Discussion of 'Sure Independence Screening for Ultrahigh Dimensional Feature Space' by J. Fan and J. Lv

May 28, 2008

Kofi Adragni and R. Dennis Cook (*University of Minnesota*)

Fan and Lv give a compelling case for predictor screening based essentially on the strength of marginal linear relationships. We have been working on screening methodology, called Screening by Principal Fitted Components (SPFC; Cook, 2007), in the $p \gg n$ context. Instead of using marginal relations, or a forward regression of Y on $\mathbf{x}$ as in penalized least squares, we adopt an inverse regression approach regressing $\mathbf{x}$ on Y. Using Fan and Lv's assumption of spherical predictors, consider the relatively simple inverse regression model

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\lambda}\mathbf{f}_y + \boldsymbol{\varepsilon}. \tag{1}$$

The term $\mathbf{f}_y \in \mathbb{R}^r$ is a user-selected function of y, $\boldsymbol{\mu} \in \mathbb{R}^p$, $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times d}$, $\boldsymbol{\lambda} \in \mathbb{R}^{d\times r}$ has rank $d \leq r$, and $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I})$. Cook (2007) showed that $Y \perp\!\!\!\perp \mathbf{x}|\boldsymbol{\Gamma}^T\mathbf{x}$. With $d = 1$, estimating the sparse subspace span$(\boldsymbol{\Gamma})$ is equivalent to estimating the sparse subspace span$(\boldsymbol{\beta})$ in equation (1) of the paper, the zero elements of $\boldsymbol{\Gamma}$ identifying the predictors to be eliminated. A value of $p > n$ need not severely hinder the estimation of span$(\boldsymbol{\Gamma})$ in inverse regression models, particularly when var$(\mathbf{x}|Y)$ is a diagonal matrix.

Let $\boldsymbol{\Gamma} \in \mathbb{R}^p$, $\mathbf{f}_y = y - \bar{y}$, $\boldsymbol{\Phi} = \boldsymbol{\Gamma}\boldsymbol{\lambda}$ and $\mathbf{X}^T = (\mathbf{x}_1, ..., \mathbf{x}_n)$. The MLE under the inverse model (1) of the $p \times 1$ vector $\boldsymbol{\Phi} = (\phi_1, \phi_2, ..., \phi_p)^T$ is $\mathbf{X}^T(y - \bar{y})$. After columnwise standardization of $\mathbf{X}$ this corresponds to the p-vector $\boldsymbol{\omega}$ of expression (2) in the paper. Consequently, SPFC reduces to SIS with $\mathbf{f}_y$ restricted to $y - \bar{y}$. Following Fan and Lv, we can select predictors by taking the first $[\gamma n]$ with the largest standardized $|\phi_i|$. But we can also tie the selection to a test statistic for $\phi_i = 0$, which automatically gives the same ordering.

The inverse regression approach is more flexible than forward regressions. Unlike SIS or penalized least squares, inverse regression models can easily accommodate a categorical response, nonlinearities and non-constant variance $\text{var}(Y|\mathbf{x})$ and still perform well. As an example, we generated $n = 70$ observations on $p = 500$ independent predictors $\mathbf{x} = (X_1, ..., X_p)^T$ with $X_1 \sim \text{Unif}(1, 10)$ and $X_i \sim N(0, 4)$, $i = 2, ..., p$. The response was generated as $y = (5X_1)\epsilon$ where $\epsilon \sim N(0, 1)$. We used SIS and generated 200 datasets to estimate the frequency that the only active predictor X_1 is among the first 35 (Fan and Lv's $\gamma = 0.5$) with the largest standardized $|\phi_i|$. The result was as expected under random selection: SIS included X_1 in the first 35 predictors 12% of time. On the other hand, SPFC with a piecewise linear basis $\mathbf{f}_y$ captured X_1 among the first *two* predictors 98% of the time. We have obtained qualitatively similar differences with nonlinear mean functions and a constant variance. Our results so far suggest that SPFC effective subsumes SIS.

References

Cook, R.D. (2007) Fisher Lecture: Dimension Reduction in Regression. *Statistical Science* **22**, 1-26