Advanced Bayesian Computation

Rajarshi Guhaniyogi
Winter 2016

February 1, 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 $\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}$.
- Note that $E(\hat{\beta}) = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}\beta \neq \beta$.
- $Var(\hat{\beta}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}$.
- 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

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

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

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

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 $\boldsymbol{\beta}$

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

  The above is equivalent to the optimization problems

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

- 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.

$$\boldsymbol{X}^*_{(n+p)\times p} = \frac{1}{\sqrt{(1+\lambda_2)}} \begin{pmatrix} \boldsymbol{X} \\ \sqrt{\lambda_2}\boldsymbol{I} \end{pmatrix}, \boldsymbol{y}^* = (\boldsymbol{y}, \mathbf{0})',$$
$$\gamma = \frac{\lambda_1}{\sqrt{1+\lambda_2}}, \boldsymbol{\beta}^* = \sqrt{1+\lambda_2}\boldsymbol{\beta}.$$

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

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

- 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

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

- Although such estimators correspond to the mode of a Bayesian posterior, they are typically not viewed as Bayesian.
- Bayes estimators $\hat{\boldsymbol{\beta}}_{Bayes}$ are defined as the value that minimizes the Bayes risk.
- Bayes risk is the expectation of a loss $L(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ averaged over the posterior of $\boldsymbol{\beta}$.
- For example, if we choose squared error loss, $\hat{\boldsymbol{\beta}}$ 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.

Bayesian Approach in High Dimensions

- Bayesians choose a prior distribution $\pi(\boldsymbol{\beta}, \sigma^2)$ and calculate the posterior

$$\pi(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}, \boldsymbol{X}) = \frac{\pi(\boldsymbol{\beta}, \sigma^2) N(\boldsymbol{y} | \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})}{\int \pi(\boldsymbol{\beta}, \sigma^2) N(\boldsymbol{y} | \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}) d\boldsymbol{\beta} d\sigma^2}$$

- When $n >> p$, $\pi(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}, \boldsymbol{X}) \approx N(\boldsymbol{\beta} | \hat{\boldsymbol{\beta}}, \boldsymbol{I}(\boldsymbol{\beta})^{-1})$, where $\boldsymbol{I}(\boldsymbol{\beta})$ is the Fisher information matrix.
- The above is called the Bernstein-Von Mises theorem or the Bayesian central limit theorem.
- This essentially means that when $n >> p$, prior does not have much role in determining the posterior. In fact, the likelihood swamps the prior and we essentially get equivalent results from frequentist and Bayesian.
- This rosy picture breaks down when p is large.
- Prior has profound effect for large p and it is essential to carefully design the prior.

Prior Design

- Priors should be designed in such a way that the posterior of $\boldsymbol{\beta}$ concentrates around the "true" $\boldsymbol{\beta}_0$.
- Prior should have sufficient information. Flat prior on $\boldsymbol{\beta}$ gives inconsistencies.
- Motivated by the idea of sparsity, one popular approach is to impose sparsity on $\boldsymbol{\beta}$ through prior distributions.
- Later we will see that designing prior on $\boldsymbol{\beta}$ can also be governed by other considerations.

Bayesian Variable Selection by Sparsity

- Spike and slab prior

$$\beta_j \overset{iid}{\sim} \pi_0\delta_0 + (1-\pi_0)g.$$

One popular choice of g is $N(0, c)$.
π_0 is the prior probability of excluding a predictor.
δ_0 is the degenerate distribution at 0.
Prior on the nonzero coefficients are given by g.

More into Spike and Slab

- Define the variable inclusion indicator by $\gamma_j = I(\beta_j \neq 0)$.
- Therefore, $\gamma_1, ..., \gamma_p$ indicate which predictors are included in the model, $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_p)' \in \{0, 1\}^p$.
- Note that, depending on whether a variable is included or excluded, the total number of candidate models is 2^p.
- A candidate model is represented by $\boldsymbol{\gamma}$.
- The size of this model $p_\gamma = \sum_{j=1}^{p} \gamma_j$, $p_\gamma \sim Binomial(p, 1 - \pi_0)$.
- Thus the expected model size is $p(1 - \pi_0)$.
- Clearly, if we fix π_0 and p is big, it gives a lot of prior information on the model size.
- π_0 is an important parameter and generally assigned a beta prior.

Posterior Probability of γ

- Let $\boldsymbol{\beta}_\gamma = \{\beta_j : \gamma_j = 1, j = 1, ..., p\}$.
- Marginal likelihood of the model γ is

$$L(\gamma|\boldsymbol{y}, \boldsymbol{X}) = \int N(\boldsymbol{y}|\boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma, \sigma^2 \boldsymbol{I})\pi(\boldsymbol{\beta}_\gamma, \sigma^2) d\boldsymbol{\beta}_\gamma d\sigma^2.$$

- The posterior probability of model γ is given by

$$\pi(\gamma|\boldsymbol{y}, \boldsymbol{X}) = \frac{L(\gamma|\boldsymbol{y}, \boldsymbol{X})\pi(\gamma)}{\sum_{\boldsymbol{\gamma}^*} L(\gamma^*|\boldsymbol{y}, \boldsymbol{X})\pi(\gamma^*)}.$$

- Not feasible to compute posterior probability of each model since there are 2^p of them.

Stochastic Search Variable Selection

- Due to the intractability of calculating the posterior probabilities exactly, stochastic search is often used.
- Stochastic Search Variable Selection (SSVS) moves between multiple models and comes back to models which are more representative of the data.
- SSVS (George & McCulloch, 1997, *Statistica Sinica*) rely on MCMC to conduct this search.
- $\beta_j \sim (1-\gamma_j)N(0, v_{0j}) + \gamma_j N(0, v_{1j})$, $\gamma_j \overset{ind.}{\sim} Ber(w_j)$.
- v_{0j} small, v_{1j} "reasonably" big (away from 0).
- George & McCulloch suggested taking $v_{0j} = \tau_j^2$, $v_{1j} = g\tau_j^2$, g big, τ_j^2 small.
- $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$, $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_p)'$.
- $\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2) = \left[\prod_{j=1}^{p} \pi(\beta_j|\sigma^2, \gamma_j)\pi(\gamma_j)\right] \pi(\sigma^2)$.
- $\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2|\boldsymbol{y}) \propto N(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2)$.

Problems with SSVS

- MCMC runs for a large number of iterations and hops between different models. Posterior probability of a model is estimated by the proportion of times the model has been visited by the Markov chain.
- Suffers when there are high correlations between variables.
- Not useful if one wants to add a flat prior to the β_j's.
- Choice of g gives headache.
- Often viewed as not scalable to really big p but use of GPUs & other tricks helps.

Inference from SSVS

- Huge advantage of Bayes is the ability to quantify uncertainty.
- Bayes allows estimation of marginal inclusion probabilities $P(\gamma_j = 1|\boldsymbol{y}, \boldsymbol{X})$. It is the proportion of times MCMC iteration visits a model with jth variable included.
- It is an indication of how important a predictor is.
- One might employ selection of predictors by thresholding marginal inclusion probability at 0.5.
- The above gives rise to the median probability model which enjoys predictive optimality properties.

More on SSVS

- SSVS is appealing for its ability to select variables.
- We will discuss its theoretical optimality properties later.
- A major drawback of the SSVS is the combinatorial search for big p. This is computationally cumbersome for big p.
- If a few predictors are highly correlated, SSVS tends to miss all of them.
- It is sometimes appealing computationally & philosophically to relax assumption of exact zeros.
- That is sparsity can be introduced in a "weaker sense".
- " This view of sparsity may appeal to Bayesians who oppose testing point null hypotheses, and would rather shrink than select".
- Instead, we want coefficients corresponding to the noisy predictors are approximately zero while leaving signals alone.

Continuous Shrinkage Priors

- Shrinkage priors are continuous prior distributions which pulls the unimportant predictor coefficients to zero while keeping the important predictor coefficient unshrunk.
- Predictor coefficients are not exactly zero but close to zero.
- Rich literature on shrinkage priors - Laplace (Bayes Lasso), Cauchy, horseshoe, generalized double Pareto, etc.
- Priors should concentrate at zero with heavy tails.

Global Local Representation

- Polson and Scott (2010) show that essentially all shrinkage priors can be represented as

$$\beta_j|\psi_j, \tau \overset{ind.}{\sim} N(0, \psi_j \tau), \;\; \psi_j \overset{iid}{\sim} g, \;\; tau \sim f.$$

- Global-scale τ facilitates concentration near zero.
- Local-scales highly variable to avoid over-shrinking β_j's corresponding to important predictors.
- Scale mixtures of Gaussians allow simple Gibbs sampler form in most cases and hence computationally appealing.
- In this formulation, the sparseness problem is the mirror image of the outlier problem.
- Strong global shrinkage handles the noise; the local ψ_j's act to detect the signals, which are outliers relative τ.

Early Shrinkage Priors

- $\beta_j|\tau,\psi_j \sim N(0,\tau^2\psi_j^2), \psi_j^2 \sim IG(\zeta/2,\zeta/2), \tau^2 \sim IG(a,b)$. This gives rise to $\beta_j|\tau \sim t_\zeta(0,\tau)$.
- Strawderman-Berger prior: $\beta_j|\psi_j \sim N(0,\lambda_j^{-1}-1)$, $\psi_j \sim Beta(1/2,1)$.

Early Shrinkage Priors

- Bayesian Lasso (Park and Casella, 2008; Hans 2009).

$$\pi(\boldsymbol{\beta}|\sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} \exp(-\lambda|\beta_j|/\sigma).$$

  Double exponential can be written as a scale-mixture of normal distributions.

$$\begin{gathered} \boldsymbol{\beta}|\tau_1^2, ..., \tau_p^2 \sim N(\mathbf{0}, \sigma^2 diag(\tau_1^2, ..., \tau_p^2)), \\ \pi(\tau_j^2) = \frac{\lambda^2}{2} \exp(-\frac{\lambda^2\tau_j^2}{2}) \\ \sigma^2 \sim \pi_\sigma. \end{gathered}$$

- Closely resembles Frequentist lasso.
- $\pi(\lambda^2) = \frac{\delta^r}{\Gamma(r)}(\lambda^2)^{r-1}e^{-\delta\lambda^2}$.

Full conditionals

Let $(\boldsymbol{X}'\boldsymbol{X} + diag(\tau_1^2, ..., \tau_p^2)) = \boldsymbol{A}$,

- $\boldsymbol{\beta}|- \sim N(\boldsymbol{A}^{-1}\boldsymbol{X}'\boldsymbol{y}, \sigma^2\boldsymbol{A}^{-1})$
- $\sigma^2|- \sim IG(\frac{n-1+p}{2}, \frac{(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})}{2})$
- $\frac{1}{\tau_j^2}|- \sim InvGaussian(\mu', \lambda')$, $\mu' = \sqrt{\frac{\lambda^2\sigma^2}{\beta_j^2}}, \lambda' = \lambda^2$.
- $\lambda^2|- \sim IG(p + r, \delta + \sum_{j=1}^{p} \tau_j^2/2)$.
- Other priors on λ^2 has also been used.
- One can also apply empirical Bayes technique to estimate the parameter λ.

Horseshoe

Horseshoe was proposed for the normal means problem, i.e. $y_i = \theta_i + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$.

$$\theta_i | \lambda_i \sim N(0, \tau^2 \lambda_i^2), \lambda_i \sim C^+(0,1), \tau \sim C^+(0,1).$$

- The same prior can be applied to the regression coefficients.
- Let $\kappa_i = \frac{1}{1+\lambda_i^2 \tau^2}$, then $E(\beta_i | \boldsymbol{y}, \lambda_i, \tau) = (1 - \kappa_i) y_i$.
- We expect κ_i to be either close to zero or close to 1.
- Show figures of κ_i for some priors.
- In Horseshoe $\kappa_i \sim Beta(1/2, 1/2)$.

κ_i Distribution of Different Priors

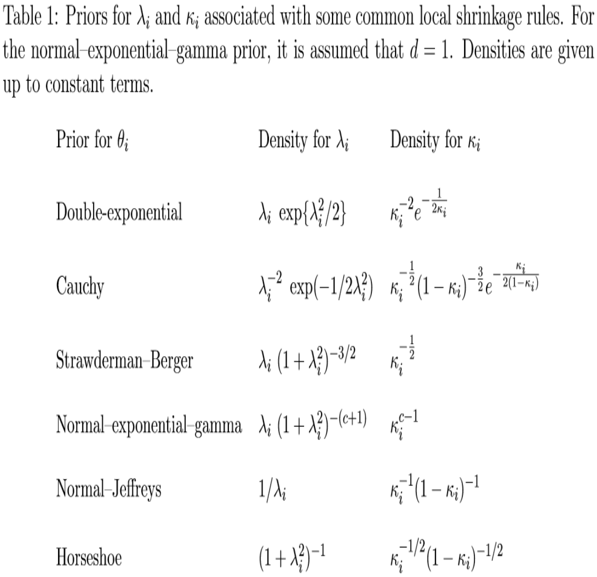


Table 1: Priors for λ_i and κ_i associated with some common local shrinkage rules. For the normal–exponential–gamma prior, it is assumed that $d = 1$. Densities are given up to constant terms.

Prior for θ_i	Density for λ_i	Density for κ_i
Double-exponential	$\lambda_i \ \exp\{\lambda_i^2/2\}$	$\kappa_i^{-2} e^{-\frac{1}{2\kappa_i}}$
Cauchy	$\lambda_i^{-2} \ \exp(-1/2\lambda_i^2)$	$\kappa_i^{-\frac{1}{2}}(1-\kappa_i)^{-\frac{3}{2}} e^{-\frac{\kappa_i}{2(1-\kappa_i)}}$
Strawderman–Berger	$\lambda_i \ (1+\lambda_i^2)^{-3/2}$	$\kappa_i^{-\frac{1}{2}}$
Normal–exponential–gamma	$\lambda_i \ (1+\lambda_i^2)^{-(c+1)}$	κ_i^{c-1}
Normal–Jeffreys	$1/\lambda_i$	$\kappa_i^{-1}(1-\kappa_i)^{-1}$
Horseshoe	$(1+\lambda_i^2)^{-1}$	$\kappa_i^{-1/2}(1-\kappa_i)^{-1/2}$

Generalized Double Pareto

- $f(\beta|\alpha,\zeta) = \frac{1}{2\zeta}\left(1 + \frac{|\beta|}{\alpha\zeta}\right)^{-\alpha+1}$.
- $\zeta = \alpha = 1$ is the standard double pareto distribution.
- $\beta \sim N(0,\tau), \tau \sim Exp(\lambda^2/2), \lambda \sim Ga(\alpha,\eta)$, then $\beta \sim GDP(\frac{\eta}{\alpha},\alpha)$.
- If α grows, density becomes lighter tailed. If η grows, density becomes flatter and variance increases.
- If α and η both grows at the same rate, variance becomes constant but tails become lighter. It reaches to a laplace density.
- Default is $\alpha = \eta = 1$. It gives a cauchy like tail behavior.

Conditional Posterior Distributions

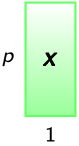
$\beta_j|\tau_j, \sigma^2 \sim N(0, \sigma^2\tau_j), \tau_j \sim Exp(\lambda_j^2/2), \lambda_j \sim Gamma(\alpha, \eta)$.

- $\boldsymbol{\beta}|- \sim N((\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{T}^{-1})^{-1}\boldsymbol{X}'\boldsymbol{y}, \sigma^2(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{T}^{-1})^{-1})$
- $\lambda_j|- \sim Gamma(\alpha + 1, |\beta_j|/\sigma + \eta)$
- $\frac{1}{\tau_j}|- \sim Inv - Gaussian(\sqrt{(\lambda_j^2\sigma^2)/\beta_j^2}, \lambda_j^2)$
- $\sigma^2|- \sim IG((n+p)/2, (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})/2 + \boldsymbol{\beta}'\boldsymbol{T}^{-1}\boldsymbol{\beta}/2)$

where $\boldsymbol{T} = diag(\tau_1, ..., \tau_p)$.

- Put prior on α and η, $\pi(\alpha) = \frac{1}{(1+\alpha)^2}$, $\pi(\eta) = \frac{1}{(1+\eta)^2}$. They are both centered at $\alpha = \eta = 1$.

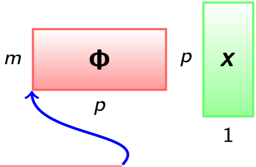
Can Compressing Predictors Help?



- $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are of mammoth size.
- Storage is highly prohibitive, let alone computation.

Idea

Compressing predictors randomly in low dimension helps solve our problem.



- Random Projection Matrix

Random Projection Matrix: Lots of Zeroes Required in $\mathbf{\Phi}$

- Overwhelming literature on scaled Gaussian projection matrix ($\Phi_{ij} \sim N(0, 1/m)$)→ **popular choice**.

Random Projection Matrix: Lots of Zeroes Required in $\mathbf{\Phi}$

- Overwhelming literature on scaled Gaussian projection matrix ($\Phi_{ij} \sim N(0, 1/m)$)→ **popular choice**.
- We anticipate sparsity in the true predictor coefficients.

Random Projection Matrix: Lots of Zeroes Required in $\mathbf{\Phi}$

- Overwhelming literature on scaled Gaussian projection matrix ($\Phi_{ij} \sim N(0, 1/m)$)→ **popular choice**.
- We anticipate sparsity in the true predictor coefficients.
- Might be important to have lots of zero entries in $\mathbf{\Phi}$.

Random Projection Matrix: Lots of Zeroes Required in $\mathbf{\Phi}$

- Overwhelming literature on scaled Gaussian projection matrix ($\Phi_{ij} \sim N(0, 1/m)$)$\rightarrow$ **popular choice**.
- We anticipate sparsity in the true predictor coefficients.
- Might be important to have lots of zero entries in $\mathbf{\Phi}$.

Random Projection (Dasgupta 2003, 2013)→ Our Choice

$$\Phi_{ij} = \begin{cases} -1/\sqrt{\psi} & w.p.\ \psi^2 \\ 0 & w.p.\ 2\psi(1-\psi) \\ 1/\sqrt{\psi} & w.p.\ (1-\psi)^2 \end{cases}$$

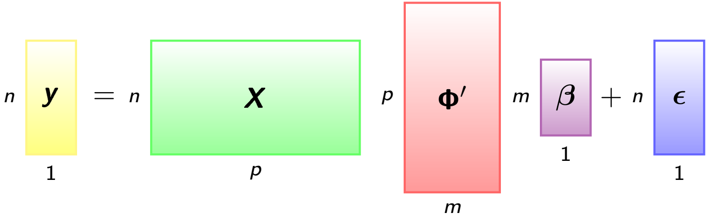
- Rows of $\mathbf{\Phi}$ are orthonormalized using Gram-Schmidt orthogonalization procedure.

Bayesian Compressed Regression

Compressed Regression

$$y = (\mathbf{\Phi} \boldsymbol{x})'\beta + \epsilon, \; \epsilon \sim N(0, \sigma^2)$$

β is the low dimensional coefficients on compressed predictors.



- No longer in the high-dimensional setting, use conjugate prior.

Conjugate Prior

$$\beta \,|\, \sigma^2 \sim N(\mathbf{0}, \sigma^2 \mathbf{\Sigma}_\beta), \; \sigma^2 \sim IG(a, b).$$

Analytical Posteriors & Posterior Predictive Distributions

Posteriors

$$\beta \,|\, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Phi} \sim t_n\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right), \;\; \sigma^2 \,|\, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Phi} \sim IG(a_1, b_1)$$

Posterior Predictive Distribution

$$y_{n+1} \,|\, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Phi}, \boldsymbol{x}_{n+1} \sim t_n\left(\mu_{pred}, \sigma^2_{pred}\right)$$

- Only needs sufficient statistics $\boldsymbol{X}'\boldsymbol{X}$, $\boldsymbol{X}'\boldsymbol{y}$ and $\boldsymbol{y}'\boldsymbol{y}$.
- Only matrix operation required $m \times m$ matrix inversion and $m \times p$, $p \times n$ matrix multiplication.

Analytical Posteriors & Posterior Predictive Distributions

Posteriors

$$\beta \,|\, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Phi} \sim t_n\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right), \;\; \sigma^2 \,|\, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Phi} \sim IG(a_1, b_1)$$

Posterior Predictive Distribution

$$y_{n+1} \,|\, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Phi}, \boldsymbol{x}_{n+1} \sim t_n\left(\mu_{pred}, \sigma^2_{pred}\right)$$

- Only needs sufficient statistics $\boldsymbol{X}'\boldsymbol{X}$, $\boldsymbol{X}'\boldsymbol{y}$ and $\boldsymbol{y}'\boldsymbol{y}$.
- Only matrix operation required $m \times m$ matrix inversion and $m \times p$, $p \times n$ matrix multiplication.

g-Prior

- g-prior was another class of approach that has surfaced long back due to its computational ease.
- Let ϕ be the precision parameter. The formulations of g-prior is

$$\boldsymbol{\beta}|\phi \sim N(\mathbf{0}, \frac{g}{\phi}(\boldsymbol{X}'\boldsymbol{X})^{-1}), \;\; \pi(\phi) \propto \frac{1}{\phi}$$

- How to choose g? Can a fixed g be used?
- Let the class of models be given by $\{\mathscr{M}_{\boldsymbol{\gamma}} : \boldsymbol{\gamma} = (\gamma_1, ..., \gamma_p)\}$.
- The marginal likelihood is given by

$$\pi(\boldsymbol{y}|\mathscr{M}_{\boldsymbol{\gamma}}) = \frac{\Gamma((n-1)/2)}{\sqrt{\pi}^{n-1}\sqrt{n}}||\boldsymbol{y} - \bar{\boldsymbol{y}}||^{-(n-1)} \frac{(1+g)^{(n-1-p_{\boldsymbol{\gamma}})/2}}{[1+g(1-R^2_{\boldsymbol{\gamma}})]^{(n-1)/2}}.$$

- $p_{\boldsymbol{\gamma}}$ is the number of nonzero $\boldsymbol{\gamma}$ in $\mathscr{M}_{\boldsymbol{\gamma}}$.
- $R^2_{\boldsymbol{\gamma}}$ is the R^2 statistics for the model $\mathscr{M}_{\boldsymbol{\gamma}}$.

Mixture of g-prior

- *Zellner-Siow prior*: $\pi(g) = \sqrt{(n/2)}\Gamma(1/2)g^{-3/2}e^{-n/(2g)}, g > 0$
- *Hyper-g prior*: $\pi(g) = \frac{a-2}{2}(1+g)^{-a/2}, g > 0.$

Big data, large sample size n

- In many machine learning or environmental applications number of predictors is small.
- Sample size is massive.
- Important data applications.
- It is a wide area with different strategies applied to different models.
- We will see a few strategies.

Divide and Conquer

- The idea is to divide the data into subsamples.
- Sequentially feed subsamples to the model.
- If the posterior distribution is dependent on the data only through sufficient statistics, it is a good strategy.