Takehome Midterm

Due at 2024, 4/27 (Sat) 23:59

Job Description

Please conduct a short simulation study to compare the performance of various estimators for the Poisson Regression that assumes

$$y_i \mid \mathbf{x}_i \sim \text{Poisson}(\mu_{\boldsymbol{\beta}}(\mathbf{x}_i)), \qquad i = 1, \cdots, n \tag{1}$$

where

$$\log\{\mu_{\boldsymbol{\beta}}(\mathbf{x}_i)\} = \beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i.$$

Parameter Setting

To generate data from (1), set related parameters as follows:

- Intercept: $\beta_0 = 1$
- Coefficients: $\boldsymbol{\beta} = (1, 1, 1, 0, \cdots 0)^T \in \mathbb{R}^p$ with $p = 3$ and $p = 50$
- Predictors: $\mathbf{x}_i \overset{iid}{\sim} N_p(\mathbf{0}, \boldsymbol{\Sigma})$ where $\{\boldsymbol{\Sigma}\}_{ij} = \sigma_{ij} = \rho^{|i-j|}$ with $\rho = 0$ and $\rho = 0.7$.
- Training sample size: $n = 500$.
- Test sample size: $n_{ts} = 5,000$
- Number of independent repetitions: 100
- Seed number for the reproducibility: your student ID.

Methods in Comparison

Competing methods you should include for the comparison are:

- (POIS) unpenalized POISson Regression
- (RIDG) RIDGe-penalized Poisson Regression
- (LASS) LASSo-penalized Poisson Regression
- (ENET) Elastic-NET-penalized Poisson Regression (with $\alpha = 0.5$)
- (SCAD) SCAD-penalized Poisson Regression

– You must use your own code (either in `R` or `Python`) to implement the methods above.
– Select tuning parameter λ that maximizes the test classification error rate.

Performance Measure Corrected

We consider the following measurements for comparison.

- Estimation Accuracy: Let $\hat{\beta}_{j,k}$ denotes the estimator of β_j at the kth iteration $k = 1, \cdots, N$ with $N = 100$.
  - MC MSE: $\frac{1}{N}\sum_{k=1}^{N}(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta})^T(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta})$
  - MC Variance: $\text{tr}\{\frac{1}{N}\sum_{k=1}^{N}(\hat{\boldsymbol{\beta}}_j - \bar{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}_j - \bar{\boldsymbol{\beta}})^T\}$ where $\bar{\boldsymbol{\beta}} = \frac{1}{N}\sum_{k=1}^{N}\hat{\boldsymbol{\beta}}_k$.
  - MC Bias: $\mathbf{1}^T|\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}|$
- Variable Selection Performance
  - CS: number of correctly selected variables
  - IS: number of incorrectly selected variables
  - AC: 1 when the variable selection result is perfect, and 0 otherwise
- Averaged Computing Time

What you should Submit

You have to submit

- A short manuscript to report your work. (No more than 5 pages, LaTex template available)
- A single program file (either in `R` or `Python`) that can reproduce your result (It must include all functions you coded and main part for the simulation)

Submission Rules (Important!)

- You must send me the following by e-mail (`sjshin@korea.ac.kr`, and cc to your email).

  i) "Report (in pdf)" that the final results in a given form. Possible sections you may consider include
  1. Introduction
  2. Competing Methods
  3. Simulation Set Up
  4. Result
  5. Discussion

  ii) "single R file" ready-to-run file to recover your results given above.
- **Due date: 4/27 (Sat) 23:59**
  - If I get your mail after 4/28 (Sun) 00:15, you will lose 30% of the credits you earned.
  - If I get your mail after 4/28 (Sun) 00:30, NO credit!
- Additional rules:
  - Subject line of the email: `ST509_Midterm_StuduentID` (ex: `ST509_Final_2019150010`)
  - File name of your report: `ST509_Midterm_StuduentID.pdf`
  - File name of your code: `ST509_Midterm_StuduentID.r` or `.ipynb`
  - All functions and codes must be be included in a single program file.
- If you do NOT strictly follow these rules above, you additionally lose 5% of your credits.