We review several numerical methods for making inference from the posterior distribution if either the prior or the likelihood function (when viwed as a function of the parameter of interest and not as a fucntion of the data) is a normal distribution. We consider the following techniques:

- quadrature methods
- adaptive quadrature methods
- monte-carlo integration
- rejection sampling
- random-walk Metropolis

One of the contentions of frequentists against the Bayesians is the need to calculate multidimentional integrals which are difficult to compute and analyze. However, with the advent of faster chips, computional power is no longer a scarce recourse. Hence, there is a shift (in my view) in computational statistics from providing computationally feasible solutions to obtaining numerically accurate results.

Notation:

- y data
- θ parameter vector
- $L(y;\theta)$ Likelihood function, but will write $L(\theta)$ to emphasize that we are viewing it as a function of θ rather than of data
- $\pi \sim \mathcal{N}(\mu, \Sigma)$ prior density normally distributed with mean μ and covariance Σ
- p posterior density
- E expectiotion w.r.t to the posterior distribution
- E_π expectiotion w.r.t to the prior distribution
- $g(\theta)$ a function of θ that is of interest for inference

Quadrature methods

Please see [1] for an introduction to numerical integration methods. If the weight function is in the form $\exp(-\theta^2)$, one can use Guass-Hermite quadrature method of numerical integration. For demonstration purpose, we assume a bivariate normal prior.

The posterior expectation of a function $g(\theta)$ is given as :

$$E[g(\theta)] = \int g(\theta) L(\theta)\, \pi(\theta)\, d\boldsymbol{\theta} \tag{1}$$

Upon expanding, we get,

$$= \int \pi(\theta_2) \int g(\theta)\, \pi(\theta_1|\theta_2)\, L(\theta) d\theta_1\, d\theta_1 \tag{2}$$

where we have written the joint density of θ_1 and θ_2 as

$$\pi(\theta_1, \theta_2) = \pi(\theta_2)\pi(\theta_1|\theta_2) \tag{3}$$

A numerical approximation to the above equation is sought using the Hermite polynomials [2] which leads us to the following:

$$E[g(\theta_1, \theta_2)] \approx \sum_{i=1}^{N1} m_{1,p}\pi(z_{2,i}) \sum_{j=1}^{N2} m_{2,j}\, g(z_{1,i}, z_{2,j}) L(z_{1,i}, z_{2,j}) \pi(z_{2,j}|z_{1,i}) \tag{4}$$

where

$$\begin{aligned} \mu_1' &= \mu_1 + \rho\tfrac{\sigma_1}{\sigma_2}(\theta_2 - \mu_2) \\ \sigma_1' &= \sigma_1(1-\rho^2)^{0.5} \\ m_{1,i} &= w_{1,i}\, \exp(t_{1,i})\sqrt{2}\sigma_2 \\ z_{1,i} &= \mu_2 + \sqrt{2}\,\sigma_2\, t_{1,i} \\ m_{2,j} &= w_{2,j}\, \exp(t_{2,j})\sqrt{2}\,\sigma_1' \\ z_{2,j} &= \mu_1' + \sqrt{2}\,\sigma_1'\, t_{2,j} \end{aligned} \tag{5}$$

and $w_{1,i}, t_{1,i}, w_{2,j}, t_{2,j}$ are the weights and abscissa of Hermite polynomials of orders $N1$ and $N2$, respectively. We can estimate posterior mean vector and covariance matrix by choosing an appropriate $g(\theta_1, \theta_2)$. To be specific,

$$\begin{aligned} \hat{\theta_1} &= E[\theta_1] \\ \hat{\theta_2} &= E[\theta_2] \\ \hat{\sigma_1^2} &= E[\theta_1^2] - \hat{\theta_1}^2 \\ \hat{\sigma_2^2} &= E[\theta_2^2] - \hat{\theta_2}^2 \\ \hat{\rho} &= \tfrac{E[\theta_1\theta_2] - \hat{\theta_1}\hat{\theta_2}}{\hat{\sigma_1}\hat{\sigma_2}} \end{aligned} \tag{6}$$

This approximation will produce exact results if the function multiplying the prior is a polynomial of order less than or equal to $2N-1$ (in the univariate case). If extended to the bivariate case, this condition becomes even stricter. It would produce decent results if the function is smooth or slowly varying. It is also argued that this is in fact a *deterministic* version of the Monte Carlo integration. On the other side, if the likelihood falls somewhere in between the

sampling points (abscissa of the Hermite polynomials), the results are disastrous. For example consider a function of this form:

$$L(\theta_1, \theta_2) \propto \exp\left(-\frac{1}{2\sigma^2}(y - a(\theta_1^2 + \theta_2^2)^{-1})^2\right) \tag{7}$$

The loci of θ maximizing this function is a circle with center at the origin and radius $\sqrt{\frac{a}{y}}$. Now it is easy to see a case where this radius is enclosed by two abscissa. If the variance is too small, then it is very likely that sampling grid misses the ML solution and in fact we may get numerically unstable results which is somewhat paradoxical. Adaptive quadrature methods [3] can be used to circumvent such difficulties.

adaptive Quadrature method

In the adaptive quadrature methods, we approximate the posterior with a normal near the mode and use this normal approximation as the quadrature kernel. More precisely,

$$E[g(\theta)] = \int g(\theta)\frac{L(\theta)\,\pi(\theta)}{\psi(\theta)}\psi(\theta)\,d\theta \tag{8}$$

where $\psi(\theta)$ is a normal approximation to the posterior at the MAP estimate. Let the posterior density be

$$p(\theta) = L(\theta)\,\pi(\theta) \tag{9}$$

Then, $\psi(\theta)$ is obtained as follows:

$$\theta_{MAP} = \arg\max_{\theta} p(\theta) \tag{10a}$$

$$\psi(\theta) = \mathcal{N}(\theta_{MAP}, -\left[\frac{\partial^2 \log\, p(\theta)}{\partial\theta\partial\theta^T}|\theta = \theta_{MAP}\right]^{-1}) \tag{10b}$$

We can use gauss-hermite quadrature methods by replacing $\pi(\theta)$ with $\psi(\theta)$.

Monte carlo Integration

In general, the approximation error in Monte-carlo integration is independent of the dimensionality of the parameters and is only dependent on the sample size. The very assumption is that we should be able to generate random samples from some distribution efficiently. If either the prior or the likelihood is normal, we have every reason to follow this approach since efficient generators exist for multivariate (pseudo) random numbers.

$$E[g(\theta)] = E_\pi[g(\theta)L(\theta)] \tag{11}$$

$$E_\pi[g(\theta)L(\theta)] = \int (g(\theta)L(\theta))\,\pi(g\theta)\,d\theta \tag{12a}$$

$$\approx \frac{1}{N}\sum_{i=1}^{N} g\left(\theta(i)\right)L(\theta(i)) \tag{12b}$$

where $\theta(i) \sim \pi$. The Monte-carlo error is inversely proportional to $\sqrt{N}$. If we choose to visit certain points (the abscissa of a certain polynomial) only once, it reduces to a quadrature method and the convergence is a function of N rather than that of $\sqrt{N}$. Thus, we can consider quadrature methods as deterministic version of Monte-carlo. But these methods are more general than quadrature methods and would tend to produce accurate results as long as the proposal distribution (in this case the prior distribution) encompasses the target distribution (for example the posterior distribution) and N is sufficiently large. In the case of posterior distribution, we don't need to impose such a restriction because an unlikely event (with probability zero) is also unlikely in the posterior.

The above two methods can be used to calculate (or approximate)a function of the parameter from the posterior distribution like the normalizing constant, posterior momements etc.. However, we can not obtain the posterior distribution as such. For example, we can not (as far as I know) calculate the mode or construct posterior confidence intervals. Where as, if we draw samples from the posterior, not only we can evaluate functions of the parameter but also obtain summary statistics including confidence intervals or even construct an empirical cdf or pdf. We review rejection sampling and Markov chain Monte-carlo techniques in this light.

Rejection sampling

We can view the Bayes relation as a way to generate posterior samples from the prior [4]. In this perspective, the likelihood acts as resampler. We draw samples from the prior and based on the likelihood of this sample coming from the likelihood we accept or reject the sample. The rejection sampling mechanism is outlined below:

- step-1: draw θ from π
- step-2: draw u from $\mathcal{U}[0,1]$
- step-3: accept θ if $u \leq \frac{p(\theta)}{M\pi(\theta)}$ otherwise repeat steps-1 to 2 until desired number of samples are generated.

where $M = \max \frac{p(\theta)}{\pi(\theta)}$. Since $p(\theta) = L(\theta)\pi(\theta)$ the rejection reduces to a very simple form:

- step-3: accept θ if $u \leq \frac{L(\theta)}{L(\hat{\theta}_{ML})}$

This method is very useful if $\hat{\theta}_{ML}$, the ML estimate of θ, can be found efficiently. This method can be easily modified to accommodate importance sampling as well by bootstrapping the samples generated thus far. Besides that, this interpretation is helpful if one were evaluating different likelihoods with the same prior or vice versa [4].

Random-walk Metropolis

Markov-Chain Monte carlo (MCMC) methods have revolutionized applied statistics at least in Bayesian analysis. In MCMC, we simulate an ergodic Markov chain whose samples are distributed as the intended one, asymptotically. For a review of MCMC techniques please refer to [5, 6]. Terminology:

- $p(\theta)$ target distribution: the distribution from which we *want* to draw the samples
- $J(\theta^*|\theta)$ proposal density: the distribution from which we *will* draw samples

Based on some strategy we accept new draws or retain old draws. Algorithms would differ in the way they choose the proposal. Though some guidelines exist on how to choose a proposal distribution, I feel it is an art rather than a science (no offense meant to die-hard Bayesians).

Metropolis-Hasting algorithm:

- step-1: draw a sample from the proposal density conditioned on the previous draw
  $\theta^* \sim J(\theta|\theta_t)$
- step-2: calculate the acceptance probability
  $\alpha(\theta^*, \theta_t) = \min\left(\frac{p(\theta^*)J(\theta^*|\theta_t)}{p(\theta_t)J(\theta_t|\theta^*)}, 1\right)$
- ste-3: $u \sim \mathcal{U}[0,1]$ if $u \leq \alpha(\theta^*, \theta_t)$, then $\theta_{t+1} = \theta^*$ else $\theta_{t+1} = \theta_t$

If the proposal density is symmetric, then it is called Metropolis algorithm. Even the Gibbs sampler is a special case of the M-H algorithms in which the vector parameter is sampled on conditional marginal densities on at a time.

random notes: We throw away certain number of samples (called *burn-in*) since MCMC samples converge to the stationary distribution (target distribution) only asymptotically. Also, we may need to start multiple parallel chains to make sure that all the chains merge after the *burn-in*. Another check for convergence is the Gelman -Rubin statisitcs which should approach to one near convergence. In my view this has to happen because, by construction Markov Chain is ergodic and hence the ensemble (across chain) variance and sample (with-in) variance must be (approximately) the same. Hence the ratio of *with-in* to *across* variance should be unity.

Random-walk Metropolis algorithm: The proposal density is chosen as:

$J(\theta^*|\theta_t) \sim \mathcal{N}(\theta_t, \Sigma)$.

By construction this distribution is symmetric and hence step-2 of the M-H algorithm simplifies to

$\alpha(\theta^*, \theta_t) = \min\left(\frac{p(\theta^*)}{p(\theta_t)}, 1\right)$

random trick: Suppose that the likelihood is in the form of (7), then the posterior will be in the form:

$p(\theta) \propto \exp(f(\theta))$ for some function f. Step-2 and 3 of the M-H reduce to:

- step-2: calculate the acceptance probability
  $\alpha(\theta^*, \theta_t) = f(\theta^*) - f(\theta_t)$
- step-3: $u \sim \mathcal{U}[0, 1]$ if $\log(u) \leq \alpha(\theta^*, \theta_t)$, then $\theta_{t+1} = \theta^*$ else $\theta_{t+1} = \theta_t$

We can avoid expensive multiplications and perform operations with additions and one logrithm which also leads to some numerical stability. For example if the argument of the exp(.) is large, then the arithmetic calculations may overflow.

random notes: The variance of the proposal distribution should be chosen such that entire parameter space is traversed. If the variance is too large (reckless proposal as Dr. AC calls it), we reject too many samples. In this case, MCMC chain might have to be thinned-down as the auto-correlation of the chain might be slowly decaying upto certain lags. Whereas if we choose small proposal (a timid proposal), it might a take a really long time to traverse the parameter space. In order to obtain a reasonable estimate of the covariance, we can use Quadrature methods to obtain an estimate of the posterior covariance Σ and then use a scaled version of this covariance matrix for the proposal as suggested by [7, Section 11.9, pp-306].

$$J(\theta^*|\theta_t) \sim \mathcal{N}(\theta_t, (2.4/d)^2\Sigma) \tag{13}$$

where d is the dimensionality of the parameter space.

References

[1] Brian Smith, "Introduction to numerical integration," Spring, 1998, CS 375 class notes , CS Dept., UMN, url: www.arc.unm.edu/ acpineda/cs375/cs375.html.

[2] J. C. Naylor and A. F. M. Smith, "Applications of a method for the efficient computation of posterior distributions," *Applied Statistics*, vol. 31, 1982.

[3] Sophia Rabe-Hesketh and Anders Skrondal, "Reliable estimation of generalized linear mixed models using adaptive quadrature," *The Stata Journal*, vol. 2, no. 1, pp. 1–21, 2002.

[4] A. F. M. Smith and A. E. Gelfand, "Bayesian statistics with no tears: A sampling-resampling perspective," *The American Statistician*, vol. 46, 1992.

[5] Dr. Alicia Carriquiry, "Bayesian analysis," Spring, 2005, class notes , Stat. Dept., Iowa State University, url: www.stat.iastate.edu/stat544/homepage.html.

[6] Siddhartha Chib and Edward Greenberg, "Understanding the metropolis-hastings algorithm," *The American Statistician*, vol. 49, 1995.

[7] Andrew Gelman, Hal S. Stern John B. Carlin, and Donald B. Rubin, *Bayesian Data Analysis*, Chapman & Hill/CRC, second edition, 2003.