J. R. Statist. Soc. B (2003)
65, *Part* 3, *pp.* 585–618

A theory of statistical models for Monte Carlo integration

A. Kong,

deCODE Genetics, Reykjavik, Iceland

P. McCullagh,

University of Chicago, USA

X.-L. Meng

Harvard University, Cambridge, USA

and D. Nicolae and Z. Tan

University of Chicago, USA

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, December 11th, 2002*, Professor D. Firth *in the Chair*]

Summary. The task of estimating an integral by Monte Carlo methods is formulated as a statistical model using simulated observations as data. The difficulty in this exercise is that we ordinarily have at our disposal all of the information required to compute integrals exactly by calculus or numerical integration, but we choose to ignore some of the information for simplicity or computational feasibility. Our proposal is to use a semiparametric statistical model that makes explicit what information is ignored and what information is retained. The parameter space in this model is a set of measures on the sample space, which is ordinarily an infinite dimensional object. None-the-less, from simulated data the base-line measure can be estimated by maximum likelihood, and the required integrals computed by a simple formula previously derived by Vardi and by Lindsay in a closely related model for biased sampling. The same formula was also suggested by Geyer and by Meng and Wong using entirely different arguments. By contrast with Geyer's retrospective likelihood, a correct estimate of simulation error is available directly from the Fisher information. The principal advantage of the semiparametric model is that variance reduction techniques are associated with submodels in which the maximum likelihood estimator in the submodel may have substantially smaller variance than the traditional estimator. The method is applicable to Markov chain and more general Monte Carlo sampling schemes with multiple samplers.

Keywords: Biased sampling model; Bridge sampling; Control variate; Exponential family; Generalized inverse; Importance sampling; Invariant measure; Iterative proportional scaling; Log-linear model; Markov chain Monte Carlo methods; Multinomial distribution; Normalizing constant; Semiparametric model; Retrospective likelihood

1. Normalizing constants and Monte Carlo integration

Certain inferential problems arising in statistical work involve awkward summation or high dimensional integrals that are not analytically tractable. Many of these problems are such that

Address for correspondence: P. McCullagh, Department of Statistics, University of Chicago, 5734 University Avenue, Chicago, IL 60657-1514, USA.
E-mail: pmcc@galton.uchicago.edu

 1369–7412/03/65585

only ratios of integrals are required, and it is this class of problems with which we shall be concerned in this paper. To establish the notation, let Γ be a set, let μ be a measure on Γ, let $\{q_\theta\}$ be a family of functions on Γ and let

$$c(\theta) = \int_\Gamma q_\theta(x)\, \mathrm{d}\mu.$$

Our goal is ideally to compute exactly, or in practice to estimate, the ratios $c(\theta)/c(\theta')$ for all values θ and θ' in the family. The family may contain a reference function q_{θ_0} whose integral is known, in which case the remaining integrals are directly estimable by reference to the standard. Our theory accommodates but does not require such a standard. For an estimator to be useful, an approximate measure of estimation error is also required.

We refer to $c(\theta)$ as the normalizing constant associated with the function $q_\theta(x)$. In particular, if q_θ is non-negative and $0 < c(\theta) < \infty$,

$$\mathrm{d}P_\theta(x) = q_\theta(x)\, \mathrm{d}\mu/c(\theta)$$

is a probability distribution on Γ. For the method to work, the family must contain at least one non-negative function q_θ, but it is not necessary that all of them be non-negative. Depending on the context and on the functions q_θ, the normalization constant might represent anything from a posterior expectation in a Bayesian calculation to a partition function in statistical physics.

Observations simulated from one or more of these distributions are the key ingredient in Monte Carlo integration. We assume throughout the paper that techniques are available to simulate from P_θ without computing the normalization constant. At least initially, we assume that these techniques generate a stream of independent observations from P_θ.

At first glance, the problem appears to be an exercise in calculus or numerical analysis, and not amenable to statistical formulation. After all, statistical theory does not seek to avoid estimators that are difficult to compute; nor is it inclined to opt for inferior estimators because they are convenient for programming. So it is hard to see how any efficient statistical formulation could avoid the obvious and excellent estimator $c(\theta) = \int q_\theta(x)\, \mathrm{d}\mu$, which has zero variance and requires no simulated data.

This paper demonstrates that the exercise can nevertheless be formulated as a model-based statistical estimation problem in which the parameter space is determined by how much information we choose to *ignore*. In effect, the statistical model serves to estimate that part of the information that is ignored and uses the estimate to compute the required integrals in a manner that is asymptotically efficient given the information available. Neither the nature nor the extent of the ignored information is predetermined. By judicious use of group invariant submodels and other submodels, the amount of information ignored may be controlled in such a way that the simulation variance is reduced with little increase in computational effort.

The literature on Monte Carlo estimation of integrals is very extensive, and no attempt will be made here to review it. For good summaries, see Hammersley and Hanscomb (1964), Ripley (1987), Evans and Swartz (2000) or Liu (2001). For overviews on the computation of normalizing constants, see DiCiccio *et al.* (1997) or Gelman and Meng (1998).

2. Illustration

The following example with sample space $\Gamma = \mathcal{R} \times \mathcal{R}^+$ is sufficiently simple that the integrals can be computed analytically. Nevertheless, it illustrates the gains that are achievable by choice of

design and by choice of submodel. Three Monte Carlo techniques are described and compared.

Suppose that we need to evaluate the integrals over the upper half-plane

$$c_\sigma = \int_\Gamma \frac{\mathrm{d}x_1\,\mathrm{d}x_2}{\{x_1^2 + (x_2+\sigma)^2\}^2}$$

for $\sigma \in \{0.25, 0.5, 1.0, 2.0, 4.0\}$. It is conventional to take μ to be Lebesgue measure, so that $q_\sigma(x) = 1/\{x_1^2 + (x_2+\sigma)^2\}^2$. As it happens, the distribution P_σ has mean $(0, \sigma)$ with infinite variances and covariances. Consider first an importance sampling design in which a stream of independent observations $x_1, \ldots, x_n$ is made available from the distribution P_1. The importance sampling estimator of c_σ/c_1 is

$$\hat{c}_\sigma/\hat{c}_1 = n^{-1} \sum q_\sigma(x_i)/q_1(x_i).$$

By applying the results from Section 4, we find on the basis of $n = 500$ simulations that the matrix

$$\hat{V} = n^{-1} \begin{pmatrix} 4.411 & 1.491 & 0.000 & -0.601 & -0.821 \\ 1.491 & 0.641 & 0.000 & -0.383 & -0.582 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ -0.601 & -0.383 & 0.000 & 0.578 & 1.273 \\ -0.821 & -0.582 & 0.000 & 1.273 & 3.591 \end{pmatrix}$$

is such that the asymptotic variance of $\log(\hat{c}_r/\hat{c}_s)$ is equal to $\hat{V}_{rr} + \hat{V}_{ss} - 2\hat{V}_{rs}$ for $r, s \in \{0.25, 0.5, 1.0, 2.0, 4.0\}$. The individual components c_r are not identifiable in the model and the estimates do not have a variance, asymptotic or otherwise. The estimated variances of the 10 pairwise logarithmic contrasts range from $0.6/n$ to $9.6/n$, with an average of $3.6/n$. The matrix $\hat{V}$ is obtained from the observed Fisher information, so a different simulation will yield a slightly different matrix.

Suppose as an alternative that it is feasible to simulate from any or all of the distributions P_σ, as in Geyer (1994), Hesterberg (1995), Meng and Wong (1996) or Owen and Zhou (2000). Various simulation designs, also called defensive importance sampling or bridge sampling plans, can now be considered in which n_r observations are generated from P_r. These are called the design weights, or bridge sampling weights. For simplicity we consider the uniform design in which $n_r = n/5$. The importance sampling estimator must now be replaced by the more general maximum likelihood estimator derived in Section 3, which is obtained by solving

$$\hat{c}_\sigma = \sum_{i=1}^{n} \frac{q_\sigma(x_i)}{\sum_s n_s \hat{c}_s^{-1} q_s(x_i)}. \tag{2.1}$$

The asymptotic covariance matrix of $\log(\hat{c})$ obtained from a sample of $n = 500$ simulated observations using equation (4.2) is

$$\hat{V}' = n^{-1} \begin{pmatrix} 2.298 & 0.974 & -0.263 & -1.197 & -1.811 \\ 0.974 & 0.668 & 0.077 & -0.588 & -1.131 \\ -0.263 & 0.077 & 0.239 & 0.117 & -0.170 \\ -1.197 & -0.588 & 0.117 & 0.690 & 0.979 \\ -1.811 & -1.131 & -0.170 & 0.979 & 2.132 \end{pmatrix}.$$

In using this matrix, it must be borne in mind that $c \equiv \lambda c$ for every positive scalar λ, so only contrasts of $\log(c)$ are identifiable. The estimated variances of the 10 pairwise logarithmic

contrasts range from $0.7/n$ to $8.1/n$, with an average of $3.0/n$. The average efficiency factor of the uniform design relative to the preceding importance sampling design is conveniently measured by the asymptotic variance ratio $3.6/3.0 = 1.2$.

A third Monte Carlo variant uses a submodel with a reduced parameter space consisting of measures that are invariant under group action. Details are described in Section 3.3, but the operation proceeds as follows. Consider the two-element group $\mathcal{G} = \pm 1$ in which the inversion $g = -1$ acts on Γ by reflection in the unit circle

$$g : (x_1, x_2) \mapsto (x_1, x_2)/(x_1^2 + x_2^2).$$

By construction, $g^2 = 1$ and $g^{-1} = g$, so this is a group action. Lebesgue measure is not invariant and is thus not in the parameter space as determined by this action. However, the measure $\rho(\mathrm{d}x) = \mathrm{d}x_1\, \mathrm{d}x_2/x_2^2$ is invariant, so we compensate by writing

$$q(x; \sigma) = x_2^2/\{x_1^2 + (x_2 + \sigma)^2\}^2$$

for the new integrand, and $c_\sigma = \int_\Gamma q(x; \sigma)\, \mathrm{d}\rho$. The submodel estimator is the same as equation (2.1), but with q replaced by the group average

$$\bar{q}(x; \sigma) = \tfrac{1}{2}\, q(x; \sigma) + \tfrac{1}{2}\, q(gx; \sigma).$$

With a uniform design and $n = 500$ observations, the estimated variance matrix of $\log(\hat{c})$ is

$$\hat{V}'' = n^{-1} \begin{pmatrix} 0.202 & -0.093 & -0.216 & -0.093 & 0.202 \\ -0.093 & 0.045 & 0.097 & 0.045 & -0.093 \\ -0.216 & 0.097 & 0.239 & 0.097 & -0.216 \\ -0.093 & 0.045 & 0.097 & 0.045 & -0.093 \\ 0.202 & -0.093 & -0.216 & -0.093 & 0.202 \end{pmatrix}.$$

The variances of the 10 pairwise logarithmic contrasts range from 0 to $0.87/n$ with an average of $0.37/n$. For reasons described in Section 3.3, the two ratios $c_{0.25}/c_4$ and $c_{0.5}/c_2$ are estimated exactly with *zero* variance. Relative to the preceding Monte Carlo estimator, group averaging reduces the average simulation variance of contrasts by an efficiency factor of 8.1. By this comparison, n simulated observations using the group-averaged estimator are approximately equivalent to $8n$ observations using the estimator (2.1) with the same design weights, but without group averaging.

To achieve efficiency gains of this magnitude, it is not necessary to use a large group, but it is necessary to have a good understanding of the integrands in a qualitative sense, and to select the group action accordingly. If the group action had been chosen so that $g(x_1, x_2) = (-x_1, x_2)$ the gain in efficiency would have been zero. Arguably, the gain in efficiency would be negative because of the slight increase in computational effort.

Given observations simulated by any of the preceding schemes, equation (2.1), or the group-averaged version, can also be used for the estimation of integrals such as

$$c'_\sigma = \int_\Gamma \frac{\log(x_1^2 + x_2^2)\, \mathrm{d}x_1\, \mathrm{d}x_2}{\{x_1^2 + (x_2 + \sigma)^2\}^2}$$

in which the integrand is not positive, and there is no associated probability distribution. In this extended version of the problem, 10 integrals are estimated simultaneously, and c'_σ/c_σ is the expected value of $\log|X_1^2 + X_2^2|$ when $X \sim P_\sigma$. For this example, $c_\sigma = \pi/4\sigma^2$, and $c'_\sigma/c_\sigma = 2\, \log(\sigma)$. The general theory described in the next section covers integrals of this sort, and the Fisher information also provides a variance estimate.

3. A semiparametric model

3.1. Problem formulation

The statistical problem is formulated as a challenge issued by one individual called the simulator and accepted by a second individual called the statistical analyst. In practice, these typically are two personalities of one individual, but for clarity of exposition we suppose that two distinct individuals are involved. The simulator is omniscient, honest and secretive but willing to provide data in essentially unlimited quantity. Partial information is available to the analyst in the form of a statistical model and the data made available by the simulator.

Let $q_1, \ldots, q_k$ be real-valued functions on Γ, known to the analyst, and let μ be any non-negative measure on Γ. The challenge is to compute the ratios c_r/c_s, where each integral $c_r = \int_\Gamma q_r(x)\, \mathrm{d}\mu$ is assumed to be finite, i.e. we are interested in estimating all ratios simultaneously, where $q_r(x) = q_{\theta_r}(x)$, using the notation of Section 1. Assume that there is at least one non-negative function q_r such that $0 < c_r < \infty$, and that n_r observations from the weighted distribution

$$P_r(\mathrm{d}x) = c_r^{-1}\, q_r(x)\, \mu(\mathrm{d}x) \tag{3.1}$$

are made available by the simulator at the behest of the analyst. The analyst's design vector $(n_1, \ldots, n_k)$ has at least one r such that $n_r > 0$. Typically, however, many of the functions q_r are such that $n_r = 0$. For these functions, the non-negativity condition is not required.

Different versions of the problem are available depending on what is known to the analyst about μ. Four of these are now described.

(a) If μ is known, e.g. if μ is Lebesgue measure, the constants can, in principle, be determined exactly by integral calculus.
(b) If μ is known up to a positive scalar multiple, the constants can be determined modulo the same scalar multiple, and the ratios can be determined exactly by integral calculus.
(c) If μ is completely unknown, neither the constants nor their ratios can be determined by calculus alone. Nevertheless, the ratios can be estimated consistently from simulated data by using a slight modification of Vardi's (1985) biased sampling model.
(d) If partial information is available concerning μ, neither the constants nor their ratios can be determined by calculus. Nevertheless, partial information may permit a substantial gain of efficiency by comparison with (c).

As an exercise in integral calculus, the first and second versions of the problem are not considered further in this paper.

3.2. A full exponential model

In this section, we focus on a semiparametric model in which the parameter μ is a measure or distribution on Γ, completely unknown to the analyst. The simulator is free to choose any measure μ, and the analyst's estimate must be consistent regardless of that choice. The parameter space is thus the set Θ of all non-negative measures on Γ, not necessarily probability distributions, and the components of interest are the linear functionals

$$c_r = \int_\Gamma q_r(x)\, \mathrm{d}\mu. \tag{3.2}$$

The state space Γ in the following analysis is assumed to be countable; the more general argument covering uncountable spaces is given by Vardi (1985) in a model for biased sampling.

We suppose that simulated data are available in the form $(y_1, x_1), \ldots, (y_n, x_n)$, in which the pairs are independent, $y_i \in \{1, \ldots, k\}$ is determined by the simulation design and x_i is a random draw from the distribution P_{y_i}. Then for each draw x from the distribution P_r the likelihood contribution is

$$P_r(\{x\}) = c_r^{-1}\, q_r(x)\, \mu(\{x\}).$$

The full likelihood at μ is thus

$$L(\mu) = \prod_{i=1}^{n} P_{y_i}(\{x_i\}) = \prod_{i=1}^{n} \mu(\{x_i\}) c_{y_i}^{-1}\, q_{y_i}(x_i).$$

It is helpful at this stage to reparameterize the model in terms of the canonical parameter $\theta \in \mathcal{R}^{\Gamma}$ given by $\theta(x) = \log[\mu(\{x\})]$. Let $\hat{P}$ be the empirical measure on Γ placing mass $1/n$ at each data point. Ignoring additive constants, the log-likelihood at θ is

$$\sum_{i=1}^{n} \theta(x_i) - \sum_{s=1}^{k} n_s\, \log\{c_s(\theta)\} = n \int_{\Gamma} \theta(x)\, \mathrm{d}\hat{P} - \sum_{s=1}^{k} n_s\, \log\{c_s(\theta)\}. \tag{3.3}$$

Although it may appear paradoxical at first, the canonical sufficient statistic $\hat{P}$, the empirical distribution of the simulated values $\{x_1, \ldots, x_n\}$, ignores a part of the data that might be considered highly informative, namely the association of distribution labels with simulated values. In fact, all permutations of the labels y give the same likelihood. The likelihood function is thus unaffected by reassignment of the distribution labels y to the draws x. This point was previously noted by Vardi (1985), section 6. Thus, under the model as specified, or under any submodel, the association of draws with distribution labels is uninformative. The reason for this is that all the information in the labels for estimating the ratios is contained in the design constants $\{n_1, \ldots, n_k\}$.

As is evident from equation (3.3), the model is of the full exponential family type with canonical parameter θ, and canonical sufficient statistic $\hat{P}$. The maximum likelihood estimate of μ, obtained by equating the canonical sufficient statistic to its expectation, is

$$n\, \hat{P}(\mathrm{d}x) = \sum_{s=1}^{k} n_s \hat{c}_s^{-1}\, q_s(x)\, \hat{\mu}(\mathrm{d}x),$$

where $\mathrm{d}x = \{x\}$. Thus

$$\hat{\mu}(\mathrm{d}x) = n\, \hat{P}(\mathrm{d}x) \Big/ \sum_{s=1}^{k} n_s \hat{c}_s^{-1}\, q_s(x), \tag{3.4}$$

where $\hat{c}_s$ is the maximum likelihood estimate of c_s. Note that $\hat{\mu}$ is supported on the data values $\{x_1, \ldots, x_n\}$, but the atoms at these points are not all equal. From the integral definition of c_r, we have

$$\hat{c}_r = \int_{\Gamma} q_r(x)\, \mathrm{d}\hat{\mu} = \sum_{i=1}^{n} \frac{q_r(x_i)}{\sum_{s=1}^{k} n_s \hat{c}_s^{-1}\, q_s(x_i)}. \tag{3.5}$$

In principle, it is necessary to check that there exists in the parameter space a measure $\hat{\mu}$ such that these equations are satisfied, which is a non-trivial exercise for certain extreme configurations of the observations. Fortunately existence and uniqueness have been studied in great detail for Vardi's biased sampling model (Vardi, 1985), which is essentially mathematically equivalent.

Although $\hat{P}$ is uniquely determined, $\hat{\mu}$ is determined modulo an arbitrary positive multiple, and the constants $\hat{c}_r$ are determined modulo the same positive multiple. In other words, the set of estimable parametric functions may be identified with the space of logarithmic contrasts $\Sigma a_r \log(c_r)$ in which $\Sigma a_r = 0$. This is clear from equation (3.1), which implies that the problem of ratio estimation is invariant under the transformation $q_r(x) \mapsto \alpha(x)\, q_r(x)$, where α is strictly positive on Γ.

The computational algorithm suggested by equation (3.5) is iterative proportional scaling (Deming and Stephan, 1940; Bishop *et al.*, 1975) applied to the $n \times k$ array $\{n_r\, q_r(x_i)\}$ in such a way that, after rescaling,

$$n_r\, q_r(x_i) \mapsto n_r\, q_r(x_i)\, \hat{\mu}(x_i)/\hat{c}_r,$$

each row total is 1 and the rth column total is n_r. The special case $k = 1$, called importance sampling, corresponds to the Horvitz–Thompson estimator (Horvitz and Thompson, 1952), which is widely used in survey sampling to correct for unequal sampling probabilities. We might expect to find the more general estimator (3.5) in use for combining data from one or more surveys that use different known sampling probabilities. Possibly this exercise is not of interest in survey sampling. In any event, the estimator does not occur in Firth and Bennett (1998) or in Pfeffermann *et al.* (1998), which are concerned with unequal selection probabilities in surveys.

The normalized version of equation (3.4) has been obtained by Vardi (1985) and by Lindsay (1995) in a nonparametric model for biased sampling. In his discussion of Vardi (1985), Mallows (1985) pointed out the connection with log-linear models and explained why the algorithm converges. These biased sampling models are not posed as Monte Carlo estimation, but the models are equivalent apart from the restriction of the parameter space to probability distributions. For further details, including connectivity and support conditions for existence and uniqueness, see Vardi (1985) or Gill *et al.* (1988). These conditions are assumed henceforth.

The preceding derivation assumes that Γ is countable, so counting measure dominates all others. If Γ is not countable, no dominating measure exists. Nevertheless, the likelihood has a unique maximum given by equation (3.4) provided that the connectivity and support conditions are satisfied. This maximizing measure has finite support, and the maximum likelihood estimate of c is given by equation (3.5).

3.3. Symmetry and group invariant submodels

In practice, we invariably 'know' that the base-line measure is either counting measure or Lebesgue measure. The method described above completely ignores such information. As a result, the conclusions apply equally to discrete sample spaces, finite dimensional vector spaces, metric spaces, product spaces and arbitrary subsets thereof. On the negative side, if the base-line measure does have symmetry properties that are easily exploited, the estimator may be considerably less efficient than it need be.

To see how symmetries might be exploited, let $\mathcal{G}$ be a compact group acting on Γ in such a way that the base-line measure μ is invariant: $\mu(gA) = \mu(A)$ for each $A \subset \Gamma$ and each $g \in \mathcal{G}$. For example, $\mathcal{G}$ might be the orthogonal group, a permutation group or any subgroup. In this reduced model, the parameter space consists only of measures that are invariant under $\mathcal{G}$. The log-likelihood function (3.3) simplifies because $\theta(x) = \theta(gx)$ for each $g \in \mathcal{G}$, and the minimal sufficient statistic is reduced to the symmetrized empirical distribution function $\hat{P}^{\mathcal{G}}$

$$\hat{P}^{\mathcal{G}}(A) = \operatorname*{ave}_{g \in \mathcal{G}} \{\hat{P}(gA)\}$$

for each $A \subset \Gamma$. If $\mathcal{G}$ is finite and acts freely, $\hat{P}^{\mathcal{G}}$ has mass $1/n|\mathcal{G}|$ at each of the transformed

sample points gx_i with $g \in \mathcal{G}$. In a rough sense, the effective sample size is increased by a factor equal to the average orbit size. The maximum likelihood estimate of μ, obtained by equating the minimal sufficient statistic to its expectation, is

$$n\,\hat{P}^{\mathcal{G}}(\mathrm{d}x) = \sum_{s=1}^{k} n_s \hat{c}_s^{-1}\,\bar{q}_s(x)\,\hat{\mu}(\mathrm{d}x),$$

where

$$\bar{q}_s(x) = \operatorname*{ave}_{g\in\mathcal{G}}\{q_s(gx)\}. \tag{3.6}$$

In other words, the estimates from the submodel are still given by equations (3.4) and (3.5), but with q_s replaced by the group average $\bar{q}_s$, and $\hat{P}$ replaced by $\hat{P}^{\mathcal{G}}$. The group-averaged estimator may be interpreted as Rao–Blackwellization given the orbit, so group averaging cannot increase the variance of $\hat{\mu}$ or of the linear functionals $\hat{c}_r$ (Liu (2001), section 2.5.5).

From the estimating equation point of view, the submodel replaces equation (3.2) by

$$c_r = \int_{\Gamma} \bar{q}_r(x)\,\mathrm{d}\mu, \tag{3.7}$$

which is a consequence of the assumption that μ is $\mathcal{G}$ invariant. However, if we proceed with equation (3.7) directly as with equation (3.2), it would appear that we need draws from

$$\bar{P}_r(\mathrm{d}x) = c_r^{-1}\,\bar{q}_r(x)\,\mu(\mathrm{d}x),$$

rather than $P_r(\mathrm{d}x)$. Although we can easily draw $\bar{x}_i$ from $\bar{P}_{y_i}$ by randomly drawing a g from $\mathcal{G}$ and setting $\bar{x}_i = gx_i$, this step is unnecessary because $\bar{P}_{y_i}$ is invariant under $\mathcal{G}$, and thus $\bar{P}_{y_i}(\{\bar{x}_i\}) = \bar{P}_{y_i}(\{x_i\})$. Provided that $\mathcal{G}$ is sufficiently small that the group averaging in equation (3.6) represents a negligible computational cost, the submodel estimator is no more difficult to compute than the original $\hat{c}$. Consequently, the submodel is most useful if $\mathcal{G}$ is a small finite group. If $\mathcal{G}$ is the orthogonal group, $\operatorname{ave}_{g\in\mathcal{G}}\{q_s(gx)\}$ is the average with respect to Haar measure over an infinite set. This is usually a non-trivial exercise in calculus or numerical integration, precisely what we had sought to avoid by simulation.

With a judicious choice of group action, the potential gain in efficiency can be very large. As an extreme example, suppose that the distributions P_r are such that for each r there exists a $g \in \mathcal{G}$ such that $P_r(A) = P_1(gA)$ for every measurable $A \subset \Gamma$. Then $\bar{q}_r(x)/\bar{q}_s(x)$ is a constant independent of x for all r and s, and the ratios of normalizing constants are estimated exactly with zero variance. This effect is evident in the example in Section 2 in which $X \sim P_\sigma$ implies $gX \sim P_{1/\sigma}$. In practice, such a group action may be hard to find, but it is often possible to find a group such that there is substantially more overlap among the symmetrized distributions $\bar{P}_r(A) = \operatorname{ave}_{g\in\mathcal{G}}\{P_r(gA)\}$ than among the original $\{P_r\}$. For location–scale models with parameter (μ, σ), reflection in the circle of radius $\hat{\sigma}$ centred at $(\hat{\mu}, 0)$ is sometimes effective for integrals of likelihood functions or posterior distributions.

Although a symmetrized estimator using a reflection such as $\bar{q}(x) = \{q(x) + q(-x)\}/2$ may remind us of the *antithetic principle* to reduce Monte Carlo error, these two methods are fundamentally different. The antithetic principle exploits symmetry in the sampling distributions, whereas group averaging utilizes the symmetry in the base-line measure. In addition, the effectiveness of using antithetic variates depends on the form of the integrand (e.g. q_2/q_1), as a non-linear function can make antithetic variates worse than the standard estimator (3.5) (see Craiu and Meng (2004)). In contrast, group averaging can do no harm regardless of the form of the integrand.

The importance link function method of MacEachern and Peruggia (2000) has some features in common with group averaging, but the construction and the implementation are different in major ways. Group structure has also been used by Liu and Sabatti (2000), but for a different purpose, to improve the rate of mixing in Gibbs sampling.

Group averaging has been discussed by Evans and Swartz (2000), page 191, as a method of variance reduction for importance sampling. Although the aims are similar, the details are entirely different. Evans and Swartz considered only subgroups of the symmetry group of the importance sampler. That is to say, the group action preserves both Lebesgue measure and the importance sampling distribution. By contrast, our method is geared towards more general bridge sampling designs, and the group action is not on the sampling distributions but on the base-line measures. In our model, no preferential status is accorded to Lebesgue measure or to any particular sampler, so it is not necessary that the group action should preserve either. On the contrary, it is desirable that the group action should mix the distributions thoroughly to make the averaged distributions as similar as possible.

3.4. Projection and linear submodels

Up to this point, the analysis has treated the k functions $q_1, \ldots, q_k$ in a symmetric manner even where the design constants $\{n_1, \ldots, n_k\}$ are not equal. In practice, there may be substantial asymmetries that can be exploited to reduce simulation error. In the simplest case, it may be known that two of the normalizing constants are equal, say $c_2 = c_3$. The reduced parameter space is then the set of measures μ such that $\int (q_2 - q_3)\, \mathrm{d}\mu = 0$. Ideally, we would like to estimate μ by maximum likelihood subject to this homogeneous linear constraint. Even when it exists and is unique, the maximum likelihood estimator in this submodel is unlikely to be cost effective, so we seek a simple one-step alternative by linear projection.

Let $\tilde{c}$ be the unconstrained estimator from equation (3.5) or the group-averaged version in Section 3.3, and let $\tilde{V}$ be the asymptotic variance matrix of $\log(\tilde{c})$ as given in equation (4.2). We consider a submodel in which c lies in the subspace $\mathcal{X} \subset \mathcal{R}^k$. For example, a single homogeneous constraint among the constants gives rise to a subspace $\mathcal{X}$ of dimension $k - 1$, and a matrix X of order $k \times (k - 1)$ whose columns span $\mathcal{X}$.

Ignoring statistical error in the asymptotic variance matrix $\mathrm{cov}(\tilde{c}) = C\tilde{V}C$, where $C = \mathrm{diag}(\tilde{c})$, the weighted least squares projection is

$$\hat{c} = X(X^{\mathrm{T}}C^{-1}\tilde{V}^{-}C^{-1}X)^{-}X^{\mathrm{T}}C^{-1}\tilde{V}^{-}\mathbf{1}, \tag{3.8}$$

where $\mathbf{1}$ is the constant vector with k components. See, for example, Hammersley and Handscomb (1964), section 5.7. Provided that all generalized inverses are reflexive, i.e. $\tilde{V}^{-}\tilde{V}\tilde{V}^{-} = \tilde{V}^{-}$, the asymptotic variance matrix is

$$\mathrm{cov}(\hat{c}) = X(X^{\mathrm{T}}C^{-1}\tilde{V}^{-}C^{-1}X)^{-}X^{\mathrm{T}}.$$

As always, only ratios of cs are estimable in the submodel.

It is perhaps worth mentioning by way of clarification the precise role of the control variates when the objective is to estimate a single ratio c_1/c_2. Suppose that $k = 3$ and that the design constants are $(0, n, 0)$, so all observations are generated from P_2. Then the importance sampling estimator $\tilde{c}_1/\tilde{c}_2$ from equation (3.5) has asymptotic variance $O(n^{-1})$. Suppose now that P_1 is in fact a mixture of P_2 and P_3, so $q_1 = \alpha_2 q_2 + \alpha_3 q_3$, this being the reason for including q_3 as a control variate such that $\int (q_2 - q_3)\, \mathrm{d}\mu = 0$. Then

$$\tilde{c}_1 = \int q_1(x)\,\mathrm{d}\tilde{\mu} = \int (\alpha_2 q_2 + \alpha_3 q_3)\,\mathrm{d}\tilde{\mu}$$
$$= \alpha_2 \tilde{c}_2 + \alpha_3 \tilde{c}_3$$

is a linear combination of $\tilde{c}_2$ and $\tilde{c}_3$, so the covariance matrix of $\tilde{c}$ has rank 2. After projection, $\hat{c}_2 = \hat{c}_3$ and $\hat{c}_1 = (\alpha_2 + \alpha_3)\hat{c}_2$, so $\hat{c}_1/\hat{c}_2 = \alpha_2 + \alpha_3$ is estimated with zero error. The coefficients α_2 and α_3 are immaterial and need not be positive.

It is evident that projection cannot increase simulation variance, but it is not evident that the potential for reduction in simulation error by projection is very great. Indeed, if we are interested primarily in estimating c_1/c_0, the reduction is typically not worthwhile unless the control variates $q_2, q_3, \ldots$ are chosen carefully. The way in which this may be done for Bayesian posterior calculations is discussed in Section 5. Efficiency factors of the order of 5–10 appear to be routinely achievable.

The discussion of control variates by Evans and Swartz (2000) involves subtraction rather than projection, so the technique is different from that proposed above. The algebra in section 5.7 of Hammersley and Hanscomb (1964) and in Rothery (1982), Ripley (1987) and Glynn and Szechtman (2000) is, in most respects, equivalent to the projection method. There are some differences but these are mostly superficial, starting with the complication that only ratios are identifiable in our models and submodels. The main difference in implementation is that our likelihood method automatically provides a variance matrix, so no preliminary experiment is required to estimate the coefficients in the projection. Glynn and Szechtman (2000), section 8, also noted that the required projection is a linear approximation to the nonparametric maximum likelihood estimator.

3.5. Log-linear submodels

Most sets that arise in statistical applications have a large amount of structure that can potentially be exploited in the construction of submodels. For example, the spaces arising in genetic problems related to the coalescent model have a tree structure with measured edges. A submodel may be useful for Monte Carlo purposes if the estimate under the submodel is easy to compute and has substantially reduced variance. The following example illustrates the principle as it applies to spaces having a product structure.

If $\Gamma = \Gamma_1 \times \ldots \times \Gamma_l$ is a product set, it is natural to consider the submodel consisting of product measures only, i.e. $\mu = \mu_1 \times \ldots \times \mu_l$, in which μ_j is a measure on Γ_j. Then each $x \in \Gamma$ has components $(x_1, \ldots, x_l)$, and $\theta(x) = \theta_1(x_1) + \ldots + \theta_l(x_l)$ in an extension of the notation of Section 3.2. The sufficient statistic in equation (3.3) is reduced to the list of l marginal empirical distribution functions, and the resulting model is equivalent to the additive log-linear model (main effects only) for an l-dimensional contingency table, or $l+1$ dimensional if the design has more than one sampler. No closed form estimator is available unless the design is such that, for each $n_r > 0$, the function $q_r(x)$ is expressible as a product $q_r(x) = q_{r1}(x_1) \ldots q_{rl}(x_l)$. Then each sampler generates observations with independent components, so $\hat{\mu}_j$ is given by equation (3.4), applied to the jth component of x. The component measures $\hat{\mu}_1, \ldots, \hat{\mu}_l$ are then independent.

In certain circumstances, the component sets $\Gamma_1, \ldots, \Gamma_l$ are isomorphic, in which case we write Γ^l instead of Γ. It is then natural to restrict the parameter space to symmetric product measures of the form μ^l. Suppose, for example, that we wish to compute the integrals

$$c_\theta = \int_{\mathcal{R}^2} \exp\{-(x_1 - \theta)^2/2 - (x_2 - \theta)^2/2 - x_1^2 x_2^2/2\}\,\mathrm{d}x_1\,\mathrm{d}x_2$$

for various values of θ in the range $(0, 4)$. These constitute a subfamily of distributions consid-

ered by Gelman and Meng (1991), whose conditional distributions are Gaussian. The parameter space in the submodel is the set of symmetric product measures $\mu \times \mu$, so μ is a measure on $\mathcal{R}$. For a sampler, it is convenient to take any distribution with density f on $\mathcal{R}$, or the product distribution on $\mathcal{R}^2$. The numerical values reported here are based on the standard Cauchy sampler. The maximum likelihood estimate of μ on $\mathcal{R}$ has mass $1/n\,f(x_i)$ at each data point, so the maximum likelihood estimate of μ^2 on $\mathcal{R}^2$ has mass $\{n^2 f(x_i)\ f(x_j)\}^{-1}$ at each ordered pair (x_i, x_j) in the sample. We find that the submodel estimator based on n simulated scalar observations is roughly equivalent in terms of statistical efficiency to $3n/2$ bivariate observations in the unconstrained model. In principle, therefore, the crude importance sampling estimator can be improved by a factor of 3 without further data. On the negative side, the submodel estimator of c_θ

$$\hat{c}_\theta = \int q_\theta(x, x')\, \mathrm{d}\hat{\mu}(x)\, \mathrm{d}\hat{\mu}(x')$$

is the sum of n^2 terms, as opposed to n terms in the unconstrained model. In terms of computational effort, therefore, the importance sampling estimator is superior. The submodel estimator is not cost effective unless the simulations are the dominant time-consuming part of the calculation.

There is one additional circumstance in which the submodel estimator achieves a gain in statistical efficiency that is sufficient to offset the increase in computational effort. If two functions q_r and q_s are such that the one-dimensional marginal distributions of P_r are the same as the one-dimensional marginal distributions of P_s, the estimated ratio $\hat{c}_r/\hat{c}_s$ has a variance that is $o(n^{-1})$, i.e.

$$n\ \mathrm{var}\{\log(\hat{c}_r/\hat{c}_s)\} \to 0$$

as $n \to \infty$. Simulation results indicate that the rate of decrease is $O(n^{-2})$. This phenomenon can be observed if we replace the preceding family of integrands by the family $\exp(-x_1^2 - x_2^2 + 2\theta x_1 x_2)$ for $-1 < \theta < 1$.

3.6. Markov chain models

The model in this section assumes that a sequence of draws constitutes an irreducible Markov chain having known transition density $q(\cdot; x)$ with respect to the unknown measure μ on Γ. If the design calls for multiple chains, the transition densities are denoted by $q_r(\cdot; x)$ for $r = 1, \ldots, k$. It is not necessary that the chain be in equilibrium; nor is it necessary that the chain be constructed to have a particular stationary distribution. Under this new model, a draw is a chain of length l with distribution $P_r^{(l)}$. The likelihood may be expressed as the product of l factors, the first three of which are

$$\begin{aligned} P_r(\mathrm{d}x_1) &= c_r^{-1}\, q_r(x_1)\, \mu(\mathrm{d}x_1), \\ P_r(\mathrm{d}x_2|x_1) &= c_r^{-1}(x_1)\, q_r(x_2; x_1)\, \mu(\mathrm{d}x_2), \\ P_r(\mathrm{d}x_3|x_2) &= c_r^{-1}(x_2)\, q_r(x_3; x_2)\, \mu(\mathrm{d}x_3). \end{aligned}$$

If the chain is not in equilibrium, the first factor is ignored. In effect, we now have l 'independent' observations from l distinct distributions, each with its own normalizing constant. The log-likelihood function for θ contributed by a single sequence of length l from $P_r^{(l)}$ is then given by

$$\sum_{t=1}^{l} \theta(x_t) - \log\{c_r(x_{t-1}; \theta)\} = l \int_\Gamma \theta(x)\, \mathrm{d}\hat{P} - \sum_{t=1}^{l} \log\{c_r(x_{t-1}; \theta)\},$$

which is a function of the entire sequence. Although the form of the likelihood is very similar to equation (3.3), the empirical distribution function $\hat{P}$ is no longer sufficient. The log-likelihood equation from k independent chains, in which the chain from P_s has length n_s, is given by

$$n\,\hat{P}\,(\mathrm{d}x) = \sum_{s=1}^{k}\sum_{t=1}^{n_s} \hat{P}_s(\mathrm{d}x|x_{t-1}) = \sum_{s=1}^{k}\sum_{t=1}^{n_s} \hat{c}_s^{-1}(x_{t-1})\, q_s(x; x_{t-1})\, \hat{\mu}(\mathrm{d}x), \tag{3.9}$$

where $n = \Sigma_s n_s, c_s(x_0) \equiv c_s, q_s(x; x_0) \equiv q_s(x)$ and $P_s(\mathrm{d}x|x_0) \equiv P_s(\mathrm{d}x)$. Consequently, for each $r = 1, \ldots, k$ and $t = 0, \ldots, n_r - 1$, we have

$$\hat{c}_r(x_t) = \int q_r(x; x_t)\, \mathrm{d}\hat{\mu}(x) = \sum_{i=1}^{n} \frac{q_r(x_i; x_t)}{\sum_{s=1}^{k}\sum_{j=1}^{n_s} \hat{c}_s^{-1}(x_{j-1})\, q_s(x_i; x_{j-1})}, \tag{3.10}$$

which can be solved for $\{\hat{c}_r(x_t), t = 0, \ldots, n_r - 1; r = 1, \ldots, k\}$ using the Deming–Stephan algorithm. When all draws are independent and the margins are equal, $q_s(x; x_{t-1}) = q_s(x)$ does not depend on x_{t-1}, and equation (3.10) reduces to equation (3.5).

At first sight, we might doubt that equation (3.10) could provide anything useful because we have at most one draw from each of the targeted P_r, namely the first component of each chain—recall that we have purposely ignored the information that all the margins are the same. That is, the transition probabilities $q_r(x_t; x_{t-1})$ can be arbitrary, and in fact they can even be time inhomogeneous (i.e. $q_r(x_t; x_{t-1})$ can be replaced by $q_{r,t}(x_t; x_{t-1})$), as long as the chain is not reducible. Furthermore, it appears that the number of 'parameters' $\{\hat{c}_r(x_t), t = 0, \ldots, n_r - 1; r = 1, \ldots, k\}$ is always the same as the number of data points. However, we must keep in mind that the model parameter is not c, but the base-line measure μ, and as long as a draw is from a *known* density with respect to μ it provides information about μ. This is in fact the fundamental reason that importance sampling can provide consistent estimators when draws are taken from an unrelated trial density. The information from the trial distributions about the base-line measure must be adequate to estimate the base-line measure of the target distribution. In particular, the union of the supports of the trial densities must cover the support of the target density.

4. Asymptotic covariance matrix

4.1. Multinomial information measure

The log-likelihood (3.3) is evidently a sum of k multinomial log-likelihoods, sharing the same parameter θ. The Fisher information for θ is best regarded explicitly as a measure on $\Gamma \times \Gamma$ such that $\mathcal{I}(A, B) = \mathcal{I}(B, A)$ and $\mathcal{I}(A, \Gamma) = 0$ for $A, B \subset \Gamma$. In particular, the multinomial information measure associated with the distribution P_r on Γ is given by $P_r(A \cap B) - P_r(A)\, P_r(B)$. In the log-likelihood (3.3), the total Fisher information measure for θ is

$$n\,\mathcal{I}(A, B) = \sum_{r=1}^{k} n_r \{P_r(A \cap B) - P_r(A)\, P_r(B)\}.$$

At least formally, the asymptotic covariance matrix of $\hat{\theta}$ is the inverse Fisher information matrix $n^{-1}\mathcal{I}^{-}$, and the asymptotic covariance matrix of $\mathrm{d}\hat{\mu}$ is

$$n^{-1}\, \mathrm{d}\mu(x)\, \mathrm{d}\mu(y)\, \mathcal{I}^{-}(x, y),$$

where $\mathcal{I}^-(x, y)$ is the (x, y) element of $\mathcal{I}^-$, indexed by $\Gamma \times \Gamma$. From expression (3.5) for $\hat{c}_r$, we find that the asymptotic covariance of $\hat{c}_r$ and $\hat{c}_s$ is $c_r c_s V_{rs}$, where

$$V_{rs} = \text{cov}\{\log(\hat{c}_r), \log(\hat{c}_s)\} = n^{-1} \int_{\Gamma\times\Gamma} \mathcal{I}^-(x, y)\, \mathrm{d}P_r(x)\, \mathrm{d}P_s(y). \tag{4.1}$$

As always, only logarithmic contrasts have variances. In this expression P_r is defined by equation (3.1) for each r, provided that c_r is finite and non-zero. There may be integrands q_r that take both positive and negative values, in which case P_r is not a probability distribution.

For the log-linear submodel discussed in Section 3.5 in which each sampler has independent components, it is necessary to replace $\mathcal{I}^-(x, y)$ in equation (4.1) by the sum $\mathcal{I}_1^-(x_1, y_1) + \ldots + \mathcal{I}_l^-(x_l, y_l)$, where $\mathcal{I}_r$ is the Fisher information measure for θ_r. Expression (4.1), or its generalization, gives the $O(n^{-1})$ term in the asymptotic variance, but this term may be 0. Examples of this phenomenon are given in Sections 3.5 and 5.2. In such cases, more refined calculations are required to find the asymptotic distribution of $\log(\hat{c})$.

4.2. Matrix version

The results of the preceding section are easily expressed in matrix notation, at least when all calculations are performed at the maximum likelihood estimate. Let $W = \text{diag}(n_1, \ldots, n_k)$, and let $\hat{P}$ be the $n \times k$ matrix whose (i, r) element is

$$\hat{P}_r(x_i) = \frac{q_r(x_i)/\hat{c}_r}{\sum_s n_s \hat{c}_s^{-1}\, q_s(x_i)}.$$

The matrix $\hat{P}W$ arises naturally in applying the Deming–Stephan algorithm to solve the maximum likelihood equation. Note that the column sums of $\hat{P}$ are all 1, whereas the row sums satisfy $\Sigma_r\, n_r\, \hat{P}_r(x_i) = 1$ for each i.

The Fisher information for θ at $\hat{\theta}$ is (the measure whose density is represented by the matrix) $I_n - \hat{P}\, W\, \hat{P}^{\mathrm{T}}$, and the asymptotic covariance matrix of $\log(\hat{c})$ is given by

$$\hat{V} = \hat{P}^{\mathrm{T}}(I_n - \hat{P}\, W\, \hat{P}^{\mathrm{T}})^- \hat{P} \tag{4.2}$$

where I_n is the identity matrix of order n. Typically, the matrix $I_n - \hat{P}\, W\, \hat{P}^{\mathrm{T}}$ has rank $n - 1$ with kernel equal to $\mathbf{1}$, the set of constant vectors. Then $I_n - \hat{P}\, W\, \hat{P}^{\mathrm{T}} + \mathbf{1}\mathbf{1}^{\mathrm{T}}/n$ is invertible with approximately unit eigenvalues, and the inverse matrix is also a generalized inverse of $I_n - \hat{P}\, W\, \hat{P}^{\mathrm{T}}$ suitable for use in equation (4.2). Although the inversion of $n \times n$ matrices can be avoided, all the numerical calculations reported in this paper use this variance formula.

The eigenvalues of $I_n - \hat{P}\, W\, \hat{P}^{\mathrm{T}} + \mathbf{1}\mathbf{1}^{\mathrm{T}}/n$ are in fact all less than or equal to 1, with equality for simple Monte Carlo designs in which all observations are generated from a single sampler. For more general designs having more than one sampler, the approximate variance formula $\hat{V} \simeq \hat{P}^{\mathrm{T}} \hat{P}$ is anticonservative. That is to say, $\hat{P}^{\mathrm{T}}(I_n - \hat{P}W\, \hat{P}^{\mathrm{T}})^- \hat{P} \geqslant \hat{P}^{\mathrm{T}} \hat{P}$ in the sense of Löwner ordering. In practice, if all the samplers have support equal to Γ, the underestimate is frequently negligible. The approximate variance formula is easier to compute and may be adequate for projection purposes described in Section 3.4.

5. Applications to Bayesian computation

5.1. Posterior probability calculation

Consider a regression model in which the component observations $y_1, \ldots, y_m$ are independent

and exponentially distributed with means such that $\log\{E(Y_i)\} = \beta_0 + \beta_1 x_i$, where $x_1, \ldots, x_m$ are known constants. For illustration $m = 10$, $x_i = i$ and the values y_i are

$$2.28, 1.46, 0.90, 0.19, 1.88, 0.72, 2.06, 4.21, 2.90, 7.53.$$

The maximum likelihood estimate is $\hat{\beta} = (-0.0668, 0.1494)$. The asymptotic standard error of $\hat{\beta}_1$ is 0.1101 using the expected Fisher information, and 0.0921 using the observed Fisher information.

Let $\pi(\cdot)$ be a prior distribution on the parameter space. The posterior probability $\mathrm{pr}(\beta_1 > 0|y)$ is the ratio of two integrals. In the denominator the integrand is the product of the likelihood and the prior. The numerator has an additional Heaviside factor taking the value 1 when $\beta_1 > 0$ and 0 otherwise. One way to approximate this ratio is to simulate observations from the posterior distribution and to compute the fraction that have $\beta_1 > 0$. However, this exercise is both unnecessary and inefficient.

Let $q_0(\beta) = L(\beta)\,\pi(\beta)$ be the product of the likelihood and the prior at β, and let $q_1(\beta) = q_0(\beta)\, I(\beta_1 > 0)$ be the integrand for the numerator. For auxiliary functions, we choose $q_2(\beta)$ to be the bivariate normal density at β with mean $\hat{\beta}$ and inverse covariance matrix equal to the Fisher information at $\hat{\beta}$. It is marginally better to use the observed Fisher information for the inverse variance matrix in q_2, but in principle we could use either or both. Let $q_3(\beta)$ be the product $q_2(\beta)\, I(\beta_1 > 0)/K$, where $K = \mathrm{pr}(\beta_1 > 0)$, computed under the normal distribution q_2. In this example,

$$K = \Phi(0.1494/0.0921) = 0.9476$$

for the observed information approximation, or 0.9127 for the expected information approximation. The common normalizing constant $2\pi|\hat{I}|^{1/2}$ for q_2 and q_3 may be ignored. By construction, therefore, $\int q_2(\beta)\,\mathrm{d}\beta = \int q_3(\beta)\,\mathrm{d}\beta$, not necessarily equal to 1.

The numerical calculations that follow use the improper uniform prior and $n = 400$ simulations from the normal proposal density q_2, so the design constants are $n = (0, 0, 400, 0)$. The unconstrained maximum likelihood estimates obtained from equation (3.5) were

$$\log(\tilde{c}_1/\tilde{c}_0) = -0.0525 \pm 0.0137,$$
$$\log(\tilde{c}_3/\tilde{c}_2) = 0.0078 \pm 0.0108$$

with correlation 0.901. Note, however, that $c_2 = c_3$ by design, but the estimator is not similarly constrained at this stage. The estimated posterior probability $\mathrm{pr}(\beta_1 > 0|y)$ is thus $\exp(-0.0525) = 0.9489$, with an approximate 90% confidence interval (0.928, 0.970).

By imposing the constraint $c_2 = c_3$ on the parameter space we obtain a new estimator by weighted least squares projection $\log(\hat{c}) = X(X^{\mathrm{T}}\tilde{V}^{-}X)^{-}X^{\mathrm{T}}\tilde{V}^{-}\log(\tilde{c})$. Here $\tilde{c}$ is the unconstrained estimator, $\tilde{V}$ is the estimated variance matrix of $\log(\tilde{c})$ and X is the model matrix

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

The resulting estimate and its standard error are

$$\log(\hat{c}_1/\hat{c}_0) = \log(\tilde{c}_1/\tilde{c}_0) - 1.136\,\log(\tilde{c}_3/\tilde{c}_2) = -0.0613 \pm 0.0059.$$

The alternative projection (3.8) is preferable in principle, but in this case the two projections yield indistinguishable estimates. The point estimate of the posterior probability is 0.940, with an approximate 90% confidence interval (0.931, 0.950). The efficiency factor is roughly 5, which

is similar to the factors obtained by Rothery (1982) using a similar technique in a problem concerning power calculation.

Much of the gain in efficiency in this example could be achieved by taking the projection coefficient to be -1 on the log-scale, i.e. by subtraction using the control variate in the traditional way. However, the real applications that we have in mind are those in which a moderately large number of integrals is to be computed simultaneously, as in likelihood calculations for pedigree analysis, or computing the entire marginal posterior for β_1. It is then desirable to include several control variates and to compute the projection using equation (3.8). In the present example, the posterior probabilities $\mathrm{pr}(\beta_1 \leqslant b|y)$ for six equally spaced values of b in $(0, 0.25)$ were computed simultaneously by this method using the corresponding six normal control variates. The six efficiency factors were 5.3, 6.2, 8.4, 10.5, 8.4 and 5.9.

This is a rather small example with $m = 10$ data points in the regression, in which an appreciable discrepancy might be expected between the posterior and the normal approximation. Numerical investigations indicate that the efficiency factor tends to increase with larger m, as might be expected. The approximate variance formula mentioned at the end of Section 4.3 is exact in this case, and effectively exact when observations are simulated from both of the normal approximations.

5.2. Posterior integral for probit regression

Consider a probit regression model in which the responses are independent Bernoulli variables with $\mathrm{pr}(y_i = 1) = \Phi(x_i^{\mathrm{T}}\beta)$, where x_i is a vector of covariates and β is the parameter. The aim of this exercise is to calculate the integral of the product $L(\beta; y)\,\pi(\beta)$, in which $L(\beta; y)$ is the likelihood function and $\pi(\cdot)$ is the prior, here taken to be Gaussian. This is done using a Markov chain constructed to have stationary distribution equal to the posterior.

The chain is generated by the standard technique of Gibbs sampling. Following Albert and Chib (1993), the parameter space is augmented to include latent variables $z_i \sim N(x_i^{\mathrm{T}}\beta, 1)$ such that y_i is the sign of z_i. A cycle of the Gibbs sampler is completed in two steps: $\beta|y, z$ and $z|y, \beta$, which are respectively multivariate normal and a product of independent truncated normal distributions. The transition probability from $\theta' = (\beta', z')$ to $\theta = (\beta, z)$ of the generated Markov chain is

$$P(\mathrm{d}\theta|\theta') = c^{-1}(\theta')\; p(z|y, \beta)\; p(\beta|y, z')\; \mu(\mathrm{d}\theta),$$

where $p(z|y, \beta)$ and $p(\beta|y, z')$ are the full conditional densities. By construction, the normalizing constants $c(\theta)$ are known to be equal to 1 for each θ.

Let $\{(\beta_t, z_t)\}_{t=1}^n$ be the simulated values from the Markov chain. The required integral is estimated by the ratio of

$$\int L(\beta)\,\pi(\beta)\,\mathrm{d}\hat{\mu} = \sum_{i=1}^{n} \frac{L(\beta_i)\,\pi(\beta_i)}{\sum_{j=1}^{n} p(\beta_i|y, z_j)} \tag{5.1}$$

to the known value $c(\theta_t) = 1$. In effect, $\hat{c}(\theta_t)$ in equation (3.10) is replaced by the known value, which is then used to compute the approximate maximizing measure $\hat{\mu}$. The resulting estimator, which may be interpreted as importance sampling with the semideterministic mixture

$$n^{-1} \sum_{j=1}^{n} p(\cdot|y, z_j)$$

as the sampler, is consistent but may not be fully efficient. For large n, n times the summand evaluated at any fixed β is a consistent estimate of the integral. Such an estimate was suggested by Chib (1995), who recommended choosing a value β^* of high posterior density, taken to be the mean value in the calculations below. The summand as a function of β also appears in calculations by Ritter and Tanner (1992) and Cui *et al.* (1992), but the purpose there is to monitor convergence of the Gibbs sampler, either with multiple parallel chains or a single long chain divided into batches.

For numerical illustration and comparisons, we use Chib's (1995) example, taken from a case-study by Brown (1980). The data were collected on 53 prostate cancer patients to predict the nodal involvement. There are five predictor variables and a binary response taking the value 1 if the lymph nodes were affected. For the results reported here, three covariates are used: the logarithm of the level of serum acid phosphatase, X-ray reading and stage of the tumour, so the model matrix X is of order 53×4 with a constant column. The prior is centred at $(0.75, 0.75, 0.75, 0.75)$ with variance $A = \mathrm{diag}(5^2, 5^2, 5^2, 5^2)$, as in Chib (1995). The Gibbs sampler was started at $\beta = \tilde{A}X^{\mathrm{T}}y$, where $\tilde{A} = (A^{-1} + X^{\mathrm{T}}X)^{-1}$, and run for a total of $N = n_0 + n$ cycles with the first n_0 discarded. The process was repeated 1000 times for several values of N. The mean and standard deviation of 1000 estimates of the logarithm of the integral are given in Table 1. The central processor unit (CPU) time in seconds was measured separately for Gibbs sampling and for subsequent integral evaluations. For $N = 500 + 5000$, the results given here for Chib's method are in close agreement with those reported by Chib (1995). All programming was done in C.

Over the range studied, the statistical efficiency of the likelihood method over Chib's estimator with $n_0 + n$ draws is not constant, but roughly $n/12$, i.e. increasing with n. For example, the efficiency factor at $N = 500 + 5000$ is estimated by

$$(0.0211/0.00103)^2 = 420.$$

To achieve the same accuracy as the likelihood method using 5000 draws, Chib's method requires 420×5000 Gibbs draws, with CPU time $420 \times 0.49 = 206$ s for the draws alone. The final row in Table 1 gives the precision per CPU second, defined as the reciprocal of the product of the total time and the variance. Over the range studied, the new method achieves a precision per CPU second of roughly 8.5–9.5 times that of Chib's method. For fixed n, the likelihood estimator is computationally more demanding, but the additional effort is worthwhile by a substantial factor.

Table 1. Numerical comparison of two integral estimators (log-scale)

	Results for the following Gibbs sampler cycles and methods:							
	$N = 50 + 500$, 0.06 CPU seconds		*$N = 100 + 1000$, 0.11 CPU seconds*		*$N = 250 + 2500$, 0.25 CPU seconds*		*$N = 500 + 5000$, 0.49 CPU seconds*	
	Chib	*Likelihood*	*Chib*	*Likelihood*	*Chib*	*Likelihood*	*Chib*	*Likelihood*
Mean + 34	−0.5693	−0.5796	−0.5588	−0.5661	−0.5542	−0.5569	−0.5510	−0.5534
Standard deviation	0.0652	0.00937	0.0475	0.00505	0.0299	0.00204	0.0211	0.00103
CPU seconds	<0.01	0.28	<0.01	1.03	<0.01	5.87	0.01	21.94
Precision per CPU second	3921	33500	4029	34396	4474	39263	4492	42024

As it happens, this is one of those problems in which the likelihood estimator of μ converges at the standard rate, but the estimate (5.1) of the integral converges at rate n^{-1}. The bias and standard deviation are both $O(n^{-1})$: in this example they appear to be approximately equal in magnitude. By contrast, Chib's estimator converges at the standard $n^{-1/2}$-rate.

6. Retrospective formulation

It is possible to give a deceptively simple derivation of equation (3.5) by a retrospective argument as follows. Regardless of how the design was in fact selected, we may regard the sample size vector $(n_1, \ldots, n_k)$ as the observed value of a multinomial random vector with index n and parameter vector $(\pi_1, \ldots, \pi_k)$. This assumption is innocuous provided that $(\pi_1, \ldots, \pi_k)$ are treated as free parameters to be estimated from the data. Evidently, $\hat{\pi}_r = n_r/n$ is the maximum likelihood estimate.

Mimicking the argument that is frequently employed in retrospective designs, we argue as follows. Given that the point x has been observed, what is the probability that this point was generated from distribution P_r rather than from one of the other distributions? A simple calculation using Bayes's theorem shows that the required conditional probability vector is

$$p(x) = \left(\frac{q_1(x)\pi_1/c_1}{\sum_s q_s(x)\pi_s/c_s}, \ldots, \frac{q_k(x)\pi_k/c_k}{\sum_s q_s(x)\pi_s/c_s} \right).$$

These conditional probabilities depend only on the ratios π_r/c_r, and not otherwise on the base-line measure μ. Conditioning on x does not eliminate the base-line measure entirely, for c_r is a linear function of μ. The conditional likelihood associated with the single observation (y, x) is thus

$$\frac{(\pi_y/c_y)\, q_y(x)}{\sum_r (\pi_r/c_r)\, q_r(x)},$$

and the log-likelihood is

$$\sum_r n_r \log(\pi_r/c_r) - \sum_{i=1}^{n} \log \left\{ \sum_r (\pi_r/c_r)\, q_r(x_i) \right\}. \tag{6.1}$$

Once again, the observed count vector $(n_1, \ldots, n_k)$ is the complete sufficient statistic, and the association of y-values with x-values is not informative.

Differentiation with respect to the parameter $\log(c_r)$ gives

$$\frac{\partial l}{\partial \{\log(c_r)\}} = c_r \frac{\partial l}{\partial c_r} = -n_r + \sum_{i=1}^{n} \frac{(\pi_r/c_r)\, q_r(x_i)}{\sum_s (\pi_s/c_s)\, q_s(x_i)}.$$

By substituting the known value $\hat{\pi}_r = n_r/n$ and setting the derivative to 0, we obtain

$$\hat{c}_r = \sum_{i=1}^{n} \frac{q_r(x_i)}{\sum_s n_s\, q_s(x_i)/\hat{c}_s}, \tag{6.2}$$

which is identical to equation (3.5). That is to say, the retrospective argument, previously put forward by Geyer (1994), gives exactly the right point estimator of c.

The astute reader will notice that, when we substitute n_r/n for π_r in the retrospective likelihood, the resulting function depends only on those cs for which $n_r > 0$, and this restriction also

applies to the maximum likelihood equation (6.2). The apparent equivalence of equations (6.2) and (3.5) is thus an illusion. By contrast with the model in Section 3, the retrospective argument does not lead to the conclusion that equation (6.2) is the maximum likelihood estimate of an integral for which $q_r(\cdot)$ takes negative values.

Even if we are willing to overlook the remarks in the preceding paragraph and to assume that $n_r > 0$ for each r, the objections are not easily evaded. The difficulty at this point is that the conditional likelihood is a function of the ratios $\phi_r = \log(\pi_r/c_r)$, so the vectors π and c are not separately estimable from the conditional likelihood. It is tempting, therefore, to substitute n_r/n for π_r, treating this as a known prior probability. After all, who can tell how the sample sizes were chosen? However plausible this argument may sound, the resulting 'likelihood' does not give the correct covariance matrix for $\log(\hat{c})$. The components of the negative logarithmic second-derivative matrix are

$$-\frac{\partial^2 l}{\partial\{\log(c_r)\}\partial\{\log(c_s)\}} = \sum_{i=1}^{n} \delta_{rs} \frac{(\pi_r/c_r)\, q_r(x_i)}{\sum_t n_t\, q_t(x_i)/c_t} - \sum_{i=1}^{n} \frac{(\pi_r/c_r)(\pi_s/c_s)\, q_r(x_i)\, q_s(x_i)}{\{\sum_t n_t\, q_t(x_i)/c_t\}^2}. \tag{6.3}$$

At $(\hat{\pi}, \hat{c})$, the first term is equal to the diagonal matrix $n_r\delta_{rs}$. The second term is non-negative definite. To put this in an alternative matrix form, write p_i for the conditional probability vector $p(x_i)$. Then the negative second-derivative matrix shown above is

$$\mathcal{I}_\phi = \sum_{i=1}^{n} (\mathrm{diag}\{p_i\} - p_i p_i^{\mathrm{T}}) \simeq W - W\hat{P}^{\mathrm{T}}\hat{P}W,$$

using the matrix notation of Section 4.2.

To see that the inverse of this matrix cannot be the correct asymptotic variance of $\log(\hat{c})$, consider the limiting case in which $q_1 = q_2 = \ldots = q_k$ are all equal. It is then known with certainty that $c_1 = \ldots = c_k$, even in the absence of data. But, in the second-derivative matrix shown above, all the conditional probability vectors p_i are equal to $(\pi_1, \ldots, \pi_k)$. The second-derivative matrix is in fact the multinomial covariance matrix with index n and probability vector π. The generalized inverse of this matrix does give the correct asymptotic variances and covariances for contrasts of $\hat{\phi} = \log(\hat{\pi}/\hat{c})$, as the general theory requires. But it does not give the correct asymptotic variances for $\log(\hat{c})$, or contrasts thereof.

This line of argument can be partly rescued, but to do so it is necessary to show that $\hat{\pi}$ and $\hat{c}$ are asymptotically independent. This is not obvious and will not be proved here, but it is a consequence of orthogonality of parameters in exponential family models. By standard properties of likelihoods, $\mathrm{cov}\{\log(\hat{\pi}) - \log(\hat{c})\} = \mathcal{I}_\phi^-$ asymptotically. On the presumption that $\hat{\pi}$ and $\hat{c}$ are asymptotically uncorrelated, we deduce that

$$\mathrm{cov}\{\log(\hat{c})\} = \mathcal{I}_\phi^- - \mathrm{cov}\{\log(\hat{\pi})\} = \mathcal{I}_\phi^- - \mathrm{diag}(1/n\pi) + \mathbf{1}\mathbf{1}^{\mathrm{T}}/n. \tag{6.4}$$

The term $\mathbf{1}\mathbf{1}^{\mathrm{T}}/n$ does not contribute to the variance of contrasts of $\log(\hat{c})$ and can therefore be ignored. When evaluated at $(\hat{c}, \hat{\pi})$, the resulting expression $(W - W\hat{P}^{\mathrm{T}}\hat{P}W)^- - W^{-1}$, involving no $n \times n$ matrices, is identical to equation (4.2), provided that each component n_r of W is strictly positive.

7. Conclusions

The key contribution of this paper is the formulation of Monte Carlo integration as a statistical model, making explicit what information is available for use and what information is 'out of bounds'. Given that agreement is reached on the information available, it is now possible

to say whether an estimator is or is not efficient. Likelihood methods are thus made available not only for parameter estimation but also for the estimation of variances and covariances for various simulation designs, of which importance sampling is the simplest special case. More interestingly, however, three classes of submodel are identified that have substantial potential for variance reduction. The associated operations are group averaging for group invariant submodels, linear projection for linear submodels or mixtures and Markov chain models for Markov chain Monte Carlo schemes. To achieve worthwhile gains in efficiency, it is necessary to exploit the structure of the problem, so it is not easy to give universally applicable advice. None-the-less, three simple examples show that efficiency factors in the range 5–10, and possibly larger, are routinely achievable in certain types of statistical computations. We believe that such factors are not exceptional, particularly for Bayesian posterior calculations.

It would be remiss of us to overlook the peculiar dilemma for Bayesian computation that inevitably accompanies the methods described here. Our formulation of all Monte Carlo activities is given in terms of parametric statistical models and submodels. These are fully fledged statistical models in the sense of the definition given by McCullagh (2002), no more or no less artificial than any other statistical model. Given that formulation, it might seem natural to analyse the model by using modern Bayesian methods, beginning with a prior on Θ. If we adopt the orthodox interpretation of a prior distribution as the one that summarizes the extent of what is known about the parameter, we are led to the Dirac prior on the true measure, which is almost invariably Lebesgue measure. For once, the prior is not in dispute. This choice leads to the logically correct, but totally unsatisfactory, conclusion that the simulated data are uninformative. The posterior distribution on Θ is equal to the prior, which is unhelpful for computational purposes. It seems, therefore, that further progress calls for a certain degree of pretence or pragmatism by selecting a non-informative, or at least non-degenerate, prior on Θ. Given such a prior distribution, the posterior distribution on Θ can be obtained, and the posterior moments of the required integrals computed by standard formulae. Although these operations are straightforward in principle, the computations are rather forbidding, so much so that it would be impossible to complete the calculations without resorting to Monte Carlo methods! This computational black hole, an infinite regress of progressively more complicated models, is an unappealing prospect, to say the least. With this in mind, it is hard to avoid the conclusion that the old-fashioned maximum likelihood estimate has much to recommend it.

Acknowledgements

The comments of four referees on an earlier draft led to substantial improvements in presentation. We would like to thank Peter Bickel for pointing out the connection with biased sampling models, and Peter Donnelly for discussions on various points.

Support for this research was provided in part by National Science Foundation grants DMS-0071726 (for McCullagh and Tan) and DMS-0072510 (for Meng and Nicolae).

References

Albert, J. and Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *J. Am. Statist. Ass.*, **88**, 669–679.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975) *Discrete Multivariate Analyses: Theory and Practice*. Cambridge: Massachusetts Institute of Technology Press.

Brown, B. W. (1980) Prediction analyses for binary data. In *Biostatistics Casebook* (eds R. J. Miller, B. Efron, B. W. Brown and L. E. Moses). New York: Wiley.

Chib, S. (1995) Marginal likelihood from the Gibbs output. *J. Am. Statist. Ass.*, **90**, 1313–1321.

Craiu, R. V. and Meng, X.-L. (2004) Multi-process parallel antithetic coupling for backward and forward Markov chain Monte Carlo. *Ann. Statist.*, to be published.
Cui, L., Tanner, M. A., Sinha, D. and Hall, W. J. (1992) Monitoring convergence of the Gibbs sampler: further experience with the Gibbs stopper. *Statist. Sci.*, **7**, 483–486.
Deming, W. E. and Stephan, F. F. (1940) On a least-squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Statist.*, **11**, 427–444.
DiCiccio, T. J., Kass, R. E., Raftery, A. and Wasserman, L. (1997) Computing Bayes factors by combining simulation and asymptotic approximations. *J. Am. Statist. Ass.*, **92**, 903–915.
Evans, M. and Swartz, T. (2000) *Approximating Integrals via Monte Carlo and Deterministic Methods*. Oxford: Oxford University Press.
Firth, D. and Bennett, K. E. (1998) Robust models in probability sampling. *J. R. Statist. Soc.* B, **60**, 3–21; discussion, 41–56.
Gelman, A. and Meng, X.-L. (1991) A note on bivariate distributions that are conditionally normal. *Am. Statistn*, **45**, 125–126.
Gelman, A. and Meng, X. L. (1998) Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statist. Sci.*, **13**, 163–185.
Geyer, C. J. (1994) Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo. *Technical Report 568*. School of Statistics, University of Minnesota, Minneapolis.
Gill, R., Vardi, Y. and Wellner, J. (1988) Large sample theory of empirical distributions in biased sampling models. *Ann. Statist.*, **16**, 1069–1112.
Glynn, P. W. and Szechtman, R. (2000) Some new perspectives on the method of control variates. In *Monte Carlo and Quasi-Monte Carlo Methods* (eds K.-T. Fang, F. J. Hickernell and H. Niederreiter), pp. 27–49. New York: Springer.
Hammersley, J. M. and Hanscomb, D. C. (1964) *Monte Carlo Methods*. London: Chapman and Hall.
Hesterberg, T. (1995) Weighted average importance sampling and defensive mixture distributions. *Technometrics*, **37**, 185–194.
Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *J. Am. Statist. Ass.*, **47**, 663–683.
Lindsay, B. (1995) *Mixture Models: Theory, Geometry and Applications*. Hayward: Institute of Mathematical Statistics.
Liu, J. S. (2001) *Monte Carlo Strategies in Scientific Computing*. New York: Springer.
Liu, J. S. and Sabatti, C. (2000) Generalized Gibbs sampler and multigrid Monte Carlo for Bayesian computation. *Biometrika*, **87**, 353–369.
MacEachern, S. N. and Peruggia, M. (2000) Importance link function estimation for Markov Chain Monte Carlo methods. *J. Comput. Graph. Statist.*, **9**, 99–121.
Mallows, C. L. (1985) Discussion of 'Empirical distributions in selection bias models' by Y. Vardi. *Ann. Statist.*, **13**, 204–205.
McCullagh, P. (2002) What is a statistical model (with discussion)? *Ann. Statist.*, **30**, 1225–1310.
Meng, X.-L. and Wong, W. H. (1996) Simulating ratios of normalizing constants via a simple identity: a theoretical explanation. *Statist. Sin.*, **6**, 831–860.
Owen, A. and Zhou, Y. (2000) Safe and effective importance sampling. *J. Am. Statist. Ass.*, **95**, 135–143.
Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H. and Rasbash, J. (1998) Weighting for unequal selection probabilities in multilevel models (with discussion). *J. R. Statist. Soc.* B, **60**, 23–40; discussion, 41–56.
Ripley, B. D. (1987) *Stochastic Simulation*. New York: Wiley.
Ritter, C. and Tanner, M. A. (1992) Facilitating the Gibbs sampler: the Gibbs stopper and the griddy-Gibbs sampler. *J. Am. Statist. Ass.*, **97**, 861–868.
Rothery, P. (1982) The use of control variates in the Monte Carlo estimation of power. *Appl. Statist.*, **31**, 125–129.
Vardi, Y. (1985) Empirical distributions in selection bias models. *Ann. Statist.*, **13**, 178–203.

Discussion on the paper by Kong, McCullagh, Meng, Nicolae and Tan

Michael Evans (*University of Toronto*)

This is an interesting and stimulating paper containing some useful clarifications. The most provocative part of the paper is the claim that treating the approximation of an integral

$$I = \int f(x)\, \mu(\mathrm{d}x) \tag{1}$$

as a problem of statistical inference gives practically meaningful results. Although the paper does effectively argue this, as my discussion indicates, I still retain some doubt about the necessity of adopting this point of view.

In the paper we have a sequence of integrals

$$c_i = \int q_i(x)\,\mu(\mathrm{d}x),$$

for $i = 1, \ldots, r$. We want to approximate the ratios c_i/c_j based on samples of size n_i from the normalized densities specified by the q_i ($n_i = 0$ whenever q_i takes negative values and we assume at least one $n_i > 0$).

If only one n_i is non-zero, say n_1, then the estimators are given by the importance sampling estimates

$$\widehat{\left(\frac{c_i}{c_j}\right)} = \sum_{k=1}^{n} \frac{q_i(x_k)}{w(x_k)} \Big/ \sum_{k=1}^{n} \frac{q_j(x_k)}{w(x_k)} \tag{2}$$

where $w = q_1/c_1$. Note that c_1 need not be known to implement equation (2). As in the general problem of importance sampling, q_1 could be a bad sampler and require unrealistically large sample sizes to make these estimates accurate. In general, it is difficult to find samplers that can be guaranteed to be good in problems.

If we have several $n_i > 0$, then the problem is to determine how we should combine these samples to estimate the ratios. Perhaps an obvious choice is the importance sampling estimate (2) where w is now the mixture

$$w(x) = \sum_{i=1}^{r} \frac{n_i}{n} \frac{q_i(x)}{c_i}. \tag{3}$$

Of course there is no guarantee that this will be a good importance sampler but, more significantly, we do not know the c_i and so cannot implement this directly. Still, if we put equation (3) into equation (2), we obtain a system of equations in the unknown ratios and, as the paper points out, this system can have a unique solution for the ratios. These solutions are the estimates that are discussed in the paper.

The paper justifies these estimates as maximum likelihood estimates and more importantly uses likelihood theory to obtain standard errors for the estimates. The above intuitive argument for the estimates does not seem to lead immediately to error estimates. One might suspect, however, that an argument based on the delta theorem should generate error estimates. Since I shall not provide such an argument, we must acknowledge the accomplishment of the paper in doing this. There are still some doubts, however, about the necessity for the statistical formulation and the likelihood arguments.

The group averaging that is discussed in the paper seems closely related to the use of this technique in Evans and Swartz (2000). There it is shown that, when G is a finite group of volume preserving symmetries of the importance sampler w, then we can replace the basic importance sampling estimate $f(x)/w(x)$, with $x \sim w$, by $f^G(x)/w(x)$ where

$$f^G(x) = \frac{1}{|G|} \sum_{g \in G} f(gx).$$

This is unbiased for integral (1) and satisfies

$$\mathrm{var}_w\left(\frac{f^G}{w}\right) = \mathrm{var}_w\left(\frac{f}{w}\right) - E_w\left\{\left(\frac{f - f^G}{w}\right)^2\right\}. \tag{4}$$

This shows that the group-averaged estimator always has variance smaller than f/w. This is because f^G and w are more alike than f and w as f^G and w now share the group of symmetries G. Some standard variance reduction methods, such as antithetic variates, can be seen to be examples of this approach.

Although equation (4) seems to indicate that we always obtain an improvement with group averaging, a fairer comparison takes into account the additional function evaluations in f^G/w. In that case, it was shown in Evans and Swartz (2000) that f^G/w is a true variance reducer if and only if

$$\frac{|G| - 1}{|G|}\mathrm{var}_w\left(\frac{f}{w}\right) < E_w\left\{\left(\frac{f - f^G}{w}\right)^2\right\} \leqslant \mathrm{var}_w\left(\frac{f}{w}\right).$$

Noting that f^G/w is the orthogonal projection of f/w onto the $L^2(w)$ space of functions invariant under G, we see that we have true variance reduction if and only if the residual $(f - f^G)/w$ accounts for a proportion that is equal to $(|G| - 1)/|G|$ of the total variation in f/w. This implies that the larger the group is the more stringent is the requirement for true variance reduction. Also, if this technique is to be effective, then f and w must be very different, at least with respect to the symmetries in G, i.e. w must be

a poor importance sampler for the original problem. This leads to some reservations about the general utility of the method.

The group averaging technique in the paper can be seen as a generalization of this as G is not required to be a group of symmetries of w. Rather we symmetrize both the importance sampler and the integrand so that the basic estimator is f^G/w^G. It may be that w^G is a more effective importance sampler than w but still we can see that the above analysis will restrict the usefulness of the method.

Although I have expressed some reservations about some aspects of the paper, overall I found it to be thought provoking as it introduces a different way to approach the analysis of integration problems and this leads to some interesting results. I am happy to propose the thanks and congratulations to the authors.

Christian P. Robert (*Centre de Recherche en Economie et Statistique and Université Dauphine, Paris*)

Past contributions of the authors to the Monte Carlo literature, including the notion of *bridge sampling*, are noteworthy, and I therefore regret that this paper does not have a similar dimension for our field. Indeed, the 'theory of Monte Carlo integration' that is advertised in the title reduces to a formal justification of bridge sampling, via nonparametric maximum likelihood estimation, and the device of pretending to estimate the dominating measure allows in addition for an asymptotic approximation of the Monte Carlo error through the corresponding Fisher information matrix. Although this additional level of interpretation of importance and bridge sampling Monte Carlo approximations (rather than estimations) of integrals is welcome, and quite exciting as a formal exercise, it seems impossible to derive a *working principle* out of the paper.

A remark of interest related to the supposedly unknown measure is that groups can be seen as acting on the measure rather than on the distribution. This brings much more freedom (but also the embarrassment of wealth) in looking for group actions, since

$$\int_\Gamma q_s(x)\,\mathrm{d}\lambda(x) = \int_{\Gamma/\mathcal{G}} \int_{\mathcal{G}} q_s(gx)\,\mathrm{d}v(g)\,\mathrm{d}\lambda(x) = \int_{\Gamma/\mathcal{G}} |\mathcal{G}|\,\bar{q}_s(x)\,\mathrm{d}\lambda(x)$$

is satisfied for all groups $\mathcal{G}$ such that $\mathrm{d}\lambda(gx) = \mathrm{d}\lambda(x)$. However, this representation also exposes the confusion that is central to the paper between variance reduction, which is obvious by a Rao–Blackwell argument, and Monte Carlo improvement, which is unclear since the computing cost is not taken into account. For instance, Fig. 1 analyses the example of Section 2 based on a *single realization* from P_1, reparameterized to $[0,1]^2$, where larger groups acting on $[0,1]^2$ bring better approximations of $c_\sigma/c_1 = 1/\sigma^2$. However, this improvement does not directly pertain to Monte Carlo integration, but rather to numerical integration (with the curse of dimensionality lurking in the background). Although the paper shows examples where using the group structure clearly brings substantial agreement, it does not shed much light on the comparison of different groups from a Monte Carlo point of view.

The *numerical* nature of the improvement is also clear when we realize that, since the group action is on the measure rather than on the distribution, it is often unlikely to be related to the geometry of this distribution and thus can bring little improvement when $q(gx_i) = 0$ for most gs and generated x_is. Fig. 2 shows the case of a Monte Carlo experiment on a bivariate normal distribution with the same groups as above: the concentration of the logit transform of the $\mathcal{N}\{(-5,10), \sigma^2 I_2\}$ distribution on a small area of the $[0,1]^2$ square (Fig. 2(a)) prevents good evaluations even after 4^{10} computations (Fig. 2(b)).

A second opening related to the likelihood representation of the weight evaluation is that a Fisher information matrix can be associated with this problem and thus variance-like figures are produced in the paper. Although these matrices stand as a formal ground for comparison of Monte Carlo strategies, I have difficulties with these figures given that they do not necessarily provide a good approximation to the Monte Carlo errors. See for instance the example of Section 2: we could apply Slutsky's lemma with the transform $c_\sigma = \exp\{\log(c_\sigma)\}$ to obtain the variance of the $\hat{c}_\sigma$s as

$$\mathrm{diag}(c_\sigma)\mathbf{V}\,\mathrm{diag}(c_\sigma),$$

but this approximation is invalidated by the absence of variance of the $\hat{c}_\sigma$s. (See also the normal range confidence lower bound on Fig. 2(b) which is completely unrelated to the range of the estimates.) Given that importance sampling estimators are quite prone to suffer from infinite variance (Robert and Casella (1999), section 3.3.2), the appeal of using Fisher information matrices is somewhat spurious as it gives a false confidence in estimators that should not be used.

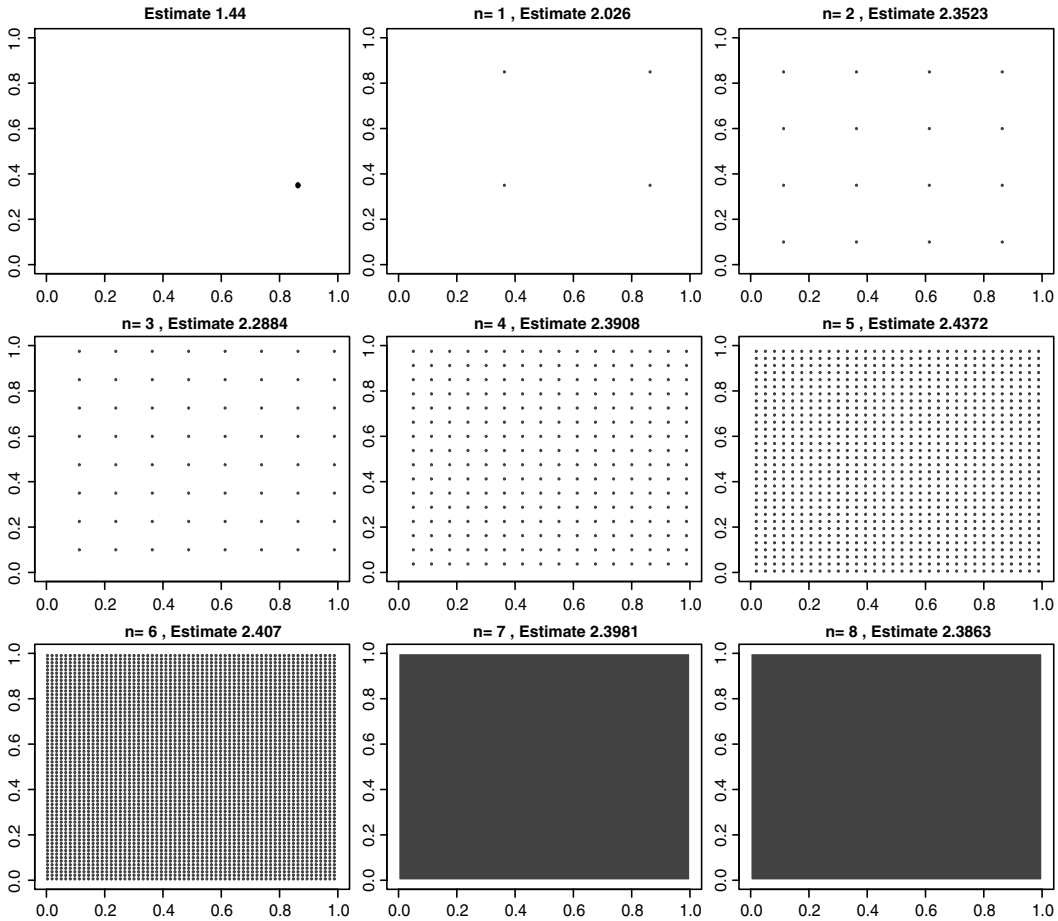


Fig. 1. Successive actions of the groups $\mathcal{G}_n$ on the simulated value $(x_1, x_2) \sim P_1$ (top left-hand corner), where $\mathcal{G}_n$ operates on $(z_1, z_2) = (\exp(x_1)/\{1 + \exp(x_1)\},\ x_2/(1 + x_2)) \in [0, 1]^2$ by changing some of the first n bits in the binary representation of the decimal part of z_1 and z_2: the estimates of $1/\sigma^2 = 2.38$ are given above each graph; the size of the orbit of $\mathcal{G}_n$ is 4^n

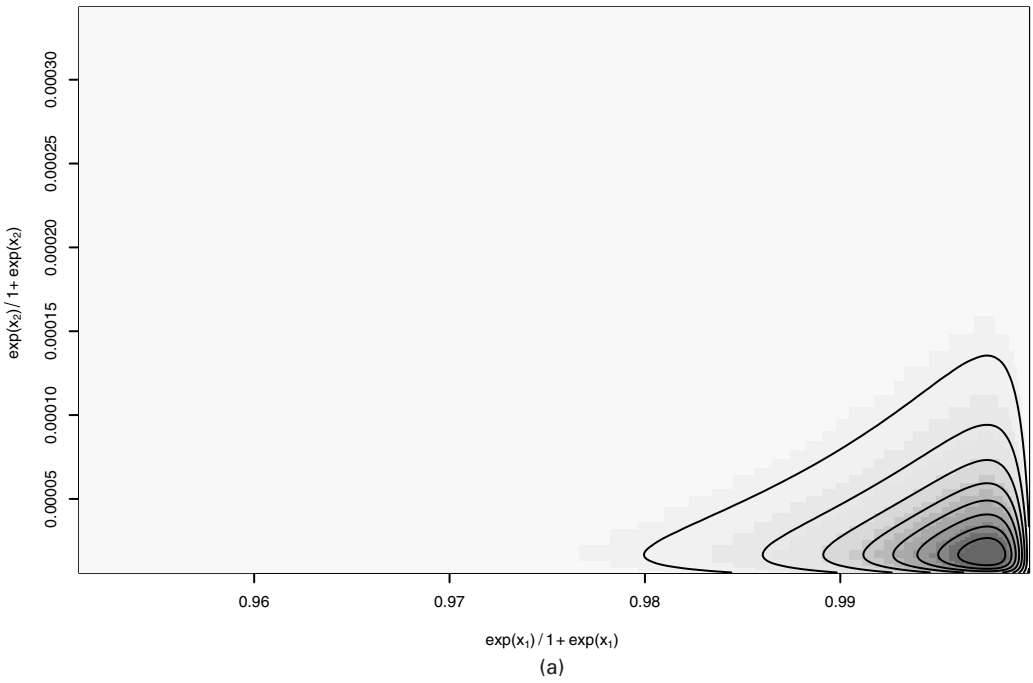
The extension to Markov chain Monte Carlo settings is another interesting point in the paper, in that estimator (3.10) reduces to

$$\hat{c}_r = \sum_{t=1}^{n} q_r(x_t) \Big/ \sum_{j=1}^{n} q_1(x_t|x_j)$$

if we use a single chain based on a transition q_1 (with known constant). Although this estimator only seems to apply to the Gibbs sampler, given that the general Metropolis–Hastings algorithm has no density against the Lebesgue measure, it is a special case of Rao–Blackwellization (Gelfand and Smith, 1990; Casella and Robert, 1996) applied to importance sampling. As noted in Robert (1995), the naïve importance sampling alternative

$$\tilde{c}_r = \frac{1}{n} \sum_{t=1}^{n} q_r(x_t)/q_1(x_t|x_{t-1}),$$

where the 'true' distribution of x_t is used instead, is a poor choice, since it most often suffers from infinite variance. (The notation in Section 3.6 is mildly confusing in that $c_r(x)$ is unrelated to c_r and is also most often known, in contrast with c_r. It thus seems inefficient to estimate the $c_r(x_t)$s.)



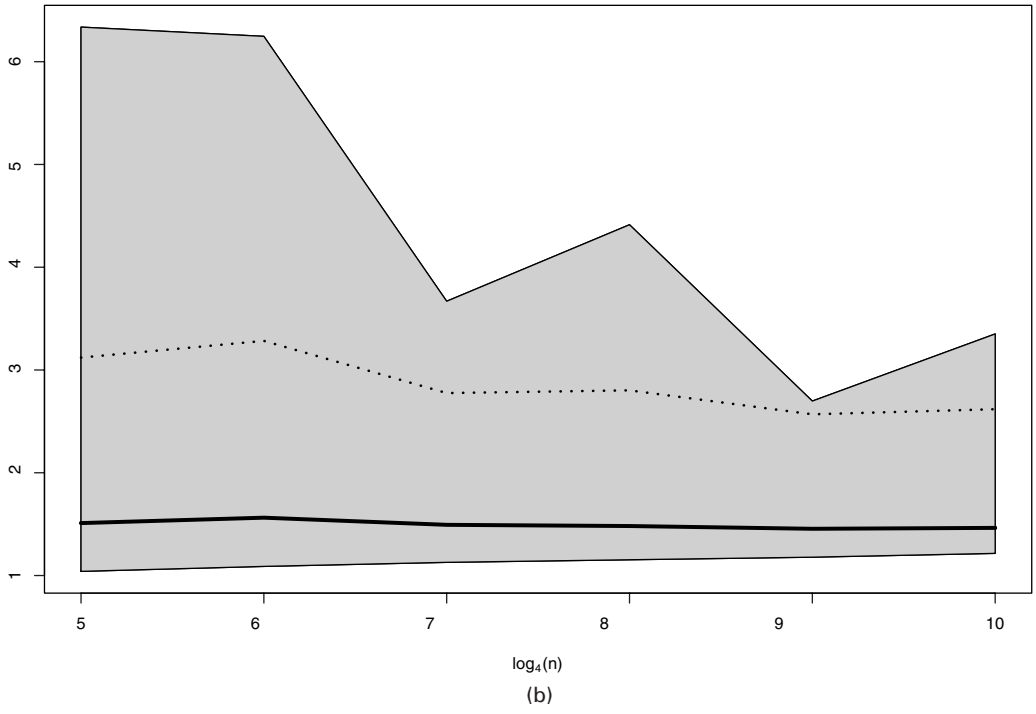


Fig. 2. (a) Contour plot of the density of the logit transform of the $\mathcal{N}\{(-5, 10), \sigma^2 I_2\}$ distribution and (b) summary of the successive evaluations of $c_\sigma / c_1 = \sigma^2 = 1.5$ by averages over the groups $\mathcal{G}_n$ for the logit transform of a single simulated value $(x_1, x_2) \sim \mathcal{N}\{(-5, 10), \sigma^2 I_2\}$: ———, average over 100 replications; ■, range of the evaluations; · · · · · · · , 2-standard-deviations upper bound

Except for the misplaced statements of the last paragraph about 'orthodox Bayesians', which simply bring to light the fundamental difference between designed Monte Carlo experiments and statistical inference, I enjoyed working on the authors' paper and thus unreservedly second the vote of thanks!

The vote of thanks was passed by acclamation.

A. C. Davison (*Swiss Federal Institute of Technology, Lausanne*)
The choice of the group is a key aspect of the variance reduction by group averaging that is suggested in the paper, and I would like to ask its authors whether they have any advice for the bootstrapper.

The simplest form of nonparametric bootstrap entails equal probability sampling with replacement from data $y_1, \ldots, y_n$, which is used to approximate quantities such as

$$m = n^{-1} \sum t(y_1^*, \ldots, y_n^*),$$

where the sum is over the n^n possible resamples $y_1^*, \ldots, y_n^*$ from the original data. If the statistic under consideration is symmetric in the y_j then the number of summands may be reduced to $\binom{2n-1}{n}$ or fewer, but it is usually so large that m is approximated by taking a random subset of the resamples. The most natural group in this context arises from permuting $y_1, \ldots, y_n$ and has size $n!$, which is far too large for use in most situations. One could readily construct smaller groups, obtained for example by splitting the data into blocks of size k and permuting within the blocks, but they seem somewhat arbitrary. How should the group be chosen in this context, where the cost of function evaluations vastly outweighs the cost of sampling?

Wenxin Jiang and Martin A. Tanner (*Northwestern University, Evanston*)
One way to rederive equation (2.1) is to regard it as importance sampling using a sample x_1^n with a 'base density' of the mixture

$$f(x) = \sum_s \frac{n_s}{n} \frac{q_s(x)}{c_s}.$$

The estimator of $c_\sigma = \int_\Gamma q_\sigma(x)\, \mathrm{d}\mu$ would be $n^{-1} \Sigma_{i=1}^n q_\sigma(x_i)/f(x_i)$. Replacing the unknown constant c_s in the base density by $\hat{c}_s$ then results in a self-consistency equation that is identical to equation (2.1). (Similarly, with group averaging the functions $q_\sigma(x)$ are essentially replaced by $\bar{q}(x; \sigma)$ and the integrals $c_\sigma = \int_\Gamma \bar{q}(x; \sigma)\, \mathrm{d}\rho$ are estimated by using an importance sample x_1^n having a base density $\Sigma_s\, (n_s/n)\, \bar{q}(x; s)/c_s$.)

The goal here is to achieve a small 'average asymptotic variance'

$$\mathrm{AAVAR} = \int_\Xi \int_\Xi \mathrm{var}\{\log(\hat{c}_\sigma/\hat{c}_\tau)\}\, \mathrm{d}\pi(\sigma)\, \mathrm{d}\pi(\tau).$$

(The paper actually uses (25/20) AAVAR as the criterion function, averaging over 10 different pairs of $\sigma, \tau \in \Xi = \{0.25, 0.5, 1, 2, 4\}$.)

What is the optimal base density f_o and the corresponding optimal $\mathrm{AAVAR}(f_\mathrm{o})$? The following is a result that is obtained by applying variational calculus.

Proposition 1 (average optimality). Let $\hat{c}_\sigma(f) = n^{-1} \Sigma_{i=1}^n q_\sigma(x_i)/f(x_i)$ be an estimator of $c_\sigma = \int_\Gamma q_\sigma(x)\, \mathrm{d}\mu$, where x_1^n is an independent and identically distributed sample from a probability measure $f(x)\, \mathrm{d}\mu$. Let

$$\mathrm{AAVAR}(f) = \int_\Xi \int_\Xi \mathrm{var}\left[\log\left\{\frac{\hat{c}_\sigma(f)}{\hat{c}_\tau(f)}\right\}\right] \mathrm{d}\pi(\sigma)\, \mathrm{d}\pi(\tau)$$

for some prechosen averaging scheme defined by measure π over a set of parameters Ξ. Then

$$\mathrm{AAVAR}(f_\mathrm{o}) = \inf_f \{\mathrm{AAVAR}(f)\} = n^{-1} \left\{\int_\Gamma q_\mathrm{o}(x)\, \mathrm{d}\mu(x)\right\}^2,$$

where

$$f_\mathrm{o}(x) \propto q_\mathrm{o}(x) = \sqrt{\left[\int_\Xi \int_\Xi \left\{\frac{q_\sigma(x)}{c_\sigma} - \frac{q_\tau(x)}{c_\tau}\right\}^2 \mathrm{d}\pi(\sigma)\, \mathrm{d}\pi(\tau)\right]}.$$

For the uniform–average measure π on the space $\Xi = \{0.25, 0.5, 1, 2, 4\}$, using $q_\sigma(x) = \{x_1^2 + (x_2 + \sigma)^2\}^{-2}$ and $\Gamma = \mathcal{R} \times \mathcal{R}^+$ gives (25/20) $\mathrm{AAVAR}(f_\mathrm{o}) = 1.2/n$, with the required integral computed

numerically. The paper uses the base densities $f_1(x) = q_1(x)/c_1$ and $f_2(x) = \Sigma_s\ (1/5)\ q_s(x)/c_s$, achieving (25/20) AAVAR$(f_{1,2}) = 3.6/n$ and $3.0/n$, with efficiency factors AAVAR$(f_{1,2})^{-1}$/AAVAR$(f_o)^{-1}$ being 33% and 40% respectively, compared with the optimal f_o.

It is noted that the optimal AAVAR(f_o) will be smaller if the densities $\{q_\sigma(\cdot)/c_\sigma\}_{\sigma\in\Xi}$ are more similar. A parallel computation can be done with group averaging. Then q_σ/c_σ is replaced by the averaged density $\bar{p}_\sigma = \frac{1}{2}(q_{1/\sigma}/c_{1/\sigma} + q_\sigma/c_\sigma)$ when computing (25/20) AAVAR(f_o), which results in a much smaller value $0.20/n$ since the averaged densities are more similar to each other. With group averaging, the choice of the base density in the paper is essentially $\Sigma_s\ (1/5)\ \bar{p}_s(x) \equiv f_3$, with (25/20) AAVAR$(f_3) = 0.37/n$ and an efficiency factor AAVAR$(f_3)^{-1}$/AAVAR$(f_o)^{-1} = 54\%$.

The paper did not use the exact base densities $f_{1,2,3}$, but rather those using the *estimated* normalizing constant $\hat{c}_s$, which may have added to the variance. Nevertheless, we see that the use of group averaging seems more effective for reducing AAVAR than the use of *any possible* importance samples or design weights.

Hani Doss (*Ohio State University, Columbus*)

An interesting application of the ideas in the paper arises in some current joint work with B. Narasimhan. In very general terms, our set-up is as follows. As in the usual Bayesian paradigm, we have a data vector D, distributed according to some probability distribution P_ϕ, and we have a prior ν_h on ϕ. This prior is indexed by some hyperparameter $h \in \mathcal{H}$, a subset of Euclidean space, and we are interested in $\nu_{h,D}$, the posterior distribution of ϕ given the data, as h ranges over $\mathcal{H}$, for doing sensitivity analyses. In our set-up, we select k values $h_1, \ldots, h_k \in \mathcal{H}$, and we assume that, for each $r = 1, \ldots, k$, we have a sample $\phi_1^{(r)}, \ldots, \phi_{n_r}^{(r)}$ from $\nu_{h_r,D}$. We have in mind situations in which $\phi_1^{(r)}, \ldots, \phi_{n_r}^{(r)}$ are generated by Markov chain Monte Carlo sampling, and in which the amount of time that it takes to generate these is non-negligible. We wish to use the k samples to estimate $\nu_{h,D}$, and we need to do this in realtime. (Markov chain Monte Carlo sampling gives rise to dependent samples, but to keep this discussion focused we ignore this difficulty entirely.)

In standard parametric models, the posterior is proportional to the likelihood times the prior density, i.e.

$$\nu_{h,D}(\mathrm{d}\phi) = c_h^{-1}\ l_D(\phi)\ \nu_h'(\phi)\ \mu(\mathrm{d}\phi),$$

where $c_h = c_h(D)$ is the marginal density of the data when the prior is ν_h, $l_D(\phi)$ is the likelihood function and μ is a common carrier measure for all the priors. In the context of this paper, we have an infinite family of probability distributions $\nu_{h,D}$, $h \in \mathcal{H}$, each known except for a normalizing constant, and we have a sample from a finite number of these. The q_hs may be taken to be $l_D(\phi)\ \nu_h'(\phi)$. The issue that for many models $l_D(\phi)$ is too complicated to be 'known' is irrelevant, since all formulae require only ratios of qs, in which l_D cancels. Equation (3.4) gives the maximum likelihood estimate of μ, and from this we estimate $\nu_{h,D}(\mathrm{d}\phi)$ *directly* via

$$\hat{\nu}_{h,D}(\mathrm{d}\phi) = \frac{1}{\hat{c}_h}\ q_h(\phi)\ \hat{\mu}(\mathrm{d}\phi) = \frac{1}{\hat{c}_h} \sum_{r=1}^{k} \sum_{i=1}^{n_r} \frac{q_h(\phi_i^{(r)})}{\sum_{r=1}^{k} n_r\ q_{h_r}(\phi_i^{(r)})/\hat{c}_{h_r}} \delta_{\phi_i^{(r)}}(\mathrm{d}\phi)$$

(here δ_a is the point mass at a), i.e. $\hat{\nu}_{h,D}$ is the probability measure which gives mass

$$w_h(i,r) = \frac{q_h(\phi_i^{(r)})/\hat{c}_h}{\sum_{s=1}^{k} n_s\ q_{h_s}(\phi_i^{(r)})/\hat{c}_{h_s}} = \left\{ \sum_{s=1}^{k} n_s \frac{\nu_{h_s}'(\phi_i^{(r)})\hat{c}_h}{\nu_h'(\phi_i^{(r)})\hat{c}_{h_s}} \right\}^{-1}$$

to $\phi_i^{(r)}$. A potentially time-consuming computation needs to be used to solve the system of equations (3.5); however, as noted by the authors, only the system involving $c_{h_1}, \ldots, c_{h_k}$ needs to be solved, since, once these have been obtained, for all other hs the calculation of $\hat{c}_h$ is immediate. The weights are simply the sequence

$$\left\{ \sum_{s=1}^{k} n_s \frac{\nu_{h_s}'(\phi_i^{(r)})}{\nu_h'(\phi_i^{(r)})\hat{c}_{h_s}} \right\}^{-1}$$

normalized to sum to 1. These weights can usually be calculated very rapidly. This is exactly the estimator that was used in Doss and Narasimhan (1998, 1999) and obtained by other methods (Geyer, 1994). An expectation $\int f(\phi)\,\hat{\nu}_{h,D}(\mathrm{d}\phi)$ is estimated via

$$\sum_{r=1}^{k}\sum_{i=1}^{n_r} f(\phi_i^r)\,w_h(i,r). \tag{5}$$

It is very useful to note that the development in Section 4 enables estimation of the variance of expression (5). We add the function $f(\phi)\,q_h(\phi)$ to our system. Letting

$$c_h^f = \int f(\phi)\,q_h(\phi)\,\mu(\mathrm{d}\phi),$$

we see that expression (5) is simply $\hat{c}_h^f/\hat{c}_h$. If we take $\tilde{P}$ to be the matrix $\hat{P}$ in Section 4.2 of the paper, augmented with the two columns

$$\left(\frac{q_h(\phi_i^{(r)})/\hat{c}_h}{\sum_{s=1}^{k} n_s\,q_{h_s}(\phi_i^{(r)})/\hat{c}_{h_s}}\right)_{\text{all } i,\text{all } r},$$

$$\left(\frac{f_h(\phi_i^{(r)})\,q_h(\phi_i^{(r)})/\hat{c}_h^f}{\sum_{s=1}^{k} n_s\,q_{h_s}(\phi_i^{(r)})/\hat{c}_{h_s}}\right)_{\text{all } i,\text{all } r}$$

(so $\tilde{P}$ is an $n \times (k+2)$ matrix), then equation (4.2) gives an estimate of the variance of $\hat{c}_h^f/\hat{c}_h$.

J. Qin (*Memorial Sloan–Kettering Cancer Center, New York*) **and K. Fokianos** (*University of Cyprus, Nikosia*)

We congratulate the authors for their excellent contribution which brings together Monte Carlo integration theory and biased sampling models. Our discussion is restricted to the following simple observation which might be utilized in connection with the theory that is developed in the paper for attacking a larger class of problems.

Suppose that $X_1,\ldots,X_{n_1}$ are independent random variables from a density function $p_\theta(x)$ which can be written as

$$p_\theta(x) = q_\theta(x)/c(\theta),$$

where $q_\theta(x)$ is a given non-negative function, θ is an unknown parameter and $c(\theta) = \int q_\theta(x)\,\mathrm{d}x$ is the normalizing constant. Assume that the integration has no closed form and let $p_0(x)$ be a density function from which simulated data $Z_1,\ldots,Z_{n_2}$ are easily drawn. Then it follows that $p_\theta(x)$ can be rewritten as

$$p(x) = \exp(\alpha + [\log\{q_\theta(x)\} - \log\{p_0(x)\}])\;p_0(x) = \exp\{\alpha + \phi(x,\theta)\}\;p_0(x),$$

where $\phi(x,\theta) = \log\{q_\theta(x)\} - \log\{p_0(x)\}$ and $\alpha = -\log\{\int \phi(x,\theta)\;p_0(x)\;\mathrm{d}x\} = -\log\{c(\theta)\}$, assuming that both of the densities are supported on the same set which is independent of θ. Hence the initial model is transformed to a two-sample semiparametric problem where the ratio of the two densitites $p(x)$ and $p_0(x)$ is known up to a parameter vector (α,θ). In addition the sample size n_1 of Xs is fixed as opposed to n_2—the sample size of the simulated data—which can be made arbitrarily large. Let $n = n_1 + n_2$ and denote the pooled sample $(T_1,\ldots,T_n) = (X_1,\ldots,X_{n_1};Z_1,\ldots,Z_{n_2})$. Following the results of Anderson (1979) and Qin (1998), inferences for the vector (α,θ) can be based on the profile log-likelihood

$$l(\alpha,\theta) = \sum_{i=1}^{n_1}\{\alpha + \phi(x_i,\theta)\} - \sum_{i=1}^{n}\log[1 + \rho\,\exp\{\alpha + \phi(t_i,\theta)\}],$$

where $\rho = n_1/n_2$. The parameter vector (α,θ) can be estimated by maximizing the profile log-likelihood $l(\alpha,\theta)$ with respect to α and θ, and the likelihood ratio test regarding θ has an asymptotic χ^2-distribution under mild conditions (Qin, 1999). The idea can be generalized further by considering m-samples along the lines of Fokianos *et al.* (2001).

Steven N. MacEachern, Mario Peruggia and Subharup Guha (*Ohio State University, Columbus*)

The authors construct an interesting framework for Monte Carlo integration. The perspective that they provide will improve simulation-based estimation. The example of Section 5.2 makes a clear case for the benefits of subsampling the output of a Markov chain.

To evaluate the efficiency of an estimation procedure, we must consider the cost of creating the estimator along with its precision. Here, the cost of generating the sample is $O(n)$, which we approximate by $c_1 n$. The additional time that is required to compute Chib's estimator is linear in n with an ignorable constant that is considerably smaller than c_1. The precision of his estimator is $O(n)$, say $p_1 n$, resulting in an asymptotic effective precision p_1/c_1. The new estimator relies on the same sample of size n. The cost to compute the estimator is $O(n^2)$, with leading term $c_2 n^2$. The new estimator has precision $O(n^2)$, with leading term $p_2 n^2$. The asymptotic effective precision of the estimator is $\lim_{n\to\infty}\{p_2 n^2/(c_2 n^2 + c_1 n)\} = p_2/c_2$. The authors report simulations that enable us to compare the two estimators, leading to the conclusion that $p_2 c_1/p_1 c_2$ is roughly 8–10. This gain in asymptotic effective precision suggests using the new estimator.

Consider a 1-in-k systematic subsample of the Markov chain of length n (total sample length nk). The subsampled version of equation (5.1) has a cost of $c_1 nk + c_2 n^2$. The precision of the estimator is $O(n^2)$ with leading term $p_{2k} n^2$. Its asymptotic effective precision is p_{2k}/c_2. This example, with positive dependence between successive parameter vectors, is typical of Gibbs samplers. Positive dependence suggests $p_{2k} > p_2$, and so it is preferable to subsample the output of the Markov chain. This benefit for subsampling should hold whenever the cost of computing the estimator is worse than $O(n)$.

We replicated the authors' simulation (500 + 5000 case). A 1-in-10 subsample reduced the standard deviation of integral (5.1) by 10%. We also drew the vector β as four successive univariate normals. For this sampler with stronger, but still modest, serial dependence, a 1-in-10 subsample reduced the standard deviation by 43% compared with the unsubsampled estimator.

In recent work (Guha and MacEachern, 2002; Guha *et al.*, 2002), we have developed *bench-mark estimation*, a technique that allows us to base our estimator primarily on the subsample while incorporating key information from the full sample. Bench-mark estimators, compatible with importance sampling and group averaging, can be far more accurate than subsampled estimators. As this example shows, subsampled estimators can be far more accurate than full sample estimators.

The following contributions were received in writing after the meeting.

Siddhartha Chib (*Washington University, St Louis*)

I view this paper as a contribution on ways to reduce the variability of estimates derived from Monte Carlo simulations. For obvious reasons, I was drawn to the discussion in Section 5.2 where the authors apply their techniques to the problem of estimating the marginal likelihood of a binary response probit model, given Markov chain Monte Carlo draws from its posterior density sampled by the approach of Albert and Chib (1993). Their estimator of the marginal likelihood is closely related to that from the Chib approach. In particular, focusing on expression (5.1), from Chib (1995)

$$L(\beta_i)\,\pi(\beta_i) \Big/ n^{-1}\sum_{j=1}^{n}\pi(\beta_i|y, z_j)$$

is an estimate of the marginal likelihood for any β_i and the numerical efficiency of this estimate is high, for a given n, when β_i is a high posterior density point.

In contrast, expression (5.1) dictates that we compute the Chib estimate at each sampled β_i and then average those estimates. Not surprisingly, by averaging, and thereby incurring a bigger computational load, we obtain a theoretically more efficient estimate. Interestingly, however, the gains documented in Table 1 are not substantial: with $n = 5000$, the averaging affects the mean only in the third decimal place. After adding -34 to each mean and observing that the numerical standard error of the Chib estimate is 0.02, it is clear that the reduction in the numerical standard error achieved by averaging will have no practical consequences for model choice in this problem.

The authors do not discuss the application of their method to models that are more complex than the binary probit. There are, for example, many situations where the likelihood is difficult to calculate, and the posterior ordinate is not available in as simple a form as in the binary probit. In such cases, the key features of the Chib method—that the likelihood and posterior ordinate must be evaluated just once, that tractable and efficient ways of estimating each can be separately identified and that this is sufficient for producing reliable and accurate estimates—have proved invaluable. Illustrations abound, e.g. to multivariate probit

models, longitudinal count data models, multivariate stochastic volatility models, non-linear stochastic diffusion models and semiparametric models constructed under the Dirichlet process framework.

If the general approach described in this paper also amounts to averaging the Chib estimate in more complex models then it is clear that the computational burden of the suggested averaging will be prohibitive in many cases and, therefore, the theoretical efficiency gain, however small or large, will be practically infeasible to realize. Coupled with the virtual equality of the estimates in a case where averaging is feasible, it is arguable whether the additional computing effort will be practical or justifiable in more complex problems.

Ya'acov Ritov (*Hebrew University of Jerusalem*)

The paper is an interesting bridge between the semiparametric theory of stratified biased samples and Monte Carlo integration. In this comment, I concentrate on the theoretical aspects of the problem, leaving the numerical analysis side to others.

The asymptotics of the biased sample model have been discussed previously, e.g. Gill *et al*. (1988), Pollard (1990) and Bickel and Ritov (1991). Bickel *et al*. (1993) discussed the information bounds and efficiency considerations for a *pot-pourri* of models under the umbrella of biased sampling. These include, in particular, stratified sampling, discussed already in Gill *et al*. (1988), with or without known total stratum probability. Case–control studies are the important application of the known probabilities model.

Contrary to the last statement of Section 3.1, the paper deals essentially only with applications in which the 'unknown' measure μ is actually known, at least up to its total mass. This covers most applications of Markov chain Monte Carlo sampling, as well as integration with respect to the Gaussian or Lebesgue measures. It is possible to think of other problems, but this is not the main emphasis of the text or of the examples. The statistician can therefore considerably improve the efficiency of the estimator by using the *known* values of different functionals such as moments and probabilities of different sets. The algorithm becomes increasingly efficient as the number of functionals becomes larger. The result, however, is an extremely complicated algorithm, which is not necessarily faster. For example, if integration according to a Gaussian measure on the line is considered, we can use the fact that all quantiles are known. In essence, we have stratified sampling. However, we can increase efficiency by using the known mean, variance and higher moments of the distribution.

A similar consideration applies to the group models discussed by the authors. The more symmetry (i.e. group structure) we use, the more efficient the estimator. Practical considerations may prevent us from using all possible symmetries, and the actual level of symmetry used amounts to the trade-off between algorithmic complication and statistical efficiency. The latter is, of course, quite different from algorithmic efficiency.

I miss an example in the paper in which the stratified biased sample model could be used in full. There is no reason, to use, for example, only a single Gaussian measure. For example, a variance slightly smaller than intended improves the efficiency of the evaluation of the integral in the centre, whereas a variance slightly larger than intended improves the evaluation in the tail. Practical guidance for the design problem could be helpful. One practical solution would be the use of an adaptive design when different designs are evaluated by the central processor unit time needed for a fixed reduction in estimation error, and the design actually used is selected accordingly.

James M. Robins (*Harvard School of Public Health, Boston*)

I congratulate the authors on a beautiful and highly original paper. In Robins (2003), I describe a substantively realistic model which demonstrates that successful Monte Carlo estimation of functionals of a high dimensional integral can sometimes require, not only that,

(a) as emphasized by the authors, the simulator hides knowledge from the analyst, but also that
(b) the analyst must violate the likelihood principle and eschew semiparametric, nonparametric or fully parametric maximum likelihood estimation in favour of non-likelihood-based locally efficient semiparametric estimators.

Here, I demonstrate point (b) with a model that is closer to the authors' but is less realistic.

Let L encode base-line variables, A with support $\{0, 1\}$ encode treatment and X encode responses, where L and X have many continuous components. Let $\theta = (l, a)$, $q_\theta(x)$ be a positive utility function and the measure u_θ depend on θ so that $u_{(l,a)}$ is the conditional probability measure for X given $(L, A) = (l, a)$. Then, among subjects with $L = l$, $c(l, a) = c(\theta) = c_\theta = \int q_\theta(x)\, \mathrm{d}u_\theta$ is the expected utility associated with treatment a and $d_{\mathrm{op}}(l) = \arg\max_{a \in \{0,1\}} \{c(l, a)/c(l, 0)\}$ is the optimal treatment strategy. Suppose that the

simulated data are $(\Theta_i, X_i), i = 1, \ldots, n$, with $\Theta_i = (L_i, A_i)$ sampled independently and identically distributed under the random design $f(a, l) = f(a|l) f(l)$ and X_i now biasedly sampled from $c_\Theta^{-1} q_{\Theta_i}(x) u_{\Theta_i}(\mathrm{d}x)$. Suppose that the information available to the analyst is $f^*(a|l)$ and $q_\theta^*(x)$, and that $c(l, a)/c(l, 0) = \exp\{-\gamma^*(l, a, \psi^\dagger)\}$ where $\psi^\dagger \in R^k$ is an unknown parameter that determines $d_{\mathrm{op}}(l), \gamma^*(l, a, \psi)$ is a known function satisfying $\gamma^*(l, 0, \psi) = \gamma^*(l, a, 0) = 0$ and $*$ indicates a quantity that is known to the analyst. Then the likelihood is $\mathcal{L} = \mathcal{L}_1 \mathcal{L}_2$ with

$$\mathcal{L}_1 = \prod_{i=1}^{n} \frac{f(L_i)\, q^*_{\Theta_i}(X_i)\, u_{\Theta_i}(\mathrm{d}X_i)}{\exp\{-\gamma^*(L_i, A_i, \psi)\}\, c(L_i, 0)}, \qquad \mathcal{L}_2 = \prod_{i=1}^{n} f^*(A_i|L_i). \tag{6}$$

Then with $Y = 1/q^*_{(L,A)}(X)$ and $b(L) = 1/c(L, 0)$ the analyst's model is precisely the semiparametric regression model characterized by $f^*(A|L)$ known and $E(Y|L, A) = \exp\{\gamma^*(L, A, \psi^\dagger)\}\, b(L)$ with $\psi^\dagger$ and $b(L)$ unknown. In particular, the analyst does not know whether $b(L)$ is smooth in L.

In this model, as discussed by Ritov and Bickel (1990), Robins and Ritov (1997), Robins and Wasserman (2000) and Robins *et al.* (2000), because of the factorization of the likelihood and lack of smoothness,

(a) any estimator of $\psi^\dagger$, such as a maximum likelihood estimator, that obeys the likelihood principle must ignore the known design probabilities $f^*(a|l)$,
(b) estimators that ignore these probabilities cannot converge at rate n^α over the set of possible pairs $(f^*(a|l), c(l, 0))$ for any $\alpha > 0$ but
(c) semiparametric estimators that depend on the design probabilities can be $n^{1/2}$ consistent. An example of an $n^{1/2}$ consistent estimator is $\hat{\psi}(g, h)$ solving

$$0 = \sum_{i=1}^{n} [Y_i \exp\{-\gamma^*(L_i, A_i, \psi)\} - h(L_i)]\, g(L_i)\{A_i - \mathrm{pr}^*(A = 1|L_i)\}$$

for user-supplied functions $h(l) \in R^1$ and $g(l) \in R^k$. Define $h_{\mathrm{opt}}(L) = b(L)$ and $g_{\mathrm{opt}}(L) = v(1, L) - v(0, L)$, where

$$v(A, L) = W(\psi^\dagger)[R(\psi^\dagger) - E\{W(\psi^\dagger)|L\}^{-1}\, E\{W(\psi^\dagger) R(\psi^\dagger)|L\}],$$

$$R(\psi^\dagger) = b(L) \partial \gamma^*(L, A, \psi^\dagger)/\partial \psi$$

and

$$W(\psi^\dagger) = \exp\{2\, \gamma^*(L, A, \psi^\dagger)\}\, \mathrm{var}(Y|A, L)^{-1}.$$

Let $(\hat{g}_{\mathrm{opt}}, \hat{h}_{\mathrm{opt}})$ be estimates of $(g_{\mathrm{opt}}, h_{\mathrm{opt}})$ obtained under 'working' parametric submodels for $b(L)$ and $\mathrm{var}(Y|A, L)$. Then $\hat{\psi}(\hat{g}_{\mathrm{opt}}, \hat{h}_{\mathrm{opt}})$ is locally efficient, i.e. it attains the semiparametric variance bound if the submodels are correct and remains consistent asymptotically normal under their misspecification (Chamberlain, 1990; Newey, 1990; Robins *et al.*, 2000).

Yehuda Vardi (*Rutgers University, Piscataway*)

The paper is a significant bridge between nonparametric maximum likelihood estimation under biased sampling and Monte Carlo integration. Estimation under biased sampling is a fairly mature methodology, including asymptotic results, and the authors have overlooked considerable published knowledge that is relevant to their work. My comments mainly attempt to narrow this gap.

(a) The problem formulation in Section 3 is equivalent to Vardi (1985) with qs, cs and μ here being ws, Ws and F in Vardi (1985). The important special case of two samples (specifically $q_\theta(x) = x^\theta, \theta = 0, 1$) has been treated even earlier (Vardi, 1982), including the nonparametric maximum likelihood estimator and its asymptotic behaviour.
(b) The algorithm proposed in Vardi (1982, 1985) for solving equation (3.5) of this paper is different from that of Deming and Stephan (1940) and is often more efficient (see Bickel *et al.* (1993), pages 340–343, and Pollard (1990), page 83). This is important in computationally intensive appplications where the Deming–Stephan algorithm is slow to converge.

(c) Regarding the asymptotics, for two samples, equation (4.2) should match theorem 3.1 of Vardi (1982) and it would be constructive to show this. For more than two samples, equation (4.2) should match proposition 2.3 in Gill *et al.* (1988) and similar results in Bickel *et al.* (1993).
(d) More on the asymptotics: 'lifting' the limiting behaviour of the estimated c_rs from standard properties of the exponential family, although a useful shortcut, seems mathematically heuristic. A good discussion of the interplay between the parametric and nonparametric aspects of the model is in chapter 14 of Pollard (1990).
(e) The methodology calls for generating samples from weighted distributions, but this is often a difficult problem with no satisfactory solution, so there is a computational trade-off here. Note problem 6.1 of Gill *et al.* (1988), which also connects biased sampling methodology with design issues and importance sampling.
(f) Gilbert *et al.* (1999) extended the model to include weight functions q_rs, which share a common unknown parameter. This might be useful in future applications of your methodology.
(g) You allow negative q_rs but assume zero observations for such samples ($n_r = 0$ in the discussion following equation (3.1)). This is confusing. A simple example with negative weight function(s) and an explanation of how to estimate the ratio of integrals in this case would help.
(h) Woodroofe (1985) showed that the nonparametric maximum likelihood estimator from truncated data is consistent only under certain conditions on the probabilities that generate the data. Woodroofe's problem has a similar data structure to your Markov chain models of Section 3.6, where each targeted distribution has a sample of size 1 or less. This leads me to think that further conditions might be necessary to achieve consistency.

The **authors** replied later, in writing, as follows.

General remarks

The view presented in the paper is that essentially every Monte Carlo activity may be interpreted as parameter estimation by maximum likelihood in a statistical model. We do not claim that this point of view is necessary; nor do we seek to establish a working principle from it. By analogy, the geometric interpretation of linear models is not necessary for fitting or understanding; nor is there a geometric working principle. The fact that the estimators are obtained by maximum likelihood does not mean that they cannot be justified or derived by heuristic arguments, even by flawed arguments. The crux of the matter is not necessity or working principle, but whether the new interpretation is helpful or useful in the sense of leading to new understanding, better simulation designs or more efficient algorithms. Jiang and Tanner's optimal sampler, although not readily computable and not from the mixture class, is a step in this direction. The remark by MacEachern, Peruggia and Guha is another step showing that, for slow mixing Markov chains, the numerical efficiency of superefficient estimators, such as those described in Section 5.2, may be improved by subsampling.

Several discussants (Evans, Robert, Davison, ...) raise questions about group averaging, its effectiveness, how to choose the group, and so on. As Robert correctly points out, having the group act on the parameter space rather than on the sampler provides great latitude for choice of group actions. Group averaging can be very effective with a small group or very ineffective with a large group, the latter being counter-productive. The group actions that we have in mind involve very small groups, so the additional computational effort is small, if not quite negligible. To choose an effective group action, it is usually necessary to take advantage of the structure of the problem, the symmetries or geometry of the model or the topology of the space. Euclidean spaces have ample structure for interesting group actions. The parameter space in a parametric mixture model has obvious symmetries that can potentially be exploited. In the nonparametric bootstrap, unfortunately, these useful pieces of structure have been stripped away, the space that remains being a forlorn finite set with no redeeming structure. In the examples that we have explored, effective group action carries each sample point to a different point that is not usually in the sample. The bootstrap restriction to the observed sample points precludes group actions other than finite permutations. If the retention of structure is not contrary to bootstrap principles, it may be possible to use temporal order, or factor levels or covariate values, and to permute the covariates or the design. Of course, this may be precisely what you do not want to do.

It is true, as Ritov points out, that maximum likelihood becomes more efficient as the parameter space is reduced by the inclusion of symmetries or additional information such as the values of certain integrals. Our submodels are designed to exploit exactly that. The hard part of the exercise is to construct a submodel such that the gain in precision is sufficient to justify the additional computational effort. Evans's remarks

show that the potential gain from the traditional type of group action is very limited. We doubt that it is possible to put a similar bound on the effectiveness of the group action on the parameter space.

Mixtures versus strata

Remarks by Evans and Jiang and Tanner suggest that it may be helpful to explain why four superficially similar estimators can have very different variances:

(a) the biased sampling estimator or bridge sampling estimator (BSE);
(b) the importance sampling estimator in which a single sample is drawn from the mixture;
(c) the post-stratified importance sampling estimator with actual sample sizes as weights;
(d) the stratified estimator in which the relative normalizing constants for the samplers are given.

We have listed these in increasing order of available information. The principal distinction is that the BSE does not use the normalizing constants for the samplers. The fourth estimator is maximum likelihood subject to certain homogeneous linear constraints of the type described in Section 3.4. For (a)–(c), the estimated mass at the sample point x_i is

$$\mathrm{d}\hat{\mu}(x_i) \propto \frac{n}{\sum_s n_s\, q_s(x_i)/\hat{c}_s},$$

$$\mathrm{d}\hat{\mu}(x_i) \propto \frac{1}{\sum_s \pi_s\, q_s(x_i)/c_s}$$

and

$$\mathrm{d}\hat{\mu}(x_i) \propto \frac{n}{\sum_s n_s\, q_s(x_i)/c_s}.$$

For (d) with the restriction $c_1 = \ldots = c_k$, the estimator obtained by constrained maximization over the set of measures supported on the data is

$$\mathrm{d}\hat{\mu}(x_i) \propto \frac{1}{\sum \lambda_r\, q_r(x_i)},$$

the ratios λ_r/λ_s being determined by the condition $\hat{c}_r = \hat{c}_s$, where $\hat{c}_r = \int q_r\, \mathrm{d}\hat{\mu}$. The constrained maximum likelihood estimator is in a mixture form even though there may be draws from only one of the distributions.

Two designs with $\Gamma = [0, 2]$ show that the variances are not necessarily comparable:

$$q_1(x) = I(0 \leqslant x \leqslant 1); \qquad q_2(x) = I(1 \leqslant x \leqslant 2); \qquad n_1 = 2n_2 \text{ (design I)};$$

$$q_1(x) = 0.5\, I(0 \leqslant x \leqslant 2); \qquad q_2(x) = I(0 \leqslant x \leqslant 1); \qquad n_1 = 2n_2 \text{ (design II)};$$

plus the mixture with weights $(\frac{2}{3}, \frac{1}{3})$. The absence of an overlap between q_1 and q_2 in design I means that μ is not estimable. The BSE of an integral such that the integrand is non-zero on both intervals does not exist. In practical terms, the variance is infinite. However, the restriction of μ to each interval is estimable, and the additional information that $c_1 = c_2$ means that these estimated measures can be combined in such a way that $\hat{\mu}([0, 1]) = \hat{\mu}([1, 2])$. The stratified maximum likelihood estimator (d) puts mass 1 at each sample point in the first interval and 2 at each point in the second. In an importance sampling design with independent and identically distributed observations from the mixture, the estimator (b) has the same form with weights 1 or 2 at each sample point. But the number of observations falling in each interval is binomial, so the measure does not have equal mass on the two intervals. This defect is corrected by a post-stratification adjustment. Estimator (d) has the natural property that, for each function q_3 that is a linear combination of q_1 and q_2, the ratio $\hat{c}_3/\hat{c}_1$ has zero variance even if the draws come entirely from P_2.

The main point of this example is that, although the formulae in all cases look like mixture sampling, the estimators can have very different variances. As Jiang and Tanner note, the use of estimated normalizing constants in the BSE may increase the variance, potentially by a large factor. Also, their optimal sampler is not in the mixture family, so the comparison is not entirely fair.

Variances

Vardi makes the point that, for a given sequence of designs, only certain integrals are consistently estimable. It would be good to have a clean description of this class, how it depends on the design and on

the group action. The asymptotic variance estimates reported in the paper are in agreement with known results such as proposition 2.3 in Gill *et al.* (1988), but they apply only to integrals that are consistently estimable. Our experience has been that the Fisher information calculation usually matches up well with empirical simulation variances. At the same time, we have not looked at extreme cases near the boundary of the consistently estimable class, where the approximation might begin to break down. Robert's example with $n = 1$ sheds little light on the adequacy of the asymptotic variance formulae. In applications of this sort, where the log-ratios $\log(\hat{c}_r/\hat{c}_s)$ do not have moments, the distinction between variance and asymptotic variance is critical. The asymptotic variance is well defined and finite, and this is the relevant number for the assessment of precision in large samples.

Retrospective, empirical and profile likelihood

The simplest way to phrase the remarks by Qin and Fokianos is to consider a regression model with response Y and covariate x in which the components of Y are independent with distribution

$$q_\theta(y)\,\mathrm{d}\mu(y)/c_\theta(\mu)$$

where $\theta_i = x_i\beta$, and μ is an arbitrary measure. The likelihood maximized over μ for fixed β is a function of β alone, here assumed to be a scalar. If the covariate is binary, as in a two-sample problem, the two-parameter function given by Qin and Fokianos is such that $l(\hat{\alpha}_\beta, \beta)$ is the profile log-likelihood for β. An equivalent result can be obtained by a conditional retrospective argument along the lines of Armitage (1971). In particular, if $\log\{q_\theta(y)\}$ is linear in θ, both arguments lead to a likelihood constructed as if the conditional distribution given y satisfies a linear logistic model with independent components.

Comparisons of precision

Table 1 shows that the precision previously achieved in nine times units can now be achieved in one time unit. Professor Chib argues unconvincingly that this is not a substantial factor. Possibly he has too much time on his hands. His subsequent remark about the addition of -34 to each value in Table 1 can only be interpreted tongue in cheek, so the comment about insubstantial factors may be meant likewise.

Likelihood principle

Although it appears to have no direct bearing on any of our results, Robins's example raises interesting fundamental issues connected with the likelihood function, its definition and its interpretation. As we understand it, the likelihood principle does not address matters connected with asymptotics; nor does it promote point estimation as an inferential activity. In particular, the non-existence of a maximum or supremum is not regarded as a contradiction of the likelihood principle; nor is the existence of multiple maxima. By convention, an estimator is said to 'obey the likelihood principle' if it is a function of the sufficient statistic.

In writing this paper, we had given no thought to the applicability of the likelihood principle in infinite dimensional models. On reflection, however, it is essential that likelihood ratios be well defined, and, to achieve this, topological conditions are unavoidable. Such conditions are natural in statistical work because an observation is a point in a topological space, in which the σ-field is generated by the open sets. Time and space limitations preclude more detailed discussion, but an example in which Θ is the set of Lebesgue measurable functions $[0, 1] \to [0, 1]$ illustrates one aspect of the matter. If the model is such that Y_i is Bernoulli with parameter $\theta(x_i)$, the 'likelihood' expression

$$L(\theta) = \prod \theta(x_i)^{y_i}\{1 - \theta(x_i)\}^{1-y_i}$$

is not a function on Θ as a topological space. Two functions θ and θ' differing on a null set represent the same point in the topological space Θ, but they may not have the same 'likelihood'. In other words, $\theta = \theta'$ does not imply $L(\theta) = L(\theta')$, so L is not a function on Θ. None-the-less, there are non-trivial submodels for which the likelihood exists and factors, and the Robins phenomenon persists.

References in the discussion

Albert, J. and Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *J. Am. Statist. Ass.*, **88**, 669–679.

Anderson, J. A. (1979) Multivariate logistic compounds. *Biometrika*, **66**, 17–26.

Armitage, P. (1971) *Statistical Methods in Medical Research*. Oxford: Blackwell.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993) *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins.

Bickel, P. J. and Ritov, Y. (1991) Large sample theory of estimation in biased sampling regression, model I. *Ann. Statist.*, **19**, 797–816.
Casella, G. and Robert, C. P. (1996) Rao-Blackwellisation of sampling schemes. *Biometrika*, **83**, 81–94.
Chamberlain, G. (1988) Efficiency bounds for semiparametric regression. *Technical Report*. Department of Economics, University of Wisconsin, Madison.
Chib, S. (1995) Marginal likelihood from the Gibbs output. *J. Am. Statist. Ass.*, **90**, 1313–1321.
Deming, W. E. and Stephan, F. F. (1940) On a least-squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Statist.*, **11**, 427–444.
Doss, H. and Narasimhan, B. (1998) Dynamic display of changing posterior in Bayesian survival analysis. In *Practical Nonparametric and Semiparametric Bayesian Statistics* (eds D. Dey, P. Müller and D. Sinha), pp. 63–87. New York: Springer.
Doss, H. and Narasimhan, B. (1999) Dynamic display of changing posterior in Bayesian survival analysis: the software. *J. Statist. Softwr.*, **4**, 1–72.
Evans, M. and Swartz, T. (2000) *Approximating Integrals via Monte Carlo and Deterministic Methods*. Oxford: Oxford University Press.
Fokianos, K., Kedem, B., Qin, J. and Short, D. A. (2001) A semiparametric approach to the one-way layout. *Technometrics*, **43**, 56–65.
Gelfand, A. E. and Smith, A. F. M. (1990) Sampling based approaches to calculating marginal densities. *J. Am. Statist. Ass.*, **85**, 398–409.
Geyer, C. J. (1994) Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo. *Technical Report 568*. School of Statistics, University of Minnesota, Minneapolis.
Gilbert, P. B., Lele, S. R. and Vardi, Y. (1999) Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials. *Biometrika*, **86**, 27–43.
Gill, R., Vardi, Y. and Wellner, J. (1988) Large sample theory of empirical distributions in biased sampling models. *Ann. Statist.*, **16**, 1069–1112.
Guha, S. and MacEachern, S. N. (2002) Benchmark estimation and importance sampling with Markov chain Monte Carlo samplers. *Technical Report*. Ohio State University, Columbus.
Guha, S., MacEachern, S. N. and Peruggia, M. (2002) Benchmark estimation for Markov chain Monte Carlo samples. *J. Comput. Graph. Statist.*, to be published.
Newey, W. (1990) Semiparametric efficiency bounds. *J. Appl. Econometr.*, **5**, 99–135.
Pollard, D. (1990) *Empirical Processes: Theory and Applications*. Hayward: Institute of Mathematical Statistics.
Qin, J. (1998) Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, **85**, 619–630.
Qin, J. (1999) Case-control studies and Monte Carlo methods. Unpublished.
Ritov, Y. and Bickel, P. (1990) Achieving information bounds in non- and semi-parametric models. *Ann. Statist.*, **18**, 925–938.
Robert, C. P. (1995) Convergence control techniques for MCMC algorithms. *Statist. Sci.*, **10**, 231–253.
Robert, C. P. and Casella, G. (1999) *Monte Carlo Statistical Methods*. New York: Springer.
Robins, J. M. (2003) Estimation of optimal treatment strategies. In *Proc. 2nd Seattle Symp. Biostatistics*, to be published.
Robins, J. M. and Ritov, Y. (1997) A curse of dimensionality appropriate (CODA) asymptotic theory for semiparametric models. *Statist. Med.*, **16**, 285–319.
Robins, J. M., Rotnizky, A. and van der Laan, M. (2000) Comment on 'On profile likelihood' (by S. A. Murphy and A. W. van der Vaart). *J. Am. Statist. Ass.*, **95**, 431–435.
Robins, J. M. and Wasserman, L. (2000) Conditioning, likelihood, and coherence: a review of some foundational concepts. *J. Am. Statist. Ass.*, **95**, 1340–1346.
Vardi, Y. (1982) Nonparametric estimation in the presence of length bias. *Ann. Statist.*, **10**, 616–620.
Vardi, Y. (1985) Empirical distributions in selection bias models. *Ann. Statist.*, **13**, 178–203.
Woodroofe, M. (1985) Estimating a distribution function with truncated data. *Ann. Statist.*, **13**, 163.