MCMC Algorithms For The Gibbs Posterior

Kun Chen[a], Wenxin Jiang[b], Martin A. Tanner[b]
[a]*Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA 52242, USA*
[b]*Department of Statistics, Northwestern University, Evanston, IL 60208, USA*

Contribution

Jiang and Tanner (2008) consider a method of classification using the Gibbs posterior which is directedly constructed from the empirical classification errors. In this paper, we address the computational aspect of the Gibbs posterior.
We note some drawbacks of the original algorithm based on the Gibbs sampler with augmented latent variables, and propose an alternative method based on the Metropolis algorithm. Numerical performance of the algorithms are examined and compared via simulated data. We find that the Metropolis algorithm produces good classification results at an improved speed of computation.

Gibbs Posterior

Problem: Predict $y \in \{0,1\}$ based on $x \in R^K$ with iid data $D^n = (y^{(i)}, x^{(i)})_1^n$, $K >> n$.
Classification rule: $I[x^T\beta > 0]$
Risk: $R(\beta) = P^*\{y \neq I(x^T\beta > 0)\}$

Standard Bayesian method, e.g. Lee et. al. (2003): generating $\beta \in \Omega$ from

$$e^{-n\{-\frac{1}{n}l(D^n|\beta)\}}\pi(d\beta)/\int_\Omega e^{-n\{-\frac{1}{n}l(D^n|\beta)\}}\pi(d\beta)$$

where π is a prior, and $l(D^n|\beta)$ is the log likelihood function based on, e.g., probit linear regression.

Gibbs posterior: generating $\beta \in \Omega$ from

$$\pi(d\beta|D^n) = e^{-n\psi R_n(\beta)}\pi(d\beta)/\int_\Omega e^{-n\psi R_n(\beta)}\pi(d\beta)$$

where $\psi > 0$ is the inverse temperature, and R_n is a sample analog of the misclassification risk R, e.g.

$$R_n = \frac{1}{n}\sum_{i=1}^n I[y^{(i)} \neq A_i]$$
$$= -\frac{1}{\psi n}\sum_{i=1}^n \ln(A_i e^{\psi(y^{(i)}-1)} + (1-A_i)e^{-\psi y^{(i)}})$$

where $A_i = I[(x^{(i)})^T\beta > 0]$.

Sec7 Algorithm (Old)

Smoothed risk: Replace A_i by $\Phi_i = \Phi(\sigma^{-1}(x^{(i)})^T\beta)$.
Prior: Normal binary prior on $(\gamma, \beta_1, \tilde{\beta}_\gamma)$ where $\gamma_j = I[\beta_j \neq 0]$, $\beta_1 \in \{-1, +1\}$ and $\tilde{\beta}_\gamma = (\beta_j)_{j>1, \gamma_j=1}$.

- $\beta_1|\gamma = \pm 1$ with $p = 0.5$, $\tilde{\beta}_\gamma|\gamma \sim N(0, V_\gamma)$;
- $\gamma_1 = 1$, $\gamma_2 = 1$ (Intercept), and $(\gamma_j)_{j=3}^K$ are $iid\ Bernolli(\lambda)$ with size restriction $\bar{r}$.

Posterior Structure: Can be viewed as likelihood for a mixture of two binary models.
$\Longrightarrow$ **Gibbs sampler with latent variables**.

Drawbacks:

- σ large $\Rightarrow$ R_n not close to empirical risk $\Rightarrow$ Bad classification performance;
- σ small $\Rightarrow$ Very slow convergence.

Metropolis Algorithm (New)

- Works for unsmoothed empirical risk.
- Classical "between" steps to propose deletion, addition or swapping of variables. Incorporate "within" step for updating parameters that can't be integrated away.

BETWEEN steps I and II: Update β to β' with model indexes changing.
I: (add/delete): Randomly choose an index $j \in \{3, ..., , K\}$.
I(i) ($\gamma_j = 1$,delete) Propose $\gamma_j' = 0$ with acceptance prob. $= \min\left\{1, \frac{\pi(\beta'|D^n)q(\beta_j)}{\pi(\beta|D^n)}\right\}$.
I(ii) ($\gamma_j = 0$,add) Propose $\gamma_j' = 1$ with acceptance prob. $= \min\left\{1, \frac{\pi(\beta'|D^n)}{\pi(\beta|D^n)q(\beta_j')}\right\}$.
II: (swap): Randomly choose a $\gamma_k' = 0$ and $\gamma_l' = 1$. Propose $\gamma_k' = 1$ and $\gamma_l' = 0$, with acceptance prob. $= \min\left\{1, \frac{\pi(\beta'|D^n)q(\beta_l)}{\pi(\beta|D^n)q(\beta_k')}\right\}$.
WITHIN step III: Update β' to β^* with model indexes fixed and with the nonzero values of β''s changed.
III: (within): Propose a move from β_1' to $\beta_1^* \sim Bin(1, 0.5)$, as well as a move from $\tilde{\beta}_\gamma'$ to $\tilde{\beta}_\gamma^* \sim N(\tilde{\beta}_\gamma', \sigma_q^2 I_{(\sum_{j=2}^K \gamma_j)\times(\sum_{j=2}^K \gamma_j)})$, with acceptance prob $= \min\left\{1, \frac{\pi(\beta^*|D^n)}{\pi(\beta'|D^n)}\right\}$.

Simulation

3 methods for comparison:

- Sec7 algorithm for $\sigma = 0.2$ (Sec7.2) and $\sigma = 0.02$ (Sec7.02);
- Metropolis algorithm;
- Lee.et.al. algorithm.

2 models for data generating with $n_{train} = 30$, $n_{test} = 200$, $K = 50$, and $x_2 = 1$ (Intercept):

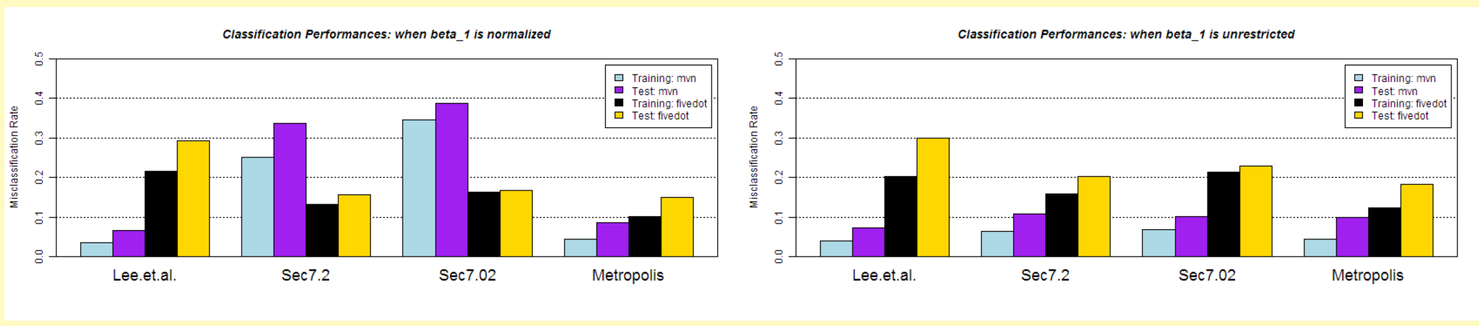
- "mvn" model: generated from multivariate normal distribution with 2 informative predictors $(x_1; x_3)$.
  $\longleftrightarrow$ a "correctly" specified model;
- "$fivedot$" model: the design points of the 2 informative predictors $(x_1; x_3)$ form a set of five points.
  $\longleftrightarrow$ a "misspecified" model.

2 measures for performance:

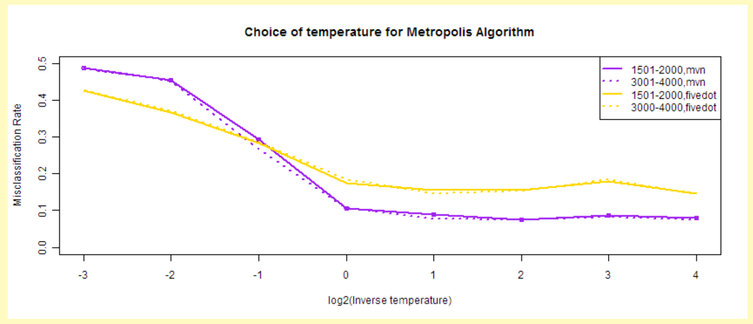
- Misclassification rate: Calculate mean error over 500 iterations after 1500 burn-in, then average over $n_{rep} = 50$ simulated datasets.
- Computation time.

Results

We used Octave on a linux machine (Pentium 4HT, 3.2 GHz, 512 MB RAM). It took per unit time about 7 min (for 2000 iterations) for the Sec7 methods, and about 5 min for the Lee.et.al. method. The Metropolis method takes about 2 min when steps I,II, III are cycled in the iterations. (The time will decrease when I or II is randomly chosen in each iteration.)



- Under the "$fivedot$" model, the Lee.et.al. algorithm performs the worst, while all other algorithms have smaller testing errors. The performance is different because Lee.et.al. method is likelihood-based which requires the correct model specification. Here, however, the linear classification rules are misspecified that it will not generate the best possible Bayes rule.
- Under the "mvn" model, the Lee.et.al. algorithm performs well, since its probability model is very close to the true model of logistic regression. The performance of the Gibbs posterior using Metropolis algorithm is still comparable, which directly uses the empirical classification error to construct the posterior.
- We conclude that the Metropolis method is generally preferable to the Sec7 methods, and produces good classification results much faster than all other methods. It can also work much better than the Lee.et.al. method when there is model misspecification and is still competitive when the model is correctly specified.



References

Brown, P., Vannucci, M. and Fearn (2002). Bayesian model averaging with selection of regressors. Journal of Royal Statistical Society B, 64, 519-536.

Jiang, W. and Tanner, M. A. (2008). Gibbs posterior for variable selection in high dimensional classification and data mining. Annals of Statistics. 36, 2207-2231.

Lee, K. E., Sha, N., Dougherty, E. R., Vannucci, M. and Mallick, B. K. (2003). Gene selection: a Bayesian variable selection approach. Bioinformatics 19, 90-97.