Computational Statistics and Data Analysis 92 (2015) 40–52

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Studies of the adaptive network-constrained linear regression and its application

Hu Yang [a,*], Danhui Yi [b,c,d]

[a] Department of e-Business, School of Information, Central University of Finance and Economics, Beijing, 100081, PR China
[b] School of Statistics, Renmin University of China, Beijing, 100872, PR China
[c] Center for Applied Statistics, Renmin University of China, Beijing, 100872, PR China
[d] Consulting Center of Statistics, Renmin University of China, Beijing, 100872, PR China

ARTICLE INFO

Article history:
Received 31 January 2014
Received in revised form 23 January 2015
Accepted 15 June 2015
Available online 2 July 2015

Keywords:
Adaptive network penalty
Variable selections
P/E ratio

ABSTRACT

The network-constrained criterion is one of the fundamental variable selection models for high-dimensional data with correlated features. It is distinguished from others in that it can select features and simultaneously encourage global smoothness of the coefficients over the network via penalizing the weighted sum of squares of the scaled difference of the coefficients between neighbor vertices. However, because more features were selected while it was applied for the process of analysis of the "*China Stock Market Financial Database—Financial Ratios*", the so-called adaptive network-constrained criterion was proposed to tackle the problem via assigning various weights to the lasso penalty. Similar to the adaptive lasso, the proposed model enjoys consistency in variable selection if the weights have been given correctly in advance. The simulations show that the proposed model performed better than the other variable selection techniques mentioned in the paper with regards to model fitting; meanwhile, it selected fewer features than the network-constrained criterion. Furthermore, the mean value of the cross-validation likelihood and the number of selected features are tested to be accurate enough for practical applications.

1. Introduction

High-dimensional data analysis has become increasingly popular and important in diverse fields of sciences, engineering and humanities; ranging from genomics and health sciences to economics, finance and machine learning. It characterizes many contemporary problems in statistics (Hastie et al., 2005). Under the sparsity assumption, which defines estimates as sparse if many of their components are zero or approximate to zero, the shrinkage methods are the useful and efficient techniques to analyze high-dimensional data by shrinking some coefficients to zero (also called variable selection). Among these methods, the lasso (l_1 penalty) is one of the most popular and basic ones proposed by Tibshirani (1996).

Following the idea of the lasso, approaches have been developed to tackle high-dimensional problems and improve them, with greater accuracy, easier explanation and cheaper computation (Fan and Lv, 2010). Some modified the penalty function to advocate estimators with three properties – sparsity, unbiasedness and continuity – such as SCAD penalty (Fan and Li, 2001) and the adaptive lasso (Zou, 2006); or to improve the efficiency of the computation, such as the least angle regression (Efron et al., 2004); or to encourage the grouping effect by combining l_1 and l_2, such as the elastic net penalty (Zou and Hastie, 2005); or to utilize the characteristics of the ordinary coefficients, such as the fused lasso (Tibshirani et al., 2005). Meanwhile,

* Corresponding author. Tel.: +86 0 10 62288896.
E-mail address: hu.yang@cufe.edu.cn (H. Yang).

http://dx.doi.org/10.1016/j.csda.2015.06.008
0167-9473/

others modified the loss function to satisfy the requirement in different application scenarios, so as to model binomial data, counting data and even time series data. Some examples are Cox Regression (Tibshirani, 1997), Logistic regression (Hastie et al., 2009), Poisson regression (Guisan, Edwards Jr et al. 2002) and even the support vector machines (Zhu et al., 2004).

Recently, some researchers began to focus on studying the approaches to select groups of relevant features. The feasible ways are based on modifying the penalties so that the models hold the so-called 'group property' (or group effect). The elastic net penalty (Zou and Hastie, 2005) is one such approach, which penalizes the sum of squares of the coefficients so that more correlated coefficients were selected. The network penalty (Li and Li, 2008) is another example, which can be viewed as an extension of the former. It performs better than the elastic net penalty because it penalizes the weighted sum of squares of the scaled difference of the coefficients between neighbor vertices, and so obtains the global smoothness coefficients. However, these two criterions did not work as well on the variable selections and the prediction problems considered here. Both of them tend to select more related variables than lasso, meanwhile lead to interpret difficultly. Other approaches, such as the group lasso were also introduced to for variable selection on groups of variables in linear regression (Meier et al., 2008; Ma et al., 2007).

In the article, we aim to analyze the "*China Stock Market Financial Database—Financial Ratios*" and select the important financial ratios related to the *Price Earning* (P/E) ratio. Considering the correlated features which exist in the database and can be categorized into several groups empirically, we applied the network-constrained criterion to the data and wanted to select the correlated features simultaneously. However, as the network-constrained criterion suffers from estimation inefficiency and selection inconsistency, more non-significant features were selected; we referred the idea of the adaptive lasso (Zou, 2006) and propose the new approach to improve the adaptive network-constrained criterion so as to select fewer features. The proposed method has the properties to select less correlated features and obtains the global smoothness coefficients.

Before the beginning of our study, some useful notations should be introduced. Consider the samples $\mathbf{S} = (\mathbf{Y}, \mathbf{X})$ with n observations where $\mathbf{Y}$ is the response $n \times 1$ vector $(Y_1, Y_2, \ldots, Y_n)^T$ and $\mathbf{X}$ is a $n \times p$ design matrix with p predictors $(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_p)$. All of the predictors $(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_p)$ and $\mathbf{Y}$ are centered and standardized respectively, which satisfy $\bar{\mathbf{Y}} = \bar{\boldsymbol{X}}_j = 0$ and $\boldsymbol{X}_j^T \boldsymbol{X}_j = 1$, $j = 1, 2, \ldots, p$. The Lasso is the regularization technique for simultaneous estimation and variable selection (Tibshirani, 1996), defined as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left(\frac{1}{n} \left\| \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \right\|^2 + \lambda \left\| \boldsymbol{\beta} \right\|_1 \right), \tag{1.1}$$

where the $1 \times p$ coefficients vector and its estimates are $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ and $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)^T$ respectively, and λ is nonnegative regularization parameter. The second term $\|\boldsymbol{\beta}\|_1$ in (1.1) is the so-called 'lasso penalty' (l_1 penalty), which is the sum of the absolute value of coefficients of predictors. If we change the sum of the absolute value of coefficients of predictors to the sum of the square of the coefficient of predictors, the penalty is the so-called 'ridge penalty' (l_2 penalty), denoted as $\sum_{j=1}^{p} \beta_j^2$ or $\|\boldsymbol{\beta}\|^2$.

The rest of the article is organized as follows: in Section 2, the adaptive network-constrained criterion and its properties are introduced. Section 3 introduces the gradient ascent optimization with the Newton–Raphson (Goeman, 2010) to estimate the parameters; additionally the strategies to choose the tuning parameters and the weight of the lasso are also mentioned. In Section 4, simulations are presented to indicate the performance of the proposed model. In Section 5, the proposed model is applied to analyze the "*China Stock Market Financial Database—Financial Ratios*" and select the important factors related to *Price Earning* (P/E) ratio. Section 6 includes a summary and a discussion.

2. The adaptive network-constrained regression model

Consider a network that is represented by a weighted graph $\mathbf{G} = (\mathbf{V}, \mathbf{E}, \mathbf{W})$, where $\mathbf{V}$ is the set of vertices that correspond to the p predictors $\mathbf{X}$ and $\mathbf{E} = \{e_{u\sim v} | u, v \in (1, 2, \ldots, p)\}$ is the set of edges. $e_{u\sim v}$ is the edge between u vertex (predictor) and v vertex which indicates they are linked together on the network. The weight corresponding to the edge $e_{u\sim v}$ is used to measure the uncertainty of the edge, denoted as w_{uv}. Meanwhile, the degree d_u was introduced to measure how many edges connected to the u and $d_u = \sum_{e_{u\sim v} \in \mathbf{E}} w_{uv}$ for $v \in (1, 2, \ldots, p)$. Thus we know that to make $d_u > 0$, at least one edge must be connected to vertex u; otherwise there are no connections from u. In this paper, the weight w_{uv} is estimated by the correlation coefficient between u and v. In this paper, the weight w_{uv} for edge $e_{u\sim v}$ is defined by: $w_{uv} = 0$ if $u = v$; $w_{uv} = \rho_{uv}$ if $u \neq v$ (ρ_{uv} is the correlation coefficient between variables u and v).

Based on the definition of the network $\mathbf{G}$, the Laplacian matrix $\mathbf{L}$ was introduced. Following the studies of Chung (1997), the element l_{uv} of the Laplacian matrix was determined,

$$l_{uv} = \begin{cases} 1 & \text{if } u = v \\ -w_{uv}/\sqrt{d_u d_v} & \text{if } u \text{ and } v \text{ are adjacent} \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

Inducing the Laplacian matrix to construct the new penalty and combining the lasso penalty, the so-called 'network penalty' was proposed for simultaneous estimation and selection of group variables (Li and Li, 2008),