Rejoinder: Sure independence screening for ultrahigh dimensional feature space *

Jianqing Fan

Department of Operations Research and Financial Engineering

Princeton University

Jinchi Lv

Information and Operations Management Department

Marshall School of Business

University of Southern California

June 30, 2008

We are very grateful to all contributors for their stimulating comments and questions on the role of variable screening and selection on high-dimensional statistical modeling. This paper would not have been in the current form without the benefits of private communications with Professors Peter Bickel, Peter Bühlmann, Eitan Greenshtein, Qiwei Yao, Cun-Hui Zhang and Wenyang Zhang at various stages of this research. We shall not be able to resolve all points in a brief rejoinder — indeed, the discussion can be seen as a collective research agenda for the future and some of the agendas have already been undertaken by the discussants.

Independent learning

We would like to point out that the correlation learning is a specific case of independent learning that we advocate, which ranks the features according to the marginal utility of each feature. Correlation ranking is the same as feature ranking according to the reduction of RSS in the least-squares setting. In general, the marginal utility can be the quasi-likelihood or classification margin, contributed by each individual feature. This has been made more explicit in the proceeding paper by Fan (2007) and the manuscript by Fan, Samworth and Wu (2008). We do not claim that independent learning can solve all high-dimensional problems, but indicate its power for some class of problems with ultra-high dimensionality. The computational expediency and stability are prominently featured in the independent learning.

We are very pleased to see that the independent learning can indeed be derived from an empirical

*Prepared as a rejoinder to the discussion paper by the authors forthcoming in the Journal of the Royal Statistical Society Series B. Financial support from the NSF grant DMS-0704337 and the NIH grant R01-GM072611 is gratefully acknowledged. Lv's research was partially supported by National Science Foundation grant DMS-0806030 and the 2008 Zumberge Individual Award from the James H. Zumberge Faculty Research and Innovation Fund at the University of Southern California. Address for correspondence: Jianqing Fan, Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544. Phone: (609) 258-7924. E-mail: jqfan@princeton.edu.

likelihood viewpoint as elucidated by Hall, Titterington and Xue. An added feature of the empirical likelihood approach is that the classifier is automatically built based on the selected features. The idea of independent learning is also applicable to generalized additive model as discussed by Helen Zhang. The critical aspect is that the degrees of freedom for each component should be comparable or adjusted as elaborated in the generalized likelihood ratio tests by Fan, Zhang and Zhang (2001) and Fan and Jiang (2007). This also partially responds to the question raised by Keming Yu on nonlinear regression and categorical covariates. While our theory does not cover the case with categorical variables, our method does. The discussion by Runze Li suggested further that the correlation learning is applicable to the nonlinear single-index model, which is more general than single layer neural network model. This also answers partially the question raised by Keming Yu.

Relation to multiple testing and selection consistency

Several discussants (Bickel, Bühlmann, Marron, Luo, Baxter and Taylor, and Helen Zhang) link independent learning with multiple testing. Bickel raises several important theoretical questions from different perspectives and Bühlmann provides nice ROC curves, both for further understanding of SIS. As Bickel correctly points out, our procedure is similar to multiple testing problem to see if each feature is correlated with the response variable. Translating the test statistics into P-values puts them into the same scale, as Adragni and Cook, Richardson, and other discussants correctly point out. Helen Zhang mentions some existing work that answers the selection consistency question raised by Bickel. However, sure screening and multiple testing have different philosophy and evaluation criterion. Multiple testing aims at controlling the false discovery rate (FDR) while screening focuses on missed discoveries. In the simulated example II in §4.2.2, for example, failing to discover the variable X_4, which is uncorrelated with Y having marginal regression coefficient 0, is regarded as a serious mistake, while in the multiple testing problem, this would even be regarded as a correct decision. Hence, the evaluation criterion is also different from the multiple testing problem.

A number of contributors (Bühlmann, Qiwei Yao, Cun-Hui Zhang, Leng and Wang) address the issue of selection consistency. This corresponds to no false and missed discoveries in variable selection, if the evaluation criterion for model selection is used. While this is a very nice property, the selection consistency in high-dimensional space is a stringent requirement. The selection consistency is usually achieved by more complicated procedures than independent learning. For example, Bühlmann explores the idea of partial faithfulness; Qiwei Yao suggests stepwise regression procedures using modified information criteria; Leng and Wang discuss the penalized likelihood methods (Fan and Li, 2001); and so does Cun-Hui Zhang. However, in high-dimensional statistical endeavors, a procedure with low FDR and no missed variables is already remarkable, if the procedure is computationally expedient and stable. Such a procedure can indeed be constructed by using SIS below.

As discussed above, SIS is not designed to control FDR. But, it can easily be used to reduce FDR with no missed variables. The idea is very simple. Split the data randomly into two halves and apply SIS separately to both subsets of the data to obtain two submodels. Since we assume that the method has a sure screening property, both submodels contain all relevant variables in the model. Therefore, we take the common variables in these two submodels as the selected model. This selected model should have a

low FDR, as a falsely discovered variable has to appear independently twice in the selected model. The probability of such an event is merely of $(n/p)^2$ under the mild condition on exchangeability, thanks to the "blessing of dimensionality". See Fan, Samworth and Wu (2008) for details and extensions. In particular, they showed that the probability of choosing r extra variables is bounded by $(n^2/p)^r/r!$.

Tuning parameters

A number of contributors (Bühlmann, James and Radchenko, Leng and Wang, Runze Li, Wengyang Zhang, Zhou and Lin) discuss the need on a data-driven method for choosing the tuning parameters in both stages 1 and 2. We agree wholeheartedly. In the first stage, our preference is to select sufficiently many features, such as $d = n$ or $d = n/\log n$, though one can easily use a two-fold cross-validation to choose the number of features d. This answers partially the question raised by Zhang and Xia. In the second or final stage, Leng and Wang suggest using a BIC-type of criterion and provide related references. Bühlmann comments correctly that a tuning scheme is needed for the second stage of ISIS. Since the ISIS means to be a simple screening procedure, a simple selection scheme suffices in many situations. The predetermined parameters $k_1, \cdots, k_\ell$ in the second stage should be a decreasing sequence and the geometric sequence is more appropriate. Suppose, for example, we wish to run ISIS five iterations ($\ell = 5$) and to decrease the number of selected variables at each stage by a factor of θ (0.75, say). Then, the first iteration should choose $k_1 = n(1-\theta)/(1-\theta^5)$, the second iteration picks θk_1 variables, and the third stage selects $\theta^2 k_1$ variables, and so on. This avoids the ambiguity of variable selection of ISIS.

Clarifications

Some of our concepts are poorly presented and cause confusions as seen in some of discussions. First of all, the paper stressed the simplicity and utility of independent learning rules in high-dimensional feature screening. While correlation learning is an important specific example, we stress in fact the independent screening. This is why we choose the title with sure independent learning. Secondly, we would like to clarify that Condition 4 in the paper indicates a constraint on the population collinearity, whereas usual conditions on the design talk about the sample collinearity. The difference between these two types of collinearity could be severe when the dimensionality is much larger than the sample size, as illustrated by Figures 1 and 4 in the paper. Condition 4 accommodates the situation in which the features can be divided into several uncorrelated groups, each satisfying Condition 4. Thirdly, while $d = n - 1$ is our default in the screening stage, we do not rule out the possibility of selecting more features in the first stage. This partially answers the concerns by Morris and Richardson that $d = n/\log n$ or $d = n - 1$ can be too small in the first stage for some applications. In other words, we do not disagree with the comments made by Greentshein and Marron, who have in mind to construct as effective a method as possible to predict future observations (the first goal in Bickel's comment). However, for the second goal in Bickel's comment (To gain insight into the relationship between features and response for scientific purposes, as well as, hopefully, to construct an improved prediction method), selection of smaller number of features without too much compromise of prediction errors is also an important object and hence the default that $d = n - 1$ makes sense. Lastly, our goal is feature screening. The number of selected features is an order of magnitude larger than the number of active features. This makes sure screening much more easily and feasible.

Questions

Various contributors raise excellent questions that can be seen as research agenda. Due to limited space, we are not able to respond most of them.

Bickel's questions 2 and 3 touch the foundation of feature selection. We are pleased that he extends the concept of sure screening to the situation when more than one parsimonious models fit approximately equally well. In such a situation, the independence screening would be more likely to obtain all important factors than more sophisticated variable selection approaches, as the former is more likely to retain highly correlated variables that would fit approximately equally well in the final model. In addition, among those parsimonious models, the variables with a large marginal utility are usually preferred. Apparently much work is needed to compare screen-first-fit-after type methods with fit-first-screen-after type methods in terms of consistency and oracle properties, but the former would be faster and can deal with higher dimensionality.

Bühlmann questions the advantages of ISIS in comparison with the boosting algorithm. With the predetermined variable selection schemes at stage 2 shown above, ISIS chooses multiple features in the second stage using the joint information of important covariates. It is less greedy than the boosting algorithm. At the same time, it avoids solving large optimization problems as SCAD or LASSO would have done without pre-screening. In other words, it bridges the gap between these two extreme methods. It would be interesting to study and compare consistency properties of ISIS and the boosting approach, as commented by Bühlmann. The concentration property in (16) always holds for the Gaussian case, as shown in our theoretical study. Bühlmann, and Zhou and Lin both raise an excellent question about the upper bound of dimensionality for our theoretical results. Inspecting our technical proofs more carefully yields the need of such an upper bound, which is now stated in Condition 1, thanks to their their comments.

Leng and Wang, and Zhang and Xia raise the question of how the stochastic error in the screening stage impacts on the second stage of estimation. For many applications, this would not be severe. As commented in the paper, to avoid the selection bias in the screening stage, we can split the sample into two portions, where the first portion is used to screen variables and the second portion is used for shrinkage estimation.

Comments

Johnstone provides beautiful theoretical results on the distribution of the largest sample canonical correlation which gives us a better idea of how the problem of collinearity becomes severe when we have only a modest sample size compared with dimensionality. We agree with Samworth that the concentration property should hold for a broader class of spherically symmetric distributions, as he shows via careful simulations, and that it is important to derive the theoretical properties of independence screening for heavier tailed distributions that may have no concentration property.

We appreciate the remark by Hall, Titerington and Xue that the correlation coefficient for binary data becomes a t-statistic for any sample size provided that the class 1 is scored by n_1^{-1} and class 2 is scored with $-n_2^{-1}$. The idea is related to assigning an empirical prior to the class labels in a reverse order. This also provides some insights to the question raised by Yufeng Liu, who would like to know

the relative merits between the correlation ranking and t-statistic ranking. Yufeng Liu discusses several scenarios in classification problems in which SIS deserves further development. One possible method of feature ranking in multiclass problems is to rank them according to the F-statistics or its variants. An alternative is to regard it as a sequence of two-class problems and to use ISIS to select all relevant features.

We appreciate the connections between SIS and SPFC using the inverse regression made by Adragni and Cook. In their simulation, the response is uncorrelated with all predictors, and this explains why SIS performs poorly. It is unclear to us how the piecewise linear basis $\mathbf{f}_y$ was constructed, but SPFC in this simulated example is the same as the correlation learning based on a nonlinear transform $\mathbf{f}_y$ of response, which is now correlated with the relevant predictor. This gives advantages over plain SIS, which does not use any transform, for this simulated model.

Greentshtein remarks that the nonparametric empirical Bayes method performs very well for classifications and works well also for sparse situations. As the method does not explicitly explore the sparsity, we would not expect it adapts to the sparse setting as well as the methods that are tailored for this setting. The nonparametric empirical Bayes method is useful for constructing an effective method for prediction class labels, without selecting features. This would not be suitable for achieving aim ii) of Bickel's discussion.

Levina and Zhu look at the performance of ISIS with the Lasso plug-in under lower SNR levels through a simulation study. We agree that incorporating SNR to the convergence rate will give us a better picture of its impact. Of importance is to develop extensions of ISIS that are robust to low SNR. Some related questions have also been addressed by Luo, Baxter and Taylor.

Several discussants, including Anagnostopoulos and Tasoulis, bring up questions of relaxing the technical assumptions in the paper to give more insights into the applicability of SIS. We believe that these questions will certainly stimulate much new research on variable screening and selection. For the Leukaemia data analysis, we can not compare the overlap of the selected genes since we do not have the keys to check that.

Robustness

A number of contributors, including Gather and Guddat, Hui Zhou, and Luo, Baxter and Taylor, bring the issue of robustness to outliers and model assumptions. We appreciate their efforts to make the procedure more robust to those assumptions. In particular, Gather and Guddat, and Hui Zhou both propose more robust procedures to the outliers. We agree with all discussants that robustness to outliers and to model assumptions are important issues and they have addressed some of those. Independent learning is still in the infancy and certainly needs more researchers to nurture and understand it.

Critisms

Many discussants give very critical scrutiny of independent learning in high-dimensional modeling. SIS and ISIS are simple procedures and can not expect to address all the needs.

Robert casts doubts on the assumption of the existence of a single true model when $p \gg n$, as Bickel does. We acknowledge that there are many models that are statistically indifferentiable given the limited amount of information, but some are more useful. SIS and ISIS are procedures to pick some submodels

that have large marginal contributions. The asymptotic results provide merely an ideal situation under which our common sense of independent screening works. On the other hand, Bayesian methods are viable tools for selecting a family of submodels that have similar performance.

Longford comes up with a pyramid view of the importance of the variables which leads to a weaker assumption than sparsity in the narrow sense. In a sense, the classical best subset selection provides such a pyramid view on the most important k-variable models. However, such a best set selection is an NP-hard problem and classical stepwise addition or stepwise algorithms provide a useful proxy. In high-dimensional endeavors, however, the noise accumulation and computational cost make these methods more challenging to use and to understand. The penalized least-squares methods provide an alternative solution to these traditional methods with more efficient computation and easier structure to understand its statistical properties. The SCAD+ and MCP+ (Zhang, 2007) or LASSO solution paths provide, in a sense, to such a pyramid view. SIS and ISIS purely assist in reducing the dimensionality so that a more efficient solution paths can be constructed.

Richardson speculates that the sure screening can be elusive in some correlated cases. The poor performance of SIS in her simulation is related to the selection of the tuning parameter and leakage effect from the peaks. If a larger d such as $d = n - 1$ is used in the first stage of screening and if the leakage issue is addressed, then the results of SIS can be significantly improved, by looking at her figures. In other words, sure screening property still holds in her simulated example. In searching for quantitative trait loci or other similar biological endeavors, the leakage issue should be addressed. The peak locations are often of interest and large values around the peak are regarded as the leakage from the peaks and correspond to the same genetic locus.

Computation

Many contributors touch the issue of computation, including Bühlmann, Runze Li, and Cun-Hui Zhang, who address different computation algorithms and their computational complexity. We agree with those discussants that SIS has the smallest computational cost at the screening stage. The PC-algorithm for exploring partial faithfulness is certainly very stimulating and useful. The PLUS algorithm proposed by Zhang (2007) is creative for effectively finding the solution paths to the folded-concave penalized least-squares problems and is backed by asymptotic theory. Alternative algorithms are iteratively reweighted penalized L_1 regression proposed by Zou and Li (2008) and elaborated further in the paper, the iterative coordinatewise minimizaton discussed by Runze Li, and local quadratic approximation (Fan and Li, 2001). With these, we agree with Cun-Hui Zhang and Runze Li that the implementation of folded-concave penalized least-squares problems are not much harder or slower to compute than LASSO. However, the gains in bias reduction can be substantial in the high-dimensional setting. We believe that with better understanding and implementation, the folded-concave penalized likelihood (Fan and Li, 2001) will play even more important roles in high-dimensional statistical modeling and feature selection.

Conclusion

Taken together, the discussants cover a wide range of topics, from foundation, philosophy and theory to methods, computation and applications. The wide interest in high-dimensional learning and related methods in many fields, from bioinformatics and genetics to climatology and finance, clearly presents

exciting opportunities for interdisciplinary collaborations and expanded exchanges in ideas and tools between statistics and other disciplines. We are very pleased to conclude by reiterating our thanks to all contributors, and to the Royal Statistical Society and the journal for hosting this forum.

References

[1] Fan (2007). Variable screening in high-dimensional feature space. *The proceedings for International Congress of Chinese Mathematicians*.

[2] Fan, J., Samworth, R. and Wu, Y. (2008). Ultra-dimensional variable selection: Beyond the linear model. *Manuscript*.

[3] Fan, J. and Jiang, J. (2007). Nonparametric inference with generalized likelihood ratio tests (with discussion), *Test*, **16**, 409–478.

[4] Fan, J., Zhang, C. and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *The Annals of Statistics*, **29**, 153–193.