Seconding the vote of thanks: Sure Independence Screening by Jianqing Fan & Jinchi Lv

Peter Bühlmann

Seminar für Statistik, ETH Zurich, CH-8092 Zurich, Switzerland

I congratulate Fan and Lv for their stimulating and thought-provoking paper. Variable screening is among the primary goals in high-dimensional data analysis. Having a computationally efficient and statistically accurate method for retaining relevant and deleting thousands of irrelevant variables is highly desirable.

Sure independence screening (SIS) is a marginal method. This makes it very easy to use. In order to understand the properties of a marginal view, consider the well known relation for a linear model of the form $Y = \sum_{j=1}^{p} \beta_j X^{(j)} + \varepsilon$:

$$\beta_j \neq 0 \iff \mathrm{Parcorr}(Y, X^{(j)} | \{X^{(k)};\ k \neq j\}) \neq 0.$$

Of course, if $\mathrm{Corr}(X^{(j)}, X^{(k)}) = 0$ for $j \neq k$ there is an exact correspondence to the marginal view:

$$\beta_j \neq 0 \iff \mathrm{Corr}(Y, X^{(j)}) \neq 0.$$

And in fact, Fan and Lv justify SIS for the situation with fairly uncorrelated variables: their discussion of Condition 4 in Section 5.1 implies (for large p) that the correlation matrix among the X-variables is “not too far away” from the identity. In contrast to the purely marginal view, it is possible to start with the marginal approach and then gradually consider partial correlations from low to higher order. This can be achieved within the framework of so-called faithful distributions, a concept which is mainly used in the literature about graphical modeling. For linear models, Bühlmann and Kalisch (2008) introduce partial faithfulness which holds if and only if for every j:

$$\mathrm{Parcorr}(Y, X^{(j)} | X^{(\mathcal{S})}) = 0 \text{ for some } \mathcal{S} \subseteq \{1, \ldots, p\} \setminus j \Longrightarrow \beta_j = 0.$$

Bühlmann and Kalisch (2008) argue that the class of linear models satisfying this condition is quite broad. Roughly speaking, the partial faithfulness assumption implies that a large (in absolute value) marginal or partial correlation does not tell us much, but a zero (partial) correlation says a lot. The idea of SIS is the other way round: a large marginal correlation is interpreted as importance for the corresponding variable while no decision is taken for small correlations. The PC-algorithm (Spirtes et al., 2000) exploits the partial faithfulness assumption. Instead of assuming fairly uncorrelated covariates (as in SIS) or partial faithfulness, the Lasso (Tibshirani, 1996) is another alternative which requires some coherence assumptions for the design matrix ruling out cases with too strong linear dependence (of certain design sub-matrices).

method	assumption	computational complexity
SIS	“fairly” uncorrelated covariates	$O(np)$
Lasso	coherence conditions for design	$O(np\min(n,p))$
PC-algo	partial faithfulness	$O(np^{\gamma})$ $(1 \le \gamma \le C)$

The exponent γ in the computational complexity of the PC-algorithm depends on the underlying sparsity. Asymptotic theory for high-dimensional settings include: for the Lasso (Meinshausen and Bühlmann, 2006; van de Geer, 2008; Meinshausen and Yu, 2007; Zhang and Huang, 2007; Bickel et al., 2007); for the PC-algorithm (Kalisch and Bühlmann, 2007; Bühlmann and Kalisch, 2008). For finite samples, we consider two simulation models:

Model (1): example III from Section 4.2.3; $p = 1000,\ n = 50;\ \rho = 0.5$
Model (2): $Y = \sum_{j=1}^{5} X^{(j)} + \varepsilon$; $p = 1000,\ n = 50;\ \rho = 0.5$

For model (1), Fan and Lv report that $\mathbb{P}[\mathcal{M}_{\mathrm{true}} \subseteq \widehat{\mathcal{M}}] = 0$ for SIS and the Lasso when using $|\widehat{\mathcal{M}}| = n - 1$. But some differences between the methods can be easily detected in Figures 1 and 2. In addition to the single number $\mathbb{P}[\mathcal{M}_{\mathrm{true}} \subseteq \widehat{\mathcal{M}}]$, it is important to report also performance measures such as the number of true positives or ROC-curves to get a more complete picture. In our small simulation study we see that the Lasso (and also the PC-algorithm) has a better “global” accuracy than SIS. The price to pay for this higher accuracy is a more complicated procedure although we should note that the Lasso has also linear computational complexity in dimensionality p if $p \gg n$. Interestingly, we note that SIS does well in the conservative domain where the false positive rate is very low. I do not know whether we can expect such a behavior in a wide variety of scenarios: if such findings would be true in general, this would indeed be a strong argument in favor of the simple SIS method for detecting very few but most relevant variables among say thousands of others. A (presumably difficult) theory which would support such a finding is lacking though.

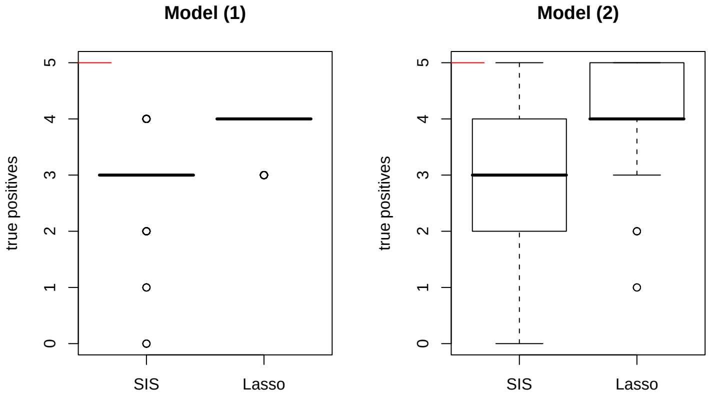

Fig. 1. Number of true positives $\sum_{j=1}^{p} I(\hat{\beta}_j \neq 0, \beta_j \neq 0)$ (with $|\widehat{\mathcal{M}}| = n - 1$) for 100 simulations from model (1) (left) and model (2) (right). The number of effective variables equals 5.

I agree with the authors that iterative SIS (ISIS) mitigates many of the problems occurring with the marginal approach of SIS. However, we need to choose a tuning parameter

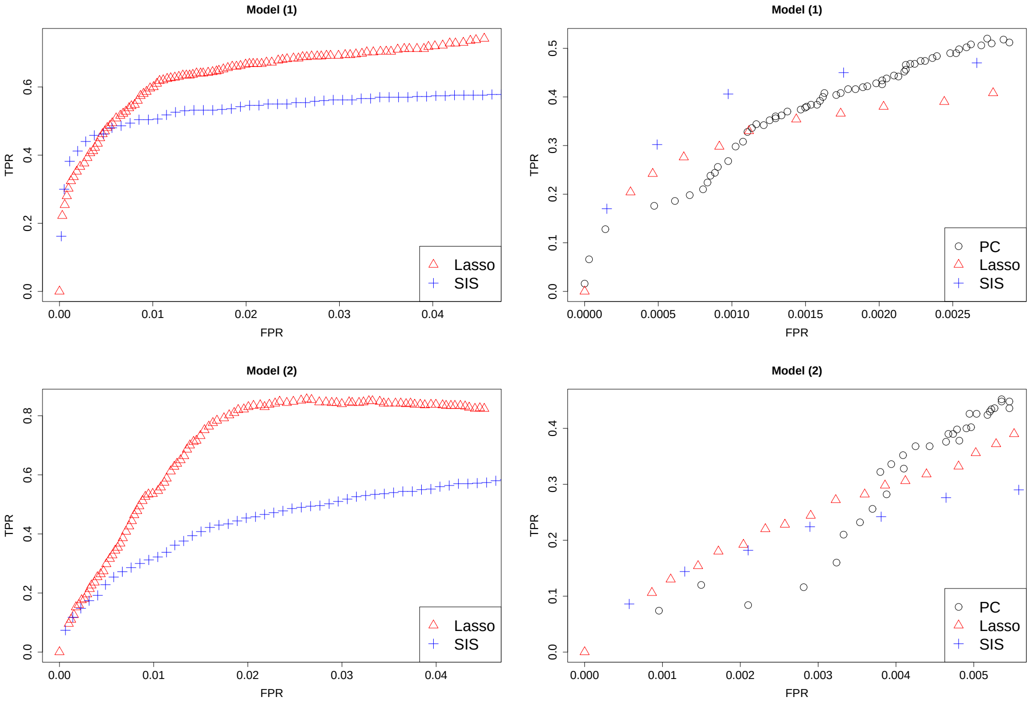

Fig. 2. ROC-curves for model (1) (top) and model (2) (bottom). Right: zoom-in for the domain with small false positive rate.

k (or denoted in the paper by $k_1, k_2, \ldots, k_\ell$) which is really unpleasant: ideally, for some rough sort of variable screening, there should be no other tuning parameter involved except the number of variables which are to selected from screening. When using $k = 1$ in ISIS, we end up with a procedure which is somewhere between orthogonal matching pursuit (which is almost identical to forward variable selection), and matching pursuit which is the same as L_2Boosting with componentwise linear least squares (Friedman, 2001; Bühlmann and Yu, 2003; Bühlmann, 2006). In particular in the high-dimensional setting with fairly low signal to noise ratio, the boosting approach is in our experience often better than orthogonal matching pursuit or forward variable selection. Why ISIS? Why not using the established boosting approach for variable screening (which is presumably not so different from Lasso, see (Efron et al., 2004))? And if there would be strong reasons for ISIS, how should we select the tuning parameter k for screening whose optimal choice may be in conflict with accurate prediction?

Finally, the authors stress the fact about ultra high-dimensionality. In their framework, the dimensionality $p = p_n$ is a function of sample size such that

$$\log(p_n) = O(n^\xi) \text{ for some } \xi > 0.$$

The usual approaches in asymptotic analysis (exponential inequalities, entropy arguments) would require that $\xi < 1$ which is equivalent to $\log(p_n)/n \to 0$ $(n \to \infty)$. Fan and Lv write in Section 5.1 (discussion of Condition 1) that "the concentration property in (16) makes restriction on ξ". What is the upper bound for $\xi > 0$, e.g. in the Gaussian case? Do we see here another range of high-dimensionality, or is ultra high-dimensionality the same as high-dimensionality where $\log(p_n)/n \rightarrow 0$?

It is my pleasure to second the vote of thanks: this paper will stimulate a lot of future research.

References

Bickel, P., Y. Ritov, and A. Tsybakov (2007). Simultaneous analysis of Lasso and Dantzig selector. Technical report, Laboratoire de Statistique, CREST.

Bühlmann, P. (2006). Boosting for high-dimensional linear models. *The Annals of Statistics 34*, 559–583.

Bühlmann, P. and M. Kalisch (2008). Variable selection for high-dimensional models: partial faithful distributions, strong associations and the PC-algorithm. Technical report.

Bühlmann, P. and B. Yu (2003). Boosting with the L_2 loss: Regression and classification. *Journal of the American Statistical Association 98*, 324–339.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression (with discussion). *The Annals of Statistics 32*, 407–451.

Friedman, J. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics 29*, 1189–1232.

Kalisch, M. and P. Bühlmann (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research 8*, 613–636.

Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics 34*, 1436–1462.

Meinshausen, N. and B. Yu (2007). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics* , to appear.

Spirtes, P., C. Glymour, and R. Scheines (2000). *Causation, Prediction, and Search* (2nd ed.). The MIT Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B 58*, 267–288.

van de Geer, S. (2008). High-dimensional generalized linear models and the Lasso. *The Annals of Statistics 36*, 614–645.

Zhang, C.-H. and J. Huang (2007). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *The Annals of Statistics* , to appear.