Discriminative Lasso

Zhihong Zhang [1], Jianbing Xiahou [1] ⋆, Zheng-Jian Bai [2], Edwin R. Hancock [4], Da Zhou [2], Si-Bao Chen [3], and Liyan Chen [1]

[1] Software School, Xiamen University, Xiamen, Fujian, China
[2] School of Mathematical Sciences, Xiamen University, Xiamen, Fujian, China
[3] School of Computer Science and Technology, Anhui University, Hefei, Anhui, China
[4] Department of Computer Science, University of York, York, UK

Abstract. Lasso-type variable selection has been demonstrated to be effective in handling high dimensional data. From the biological perspective, traditional Lasso-type models are capable of learning which stimuli are valuable while ignoring the many that are not, and thus perform feature selection. Traditional Lasso has the tendency to over-emphasize sparsity and to overlook the correlations between features. These drawbacks have been demonstrated to be critical in limiting its performance on real-world feature selection problems. Although some work has considered the problem of correlation, the issue of discriminative ability resulting from sparsity has been overlooked. To overcome this shortcoming, we propose a discriminative Lasso (referred to as dLasso) in which sparsity and correlation are jointly considered. Specifically, the new method can select features (or stimuli) that are correlated more strongly with the response but are less correlated with each other. Moreover, an efficient alternating direction method of multipliers (ADMM) is presented to solve the resulting sparse non-convex optimization problem. Extensive experiments on different data sets show that although our proposed model is not a convex problem, it outperforms both its approximately convex counterparts and a number of state-of-the-art feature selection methods.

Keywords: Lasso, Feature selection, Feature graph, ADMM

1 Introduction

High-dimensional regression/classification is a challenging problem due to the curse of dimensionality. An effective way to resolve this problem is by feature selection, that is to select a subset of the most informative or discriminative predictors from the input predictor set. The pivotal requirement is that the predictor set contains elements that are not only jointly informative with the response, but also have little redundancy with each other. This results in a classifier that performs accurately on the learning samples.

Recently, sparse coding has been demonstrated to be effective in handling high dimensional data [10,27]. In the biological sense, a sparse code generally refers to a representation where a small number of neurons are active with majority of neurons being inactive or showing low activity [19]. It is widely believed that the mammalian visual cortex uses a sparse code to efficiently represent natural images[24,29], where the redundancy of the relayed input is reduced as it passes from the retina. The application

⋆ Corresponding author: jbxiahou@xmu.edu.cn

of sparse coding in feature selection is also inspired by this mechanism. Here each stimulus is encoded by a small subset of neurons. This enables simultaneous parameter estimation and variable selection. A well-known example is the penalization of the ℓ_1-norm of the estimator, known as the Lasso (Least Absolute Shrinkage and Selection Operator) [23].

Lasso assumes that the input variables are nearly independent, i.e., they are not highly correlated, while in most real-life data, the variables are often correlated. For example, the functionality of the human brain typically involves more than one cerebral component. By investigating each individual regional brain phenotype separately will lead to a loss of informative interactions between them [25]. Furthermore, in the presence of highly correlated features Lasso tends to only select one of these features, resulting in suboptimal performance [33]. For this reason, the Elastic Net [5] uses an additional ℓ_2-regularization to promote a grouping effect. This method permits groups of correlated features to be selected when the groups are not known in advance. While promising, these methods do not incorporate prior knowledge into the regression/classification process, which is critical in many applications.

Given feature grouping information, the group Lasso [30] is a refinement in which variables are organized into groups and each group of variables is penalized based on a combination of the ℓ_1-norm and the ℓ_2-norm. If there is a group of variables in which the pairwise correlations are relatively high, Lasso tends to select only one variable from the group and is not sensitive to the feature selected. By contrast, the group Lasso considers this group in a holistic way and determines whether it is important to the problem at hand. If this is the case, then each variable in the group is selected, otherwise none are selected. However, the requirement of a non-overlapping group structure in group Lasso limits its practical applicability. For example, in microarray gene expression data analysis, genes may form overlapping groups since each gene may participate in multiple pathways [13]. A further extension of the group Lasso, namely sparse group Lasso, yields sparsity at both the group and individual feature levels. By contrast, it not only determines which groups are selected, but also further selects some of the most important feature variables from each selected group. The coefficients exhibit sparsity not only between groups, but also within each group [32].

From the above review of the literature, it is clear that traditional Lasso-type models assume conditional independence among the variables, and their aim is to conduct regression individually for each response vector rather than performing it jointly for all the response vectors. As a result they perform only data approximation and representation. In feature selection, they do not explicitly incorporate correlation information either between the response vectors and variables (referred to as relevant information) or the variable correlation (referred to as redundant information). Some recent work has addressed the correlation problem. For instance, Chen et al. [3] proposed an uncorrelated Lasso (unLasso) for variable selection, where variable de-correlation is considered simultaneously with variable selection, so that the selected variables are uncorrelated as much as possible. Jiag et al. [14] proposed a covariate-correlated Lasso (ccLasso) that selects the covariates that correlate most strongly with the response variable. Another popular approach, known as graph sparsity [21,11,22], is to consider the implicit relations between different features. The implicit feature relations can be represented as

a graph, where the nodes represent the features, and the edges imply the relationships between features. By enforcing sparsity in the feature connectivity, the estimation accuracy can be improved using a subgraph containing small number of connected features.

Although much improvement has been achieved in the work mentioned above, the selected features might not be optimal. Intuitively, if we select a few variables to form a linear combination that best approximates the response vector, then the variables correlate strongly with the response vector while there will be little correlation between variables. To our knowledge, existing Lasso-type of variable selection methods have not simultaneously considered the correlation between the response and the variables together with the correlation between the variables. We refer to these two processes as 'response-variable correlation' and 'variable-variable correlation', respectively.

In order to solve the aforementioned problem in existing Lasso-type variable selection methods, we propose a discriminative Lasso (referred to as dLasso). This not only discovers the correlations between the variables and the response, but also discriminates similar features. This distinguishes it from most of the existing work, which uses convex methods and which may be suboptimal in terms of the accuracy of both feature selection and parameter estimation. Specifically, we develop a non-convex paradigm for sparse group feature selection, which is motivated by graph-based clustering methods and which group the dominant vertices into clusters. The proposed dLasso method uses a novel graph regularizer on the feature coefficients which simultaneously considers the 'response-variable correlation' and the 'variable-variable correlation' in the data. Consequently, the proposed regularizer can trade off between feature relevance and feature redundancy. For the purposes of optimization, we employ the alternating direction method of multipliers (ADMM) to solve the proposed method efficiently. Because of the graph-based representation, our method is connectionist in approach, and this potentially provides a biologically plausible route to its implementation.

The remainder of this paper is organized as follows. We briefly review the normal Lasso and Elastic Net in Section 2 and introduce our formulation of discriminative Lasso in Section 3. Then an effective iterative algorithm is presented to solve the sparse optimization problem in Section 4. Experimental results and performance comparisons with competing method are presented in Section 5. We conclude this paper by summarizing the proposed method in Section 6.

2 Brief Review of Sparse Learning Based Feature Selection

According to the structure of the norm, sparsity can be obtained from the following two types of regularization terms for feature selection, namely a)lat sparsity, where the sparsity is often achieved by ℓ_1-norm or ℓ_0-norm regularizer to select individual feature, and b) structural sparsity, where the $\ell_{2,1}$-norm or $\ell_{2,0}$-norm are imposed to select group features.

Typically we have a set of training data $(x_1, y_1), \ldots, (x_n, y_n)$ from which to estimate the parameters β. Each $x_i = \{f_1^i, f_2^i, \ldots, f_d^i\}^T \in \Re^{d\times 1}$ is a predictive vector of feature measurements for the i-th case. The most popular estimation method is least squares, in which we select the coefficients $\beta = \{\beta_1, \ldots, \beta_d\}^T$ to minimize the residual sum of squares

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{d} \beta_j f_j^i)^2 = \min_{\beta} \|\mathbf{y}^T - \beta^T \mathbf{X}\|_2^2$$
$$s.t. \quad \sum_{j=1}^{d} \|\beta\|_0 = k \tag{1}$$

where $\mathbf{y} \in \Re^{n \times 1}$ is the label vector, $\mathbf{X} \in \Re^{d \times n}$ is the training data, and k is the number of features selected. Solving Eq.1 directly has been proved to be NP-hard and hence very difficult to solve via optimization [4]. In many practical situations it is convenient to allow for a certain degree of error, and we can relax the optimization constraint using the following formulation

$$\min_{\beta} \|\mathbf{y}^T - \beta^T \mathbf{X}\|_2^2 + \lambda \|\beta\|_0 \tag{2}$$

where $\lambda \geq 0$ is the regularization parameter. Unfortunately Eq.2 is still challenging, and for practical purposes an alternative formulation using ℓ_1-norm regularization instead of ℓ_0-norm has been proposed

$$\min_{\beta} \|\mathbf{y}^T - \beta^T \mathbf{X}\|_2^2 + \lambda \|\beta\|_1 \tag{3}$$

where $\|\beta\|_1$ is ℓ_1-norm of vector β (sum of absolute elements), $\|\beta\|_1 = \sum_{j=1}^{d} |\beta_j|$. The tuning parameter $\lambda \geqslant 0$ controls the amount of regularization applied to the estimate. The larger λ, the larger the number of zeros in β. The nonzero components give the selected variables. After we obtain β^*, we choose the feature indices corresponding to the top k largest values of the summation of the absolute values along each column. In statistics, Eq.3 is referred to as the regularized counterpart of the Lasso problem [23] and has been widely studied (e.g. [6,16,17]). However, one of the main limitations of ℓ_1-norm feature selection is that it focuses on estimating the response vector for each variable individually without considering relations with the remaining variables. Moreover, the ℓ_1-minimization algorithm is not stable when compared with ℓ_2-minimization [26].

The Elastic Net [33] adds an ℓ_2-minimization term into the Lasso objective function, which can then be formulated as

$$\min_{\beta \in \Re^d} \|\mathbf{y}^T - \beta^T \mathbf{X}\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2, \tag{4}$$

where $\lambda_1, \lambda_2 \geqslant 0$ are tuning parameters. Apart from enjoying a similar sparsity of representation of Lasso, the Elastic Net encourages a grouping effect, where strongly correlated predictors tend to be in or out of the model together [33].

Predictors with high correlation contain similar properties, and contain some overlapped information. In some cases, especially when the number of selected predictors is very limited, more information needs to be contained in the selected predictors. Strongly correlated predictors should not participate in the model together. When strongly correlated predictors are present, then only one is selected. As a result the limited selected predictors will contain more information.

3 Discriminative Lasso

In this section, we develop a new feature selection method by simultaneously considering the 'response-variable correlation' and 'variable-variable correlation' information. This leads to a new Lasso-type variable selection approach. Here the regression coefficients associated with large 'response-variable correlations' together with those associate with small 'variable-variable correlation' are encouraged. As a result, the selected variables are jointly informative with respect to the response, while allowing only limited redundancy among them.

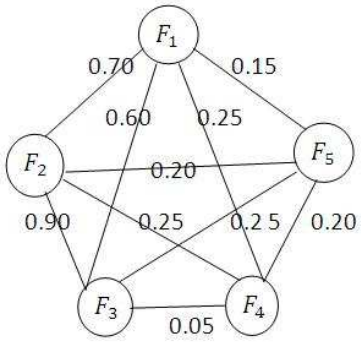


Fig. 1. The subset of features $\{F_1, F_2, F_3\}$ is relevant feature subset

3.1 'Response-Variable' and 'Variable-Variable' Correlation

Suppose that $\mathbf{X} = [x_1, x_2, \cdots, x_n] \in \Re^{d\times n}$ is the matrix of predictors with n observations of d predictors, and the corresponding response vector is $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T \in \Re^{n\times 1}$. Moreover, suppose that the responses and the d predictors are preprocessed so that they have zero mean and unit variance, i.e. $\|\mathbf{y}\| = 1$ and that each feature measurement of a predictor vector is normalized so that $\|f_i\| = 1$.

With these ingredients, our discriminative Lasso (referred to as dLasso) is motivated by the following observation: If we select a relatively small number of variables to form a linear combination that best approximates the response vector, then the variables correlated more with the response vector while little correlation between variables would be good choices. To our knowledge, existing Lasso-type variable selection methods have not simultaneously considered the correlations between the response and the variables, together with the correlation between variables. We refer to these two sources of correlation as the i) 'response-variable correlation' and ii) the 'variable-variable correlation'. Therefore, the discriminating power of a feature pair $\{f_i, f_j\}$ is estimated through Pearson's correlation coefficient [8], and it is defined as

$$S_{i,j} = \frac{1}{2}\rho(f_i, \mathbf{y}) + \frac{1}{2}\rho(f_j, \mathbf{y}) - \rho(f_i, f_j). \tag{5}$$

The definition of feature relevance consists of three terms. The first two terms, which together are referred to as the relevancy, indicate the individual relevance of the feature with respect to the response. The final term is used to measure the redundancy between features. A large value of $S_{i,j}$ means that

a) both $\rho(f_i, \mathbf{y})$ and $\rho(f_j, \mathbf{y})$ are large, and this in turn indicates that features $\{f_i, f_j\}$ are informative themselves with respect to the response $\mathbf{y}$ and
b) $\rho(f_i, f_j)$ is small indicating that features $\{f_i, f_j\}$ are not redundant.

To ensure that the selected predictors of Lasso-type ℓ_1-minimization are discriminative, the regression coefficient vector β should satisfy the condition

$$\max_{\beta \in \Re^d} \beta^T \mathbf{S} \beta, \tag{6}$$

where β is a d-dimensional indicator vector and $S_{ii} = 0$, i.e., all diagonal entries of $\mathbf{S}$ are set to zero. Our idea is motivated by the graph-based clustering method which groups the dominant vertices into a single cluster [31]. Features will be selected if and only if $\beta_i > 0$, and will have maximum internal homogeneity with respect to feature relevance, see Eq.5. In fact, the main property of the selected feature subset $\{f_i | 1 \leqslant i \leqslant d, \beta_i > 0\}$ is that overall their relevance is greater than that for the features which are not selected From graph theory, the set of selected feature turns out to be equivalent to a maximal clique [18]. To provide an example, see Fig. 1. Here, there are N training samples, each having 5 feature vectors. In order to capture the discriminative features from these 5 features (represented as $F_1, \ldots, F_5$), we construct a graph $G = (V, E)$ with node-set V, edge-set $E \subseteq V \times V$ and edge weight matrix $\mathbf{S}$ computed using Eq.5. Each vertex represents a feature and the edge between two features represents their pairwise relationship (or affinity). The weight on the edge reflects the degree of relevance between two features. In our example, in Fig. 1, features $\{F_1, F_2, F_3\}$ form the discriminative feature subset, since the edge weights "internal" to that set (0.6, 0.7 and 0.9) are larger than the sum of those between the internal and external features (which is between 0.05 and 0.25).

3.2 Discriminative Lasso for Feature Selection

Our discriminative feature subset selection method is motivated by the desire to encourage the selected features to correlate more with the response while resulting in low redundancy between them. Therefore, we unify Equations Eq.3 and Eq.6, and propose the dLasso method for feature representation and variable selection, which is formulated as

$$\min_{\beta \in \Re^d} \frac{1}{2} \|\mathbf{y}^T - \beta^T \mathbf{X}\|_2^2 + \lambda_1 \|\beta\|_1 - \lambda_2 \beta^T \mathbf{S} \beta, \tag{7}$$

where $\lambda_1, \lambda_2 \geq 0$ are tuning parameters. Note that $\beta^T \mathbf{S} \beta$ is a nonconvex constrain.

Our dLasso contrasts with previous Lasso-type feature selection methods, which use convex methods and which may be suboptimal in terms of the accuracy of feature selection and parameter estimation. Here, the proposed dLasso method imposes stricter

nonconvex constraints, i.e., 'variable-response correlations' and 'variable-variable correlations', in locating the optimal regression β. Once the solution β^* of Eq.7 is obtained, we can easily recover the number of the selected features and index them.A feature f_i is selected if and only if $\beta_i^* > 0$. Consequently, the number of selected features is determined by the number of positive elements of the indicator vector β^*.

4 Optimization Algorithm

We have proposed a dLasso method to solve the non-convex problem 7 by using alternating direction method of multipliers (ADMM) [1]. The basic idea of the ADMM approach is to decompose a hard problem into a set of simpler ones. ADMM attempts to combine the benefits of augmented Lagrangian methods and with those of dual decomposition for the constrained optimization problem [1]. By introducing an auxiliary variable γ into the objective function Eq.7, the problem solved by ADMM takes the following form:

$$\begin{aligned} &\min_{\beta,\gamma\in\Re^d} f(\beta)+g(\gamma) := \frac{1}{2}\|\mathbf{y}^T-\beta^T\mathbf{X}\|_2^2-\lambda_2\beta^T\mathbf{S}\beta+\lambda_1\|\gamma\|_1, \\ &s.t. \quad \beta-\gamma=0 \end{aligned} \tag{8}$$

which is clearly equivalent to the problem in Eq.7. We can regard γ as a proxy for β. The augmented Lagrangian associated with the constrained problem 8 is given by

$$\begin{aligned} L(\beta,\gamma,z) = \frac{1}{2}\|\curvearrowright^T-\beta^T\mathbf{X}\|_2^2-\lambda_2\beta^T\mathbf{S}\beta+\lambda_1\|\gamma\|_1 \\ +\langle\beta-\gamma,z\rangle+\frac{\rho}{2}\|\beta-\gamma\|_2^2 \end{aligned} \tag{9}$$

Here ρ is a positive penalty parameter (or dual update length) and z is a dual variable (i.e. the Lagrange multiplier) corresponding to the equality constraint $\beta=\gamma$. By introducing an additional variable γ and an additional constraint $\beta-\gamma=0$, we have simplified the problem 7 by decoupling the objective function into two parts that depend on two different variables.

The alternating direction method of multipliers (ADMM) that solves our original problem 7 searches for a saddle point of the augmented Lagrangian by iteratively minimizing $L(\beta,\gamma,z)$ over β and γ, and then updating z according to the following update rule:

1) β-minimization: $\beta^{k+1}=\arg\min_{\beta\in\Re^d} L(\beta,\gamma^k,z^k)$

2) γ-minimization: $\gamma^{k+1}=\arg\min_{\gamma\in\Re^d} L(\beta^{k+1},\gamma,z^k)$

3) z-update: $z^{k+1}=z^k+\rho(\beta^{k+1}-\gamma^{k+1})$

The algorithm iterates until a stopping criterion is satisfied. Applying ADMM, we carry out the following steps at each iteration:

Update β: In the (k+1)-th iteration, β^{k+1} is computed by minimizing $L(\beta, \gamma, z)$ with γ^k and z^k fixed. Then we need to solve the following subproblem:

$$\min_{\beta \in \Re^d} \frac{1}{2}\|\mathbf{y}^T - \beta^T \mathbf{X}\|_2^2 - \lambda_2 \beta^T \mathbf{S}\beta + \langle \beta - \gamma^k, z^k \rangle + \frac{\rho}{2}\|\beta - \gamma^k\|_2^2 \tag{10}$$

Taking the derivatives with respect to elements of the vector β and setting them to zero, we have

$$\begin{aligned}
&\frac{\partial}{\partial \beta}\left[\frac{1}{2}\|\mathbf{y}^T - \beta^T \mathbf{X}\|_2^2 - \lambda_2 \beta^T \mathbf{S}\beta + \langle \beta - \gamma^k, z^k \rangle + \frac{\rho}{2}\|\beta - \gamma^k\|_2^2\right] = 0 \\
&\Rightarrow \begin{cases}
\frac{\partial}{\partial \beta}\frac{1}{2}\|\mathbf{y}^T - \beta^T \mathbf{X}\|_2^2 = -\mathbf{X}\mathbf{y} + \mathbf{X}\mathbf{X}^T\beta, \\
\frac{\partial}{\partial \beta}(-\lambda_2 \beta^T \mathbf{S}\beta) = -2\lambda_2 \mathbf{S}\beta, \\
\frac{\partial}{\partial \beta}\langle \beta - \gamma^k, z^k \rangle = z^k, \\
\frac{\partial}{\partial \beta}(\frac{\rho}{2}\|\beta - \gamma^k\|_2^2) = \rho(\beta - \gamma^k).
\end{cases} \\
&\Rightarrow \beta^{k+1} = (\rho\mathbf{I} + \mathbf{X}\mathbf{X}^T - 2\lambda_2 \mathbf{S})^{-1}[\mathbf{X}\mathbf{y} - z^k + \rho\gamma^k]
\end{aligned} \tag{11}$$

Update γ: Now supposing that β_i^{k+1} and the Lagrangian multipliers z_i^k, $i = 1, \cdots, d$ are fixed in the Lagrangian, the optimization problem related to γ_i^{k+1}, $i = 1, \cdots, d$ boils down to:

$$\min_{\gamma_i} \lambda_1 \sum_{i=1}^{d} |\gamma_i|_1 - \sum_{i=1}^{d} \langle \gamma_i, z_i^k \rangle + \frac{\rho}{2}\sum_{i=1}^{d}(\beta_i^{k+1} - \gamma_i)^2 \tag{12}$$

Taking the derivative with respect to γ_i and setting it to zero, we have

$$\begin{aligned}
&\frac{\partial}{\partial \gamma_i}\left[\lambda_1 \sum_{i=1}^{d} |\gamma_i|_1 - \sum_{i=1}^{d} \langle \gamma_i, z_i^k \rangle + \frac{\rho}{2}\sum_{i=1}^{d}(\beta_i^{k+1} - \gamma_i)^2\right] = 0 \\
&\Rightarrow \frac{\partial(\lambda_1|\gamma_i|)}{\partial \gamma_i} = z_i^k - \rho(\gamma_i - \beta_i^{k+1}) \\
&\Rightarrow \gamma_i^{k+1} = \begin{cases}
\frac{1}{\rho}(z_i^k + \rho\beta_i^{k+1} - \lambda_1), \text{ if } z_i^k + \rho\beta_i^{k+1} > \lambda_1 \\
\frac{1}{\rho}(z_i^k + \rho\beta_i^{k+1} + \lambda_1), \text{ if } z_i^k + \rho\beta_i^{k+1} < -\lambda_1 \\
0 \qquad\qquad\qquad\qquad \text{if } z_i^k + \rho\beta_i^{k+1} \in [-\lambda_1, \lambda_1].
\end{cases}
\end{aligned} \tag{13}$$

Update z: Update z_i^{k+1}, $i = 1, \cdots, d$:

$$z_i^{k+1} = z_i^k + \rho(\beta_i^{k+1} - \gamma_i^{k+1}). \tag{14}$$

A summary of the proposed method is shown in Algorithm 1 below
On the convergence of Algorithm 1, we have the following result.

Algorithm 1: The proposed ADMM algorithm for dLasso

Input: X,y,β^0,z^0, λ_1,λ_2 and ρ
Output: β
1: **while** not converge **do**
2: Update β^{k+1} according to Eq.11;
3: Update γ_i^{k+1}, $i = 1, \cdots, d$ according to Eq.13;
4: Update z_i^{k+1}, $i = 1, \cdots, d$ according to Eq.14.
5: **end while**

Theorem 1. *Let $\{\beta^k\}, \{\gamma^k\}, \{z^k\}$ be the iterative sequences generated by Algorithm 1. Suppose that the sequence $\{z^k\}$ converges to a point, i.e., $\lim_{k\to\infty} z^k = \bar{z}$ for some $\bar{z}$. Then every limit point $(\bar{\beta}, \bar{\gamma})$ of the sequence $\{(\beta^k, \gamma^k)\}$, together with $\bar{z}$, satisfy the necessary first order conditions of the problem 8: 1) Primal feasibility: $\bar{\beta} - \bar{\gamma} = 0$. 2) Dual feasibility: $\nabla f(\bar{\beta}) + \bar{z} = 0$ and $0 \in \partial g(\bar{\gamma}) - \bar{z}$, where ∂ denotes the sub-differential operator (see [20]).*

One can easily prove Theorem 1 by following a proof similar to that of Proposition 3 in [15]. We observe from Theorem 1 that, in general, Algorithm 1 converges to a local solution to problem 8.

The algorithm stops when the primal and dual residuals [1] satisfy a stopping criterion. The stopping criterion can be specified by two thresholds namely a) the absolute tolerance ε_{abs} and b) the relative tolerance ε_{rel} (see Boyd et al.[1] for more details). The penalty parameter ρ affects the primal and dual residuals, and hence affects the termination of the algorithm. A large ρ tends to produce small primal residuals, but increases the dual residuals [1]. A fixed ρ (say 10) is commonly used. But there are some schemes for varying the penalty parameter which achieve better convergence [28].

5 Experiments and Comparisons

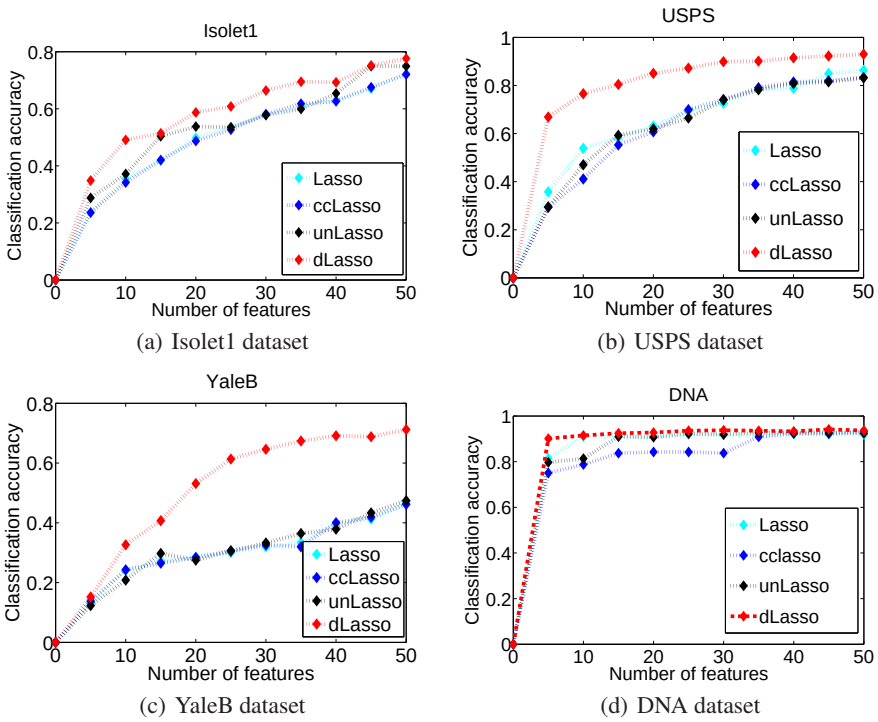
To demonstrate the effectiveness of the proposed approach, we conduct experiments on four benchmark data sets, i.e., the USPS handwritten digit data set [12], Isolet speech data set and DNA data set from the UCI Machine Learning Repository [7], YaleB face data set [9]. Table. 1 summarizes the extents and properties of the four data-sets.

Table 1. Summary of four benchmark data sets

Data-set	Sample	Features	Classes
Isolet1	1560	617	26
USPS	9298	256	10
YaleB	2414	1024	38
DNA	2000	180	3

5.1 Classification Comparison

In order to explore the discriminative capabilities of the information captured by our method, we use the selected features for further classification. We compare the classification results from our proposed method (dLasso) with three representative Lasso-type feature selection algorithms. These methods are the Lasso [23], unLasso [3] and ccLasso [14]. A 10-fold cross-validation strategy using the C-Support Vector Machine (C-SVM) [2] is employed to evaluate the classification performance. Specifically, the entire sample is randomly partitioned into 10 subsets and then we choose one subset for test and use the remaining 9 for training, and this procedure is repeated 10 times. The final accuracy is computed by averaging of the accuracies from all experiments.



(a) Isolet1 dataset (b) USPS dataset

(c) YaleB dataset (d) DNA dataset

Fig. 2. Accuracy rate vs. the number of selected features on four benchmark datasets

The classification accuracies of different algorithms obtained with different feature subsets are shown in Fig.2. From the figure, it is clear that our proposed method dLasso is, by and large, superior to the alternative Lasso-type feature selection methods on all four benchmark datasets. As Fig.2 (a) and (b) shows, when the number of selected features is small, the dLasso performs much better than other Lasso-type feature selection methods. The results verify that dLasso can select more discriminative feature subset than baselines. However, we observed that the advantage of the proposed algorithm

Table 2. Classification results of different feature selection algorithms on different datasets when different number of features are selected. The best results are highlighted in bold.

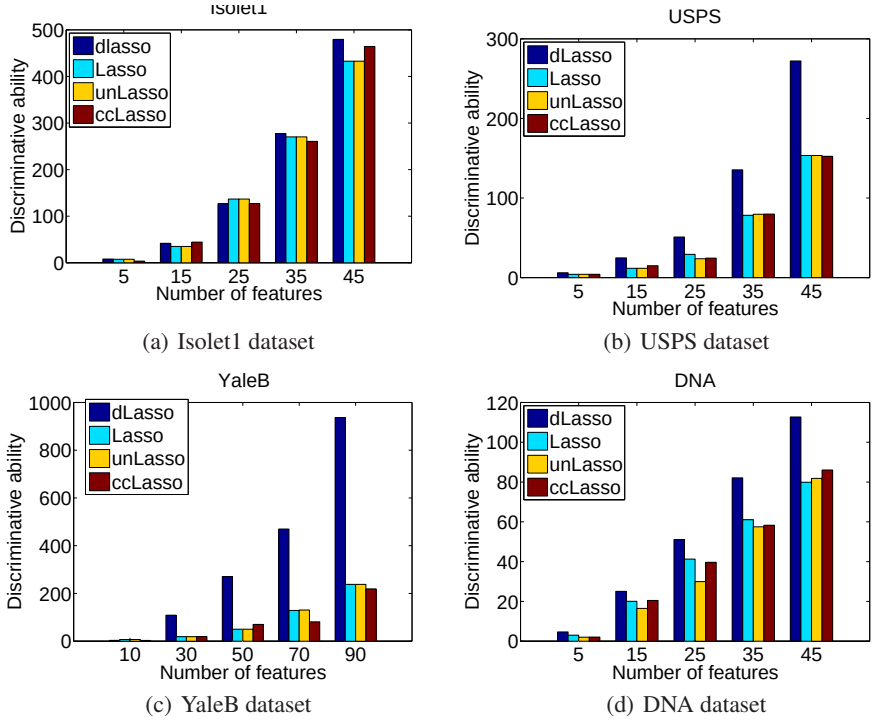
Method	Isolet1	USPS	YaleB	DNA
Lasso	23.59% ± 2.75 (5)	35.00% ± 2.78 (5)	24.40% ± 2.46 (10)	81.45% ± 2.36 (5)
	41.86% ± 4.07 (15)	50.64% ± 3.42 (15)	33.65% ± 2.16 (30)	90.90% ± 2.48 (15)
	53.59% ± 5.11 (25)	56.99% ± 2.58 (25)	46.85% ± 2.68 (50)	92.00% ± 1.90 (25)
	60.64% ± 5.36 (35)	62.82% ± 2.14 (35)	48.63% ± 3.23 (70)	91.25% ± 2.68 (35)
	67.05% ± 3.73 (45)	71.86% ± 1.20 (45)	56.22% ± 2.93 (90)	92.00% ± 1.49 (45)
ccLasso	23.65% ± 2.64 (5)	29.43% ± 1.55 (5)	24.27% ± 2.31 (10)	75.05% ± 2.99 (5)
	42.05% ± 3.21 (15)	55.27% ± 1.86 (15)	31.29% ± 2.02 (30)	83.70% ± 2.76 (15)
	52.69% ± 4.13 (25)	69.95% ± 1.35 (25)	47.18% ± 4.01 (50)	84.25% ± 3.13 (25)
	61.79% ± 2.76 (35)	79.07% ± 1.62 (35)	48.92% ± 2.50 (70)	90.85% ± 1.53 (35)
	67.56% ± 3.81 (45)	81.96% ± 1.40 (45)	55.93% ± 3.14 (90)	92.25% ± 1.51 (45)
unLasso	28.78% ± 2.98 (5)	29.47% ± 1.38 (5)	29.25% ± 2.82 (10)	79.70% ± 2.26 (5)
	50.38% ± 2.99 (15)	59.27% ± 2.06 (15)	38.30% ± 3.76 (30)	91.00% ± 2.36 (15)
	53.53% ± 3.19 (25)	66.48% ± 2.18 (25)	46.89% ± 3.19 (50)	92.25% ± 1.64 (25)
	59.94% ± 2.64 (35)	78.23% ± 1.58 (35)	50.71% ± 3.54 (70)	92.65% ± 2.21 (35)
	74.94% ± 4.06 (45)	81.49% ± 1.30 (45)	53.11% ± 2.04 (90)	92.80% ± 1.27 (45)
dLasso	**34.87% ± 3.97(5)**	**66.89% ± 1.98(5)**	**33.11% ± 2.79(10)**	**90.10% ± 1.79(5)**
	51.41% ± 3.38(15)	**80.43% ± 2.60(15)**	**64.85% ± 3.27(30)**	**92.45% ± 2.42(15)**
	60.83% ± 4.82(25)	**87.20% ± 1.59(25)**	**71.87% ± 2.67(50)**	**93.55% ± 1.95(25)**
	69.49% ± 3.93(35)	**90.13% ± 1.53(35)**	**73.15% ± 3.65(70)**	**93.55% ± 2.05(35)**
	75.19% ± 3.07(45)	**92.27% ± 2.26(45)**	**76.56% ± 3.09(90)**	**94.15% ± 1.83(45)**

over the other three comparative methods tends to diminish as the selected number of features is increased. This is within our expectation, as any feature selection method will work well if we aim to select most of features.

For clear comparison, we summarize the classification accuracy of different methods when different number of features are selected in Table 2. In the table, the classification accuracy (MEAN ± STD) is shown first and the number of features selected is reported in brackets. As can be observed, our method dLasso improved the classification accuracy in the range from 0.25% to 7.7% (Isolet1), 10.31% to 31.89% (USPS), 3.86% to 26.55% (YaleB) and 0.9% to 9.65% (DNA), respectively, compared to the best performances among the competing methods. Comparatively, ccLasso [14] gives the worst performance. This may be explained by our observation that it is unable to handle feature redundancy and is prone to select redundant features. The advantage of the proposed dLasso algorithm over unLasso [3] is that the former not only discovers the correlations between the variables and the response, but also discriminates similar features. As validated by the experiment results, our proposed dLasso method can select the most informative feature subset.

5.2 Discriminative ability

The aim of this experiment is to compare the discriminative ability of selected features by different methods. As mentioned before, if we select a few variables to form



(a) Isolet1 dataset (b) USPS dataset

(c) YaleB dataset (d) DNA dataset

Fig. 3. Discriminative ability of selected features by different methods

a linear combination that best approximates the response vector, then the selected variables correlate more with the response vector while little correlation between variables would be good choice. To further illustrate these, we compute the discrimination of a feature subset by Eq.5. Figure 3 shows the comparison results across different number of selected features. Here we can note that for small number of selected features, all methods have the comparable performance. However, if we select large number of features, the discriminative ability of dLasso is clearly larger than the alternative Lasso-type feature selection methods. The results further verify that dLasso can select more informative feature subset than baselines. The improvement mainly derives from incorporating correlation information between response vectors and variables as well as the variable correlation into the regression model.

6 Conclusion

In this paper, we simultaneously consider the 'response-variable correlation' and 'variable-variable correlation' information in our Lasso-type variable selection approach, where regression coefficients associated with larger 'response-variable correlation' as well as smaller 'variable-variable correlation' are penalized less. Therefore, the selected variables are jointly informative with the response while little redundancy among them. We

employ an efficient ADMM algorithm to solve the proposed formulation. Numerical experiments on real data demonstrate the effectiveness of the proposed method.

Acknowledgments. This work is supported by National Natural Science Foundation of China (Grant No.61402389, 11271308 and 11401499), the Fundamental Research Funds for the Central Universities (No. 20720160073,20720150001,20720140524 and 20720150098) and Fujian Province Soft Sciences Foundation of China (No. 2014R0091).

7 Compliance with Ethical Standards

7.1 Funding

This work is supported by National Natural Science Foundation of China (Grant No.61402389, 11271308 and 11401499), the Fundamental Research Funds for the Central Universities (No. 20720160073,20720150001,20720140524 and 20720150098) and Fujian Province Soft Sciences Foundation of China (No. 2014R0091).

7.2 Disclosure of potential conflicts of interest

I have no potential conflict of interest.

7.3 Research involving human participants and/or animals

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends® in Machine Learning 3(1), 1–122 (2011)
2. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2(3), 1–27 (2011)
3. Chen, S.B., Ding, C., Luo, B., Xie, Y.: Uncorrelated lasso. In: Twenty-Seventh AAAI Conference on Artificial Intelligence (2013)
4. Davis, G., Mallat, S., Avellaneda, M.: Adaptive greedy approximations. Constructive approximation 13(1), 57–98 (1997)
5. De Mol, C., De Vito, E., Rosasco, L.: Elastic-net regularization in learning theory. Journal of Complexity 25(2), 201–230 (2009)
6. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al.: Least angle regression. The Annals of statistics 32(2), 407–499 (2004)
7. Frank, A., Asuncion, A.: Uci machine learning repository (2010)
8. Gan, G., Ma, C., Wu, J.: Data clustering: theory, algorithms, and applications. ASA-SIAM Series on Statistics and Applied Probability,SIAM,Philadelphia, ASA, Alexandria, VA (2007)

9. Georghiades, A.S., Belhumeur, P.N., Kriegman, D.J.: From few to many: Illumination cone models for face recognition under variable lighting and pose. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(6), 643–660 (2001)
10. He, B., Xu, D., Nian, R., van Heeswijk, M., Yu, Q., Miche, Y., Lendasse, A.: Fast face recognition via sparse coding and extreme learning machine. Cognitive Computation 6(2), 264–277 (2014)
11. Huang, J., Zhang, T., Metaxas, D.: Learning with structured sparsity. The Journal of Machine Learning Research 12, 3371–3412 (2011)
12. Hull, J.J.: A database for handwritten text recognition research. IEEE Transactions on Pattern Analysis and Machine Intelligence 16(5), 550–554 (1994)
13. Jacob, L., Obozinski, G., Vert, J.P.: Group lasso with overlap and graph lasso. In: Proceedings of the 26th Annual International Conference on Machine Learning. pp. 433–440. ACM (2009)
14. Jiang, B., Ding, C., Luo, B.: Covariate-correlated lasso for feature selection. In: Machine Learning and Knowledge Discovery in Databases, pp. 595–606. Springer (2014)
15. Magnússon, S., Weeraddana, P.C., Rabbat, M.G., Fischione, C.: On the convergence of alternating direction lagrangian methods for nonconvex structured optimization problems. arXiv preprint arXiv:1409.8033 (2014)
16. Osborne, M.R., Presnell, B., Turlach, B.A.: A new approach to variable selection in least squares problems. IMA journal of numerical analysis 20(3), 389–403 (2000)
17. Osborne, M.R., Presnell, B., Turlach, B.A.: On the lasso and its dual. Journal of Computational and Graphical statistics 9(2), 319–337 (2000)
18. Pavan, M., Pelillo, M.: A new graph-theoretic approach to clustering and segmentation. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. vol. 1, pp. 145–152. IEEE (2003)
19. Pichevar, R., Lahdili, H., Najaf-Zadeh, H., Thibault, L.: New Trends in Biologically-Inspired Audio Coding. INTECH Open Access Publisher (2010)
20. Rockafellar, R.: Convex analysis (1970)
21. Schmidt, M.: Graphical model structure learning with l1-regularization. Ph.D. thesis, UNIVERSITY OF BRITISH COLUMBIA (Vancouver (2010)
22. Shervashidze, N., Bach, F.: Learning the structure for structured sparsity. IEEE Transactions on Signal Processing 63(18), 4894–4902 (2015)
23. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) pp. 267–288 (1996)
24. Vinje, W.E., Gallant, J.L.: Sparse coding and decorrelation in primary visual cortex during natural vision. Science 287(5456), 1273–1276 (2000)
25. Wang, H., Nie, F., Huang, H., Kim, S., Nho, K., Risacher, S.L., Saykin, A.J., Shen, L., Initiative, A.D.N., et al.: Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the adni cohort. Bioinformatics 28(2), 229–237 (2012)
26. Xu, H., Caramanis, C., Mannor, S.: Sparse algorithms are not stable: A no-free-lunch theorem. Pattern Analysis and Machine Intelligence, IEEE Transactions on 34(1), 187–193 (2012)
27. Xu, J., Yang, G., Yin, Y., Man, H., He, H.: Sparse-representation-based classification with structure-preserving dimension reduction. Cognitive Computation 6(3), 608–621 (2014)
28. Yang, S., Wang, J., Fan, W., Zhang, X., Wonka, P., Ye, J.: An efficient admm algorithm for multidimensional anisotropic total variation regularization problems. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 641–649. ACM (2013)
29. Yao, Y., Guo, P., Xin, X., Jiang, Z.: Image fusion by hierarchical joint sparse representation. Cognitive Computation 6(3), 281–292 (2014)

30. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68(1), 49–67 (2006)
31. Zhang, Z.: Feature selection from higher order correlations (2012)
32. Zhao, P., Rocha, G., Yu, B.: The composite absolute penalties family for grouped and hierarchical variable selection. The Annals of Statistics pp. 3468–3497 (2009)
33. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67(2), 301–320 (2005)