Lasso on Categorical Data

Yunjin Choi, Rina Park, Michael Seo

December 14, 2012

1 Introduction

In social science studies, the variables of interest are often categorical, such as race, gender, and nationality. However, it is often difficult to fit a linear model on such data, especially when some or all of the explanatory variables are categorical. When the response variable is the only categorical variable, it is common to use the logit model to overcome the defects of ordinary least squares. However, when the covariates are also categorical, corresponding variables are coded using dummy variables into our design matrix. In this approach, the data matrix becomes sparse, the column dimension increases, and columns might be highly correlated. This might result in a singular data matrix causing coefficients of Linear Square Estimation(LSE) impossible. In order to avoid this pitfall, researchers use Group Lasso(GL). Though GL has beneficial properties when dealing with categorical data, it is not devised specifically to analyze factor variables and there still remains room for improvement. In this project, we propose Modified Group Lasso(MGL) for improvements in categorical explanatory data. It performs better than Lasso or GL in various settings, particularly for large column dimension and big group sizes. Also MGL is robust to parameter selection and has less risk of being critically damaged by biased trial samples.

2 Background

2.1 Lasso

Lasso is a penalized linear regression. In linear regression, the underlying assumption is $E(Y|X = x) = x^T\beta$ (where, $Y \in \mathbb{R}^N$ and $X \in \mathbb{R}^{N\times p}$). While LSE minimizes $\frac{1}{2}\|Y - X\beta\|_2^2$ with respect to β, Lasso minimizes the penalty function $\frac{1}{2}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$. Since the additional L_1 penalty term is non-smooth, Lasso has variable selection property that can deal with the multicollinearity in data matrix. This property suggests that only the selected variables via the procedure will be included in the model. In this sense, Lasso is a proper method for factor data analysis, as it takes care of difficulties described above. However, the variable selection property of Lasso yields a new problem: partial selection of dummy variables. It is not reasonable to select only a portion of dummy variables derived from one categorical variable. Many researchers use Group Lasso to bypass this problem.

2.2 Group Lasso(GL)

Group Lasso performs similarly to Lasso except that it selects a group of variables rather than a single variable at each step of selection. The groups were pre-assigned on covariates. Therefore, in categorical case, we group the stack of dummy variables originated from a factor variable and each group represents a corresponding factor variable. This group selection property of GL has been facilitated by minimzing:

$$\frac{1}{2}\|Y - \sum_{l=1}^{L} X^{(l)}\beta^{(l)}\|_2^2 + \lambda \sum_{l=1}^{L} \sqrt{p_l}\|\beta^{(l)}\|_2 \tag{1}$$

Here $l \in \{1, 2, ..., L\}$ denotes the index of a group, p_l, the size of the l-th group, $X^{(l)}$, corresponding submatrix and $\beta^{(l)}$, the corresponding coefficient vector. The additional L_2 terms(not squared) in (1) takes the role of L_1 term in Lasso.

3 Modified Group Lasso for Categorical Data Matrix(MGL)

GL has been developed in order to select an assigned group of variables at each selection stage, regardless of the data type. However, GL is not specialized for the categorical/mix data case. Note that the value of (1) becomes bigger as $\sqrt{p_l}$ increases. As a consequence, GL tends to exclude groups with big dimensions. As for categorical data, all the variables in a same group basically represents one explanatory variable. Thus, favoring small groups can cause severe bias. For example, nationality variable which has more than 150 levels(US, Canada, Mexico, etc.,) and gender variable with two levels(Male and Female) should have equal chance of getting in the model when other conditions are the same. Our Modified Group Lasso gets rid of this effect by minimizing:

$$\frac{1}{2}\|Y - \sum_{l=1}^{L} X^{(l)}\beta^{(l)}\|_2^2 + \lambda \sum_{l=1}^{L} \|\beta^{(l)}\|_2. \tag{2}$$

Unlike (1), the additional L_2 terms in (2) are equally weighted. This approach ignores the size of groups in the selection procedure, and agrees with Ravikumar's sparse additive model fitting idea$^[Ravi]$: $\min\left(\frac{1}{2}\|Y - \sum_{l=1}^{L} f_l(X)\|_2^2 + \lambda \sum_{l=1}^{L} \|f_l(X)\|_2\right)$
One great advantage in this method is that we can still use the optimization algorithms in GL. GL and MGL have functionally the same form of target function to minimize and adaptive LARS can be adopted in both cases[3]. In this project, we implemented modified LARS[8].

4 Application

4.1 Simulation

We have applied MGL under various settings. In each case, N, the total number of observations is 200. Each row of the data sub-matrix follows multinomial distribution with equal probability when the corresponding group size is bigger than 1, otherwise, follows standard normal distribution. Thus, a multinomial distributed row of sub-matrix $X^{(l)}$ represents dummy variables of the categorical variables. Using this categorical data matrix X, the response vector Y was generated as $Y = X\beta + \epsilon = \sum_{l=1}^{L} X^{(l)}\beta(l) + \epsilon$. Here the coefficient vector β and the noise vector ϵ is adjusted to have Signal-to-Noise-Ratio=3. Also, 30 percent of β elements are set to be 0 to observe the model selection property.
Estimation of models were conducted by comparing two types of errors: estimation error:= $\|y - \hat{y}\|_2^2$

and coefficients error:= $\|\beta - \hat{\beta}\|_1$
In the figures below, black, red and green lines represent MGL, GL and Lasso respectively. The estimation error and coefficient error were provided in first row and second row respectively. In every graph, $x - axis$ is the constraint coefficient λ. Each column in the figures implies two error estimations from the same simulation run.

For the model fitting, following the convention, we choose the lambda which minimizes the estimation error. Thus to compare the performance of each models, it is reasonable to compare the minimum values over lambda. In this simulation, with respect to the estimation error, it seems that MGL surpasses other methods when the number of covariates is large and the size of groups is big. This coincides with our minimization criteria since when the size of groups are small, the distinction between MGL, GL, and Lasso would vanish. Except for several exceptions MGL performs better than GL ans Lasso.
In terms of coefficient error, it seems to be unstable and no method has dominance over others. This can be explained by the multicollinearity in the data matrix. As mentioned above, using dummy variables can introduce severe multicollinearity as the number of covariates and the size of p grow. When there exists multicollinearity in the data matrix,it is possible that there exists a linear model other than the true model which explains the data as good as the true one. This can explain the poor outcome in coefficient error while the estimation performance is nice.
One remarkable thing about MGL is that it is robust to the choice of λ. We observe from the figures that MGL has the smallest curvature in any case. As a consequence, even though we choose the wrong lambda in application, the fitted y would not be catastrophically deviated from the true y.

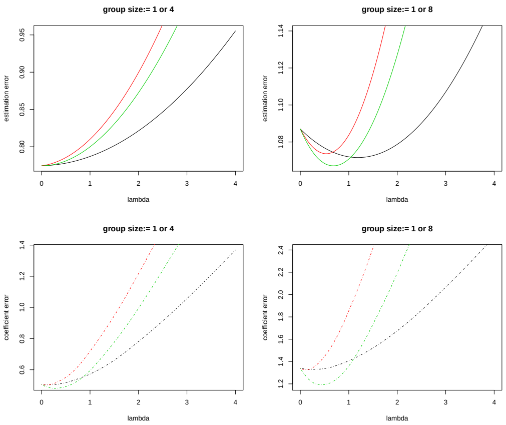

Figure 1: Number of covariates=2:from left to right, categorical variables with level 1)1 and 4, 2)1 and 8

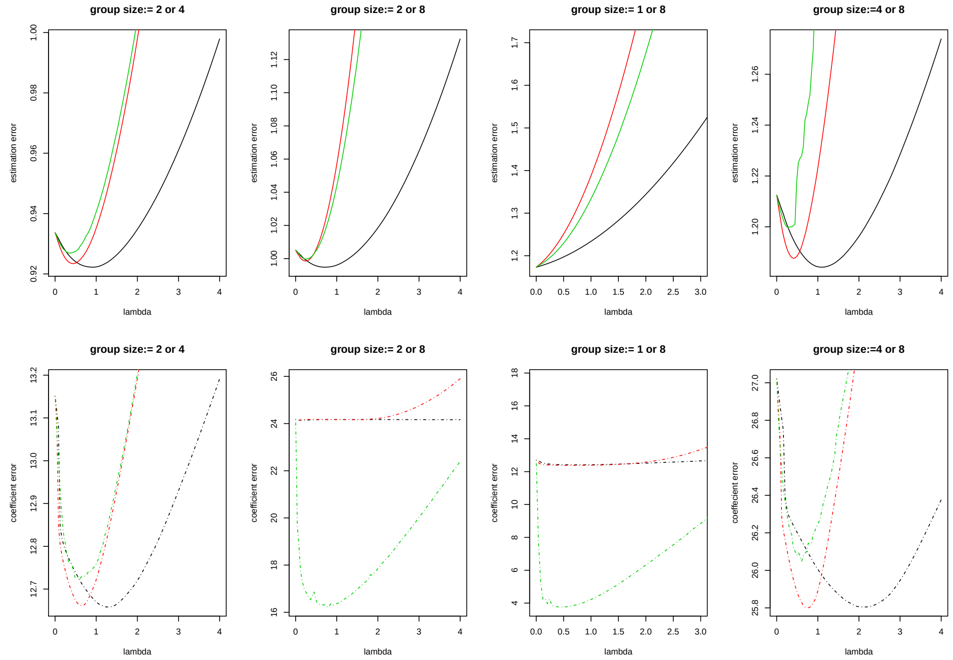

Figure 2: Number of covariates=4:from left to right, categorical variables with level 1) 2 and 4, 2) 2 and 8, 3) 1 and 8, 4) 4 and 8

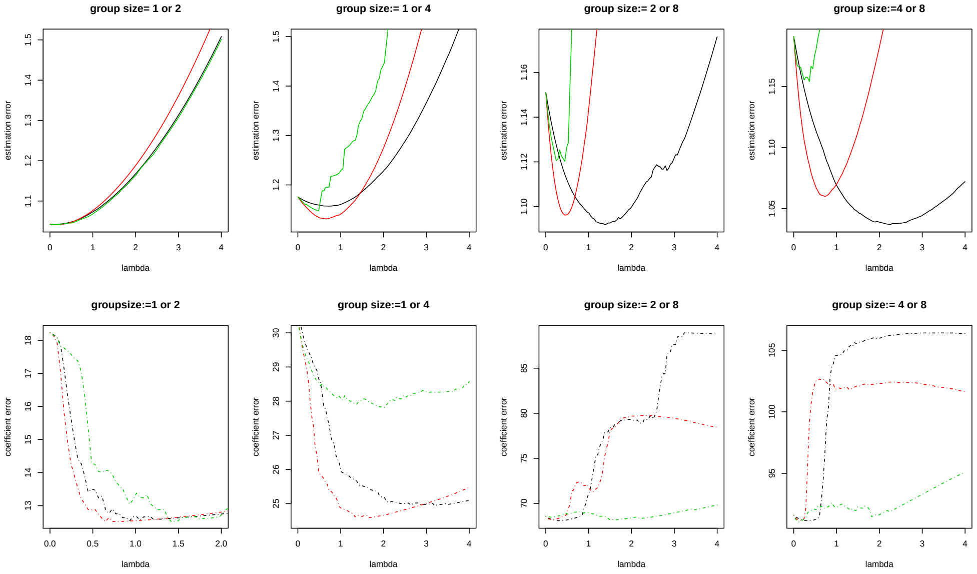

Figure 3: Number of covariates=4:from left to right, categorical variables with level 1) 1 and 2, 2) 1 and 4, 3) 2 and 8, 4) 4 and 8

4.2 Real Data

MGL, GL and Lasso are applied on the real mixed data in education. The data was collected from 1988 - 92 National Education Longitudinal Study (NELS). Data consists of 278 observations with 31 mixed types of covariates; such as gender and race for the factor variables, and units of math courses and average science score as quantitative variables. The response variable is the average math score.

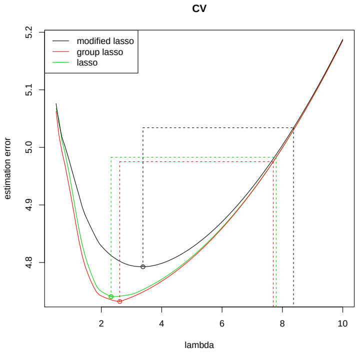

Figure 4: 20-fold Cross Validation on Educational Study Data for estimation error

Out of 278 data points, we chose 200 random points as a trial set and tested on the remaining. Figure 4 is a 20-fold Cross Validation estimation error result. Broken lines represent one standard deviation of the error for each method. The tuning parameter λs were chosen via one-standard-deviation rule following the convention of Lasso regression. The convention is to choose the biggest lambda within the range of one-standard-deviation from the minimum value in order to avoid overfitting and bias in the test data.

MGL achieves the smallest error on the test set, even though its CV estimation error is the worst. This can be a good example of robustness in MGL. On the test set, the estimation error ratio showed a significant improvement for MGL to Lasso, MGL to GL, and GL to Lasso, which was 0.947, 0.951 and 1.005 respectively. We can improve the performance on CV with more factor variables.

5 Conclusion and Discussion

Among previously introduced linear models, MGL performs relatively well. It has two main advantages: small estimation error and robustness to parameter selection in categorical data. With these advantages, we can apply to a wide academic realm dealing with categorical data, such as social science and education. Future research goal is to 1) analyze the case in which MGL has dominance both theoretically and empirically and 2) develop an analogous approach in logistic regression.

References

[1] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58:267288,1996.

[2] Park, M. and Hastie, T. Regularization path algorithms for detecting gene interactions. Available at http://www.stat.stanford.edu/ hastie/pub.htm, 2006.

[3] Yuan, M. and Lin, Y. Model Selection and Estimation in Regression with Grouped Variables. *Journal of Royal Statistical Society, Series B*, 68(1):49-67, 2007.

[4] Chesneau, Ch. and Hebiri, M. Some Theoretical Results on the Grouped Variables Lasso. *Mathematical Methods of Statistics*, 17:317-326, 2008.

[5] Freidman, J. H., Hastie, T. and Tibshirani, R. Regularized paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 2008.

[6] Wang, H. and Leng, C. A note on adaptive group lasso. *Computational Statistics and Data Analysis*, 2008.

[7] Zou, H. and Hastie, T. Regularization and Variable Selection via the Elastic Net. *Journal of Royal Statistical Society, Series B*, 67(2):301-320, 2008.

[8] Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. Sparse additive models. *Journals of the Royal Statistical Society: Series B*, 71(5): 10091030, 2009. ISSN 14679868.

[9] Kim, S. and Xing, E. Tree-Guided Group Lasso for Multi-Task Regression with Structured Sparsity. *Proceedings of the 27th International Conference on Machine Learning*, 2010.

[10] Simon, N. and Tibshirani, R. Standardization and the Group Lasso Penalty. *Statistica Sinica*, 22:983-1001, 2012