Advanced Bayesian Computation

Rajarshi Guhaniyogi
Winter 2016

January 8, 2016

Course Information

- Lectures: MWF 9:30-10:40
- Lecture notes or relevant study materials will be posted every week.
- The course will be graded on two homeworks and one end term project.
- Homework 1: 25%, Homework 2: 25% and End Term: 50%.
- Students taking Satisfactory/Unsatisfactory are required to submit all the homeworks and the final project.
- There will be a 23 minutes presentation for the end term project. I would encourage you to work on the end term project from the late January.
- Lectures will be delivered for 9 weeks. Last week is reserved for the end term presentation.

Overview

High dimensional regression with an emphasis on Bayesian methodology

- Penalized optimization: Ridge regression, lasso, elastic net, adaptive lasso, group lasso.
- Bayesian high dimensional regression:
  (i) g-prior, two paradoxes, connection with model selection, mixture of g-priors.
  (ii) Spike and slab prior, detailed discussion, problem with model selection and computation, stochastic search variable selection, issues.
  (iii) Median probability model in connection with spike and slab prior.
  (iv) shrinkage estimation, how the name has appeared, motivation, some of the prominent shrinkage priors, Polson and Scott representation.
  (v) Briefly describe a theoretical result for shrinkage priors.

Overview Cont.

Modeling big data

(i) Divide and conquer technique in big data, finding sufficient statistic.
(ii) Sequential Monte Carlo.
(iii) Assumed density filtering.
(iv) Stochastic gradient decent and other applications through stochastic gradient Langevin dynamics.

Approximate Bayes method

(i) Variational Bayes: Definition, how to compute it.
(ii) Variational Bayes in nonparametric models.
(iii) Stochastic variational inference.

Regression Analysis: An old tool

- Statistical regression is occupying the literature from early 19th century.
- The entire strength of statistics comes from regression analysis.
- With the advancements in computation techniques and various sources of data, regression analysis has been extended to model various situations.
- Our motto is to discuss techniques that makes us up to date with the modern techniques in regression analysis.
- In particular, we will discuss situations where the number of predictors is large.
- Such things typically occur in biomedical applications.

Linear Regression: Formulation

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon, \;\; \epsilon \sim N(0, \sigma^2)$$

- Different structures of ϵ can be accommodated.
- We minimize sum of squared errors to estimate the regression coefficients.

Understanding Error

- Sum of squared error is a representation of the error in the OLS.
- Sum of squared prediction error is the sum of variance and square of bias.
- Though we only care about the squared prediction error, it becomes helpful to individually understand variance and squared bias.

Tradeoff between Bias and Variance

- There is a tradeoff between bias and variance in the sense that if model complexity increases, bias decreases, variance increases.
- It is always important to protect from under and over-fitting.
- Important to hit the point with lowest prediction error.

Gauss Markov Theorem

- Gauss Markov theorem states that among all linear unbiased estimates, OLS has the smallest error.
- There can be some BIASED estimator which is able to provide lower MSE.

Shrinkage Estimation

- Let OLS estimate is $\hat{\beta}_j$. What happens to the MSE if we use an estimator $\tilde{\beta}_j = \frac{\hat{\beta}_j}{1+\lambda}$?

Shrinkage Estimation

- Let OLS estimate is $\hat{\beta}_j$. What happens to the MSE if we use an estimator $\tilde{\beta}_j = \frac{\hat{\beta}_j}{1+\lambda}$?
- **Initially looks like a crazy idea, but lets give it a shot**.
- In particular, can we achieve lower MSE than OLS?
- Yes, we can. But the resulting estimator has to be biased. Whatever we pay for bias is compensated by the variance.
- λ that minimizes the error is $\lambda = \frac{p\sigma^2}{\sum_{j=1}^{p} \hat{\beta}_j^2}$.
- Note: As λ becomes big this estimator approaches to 0.

Shrinkage Estimation

- Charles Stein with his student James found that the estimator $\beta_j' = \left(1 - \frac{(p-2)\sigma^2}{\sum \hat{\beta}_j^2}\right) \hat{\beta}_j$ has less MSE when σ^2 is known.
- Stanley Sclove proposed to shrink the estimator close to zero if we find negative value, i.e. $\left(1 - \frac{(p-2)\sigma^2}{\sum \hat{\beta}_j^2}\right)^+ \hat{\beta}_j$.
- If σ^2 is unknown, he proposed taking $\beta_j' = \left(1 - \frac{cRSS}{\sum \hat{\beta}_j^2}\right)^+ \hat{\beta}_j$, for some constant c.

Shrinkage Estimation Contd..

- Note that the F-statistic is given by $F = \frac{\sum \hat{\beta}_j^2 / p}{RSS/(n-p)}$.
- Expressing Sclove estimator as $\beta_j' = \left(1 - \frac{c(n-p)}{pF}\right)^+ \hat{\beta}_j$, it seems that if the F test statistic is greater than c then all estimators are set to zero.

Shrinkage Estimation Contd...

- The above estimation sets all elements to either zero or nonzero.
- Stepwise regression adds or subtracts new variables in the regression if there is an improvement in terms of AIC or BIC. AIC $=$ n RSS $+2$ df, AIC $=$ n RSS $+\log(n)$ df.
- But this is not automated. Is there any method that automates shrinkage?
- What about the shrinkage parameter. Can we use it to estimate stuff?

Ridge Regression

- In statistical literature, ridge regression was introduced from a completely different perspective.
- Remember, if $\boldsymbol{X}$ is the $n \times p$ matrix and $\boldsymbol{y}$ is the $n \times 1$ responser vector, OLS estimator is given by the solution to the equation $\boldsymbol{X}'\boldsymbol{X}\beta = \boldsymbol{X}'\boldsymbol{y}$.
- Suppose $\boldsymbol{X}'\boldsymbol{X}$ does not have an inverse or the inverse is highly unstable.
- Can happen when $n < p$ or when columns are highly correlated.
- One idea is to solve $(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})\beta = \boldsymbol{X}'\boldsymbol{y}$, with small λ.

Ridge Regression

- For ridge regression $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}$.
- Note that $E(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \neq \boldsymbol{\beta}$.
- $Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}$.
- λ is the key parameter. How to choose λ?

Generalized Cross Validation to Choose λ

- **k fold:**
  (i) Divide the data into ten (equal) parts, $\mathscr{S}_1, ..., \mathscr{S}_k$.
  (ii) Set λ on a grid, say $\lambda \in \{\lambda_1, ..., \lambda_s\}$.
  (iii) For every λ_j, use $\mathscr{S}_{-i_1}$ to fit the model and $\mathscr{S}_{i_1}$ to calculate model fitting error for $i_1 = 1, ..., 10$.
  (iv) Find the average mean squared error.
  (v) Choose that λ_j which minimizes this error.
  (vi) In general, $k = 10$ is used.
- **leave one out:**
  (i) When n is small, generally leave one out cross validation is preferred over the k fold.
  (ii) Fit the model with $n - 1$ data points and validate with the nth one.
  (iii) Repeat it for all sample points to calculate the mean squared error.
  (iv) Choose λ_j that minimizes the error.

More on Ridge Regression

- Ridge regression will ensure that the coefficients decrease in size.
- In Ridge regression, one does not penalize the intercept as it is in the same scale as the predictors.
- Also predictors can be of vastly different scales. To ensure fair shrinkage to all, generally predictors are standardized.
- This also sets the intercept to zero.
- R code to compute ridge regression is attached.

Variable Selection

- Variable selection means to select important variables which are affecting the response under the regression model.
- For example, there may be a subset of coefficients which are identically zero. The corresponding predictors have no effect on the regression.
- For ridge regression the coefficients are zero only when $\lambda = \infty$.
- Therefore ridge regression **can't select variables**.
- It is useful when a lot of coefficients are close to zero.
- It also does not perform well when a lot of coefficients are moderately large.
- Some post-processing steps may be taken to select variables. But is there any model based straightforward way to select variables?

Lasso

- Lasso is an acronym for least absolute selection and shrinkage operator.
- It combines the good features of ridge regression with variable selection.
- It is competitive in terms of prediction error w.r.t ridge regression.
- Note that the formulation of ridge regression is

$$arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j|^2$$

Lasso replaces l_2 penalty by the l_1 penalty, i.e.

$$arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$