Journal of the Korean Statistical Society 44 (2015) 352–365

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

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

Some properties of generalized fused lasso and its applications to high dimensional data

Woncheol Jang [a,*], Johan Lim [a], Nicole A. Lazar [b], Ji Meng Loh [c], Donghyeon Yu [d]

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

[b] Department of Statistics, University of Georgia, Athens, GA, USA

[c] Department of Mathematical Sciences, New Jersey Institute of Technology, Newark, NJ, USA

[d] Department of Statistics, Keimyung University, Daegu, Republic of Korea

ARTICLE INFO

Article history:
Received 13 August 2014
Accepted 7 October 2014
Available online 29 October 2014

AMS 2000 subject classifications:
primary 62J05
secondary 62H30

Keywords:
Prediction
Regularization
Spatial correlation
Supervised clustering
Variable selection

ABSTRACT

Identifying homogeneous subgroups of variables can be challenging in high dimensional data analysis with highly correlated predictors. The generalized fused lasso has been proposed to simultaneously select correlated variables and identify them as predictive clusters (grouping property). In this article, we study properties of the generalized fused lasso. First, we present a geometric interpretation of the generalized fused lasso along with discussion of its persistency. Second, we analytically show its grouping property. Third, we give comprehensive simulation studies to compare our version of the generalized fused lasso with other existing methods and show that the proposed method outperforms other variable selection methods in terms of prediction error and parsimony. We describe two applications of our method in soil science and near infrared spectroscopy studies. These examples having vastly different data types demonstrate the flexibility of the methodology particularly for high-dimensional data.

1. Introduction

Suppose that we observe $(x_1, y_1), \ldots, (x_n, y_n)$, where $x_i = (x_{i1}, \ldots, x_{ip})^T$ is a p-dimensional predictor and y_i is the response variable. We consider a standard linear model for each of n observations

$$y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i, \quad \text{for } i = 1, \ldots, n,$$

with $\mathrm{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$. We also assume that the predictors are standardized and the response variable is centered,

$$\sum_{i=1}^{n} y_i = 0, \qquad \sum_{i=1}^{n} x_{ij} = 0 \quad \text{and} \quad \sum_{i=1}^{n} x_{ij}^2 = 1 \quad \text{for } j = 1, \ldots, p.$$

The dramatic increase in the amount of data collected in many fields comes with a corresponding increase in the number of predictors p available in data analyses. For simpler interpretation of the underlying processes generating the data, it is

* Corresponding author.
E-mail addresses: wcjang@snu.ac.kr (W. Jang), johanlim@snu.ac.kr (J. Lim), nlazar@stat.uga.edu (N.A. Lazar), loh@njit.edu (J.M. Loh), dhyeon.yu@gmail.com (D. Yu).

http://dx.doi.org/10.1016/j.jkss.2014.10.002
1226-3192/

often desired to have a relatively parsimonious model. This in turn creates the challenge of identifying important predictors out of the many that are available.

As a motivating example, we consider a study involving near infrared (NIR) spectroscopy data measurements of cookie dough (Osborne, Fearn, Miller, & Douglas, 1984). Near infrared reflectance spectral measurements were made at 700 wavelengths from 1100 to 2498 nanometers (nm) in steps of 2 nm for each of 72 cookie doughs made with a standard recipe. The study aims to predict dough chemical composition using the spectral characteristics of NIR reflectance wavelength measurements. Here, the number of wavelengths p is much bigger than the sample size n.

One possible approach is to cluster predictors based on the correlation structure and to use averages of the predictors in each cluster as new predictors. Park, Hastie, and Tibshirani (2007) use this approach for gene expression data analysis and introduce the concept of a *super gene*. However, NIR spectroscopy data are well known to have measurement errors which induce positive correlations among the wavelengths. Ideally, we would like to keep all relevant (possibly correlated) wavelengths while achieving better predictive performance. The hierarchical clustering used in Park et al. (2007) for grouping does not account for the correlation structure of the predictors.

While variable selection in regression is an increasingly important problem, it is also very challenging, particularly when there is a large number of highly correlated predictors. Since the important contribution of the least absolute shrinkage and selection operator (*lasso*) method by Tibshirani (1996), many other methods based on regularized or penalized regression have been proposed for parsimonious model selection, particularly in high dimensions, e.g. elastic net, fused lasso, OSCAR and generalized lasso (Bondell & Reich, 2008; Tibshirani, Saunders, Rosset, Zhu, & Knight, 2005; Tibshirani & Taylor, 2011; Zou & Hastie, 2005). Briefly, these methods involve penalization to fit a model to data, resulting in shrinkage of the estimators. Many methods have focused on addressing various possible shortcomings of the lasso method, for instance when there is dependence or collinearity between predictors.

In the lasso, a bound is imposed on the sum of the absolute values of the coefficients:

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\| y - \sum_{j=1}^{p} \beta_j x_j \right\|^2 \quad \text{subject to} \sum_{j=1}^{p} |\beta_j| \le t,$$

where $y = (y_1, \ldots, y_n)$ and $x_j = (x_{1j}, \ldots, x_{nj})$.

The lasso method is a shrinkage method, like ridge regression (Hoerl & Kennard, 1970), with automatic variable selection. Due to the nature of the L_1 penalty term, the lasso shrinks each coefficient and selects variables simultaneously. However, a major drawback of the lasso is that if there exists collinearity among a subset of the predictors, it usually only selects one to represent the entire collinear group. Furthermore, the lasso cannot select more than n variables when $p > n$.

Penalized regression methods have also been proposed for grouped predictors (Bondell & Reich, 2008; She, 2010; Tibshirani et al., 2005; Zou & Hastie, 2005). All these methods work by introducing a new penalty term in addition to the L_1 penalty term of the lasso to account for correlation structure. For example, based on the fact that ridge regression tends to shrink the correlated predictors toward each other, the elastic net (Zou & Hastie, 2005) uses a linear combination of the ridge and lasso penalties for group predictor selection; elastic net solves the following constrained least squares optimization problem,

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\| y - \sum_{j=1}^{p} \beta_j x_j \right\|^2 \quad \text{subject to } \alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2 \le t.$$

The second term forces highly correlated predictors to be averaged while the first term leads to a sparse solution of these averaged predictors.

Bondell and Reich (2008) propose OSCAR (Octagonal Shrinkage and Clustering Algorithm for Regression), which is defined by

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\| y - \sum_{j=1}^{p} \beta_j x_j \right\|^2 \quad \text{subject to} \sum_{j=1}^{p} |\beta_j| + c \sum_{j<k}^{p} \max\{|\beta_j|, |\beta_k|\} \le t.$$

By using a pairwise L_∞ norm as the second penalty term, OSCAR encourages equality of coefficients.

Unlike the elastic net and OSCAR, the fused lasso (Tibshirani et al., 2005) accounts for *spatial* correlation of predictors. A key assumption in the fused lasso is that the predictors have a certain type of ordering. The fused lasso solves

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\| y - \sum_{j=1}^{p} \beta_j x_j \right\|^2 \quad \text{subject to} \sum_{j=1}^{p} |\beta_j| \le t_1 \quad \text{and} \quad \sum_{j=2}^{p} |\beta_j - \beta_{j-1}| \le t_2.$$

The second constraint, called a *fusion penalty*, encourages sparsity in the differences of coefficients. The method can theoretically be extended to multivariate data, although with a corresponding increase in computational requirements.

She (2010) introduces the clustered lasso (*classo*), a generalization of the fused lasso. Without the ordering restriction on predictors, the classo is defined by

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\| y - \sum_{j=1}^{p} \beta_j x_j \right\|^2 \quad \text{subject to} \sum_{j=1}^{p} |\beta_j| \le t_1 \quad \text{and} \quad \sum_{j<k}^{p} |\beta_j - \beta_k| \le t_2.$$