Balancing and Elimination of Nuisance Variables

Siamak Noorbaloochi, Minneapolis VA Medical Center and University of Minnesota
David Nelson, Minneapolis VA Medical Center and University of Minnesota
Masoud Asgharian, McGill University

Recommended Citation:
DOI: 10.2202/1557-4679.1209
Balancing and Elimination of Nuisance Variables

Siamak Noorbaloochi, David Nelson, and Masoud Asgharian

Abstract

Addressing covariate imbalance in causal analysis will be reformulated as an elimination of the nuisance variables problem. We show, within a counterfactual balanced setting, how averaging, conditioning, and marginalization techniques can be used to reduce bias due to a possibly large number of imbalanced baseline confounders. The notions of X-sufficient and X-ancillary quantities are discussed and, as an example, we show how sliced inverse regression and related methods from regression theory that estimate a basis for a central sufficient subspace provide alternative summaries to propensity based analysis. Examples for exponential families and elliptically symmetric families of distributions are provided.

KEYWORDS: confounding, dimension reduction, sufficient summary, ancillarity

Author Notes: Research supported in part by VAHSR&D Grant IIR 07-229.
1 Introduction

A common problem of causal inference in observational studies is covariate imbalance or confounding. Covariate imbalance occurs when we are interested in estimating the effect of a variable of interest, typically an intervention or treatment, on a response but the distribution of the intervention and the outcome both vary with the same set of covariates. In this paper, we study this problem by treating intervention assignment as a nuisance variable in the distribution of the covariates and examine how to eliminate this nuisance variable. The elimination of nuisance parameters in statistical inference is a classic, well studied area. We find that, by treating the intervention variable as a nuisance variable, we are able to use the ideas underlying classical methods for eliminating nuisance parameters to readily address covariate imbalance.

Two approaches we consider are marginalization and conditioning. With marginalization we try to study the effect of the intervention on the outcome by using a distribution for a function of the covariates that does not depend upon. With conditioning we find a covariate-based statistic, called a sufficient summary, for which the conditional distribution of the covariates given the statistic is independent of the intervention and then study the effect of the intervention on the responses conditional on this statistic. Both approaches are based on a compatible pair of conditional models: one for the covariates given the intervention and the other for the intervention given the covariates. The counterparts of these two methods in the statistical theory for the elimination of nuisance parameters are Fisher’s classical ideas of ancillarity and sufficiency. However, in this traditional statistical inference, sufficiency is used to marginalize and ancillarity is used to perform a conditional analysis.

A critical issue in a causal analysis is to actually define the causal effect. Indeed, for the conditional elimination of the covariate imbalance the issue of how to aggregate conditional inferences is central to the definition and estimation of causal effects. Rosenbaum and Rubin (1983) provide a Neyman-type counterfactual model to give meaning to the average of the conditional effects, here averaged with respect to the marginal distribution of the covariates. Note, this specific averaging is but one potential averaging proposed by Cochran (1968) for defining unbiased group differences in outcomes. In the following section we outline an alternative decision theoretic justification, using squared error loss, for using this specific form of averaging of the conditional distributions. The third section introduces the conditional elimination method wherein we define the notion of sufficient summaries which are analogous to sufficient statistics. We show that propensity scores and related balancing scores are all sufficient summaries. In addition, we show how the effective dimension reduction directions discussed by Li (1994) and others in the theory
of dimension reduction for regression can be used as linear sufficient summaries. These summaries were devised chiefly with the aim of reducing the dimension of the predictor space. We discuss how they may be used to reduce bias due to covariate imbalance. In the last section, we introduce ancillary quantities and provide conditions under which these quantities can be used for the elimination of covariate imbalance. Finally, we demonstrate how these ideas can be applied to real data sets. First, however, we need to define the notation and additional concepts we will use in the subsequent discussion.

Let \((Y, T, X)\) denote a vector of observable variables where \(Y\) is a response variable, \(T\) is a random variable whose effect on the distribution of \(Y\) is of interest, and \(X\) is a set of \(p\) confounding covariates. The sets of possible values of \(Y\), \(T\), and \(X\), respectively, will be denoted by \(\mathcal{Y}\), \(\mathcal{T}\), and \(\mathcal{X}\). For simplicity, we assume the families of distributions for these measures are dominated with respect to Lebesgue measure or a counting measure, and hence, densities and probability mass functions will be used to describe the models.

Consider a family of conditional covariate models \(\mathcal{F} = \{f_\theta(x|t) : \theta \in \Theta\}\) and a family of propensity models \(\mathcal{G} = \{f_\beta(t|x) : \beta \in B, t \in \mathbb{R}, x \in \mathbb{R}^p\}\). We assume the propensity models and the conditional covariate models are compatible (Arnold and Press, 1989). A pair of conditional distributions, \(f_\theta(t|x)\) and \(f_\beta(x|t)\), are compatible if there exists a joint distribution \(f_{\beta,\theta}(t,x)\) having these pairs of distributions as conditional distributions. The family \(\mathcal{R} = \{f_\gamma(y|t,x) : \gamma \in \Gamma\}\) is the regression model. The indexing sets \(\Theta\), \(\mathcal{B}\), and \(\Gamma\) are either subsets of \(\mathbb{R}^d\) or specify larger semi- or non-parametric families. Let \(f_{\gamma,\theta}(y,t,x)\) denote the density for the joint distribution of the random vector \((Y,T,X)\). The available sample is \(\{(y_i,t_i,x_i)\}, i = 1, \ldots, n\) with the observations drawn from this joint distribution \(f_{\gamma,\theta}(y,t,x)\).

For situations where \(\mathcal{T}\) is a finite set, a common practice in propensity analysis is to assume a parametric form, such as

\[
Pr(T = t | X = x) = \exp\left\{ \sum_{j=0}^{d} \beta h_j(x) \right\} \left\{ 1 + \exp\left\{ \sum_{j=0}^{d} \beta h_j(x) \right\} \right\}^{-1}
\]

for some set of \(d + 1\) known functions \(h_j\). However, compatibility requires that this “working” logistic model yields an equivalent parametric exponential family conditional models for the covariates of the form

\[
f_\theta(x|T = t) = \beta(\theta, x) \exp\{Q'(\theta, t) T(x)\} c(\theta, t).
\]

For details on this necessary equivalence see Kay and Little (1987) and Arnold and Press (1989).
A Counterfactual Setting. The question “If the treatment and the covariates were independent, what would be the difference in the distributions of $Y$ across levels of $T$?” is a basic conditional question in defining and estimating causal effects of $T$ on $Y$. Indeed, to study the unconfounded effect of $T$ on $Y$ one would like the process generating the observed data to follow a member of the family of distributions:

$$
\mathcal{J} = \{(v(y|x,t), \eta(x), \pi(t)) : v(y|x,t) \geq 0, \\
\eta(x) \geq 0, \pi(t) \geq 0, \int_{\mathcal{Y}, \mathcal{X}} v(y|x,t) \eta(x) \, dx dy = 1, t \in \mathcal{T}\},
$$

where, for each member, $X$ is independent of $T$ with marginal distributions given by $\eta$ and $\pi$. Note that we intentionally use new notation to denote the relevant distributions in these unrealized worlds to put emphasis on the fact that were $T$ and $X$ independent, the data generation mechanism would be different from the corresponding mechanism in the realized world. The assignment mechanism $T$ and $X$ are dependent in the factual world generating the data with the distributions $f_T(y|x,t), f_0(x|t)$ and $f_\pi(t)$ modeled by $\mathcal{R}, \mathcal{F}$ and $\mathcal{G}$. Each choice of $v(y|x,t)$ and $\eta(x)$ defines a conditional joint distribution of the response and the covariates in a counterfactual setting where $T$ and $X$ are independent. To study the causal effect of $T$ on $Y$ using such counterfactual models, we want to choose a member of $\mathcal{J}$ which is as “close” as possible to the factual world. Therefore, we make the explicit assumptions: (i) $\pi(t) = f(t)$, (ii) for each given $x$ and $t$, the conditional distribution of $Y$ is the same as that for the observed factual universe, namely, $v_T(y|x,t) = f_T(y|x,t)$. We refer to this assumption as the regression assumption.

Under these assumptions, the derived family of counterfactual response models given $T = t$ is:

$$
\{(v_T(y|t) : v_T(y|t) = \int_{\mathcal{X}} f_T(y|x,t) \eta(x) \, dx, \eta \in \mathcal{J}, \gamma \in \Gamma\}.
$$

Corresponding to each of these counterfactual models, which are defined by different choices for $\eta(x)$, we can define a set of potential outcomes $\{Y_t(\gamma, \eta); t \in \mathcal{T}\}$. Different functional of these models then could be used for a counterfactual assessment. For example, for a dichotomous treatment $T$, an average causal treatment effect in the counterfactual setting indexed by $(\gamma, \eta)$ could be defined as

$$
\kappa_T(\eta) = E[Y_t(\gamma, \eta) - Y_0(\gamma, \eta)] = E[\eta][E_T(Y|X, T = 1) - E_T(Y|X, T = 0)].
$$

There are an infinite number of possible specifications for $\eta(x)$. Typically, a version of the ignorability assumption implies that there is just one “true” model, and one “true” causal effect of the treatment, and this is the quantity that causal inference procedures attempt to estimate. Without such an assumption, we need to consider how to choose $\eta$ in a sensible, perhaps optimal, manner.
2 Elimination of Nuisance Variables

The effect of confounding on the estimation of the effect of $T$ on $Y$ can be addressed at the design stage (using randomization, blocking, etc.) or at the analysis stage. In observational studies we are typically limited to addressing confounding using analytic methods. In doing so, one tacitly accepts the regression assumption and then attempts to control or eliminate the dependence between $X$ and $T$. Elimination of nuisance parameters is a classical problem in statistical inference. For a review of these methods, see Basu (1977) and Severini (2000). One potential approach to addressing confounding is to treat $X$ as a “nuisance” variable in the propensity model, $G$, or, equivalently, treat $T$ as a “nuisance” variable in the covariate model, $F$.

Methods to eliminate nuisance variables, whether exact or approximate, are based on factorization and summarization of the likelihood function. Classical summarization methods are marginalization, maximization, and conditioning. The factual conditional likelihood is given by $f_y(y|x,t)f_\theta(x|t)$. Under the regression assumption, to construct a counterfactual distribution for the potential outcomes one should concentrate on the forward regression model, $\mathbb{F} = \{f_\theta(x|t) : \theta \in \Theta\}$. In this model, $t$ is an unwanted index that is a nuisance in defining a measure of the direct effect of $T$ on $Y$. Therefore, one could look for an optimal method to eliminate the dependence of the covariates and $T$, which is very similar to the classical methods for elimination of nuisance parameters in the development of inference procedures. Indeed, this outlook sheds light on a number of existing methods of estimation and opens the door to a host of new procedures for causal analysis.

“Bayesian” Marginalization. Bayesian Marginalization here could be defined as a method of eliminating the nuisance variable by averaging over the nuisance variable using a well chosen weight function. However, since $T$ is an observable random variable, $f(t)$ is a natural choice for this weight function. This suggests $\eta(x)$ should be chosen to be the marginal distribution $f_\theta(x) = \int_{\mathbb{F}} f_\theta(x|t)f(t)dt$. Note that in the factual world, $T$ and $X$ are not independent and $f_\theta(x)$ is indeed a counterfactual choice.

A nice property of $f_\theta(x)$ is it yields the counterfactual balanced model closest to the factual world in the sense of squared error loss, that is

\[ f_\theta(x) = \arg\min_{\mathcal{D}} E_T (f_\theta(x|T=t) - \nu(x))^2 \]

where $\mathcal{D} = \{\nu(x) : \nu(x) \geq 0 \int \nu(x) dx = 1\}$ is the class of possible densities over the covariate space. Further,

\[ f_\theta(x) = \arg\min_{\mathcal{D}} \int_{\mathcal{D}} \int_{\mathbb{Y}} f_y(y|x,t)(f_\theta(x|T=t) - \nu(x))^2 f(t)dydt \]
which implies the family of distributions
\[
\{ \nu_{\gamma, \theta}(y \mid T = t) = \int_{\mathcal{X}} f_{\gamma}(y \mid x, t) f_{\theta}(x) \, dx, \ t \in \mathcal{T}, \gamma \in \Gamma, \theta \in \Theta \}
\] (1)
could be used as the conditional distributions of \( Y \) in counterfactual settings having the same forward regression model as the factual world but with independence between \( T \) and \( X \) with closest resemblance to the factual setting generating the data as measured by squared distance. Obviously, other measures of the closeness of a balanced \( \nu(x) \) to the actual covariate models will result in a different “counterfactual working models”.

The well known causal model discussed by Rubin (1974) and by Rosenbaum and Rubin (1983) provides another basis for this choice of \( \eta(x) \). In this model, each unit is associated with a collection of random variables called potential outcomes, \( \{ Y_t : t \in \mathcal{T} \} \), which are defined as “the outcome if the unit were to receive treatment level \( t \)”. Consider a pair of potential outcomes \( Y_0 \) and \( Y_1 \), \( T = 0, 1 \), and the observed response, \( Y = Y_0 T + Y_1 (1 - T) \). From this definition of potential outcomes one can immediately deduce that
\[
\nu_{Y_t}(y \mid x, T = t) = f(y \mid x, T = t), \ t = 0, 1
\]
which is the regression assumption. The strong ignorability assumption
\[
\nu_{Y_0, Y_1}(y_0, y_1 \mid X = x, T = 1) = \nu_{Y_0, Y_1}(y_0, y_1 \mid X = x, T = 0)
\]
immediately implies
\[
\nu_{Y_t}(y) = \int_{\mathcal{X}} \nu_{Y_t}(y \mid X = x, T = t) f_{\theta}(x) \, dx
\]
and, hence, a member of the family of distributions in (1) forms the basis for the potential outcome models in these settings. Indeed, the main task of the strong ignorability assumption is to justify use of the above model for the conditional responses.

Another argument justifying the use of this family is based on the notion of the \( do(x) \) operator introduced by Pearl. The details of this argument are given in Pearl (2009). The stratification methods described in Cochran (1968) also are based on using this marginal distribution.

**Conditioning.** In the following discussion, ancillarity and sufficiency will be highly relevant. Conditioning is based on finding a \( U(x_1, x_2, \ldots, x_p) \) such that \( X \) and \( T \) are conditionally independent given \( U \). Hence, inference can be carried
out conditional on the values of $U$. These conditional inferences can then be combined in some sensible manner. An important point to be emphasized is that, for classical methods of statistical inference, we use ancillarity in conditional modeling and sufficiency in informative marginalization. For creating balance, however, we use sufficiency for conditioning and ancillarity for the resulting non-informative distributions.

Note that in classical statistical inference, where $U$ is a sufficient statistic for a model parameter, the sufficiency principle recommends replacing the likelihood with the derived marginal likelihood for the observed sufficient statistic. The crucial property of $U$ is that, within any condition set defined by $U$, the conditional models for $X$ are free from the indexing parameter. That is for all the values of the index, the conditional distributions for $X$ are the same. Therefore, when treating $T$ as an index parameter, we can look for a quantity sufficient for the family indexed by $T$ and look within slices constructed by the sufficient quantities.

The conditionality principle advocates restricting the inferences to conditional models given maximal ancillaries. If we treat $T$ as a “parameter”, the marginal distribution of the ancillary becomes invariant with respect to $T$. Therefore, to create balance, one may identify situations where an ancillary quantity for $T$ exists and the original observed model replaced with one based on this ancillary quantity.

Using the notion of partial sufficiency from Cox and Hinkley (1974), we formally define the following notion of a balancing condition. See also Noorbaloochi and Nelson (2008) and Nelson and Noorbaloochi (2009).

**Sufficient Summaries.** For the covariate model $\mathcal{F}$, the statistic $S(\theta, X)$ is an $X$-sufficient summary for $T$ if

$$f_\theta(x|S(X), T) = f_\theta(x|S(X)),$$

that is, the distribution of $X$ over subpopulations identified by $S(X) = s$ does not depend on $T$. For the compatible propensity model $\mathcal{G}$, the statistic $S(X)$ is an $X$-sufficient summary if, for any $f_T$ in $\mathcal{G}$,

$$f_T(t|S(X)) = f_T(t|X).$$

If we think of $T$ as a “parameter” and $X$ as the data, this is similar to the notion of Bayesian sufficiency, where $S(data)$ is sufficient if $f(\text{parameter}|S(data)) = f(\text{parameter}|data)$. Note that to identify, or estimate, a sufficient summary, one does not need to know or use the propensity model $\mathcal{G}$. In the above, we intentionally suppressed the parameter $\theta$, (real-valued or otherwise), assuming it is known or can be estimated via a consistent estimator in an available large sample.

Sets of sufficient covariates discussed by Dawid (1979), Robins and Morgenstern (1987), Pearl (2009) and Greenland (2003) are special cases of sufficient
summaries. In addition, as discussed in Nelson and Noorbaloochi (2009), propensity theory can be subsumed into the theory of $X$-sufficient summaries. Given the elementary nature of these sufficient summaries and the vast literature on sufficient statistics, a host of methodologies for causal inference can be recast as methods for introducing covariate balance through conditioning. Further, many of the misrepresentations and the purported limitations of propensity theory can be clarified or refuted if the connection to sufficiency is more fully emphasized. The following elementary theorem underlies the theory latent in Rosenbaum’s and Rubin’s propensity score development (1983), Imbens (2000) and Hirano and Imbens (2004) development of generalized propensity scores, and Imai’s and Van dyke’s (2004) development of propensity theory for continuous treatment regimens.

**Overlap Assumption**: For each $x \in X$, assume $f_\theta(x|T=t) > 0$ for all $t \in T$ and $\theta \in \Theta$. Therefore, it is assumed that all the members of $\mathcal{F}$ have common support.

**Theorem 1.** Given the overlap assumption, if $T$ takes a finite number of values, say $k$, then the following versions of $X$-sufficient summaries are equivalent:

i. Density ratios: $\left( \frac{f_{t_2}(x)}{f_{t_1}(x)}, \frac{f_{t_3}(x)}{f_{t_1}(x)}, \ldots, \frac{f_{t_k}(x)}{f_{t_1}(x)} \right)$

ii. Odds: $\left( \frac{P\{T=t_2|X=x\}}{P\{T=t_1|X=x\}}, \ldots, \frac{P\{T=t_k|X=x\}}{P\{T=t_1|X=x\}} \right)$

iii. Propensity Vector: $(P\{T=t_1|X=x\}, \ldots, P\{T=t_{k-1}|X=x\})$, Provided $T$ has marginal probabilities $\pi(t)$ with $\pi(t_i) > 0$ for all $i$.

This implies that $d$, the linear dimension of the summary, is at most $k - 1$. If $T$ is countable, the summaries above can be restated in terms of a series of density ratios, odds, and propensities that yield equivalent sufficient summaries. For a continuous $T$, the following $X$-sufficient summaries are equivalent:

iv. Density ratio functional, $S: t \rightarrow \frac{f(x|T=t)}{f_t(x)}$

v. Propensity functional, $S: t \rightarrow f(t|x)$ where $f(t|x)$ is the conditional density of $t$ given the covariates.

The proof of sufficiency of density ratios is elementary and is exactly similar to that for sufficiency of likelihood ratios. Equivalence of odds ratios and propensity vectors to density ratios, for the finite case is based on noting that the transformation between any pair of the three vectors is an isomorphism. For the non-finite case, for the sufficiency of the density ratio functionals see Rao (2002) and for the equivalence of propensity functional and the density ratios for our dominated families see Blackwell and Ramamoorthi (1982).
Some immediate remarks are in order. In general, there is no univariate propensity score. In the special case of a binary \( T \), the density ratio \( r(x) = \frac{f_1(x)}{f_0(x)} \) is either one-dimensional or zero dimensional. For example, if \( f_1(x) = f_0(x) \), then \( r(x) = 1 \) and hence is zero-dimensional. However, as soon as \( k > 2 \), one generally has to use a propensity vector, sequence, or function of \( T \) to balance the confounding covariates. One may be able to generate balance using an equivalent unidimensional summary under certain underlying models. For example, proportional odds models ensure unidimensionality. A sufficient condition to ensure unidimensionality of the density ratios is the monotone density ratio property, which holds for the covariate model \( F \) if there is a monotone function \( h_{\theta}(\cdot) \) such that, for all \( t \in T \),

\[
ft(x) = h_{\theta}(x).
\]

For further illustration, we provide some examples of sufficient summaries in some common situations.

**Example 1. Exponential Family.** Assume, in each treatment group, the covariates have a conditional distribution in an exponential family with densities

\[
f_\theta(x | t) = h(\theta, x) \exp \left( \sum_j G_j(\theta, x)Q_j(\theta, t) \right) C(\theta, t).
\]

(4)

Then, for any value \( t_0 \), the density ratios

\[
S : x \rightarrow r(t) = S_t(\theta, x) = \frac{f_\theta(x | T = t)}{f_\theta(x | T = t_0)}
\]

form a sufficient summary. Further, this summary is equivalent to

\[
\sum_j d G_j(\theta, x)(Q_j(\theta, t) - Q_j(\theta, t_0)).
\]

(5)

Note that when \( T \) takes \( k \) distinct values, \( k < \infty \), a version of sufficient summary is \( \Lambda G \), where \( \Lambda = [Q_j(\theta, t) - Q_j(\theta, t_0) : t = 1, \ldots, k, j = 1, \ldots, d] \) is a \( k \times d \) matrix and \( G = (G_1(\theta, x), \ldots, G_d(\theta, x))^t \). The dimension of the summary is equal to the rank of \( \Lambda \) provided the \( G_j(\theta, x) \) are linearly independent. For example, if \( Q_j(\theta, t) = \lambda_j(\theta)b(t) \) for some functions \( \lambda_j \) and \( b \) then the summary in Equation (5) is equivalent to the unidimensional summary \( \sum_j d \lambda_j(\theta)T_j(\theta, x) \). Noorbaloochi and Nelson (2008) outline a test of the dimensionality of \( \Lambda T \) when \( \theta \) is estimated by an asymptotically normal estimator.

**Example 2. Elliptically Symmetric Covariate Models.** As evident from the previous discussion, to construct a balancing score, or sufficient summary, one needs to estimate some form of conditional density (with respect to some dominating measure). In the case of a low-dimensional covariate space, matching or
stratifying on the values of the $X$, which are in themselves an $X$-sufficient summary, may be used. Families of elliptically symmetric distributions have received much attention in the covariate dimension reduction and balancing literature due to their rather tractable estimation procedures and the optimal bias reduction properties. Specifically, under some simple conditions, matching by the values of $X$, with an underlying elliptically symmetric distribution yields equal percent bias reduction. If a matching method is not equal percent bias reducing, then matching may increase the bias for some functions of $X$. Rubin (1976) details these results. Rubin and Thomas (1992b) consider the multivariate normal distribution and linear propensity scores, and Rubin and Thomas (1992a) extend these results to elliptically symmetric distributions. Rubin and Stuart (2006) further extend these results to the family of discriminant mixtures of proportional elliptically symmetric distributions.

Assume members of the covariate model have elliptically symmetric conditional distributions given by the density

$$f_t(x) = h_t((x - \mu_t)^\Sigma_t^{-1}(x - \mu_t))$$

where $\mu_t$ and $\Sigma_t$ are location and scale parameters. It is not difficult to see that $\{\Sigma_t^{-1/2}x : t \in T\}$ jointly form an $X$-sufficient summary. For a fixed function $m$, consider the set

$$C_m = \bigcap_{t \in T} \{x : (x - \mu_t)^\Sigma_t^{-1}(x - \mu_t) = m(t)\}.$$ 

Note, for these elliptically symmetric distributions, $f_t(x \mid x \in C_m)$ is a uniform distribution and, hence, is independent of $T$. Now, note that $\Sigma_t^{-1/2}x = c(t)$ for some function $c$ implies $\Sigma_t^{-1/2}(x - \mu_t) = m(t)$ for $m(t) = c(t) - \Sigma_t^{-1/2}\mu_t$.

A useful property of this family of distributions is the resulting linear sufficient summaries for reducing the dimension and balancing the confounding covariates. Thus propensity scores, vectors, series, or functionals can be defined as linear combinations of the covariates and, hence, available linear analysis theory can be utilized in density estimation, dimensionality reduction, and balancing.

### 3 Dimensionality Reduction

In principal, dimensionality reduction (or in common parlance, dimension reduction) of the covariate space is technically an issue separate from bias reduction and balancing. However, in practice the same approach is often used to address...
both issues. Deterministic and probabilistic methods comprise the two different approaches of attacking dimensionality. While not a topic of discussion here, variations of principal component analysis and applications of the Karhunen-Loeve theorem, Fourier series analysis, wavelets, and other orthogonal basis transformations, and a number of approximation methods are classical deterministic methods directly applied to reduce the the p-dimensional random variables. Kirby (2001) presents a detailed discussion of these methods.

It is well-known that stratification and matching, while useful and simple for confounders of small dimension, quickly break down as the dimensionality of the confounders increases. Increases in bias and variance of Horvitz-Thompson type estimators are some well-known problems with a large \( p \). Curse of dimensionality results in the violation of the likelihood principle and the loss of efficiency in the inverse propensity weighted estimators (Robins and Ritov (1997)) See also Robins and Wasserman (2000).

Obviously, if \( p \) is large, estimation of the densities \( f(x|T = t) \) will require high dimensional density estimation. Consider then the propensity model \( \mathcal{G} = \{ f_{\beta}(t|x) : \beta \in \mathcal{B}, t \in \mathbb{R}, x \in \mathbb{R}^p \} \). To construct a sufficient summary or a propensity element using these models, we again may be directly faced with the potential need to implement some dimensionality reduction of the confounding covariates as we are faced with a regression with numerous covariates. In addition to these deconfounding procedures, the original regression model \( \mathcal{R} \) also runs into dimensionality issues.

The benefit or, perhaps necessity of reducing the dimension of numerous covariates for the estimation of regression functions is the basis for the vast literature on regression graphics. This literature offers an approximate solution with construction of linear sufficient summaries, \( (\beta_1'x, \beta_2'x, \ldots, \beta_k'x) \) (Li, 1991), under the assumption that

\[
f(t | \beta_1'x, \ldots, \beta_k'x) = f(t | x)
\]

(7)

where \( (\beta_1, \beta_2, \ldots, \beta_k) \) are unknown projection vectors and \( k \), while less than \( p \), is unknown. If (7) holds for a particular \( B = (\beta_1, \beta_2, \ldots, \beta_k) \) then it also holds for \( AB \) where \( A \) is any full rank matrix. Hence, finding a basis for the subspace spanned by columns of \( B \) is of primary interest. These unknown \( \beta \) are the effective dimension reduction directions (EDR-directions) (Li, 1991). The span of the \( \beta \) is also referred to as the effective dimension reduction space (EDR-space). Our immediate aim then is to estimate such basis vectors and to subsequently use these estimates, \( \hat{\beta}'X \), for further data analysis. Note, as discussed above then, by working within subpopulations defined by \( Bx = s \), these sufficient summaries simultaneously can be used to reduce the dimension of the covariates and, as given by the correspondence between Equation (3) and Equation (7), reduce the bias of estimation.
Regression dimension reduction was introduced by Li (1991). Li (1997) studies some confounding issues for high-dimensional data. Hall and Li (1993) identify problems where the linear sufficient summaries are good approximate sufficient quantities. Carroll and Li (1995) use the dimension reduction ideas for treatment-control comparisons. Cook (1996) and Cook and Lee (1999) also consider the binary response variable. The test of dimensionality in these papers can be used to see if confounding is present and how to construct linear propensity scores. Cook and Weisberg (1994) provide a detailed application of sufficient summaries in regression graphics and a comprehensive account of linear sufficient summaries and different estimation methods for semi-parametric models satisfying the linear design condition defined below. Chiaromonte, Cook, and Li (2002) address dimension reduction when some of the covariates are categorical. Cook (2007) provides a nice introduction to the methodology and Cook and Forzani (2009) develop a likelihood-based dimension reduction method. Fukumizu, Bach, and Jordan (2003) and Fukumizu, Bach, and Jordan (2009) use kernel methods to develop a methodology for constructing approximately linear sufficient directions for models that do not necessarily satisfy this condition.

Sliced inverse regression (SIR) is one of the oldest estimation procedures for constructing linear summaries. SIR finds a $k$-dimensional basis for the sufficient subspace of $\mathbb{R}^p$ using the mean regression for members of $\mathcal{F}$, $E(X|T = t)$, which is based on $p$ one-dimensional regressions. The connection between the model $\mathcal{F}$ and the model given by Equation 7 is given in the following theorem from Li (1991).

**Theorem 2.** Suppose Equation (7) and the Linear Design Condition,

$$\forall b \in \mathbb{R}^p : E\left(b'X|\beta'x\right) = c_0 + \sum_{i=1}^{k} c_i \beta_i'x,$$  \hspace{1cm} (8)

for a set of constants $c_i$ (dependent upon $b$) hold, then the centered inverse regression curve $E(X|T = t) - E(X)$ lies in the linear subspace spanned by the vectors $\Sigma \beta_i$, $i = 1, \ldots, k$, where $\Sigma = Cov(X)$.

SIR estimates the EDR directions, $\beta$, using the following simple results. For a set of vectors $\eta_1, \ldots, \eta_k$, define $\text{span}(\eta_1, \ldots, \eta_k)$ to be the subspace of all linear combinations, of the $\eta_i$ and standardize the covariates, $Z = \Sigma^{-1/2}\{X - E(X)\}$. The inverse regression curve $m_1(t) = E(Z|T = t)$ lies in $\text{span}(\eta_1, \ldots, \eta_k)$ for $\eta_i = \Sigma^{1/2}\beta_i$. With $b$ orthogonal to $\text{span}(\eta_1, \ldots, \eta_k)$, it follows that $b'm_1(t) = 0$ and, further, that $m_1(t)m_1(t)'b = Cov\{m_1(t)\}b = 0$. As a consequence, $Cov\{E(Z|T = t)\}$ is degenerate in all directions orthogonal to the EDR-directions $\eta_i$ of $Z$.

These results suggest the following algorithm for estimating the $\beta$. First, standardize the observed covariates, slice the observed values for $T$ into $S$ disjoint
intervals, and then within slices find the mean of the standardized covariates. Construct the observed covariance matrix of the within slice averages as an estimate of $\text{Cov}\{m_i(t)\}$. Find the eigenvectors for this matrix. In general, this estimated covariance matrix will have full rank because of random variability in the data generation. Therefore, we can use the eigenvectors, $\hat{\eta}_i$, of this matrix with correspondingly large eigenvalues as estimates for the EDR-direction $\eta_i$. We can rescale these to estimate $\hat{\beta}_i = \hat{\Sigma}^{-1/2}\hat{\eta}_i$ for the EDR-directions of $X$. See Li (1991) for additional details. It is important to remember that $k$, the number of linearly independent EDR-directions, is the minimum number of distinct propensity scores one needs for balancing when $T$ is multi-valued or continuous. In the continuous case, $k$ is an approximation obtained based on the slicing of $T$. The number of slices will depend on the sample size and the underlying distribution; however, five slices seems to provide a good initial approximation.

Causal Analysis Using Sufficient Summaries. As discussed in the first section, under the strong ignorability assumption when estimating expectations of $Y_t$, and in certain other scenarios, the indexed distribution

$$v_{\gamma, \theta}(y | T = t) = \int_X f_{\gamma}(y | x, t) f_{\theta}(x) \, dx$$

forms a causal response distribution using marginalization with respect to $f_{\theta}(x)$. Let $\mathcal{S}_s = \{x : S(\theta, x) = s_\theta\}$ where $S(\theta, x)$ is an $X$-sufficient summary for $T$. For any fixed set of parameters

$$v_{\gamma, \theta}(y | T = t) = \int_\mathcal{S} \int_{\mathcal{S}_s} f_{\gamma}(y | x, t) f_{\theta}(x | s) dx f_{\theta}(s) ds.$$ 

As $\int_{\mathcal{S}_s} f_{\gamma}(y | x, t) f_{\theta}(x | s) dx = f_{\gamma}(y | S = s, T = t)$ and $f_{\theta}(x) = f_{\theta}(x | S = s)f_{\theta}(s)$, then

$$v_{\gamma, \theta}(y | T = t) = \int_{\mathcal{S}} f_{\gamma}(y | S = s, T = t) f_{\theta}(S = s) ds \quad (9)$$

where $f_{\theta}(s) = \int_{\mathcal{S}} f_{\theta}(s | T = t)f(t) \, dt$. Equation (9) shows that by using the possibly lower-dimensional sufficient summaries, one can derive the same causal response model that would have been constructed from the original covariates. If equation (7) holds, we can use $k$ linear summaries estimated using SIR or some similar technique to reduce the bias in estimating the effect of $T$ on $Y$. We can investigate this in a number of different frameworks. In the following discussion, we will consider a standard counterfactual framework for a binary $T$. As will be clear, generalization to multi-valued and continuous interventions is not difficult. Again, under the strong ignorability assumption,

$$E(Y_t | S = s, T = 1) = E(Y_t | S = s, T = 0) = E(Y | S = s, T = t).$$

DOI: 10.2202/1557-4679.1209

12
Conditional on $S(X) = B'X = s$, the treatment and the control groups are balanced, that is, $X \perp T \mid B'X = s$. As discussed above, an average treatment effect can be estimated using $E(Y_1) - E(Y_0) = ES\{E(Y_1 \mid S, T = 1) - E(Y_0 \mid S, T = 0)\}$ Consider also that the average treatment effect among the treated can be estimated using the result

$$E(Y_1 \mid T = 1) - E(Y_0 \mid T = 1) = ES_{T=1}\{E(Y_1 \mid S, T = 1) - E(Y_0 \mid S, T = 0)\}$$

where now the outer expectation is taken over the conditional distribution of $S$ given $T = 1$, namely the distribution of baseline variables in the treated group. Other interesting causal parameters can be estimated in similar fashion.

## 4 Marginalization: $X$-ancillary Summaries

When balance in the covariates is established through randomization or design, the causal response distributions, $\nu_{\gamma, \theta}(y \mid T = t)$, are the same as the distributions of the response in each treatment group, $f_{\gamma, \theta}(y \mid T = t)$. In this case, the distribution of the covariates is the same within each treatment group. For situations where this is not the case, a potentially useful line of inquiry would be to investigate functions of the covariates (and possibly parameters) that have the same distribution for all levels of $T$.

Recall that an ancillary statistic is one with a parameter-free distribution. All information in the likelihood about the parameters lies in the conditional likelihood given the ancillary. This is in complete contrast to sufficiency where the conditional models have no information about the parameters. As discussed in previous sections, propensity scores and, more generally, $X$-sufficient summaries are similar to sufficient statistics in that after conditioning or partitioning the population, within partition sets the distribution of the covariates is constant across levels of $T$.

**Definition 2.** Within the covariate model, $\mathcal{F}$, the function $A(\theta, X)$ is an $X$-ancillary summary for the observable $T$ if, for all $(t, \theta) \in T \times \Theta$

$$f_\theta(x|t) = f(x|A(\theta, x) = a(\theta), t)f_\theta(a)$$

If $T$ and $X$ are independent then $X$ itself is $X$-ancillary. It is assumed that $T$ and $\theta$ are variation-independent, that is the distribution of $T$ does not depend on $\theta$ and $\Theta$ is not defined via $T$. In contrast to the use of ancillary statistics in inference whereby conditioning on the observed value of the ancillary statistic restricts attention to the conditional part of the factorized likelihood for introducing balance of the covariates across levels of $T$, the second factor is of prime interest. The
above definition parallels the notion of partial ancillarity, Cox and Hinkley (1974); $M$-ancillarity, Barndorff-Nielsen (1973); and especially, $S$-ancillarity, Barndorff-Nielsen and Blaesild (1975); the notion of ancillarity proposed by Sandved (1968) and further explored in Sandved (1972).

For simplicity in the following, we ignore $\theta$ in $A(\theta, x)$. The $X$-ancillary summary $A(X)$ and $T$ are independent, hence

\[
  f_{\gamma, \theta}(y|T = t) = \int_{\mathcal{A}} f_{\gamma}(y|A(X) = a, T = t) f_{\theta}(a) da
\]

In addition, if $f_{\gamma}(y|x, t) = f_{\gamma}(y|A(x), t)$, (i.e., if $A(x)$ which is ancillary for $T$ in $\mathcal{F}$ and is also an $X$-sufficient summary for $Y$ in the regression model $\mathcal{R}$) then averaging the usual conditional regression model results will not introduce any confounding due to imbalance in the covariates. Note that in this situation, in the presence of an $X$-ancillary and the dependence of the regression model on an $X$-ancillary, the indirect effect of $T$ on $Y$ that was exerted through $X$ will be blocked by the $X$ ancillary $A(X)$ and the confounding effect of $X$ has been eliminated in the slices defined by $A$. After characterizing such summaries, the dependence of the regression model on the ancillary summary may be empirically investigated. For example, if $E(Y|x, t) = \mu + \alpha_t + h(A(x))$ holds, then

\[
  \int_{\mathcal{X}} (\mu + \alpha_t + h(A(\theta, x))) f_{\theta}(x|T = t) dx = \mu + \alpha_t + \int_{\mathcal{A}} h_{\theta}(a)) f_{\theta}(a) da
\]

suggesting that in order to find an approximately unbiased estimator for, say, $\alpha_1 - \alpha_0$, one should first slice the population by the values of the ancillary, then within each slice construct an unbiased estimator for $\alpha_1 - \alpha_0$. Averaging these estimators with respect to the $t$-free distribution of $A$ gives an approximate unconfounded estimator. In the following example, we provide a sufficient condition under which, for a jointly normal set of covariates, $X$-ancillaries exist.

**Example 3.** Let $X|T = t \sim MVN(\mu_t, \Sigma_t)$ for $t = 0, 1$. Let $(\sigma_{1j}, \sigma_{2j}, \ldots, \sigma_{pj})$ denote the $p$ linearly independent columns of $\Sigma_t$ and

\[
  \mathcal{M} = sp\{\sigma_{11} - \sigma_{01}, \sigma_{12} - \sigma_{02}, \ldots, \sigma_{1p} - \sigma_{0p}, \mu_1 - \mu_0\}
\]

be the linear subspace generated by these column vectors. Assume, $dim(\mathcal{M}) < p$, then for any $c$ in the orthogonal subspace of $\mathcal{M}$, $c^t X$ is an $X$-ancillary quantity for $T$.

Note that $c^t X$ given $T = t$ is distributed normally with mean $c^t \mu_t$ and variance $c^t \Sigma_t c$. But $(\Sigma_1 - \Sigma_0)c = 0$ and, hence, $c^t \Sigma_1 c = c^t \Sigma_0 c$. The same holds for the means. That is, the distribution for $c^t X$ is independent of $T$. It is interesting to note that, intuitively, $\mathcal{M}$ is a space of all possible differences between the distributions across
the treatment groups. Its orthogonal complement characterizes all linear ancillary quantities. It is not surprising that, in this situation, one can show that the linear $X$-sufficient summaries span a subspace of $\mathcal{M}$. For the commonly considered case where the covariance matrices are equal, $\Sigma_t = \Sigma$, we have $\mathcal{M} = sp\{\mu_t - \mu_0, \ t \in T\}$ and, if the number of covariates is larger than the number of the treatment groups, a linear $X$-ancillary random variable exists. For the usual $T = 0, 1$, when $\Sigma$ is full rank, there are $p - 1$ linearly independent $X$-ancillary quantities. In general, there are $p - \dim(\mathcal{M})$ independent non-zero vectors that span the orthogonal complement of $\mathcal{M}$. Obviously, the larger $p - \dim(\mathcal{M})$ is the more choices of $X$-ancillaries there are. The question of the choice of an optimal ancillary issues need further investigations.

Next we apply the preceding results to the Australian Institute of Sport Data, Cook and Weisberg (1994). The data, obtained from 102 male and 100 female athletes at the Australian Institute of Sport, comprise 13 variables: Sex, red cell count (RCC), white cell count (WCC), hematocrit (Hc), hemoglobin (Hg), plasma ferritin concentration (Fe), body mass index (BMI), sum of skin folds (SSF), Body fat percentage (Bfat), lean body mass (LBM), height (Ht) in cm, weight (Wt) in kg, and sport in which the athlete competed (Sport). For illustration of the methodology discussed above we treat Sex as a grouping variable of interest and attempt to balance the eight covariates RCC, WCC, Hc, Hg, Fe, SSF, Ht and Wt. As suggested by Cook and Weisberg (1994), we log transform the covariates to better meet the assumptions of elliptically symmetric distributions underlying the use of effective dimension reduction directions and related methodologies, hence here

$$X = (\log(SSF), \log(Wt), \log(Hg), \log(Ht),$$
$$\log(WCC), \log(RCC), \log(Hc), \log(Ferr)).$$

The distributions of these variables differ by sex. An assumption of conditional multivariate normality given sex did not seem to be contradicted by the data and there appeared to be little difference between the sexes in the sample covariance matrices for the eight covariates.

**Sufficient Quantity.** By Theorem 1, under conditional multivariate normal models with means $\mu_i$, $i = 0, 1$, and common covariance matrix $\Sigma$, the log density ratios form a unidimensional sufficient quantity proportional to $S = (\mu_1 - \mu_0)'\Sigma^{-1}x$. The maximum likelihood estimate of this version of the sufficient quantity is given by

$$\hat{S}(x) = 2.32\log(SSF) - 3.97\log(Wt) - 5.02\log(Hg) - 4.11\log(Ht)$$
$$-0.161\log(WCC) - 3.27\log(RCC) + 3.21\log(Hc) - 0.55\log(Ferr)$$
### Table 1: Estimated EDR directions for Australian Athlete Data

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Dir1</th>
<th>Dir2</th>
<th>Dir3</th>
<th>Dir4</th>
<th>Dir5</th>
<th>Dir6</th>
<th>Dir7</th>
<th>Dir8</th>
</tr>
</thead>
<tbody>
<tr>
<td>log(SSF)</td>
<td>0.252</td>
<td>-0.419</td>
<td>0.097</td>
<td>-0.002</td>
<td>-0.031</td>
<td>0.012</td>
<td>0.016</td>
<td>0.018</td>
</tr>
<tr>
<td>log(Wt)</td>
<td>-0.432</td>
<td>-0.405</td>
<td>-0.327</td>
<td>0.036</td>
<td>0.238</td>
<td>-0.048</td>
<td>0.114</td>
<td>0.107</td>
</tr>
<tr>
<td>log(Hg)</td>
<td>-0.546</td>
<td>-0.512</td>
<td>-0.012</td>
<td>-0.525</td>
<td>-0.255</td>
<td>0.727</td>
<td>-0.063</td>
<td>0.865</td>
</tr>
<tr>
<td>log(Ht)</td>
<td>-0.447</td>
<td>-0.419</td>
<td>0.211</td>
<td>0.018</td>
<td>-0.891</td>
<td>0.152</td>
<td>-0.408</td>
<td>-0.345</td>
</tr>
<tr>
<td>log(WCC)</td>
<td>-0.018</td>
<td>-0.016</td>
<td>-0.029</td>
<td>-0.024</td>
<td>0.013</td>
<td>-0.046</td>
<td>-0.211</td>
<td>-0.001</td>
</tr>
<tr>
<td>log(RCC)</td>
<td>-0.356</td>
<td>-0.334</td>
<td>-0.128</td>
<td>-0.254</td>
<td>-0.077</td>
<td>-0.659</td>
<td>0.696</td>
<td>-0.296</td>
</tr>
<tr>
<td>log(Hc)</td>
<td>0.349</td>
<td>0.328</td>
<td>0.906</td>
<td>0.811</td>
<td>0.278</td>
<td>-0.099</td>
<td>-0.536</td>
<td>-0.145</td>
</tr>
<tr>
<td>log(Ferr)</td>
<td>-0.060</td>
<td>-0.056</td>
<td>0.027</td>
<td>-0.002</td>
<td>0.001</td>
<td>0.014</td>
<td>-0.007</td>
<td>-0.107</td>
</tr>
</tbody>
</table>

**Effective Dimension Reduction (EDR) Directions.** Note that to derive these directions, conditional multivariate normality is not needed. The only condition required for estimating the directions is the linear design condition which holds for multivariate normal families and other elliptically symmetric distributions. We used the R package dr, Weisberg (2002) to apply sliced inverse regression to estimate EDR directions. In the present case, two slices of sizes 102 and 100 have been used. The coefficients for the eight effective directions, Dir1 through Dir8, ordered by their corresponding eigenvalues are presented in the Table 1.

A test of dimensionality indicates that a one-dimensional summary is adequate for informative summarization. The coefficients for the first EDR direction, up to a scale factor, closely resemble the coefficients estimated for the sufficient quantity above. The correlation between the scores given by this first EDR direction and the estimated sufficient quantity above is 0.97. Note that, under the conditional multivariate Normal assumption the sufficient summary is a linear function of the original covariates but in general the X-sufficient quantities are not necessarily linear functions of the covariates.

**X-Ancillary Quantities.** Here we find seven orthogonal vectors, \( \mathbf{u}_1, \ldots, \mathbf{u}_7 \), each of which is orthogonal to the difference in covariate means \( \hat{\mu}_1 - \hat{\mu}_0 \). These then yield seven vectors \( \text{anc}_i = \mathbf{u}_i^{\prime} \mathbf{x}, \ i = 1, 2, \ldots, 7 \). Table 4 presents the three first moments of these X-ancillary quantities for each sex. The balance is evident. We also find that the correlations between the sufficient dimension reduction score corresponding to the first direction and these seven ancillary quantities are zero. Indeed, we have the following result.
Second Moment  Third Moment
---  
4  
0  
4  
1  
0  
1  
18  
1  
221  
1  
5  
36  
6  
2  
20  
1  
3  
0  
6  
94  
2

The linear regression of LBM on Sex and the covariates above co-
v a r i a t e s yields -1.7481 as the least square estimate of this gender effect. However,
mas (LBM). The linear regression of LBM on Sex and the covariates above co-
of Sport assume we are interested in studying the effect of gender on lean body
v a r i a t e s through the ancillary quantities. W i t h the data from the Australian Institute

\[ \text{Table 2: Moments of } X \text{-Ancillaries for Australian Athlete Data} \]

<table>
<thead>
<tr>
<th>Ancillary</th>
<th>Male</th>
<th>Female</th>
<th>Male</th>
<th>Female</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\text{anc}_1)</td>
<td>6.03</td>
<td>6.03</td>
<td>36.5</td>
<td>36.5</td>
<td>222</td>
<td>221</td>
</tr>
<tr>
<td>(\text{anc}_2)</td>
<td>1.76</td>
<td>1.76</td>
<td>3.16</td>
<td>3.15</td>
<td>5.72</td>
<td>5.70</td>
</tr>
<tr>
<td>(\text{anc}_3)</td>
<td>1.14</td>
<td>1.14</td>
<td>1.31</td>
<td>1.31</td>
<td>1.53</td>
<td>1.54</td>
</tr>
<tr>
<td>(\text{anc}_4)</td>
<td>4.55</td>
<td>4.55</td>
<td>20.7</td>
<td>20.7</td>
<td>94.0</td>
<td>94.0</td>
</tr>
<tr>
<td>(\text{anc}_5)</td>
<td>1.67</td>
<td>1.67</td>
<td>2.84</td>
<td>2.85</td>
<td>4.92</td>
<td>4.96</td>
</tr>
<tr>
<td>(\text{anc}_6)</td>
<td>0.29</td>
<td>0.29</td>
<td>0.10</td>
<td>0.11</td>
<td>0.04</td>
<td>0.05</td>
</tr>
<tr>
<td>(\text{anc}_7)</td>
<td>2.63</td>
<td>2.63</td>
<td>6.92</td>
<td>6.92</td>
<td>18.3</td>
<td>18.3</td>
</tr>
</tbody>
</table>

**Theorem 3.** Let \(T\) be a binary grouping variable. Assume the conditional distribution of \(X\) given \(T = t\) is multivariate normal with mean \(\mu_t\) and covariance matrix \(\Sigma_t\). Then the \(X\)-sufficient quantities are uncorrelated with the \(X\)-ancillary quantities.

To see this note that the \(X\)-sufficient summary is equivalent to the log density ratios which are proportional to \(x'(\Sigma_1^{-1} - \Sigma_0^{-1})x - 2x'\Sigma_1^{-1}\mu_1 - \Sigma_0^{-1}\mu_0\). Let \(c'x\) be an \(X\)-ancillary quantity, that is, \(c\) satisfies \(\Sigma_1c = \Sigma_0c\) and \(\mu_1'c = \mu_0'c\). Then from the known identity \(\text{Cov}(Ax, x'Bx) = 2A\Sigma B\mu\) and the covariance between linear forms we have

\[
\text{Cov}_{T=1}(x'(\Sigma_1^{-1} - \Sigma_0^{-1})x - 2x'\Sigma_1^{-1}\mu_1 - \Sigma_0^{-1}\mu_0, c'x) = \\
2c'\Sigma_1(\Sigma_1^{-1} - \Sigma_0^{-1})\mu_1 - 2c'\Sigma_1(\Sigma_1^{-1}\mu_1 - \Sigma_0^{-1}\mu_0) = \\
2c'\Sigma_1\Sigma_0^{-1}(\mu_0 - \mu_1) = 0
\]

The same result holds when \(T = 0\). Thus, with \(\mu_1'c = \mu_0'c\) yielding a zero covariance for the conditional expectations, we have the posited result for the marginal distribution of \(X\).

**Corollary 1.** Under the conditions of Theorem 2, if the first \(q\) EDR directions, \((d_1, d_2, \ldots, d_q)\), jointly form a sufficient summary for the grouping measure or treatment indicator then the remaining directions, \((d_{q+1}, d_{q+2}, \ldots, d_p)\), yield coefficients for \(X\)-ancillary quantities.

In the above example, directions Dir2 through Dir7 form a basis for the linear space spanned by \(\{u_1, u_2, \ldots, u_7\}\).

A use of ancillary quantities is to estimate the adjusted effect of the treatment on the response, provided the response regression model depends on the co-
v a r i a t e s through the ancillary quantities. With the data from the Australian Institute of Sport assume we are interested in studying the effect of gender on lean body mass (LBM). The linear regression of LBM on Sex and the covariates above co-
v a r i a t e s yields -1.7481 as the least square estimate of this gender effect. However,
the estimated coefficient for sex from a multiple regression of LBM on Sex and the balanced ancillary measures \( \text{anc}_1, \text{anc}_2, \ldots, \text{anc}_7 \) yields -19.762 as the estimate of the effect of gender on LBM.

More elaborate, sophisticated analysis should provide more precise estimates. However, the purpose of the discussions here is simply to illustrate the use of the concepts, particularly the estimation of the sufficient and ancillary quantities. In this simple analysis we did not use matching and stratification techniques. These methods are applicable when conditioning methods, such as propensity score, sufficient quantities, and EDR scores are used to estimate treatment effects. Here, the main requirement is that the regression model should be expressible as a function of ancillary quantities.

The binary assumption on the intervention variable and many of the working assumptions in the above discussion are really not essential. More general theory for the construction of the ancillary and sufficient quantities and for their use in developing estimators for causal inference need to receive more attention and effort. In this paper we have presented some initial results in these areas that we hope will attract further discussion and research.

References


