Advanced Bayesian Computation

Rajarshi Guhaniyogi
Winter 2016

January 15, 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.

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

Lasso Contd..

- As λ increases less variables are included, might have higher prediction error after certain λ.
- The idea is to choose λ so as to have proper model fit as well as variable selection.
- λ is again chosen using generalized cross validation.
- Code for lasso.
- Great thing about lasso is its property of variable selection. Why it happens to lasso and not to ridge?

Insight into the Geometry of Lasso and Ridge

- the ridge and lasso optimization can be written as the minimization over $\boldsymbol{\beta}$

$$||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 \quad \text{subject to} \quad ||\boldsymbol{\beta}||_2^2 \leq \lambda$$
$$||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 \quad \text{subject to} \quad ||\boldsymbol{\beta}||_1 \leq \lambda.$$

  The above is equivalent to the optimization problems

$$(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{OLS})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{OLS}) \quad \text{subject to} \quad ||\boldsymbol{\beta}||_2^2 \leq \lambda$$
$$(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{OLS})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{OLS}) \quad \text{subject to} \quad ||\boldsymbol{\beta}||_1 \leq \lambda.$$

- OLS corresponds to the unconstrained optimization.
- The shapes of ridge and lasso are discussed in class.

Elastic net: Motivation

- Variable selection with lasso has two shortcomings.
  (i) The number of variables selected is bounded by the total number of samples in the dataset.
  (ii) Lasso fails to perform group variable selection, i.e. if a group of variables are correlated, lasso tends to select only one of them.
- Elastic net is motivated by the above two shortcomings.
- You are throwing a net to catch multiple fishes together.

Theorem: Suppose $x_i = x_j$ and $J(\boldsymbol{\beta})$ is a strictly convex function. Suppose $\hat{\boldsymbol{\beta}}$ is obtained by optimizing the objective function $||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \lambda J(\boldsymbol{\beta})$. Then $\hat{\beta}_i = \hat{\beta}_j$.

- Since elastic net penalty is strictly convex, elastic net achieves group variable selection.

Elastic Net

- The elastic net forms a hybrid of the l_1 and l_2.
- The l_1 part of the penalty generates a sparse model.
- The quadratic part of the penalty
  (i) removes limitation on the number of selected variables;
  (ii) encourages grouping effect.

$$\boldsymbol{X}^*_{(n+p)\times p} = \frac{1}{\sqrt{(1+\lambda_2)}} \begin{pmatrix} \boldsymbol{X} \\ \sqrt{\lambda_2}\boldsymbol{I} \end{pmatrix}, \boldsymbol{y}^* = (\boldsymbol{y}, \boldsymbol{0})',$$

$$\gamma = \frac{\lambda_1}{\sqrt{1+\lambda_2}}, \boldsymbol{\beta}^* = \sqrt{1+\lambda_2}\boldsymbol{\beta}.$$

The elastic net objective function can be written as

$$||\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*||^2 + \gamma||\boldsymbol{\beta}^*||_1.$$

Thus elastic net can select all p predictors.

Connections Between Lasso, Elastic Net and Ridge

- Naive elastic net is given by

$$\hat{\beta}_{elastic} = arg\,\min ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \lambda_1||\boldsymbol{\beta}||^2 + \lambda_2||\boldsymbol{\beta}||_1$$

- Elastic net penalty can be viewed as $\sum_{j=1}^{p}\left[(1-\alpha)|\beta_j| + \alpha|\beta_j|^2\right]$.
- $\alpha = 0$ gives lasso, $\alpha = 1$ gives ridge.
- Solution to the above elastic net penalty is known as the naive elastic net. Unfortunately it does not perform well in practice.
- The intuitive reason being double penalization.
- Actual elastic net is scaled naive elastic net estimates, $\beta(enet) = (1+\lambda_2)\beta(naive\ \ enet)$.

Adaptive Lasso

- The *adaptive lasso* uses a weighted penalty of the form $\sum_{j=1}^{p} w_j|\beta_j|$ where $w_j = 1/|\hat{\beta}_j|^{\nu}$, $\hat{\beta}_j$ is the ordinary least squares estimate and $\nu > 0$.
- The adaptive lasso yields consistent estimates of the parameters while retaining the attractive properties of lasso. Idea is to favor predictors with univariate strength, to avoid spurious selection of noise predictors.
- When $p > n$, can use univariate regression coefficients in place of full least squares estimates.
- In general, when the predictors are correlated it is a good practice to use univariate regression coefficients.
- Adaptive lasso recovers the correct model under milder condition than lasso.
- Computationally it does not add any extra significant burden to lasso computation.

Group Lasso

- In some problems, the predictors belong to pre-defined groups.
- In this situation it may be desirable to shrink and select the members of a group together. The *group lasso* in one way to achieve this.
- Suppose p predictors are divided into m groups, with p_j number of predictors in group j, $j = 1, ..., m$; $p_1 + \cdots + p_m = p$.
- $\boldsymbol{X}_j$ matrix corresponding to the jth group of predictors.
- $\boldsymbol{\beta}_j$ is the vector coefficient corresponding to $\boldsymbol{X}_j$.
- Group lasso minimizes

$$arg \min_{\boldsymbol{\beta} \in \mathscr{R}^p} \left[||\boldsymbol{y} - \beta_0 \mathbf{1} - \sum_{j=1}^{m} \boldsymbol{X}_j \boldsymbol{\beta}_j||^2 + \lambda \sum_{j=1}^{m} \sqrt{p_j} ||\boldsymbol{\beta}_j||_2 \right]$$

Clustering in High Dimensions: Nonnegative Matrix Factorization

- Given a matrix $\boldsymbol{M}_{p\times n}$ and a desired rank $k << min(n, p)$, find $\boldsymbol{W}_{p\times k}$ and $\boldsymbol{H}_{k\times n}$ s.t. $\boldsymbol{M} \approx \boldsymbol{W}\boldsymbol{H}$ by solving an optimization problem $min_{\boldsymbol{W}>0,\boldsymbol{H}>0}||\boldsymbol{M} - \boldsymbol{W}\boldsymbol{H}||^2$.
- Why do this when SVD does a better job in approximating $\boldsymbol{M}$.
- If $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}$, then $||\boldsymbol{M} - \boldsymbol{U}_k\boldsymbol{\Sigma}_k\boldsymbol{V}_k|| \leq ||\boldsymbol{M} - \boldsymbol{W}\boldsymbol{H}||$.
- Reason to do NMF: For nonnegative data NMF approximation provides better interpretation.

NMF and K-Means Clustering

- k-means clustering can be written as $||\boldsymbol{M} - \boldsymbol{W}\boldsymbol{H}||^2$.
- Columns of $\boldsymbol{H}$ gives us the cluster membership indicators.
- Look at the largest element in each column of $\boldsymbol{H}$.
- That sample is included in the corresponding cluster.
- Sometimes to make it similar to the K-means, sparse NMF is employed.

Penalized Optimization: Unsatisfactory in Predictive Inference

- Penalized optimization is unable to provide predictive inference. Only provides point prediction.
- Typical focus in many scientific applications is uncertainty characterization.
- Different choices of tuning parameters may affect inference considerably.

Bayesian Approach

- If loss function corresponds to a likelihood & penalty to the log prior (up to normalizing constants), then estimates correspond to mode of a Bayesian posterior (MAP estimates).
- Consider the linear regression model with known σ^2 and with prior

$$y_i \sim N(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2), \;\; \beta_j \sim \pi_\beta.$$

- The log posterior of $\boldsymbol{\beta}$ upto a constant is

$$-\frac{1}{2\sigma^2}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \sum_{j=1}^{p} \log(\pi_\beta(\beta_j)$$

- Although such estimators correspond to the mode of a Bayesian posterior, they are typically not viewed as Bayesian.
- Bayes estimators $\hat{\beta}_{Bayes}$ are defined as the value that minimizes the Bayes risk.
- Bayes risk is the expectation of a loss $L(\hat{\beta}, \beta)$ averaged over the posterior of β.
- For example, if we choose squared error loss, $\hat{\beta}$ is the posterior mean.
- MAP is not a Bayes estimator for a reasonable choice of loss function.
- Also, we would like to utilize the whole posterior instead of just using a point estimate.