1 Bayesian Computation.

If the selection of an adequate prior was the major conceptual and modeling challenge of Bayesian analysis, the major implementational challenge is computation. As soon as the model deviates from the conjugate structure, finding the posterior (first the marginal) distribution and the Bayes rule is all but simple. A closed form solution is more an exception than the rule, and even for such closed form solutions, lucky mathematical coincidences, convenient mixtures, and other "tricks" are needed. Up to this point I believe you got a sense of this calculational challenge.

If classical statistics relies on optimization, Bayesian statistics relies on integration. The marginal needed for the posterior is an integral

$$m(x) = \int_\Theta f(x|\theta)\pi(\theta)d\theta,$$

and the Bayes estimator of $h(\theta)$, with respect to the squared error loss is a ratio of integrals,

$$\delta_\pi(x) = \int_\Theta h(\theta)\pi(\theta|x)d\theta = \frac{\int_\Theta h(\theta)f(x|\theta)\pi(\theta)d\theta}{\int_\Theta f(x|\theta)\pi(\theta)d\theta}.$$

The difficulties in calculating the above Bayes rule are that (i) the posterior cannot be represented in a finite form, and (ii) the integral of $h(\theta)$ does not have a closed form integral under the possibly closed form posterior distribution. Adopting a different loss function usually makes calculation even more difficult. An exception is absolute loss for which the Bayes rule is the mode of the posterior, and the mode is not influenced by normalizing (trouble making) constant, $m(x)$.

The last two decades of research in Bayesian statistics contributed to tremendous broadening of the scope of Bayesian models. Models that could not be handled before are now routinely solved. This is done by Markov Chain Monte Carlo (MCMC) Methods, and their introduction to the field of statistics revolutionized Bayesian statistics.

This handout overviews pre MCMC techniques: Monte Carlo Integration, Importance Sampling, and Analytic Approximations (Riemann, Laplace, and Saddlepoint).

1.1 Bayesian CLT

Suppose that $X_1, X_2, \ldots, X_n \sim f(x|\theta)$, where θ is p-dimensional parameter, and that the prior on θ is $\pi(\theta)$. The prior $\pi(\theta)$ could be improper, but we assume that the posterior is proper and that its mode exists. Then, when $n \to \infty$,

$$[\theta|x] \to \mathcal{MVN}_p(\theta_M, H^{-1}(\theta_M)),$$

where θ_M is posterior mode, i.e., a solution of

$$\frac{\partial \pi^*(\theta|x)}{\partial \theta_i} = 0, \; i = 1, \ldots, p,$$

where $\pi^*(\theta|x) = f(x|\theta)\pi(\theta)$ is non-normalized posterior. Let H be the Hessian defined as

$$H(\theta) = -\left(\frac{\partial^2 \pi^*(\theta|x)}{\partial\theta_i \partial\theta_j}\right).$$

The asymptotic covariance matrix is

$$H^{-1}(\theta_M) = (H(\theta))^{-1}\,|_{\theta=\theta_M}$$

The proof can be found in standard texts on asymptotic theory.

Example: Bernoulli's. Assume that $X_1, \ldots, X_n \sim \mathcal{B}er(\theta)$ and that the prior on θ is 1. Show that $\theta_M = \bar{X}$, $H(\theta) = \frac{\sum X_i}{\theta^2} + \frac{n-\sum X_i}{1-\theta^2}$. This gives $H^{-1}(\theta_M) = \frac{\theta_M(1-\theta_M}{n}$. The Bayesian CLT gives expected result,

$$[\theta|X_1, \ldots, X_n] \to \mathcal{N}(\theta_M, \frac{\theta_M(1-\theta_M)}{n}).$$

Example: Poisson/Gamma. Let $X_1, \ldots, X_n \sim \mathcal{P}oi(\theta)$ and $\theta \sim \theta^{\alpha-1}exp\{-\beta\theta\}$. Then,

$$[\theta|x] \to \mathcal{N}\left(\frac{\alpha + \sum X_i - 1}{n+\beta}, \frac{\alpha + \sum X_i - 1}{(n+\beta)^2}\right).$$

This follows from the fact that the mode is $\theta_M = \frac{\alpha+\sum X_i - 1}{n+\beta}$ and that $H(\theta) = \frac{\alpha+\sum X_i - 1}{\theta^2}$.

Posterior approximations by using Bayesian Central Limit Theorem are called first order approximations or modal approximations. Since the posterior is approximated by normal distribution, this approximation may be poor if the true posterior is skewed or if the sample size is small.

1.2 Laplace Approximation

Suppose we are interested in finding $\int_A f(x|\theta)dx$ for a particular value of θ.

Let $f(x|\theta)$ be represented as $\exp\{-nh(x|\theta)\}$. Let x_θ is the value of x that minimizes $h(x|\theta)$ (or equivalently maximizes $f(x|\theta)$). If $h''(x_\theta|\theta) = \left(\frac{\partial^2 h(x|\theta)}{\partial x^2}\right)_{x=x_\theta}$, then

$$\int_a^b e^{-nh(x|\theta)}dx \approx e^{-nh(x_\theta|\theta)}\sqrt{\frac{2\theta}{nh''(x_\theta|\theta)}}\left[\Phi(\sqrt{nh''(x_\theta|\theta)}(b-x_\theta) - \Phi(\sqrt{nh''(x_\theta|\theta)}(a-x_\theta)\right].$$

Using Laplace method, approximate Gamma integral $\int_a^b \frac{x^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha}e^{-x/\beta}dx$, for $\alpha = 3, \beta = 3$ and $(a,b) = (3,5), (7,12)$, and $(5,\infty)$. Compare with exact values and discuss.

Consider the posterior expectation of interest,

$$E^{\theta|x}(g(\theta)) = \frac{\int_\Theta g(\theta)f(x|\theta)\pi(\theta)d\theta}{\int_\Theta f(x|\theta)\pi(\theta)d\theta} = \frac{\int_\Theta b_N(\theta)\exp\{-nh_N(\theta)\}d\theta}{\int_\Theta b_D(\theta)\exp\{-nh_D(\theta)\}d\theta}$$

If $h_N(\theta) = h_D(\theta)$, it is said that the representation is in *standard* form. If, on the other hand, $b_N(\theta) = b_D(\theta)$, it is said that the representation is in *fully exponential* form.

If $E^{\theta|x}(g(\theta))$ can be written in the standard form,

$$E^{\theta|x}(g(\theta)) = \hat{g} + \frac{\sigma_D^2 \hat{b_D'}\hat{g'}}{n\hat{b_D}} + \frac{\sigma_D^2 \hat{g''}}{2n} + \frac{\sigma_D^4 \hat{h'''}\hat{g'}}{2n} + O(n^{-2}).$$

For the fully exponential form, if g is positive and $g(\theta_D)$ is uniformly bounded away from zero,

$$E^{\theta|x}(g(\theta)) = \frac{\hat{b}_N}{\hat{b}_D}\frac{\sigma_N^2}{\sigma_D^2}\exp\{-n(\hat{h}_N - \hat{h}_D)\} + O(n^{-2}).$$

To illustrate the above formulas, let's find approximation to expectation of Beta $\mathcal{B}(\alpha,\beta)$ distribution, $\frac{\alpha}{\alpha+\beta}$.

1.3 Classical Monte Carlo Integration

Suppose $F(x)$ is a probability distribution, and $h(x)$ is a measurable function for which $Eh(X) < \infty$ when $X \sim F$. Let $X_1, \ldots, X_n$ be a sample from F. Then

$$\int_{\mathcal{X}} h(x)dF(x) = E^X h(X) \approx \frac{1}{n}\sum_{i=1}^{n} h(X_i).$$

This is the same as if we wrote

$$\int_{\mathcal{X}} h(x)dF(x) \approx \int_{\mathcal{X}} h(x)dF_n(x|X_1, \ldots, X_n),$$

where $F_n(x|X_1, \ldots, X_n) = \frac{\sum_{i=1}^{n} \mathbf{1}(X_i \leq x)}{n}$ is the empirical distribution function. For simplicity of notation, assume that F is continuous and that density is f, although all results remain valid for general probability distributions. Since, by assumption, $Eh(X) = \int_{\mathcal{X}} h(x)f(x)dx$ is finite for $X \sim f$, from the strong law of large numbers (SLLN) it follows

$$I_n = \frac{1}{n}\sum_{i=1}^{n} h(X_i) \overset{a.s.}{\to} \int_{\mathcal{X}} h(x)f(x)dx = I.$$

The symbol $\overset{a.s.}{\to}$ stands for almost sure convergence, meaning that the the convergence fails on the set that has probability 0. The speed of this almost sure convergence is measured by the speed of decay of variance of the Monte Carlo approximations. Theoretically, $Var(I_n) = \frac{I(1-I)}{n}$.

Example: Normal Likelihood/Cauchy Prior.

The following model is given:

$$\begin{aligned} X|\theta &\sim \mathcal{N}(\theta, 1) \\ \theta &\sim \mathcal{C}a(0, 1). \end{aligned}$$

The Bayes rule

$$\delta_\pi(x) = \frac{\int_{\mathbb{R}} \frac{\theta}{1+\theta^2} \exp\{-1/2(x-\theta)^2\}d\theta}{\int_{\mathbb{R}} \frac{1}{1+\theta^2} \exp\{-1/2(x-\theta)^2\}d\theta}$$

is approximately equal to

$$\delta_\pi(x) \approx \frac{\sum_{i=1}^{n} \frac{\eta_i}{1+\eta_i^2}}{\sum_{i=1}^{n} \frac{1}{1+\eta_i^2}},$$

where $\eta_i \sim \mathcal{N}(X, 1)$.

Assume that $X = 2$. Then high precision numerical algorithms (up to 20 decimal places) in MATHEMATICA give (a) $\delta(2) = 1.2821951026935283611$, and (b) $P^{\theta|2}(\theta \geq 1) = 0.58830709746541437673$. Consider those *exact* values and apply the simulation to check the performance of Monte Carlo.

Uniform on n-Sphere.

This is an example of Christian Robert that is a multivariate generalization of the representation of symmetric distributions as (scale) mixture of uniforms.

Assume

$$\begin{aligned} \boldsymbol{X}|\boldsymbol{\theta} &\sim MVN_p(\boldsymbol{\theta}, I), \\ \boldsymbol{\theta}|c &\sim \mathcal{U}(||\boldsymbol{\theta}||^2 = c), \\ c &\sim \mathcal{G}a(\alpha, \beta). \end{aligned}$$

Assume that for $\alpha = 2, \beta = 3$ and $\boldsymbol{x} = (0, 0, 0, 0, 0, 1)'$ we want to find Bayes rule, $\delta(\boldsymbol{x})$.

Robert (1992, 2001) shows that the Bayes rule can be expressed in almost closed form (up to confluent hypergeometric functions),

$$\delta(\boldsymbol{x}) = \frac{2\alpha}{p} \frac{1}{1+2\beta} \frac{{}_1F_1(\alpha+1; (p+2)/2; ||\boldsymbol{x}||^2/(2+4\beta))}{{}_1F_1(\alpha; p/2; ||\boldsymbol{x}||^2/(2+4\beta))},$$

where ${}_1F_1(a, b; z) = \sum_{k=1}^{\infty} (a)_k/(b)_k z^k/k!$, and $(a)_k = a(a+1)\dots(a+k-1)$. Thus, $\delta(\boldsymbol{x}) = (0, 0, 0, 0, 0, 0.134123)'$.

To approximate this rule by MC method, one may do prior simulation.

1. Generate c from $\mathcal{G}a(2, 3)$. [`BayesLab function rand_gamma`]

2. Simulate M uniform variates $\boldsymbol{\theta}$ on 6-dimensional sphere.

The polar representation of an element $\boldsymbol{\theta}$ from sphere is:

$$\begin{aligned} \theta_1 &= \sqrt{c}\, \cos\varphi_1 \\ \theta_2 &= \sqrt{c}\, \sin\varphi_1 \cos\varphi_2 \\ \theta_3 &= \sqrt{c}\, \sin\varphi_1 \sin\varphi_2 \cos\varphi_3 \\ \dots \\ \theta_{p-1} &= \sqrt{c}\, \sin\varphi_1 \sin\varphi_2 \dots \cos\varphi_{p-1} \\ \theta_p &= \sqrt{c}\, \sin\varphi_1 \sin\varphi_2 \dots \sin\varphi_{p-1} \end{aligned}$$

where $0 \le \varphi_1, \dots, \varphi_{p-2} \le \pi$ and $0 \le \varphi_{p-1} \le 2\pi$. If φ-angles are selected uniformly from their respective domains, the $\boldsymbol{\theta}$ is uniformly distributed on the sphere of radius $\sqrt{c}$.

```
angles = pi*rand(1,p-1); angles(p-1) = 2*angles(p-1); angles(p)=0;
theta(1)=sqrt(c) * cos(angles(1));
for i=2:p
    theta(i)=theta(i-1)*sin(angles(i-1))/cos(angles(i-1))*cos(angles(i));
end
```

3. Approximate δ by

$$\frac{\sum_{i=1}^{M} \boldsymbol{\theta}_i \exp\{-||\boldsymbol{x} - \boldsymbol{\theta}_i||^2/2\}}{\sum_{i=1}^{M} \exp\{-||\boldsymbol{x} - \boldsymbol{\theta}_i||^2/2\}}.$$

Ripley's Example. Consider $p = \int_2^{\infty} \frac{dx}{\pi(1+x^2)}$, the tail of the standard Cauchy distribution. Of course, a Cauchy random variable has an explicit cumulative distribution function, $F(x) = \frac{1}{2} + \frac{1}{\pi}\arctan(x)$ and the above tail probability is $p = \frac{1}{2} - \frac{1}{\pi}\arctan(2) = \frac{1}{\pi}\arctan(1/2) = 0.147584$.

Discuss!

1.4 Importance Sampling

Importance sampling, or weighted sampling, is a Monte Carlo technique in which the integral of interest is transformed in a convenient way to enhance the simulation. Suppose f is a density and $\int_{\mathcal{X}} h(x)f(x)dx$ is of interest. Assume that sampling from f is either difficult or impossible and direct application of Monte Carlo method is hard. The idea of importance sampling is to multiply and divide the expression $h(x)f(x)$ by a convenient density $g(x)$, $\frac{h(x)f(x)}{g(x)}g(x)$. Then,

$$\int_{\mathcal{X}} h(x)f(x)dx = \int_{\mathcal{X}} \frac{h(x)f(x)}{g(x)}g(x)dx.$$

Then conditionally, the density g is easy to sample from, and the Monte Carlo approximation to the integral is

$$\int_{\mathcal{X}} \frac{h(x)f(x)}{g(x)}g(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} \frac{h(X_i)f(X_i)}{g(X_i)}, \quad X_i \overset{i.i.d.}{\sim} g. \tag{1}$$

The density g is called the importance density and its choice depends on f.

There are several guidelines to how to select the importance density g. An obvious requirement is that support of f has to be subset of support of g, $supp(f) \subset supp(g)$ since otherwise we may have an undefined integrand.

Several authors investigated the form of importance density that minimizes the variance of the simulations. Theoretical results are available but the optimal density requires knowledge of $\int h(x)f(x)dx$, the integral we are approximating, and has no practical value. However, from the form of the optimal density g^*, one concludes that for importance densities for which $|h|f/g$ is almost constant and has finite variance, the importance scheme works well.

An attractive feature of importance sampling is that a single random sample from g can be used for different f and h.

If the ratio f/g in (1) is known up to a constant, which may often be the case in Bayesian calculations, the following approximation is used,

$$\int_{\mathcal{X}} \frac{h(x)f(x)}{g(x)}g(x)dx \approx \sum_{i=1}^{n} \frac{h(X_i)f(X_i)}{g(X_i)} / \sum_{i=1}^{n} \frac{f(X_i)}{g(X_i)}, \quad X_i \overset{i.i.d.}{\sim} g.$$

References

[1] Robert, C. (2001). *Bayesian Choice*, Second Edition, Springer Verlag.

Appendix: Simulation of Random Numbers

The BayesLab©matlab suite contains random number generators for major discrete and continuous probability laws. Here we give a couple of general methods that may be of help if the distribution is out of list.

Theorem 1.1 *(Inverse Transformation Method) Let U be the uniform (0,1) random variable and F a cdf for which F^{-1} exists. Then $F^{-1}(U)$ is a draw from distribution F.*

Exponential Random Variates. Since for the exponential $\mathcal{E}(\lambda)$ distribution $y = F(x) = 1 - e^{-\lambda x}, \lambda, x \geq 0$, the inverse function is $x = -\frac{1}{\lambda}\log(1-y)$. Thus

$$X = -\frac{1}{\lambda}\log(1-U),$$

has $\mathcal{E}(\lambda)$ distribution. In fact, since $U \stackrel{d}{=} (1-U)$ one can use

$$X = -\frac{1}{\lambda}\log U.$$

Accept-Reject Method (ARM). This method was originally proposed by von Neumann. Given the density of interest (target density), f, find proposal density (envelope density, instrumental density) g such that

$$(\forall x \in supp(f))\ f(x) \leq Mg(x).$$

The algorithm is:
Step 1. Generate a candidate $X \sim g$. Generate $U \sim \mathcal{U}(0,1)$.
Step 2. Accept $Y = X$ if $U \leq \frac{f(X)}{Mg(X)}$;
Step 3. Return to Step 1.
Indeed, this generates $Y \sim f$. The distribution of Y is given by

$$P(Y \leq y) = P\left(X \leq y | U \leq \frac{f(X)}{Mg(X)}\right) = \frac{P(X \leq y,\ U \leq \frac{f(X)}{Mg(X)})}{P(U \leq \frac{f(X)}{Mg(X)})}.$$

This ratio is,

$$P(Y \leq y) = \frac{\int_{-\infty}^{y}\int_0^{f(x)/Mg(x)} du\ g(x)dx}{\int_{-\infty}^{y}\int_0^{\infty} du\ g(x)dx} = \frac{1/M\int_{-\infty}^{y} f(x)dx}{1/M\int_{-\infty}^{\infty} f(x)dx} = \int_{-\infty}^{y} f(x)dx.$$

Since each proposal will be accepted with probability $P(U \leq \frac{f(X)}{Mg(X)}) = 1/M$, ("success") the number of trials necessary to produce a draw from f is geometric $\mathcal{G}e(1/M)$. The tight bound on M will increase efficiency as well as proposal densities with bound M close to 1.

Exercises:

1. Using ARM, generate from $calBe(2,4)$ using uniform $\mathcal{U}(0,1)$ proposals. Show first that $20x(1-x)^3$ is maximized at $x = 1/4$ and that $M = 135/64$.

2. Generating random normal is of great interest and several well established methods exist. Here is an ARM version. Take proposals from $\mathcal{DE}(1)$, $g(x) = \frac{1}{2}e^{-|x|}$. They are simply exponentials $\mathcal{E}(1)$ multiplied by "random sign," i.e., $S = 2B - 1$, where $B \sim \mathcal{B}er(1/2)$.
Estimate M. ($M \leq \sqrt{2e/\pi} \approx 1.3155$).

3. The density $f(x) = \left(\cos\left(\frac{\pi x}{2}\right)\right)^2, -1 \leq x \leq 1$, called Bickel-Levit prior is of interest in some areas of decision theory (approximation to the least favorable prior in estimating a bounded normal mean). Propose ARM method for simulating from Bickel-Levit prior.

4. Let $f(x|\theta)$ be $\mathcal{DE}(\theta,1)$, and let the prior distribution for θ is a symmetric two-point distribution (concentrated at $-\mu$ and μ, $\mu > 0$).

(a) Find the marginal distribution, $m(x)$.
(b) Propose a sampling scheme to draw from $m(x)$.

5. Suppose that $Y_1, \ldots, Y_n \sim \mathcal{B}er(\theta_i),\ i = 1, \ldots, n$. Suppose that vector of covariates $X_i = (X_{i1}, \ldots, X_{ip})$ correspond to each Y_i, and that

$$\theta_i = \frac{\exp\{X_i'\beta\}}{1 + \exp\{X_i'\beta\}},$$

where β is the vector of regression coefficients.

(a) Show that the likelihood is

$$f(x|\beta) = \exp\left\{\sum_{i=1}^{n}[Y_i X_i'\beta - \log(1 + e^{X_i'\beta})]\right\}.$$

(b) Assume $\pi(\beta) = 1$. Show that the posterior mode is the MLE for β, and

$$-\frac{\partial^2 log\pi^*(\beta|y)}{\partial\beta_i\partial\beta_j} = \sum_{k=1}^{n} X_{ki}X_{kj}\frac{\exp\{X_i'\beta\}}{(1 + \exp\{X_i'\beta\})^2},\ 1 \le i, j \le p.$$

(c) Show

$$[\beta|Y] \to \mathcal{MVN}_p(\hat{\beta}_{mle}, (X'VX)^{-1}),$$

where

$$V = diag\left(\frac{\exp\{X_1'\hat{\beta}_{mle}\}}{(1 + \exp\{X_1'\hat{\beta}_{mle}\})}, \ldots, \frac{\exp\{X_n'\hat{\beta}_{mle}\}}{(1 + \exp\{X_n'\hat{\beta}_{mle}\})}\right).$$

5. Show that correlated random variables $F^{-1}(U)$ and $G^{-1}(U)$ have maximum positive correlation and have the distributions F and G; for maximal negative correlation it is enough take $G^{-1}(-U)$.

Uniform on Sphere

This MATHEMATICA code gives the exact value of the Bayes rule for $\boldsymbol{x} = (0, 0, 0, 0, 0, 1)$. The code

```
alpha=2; beta=3;
p=6;x = Table[0,{p}]; x[[6]]=1;
delta = 2 * alpha/p    1/(2 + beta) Hypergeometric1F1[alpha+1,
          (p+2)/2, Norm[x]^2/(2 + 4 beta)]/
        Hypergeometric1F1[alpha,p/2, Norm[x]^2/(2 + 4 beta)]  x//N
```

results in $\delta = \{0, 0, 0, 0, 0, 0.134123\}$.