Bayesian Statistics

Debdeep Pati
Florida State University

March 10, 2017

True model unknown

- In the DDE application, we assumed that we knew in advance that the probit model with pre-specified predictors was appropriate.
- There is typically substantial uncertainty in the model & it is more realistic to suppose that there is a list of a priori plausible models.
- Typical Strategy: sequentially change model until a good fit is produced, and then base inferences/predictions on the final selected model.
- Strategy is flawed in ignoring uncertainty in the model selection process - leads to major bias in many cases.

Bayes Model Uncertainty

- Let $M \in \mathcal{M}$ denote a model index, with $\mathcal{M}$ a list of possible models.
- To allow for model uncertainty, Bayesians first choose:
  - A prior probability for each model: $P(M = m) = \pi_m, m \in M$.
  - Priors for the coefficients within each model, $\pi(\theta_m), m \in M$.
- Given data, y, the posterior probability of model $M = m$ is

$$\hat{\pi}_m = P(M = m \mid y) = \frac{\pi_m L_m(y)}{\sum_{l \in \mathcal{M}} \pi_l L_l(y)}$$

where $L_m(y) = \int L(y \mid M = m, \theta_m)\pi(\theta_m)d\theta_m$ is the marginal likelihood for the model $M = m$.

Some comments

- In the absence of prior knowledge about which models in the list are more plausible, one often lets $\pi_m = 1/|\mathcal{M}|$, with $|\mathcal{M}|$ the number of models.
- The highest posterior probability model is then the model with the highest marginal likelihood.
- Unlike the maximized likelihood, the marginal likelihood has an implicit penalty for model complexity.
- This penalty is due to the integration across the prior, which is higher in larger models.

Bayes factors

- The Bayes factor (BF) can be used as a summary of the weight of evidence in the data in favor of model m_1 over model m_2.
- The BF for model m_1 over m_2 is defined as the ratio of posterior to prior odds, which is simply:

$$BF_{12} = \frac{L_1(y)}{L_2(y)}$$

  a ratio of marginal likelihoods.
- Values of $BF_{12} > 1$ suggest that model m_1 is preferred, with the weight of evidence in favor of m_1 increasing as BF_{12} increases.

Bayesian model averaging

- Posterior model probabilities can be used for model selection and inferences.
- When focus is on prediction, BMA preferred to model selection (Madigan and Raftery, 1994)
- To predict y_{n+1} given x_{n+1}, BMA relies on:

$$f(y_{n+1} \mid x_{n+1}, \mathbf{y}, \mathbf{x}) = \sum_{m \in \mathcal{M}} \hat{\pi}_m \int L(y_{n+1} \mid x_{n+1}, M = m, \theta_m) \times \pi(\theta_m \mid M = m, \mathbf{y}, \mathbf{x}) d\theta_m$$

where $\hat{\pi}_m = P(M = m \mid \mathbf{y}, \mathbf{x})$ is the posterior probability of the model.

Bayesian model averaging

- Computation of the posterior model probabilities requires calculation of the marginal likelihoods, $L_m(y)$
- These marginal likelihoods are not automatically produced by typical MCMC algorithms
- Routine implementations rely on the Laplace approximation (Tierney and Kadane, 1986; Raftery, 1996)
- In large model spaces, it is not feasible to do calculations for all the models, so search algorithms are used.
- Refer to Hoeting et al. (1999) for a tutorial on BMA

Bayesian variable selection

- Suppose we start with a vector of p candidate predictors, $x_i = (x_{i1}, \ldots, x_{ip})'$.
- A very common type of model uncertainty corresponds to uncertainty in which predictors to include in the model.
- In this case, we end up with a list of 2^p different models, corresponding to each of the p candidate predictors being excluded or not.

Stochastic Search variable selection (SSVS)

- George and McCulloch (1993, 1997) proposed a Gibbs sampling approach for the variable selection problem.
- Similar approaches have been very widely used in applications.
- The SSVS idea will be illustrated through a return to the DDE and preterm birth application

Bayes Variable Selection in Probit Regression

- Earlier we focused on the model, $P(y_i = 1 \mid x_i, \beta) = \Phi(x_i'\beta)$, with y_i an indicator of premature delivery.
- Previously, we chose a $\mathrm{N}_7(0, 4I)$ prior for β, assuming all 7 predictors were included.
- To account for uncertainty in subset selection, choose a mixture prior:

$$\pi(\beta) = \prod_{j=1}^{p} \{\delta_0(\beta_j) p_{0j} + (1 - p_{0j}) N(\beta_j; 0, c_j^2)\}$$

where p_{0j} is the prior probability of excluding the j-th predictor by setting its coefficient to 0

SSVS - Full conditional distributions

- Data augmentation Gibbs sampler described earlier easily adapted
- Sample from conditional posterior of β_j , for $j = 1, \ldots, p$,

$$\pi(\beta_j \mid \beta_{-j}, \mathbf{z}, \mathbf{y}, \mathbf{x}) = \hat{p}_j \delta_0(\beta_j) + (1 - \hat{p}_j) \mathrm{N}(\beta_j; E_j, V_j)$$

where $V_j = (c_j^{-2} + X_j' X_j)^{-1}, E_j = V_j X_j' (\mathbf{z} - X_{-j}\beta_{-j}), X_j = j$ th column of X, $X_{-j} = X$ with j th column excluded, $\beta_{-j} = \beta$ with j th element excluded, and

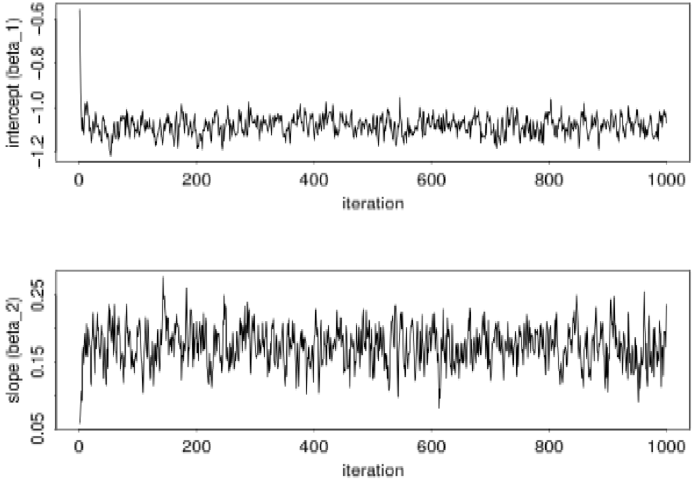
$$\hat{p}_j = \frac{p_{0j}}{p_{0j} + (1 - p_{0j}) \frac{N0;0,c_j^2)}{N(0;E_j,V_j)}}$$

is the conditional probability of $\beta_j = 0$.

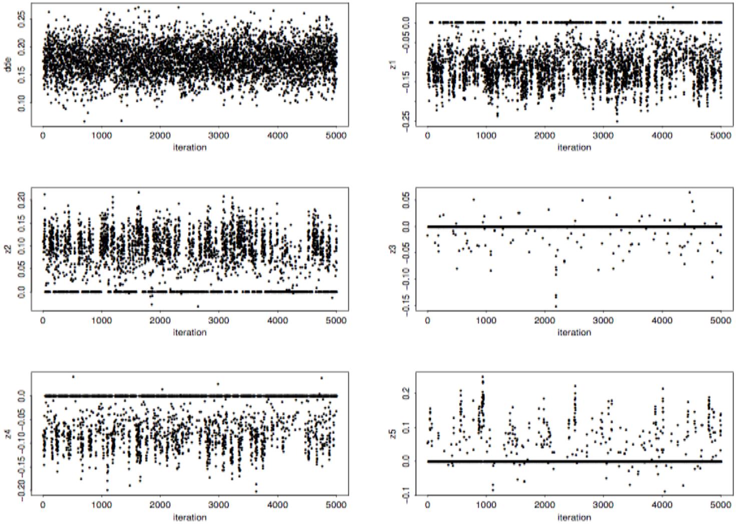
SSVS - Comments

- After convergence, generates samples of models, corresponding to subsets of the set of p candidate predictors, from the posterior distribution.
- Based on a large number of SSVS iterations, we can estimate posterior probabilities for each of the models.
- For example, the full model may appear in 10% of the samples collected after convergence, so that model would be assigned posterior probability of 0.10.
- To summarize, one can present a table of the top 10 or 100 models
- Potentially more useful to calculate marginal inclusion probabilities.

Samples from the posterior (normal prior)



Samples from the posterior (SSVS)



Comments

- Samples congregate on 0 for the regression coefficient for predictors that are not as important.
- Such samples correspond to models with that predictor excluded.
- Even though the prior probabilities of exclusion are the same, posterior probabilities vary greatly for the different predictors.

Posterior summaries - Normal prior Analysis

Parameter	Mean	Median	SD	95% credible interval
β_1	-1.08	-1.08	0.04	(-1.16, -1.01)
β_2	0.17	0.17	0.03	(0.12, 0.23)
β_3	-0.13	-0.13	0.04	(-0.2, -0.05)
β_4	0.11	0.11	0.03	(0.05, 0.18)
β_5	-0.02	-0.02	0.03	(-0.08, 0.05)
β_6	-0.08	-0.08	0.04	(-0.15, -0.02)
β_7	0.05	0.06	0.06	(-0.07, 0.18)

Posterior summaries - Mixture prior Analysis

Parameter	Mean	Median	SD	95% CI	$\Pr(\beta_j = 0 \mid \text{data}$
β_1	-1.05	-1.05	0.03	(-1.12, -0.99)	0.00
β_2	0.18	0.18	0.03	(0.12, 0.23)	0.00
β_3	-0.08	-0.09	0.06	(-0.19, 0.00)	0.36
β_4	0.05	0.00	0.06	(0.00, 0.16)	0.50
β_5	0.00	0.00	0.01	(0.00, 0.00)	0.98
β_6	-0.02	0.00	0.04	(-0.13, 0.00)	0.72
β_7	0.01	0.00	0.02	(0.00, 0.1)	0.93

Posterior probabilities of visited models

	$\widehat{\pi}_m$	Model Indicator
1	0.24981301421092	1 1 0 0 0 0 0
2	0.225878833208676	1 1 1 1 0 0 0
3	0.196958364497632	1 1 1 1 0 1 0
4	0.139865370231862	1 1 1 0 0 0 0
5	0.0363999002742458	1 1 0 0 0 1 0
6	0.0304163550236849	1 1 0 1 0 0 0
7	0.0274245823984044	1 1 1 0 0 1 0
8	0.0206930939915233	1 1 0 0 0 0 1
9	0.0177013213662428	1 1 1 1 0 0 1
10	0.012216404886562	1 1 1 0 0 0 1

SSVS - Comments

- In 4,000 Gibbs iterations only $26/128 = 20.3\%$ of the models were visited
- There wasnt a single dominant model, but none of the models excluded the intercept or DDE slope.
- All of the better models included the 3rd & 5th of the 5 possible confounders

Some Limitations of SSVS

- High autocorrelation in model search - sampling from conditional posterior given other predictors currently in model
- No guarantee of finding the best model - may remain for long intervals in local regions of the model space
- As the number of predictors increases, model space enormous - convergence of MCMC may be effectively impossible
- May obtain poor estimates of posterior model probabilities - most models are never visited & many are only visited once

Issues in Large Model Spaces

- Results from SSVS can be difficult to interpret - there may be 100s or 1000s of models with very similar posterior probabilities
- Makes it clear that it is problematic to base inferences on any one selected model in a large model space
- Seldom enough information in the data to definitively conclude in favor of one model
- This issue is swept under the rug by optimization-based approaches
- Critical to account for model uncertainty in inferences

Marginal Inclusion Probabilities

- As posterior model probabilities are not very helpful, one often focuses instead on marginal inclusion probabilities
- Provides a weight of evidence that a predictor should be included
- Can be artificially small if there are several correlated predictors - correlated predictors are big problem in general!
- Even in the absence of correlation, Bayes multiplicity adjustments can lead to smallish inclusion probabilities for important predictors

Some Comments on Multiplicity Adjustments

- If independent Bernoulli priors are chosen for the variable inclusion indicators, no adjustment for multiplicity
- In such a case, one gets more and more false positives as the number of candidate predictors increases
- A Bayes adjustment for multiplicity can proceed by including dependence in the hypotheses
- Most common approach is to choose a beta hyperprior for probability of including a variable

Multiplicity Adjustments (Continued)

- If a beta hyperprior is chosen, then the incorporation of a large number of “null” predictors, will lead to updating
- The inclusion probability will then have a posterior that is increasing concentrated near zero
- The more null predictors that are included, the more one needs overwhelming evidence in the data to assign a high inclusion probability to an important predictor.
- Tends to lead to very small inclusion probabilities for the vast majority of predictors, with the small number of important predictors assigned probabilities higher but not close to one.