

# Topics in Dimension Reduction

Fall 2015

Instructor: R. D. Cook

In this course we will study dimension reduction in three distinct but related regression contexts: Dimension reduction for graphics in regression, sufficient dimension reduction and envelopes. Broadly viewed, dimension reduction has always been a central statistical concept. In the second half of the nineteenth century ‘reduction of observations’ was widely viewed as a core goal of statistical methodology, and in 1922 Fisher formulated the concept of a sufficient statistic as a means of reduction without loss of information.

**Dimension reduction for graphics.** This topic will be driven by asking if we can construct a few low-dimensional *sufficient summary plots* that contain all or nearly all of the relevant regression information without pre-specifying a parsimoniously parameterized model. If so, then the plots themselves can be used as a guide to understanding the regression. (two-three weeks)

**Sufficient dimension reduction** In recent years sufficient dimension reduction (SDR) has been used to denote a body of new ideas and methods for dimension reduction. Like Fisher’s classical notion of a sufficient statistic, SDR strives for reduction without loss of information. But unlike sufficient statistics, sufficient reductions may contain unknown parameters and thus need to be estimated. In the context of regression, a reduction  $R(X)$  of the  $p$ -dimensional predictor  $X$  is sufficient if the conditional distribution of the response  $Y$  given  $X$  is the same as the distribution of  $Y$  given  $R(X)$ .

The modern meaning of SDR was introduced in the late 1990’s (Cook 1998; Cook and Yin, 1999, JASA, p. 1187-1200) in the context of regression graphics, where the overarching goal is to find a sufficient summary plot. Since that time the phrase and the associated ideas have been used with increasing frequency in the statistics literature, with ever more ambitious goals. SDR is now serviceable outside the context of regression graphics. (two-three weeks)

**Envelopes.** An envelopes is a nascent construct for increasing efficiency in multivariate statistics without altering the traditional goals. Essentially a form of targeted dimension reduction that is a descendent of regression graphics and SDR, envelope estimators and predictors have the potential to be substantially less variable than standard estimators, sometimes equivalent to taking thousands of additional observations.

Envelopes also link with some standard multivariate methodology. For instance, partial least squares regression depends fundamentally on an envelope, which can be used as a well-defined parameter that characterizes partial least squares. The establishment of an envelope as the nucleus of partial least squares then opens the door to pursuing the same goals but using envelope estimators that can significantly improve upon partial least squares predictions.

Envelopes are not limited to linear models and can work well for predictions based on high dimensional regressions. (rest of the semester)