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 xxxx

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

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

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

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

Please cite this article in press as: Jang, W., et al., Some properties of generalized fused lasso and its applications to high dimensional data. Journal of the Korean Statistical Society (2015), http://dx.doi.org/10.1016/j.jkss.2014.10.002

All of these aforementioned methods can be encapsulated in the framework of the generalized lasso (Tibshirani & Taylor, 2011):

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\| y - \sum_{j=1}^{p} \beta_j x_j \right\|^2 \quad \text{subject to } \|D\beta\|_1 \le t$$

where $D \in \mathbb{R}^{m \times p}$ is a specified penalty matrix.

Assuming

$$D = \begin{pmatrix} I_n \\ \lambda F \end{pmatrix}, \qquad F = \begin{pmatrix} -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & -1 & 1 \end{pmatrix},$$

the generalized lasso becomes the fused lasso.

All the above methods, however, have some limitations when it comes to applying them to high dimensional data such as the cookie dough data set. For example, OSCAR cannot handle high dimensional data because it considers all pairwise comparisons and solves the optimization problem with quadratic programming, which additionally requires a number of auxiliary variables. The pathwise algorithm for the generalized lasso by Tibshirani and Taylor (2011) is not computationally efficient for high-dimensional data with numerous penalty terms like the fused lasso ($m = 2p - 1$), since the path algorithm solves its dual problem whose dimension is the number of penalty terms.

In this paper, we introduce a variant of the generalized 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. Our method is similar to classo in terms of penalty form, so we call procedures with L_1 and fusion penalty terms the generalized fused lasso and consider both methods as variants of the generalized fused lasso. The differences between the two methods will be explained in Section 2. 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 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, 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. We conclude the paper with a discussion in Section 6.

2. Generalized fused lasso with its geometric view

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. Computation can be done by solving a constrained least-squares problem.

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. Fig. 1 shows the shape of the constraint regions for the elastic net, OSCAR and HORSES methods. The shape of the constraint region for HORSES is hexagonal. Compared to the elastic net and OSCAR, HORSES encourages equality of coefficients only in the direction of $y = x$.

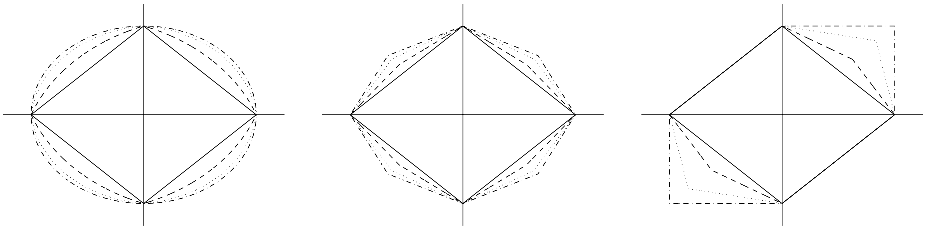
HORSES yields estimates using

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

where $d^{-1} \le \alpha \le 1$ and d is a thresholding parameter. Our penalty term is mathematically equivalent to the classo 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 cannot be presented with the form of the classo penalty term. Note that one can write the HORSES optimization problem in the equivalent Lagrangian form

$$\underset{\beta}{\operatorname{argmin}} \left\{ \left\| y - \sum_{j=1}^{p} \beta_j x_j \right\|^2 + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j<k} |\beta_j - \beta_k| \right) \right\}.$$

Please cite this article in press as: Jang, W., et al., Some properties of generalized fused lasso and its applications to high dimensional data. Journal of the Korean Statistical Society (2015), http://dx.doi.org/10.1016/j.jkss.2014.10.002



(a) elastic net. (b) OSCAR. (c) HORSES.

Fig. 1. Graphical representation of the constraint region in the (β_1, β_2) plane for (a) elastic net, (b) OSCAR, and (c) HORSES.

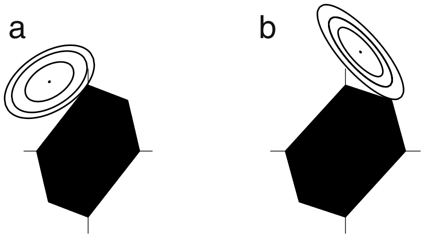


Fig. 2. Graphical representation in the (β_1, β_2) plane. HORSES solutions are the first time the contours of the sum of squares function hit the hexagonal constraint region. (a) Contours centered at OLS estimate with a negative correlation. Solution occurs at $\widehat{\beta}_1 = 0$; (b) Contours centered at OLS estimate with a positive correlation. Solution occurs at $\widehat{\beta}_1 = \widehat{\beta}_2$.

With the HORSES formulation, instead of an octagon as in Fig. 1(b) for OSCAR, constraint regions of the generalized fused lasso are represented by a hexagon (Fig. 1(c)), 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.

In a graphical representation in the (β_1, β_2) plane, the solution is the first time the contours of the sum of squares loss function hit the constraint regions. Fig. 2 gives a schematic view. Fig. 2(a) shows the solution for HORSES when there is negative correlation between predictors. HORSES treats them separately by making $\widehat{\beta}_1 = 0$. On the other hand, HORSES yields $\widehat{\beta}_1 = \widehat{\beta}_2$ when predictors are positively correlated, as in Fig. 2(b).

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 those of the L_1 norm and 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 & 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.

Definition 2.1 (*Greenshtein & Ritov, 2004*)**.** $\widehat{\beta}$ is persistent if $\widehat{R}(\widehat{\beta}) - R(\beta^*) \xrightarrow{P} 0$ where

$$R(\beta) = \frac{1}{n}\mathrm{E}\left\| y - \sum_{j=1}^{p} \beta_j x_j \right\|^2, \qquad \widehat{R}(\beta) = \frac{1}{n}\left\| y - \sum_{j=1}^{p} \beta_j x_j \right\|^2,$$

and $\beta^* = \operatorname{argmin}_{\beta \in \mathcal{M}_n} R(\beta)$. Here $\mathcal{M}_n$ is a constraint region of β.

Following Greenshtein and Ritov (2004), define $z = (Z_0, \ldots, Z_p) = (Y, X_1, \ldots, X_p)$. Then

$$R(\beta) = \sum_{j=0}^{p}\sum_{k=0}^{p} \beta_j \beta_k \mathrm{E}\,(Z_j Z_k), \qquad \widehat{R}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{p}\sum_{j=0}^{p} \beta_j \beta_k Z_{ij} Z_{ik}$$

where $\beta_0 = -1$.

Hence

$$\begin{aligned} |\widehat{R}(\beta) - R(\beta^*)| &\leq \max_{j,k}\left| \frac{1}{n}\sum_i Z_{ij}Z_{ik} - \mathrm{E}\,(Z_j Z_k) \right| \sum_j \sum_k |\beta_j \beta_k| \\ &= \|\beta\|^2 \max_{j,k}\left| \frac{1}{n}\sum_i Z_{ij}Z_{ik} - \mathrm{E}\,(Z_j Z_k) \right|. \end{aligned}$$

From Greenshtein and Ritov (2004), one can assume that

$$\max_{j,k}\left|\frac{1}{n}\sum_{i=1}^{n}Z_{ij}Z_{ik}-\mathrm{E}\,(Z_jZ_k)\right|=O\left(\sqrt{\frac{\log n}{n}}\right).$$

Therefore,

$$\sup_{\beta\in\mathcal{M}_n}|\widehat{R}(\beta)-R(\beta^*)|\le\sup_{\beta\in\mathcal{M}_n}\|\beta\|_1^2\cdot O\left(\sqrt{\frac{\log n}{n}}\right).$$

For the lasso, the sufficient condition for persistence is

$$\sup_{\beta\in\mathcal{M}_n^0}\|\beta\|_1^2=t_n^2=o\left(\frac{n}{\log n}\right)^{1/2},$$

where $\mathcal{M}_n^0=\left\{\beta:\sum_{j=1}^p|\beta_j|\le t_n\right\}$. Define constraint regions of HORSES, OSCAR and the elastic net as follows:

$$\mathcal{M}_n=\left\{x^T\beta:\alpha\sum_{j=1}^{p}|\beta_j|+(1-\alpha)\sum_{j<k}|\beta_j-\beta_k|\le t_n,\ \frac{1}{\sqrt{p}}\le\alpha\le 1\right\},$$

$$\mathcal{M}_n'=\left\{x^T\beta:\alpha\sum_{j=1}^{p}|\beta_j|+(1-\alpha)\sum_{j<k}\max\{|\beta_j|,|\beta_k|\}\le t_n,\ 0\le\alpha\le 1\right\},$$

$$\mathcal{M}_n''=\left\{x^T\beta:\alpha\sum_{j=1}^{p}|\beta_j|+(1-\alpha)\sum_{j=1}^{p}\beta_j^2\le t_n\right\}.$$

For OSCAR, elastic net and HORSES, one can easily show that $\|\beta\|_1^2$ attains the maximum when $\beta_1=\beta_2=\cdots=\beta_p$. Hence

$$\sup_{\beta\in\mathcal{M}_n'}\|\beta\|_1^2=(pt_n)^2\quad\text{and}\quad\sup_{\beta\in\mathcal{M}_n''}\|\beta\|_1^2=pt_n^2.$$

Hence, elastic net is persistent if $t_n=o\left(\frac{n}{p^2\log n}\right)^{1/2}$ while OSCAR needs the stronger condition $t_n=o\left(\frac{n}{p^4\log n}\right)^{1/2}$.

HORSES achieves the maximum with

$$\beta_1=\beta_2=\cdots=\beta_p=\frac{t_n}{p\alpha}.$$

As a result,

$$\sup_{\beta\in\mathcal{M}_n}\|\beta\|^2=\left(\frac{t_n}{\alpha}\right)^2=pt_n^2,$$

when $\alpha=1/\sqrt{p}$. Therefore, if t_n in HORSES is of the same order of t_n in elastic net, HORSES is persistent.

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, that is, each covariate will be assigned to the same coefficient within clusters.

Theorem 2.1. *Let $\lambda_1=\lambda\alpha$ and $\lambda_2=\lambda(1-\alpha)$ be the two tuning parameters in the HORSES criterion. Given data (y,X) with centered response y and standardized predictors $X=(x_1,\ldots,x_p)^t$, let $\widehat{\beta}(\lambda_1,\lambda_2)$ be the HORSES estimate using the tuning parameters (λ_1,λ_2). Let $\rho_{ij}=x_i^Tx_j$ 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 but adjacent in the sense that, for all $k\ne i,j$, $\widehat{\beta}_k\notin\left[\widehat{\beta}_i(\lambda_1,\lambda_2),\widehat{\beta}_j(\lambda_1,\lambda_2)\right]$. Then there exists $\lambda_0\ge 0$ such that if $\lambda>\lambda_0$ then

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

Furthermore, it must be that

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

Proof. Suppose the covariates $(x_1, x_2, \ldots, x_p)$ are ordered such that their corresponding coefficient estimates satisfy

$$\widehat{\beta}_1 \le \widehat{\beta}_2 \le \cdots \le \widehat{\beta}_{\mathrm{L}} < 0 < \widehat{\beta}_{\mathrm{L+1}} \cdots \le \widehat{\beta}_{\mathrm{Q}}$$

and $\widehat{\beta}_{Q+1} = \cdots = \widehat{\beta}_p = 0$.

Let $\widehat{\theta}_1, \ldots, \widehat{\theta}_G$ denote the G unique nonzero values of the set of $\widehat{\beta}_j$, so that $G \le Q$.

For each $g = 1, 2, \ldots, G$, let

$$\mathcal{G}_g = \{j : \widehat{\beta}_j = \widehat{\theta}_g\}$$

denote the set of indices of the covariates that correspond to those values for the coefficients. Now construct the grouped $n \times G$ covariate matrix $X^* \equiv \left[x_1^* \cdots x_G^*\right]$ with

$$x_g^* = \sum_{j \in \mathcal{G}_g} x_j.$$

The optimization problem can be rewritten with the above active set notation as:

$$\underset{\theta_g}{\operatorname{argmin}} \left\{ \left\| y - \sum_{g=1}^{G} \theta_g x_g^* \right\|^2 + \lambda_1 \sum_{g=1}^{G} w_g \theta_g + \lambda_2 \sum_{g_1 < g_2} v_{g1,g2}\big(\theta_{g2} - \theta_{g1}\big) \right\}.$$

Here w_g is the number of elements in the set $\mathcal{G}_g$ and $v_{g1,g2}$ is the number of pairs (j_1, j_2) where $j_1 \in \mathcal{G}_{g1}$ and $j_2 \in \mathcal{G}_{g2}$.

Suppose we consider the derivative of the objective function with respect to θ_g which is

$$-2x_g^{*T}\big(y - X^*\widehat{\theta}\big) + \lambda_1 w_g + \lambda_2 \Big\{ \sum_{g1<g} v_{g1,g} - \sum_{g<g2} v_{g,g2} \Big\} = 0.$$

We also consider the derivative with respect to θ_h:

$$-2x_h^{*T}\big(y - X^*\widehat{\theta}\big) + \lambda_1 w_h + \lambda_2 \Big\{ \sum_{g1<h} v_{g1,h} - \sum_{h<g2} v_{h,g2} \Big\} = 0.$$

Subtracting the latter from the former gives

$$-2\big(x_g^{*T} - x_h^{*T}\big)\big(y - X^*\widehat{\theta}\big) + \lambda_1\big(w_g - w_h\big) + \lambda_2 \Big\{ (v_{+,g} - v_{g,+}) - (v_{+,h} - v_{h,+}) \Big\} = 0, \tag{1}$$

where $\sum_{g1<a} v_{g1,a} = v_{+,a}$ and $\sum_{b<g2} v_{b,g2} = v_{b,+}$. The Eq. (1) is equivalent to

$$-2\big(x_g^{*T} - x_h^{*T}\big)\big(y - X^*\widehat{\theta}\big) - \lambda w_h + \lambda\big(1 - \alpha\big) \sum_{k=g+1}^{h-1} w_k = 0.$$

In the theorem, we assume $\widehat{\beta}_i$ and $\widehat{\beta}_j$ are adjacent, we have $h = g + 1$ (we assume $g < h$ without loss of generality) and $\sum_{k=g+1}^{h-1} w_k = 0$. Thus,

$$\lambda w_h = 2\big(x_g^{*T} - x_h^{*T}\big)\big(y - X^*\widehat{\theta}\big).$$

Since X is standardized, $\|x_i^T - x_j^T\|^2 = 2(1 - \rho_{ij})$. This together with the fact that $\|y - X\widehat{\beta}\|^2 \le \|y\|^2$ gives

$$w_h \le 2\lambda^{-1}\|y\|\sqrt{2(1 - \rho_{ij})}.$$

Therefore, if $2\lambda^{-1}\|y\|\sqrt{2(1 - \rho_{ij})} < 1$, then we encounter a contradiction to the fact that $w_h \ge 1$. □

Theorem 2.1 shows that, for a given λ, if any two adjacent estimates $\widehat{\beta}_i$ and $\widehat{\beta}_j$ are not same, the given λ should be larger than $2\|y\|\sqrt{2(1 - \rho_{ij})}$. This indirectly implies that, for any adjacent pairs x_i and x_j, the required magnitude of penalization (the value of λ) to make their coefficients be equal is inversely proportional to the correlation coefficient ρ_{ij}. Thus, we need a lower (or higher, respectively) value of λ if ρ_{ij} is positive (or negative, respectively) to make their coefficient be equal.

3. Computation and tuning

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, Hastie, Höfling, and Tibshirani (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 does not 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 path algorithm for the fused lasso but this has difficulty for the generalized fused lasso due to the number of penalty terms ($m = p(p+1)/2$). 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, Pham, and Ruszczynski (2011) propose the alternating linearization algorithm which solves two linearized sub-problems derived from the original problem. Liu, Yuan, and Ye (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.

3.1. Computation

Recall that solving the equations for the HORSES procedure,

$$\text{minimize} \quad \left\| y - \sum_{j=1}^{p} \beta_j x_j \right\|^2$$

$$\text{subject to } \alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j<k} |\beta_j - \beta_k| \le t,$$

is equivalent to solving its Lagrangian counterpart

$$f(\beta) = \frac{1}{2} \left\| y - \sum_{j=1}^{p} \beta_j x_j \right\|^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j<k} |\beta_j - \beta_k|, \tag{2}$$

where $\lambda_1 = \lambda\alpha$ and $\lambda_2 = \lambda(1-\alpha)$ with $\lambda > 0$.

Solving (2) to obtain estimates for the HORSES procedure, we modify the pathwise coordinate descent algorithm of Friedman et al. (2007). The pathwise coordinate descent algorithm is an adaptation of the coordinate-wise descent algorithm for solving the 2-dimensional fused lasso problem with a non-separable penalty (objective) function. Our extension involves modifying the pathwise coordinate descent algorithm to solve the regression problem with a fusion penalty. As shown in Friedman et al. (2007), the proposed algorithm is much faster than a general quadratic program solver. Furthermore, it allows the HORSES procedure to run in situations where $p > n$.

Our modified pathwise coordinate descent algorithm has two steps, the descent and the fusion steps. In the descent step, we run an ordinary coordinate-wise descent procedure to sequentially update each parameter β_k given the others. The fusion step is considered when the descent step fails to decrease the objective function. In the fusion step, we add an equality constraint on pairs of β_ks to take into account potential fusions and do the descent step along with the constraint. In other words, the fusion step moves given pairs of parameters together under equality constraints to decrease the objective function. The details of the algorithm are as follows:

- Descent step:

  The derivative of (2) with respect to β_k given $\beta_j = \tilde{\beta}_j, j \ne k$, is

  $$\frac{\partial f(\beta)}{\partial \beta_k} = x_k^T x_k \beta_k - \Bigl(y - \sum_{j \ne k} \tilde{\beta}_j x_j\Bigr)^T x_k + \lambda_1 sgn(\beta_k) + \lambda_2 \sum_{j=1}^{k-1} sgn(\tilde{\beta}_j - \beta_k) + \lambda_2 \sum_{j=k+1}^{p} sgn(\beta_k - \tilde{\beta}_j), \tag{3}$$

  where the $\tilde{\beta}_j$s are current estimates of the β_j's and $sgn(x)$ is a subgradient of $|x|$. The derivative (3) is piecewise linear in β_k with breaks at $\{0, \tilde{\beta}_j, j \ne k\}$ unless $\beta_k \notin \{0, \tilde{\beta}_j, j \ne k\}$.

  - If there exists a solution to $\bigl(\partial f(\beta) / \partial \beta_k\bigr) = 0$, we can find an interval (c_1, c_2) which contains it, and further show that the solution is

    $$\tilde{\beta}_k = sgn\Bigl\{\tilde{y}^T x_k - \lambda_2 \Bigl(\sum_{j<k} s_{jk} + \sum_{j>k} s_{kj}\Bigr)\Bigr\} \times \frac{\Bigl(\Bigl|\tilde{y}^T x_k - \lambda_2 \Bigl(\sum_{j<k} s_{jk} + \sum_{j>k} s_{kj}\Bigr)\Bigr| - \lambda_1\Bigr)_+}{x_k^T x_k},$$

    where $\tilde{y} = y - \sum_{j \ne k} \tilde{\beta}_j x_j$, and $s_{jk} = sgn(\tilde{\beta}_j - \frac{c_1 + c_2}{2})$.

– If there is no solution to $\left(\partial f(\beta)/\partial\beta_k\right) = 0$, we let

$$\tilde{\beta}_k = \begin{cases} \tilde{\beta}_l & \text{if } f(\tilde{\beta}_l) = \min\left\{f(0), f(\tilde{\beta}_j),\ \text{for } j \neq k\right\} \\ 0 & \text{if } f(0) \le f(\tilde{\beta}_j),\ \text{for every } j \neq k. \end{cases}$$

- Fusion step:

  If the descent step fails to decrease the objective function $f(\beta)$, we consider the fusion of pairs of β_ks. For every single pair (k, l), $l \neq k$, we consider the equality constraint $\beta_k = \beta_l = \gamma$ and try a descent move in γ. The derivative of (2) with respect to γ becomes

$$\frac{\partial f(\beta)}{\partial\gamma} = (x_k^T x_k + x_l^T x_l)\gamma - \tilde{y}^T(x_k + x_l) + 2\lambda_1 sgn(\gamma) + 2\lambda_2 \sum_{j<k,l} sgn(\tilde{\beta}_j - \gamma) + 2\lambda_2 \sum_{j>k,l} sgn(\gamma - \tilde{\beta}_j),$$

  where $\tilde{y} = y - \sum_{j\neq k,l} \tilde{\beta}_j x_j$. If the optimal value of γ obtained from the descent step decreases the objective function, we accept the move $\beta_k = \beta_l = \gamma$.

3.2. *Choice of tuning parameters*

Estimation of the tuning parameters α and λ used in the algorithm above is very important for its successful implementation, as it is for the other methods of penalized regression. Several methods have been proposed in the literature, and any of these can be used to tune the parameters of the HORSES procedure. K-fold cross-validation (CV) randomly divides the data into K roughly equally sized and disjoint subsets D_k, $k = 1, \ldots, K$; $\bigcup_{k=1}^K D_k = \{1, 2, \ldots, n\}$. The CV error is defined by

$$\text{CV}(\alpha, \lambda) = \sum_{k=1}^{K} \sum_{i\in D_k} \left(y_i - \sum_{j=1}^{p} \widehat{\beta}_j^{(-k)}(\alpha, \lambda) x_{ij} \right)^2,$$

where $\widehat{\beta}_j^{(-k)}(\alpha, \lambda)$ is the estimate of β_j for a given α and λ using the data set without D_k.

Generalized cross-validation (GCV) and Bayesian information criterion (BIC) (Tibshirani, 1996; Tibshirani et al., 2005; Zou, Hastie, & Tibshirani, 2007) are other popular methods. These are defined by

$$\text{GCV}(\alpha, \lambda) = \frac{\text{RSS}(\alpha, \lambda)}{n - \text{df}},$$
$$\text{BIC}(\alpha, \lambda) = n \times \log\left(\text{RSS}(\alpha, \lambda)\right) + \log n \times \text{df}$$

where $\widehat{\beta}_j\left(\alpha, \lambda\right)$ is the estimate of β_j for a given α and λ, df is the degrees of freedom and

$$\text{RSS}(\alpha, \lambda) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \widehat{\beta}_j(\alpha, \lambda) x_{ij} \right)^2.$$

Here, the degrees of freedom is a measure of model complexity. To apply these methods, one must estimate the degrees of freedom (Efron, Hastie, Johnstone, & Tibshirani, 2004). Following Tibshirani et al. (2005) for the fused lasso, we use the number of distinct groups of non-zero regression coefficients as an estimate of the degrees of freedom.

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 two more scenarios 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 data are generated from the model

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

where $\epsilon \sim N\left(0, \sigma^2\right)$. Except for scenario **C5** and **C7**, we generate predictors $x_i = (x_{i1}, \ldots, x_{ip})^t$ from a multivariate normal distribution with mean 0 and covariance Σ where $\Sigma_{j,j} = 1$ for $j = 1, \ldots, p$.

(**C1**) $n = 20$, $p = 8$, $\Sigma_{i,j} = 0.7^{|i-j|}$, $\sigma = 3$ and

$$\beta = (3, 2, 1.5, 0, 0, 0, 0, 0)^T.$$

(**C2**) $n = 20$, $p = 8$, $\Sigma_{i,j} = 0.7^{|i-j|}$, $\sigma = 3$ and

$$\beta = (3, 0, 0, 1.5, 0, 0, 0, 2)^T.$$

Table 1
True number of groups in each scenario.

C1	C2	C3	C4	C5	C6	C7
3	3	1	1	3	4	2

(**C3**) $n = 20, p = 8, \Sigma_{i,j} = 0.7^{|i-j|}, \sigma = 3$ and

$$\beta = (0.85, 0.85, 0.85, 0.85.0.85, 0.85, 0.85, 0.85)^T.$$

(**C4**) $n = 100, p = 40, \Sigma_{i,j} = 0.5, \sigma = 15$ and

$$\beta = (\underbrace{0, \dots, 0}_{10}, \underbrace{2, \dots, 2}_{10}, \underbrace{0, \dots, 0}_{10}, \underbrace{2, \dots, 2}_{10})^T.$$

(**C5**) $n = 50, p = 40, \sigma = 15$ and

$$\beta = (\underbrace{3, \dots, 3}_{15}, \underbrace{0, \dots, 0}_{25})^T.$$

The predictors for scenario **(C5)** are generated as follows:

$$\begin{aligned}
&x_i = Z_1 + \eta_i^x, Z_1 \sim N(0, 1), \quad i \in G_1 = \{1, \dots, 5\}\\
&x_i = Z_2 + \eta_i^x, Z_2 \sim N(0, 1), \quad i \in G_2 = \{6, \dots, 10\}\\
&x_i = Z_3 + \eta_i^x, Z_3 \sim N(0, 1), \quad i \in G_3 = \{11, \dots, 15\}\\
&x_i \sim N(0, 1), \quad i = 16, \dots, 40,
\end{aligned}$$

where $\eta_i^x \sim N(0, 0.16), i = 1, \dots, 15$. Then $\mathrm{Corr}(x_i, x_j) \approx 0.85$ for $i, j \in G_k$ for $k = 1, 2, 3$.

(**C6**) $n = 50, p = 100, \Sigma_{i,j} = 0.7^{|i-j|}, \sigma = 3$ and

$$\beta = (\underbrace{3, \dots, 3}_{5}, \underbrace{0, \dots, 0}_{10}, \underbrace{2, \dots, 2}_{5}, \underbrace{0, \dots, 0}_{10}, \underbrace{-1.5, \dots, -1.5}_{5}, \underbrace{0, \dots, 0}_{10}, \underbrace{1, \dots, 1}_{5}, \underbrace{0, \dots, 0}_{50})^T.$$

(**C7**) $n = 100, p = 40, \sigma = 15$ and

$$\beta = (\underbrace{2, \dots, 2}_{10}, \underbrace{-2, \dots, -2}_{10}, \underbrace{0, \dots, 0}_{20})^T$$

The predictors for scenario (**C7**) are generated as follows:

$$\begin{aligned}
&Z_1 \sim N(0, 1)\\
&x_i = Z_1 + \eta_i^x, \quad i \in G_1 = \{1, \dots, 10\}\\
&x_i = -Z_1 + \eta_i^x, \quad i \in G_2 = \{11, \dots, 20\}\\
&x_i \sim N(0, 1), \quad i = 21, \dots, 40
\end{aligned}$$

where $\eta_i^x \sim N(0, 0.16), i = 1, \dots, 20$. Then $\mathrm{Corr}(x_i, x_j) \approx 0.85$ for $i, j \in G_k$ for $k = 1, 2$ and $\mathrm{Corr}(x_i, x_j) \approx -0.85$ for $i \in G_k$ and $j \in G_l$ for $k \neq l$. In other words, there are two blocks of non-zero coefficients and their corresponding variables are positively correlated in the same block and negatively correlated in the different block.

We generate 100 data sets of size $2n$ for each scenario **C1**–**C7**. In each data set, the final model is estimated as follows: (i) For each (α, λ), we use the first n observations as a training set to estimate the model and use the other n observations as a validation set to compute the prediction error $\mathrm{PE}(\alpha, \lambda)$; (ii) We set the tuning parameters to be the values (α^*, λ^*) that minimize the prediction error $\mathrm{PE}(\alpha, \lambda)$; (iii) The final model is estimated using the training set with $(\alpha, \lambda) = (\alpha^*, \lambda^*)$.

We compare the mean square error (MSE) and the model complexity of the five penalized methods. The MSE is calculated as in Tibshirani (1996) via

$$\mathrm{MSE} = (\widehat{\beta} - \beta)^T \mathrm{V} (\widehat{\beta} - \beta),$$

where V is the population covariance matrix for X. The model complexity is measured by the number of groups. Based on the coefficient values and correlation structure, Table 1 shows the true number of groups for each of the seven scenarios. Note that the true number of groups is not always the same as the degrees of freedom. For example, we note that the true number of groups in scenario **C5** is three based on the correlation structure although all nonzero coefficients have the same value. On the other hand, scenario **C4** assumes a compound symmetric covariance structure, therefore the number of groups only depends on the coefficient values. Hence, the order of the coefficients does not matter and we can consider scenario

Table 2
MSE and model complexity.

Case	Method	MSE Med.	MSE 10th perc.	MSE 90th perc.	DF Med.	DF 10th perc.	DF 90th perc.
C1	Ridge	2.31	0.98	4.25	8	8	8
	Lasso	1.92	0.68	4.02	5	3	8
	Elastic net	1.64	0.49	3.26	5	3	7.5
	OSCAR	1.68	0.52	3.34	4	2	7
	HORSES	1.85	0.74	4.40	5	3	8
C2	Ridge	2.94	1.36	4.63	8	8	8
	Lasso	2.72	0.98	5.50	5	3.5	8
	Elastic net	2.59	0.95	5.45	6	4	8
	OSCAR	2.51	0.96	5.06	5	3	8
	HORSES	2.21	1.03	4.70	5	2	8
C3	Ridge	1.48	0.56	3.39	8	8	8
	Lasso	2.94	1.39	5.34	6	4	8
	Elastic net	2.24	1.02	4.05	7	5	8
	OSCAR	1.44	0.51	3.61	5	2	7
	HORSES	0.50	0.02	2.32	2	1	5.5
C4	Ridge	27.4	21.2	36.3	40	40	40
	Lasso	45.4	32	56.4	21	16	25
	Elastic net	34.4	24	45.3	25	21	28
	OSCAR	25.9	19.1	38.1	15	5	19
	HORSES	21.2	19.3	33.0	3.5	1	19.5
C5	Ridge	70.2	41.8	103.6	40	40	40
	Lasso	64.7	27.6	116.5	12	9	18
	Elastic net	40.7	17.3	94.2	17	13	25
	OSCAR	51.8	14.8	96.3	12	9	18
	HORSES	46.1	18.1	92.8	11	5.5	19.5
C6	Ridge	27.71	19.53	38.53	100	100	100
	Lasso	13.36	7.89	20.18	31	24	39.1
	Elastic net	13.57	8.49	25.33	30	23.9	37
	OSCAR	13.16	8.56	19.16	50.00	35.9	83.7
	HORSES	12.20	7.11	22.02	33.5	24	66.3
C7	Ridge	18.29	7.85	26.61	40	40	40
	Lasso	27.67	18.70	42.73	15	11	24
	Elastic net	24.69	16.99	36.94	14	11	21
	OSCAR	27.17	16.70	43.25	15	10	23
	HORSES	16.40	6.58	32.72	16	5	18.1

C4 as having only one group of non-zero coefficients. We take the model complexity of scenario **C6** to be four, based on the coefficient values. However, it is possible that some of the zero coefficients might be included as signals because of strong correlations and relatively small differences in coefficient values in this case. For example, the correlation between $x_{50}(\beta_{50} = 1)$ and $x_{51}(\beta_{51} = 0)$ is 0.7. Therefore it is possible that the true model complexity in this case may be bigger than four. **C7** is similar to **C5**, hence it is not straightforward to determine the true model complexity in **C7**.

The simulation results are summarized in Table 2. The HORSES procedure reports the smallest dfs except for scenarios **C1** and **C6**. In both scenarios, the differences of df between the least complex model and HORSES is marginal (4 vs 5 in **C1** and 30 vs 33.5 in **C6**). The HORSES procedure is also very competitive in the MSE comparison. Its MSE is the smallest in scenarios **C2–4** and **C6–7** and the second or third smallest in scenarios **C1** and **C5**.

It is interesting to observe that HORSES is the best in scenario **C2**, but third in scenario **C1** although the differences in MSE and df of the elastic net and HORSES in scenario **C1** are minor. The values of the parameters are the same in both scenarios, but variables with similar coefficients are highly correlated in scenario **C1**, while these variables have little correlation with each other in scenario **C2**. Hence we can consider the grouping of predictors as mainly determined by coefficient values in scenario **C2** while in scenario **C1**, the correlation structure may have an important role in the grouping. This can be confirmed by comparing the median MSEs of each method in the two scenarios **C1** and **C2**. As expected, the median MSE in scenario **C1** is always smaller than the median MSE in scenario **C2**. The difference in the median MSEs can be interpreted as the gain achieved by using the correlation structure when grouping. Because of the explicit form of the fusion penalty in HORSES, our procedure seems to give more weights to differences among the coefficient values while still accounting for correlations. As a result, HORSES effectively groups in scenario **C2**.

Not surprisingly, the HORSES procedure is much more successful than the other procedures in finding the correct model in scenario **C3**, where it gives a higher weight to the fusion penalty (α close to $1/\sqrt{p}$). In our simulation, the average of chosen αs in scenario **C3** is 0.5546 that is closer to $1/\sqrt{8} \approx 0.3536$ than 1. HORSES also has the smallest MSE among the methods. In this case, the true model is not sparse and the lasso and elastic net methods fail.

Please cite this article in press as: Jang, W., et al., Some properties of generalized fused lasso and its applications to high dimensional data. Journal of the Korean Statistical Society (2015), http://dx.doi.org/10.1016/j.jkss.2014.10.002

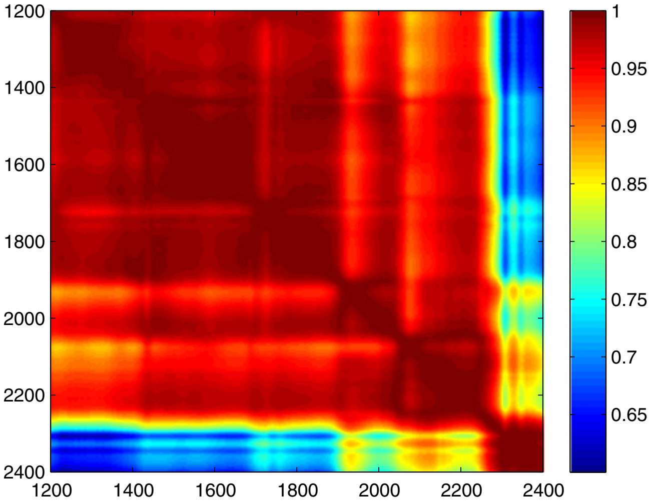


Fig. 3. Graphical representation of the correlation matrix of the 300 wavelengths of the cookie dough data.

HORSES outperforms the other methods again in scenario **C4**. Since the model assumes the compound symmetric covariance structure, the grouping is solely based on the coefficient values. Because of the fusion penalty, the HORSES procedure is very effective in grouping and produces 3.5 as the median df while the second smallest df is 15 with OSCAR.

In scenario **C5**, HORSES has the second smallest median MSE (=46.1) with the elastic net's median MSE smallest at 40.7. However, HORSES chooses the least complex model and shows better grouping compared to the elastic net.

Scenario **C6** considers a large p and small n case. The HORSES procedure reports the smallest MSE while the elastic net chooses the least complex model. However we notice that all methods report at least 30 as the df. This might be due to the fact that the true model complexity in this case is not clear, as we point out above.

In scenario **C7**, HORSES has the smallest median MSE (=16.4). However surprisingly, HORSES also has the smallest median df while elastic net and the OSCAR have slightly higher dfs in **C7** in which they are expected to outperform the HORSES. This seems due to the complicated correlation structure of **C7**. In summary, HORSES procedure outperforms the other methods in choosing the least complex model and attaining the best grouping, while also providing competitive results in terms of MSE.

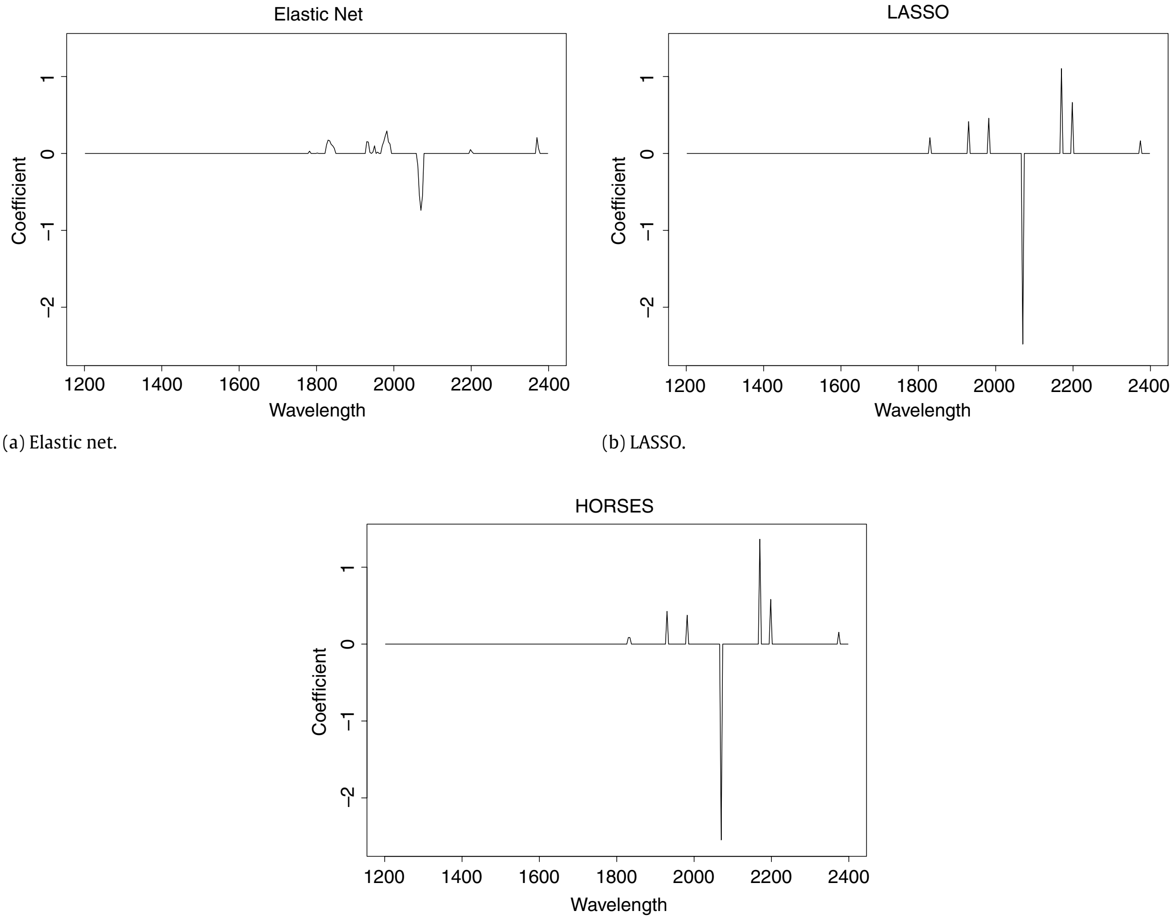
5. Data analysis

In this section, we consider two applications. The first one is a high dimensional data example where we show how the proposed method achieves sparsity and clustering simultaneously. The other one is analysis of a small data set in which we compare HORSES with other methods in detail.

5.1. Cookie dough data

In this case study, we consider the cookie dough data set from Osborne et al. (1984), which was also analyzed by Brown, Fearn, and Vannucci (2001), Caron and Doucet (2008), Griffin and Brown (2012), 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 data set randomly into a training set with 39 observations and a test set with 31 observations. Fig. 3 shows the correlations between NIR reflectance measurements based on all observations. There are very strong correlations between any pair of predictors in the range of 1200–2200 and 2200–2400. Note however that strong correlations do not necessarily imply strong signals in this case since the correlations can be due to measurement error.

With the training data set, tuning parameters of HORSES are computed to be $\alpha = 0.999$ and $\lambda = 0.1622$ (equivalently, $\lambda_1 = 0.1620$ and $\lambda_2 = 0.00016$). Since the L_1 penalty dominates the penalty function, we expect that both HORSES and the lasso will yield very similar results. We compare the lasso, elastic net and HORSES via the prediction mean squared error and degrees of freedom on the test data. The OSCAR method is not included in this comparison because we are not able to apply it due to the high dimension of the data. Table 3 presents the prediction mean squared error and degrees of freedom of each method. The elastic net has the smallest MSE, but the differences in MSE across the three methods are small. On the other hand, the lasso and HORSES methods provide parsimonious models with small degrees of freedom. The estimated coefficients for the lasso, elastic net and HORSES methods are presented in Fig. 4. The elastic net produces 11 peaks while



(a) Elastic net.

(b) LASSO.

(c) HORSES.

Fig. 4. Coefficient estimates for the 300 predictors of the cookie dough data.

Table 3
Cookie dough data results.

	Elastic net	HORSES	Lasso
Mean squared error	2.442	2.586	2.556
Degrees of freedom	11	7	7

both the lasso and HORSES have 7 peaks. The estimated spikes from the lasso and HORSES are consistent with the results obtained in Caron and Doucet (2008). The main difference between the two methods is at wavelengths 1832 and 1836, where the lasso estimates are 0.204 and 0 while the HORSES estimates are 0.0853 at both wavelengths. The elastic net has peaks at wavelength 1784 and 1804 but the other two methods do not yield a peak at those wavelengths. We observe a reverse pattern at wavelength 2176.

5.2. Appalachian Mountains Soil Data

Our next example is the Appalachian Mountains Soil Data from Bondell and Reich (2008). Fig. 5 shows a graphical representation of the correlation matrix of 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.

As can be seen from Fig. 5, there are several highly correlated predictors. Note that our correlation graphic shows the signed correlation values and is thus different from the one in Bondell and Reich (2008) showing the *absolute* value of

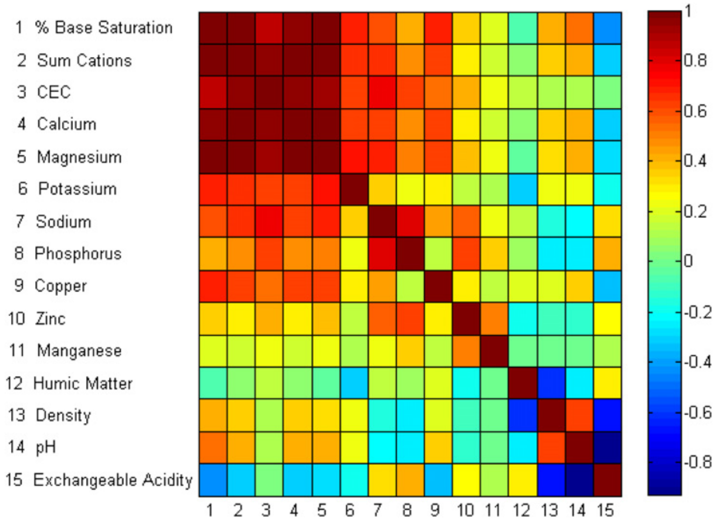


Fig. 5. Graphical representation of the correlation matrix of the 15 predictors of the Appalachian soil data.

Table 4
Results of analyzing the Appalachian soil data using OSCAR and HORSES, and two different methods for choosing the tuning parameters.

Variable	OSCAR (5-fold CV)	OSCAR (GCV)	HORSES (5-fold CV)	HORSES (GCV)
% Base saturation	0	−0.073	0	−0.1839
Sum cations	−0.178	−0.174	−0.1795	−0.1839
CEC	−0.178	−0.174	−0.1795	−0.1839
Calcium	−0.178	−0.174	−0.1795	−0.1839
Magnesium	0	0	0	0
Potassium	−0.178	−0.174	−0.1795	−0.1839
Sodium	0	0	0	0
Phosphorus	0.091	0.119	0.0803	0.2319
Copper	0.237	0.274	0.2532	0.3936
Zinc	0	0	0	−0.0943
Manganese	0.267	0.274	0.2709	0.3189
Humic matter	−0.541	−0.558	−0.5539	−0.6334
Density	0	0	0	0
pH	0.145	0.174	0.1276	0.2319
Exchangeable acidity	0	0	0	0.0185
Degrees of Freedom	6	5	6	7

correlation. The first seven covariates are closely related. Specifically they concern positively charged ions (cations). The predictors named "calcium", "magnesium", "potassium", and "sodium" are all measurements of cations of the corresponding chemical elements, while "% Base Saturation", "Sum Cations" and "CEC" (cation exchange capacity) are all summaries of cation abundance. The correlations between these seven covariates fall in the range (0.360, 0.999). There is a very strong positive correlation between percent base saturation and calcium ($r = 0.98$), but the correlation between potassium and sodium ($r = 0.36$) is not quite as high as the others. Of the remaining eight variables, the strongest negative correlation is between soil pH and exchangeable acidity ($r = -0.93$). Since both of these are measures of acidity, this appears surprising. However, exchangeable acidity measures only a subset of the acidic ions measured in pH, this subset being of more significance only at low pH values.

Note that because "Sum Cations" is the sum of the other four cation measurements the design matrix for these predictors is not full rank.

We analyze the data with the HORSES and OSCAR procedures and report the results in Table 4. Although OSCAR and HORSES use the same definition of df, the OSCAR procedure groups predictors based on the *absolute* values of the coefficients. Therefore the number of groups is not the same as the df in OSCAR. The results for the lasso using the 5-fold cross-validation and GCV can be found in Bondell and Reich (2008). The 5-fold cross-validation OSCAR and HORSES solutions are similar. They select the exact same variables, but with slightly different coefficient estimates. Since the sample size is only 20 and the number of predictors is 15, the 5-fold cross-validation method may not be the best choice for selecting tuning parameters. However, using GCV, OSCAR and HORSES provide different answers. Compared to the 5-fold cross-validation solutions, the OSCAR solution has one more predictor (% Base saturation) while the HORSES solution has 3 additional predictors (% Base saturation, Zinc, Exchangeable acidity). More interestingly, in the OSCAR solution, % Base saturation is not in the group measuring *abundance of cations*, while pH is.

On the other hand, the % Base saturation variable is included in the *abundance of cations* group. The HORSES solution also produces an additional group of variables consisting of Phosphorus and pH.

6. Conclusion

We proposed a new group variable selection procedure in regression that produces a sparse solution and also groups positively correlated variables together. We developed a modified pathwise coordinate optimization for applying the procedure to data. Our algorithm is much faster than a quadratic program solver and can handle cases with $p > n$. For much bigger data, we may consider the majorization–minimization (MM) algorithm proposed by Yu, Won, Lee, Lim, and Yoon (in press).

Such a procedure is useful relative to other available methods in a number of ways. First, it selects groups of variables, rather than randomly selecting one variable in the group as the lasso method does. Second, it groups positively correlated rather than both positively and negatively correlated variables. This can be useful when studying the mechanisms underlying a process, since the variables within each group behave similarly, and may indicate that they measure characteristics that affect a system through the same pathways. Third, the penalty function used ensures that the positively correlated variables do not need to be spatially close. This is particularly relevant in applications where spatial contiguity is not the only indicator of functional relation, such as brain imaging or genetics.

A simulation study comparing the HORSES procedure with the ridge regression, lasso, elastic net and OSCAR methods over a variety of scenarios showed its superiority in terms of sparsity, effective grouping of predictors and MSE.

It is desirable to achieve a theoretical optimality such as the oracle property of Fan and Li (2001) in high dimensional cases. One possibility is to extend the idea of the adaptive elastic net (Zou & Zhang, 2009) to the HORSES procedure. Then we may consider the following penalty form:

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

where $\widehat{w}_j$ are the adaptive data-driven weights.

Investigating theoretical properties of the above estimator will be a topic of future research.

Acknowledgments

Woncheol Jang's research was supported by the Basic Science Research Program through the National Research Foundation (NRF) of Korea grant by the Ministry of Education (No. 2013R1A1A2010065) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2014R1A4A1007895). Johan Lim's work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2011-0030810).

References

Bondell, H. D., & 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., & Vannucci, M. (2001). Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *Journal of the American Statistical Association*, *96*, 398–408.

Caron, F., & Doucet, A. (2008). Sparse Bayesian nonparametric regression. In *Proceedings of the 25th international conference on machine learning. (ICML)*, Helsinki, Finland (pp. 88–95).

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, *32*, 407–499.

Fan, J., & Li, R. (2001). Variable selection via non concave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360.

Friedman, J., Hastie, T., Höfling, H., & Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics*, *1*, 302–332.

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

Griffin, J., & Brown, P. (2012). Bayesian hyper-lassos with non-convex penalization. *Australian & New Zealand Journal of Statistics*, *53*, 423–442.

Hans, C. (2011). Elastic net regression modeling with the orthant normal prior. *Journal of the American Statistical Association*, *106*, 1383–1393.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, *12*, 55–67.

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

Liu, J., Yuan, L., & 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* (pp. 323–332).

Osborne, B. G., Fearn, T., Miller, A. R., & Douglas, S. (1984). Application of near infrared reflectance spectroscopy to compositional analysis of biscuits and biscuit doughs. *Journal of the Science of Food and Agriculture*, *35*, 99–105.

Park, M. Y., Hastie, T., & Tibshirani, R. (2007). Averaged gene expressions for regression. *Biostatistics*, *8*, 212–227.

She, Y. (2010). Sparse regression with exact clustering. *Electronic Journal of Statistics*, *4*, 1055–1096.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, *58*, 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fussed lasso. *Journal of the Royal Statistical Society: Series B*, *67*, 91–108.

Tibshirani, R. J., & Taylor, J. (2011). The solution path of the generalized lasso. *Annals of Statistics*, *39*, 1335–1371.

Ye, G.-B., & Xie, X. (2011). Split Bregman method for large scale fused lasso. *Computational Statistics & Data Analysis*, *55*, 1552–1569.

Yu, D., Won, J.-H., Lee, T., Lim, J., & Yoon, S. (2014). High-dimensional fused lasso regression using majorization-minimization and parallel processing. *Journal of Computational and Graphical Statistics*, http://dx.doi.org/10.1080/10618600.2013.878662 (in press).

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, *67*, 301–320.

Zou, H., Hastie, T., & Tibshirani, R. (2007). On the degrees of freedom of the lasso. *Annals of Statistics*, *35*, 2173–2192.

Zou, H., & Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics*, *37*, 1733–1751.