Covariate-Correlated Lasso for Feature Selection

Bo Jiang[1], Chris Ding[1,2], and Bin Luo[1]

[1] School of Computer Science and Technology, Anhui University, Hefei, 230601, China
[2] CSE Department, University of Texas at Arlington, Arlington, TX 76019, USA
{jiangbo,luobin}@ahu.edu.cn, chqding@uta.edu

Abstract. Lasso-type variable selection has been increasingly adopted in many applications. In this paper, we propose a covariate-correlated Lasso that selects the covariates correlated more strongly with the response variable. We propose an efficient algorithm to solve this Lasso-type optimization and prove its convergence. Experiments on DNA gene expression data sets show that the selected covariates correlate more strongly with the response variable, and the residual values are decreased, indicating better covariate selection. The selected covariates lead to better classification performance.

1 Introduction

In many regression applications, there are too many unrelated predictors which may hide the relationship between the response and the most related predictors. A common way to resolve this problem is variable selection, that is to select a subset of the most representative or discriminative predictors from the input predictor set. In machine learning and data mining tasks, the main challenge of variable selection is to select a set of predictors, as small as possible, that help the classifier to accurately classify the learning examples. Various kinds of variable selection methods have been proposed to tackle the issue of high dimensionality. One major type of variable selection methods is to use the filter methods, such as: t-test, F-statistic [5], ReliefF [10], mRMR [12] and mutual information [13]. These methods are usually independent of classifiers. Another wrapper-type of variable selection methods is to take classifiers to evaluate subsets of predictors [9]. In addition, some stochastic search techniques have also been used for variable selection [16].

Recently, sparsity regularization receives increasing in variable selection. The well known Lasso (Least Absolute Shrinkage and Selection Operator) is a penalized least square method with ℓ_1-regularization, which is used to shrink/suppress variables to achieve variable selection [3,14,19,17,18]. However, ℓ_1-minimization algorithm is not stable compared with ℓ_2-minimization. Elastic Net added ℓ_2-regularization in Lasso to make the regression coefficients more stable [19]. Group Lasso was proposed where the covariates are assumed to be clustered in groups, and the sum of Euclidean norms of the loadings in each group is used [17]. Supervised Group Lasso performed K-means clustering before Group Lasso [11]. From the covariate point of view, the aim of traditional Lasso-type models is to select a set of covariates from the input covariate set that linearly represent the response approximately. However, they consider data approximation and representation only, without explicitly incorporating the correlation between the response and covariates.

T. Calders et al. (Eds.): ECML PKDD 2014, Part I, LNCS 8724, pp. 595–606, 2014.
© Springer-Verlag Berlin Heidelberg 2014

In this paper, correlation information is considered into the Lasso-type variable selection, where regression coefficients associated with larger correlations between the response and covariates are penalized less. Therefore, the selected covariates are highly correlated with the response, i.e., the response can be sparsely approximated (represented) by its closer covariates. In the following, we firstly briefly review the normal Lasso and Elastic Net, then present our covariate-correlated Lasso (ccLasso) model. An efficient iterative algorithm, with its proof of convergence, is presented to solve the proposed ccLasso optimization problem. Promising experimental results show the benefits of the proposed ccLasso model.

2 Brief Review of Lasso and Elastic Net

Let $(x_1, y_1), \cdots, (x_n, y_n)$ be the input data, where $x_i \in \mathcal{R}^p$ is a vector of predictors and $y_i \in \mathcal{R}$ is a scalar response for x_i. Formulate them in matrix form $X = (x_1, x_2, \cdots, x_n)^T \in \mathcal{R}^{n\times p}$ and $y = (y_1, y_2, \cdots, y_n) \in \mathcal{R}^n$. Here we adopt the language of LARS (covariate point of view) [6]. The j-th column of X (e.g., j-th dimension or feature throughout the n data points) is the j-th covariate, denoted as a column vector $a_j \in \mathcal{R}^n$. Let $A = (a_1, a_2, \cdots, a_p)$. The goal of Lasso is variable (covariate) selection. It selects a subset of $k < p$ covariates from the p covariates $a_1, \cdots, a_p$ (remember p is the dimension of x_i) that best approximate the response vector y. Lasso minimizes

$$\min_{\beta} \|y - \sum_{j=1}^{p} \beta_j a_j\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| = \|y - A\beta\|^2 + \lambda\|\beta\|_1, \tag{1}$$

Here ℓ_q-norm of vector v is defined as $\|v\|_q = \left[\sum_{i=1}^{n} |v_i|^q\right]^{1/q}$. For simplicity, we ignore the subscript 2 for the Euclidean distance $q = 2$: $\|v\| = \|v\|_2$. $\lambda \geq 0$ is a penalty parameter. When λ is large, many components of β are zero. The nonzero components give the selection of covariates. This covariate point of view is identical to the compressed sensing of Donoho et al[4].

In general, ℓ_1-minimization is not stable compared with ℓ_2-minimization [15]. To compensate for this, Elastic Net [19] further adds the ridge regression penalty term into Lasso objective function, which can be formulated as

$$\min_{\beta\in\mathcal{R}^p} \|y - A\beta\|^2 + \lambda\|\beta\|_1 + \zeta\|\beta\|^2, \tag{2}$$

where $\lambda, \zeta \geq 0$ are model parameters. Apart the sparsity, Elastic Net usually encourages a grouping effect, i.e., strongly correlated covariates tend to be in or out of the model together.

3 Covariate-Correlated Lasso

In this section, we present our covariate-correlated Lasso (ccLasso).

3.1 Covariate-Response Vector Correlation

First, we rescale the data. Note that in Lasso we can normalize y such that $\|y\| = 1$; This change is absorbed by β through an overall proportional constant. Second, we can also normalize each covariate a_j to $\|a_j\| = 1$. The difference is absorbed into β_j. Our covariate-correlated Lasso (ccLasso) is motivated by the following two observations.

First, since each covariate a_j has the same dimension as y, we may consider the correlation between y and covariate a_j. This is useful information. Intuitively, if we select a few covariates to form a linear combination that best approximates the response vector y, then the covariates correlated more with y would be good choices. In fact, this correlation information has been emphasized and successfully used in analysis of gene expression microarray data of gene selection [8]. To the best of our knowledge, this correlation information has not been explored or emphasized in Lasso-type covariate selection.

Then, we can prove that if we restrict β to have only one nonzero component, the selected covariate must be the covariate which correlates with y the most, i.e., the highest correlation coefficient w.r.t. y.

Lemma 1. *If we select one covariate among the p covariates, the selected one has the highest correlation coefficient with y.*

Proof. Selecting one covariate a_j that minimizes the error most is the following minimization problem,

$$\min_{j,\beta_j} J = \|y - \beta_j a_j\|^2 = y^T y + \beta_j^2 a_j^T a_j - 2\beta_j y^T a_j. \tag{3}$$

Since y and a_j are normalized, i.e., $y^T y = 1$, $a_j^T a_j = 1$ and y and a_j are already centered as in standard regression, the correlation coefficient is

$$\rho(y, a_j) = \frac{y^T a_j}{\|y\|\|a_j\|} = y^T a_j.$$

Thus $J = 1 + \beta_j^2 - 2\beta_j \rho(y, a_j)$. Setting the derivative w.r.t. β_j to zero, we obtain $\beta_j = \rho(y, a_j)$. Thus, $J = 1 - [\rho(y, a_j)]^2$ and the selection problem becomes

$$\min_j \ 1 - [\rho(y, a_j)]^2. \tag{4}$$

Therefore the selected one must has the highest (absoluate value) correlation coefficient with y. □

The above result is intuitively appealing: if we select one covariate to represent y approximately, the selected covariate must be the one closest (most correlated) to y. If we select two covariates to represent y, the standard LASSO results are not necessarily the two covariates most correlated to y. Our covariate-correlated Lasso (ccLasso) is motivated by the desire to encourage the selected covariates to correlate more with y.

3.2 Covariate-Correlated Lasso

By imposing the correlation information into the variable selection, our covariate-correlated Lasso can be formulated as follows,

$$\min_{\beta} \ \|y - A\beta\|^2 + \lambda \sum_{j=1}^{p} \mu_j |\beta_j|, \tag{5}$$

where

$$\mu_j = (1 - |\rho(y, a_j)|)^2. \tag{6}$$

The intuition is that when a_j correlates strongly (either positively or negatively) with y, μ_j is close to zero, thus a small penalty. As λ increases, β_j with large penalty tend to go to zero. Thus the final selected covariates tend to have larger correlation with y.

We now use α to replace β to denote/emphasize the regression coefficients obtained from ccLasso. Let $D = \mathrm{diag}(\mu_1, \cdots, \mu_p)$, then ccLasso can be written compactly as

$$\min_{\alpha} \ \|y - A\alpha\|^2 + \lambda \|D\alpha\|_1. \tag{7}$$

4 Computational Algorithm

4.1 Update Algorithm

Problem Eq.(7) is a convex formulation and we seek the global optimal solution. In this section, an efficient algorithm is derived to solve this problem. The detailed algorithm is given in Algorithm 1. In Algorithm 1, the initialization is the solution of the ridge regression problem

$$\min_{\alpha} \ \|y - A\alpha\|^2 + \lambda \alpha^T D\alpha. \tag{8}$$

The solution of this ridge regression problem is given by

$$\alpha^{(0)} = (A^T A + \lambda D/2)^{-1} A^T y. \tag{9}$$

4.2 Convergence Analysis

In this section, we provide a convergence analysis for Algorithm 1. Since $L(\alpha)$ is a convex function of α, thus, we only need to prove that the objective function value $L(\alpha)$ is non-increasing in each iteration in Algorithm 1. This is summarized in Theorem 1.

Theorem 1. *The objective function value $L(\alpha)$ of Eq.(7) for ccLasso minimization problem is non-increasing,*

$$L(\alpha^{t+1}) \leq L(\alpha^t), \tag{12}$$

upon the updating formulae Eq.(11) in Algorithm 1.

To prove Theorem 1, we need the help of the following two Lemmas, which are needed to be proved firstly.

Algorithm 1. Algorithm for covariate-correlated Lasso

1: **Input:** Training data $A \in \mathcal{R}^{n\times p}$ and corresponding response $y \in \mathcal{R}^n$, parameters λ, maximum number of iteration t_{max}, and convergence tolerance $\epsilon > 0$;
2: Compute $B = A^T A$, D, μ_j as in Eqs.(6,7)
3: Initialize $t = 0$, $\alpha^{(t)} = (A^T A + \lambda D/2)^{-1} A^T y$.
4: Update diagonal matrix

$$M^{(t)} = \text{diag}\left(\sqrt{|\alpha_1^{(t)}|}, \cdots, \sqrt{|\alpha_p^{(t)}|}\right); \tag{10}$$

5: Update combination coefficients

$$\alpha^{(t+1)} = M^{(t)}\left[M^{(t)} B M^{(t)} + \frac{\lambda}{2} D\right]^{-1} M^{(t)} A^T y; \tag{11}$$

6: If $t > t_{max}$ or $\|\alpha^{(t+1)} - \alpha^{(t)}\| < \epsilon$, go to step 7; otherwise, set $t = t + 1$ and go to step 4;
7: **Output:** The converged regression coefficients $\alpha^* = \alpha^{(t+1)}$.

Lemma 2. *Define an auxiliary function*

$$G(\alpha) = \|y - A\alpha\|^2 + \lambda \sum_{i=1}^{p} \frac{\alpha_i^2}{2|\alpha_i^{(t)}|} d_i. \tag{13}$$

Along with the $\{\alpha^{(t)}, t = 0, 1, 2, \cdots\}$ *sequence obtained in Algorithm 1, the following inequality holds,*

$$G(\alpha^{(t+1)}) \leq G(\alpha^{(t)}). \tag{14}$$

Proof. Since both two terms in auxiliary function $G(\alpha)$ are semi-definite programming (SDP) problems, we can obtain the global optimal solution of $G(\alpha)$ by taking the derivatives and let them equal to zero.

Making use of $M^{(t)}$ denotation in Eq.(10), the auxiliary function $G(\alpha)$ can be rewritten as

$$G(\alpha) = \|y - A\alpha\|^2 + \frac{\lambda}{2} \alpha^T (M^{(t)})^{-2} D\alpha. \tag{15}$$

Take the derivative of Eq.(15) with respect to α, and we get

$$\frac{\partial G(\alpha)}{\partial \alpha} = 2A^T A\alpha - 2A^T y + \lambda (M^{(t)})^{-2} D\alpha. \tag{16}$$

The second order derivatives are

$$\frac{\partial^2 G(\alpha)}{\partial \alpha_i \partial \alpha_j} = 2A^T A + \lambda (M^{(t)})^{-2} D. \tag{17}$$

This is clearly a positive semi-definite matrix. Thus function $G(\alpha)$ is a convex function and its global optimal solution α^* is unique. By setting $\frac{\partial G(\alpha)}{\partial \alpha} = 0$, we obtain

$$\alpha^* = \left[A^T A + \frac{\lambda}{2}(M^{(t)})^{-2} D\right]^{-1} A^T y \tag{18}$$

$$= M^{(t)} \left[M^{(t)} B M^{(t)} + \frac{\lambda}{2} D\right]^{-1} M^{(t)} A^T y. \tag{19}$$

The solution α^* is the global optima of $G(\alpha)$. Thus $G(\alpha^*) \leq G(\alpha)$ for any α. In particular, $G(\alpha^*) \leq G(\alpha^{(t)})$. Comparing Eq.(11) with Eq.(19), $\alpha^{(t+1)} = \alpha^*$. This completes the proof of Lemma 2.

Remark. It is important to note that we use Eq.(19) instead of the seemingly simpler Eq.(18). This is because as iteration progresses, some elements of $\alpha^{(t)}$ could become zero due to the sparsity of l_1-penalty. This causes the failure of the inverse of $M^{(t)}$ in Eq.(18). Thus Eq.(18) is ill-defined. However, $M^{(t)}$ is well-defined. Thus Eq.(19) is well-defined, which is chosen as the updating rule Eq.(11) in Algorithm 1.

Lemma 3. *The $\{\alpha^{(t)}, t = 0, 1, 2, \cdots\}$ sequence obtained by iteratively computing Eqs.(10,11) in Algorithm 1 has the following property*

$$L(\alpha^{(t+1)}) - L(\alpha^{(t)}) \leq G(\alpha^{(t+1)}) - G(\alpha^{(t)}). \tag{20}$$

Proof. Setting $\Delta = (L(\alpha^{(t+1)}) - L(\alpha^{(t)})) - (G(\alpha^{(t+1)}) - G(\alpha^{(t)}))$, substitute Eq.(7) and Eq.(13) in it, we obtain

$$\begin{aligned}
\Delta &= (\lambda \|D\alpha^{(t+1)}\|_1 - \lambda \|D\alpha^{(t)}\|_1) - \Big(\lambda \sum_{i=1}^{p} d_i \frac{(\alpha_i^{(t+1)})^2}{2|\alpha_i^{(t)}|} - \lambda \sum_{i=1}^{p} d_i \frac{(\alpha_i^{(t)})^2}{2|\alpha_i^{(t)}|}\Big) \\
&= -\frac{\lambda}{2} \sum_{i=1}^{p} \frac{d_i}{|\alpha_i^{(t)}|} \Big(-2|\alpha_i^{(t+1)}||\alpha_i^{(t)}| + 2|\alpha_i^{(t)}|^2 + (\alpha_i^{(t+1)})^2 - (\alpha_i^{(t)})^2 \Big) \\
&= -\frac{\lambda}{2} \sum_{i=1}^{p} \frac{d_i}{|\alpha_i^{(t)}|} \Big(|\alpha_i^{(t+1)}| - |\alpha_i^{(t)}| \Big)^2 \leq 0.
\end{aligned} \tag{21}$$

This completes the proof of Lemma 3.

Proof of Theorem 1. From Lemma 2 and Lemma 3, we have,

$$L(\alpha^{(t+1)}) - L(\alpha^{(t)}) \leq G(\alpha^{(t+1)}) - G(\alpha^{(t)}) \leq 0, \tag{22}$$

which is to say

$$L(\alpha^{(t+1)}) \leq L(\alpha^{(t)}). \tag{23}$$

This completes the proof of Theorem 1. Therefore, Algorithm 1 converges to the global optimal solution of ccLasso model starting from any initial coefficient $\alpha^{(0)}$, due to the convexity of optimization problem. Setting $d_i = 1$, the same algorithm can solve the standard Lasso problem.

5 Experiments

We evaluate the effectiveness of the proposed covariate-correlated Lasso (ccLasso) on the two well known data sets: Colon Cancer Data [1] and Leukemia Dataset [8]. The performance in variable selection and classification accuracy of the ccLasso will be compared with other methods. Once the variables are selected by our ccLasso method, the standard regression has been used to achieve classification [3].

5.1 Colon Cancer Data

The data is Affymetrix Oligonucleotide Array measurements of gene expression levels of 40 tumor and 22 normal colon tissues for 6500 human genes [1]. A subset of 2000 genes based on highest minimal intensity across the samples was selected[1]. These data were first preprocessed by taking a base 10 logarithmic of each expression level, and then each sample is centerized and normalized to zero mean and unit variance across the genes [3].

Classification Comparison. Because this dataset does not contain test set, we use the leave-one-out cross validation (LOOCV) method to evaluate the performance of the classification methods on a selected subset of genes [3]. The external LOOCV procedure is performed as follows: 1) remove one observation from the training set; 2) Select top 150 genes as ranked in terms of the t statistic; 3) Re-selected the k most important genes from the 150 genes by the proposed ccLasso algorithm; 4) Use these k genes to classify the left out sample. This process was repeated for all observations in the training set, and the average classification performance has been computed. Figure 1 shows the comparison results across different k genes selected out. Here, we can note that (1) the performances of all three methods are better as more genes are picked out for classification. (2) Lasso performs better than Elastic Net in this dataset. (3) The proposed ccLasso shows consistent superiority over the Lasso and Elastic Net. The best classification accuracy and its corresponding genes are summarized in Table 1. The proposed ccLasso is compared with the following classification methods: SVM [7], MAVE-LD [2], gsg-SSVS [16], Lasso [14] and Elastic Net [19]. It is clear demonstrated that the proposed ccLasso is better than the other popular classification methods using only moderate number of genes.

Table 1. Classification results on Colon Cancer Data

Method	No. of genes	LOOCV accuracy
SVM	1000 or 2000	0.9032
MAVE-LD	50	0.8387
gsg-SSVS	10/14	0.8871
Lasso	18	0.8316
Elastic Net	18	0.9510
ccLasso	18	0.9755

[1] `http://microarray.princeton.edu/oncology/affydata/`

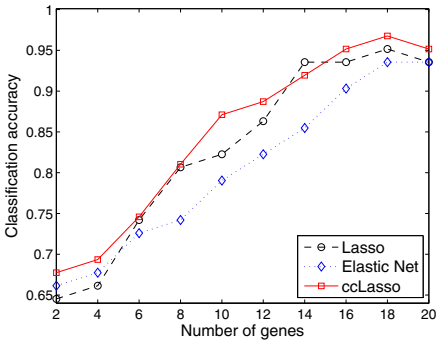


Fig. 1. External LOOCV classification accuracy of Lasso, Elastic Net and ccLasso on Colon Cancer Data

Average Correlation. As discussed in Lemma 1 in Section 3.1, we show that if we select only one covariate using Lasso model, this covariate must be the one with the highest correlation with y. From this, we expect that for small number of selected covariates, their average correlation with y will be high. But if we select larger number of covariates using Lasso, their average correlation with y will be smaller. In contrast, our ccLasso can select the large number of desired covariates that are highly correlated with the response y. To further illustrate these, we compute the average correlation coefficients between the selected covariates (genes) and y across different number of genes. Figure 2 (a) shows the comparison results. Here we can noted that for small number of selected genes, both Lasso and ccLasso can select the genes that are highly correlated with y. However, if we select large number of genes, the average correlation coefficients for ccLasso are clearly larger than that for Lasso model.

Residual Comparison. Both Lasso and ccLasso are the approximation models for solving the following problem

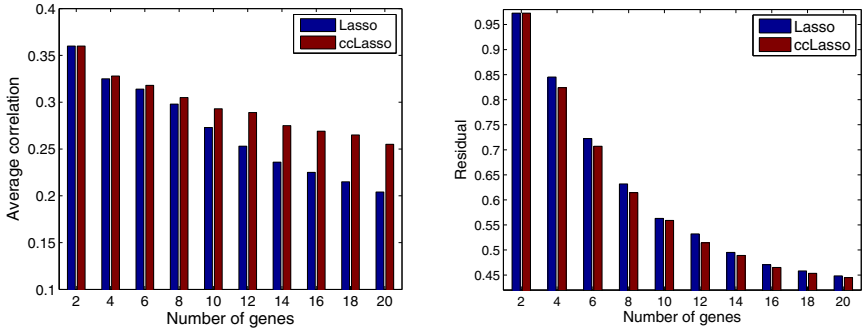
$$\min_{\beta} \|y - A\beta\|, \quad s.t. \quad \|\beta\|_0 = k. \tag{24}$$

In other words, we select a subset A_S of the covariates A with k entries (training samples) such that we achieve the best representation using A_S. This is a discrete selection problem and is well known to be is NP hard.

When using Lasso and ccLasso, this is done as follows,
(A0) Tuning the model parameter λ such that the optimal solution α contains k nonzero entries. (B0) Select the co-variates corresponding to the nonzero entries of α. This gives the subset A_S. (C0) Compute the optimal representation β by solving the linear regression problem,

$$J_{\text{residual-error}} = \min_{\beta} \|y - A_S\beta\|. \tag{25}$$

We compare the covariate subset A_S selected by ccLasso and Lasso, and then compute the residual errors. The results are shown in Figure 2 (b). It is clear that ccLasso selected



(a) Average correlation coefficients (b) Residual errors

Fig. 2. Average correlation between selected covariates and y and residual errors for Lasso and ccLasso on Colon dataset

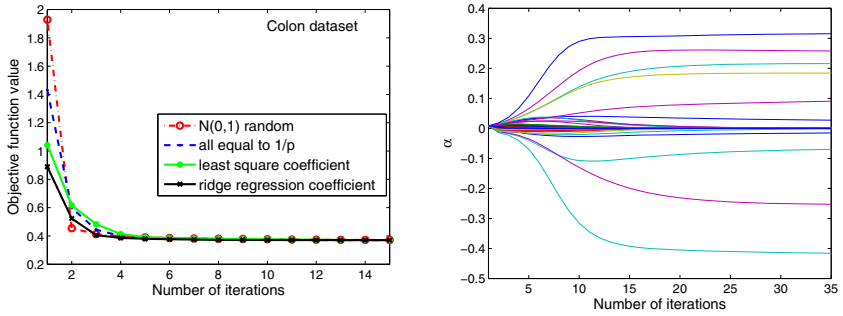


Fig. 3. LEFT: Objective function convergence with different initializations on Colon Dataset; RIGHT: Coefficient vector α during iterations (different colors denote different elements of α)

variables lead to lower residual error than Lasso. Thus, ccLasso model provides better approximate solutions for the discrete selection problem as compared to Lasso model.

Convergence of ccLasso. Figure 3 shows the variation of objective function across the iterations with different initializations in Algorithm 1. We can see that Algorithm 1 converges very quickly and the maximum iteration number is fewer than 30. Regardless of the initializations, the final objective function values are the same and converge almost at the same time, indicting the efficiency and effectiveness of the proposed ccLasso algorithm.

5.2 Leukemia Dataset

The leukaemia data contains DNA gene expressions of 72 tissue samples [8][2]. Following previous work, these tissue samples are divided into the training set of 38 samples

[2] `http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi`

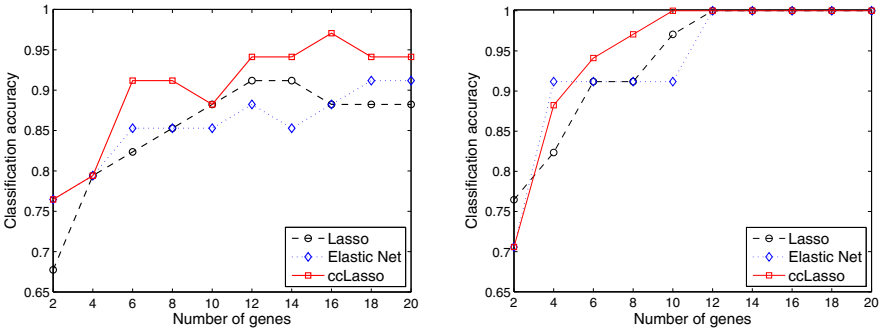
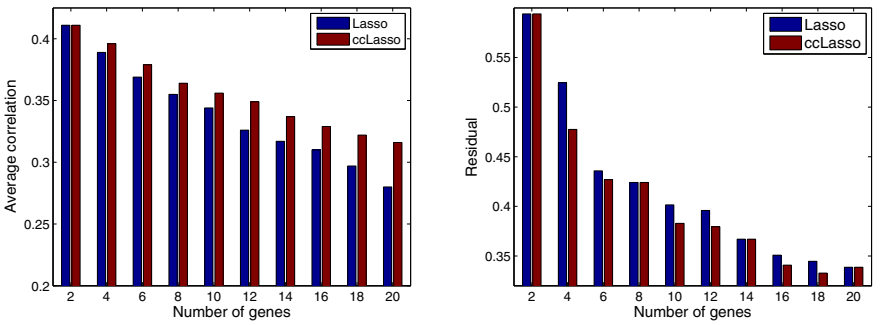


Fig. 4. Classification accuracy of Lasso, Elastic Net and ccLasso on Leukemia testing set (left) and training set (right)

Table 2. Classification results on Leukemia Dataset

Method	No. of genes	Training accuracy	Test accuracy
SVM	25∼2000	0.9474	0.8824∼0.9412
MAVE-LD	50	0.9737	0.9706
gsg-SSVS	14	0.9737	0.9706
Lasso	20	1.0000	0.8824
Elastic Net	20	1.0000	0.9118
ccLasso	20	1.0000	0.9412



(a) Average correlation coefficients (b) Residual errors

Fig. 5. Average correlation between selected covariates and y and residual errors for Lasso and ccLasso on Leukemia dataset

and the test set of 34 samples. The data gives expression levels of 7129 genes and DNA products. We use the preprocess method suggested by [3,5].Figure 4 shows the comparison classification accuracy results on training and testing sets across different number of genes, respectively. Here we can note that (1) all of the three methods perform well

on training set (all classify correct). (2) On test set, Elastic Net generally performs better than Lasso with the same number of genes selected out. (3) Our ccLasso consistently outperforms the other two methods. Table 2 summarizes the comparison results. Noted that the proposed ccLasso performs better than other methods with moderate number of genes. Figure 5 shows the average correlation coefficients and residual error results, respectively. Noted that as the number of selected genes increases, ccLasso model can return the genes that are more correlated with y (Fig. 5 (a)). Also it returns lower residual errors and thus approximates the discrete variable selection problem more closely than Lasso model (Fig. 5 (b)). Figure 6 shows the variation of objective function across the iterations with different initializations on this dataset. We can see that Algorithm 1 converges very quickly regardless of the different initializations. The above results are general consistent with that on Colon data, and further demonstrates the benefits of the proposed ccLasso.

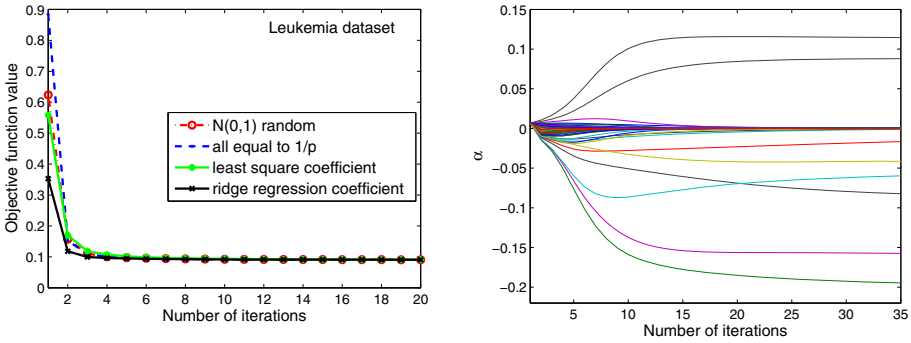


Fig. 6. LEFT: Objective function convergence with different initializations on Leukemia Dataset; RIGHT: Coefficient vector α during iterations (different colors denote different elements of α)

6 Conclusion

Covariate-correlated Lasso (ccLasso) naturally promotes correlation of the selected variable (covariate) with response y; this leads to smaller residual values, indicating a better solution to the discrete variable selection problem. The model achieves this with no extra parameters and same level of computation as standard Lasso. An efficient algorithm has been derived to solve ccLasso. Experiments on two well known gene datasets show that the proposed ccLasso consistently outperforms several state-of-the-art feature selection methods.

Acknowledgments. This work was supported in part by the National High Technology Research and Development Program of China (863 Program) under Grant 2014AA012204, and by the National Natural Science Foundation of China under Grant 61202228.

References

1. Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., Levine, A.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences 96(12), 6745–6750 (1999)
2. Antoniadis, A., Lambert-Lacroix, S., Leblanc, F.: Effective dimension reduction methods for tumor classification using gene expression data. Bioinformatics 19(5), 563–570 (2003)
3. Chen, S., Ding, C., Luo, B., Xie, Y.: Uncorrelated lasso. In: AAAI, pp. 166–172 (2013)
4. Donoho, D.: Compressed sensing. Technical Report, Stanford University (2006)
5. Dudoit, S., Fridlyand, J., Speed, T.P.: Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association 97(457), 77–87 (2002)
6. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. Annals of Statistics 32, 407–451 (2004)
7. Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M., Haussler, D.: Support vector machine classification and validation of cancer tissue samples using microarray expression data. BMC Bioinformatics 16(10), 906–914 (2000)
8. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., et al.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439), 531–537 (1999)
9. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artif. Intell. 97(1-2), 273–324 (1997)
10. Kononenko, I.: Estimating attributes: Analysis and extensions of RELIEF. In: Bergadano, F., De Raedt, L. (eds.) ECML 1994. LNCS, vol. 784, Springer, Heidelberg (1994)
11. Ma, S., Song, X., Huang, J.: Supervised group lasso with applications to microarray data analysis. BMC Bioinformatics 8 (2007)
12. Peng, H., Long, F., Ding, C.H.Q.: Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. 27(8), 1226–1238 (2005)
13. Raileanu, L.E., Stoffel, K.: Theoretical comparison between the gini index and information gain criteria. Ann. Math. Artif. Intell. 41(1), 77–93 (2004)
14. Tibshirani, R.: Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological) 58(1), 267–288 (1996)
15. Xu, H., Caramanis, C., Mannor, S.: Sparse algorithms are not stable: A no-free-lunch theorem. IEEE Trans. Pattern Anal. Mach. Intell. 34(1), 187–193 (2012)
16. Yang, A.J., Song, X.Y.: Bayesian variable selection for disease classification using gene expression data. Bioinformatics 26(2), 215–222 (2010)
17. 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)
18. Zou, H.: The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101, 1418–1429 (2006)
19. 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)