Therefore, the posterior expectation of the function

$$n\nabla\psi(\theta) = \sum_{i=1}^{n} \frac{\exp(x_i^t\theta)}{1+\exp(x_i^t\theta)}\, x_i$$

is known and equal to $(\sum_i Y_i x_i + \zeta)/(\lambda+1)$ under the prior distribution $\pi(\theta|\zeta,\lambda)$. Unfortunately, a control variate version of

$$\delta_1 = \frac{1}{m}\sum_{j=1}^{m} \theta_j$$

is not available since the optimal constant β^* (or even its sign) cannot be evaluated, except by the regression of the θ_j's upon the

$$\sum_i \frac{\exp x_i^t\theta_j}{1+\exp x_i^t\theta_j}\, x_i\,.$$

Thus the fact that the posterior mean of $\nabla\psi(\theta)$ is known does not help us to establish a control variate estimator. This information can be used in a more informal way to study convergence of δ_1 (see, for instance, Robert 1993). ‖

In conclusion, the technique of control variates is manageable only in very specific cases: the control function h must be available, as well as the optimal weight β^*. See, however, Brooks and Gelman (1998b) for a general approach based on the score function (whose expectation is null under general regularity conditions).

4.5 Problems

4.1 (Chen and Shao 1997) As mentioned, normalizing constants are superfluous in Bayesian inference except in the case when several models are considered at once (as in the computation of Bayes factors). In such cases, where $\pi_1(\theta) = \tilde{\pi}_1(\theta)/c_1$ and $\pi_2(\theta) = \tilde{\pi}_2(\theta)/c_2$, and only $\tilde{\pi}_1$ and $\tilde{\pi}_2$ are known, the quantity to approximate is $\varrho = c_1/c_2$ or $\xi = \log(c_1/c_2)$.

(a) Show that the ratio ϱ can be approximated by

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\tilde{\pi}_1(\theta_i)}{\tilde{\pi}_2(\theta_i)}, \qquad \theta_1,\ldots,\theta_n \sim \pi_2.$$

(*Hint:* Use an importance sampling argument.)

(b) Show that

$$\frac{\int \tilde{\pi}_1(\theta)\alpha(\theta)\pi_2(\theta)d\theta}{\int \tilde{\pi}_2(\theta)\alpha(\theta)\pi_1(\theta)d\theta} = \frac{c_1}{c_2}$$

holds for every function $\alpha(\theta)$ such that both integrals are finite.

(c) Deduce that

$$\frac{\frac{1}{n_2}\sum_{i=1}^{n_2}\tilde{\pi}_1(\theta_{2i})\alpha(\theta_{2i})}{\frac{1}{n_1}\sum_{i=1}^{n_1}\tilde{\pi}_2(\theta_{1i})\alpha(\theta_{1i})},$$

with $\theta_{1i}\sim\pi_1$ and $\theta_{2i}\sim\pi_2$, is a convergent estimator of ϱ.

(d) Show that part (b) covers the case of the Newton and Raftery (1994) representation

$$\frac{c_1}{c_2}=\frac{\mathbb{E}^{\pi_2}[\tilde{\pi}_2(\theta)^{-1}]}{\mathbb{E}^{\pi_1}[\tilde{\pi}_1(\theta)^{-1}]}.$$

(e) Show that the optimal choice (in terms of mean square error) of α in part (c) is

$$\alpha(\theta)=c\frac{n_1+n_2}{n_1\pi_1(\theta)+n_2\pi_2(\theta)},$$

where c is a constant. (*Note:* See Meng and Wong 1996.)

4.2 (Continuation of Problem 4.1) When the priors π_1 and π_2 belong to a parameterized family (that is, $\pi_i(\theta)=\pi(\theta|\lambda_i)$), the corresponding constants are denoted by $c(\lambda_i)$.

(a) Verify the identity

$$-\log\left(\frac{c(\lambda_1)}{c(\lambda_2)}\right)=\mathbb{E}\left[\frac{U(\theta,\lambda)}{\pi(\lambda)}\right],$$

where

$$U(\theta,\lambda)=\frac{d}{d\lambda}\log(\tilde{\pi}(\theta|\lambda))$$

and $\pi(\lambda)$ is an arbitrary distribution on λ.

(b) Show that ξ can be estimated with the *bridge estimator* of Gelman and Meng (1998),

$$\hat{\xi}=\frac{1}{n}\sum_{i=1}^{n}\frac{U(\theta_i,\lambda_i)}{\pi(\lambda_i)},$$

when the (θ_i,λ_i)'s are simulated from the joint distribution induced by $\pi(\lambda)$ and $\pi(\theta|\lambda_i)$.

(c) Show that the minimum variance estimator of ξ is based on

$$\pi(\lambda)\propto\sqrt{\mathbb{E}_\lambda[U^2(\theta,\lambda)]}$$

and examine whether this solution gives the Jeffreys prior.

4.3 For the situation of Example 4.2:

(a) Show that $\hat{\delta}_m^\pi(x)\to\delta^\pi(x)$ as $m\to\infty$.

(b) Show that the Central Limit Theorem can be applied to $\hat{\delta}_m^\pi(x)$.

(c) Generate random variables $\theta_1,\ldots,\theta_m\sim\mathcal{N}(x,1)$ and calculate $\hat{\delta}_m^\pi(x)$ for $x=0,1,4$. Use the Central Limit Theorem to construct a measure of accuracy of your calculation.

4.4 Verify equation (4.10), that is, show that

$$\begin{aligned}\mathbb{E}\left[e^{-X^2}|y\right]&=\frac{1}{\sqrt{2\pi\sigma^2 y}}\int_{-\infty}^{\infty}e^{-x^2}e^{-(x-\mu)^2/2\sigma^2 y}dx\\&=\frac{1}{\sqrt{2\sigma^2 y+1}}\exp-\frac{\mu^2}{1+2\sigma^2 y},\end{aligned}$$

by completing the square in the exponent to evaluate the integral.

4.5 In simulation from mixture densities, it is always possible to set up a Rao–Blackwellized alternative to the empirical average.

(a) Show that, if $f(x) = \int g(x|y)h(y)dy$, then

$$\frac{1}{M}\sum_{i=1}^{M} h(X_i), \quad X_i \sim f \text{ and } \frac{1}{M}\sum_{i=1}^{M} \mathbb{E}(h(X)|Y_i), \quad Y_i \sim g,$$

each converge to $\mathbb{E}_f(X)$.

(b) For each of the following cases, generate random variables X_i and Y_i, and compare the empirical average and Rao–Blackwellized estimator of $\mathbb{E}_f(X)$ and $\text{var}_f(X)$:

a) $X|y \sim \mathcal{P}(y)$, $Y \sim \mathcal{G}a(a, b)$ (X is negative binomial);

b) $X|y \sim \mathcal{N}(0, y)$, $Y \sim \mathcal{G}a(a, b)$ (X is a generalized t);

c) $X|y \sim \mathcal{B}in(y)$, $Y \sim \mathcal{B}e(a, b)$ (X is beta-binomial).

4.6 For the estimator δ_2 of Section 4.2:

(a) Verify the expression (4.11) for ρ_i.

(b) Verify the recursion (4.12).

(c) Prove Proposition 4.6. (*Hint:* Show that $\mathbb{E}[\delta_2] = \mathbb{E}[\delta_1]$ and apply the Rao–Blackwell Theorem.)

4.7 Referring to Lemma 4.3, show how to use (4.5) to derive the expression for the asymptotic variance of δ_h^n.

4.8 Given an Accept–Reject algorithm based on (f, g, ρ), we denote by

$$b(y_j) = \frac{(1-\rho)f(y_j)}{g(y_j) - \rho f(y_j)}$$

the importance sampling weight of the rejected variables $(Y_1, \ldots, Y_t)$, and by $(X_1, \ldots, X_n)$ the accepted variables.

(a) Show that the estimator

$$\delta_1 = \frac{n}{n+t}\,\delta^{AR} + \frac{t}{n+t}\,\delta_0,$$

with

$$\delta_0 = \frac{1}{t}\sum_{j=1}^{t} b(Y_j)h(Y_j)$$

and

$$\delta^{AR} = \frac{1}{n}\sum_{i=1}^{n} h(X_i),$$

does not uniformly dominate δ^{AR}. (*Hint:* Consider the constant functions.)

(b) Show that

$$\delta_{2w} = \frac{n}{n+t}\,\delta^{AR} + \frac{t}{n+t}\sum_{j=1}^{t} \frac{b(Y_j)}{S_t}h(Y_j)$$

is asymptotically equivalent to δ_1 in terms of bias and variance.

(c) Deduce that δ_{2w} asymptotically dominates δ^{AR} if (4.20) holds.

4.9 Referring to Section 4.1.2:

(a) Show that $\text{cov}(\bar{\mathbf{X}}_k, \bar{\mathbf{X}}_{k'}) = \sigma^2/\max\{k, k'\}$, regardless of the distribution of X_i.

(b) Verify that Σ^{-1} is given by (4.7) for $n = 3, 10, 25$.
(c) Establish a recursion relation for calculating the elements a_{ij} of Σ^{-1}:

$$a_{ij} = \begin{cases} 2i^2 & \text{if } i = j < n \\ n^2 & \text{if } i = j = n \\ -ij & \text{if } |i-j| = 1 \\ 0 & \text{otherwise.} \end{cases}$$

(d) If we denote by Σ_k^{-1} the inverse matrix corresponding to $(\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k)$, show that to get Σ_{k+1}^{-1} you only have to change one element, then add one row and one column to Σ_k^{-1}.

4.10 (a) Establish (4.8) by showing that (i) $\mathbf{1}'\Sigma^{-1}\bar{\mathbf{x}} = n$ and (ii) $\mathbf{1}'\Sigma^{-1}\mathbf{1} = n$.
(b) Show that a reasonable approximation for d_n is $\hat{\sigma}^2(n + 2\sqrt{2n})$. (*Hint:* Consider the mean and variance of the χ_n^2 distribution.)

4.11 Referring to Example 4.2:
(a) Compare a running mean plot with ordinary univariate normal error bars to the variance assessment of (4.9). Discuss advantages and disadvantages.
(b) Compare a running mean plot with empirical error bars to the variance assessment of (4.9). Use 500 estimates to calculate 95% error bars. Discuss advantages and disadvantages.
(c) Repeat parts (*a*) and (*b*) for $h(x) = x^2$ and thus assess the estimate of the posterior variance of θ.

4.12 In the setting of Section 4.3, examine whether the substitution of

$$\sum_{i=1}^{n-1} (X_{(i+1)} - X_{(i)}) \, \frac{h(X_{(i)}) + h(X_{(i+1)})}{2}$$

into

$$\sum_{i=1}^{n-1} (X_{(i+1)} - X_{(i)}) \, h(X_{(i)})$$

improves the speed of convergence. (*Hint*: Examine the influence of the remainder terms

$$\int_{-\infty}^{X_{(1)}} h(x)f(x)dx \quad \text{and} \quad \int_{X_{(n)}}^{+\infty} h(x)f(x)dx \, . \Bigg)$$

4.13 Show that it is always possible to express an integral $\mathfrak{I}$ as $\mathfrak{I} = \int_0^1 \tilde{h}(y)\tilde{f}(y)dy$, where the integration is over $(0, 1)$ and $\tilde{h}$ and $\tilde{f}$ are transforms of the original functions.

4.14 In this problem we will prove Proposition 4.9
(a) Define $U_{-1} = 0$ and $U_{m+1} = 1$, and show that δ can be written

$$\delta = \sum_{i=-1}^{m} h(U_i)(U_{i+1} - U_i) = \sum_{i=-1}^{m} \int_{U_i}^{U_{i+1}} h(U_i)du,$$

and thus the difference $(\mathfrak{I} - \delta)$ can be written as $\sum_{i=-1}^{m} \int_{U_i}^{U_{i+1}} (h(u) - h(U_i)) \, du$.
(b) Show that the first-order expansion $h(u) = h(U_i) + h'(\zeta)(u - U_i)$, $\zeta \in [U_i, u]$, implies that $|h(u) - h(U_i)| < c(u - U_i)$, with $c = \sup_{[0,1]} |h'(X)|$.

(c) Show that

$$\text{var}(\delta) = \mathbb{E}[(\mathfrak{I} - \delta)^2] < c^2 \left\{ (m+2)\, \mathbb{E}[Z_i^4] + (m+1)(m+2)\, \mathbb{E}[Z_i^2 Z_j^2] \right\},$$

where $Z_i = U_{i+1} - U_i$.

(d) The U_i's are the order statistics of a uniform distribution. Show that (i) the variables Z_i are jointly distributed according to a Dirichlet distribution $\mathcal{D}_m(1, \ldots, 1)$, with $Z_i \sim \mathcal{B}e(1, m)$ and $(Z_i, Z_j, 1 - Z_i - Z_j) \sim \mathcal{D}_3(1, 1, m-1)$, (ii) $\mathbb{E}[Z_i^4] = m$, $\int_0^1 z^4 (1-z)^{m-1} dz = \frac{24\, m!}{(m+4)!}$, and $\mathbb{E}[Z_i^2 Z_j^2] = \frac{4\, m!}{(m+4)!}$.

(e) Finally, establish that

$$\text{var}(\delta) \leq c \left\{ \frac{24}{(m+4)(m+3)(m+1)} + \frac{4}{(m+4)(m+3)} \right\} = \mathcal{O}(m^{-2}),$$

proving the proposition.

4.15 For the situation of Example 4.14:

(a) Verify (4.18).

(b) Verify the conditions on β in order for δ_2 to improve on δ_1.

(c) For f the density of $\mathcal{N}(0,1)$, find $P(X > a)$ for $a = 3, 5, 7$.

(d) For f the density of $\mathcal{T}_5$, find $P(X > a)$ for $a = 3, 5, 7$.

(e) For f the density of $\mathcal{T}_5$, find a such that $P(X > a) = .01, .001, .0001$.

4.16 A naïve way to implement the antithetic variable scheme is to use both U and $(1 - U)$ in an inversion simulation. Examine empirically whether this method leads to variance reduction for the distributions (i) $f_1(x) = 1/\pi(1 + x^2)$, (ii) $f_2(x) = \frac{1}{2} e^{-|x|}$, (iii) $f_3(x) = e^{-x} \mathbb{I}_{x>0}$, (iv) $f_4(x) = \frac{2}{\pi\sqrt{3}} \left(1 + x^2/3\right)^{-2}$, and (v) $f_5(x) = 2x^{-3} \mathbb{I}_{x>1}$. Examine variance reductions of the mean, second moment, median, and 75th percentile.

To calculate the weights for the Rao–Blackwellized estimator of Section 4.2, it is necessary to derive properties of the distribution of the random variables in the Accept–Reject algorithm [*A*.4]. The following problem is a rather straightforward exercise in distribution theory and is only made complicated by the stopping rule of the Accept–Reject algorithm.

4.17 This problem looks at the performance of a *termwise* Rao–Blackwellized estimator. Casella and Robert (1998) established that such an estimator does not sacrifice much performance over the full Rao–Blackwellized estimator of Proposition 4.6. Given a sample $(Y_1, \ldots, Y_N)$ produced by an Accept–Reject algorithm to accept m values, based on (f, g, M):

(a) Show that

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\mathbb{I}_{U_i \leq \omega_i} | Y_i] h(Y_i) = \frac{1}{n-m} \left(h(Y_n) + \sum_{i=1}^{n-1} b(Y_i) h(Y_i) \right),$$

with

$$b(Y_i) = \left(1 + \frac{m(g(Y_i) - \rho f(Y_i))}{(n-m-1)(1-\rho) f(Y_i)} \right)^{-1}.$$

(b) If $S_n = \sum_1^{n-1} b(y_i)$, show that

$$\delta = \frac{1}{n-m}\left(h(Y_n) + \frac{n-m-1}{S_n}\sum_{i=1}^{nt-1} b(Y_i)h(Y_i)\right)$$

asymptotically dominates the usual Monte Carlo approximation, conditional on the number of rejected variables m under quadratic loss. (*Hint:* Show that the sum of the weights S_n can be replaced by $(n-m-1)$ in δ and assume $\mathbb{E}_f[h(X)] = 0$.)

4.18 Strawderman (1996) adapted the control variate scheme to the Accept–Reject algorithm. When $Y_1, \ldots, Y_N$ is the sample produced by an Accept–Reject algorithm based on g, let m denote the density

$$m(y) = \frac{t-1}{n-1}f(y) + \frac{n-t}{n-1}\frac{g(y)-\rho f(y)}{1-\rho}\ ,$$

when $N = n$ and $\rho = \dfrac{1}{M}$.

(a) Show that

$$\mathfrak{I} = \int h(x)f(x)dx = \mathbb{E}_N\left[\mathbb{E}\left[\frac{h(Y)f(Y)}{m(Y)}\middle| N\right]\right]\ ,$$

where m is the marginal density of Y_i (see Problem 3.29).

(b) Show that for any function $c(\cdot)$ and some constant β,

$$\mathfrak{I} = \beta\mathbb{E}[c(Y)] + \mathbb{E}\left[\frac{h(Y)f(Y)}{m(Y)} - \beta c(Y)\right]\ .$$

(c) Setting $d(y) = h(y)f(y)/m(y)$, show that the optimal choice of β is

$$\beta^* = \mathrm{cov}[d(Y), c(Y)]/\mathrm{var}[c(Y)].$$

(d) Examine choices of c for which the optimal β can be constructed and, thus, where the control variate method applies.

(*Note:* Strawderman 1996 suggests estimating β with $\hat{\beta}$, the estimated slope of the regression of $d(y_i)$ on $c(y_i)$, $i = 1, 2, \ldots, n-1$.)

4.19 For $t \sim \mathcal{G}eo(\rho)$, show that

$$\mathbb{E}[t^{-1}] = -\frac{\rho\log\rho}{1-\rho}\ , \qquad \mathbb{E}[t^{-2}] = \frac{\rho\mathrm{Li}(1-\rho)}{1-\rho}\ ,$$

where $\mathrm{Li}(x)$ is the dilog function (see Note 4.6.2).

4.20 (Continuation of Problem 4.19) If $N \sim \mathcal{N}eg(n, \rho)$, show that

$$\mathbb{E}[(N-1)^{-1}] = \frac{\rho}{n-1}\ , \quad \mathbb{E}[(N-2)^{-1}] = \frac{\rho(n+\rho-2)}{(n-1)(n-2)}\ ,$$

$$\mathbb{E}[((N-1)(N-2))^{-1}] = \frac{\rho^2}{(n-1)(n-2)}\ ,$$

$$\mathbb{E}[(N-1)^{-2}] = \frac{\rho^t{}_2F_1(n-1, n-1; n; 1-\rho)}{(n-1)^2}\ .$$

4.21 (Continuation of Problem 4.19) If Li is the dilog function, show that

$$\lim_{\rho\to 1}\frac{\rho\mathrm{Li}(1-\rho)}{1-\rho}=1 \qquad \text{and} \qquad \lim_{\rho\to 1}\log(\rho)\mathrm{Li}(1-\rho)=0\,.$$

Deduce that the domination corresponding to (4.20) occurs on an interval of the form $[\rho_0, 1]$.

4.22 Given an integral

$$\mathfrak{I}=\int_{\mathcal{X}} h(x)f(x)dx$$

to be evaluated by simulation, compare the usual Monte Carlo estimator

$$\delta_a=\frac{1}{np}\sum_{i=1}^{np}h(x_i)$$

based on an iid sample $(X_1,\ldots,X_{np})$ with a stratified estimator (see Note 4.6.3)

$$\delta_a=\sum_{j=1}^{p}\frac{p_j}{n}\sum_{i=1}^{n}h(Y_i^p)\,,$$

where $p_j=\displaystyle\int_{\mathcal{X}_j} f(x)dx$, $\mathcal{X}=\displaystyle\bigcup_{j=1}^{p}\mathcal{X}_j$ and $(Y_1^j,\ldots,Y_n^j)$ is a sample from $f\mathbb{I}_{\mathcal{X}_j}$. Show that δ_2 does not bring any improvement if the p_j's are unknown and must be estimated.

4.6 Notes

4.6.1 Monitoring Importance Sampling Convergence

With reference to convergence control for simulation methods, importance sampling methods can be implemented in a *monitored* way; that is, in parallel with other evaluation methods. These can be based on alternative instrumental distributions or other techniques (standard Monte Carlo, Markov chain Monte Carlo, Riemann sums, Rao–Blackwellization, etc.). The respective samples then provide separate evaluations of $\mathbb{E}[h(X)]$, through (3.8), (3.11), or yet another estimator (as in Section 3.3.3), and the convergence criterion is to stop when most estimators are close enough. Obviously, this empirical method is not completely foolproof, but it generally prevents pseudo-convergences when the instrumental distributions are sufficiently different. On the other hand, this approach is rather *conservative*, as it is only as fast as the slowest estimator. However, it may also point out instrumental distributions with variance problems. From a computational point of view, an efficient implementation of this control method relies on the use of *parallel programming* in order to weight each distribution more equitably, so that a distribution f of larger variance, compared with another distribution g, may compensate this drawback by a lower computation time, thus producing a larger sample in the same time.[5]

[5] This feature does not necessarily require a truly parallel implementation, since it can be reproduced by the cyclic allocation of uniform random variables to each of the distributions involved.