Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Screening active factors in supersaturated designs

Ujjwal Das[a], Sudhir Gupta[a,*], Shuva Gupta[b]

[a] Division of Statistics, Northern Illinois University, DeKalb, IL 60115, USA

[b] Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA

ARTICLE INFO

Article history:
Received 17 June 2013
Received in revised form 28 February 2014
Accepted 28 February 2014
Available online xxxx

Keywords:
Corrected AIC
Dantzig selector
Effect heredity
Nonconvex penalty
Shrinkage estimation
SCAD
Smoothly clipped absolute deviation
Sparsity tuning parameter

ABSTRACT

Identification of active factors in supersaturated designs (SSDs) has been the subject of much recent study. Although several methods have been previously proposed, a solution to the problem beyond one or two active factors still seems to be unsatisfactory. The smoothly clipped absolute deviation (SCAD) penalty function for variable selection has nice theoretical properties, but due to its nonconvex nature, it poses computational issues in model fitting. As a result, so far it has not shown much promise for SSDs. Another issue regarding its inefficiency, particularly for SSDs, has been the method used for choosing the SCAD sparsity tuning parameter. The selection of the SCAD sparsity tuning parameter using the AIC and BIC information criteria, generalized cross-validation, and a recently proposed method based on the norm of the error in the solution of systems of linear equations are investigated. This is performed in conjunction with a recently developed more efficient algorithm for implementing the SCAD penalty. The small sample bias-corrected cAIC is found to yield a model size closer to the true model size. Results of the numerical study and real data analyses reveal that the SCAD is a valuable tool for identifying active factors in SSDs.

1. Introduction

In the preliminary phase of experiments in industry, simulation experiments or exploratory studies, it is often necessary to investigate a large number of factors. The main purpose of such experiments is to identify a few important factors with non-negligible effects (effect sparsity) from a large number of potentially relevant factors. Due to constraints of cost and time, the number of runs or experiments has to be kept much smaller than the number of potential factors. Designs requiring fewer runs than the number of effects to be estimated are called *supersaturated designs*. If p is the number of factors and n is the number of runs, then for a 2-level supersaturated design, $n < p + 1$. Effect sparsity and absence of interactions in general form the basis of these experiments. Sometimes few interactions may also have to be entertained as potential effects. In that situation at least one of the main effects from the interactions should be present in the model. This characteristic of the model is known as "effect heredity" and will be discussed in Section 4.

Although the history of SSDs dates back to Satterthwaite (1959) and to Booth and Cox (1962), many developments in this area have taken place over the last two decades. Renewed interest in SSDs was initiated by Lin (1993) and Wu (1993). Further constructions of SSDs are due to Nguyen (1996); Cheng (1997); Butler et al. (2001); Liu and Dean (2004), among several others. SSDs for factors with more than two levels have also been studied in the literature. For further details we refer to Gilmour (2006) and Georgiou (2014).

[*] Corresponding author. Tel.: +1 815 501 3988.
E-mail addresses: dsujjwal@gmail.com (U. Das), sudhir@math.niu.edu, sudhirgupta.stat@gmail.com (S. Gupta), sgupta22@ncsu.edu (S. Gupta).

http://dx.doi.org/10.1016/j.csda.2014.02.023
0167-9473/© 2014 Elsevier B.V. All rights reserved.

The analysis of SSDs is complicated due to the inherent non-full rank nature of the design matrix, and the fact that the columns of the model matrix are correlated. As a result, the effects of different factors are aliased with one another making it very difficult to identify the non-negligible or active factors correctly. Wu (1993); Westfall et al. (1998) and Liu et al. (2007) studied classical stepwise and all subset regression methods for the analysis of SSDs. Chipman et al. (1997) developed a Bayesian approach based on a stochastic search variable selection (SSVS) method. Beattie et al. (2002) proposed a two-stage Bayesian model selection strategy, combining SSVS and the intrinsic Bayes factor method. Li and Lin (2002) applied the shrinkage estimation and selection method of Fan and Li (2001) with the SCAD penalty. Zhang et al. (2007) introduced a new method, called PLSVS, based on partial least squares. Koukouvinos and Mylona (2008) proposed a method for SSDs having a specific block orthogonal structure, while Koukouvinos and Mylona (2009) applied a group screening method for the analysis of a class of optimal multi-level SSDs. Phoa et al. (2009) applied the Dantzig selector (DS) method of Candes and Tao (2007) for analyzing SSDs and, through simulations, demonstrated it to be more efficient than the method discussed by Chipman et al. (1997), the SCAD method of Li and Lin (2002), and PLSVS method of Zhang et al. (2007). Marley and Woods (2010) proposed a model averaging approach and compared it to the forward selection and DS methods. Based on an extensive simulation study, they found the performance of DS to be better than the other two methods overall. They also concluded that in SSDs, (a) the number of factors should be less than twice the number of runs, and (b) the number of runs should be at least three times the number of active effects. Edwards and Mee (2011) proposed a global model randomization test with all subset regression. Their findings also support Marley and Woods (2010) conclusions in (a) and (b) above. Koukouvinos et al. (2011) studied the use of information theoretic entropy measures for analysis of SSDs under the set-up of generalized linear models.

Kim et al. (2008) proved that, under effect sparsity, the SCAD estimator possesses the oracle property even in high-dimensional settings where the number of covariates p is much larger than the sample size n. The oracle property means that asymptotically the SCAD estimator is obtained as if the non-active effects were known in advance. More recently, Breheny and Huang (2011) proposed a coordinate descent algorithm for the SCAD optimization problem and showed its superiority over the other algorithms. In this paper, we study the efficacy of SCAD for identifying active effects in SSDs utilizing the algorithm of Breheny and Huang (2011). Unlike the two-stage SCAD procedure of Li and Lin (2002), the method discussed can be completed in the first stage itself and does not require good initial estimates of parameters. Cross-validation (CV) and generalized cross-validation (GCV) of Craven and Wahba (1979) are widely used for controlling the extent of SCAD penalty. In CV, some observations are left out for the purpose of validating the fitted model, while the remaining part of the sample (training sample) is used for model fitting. In small sample setting of SSDs, the loss of even one observation can lead to substantial adverse effect on the quality of the fitted model. Furthermore, loss of an observation, in other words, a treatment combination, jeopardizes the optimality properties of a SSD and hence the quality of the fitted model. Thus, CV has limited utility for analyzing SSDs and is not discussed further in this paper in view of its poor performance in our numerical study.

We propose the use of the Akaike (AIC), corrected Akaike (cAIC), and Bayesian (BIC) information criteria in conjunction with the algorithm of Breheny and Huang (2011) to regulate the sparsity in SCAD penalty. Wang et al. (2007) showed that GCV leads to nonignorable overfitting of the model even if the sample size goes to infinity. They also showed that AIC provides a good approximation to log (GCV). In our numerical study, the performance of GCV in terms of identifying the correct model was similar to AIC. However, GCV was found to yield model size somewhat larger than that obtained using AIC. In several cases the GCV mean model size was larger by 0.5–0.6. Since our results are in line with Wang et al. (2007), to avoid redundancy GCV results are not included in the tables in Sections 3 and 4. The recently proposed method of Androulakis et al. (2011), here referred to as AKM, is also studied. The results and the findings of Phoa et al. (2009) and Marley and Woods (2010) lead us to include DS in our numerical study. Like Marley and Woods (2010), our numerical results also show that the performance of the methods tends to deteriorate as the magnitude of effect sizes is made smaller. Therefore, for a more informative comparison, we study the performance of the methods especially for relatively smaller effect sizes.

Section 2 provides a brief overview of the estimation methods considered in this paper. The methods are compared through simulation in Section 3. Yuan et al. (2007) used the LARS algorithm of Efron et al. (2004) and proposed an extension to address effect heredity in model selection. They explicitly mention that despite its nice theoretical properties, the SCAD approach does not incorporate the effect heredity principle. In Section 4, the efficacy of SCAD for effect heredity models is studied numerically. Note that the extensive numerical study of effect heredity models has not received much attention in the literature. The proposed methods are illustrated on some real world data in Section 5. These real examples also serve to further illustrate the efficacy of SCAD for effect heredity models. Finally, some concluding remarks are made in Section 6.

2. Penalizing methods and information criteria

In this section we briefly describe the SCAD method under AIC, cAIC, BIC and AKM criteria, and the DS method. Let $\boldsymbol{X}_0$ be the $n \times k$ design matrix of $+1$'s and -1's for two-level factors and $\mathbf{1}_n$ be the $n \times 1$ vector of 1's. Consider the linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{y}$ is the $n \times 1$ response vector, $\boldsymbol{X} = [\mathbf{1}_n,\ \boldsymbol{X}_0]$ is the $n \times (k+1)$ model matrix, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ is the usual vector of regression parameters, and $\boldsymbol{\epsilon}$ is the vector of independently distributed normal random errors with mean zero and constant variance σ^2. Note that k is greater than p when some interactions may also be active, where p is the number of factors. Under model (1), the DS of the regression coefficients $\boldsymbol{\beta}$ is obtained from

$$\min_{\boldsymbol{\beta} \in R^k} \|\boldsymbol{\beta}\|_{l_1}$$