Some Properties of Generalized Fused Lasso and Its Applications to High Dimensional Data

Woncheol Jang, Johan Lim
Seoul National University, Seoul, Korea,
wjang,johanlim@snu.ac.kr

Ji Meng Loh
New Jersey Institute of Technology,
i.m.loh@njit.edu

Nicole Lazar

University of Georgia, GA, USA,
nlazar@stat.uga.edu

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. In this article, we study several properties of generalized fused lasso. First, we present a geometric interpretation of the generalized fused lasso along with a discussion of its persistency. Second, we analytically show its grouping property. Third, we introduce a modified version of the generalized fused lasso and perform comprehensive simulation studies to compare our version of the generalized fused lasso with other existing methods, showing that the generalized fused lasso 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.

Keywords: Fused lasso regression; ℓ_1 regularization; grouping property; persistency.

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

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

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

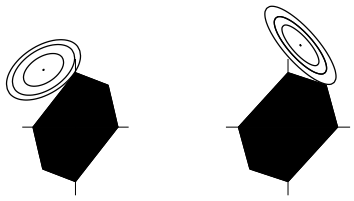
$$\sum_{i=1}^{n} y_i = 0, \quad \sum_{i=1}^{n} x_{ij} = 0 \text{ and } \sum_{i=1}^{n} x_{ij}^2 = 1 \quad \text{for } j = 1, \ldots, p.$$

The ℓ_1 regularization methods are commonly used to analyze high dimensional data in recent years. We can obtain sparse estimates by using ℓ_1 regularization methods. In this paper, we consider the fused lasso regression with the generalized fusion penalty. Let $y = (y_1, y_2, \ldots, y_n) \in \mathrm{R}^n$ and $x_j = (x_{1j}, \ldots, x_{nj}) \in \mathrm{R}^n$. The fused lasso regression (FLR) with the generalized fusion penalty minimizes

$$f(\beta) = \frac{1}{2}\big\|y - -\sum_{j=1}^{p} \beta_j x_j\big\|_2^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{(k,l)\in E} |\beta_k - \beta_l|, \tag{1}$$

where λ_1 and λ_2 are non-negative tunning parameters, E is the index set of adjacent pairs of variables specified in model. We introduce a variant of the generalized fused lasso that effectively selects positively correlated variables in high dimension with an exact grouping property which we will explain in Section 2. We call this procedure a *Hexagonal Operator for Regression with Shrinkage and Equality Selection*, or HORSES for short. We study several interesting properties of the generalized fused lasso. First, HORSES representation provides a better geometrical view of the generalized fused lasso and a better understanding to its persistency. Second, we analytically show that HORSES (equivalently, generalized fused lasso) is a variable selection method that is specifically tailored to the situation in which there are strong positive correlations between predictors. HORSES finds a homogeneous subgroup structure within the high dimensional predictor space. In addition, we implement comprehensive simulation studies to compare our version of the generalized fused lasso with other existing methods and show that the generalized fused lasso outperforms other variable selection methods in terms of prediction error and parsimony.

The remainder of the paper is organized as follows. In Section 2, the HORSES representation of the generalized fused lasso is introduced. In Section 3, we briefly review the recent advances in algorithms for solving the generalized fused lasso and introduce our algorithm for HORSES. In addition, we provide procedures to select tuning parameters. Simulation studies to show the performances of variable selection and prediction are presented in Section 4. Two data analyses using HORSES are presented in Section 5.



(a) (b)

Figure 1: Graphical representation in the (β_1, β_2) plane. HORSES solutions are the first time the contours of the sum of squares function hit the hexagonal constraint region. (a) Contours centered at OLS estimate with a negative correlation. Solution occurs at $\widehat{\beta}_1 = 0$; (b) Contours centered at OLS estimate with a positive correlation. Solution occurs at $\widehat{\beta}_1 = \widehat{\beta}_2$.

2 Model

In this section we describe our variant of the generalized fused lasso. Following the formulation of the elastic net, our penalty term is a linear combination of an L_1 penalty for the coefficients and another L_1 penalty for pairwise differences of coefficients.

The novelty of the generalized fused lasso is that it encourages grouping of *positively* correlated predictors with a sparsity solution. While the elastic net and OSCAR have a similar feature, these methods can put negatively correlated predictors into the same group.

HORSES yields estimates using

$$\begin{aligned} \widehat{\beta} &= \underset{\beta}{\operatorname{argmin}} \|y - \sum_{j=1}^{p} \beta_j x_j\|^2 \text{ subject to} \\ & \alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j<k} |\beta_j - \beta_k| \leq t, \end{aligned}$$

where $d^{-1} \leq \alpha \leq 1$ and d is a thresholding parameter. Our penalty term is mathematically equivalent to the classo by She (2010) except for the thresholding parameter d. However, our specification of the penalty terms using a linear combination provides a geometric interpretation of the constraint region which can not be presented with the form of the classo penalty term.

With the HORSES formulation, instead of an octagon of OSCAR by Bondell and Reich (2008), constraint regions of the generalized fused lasso are represented by a hexagon (Figure 1(a) and 1(b)), which focuses on selection of groups of predictors that are positively correlated. This explains why the generalized fused lasso works better when there are strong positive correlations among the predictors.

What distinguishes HORSES from other generalized fused lasso is the thresholding parameter d. With d, one can prevent the estimates from being a solution only via the second

penalty function, so the HORSES method always achieves sparsity. We recommend $d = \sqrt{p}$, where p is the number of predictors. This ensures that the constraint parameter region lies between that of the L_1 norm and of the elastic net method, i.e. the set of possible estimates for the HORSES procedure is a subset of that of the elastic net. As a result, we can show that HORSES is persistent (Greenshtein and Ritov, 2004) when the elastic net is persistent. If we choose $d = p$, the HORSES parameter region lies within that of the OSCAR method, but requires a much stronger condition for persistence. We will show details in the talk.

As the correlation between two predictors increases, the predictors are more likely to be grouped together. The elastic net also has a grouping property, but does not assign identical coefficients to predictors within groups. The following theorem shows that HORSES has the exact grouping property. Our proof follows closely the proof of Theorem 1 in Bondell and Reich (2008).

Theorem 1. *Let $\lambda_1 = \lambda\alpha$ and $\lambda_2 = \lambda(1-\alpha)$ be the two tuning parameters in the HORSES criterion. Given data $\big(y, X\big)$ with centered response y and standardized predictors $X = (x_1, \ldots, x_p)^t$, let $\widehat{\beta}\big(\lambda_1, \lambda_2\big)$ be the HORSES estimate using the tuning parameters $\big(\lambda_1, \lambda_2\big)$. Let $\rho_{ij} = x_i^{\mathrm{T}} x_j$ be the sample correlation between covariates x_i and x_j.*

For a given pair of predictors x_i and x_j, suppose that both $\widehat{\beta}_i(\lambda_1, \lambda_2)$ and $\widehat{\beta}_j(\lambda_1, \lambda_2)$ are distinct from the other $\widehat{\beta}_k$. Then there exists $\lambda_0 \geq 0$ such that if $\lambda > \lambda_0$ then

$$\widehat{\beta}_i\big(\lambda_1, \lambda_2\big) = \widehat{\beta}_j\big(\lambda_1, \lambda_2\big), \quad \textit{for all } \alpha \in [d^{-1}, 1].$$

Furthermore, it must be that

$$\lambda_0 \leq 2\|y\|\sqrt{2(1-\rho_{ij})}.$$

3 Computation

Developing an efficient algorithm to implement generalized fused lasso procedures is critical for its application to high dimensional data. Recent advances of such algorithms for the generalized fused lasso are twofold: the pathwise algorithm and the optimization procedure for a given set of tuning parameters. The pathwise algorithm for the generalized fused lasso was first discussed by Friedman et al. (2007). They consider the pathwise coordinate descent algorithm, which sequentially solves a series of the coordinate descent (CD) algorithm. However, monotonicity of the solution path doesn't hold for general design matrix. Furthermore, the CD algorithm for non-separable penalty term may not converge. Hence a modified CD algorithm is used in their procedure. Tibshirani and Taylor (2011) propose a pathwise algorithm for generalized fused lasso but this still has difficulty for singular design matrix and high-dimensional data. Recently, another set of algorithms based on the optimization technique called the *first order method*, is introduced. For example, Ye and Xie (2011) introduce an algorithm based on split-Bregman iteration, which iteratively solves an augmented Lagrangian function having additional least square penalties for the violation of linear constraints. Lin et al. (2011) propose the alternating linearization algorithm which solves two

linearized sub-problems derived from the original problem. Liu et al. (2010) rewrite the generalized fused lasso as the fused lasso signal approximator (FLSA) with an identity design matrix and further reformulate the FLSA as a problem of finding an appropriate subgradient of the fused penalty at the minimizer.

To implement HORSES, we use the modified CD algorithm by Friedman et al. (2007) but do not apply the pathwise step since the monotonicity of the solution path does not hold for a general design matrix. Instead, we estimate tuning parameters by minimizing the prediction error with cross-validation. The code is implemented in C and the R statistical package. Example code is available from the second author upon request.

4 Simulations

We numerically compare the performance of HORSES and several other penalized methods: ridge regression, LASSO, elastic net, and OSCAR. The first five scenarios are very similar to those in Zou and Hastie (2005) and Bondell and Reich (2008). We also consider one more scenario for $p > n$ where we choose $p = 100$ because this is the maximum number of predictors that can be handled by the quadratic programming used in OSCAR. The details of results will be given in the talk.

5 Data Analysis

Two data sets are analyzed. The first data set is the cookie dough dataset from Osborne et al. (1984), which was also analyzed by Brown et al. (2001) and Hans (2011). Brown et al. (2001) consider four components as response variables: percentage of fat, sucrose, flour and water associated with each dough piece. Following Hans (2011), we attempt to predict only the flour content of cookies with the 300 NIR reflectance measurements at equally spaced wavelengths between 1200 and 2400 nm as predictors (out of the 700 in the full data set). Also as in Hans (2011) we remove the 23rd and 61st observations as outliers. Then we split the dataset randomly into a training set with 39 observations and a test set with 31 observations. The correlations between NIR reflectance measurements show strong correlations between any pair of predictors in the range of 1200-2200 and 2200-2400. The second data example is the Appalachian Mountains Soil Data from Bondell and Reich (2008), which has 15 soil characteristics computed from measurements made at twenty 500-m^2 plots located in the Appalachian Mountains of North Carolina. The data were collected as part of a study on the relationship between rich-cove forest diversity and soil characteristics. Forest diversity is measured as the number of different plant species found within each plot. The values in the soil data set are averages of five equally spaced measurements taken within each plot and are standardized before the data analysis. These soil characteristics serve as predictors with forest diversity as the response. We apply Elastic Net, OSCAR, and HORSES to analyzing the data and compare the results. Details will be given at the talk.

References

Bondell, H. D. and Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics*, **64** 115–123.

Brown, P.J., Fearn T., and Vannucci, M. (2001). Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *J. Amer. Statist. Assoc.*, **96** 398-408.

Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R.(2007). Pathwise coordinate optimization. *Ann. Appl. Statist.*, **1** 302–332.

Greenshtein, E. and Ritov, Y.(2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, **10** 971–988.

Hans, C. (2011). Elastic net regression modeling with the orthant normal prior. *J. Amer. Statist. Assoc.*, **106** 1383–1393.

Lin, X., Pham, M. and Ruszczynski, A.(2011). Alternating linearization for structured regularization problems. *arXiv:1201.0306*.

Liu, J., Yuan, L. and Ye, J.(2010). An efficient algorithm for a class of fused lasso problems. In *the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 323–332.

Osborne, B.G., Fearn, T., Miller, A.R., and Douglas, S. (1984). Application of near infrared reflectance spectroscopy to compositional analysis of biscuits and biscuit doughs. *J. Sci. Food Agr.*, **35** 99-105.

She, Y. (2010). Sparse regression with exact clustering. *Electron J. Statist.*, **4** 1055-1096.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc., Ser. B*, **58** 267–288.

Tibshirani, R., Saunders, M., Rosset, S. Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fussed lasso. *J. Roy. Statist. Soc. Ser. B*, **67**, 91–108.

Tibshirani, R. J. and Taylor, J..(2011). The solution path of the generalized lasso. *Ann. Statist.*, **39** 1335–1371.

Ye, G.-B. and Xie, X.(2011). Split Bregman method for large scale fused lasso. *Comput. Stat. Data. Anal.*, **55** 1552–1569.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B*, **67** 301–320.