Sufficient Dimension Reduction with Multiple Populations

Xuerong Wen

November 29, 2016

Abstract

The importance of dimension reduction becomes more critical nowadays since modern scientific innovations allow scientists to collect massive and high-dimensional data at a rapid rate. For these high-dimensional data, traditional methods developed for smaller data sets can break down due to the curse of dimensionality (Friedman, 1994). Developing new ways to extract relevant information from high dimensional data sets, while grasping the important features or patterns in the data, is of great practical importance. Li (1991) and Cook (1998) proposed sufficient dimension reduction that aims at reducing the dimension of $X$ while preserving the regression relationship between the response $Y$ and $p$-dimensional predictor $X$.

In this talk, we explore sufficient dimension reduction methods for multiple populations, which is of special interest in applications. Consider the AIS dataset discussed by Weisberg (2005), which contains information on the lean body mass $L$ and other physical and hematological measurements ($X$), from 102 male and 100 female elite Australian athletes who trained at the Australian Institute of Sport. We investigate how the relationship between the body fat and various predictors varies with gender. Suppose that subject matter knowledge and prior modeling experience suggest that a $d$-dimensional multi-index model of the form $Y = g(\beta_1^T X, \beta_2^T X, ..., \beta_d^T X)$ applies to both female and male groups, naturally, we would like to know if the equivalent set of indices of the hematological measurements serve for both genders. Informal comparisons such as those based upon graphical methods can be carried out. However, such comparisons might become unwieldy when $d$ is greater than 2, and the resulting conclusions could be overly subjective. Hence, a formal test seems necessary here.

We propose two link-free procedures for testing whether two or more multi-index models share identical indices via the sufficient dimension reduction approach. The asymptotic null distributions of our test statistics are derived. Monte Carlo studies are performed to investigate the efficacy of our proposed methods. A real-world application is also considered.