Statistics 700 Homework 4

SMC and MCMC

Due date: 6:00 pm (EST) Oct. 31, 2017

Bayesian Lasso. The Lasso method estimates linear regression coefficients through L_1 constrained least squares, also known as the penalized regression. It provides a way of obtaining a sparse solution for high-dimensional regression problems, which are prevalent in modern applications such as genetics. Here we investigate on the Bayesian interpretation and computation of the LASSO method.

Consider a regression model

$$\boldsymbol{y} = \mu \mathbf{1}_n + X\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{y}$ is the $n \times 1$ vector of responses, $\mathbf{1}_n$ is the $n \times 1$ vector of 1s, μ is the overall mean, $\boldsymbol{\beta}$ is $p \times 1$ vector of regression coefficients, X is the $n \times p$ matrix of standardized regressors, and $\boldsymbol{\epsilon}$ is the $n \times 1$ vector of independent and identically distributed normal errors with mean 0 and unknown variance σ^2. Lasso estimates achieve

$$\min_{\beta} \left\{ (\tilde{\boldsymbol{y}} - X\boldsymbol{\beta})^T (\tilde{\boldsymbol{y}} - X\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

for some $\lambda \geq 0$, where $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \bar{y}\mathbf{1}_n$, $\bar{y}$ is sample average of $\boldsymbol{y}$.

We use the Diabetes Data from Efron et al. (2004). It contains $p = 10$ variables and $n = 442$ measurements. The data is standardized such that the means of all variables are zero, and all variances are equal to one. Refer to `https://artax.karlin.mff.cuni.cz/r-help/library/care/html/efron2004.html` for detailed information about loading the data. References: Efron, B., et al. 2004. Least angle regression (with discussion). Ann. Statist. 32:407–499.

1. Show that the Lasso estimate is equivalent to the Bayes posterior mode with conditional Laplace prior on $\boldsymbol{\beta}$ and Jeffrey's prior on σ^2, i.e.

$$\pi(\boldsymbol{\beta}|\sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} \exp\left[-\lambda \frac{|\beta_j|}{\sqrt{\sigma^2}}\right], \ \pi(\sigma^2) \propto \sigma^{-2}. \tag{1}$$

2. The Laplace distribution can be represented as a scale mixture of normals, i.e.

$$\frac{a}{2} \exp\{-a|z|\} = \int_0^\infty \frac{1}{\sqrt{2\pi s}} e^{-z^2/(2s)} \frac{a^2}{2} e^{-a^2 s/2} ds. \tag{2}$$

Then verify that we can represent the Bayesian Lasso model as

$$\begin{aligned}
y_i|\mu, X_i, \boldsymbol{\beta}, \sigma^2 &\overset{iid}{\sim} N(\mu + X_i\boldsymbol{\beta}, \sigma^2), 1 \le i \le n; \\
\beta_j|\sigma^2, \tau_j^2 &\overset{iid}{\sim} N(0, \sigma^2\tau_j^2), 1 \le j \le p; \\
\tau_j^2 &\overset{iid}{\sim} \frac{\lambda^2}{2} \exp\left\{-\frac{\lambda^2\tau_j^2}{2}\right\}, 1 \le j \le p; \\
\sigma^2 &\sim \pi(\sigma^2).
\end{aligned}$$

Hint: integrate out $\{\tau_j^2\}_{1\le j\le p}$.

3. Assume a flat prior on μ, write down the posterior distribution of $(\boldsymbol{\beta}, \{\tau_j^2\}_{1\le j\le p}, \sigma^2|X, \boldsymbol{y})$ after integrating out μ. Describe and implement a Gibbs sampler of the posterior using the diabetes data ($\lambda = 0.237$).

4. Implement a Metropolis-Hastings algorithm to sample from the posterior of $(\boldsymbol{\beta}, \sigma^2|X, \boldsymbol{y})$ with the diabetes data, using a conditional Laplace prior on $\boldsymbol{\beta}$ and a Jeffrey's prior on σ^2 given in Equation (1). Again, we take $\lambda = 0.237$.

   - Optional (≤ 5 bonus points). Try a reparametrization, e.g. $\log(\sigma^2)$, and discuss whether or not the reparametrization improves the sampling.

5. Compare the results from 3 (Gibbs sampler) and 4 (Metropolis-Hastings) with the Lasso estimates and ordinary least squares estimates. What do you find?

6. In practice, the value of λ is not given. Now implement and compare the following methods of choosing λ in the Bayesian Lasso and the ordinary Lasso.

   - K-fold cross validation for ordinary Lasso.

- (Optional) Empirical Bayes for Bayesian Lasso.
  - At each iteration, run Gibbs sampler using a λ value estimated from the sample of the previous iteration: iteration k uses the Gibbs sampler with $\lambda^{(k-1)}$ and update

$$\lambda^{(k)} = \sqrt{\frac{2p}{\sum_{j=1}^{p} E_{\lambda^{(k-1)}}(\tau_j^2|\tilde{\boldsymbol{y}})}}, \tag{3}$$

replacing $E_{\lambda^{(k-1)}}(\tau_j^2|\tilde{\boldsymbol{y}})$ with averages from the Gibbs sample. Set initial value as

$$\lambda^{(0)} = p\sqrt{\hat{\sigma}_{LS}^2} / \sum_{j=1}^{p} |\hat{\beta}_j^{LS}|,$$

where $\hat{\sigma}_{LS}^2$ and $\hat{\beta}_j^{LS}$ are estimates from the usual least squares procedure. What is your reasoning behind Equation 3? Can you track the $\lambda^{(k)}$ throughout the Gibbs sampler? What do you find?

- Full Bayes for Bayesian Lasso.
  - Put a Gamma prior on λ^2, i.e.

$$\pi(\lambda^2) = \frac{\delta^r}{\Gamma(r)}[\lambda^2]^{r-1}\exp(-\delta\lambda^2),$$

where $r = 1, \delta = 1.78$ for the diabetes data.

Do you obtain similar results from the three methods above? Explain.

Remark: This homework problem is based on "The Bayesian Lasso" by Trevor PARK and George CASELLA, Journal of the American Statistical Association, June 2008, Vol. 103, No. 482, pp. 681-686. **Please work out your own solutions before referring to the original paper.**

Optional Reading. Read at least one of the following papers and post your summary and thoughts on Canvas. Bonus points up to 5 will be rewarded.

1. Cowles, Mary Kathryn, and Bradley P. Carlin. "Markov chain Monte Carlo convergence diagnostics: a comparative review." Journal of the American Statistical Association 91.434 (1996): 883-904.

2. Neal, Radford M. "Probabilistic inference using Markov chain Monte Carlo methods." (1993).

3. Kass, Robert E., et al. "Markov chain Monte Carlo in practice: a roundtable discussion." The American Statistician 52.2 (1998): 93-100.

4. Betancourt, Michael. "The Convergence of Markov chain Monte Carlo Methods: From the Metropolis method to Hamiltonian Monte Carlo." arXiv preprint arXiv:1706.01520 (2017).

5. Salimans, Tim, Diederik Kingma, and Max Welling. "Markov chain monte carlo and variational inference: Bridging the gap." Proceedings of the 32nd International Conference on Machine Learning (ICML-15). 2015.

6. Rosenthal, Jeffrey S. "Minorization conditions and convergence rates for Markov chain Monte Carlo." Journal of the American Statistical Association 90.430 (1995): 558-566.

7. Rosenthal, Jeffrey S. "Asymptotic variance and convergence rates of nearly-periodic Markov chain Monte Carlo algorithms." Journal of the American Statistical Association 98.461 (2003): 169-177.

8. Walker, Stephen, and Nils Lid Hjort. "On Bayesian consistency." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63.4 (2001): 811-821.

9. Walker, Stephen G. "Modern Bayesian asymptotics." Statistical Science (2004): 111-117.

10. Walker, Stephen. "New approaches to Bayesian consistency." Annals of Statistics (2004): 2028-2043.

11. De Blasi, Pierpaolo, and Stephen G. Walker. "Bayesian asymptotics with misspecified models." Statistica Sinica (2013): 169-187.

12. Gelfand, Alan E., and Dipak K. Dey. "Bayesian model choice: asymptotics and exact calculations." Journal of the Royal Statistical Society. Series B (Methodological) (1994): 501-514.

13. Carlin, Bradley P., and Siddhartha Chib. "Bayesian model choice via Markov chain Monte Carlo methods." Journal of the Royal Statistical Society. Series B (Methodological) (1995): 473-484.