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

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

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

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

- Although such estimators correspond to the mode of a Bayesian posterior, they are typically not viewed as Bayesian.
- Bayes estimators $\hat{\beta}_{Bayes}$ 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.

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 Bernstain-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.

- One natural prior to consider is

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

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

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

- 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, 1993, *JASA*) 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 $v_{0j} = \tau_j^2$, $v_{1j} = g_j^2\tau_j^2$, g_j big, τ_j^2 small. Choice of g_j and τ_j?
- $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$, $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_p)'$.
- $\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2) = \left[\prod_{j=1}^{p} \pi(\beta_j|\sigma^2, \gamma_j)\pi(\gamma_j)\right] \pi(\sigma^2)$.
- $\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2|\boldsymbol{y}) \propto N(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2)$.

Updates in George and McCulloch, 1993 JASA

- Note that $\boldsymbol{\beta}|\boldsymbol{\gamma} \sim N(\mathbf{0}, \boldsymbol{D})$ where $\boldsymbol{D} = diag(a_1\tau_1^2, ..., a_p\tau_p^2)$ where $a_j = 1$ if $\gamma_j = 0$ and $a_j = g_j^2$ if $\gamma_j = 1$.
- Thus $\pi(\boldsymbol{\beta}|-) \propto N(\boldsymbol{y}|\boldsymbol{X}\beta, \sigma^2\boldsymbol{I})N(\boldsymbol{\beta}|\mathbf{0}, \boldsymbol{D})$
- $P(\gamma_j = 1|-) = h_1/(h_1 + h_2)$, where $h_1 = w_j N(\beta_j|0, g^2\tau_j^2)$, $h_2 = (1 - w_j)N(\beta_j|0, \tau_j^2)$
- If prior of $\sigma^2 \sim IG(a_\sigma, b_\sigma)$, then posterior of σ^2 is also Inverse Gamma.
- If additionally w_j is assigned a Beta(a_{w_j}, b_{w_j}) prior, then $\pi(w_j|-) \propto w_j^{\gamma_j}(1 - w_j)^{1-\gamma_j} Beta(w_j|a_{w_j}, b_{w_j})$. This is also a Beta distribution.

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.

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.
- Often viewed as not scalable to really big p but use of GPUs & other tricks helps.

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.

Spike and Slab LASSO

- We have seen penalized optimization with convex and separable penalty functions.
- Some non-convex and non-separable penalties can have desirable properties, however convex optimization can't be used for them.
- A few examples are MCP penalty of Zhang (2010), SCAD penalty of Fan and Li (2001).
- These penalties have the ability to threshold (select) and, at the same time, diminish the well-known estimation bias of the LASSO.
- Any penalized likelihood estimator may be seen as a posterior mode under a prior $\pi(\beta|\lambda)$, where $J(\beta) = log(\pi(\beta|\lambda))$.
- In particular, separable penalties stem from independent product priors.

Spike and Slab LASSO Contd.

- For the spike and slab prior
$\pi(\boldsymbol{\beta}|\boldsymbol{\gamma}) = \prod_{j=1}^{p} [\gamma_j \psi_1(\beta_j) + (1-\gamma_j)\psi_0(\beta_j)]$, $\boldsymbol{\gamma} \sim \pi(\boldsymbol{\gamma})$.
- Rockova (2015) deploys $\psi_1(\beta_j) = \frac{\lambda_1}{2} \exp(-\lambda_1|\beta_j|)$ and $\psi_0(\beta_j) = \frac{\lambda_0}{2} \exp(-\lambda_0|\beta_j|)$.
- Let $\gamma_j \sim Ber(\theta)$, then $\pi(\boldsymbol{\beta}|\theta) = \prod_{j=1}^{p} [\theta\psi_1(\beta_j) + (1-\theta)\psi_0(\beta_j)]$
- When $\psi_1(\cdot) = \psi_0(\cdot)$, we get back the LASSO penalty.
- Letting $\lambda_0 \to \infty$ and $\lambda_1 \to 0$ gives back l_0 penalty.
- Thus a continuum of non-convex penalties can be created between these two extremes.

Spike and Slab LASSO Contd.

- The spike and slab LASSO penalty $-\frac{\pi(\boldsymbol{\beta}|\theta)}{\pi(\mathbf{0}|\theta)}$.
- This penalty is the sum of the LASSO penalty and a non convex penalty.
- Use EM algorithm coordinatewise to get the maximum.
- The parameter expanded version of the prior is easy to find, thus EM algorithm can be easily employed.

Bayes Factor

- Bayes factor is a popular technique for hypothesis testing in the Bayesian paradigm.
- Suppose $\boldsymbol{y}$ is the data and we are to test hypotheses H_1 vs. H_2.
- The Bayes factor $B_{12} = \frac{P(\boldsymbol{y}|H_1)}{P(\boldsymbol{y}|H_2)}$.
- Clearly, $\frac{P(H_1|\boldsymbol{y})}{P(H_2|\boldsymbol{y})} = \frac{P(\boldsymbol{y}|H_1)P(H_1)}{P(\boldsymbol{y}|H_2)P(H_2)}$.
- $P(\boldsymbol{y}|H_k)$, $k = 1, 2$ is obtained by integrating over the parameter space

$$P(\boldsymbol{y}|H_k) = \int P(\boldsymbol{y}|\boldsymbol{\theta}_k, H_k)\pi(\boldsymbol{\theta}_k|H_k)d\boldsymbol{\theta}_k,$$

$\boldsymbol{\theta}_k$ is the parameter corresponding to the hypothesis H_k.

Bayes Factor Contd..

- $3.2 > B_{12} > 1$: not more than a bare mention.
- $10 > B_{12} > 3.2$: substantial.
- $100 > B_{12} > 10$: strong.
- $B_{12} > 100$: decisive.
- The cut-off, however, is context specific.

Bayes Factor Contd..

- For some models, Bayes factor has closed form.
- However, in many models, Bayes factor does not come in closed form.
- *Never try to approximate the integral with the MCMC samples.*
- Rather, a suggestion is to use the Laplace approximation of the integral.
- Otherwise, one can use Gaussian quadrature to evaluate the integral.

g-Prior

$$y_i \sim N(\mu_i, 1/\phi), \;\; i = 1, ..., n.$$

- $\boldsymbol{x}_1, ..., \boldsymbol{x}_p$ correspond to p columns each of length n.
- Let $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_p) \in \{0, 1\}^p$.
- $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)'$ and $\boldsymbol{X}_\gamma$ is an $n \times p_\gamma$ dimensional matrix that includes columns corresponding to $\gamma_i = 1$.
- $\mathscr{M}_\gamma : \boldsymbol{\mu} = \mathbf{1}_n \alpha + \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma$.
- $\boldsymbol{\beta}_\gamma$ is p_γ-dimensional.
- $\boldsymbol{\Theta}_\gamma = \{\boldsymbol{\beta}_\gamma, \alpha, \phi\}$.

g-Prior Contd..

- 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

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

- Let $\mathscr{M}_b$ be any base model. Then

$$BF[\mathscr{M}_{\gamma} : \mathscr{M}_{\zeta}] = \frac{BF[\mathscr{M}_{\gamma} : \mathscr{M}_{b}]}{BF[\mathscr{M}_{\zeta} : \mathscr{M}_{b}]}$$

- The marginal likelihood is given by

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

g-Prior Contd..

- When $\mathscr{M}_b$ is the null model, denoted by $\mathscr{M}_N$

$$BF[\mathscr{M}_\gamma : \mathscr{M}_N] = (1+g)^{\frac{n-p_\gamma-1}{2}}[1+g(1-R_\gamma^2)]^{-(n-1)/2}.$$

- When $\mathscr{M}_b$ is the full model, denoted by $\mathscr{M}_F$

$$BF[\mathscr{M}_\gamma : \mathscr{M}_F] = (1+g)^{\frac{-n+p+1}{2}}[1+g\frac{(1-R_F^2)}{(1-R_\gamma^2)}]^{(n-p_\gamma-1)/2}.$$

- R_γ^2 is the R^2 statistics for the model $\mathscr{M}_\gamma$.
- How to choose g? Can a fixed g be used?
- Barlett paradox and information paradox.