Sufficient Dimension Reduction Summaries

David Nelson, PhD

January 18, 2008

Joint work with Siamak Noorbaloochi, PhD

Research supported by VA Health Services Research IIR 03-005

Outline

Dimension Reduction in Standard Observational Framework

- Standard Observational Study Framework
- Role and Benefits of Covariate Reduction
- Sufficient Dimension Reduction Summaries
- SDR Summary Estimation

A Standard Observational Study Framework

Population P with k subgroups P_j indexed by $Z \in \{1, \ldots, k\}$

Each unit in P has values for Z, outcome Y, p covariates X

Subgroup membership, Z, may be random or fixed

Our goal is to estimate differences in distributions for Y given Z but X may play confounding role

A Standard Observational Study Framework

E.g., average difference in conditional expectations

$$\int_X \Big(E(Y \,|\, Z = 1,\, X = x) - E(Y \,|\, Z = 0,\, X = x) \Big) f_X(x) dx$$

E.g., for $Y \in \{0, 1\}$ estimate average risk ratio

$$\int_X \frac{f_Y(1 \,|Z = 1,\, X = x)}{f_Y(1 \,|Z = 0,\, X = x)} f_X(x \,|\, Z = 1) dx$$

A Standard Observational Study Framework

In general consider estimating a balanced group comparison

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

for a known functional $\mathcal{F}$ and a weighting density function

$$g_X(x) = \sum_{j=1}^{k} w_j f_X(x \,|\, Z = j)$$

Assume $\mathcal{F}$ depends on X only through the $f(y \,|\, Z = i, X = x)$

Why Reduce Covariate Dimension?

To estimate these group comparisons we need to estimate conditional distribution functions, or regression functions,

$$f(y \mid Z = i,\, X = x)$$

Several issues then arise:

- Curse of Dimensionality and Overfitting
- Difficulty in model identification
- Potential bias in model identification

These issues greatly diminished if wisely transform covariates to handful of new covariates

One Way to Reduce Covariate Dimension

Consider the balancing property from Propensity Theory

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

In the potential outcomes framework this assumption together with the strong ignorability assumption

$$(Y_0, Y_1) \perp Z \mid X$$

leads to

$$(Y_0, Y_1) \perp Z \mid T(X)$$

Sufficient Dimension Reduction Summaries wrt (Y, Z)

Assume

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

or equivalently

$$f_X(x \mid T(X) = t, Z, Y) = f_X(x \mid T(X) = t)$$

$T(X)$ analogous to Sufficient Statistic

Minimal SDR Summaries wrt (Y, Z)

Conditional density ratio vector for categorical $Y \in \{1, \ldots, m\}$

$$L(X) = \left(\frac{f_X(X \mid Z = 1, Y = 2)}{f_X(X \mid Z = 1, Y = 1)}, \ldots, \frac{f_X(X \mid Z = k, Y = m)}{f_X(X \mid Z = 1, Y = 1)} \right)$$

Conditional density ratio vector function for continuous Y, for a given y_0

$$L(X;\, y) = \frac{f_X(X \mid Z = 1, Y = y)}{f_X(X \mid Z = 1, Y = y_0)}, \ldots, \frac{f_X(X \mid Z = k, Y = y)}{f_X(X \mid Z = 1, Y = y_0)}$$

Sufficient Dimension Reduction Summaries wrt (Y, Z)

Easily rewrite

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

as

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

for weighting density function

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

Sufficient Dimension Reduction Summaries wrt (Y, Z)

Let $T(X)$ be SDR summary for X wrt to Y and Z

Let $\widehat{\delta}(Y, Z, X)$ be an estimator of some θ

Let

$$\phi(Y, Z, T) = E(\delta(Y, Z, X) \,|\, Y,\, Z,\, T(X))$$

Then $E(\widehat{\delta}(Y, Z, X)) = E(\phi(Y, Z, T))$ and

$$Var(\phi(Y, Z, T)) \leq Var(\widehat{\delta}(Y, Z, X))$$