Math 459 Lecture 19

Todd Kuffner

Last Time

Bayesian approach to hypothesis testing

Introduction to Bayes Factors

The importance of the marginal likelihood

Today: computing Bayes factors and marginal likelihoods via Laplace approximation and MCMC

Reminder

Consider a model $M_\alpha = \{\mathcal{F}_\alpha, \Pi_i, \lambda_\alpha\}$ for some α in a countable index set A.

- usual way a Bayesian assesses evidence for or against some model is by computing the Bayes factor for the model of interest and some alternative model under consideration.

$$BF_{kl} = \frac{m(y|M_k)}{m(y|M_l)}$$

where the marginal likelihood for model M_i is

$$m(y|M_i) = \int f(y|\theta_i, M_i)\pi(\theta_i|M_i)d\theta_i,$$

where $f(y|\theta_i, M_i)$ is the likelihood under model i and $\pi(\theta_i|M_i)$ is the prior under model i.

Key issue: how to compute or accurately estimate the marginal likelihoods

Strength of Evidence

Let K be the Bayes factor for H_0 relative to H_1.

Jeffreys scale

K	**Strength of evidence for H_0**
< 1	negative (supports H_1)
$(1, \sqrt{10})$	weak
$(\sqrt{10}, 10)$	substantial
$(10, 10\sqrt{10})$	strong
$(10\sqrt{10}, 100)$	very strong
> 100	decisive

Kass & Raftery (1995 survey)

$2 \log K$	K	**Strength of evidence for H_0**
$(0, 2)$	$(1, 3)$	weak
$(2, 6)$	$(3, 20)$	positive
$(6, 10)$	$(20, 150)$	strong
> 10	> 150	very strong

Problem

Often the integral

$$m(y|M_i) = \int f(y|\theta_i, M_i)\pi(\theta_i|M_i)d\theta_i,$$

cannot be evaluated analytically.

- some common numerical methods are inefficient because when sample sizes are moderate or large, the integrand becomes highly peaked around its maximum, which *can be found by other methods*
- for example, quadrature methods which are initialized without knowing the maximum can encounter difficulty finding the region where the integrand mass is accumulating
- moreover in high dimensions, MCMC is often more efficient (see below)

In fact, the commonly-encountered examples for which the marginal likelihood can be evaluated analytically are restricted to exponential family models with conjugate priors (e.g. normal linear models)

Proposal 1: Laplace's method

- Tierney, Kadane (1986) Accurate approximations for posterior moments and marginal densities. *JASA* **81**, 82–86.
- Tierney, Kass, Kadane (1989) Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *JASA* **84**, 710–716.

Laplace approximation of integrals

Motivation

Suppose we have an integral of the form

$$I(x) = \int_a^b e^{xg(t)} f(t)dt$$

for large x (i.e. as $x \to \infty$).

- we want an accurate approximation $\hat{I}(x)$ s.t. the ratio converges to 1 as $x \to \infty$
- clearly $\hat{I}(x)$ will depend on the two functions f, g, e.g. whether g is monotone or not, the boundary behaviors of f, g, etc.

Example

Suppose g is strictly monotone and has nonzero derivative in the interval (a, b), then *integration by parts* yields

$$I(x) = \frac{1}{x}\frac{f(t)}{g'(t)}e^{xg(t)}\Big|_a^b - \frac{1}{x}\int_a^b \frac{d}{dt}\frac{f(t)}{g'(t)}e^{xg(t)}dt$$

- if one of $f(a)$, $f(b)$ is nonzero, the first term will dominate and then

$$I(x) \sim \frac{1}{x}\frac{f(b)}{g'(b)}e^{xg(b)} - \frac{1}{x}\frac{f(a)}{g'(a)}e^{xg(a)}$$

This is the simplest situation.

More realistically

In practical situations, g may not be strictly monotone and may have zero derivative at one or more points in the interval (a, b).

- then we can't integrate by parts

Idea of Laplace's method: if g has a maximum at some unique c in (a, b), and if $f(c) \neq 0$, $g'' \neq 0$, then due to the large magnitude of the parameter x, the *dominant part* of the integral will come from a *neighborhood of c*.

(continued)

Taylor expansion of g around c up to a quadratic term yields a normal kernel (density).

- the integral will not change much if the range of integration is changed from (a, b) to the real line
- upon normalizing the normal density, the factor $\sqrt{2\pi}$ appears and we have the approximation

$$\hat{I}(x) = \frac{\sqrt{2\pi} f(c) e^{xg(c)}}{\sqrt{-xg''(c)}}$$

  and as $x \to \infty$, $\hat{I}(x) \sim I(x)$
- intermediate steps of this derivation require that a, b are not stationary points of g, that $g'(c) = 0$ and that $g''(c) < 0$
- if either of the two boundary points a, b is a stationary point of g, then the approximating function will change to accomodate contributions from the boundary stationary points

If the interior local maximum of g is *not unique*, then $I(x)$ must be partitioned into subintervals separating the different maxima and summing over the terms to obtain a final approximation to $I(x)$.

Example 1

Let

$$I(x) = \int_{-\infty}^{\infty} e^{x(t-e^t)} dt,$$

(which equals $\Gamma(x)/x^x$ for a suitable change of variable in $I(x)$).

- try an asymptotic approximation for $I(x)$ as $x \to \infty$ by Laplace's method with $f(t) = 1$ and $g(t) = t - e^t$
- the only saddlepoint of g is $t = 0$, and $g''(t) = -e^t$

Therefore, the Laplace approximation is

$$\hat{I}(x) = \frac{\sqrt{2\pi}e^{-x}}{x}$$

Example 2

Consider approximation of $n!$ by writing it directly as a Gamma function:

$$n! = \int_0^\infty e^{-z} z^n dz = \int_0^\infty e^{n \log z} e^{-z} dz = n^{n+1} \int_0^\infty e^{n(\log t - t)} dt$$

- consider Laplace approximation with $f(t) = 1$, $g(t) = \log t - t$
- plugging in g'' yields

$$n! \sim e^{-n} n^{n+\frac{1}{2}} \sqrt{2\pi}$$

which is the usual *Stirling approximation*

Application to Marginal Likelihood

Let $\tilde{\ell}(\theta) = \frac{1}{n}\log L(\theta) + \frac{1}{n}\log \pi(\theta)$, so that

$$m(y) = \int e^{n\tilde{\ell}(\theta)} d\theta$$

Laplace's method: let θ_m be the posterior mode

$$m(y) \approx \int \exp[n\tilde{\ell}(\theta_m) - n(\theta - \theta_m)^2/(2\sigma^2)] d\theta$$

with $\sigma^2 = -1/\tilde{\ell}''(\theta_m)$

(continued)

$$
\begin{aligned}
m(y) &\approx \int \exp[n\tilde{\ell}(\theta_m) - n(\theta - \theta_m)^2/(2\sigma^2)]d\theta \\
&\approx \sqrt{2\pi}\sigma n^{-1/2} \exp\{n\tilde{\ell}(\theta_m)\}
\end{aligned}
$$

- the error in this approximation is $O(n^{-1})$ in the sense that $m(y) = \hat{m}(y)(1 + O(n^{-1}))$, and the same error holds if using MLE instead of posterior mode for θ
- prior must be explicitly specified
- using the expected information matrix (instead of observed) in σ is less accurate by conveniently implemented

Common Presentation

Expanding $\tilde{\ell}(\theta)$ as a quadratic about θ_m and then exponentiating yields an approximation to the $f(y|\theta)\pi(\theta)$ having the form of a normal density with mean θ_m and covariance matrix

$$\tilde{\Sigma} = (-D^2\tilde{\ell}(\theta_m))^{-1},$$

where $D^2\tilde{\ell}(\theta_m)$ is the Hessian matrix of second derivatives.

- integrating this approximation yields

$$\hat{m}(y) = (2\pi)^{d/2}|\tilde{\Sigma}|^{1/2} f(y|\theta_m)\pi(\theta_m),$$

  with $\dim(\theta) = d$

Usage in Model Comparison

Consider model comparison when model M_0 is **nested** in model M_1, e.g.

$$M_0: \;\; Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon,$$

$$M_1: \;\; Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

with the same error distribution in both models

- nested means it is possible to obtain any joint distribution for the data (Y, X) in model M_0 using a restriction on the parameters in M_1
- here this means the predictors in M_0 are a subset of the predictors in M_1

More generally, suppose M_0 is nested in M_1 and that M_0 has d_0 parameters and M_1 has $d_1 \geq d_0$ parameters.

- If we use the Laplace approximation to the marginal likelihoods in the Bayes factor, evaluated at the MLE in each model, we arrive at an approximation of the Bayes factor:

$$\begin{aligned} 2\log B_{10} \approx \Lambda &+ \log|\tilde{\Sigma}_1| - \log|\tilde{\Sigma}_0| + \log\pi(\hat{\theta}_1|M_1) \\ &- \log\pi(\hat{\theta}_0|M_0) + (d_1 - d_0)\log 2\pi \end{aligned}$$

  where $\Lambda = 2[\log f(y|\hat{\theta}_1, M_1) - \log f(y|\hat{\theta}_0, M_0)]$
- $\hat{\theta}_0$ means the MLE of the parameters in model M_0 and $\hat{\theta}_1$ is the MLE of the parameters in M_1

Simple Monte Carlo estimation of marginal likelihood

Basic idea: since

$$m(y) = \int f(y|\theta)\pi(\theta)d\theta,$$

draw $\theta_1, \ldots, \theta_m$ i.i.d. from $\pi(\theta)$.

- a Monte Carlo estimator of $m(y)$ is then

$$\frac{1}{m}\sum_{i=1}^{m} f(y|\theta^{(i)})$$

Common problem is large variance of this estimator.

Marginal likelihood from the Gibbs Sampler Output

- Chib (1995) Marginal likelihood from the Gibbs output. *JASA* **90**(432), 1313–1321.

Note the **basic marginal likelihood identity** (rearranging definition of posterior):

$$m(y) = \frac{f(y|\theta)\pi(\theta)}{\pi(\theta|y)}$$

- an identity since it holds *for any* θ (in the parameter space)
- easy to evaluate $f(y|\theta)$ and $\pi(\theta)$
- so to estimate $m(y)$ we only need to estimate the posterior $\pi(\theta|y)$

(continued)

Decompose θ into two blocks (θ_1, θ_2) such that $\pi(\theta_1|\theta_2, y)$ and $\pi(\theta_2|\theta_1, y)$ are known (completely specified).

$$\pi(\theta_1, \theta_2|y) = \pi(\theta_2|\theta_1, y)\pi(\theta_1|y)$$

- Gibbs sampler gives dependent draws from joint posterior $\pi(\theta_1, \theta_2|y)$ and therefore marginally from $\pi(\theta_2|y)$; find the marginal posterior for θ_1 by

$$\begin{aligned}\pi(\theta_1|y) &= \int \pi(\theta_1|\theta_2, y)\pi(\theta_2|y)d\theta_2 \\ &\approx \frac{1}{G}\sum_{g=1}^{G} \pi(\theta_1|\theta_2^{(g)}, y)\end{aligned}$$

Such a marginal posterior estimator is (simulation) consistent; under regularity conditions $\hat{\pi}(\theta|y) \to \pi(\theta|y)$ almost surely as $G \to \infty$ (due to ergodic theorem)

General case (arbitrary number of blocks)

Decompose posterior at the point θ as

$$\pi(\theta|y) = \pi(\theta_1|y) \times \pi(\theta_2|\theta_1, y) \times \cdots \times \pi(\theta_B|\theta_1, \ldots, \theta_{B-1}, y)$$

where the last term is the marginal ordinate and can be estimated from the draws of the initial Gibbs run

- the other terms are the reduced conditional ordinates $\pi(\theta_r|\theta_1, \ldots, \theta_{r-1}, y)$ given by

$$\int \pi(\theta_r|\theta_1, \ldots, \theta_{r-1}, \theta_l(l > r), y) d\pi(\theta_{r+1}, \ldots, \theta_B|\theta_1, \ldots, \theta_{r-1}, y)$$

Estimate this by

$$\hat{\pi}(\theta_r|\theta_s(s < r), y) = G^{-1} \sum_{j=1}^{G} \pi(\theta_r|\theta_1, \ldots, \theta_{r-1}, \theta_l^{(j)}(l > r), y)$$

and estimate the joint density by $\prod_{r=1}^{B} \hat{\pi}(\theta_r|\theta_s(s < r), y)$

Bayes factor estimate

Typically the log of the marginal likelihood is estimated using the above method, yielding the estimate of B_{kl}

$$\hat{B}_{kl} = \exp\{\log \hat{m}(y|M_k) - \log \hat{m}(y|M_l)\}$$

What about Metropolis-Hastings?

- Chib, Jeliazkov (2001) Marginal likelihood from the Metropolis-Hastings output. *JASA* **96**(453), 270–281.

Goal is again to estimate posterior $\pi(\theta|y)$ given a posterior sample $\{\theta^{(1)}, \ldots, \theta^{(M)}\}$.

Let $q(\theta, \theta'|y)$ denote the proposal (candidate generating) density for the transition from θ to θ', which can depend on data y.

- let
$$\alpha(\theta, \theta'|y) = \min\left\{1, \frac{f(y|\theta')\pi(\theta')}{f(y|\theta)\pi(\theta)} \frac{q(\theta', \theta|y)}{q(\theta, \theta'|y)}\right\}$$
denote the probability of a move (i.e. probability of accepting the proposed value)
- letting $p(\theta, \theta'|y) = \alpha(\theta, \theta'|y)q(\theta, \theta'|y)$ denote the subkernel of the M-H algorithm, then from the reversibility (detailed balance) of the subkernel, we have for any point θ^*
$$p(\theta, \theta^*|y)\pi(\theta|y) = \pi(\theta^*|y)p(\theta^*, \theta|y)$$

Integrating both sides w.r.t. θ, where $\theta \in \Theta \subset \mathbb{R}^d$, we find that the posterior ordinate is given by

$$\pi(\theta^*|y) = \frac{\int \alpha(\theta, \theta^*|y) q(\theta, \theta^*|y) \pi(\theta|y) d\theta}{\int \alpha(\theta^*, \theta|y) q(\theta^*, \theta|y) d\theta}$$

To clarify the estimation procedure, write this in a different form:

$$\pi(\theta^*|y) = \frac{E_1\{\alpha(\theta, \theta^*|y) q(\theta, \theta^*|y)\}}{E_2\{\alpha(\theta^*, \theta|y)\}}$$

where the numerator expectation E_1 is w.r.t. the distribution $\pi(\theta|y)$ and the denominator expectation E_2 is w.r.t. $q(\theta^*, \theta|y)$.
$\Rightarrow$ a simulation-consistent estimate is

$$\hat{\pi}(\theta^*|y) = \frac{M^{-1} \sum_{g=1}^{M} \alpha(\theta^{(g)}, \theta^*|y) q(\theta^{(g)}, \theta^*|y)}{J^{-1} \sum_{j=1}^{J} \alpha(\theta^*, \theta^{(j)}|y)}$$

where $\{\theta^{(g)}\}$ are the M-H samples from the posterior and $\{\theta^{(j)}\}$ are draws from $q(\theta^*, \theta|y)$, given the fixed value of θ.