applies, so allowing us to identify ACE as the coefficient δ of T in the population linear regression of Y on (T, LD) in the observational regime. Its sample analogue δ^*, obtained by running a linear regression of Y on T and the sample linear discriminant LD^*, will then be an unbiased estimator of ACE — indeed, will be identical with the straightforward estimator obtained by regressing Y on (T, X).

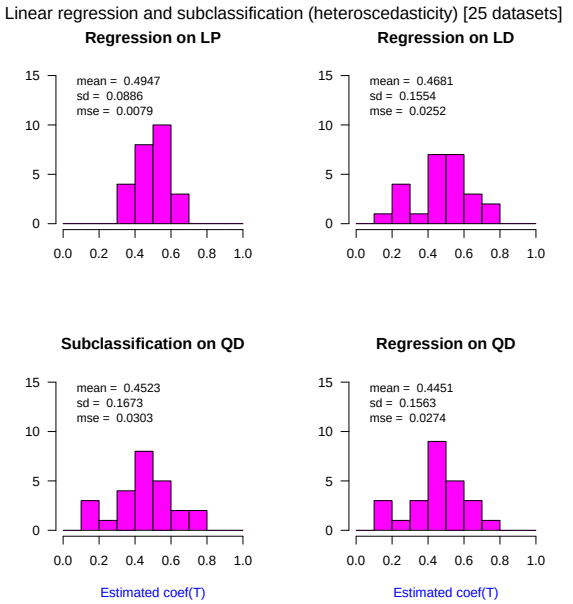

Figure 5: (Clockwise from top left.) Estimates of ACE = 0.5 from regressing Y on T and: linear predictor $\mathrm{LP} = X_1$; population linear discriminant $\mathrm{LD} \propto X_2$; population quadratic discriminant (propensity variable) QD. Fourth plot: by subclassification on QD.

5.1 Simulations

We simulated 25 datasets of size 400 from the above model with the following parameter values: $p = 20$; $d = 0$, $\delta = 0.5$, $b = (1, 0, \ldots, 0)'$; $\mu_0 = (0, \ldots, 0)'$, $\mu_1 = (0, 0.5, 0, \ldots, 0)'$, Σ_0 diagonal with 0.8 for the first ten entries and 1.3 for the rest, Σ_1 the identity matrix; $\pi = 0.5$.

Figure 5 gives estimates of ACE obtained by adjusting for known population quantities. The first three plots arise from linear regression of Y on, respectively (T, LP), (T, LD), and (T, QD), where $\mathrm{LP} = X_1$, and (from (24) and (23)) $\mathrm{LD} = (5/9)X_2$, $\mathrm{QD} = (1/2)X_2 + (1/8)\sum_{i=1}^{10} X_i^2 - (3/26)\sum_{j=11}^{20} X_j^2$. The last plot is computed by subclassification on the population propensity variable QD (with classes formed by dividing its empirical distribution into 5 equal parts).

Figure 6 gives the corresponding results when these population variables are replaced by their sample estimates. (The first plot is from the regression on all the predictors).

Again regression adjustment for the true linear predictor is, unsurprisingly, the best procedure. Next comes the sample regression on all predictors or, equivalently, on the sample linear discriminant. Then, roughly

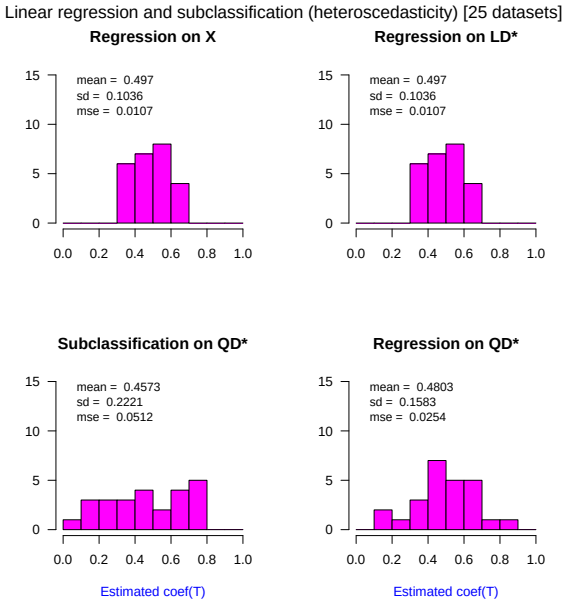

Figure 6: (Clockwise from top left.) Estimates of ACE = 0.5 from regressing Y on T and: all covariates X; sample linear discriminant LD^*; sample quadratic discriminant (propensity variable) QD^*. Fourth plot: by subclassification on QD^*.

equal, are regression on the population linear discriminant, on the sample quadratic discriminant, and on the population quadratic discriminant. Last comes subclassification on the quadratic propensity variable, the performance being particularly dismal when this is estimated.

6 Discussion

We have identified a propensity variable as a minimal treatment-sufficient reduction of a sufficient covariate X. For a simple normal linear model, this can be taken as the linear discriminant LD between the two observational distributions of X given treatment. We have shown that linear adjustment for the sample estimate LD^* of LD yields exactly the same estimate of ACE as adjustment for all of X, and so can neither increase nor decrease precision. Typically, though not universally, adjustment for the true LD will deliver worse accuracy. Even when, in the heteroscedastic case, LD is not a genuine propensity variable, adjusting for the estimated LD^* will still be equivalent to running a full regression on all predictors, and hence still yield an unbiased (but not more—nor less—accurate) estimator of ACE. In contrast, adjustment (be it parametric or nonparametric, population-based or sample-based) for a genuine propensity variable (the quadratic discriminant) performs poorly.

Our investigations, limited though they are to a very simple model, add weight to the accruing evidence (Hahn 1998; Robins *et al.* 1992; Hirano *et al.* 2003; Winkelmayer and Kurth 2004; Senn *et al.* 2007) that propensity analysis has little to contribute to improving the estimation of causal effects.

Acknowledgements

Hui Guo was supported through an EPSRC Science and Innovation Award to the University of Cambridge.

References

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

Dawid, A. P. (1979b). Some misleading arguments involving conditional independence. *Journal of the Royal Statistical Society, Series B*, **41**, 249–52.

Dawid, A. P. (1980). Conditional independence for statistical operations. *Annals of Statistics*, **8**, 598–617.

Dawid, A. P. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review*, **70**, 161–89. Corrigenda, *ibid.*, 437.

Geneletti, S. and Dawid, A. P. (2010). Defining and identifying the effect of treatment on the treated. In *Causality in the Sciences*, (ed. P. M. Illari, F. Russo, and J. Williamson). Oxford University Press. To appear.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, **66**, 315–31.

Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika*, **95**, 481–8.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, **71**, 1161–89.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N., and Leimer, H.-G. (1990). Independence properties of directed Markov fields. *Networks*, **20**, 491–505.

Pearl, J. (1986). A constraint–propagation approach to probabilistic reasoning. In *Uncertainty in Artificial Intelligence*, (ed. L. N. Kanal and J. F. Lemmer), pp. 357–70. North-Holland, Amsterdam.

Pearl, J. (1993). Comment: Graphical models, causality and intervention. *Statistical Science*, **8**, 266–9.

Robins, J. M., Mark, S. D., and Newey, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, **48**, 479–95.

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

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, **79**, 516–24.

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

Rubin, D. B. (1978). Bayesian inference for causal effects: the role of randomization. *Annals of Statistics*, **6**, 34–68.

Senn, S., Graf, E., and Caputo, A. (2007). Stratification for the propensity score compared with linear regression techniques to assess the effect of treatment or exposure. *Statistics in Medicine*, **26**, 5529–44. Erratum: *Stat. Med.* **27** (2008), 4615.

Verma, T. and Pearl, J. (1990). Causal networks: Semantics and expressiveness. In *Uncertainty in Artificial Intelligence 4*, (ed. R. D. Shachter, T. S. Levitt, L. N. Kanal, and J. F. Lemmer), pp. 69–76. North-Holland, Amsterdam.

Winkelmayer, W. C. and Kurth, T. (2004). Propensity scores: Help or hype? *Nephrology Dialysis Transplantation*, **19**, 1671–3.