Published in final edited form as:
J R Stat Soc Series B Stat Methodol. 2008 November ; 70(5): 903–.

Discussion of “Sure Independence Screening for Ultra-High Dimensional Feature Space

Hao Helen Zhang
Campus Box 8203, North Carolina State University, Raleigh, NC 27695-8203, U.S.A

Discussion

We congratulate the authors for their thought-provoking and fascinating work on a fundamental yet challenging topic in variable selection. Driven by the pressing need of high dimensional data analysis in many fields, the problem of dimension reduction without losing relevant information becomes increasingly important. Fan and Lv successfully tackled the extremely challenging case, where $\log(p) = O(n^{\xi})$, $\xi > 0$. The proposed Sure Independence Screening (SIS) is a state of the art method for high dimensional variable screening: simple, powerful, and having optimal properties. This work is a substantial contribution to the area of variable selection and will also make a significant impact in other scientific fields..

Extension to nonparametric models

In linear models, marginal correlation coefficients between linear predictors and the response are effective measures to capture strength of their linear relationship. However, correlation coefficients generally do not work for ranking nonlinear effects. Consider the additive model,

$$Y_i = \sum_{j=1}^{p} f_j(X_{ij}) + \varepsilon_i, \quad i = 1, \cdots, n,$$

where f_j takes an arbitrary nonlinear function form. Motivated by the ranking idea of the SIS, one could first fit a univariate smoother for each predictor and then use some marginal statistics to rank the covariates. Many interesting questions arise in this approach. Firstly, what are good measures to characterize the strength of the nonlinear relationship fully? Possible choices include nonparametric test statistics, p-values, and goodness-of-fit statistics like R^2. But which is best? Also, how do we develop the consistent selection theory for the procedure of screening nonlinear effects? All of these questions are challenging because of the complicated estimation that is involved in nonparametric modeling. It would be interesting to explore whether and how the SIS can be extended to this context.

Connection to multiple hypotheses testing and false discovery rate control

The variable selection problem can be regarded as the problem of testing multiple hypotheses: $H_1 : \beta_1 = 0, \ldots, H_p : \beta_p = 0$. Screening important variables is hence equivalent to identifying the hypotheses to be rejected. The false discovery rate (Benjamini and Hochberg,

1995) has been developed to control the proportion of false rejections. Some consistent procedures based on individual tests of each parameter have been developed (Potscher 1983; Bauer et al. 1988). Recently, Bunea et al., (2006) considered the case when *p* increases with *n*, and showed the false discover rate or Bernoulli adjustment can lead to consistent selection of variables under certain conditions. Their method is based on the ordered *p*-values of individual *t*-statistics for testing $H_j : \beta_j = 0$, $j = 1, \ldots, p$. It would be interesting to compare the SIS with these adjusted multiple hypotheses testing approaches.

References

Bauer P, Potscher BM, Hackl P. 1988; Model selection by multiple test procedures. Statistics. 19:39–44.

Benjamini Y, Hochberg Y. 1995; Controlling the false discovery rate: a practical and powerful approach to multiple hypotheses testing. Journal of Royal Statistical Society, B. 57:289–300.

Bunea F, Wegkamp M, Auguste A. 2006; Consistent variable selection in high dimensional regression via multiple testing. Journal of Statistical Planning and Inference. 136:4349–4364.

Potscher B. 1983; Order estimation in ARMA models by Lagrange multiplier tests. Annals of Statistics. 11:872–885.