Topics in Dimension Reduction

Spring 2019
R. D. Cook
email: dennis@stat.umn.edu

In this course we will study dimension reduction in three distinct but related regression contexts: Envelopes, sufficient dimension reduction (SDR) and dimension reduction for regression graphics. These overarching topics cover a number of standard methods, including principal component regression, partial least squares, path models, and reduced-rank regression. We will attempt to construct common frameworks that may suggest new theory and methods.

Envelopes. Envelopes, which were introduced by Cook, Li and Chiaromonte (2007, 2010), are relatively new constructs for increasing efficiency in multivariate statistics without altering the traditional goals. Essentially a form of targeted dimension reduction that is a descendent of SDR, envelope estimators have the potential to be substantially less variable than standard estimators, sometimes equivalent to increasing the sample size many times over. 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.

Sufficient dimension reduction. Sufficient dimension reduction (SDR) denotes a body of 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.

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. This topic will be addressed in concert with envelopes and SDR.

Primary references. In addition to readings from the literature, the following three books will form the primary references for the course. The first (Cook 2018) is required.

Cook, R. D. (2018). *An Introduction to Envelopes.* New York: Wiley.

Li, B. (2018). *Sufficient Dimension Reduction.* Boca Raton, FL: Chapman and Hall/CRC press.

Cook, R. D. (1998). *Regression Graphics.* New York: Wiley.