Reweighting Monte Carlo Mixtures

By
Charles J. Geyer[1].
Technical Report No. 568
School of Statistics
University of Minnesota
December 6, 1991

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

Abstract

Markov chain Monte Carlo (e. g., the Metropolis algorithm, Hastings algorithm, and Gibbs sampler) is a general multivariate simulation method applicable to a wide range of problems. It permits sampling from any stochastic process whose density is known up to a constant of proportionality. The Gibbs sampler has recently received much attention as a method of simulating from posterior distributions in Bayesian inference, but Markov chain Monte Carlo is no less important in frequentist inference with applications in maximum likelihood, hypothesis testing, and the parametric bootstrap. It 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 may perform much better. But using such a mixture gives rise to another problem when the densities are known only up to constants of proportionality. These normalizing constants must be calculated to obtain the mixture density. Direct Monte Carlo estimation, though possible, is very inefficient. A new method, reverse logistic regression, accurately estimates these constants, permitting the use of these mixture estimates in Markov chain Monte Carlo.

1 Introduction

Markov chain Monte Carlo methods are general multivariate simulation tools applicable to a wide range of statistical inference problems. In ordinary (independent sample) Monte Carlo one estimates an integral

$$Pg = \int g(x)\, dP(x)$$

by averaging over independent, identically distributed samples X_1, X_2, ... from P

$$\mathbb{P}_n g = \frac{1}{n} \sum_{i=1}^{n} g(X_i).$$

Consistency and asymptotic normality of the estimate follow from the law of large numbers and the central limit theorem

$$\mathbb{P}_n g \xrightarrow{\text{a. s.}} Pg \tag{1}$$

and

$$\sqrt{n}\,(\mathbb{P}_n - P)\, g \xrightarrow{\mathcal{D}} N(0, \sigma_g^2) \tag{2}$$

where $\sigma_g^2 = \operatorname{Var} g(X)$. Markov chain Monte Carlo is the notion of replacing independent samples with a Markov chain X_1, X_2, ... having P as a stationary distribution. Then if the chain is irreducible (1) will hold, and under further regularity conditions (Shervish and Carlin, 1990; Chan, 1991; Liu, Wong and Kong, 1991; Tierney, 1991) a central limit theorem (2) will hold as well, the only difference being that σ_g^2 will now have the form

$$\sigma_g^2 = \sum_{t=-\infty}^{\infty} \gamma_t \tag{3}$$

where

$$\gamma_t = \gamma_{-t} = \operatorname{Cov}\Big(g(X_0), g(X_t)\Big),$$

X_0 here having the stationary distribution (see Hastings, 1970, Geyer 1991, or standard works on time series, for details).

Another way to look at Markov chain Monte Carlo is the following. Note that (1) holds simultaneously for all functions g in any countable family (since a countable union of null sets is a null set). Hence if the sample space is a second countable topological space (e. g. $\mathbb{R}^d$ or any separable metric space) and the countable family of functions is taken to be indicators of open sets in the countable base and their finite intersections, then, for almost all sample paths of the Markov chain,

$$\mathbb{P}_n 1_B \xrightarrow{\text{a. s.}} P 1_B, \qquad \text{for all open sets } B,$$

which implies

$$\mathbb{P}_n \xrightarrow{\mathcal{D}} P \tag{4}$$

(Billingsley, 1968, Theorem 2.2). That is, just as with independent sampling, the empirical converges in distribution to the truth for almost all sample paths. Even

though the samples are not independent and their marginal distributions are never exactly the equilibrium distribution, the cloud of sample points looks like the equilibrium distribution for large sample sizes.

The first Markov chain Monte Carlo algorithm was given by Metropolis, et. al (1953). This was generalized by Hastings (1970). The Hastings algorithm is apparently the most general Markov chain Monte Carlo scheme that is actually useful. It can be used to efficiently simulate almost any stochastic process. More precisely, it simulates distributions whose densities are known "up to a constant of proportionality." Given a function h that is nonnegative, integrable, and not zero almost everywhere, the Hastings algorithm simulates a Markov chain having a stationary distribution with whose density is proportional to h.

A special case of the Hastings algorithm was described under the name "Gibbs sampler" by Geman and Geman (1984) and was used by them and by others (see Besag, York, and Mollié, 1991) for Bayesian image reconstruction and related problems in spatial statistics. The Gibbs sampler has recently received widespread attention as a general method for Bayesian inference following the paper of Gelfand and Smith (1990). Despite the trendiness of the Gibbs sampler, it often leads to severe difficulties that are easily handled by the Hastings algorithm. Hence it should not be considered the method of choice but merely one form of the Hastings algorithm to be used only when the sampling it requires from one-dimensional conditional distributions is easy.

Markov chain Monte Carlo is not limited to Bayesian inference. It has also been used for Monte Carlo hypothesis testing (Besag and Clifford, 1989, 1991) and for Monte Carlo maximum likelihood (Ogata and Tanemura, 1981, 1984, 1989; Penttinen, 1984; Strauss, 1986; Geyer and Thompson, 1992).

In the maximum likelihood problem, one is not interested in just one distribution. There is a parametric family of distributions known up to a constant of proportionality

$$f_\theta(x) = \frac{1}{z(\theta)} h_\theta(x) \tag{5}$$

where the functions h_θ are taken to be known, but the normalizing constant

$$z(\theta) = \int h_\theta(x)\, d\mu(x)$$

is intractable and must be estimated by Monte Carlo using

$$\frac{z(\theta)}{z(\psi)} = \int \frac{h_\theta(x)}{h_\psi(x)} f_\psi(x)\, d\mu(x) = E_\psi \frac{h_\theta(X)}{h_\psi(X)} \tag{6}$$

which is valid for any θ and ψ. Taking θ as a variable and ψ as fixed, this expresses the function $z(\theta)$ up to a constant of proportionality as an expectation with respect to P_ψ. Hence it can be calculated by averaging over a Markov chain with equilibrium distribution P_ψ, which can be generated by the Hastings algorithm without knowing the value of $z(\psi)$. This gives an estimate of the log likelihood from which maximum likelihood estimates can be determined

$$l_n(\theta) = \log \frac{h_\theta(x)}{h_\psi(x)} - \log\left(\frac{1}{n} \sum_{i=1}^{n} \frac{h_\theta(X_i)}{h_\psi(X_i)} \right) \tag{7}$$

The maximizer of (7) is the Monte Carlo approximant of the MLE (see Geyer and Thompson, 1992 for details).

The formula (6) is closely related to the importance sampling formula. Define importance weights by

$$w_{n,\theta}(x) = \frac{h_\theta(x)/h_\psi(x)}{\sum_{i=1}^n h_\theta(X_i)/h_\psi(X_i)}. \tag{8}$$

and a weighted empirical distribution $\mathbb{P}_{n,\theta}$ as the distribution that puts mass $w_{n,\theta}(X_i)$ at point X_i. Then, under the assumption that all of the h_θ are continuous and that h_ψ is strictly positive, for almost all sample paths of the Markov chain Monte Carlo,

$$\mathbb{P}_{n,\theta} \xrightarrow{\mathcal{D}} P_\theta, \qquad \forall \theta.$$

That is we have (4) not just for P_ψ but for all distributions in the family.

This principle of estimating everything of interest from one run of the Markov chain applies to Bayesian inference as well as maximum likelihood. If we think of the state variable x as being the parameter of a model and θ as being a hyperparameter of the prior, this says that in the typical Bayesian applications of Markov chain Monte Carlo we can estimate the posterior under many different priors from one Monte Carlo run.

Not all of the estimates will be equally good, of course. The farther P_θ is from P_ψ the worse the approximation will be. The same effect occurs in likelihood inference. The farther θ is from ψ, the worse (7) approximates the actual log likelihood. This leads to the conclusion that it is rarely efficient to reweight a distribution in the family of interest. We have not, after all, used any properties of h_ψ other than continuity and positivity. It could have been any function which is proportional to a probability density.

In order to do well for a wide range of parameter values, h_ψ should be chosen so that it puts appreciable mass under each h_θ of interest, that is it should be a mixture of all of the f_θ. There seems, however, to be no easy way to discover such a mixture. The f_θ are generally unknown because their normalizing constants $z(\theta)$ are unknown. One could, of course, just try out Markov chains for various functions h until a chain is found that spreads out under all distributions of interest. This may, however, be extremely difficult to do, since the state space may be of large dimension in problems of interest, and finding a distribution that spreads out in just the right way may be extremely hard.

Typically, in problems attacked by Markov chain Monte Carlo, one has no idea what any of the distributions of interest look like except from Monte Carlo experiments. So to have any idea where h_ψ should put mass, one needs to collect samples from a number of distributions in the model, say we have collected a sample X_{1j}, ..., $X_{n_j j}$ from P_{θ_j} for $j = 1, \ldots, m$. Then the pooled samples can be thought of as a sample from

$$f_{\text{mix}} = \sum_{j=1}^m \frac{n_j}{|n|} f_{\theta_j} = \sum_{j=1}^m \frac{n_j}{|n|} \frac{1}{z(\theta_j)} h_{\theta_j} \tag{9}$$

where $n = (n_1, \ldots, n_m)$ and $|n| = n_1 + \cdots + n_m$.

This distribution spreads out under the distributions it is composed of, and ones nearby. If the distributions sampled cover the interesting region of the state space, it will be a good idea to use f_{mix} as the h_ψ in the formulas (7) and (8). But to do this we need to know f_{mix} or rather h_{mix} which is f_{mix} multiplied by an unknown constant. But we don't know h_{mix} exactly because we don't know the $z(\theta_j)$. To complete this program, we need a method of estimating the $z(\theta_j)$ up to a constant of proportionality using the samples already collected.

We could use (6) or rather its Monte Carlo analogue

$$\frac{z(\theta_k)}{z(\theta_j)} \approx \frac{1}{n_j} \sum_{i=1}^{n_j} \frac{h_{\theta_k}(X_{ij})}{h_{\theta_j}(X_{ij})} \tag{10}$$

but this has several drawbacks. It gives $m(m-1)$ incompatible estimates for the $m-1$ quantities of interest and none of them use all of the available data. Moreover, the estimates (10) suffer from the very problem we are trying to cure. They will be bad whenever θ_j and θ_k are far apart.

What is needed is a way to estimate the $z(\theta_j)$ using all of the data. Such a method is explained in the next section.

2 Reverse Logistic Regression

In order to simplify notation, let us omit the θ's giving h_j for h_{θ_j}, z_j for $z(\theta_j)$, P_j for P_{θ_j} and so forth.

Rewrite (9) as

$$f_{\text{mix}}(x) = \sum_{j=1}^{m} h_j(x) e^{\psi_j + a_{nj}} \tag{11}$$

where for convenience we have defined

$$\psi_j = -\log z_j \tag{12}$$

and

$$a_{nj} = \log \frac{n_j}{|n|}.$$

Let

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

Then, given that the value x was observed in the mixture sample, the probability that it occurred in the jth chain is $p_j(x, \psi + a_n)$, where a_n denotes the vector $(a_{n1}, \ldots, a_{nm})$.

Considering both ψ and the measure μ as unknown parameters, the log likelihood for estimating the mixture density is

$$\begin{aligned} l_{\text{full}}(\psi, \mu) &= \sum_{j=1}^{m} \sum_{i=1}^{n_j} \log[f_j(X_{ij}) \mu(\{X_{ij}\})] \\ &= \sum_{x \in S} \sum_{j=1}^{m} k_j(x) \log \left[p_j(x, \psi + a_n) p_{\text{mix}}(x) e^{-a_{nj}} \right]. \end{aligned}$$

where

$$S = \{ X_{ij} : j = 1, \ldots, m, \ i = 1, \ldots, n_j \} \tag{13}$$

is the combined sample, where $k_j(x)$ is the number of points in the jth sample having the value x, and where

$$p_{\text{mix}}(x) = f_{\text{mix}}(x)\mu(\{x\})$$

is the probability of the point x under the mixture distribution.

Consider maximizing l_{full} as a function of μ holding ψ fixed. Since μ is arbitrary so is p_{mix}, hence the nonparametric maximum likelihood estimate of p_{mix} is obviously the empirical distribution

$$\hat{p}_{\text{mix}}(x) = \frac{1}{|n|} \sum_{j=1}^{m} k_j(x),$$

which does not depend on the value of ψ. Plugging this back into the log likelihood gives the profile likelihood for ψ

$$l_n(\psi) = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \log p_j(X_{ij}, \psi + a_n) + \sum_{x \in S} \sum_{j=1}^{m} \log \left[\hat{p}_{\text{mix}}(x) e^{-a_{nj}} \right].$$

Since the second term does not contain the parameter ψ, it is irrelevant to inference about ψ, and may be ignored, which gives

$$l_n(\psi) = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \log p_j(X_{ij}, \psi + a_n) \tag{14}$$

as the profile likelihood for ψ having supped out the infinite dimensional nuisance parameter μ.

Treating μ as a nuisance parameter may seem strange, since it is known. But it is not "known" in the sense that we can do with it what we want, in this case calculate p_{mix}. Considering it a nuisance parameter yields a procedure that works without knowledge of any specific properties of p_{mix}.

The likelihood (14) is easily maximized since it is arithmetically equivalent to a logistic regression for $m = 2$ and fitting a "log-linear" or "multinomial response" model for $m > 2$. It is not, however, statistically equivalent. The "response" (which sample a point is in) is fixed and the "predictor" (the position of a point) is random. The regression is reversed. We shall take as our estimator of ψ the maximizer of (14). This will be referred to as the reverse logistic regression estimator (even for $m > 2$).

This estimation procedure is very similar in spirit to logistic discrimination. The argument given here for reverse logistic regression being maximum likelihood follows the argument of Anderson (1972, 1982) for separate sample logistic discrimination, though it can be somewhat simplified in this context.

A further simplification ensues if we consider maximizing the function

$$g_n(\eta) = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \log p_j(X_{ij}, \eta). \tag{15}$$

This yields an equivalent problem since $l_n(\psi) = g_n(\psi + a_n)$ so $\hat{\psi}_n$ is a maximizer of l_n if and only if $\hat{\eta}_n = \hat{\psi}_n + a_n$ is a maximizer of g_n. Since

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

when $\eta = \psi + a_n$, it is enough for most practical purposes to only estimate η. The corresponding estimate of ψ is only of interest when the sample size is varied.

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 ψ only up to an additive constant, and hence determines an estimate of f_{mix} only up to a constant of proportionality.

This, however, does not settle the question of identifiability. There are further complexities. Consider the following condition

Condition A *Let*

$$A_{jk} = \{\, x : h_j(x) > 0 \text{ and } h_k(x) > 0 \,\}.$$

Then there do not exist disjoint sets J and K such that

$$\mu(A_{jk}) = 0, \qquad \text{whenever } j \in J \text{ and } k \in K.$$

and its finite-sample analogue

Condition B *Let*

$$A'_{jk} = \{\, x \in S : h_j(x) > 0 \text{ and } h_k(x) > 0 \,\}.$$

Then there do not exist disjoint sets J and K such that

$$A'_{jk} = \varnothing, \qquad \text{whenever } j \in J \text{ and } k \in K.$$

When condition A fails to hold we say the problem is *separated*, and when condition B fails to hold we say the finite-sample problem is separated.

If Condition A does not hold, the sample space can be divided into disjoint sets A and A^c such that every distribution P_j is concentrated on one or the other. Then the problem divides into two completely independent problems, and there is no loss of generality in adopting it. As will be seen, Condition A guarantees asymptotic identifiability of ψ or η up to an additive constant. Condition B guarantees the identifiability of ψ or η for finite sample sizes.

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 than the logistic discrimination model.

2.2 Consistency

Turning to a different issue, we do not assume that the samples X_{1j}, X_{2j}, ... are i. i. d. P_j, since this will not be the case in Markov chain Monte Carlo. It will be enough if an ergodicity condition holds

Condition C *For any P_j-integrable function h*

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

With these preliminaries out of the way, we can now state a theorem on the consistency of our proposed estimator.

Theorem 1 *If condition B holds, the function l_n (resp. g_n) has a unique maximizer subject to the constraint that the ψ_k (resp. η_n) sum to zero. Suppose*

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

Then, under Conditions A and C, the maximizer of l_n is a strongly consistent estimator of the true ψ (up to an additive constant).

A proof is given in the appendix.

2.3 Asymptotic Normality

For simplicity in discussing asymptotic normality we assume that the sampling fractions converge to a nonzero limit

Condition D

$$\frac{n_j}{n_1 + \cdots + n_m} \to \nu_j > 0.$$

Under this condition if we let ψ_0 be the true value of ψ, let $a = \log \nu$, and let $\eta_0 = \psi_0 - a$, then if $\sqrt{n}(\hat{\eta}_n - \eta_0)$ has a central limit theorem, $\sqrt{n}(\hat{\psi}_n - \psi_0)$ converges to the same limit.

Such a central limit theorem will hold whenever, whenever the Markov chain Monte Carlo has a central limit theorem for the function ∇g_n, that is whenever

Condition E

$$\frac{1}{\sqrt{n}} \nabla g_n(\eta_0) \xrightarrow{\mathcal{D}} N(0, A) \tag{16}$$

holds. The variance matrix A has the form given by (3) for the diagonal entries and by a similar formula with cross-correlations replacing autocorrelations for the off-diagonal terms. It 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, 1991).

Theorem 2 *If conditions A, C, D, and E hold, then*

$$-\frac{1}{n}\nabla^2 g_n(\eta_0) \xrightarrow{\text{a. s.}} B \tag{17}$$

where

$$\begin{aligned} B_{rr} &= \sum_{j=1}^{m} \nu_j E_j p_r(X,\eta)[1 - p_r(X_{ij},\eta)] \\ B_{rs} &= -\sum_{j=1}^{m} \nu_j E_j p_r(X,\eta) p_s(X,\eta), \qquad r \neq s \end{aligned}$$

and

$$\sqrt{n}(\hat{\psi}_n - \psi_0) \xrightarrow{\mathcal{D}} N\left(0, B^+ A B^+\right) \tag{18}$$

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

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

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

A proof is given in the appendix.

3 Examples

Examples demonstrating the method will be taken from a two-parameter Ising model on a 32×32 square lattice with periodic boundary conditions. This model is formally described as follows. Given an $n \times n$ square lattice, indexed according to some scheme, let $i \sim j$ denote that sites i and j are nearest neighbors. The lattice is taken to be glued at the edges to make a torus so that every site has four nearest neighbors. The state variable of the random field is a vector $x = \{x_i\}$ of random variables taking values in $\{-1, 1\}$, one random variable for each lattice site. The statistical model is a two-parameter exponential family with natural statistics $t_1(x) = \sum_i x_i$ and $t_2(x) = \sum_i \sum_{j \sim i} x_i x_j$.

For concreteness we will call the lattice sites with $x_i = 1$ "white pixels" and the rest "black pixels" following the language of image processing. In this language t_1 is the excess of white over black pixels, and t_2 is the excess of concordant nearest neighbor pairs (both $+1$ or both -1) over discordant pairs.

The probability of a point x in the sample space is

$$f_\theta(x) = \frac{1}{z(\theta)} e^{\langle t(x), \theta \rangle}$$

where $\langle t, \theta \rangle = t_1\theta_1 + t_2\theta_2$ and z is the normalizing constant (or partition function)

$$z(\theta) = \sum_{x \in \mathcal{S}} e^{\langle t(x), \theta \rangle}, \tag{19}$$

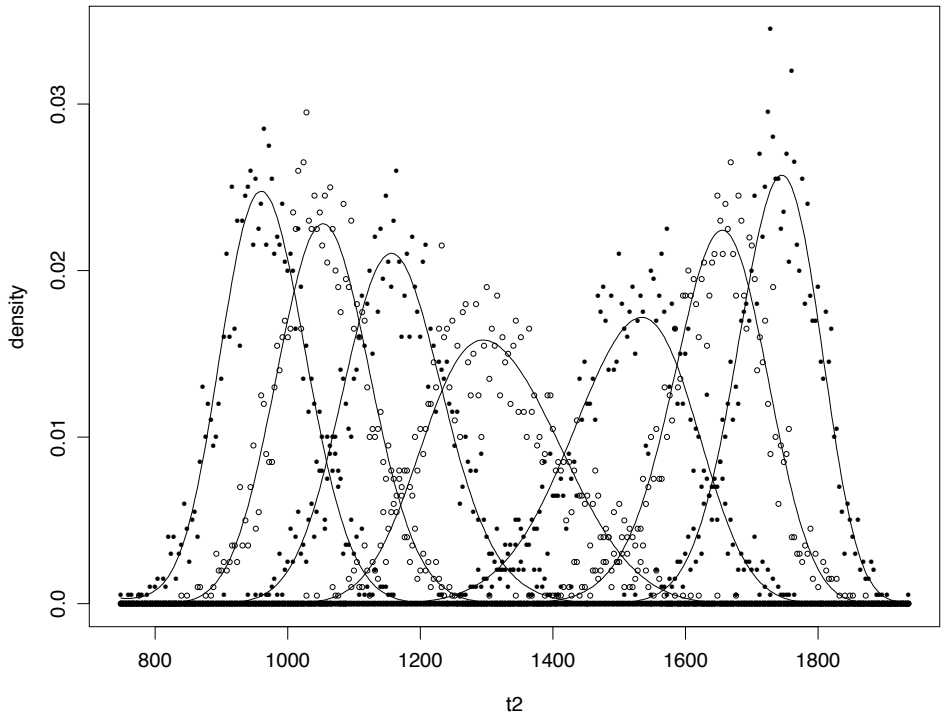

Figure 1: Empirical and smoothed densities for the second canonical statistic of an Ising model, for values of the corresponding canonical parameter 0.365, 0.385, 0.405, 0.425, 0.445, 0.465, 0.485. Curves alternate black and white dots. The sample size is 2000 for each sample.

$\mathcal{S}$ being the state space of $2^{32\times 32}$ possible values of x. The parameters θ_1 and θ_2 are referred to here as the "level" parameter and "dependence" parameter respectively. We shall also use the notation α for θ_1 and β for θ_2. This is a family in which no analytic formula is known for the normalizing constants $z(\theta)$, but samples can be simulated by the Hastings algorithm. Here a Metropolis algorithm accelerated by the method of "symmetry swaps" (Geyer, 1991) was used to obtain rapid mixing for all parameter values.

Consider the distributions shown in Figure 1, which are for an Ising model with $\alpha = 0$ so $t_2(x)$ is the natural sufficient statistic. They are all absolutely continuous with respect to each other, so in principle, any one can be used to estimate any other via importance reweighting. For practical sample sizes, however, in this case 2000, the ranges of some of the samples do not even overlap, it is completely unreasonable to attempt to estimate the distribution at one end by reweighting the sample at the other end. They are, however, all well estimated by reweighting the mixture distribution. The smooth curves in Figure 1 are all reweighted versions of a smooth density estimate of the mixture distribution. Note how well each of the empirical curves is fitted.

Figure 2 compares the reverse logistic regression estimates of ψ obtained by maximizing (14) with the direct Monte Carlo integration estimates obtained from

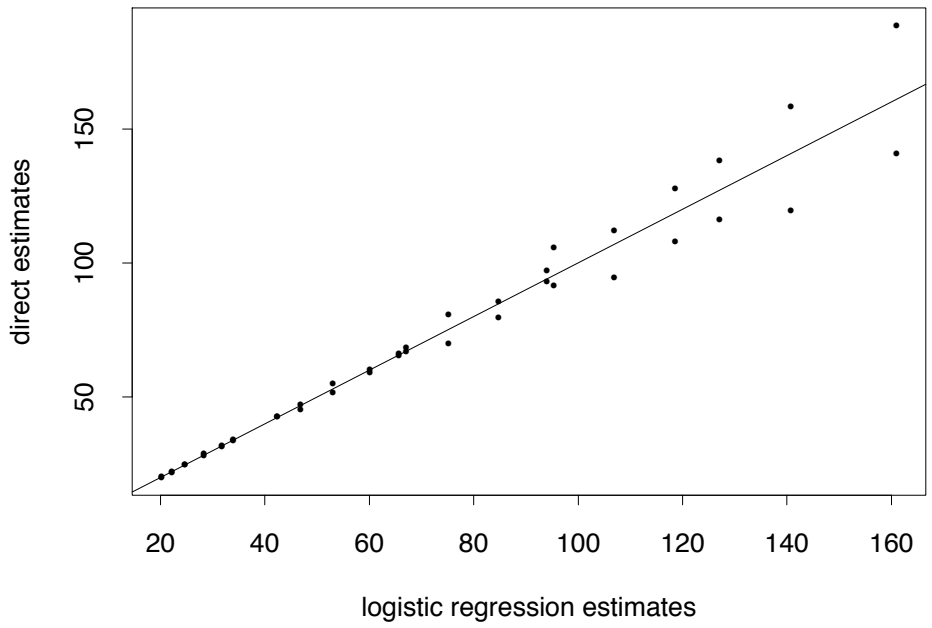

Figure 2: Comparison of Reverse Logistic Regression Estimates and Direct Estimates. Points are differences, estimates of $\psi_j - \psi_k$ for all pairs j, k. Solid line is equality.

(10). For each difference $\psi_j - \psi_k$ there are one logistic regression estimate and two direct estimates. Note that the logistic regression estimate is usually right in the middle of the direct estimates. Some of the errors in the direct estimates are very large. At the right end there is a difference of 47.7 between the two estimates of ψ, which corresponds to a ratio of $e^{47.7} = 2 \times 10^{21}$ disagreement in the estimates of the ratio of the z's. This worst case occurs for the two distributions that are farthest apart. The best cases are the six points at the left side whose differences and differences from the reverse logistic regression estimates are too small to see on the scale of the plot. The differences are still not negligible, however. The largest two of the six are 0.38 and 0.52 corresponding to factors of 1.46 and 1.68 disagreement about the z's.

That the errors for adjacent distributions are much smaller suggests estimating the very large differences by adding the small ones. This is better, though since we have six differences to add and two estimates for each, we get $2^6 = 64$ disagreeing estimates, the mean of which, 160.95, agrees almost exactly with the reverse logistic regression estimator, 160.97. The standard error of the 64 estimates is, however, not negligible, 0.35. Since this adding of estimates does not generalize to higher dimensions, we will say no more of it. The only point was to show how well the reverse logistic regression procedure works.

3.1 Maximum Likelihood

Now consider doing maximum likelihood in the one-parameter family obtained by setting $\alpha = 0$ in the Ising model. Figure 1 shows the distributions of t_2 for seven different values of β, 0.365, 0.385, 0.405, 0.425, 0.445, 0.465, 0.485, which span the critical value 0.44069 at which an infinite lattice model "freezes." As β goes through the critical value, the distribution of t_1 goes from being unimodal (at low β) to strongly bimodal (at high β).

Suppose we wish to calculate the maximum likelihood estimate of θ given an observation x. Since this is an exponential family the maximum likelihood estimate (MLE) is obtained by finding the θ that satisfies $E_\theta(X) = x$. Let $\tau(\theta) = E_\theta(X)$ denote the mapping from the canonical parameter to the mean value parameter, so the MLE is the solution of $\tau(\theta) = x$.

The Monte Carlo analog solves $\tau_n(\theta) = x$ where

$$\tau_n(\theta) = \frac{\sum_{j=1}^n t(X_j) e^{\langle t(X_j), \theta - \theta_j \rangle}}{\sum_{j=1}^n e^{\langle t(X_j), \theta - \theta_j \rangle}} \tag{20}$$

which is obtained by differentiating (7) where $h_\theta(x) = e^{\langle t(x), \theta \rangle}$.

The estimate $\tau_n(\theta)$ is, however, accurate only for θ near θ_j. This is illustrated by Figure 3, which shows (dotted lines) the curve τ_n estimated from each of the samples shown in Figure 1. Note that each of the samples does well (is close to the solid line, which was obtained by tilting the mixture) over a small range of θ values near the θ_j for that curve. Elsewhere the curves do very poorly. This is only to be expected; as θ goes from $-\infty$ to ∞ the value of $\tau_n(\theta)$ goes from $\min_i X_{ij}$ to $\max_i X_{ij}$. Since the samples do not cover the whole sample space, neither can the range of the τ_n curves.

The solid curve in Figure 3 is estimated by replacing f_{θ_j} by f_{mix} and the jth sample by the mixture sample in (20). Specifically, with S given by (13) and f_{mix} defined by (11) with ψ estimated by the reverse logistic regression procedure

$$\tau_{n,\text{mix}}(\theta) = \frac{\sum_{x \in S} t(x) f_\theta(x) / f_{\text{mix}}(x)}{\sum_{x \in S} f_\theta(x) / f_{\text{mix}}(x)}$$

As can be seen from Figure 3, it is a much better estimator than any of the individual curves.

How this applies to maximum likelihood is shown in Figures 4 and 5. In Figure 4 the MLE derived using the solid curve in Figure 3 is compared with an MLE calculated using one sample from the distribution in the middle of the range in Figure 1. In order to make a fair accuracy comparison, the same number of points in total were used for both estimates, seven samples of 2,000 for the first and one sample of 14,000 for the second. This shows that the single-sample MLE is not as accurate, and that the accuracy decreases in the tails of the distribution (for estimates far from the parameter value of the simulations). The single-sample MLE does not do badly; of the 500 points there are only four with relative errors of more than half a percent and one with relative error more than one percent. But even at these relatively large sample sizes there is noticeable error.

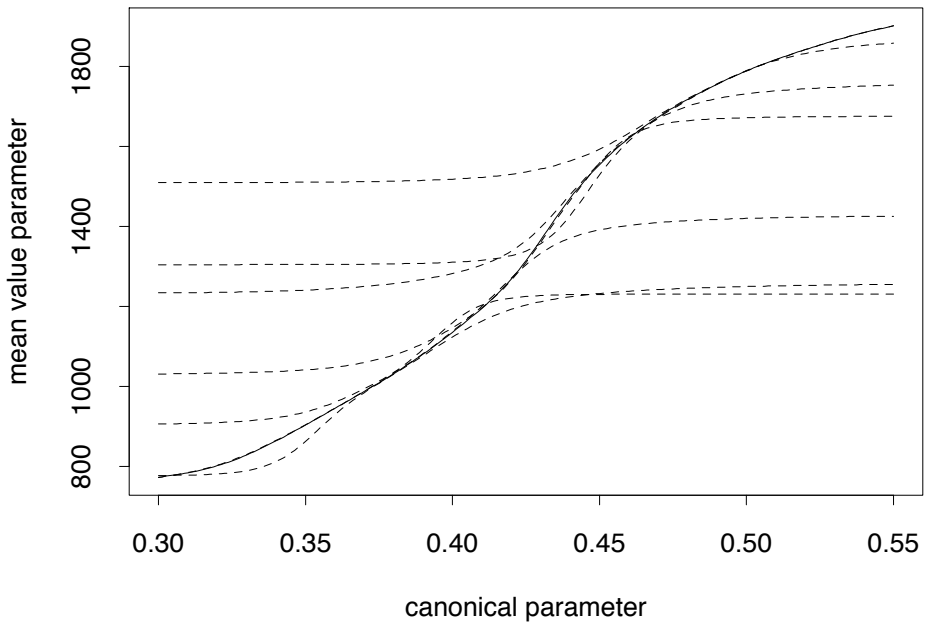

Figure 3: Estimates of the curve mapping the canonical parameter to mean value parameter. Dashed lines are the curves estimated from the seven separate samples shown in Figure 1. Solid line is the curve estimated from their mixture.

Figure 5 compares a mixture MLE and single-sample MLE derived from much smaller sample sizes, only 1,000 points. The 14,000-point mixture MLE is taken as the standard of comparison. For this example we experimented with unequal sample sizes in the mixture. Subsamples of sizes 0, 100, 200, 400, 200, 100, and 0, going from left to right, were taken from the samples shown in Figure 1. The MLE with the mean value map τ estimated from this mixture shown by the white dots in Figure 5. Note that even with light sampling in the tails the Monte Carlo error does not grow rapidly as one moves away from the center. The black dots show an estimate using a subsample from the large sample used for the single-sample estimate of Figure 4. Now there is a large difference between the two methods. The 1,000-point mixture MLE does almost as well as the 14,000-point single-sample MLE. In fact its worst relative error is smaller. The single-sample MLE now has very large errors, the largest relative error being almost five percent.

One-dimensional problems, though easy to visualize, do not reveal the full advantages of mixture estimates. Both mixture estimates and single-sample estimates suffer from the "curse of dimensionality." But single-sample estimates suffer much more. Single-sample estimation goes bad near the "boundary" of the sample (which for definiteness we may take to be its convex hull). As the dimension of the sample space increases, an increasing fraction of the sample is near the boundary. So for high dimensional problems a procedure like the preceding examples of simulating the sampling distribution of the MLE using just one sample to estimate the likelihood is unworkable. Mixture methods promise to be better behaved. It does seem that

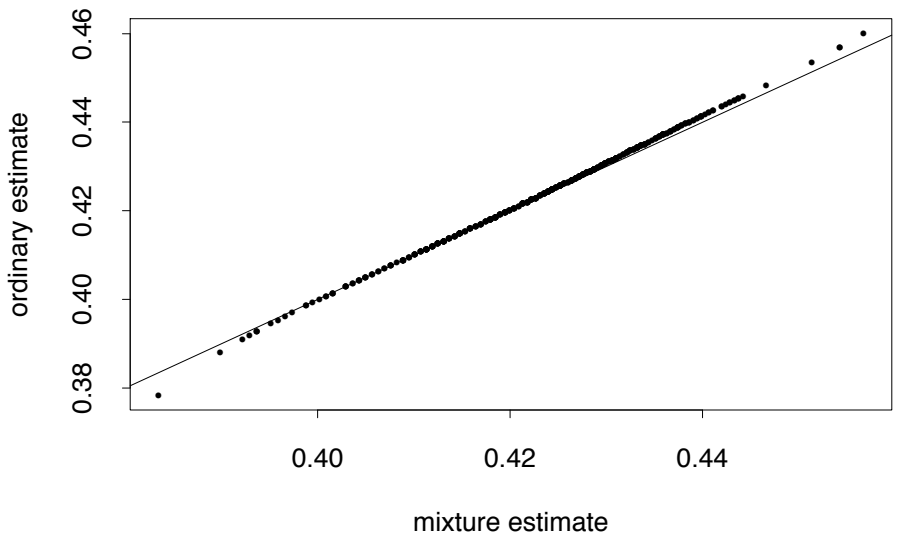

Figure 4: Distribution of mixture and non-mixture Monte Carlo maximum likelihood estimates. Distribution of MLE's for 500 samples from an Ising model with parameters $\alpha = 0$ and $\beta = .425$. The mixture estimate used the seven samples of 2,000 points shown in Figure 1 to estimate the function τ. The ordinary estimate used 14,000 points in one sample from the distribution with parameters $\alpha = 0$ and $\beta = .425$. Solid line is equality.

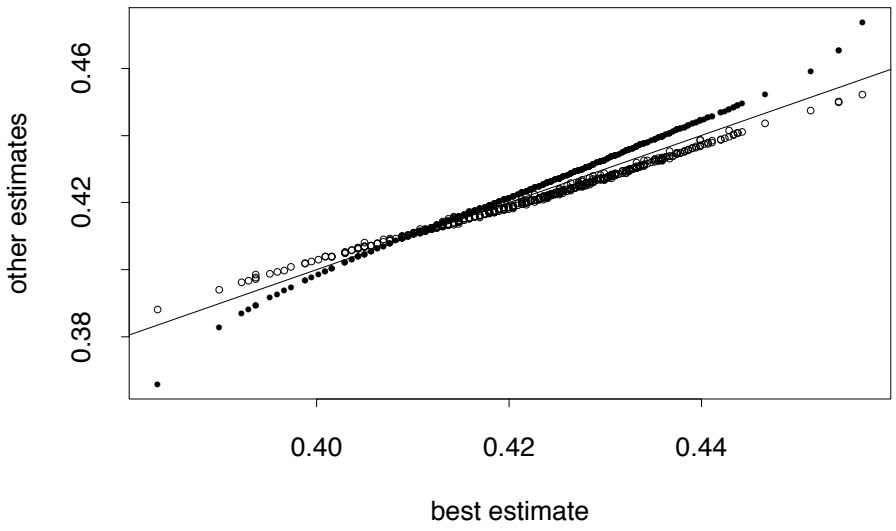

Figure 5: Distribution of mixture and non-mixture Monte Carlo maximum likelihood estimates using small background samples. White dots compare the mixture MLE of Figure 4 with an MLE calculated using sample sizes of 100, 200, 400, 200, 100 rather than 2,000 each. Black dots compare mixture MLE of Figure 4 with an MLE calculated one sample of size 1,000 from the distribution with parameters $\alpha = 0$ and $\beta = .425$. Solid line is equality.

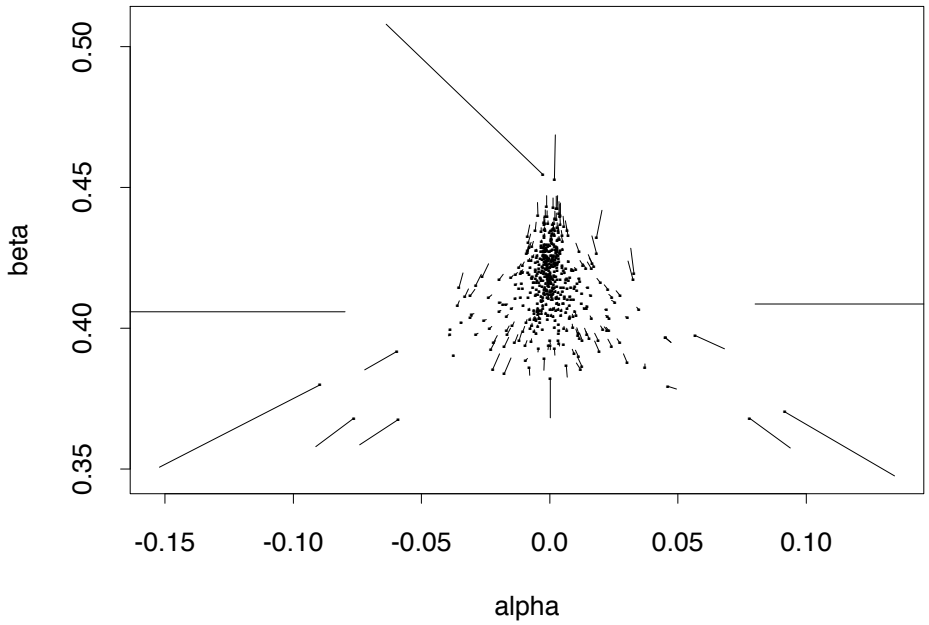

Figure 6: Distribution of mixture and non-mixture Monte Carlo maximum likelihood estimates in two dimensions. Dots are mixture MLE estimates for both parameters of the Ising model corresponding to a sample of 500 realizations of an Ising model with parameters $\alpha = 0$ and $\beta = .425$. The mixture had five components that were actual samples with parameter vectors (0,0.425), (0.005,0.420), (0,0.410), (0,0.435), and (0.010,0.410), each with 400 points (2,000 points in all). Each of these samples was "doubled" by changing α to $-\alpha$ and all values of t_1 to their negatives. Estimates were also calculated using one sample of size 2,000 with parameter (0,0.425). This was also "doubled" by changing all t_1 values to their negatives. The black lines connect the mixture estimates to the one-sample estimates. The two lines leaving the figure indicate one-sample esimates that do not exist (are "at infinity").

the number of components in the mixture would have to increase with dimension, but at least mixture methods can be used, albeit at some additional cost, when single-sample methods fail completely. These points remain largely unexplored. For now we shall end with a two-dimensional example.

This is again from the Ising model, now with both parameters α and β being estimated. Again 500 MLE's were calculated using both mixture and single-sample Monte Carlo schemes. The results are shown in Figure 6. The parameter values for the mixture were chosen with knowledge (from earlier single-sample studies) of the distribution of the MLE, but no attempt was made to get an optimal mixture. Each component of the mixture was "doubled" using the natural symmetry of the Ising model. The model is symmetric to change of signs of the x_i and simultaneous change of sign of α. This change of sign for each of the x_i changes $t_1(x)$ to $-t_1(x)$ and leaves $t_2(x)$ unchanged. The procedure adds only a slight complication to the estimation and has the desirable property that the Monte Carlo MLE's share the symmetry of the model. This symmetry trick was used for both the single-sample and mixture methods; it helps both equally well.

Going from one to two dimensions brings qualitatively new behavior. Two of the single-sample MLE's do not exist. This occurs whenever an observation for which we wish to calculate an MLE lies outside the convex hull of the sample. Unless the "foreground" sample of points for which we calculate MLE's is much smaller than the "background" sample which we use to calculate the (Monte Carlo) likelihood, foreground points outside the convex hull of the background points will occur with high probability, especially if the dimensionality is large. Besides the two points with undefined single-sample MLE's there are another three with huge errors. All of the points near the boundary of the background sample are poorly estimated by the single-sample method. In order to check that in fact it was the single-sample estimate that was wrong when the two methods were in disagreement, different single-sample estimates were calculated using new single samples of size 4000 with parameter values very near those being estimated. These showed that, as expected, the mixture MLE's were correct, not the single-sample MLE's.

4 Discussion

Markov chain Monte Carlo can be used to simulate any stochastic process whose densities are known up to the normalizing constants. In order to carry out statistical inference about parameters of the model it may be necessary to simulate from many distributions in the model. This is obvious for likelihood-based methods, but may also be true in Bayesian methods if more than one prior is under consideration (hence more than one posterior) or if inference about many quantities is contemplated (hence many useful importance sampling distributions).

In such cases it is very inefficient to simply sample from each distribution of interest (possibly infinitely many of them). Importance weighting must be used to reweight some samples to other distributions, which allows one sample to provide information about many distributions. This only works well if the distribution

being reweighted places mass in all regions of interest, i. e., under all distributions of interest. One way to accomplish this is to use a mixture of distributions in the model as the distribution to reweight. For complex models about which little is known except from Monte Carlo, this may be the only useful method. Any improvements would use detailed knowledge about the model above and beyond what is required for Markov chain Monte Carlo simulation to work.

In order to carry out the reweighting, the mixture density must be estimated, which, since the individual densities are "known up to constants of proportionality," means estimating these proportionality constants. A good estimation method that requires no knowledge beyond that required for Markov chain Monte Carlo is reverse logistic regression. This method uses all the data to produce a single estimate of the proportionality constants. The method seems to work well. It is not clear that improvements are needed or would be worth the cost of the regularity conditions that would be necessary to get improvements.

Another way to justify the use of mixture distributions is the following. Suppose that one does not follow the advice given above to make all inferences using just one Monte Carlo sample from just one distribution (possibly a mixture) but instead uses several samples from several different distributions. Then these inferences are not as good as they could be, since they do not use "all the data," in this case all the Monte Carlo samples, in each inference. The way to use all the data is to derive each inference from the mixture distribution of the samples, i. e. to follow the advice.

This does not argue against using one sample from a distribution that is not a mixture. If one could discover a distribution that is easy to sample from and approximates mixtures of distributions in the model, it would be preferable to use it. There doesn't, however, seem to be any systematic way to discover such distributions. There is no available theory to use as a guide, and haphazard experimentation does not produce better results than using mixtures, at least in the author's experience.

For maximum likelihood estimation, these methods satisfactorily solve one problem that was left open in Geyer and Thompson (1992). The Monte Carlo log likelihood (7) is a good approximation only for θ such that the distribution determined by h_ψ puts appreciable mass under the distribution determined by h_θ. When h_ψ is a distribution in the model, this usually means when θ is near ψ. But for many purposes, such as doing a "parametric bootstrap" of the MLE as done in our last example, one wants the approximation to hold over a wide range of θ, rather far out in the tails of the sampling distribution of $\hat{\theta}_n$. This cannot be accomplished for any h_ψ in the model. It is necessary to use a mixture.

Though it has not been discussed or used in this paper, it should said that the cost of using mixtures can be drastically reduced. It costs as much to generate a mixture of m components each with n sample points as to generate a sample of size mn from a single distribution. The method of Metropolis-coupled chains (Geyer, 1991) typically gives each of m coupled chains of length n the accuracy of one chain of length mn. When this is true, using Metropolis-coupled Markov chain Monte Carlo to generate the mixtures means that the mixture comes "for free". One gets the whole mixture for the cost of estimating any one of its components with the

same accuracy. There is (approximately) no extra cost of sampling. There is the cost of the logistic regression to determine the normalizing constants, but this is negligible. Just to reiterate, this method has not been used in the comparisons in our examples. Mixtures do better than single samples without use of coupled chains. Using coupled chains only increases what is already an advantage to the use mixtures.

Acknowledgements

Much of this paper was written during a year at the University of Chicago supported by an NSF Postdoctoral Fellowship. Peter McCullagh pointed out the connection between the reverse logistic regression estimator described here and separate sample logistic discrimination. He also suggested that (14) might be a partially maximized likelihood and made other helpful comments. R. R. Bahadur suggested the approach that makes Theorem 1 work without assuming convergence of the sampling fractions.

A Proofs

Proof of Theorem 1. First suppose that there is a vector ν such that $n_j/|n| \to \nu_j$ as $|n| \to \infty$. Then by the ergodicity condition

$$\frac{1}{|n|} \sum_{j=1}^{m} p_j(X_{ij}, \eta) \to \nu_j E_j \log p_j(X, \eta) \tag{21}$$

almost surely $[P_j]$ for any fixed η. Hence, a countable union of null sets being a null set, (21) 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

$$\frac{1}{|n|} g_n(\eta) \to \gamma(\eta) = \sum_{j=1}^{m} \nu_j E_j \log p_j(X, \eta)$$

simultaneously for all η in the countable dense set.

An outline of the proof goes as follows. Both g_n and γ are concave functions, finite and twice differentiable everywhere. Direct calculation shows that the gradient of γ is zero at $\eta = \psi + a$, where $a_j = \log \nu_j$ and ψ is the truth (12). Direct calculation also shows that the Hessians of g_n and γ have no null eigenvectors except $u = (1, 1, \ldots, 1)$, which establishes the uniqueness assertions. From well known theorems of convex analysis (see Rockafellar and Wets, forthcoming, or Haberman, 1989, for details) convergence on a dense set implies uniform convergence on compact sets implies convergence of maximizers (provided that the limit, here γ, has a unique maximizer, which it does). That is, $\hat{\eta}_n \to \eta$ which implies $\hat{\psi}_n \to \psi$. This proves that ψ_n is consistent since for any subsequence n_k there is a further subsequence n'_l along which $a_{n'_l}$ converges to some a, which implies that $\hat{\psi}_{n'_l} \to \psi$. 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, i. e. $\hat{\psi}_n \to \psi$.

We now begin filling in the details of these assertions. To check concavity, we calculate derivatives.

$$\frac{\partial p_j(x,\eta)}{\partial \eta_j} = p_j(x,\eta)[1 - p_j(x,\eta)] \tag{22a}$$

$$\frac{\partial p_j(x,\eta)}{\partial \eta_r} = -p_j(x,\eta)p_r(x,\eta), \qquad j \neq r \tag{22b}$$

Hence

$$\begin{aligned} \frac{\partial g_n(\eta)}{\partial \eta_r} &= \sum_{i=1}^{n_r} [1 - p_r(X_{ir},\eta)] - \sum_{j \neq r} \sum_{i=1}^{n_j} p_r(X_{ij},\eta) \\ &= n_r - \sum_{j=1}^{m} \sum_{i=1}^{n_j} p_r(X_{ij},\eta) \end{aligned} \tag{23a}$$

$$-\frac{\partial^2 g_n(\eta)}{\partial \eta_r^2} = \sum_{j=1}^{m} \sum_{i=1}^{n_j} p_r(X_{ij},\eta)[1 - p_r(X_{ij},\eta)] \tag{23b}$$

$$-\frac{\partial^2 g_n(\eta)}{\partial \eta_r \partial \eta_s} = -\sum_{j=1}^{m} \sum_{i=1}^{n_j} p_r(X_{ij},\eta)p_s(X_{ij},\eta), \qquad r \neq s \tag{23c}$$

(Note the similarity to logistic regression.) Since $\sum_r p_r(X_{ij},\eta) = 1$ by definition, the matrix $-\partial^2 g_n(\eta)/\partial \eta_r \partial \eta_s$ is the sum of positive semi-definite matrices, which have the form of covariance matrices of multinomials. Hence g_n has a negative semi-definite Hessian and is concave.

The constrained maximizer of g_n will be unique if u is its only null eigenvector. For any vector φ the bilinear form with the Hessian satisfies

$$-\sum_{r=1}^{m} \sum_{s=1}^{m} \frac{\partial^2 g_n(\eta)}{\partial \eta_r \partial \eta_s} \varphi_r \varphi_s = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \left[p_r(X_{ij},\eta)\varphi_r^2 - \left(\sum_{s=1}^{m} p_s(X_{ij},\eta)\varphi_s \right)^2 \right] \geq 0,$$

and is zero when φ is a null eigenvector. Moreover, this is true term by term

$$p_r(X_{ij},\eta)\varphi_r^2 - \left(\sum_{s=1}^{m} p_s(X_{ij},\eta)\varphi_s \right)^2 \geq 0 \tag{24}$$

for any vector φ and is zero when φ is a null eigenvector because each term is a bilinear form for the covariance matrix of some multinomial. Since (24) is zero only where it achieves its minimum, φ is a null eigenvector if and only if the gradient of (24) is zero, that is if

$$p_r(X_{ij},\eta) \left[\varphi_r - \sum_{s=1}^{m} p_s(X_{ij},\eta)\varphi_s \right] = 0, \qquad \forall i, j \tag{25}$$

If for any indices r and r' there is some X_{ij} such that both $p_r(X_{ij},\eta)$ and $p_{r'}(X_{ij},\eta)$ are nonzero (which occurs when both $h_r(X_{ij})$ and $h_{r'}(X_{ij})$ are both nonzero), then

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

Hence whenever $A'_{rr'}$ as defined in Condition B is nonempty $\varphi_r = \varphi_{r'}$. Hence under under Condition B all of the φ_r are the same (if for disjoint sets J and K such that $\varphi_j \neq \varphi_k$, $j \in J$ and $k \in K$, then $A'_{jk} = \varnothing$ for each such j and k and Condition B fails). So u is the only null eigenvector. Thus g_n has a unique maximizer subject to the constraint (and the same is true of l_n).

Clearly γ is concave since it is the expectation of g_n. We begin our demonstration that γ is everywhere finite by showing that it is finite at the true ψ. Clearly $\gamma \leq 0$, and at the true ψ

$$-\log p_j(x, \psi) = \log\left(1 + \sum_{k \neq j} \frac{f_k(x)}{f_j(x)}\right) \leq \sum_{k \neq j} \frac{f_k(x)}{f_j(x)}$$

so

$$-E_j \log p_j(X, \psi) \leq \sum_{k \neq j} \int \frac{f_k(x)}{f_j(x)}\, dP_j(x) = \sum_{k \neq j} \int f_k(x)\, d\mu(x) = m - 1$$

Next we calculate the gradient of γ at any point η where γ is finite. Since the difference quotients for directional derivatives of a concave function converge monotonely (a property of convexity, see Rockafellar, 1970, Theorem 23.1) directional derivatives may be commuted with expectations by monotone convergence.

$$\begin{aligned}
\gamma'(\eta; \varphi) &= \lim_{h \downarrow 0} \frac{\gamma(\eta + h\varphi) - \gamma(\eta)}{h} \\
&= \sum_{j=1}^{m} \nu_j E_j \left(\frac{d}{dh} \log p_j(X, \eta + h\varphi) \bigg|_{h=0} \right) \\
&= \sum_{r=1}^{m} \varphi_r \left(\nu_r - \sum_{j=1}^{m} \nu_j E_j p_r(X, \eta) \right)
\end{aligned}$$

Since $\gamma'(\eta; -\varphi) = -\gamma'(\eta; \varphi)$, it follows that γ is differentiable at each point where it is finite, and the gradient is defined by

$$\frac{\partial \gamma(\eta)}{\partial \eta_r} = \nu_r - \sum_{j=1}^{m} \nu_j E_j p_r(X, \eta) \tag{26}$$

wherever $\gamma(\eta)$ is finite. But since $p_r(X, \eta)$ is uniformly bounded between 0 and 1, so is (26), hence γ must be finite and differentiable everywhere.

Because the integrand in (26) is uniformly bounded we can again commute expectations and derivatives by dominated convergence, so the Hessian of γ is given by

$$\begin{aligned}
-\frac{\partial^2 \gamma(\eta)}{\partial \eta_r^2} &= \sum_{j=1}^{m} \nu_j E_j p_r(X, \eta)[1 - p_r(X_{ij}, \eta)] \\
-\frac{\partial^2 \gamma(\eta)}{\partial \eta_r \partial \eta_s} &= -\sum_{j=1}^{m} \nu_j E_j p_r(X, \eta) p_s(X, \eta), \qquad r \neq s
\end{aligned}$$

By arguments similar to those applied to the Hessian of g_n we get the analogue of (25)

$$E_j p_r(X,\eta)\left[\varphi_r - \sum_{s=1}^{m} p_s(X_{ij},\eta)\varphi_s\right] = 0, \qquad \forall j$$

for any null eigenvector φ. Hence with A_{jk} as defined in Condition A and $x \in A_{rr'}$

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

Thus Condition A implies that ϕ is proportional to u and that γ has a unique maximum subject to the constraint.

It remains only to be shown that $\psi + a$ maximizes γ, i. e., that (26) has a zero there.

$$\begin{aligned}\frac{\partial\gamma(\psi+a)}{\partial\psi_r} &= \nu_r - \sum_{j=1}^{m}\nu_j E_j \frac{\nu_r \frac{1}{z_r} h_r(X)}{\sum_k \nu_k \frac{1}{z_k} h_k(X)} \\ &= \nu_r - \sum_{j=1}^{m}\nu_j \int \frac{\nu_r f_r(x)}{\sum_k \nu_k f_k(x)} f_j(x)\, d\mu(x) \\ &= \nu_r - \int \nu_r f_r(x)\, d\mu(x) = 0\end{aligned}$$

So $\nabla\gamma(\psi + a) = 0$ at the true ψ. This concludes the proof of the theorem.

Proof of Theorem 2. From (22) it is clear that differentiating $\nabla^2 g_n(\eta_0)$ 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 $\nabla g_n(\eta_0)$ is

$$\nabla g_n(\eta) = \nabla g_n(\eta_0) + \nabla^2 g_n(\eta_0)(\eta - \eta_0) + nO(\|\eta - \eta_0\|^2)$$

Hence defining B_n by

$$-\frac{1}{n}\left(\nabla g_n(\hat\eta_n) - \nabla g_n(\eta_0)\right) = B_n(\hat\eta_n - \eta_0)$$

and using the consistency of $\hat\eta_n$, the ergodicity condition C, and dominated convergence

$$B_n = -\frac{1}{n}\nabla^2 g_n(\eta_0) + O(\|\hat\eta_n - \eta_0\|) = -\frac{1}{n}\nabla^2 g_n(\eta_0) + o_p(1) \xrightarrow{P} B. \tag{27}$$

Now a slight problem arises because of the nonidentifiability of the sum of the η_i, i. e. $u'\eta$. Note that $Au = Bu = 0$ by algebraic identities. So we take as the likelihood equations, imposing the constraint $u'\eta = 0$

$$\begin{pmatrix} B_n \\ u' \end{pmatrix}\sqrt{n}(\hat\eta_n - \eta_0) = \begin{pmatrix} \frac{1}{\sqrt{n}}\nabla g_n(\eta_0) \\ 0 \end{pmatrix} \tag{28}$$

Hence applying Condition E, (27), and Lemma 6.4.1 in Lehmann (1983)

$$\sqrt{n}(\hat\eta_n - \eta_0) \xrightarrow{\mathcal{D}} Y \tag{29}$$

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 is the Moore-Penrose inverse of B (Rao and Mitra, 1971, p. 51 ff.) Hence Y is distributed $N(0, B^+ A B^+)$, which concludes the proof of the theorem.

References

Anderson, J. A. (1972) Separate sample logistic discrimination. *Biometrika*, **59**, 19–35.

Anderson, J. A. (1982) Logistic discrimination. In P. R. Krishnaiah and L. N. Kanal, eds., *Handbook of Statistics*, Vol. 2. Amsterdam: North-Holland, 169–191.

Besag, J. and Clifford, P. (1989) Generalized Monte Carlo significance tests. *Biometrika* **76**, 633–642.

Besag, J. and Clifford, P. (1991) Sequential Monte Carlo p-values. *Biometrika* **78**, 301–304.

Besag, J., York, J. and Mollié, A. (1991) Bayesian image restoration, with two applications in spatial statistics (with discussion). *Ann. Inst. Statist. Math.* **43**, 1–59.

Billingsley, P. (1968) *Convergence of Probability Measures*. New York: Wiley.

Chan, K. S. (1991) Asymptotic behavior of the Gibbs sampler. Technical Report No. 294, Department of Statistics, University of Chicago.

Gelfand, A. E. and Smith A. F. M. (1990) Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Assoc.*, **85**, 398–409.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell.*, **6**, 721–741.

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

Geyer, C. J. and Thompson, E. A. (1992) Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. R. Statist. Soc.* B, to appear.

Haberman, S. J. (1989) Concavity and estimation. *Ann. Statist.* **17**, 1631–1661.

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

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

Liu, J., Wong, W. H., and Kong, A. (1991) Correlation structure and convergence rate of the Gibbs sampler with various scans. Technical Report No. 304, Department of Statistics, University of Chicago.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953) Equation of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1092.

Ogata, Y. and Tanemura M. (1981) Estimation of interaction potentials of spatial point patterns through the maximum likelihood procedure. *Ann. Inst. Statist. Math.*, **33**, Part B, 315-338.

Ogata, Y. and Tanemura M. (1984) Likelihood analysis of spatial point patterns. *J. R. Statist. Soc.* B, **46**, 496–518.

Ogata, Y. and Tanemura M. (1989) Likelihood estimation of soft-core interaction potentials for Gibbsian point patterns. *Ann. Inst. Statist. Math.*, **41**, 583–600.

Penttinen, A. (1984) Modelling interaction in spatial point patterns: Parameter estimation by the maximum likelihood method. *Jyväskylä Studies in Computer Science, Economics, and Statistics*, **7**.

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. (forthcoming) *Variational Analysis*. New York: Springer-Verlag.

Shervish, M. J. and Carlin, B. P. (1990) On the convergence rate of successive substitution sampling. Technical Report, No. 492, Department of Statistics, Carnegie-Mellon University.

Strauss, D. (1986) A general class of models for interaction. *SIAM Rev.*, **28**, 513–527.

Tierney, L. (1991) Markov chains for exploring posterior distributions. Technical Report No. 560, School of Statistics, University of Minnesota.