Computational Statistics and Data Analysis 54 (2010) 2230–2243

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Variable selection via combined penalization for high-dimensional data analysis

Xiaoming Wang[a], Taesung Park[b], K.C. Carriere[c,*]

[a] School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, 200433, China

[b] Department of Statistics, Seoul National University, Seoul, Republic of Korea

[c] Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, AB, T6G 2G1, Canada

ARTICLE INFO

Article history:
Received 10 February 2009
Received in revised form 4 February 2010
Accepted 29 March 2010
Available online 18 April 2010

Keywords:
Linear model
Combined penalization
SCAD
Ridge penalty
GCV
Variable selection
Microarray classification

ABSTRACT

We propose a new penalized least squares approach to handling high-dimensional statistical analysis problems. Our proposed procedure can outperform the SCAD penalty technique (Fan and Li, 2001) when the number of predictors p is much larger than the number of observations n, and/or when the correlation among predictors is high. The proposed procedure has some of the properties of the smoothly clipped absolute deviation (SCAD) penalty method, including *sparsity* and *continuity*, and is asymptotically equivalent to an oracle estimator. We show how the approach can be used to analyze high-dimensional data, e.g., microarray data, to construct a classification rule and at the same time automatically select significant genes. A simulation study and real data examples demonstrate the practical aspects of the new method.

1. Introduction

Variable selection is fundamental to statistical modeling, especially in high-dimensional statistical problems such as microarray data analysis. In practice, investigators usually begin by introducing a large number of possible predictors to avoid modeling biases. In this situation, however, ordinary least squares (OLS) are often a poor choice for both prediction and interpretation.

To enhance predictability and make the model more parsimonious, analysts typically use stepwise selection procedures. Although these methods may be used effectively to identify good models, they have several drawbacks: (a) theoretical properties can be hard to understand (Fan and Li, 2001); (b) estimates are extremely variable because of their inherent discreteness, as addressed by Breiman (1996); and (c) the computational costs are high, as p gets large.

We advocate a penalization technique on the general form of *loss* + *penalty*. Ridge regression (Hoerl and Kennard, 1970) minimizes the residual sum of squares, imposing a quadratic penalty on the coefficients. Ridge regression, a continuous shrinkage method, achieves better prediction performance through a bias–variance trade-off. However, ridge regression does not result in a parsimonious model; it always keeps all the predictors in the model.

A key aspect of a penalization technique is the penalty function. Many studies have focused on this problem: for example, the L_2 penalty that results in a ridge regression (Hoerl and Kennard, 1970), the L_q penalty that leads to a bridge regression (Frank and Friedman, 1993; Fu, 1998), the L_1 penalty that yields a soft threshold rule (Donoho and Johnstone, 1994), the L_1 penalty that leads to the least absolute shrinkage and selection operator (LASSO) introduced by Tibshirani (1996, 1997)

[*] Corresponding author. Tel.: +1 780 492 4230; fax: +1 780 492 6826.
E-mail addresses: xmwang492@hotmail.com (X. Wang), tspark@snu.ac.kr (T. Park), kccarrie@ualberta.ca (K.C. Carriere).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.03.026

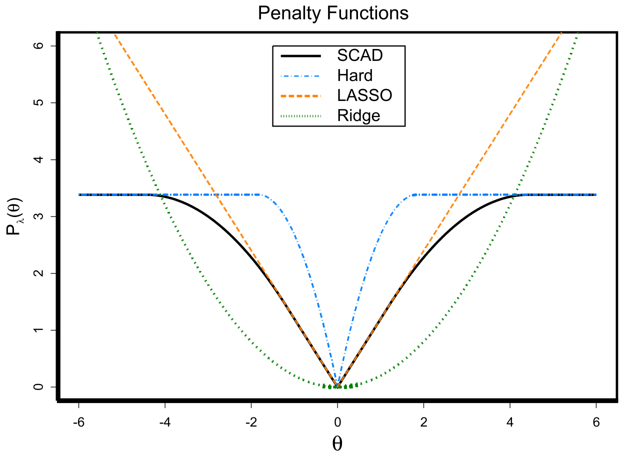

Fig. 1. The four penalty functions of the ridge, hard threshold, LASSO and SCAD techniques.

and further studied by Efron et al. (2004), and the smoothly clipped absolute deviation penalty (SCAD) proposed by Fan and Li (2001). Fig. 1 depicts some of the penalty functions including ridge regression, the hard thresholding rule (Donoho and Johnstone, 1994), LASSO and SCAD.

Fan and Li (2001) point out that a good penalty function results in an estimator with the following three properties.

1. *Unbiasedness*: The resulting estimator is unbiased when the unknown coefficient of the regression is large.
2. *Sparsity*: The resulting estimator has an automatic thresholding rule that sets some estimated coefficients to zero to reduce the complexity of the model if the estimated values are small.
3. *Continuity*: The resulting estimator is a continuous function of data to avoid instability in model prediction.

Fan and Li (2001) discuss various penalty functions in terms of the above three properties and establish the conditions for a penalty function that meets the requirements. Their work led to the development of a new penalty function, SCAD, a non-concave penalty function widely used in penalization techniques in various statistical contexts (see Antoniadis and Fan, 2001; Fan and Li, 2004, 2006, among others). Many authors believe that the penalization techniques related to SCAD are less biased than LASSO (Tibshirani, 1996).

Although the SCAD penalty and its related non-concave penalization techniques have produced some remarkable results, there are still concerns about applying SCAD to high-dimensional problems. For example, a typical microarray data set often has many thousands of predictors (up to 500,000 genes) and small samples. The number of predictors in this case is much larger than the number of observations (i.e., $p \gg n$), and correlations among the predictors could be very high. This concern has led us to conduct the current study.

We propose a new regularization technique that modifies the SCAD penalty by adding a quadratic penalty item. We call this technique "combined penalization (CP)". Our goal is to find a new variable selection approach that can work as well as the SCAD technique when this technique is the best choice, and that can resolve the problem highlighted above. We anticipate that the proposed variable selection technique will work well when $p \gg n$, for example with microarray data, and provides better prediction performance when the predictors are highly correlated.

This paper is organized as follows. In Section 2, we outline the basic idea underlying our variable selection method and discuss its related sampling properties. In Section 3, we provide details of the implementation of our method, including the iterative algorithm for solving least squares with combined penalization, a special consideration of how to overcome the computational burden. We also discuss the problems related to choosing the regularization parameters in the model. In Section 4, we discuss extending the combined penalization technique to general models, including logistic regression. In Section 5, computer simulations and real data examples demonstrate the utility of our method. We present our conclusions in Section 6. Theoretic proofs are gathered in the Appendix.

2. Combined penalization

2.1. Least squares with combined penalization

The penalized least squares method and variable selection in linear regression are intimately connected. Consider a general linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\mathbf{y}$ is an $n \times 1$ response vector and $\mathbf{X} = [X_1|\cdots|X_p]$ is an $n \times p$ design matrix. Without loss of generality, we assume that the response is centered and that the predictors are normalized.