A Sparse Group LASSO

Saihiel Bakshi

University of Toronto Mississauga

March 31, 2020

Outline

Outline

Background

- In Supervised Learning, both predictors (input) and response (output) variables are observed
  - Want a model to understand/predict the relationship between the two
  - When huge amount of predictors are present all variables are not equally important to this relationship
- Variable selection is the process of selecting subsets of predictors with the most predictive power
  - In high dimensional settings, this is crucial
  - Increases interpretability of models
  - Need to find a balance between model's generalizability and computational costs
- Variable selection through shrinkage is preferred to automatic subset selection methods
  - Shrinkage methods are more stable
  - Continuous process, and less variable as a result

Regularization and Sparsity

- Regularization methods control model complexity by shrinking coefficient estimates
  - Penalize the model for have large coefficients
  - Coefficients shrink towards 0
- Sparsity regularization selects the input variables that best describe the response by setting coefficients to exactly 0
- In high-dimensional learning exploiting problem specific assumptions can lead to higher accuracy
- Structured sparsity regularization uses prior assumptions of predictors, such as groupings, to select optimal parameters

Outline

Data and Notation

- Data: $(X_i, Y_i), i = 1, \ldots, n$
- Y is the response vector of size n
- **X** is an n by p feature matrix
  - n is the sample size
  - p is number of predictors
  - $p >> n$, a case where standard linear regression fails
  - Note: we fit the model without an intercept and standardize the inputs before applying shrinkage methods
- Further, **X** is divided into m different groups
  - Such as, factor level indicators in categorical data
  - $X^{(\ell)}$ is the submatrix of **X** with columns corresponding to the predictors in group ℓ
- β is the coefficient vector, and $\beta^{(\ell)}$is the coefficient vector of group ℓ
  - p_ℓ is the length of $\beta^{(\ell)}$

Grouped Data

In our feature matrix $\mathbf{X}$

- Non-overlapping groups within feature matrix
- Simply apply shrinkage methods on such data is not very useful
- Still need to use all the input features

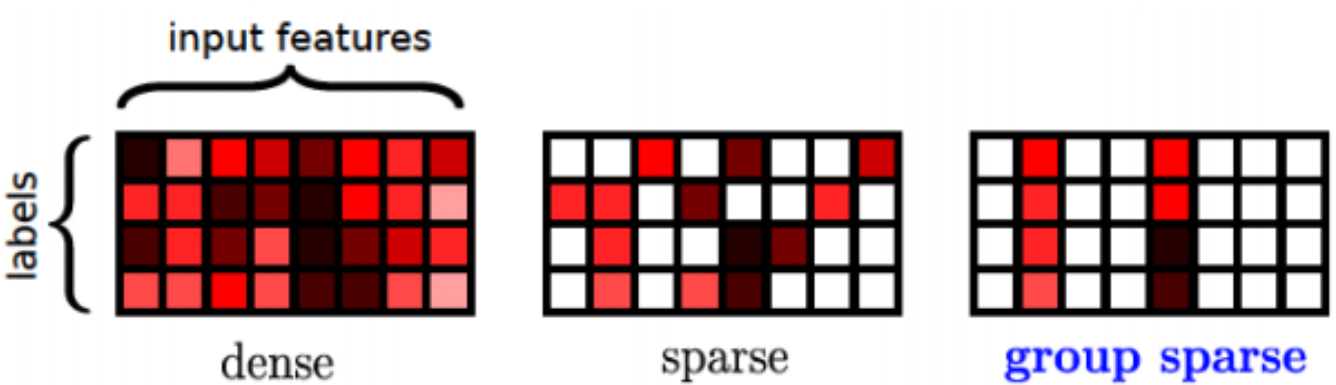


Figure 1: Visualizing Sparsity in Grouped Data

Outline

Motivation

- Standard L_1 regularization (LASSO) cannot yield sparsity at a group level
- Applying a variation of LASSO with a Euclidean norm penalty gives group lasso
  - The group lasso does not, however, yield sparsity within a group
  - That is, if a group of parameters is non-zero, they will all non-zero

A Sparse-Group Lasso

A more general penalty that yields sparsity at both the group and individual feature levels, in order to select optimal groups and predictors within a group.

Outline

L_1 Regularization Penalty

- When $p >> n$ linear regression lacks a unique solution
- Tibshirani (1996) solved this problem by bounding the l_1 norm of the solution.
- This approach, known as Lasso, minimizes:

$$\frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

- Lasso finds a solution with few nonzero entries in β
- Since this is the sum of convex problems, this is still a convex optimization problem
  - When $p >> n$ lasso selects at most n variables before saturating
  - When our data is grouped (high correlation among predictors), lasso does not apply sparsity to entire groups.

Group Lasso

- In 2007, Yuan & Lin propose the group lasso which solves the convex optimization problem:

$$\min_\beta \frac{1}{2}\left\| y - \sum_{l=1}^{m} X^{(l)}\beta^{(l)} \right\|_2^2 + \lambda \sum_{l=1}^{m} \sqrt{p_l}\left\|\beta^{(l)}\right\|_2$$

- This criterion exploits the non-differentiability of $\left\|\beta^{(l)}\right\|_2$ at $\beta^{(l)} = 0$
- Thus setting groups of coefficients to exactly 0
- Like in lasso, tuning parameter λ controls the sparsity of the solution.
  - When the size of each group is 1, the solution is same as lasso

Group Lasso

- The group lasso is able to give solutions with sparse sets of groups
- At a group level, this method acts like lasso
- Entire groups of predictors may drop out of the model
- This can create problems:
  - The method does not yield sparsity within a group
  - If a group of parameters is non-zero, they will all be non-zero
  - Additionally, Yuan & Lin's algorithm assumed that submatrices in each group are orthonormal
  - If the predictors are not orthonormal, one approach is to othonormalize them before apply group lasso
    - Generally, this will not provide a solution to the original problem
  - If predictors are orthonormal, then convergence is not guaranteed for their proposed method

Outline

A Sparse-Group Lasso

- Sparse-Group lasso gives sparse solutions at both a group level ("groupwise sparsity") and within groups ("within group sparsity")

Friedman et al proposed the criteria:

$$\min_\beta \frac{1}{2n}\left\| y - \sum_{l=1}^{m} X^{(l)}\beta^{(l)} \right\|_2^2 + (1-\alpha)\lambda \sum_{l=1}^{m} \sqrt{p_l}\left\|\beta^{(l)}\right\|_2 + \alpha\lambda\left\|\beta^{(l)}\right\|_1$$

where $\alpha \in [0, 1]$

- This optimization problem is a convex combination of the lasso and group lasso penalties
  - $\alpha = 0$ gives the group lasso fit
  - $\alpha = 1$ gives the standard lasso fit
- While this seems to be similar to the elastic net penalty, it differs as the $\left\|\beta^{(l)}\right\|_2$ penalty is not differentiable at $\mathbf{0}$
  - As a result some groups can be completely zeroed out
  - Within groups, the fit is the same as elastic net

A Sparse-Group Lasso

The objective function defined above is convex, so the optimal solution is characterized by the subgradient equations.
For a group k, $\hat{\beta}^{(k)}$ satisfies

$$\frac{1}{n} X^{(k)T} \left(y - \sum_{l=1}^{m} X^{(l)} \hat{\beta}^{(l)} \right) = (1-\alpha)\lambda\mu + \alpha\lambda\upsilon$$

where μ and υ are subgradients of $\|\hat{\beta}^{(k)}\|_2$ and $\|\hat{\beta}^{(k)}\|_1$ respectively, evaluated at $\hat{\beta}^{(k)}$. Then, it can be shown that the subgradient equations can be satisfied with $\hat{\beta}^{(k)} = 0$ if

$$\|S(X^{(k)T} r(-k)/n, \alpha\lambda)\|_2 \leq (1-\alpha)\lambda$$

where $r(-k)$ is the partial residual of y, subtracted from all other group fits except group k; and $S(\cdot)$ is the coordinate-wise soft thresholding operator.

Methodology

Friedman et al. proposed *pathwise coordinate gradient descent*, using accelerated generalized descent with backtracking within each group.

- Blockwise Coordinate Gradient Descent is guaranteed to converge to the global optimum
  - The criteria is the sum of a convex differential function (the loss), and a separable penalty (between groups)
- Instead of fixing the regularization parameter implement a Pathwise solution for various λ values
- Implements Nesterov's Momentum and Backtracking in generalized descent:
  - Nesterov's momentum smooths updates by taking a step in the direction of the previous accumulated gradient then corrects the velocity based on the step
  - Backtracking is a step size optimization method that maximizes the step size in the direction of steepest descent

Algorithm

Algorithm to fit a Spare-Group Lasso:

1. **[Outer loop]** Cyclically iterate over the groups; at each group k to minimize over, consider the other group coefficients as fixed
2. Check if the group's coefficients are exactly 0. If not, enter inner loop
3. **[Inner loop]** Start with $\beta^{(k,l)} = \theta^{(k,l)} = \beta_0^{(k)}$, step size $t = 1$, and counter $l = 1$. Unit convergence repeat:
   1. Update the gradient g by $g = \nabla l(r_{(-k)}, \beta^{(k,l)})$
   2. Optimize step size by iterating $t = 0.8 * t$ until
      $$l(U(\beta^{(k,l)}, t)) \leq l(\beta^{(k,l)}) + g^T \Delta_{(l,t)} + \frac{1}{2t}\|\Delta_{(l,t)}\|_2^2$$
   3. Update $\theta^{(k,l)}$ by:
      $$\theta^{(k,l+1)} \leftarrow U(\beta^{(k,l)}, t)$$
   4. Update the center via a Nesterov step by
      $$\beta^{(k,l+1)} \leftarrow \theta^{(k,l)} + \frac{l}{l+3}(\theta^{(k,l+1)} - \theta^{(k,l)})$$
   5. Set $l = l + 1$

   Note: $U(\beta_0, t)$ is the update rule, $\Delta_{(l,t)} = U(\beta^{(k,l)}, t) - \beta^{(k,l)}$

Pathwise Solutions

Iteratively fitting models over a grid of α and λ values is computationally impractical. Instead fix the mixing parameter α and compute solutions for a path of λ values using warm starts.

- Start with large values of λ to set $\hat{\beta} = 0$ and decrease λ from there
- By using the previous solution for the algorithm at the next λ value along the path, the method is made efficient
- Since α is fixed, the objective is a piecewise quadratic in λ
- Find the smallest λ_l for each group that sets that group's coefficients to 0
  - Thus begin the path search with:
    $$\lambda^{\text{max}} = \text{max}_i \lambda_i$$
  - The exact value at which the first coefficient enters the model.
  - Set λ^{min} to be a small fraction of λ^{max} [default 0.1]
- Optimal α value is problem specific

A Sparse-Group Lasso

Why pick sparse-group lasso over existing methods?

- Enables sparsity at both group level as well as predictor level
- Improves interpretability beyond existing methods such as lasso and group-lasso

A Sparse-Group Lasso

Why is the proposed algorithm significant?

- Guarantees convergence to global optima
- Enables optimal tuning of regularization hyperparameter efficiently
- Can be used to fit group-lasso with non-orthonormal predictors

A Sparse-Group Lasso

When is this method useful?

- Data with many predictors and large number of levels per predictor
- It is likely that even for informative predictors, many of the levels may not be informative
- Spare-group lasso considers this, setting coefficients for many levels to 0, even in nonzero groups

Outline

Real Data Analysis

- Regression is often run in a setting where the predictors have a natural grouping
- One such case is gene pathway research:
  - In many genetic conditions, genes do not function independently
  - In one pathway, if all genes seem moderately good at predicting outcomes, this evidence should be up-weighted over similarly predictive genes in different pathways
  - But every gene in an active pathway is not necessarily indicated in the genetic condition
  - The objective is to find pathways of interest and then select driving genes from them.
- To investigate this, sparse-group lasso, group-lasso, and lasso are compared on a real data example with gene expression data

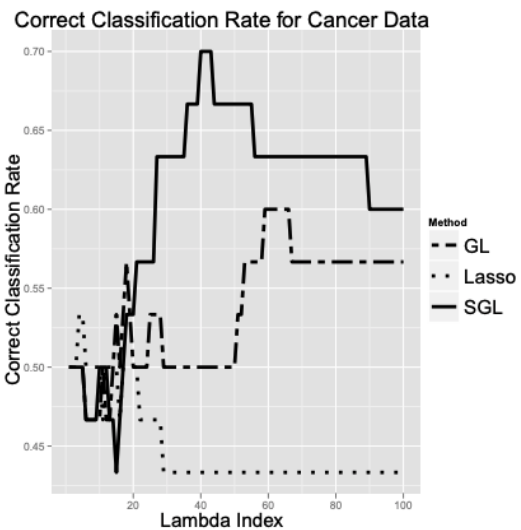
Real Data Analysis

Breast Cancer data of Ma et at. (2004):

60 total patients [28 cancer recurrences, 32 non-recurrences]

270 genetic bands containing $\sim$18.5 genes each for every patient

- The 3 models were fit on 30 randomly sampled patients
- Each model was fit with 100 λ values



Method	Accuracy	# Genes	# Bands
Lasso	53%	3	270
Sparse	70%	54	11
Group	60%	74	15

Table 1: Results for the 3 models

Figure 2: Test accuracy

Simulation Procedure

Comparison of the performance of lasso and sparse-group lasso for variable selection on simulated data. The covariate matrix for X was simulated with different numbers of covariates, observations, and groups.

- The columns of X were iid. Gaussian, and the response, y was constructed as:

$$y = \sum_{l=1}^{g} X^{(l)} \beta^{(l)} + \sigma\epsilon$$

  where $\epsilon \sim (0, I), \beta^{(l)} = (1, 2, \ldots, 5, 0, \ldots, 0)$ for $l = 1, \ldots, g$, and σ was set so the signal to noise ratio was 2
- The number of generative groups, g varied from 1 to 3 changing the amount of the sparsity
- The penalty parameters were chosen for both the lasso and sparse-group lasso (with $\alpha = 0.95$) so the number of nonzero coefficients chosen in the fits matched the true number (5, 10, or 15 corresponding to g = 1,2,3)

Simulation Results I

The proportion of correctly identified covariates averaged over 10 trials:

	Number of Groups in Generative Model		
	$n = 60,\ p = 1500,\ m = 10$		
	1 group	2 groups	3 groups
SGL	0.72	0.36	0.28
Lasso	0.60	0.38	0.31

Table 2: Results of simulation I

Simulation Results II

The proportion of correctly identified covariates averaged over 10 trials:

	Number of Groups in Generative Model		
	$n = 70$, $p = 2000$, $m = 200$		
	1 group	2 groups	3 groups
SGL	0.68	0.44	0.31
Lasso	0.54	0.30	0.26

Table 3: Results of simulation II

Simulation Results III

The proportion of correctly identified covariates averaged over 10 trials:

	Number of Groups in Generative Model		
	$n = 150$, $p = 10000$, $m = 100$		
	1 group	2 groups	3 groups
SGL	0.77	0.72	0.52
Lasso	0.76	0.62	0.43

Table 4: Results of simulation III

Simulation Results IV

The proportion of correctly identified covariates averaged over 10 trials:

	Number of Groups in Generative Model		
	$n = 200$, $p = 20000$, $m = 400$		
	1 group	2 groups	3 groups
SGL	0.92	0.78	0.68
Lasso	0.82	0.68	0.52

Table 5: Results of simulation IV

Outline

Conclusion

- In high-dimensional data exploiting problem specific assumptions can lead to higher accuracy as well as faster computations
- Structured sparsity regularization uses prior assumptions of predictors, such as groupings, to select optimal parameters
- A sparse-group lasso exploits grouping within parameter space to yield groupwise and within group sparsity in regression models

Thank you!