

Basics of Bayesian inference

- Bayes' theorem
- Bayesian inference
 - Point estimation
 - Interval estimation
 - Hypothesis testing and model choice
- Bayes computation:
 - Revisit Gibbs and Metropolis-Hasting
 - Slice sampling
 - Convergence diagnosis
 - variance estimation

Bayes' Theorem

We model the observed data and "unknown parameters" as random variables. The Bayes' theorem allows us to combine external knowledge and complex data models in estimating the "unknowns".

We specify the distributional model (LIKELIHOOD)

$$f(y|\theta)$$

y is the observed data, and θ are the unknown parameters.

We assume a PRIOR distribution for θ :

$$\pi(\theta|\lambda)$$

where λ is a hyperparameter.

Inference on θ is based on its POSTERIOR distribution:

$$p(\theta|y, \lambda) = \frac{p(y, \theta|\lambda)}{p(y|\lambda)} =$$
$$\frac{p(y, \theta|\lambda)}{\int p(y, \theta|\lambda)d\theta} = \frac{p(y|\theta)\pi(\theta|\lambda)}{\int p(y|\theta)\pi(\theta|\lambda)d\theta}$$

Since λ might be unknown (hyperprior) we need an additional step:

$$p(\theta|y) = \frac{p(y, \theta)}{p(y)} =$$
$$\frac{\int p(y|\theta)\pi(\theta|\lambda)h(\lambda)d\lambda}{\int \int p(y|\theta)\pi(\theta|\lambda)h(\lambda)d\theta d\lambda}$$

Alternatively, we can replace λ by an estimated value of λ , $\hat{\lambda}$, which could be the maximizer of $p(y|\lambda)$. Inference based on this estimated posterior $p(\theta|y, \hat{\lambda})$ is referred to as **EMPIRICAL BAYES** analysis.

Forms of π and h , known as conjugate priors, enable analytic evaluation of these integrals. But, in the presence of nuisance parameters (unknown variable) some intractable integrations still remain. Here the need of of the recently developed Markov chain Monte Carlo (MCMC) integration methods.

Bayesian inference

Point estimation

Estimates of θ :

The mean of the posterior:

$$\hat{\theta} = E(\theta|y)$$

The median of the posterior:

$$\hat{\theta} : \int_{-\infty}^{\hat{\theta}} p(\theta|y) d\theta = .5$$

The mode of the posterior:

$$\hat{\theta} : p(\hat{\theta}|y) = \sup_{\theta} p(\theta|y)$$

The last one is the easiest to compute, since it does not required any integration.

If the posterior exists under a flat prior $p(\theta) = 1$, then the posterior mode is just the MLE of θ .

The posterior median is often the best point estimate (because it represents better the center of a non-symmetric distribution).

Interval estimation

The posterior allows inference about not only the median, but any QUANTILE.

We can obtain a Bayesian confidence interval, that we generally call CREDIBLE INTERVAL. The probability that θ lies in (q_L, q_U) is $(1 - \alpha)$, where q_L and q_U satisfy:

$$\int_{-\infty}^{q_L} p(\theta|y)d\theta = \alpha/2$$

and

$$\int_{q_U}^{\infty} p(\theta|y)d\theta = 1 - \alpha/2.$$

Thus,

$$p(q_L < \theta < q_U) = 1 - \alpha.$$

The frequentist CI does not satisfy that condition. Instead, it gives an interval such that the probability that the random interval covers the TRUE parameter is $1 - \alpha$, i.e.

$$P(\theta \in (a, b) | \theta) = 1 - \alpha.$$

The interval $p(q_L < \theta < q_U) = 1 - \alpha$, is the equal tail credible set.

For symmetric unimodal posteriors, this interval will be symmetric about this mode. It will be also optimal, in the sense that it will have shortest length among sets C satisfying:

$$1 - \alpha \leq P(C|y) = \int_C p(\theta|y) d\theta$$

For posteriors that are not symmetric and unimodal, a better, shorter, interval can be obtained by taking only values of θ that have posterior greater than some cutoff. The cutoff is as large as possible while C satisfies the previous condition.

This is called the HIGHEST POSTERIOR DENSITY (HPD) confidence set. More difficult to compute but always of optimal length.

Hypothesis testing and model choice

Hypothesis testing is not very straightforward. There is not agreement between Bayesian about how to approach the problem. Deviance Information Criterion (DIC) has gained popularity (available in WinBUGS). We will discuss next BF and BIC for model choice.

Bayes factor for model choice

We replace the two hypotheses H_0 and H_1 by two candidate parametric models M_1 and M_2 with parameters θ_1 and θ_2 respectively. The priors are:

$$\pi_i(\theta_i).$$

Thus, the marginals of Y ,

$$p(y|M_i) = \int f(y|\theta_i, M_i)\pi_i(\theta_i)d\theta_i$$

Bayes' theorem can be applied to obtain the posterior of

$$P(M_1|y)$$

and

$$P(M_2|y) = 1 - P(M_1|y).$$

The quantity used to summarize these results is BAYES FACTOR BF, the ratio of the posterior odds of M_1 to the prior odds of M_1 :

$$\begin{aligned} BF &= \frac{P(M_1|y)/P(M_2|y)}{p(M_1)/p(M_2)} \\ &= \frac{P(y|M_1)}{p(y|M_2)} \end{aligned}$$

If both models have same prior, then BF is the posterior odds of M_1 .

In models such that $\theta_1 = \theta_2 = \theta$, and both hypotheses are simple:

$$M_1 : \theta = \theta^{(1)}$$

and

$$M_2 : \theta = \theta^{(2)}.$$

Then $\pi_i(\theta)$ is a point mass at $\theta^{(i)}$. And, we have,

$$BF == \frac{P(y|\theta^{(1)})}{p(y|\theta^{(1)})},$$

which is the LIKELIHOOD RATIO between the two models.

Bayesian Information Criterion (BIC)

BIC also known as Schwarz Criterion. Schwarz showed that for nonhierarchical models and large sample sizes n , BIC approximates $-2 \log BF$. BIC is a penalized likelihood ratio model choice criterion, if we think of M_2 as the "full" model and M_1 as the "reduced" model.

$$\Delta BIC = W - (p_2 - p_1) \log n$$

where p_i is the number of parameters in model M_i , and

$$W = -2 \log \left\{ \frac{\sup_{M_1} f(y|\theta)}{\sup_{M_2} f(y|\theta)} \right\}$$

the usual likelihood ratio test statistic.

An alternative to BIC is the Akaike Information Criterion (AIC),

$$\Delta AIC = W - 2(p_2 - p_1).$$

This is also a penalized likelihood ratio model choice criteria.

The more serious limitation in using BF or their approximations (BIC, AIC) are that they are not appropriate under noninformative priors. If $\pi_i(\theta_i)$ is improper then $p(y|M_i)$ is as well. A solution is to use DIC.

DIC

Spiegelhalter et al. propose a generalization of AIC (whose asymptotic justification is not appropriate for hierarchical models). It is based on the posterior of the deviance

$$D(\theta) = -2 \log f(y|\theta) + 2 \log h(y)$$

h some function of the data alone.

The fit is summarized with the mean of the posterior of D :

$$\bar{D} = E_{\theta|y}[D]$$

and the effective number of parameters p_D . In Gaussian models, p_D is the expected deviance minus D evaluated at the posterior expectations:

$$p_D = E_{\theta|y}[D] - D(E_{\theta|y}[D]) = \bar{D} - D(\bar{\theta})$$

The DIC is then defined as

$$DIC = \bar{D} + p_D = 2\bar{D} - D(\bar{\theta})$$

smaller DIC indicate a better-fitting model.

DIC has no absolute scale, only DIFFERENCES in DIC across models are meaningful.

Identification of what is a SIGNIFICANT difference in DIC is difficult, delta method approximation to the variance of DIC have not been very successful. In practice, we recompute DIC few times using different random number seeds.

It is up to the user to think carefully about which parameters ought to be "in focus" before using DIC, i.e. the likelihood function used could be marginal (only for the parameters of interest).

Bayesian computation

The most popular computing tools in Bayesian practice today are Markov chain Monte Carlo (MCMC) methods. Because their ability to enable inference from posterior distributions of large problems, by reducing the problem to one of RECURSIVELY solving a series of lower-dimensional problems.

Like traditional Monte Carlo, MCMC works by producing not a closed form for the posterior by a SAMPLE of values from this distribution.

A histogram based on such a sample is typically sufficient for reliable inference. However, unlike traditional MC methods, MCMC algorithms produce CORRELATED samples from this posteriors, since they are recursive draws from a particular Markov chain, the

stationary distribution of which is the same as the posterior.

The two most popular are Gibbs sampler and Metropolis-Hastings algorithm.

Algorithm to simulate from the posterior distribution **The Gibbs sampler**

Gibbs Sampling

We describe now another algorithm to efficiently generate the simulations from the posterior of a vector parameter ϕ . The most convenient methods are of Markov chain Monte Carlo (MCMC) type, of which Gibbs sampling is one of the most widely used.

Gibbs sampling. Start with an arbitrary initial value for the vector parameter $\phi^{(0)} = (\phi_1^{(0)}, \dots, \phi_n^{(0)})$. Generate a new value of ϕ_1 , denoted $\phi_1^{(1)}$, from the conditional distribution of Φ_1 given $\Phi_2 = \phi_2^{(0)}, \dots, \Phi_n = \phi_n^{(0)}$. Then generate a new value of ϕ_2 , denoted $\phi_2^{(1)}$, from the conditional distribution of Φ_2 given $\Phi_1 = \phi_1^{(1)}, \Phi_3 = \phi_3^{(0)}, \dots, \Phi_n = \phi_n^{(0)}$. Continue up to the generation of $\phi_n^{(1)}$ from the conditional distribution of Φ_n given

$\Phi_1 = \phi_1^{(1)}, \dots, \Phi_{n-1} = \phi_{n-1}^{(1)}$. This completes one iteration of the sampler. Then, starting from the new vector $\phi^{(1)}$, return to ϕ_1 and repeat the whole process to generate $\phi^{(2)}$.

Thus, Gibbs sampling consists purely in sampling from conditional distributions, because instead of updating ϕ *in block*, it is more computationally efficient to divide ϕ into components and then update these components one by one.

The Metropolis-Hastings algorithm.

Gibbs sampler is easy to implement but requires the ability to sample from each of the full conditional distributions.

The MH algorithm is a rejection algorithm that attacks the problem of not being able to sample from a distribution.

We start with an arbitrary $\mathbf{x}^{(0)}$ and generate a new “trial value” \mathbf{x}' from some distribution $q(\mathbf{x}'; \mathbf{x}^{(0)})$ which depends on $\mathbf{x}^{(0)}$. Then form the ratio

$$\alpha = \frac{q(\mathbf{x}^{(0)}; \mathbf{x}')l(\mathbf{x}')}{q(\mathbf{x}'; \mathbf{x}^{(0)})l(\mathbf{x}^{(0)})}.$$

where $l(\mathbf{x}) = f(y|\mathbf{x})\pi(\mathbf{x})$, the likelihood times the prior for x . If $\alpha \geq 1$ then we accept \mathbf{x}' ; in other words, set $\mathbf{x}^{(1)} = \mathbf{x}'$. If $\alpha < 1$, we perform an independent random drawing: with probability α , accept \mathbf{x}' and set $\mathbf{x}^{(1)} = \mathbf{x}'$; otherwise, reject \mathbf{x}' and set $\mathbf{x}^{(1)} = \mathbf{x}^{(0)}$.

Theoretically we can choose any proposal q , but in practice only a "good" choice will work. The usual approach is

$$q(\mathbf{x}'|\mathbf{x}^{(0)}) = \mathbf{N}(\mathbf{x}'|\mathbf{x}^{(0)}, \tilde{\Sigma})$$

where $\tilde{\Sigma}$ might be an empirical estimate of the true posterior variance, derived from a preliminary sampling run.

Accepting all candidates is usually the result of an overly narrow proposal. The ideal is to choose $\delta\Sigma$ such that around 50% of the candidates are accepted.

We can also use

$$q(\mathbf{x}'|\mathbf{x}^{(0)}) = \mathbf{q}(\mathbf{x}')$$

this is not symmetric in its arguments.

Metropolis algorithm

If we replace the acceptance ratio with

$$\alpha = l(\mathbf{x}')/l(\mathbf{x}^{(0)})$$

where q now has to be symmetric in its arguments, i.e.

$$q(\mathbf{x}'|\mathbf{x}^{(0)}) = q(\mathbf{x}^{(0)}|\mathbf{x}^{(')}),$$

we have the Metropolis algorithm.

Once we have the simulated values from the posterior, we usually ignore the ones in the burn-in period, and we can construct a histogram with the rest, and also obtain the sample mean to estimate the expected value of the posterior.

In practice, we may actually run m parallel chains instead of only 1.

Slice sampling

This is an alternative to M-H.

We seek to sample

$$\theta \sim f(\theta) \equiv h(\theta) / \left(\int h(\theta) d\theta \right)$$

where h is known. We add an AUXILIARY variable U , such that,

$$U|\theta \sim \text{Unif}(0, h(\theta))$$

Then, the joint of θ and U is

$$p(\theta, u) \propto 1 \cdot I(U < h(\theta))$$

I indicator function.

We run a Gibbs sampler from $U|\theta$ followed by $\theta|U$ at each iteration, we then can obtain samples from $p(\theta, u)$ and then get the marginal of θ .

Sampling from $\theta|U$ requires a draw from a uniform distribution from θ over

$$S_U = \{\theta : U < h(\theta)\}.$$

If $h(\theta) = h_1(\theta)h_2(\theta)$ where h_1 is a standard density, and h_2 nonstandard. Then we introduce U , such that

$$U|\theta \sim U(0, h_2(\theta))$$

Now,

$$p(\theta, u) \propto h_1(\theta) \cdot I(U < h_2(\theta))$$

I indicator function. To sample from $U|\theta$ is routine, to sample from $\theta|U$ we now draw θ from h_1 and retain it only if θ is such that $U < h_2(\theta)$.

Convergence diagnosis

The most problematic part of MCMC is deciding when is safe to stop and determine t_0 (the burn-in period).

We commonly run a few ($m = 3$) parallel sampling chains, initialized at widely disparate starting locations that are overdispersed with respect to the true posterior. We plot the chains (trace plots) and find a point t_0 such that the m chains overlap.

Difficulties:

- The posterior is unknown so we do not know if the chains are overdispersed.
- it is hard to automate this diagnosis, since it requires a subjective judgement by a human viewer.

Among the most popular diagnostics are the Gelman and Rubin (1992) approach.

Gelman and Rubin approach to determine convergence.

We run m chains, overdispersed (using a preliminary posterior-mode finding algorithm). Running each chain for $2N$ iterations, and then we try to see whether the variation within the chains for a given parameter of interest λ is approximately the same as the total variation across the chains during the latter N iterations.

Specifically, we monitor convergence by the estimated SCALE REDUCTION factor,

$$\sqrt{\hat{R}} = \sqrt{\left(\frac{N-1}{N} + \frac{m+1}{mN} \frac{B}{W} \right) \frac{df}{df-2}}$$

where B/N is the variance between the means from the m parallel chains, W is the average of the m within-chain variances, and df is the degrees of freedom of an approximating t density to the posterior distribution.

The factor must approach to 1 as $N \rightarrow \infty$.

This approach focuses only on detecting bias in the MCMC estimator, and it is a univariate quantity. Multiple modes in the posterior may easily fool most of these type of diagnostics.

Variance estimation

A criticism of MCMC is that no two analysts will obtain the same answer. Thus, assessment of the variance is crucial.

Supposed that we have a single chain of N post-burn-in samples of a parameter λ , of the posterior mean estimator

$$\hat{E}(\lambda|y) = \hat{\lambda}_N = \frac{1}{N} \sum_{t=1}^N \lambda^{(t)}.$$

Assuming the samples are independent, the variance of $\hat{\lambda}$ can be estimated

$$\hat{\text{var}}(\hat{\lambda}_n) = s_\lambda^2/N,$$

where s_n^2 is the sample variance,

$$s_N^2 = \frac{1}{N-1} \sum (\lambda^{(t)} - \hat{\lambda}_N)^2$$

This would underestimate the true variance due to the **AUTOCORRELATION** in the MCMC samples.

A possibility is **THINNING**, i.e. retaining only every *k*th sampled value. Those, this approach seems to increase the variance.

A better approach is to use all samples, but using the notion of **EFFECTIVE SAMPLE SIZE**, or **ESS**:

$$ESS = N/\kappa(\lambda)$$

where $\kappa(\lambda)$ is the "autocorrelation time" for λ , given by:

$$\kappa(\lambda) = 1 + 2 \sum_{k=1}^{\infty} \rho_k(\lambda),$$

where $\rho_k(\lambda)$ is the autocorrelation at lag k for the parameter λ , estimated with sample autocorrelations from the MCMC chain. Thus, the variance estimate for $\hat{\lambda}_N$ is

$$\hat{\text{var}}_{ESS}(\hat{\lambda}_n) = s_\lambda^2 / ESS(\lambda) =$$
$$\frac{\kappa(\lambda)}{N(N-1)} \sum_{t=1}^N (\lambda^{(t)} - \hat{\lambda}_N)^2$$

A final approach to estimate the variance is using BATCHING.

We divide the run of length N into m successive batches of length k ($N = mk$), with batch means B_1, \dots, B_m . Then,

$$\hat{\lambda}_n = \tilde{B} = 1/m \sum B_i$$

The variance

$$\hat{\text{var}}_{batch}(\hat{\lambda}_N) = \frac{1}{m(m-1)} \sum (B_i - \hat{\lambda}_N)^2$$

with k large enough so the correlation between batches is negligible, and m is large enough to reliably estimate the variance.

Regardless how we estimate the variance, \hat{V} , a 95% confidence interval for $E(\lambda|y)$ is then given by

$$\hat{\lambda}_N \pm z_{.025} \sqrt{\hat{V}}$$

where z is from the normal table. If the batching method uses fewer than 30 batches, is a good idea to replace z by t (from a t distribution with $m-1$ degrees of freedom).