STSCI 4780 Lab09
Bayesian calibration and frequentist performance of Bayesian procedures

Tom Loredo, CCAPS & DSS, Cornell University

Bayesian Inference and the Joint Distribution

Recall that Bayes's theorem comes from the *joint distribution for data and hypotheses* (parameters/models):

$$\begin{aligned} p(\theta, D|M) &= p(\theta|M)\, p(D|\theta, M) \\ &= p(D|M)\, p(\theta|D, M) \end{aligned}$$

Bayesian inference takes $D = D_{\text{obs}}$ and solves RHS for the posterior:

$$\to p(\theta|D_{\text{obs}}, M) = \frac{p(\theta|M)p(D_{\text{obs}}|\theta, M)}{p(D_{\text{obs}}|M)}$$

MCMC is nontrivial technology for building RNGs to sample θ values from the *intractable posterior*, $p(\theta|D_{\text{obs}}, M)$

Posterior sampling is hard, but sampling from the other distributions is often easy:

- Often easy to draw θ^* from $\pi(\theta)$
- Typically easy to draw D_{sim} from $p(D|\theta, M)$
- Thus we can sample the joint for (θ, D) by sequencing:

$$\theta^* \sim \pi(\theta)$$
$$D_{\text{sim}} \sim p(D|\theta^*, M)$$

- $\{D_{\text{sim}}\}$ from above are samples from prior predictive,

$$p(D|M) = \int d\theta \, \pi(\theta) p(D|\theta, M)$$

Now note that $\{D_{\text{sim}}, \theta\}$ with $\theta \sim p(\theta|D_{\text{sim}}, M)$ (via MCMC) are also samples from the joint distribution

Joint distribution methods check the consistency of these two joint samplers to validate a posterior sampler implementation

Example: "Calibration" of credible regions

How often may we expect an HPD region with probability P to include the true value if we analyze many datasets? I.e., what's the frequentist coverage of an interval rule $\Delta(D)$ defined by calculating the Bayesian HPD region each time?

Suppose we generate datasets by picking a parameter value from $\pi(\theta)$ and simulating data from $p(D|\theta)$

The fraction of time θ will be in the HPD region is:

$$Q = \int d\theta \; \pi(\theta) \int dD \; p(D|\theta) \; [\![\theta \in \Delta(D)]\!]$$

Note $\pi(\theta)p(D|\theta) = p(\theta, D) = p(D)p(\theta|D)$, so

$$Q \;\; = \;\; \int dD \int d\theta \; p(\theta|D) \, p(D) \; [\![\theta \in \Delta(D)]\!]$$

$$
\begin{aligned}
Q &= \int dD \int d\theta\; p(\theta|D)\, p(D)\; [\![\theta \in \Delta(D)]\!] \\
&= \int dD\; p(D) \int d\theta\; p(\theta|D)\; [\![\theta \in \Delta(D)]\!] \\
&= \int dD\; p(D)\; \int_{\Delta(D)} d\theta\; p(\theta|D) \\
&= \int dD\; p(D) P \\
&= P
\end{aligned}
$$

The HPD region includes the true parameters $100P\%$ of the time

This is exactly true for any problem, even for small datasets

Keep in mind it involves drawing θ from the prior; credible regions are "calibrated with respect to the prior"

A Tangent: Average Coverage

Recall the original Q integral:

$$\begin{aligned} Q &= \int d\theta\, \pi(\theta) \int dD\; p(D|\theta)\; [\![\theta \in \Delta(D)]\!] \\ &= \int d\theta\, \pi(\theta) C(\theta) \end{aligned}$$

where $C(\theta)$ is the (frequentist) coverage of the HPD region when the data are generated using θ

This indicates Bayesian regions have accurate *average coverage*

The prior can be interpreted as quantifying how much we care about coverage in different parts of the parameter space

Basic Bayesian Calibration Diagnostics

Encapsulate your sampler: Create an MCMC posterior sampling algorithm for model M that takes data D as input and produces posterior samples $\{\theta_i\}$, and a $100\,P\%$ credible region $\Delta_P(D)$

Initialize counter $Q = 0$
Repeat $N \gg 1$ times:

1. Sample a “true” parameter value θ^* from $\pi(\theta)$
2. Sample a dataset D_{sim} from $p(D|\theta^*)$
3. Use the encapsulated posterior sampler to get $\Delta_P(D_{\text{sim}})$ from $p(\theta|D_{\text{sim}}, M)$
4. If $\theta^* \in \Delta_P(D)$, increment Q

Check that $Q/N \approx P$

Easily extend the idea to check *all* credible region sizes:

Initialize a list that will store N probabilities, P
Repeat $N \gg 1$ times:

1. Sample a "true" parameter value θ^* from $\pi(\theta)$
2. Sample a dataset D_{sim} from $p(D|\theta^*)$
3. Use the encapsulated posterior sampler to get $\{\theta_i\}$ from $p(\theta|D_{\text{sim}}, M)$
4. Find P so that θ^* is on the boundary of $\Delta_P(D)$; append to list $[P = \text{fraction of } \{\theta_i\} \text{ with } q(\theta_i) > q(\theta^*)]$

Check that the Ps follow a uniform distribution on $[0, 1]$

Other Joint Distribution Tests

- Geweke 2004: Calculate means of scalar functions of (θ, D) two ways; compare with z statistics
- Cook, Gelman, Rubin 2006: Posterior quantile test, expect $p[g(\theta) > g(\theta^*)] \sim$ Uniform (HPD test is special case)

What Joint Distribution Tests Accomplish

Suppose the prior and sampling distribution samplers are well-validated

- **Convergence verification:** If your posterior sampler is bug-free but was not run long enough $\rightarrow$ unlikely that inferences will be calibrated
- **Bug detection:** An incorrect posterior sampler implementation will not converge to the correct posterior distribution $\rightarrow$ unlikely that inferences will be calibrated, even if the chain converges

Cost: Prior and data sampling is often cheap, but posterior sampling is often expensive, and joint distribution tests require you run your MCMC code *hundreds* of times

Compromise: If MCMC cost grows with dataset size, running the test with small datasets provides a good bug test, and *some* insight on convergence; could also test a simplified model

Frequentist Performance of Bayesian Procedures

Many results known for parametric Bayes performance:

- Estimates are consistent if the prior doesn't exclude the true value.
- Credible regions found with flat priors are typically confidence regions to $O(n^{-1/2})$ (Bernstein-von Mises Theorem); "reference" priors can improve their performance to $O(n^{-1})$.
- Marginal distributions have better frequentist performance than conventional methods like profile likelihood. (Bartlett correction, ancillaries are competitive but hard.)
- Bayesian model comparison is asymptotically consistent (not true of significance/NP tests, AIC).
- Misspecification: Bayes converges to the model with sampling dist'n closest to truth via Kullback-Leibler

- Frequentist behavior in nonparametric & semiparametric contexts is more complex and a topic of ongoing research; *you must be more careful with priors here*
- Wald's complete class theorem: *Optimal* frequentist methods are *Bayes rules* (equivalent to Bayes for some prior)
- ...

Parametric Bayesian methods are typically good frequentist methods.

Some references:

- "The Interplay of Bayesian and Frequentist Analysis" (Bayarri & Berger 2004) *Statistical Science*, **19**, 58–80
- "Calibrated Bayes: A Bayes/Frequentist Roadmap" (Little 2006; 2005 ASA President's Invited Address) *The American Statistician*, **60**, 213–223