Journal of Statistical Planning and Inference 145 (2014) 148–164

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Goodness-of-fit testing-based selection for large-p-small-n problems: A two-stage ranking approach ☆

Xiaobo Ding [a], Lexin Li [b], Lixing Zhu [c,d,*]

[a] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
[b] Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA
[c] Department of Mathematics, Hong Kong Baptist University, Hong Kong
[d] School of Statistics and Mathematics, Yunnan University of Finance and Economics, Wuhua, Yunnan, China

ARTICLE INFO

Article history:
Received 20 October 2012
Received in revised form
16 August 2013
Accepted 17 August 2013
Available online 31 August 2013

Keywords:
Backward screening
Linear model
Marginal effect
Sequential goodness-of-fit testing

ABSTRACT

In this paper, we investigate two-stage ranking–selection procedures for ultra-high dimensional data in the framework of goodness-of-fit testing. We develop a k-step marginal F-test (MFT_k) screening in the first stage. The MFT_1 is, as a statistic, equivalent to that used in the sure independence screening (SIS) proposed by Fan and Lv (2008). The MFT_k with $k \geq 2$ makes improvement over the MFT_1 mainly on better handling correlations among predictors. For selecting a more parsimonious working model in the first stage, we propose a soft threshold cutoff through a sequential goodness-of-fit testing. This avoids some drawbacks of the hard threshold cutoff in Fan and Lv (2008) and the extended BIC used in Wang (2009). In the second stage, we develop one-step backward screening to further remove those insignificant predictors from the model. Further, likewise as the iterative SIS, we provide the iterative versions of the proposed procedures to have more accurate variable selection. Extensive numerical studies and real data analysis are carried out to examine the performance of our proposed procedures.

1. Introduction

With the development of modern scientific research, people often collect data sets with tens of thousands of variables, while the sample sizes are relatively small. These are the so-called large-p-small-n paradigms, where great difficulties exist in statistical estimation, inference and computation. However, it is often the case that only a small number of predictors are significant for response in regression modeling, and it is expected to pick them out to form a more effective working model when the full model is of sparse structure. It is known that the classical variable selection methods such as AIC and BIC are more suitable for the cases where the number of predictors is not very large. As such, new techniques are highly demanded for the large-p-small-n paradigms.

The research in this area is already very intensive. But most of the classical variable selection methods can be considered as one-stage ranking–selection procedures: they rank predictors so as to provide a moderate number of candidate models, and then select the true model from them. Here a model refers to a subset of predictors, and the true model refers to the set of the significant predictors. There are two typical ranking methodologies in this area. One is to regard selection as a testing problem. For example, Zheng and Loh (1995), Bunea et al. (2006), Benjamini and Gavrilov (2009) and others used test statistics to rank predictors. The other is to use penalization. The solution path of the penalized regression methods can be considered as a ranking of predictors. The typical methods include LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001),

☆ The research was supported by a grant from University Research Council of Hong Kong.
* Corresponding author. Tel.: +86 852 34117016.
E-mail address: lzhu@hkbu.edu.hk (L. Zhu).

0378-3758/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2013.08.012

Dantzig selector (Candés and Tao, 2007) and so forth. When predictors are ranked, then information criteria, false discovery rate controlling or other methods can be used to select the significant predictors.

It is naturally expected that the significant predictors can be ranked prior to the insignificant ones such that the true model can then be selected consistently in certain sense. In the large-p-small-n paradigms, this is difficult to achieve especially when the predictors are correlated. A direct way is to have a larger model first, which has moderate size and contains the significant predictors, and re-rank the predictors in the model by some sophisticated approaches reported in the literature. We call such a procedure a two-stage ranking–selection procedure.

Sure independence screening (SIS) proposed by Fan and Lv (2008) is a promising two-stage ranking–selection method in the literature. This method has been extended to handle linear regression models (Fan and Lv, 2008), generalized linear models (Fan et al., 2010), parametric models (Fan et al., 2009), nonparametric additive models (Fan et al., 2011) and Cox's proportional hazard models (Fan et al., 2010; Zhao and Li, 2012). However, the SIS may also rank some significant predictors posterior to the insignificant ones, and thus these significant predictors cannot be picked out in practice, Fan and Lv (2008) then proposed the iterative SIS (ISIS) to be a remedy. An almost equivalent method is the forward regression proposed by Wang (2009).

We note that the SIS is in effect based on a F-test statistic for every predictor marginally although they did not specifically use a testing way to do so. This motivates us to further develop a goodness-of-fit testing-based two-stage ranking–selection procedure. We also desire to develop a novel methodology to better handle correlation among predictors. To well illustrate the proposed procedures, we consider the Gaussian linear model, keeping in mind that the procedures can be extended for more general models. Specifically, we develop a k-step marginal F-test (MFT_k) screening procedure with k being a pre-specified positive integer to initially rank predictors, use a sequential goodness-of-fit testing to determine the number of predictors we should select, and then propose a backward screening (BS) to further screen out the insignificant predictors.

The MFT_k is to rank predictors by F-test statistic for each set of k predictors. We will show that when $k=1$, the MFT_1 is, in the sense of ranking predictors, equivalent to the SIS. When $k \geq 2$, the MFT_k clearly takes more about correlations among predictors into account than the SIS does, making the resulting ranking more informative. In the real data example later, we will see that the MFT_1 almost fails whereas the MFT_2 and MFT_3 work well. On the other hand, although the MFT_1 is proved to be equivalent to the SIS in certain sense, goodness-of-fit testing provides us a simple but useful way of extending MFT_1 to MFT_k, whereas marginal correlation coefficients of a set of predictors are difficult to be used simultaneously in predictors ranking when the SIS is applied.

Another issue is about how many predictors will be selected in the first stage. Fan and Lv (2008) and the relevant methods set hard threshold cutoff values. However, we will see that in practice, the hard threshold cutoff either loses some of the significant predictors or involves many insignificant ones. Zhao and Li (2012) proposed a method that can control the false positive rate based on the partial orthogonality assumption, which is too restrictive. Wang (2009) used the extended BIC criterion proposed by Chen and Chen (2008). But it is known that the performance of the information criteria depends on how fast the number of predictors increases with the sample size, which cannot be defined clearly in finite sample cases. In addition, the number of predictors that we should select from all of predictors is generally larger or even much larger than the number of those significant predictors. In such cases, the above methods perform poorly. In this paper we suggest a soft threshold cutoff by sequential goodness-of-fit testing, which requires no further restrictive assumptions. The only tuning parameter that we should set is the nominal level for each goodness-of-fit test, which can be easily set and explained in the finite sample.

In the second stage, Fan and Lv (2008) and relevant methods applied the classical penalized regression methods such as the SCAD to further screen out insignificant predictors as many as possible. Of course, these penalized regression methods should be integrated with a selection method such as AIC, BIC or a cross-validation method so that they can select the true model. However, our proposed BS can rank and select predictors simultaneously, which makes it easier to be implemented with less computational workload than the penalized regression methods. In addition, it can be seen from the simulation studies that the BS has better performance.

Although the MFT_k can take more correlations into account, it may also lose some significant predictors in the initial ranking procedure. We then also propose iterative MFT_k. Interestingly, the iterative MFT_k can be seen as an extension of the forward regression, because the forward regression is equivalent to the iterative MFT_1. Nevertheless, it is very obvious that for variable selection, no method works when no constraint on the correlation among predictors is assumed. Our method is to provide a way to better deal with this important issue, rather than completely solving it. The research along this line is still worth of further study.

The rest of this paper is organized as follows. In Section 2 we propose the two-stage ranking–selection procedure. Theoretical results are given in Section 3. In Section 4 numerical studies are conducted to examine the performance of the proposed procedure. Section 5 provides a real data example. Section 6 offers some discussions. Assumptions and technical proofs are supplied in the supplementary file.

2. Goodness-of-fit testing-based selection: two-stage ranking procedure

2.1. Model and goodness-of-fit testing

Consider the Gaussian linear model

$$Y = \sum_{i=1}^{p} \beta_i X_i + \epsilon, \tag{1}$$