Estimating Posterior Model Probabilities via Bayesian Model Based Sampling

Merlise Clyde

2021-10-15

Outline

- Canonical Regression Model & Bayesian Model Averaging
- Estimation via MCMC Monte Carlo Frequencies
- Probability Proportional to Size Sampling in Finite Populations
- Adaptive Independent Metropolis/Adaptive Importance Sampling

Canonical Regression Model

- Observe response vector $\mathbf{Y}$ with predictor variables $X_1 \ldots X_p$.
- Model for data under a specific model M_γ:

$$\mathbf{Y} \mid \alpha, \beta_\gamma, \phi, M_\gamma \sim \mathsf{N}(\mathbf{1}_n \alpha + \mathbf{X}_\gamma \beta_\gamma, \mathbf{I}_n/\phi)$$

- Models M_γ encoded by $\gamma = (\gamma_1, \ldots \gamma_p)^T$ binary vector with $\gamma_j = 1$ indicating that X_j is included in model M_γ where

$$\gamma_j = 0 \Leftrightarrow \beta_j = 0$$
$$\gamma_j = 1 \Leftrightarrow \beta_j \neq 0$$

- $\mathbf{X}_\gamma$: the $n \times p_\gamma$ design matrix for model M_γ
- β_γ: the p_γ vector of non-zero regression coefficients under M_γ
- intercept α, precision ϕ common to all models

Bayesian Model Averaging (BMA)

- prior distributions on all unknowns $(M_\gamma, M_\gamma, \alpha_{M_\gamma}, \phi_{M_\gamma})$ and turn the Bayesian crank to get posterior distributions!
- for **nice** priors, we can integrate out the parameters $\theta_\gamma = (\beta_\gamma, \alpha_{M_\gamma}, \phi_{M_\gamma})$ to obtain the marginal likelihood of M_γ

$$p(\mathbf{Y} \mid M_\gamma) = \int p(\mathbf{Y} \mid \theta_\gamma, M_\gamma) p(\theta_\gamma \mid M_\gamma) d\theta_\gamma$$
$$p(M_\gamma \mid \; Y) = \frac{p(\mathbf{Y} \mid M_\gamma) p(M_\gamma)}{\sum_{\gamma \in \Gamma} p(\mathbf{Y} \mid M_\gamma) p(M_\gamma)}$$

- posterior distribution of quantities Δ of interest under BMA
$$\sum_{\gamma \in \Gamma} p(M_\gamma \mid \mathbf{Y}) p(\Delta \mid \mathbf{Y}, M_\gamma)$$
- estimation $\mathsf{E}[\mu \mid \mathbf{Y}]$, $p(\mathbf{Y}^* \mid \mathbf{Y})$, marginal inclusion probabilities $P(\gamma_j = 1 \mid \mathbf{Y})$

MCMC Sampling from Posterior Distribution

Use a sample of models from Γ to approximate the posterior distribution of models

- design a Markov Chain to transition through Γ with stationary distribution $p(M_\gamma \mid \mathbf{Y})$
$$p(M_\gamma \mid \mathbf{Y}) \propto p(\mathbf{Y} \mid M_\gamma) p(M_\gamma)$$
- propose a new model from $q(\gamma^* \mid \gamma)$
- accept moving to γ^* with probability
$$\mathsf{MH} = \max(1, \frac{p(M_\gamma * \mid \mathbf{Y}) p(M_\gamma^*)/q(\gamma^* \mid \gamma)}{p(M_\gamma \mid \mathbf{Y}) p(M_\gamma)/q(\gamma)})$$
- otherwise stay at model M_γ
- models are sampled proportional to their posterior probabilities as $T \to \infty$

Estimation in BMA

Estimate the probabilities of models via Monte Carlo frequencies of models or ergodic averages

$$\begin{aligned} p(\widehat{M_\gamma \mid} \mathbf{Y}) &= \frac{\sum_{t=1}^T I(M_t = M_\gamma)}{T} \\ &= \frac{\sum_{\gamma \in S} n_\gamma I(M_\gamma \in S)}{\sum n_\gamma} \end{aligned}$$

- $T = \#$ MCMC samples
- S is the collection of unique sampled models
- n_γ is the frequency of model M_γ in S
- $n = \sum_{\gamma \in S} n_\gamma$ total number of unique models in the sample
- asymptotically unbiased as $T \to \infty$

Monte Carlo Frequencies

- fundamentally unsound to a Bayesian ! (O'Hagan 1987, *The Statistician*)
- ignores observed information in the marginal likelihoods $\times$ prior probabilities!
- Can view MH as a form of Probability Proportional to Size Sampling (PPS) With Replacement
- can we do better using ideas from Finite Population Sampling?
  - Let $q(M_i)$ be the probability of selecting M_i
  - Goal is to estimate $C = \sum_i^N p(\mathbf{Y} \mid M_i)p(M_i)$
    - Hansen-Hurwitz (HH)
    - Horvitz-Thompson (HT)
    - Hájek
    - Basu/Bayes

Hansen-Hurwitz (HH)

- Hansen-Hurwitz (1943) may be viewed as an importance sampling estimate
$$\hat{C} = \frac{1}{n}\sum_i^n \frac{n_i p(\mathbf{Y} \mid M_i)p(M_i)}{q(M_i)}$$
- If we have "perfect" samples from the posterior then $q(M_i) = \frac{p(\mathbf{Y}|M_i)p(M_i)}{C}$ and recover C!
- Since C is unknown, apply the ratio HH estimator (or self-normalized IS)
$$\hat{C} = \frac{\frac{1}{n}\sum_i^n \frac{n_i p(\mathbf{Y}|M_i)p(M_i)}{q(M_i)}}{\frac{1}{n}\sum_i^n \frac{1}{q(M_i)}} = \left[\frac{1}{n}\sum_i \frac{n_i}{p(\mathbf{Y} \mid M_i)p(M_i)}\right]^{-1}$$

. . .

But this recovers the "infamous" harmonic mean estimator of Newton & Raftery (1994) - while unbiased, it's is highly unstable!

Horvitz-Thompson (HT)

- inclusion probability that $\gamma_i \in S$ - under sampling with replacement $\pi_i = 1 - (1 - q(M_i))^{\mathrm{T}}$

- HT estimate of normalizing constant:

$$\hat{C} = \frac{1}{n}\sum_{i\in n}\frac{p(\mathbf{Y}\mid M_i)p(M_i)}{\pi_i}$$

(dominates HH, unique hyper-admissible estimate of C)

- Hájek (1971) estimator uses an auxilary variable $A_i > 0$, where we expect $p(\mathbf{Y}\mid M_i)p(M_i) \propto A_i$, with $A \equiv \sum_{i=1}^{N} A_i$

$$\hat{C} = \frac{\sum_{i=1}^{n}\frac{p(\mathbf{Y}|M_i)p(M_i)}{\pi_i}}{\sum_{i=1}^{n}\frac{A_i/A}{\pi_i}}$$

may be preferable when $p(\mathbf{Y}\mid M_i)p(M_i)$ are weakly correlated with π_i or when n is not fixed

Basu and Bayes

Basu's (1971) famous circus example illustrated potential problems with the Horvitz-Thompson estimator (similar problem arises with IS)

- violates the likelihood principle
- once we have samples, $p(\mathbf{Y}\mid M_i)p(M_i)$ are fixed and the sampling probabilities are not relevant
- only randomness is for the remaining units that were not sampled. (which is related to the sampling design)
- Basu's estimate (using $\pi_i = A_i/A$),

$$C = \sum_{i\in S} p(\mathbf{Y}\mid M_i)p(M_i) + \frac{1}{n}\left(\sum_{i\in S}\frac{p(\mathbf{Y}\mid M_i)p(M_i)}{\pi_i}\right)\times\left(\sum_{i\notin S}\pi_i\right)$$

- conditions on the observed data sum and estimates remaining

Model Based Methods

Basu (1971)'s estimate of the total can be justified as a "super-population" Model Based approach (Meeden and Ghosh, 1983)

- Let $m_i = p(\mathbf{Y}\mid M_i)p(M_i)$

$$m_i \mid \pi_i \overset{\text{ind}}{\sim} N(\pi_i\beta, \sigma^2\pi_i^2) \quad (1)$$
$$p(\beta,\sigma^2) \propto 1/\sigma^2 \quad (2)$$

- posterior mean of β is $\hat{\beta} = \frac{1}{n}\sum_{i \in S} \frac{m_i}{\pi_i}$ (the HT of the total)
- using the posterior predictive for $m_i \notin S$, $\mathsf{E}[m_i \mid m_j \in S] = \pi_i \hat{\beta}$

$$C = \sum_{i \in \Gamma} m_i = \sum_{i \in S} m_i + \sum_{i \notin S} m_i$$

$$\hat{C} = \sum_{i \in S} m_i + \sum_{i \notin S} \hat{\beta}\pi_i = \sum_{i \in S} m_i + \left[\frac{1}{n}\sum_{i \in S}\frac{m_i}{\pi_i}\right]\sum_{i \notin S}\pi_i$$

Final Posterior Estimates

- estimate of posterior probability M_γ for $M_\gamma \in S$

$$\frac{p(\mathbf{Y} \mid M_\gamma)p(M_\gamma)}{\sum_{i \in S} p(\mathbf{Y} \mid M_i)p(M_i) + \frac{1}{n}\sum_{i \in S}\frac{p(\mathbf{Y}|M_i)p(M_i)}{\pi_i}\sum_{i \in S}(1-\pi_i)}$$

- estimate of all models in $\Gamma - S$ from the predictive distribution

$$\frac{\frac{1}{n}\sum_{i \in S}\frac{p(\mathbf{Y}|M_i)p(M_i)}{\pi_i}\sum_{i \in S}(1-\pi_i)}{\sum_{i \in S} p(\mathbf{Y} \mid M_i)p(M_i) + \frac{1}{n}\sum_{i \in S}\frac{p(\mathbf{Y}|M_i)p(M_i)}{\pi_i}\sum_{i \in S}(1-\pi_i)}$$

- Uses renormalized marginal likelihoods of sampled models
- easy to compute marginal inclusion probabilities
- Other mean/variance assumptions for the super-population model lead to other estimates for C, $p(M_\gamma \mid \mathbf{Y})$, etc
- What about $\mathsf{E}[| \mathbf{Y}]$, $\mathsf{E}[\mathbf{X} \mid \mathbf{Y}]$, $\mathsf{E}[\mathbf{Y}^* \mid \mathbf{Y}]$ or $p(\Delta \mid \mathbf{Y})$?

Choice for $q(M_\gamma)$ or $\mathbf{A}_{M_\gamma}$?

- The joint posterior distribution of γ (dropping $\mathbf{Y}$) may be factored:

$$p(M_\gamma \mid \mathbf{Y}) \equiv p(\gamma \mid \mathbf{Y}) = \prod_{j=1}^{p} p(\gamma_j \mid \gamma_{<j})$$

where $\gamma_{<j} \equiv \{\gamma_k\}$ for $k < j$ and $p(\gamma_1 \mid \gamma_{<1}) \equiv p(\gamma_1)$.

- As γ_j are binary, re-express as

$$p(\gamma \mid \mathbf{Y}) = \prod_{j=1}^{p}(\rho_{j|<j})^{\gamma_j}(1-\rho_{j|<j})^{1-\gamma_j}$$

where $\rho_{j|<j} \equiv \Pr(\gamma_j = 1 \mid \gamma_{<j})$ and $\rho_{1|<1} = \rho_1$, the marginal probability.

- Product of **Dependent** Bernoullis

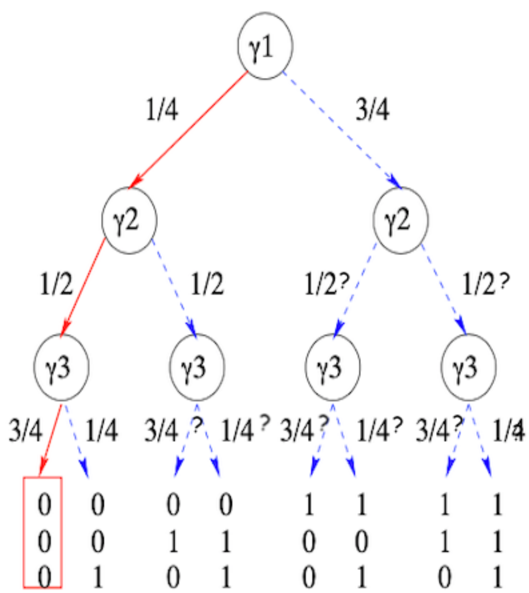
Global Adaptive MCMC Proposal

Factor proposal

$$q(\gamma) = \prod_{j=1}^{p} q(\gamma_j \mid \gamma_{<j}) = \prod_{j} \mathsf{Ber}(\hat{\rho}_{j|<j})$$

- Note: $\Pr(\gamma_j = 1 \mid \gamma_{<j}) = \mathsf{E}[\gamma_j = 1 \mid \gamma_{<j}]$
- Fit a sequence of p regressions γ_j on $\gamma_{<j}$

$$\begin{aligned}
\gamma_1 &= \mu_1 + \epsilon_1 \\
\gamma_2 &= \mu_2 + \beta_{21}(\gamma_1 - \mu_1) + \epsilon_2 \\
\gamma_3 &= \mu_3 + \beta_{31}(\gamma_1 - \mu_1) + \beta_{32}(\gamma_2 - \mu_2) + \epsilon_3 \\
&\vdots \\
\gamma_p &= \mu_p + \beta_{p1}(\gamma_1 - \mu_1) \ldots + \beta_{p-1\,p-1}(\gamma_{p-1} - \mu_{p-1}) + \epsilon_p
\end{aligned}$$



Compositional Regression

Approximate model

$$\gamma \sim \mathsf{N}(\mu, \Sigma_\gamma)$$

- Wermouth (1980) compositional regression

$$\mathbf{G} = \mathbf{1}_T\mu^T + (\Gamma - \mathbf{1}_T\mu^T)\mathbf{B} + \epsilon$$

- $\mathbf{G}$ is $T \times p$ matrix where row t is γ_t
- μ is the p dimensional vector of $\mathsf{E}[\gamma]$
- $\Sigma_\gamma = \mathbf{U}^T\mathbf{U}$ where $\mathbf{U}$ is upper triangular Cholesky decomposition of covariance matrix of γ ($p \times p$)
- $\mathbf{B}^T = \mathbf{I}_p - diag(\mathbf{U})^{-1}\mathbf{U}^{-T}$ (lower triangle)
- $\mathbf{B}$ is a $p \times p$ upper triangular matrix with zeros on the diagonal and regression coefficients for jth regression in row j

Estimators of B and μ

- OLS is BLUE and consistent, but $\mathbf{G}$ may not be full rank
- apply Bayesian Shrinkage with "priors" on μ (non-informative or Normal) and Σ (inverse-Wishart)
- pseudo-posterior mean μ is the current estimate of the marginal inclusion probabilities $\bar{\gamma} = \hat{\mu}$
- use pseudo-posterior mean for Σ
- one Cholesky decomposition provides all coefficients for the p predictions for proposing γ^*
- constrain predicted values $\hat{\rho}_{j|<j} \in (\delta, 1 - \delta)$
- generate $\gamma_j^* \mid \gamma_{<j}^* \sim \mathsf{Ber}(\hat{\rho}_{j|<j})$
- use as proposal for Adaptive Independent Metropolis-Hastings or Importance Sampling (Accept all) -or- Samping Without Replacement (todo)

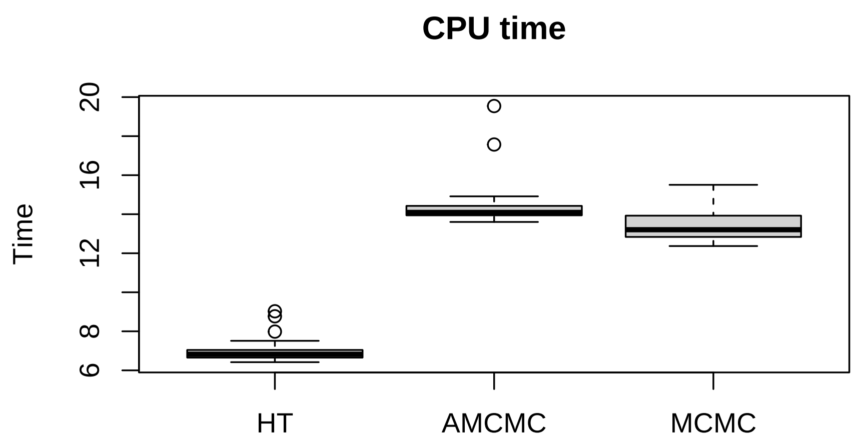
Simulation

- `tecator` data (Griffin et al (2021))
- a sample of $p = 20$ variables
- compare
  - enumeration to

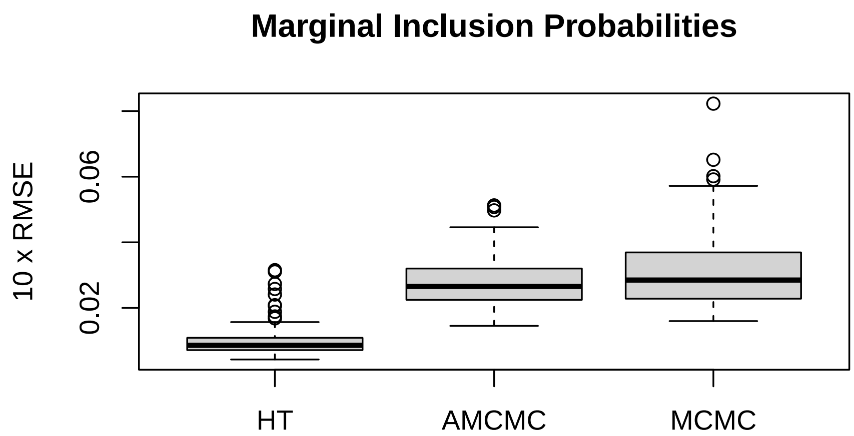
- MCMC with add, delete, and swap moves with q
- Adaptive Independent MCMC
- Importance Sampling with HT

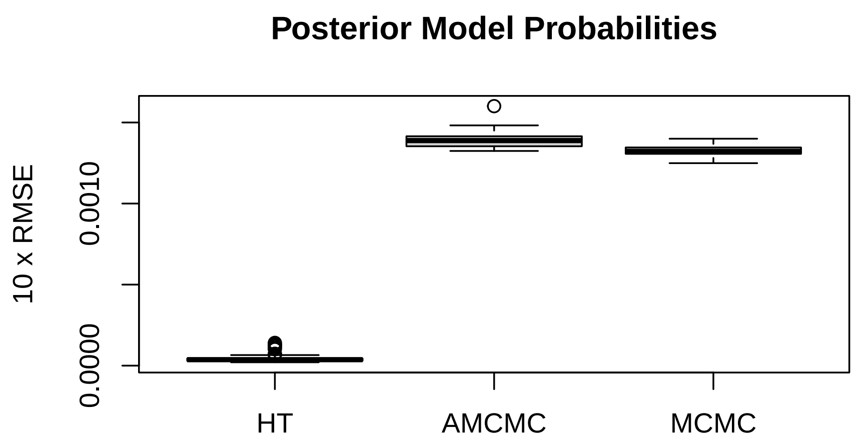
- same settings `burnin.it`, `MCMC.it`, `thin`

```
load("sim_code/tecator-time.dat")
boxplot(time, main="CPU time", ylab = "Time")
```



MSE Comparision





Continued Adaptation ?

- can update Cholesky with rank 1 updates with new models

- how to combine IS with MH samples (weighting) ?
- HT/Hajek - computational complexity involved if we need to compute inclusion probability for all models based on updates (previous models and future models)
- Basu (1971) approach works with PPS-WOR take $\pi_i \propto A_i \equiv q(\gamma_i)$ (adaptation?)

Refinements

- Want to avoid MCMC for
  - pseudo Bayesian posteriors used to learn proposal distribution in sample design for models
  - estimation of posterior model probabilities in model-based approaches (ie learning β, sampling from predictive distribution)
  - estimation of general quantities under BMA?
- avoid infinite regret
- more general models?

Summary

- Adaptive Independent Metropolis proposal for models (use in more complex IS)
- Use observed values of unique marginal likelihoods of models for estimating posterior distribution
- Bayes estimates from MC output (solution to O'Hagan '73?)

::: ::::