AD-A266 563

(12)

MAXIMUM LIKELIHOOD ESTIMATION FOR CONSTRAINED OR MISSING DATA MODELS

Alan E. Gelfand
Bradley P. Carlin

DTIC
ELECTE
JUL 12 1993
S A D

TECHNICAL REPORT No. 468
MAY 5, 1993

Prepared Under Contract
N00014-92-J-1264 (NR-042-267)
FOR THE OFFICE OF NAVAL RESEARCH

Reproduction in whole or in part is permitted
for any purpose of the United States Government.

Approved for public release; distribution unlimited

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305-4065

Reproduced From
Best Available Copy

93 7 09 082

93-15611

MAXIMUM LIKELIHOOD ESTIMATION FOR CONSTRAINED OR MISSING DATA MODELS

Alan E. Gelfand
Bradley P. Carlin

TECHNICAL REPORT No. 468
MAY 5. 1993

Prepared Under Contract
N00014-92-J-1264 (NR-042-267)
FOR THE OFFICE OF NAVAL RESEARCH

Professor Herbert Solomon, Project Director

Reproduction in whole or in part is permitted
for any purpose of the United States Government.

Approved for public release; distribution unlimited

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305-4065

DTIC QUALITY INSPECTED 5

Reproduced From
Best Available Copy

Accesion For
NTIS CRA&I ☑
DTIC TAB ☐
Unannounced ☐
Justification
By
Distribution /
Availability Codes
Dist | Avail and / or Special
A-1

Maximum Likelihood Estimation for Constrained or Missing Data Models

Alan E. GELFAND and Bradley P. CARLIN

University of Connecticut and University of Minnesota

Key words and phrases: EM algorithm; Gibbs sampler; Monte Carlo approximant.

Abstract

In statistical models involving constrained or missing data, likelihoods containing integrals emerge. In the case of both constrained and missing data, the result is a ratio of integrals, which for multivariate data may defy exact or approximate analytic expression. Seeking maximum likelihood estimates in such settings, we propose Monte Carlo approximants for these integrals, and subsequently maximize the resulting approximate likelihood. Iteration of this strategy expedites the maximization, while the Gibbs sampler is useful for the required Monte Carlo generation. As a result, we handle a class of models broader than the customary EM setting without using an EM-type algorithm. Implementation of the methodology is illustrated in two numerical examples.

1 Introduction

In challenging parametric modeling settings the maximum likelihood estimator is generally the estimator of choice. This follows from both foundational considerations (e.g. the Likelihood Principle) as well as practical ones (e.g. good large sample behavior under mild conditions). Here we propose a Monte Carlo approach for calculation of maximum likelihood estimators which handles a range of previously inaccessible problems.

The context we have in mind results in a likelihood function which is unavailable explicitly. Some likelihoods of this type have been analyzed using the EM algorithm (Dempster, Laird and

Rubin, 1977). As we clarify later in this section, however, the class of models we envision yields a likelihood which cannot be handled by the customary version of this algorithm.

Though we present our method in terms of general multivariate joint distributions, all of our illustrations and data examples assume an underlying exponential family of models. This is because the behavior of the likelihood surface and hence the properties of the MLE are perhaps best understood in such families (see e.g. Barndorff-Nielsen, 1978; Brown, 1986; Jacobsen, 1988, and references therein). We do not address theoretical concerns regarding e.g. existence, uniqueness, consistency, or asymptotic normality. Rather, we offer a method for obtaining the maximum of the likelihood when it is reasonably well behaved. Problems and remedies associated with poorly behaved likelihoods are well discussed in the literature and apply to our approach as well. In particular, the use of multiple starting points with a given maximization routine often helps avoid convergence to local, rather than global, maxima.

Our models assume the observed data to be constrained in some way. We also allow for the possibility of missing data with constraints upon the entire set of variables, both observed and unobserved. As a general version of this setting let x denote the k-dimensional observed data vector and $\boldsymbol{\theta}$ the p-dimensional parameter vector. We suppose that the likelihood takes the form

$$L(\boldsymbol{\theta}; \mathbf{x}) = \frac{c_1(\mathbf{x}; \boldsymbol{\theta})}{c_2(\boldsymbol{\theta})} , \tag{1}$$

where

$$c_1(\mathbf{x}; \boldsymbol{\theta}) = \int_{\mathbf{y} \in C(\mathbf{x})} f(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) d\mathbf{y} , \tag{2}$$

f is a parametric family of densities, and

$$c_2(\boldsymbol{\theta}) = \int_{\mathbf{x} \in S} c_1(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x} .$$

That is, as a function of x, $L(\theta; \mathrm{x})$ is a normalized density function. Here y is viewed as a t-dimensional missing (or latent) data vector constrained by the observed x to the set $C(\mathrm{x})$, with the observed x itself constrained to a set S.

In the case where there is no missing data we take $c_1(\mathrm{x}; \theta)$ to be a parametric family of densities for x with $c_2(\theta)$ being the normalizing constant arising from the restriction of x to S. Computation of the function $c_2(\theta)$ then requires a k-dimensional integration which we presume cannot be carried out explicitly. In fact with k large and S awkward, evaluation of $c_2(\theta)$ at a particular θ_0 may defy exact or approximate analytic numerical integration. Hence, we are drawn to Monte Carlo approaches. In the case of latent or missing data y, computation of the function $c_1(\mathrm{x}; \theta)$ requires a t-dimensional integration over a constrained set $C(\mathrm{x})$, which again we cannot carry out explictly. Moreover, $c_2(\theta)$ now requires a $(k+t)$-dimensional integration.

Of course, in principle one could attempt a grid search for the maximizing θ in (1). That is, at a given θ, perform Monte Carlo integrations for c_1 and c_2 to obtain L, and then search through the space of θ for a maximum L. This is in fact the approach that emerges in several papers from the econometrics literature on simulated moments estimation; see for example McFadden (1989) and Pakes and Pollard (1989). The primary concern of these papers appears to be the accuracy and precision of the simulator carrying out the integrations, rather than efficient maximization using such a simulator. These papers typically assume the observed data x to be a deterministic function of the latent data y, thus simplifying the structure of the integrals in (1). Constrained multivariate normal models for y are presumed, resulting in tailored simulators inappropriate for the broader class of statistical models we envision (see Section 2). Moreover, when p is large a naive grid search for the MLE $\hat{\theta}$ may be impractical; for smaller p our proposed method is faster.

The EM algorithm is a widely used tool for handling incomplete data problems. It cannot, however, accommodate constraints on the observed data. That is, it presumes that that $c_1(\mathrm{x}; \theta)$

is itself a normalized density in x, so that $c_2(\theta)$ in (1) disappears. To clarify, recall the general version of the EM algorithm as presented in Dempster, Laird and Rubin (1977). At the lth stage, given $\theta = \theta^{(l)}$ the E-step computes a function $Q(\theta'|\theta = \theta^{(l)})$ which the M-step then maximizes over θ'. In the case of (1),

$$Q(\theta'|\theta) = \frac{\int_{C(\mathrm{x})} f(\mathrm{x},\mathrm{y}|\theta)\log f(\mathrm{x},\mathrm{y}|\theta')d\mathrm{y}}{c_1(\mathrm{x};\theta)} - \log c_2(\theta') .$$

But since this expression still involves c_1 and c_2, it is no easier to work with than (1). In summary, if the likelihood is of the form (1) we need to approximate one or both of $c_1(\mathrm{x};\theta)$ and $c_2(\theta)$ as functions of θ. While Monte Carlo integration seems to be a natural tool in this regard, it is not immediately clear how to proceed.

We develop Monte Carlo approximants following importance sampling ideas proposed in Geyer and Thompson (1992). In their setting (an autologistic model), data vectors $\mathrm{X}_1,\ldots,\mathrm{X}_n$ come from an exponential family model where only the nonnormalized form of the exponential kernel, $\exp(\theta'\mathrm{T}(\mathrm{x}))$, is specified. That is, the vector T is chosen such that the sufficient statistic, $\sum_{i=1}^n \mathrm{T}(\mathrm{x}_i)$, is a suitable summary of the data. Hence the normalizing constant, a function of θ and therefore needed for the ML estimation, is unknown and requires integration over X to compute. Geyer and Thompson introduce a Monte Carlo approximant for this function, as well as an iterative approach for carrying out the maximization. We generalize their ideas beyond approximation of the normalizing function to a broad class of constrained or missing data problems.

The format for the remainder of the paper is as follows. In Section 2 we offer a collection of motivating examples where likelihoods of the form (1) arise. In Section 3 we formalize the Monte Carlo approach. Finally, Section 4 presents two datasets for which models of the form (1) are used and the approach of Section 3 is carried out.

2 Illustrative Examples

In order to demonstrate the range of application of our approach, in the following subsections we present three situations where the form (1) arises. Two of these are analyzed more fully using appropriate datasets in Section 4. Other illustrations include multivariate biased sampling settings, and the analysis of adaptive patient followup schemes for clinical trials data.

2.1 Patterned covariance models

An elementary illustration is provided by a constrained vector x from a multivariate normal having mean 0 and covariance $\Sigma(\theta)$, assumed to be a patterned matrix. Such structure arises in variance components models, time series models, and moving average processes. Particular forms include (a) $\Sigma_{ii}(\theta) = \sigma^2$, $\Sigma_{ij}(\theta) = \rho\sigma^2$, (b) $\Sigma_{ij}(\theta) = \sigma^2\rho^{|i-j|}$, (c) $\Sigma_{ij}(\theta) = \sigma^2 b_{|i-j|}$, and (d) $\Sigma_{ij}(\theta) = (m - |i-j|)\sigma^2$, $|i-j| \leq m$; $\Sigma_{ij}(\theta) = 0$, $|i-j| > m$. See also Rubin and Szatrowski (1982) in this regard. Constraints on x might be $|x_i| \leq c_i$, $i = 1, \ldots, k$, or $x_1 < x_2 < \cdots < x_k$. In the equicorrelated case (i) under say the latter constraints, $c_1(\mathrm{x}; \theta)$ is precisely the multivariate normal density $N_k(0, \Sigma(\theta))$ with

$$c_2(\theta) = \int_{x_1 < x_2 < \cdots < x_k} N_k(\mathrm{x}|0, \Sigma(\theta))\, dx_1 dx_2 \cdots dx_k \,.$$

Unless k is very small, calculation of c_2 at a given θ_0 is only feasible using Monte Carlo methods.

2.2 Categorical data models

Consider the case of truncated multinomial trials. For instance, it is sometimes the case that the observation of a zero count is truncated. We dichotomize according to presence or absence and then, *if present*, record how many. If we assume cell counts arise from independent Poisson

distributions in this fashion, the conditional distribution of cell counts given the total number becomes a truncated multinomial. More generally for a fixed number of trials, n, suppose the ith cell is restricted to have at least r_i observations, $i = 1, \ldots, k$. Let $\mathbf{x} = (x_1, \ldots, x_k)$ where x_i denotes the count in the ith cell, with $x_i \geq r_i$ for each i, and q_i is the probability associated with the ith cell. Then, with $\theta = (q_1, \ldots, q_k)$, $c_1(\mathbf{x}; \theta) = n! \prod_{i=1}^{k} q_i^{x_i}/x_i!$ and $c_2(\theta) = \sum_S c_1(\mathbf{x}; \theta)$ where now $S = \{\mathbf{x}\text{: each } x_i \text{ is integer-valued, } x_i \geq r_i, \text{ and } \sum_{i=1}^{k} x_i = n\}$. Other variations include the case where a particular multinomial cell is known to supply the largest count or where the counts are constrained to increase up to a particular cell and then decrease thereafter.

2.3 Compositional data models

Frequently samples are taken such that each observation is a vector whose components sum to 1. Examples include land samples described in terms of proportions of different types of vegetation, soil samples described by proportion of chemical content, and rock samples described by proportion of mineral content. Such data is referred to as compositional data (Aitchison, 1986). Distributional models are specified on the simplex $\{\mathbf{p} = (p_1, p_2, \ldots, p_k) : p_i \geq 0, \sum_{i=1}^{k} p_i = 1\}$.

Let $f(\mathbf{p}|\theta)$ denote a parametric specification of the joint density for $\mathbf{p}$. The most obvious choice of f, the Dirichlet family, is usually undesirable since it forces an assumption of negative correlation amongst every (p_i, p_j) pair, as well as certain conditional independencies. Instead, baseline logit transformations $z_i = \log(p_i/p_k)$ are often adopted, with $\mathbf{z} = (z_1, \ldots, z_{k-1})$ assumed to follow, say, a $N_{k-1}(\mu, \Sigma)$ distribution, so that $\theta = (\mu, \Sigma)$. The $(k-1)$ logits uniquely determine the composition.

The mean μ is often expressed as a parametric function of explanatory variables. Of particular interest is the covariance matrix Σ, since the nature of covariation between proportions is a primary research question. Usually MLE's are sought, so that given samples $\mathbf{p}_t$, $t = 1, \ldots, n$, we would convert to $\mathbf{z}_t$ and then obtain the customary MLE for μ and Σ.

Now suppose there are constraints on the p's. For instance, we might know that the first classification is most common, i.e. $p_1 \geq p_i$, $i = 2, \ldots, k$, or we might know that the classifications are in decreasing order of prevalence, i.e. $p_1 \geq p_2 \geq \cdots \geq p_k$. On the logit scale these convert to $z_1 \geq z_i$, $i = 2, \ldots, k$, and $z_1 \geq z_2 \geq \cdots \geq z_{k-1} > 0$, respectively. If S denotes these constraints and $f(z_1, \ldots, z_{k-1}|\theta)$ denotes the density of z then $c_2(\theta) = \int_S f(z_1, \ldots, z_{k-1}|\theta)\, dz_1 \cdots dz_{k-1}$.

Next imagine that, as often happens, the kth classification is a leftover, or "other" category. Unfortunately what classifcations comprise "other" may vary across data collection sources. That is, we can think of p as the most refined classification vector, but we actually observe q's where components of p are collapsed. Then the likelihood for the observed data $\mathbf{q}_1, \ldots, \mathbf{q}_n$ takes the form

$$L(\theta; \mathbf{q}_1, \ldots, \mathbf{q}_n) = \prod_{t=1}^{n} \int_{C(\mathbf{q}_t)} f(p_{t1}, \ldots, p_{tk}|\theta)\, dp_{t1} \cdots dp_{tk}\ , \tag{3}$$

where $C(\mathbf{q}_t)$ reflects the collapsing of $\mathbf{p}_t$ to give $\mathbf{q}_t$. Whether on the p scale or the z scale, expression (3) is of the form (2). If we also incorporate the previouly described restrictions on p, the likelihood takes the most general form (1).

3 The Monte Carlo approximant approach

Returning to (1), observe that we may write

$$c_1(\mathbf{x}; \theta) = c_1(\mathbf{x}; \theta_0) \cdot \left(\int_{C(\mathbf{x})} \frac{f(\mathbf{x}, \mathbf{y}|\theta)}{f(\mathbf{x}, \mathbf{y}|\theta_0)} f(\mathbf{y}|\mathbf{x}, \theta_0) d\mathbf{y} \right) \left(\int_{C(\mathbf{x})} f(\mathbf{y}|\mathbf{x}, \theta_0) d\mathbf{y} \right)^{-1}\ , \tag{4}$$

and similarly

$$c_2(\theta) = c_2(\theta_0) \cdot \left(\int_S \int_{C(\mathbf{x})} \frac{f(\mathbf{x}, \mathbf{y}|\theta)}{f(\mathbf{x}, \mathbf{y}|\theta_0)} f(\mathbf{x}, \mathbf{y}|\theta_0) d\mathbf{y} d\mathbf{x} \right) \left(\int_S \int_{C(\mathbf{x})} f(\mathbf{x}, \mathbf{y}|\theta_0) d\mathbf{y} d\mathbf{x} \right)^{-1}\ . \tag{5}$$

Thus if $\{\mathrm{w}_j^*,\ j = 1, \ldots, B_1\}$ are drawn from $g(\mathrm{y}|\mathrm{x}, \theta_0)$, the conditional distribution of y given x and θ_0 restricted to $C(\mathrm{x})$, then a Monte Carlo approximant for (4) is

$$\hat{c}_1(\mathrm{x}; \theta) = c_1(\mathrm{x}; \theta_0) \cdot \frac{1}{B_1} \sum_{j=1}^{B_1} \frac{f(\mathrm{x}, \mathrm{w}_j^*|\theta)}{f(\mathrm{x}, \mathrm{w}_j^*|\theta_0)} . \tag{6}$$

If instead $\{\mathrm{x}_j^*,\ j = 1, \ldots, B_2\}$ are drawn from $g(\mathrm{x}|\theta_0)$, the distribution of x given θ_0 restricted to S, and subsequently for each x_j^* a y_j^* is drawn from $g(\mathrm{y}|\mathrm{x}_j^*, \theta_0)$, then a Monte Carlo approximant for (5) is given by

$$\hat{c}_2(\theta) = c_2(\theta_0) \cdot \frac{1}{B_2} \sum_{j=1}^{B_2} \frac{f(\mathrm{x}_j^*, \mathrm{y}_j^*|\theta)}{f(\mathrm{x}_j^*, \mathrm{y}_j^*|\theta_0)} . \tag{7}$$

Hence we may approximate the log likelihood $\log L(\theta; \mathrm{x})$ by $\log \hat{c}_1(\mathrm{x}; \theta) - \log \hat{c}_2(\theta)$. This in turn implies that an approximate MLE $\hat{\theta}$ may be found by maximizing

$$\log \sum_{j=1}^{B_1} \frac{f(\mathrm{x}, \mathrm{w}_j^*|\theta)}{f(\mathrm{x}, \mathrm{w}_j^*|\theta_0)} - \log \sum_{j=1}^{B_2} \frac{f(\mathrm{x}_j^*, \mathrm{y}_j^*|\theta)}{f(\mathrm{x}_j^*, \mathrm{y}_j^*|\theta_0)} , \tag{8}$$

where we have ignored terms free of θ. Thus using the approximants (6) and (7), we have replaced an intractable form (1) with an explicit form (8). Note that the terms free of θ involve the unknown quantities $c_1(\mathrm{x}; \theta_0)$ and $c_2(\theta_0)$, but fortunately these need not be evaluated. Henceforth, we shall refer to the method of finding $\hat{\theta}$ via maximization of (8) as the *MCMLE algorithm*. Notice that if there is no missing data y, a Monte Carlo approximant is not needed for $c_1(\mathrm{x}; \theta)$ in (1), and (8) simplifies to

$$\log f(\mathrm{x}|\theta) - \log \sum_{j=1}^{B_2} \frac{f(\mathrm{x}_j^*|\theta)}{f(\mathrm{x}_j^*|\theta_0)} . \tag{9}$$

A byproduct of the explicit Monte Carlo approximant is the possibility of approximation to the asymptotic covariance matrix of the MLE. If we calculate (either analytically or numerically) the Hessian matrix of mixed partial derivatives of (8) with respect to θ evaluated at the MLE, say

$H(\hat{\theta})$, then $[-H(\hat{\theta})]^{-1}$ provides a rough indication of the uncertainty associated with $\hat{\theta}$.

Implementation of expression (8) poses two challenges: carrying out the required sampling to create the Monte Carlo approximants, and maximizing the resulting expression. With regard to the first challenge, the sampling for (6) and (7) requires making draws from a joint distribution whose form is known perhaps only up to normalizing constant and which is confined to a specified set. If the joint density for x and y is of the form $f(\mathrm{x},\mathrm{y}|\theta)$, then in (6), given x and θ_0 we need to sample from $f(\mathrm{y}|\mathrm{x},\theta_0) \propto f(\mathrm{x},\mathrm{y}|\theta_0)$ restricted to $C(\mathrm{x})$. In (7) we need to sample from $f(\mathrm{x},\mathrm{y}|\theta_0)$ restricted to the set $\{(\mathrm{x},\mathrm{y}) : \mathrm{x} \in S \text{ and } \mathrm{y} \in C(\mathrm{x})\}$.

In some cases, simple rejection sampling (e.g. generating y from $f(\mathrm{y}|\mathrm{x},\theta_0)$ and retaining it if it belongs to $C(\mathrm{x})$), though inefficient, will be easiest. Alternatively, Markov chain Monte Carlo using the Gibbs sampler (see e.g. Gelfand and Smith, 1990; Tierney, 1991) is attractive here since required sampling is from complete conditional distributions, all of which are proportional to the joint density $f(\mathrm{x},\mathrm{y}|\theta)$. Let us adopt the notation $\mathrm{y}_{-i} = (y_1,\ldots,y_{i-1},y_{i+1},\ldots,y_t)$, with a similar definition for x_{-i}. Then in (6) we need to draw from $f(y_i|\mathrm{y}_{-i},\mathrm{x},\theta)$, $i = 1,\ldots,t$, appropriately restricted. In (7) we need also to draw from $f(x_i|\mathrm{x}_{-i},\mathrm{y},\theta)$, $i = 1,\ldots,k$, again appropriately restricted. At a particular y_i or x_i the constraint sets must be viewed in univariate cross sections given the other variates. This typically results in restriction to an interval or a set of intervals. Recent work of Gelfand, Smith, and Lee (1992) is pertinent here in discussing one-for-one sampling from such univariate distributions. Though their work is in the Bayesian framework where integration and sampling is over the parameter space, it is equally well applicable for our situation where integration and sampling is over the data space.

In the approximations (6) and (7), $f(\cdot,\cdot|\theta_0)$ plays the role of an importance sampling density. More generally, for instance, the w_j^*'s in equation (6) might be drawn from any importance sampling density $h(\mathrm{y}|\mathrm{x})$ that is appropriately restricted to $C(\mathrm{x})$. But no such *single* density could possibly

perform well for all integrands over the range of possible values for θ. On the other hand, adopting an h that changed with θ would require a simulation for each θ, rendering any naive MLE grid search algorithm infeasible.

How do we avoid this problem? Our selection of $g(\mathbf{y}|\mathbf{x}, \theta_0)$ for $h(\mathbf{y}|\mathbf{x})$ suggests an iterative approach to create a sequence of importance sampling densities that improves relative to the density f at the MLE $\hat{\theta}$. This approach eliminates the need for grid search and also the costly set-up time in developing appropriate importance sampling densities. The idea follows from Geyer and Thompson (1992), who observed that starting at some $\theta^{(0)}$ if we maximize (8) to obtain $\hat{\theta}$ we can take $\theta^{(1)} = \hat{\theta}$ and rerun the entire procedure, resulting in a new $\hat{\theta}$ and an iterative version of the procedure. To understand the value of iteration we need to distinguish the maximization of (8) from the maximization of (1). Expression (8) is an approximation to (1) which depends upon the accuracy of the approximations (6) and (7). If we view $f(\mathbf{x}, \mathbf{y}|\theta_0)$ as an importance sampling density for $f(\mathbf{x}, \mathbf{y}|\hat{\theta})$, then for a given B_1 and B_2 the Monte Carlo integration for $c_1(\mathbf{x}; \hat{\theta})$ and $c_2(\hat{\theta})$ improves as $f(\mathbf{x}, \mathbf{y}|\theta_0)$ gets "closer" to $f(\mathbf{x}, \mathbf{y}|\hat{\theta})$, i.e. as θ_0 gets closer to $\hat{\theta}$. Thus the sequence $\{\theta^{(l)}\}$ produced by iteration should be getting closer to the true $\hat{\theta}$ which maximizes (1). Since our objective is only to insure a good Monte Carlo approximant, we need not, however, take more than a few iterations to obtain $\theta^{(l)}$ in the vicinity of the true $\hat{\theta}$. At this point, one final iteration with B_1 and B_2 very large (say 10,000) will produce an accurate final estimate. Thus through a small number of iterations we achieve an efficient and broadly applicable maximization strategy.

4 Numerical Examples

4.1 Truncated Correlated Normal

Consider the equicorrelated k-variate normal distribution described in Subsection 2.1 where $\Sigma_{ii} = 1, \Sigma_{ij} = \rho$ for $i \neq j$, subject to truncation to the set $S = \{\mathbf{x} : \max|x_i| < L\}$ for some

$L > 0$. We seek the MLE of $\theta \equiv \rho$. Notice that direct standardization of this likelihood involves a k-dimensional integral for each candidate ρ value, whereas our Monte Carlo approach requires only the generation of $\mathbf{x}_j^*$'s from the truncated correlated normal, given the current approximation ρ_0.

To carry out the required sampling, we note that $\Sigma_{ii}^{-1} = a_k$ and $\Sigma_{ij}^{-1} = b_k$ for $i \neq j$, where $a_k = [-(k-2)\rho - 1]/[(k-1)\rho^2 - (k-2)\rho - 1]$ and $b_k = \rho/[(k-1)\rho^2 - (k-2)\rho - 1]$ (see e.g. Rao, 1973). Hence for any i,

$$g_\rho^{(i|l\neq i)}(x_i|x_{lj,l\neq i}^*) \propto N\left(\psi_{\rho,a_{k-1}} \sum_{l\neq i} x_{lj}^* \ , \ 1 - \rho(k-1)\psi_{\rho,a_{k-1}}\right) 1_{(-L,L)}(x_i) \ , \tag{10}$$

where $\psi_{\rho,a_{k-1}} = 1 - (1-\rho)a_{k-1}$. Generation of the necessary samples may now proceed by a Gibbs sampling algorithm. That is, we successively sample from the complete conditional distributions in (10), updating the value of $\sum_{l\neq i} x_{lj}^*$ as we go. After a suitably large number of "burn-in" iterations N, the $\mathbf{x}_j^*$ values emerging from the sampler are approximately distributed according to their true joint distribution.

With regard to implementation, we first choose a starting value for $\sum_{l\neq 1} x_{lj}^*$, then run the substitution sampling chain for N "burn-in" iterations to essentially reach the chain's ergodic distribution, and finally continue for an additional B_2 iterations, now retaining the $\mathbf{x}_j^*$ values generated for use in a Monte Carlo approximant. Values obtained in this way will of course be serially correlated, and some authors thus recommend retaining only every M^{th} sample. Even for $M = 1$, however, the $\mathbf{x}_j^*$'s will still be from the correct ergodic distribution; taking B_2 large will also help to ameliorate this problem. Proper selection of N (ascertaining "convergence" of the sampler) is another important issue, and a source of much recent research interest (see for example Gelman and Rubin, 1992; Raftery and Lewis, 1992). In this case, our experience with normal sampling models convinced us that taking $N = 20$ would constitute ample burn-in.

Using our Gibbs-sampled $\tilde{\mathbf{x}}_j^*$'s, the Monte Carlo approximant to the log likelihood (9) for this model becomes

$$\begin{aligned}
&\tfrac{1}{2}\log|\Sigma^{-1}| - \tfrac{1}{2}\mathbf{x}'\Sigma^{-1}\mathbf{x} - \log\textstyle\sum_{j=1}^{B_2}\left[\frac{|\Sigma^{-1}|^{1/2}\exp(-\frac{1}{2}\mathbf{x}_j^{*\prime}\Sigma^{-1}\mathbf{x}_j^*)}{|\Sigma_0^{-1}|^{1/2}\exp(-\frac{1}{2}\mathbf{x}_j^{*\prime}\Sigma_0^{-1}\mathbf{x}_j^*)}\right] \\
&\quad\propto\ -\tfrac{1}{2}\mathbf{x}'\Sigma^{-1}\mathbf{x} - \log\textstyle\sum_{j=1}^{B_2}\exp[\frac{1}{2}\mathbf{x}_j^{*\prime}(\Sigma_0^{-1}-\Sigma^{-1})\mathbf{x}_j^*] \\
&\quad=\ -\tfrac{1}{2}\textstyle\sum_{i=1}^{k}\left[b_k x_i(\sum_{l=1}^{k}x_l) + (a_k - b_k)x_i^2\right] \\
&\qquad\qquad -\log\textstyle\sum_{j=1}^{B_2}\exp\left\{\frac{1}{2}\sum_{i=1}^{k}\left[(b_k^{(0)}-b_k)x_{ij}^*(\sum_{l=1}^{k}x_{lj}^*) + (a_k^{(0)}-a_k-b_k^{(0)}+b_k)x_{ij}^{*2}\right]\right\},
\end{aligned}$$

where a_k and b_k are defined as above, and $a_k^{(0)}$ and $b_k^{(0)}$ are defined similarly but depend on the fixed value ρ_0 instead of the unknown ρ. Note that our calculations feature two levels of iteration (Gibbs sampling within our MCMLE iterative framework).

We apply the above method to the following vector of $k = 10$ observations: $\mathbf{x}' = (-0.167, -0.934, 0.175, -0.349, -1.012, -0.378, -0.720, -1.208, -0.664, -1.435)$. This $\mathbf{x}$ was generated from the truncated correlated normal having $\rho = 0.5$ and $L = 2$. We used an initial guess of $\rho_0 = \rho = 0.5$, and a univariate maximization routine with maximum search window ± 0.1 (recall $-1/(k-1) < \rho < 1$ in order for Σ to be nonsingular). Running the MCMLE algorithm for $i = 6$ iterations (and using $B_2 = 10,000$ replications at iteration 6) we obtained the MLE $\hat{\rho} = 0.743$ and an associated approximate standard deviation (computed numerically using second differences) of 0.126. In this example, convergence is rapid even for poor initial choices of ρ_0.

4.2 Constrained 2×2 Table

One of the datasets presented by Andrews and Herzberg (1985) concerns species composition in a continuous, roughly semicircular arc of woodlands near Bradford, England. The data, collected by students at the University of Bradford, record counts of various kinds of trees at several sites within each of the woodlands. Table 1 gives the numbers of oak and sycamore trees in two such sites

from different woodlands (one in Royd's Cliffe and one in Dixon's Wood). However, this selection was not done completely arbitrarily: Royd's Cliffe sites having no oak trees were not eligible for selection.

	Oaks	Sycamores
Site 1 (Royd's Cliffe)	2	3
Site 2 (Dixon's Wood)	2	8

Table 1: Restricted multinomial data: Tree counts in two woodland sites

Assuming the observations in this table arise as independent Poisson variables, it is customary to condition on the total count n. This results in a multinomial $\mathcal{M}(n;\theta)$ model with $n = 15$ and $\theta = (q_{11}, q_{12}, q_{21}, q_{22})$, but under the restriction that $x_{11} > 0$. This is a special case of the class of restricted categorical data models considered in Subsection 2.2. Direct standardization of the likelihood via simple summation would be possible in this trivariate problem, but a very tedious accounting problem indeed. We shall use the Monte Carlo approach to find maximum likelihood estimates for the q_{ij}'s and the odds ratio $R = (q_{11}q_{22})/(q_{12}q_{21})$, and compare these results to those obtained presuming an unrestricted model.

The unstandardized likelihood for this model is

$$c_1(\mathrm{x};\theta) \propto q_{11}^{x_{11}} q_{12}^{x_{12}} q_{21}^{x_{21}} q_{22}^{n-x_{11}-x_{12}-x_{21}} \; 1_{\{1,\ldots,n\}}(x_{11}) 1_{\{0,\ldots,n\}}(x_{12}) 1_{\{0,\ldots,n\}}(x_{21}) \,,$$

so that the Monte Carlo log likelihood (9) becomes

$$x_{11}\log q_{11} + x_{12}\log q_{12} + x_{21}\log q_{21} + (n - x_{11} - x_{12} - x_{21})\log q_{22}$$
$$- \log \sum_{j=1}^{B_2} \left[\frac{q_{11}^{x^*_{11j}} q_{12}^{x^*_{12j}} q_{21}^{x^*_{21j}} q_{22}^{n-x^*_{11j}-x^*_{12j}-x^*_{21j}}}{q_{11,0}^{x^*_{11j}} q_{12,0}^{x^*_{12j}} q_{21,0}^{x^*_{21j}} q_{22,0}^{n-x^*_{11j}-x^*_{12j}-x^*_{21j}}} \right] ,$$

where the x^*_j values have been generated from our truncated multinomial model conditional on the

parameter values $q_{11,0}, q_{12,0}, q_{21,0}$, and $q_{22,0}$. Generating multinomial observations and rejecting those having $x_{11}^* = 0$ is a crude but effective way of drawing the $\mathbf{x}_j^*$.

Convergence of the MCMLE algorithm was again rapid: only $i = 3$ iterations were required to obtain the estimates $\hat{q}_{11} = 0.111, \hat{q}_{12} = 0.204$, and $\hat{q}_{21} = 0.137$, with associated standard deviations of 0.096, 0.108, and 0.091, respectively (again arising from a numerically computed Hessian). These results again used $B_2 = 10,000$ Monte Carlo replications at the final iteration. The fact that $\hat{q}_{11} < \hat{q}_{21}$ is intuitively plausible, since these two cells produced the same observed counts but, unlike x_{21}, x_{11} could not have been 0. The unrestricted MLE's in this case are of course $2/15 = 0.133$, $3/15 = 0.2$, and $2/15 = 0.133$, respectively. The discrepancy for the odds ratio R is more pronounced: while the raw sample odds ratio is 2.7, its MLE under the restricted model is only 2.2 (estimated standard deviation 3.0). The MCMLE algorithm's ability to produce results in such settings enables comparison of standard models with interesting but unwieldy truncated ones.

References

AITCHISON, J. (1986), *The Statistical Analysis of Compositional Data*, London: Chapman and Hall.

ANDREWS, D.F. and HERZBERG, A.M. (1985), *Data: A Collection of Problems From Many Fields for the Student and Research Worker*, New York: Springer-Verlag.

BARNDORFF-NEILSEN, O.E. (1978), *Information and Exponential Families in Statistical Theory*, New York: John Wiley and Sons.

BROWN, L.D. (1986), "Fundamentals of Statistical Exponential Families," Lecture Notes Monograph Series, Hayward, California: Institute of Mathematical Statistics.

DEMPSTER, A., LAIRD, N.M. and RUBIN, D.B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm" (with discussion), *J. Roy. Statist. Soc., Ser. B*, **39**, 1–38.

GELFAND, A.E. and SMITH, A.F.M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *J. Amer. Statist. Assoc.*, **85**, 398–409.

GELFAND, A.E., SMITH, A.F.M., and LEE, T-M. (1992), "Bayesian Analysis of Constrained Parameter and Truncated Data Problems Using Gibbs Sampling," *J. Amer. Statist. Assoc.*, **87**, 523–532.

GELMAN, A. and RUBIN, D.B. (1992), "A Single Series From the Gibbs Sampler Provides a False Sense of Security," in *Bayesian Statistics 4*, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, Oxford: University Press, pp. 625–631.

GEYER, C.J. and THOMPSON, E.A. (1992), "Constrained Monte Carlo Maximum Likelihood for Dependent Data" (with discussion), *J. Roy. Statist. Soc., Ser. B*, **54**, 657–699.

JACOBSEN, M. (1988), "Discrete Exponential Families: Deciding When the Maximum Likelihood Estimator Exists and is Unique," Preprint #1, Institute of Mathematical Statistics, Copenhagen.

MCFADDEN, D. (1989), "A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration," *Econometrica*, **57**, 995–1026.

PAKES, A. and POLLARD, D. (1989), "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, **57**, 1027–1057.

RAO, C.R. (1973), *Linear Statistical Models*, Second edition, New York: John Wiley and Sons.

RAFTERY, A.E. and LEWIS, S. (1992), "How Many Iterations in the Gibbs Sampler?" in *Bayesian*

Statistics 4, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, Oxford: University Press, pp. 763–773.

RUBIN, D.B. and SZATROWSKI, T.H. (1982), "Finding Maximum Likelihood Estimates for Patterned Covariance Matrices by the EM Algorithm," *Biometrika*, 69, 657–666.

TIERNEY, L. (1991), "Markov chains for exploring posterior distributions," Technical Report #560, School of Statistics, University of Minnesota.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
Technical Report No. 468

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
Maximum Likelihood Estimation for Constrained or Missing Data Models

5. TYPE OF REPORT & PERIOD COVERED
Technical

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Alan E. Gelfand and Bradley P. Carlin

8. CONTRACT OR GRANT NUMBER(s)
N0025-92-J-1264

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Department of Statistics
Stanford University
Stanford, CA 94305-4065

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
NR-042-267

11. CONTROLLING OFFICE NAME AND ADDRESS
Office of Naval Research
Statistics & Probability Program
Code 111

12. REPORT DATE
May 5, 1993

13. NUMBER OF PAGES
25

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES
THE VIEW, OPINIONS, AND/OR FINDINGS CONTAINED IN THIS REPORT ARE THOSE OF THE AUTHOR(S) AND SHOULD NOT BE CONSTRUED AS AN OFFICIAL DEPARTMENT OF THE ARMY POSITION, POLICY, OR DECISION, UNLESS SO DESIGNATED BY OTHER DOCUMENTATION.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
EM algorithm; Gibbs sampler; Monte Carlo approximant.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)
See Reverse Side

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Abstract

In statistical models involving constrained or missing data, likelihoods containing integrals emerge. In the case of both constrained and missing data, the result is a ratio of integrals, which for multivariate data may defy exact or approximate analytic expression. Seeking maximum likelihood estimates in such settings, we propose Monte Carlo approximants for these integrals, and subsequently maximize the resulting approximate likelihood. Iteration of this strategy expedites the maximization, while the Gibbs sampler is useful for the required Monte Carlo generation. As a result, we handle a class of models broader than the customary EM setting without using an EM-type algorithm. Implementation of the methodology is illustrated in two numerical examples.