

Sufficient Dimension Reduction Summaries

David Nelson*

Siamak Noorbaloochi†

Abstract

Observational studies assessing causal or non-causal relationships between an explanatory measure and an outcome can be complicated by hosts of confounding measures. Large numbers of confounders can lead to several biases in conventional regression based estimation. Inference is more easily conducted if we reduce the number of confounders to a more manageable number. We discuss use of sufficient dimension reduction (SDR) summaries in estimating covariate balanced comparisons among multiple populations. SDR theory is related to the dimension reduction considered in regression theory. SDR summaries share much with sufficient statistics and encompass propensities. A specific type of SDR summary can wholly replace the original covariates with no loss of information or efficiency. Estimators with minimal expected loss can be based on these SDR summaries rather than all of the covariates.

Key Words: Dimension Reduction, Sufficiency, Propensity Theory

1. Introduction

We consider estimating differences in the conditional distribution of an outcome Y among k populations, $\mathcal{P}_1, \mathcal{P}_2, \dots, \mathcal{P}_k$. From each population, \mathcal{P}_z , a sample of size n_z is obtained and for each sampled unit the real-valued response variable Y and p common real-valued covariates $X = (X_1, X_2, \dots, X_p)$ are measured. Unit population membership is measured in Z , taking values in $\{1, \dots, k\}$. We assume the values (Y, X) among different units are independent and their distribution may differ across populations. For convenience, we use the notation $f_{Y,X}(\cdot, \cdot | Z = j)$, $f_X(\cdot | Z = j)$ and $f_Y(\cdot | Z = j)$ to denote the joint and marginal densities or probability densities of X and Y in the j^{th} population, even though Z may not be a random variable. We let $f_Y(y | Z = j, X = x)$ represent the conditional distribution of Y given $X = x$ in population \mathcal{P}_j .

In observational settings the specification of meaningful measures of differences in the conditional distribution of Y among the different populations can be complicated by the need to consider large numbers of potential confounding covariates X . Balanced outcome comparisons are critical to inference in these situations. For an outcome Y , set of covariates X , and grouping measure $Z \in \{1, \dots, k\}$, a covariate balanced comparison takes the form

$$\theta_g = \int \mathcal{F}(Y, f(y | Z = 1, X = x), \dots, f(y | Z = k, X = x))g(x)dx \quad (1)$$

for a known functional \mathcal{F} and a weighting density g . We assume

$$\mathcal{F}(Y, f(y | Z = 1, X = x), \dots, f(y | Z = k, X = x))$$

depends on X only through the conditional density functions for Y given Z and X . In practice, the weighting function g is typically a weighted combination of the conditional distributions of the covariates within the different populations, $\sum_j w_j f_X(x | Z = j)$, such as the marginal distribution of X . See Nelson and Noorbaloochi (2009) for further discussion of these weighting densities.

*VA HSR&D and University of Minnesota, 1 Veterans Drive, Minneapolis, MN 55417, Research supported by VA HSR&D Grant 03-005

†VA HSR&D and University of Minnesota, 1 Veterans Drive, Minneapolis, MN 55417, Research supported by VA HSR&D Grant 03-005

An example of a balanced comparison is the averaged difference in expectations

$$\int \left\{ \int y \left(f_Y(y|Z=1, X=x) - f_Y(y|Z=2, X=x) \right) dy \right\} f_X(x) dx$$

frequently discussed in propensity theory. A further example, for dichotomous Y and Z taking values in $\{1, 2\}$, is given by the covariate balanced average odds ratio

$$\int \frac{f_Y(1|Z=1, X=x)}{f_Y(2|Z=1, X=x)} \frac{f_Y(2|Z=2, X=x)}{f_Y(1|Z=2, X=x)} f_X(x|Z=1) dx$$

where we average the conditional odds ratios with respect to the marginal distribution of the covariates in one of the populations.

When X has large dimension, the direct estimation of such θ_g typically requires implementation of a large dimensional regression analysis or some similar analysis. For example, to estimate balanced difference in expectations we would need to estimate the regression functions $E(Y|Z=i, X=x)$ and to estimate balanced odds ratios we would need to estimate $f_Y(1|Z=i, X=x)/(1-f_Y(1|Z=i, X=x))$. Often direct estimation of the balanced comparisons will be based upon parameters estimated in these large dimensional regression analyses. When several potential predictor measures are available, the estimation of these parameters suffers from several well-known problems.

With larger numbers of predictors we may need larger sample sizes to estimate parameters with precision, this issue is related to the curse of dimensionality. If the underlying regression function is of low dimension then we risk overfitting using all of the predictors. Without accurate prior knowledge of the form of the regression function, we need to implement a model selection process to avoid overfitting. Several standard model identification processes can result in upwardly biased parameters estimates and test statistics. See Harrell (2001) for a broad overview of the difficulties in model selection and identification processes in regression analysis. These issues can be greatly diminished, and inference more easily conducted, if we wisely replace the original covariates with a new, reduced, set of covariates using a summary function $T(X)$, $T: \mathbb{R}^p \rightarrow \mathbb{R}^d$ with $d < p$, hopefully with d much less than p .

2. Dimension Reduction and Propensities

Rosenbaum and Rubin (1983) developed propensity theory to reduce covariate dimension in these situations. Propensity theory focuses on estimation of causal effects and hence considers a potential outcomes framework in which Y_i is the outcome for an individual under treatment or intervention condition $Z=i$. In the framework considered in the rest of this discussion we can think of the outcome Y as a combination of these potential outcomes, $Y = \sum_i Y_i I_{(Z=i)}$. The literature on propensity theory tends to focus on bias control and unbiased estimation of causal effects relative to estimation ignoring the covariates. However, one can view propensity theory solely as a dimension reduction theory.

Essentially one assumption and one conditional independence property drive propensity theory. The strong ignorability assumption drives the ability to conduct inference in an unconfounded manner. This assumption stipulates the potential outcomes are independent of group membership conditional on the set of covariates,

$$(Y_1, \dots, Y_k) \perp Z \mid X.$$

The conditional independence property satisfied by balancing scores $T(X)$, defined by the property

$$Z \perp X \mid T(X),$$

essentially states that all of the information in X about the distribution of Z is contained in $T(X)$. Balancing scores with linear dimension greater than that of X are of little interest. When X has

dimension greater than k the smallest or coarsest such balancing score is the propensity vector given by, say,

$$(Prob(Z = 1 | X), \dots, Prob(Z = k - 1 | X)).$$

Similar to minimal sufficient statistics, the propensity score is the smallest such balancing function in that the propensity score is a function of any other balancing score.

Combining this conditional independence property with the strong ignorability assumption leads to dimension reduction in the X , in that unconfounded inference can now be implemented using the smaller set of covariates $T(X)$, as $(Y_1, \dots, Y_k) \perp Z | T(X)$. We can use these balancing scores in place of the original set of covariates in forming estimates of the causal effects. In this sense then propensity theory is a theory of covariate dimension reduction.

While propensity theory reduces the covariates to a smaller dimension, estimation using these dimension reduction summaries can be less precise than estimation using the full set of covariates X as information in X about Y can be lost when X is replaced by the propensities. A natural question that then arises is whether we can reduce the dimension of the covariates and not suffer as much potential loss of precision. The conditional independence condition defining the balancing scores mimics sufficiency. In essence all of the information in X about the distribution of Z is contained in the balancing score. We use this analogy to sufficiency to develop a covariate dimension reduction theory for the estimation of covariate balanced comparisons among multiple populations that retains more of the information the covariates contain about both the outcome and population membership.

3. Sufficient Dimension Reduction Summaries for Discrete Outcomes

Consider a response variable Y taking values in $\{1, \dots, m\}$. We assume that if $f_X(x | Z = j, Y = i) > 0$ for some i and j then $f_X(x | Z = j, Y = i) > 0$ for all i and j . The following results use a generalization of sufficiency to identify sufficient dimension reduction (SDR) summary functions of X which contain all of the information in X about the distribution of Y and Z . If Z is not random these summaries contain all of the information about the distribution of Y among the different levels of Z . We start with the following definition.

DEFINITION 1. $T(X)$, with $T : \mathbb{R}^p \rightarrow \mathbb{R}^d$ and $d \leq p$, is an X -sufficient summary relative to (Y, Z) if, for all z and y ,

$$f_X(x | Z = z, Y = y) = f_X(x | T(X) = T(x)) f_T(T(x) | Z = z, Y = y).$$

$T(X)$ is an X -sufficient summary relative to (Y, Z) if and only if X is conditionally independent of (Y, Z) given T . For such T ,

$$f_Y(y | Z, X) = f_Y(y | Z, T(X)).$$

Note then that $T(X)$ captures all of the information in X regarding the association of Y and Z . The following theorem characterizes these sufficient summaries.

THEOREM 1. *The conditional density ratio vector*

$$\vec{L}(X) = \left(\frac{f_X(X | Z = 1, Y = 2)}{f_X(X | Z = 1, Y = 1)}, \dots, \frac{f_X(X | Z = i, Y = j)}{f_X(X | Z = 1, Y = 1)}, \dots, \frac{f_X(X | Z = k, Y = m)}{f_X(X | Z = 1, Y = 1)} \right)$$

satisfies

$$f_X(x | Z, Y) = f_X(x | \vec{L}(X) = \vec{L}(x)) f_T(\vec{L}(x) | Z, Y).$$

In addition, any summary function T satisfying

$$f_X(x | Z, Y) = f_X(x | T(X) = T(x)) f_T(T(x) | Z, Y)$$

is finer than $\vec{L}(X)$ in the sense that $\vec{L}(X) = h(T(X))$ for some function h .

A proof of the theorem is outlined in the Appendix. Now consider the balanced comparison of outcomes, θ_g , discussed above.

THEOREM 2. *Assume, for θ_g is as specified in Equation (1), that $g(x) = \sum_{j=1}^k w_j f_X(x | Z = j)$ for a set of weights w . Let $T(X)$ be an SDR summary relative to (Y, Z) , then*

$$\theta_g = \int \mathcal{F}(Y, Z, f(y | Z = 1, T(X) = t), \dots, f(y | Z = k, T(X) = t)) g_T(t) dt$$

for $g_T(t) = \sum_{j=1}^k w_j f_T(t | Z = j)$.

The proof of this Theorem is trivial given the assumption that \mathcal{F} depends on X only through the conditional density functions $f(y | Z, X)$. We can wholly replace the covariates X with the summary $T(X)$ in the formulation and, hence, the estimation of θ_g . Inference regarding the association between Z and Y , or how the distribution of Y changes with Z , can be based on the conditional distributions of Y given Z and the $(km - 1)$ dimensional vector of conditional density ratios with no loss of information. If both Z and Y are random then the conditional density ratio vector is equivalent to the vector of conditional probabilities

$$\vec{e}_{YZ}(X) = \left(P(Z = 1, Y = 2 | X), \dots, P(Z = 1, Y = k | X), \right. \\ \left. P(Z = 2, Y = 1 | X), \dots, P(Z = k, Y = m | X) \right).$$

This is simply a propensity vector for both Y and Z . In summary, for discrete Z and Y there is an SDR summary $T(X)$ of nominal linear dimension $km - 1$ given by these conditional density ratios for which $f(y | Z, X) = f(y | Z, T(X))$. We now turn attention to sufficient summaries for continuous Y .

4. Sufficient Dimension Reduction Summaries for General Outcomes

More generally, consider a response measure $Y \in \mathcal{Y}$. Here then Y may be continuous. Let \mathcal{G} be a family of parametric models for the conditional distribution of X given (Y, Z) which is indexed by $\eta \in \Theta$. Here X -sufficient summaries can be defined as in Noorbaloochi & Nelson (2008). Let $\Theta, \mathcal{Y}, \mathcal{X}$ denote the parameter space, response space, and covariate space, respectively. We assume $f_\eta(x | Y = y, Z = j) > 0$ for all $x \in \mathcal{X}, y \in \mathcal{Y}, j \in \{1, \dots, k\}$, and $\eta \in \Theta$. For this situation we define sufficient summaries as follows.

DEFINITION 2. *The summary $S(X)$, for the function $S(x) = (S_1(x), S_2(x), \dots, S_d(x))$ mapping \mathcal{X} onto functions from $\Theta \otimes \{1, \dots, k\} \otimes \mathcal{Y}$ to \mathbb{R}^d , $d < p$, is an X -sufficient summary relative to (Y, Z) if, for all $\eta \in \Theta$,*

$$f_\eta(x | Z, Y) = f_\eta(x | S(X) = S(x)) f_\eta(S(x) | Z, Y)$$

for all x .

If $S(X)$ is an X -sufficient summary then $X \perp (Y, Z) | S(X)$. We can again characterize these sufficient summaries in terms of density ratios.

THEOREM 3. *Let y_0 be a fixed element of \mathcal{Y} . Consider the function-valued summary*

$$T(X) = \left(\frac{f_\eta(X | Z = 1, Y = y)}{f_\eta(X | Z = 1, Y = y_0)}, \dots, \frac{f_\eta(X | Z = k, Y = y)}{f_\eta(X | Z = 1, Y = y_0)} \right),$$

a random function mapping the covariate space \mathcal{X} onto functions on $\Theta \otimes \{1, \dots, k\} \otimes \mathcal{Y}$ of the form $T_x(\eta, j, y) = f_\eta(x | Z = j, Y = y) / f_\eta(x | Z = 1, Y = y_0)$. Assume $k \leq p$. The summary T then satisfies

$$f_\eta(x | Z, Y) = f_\eta(x | T(X) = T(x)) f_\eta(T(x) | Z, Y)$$

for all η .

If Y is categorical these results yield the sufficient summaries discussed in the previous section. Furthermore, the results of Theorem 2 apply here as well. Noorbaloochi & Nelson (2008) discussion X -sufficient summaries in Exponential families of distributions. For exponential families of distributions we often can find simpler, convenient forms for these SDR summary functions

In summary, for categorical Y and for frequently considered families of conditional distributions for a continuous Y given Z and X we can identify sufficient summaries satisfying $f(y|Z, X) = f(y|Z, T(X))$. The smallest or minimal summaries are characterized by the ratios of the conditional densities of X given Y and Z . For categorical Y and random Z these conditional density ratios can be viewed as an expanded propensity function, that is, these ratios are equivalent to the vector of conditional density functions for Y and Z given X .

We can reformulate the above arguments considering X -sufficient summaries with respect to just Z . This leads to minimal dimension reduction summaries given by the conditional density ratios for X given Z . These are equivalent to the propensity vector for Z . However, while using this mathematical framework we can establish the results of Theorem 2 for averaged differences in conditional expectations for Y given X and Z , for other estimands it is not clear whether such wholesale replacement of X with the dimension reduction summary holds.

5. Reduced Expected Loss and Sufficient Dimension Reduction Summaries

If $T(X)$ is an SDR summary relative to Y and Z then $T(X)$ contains all the information in X about the association between Y and Z . For $T(x) = t$,

$$f_Y(y|X = x, Z = j) = f_Y(y|T(X) = t, Z = j)$$

and the original regression functions are really regression functions of Z and $T(X)$. We then can replace the original regression problem with one involving $T(X)$ rather than X . This is not the case with sufficient summaries relative to just Z , such as propensity scores, which omit information in X about Y . As long as Y is not independent of X given Z , X will contain information about Y beyond the information contained in the propensity score and estimation using X can then be more precise. Estimation using an SDR summary relative to both Y and Z to reduce the number of covariates will not suffer from such deficiencies.

For simplicity assume Z is random. Let $T(X)$ be the conditional density ratios for X relative to Y and Z and let $\delta(Y, Z, X)$ be an estimator of some θ . Assume $L(\delta, \theta)$ is a loss function which is convex with respect to δ . Let $\phi(Y, Z, T) = E(\delta(Y, Z, X) | Y, Z, T)$. A simple application of Jensen's inequality yields

$$\begin{aligned} E\{L(\phi, \theta)\} &= \int L(\phi(Y, Z, T(X)), \theta) f_T(t | Z = z, Y = y) f_{YZ}(y, z) dt dy dz \\ &\leq \int L(\delta(Y, Z, X), \theta) f_X(x | Z = z, Y = y) f_{YZ}(y, z) dx dy dz = E\{L(\delta, \theta)\}. \end{aligned}$$

For example, for an unbiased predictor δ the predictor ϕ satisfies $E(\delta(Y, Z, X)) = E(\phi(Y, Z, T))$ and, for squared error loss, has smaller loss as $var(\phi(Y, Z, T)) \leq var(\delta(Y, Z, X))$. These results are easily extended to estimators using observations from samples of independent, identically distributed elements by demonstrating that the conditional density ratios for the vector of observations is equivalent to the vector of conditional density ratios for the independent elements. Hence for a sample, (y_i, z_i, x_i) , $i = 1, \dots, n$ it suffices to consider estimators that are functions solely of the $(y_i, z_i, T(x_i))$.

6. Final Comments

Sufficiency is a powerful statistical concept with direct interpretation as a dimension reduction theory. Here, we used sufficiency to develop the concept of sufficient dimension reduction summaries

for use in the estimation of balanced comparisons of outcomes among multiple discrete populations. For random polytomous Z , the conditional density ratios with respect to Z are invertible transformations of extended propensity scores. Conditional density ratios relative to (Y, Z) possess more optimal properties with respect to expected loss and should offer better performance in estimating outcome differences than the propensity vector.

The SDR summary results for continuous Y and polytomous Z are similar to the results presented in Noorbaloochi & Nelson (2008) where we considered more general regression settings. There we detailed how sufficient summaries, defined in a manner similar to that presented here, generalize sliced inverse regression and related dimension reduction approaches considered in regression theory (Li 1991, Cook & Weisberg 1991, Li, 1992). Briefly, the SDR summary property, $f(Y, Z|X) = f(Y, Z|T(X))$ mirrors the standard assumption in these theories that $f(Y|X) = f(Y|\Gamma X)$ for some linear transformation $\Gamma: \mathbb{R}^p \rightarrow \mathbb{R}^d$, $d < p$. Noorbaloochi and Nelson (2008) generalize this condition to $f(Y|X) = f(Y|S(X))$ for an X -sufficient summary function $S: \mathbb{R}^p \rightarrow \mathbb{R}^d$, $d < p$. We see therefore that sufficiency and these sufficient summaries underly and unify important areas of statistical theory.

7. Appendix

Proof Theorem 1. For all $x' \in \{x' : \vec{L}(x') = \vec{c}\}$ and for all i, j

$$f_X(x' | Z = i, Y = j) = c_{ij} f_X(x' | Z = 1, Y = 1)$$

and so $f_X(x | Z = i, Y = j, \vec{L}(X) = \vec{c})$ is constant across all i, j . If $T(X)$ is another summary that satisfies the definition then, given $T(x) = c$, we have that

$$\frac{f_X(x | Z = i, Y = j)}{\int_{T(q)=c} f_X(q | Z = i, Y = j) dq} = \frac{f_X(x | Z = i', Y = j')}{\int_{T(q)=c} f_X(q | Z = i', Y = j') dq}$$

for all i, j, i', j' . Hence

$$\left\{ x | T(x) = c \right\} \subseteq \left\{ x | \vec{L}(x) = \left(\frac{\int_{T(q)=c} f_X(q | Z = 1, Y = 2) dq}{\int_{T(q)=c} f_X(q | Z = 1, Y = 1) dq}, \dots, \frac{\int_{T(q)=c} f_X(q | Z = k, Y = m) dq}{\int_{T(q)=c} f_X(q | Z = 1, Y = 1) dq} \right) \right\}$$

and, therefore, T is finer than \vec{L} .

Proof Theorem 3. Let y_0 be a fixed element of \mathcal{Y} . Consider the sets

$$L(g_\eta) = \{x' : T(x') = g_\eta(y) = (g_{\eta,1}(y), \dots, g_{\eta,k}(y))\}.$$

For any $x' \in L(g_\eta)$ and for all $\eta \in \Xi$ and all $y \in \mathcal{Y}$

$$f_\eta(x' | Z = j, Y = y) = g_{\eta,j}(y) f_\eta(x' | Z = 1, Y = y_0).$$

So again the distribution of X given $T(X) = L(g_\eta)$ is constant across all values for Z and Y .

REFERENCES

- Cook, R.D., Weisberg, S. (1991), "Discussion of 'Sliced Inverse Regression' by K.C. Li," *Journal of the American Statistical Association*, 86, 328–332.
- Harrell, F.E. Jr. (2001), *Regression modeling strategies with applications to linear models, logistic regression, and survival analysis*, New York: Springer-Verlag.
- Li, K.C. (1991), "Sliced inverse regression for dimension reduction (with discussion)," *Journal of the American Statistical Association*, 86, 316–342.
- Li, K.C. (1992), "On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma," *Journal of the American Statistical Association*, 87, 1025–1039.
- Nelson, D., Noorbaloochi S. (2009), "Dimension reduction summaries for balanced contrasts" *Journal of Statistical Planning and Inference*, 139, 617–628.
- Noorbaloochi S., Nelson, D. (2008), "Conditionally specified models and dimension reduction in the exponential families," *Journal of Multivariate Analysis*, 99, 1574–1589.
- Rosenbaum, P.R., Rubin, D.B. (1983), "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70, 41–55.