Dimension Reduction in Discrete Response Regression

Hakbae Lee
American University
4400 Massachusetts Ave, NW
Washington, DC 20016-8050, USA
hblee@american.edu

1. Introduction

There has been considerable work on finding dimension-reduction subspaces for regressions with a continuous response: Sliced Inverse Regression (SIR, Li 1991), principal Hessian directions (pHd, Li 1992) and Sliced Average Variance Estimation (SAVE, Cook and Weisberg 1991). However, the direct use of those methods may not be "best" when the response is not continuous. The goal of this paper is dimension reduction in discrete response regression where the response has a finite number of values.

2. Statistical Inference

The fundamental methodology is based on the inverse regression of the predictors on the response, rather than the forward regression of the response on the predictors. First (means) and second (covariances) moments of the inverse regression are used as basic tools in this study. Cook and Lee (1999) showed that when the response has only two possible values, the SAVE paradigm is the most comprehensive. Furthermore, the SAVE paradigm is more comprehensive than SIR or DOC when the response a finite number of values.

In practice, a set of three methods is recommended: SIR, DOC (Difference Of Covariances: Cook and Lee, 1999), and SAVE. The pHd method seems to have no further contribution. If some location (mean) effect is dominat, SIR is recommended to use. If some scale (covariance) effect is dominat, DOC is useful to capture sufficient dimension reduction. Generally SAVE seems to explain both of location and scale effect.

A major part of inference is to estimate and test the dimension of the subspace constructed by each method. Consider a random symmetric matrix M_n from a method. Suppose that M_n converges in probability to a symmetric matrix M, and that the subspace formed by M is a subspace of $S_{y|x}$ where $S_{y|x}$ is called the central space which is the inferential object in this study (Cook 1994). The central space $S_{y|x}$ is spanned by left singular vectors corresponding to the singular values which are inferred to be non-zero in the population. Asymptotic derivation of asymptotic distribution of singular values was required. Generally, the test statistics are distributed as linear combinations of χ^2 random variables asymptotically. Under the assumption

of conditional normality for the predictors given the response, the linear combinations of χ^2 random variables reduce to central χ^2 random variable.

REFERENCES

Cook, R. D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association* **89**, 177–189.

Cook, R. D. and Lee, H. (1999). Dimension Reduction in Binary Response Regression. *Journal of the American Statistical Association* **94**, 1187–1200.

Cook, R. D. and Weisberg, S. (1991). Discussion of "Sliced inverse regression" by K. C. Li. *Journal of the American Statistical Association* **86**, 328–332.

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.

RESUME

Dimension Réduction à Discret Réponse Rérgression

Sans perte de renseigment original le paradigme pour dimension réduction est utile à données avec haute corrélation. Pendant la dernière deécade il y a considérable œuvre sur dimension réduction à rérgression avec la réponse continu: *Sliced Inverse Regression (SIR, Li 1991), principal Hessian directions (pHd, Li 1992), Sliced Average Variance Estimation (SAVE, Cook et Weisberg 1991)* et *Difference of Covariances (DOC, Cook et Lee 1999).* En particulier, quand une réponse est binaire, le *SAVE* paradigme est le plus compréhensif (Cook et Lee 1999). L'étude sera adressé *SIR, DOC, pHd* et *SAVE* sur rérgression avec la réponse discret.