Literature Review

Sparse regression: Scalable algorithms and empirical performance

Authors: Dimitris Bertsimas, Jean Pauphilet and Bart Van Parys

Reviewed by: Long Ngo

Table of Contents

Introduction

Features selection plays a significant role in data science. Top reasons can be mentioned are that it enhances accuracy of a model, reduce overfitting and obtain a less complexity model which lessens computational expense. Least absolute shrinkage and selection operator (Lasso) method utilizing L1 norm for penalty is widely known and used because of its practical success. However, recent researches pointed out that under certain conditions of data, Lasso fails to select true set of features.The research provides a comprehensive assessment of feature selection methods in various regimes of noise and correlation and presents alternatives to the Lasso-regularization which are not as popular in practice yet.

Methods and metrics

1. Optimization problem

The optimization problem used in this research is minimizing the loss function of model and response data.

$$\min_{w \in \mathbb{R}^p} \sum_{i=1}^{n} \ell(y_i, w^\top x_i)$$

Where:

- ℓ is an appropriate loss function
- Covariates matrix X (features)
- Coefficient matrix W
- Response data Y

In features selection, we constrain the number of features used for prediction:

$$\min_{w \in \mathbb{R}^p} \sum_{i=1}^{n} \ell(y_i, w^\top x_i) \text{ s.t. } \|w\|_0 \leqslant k$$

or the relaxation:

$$\min_{w \in \mathbb{R}^p} \sum_{i=1}^{n} \ell(y_i, w^\top x_i) + \lambda \|w\|_0, \quad (1)$$

where$\| w \|_0$is number of non-zero coefficients of W

2. Methods

There are five methods that are introduced in the research: the convex integer optimization formulation, its Boolean relaxation, l1 regularization and two methods with non-convex penalties.

2.1. Lasso – l1 relaxation

Instead of solving the NP-hard problem (1), Lasso replaces the non-convex l0-pseudo norm by the convex l1-norm:

$$\min_{w \in \mathbb{R}^p} \sum_{i=1}^{n} \ell(y_i, w^\top x_i) + \lambda \|w\|_1$$

Lasso relies on a very important assumption about data which is mutual incoherence. The impact of this assumption to performance of Lasso will be described in Result section of this review.

2.2. Non-convex penalties

Recently, other formulations have been proposed, of the form:

$$\min_{w\in\mathbb{R}^p} \sum_{i=1}^{n} \ell(y_i, w^\top x_i) + \sum_{j=1}^{p} p_{\lambda,\gamma}(|w_j|)$$

where $\rho_{\lambda\gamma}(.)$ is penalty function parametrized by λ and γ.

2.2.1. Minimax Concave Penalty (MCP)

The minimax concave penalty is defined by:

$$p_{\lambda,\gamma}(u) = \lambda \int_0^u \left(1 - \frac{t}{\gamma\lambda}\right)_+ \mathrm{d}t = \begin{cases} \lambda u - \dfrac{u^2}{2\gamma} & \text{if } u \leq \gamma\lambda. \\ \dfrac{\gamma\lambda^2}{2} & \text{if } u > \gamma\lambda. \end{cases}$$

for $\lambda \geq 0$ and $\gamma > 1$.

2.2.2. Smoothly Clipped Absolute Deviation (SCAD)

The smoothly clipped absolute deviation penalty isdefined by:

$$p_{\lambda,\gamma}(u) = \begin{cases} \lambda u & \text{if } u \leq \lambda \\ \dfrac{\gamma\lambda u - (u^2 + \lambda^2)/2}{\gamma - 1} & \text{if } \lambda < u \leq \gamma\lambda, \\ \dfrac{\lambda^2(\gamma^2 - 1)}{2(\gamma - 1)} & \text{if } u > \gamma\lambda, \end{cases}$$

for $\lambda \geq 0$ and $\gamma > 2$.

2.3. Convex integer optimization formulation (CIO)

The integer optimization formulation considers an l2-regularized version of the initial formulation:

$$\min_{w\in\mathbb{R}^p} \sum_{i=1}^{n} \ell(y_i, w^\top x_i) + \frac{1}{2\gamma}\|w\|_2^2 \text{ s.t. } \|w\|_0 \leqslant k.$$

Then a binary variable to $s \in \{0,1\}$ is introducedto encode the support of W:

$$\min_{s\in\{0,1\}^p : s^\top \boldsymbol{e} \leqslant k} \ \max_{\alpha\in\mathbb{R}^n} \ f(\alpha, s) := \left(-\sum_{i=1}^{n} \hat{\ell}(y_i, \alpha_i) - \frac{\gamma}{2}\sum_{j=1}^{p} s_j \alpha^\top X_j X_j^\top \alpha \right) \quad (2)$$

where $\hat{\ell}(y, \alpha) := \max_{u\in \mathrm{R}} u\alpha - \ell(y, u)$ is the Fenchel conjugate of the loss function ℓ.

In special case of OLS, the function f is a quadratic function of α, the inner maximization problem can be solved in closed form:

$$\max_{\alpha} f(\alpha, s) = \frac{1}{2} Y^\top (I_n + \gamma X_s X_s^\top)^{-1} Y.$$

Then cutting-plane algorithm is used to solve this convex integer optimization problem.

2.4. Boolean relaxation (BR)

The Boolean relaxation method considers relaxation of the binary variable s of the problem (2):

$$\min_{s\in[0,1]^p \,:\, \mathbf{e}^\top s \leqslant k} \ \max_{\alpha\in\mathbb{R}^n} f(\alpha, s)$$

Authors then solved this Boolean relaxation problem by using dual sub-gradient algorithm.

3. Metrics

In order to compare the performance of all methods, the research presents three metrics: accuracy, false detection rate and computational tractability.

Accuracy (A) is defined:

$$A(w) := \frac{|\{j : w_j \neq 0, \ w_{true,j} \neq 0\}|}{|\{j : w_{true,j} \neq 0\}|}$$

False Detection Rate (FDR) is defined:

$$FDR(w) := \frac{|\{j : w_j \neq 0, \ w_{true,j} = 0\}|}{|\{j : w_j \neq 0\}|}$$

Computational tractability is measured by computational time for each algorithm in relative to the time needed to compute a Lasso estimator and real computational time needed to compute each method presented in the research.

Result

The research provides an exhaustive assessment on the aforementioned methods which covers comparing these methods on synthetic data sets for both linear regression and classification problems. The authors the implications of the feature selection methods in terms of induced sparsity and prediction accuracy on a real data set from genomics. However, in the scope of this review, only results of linear regression on synthetic data are discussed. Similar behaviors of the methods are observed for synthetic and empirical data of the classification problem.

The authors first simulate synthetic data satisfying mutual coherence condition, which is required by required by l1-regularized estimators to be statistically consistent. The performances of the five methods are compared in six different regimes of noise and correlation: low, medium and high noise; low and high correlation. The accuracies of the methods are described in Fig 1 of Appendix 1 of this review, which shows that accuracies smoothly converge to 1 as sample size n increases to infinity. Noise and correlation have different impact on performance of the method. While noise hinders all method performance and reduces the gap in performance between methods, high correlation strongly decreases the performance of Lasso, moderately those of SCAD and very slightly CIO, BR and MCP methods. Among all methods, l1-regularization is the less accurate, selects fewer correct features than the four other methods and is sensitive to correlation between features. In terms of False Detection Rate (FDR), MCP and CIO are the best method when they achieve the lowest FDR while Lasso persistently returns around 80% of incorrect features regardless regime of noise and correlation. Fig 3 in Appendix 1 reports relative computational time compared to Lasso in log scale. Unsurprisingly, CIO and BR are the most computer expensive ones, which terminate in times of two orders of magnitude larger than Lasso. All things considered, CIO and the MCP penalization are the best performing method in all six regimes, with a fine advantage for CIO.

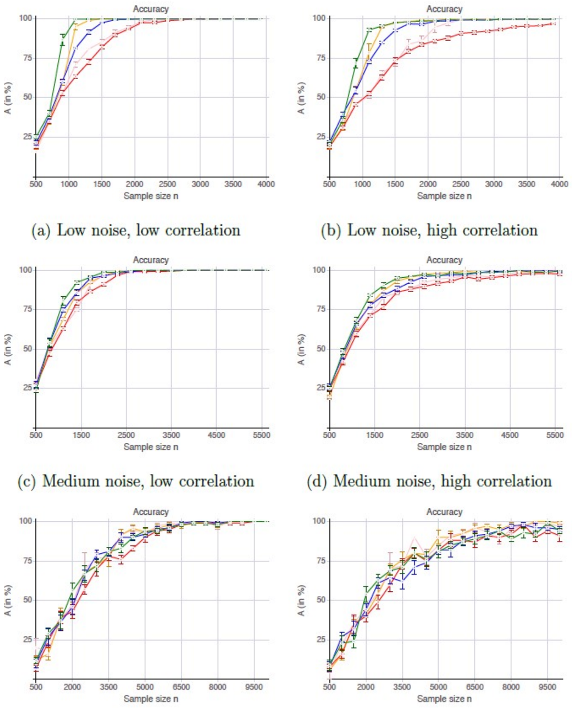
The second synthetic data implemented by the researcher is synthetic data not satisfying mutual coherence condition. A "hard" correlation structure is considered, i.e., a setting where the standard Lasso estimator is inconsistent. Therefore, only three regimes data are simulated under three levels of noise: low, medium and high noise. As observed from Fig 1 in Appendix 2, the accuracy of Lasso reach a threshold strictly lower than 1, meaning Lasso with this type of data is not able to recover a true support of features. On the other hand, the other methods experience their accuracies converges to 1 as sample size n increases to infinity. As far as accuracy is concerned, CIO outperforms the other methods. Similar patterns of FDR and computational tractability to the case where the mutual incoherence condition holds are found. Lasso as expected is the method has the highest FDR among all five while CIO and BR demonstrates the longest times of termination.

Conclusion and final project approach

A proper feature selection method should recover all and nothing but true features, as the sample size increases. When mutual incoherence condition is satisfied, all five methods attain perfect accuracy. However, when mutual incoherence fails to hold, l1-regularized estimators do not recover all true features. Lasso is the least accurate and sensitive to correlation between features; while integer optimization formulation, Boolean relaxation and the MCP non-convex estimator are the most accurate. The integer optimization formulation is the only method in the study which clearly out performs all other methods with the lowest false detection rate. Though computationally expensive, the computational cost of integer optimization is only one or two orders of magnitude higher than other alternatives and remains affordable in many real-world problems, even high-dimensional ones.

Based on progressive results that this research paper provides, I then seek to apply Lasso, one method of convex integer formulation and one method of non-convex penalty on the handwritten digits MNIST data set and give a comprehensive assessment on performance of these methods in predicting real-world handwritten data.

Appendix 1: Accuracy, False Detection Rate and computational time when synthetic data satisfying mutual incoherence condition



(a) Low noise, low correlation

(b) Low noise, high correlation

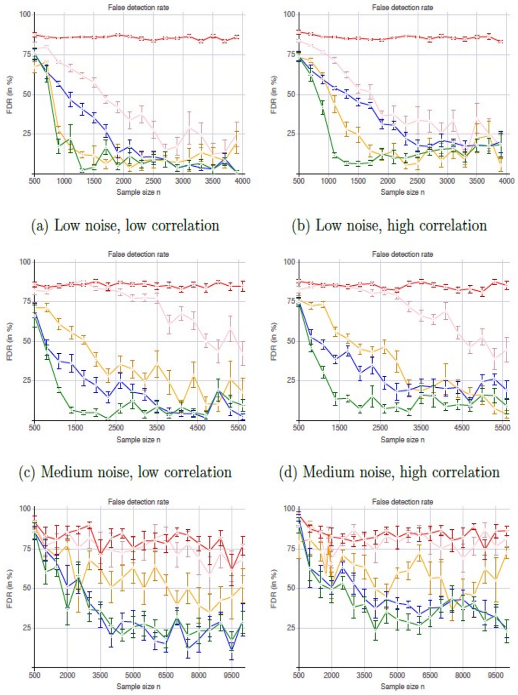
(c) Medium noise, low correlation

(d) Medium noise, high correlation

(e) High noise, low correlation

(f) High noise, high correlation

Fig 1: Accuracy



(a) Low noise, low correlation

(b) Low noise, high correlation

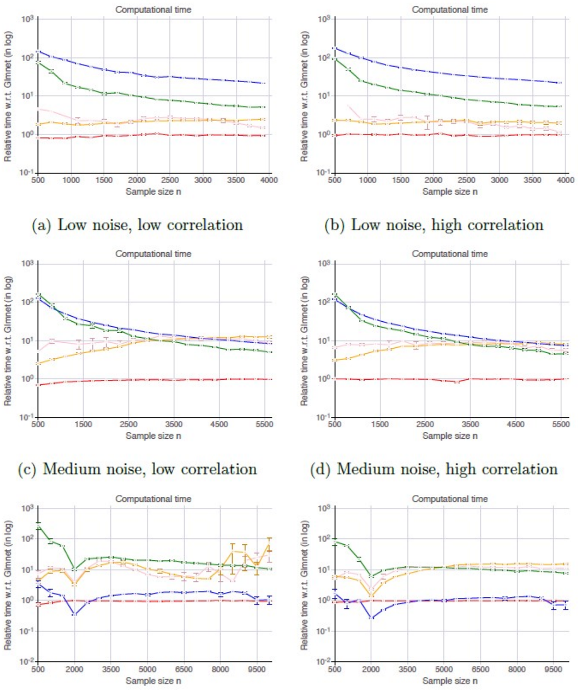
(c) Medium noise, low correlation

(d) Medium noise, high correlation

(e) High noise, low correlation

(f) High noise, high correlation

Fig 2: False Detection Rate



(a) Low noise, low correlation

(b) Low noise, high correlation

(c) Medium noise, low correlation

(d) Medium noise, high correlation

(e) High noise, low correlation

(f) High noise, high correlation

Fig 3: Computational time

---- LASSO

---- MCP

---- SCAD

---- Integer Optimization Formulation

---- Boolean Relaxation

Appendix 2: Accuracy, False Detection Rate and computational time when synthetic data not satisfying mutual incoherence condition

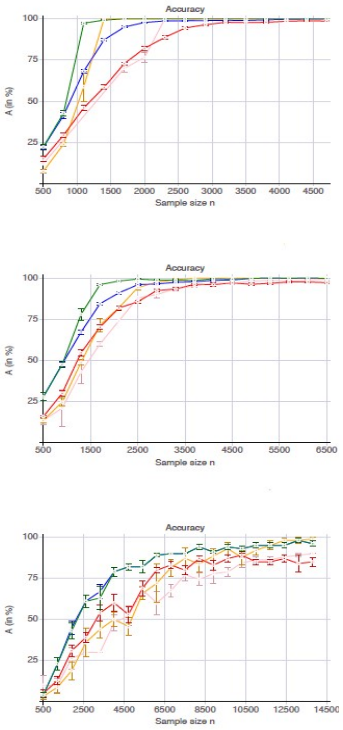


Fig 1: Accuracy

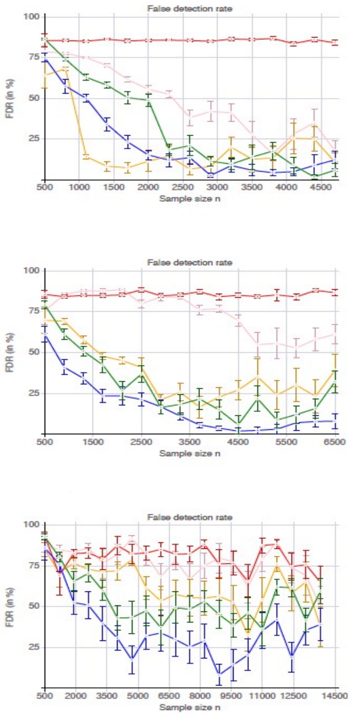
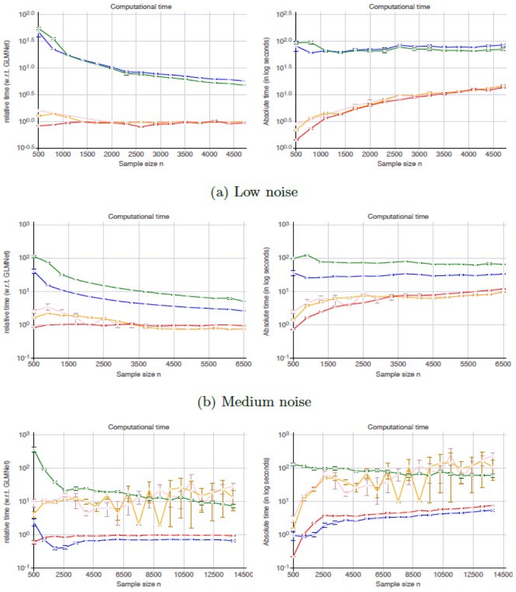


Fig 2: False Detection Rate



(a) Low noise

(b) Medium noise

(c) High noise

---- LASSO

---- MCP

---- SCAD

---- Integer Optimization Formulation

---- Boolean Relaxation

Fig 3: Computational time

Reference

Bertsimas, D., Pauphilet, J., & Van Parys, B. (2020, October 01). Institute of Mathematical Statistics: Statistical Science Future Papers. Retrieved November 01, 2020, from https://www.e-publications.org/ims/submission/STS/user/submissionFile/35175?confirm=f096a84a