Estimating Normalizing Constants and Reweighting Mixtures in Markov Chain Monte Carlo

By
Charles J. Geyer[1].
Technical Report No. 568
School of Statistics
University of Minnesota
December 6, 1991
Revised July 31, 1993
and December 20, 1993
and May 30, 1994
and July 9, 1994

[1]Research supported in part by grant DMS-9007833 from the National Science Foundation

Abstract

Markov chain Monte Carlo (the Metropolis-Hastings algorithm and the Gibbs sampler) is a general multivariate simulation method that permits sampling from any stochastic process whose density is known up to a constant of proportionality. It has recently received much attention as a method of carrying out Bayesian, likelihood, and frequentist inference in analytically intractable problems. Although many applications of Markov chain Monte Carlo do not need estimation of normalizing constants, three do: calculation of Bayes factors, calculation of likelihoods in the presence of missing data, and importance sampling from mixtures. Here reverse logistic regression is proposed as a solution to the problem of estimating normalizing constants, and convergence and asymptotic normality of the estimates are proved under very weak regularity conditions.

Markov chain Monte Carlo is most useful when combined with importance reweighting so that a Monte Carlo sample from one distribution can be used for inference about many distributions. In Bayesian inference, reweighting permits the calculation of posteriors corresponding to a range of priors using a Monte Carlo sample from just one posterior. In likelihood inference, reweighting permits the calculation of the whole likelihood function using a Monte Carlo sample from just one distribution in the model. Given this estimate of the likelihood, a parametric bootstrap calculation of the sampling distribution of the maximum likelihood estimate can be done using just one more Monte Carlo sample.

Although reweighting can save much calculation, it does not work well unless the distribution being reweighted places appreciable mass in all regions of interest. Hence it is often not advisable to sample from a distribution in the model. Reweighting a mixture of distributions in the model performs much better, but this cannot be done unless the mixture density is known and this requires knowledge of the normalizing constants, or at least good estimates such as those provided by reverse logistic regression.

1 INTRODUCTION

Markov chain Monte Carlo (MCMC) methods, including the Metropolis-Hastings algorithm (Metropolis, et al. 1953; Hastings 1970) and the Gibbs sampler (Geman and Geman 1984; Gelfand and Smith 1990), are general multivariate simulation tools applicable to a wide range of statistical inference problems. For recent reviews see Geyer (1992), Smith and Roberts (1993), Besag and Green (1993), and Tierney (in press).

An MCMC algorithm simulates an ergodic Markov chain X_1, X_2, ... having a specified stationary distribution. Expectations with respect to the stationary distribution are approximated by sample averages

$$E_n g(X) = \frac{1}{n} \sum_{i=1}^{n} g(X_i).$$

The Metropolis-Hastings algorithm, in particular, can be used to efficiently simulate any distribution whose density is known up to a constant of proportionality. Given a function h that is nonnegative, integrable, and not zero almost everywhere, the Metropolis-Hastings algorithm simulates a Markov chain having a stationary distribution whose density is proportional to h and does so without knowledge of the proportionality constant.

Although the Metropolis-Hastings algorithm enables simulation from a density with unknown normalizing constant, there are situations in which one needs to know normalizing constants. In Bayesian inference, the ratio of normalizing constants for two different posteriors is the Bayes factor for the two hypotheses. In Monte Carlo likelihood inference with missing data (Thompson and Guo 1991), the ratio of normalizing constants is the likelihood ratio for hypotheses. The method proposed here for the estimation of normalizing constants is reverse logistic regression, which is related to logistic discrimination (Anderson 1972, 1982). It is a very stable method of estimating normalizing constants, much better than direct Monte Carlo integration.

Another reason for wanting normalizing constants is an importance sampling scheme that we call "reweighting mixtures." Suppose one has samples from a distribution with unnormalized density h but wants to do integration with respect to another distribution with unnormalized density h_θ. The importance sampling formula says that such integrals can be calculated as weighted averages

$$E_{n,\theta} g(X) = \sum_{i=1}^{n} w_\theta(X_i) g(X_i), \tag{1}$$

where

$$w_\theta(x) = \frac{h_\theta(x)/h(x)}{\sum_{i=1}^{n} h_\theta(X_i)/h(X_i)}, \tag{2}$$

but this does not work well unless h is spread out so that it puts appreciable mass under all interesting h_θ. When we are interested in expectations (1) for a wide range of θ, as occurs in likelihood analysis (Geyer and Thompson 1992), no distribution in the model will serve as a good importance sampling distribution h (Green 1992).

The method of reweighting mixtures was invented (Geyer and Thompson 1992, reply to discussion) to provide good importance sampling distributions in this situation. In Bayesian inference, the same importance sampling scheme permits sensitivity analysis, determining posteriors for a range of priors using only one sample (Besag 1992, Geyer 1991b, Smith 1992).

In principle, one could use an h of any form, but in practice one may know nothing about the distributions except what is learned from sampling and cannot choose a good h without some preliminary sampling from other distributions. Suppose one starts with samples from m different distributions with unnormalized densities $h_j(x)$. Let $f_j(x) = h_j(x)/c_j$ denote the associated probability densities, where c_j is the unknown normalizing constant that makes f_j integrate to one, and let X_{ij}, $i = 1, \ldots, n_j$ denote the samples. If we replace h by h_j, n by n_j, and X_i by X_{ij} in (1) and (2), we get a different estimate of $E_\theta g(X)$ for each j. This is undesirable; what we want is a single estimate that uses all of the samples, not m different estimates each of which uses only the X_{ij} for one j.

The following somewhat counterintuitive scheme does this. We "forget" which distribution a sample point X_{ij} comes from, and consider all the X_{ij} to come from the same distribution, the mixture distribution with density

$$f_{\text{mix}}(x) = \sum_{j=1}^{m} \frac{n_j}{|n|} f_j(x) = \sum_{j=1}^{m} \frac{n_j}{|n|} \frac{1}{c_j} h_j(x) \tag{3}$$

where n now denotes the vector $(n_1, \ldots, n_m)$ and $|n| = n_1 + \cdots + n_m$. In particular, we approximate an expectation $E_\theta g(X)$ by

$$\frac{1}{|n|} \sum_{j=1}^{m} \sum_{i=1}^{n_j} g(X_{ij}) \frac{f_\theta(X_{ij})}{f_{\text{mix}}(X_{ij})}. \tag{4}$$

In MCMC, (4) is not useful as it stands because the c_j are unknown. They can, however, be estimated up to a constant of proportionality by reverse logistic regression, giving estimates $\hat{c}_j$ which are approximately an unknown constant times the c_j. Plugging the $\hat{c}_j$ into (3) gives

$$h_{\text{mix}}(x) = \sum_{j=1}^{m} \frac{n_j}{|n|} \frac{1}{\hat{c}_j} h_j(x) \tag{5}$$

(we change the letter from f to h to remind us that this density is now unnormalized because of the unknown proportionality constant in the $\hat{c}_j$). Then plugging h_{mix} into (1) and (2) gives

$$E_{n,\theta} g(X) = \sum_{j=1}^{m} \sum_{i=1}^{n_j} w_\theta(X_{ij}) g(X_{ij}). \tag{6}$$

where

$$w_\theta(x) = \frac{h_\theta(x)/h_{\text{mix}}(x)}{\sum_{j=1}^{m} \sum_{i=1}^{n_j} h_\theta(X_{ij})/h_{\text{mix}}(X_{ij})}. \tag{7}$$

Equations (5), (6), and (7) define the method of "reweighting mixtures."

We say the method is counterintuitive because it appears to improve estimation by forgetting information, i. e., which distribution an X_{ij} comes from. The procedure seems a little less paradoxical if one considers that a point that occurs in one of the samples could have appeared in any of the other samples (whether or not it actually did). So it makes sense to treat each sample point as if it could have come from any of the distributions, which is what mixture reweighting does.

Reverse logistic regression has been used to estimate likelihood ratios by Thompson, Lin, Olshen, and Wijsman (1993). Reweighting mixtures has been used to do maximum likelihood estimation by Geyer and Møller (in press). At the time this was written, the method does not seem to have been used for calculating Bayes factors or for Bayesian sensitivity analysis, but these problems are no different computationally from likelihood problems. The only requirement for using reverse logistic regression is that all distributions of interest be on the same state space, which can be arranged in Bayesian problems by putting independent proper priors on parameters that otherwise would not be in a model.

2 REVERSE LOGISTIC REGRESSION

The problem for reverse logistic regression is, given samples X_{ij} from the unnormalized densities h_j, to determine the normalizing constants c_j, at least up to a single constant of proportionality. To do this we rewrite (3) as

$$h_{\text{mix}}(x) = \sum_{j=1}^{m} h_j(x) e^{\eta_j} \tag{8}$$

where

$$\eta_j = -\log c_j + \log \frac{n_j}{|n|},$$

and let

$$p_j(x, \eta) = \frac{h_j(x) e^{\eta_j}}{\sum_{k=1}^{m} h_k(x) e^{\eta_k}}. \tag{9}$$

Given that the value x was observed in the mixture sample, $p_j(x, \eta)$ is the probability that it occurred in the jth sample. Of course we actually know which distribution each sample point came from, but our mixture estimation heuristic requires that we forget this information. We propose to estimate the η's, which determine the c's, by maximizing the log quasi-likelihood

$$l_n(\eta) = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \log p_j(X_{ij}, \eta). \tag{10}$$

This actually has an interpretation as a profile log likelihood resulting from maximizing over the unknown mixture probability distribution. The argument is similar to that of Anderson (1972, 1982) for separate sample logistic discrimination, and is given in detail in Geyer (1991b). Here we just take (9) and (10) to define an

objective function to be maximized to yield estimates of the η's. It will be referred henceforth as the "log likelihood" for reverse logistic regression.

The objective function (10) is arithmetically equivalent to the log likelihood for a logistic regression for $m = 2$ and for a "log-linear" or "multinomial response" model for $m > 2$. This is more easily seen by looking at the derivatives

$$\frac{\partial l_n(\eta)}{\partial \eta_r} = n_r - \sum_{j=1}^{m} \sum_{i=1}^{n_j} p_r(X_{ij}, \eta). \tag{11}$$

It is not statistically equivalent to a logistic regression, because the regression is reversed. The "response" n_r is nonrandom and the "predictor" X_{ij} is random. The maximizer of (10) will be referred to as the reverse logistic regression estimator of η even for $m > 2$.

Note that for each r equation (11) sums over all i and j, so the estimates would be unchanged if the X_{ij} were permuted among samples. Thus reverse logistic regression, like mixture reweighting, "forgets" the sample j to which an X_{ij} belongs.

2.1 Identifiability

Note that adding a constant to all of the η_j does not change the value of any p_j, hence maximizing l_n determines an estimate of η only up to an additive constant, and hence determines an estimate of h_{mix} only up to a constant of proportionality.

There are also other aspects to the identifiability question. Suppose there are disjoint subsets K and L of $\{1, \ldots, m\}$ such that such that for each point x in the state space and each $k \in K$ and $l \in L$ either $h_k(x)$ or $h_l(x)$ is zero. When this happens, we say the problem is *separable*, otherwise it is *inseparable*.

A separable problem has a further lack of identifiability. A constant can be added to all of the η_k, $k \in K$ and a different constant added to all of the η_l, $l \in L$ without changing any of the p_j, because at each point x either all of the $h_k(x)$ or all of the $h_l(x)$ are zero. It turns out (Theorem 1 below) that for an inseparable problem η is identifiable up to an additive constant.

A similar definition applies to the finite-sample problem. Suppose that there are disjoint subsets K and L of $\{1, \ldots, m\}$ such that for each point x in the Monte Carlo sample (each X_{ij}) and each $k \in K$ and $l \in L$ either $h_k(x)$ or $h_l(x)$ is zero. When this happens, we say the Monte Carlo sample is *separable*, otherwise it is *inseparable*. The maximum likelihood estimate of η is unique up to an additive constant if and only if the Monte Carlo sample is inseparable (Theorem 1).

These regularity conditions are similar in spirit to those given by Anderson (1972, section 4.2) though they are weaker since our model is simpler.

2.2 Convergence

For MCMC using an irreducible Markov chain, ergodicity implies

$$\frac{1}{n_j} \sum_{j=1}^{n_j} g(X_{ij}) \xrightarrow{\text{a. s.}} E_j g(X), \qquad \text{as } n_j \to \infty. \tag{12}$$

for any P_j-integrable function g. With ergodicity, the reverse logistic regression estimates converge to the true values of the normalizing constants by the following theorem, which is proved in the Appendix.

Theorem 1 *If the Monte Carlo sample is inseparable, the log likelihood l_n has a unique maximizer subject to the constraint that the η_j sum to zero. Suppose that*

$$\liminf_{|n|\to\infty} \frac{n_j}{|n|} > 0, \qquad j = 1, \ldots, m.$$

Let ψ_0 denote the true log normalizing constants normalized to add to zero

$$\psi_{0,j} = -\log c_j + \frac{1}{m}\sum_{k=1}^{m} \log c_k$$

and let $\hat{\psi}_n$ be the estimator of ψ derived from $\hat{\eta}_n$

$$\hat{\psi}_{n,j} = \hat{\eta}_{n,j} - \log \frac{n_j}{|n|} + \frac{1}{m}\sum_{k=1}^{m} \log \frac{n_k}{|n|}$$

Then if the problem is inseparable and the sampler ergodic, $\hat{\psi}_n$ converges to ψ_0 for almost all sample paths of the Monte Carlo.

2.3 Asymptotic Normality

For simplicity in discussing asymptotic normality we assume we are in one of two special cases. Case I: all of the n_j are equal for each n and the vectors $(X_{i1}, \ldots, X_{im})$, $i = 1, 2, \ldots$ form a Markov chain. This would be the case when the distributions are simulated using the Metropolis-coupled scheme in Geyer (1991a). Case II: the sampling fractions converge to a nonzero limit

$$\frac{n_j}{|n|} \to \nu_j > 0, \qquad \text{as } |n| \to \infty, \tag{13}$$

and the samplers for each of the distributions are independent Markov chains. In either case we can define an asymptotic true value of η

$$\eta_{0,j} = \psi_{0,j} + \log \nu_j - \frac{1}{m}\sum_{k=1}^{m} \log \nu_k \tag{14}$$

($\nu_j = 1/m$ in Case I). Then $\sqrt{|n|}(\hat{\eta}_n - \eta_0)$ will be asymptotically normal whenever the MCMC has a central limit theorem (CLT) for the score function ∇l_n, that is whenever

$$\frac{1}{\sqrt{|n|}} \nabla l_n(\eta_0) \xrightarrow{\mathcal{D}} N(0, A) \tag{15}$$

holds. Conditions under which a CLT holds are discussed by Hastings (1970), Kipnis and Varadhan (1986), Schervish and Carlin (1992), Liu, Wong and Kong (in press),

Chan (1993a, b), Tierney (in press), and Chan and Geyer (in press). The variance matrix A usually is the sum of autocovariance matrices. In Case I, A is usually given by

$$A_{rs} = \sum_{j=1}^{m} \sum_{l=1}^{m} \frac{1}{m} \sum_{k=-\infty}^{\infty} \operatorname{Cov}\{p_r(X_{0j}, \eta_0), p_s(X_{kl}, \eta_0)\} \tag{16}$$

and in Case II by

$$A_{rs} = \sum_{j=1}^{m} \nu_j \sum_{k=-\infty}^{\infty} \operatorname{Cov}\{p_r(X_{0j}, \eta_0), p_s(X_{kj}, \eta_0)\} \tag{17}$$

where the covariances are defined for the stationary Markov chain and are the same for lag k and $-k$.

Theorem 2 *If the sampler is irreducible and the problem inseparable and if (15) holds, then*

$$-\frac{1}{|n|} \nabla^2 l_n(\hat{\eta}_n) \xrightarrow{\text{a. s.}} B \tag{18}$$

where

$$B_{rr} = \sum_{j=1}^{m} \nu_j E_j p_r(X, \eta)[1 - p_r(X, \eta)] \tag{19a}$$

$$B_{rs} = -\sum_{j=1}^{m} \nu_j E_j p_r(X, \eta) p_s(X, \eta), \qquad r \neq s \tag{19b}$$

and

$$\sqrt{|n|}(\hat{\eta}_n - \eta_0) \xrightarrow{\mathcal{D}} N\left(0, B^+ A B^+\right) \tag{20}$$

where B^+ is the Moore-Penrose inverse of B, given by

$$B^+ = \left(B + \tfrac{1}{m} u u'\right)^{-1} - \tfrac{1}{m} u u'$$

where $u = (1, 1, \ldots, 1)$.

A proof is given in the Appendix.

2.4 Computation

The gradient (score) and Hessian (observed Fisher information) of the log likelihood are given by (11) and

$$-\frac{\partial^2 l_n(\eta)}{\partial \eta_r^2} = \sum_{j=1}^{m} \sum_{i=1}^{n_j} p_r(X_{ij}, \eta)[1 - p_r(X_{ij}, \eta)] \tag{21a}$$

$$-\frac{\partial^2 l_n(\eta)}{\partial \eta_r \partial \eta_s} = -\sum_{j=1}^{m} \sum_{i=1}^{n_j} p_r(X_{ij}, \eta) p_s(X_{ij}, \eta), \qquad r \neq s \tag{21b}$$

If the Monte Carlo sample is inseparable, the maximum likelihood estimate $\hat{\eta}$ is the unique η making (11) zero and satisfying the constraint that the η_k sum to zero. Since the log likelihood is concave, it can be maximized by many iterative schemes. For example, Newton-Raphson can be used if care is taken to guard against overflow in the calculations and against overly large steps. An alternative, suggested by Alun Thomas (personal communication) is to use successive maximization over each parameter, that is setting $\zeta_r = e^{-\eta_r}$ to

$$\frac{1}{n_r} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \frac{h_r(X_{ij})}{\sum_{k=1}^{m} h_k(X_{ik})/\zeta_k}$$

iteratively. To get convergence it is necessary to impose a constraint, the most convenient one here being to set $\zeta_1 = 1$ and only vary the other $m - 1$ parameters. Unlike Newton-Raphson, this method is globally convergent, see, e. g., Jensen, Johansen and Lauritzen (1991, Theorem 2), but it converges very slowly.

The estimate having been found, the matrix B is estimated by plugging $\hat{\eta}$ into (21a) and (21b). The matrix A defined by (16) and (17) cannot be calculated exactly but can be estimated from the Monte Carlo sample by standard time-series methods (see, for example, Hastings 1970 or Geyer 1992).

3 REWEIGHTING MIXTURES

The method of using a mixture of Monte Carlo samples is described by equations (5), (6), and (7) in the introduction. In this section we show that the asymptotics of reweighted mixture estimators do not depend on the asymptotics of the reverse logistic regression estimates. At least to first order, the estimated normalizing constants may be treated as if they were known. The asymptotic variance of (6) contains no term involving the variance of $\hat{\eta}$, as is shown by the following two theorems. Proofs are given in the Appendix.

Theorem 3 *Under the conditions of Theorem 1 the mixture estimate (6) converges to $E_\theta g(X)$ even when the normalizing constants are unknown and estimated by reverse logistic regression.*

Note that (6) is the ratio of the numerator

$$\frac{1}{|n|} \sum_{j=1}^{m} \sum_{i=1}^{n_j} g(X_{ij}) \frac{h_\theta(X_{ij})}{h_{\text{mix}}(X_{ij})} \tag{22}$$

and a denominator that is the same with $g \equiv 1$. The proof of Theorem 3 shows that the denominator converges to the constant $c(\theta)$, even when η is estimated. Suppose that $E_\theta g(X) = 0$. Then for the stationary chain (22) is the average over a mean zero time series when the normalizing constants are known and converges to zero even when η is estimated.

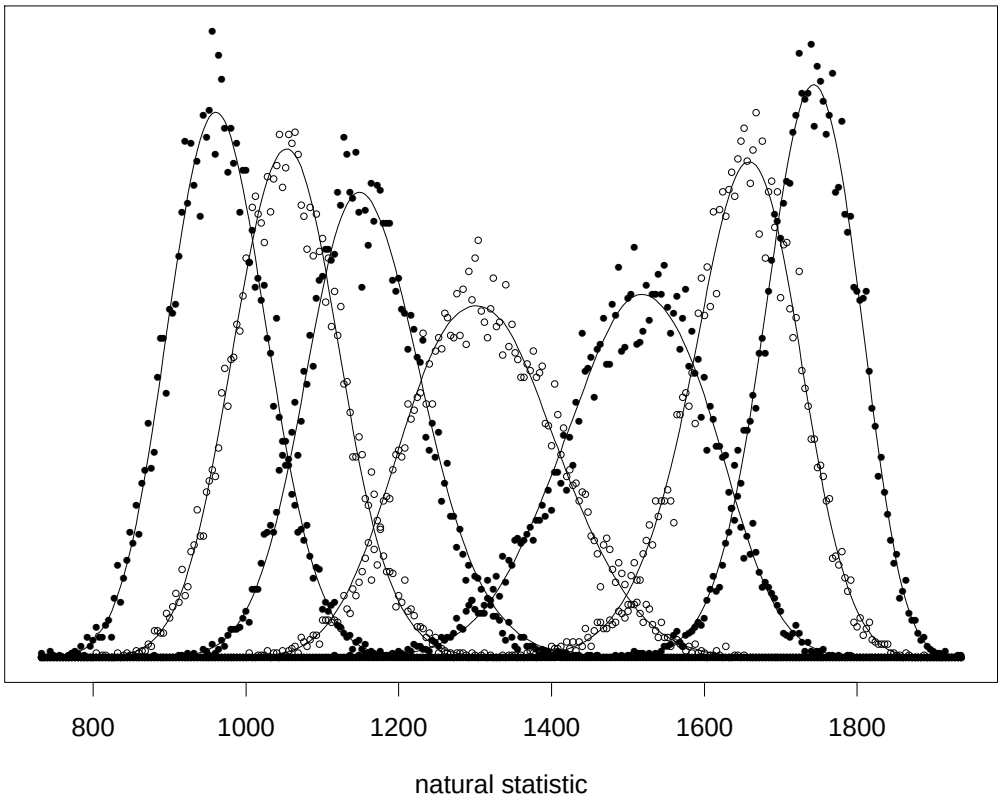

Figure 1: Empirical and smoothed densities for the natural statistic of a symmetric Ising model, for values of the natural parameter 0.365, 0.385, 0.405, 0.425, 0.445, 0.465, 0.485. The sample size is 10000 for each sample. Each sampler was run for 200 iterations before collecting samples. Empirical densities alternate black and white dots.

Theorem 4 *Suppose that*

$$\sqrt{|n|}\Big(E_{n,\theta}g(X) - E_\theta g(X)\Big) \xrightarrow{\mathcal{D}} N(0, \sigma^2) \tag{23}$$

where $E_{n,\theta}g(X)$ is defined by (6) when the normalizing constants are known, then (23) also holds (with the same variance σ^2) when the normalizing constants are estimated by reverse logistic regression.

4 EXAMPLE

As an example of the behavior of the mixture estimation we use a one-parameter Ising model on a 32×32 square lattice with periodic boundary conditions. The state variable is a vector $x = \{x_i\}$ of random variables at the lattice sites taking values ± 1. The model is an exponential family with natural statistic $t(x) = \frac{1}{2}\sum_i \sum_{j \sim i} x_i x_j$

where $i \sim j$ denotes i and j being nearest neighbors The probability of a point x is

$$f_\theta(x) = \frac{1}{c(\theta)} e^{t(x)\theta}$$

and c is the normalizing constant

$$c(\theta) = \sum_{x \in \mathcal{S}} e^{t(x)\theta}, \tag{24}$$

$\mathcal{S}$ being the state space of $2^{32 \times 32}$ possible values of x. Samples from the family were obtained using a Gibbs sampler accelerated by "symmetry swaps" (Geyer 1991a) to obtain rapid mixing for all parameter values.

All the distributions in the family are absolutely continuous with respect to each other, so in principle, any one can be used to estimate any other via importance reweighting. In practice, this may work badly. Figure 1 shows the distributions for seven different values of θ for sample size 10000. The ranges of some of the samples do not even overlap. An attempt to estimate the distribution at one end by reweighting the sample at the other end fails badly. All seven distributions, however, are all well estimated by reweighting the mixture distribution. The smooth curves in Figure 1 are a smooth density estimate of the mixture distribution reweighted to each of the individual distributions. Note how well each of the empirical curves is fitted. It is clear that reweighting the mixture will estimate well any density in this range of parameter values.

In an exponential family the maximum likelihood estimate (MLE) is obtained by finding the parameter value θ that makes the expectation of the natural statistic $E_\theta(t(X))$ equal to its observed value $t(x)$. Let $\tau(\theta) = E_\theta(t(X))$ denote the mapping from the natural parameter to the mean value parameter, so the MLE is the solution of $\tau(\theta) = t(x)$. The Monte Carlo analog (Geyer and Thompson 1992) solves $\tau_n(\theta) = t(x)$ where

$$\tau_n(\theta) = \frac{\sum_{j=1}^n t(X_j) e^{t(X_j)(\theta - \psi)}}{\sum_{j=1}^n e^{t(X_j)(\theta - \psi)}} \tag{25}$$

The estimate $\tau_n(\theta)$ is accurate only for θ near ψ. This is illustrated by Figure 2, which shows (dashed lines) the curve τ_n estimated from each of the samples shown in Figure 1, i. e., the X_j are one of the simulations and ψ is the parameter value for that simulation. Note that each of the curves agrees with the others only in a small range of θ values near the ψ for that curve. Elsewhere the curves do very poorly. This is only to be expected; as θ goes from $-\infty$ to ∞ the value of $\tau_n(\theta)$ traverses the range of the samples. Since the samples do not cover the whole sample space, neither can the range of the τ_n curves.

The solid curve in Figure 2 is estimated by replacing h_ψ by h_{mix} and the sample by the mixture sample in (25) giving

$$\tau_{n,\text{mix}}(\theta) = \frac{\sum_{j=1}^m \sum_{i=1}^{n_j} t(X_{ij}) h_\theta(X_{ij}) / h_{\text{mix}}(X_{ij})}{\sum_{l=1}^m \sum_{k=1}^{n_l} h_\theta(X_{kl}) / h_{\text{mix}}(X_{kl})} \tag{26}$$

As can be seen from Figure 2, it is a much better estimator than any of the individual curves.

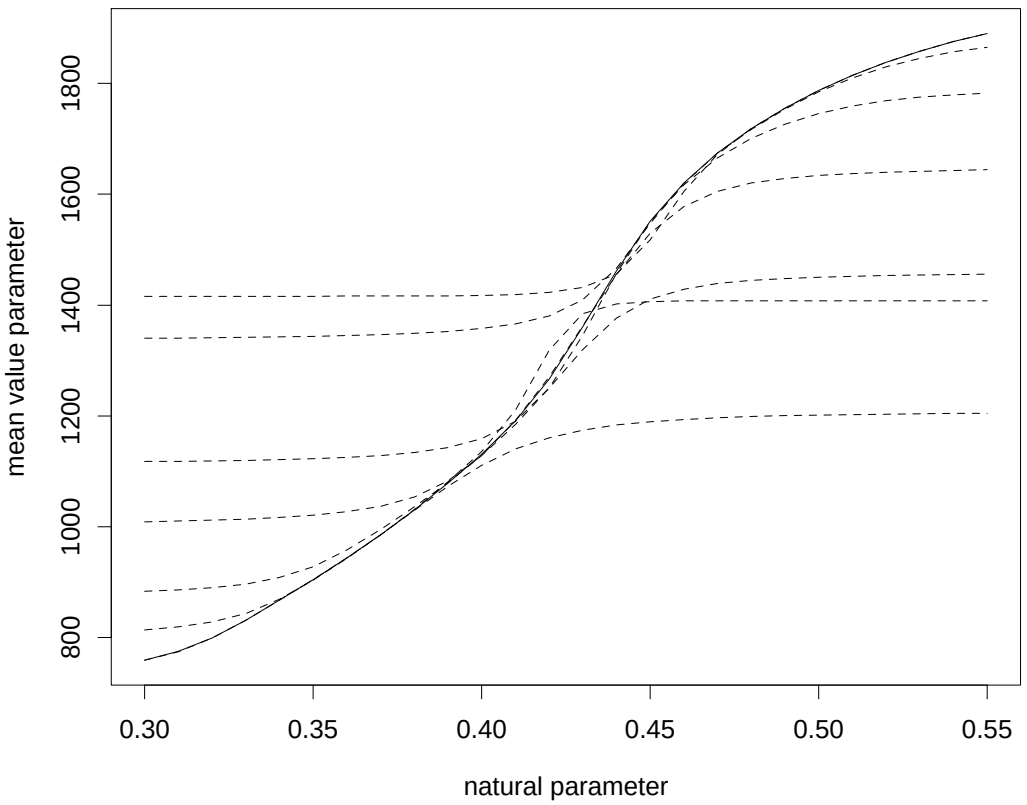

Figure 2: Estimates of the curve mapping the natural parameter to mean value parameter. Dashed lines are the curves (25) estimated from the seven separate samples shown in Figure 1. Solid line is the curve (26) estimated from their mixture.

5 DISCUSSION

Reverse logistic regression is a useful method of obtaining normalizing constants in MCMC. The estimates always converge to the actual normalizing constants and are asymptotically normal when the Markov chain is well behaved (when it is geometrically ergodic, for example). The asymptotic variance can be estimated using the formulas of Sections 2.3 and 2.4. The estimated normalizing constants are useful in three situations: calculation of Bayes factors, calculation of likelihood ratios in missing data problems, and calculating mixture densities for use in importance sampling. The mixture densities are useful in stable Monte Carlo maximum likelihood estimation, including parametric bootstrapping of the distribution of the MLE, and in Bayesian sensitivity analysis.

For maximum likelihood estimation, these methods satisfactorily solve one problem that was left open in Geyer and Thompson (1992). The Monte Carlo log likelihood based on a sample from a single distribution in the model is a good approximation only for parameter values near that distribution. Though it can be used, with some iteration, to obtain the maximum likelihood estimate for the observed data, the method is not accurate enough for calculating likelihood ratios for widely separated parameter points or for parametric bootstrapping. Mixture estimation solves both problems, though for high-dimensional problems even mixture estimation does not seem useful for bootstrapping. There is no problem in calculating likelihood ratios, regardless of dimension, since all that is required is a sequence of distributions along the line between the two points for which the ratio is to be calculated.

The cost of using mixture estimation can be substantially reduced by using the method of Metropolis-coupled chains (Geyer 1991a), which simulates a large number of distributions with about the same accuracy in the amount of time that an uncoupled sampler simulates one, if the uncoupled samplers are slowly mixing. When this is the case, the increased accuracy and stability of mixture estimation comes almost for free.

Acknowledgements

Much of this paper was written during the year 1990-91 at the University of Chicago supported by NSF Postdoctoral Fellowship DMS-9007833. Peter McCullagh pointed out the connection between the reverse logistic regression estimator described here and separate sample logistic discrimination. R. R. Bahadur suggested the approach that makes Theorem 1 work without assuming convergence of the sampling fractions. This version of the paper benefited greatly from discussions with Elizabeth Thompson, Alun Thomas, Jesper Møller, and Hani Doss, who have used the methods and found things that needed improvement or better explanation.

APPENDIX: PROOFS

Let μ be the measure on the state space with respect to which densities are defined, so $E_\theta g(X) = \int g(x) f_\theta(x)\, d\mu(x)$ and so forth.

Proof of Theorem 1. First suppose that the sampling fractions converge, i. e. that (13) holds. Then by ergodicity (12)

$$\frac{1}{|n|} l_n(\eta) \to \lambda(\eta) = \sum_{j=1}^{m} \nu_j E_j \log p_j(X, \eta) \tag{27}$$

almost surely for any fixed η. Hence, a countable union of null sets being a null set, (27) holds simultaneously for all η in a countable dense set (except along a null set of sample paths of the Monte Carlo, which will not be mentioned again). This implies (27) simultaneously for all η in the countable dense set.

An outline of the proof goes as follows. Both l_n and λ are concave functions, finite and twice differentiable everywhere. Direct calculation shows that the gradient of λ is zero at η_0 defined by (14) and that the Hessian λ has no null eigenvectors except $u = (1, 1, \ldots, 1)$ so the the maximizer of λ is unique. From standard theorems of convex analysis (see Rockafellar and Wets in press or Haberman 1989, for details) convergence on a dense set implies uniform convergence on compact sets implies convergence of maximizers (provided that the limit has a unique maximizer). That is, $\hat{\eta}_n$ is unique for all sufficiently large n and converges to η_0. This proves that the convergence of $\hat{\psi}_n$ to ψ_0 even without assuming (13), because for any subsequence there is a further subsubsequence along which (13) holds by compactness hence along which $\hat{\psi}_n$ converges to ψ_0. But if every subsequence has a further convergent subsequence, and all such subsequences converge to the same limit, the whole sequence must converge to that limit.

The function l_n is concave because it is arithmetically equivalent to the log likelihood for an exponential family, and λ is concave because it is the expectation of such a concave function. Clearly $\lambda \le 0$, and

$$\begin{aligned}
-\lambda(\eta) &= \sum_{j=1}^{m} \nu_j E_j \log \frac{1}{p_j(X, \eta)} \\
&= \sum_{j=1}^{m} \nu_j E_j \log \left(1 + \sum_{k \ne j} \frac{h_k(X)}{h_j(X)} e^{\eta_k - \eta_j} \right) \\
&\le \sum_{j=1}^{m} \nu_j E_j \sum_{k \ne j} \frac{h_k(X)}{h_j(X)} e^{\eta_k - \eta_j} \\
&\le \sum_{j=1}^{m} \sum_{k \ne j} \nu_j \frac{c_k}{c_j} e^{\eta_k - \eta_j}
\end{aligned}$$

So λ is everywhere finite.

Since the difference quotients for directional derivatives of a concave function converge monotonely (see Rockafellar 1970, Theorem 23.1) directional derivatives may be commuted with expectations by monotone convergence, and so derivatives

may be taken under the integral sign whenever the integrand is differentiable and the expectation of the derivative is finite. Hence the gradient of λ is given by

$$\frac{\partial \lambda(\eta)}{\partial \eta_r} = \nu_r - \sum_{j=1}^{m} \nu_j E_j p_r(X, \eta) \tag{28}$$

Because the integrand in (28) is uniformly bounded we can again commute expectations and derivatives by dominated convergence, so the negative Hessian of λ is given by (19) and $B = EJ(X)$, where $E = \sum_j \nu_j E_j$ and

$$\begin{aligned} J_{rr}(X) &= p_r(X, \eta)[1 - p_r(X, \eta)] \\ J_{rs}(X) &= -p_r(X, \eta) p_s(X, \eta), \qquad r \neq s \end{aligned}$$

For any vector φ, we have $\varphi' B \varphi = E \varphi' J(X) \varphi \geq 0$ with equality if and only if $J(X)\varphi = 0$ almost everywhere $[P = \sum_j \nu_j P_j]$, that is

$$p_r(X, \eta) \left[\varphi_r - \sum_{s=1}^{m} \varphi_s p_s(X, \eta) \right] = 0, \qquad a.\, e.\, [P]$$

Hence for each r and r', either $h_r(x) h_{r'}(x) = 0$ for almost all x, $[P]$, which contradicts inseparability of the problem, or for some x

$$\varphi_r = \varphi_{r'} = \sum_{s=1}^{m} \varphi_s p_s(x, \eta)$$

Thus the only eigenvector of H is u. A similar proof shows that u is the only null eigenvector of the Hessian of l_n whenever the Monte Carlo sample is inseparable.

It only remains to be shown that η_0 maximizes λ, that is

$$\begin{aligned} \nu_r - \sum_{j=1}^{m} \nu_j E_j p_r(X, \eta_0) &= \nu_r - \sum_{j=1}^{m} \nu_j E_j \frac{\nu_r f_r(X)}{\sum_k \nu_k f_k(X)} \\ &= \nu_r - \sum_{j=1}^{m} \nu_j \int \frac{\nu_r f_r(x)}{\sum_k \nu_k f_k(x)} f_j(x)\, d\mu(x) = 0 \end{aligned}$$

Proof of Theorem 2. Since differentiating $\nabla^2 l_n(\eta_0)$ again gives terms that are products of the $p_r(X_{ij}, \eta_0)$ and hence bounded, the third derivatives are uniformly $O(n)$. Thus the Taylor expansion for $\nabla l_n(\eta_0)$ is

$$\nabla l_n(\eta) = \nabla l_n(\eta_0) + \nabla^2 l_n(\eta_0)(\eta - \eta_0) + |n| O(\|\eta - \eta_0\|^2)$$

Hence defining B_n by

$$-\frac{1}{|n|} \left(\nabla l_n(\hat{\eta}_n) - \nabla l_n(\eta_0) \right) = B_n(\hat{\eta}_n - \eta_0) \tag{29}$$

and using the consistency of $\hat{\eta}_n$ and ergodicity

$$B_n = -\frac{1}{|n|} \nabla^2 l_n(\eta_0) + O(\|\hat{\eta}_n - \eta_0\|) \xrightarrow{\text{a. s.}} B. \tag{30}$$

Similarly

$$\nabla^2 l_n(\eta) = \nabla^2 l_n(\eta_0) + |n| O(\|\eta - \eta_0\|)$$

implies (18).

By algebraic identities $Au = Bu = 0$. Imposing the constraint $u'\eta = 0$ and using $\nabla l_n(\hat{\eta}_n) = 0$ equation (29) becomes

$$\begin{pmatrix} B_n \\ u' \end{pmatrix} \sqrt{|n|}(\hat{\eta}_n - \eta_0) = \begin{pmatrix} \frac{1}{\sqrt{|n|}} \nabla l_n(\eta_0) \\ 0 \end{pmatrix} \tag{31}$$

Hence applying (15), (30), and Lemma 6.4.1 in Lehmann (1983)

$$\sqrt{|n|}(\hat{\eta}_n - \eta_0) \xrightarrow{\mathcal{D}} Y \tag{32}$$

where Y is the solution of the system of equations

$$\begin{pmatrix} B \\ u' \end{pmatrix} Y = \begin{pmatrix} Z \\ 0 \end{pmatrix}$$

and Z is an $N(0, A)$ random vector. It is easily verified that the solution is $Y = B^+ Z$ where B^+ as defined in the statement of the theorem is the Moore-Penrose inverse of B (Rao and Mitra 1971, p. 51 ff.) Hence Y is distributed $N(0, B^+ A B^+)$.

Proof of Theorem 3. First let us see that (6–7) do estimate the intended quantities. Suppose that (13) holds (going to subsequences, if necessary). It is enough to consider convergence of (22). This converges to

$$\sum_{j=1}^{m} \nu_j \int g(x) \frac{h_\theta(x)}{\sum_{k=1}^{m} h_k(x) e^{\eta_k}} f_j(x)\, d\mu(x) = \int g(x) h_\theta(x)\, d\mu(x) = c(\theta) E_\theta g(X)$$

Thus (6) does converge to $E_\theta g(X)$ if η_0 is known. The next task is to show that this also happens if $\hat{\eta}_n \to \eta_0$ and $\hat{\eta}_n$ is used in calculating h_{mix}. Then

$$1 - \epsilon \le \exp\{\hat{\eta}_{n,k} - \eta_k\} \le 1 + \epsilon, \qquad k = 1, \ldots, m$$

holds eventually, and

$$\begin{aligned} &\liminf_{|n| \to \infty} \frac{1}{|n|} \sum_{j=1}^{m} \sum_{i=1}^{n_j} g(X_{ij}) \frac{h_\theta(X_{ij})}{\sum_{k=1}^{m} h_k(X_{ij}) e^{\hat{\eta}_{n,k}}} \\ &\ge \liminf_{|n| \to \infty} \frac{1}{|n|} \sum_{j=1}^{m} \sum_{i=1}^{n_j} g(X_{ij}) \frac{h_\theta(X_{ij})}{\sum_{k=1}^{m} h_k(X_{ij})(1+\epsilon) e^{\eta_k}} \\ &= \frac{1}{1+\epsilon} c(\theta) E_\theta g(X) \end{aligned}$$

with probability one and similarly for the limit superior, so

$$\frac{1}{|n|} \sum_{j=1}^{m} \sum_{i=1}^{n_j} g(X_{ij}) \frac{h_\theta(X_{ij})}{\sum_{k=1}^{m} h_k(X_{ij}) e^{\hat{\eta}_{n,k}}} \to c(\theta) E_\theta g(X) \tag{33}$$

Proof of Theorem 4. Without loss of generality assume $E_\theta g(X) = 0$. By the remarks preceding the theorem and Slutsky's theorem, (23) holds if and only if

$$\frac{1}{\sqrt{|n|}} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \frac{g(X_{ij}) h_\theta(X_{ij})}{h_{\text{mix}}(X_{ij})} \xrightarrow{\mathcal{D}} N(0, \tau^2) \tag{34}$$

where $\tau^2 = c(\theta)^2 \sigma^2$. With the normalizing constants estimated by reverse logistic regression the left hand side of (34) becomes

$$\begin{aligned} &\frac{1}{\sqrt{|n|}} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \frac{g(X_{ij}) h_\theta(X_{ij})}{\sum_{k=1}^{m} h_k(X_{ij}) e^{\hat{\eta}_{n,k}}} \\ &= \frac{1}{\sqrt{|n|}} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \frac{g(X_{ij}) h_\theta(X_{ij})}{h_{\text{mix}}(X_{ij})} \left(1 + \frac{\sum_{l=1}^{m} h_l(X_{ij}) [e^{\eta_l} - e^{\hat{\eta}_{n,l}}]}{\sum_{k=1}^{m} h_k(X_{ij}) e^{\hat{\eta}_{n,k}}} \right) \end{aligned} \tag{35}$$

and the second term in parentheses converges to zero almost surely, hence in probability, uniformly in i. The term outside the parentheses is bounded in probability. Hence (35) equals (34) plus a term that is $o_p(1)$, so they converge to the same limiting distribution.

References

Anderson, J. A. (1972), "Separate Sample Logistic Discrimination," *Biometrika*, 59, 19–35.

——— (1982), "Logistic Discrimination," in *Handbook of Statistics* (Vol. 2), eds. P. R. Krishnaiah and L. N. Kanal, Amsterdam: North-Holland, pp. 169–191.

Besag, J. (1992), "Discussion of the Paper by Geyer and Thompson," *Journal of the Royal Statistical Society*, Ser. B, 54, 690.

Besag, J. and Green, P. J. (1993), "Spatial Statistics and Bayesian Computation" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 55, 25–37.

Chan, K. S. (1993a), "Asymptotic Behavior of the Gibbs Sampler," *Journal of the American Statistical Association*, 88, 320–326.

——— (1993b), "On the Central Limit Theorem for an Ergodic Markov Chain," *Stochastic Processes and their Applications*, 47, 113–117.

Chan, K. S. and Geyer C. J. (in press), "Discussion of the Paper by Tierney," *Annals of Statistics*.

Gelfand, A. E. and Smith A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Geman, S. and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Geyer, C. J. (1991a), "Monte Carlo Maximum Likelihood for Dependent Data," *Computer Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, 156–163.

——— (1991b), "Reweighting Monte Carlo Mixtures," Technical Report No. 568, School of Statistics, University of Minnesota.

——— (1992), "Practical Markov Chain Monte Carlo" (with discussion), *Statistical Science*, 7, 473–511.

Geyer, C. J. and Møller, J. (in press), "Simulation and Likelihood Inference for Spatial Point Processes," *Scandanavian Journal of Statistics*.

Geyer, C. J. and Thompson, E. A. (1992), "Constrained Monte Carlo Maximum Likelihood for Dependent Data" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 54, 657–699.

Green, P. J. (1992), "Discussion of the Paper by Geyer and Thompson," *Journal of the Royal Statistical Society*, Ser. B, 54, 683–684.

Haberman, S. J. (1989), "Concavity and Estimation," *Annals of Statistics*, 17, 1631–1661.

Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and their applications," *Biometrika*, 57, 97–109.

Jensen, S. T., Johansen, S. and Lauritzen, S. L. (1991), "Globally Convergent Algorithms for Maximizing a Likelihood Function," *Biometrika*, 78, 867-877.

Kipnis, C. and Varadhan, S. R. S. (1986), "Central Limit Theorem for Additive Functionals of Reversible Markov Processes and Applications to Simple Exclusions," *Communications in Mathematical Physics*, 104, 1–19.

Lehmann, E. L. (1983), *Theory of Point Estimation* (2nd ed.), New York: Wiley.

Liu, J., Wong, W. H., and Kong, A. (in press), "Correlation Structure and Convergence Rate of the Gibbs Sampler with Various Scans," *Journal of the Royal Statistical Society*, Ser. B.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953), "Equation of State Calculations by Fast Computing Machines," Journal of Chemical Physics, 21, 1087–1092.

Rao, C. R. and Mitra, S. K. (1971), *Generalized Inverse of Matrices and its Applications*. New York: Wiley.

Rockafellar, R. T. (1970), *Convex Analysis*, Princeton: Princeton University Press.

Rockafellar, R. T. and Wets, R. J. B. (in press), *Variational Analysis*, New York: Springer-Verlag.

Schervish, M. J. and Carlin, B. P. (1992), "On the Convergence Rate of Successive Substitution Sampling," *Journal of Computational and Graphical Statistics*, 1, 111–127.

Smith, A. F. M. (1992), "Discussion of the Paper by Geyer and Thompson," *Journal of the Royal Statistical Society*, Ser. B, 54, 684–686.

Smith, A. F. M. and Roberts, G. O. (1993), "Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 55, 3–23.

Thompson, E. A. and Guo, S. W. (1991), "Evaluation of Likelihood Ratios for Complex Genetic Models," *IMA Journal of Mathematics Applied in Medicine and Biology*, 8, 149–169.

Thompson E. A., Lin S., Olshen A. B., and Wijsman E. M. (1993), "Monte Carlo Analysis on a Large Pedigree," *Genetic Analysis Workshop 8, Genetic Epidemiology*, 10, 677–682

Tierney, L. (in press), "Markov Chains for Exploring Posterior Distributions" (with discussion), *Annals of Statistics*.