The Bayesian Analysis of Truncated Regression Models

S. Barry
T.J. O'Neill
Department of Statistics
The Faculties, Australian National University

September 30, 1994

Abstract

The analysis of truncated data when an unobserved latent structure is assumed is considered. A Bayesian analysis of truncated data in the presence of covariates is presented. A method is derived for the analysis of truncated data when the response for the unobserved members is known. The Gibbs sampler is used to approximate the required posteriors. The method is applied to some simulated data and to a real data set.[1]

[1]**Acknowledgment**
This research was supported by a seeding grant from the Federal Office of Road Safety, Australian Department of Transport.

1 Introduction

This paper examines the regression analysis of data from a truncated Poisson distribution, where the response is truncated at zero. An example of such data is the data on fishing trips presented in Grogger & Carson (1991). This data set is based on a sample of households and records the number of fishing trips undertaken as well as a range of covariates. Inclusion in the sample is based on at least one fishing trip being undertaken, and sample selection was not affected by the number of trips taken greater than 1. This is an example of truncation where the response of the unobserved members is known, in this case zero. This can be contrasted with truncated Gaussian regression, where the value of the unobserved responses is unknown, only that observations were observed outside a certain level. The results in this paper are developed for the analysis of truncated data where the response is known for the unobserved members of the sample. This applies when only one state is unobserved. An example of this is grouped truncated binary data detailed in O'Neill & Barry (1993a). Group truncated binary data occurs when groups of binary outcomes are observed only if at least one of the responses is positive. An example of this is data on car accidents involving fatalities. In this case accidents where no fatalities result are not observed in the data set. The results of this paper can also be extended to more general forms of truncation.

Truncated data arises when the range of possible responses is restricted in some way. Methods for the regression modelling of such data have been proposed for a number of situations. For example, forms of the Tobit model (Amemiya, 1979) deal with the regression modelling of truncated Gaussian responses. There is discussion of these models in the econometrics literature. More recently, Grogger & Carson (1991) and Shaw (1988) have detailed a regression model for data with a truncated Poisson distribution, and extended it to the truncated Negative Binomial distribution. O'Neill & Barry (1993a) have recently proposed a truncated model for grouped binary data which O'Neill & Barry (1993b) extend to grouped ordinal data. Weiss (1993) has proposed a truncated model for correlated ordinal data.

A distinction should be made between data that is *observationally* truncated and data which is *distributionally* truncated. Observationally truncated data refers to data where it is natural to view the observations as being a truncated sample from a larger, unobserved, sample. An example is the group truncated data in O'Neill & Barry (1993a), or the fishing data described above. In this case questions regarding the size of this unobserved sample and the covariate distribution of the unobserved sample may be of interest. Alternately, distributionally truncated data is data where the response is truncated in some manner which does not rely on the observational mechanism for its interpretation. An example of this is data with the response taking the positive integers as values. This data can be treated as either a sample from

an unknown distribution with support the positive integers, or a truncated sample from the Poisson distribution. The mechanism which gave rise to this support is not assumed. In this case the truncation is used as a device to allow estimation within a familiar (Poisson) framework. This paper is derived assuming observationally truncated data but is applicable to both forms of truncation. This is because the analysis techniques used in the distributionally truncated case are often used to fit models to observationally truncated data.

The regression modelling of data with a Poisson distribution is routinely performed in the literature as a generalised linear model (McCullagh & Nelder, 1989). The analysis of truncated count data has previously been based on maximum likelihood methods. These all consider the distribution of observations conditional on being observed. This paper examines the use of Bayesian methods in the analysis of truncated Poisson regression, and examines some of the novel aspects of the approach. Section 2 develops the likelihood for the truncated model. In Section 3 the priors are discussed, posteriers are derived, and a Gibbs sampling algorithm to explore them is presented. Section 4 presents some numerical examples and the method is discussed in Section 5.

2 Bayesian Model

As the results in this section will be developed with respect to the Poisson model, it will be briefly reviewed. Following McCullagh & Nelder (1989), consider a sample, $(Y_1, x_1) \ldots (Y_N, x_N)$ where Y_i is the response for the ith individual and x_i are covariates measured on the ith individual. It is assumed that Y_i follows a Poisson distribution

$$Pr(Y_i = y) = \frac{e^{-\mu_i}\mu_i^y}{y!}; \quad y = 0, 1, 2, \ldots$$

where μ_i is the mean of the process. If the log link is used the mean is modelled by

$$g(\mu) = \log \mu_i = x_i^{'}\beta$$

where β is a vector of parameters and $g()$ is the link function.

We now consider the case truncated at zero. We will begin by considering the likelihood of obtaining a particular truncated sample. Consider a sample of size N from the non truncated distribution. For clarity we will use the notation found in Gelfand, Smith, & Lee (1992) for the densities. This notation denotes the density of a random variable K as $[K]$ instead of the usual functional notation $f(k)$.

Assume that the covariates x are realisations of the random variables X with density $[X]$, independent of the response. Consider the sample with n non truncated

observations. By permutation if necessary this is

$$(Y_1^*, x_1^*), \ldots, (Y_n^*, x_n), (0, x_{n+1}), \ldots, (0, x_N)$$

where n is the sample size of the truncated sample. The likelihood of the complete sample is then

$$[Y, X, n|\beta, N] = \left(\begin{array}{c} N \\ n \end{array} \right) \prod_{i=1}^{n} [Y_i|X_i, \beta][X_i] \prod_{i=n+1}^{N} [Y = 0|X_i, \beta][X_i].$$

and the posterior is thus

$$[\beta|Y, X] \propto [Y, X|\beta][\beta].$$

This is of no immediate use, as the covariate values for the truncated observations are unknown. To avoid this problem we calculate the marginal distribution of n and $(Y_1^*, x_1^*) \ldots (Y_n^*, x_n^*)$. This is

$$[Y^*, X^*, n|\beta, N] = \left(\begin{array}{c} N \\ n \end{array} \right) \prod_{i=1}^{n} [Y_i|X_i, \beta][X_i][\beta] \int \cdots \int \prod_{i=n+1}^{N} [Y = 0|X_i, \beta][X_i] \prod_{i=n+1}^{N} dx_i.$$

As the distribution of the X's are identical this reduces to

$$[Y^*, X^*, n|\beta, N] = \left(\begin{array}{c} N \\ n \end{array} \right) \prod_{i=1}^{n} [Y_i|X_i, \beta][X_i][\beta]\{1 - P(\beta)\}^{N-n} \tag{1}$$

where

$$P(\beta) = 1 - \int [Y = 0|X, \beta][X]dx, \tag{2}$$

the unconditional probability of observing a unit randomly chosen from $[X]$.

If N and $[X]$ are known Equation 1 could be used to form inferences about β, although this could prove problematic due to the (possibly) multi dimensional integral in Equation 2.

With truncated data, it is not usual for N to be known. In this case we assume that N is a random variable and postulate a distribution for N. We will denote this by $[N|\lambda]$ where λ is a fixed parameter. The joint distribution of $[Y^*, X^*, n, N|\beta, \lambda]$ is then

$$[Y^*, X^*, n, N|\beta, \lambda] = \left(\begin{array}{c} N \\ n \end{array} \right) \prod_{i=1}^{n} [Y_i|X_i, \beta][X_i]\{1 - P(\beta)\}^{N-n}[N|\lambda] \tag{3}$$

and the marginal for $[Y^*, X^*, n|\beta, \lambda]$ is

$$\sum_{j=n}^{\infty} \left(\begin{array}{c} j \\ n \end{array} \right) \prod_{i=1}^{n} [Y_i|X_i, \beta][X_i]\{1 - P(\beta)\}^{j-n} Pr(N = j|\lambda). \tag{4}$$

The choice of $[N|\lambda]$ will depend on the particular problem being considered. For example if it is assumed that N has a Poisson distribution with mean λ then Equation 4 simplifies to

$$[Y^*, X^*, n|\beta, \lambda] \propto \prod_{i=1}^{n} [Y_i|X_i, \beta][X_i]\lambda^n e^{-\lambda P(\beta)}.$$

The last issue that must be resolved is the distribution of the x_i's. It is possibly multi dimensional and with no knowledge of the distribution, the specification of a form that is flexible enough is difficult. Arguing that the observed covariates provide the only information regarding the distribution of X we propose to approximate the covariate density by using the "empirical" support of the observed covariates. By this we mean that we shall let the support of the distribution be defined by the observed covariates. Thus the empirical distribution is a multinomial distribution. We use the parameter η in this case to represent the vector of length n with components the cell probabilities of the multinomial distribution. This is loosely analogous to the use of the multinomial simplification to derive empirical likelihoods from the intractable non parametric likelihoods (see Efron & Tibshirani (1993)).

The density is then

$$[Y^*, X^*, n|\beta, \lambda, \eta] \propto \prod_{i=1}^{n} [Y_i|X_i, \beta][X_i|\eta]e^{-\lambda P(\beta|\eta)} \tag{5}$$

where $P(\beta|\eta)$ is $1 - \sum_{i=1}^{n} [Y = 0|x_i, \beta]\eta_i$.

3 Priors and Posteriors

3.1 Prior Specification

3.1.1 β, λ priors

Where there is no prior information we propose using a uniform prior for λ and β. These priors are attractive as they are dominated by the likelihoods, and produce estimates which maximise the likelihood function in Equation 5. We will now consider the prior for η.

3.1.2 η prior

The specification of the prior is complicated by the truncation. It is important to distinguish between priors at the untruncated and truncated levels of the likelihood. The correspondence between a prior $[\beta, X]$ in the untruncated space and the induced

prior $[\beta, X]_T$ in the truncated space is

$$[\beta, X]_T = \frac{[\beta, X]P(\beta, X)}{\iint [\beta, X]P(\beta, X)d\beta dX}. \tag{6}$$

So for example in situations where β and X are assumed to be apriori independent and the non-informative priors $[\beta]$ and $[X]$ are constants, then the non-informative truncated prior will be

$$[\beta, X]_T = \frac{P(\beta, X)}{\iint P(\beta, X)d\beta dX}$$

which is not of the usual form. Note that Equation 6 can be inverted to give

$$[\beta, X] = \frac{[\beta, X]_T P(\beta, X)^{-1}}{\iint [\beta, X]_T P(\beta, X)^{-1} d\beta dX}. \tag{7}$$

Equation 7 suggests that priors in the truncated situation should be tilted by $P(\beta, X)^{-1}$ to obtain the corresponding prior in the untruncated situation. Note that if β and X are assumed to be apriori independent in the untruncated situation, then $[\beta, X] = [\beta][X]$ and from Equation 7,

$$[X] = \frac{[X \mid \beta]_T [\beta]_T [\beta]^{-1} P(\beta, X)^{-1}}{\iint [\beta, X]_T P(\beta, X)^{-1} d\beta dX},$$

or

$$[X] = \frac{[X \mid \beta]_T P(\beta, X)^{-1}}{\int [X \mid \beta]_T P(\beta, X)^{-1} dX}. \tag{8}$$

So the prior $[X]$ is equal to the prior $[X \mid \beta]_T$ tilted by $P(\beta, X)^{-1}$.

In the case where X is known to be discrete and have a finite set of values, the prior $[X]$ can be specified with support the observed truncated x. Since ultimately all of the possible values of x will appear in the truncated sample, this will ultimately be equivalent to knowing the support of $[X]$ apriori and specifying the prior on that support. The nonparametric approach would assume that $[X \mid \beta]_T$ is Multinomial with support the observed x and parameters η which have a Dirichlet distribution. If a Multinomial has k points, then a non-informative self-consistent prior for the parameters of the Multinomial is Dirichlet$(2/k, 2/k, \ldots, 2/k)$. Self consistent means that when the k points are aggregated in k_0 equal sized sets, the marginal prior for the k_0 equal sized sets obtained from the n dimensional prior is Dirichlet$(2/k_0, 2/k_0, \ldots, 2/k_0)$. Note that for $k = 1$, Dirichlet$(1, 1)$ is $U(0, 1)$ which is the non-informative prior for two points. Equation 8 suggests that the location of the truncated prior for X should be tilted by $P(\beta, X)^{-1}$. This can be achieved in the non parametric setting by taking the 'non-informative' Empirical Dirichlet Prior (EDP) to be

$$Dirichlet\,(\gamma_i, i = 1, \ldots, n)$$

where

$$\gamma_i = \frac{2P(\beta, x_i)^{-1}}{\sum_j P(\beta, x_j)^{-1}}.$$

It is worth examining in detail the outcome if EDP is applied to a situation where X is actually discrete. For simplicity of presentation we will study the case where X has only two values, a and b with probabilities π_0 and π_1 respectively. Using the knowledge that X is discrete and assuming a uniform prior for π_1, we would obtain posteriors

$$[\beta \mid \lambda, \pi_1, Y^*, X^*, n] \propto [Y^* \mid X^*] \exp(-\lambda P(\beta \mid \pi_1))[\beta], \tag{9}$$

$$[\lambda \mid \beta, \pi_1, Y^*, X^*, n] \propto \lambda^n \exp(-\lambda P(\beta \mid \pi_1))[\lambda], \tag{10}$$

$$[\pi_1 \mid \lambda, \beta, Y^*, X^*, n] \propto \exp(-\lambda P(\beta \mid \pi_1))\pi_1^{n_1}\pi_0^{n_0}, \tag{11}$$

where

$$P(\beta \mid \pi_1) = \pi_0 P(\beta, a) + \pi_1 P(\beta, b),$$

n_1 is the number of b observations in the sample and $n_0 = n - n_1$. By contrast, if EDP is used and we define

$$\pi_1 = \sum_{b \text{ observations}} \eta_i$$

and $\pi_0 = 1 - \pi_1$, then the posteriors for β and λ are once again given by Equations 9 and 10. Assuming without loss of generality that the b observations are labelled $1, \ldots, n_1$ and letting

$$y_i = \begin{cases} \eta_i/\pi_1 & , \quad i = 1, \ldots, n_1 \\ \eta_i/\pi_0 & , \quad i = n_1 + 1, \ldots, n \end{cases},$$

the posterior of $\pi_1, y_1, \ldots, y_{n_1-1}, y_{n_1+1}, \ldots, y_{n-1}$ is

$$\begin{aligned} [\pi_1, y_1, \ldots, y_{n_1-1}, y_{n_1+1}, \ldots, y_{n-1} \mid \lambda, \beta, Y^*, X^*, n] \quad &\propto \quad \exp(-\lambda P(\beta \mid \pi_1)) \\ &\times \quad \pi_1^{n_1+r_1-1}\pi_0^{n_0+r_0-1}\prod_1^n y_i^{c_i} \end{aligned}$$

where

$$r_1 = \frac{2n_1 P(\beta, b)^{-1}}{n_1 P(\beta, b)^{-1} + n_0 P(\beta, a)^{-1}},$$

$$r_0 = \frac{2n_0 P(\beta, a)^{-1}}{n_1 P(\beta, b)^{-1} + n_0 P(\beta, a)^{-1}}$$

and

$$c_i = \frac{2P(\beta, x_i)^{-1}}{n_1 P(\beta, b)^{-1} + n_0 P(\beta, a)^{-1}}.$$

So π_1 and $y_1, \ldots, y_{n_1-1}, y_{n_1+1}, \ldots, y_{n-1}$ are independent in the posterior and the posterior of π_1 is

$$[\pi_1 \mid \lambda, \beta, Y^*, X^*, n] \propto \exp(-\lambda P(\beta \mid \pi_1))\pi_1^{n_1+r_1-1}\pi_0^{n_0+r_0-1}. \tag{12}$$

Note that $y_1, \ldots, y_{n_1-1}, y_{n_1+1}, \ldots, y_{n-1}$ do not appear in the posteriors for β and λ and so can be ignored unless specifically required. Consequently the only difference between the posteriors are the coefficients of π_1 and π_0 in Equations 11 and 12. Since the difference in the coefficients is at most one and converges to zero as the sample size increases, the difference in the posteriors will be negligible as the sample size increases and EDP will be essentially equivalent to the known two point case. An identical argument applies to establish the essential equivalence of EDP and k point support for X. Finally since any continuous density for X can be arbitrarily well approximated by a discrete k point support density with k sufficiently large, it follows that any non truncated prior for X can be arbitrarily well approximated by EDP for sufficiently large sample sizes.

With these priors the posterior becomes,

$$[\beta, \lambda, \eta | Y^*, X^*, n] \propto \prod_{i=1}^{n} \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \prod_{i=1}^{n} \eta_i e^{-\lambda P(\beta|\eta)} \prod_{i=1}^{n} \eta_i^{\gamma_i - 1}. \tag{13}$$

3.2 Gibbs Sampling Algorithm

To produce inference about β, λ and η requires the numerical integration of the form given in Equation 13 to obtain the required posteriors. This is computationally prohibitive in all but the most simple cases.

Instead we propose using a Gibbs sampling algorithm to approximate the posterior 13. The Gibbs sampler is described in numerous articles, for example see Tanner (1993). The implementation of the Gibbs sampler requires the decomposition of Equation 13 into a set of convenient conditional distributions. We have chosen the following.

$$[\beta | \lambda, \eta, X^*, Y^*, n] \propto \prod_{i=1}^{n} \frac{e^{-\mu}\mu^{y_i}}{y_i!} \prod_{i=1}^{n} \eta_i \lambda^n e^{-\lambda P(\beta)} \prod_{i=1}^{n} \eta_i^{\gamma_i - 1} \tag{14}$$

$$[\lambda | \beta, \eta, Y^*, X^*, n] \propto \lambda^n e^{-\lambda P(\beta)} \tag{15}$$

$$[\eta | \beta, \lambda, Y^*, X^*, n] \propto \prod_{i=1}^{n} \eta_i \lambda^n e^{-\lambda P(\beta)} \prod_{i=1}^{n} \eta_i^{\gamma_i - 1} \tag{16}$$

To sample from Equation 14, note that it consists of the usual likelihood used in the non truncated analysis, but tilted by the remaining terms. Following Zeger

& Karim (1991) we sample from Equation 14 by rejection sampling. This is done by maximising the function for β, and finding the curvature at the maximum point. A normal density with deflated curvature is then used as the rejection envelope. Although we have not proved the log concavity of Equation 14, no problems with incorrect acceptances have arisen in a range of simulations carried out. Heuristic arguments suggest it should not present difficulties. The choice of the deflation factor depends on the problem, but no difficulties were encountered in choosing a value that achieved acceptable rejection rates, while still being a valid envelope.

Examining Equation 15 it is immediately apparent that it is proportional to a Gamma density with shape and location parameters, n and $P(\beta|\eta)$ respectively. It is thus simple to generate deviates from this distribution. Of course in some applications it is of interest to make inference regarding N. This can be done by noting that the Gibbs sampling algorithm produces approximate samples from the marginal distribution $[\beta, \lambda, \eta | Y^*, X^*, n]$. Thus points sufficiently separated in the sequence can be treated as independent realizations and the conditional density for N, $[N|\beta, \lambda, \eta, Y^*, X^*, n]$ can be derived from Equation 3 and used to produce a sample from the full posterior.

Sampling from Equation 16 presents greater difficulties. One approach is to consider the reduced vector $[\eta_1, \cdots, \eta_{n-1}]$. Observing that the conditional distribution of η_i, given the other η's is

$$[\eta_i|\beta, \lambda, n, \eta_{-i}, y^*, x^*] \propto e^{-d_i\eta_i}\eta_i^{\gamma_i}\left(1 - \sum_{j \neq n} \eta_j\right)^{\gamma_n} \tag{17}$$

where

$$d_i = \lambda P(\beta, x_i) - \lambda P(\beta, x_n), \eta_{-i} = \eta_1, \ldots \eta_{i-1}, \eta_{i+1}, \ldots, \eta_{n-1}$$

and

$$0 \leq \eta_i \leq 1 - \sum_{j \neq i,n} \eta_j$$

It is therefore possible to use a rejection sampling algorithm to sample η_i. This is done using a uniform distribution over $[0, 1 - \sum_{j \neq i,n} \eta_j]$ scaled to the maximum of Equation 17. This maximum is the solution of a quadratic equation and is thus easily found. To produce a sample from the full vector a Gibbs sampling sub-chain can be used, by repeating this process until approximate convergence is reached.

A concern with this algorithm is that convergence may be slow due to the dependence between the components of η. An alternative is to consider a rejection sampling algorithm to simulate from the full η vector. This can be used either to asses the convergence of the Gibbs approximation, or if efficient enough, to directly sample η. This is detailed in the appendix. The conclusion was that the Gibbs algorithm appeared to give good approximation to a sample from $[\eta|conditional]$ if the following conditions

were met. Firstly, if it was run for n full iterations. Secondly if, by permutation if necessary, η_n was set to the η with the largest expectation. This second condition allows the easiest traversal of the simplex and thus speeds convergence.

4 Examples

4.1 Example 1

As a simple example consider a regression model with a single intercept term. In this case X is 1 with probability 1. Thus the posterior 13 reduces to

$$[\beta, \lambda, \eta | n, y^*, x^*] \propto \prod_{i=1}^{n} \frac{e^{-\mu}\mu^{y_i}}{y_i!} \prod_{i=1}^{n} \eta_i \lambda^n e^{-\lambda P(\beta)} \prod_{i=1}^{n} \eta_i^{\gamma_i - 1} \tag{18}$$

For any η this function is maximised (and hence is the posterior mode) by $\lambda = n/P(\beta)$ and the solution of

$$\frac{\sum_n y_i}{n} = \frac{\mu}{P(\beta)} \tag{19}$$

which is the same as the maximum likelihood estimate based on the conditional model.

4.2 Example 2

In this section a small scale simulation experiment is presented. The Gibbs sampling algorithm was implemented using Splus, with C code used to gain efficiency in the sampling from the η vector. The simulations consisted of the following steps:

1. A covariate matrix was generated with 100 rows and ith row

   $$(1, x_{i1}, x_{i2}])$$

   where x_{i1} had a uniform distribution over [-1,1] and x_{i2} was Bernoulli with probability .5. From this the expected values, μ were generated via the link function $g()$ and the linear predictor $X\beta$, where $\beta = (\beta_0, \beta_1, \beta_2) = (1, -2, 2)$. In the simulations the marginal probability of being observed was .598, so each data set had approximately 60 observations.

2. The following steps were iterated.

   - A sample y was generated from μ.
   - The sample was truncated to form a sample y^*, x^*.

- The Gibbs algorithm was run for 2000 iterations, and the sample path saved.

The use of simulation to assess the behaviour of a Bayesian procedure is worthy of comment. The Bayesian approach does not require the frequentist justification over repeated samples. The Bayesian could simulate from their priors and then the densities, but this would be purely an exercise in verifying the calculus. The simulations carried out here are justified in the following ways. Firstly, the use of the uniform priors for β and λ means that the posterior is approximating the likelihood function and so the posterior modes are interpretable as maximum likelihood estimates. Secondly, the use of the procedure with the Empirical Dirichlet prior in any practical problem requires confidence in its behaviour.

Each iteration of the Gibbs sampler consisted of drawing samples directly from the conditional densities for η and λ and using the Gibbs subchain described in Section 3.2 to gain an approximate sample from the η's. The number of iterations used in the Gibbs subchain was equal to the length of the observed y^*.

The Gibbs sampler was started by using the known true parameter values. It was run for 50 iterations to eliminate the effect of the initial conditions. In the preliminary investigations the posterior densities appeared unimodal and well behaved so it was considered unnecessary to use a sequence of different starting values. The simulations have confirmed this view.

For each run, the posterior mean, mode and selected quantiles were estimated for each of the λ and β marginals. The mode was estimated using kernel smoothing. The mean exhibited considerable bias in estimating the mode due to the skewed posterior distributions. This was most pronounced with respect to λ, which posessed a long right tail. For the η's the calculation of the mode could only sensibly be done by considering the joint mode. As the dimension of this is ≈ 60 it was not undertaken. The results of the simulations for the regression coefficients are presented in Figure 1. This figure shows the distribution of the estimated modes from the simulations. For comparison the distribution of the maximum likelihood estimates (MLE) based on the true truncated model and those produced by the Poisson model, ignoring the truncation, are also presented. These were produced from the same data sets generated during the simulation. Examining the figure it is seen that the estimate of the intercept term exhibits a slight negative bias over the simulations, while the estimate for the continuous covariate has an small positive bias. The exact extent of these biases would require extensive simulation. This is currently not computationally feasible. The outlying terms in these plots are produced by simulations were the response for units with $x_{i2} = 1$ were all 1. In this case β_2 can equally plausibly have a range of negative values (implying heavy to severe truncation), and this is reflected in the Gibbs sampler, which drifts over this parameter. This also causes λ to inflate

Table 1: Table 1. Estimated coverage probabilities.

parameter	β_0	β_1	β_2	λ
coverage	.93	.92	.86	.86

as there is no information regarding the extent of the truncation.

Note that the comparison between the Poisson MLE, ignoring the truncation, and the other estimates is not simple. This is due to the Poisson MLE attempting to model the mean of the observed responses, wheras the other techniques attempt to model the mean of the underlying process. Thus they are attempting to estimate different quantities. With this in mind, it is still informative to see the effect of the tilting in Equation 5 on the usual Poisson likelihood. In addition it also shows the serious errors in location and precision that can occur if the truncation is ignored.

The distribution of the estimated mode of the marginal posterior for λ is presented in Figure 2.

Table 1 estimates the coverage of the 90% Bayesian intervals calculated from the estimated 5% and 95% quantiles. Note that the coverage for λ is not a true coverage as the method in fact estimates the mean of an assumed underlying process. Although in this case the process was deterministic, the result is still interpretable.

4.3 Example 3

The Gibbs sampling algorithm was used to fit the Poisson model to some truncated count data on the abundance of Leadbeaters Possom, an Australian marsupial. This data consists of counts of possums and habitat variables from a sample of sites. This data has been analysed in the context of zero inflation. With zero inflation the number of zero responses is greater than would be expected from the Poisson model. Welsh, Cunningham, Donnelly, & Lindenmeyer (1994) analyse this data by assuming that the data is contaminated by zero responses. They proceed by modelling the probability of a positive response. For the positive responses they modelled the effect of the covariates on the response, conditional on a positive response, via the truncated model.

The Gibbs sampling algorithm allows for the fitting of the truncated model, where we may consider λ and the η's as auxillary variables. In this case the technique replaces maximum likelihood for the estimation of the regression effects. Alternatively we can consider the following approach. In this analysis we assume that there is an unobserved covariate with two levels. If the covariate is at the first level then the habitat is unsuitable for Leadbeaters Possums and none will be found. If the covariate

Table 2: Table 2. Parameter estimates and standard errors for possum data.

variable	estimated posterior mode	est. variance	Truncated MLE	est. variance
Intercept	1.08	.17	1.13	.11
log(stags+1)	.247	.019	.246	.012
bark	.040	.0003	.037	.0002
no.s	-.08	.001	-.09	.0008
slope	-.036	.0002	-.031	.00016

is at the second level the habitat is suitable and the density of the possums follows a Poisson distribution with mean depending on the habitat variables via the log link. The analysis will then enable inference to be made regarding the number of suitable sites in the sample.

The variables chosen were the same as those used by Welsh et al. (1994). These were chosen by the standard analysis of deviance used in fitting generalised linear models.

The Gibbs sampler was run for 20000 iterations, after an initial 100 iterations to limit the effect of the initial conditions. As in the previous example a Gibbs subchain was used to sample η. The initial conditions were naive estimates produced from the maximum likelihood estimates. Other starting points were tried. These had no effect on the results, with the exception of initially upsetting some of the tuning in the rejection sampling algorithm.

The results from the analysis are presented in Table 2, along with the maximum likelihood esimates. The approximate marginal posterior distribution for N is given in Figure 3. This was generated by taking samples at intervals of 50 from the Gibbs sequence and drawing from the conditional distribution of N. Note that the estimated mode is approximately 60 and thus the intensity of truncation is low.

5 Discussion

The techniques presented in this paper provide a novel approach to the estimation of regression coefficients with truncated data. They also allow for inference to be drawn regarding the truncation process. This is an extension of the usual conditional approach which considers the likelihood of the response *within* the observed sample. The results of the simulation study and example provide encouragement about the stability and interpretability of the estimators.

When the data are distributionally truncated the new method offers no significant advantage over the usual maximum likelihood estimates. This is because it is only the regression coefficients that are of interest, and the truncation is only used to produce a distribution consistent with the data. Alternately, for observationally truncated data, the Gibbs model provides a method for infering the observational process. This can be important in attempting to estimate size and distribution of covariates across the unobserved population. This is sometimes a side interest, but can also be the primary focus of the study, as in capture-recapture studies which are a form of truncated data. The Bayesian approach has considerable advantages over the frequentist analysis, as the missing components $[X]$ and N can be incorporated. In the frequentist analysis they cause major difficulties, and the estimation of N is non regular.

Over recent years, Gibbs sampling has gained wide exposure for its use in obtaining samples from posteriors such as Equation 13, thus allowing the construction of approximate marginal and joint posteriors for the parameters. In the case of missing data, the Gibbs sampling algorithm is made more attractive due to the simplifications that occur when the data is augmented by the unobserved components (Smith & Roberts, 1993). In censored regression, for example, the censored observations are included as additional parameters in the model.

It is not apparent that the simplifications found in the censored case are realised in the truncated case. Note that Equation 1 is analagous to the likelihood used in the case of censored regression. It differs in this case due to the covariates being unknown but the response known. In the censored case the data is augmented by the unknown responses, considerably simplifying the Gibbs sampling. In the truncated case it is appealing to attempt to augment the data to seek these simplifications. The obvious candidate is to augment the data with the unknown covariates and to thus avoid the integration performed in section 2. This augmentation is complicated by the unknown λ and random population size, and will not produce the simplifications in the Gibbs sampling found in the censored case. This will potentially invalidate the use of the Gibbs sampling algorithm.

The method is facilitated by the use of the empirical distribution of the covariates. Research is still required into the effect of this approximation on the procedure. It is obvious that in very sparse data sets the empirical distribution may provide a poor approximation to the true distribution. Whether there are any deeper pathological problems is unknown, although various unpublished simulations have not highlighted any. A second issue relates to how well the Dirichlet specification reflects the uncertainty in η.

The method has application to other truncation problems such as group truncated ordinal data (O'Neill & Barry (1993a),O'Neill & Barry (1993b)). and data from multiple capture/recapture experiments. It can be potentially extended to include random effects. It could also be used in sytems with more general truncation patterns by re-

placing the term $[Y = 0|x, \beta]$ with the appropriate integral. The sampling algorithms should continue to hold providing the likelihood is approximately quadratic, which is not a particularly strong assumption.

Reference

Amemiya, T. (1979). *Advanced Econometrics.* Chapman and Hall, New York.

Efron, B., & Tibshirani, R. (1993). *An Introduction to the Bootstrap.* Chapman and Hall, New York.

Gelfand, A., Smith, A., & Lee, T. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association, 87,* 523–532.

Grogger, J. T., & Carson, R. (1991). Models for truncated counts. *Journal of Econometrics, 6,* 225–238.

McCullagh, P., & Nelder, J. (1989). *Generalized Linear Models.* Chapman and Hall, New York.

O'Neill, T. J., & Barry, S. C. (1993a). Truncated logistic regression. *To appear in Biometrics.*

O'Neill, T. J., & Barry, S. C. (1993b). Truncated ordinal regression. *To appear in Statistics and Probability Letters.*

Shaw, D. (1988). On site samples' regression: Problems of non negative integers, truncation, and exogenous stratification. *Journal of Econometrics, 37,* 211–223.

Smith, A., & Roberts, O. R. (1993). Bayesian computation via the Gibbs sampler and related Markov Chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B, Methodological, 55,* 3–23.

Tanner, M. (1993). *Tools for Statistical Inference.* Springer-Verlag, New York.

Weiss, A. A. (1993). A bivariate ordered probit model with truncation: Helmet use and motorcycle injuries. *Applied Statistics, 42,* 487–499.

Welsh, A. H., Cunningham, R., Donnelly, C., & Lindenmeyer, D. (1994). Modelling the abundance of rare species: Statistical models for counts with extra zeroes. *submitted.*

Zeger, S., & Karim, M. R. (1991). Generalized linear models with random effects:a Gibbs sampling approach. *Journal of the American Statistical Association, 86*, 79–86.

A Rejection Sampling from Exponential Tilted Dirichlet Random Variables

We consider the problem of generating random variables from the density,

$$f(\eta) \propto \prod_{i=1}^{n} \eta_i^{\alpha_i} \exp(-c_i \eta_i), \sum_{i=1}^{n} \eta_i = 1, c_i \geq 1. \tag{20}$$

Since $\sum_{i=1}^{n} \eta_i = 1$ we may take in equation 20

$$f(\eta) \propto \prod_{i=1}^{n} \eta_i^{\alpha_i} \exp(-d_i \eta_i) \tag{21}$$

where

$$d_i = c_i - c_{\min},$$

and

$$c_{\min} = \text{minimum}(c_1, \ldots, c_n).$$

Without loss of generality, we may assume by permutation if necessary, that $d_n = 0$ in equation 21. Then letting x be the solution of

$$x \sum_{i=1}^{n} \frac{\alpha_i}{(\alpha_n + d_i x)^{-1}} = 1, \tag{22}$$

it follows that $\prod_{i=1}^{n} \eta_i^{\alpha_i} \exp(-d_i \eta_i)$ is a strongly unimodal function with maximum at

$$v = \left(\frac{\alpha_i x}{\alpha_n + d_i x} \right)_{i=1,\ldots,n}. \tag{23}$$

We wish to use a combination of rejection sampling and Gibbs sampling to generate observations from equation 20. To do this we require the curvature of f at v. But since v is the location of the maximum of f, the curvature of f at v is $f(v)$ by the curvature of $\log f$ at v. So

$$\left. \frac{\partial^2 f}{\partial \eta \partial \eta'} \right|_v = f(v) \left(\left. \frac{\partial^2 \log f}{\partial \eta \partial \eta'} \right|_v \right)$$

where we are differentiating with respect to

$$\eta' = (\eta_1, \ldots, \eta_{n-1}).$$

Now

$$-\left.\frac{\partial^2 f}{\partial \eta \partial \eta'}\right|_v = \text{diag}\left(\frac{1}{v_i^2}\right) + \frac{1}{v_n^2} E. \tag{24}$$

There appear to be two main options for approximating f.

The first alternative is to use an approximating Gaussian distribution for the density of $\eta' = (\eta_1, \ldots, \eta_{n-1})$. This method has the advantage of being able to exactly match the curvature of density 20 at v and therefore potentially lower the rate of rejection. The drawback is that unlike the tilted Dirichlet where the observations are constrained to lie in the simplex, the Gaussian random variables can have individual entries which are less than zero or they can sum to greater than one. Either of these will cause the random vector to be rejected. However this will not become a serious problem until $n \approx 10$. The problem with this method is that the ratios obtained can be greater than one and so the rejection sampling is not valid. A similar approach using suitably chosen independent Beta random variables was also tried but suffered from similar problems.

The other alternative is to try to approximate the tilted Dirichlet by an ordinary Dirichlet. In order to get the maximum to occur at the same location, it is necessary to take a Dirichlet with density $\propto \prod_{i=1}^n \eta_i^{c_0 v_i}$ where c_0 is a constant that should be chosen to get the curvatures to match as closely as possible at v. The best possible choice turns out to be $c_0 = 1/x$. Unfortunately when this Dirichlet is used in a rejection sampling scheme, the rate of rejection can be so large that it makes the method unusable. This situation occurs when some of the d_i are very large, say fifty or more. This can happen for $n = 2$.

In order to overcome this difficulty, the n units are split into two groups, those with d_i less than some value, c say, and those with d_i greater than c. Without loss of generality we will assume that $d_n = 0$, $d_i \leq c, i = k+1, \ldots, n$ and $d_i > c, i = 1, \ldots, k$. Then since $1 - x \leq \exp(-x)$ for $x \in (0, 1)$ it follows that

$$\begin{aligned}
\prod_{i=1}^n \eta_i^{\alpha_i} \exp(-d_i \eta_i) &= \left\{\prod_{i=1}^k \eta_i^{\alpha_i} \exp(-d_i \eta_i)\right\} \left\{\prod_{i=k+1}^{n-1} \eta_i^{\alpha_i} \exp(-d_i \eta_i)\right\} \times \\
&\quad (1 - \eta_{k+1} - \ldots - \eta_{n-1})^{\alpha_n} \left(1 - \frac{\eta_1 + \ldots + \eta_k}{1 - \eta_{k+1} - \ldots - \eta_{n-1}}\right)^{\alpha_n} \\
&\leq \prod_{i=1}^k \eta_i^{\alpha_i} \exp\left\{-\left(d_i + \frac{\alpha_i}{1 - \eta_{k+1} - \ldots - \eta_{n-1}}\right) \eta_i\right\} \times
\end{aligned}$$

$$\left\{\prod_{i=k+1}^{n-1}\eta_i^{\alpha_i}\exp(-d_i\eta_i)\right\}(1-\eta_{k+1}-\ldots-\eta_{n-1})^{\alpha_n}.$$

$$\leq \prod_{i=1}^{k}\eta_i^{\alpha_i}\exp\left\{-\left(d_i+\alpha_i\right)\eta_i\right\}\times \tag{25}$$

$$\left\{\prod_{i=k+1}^{n-1}\eta_i^{\alpha_n}\exp(-d_i\eta_i)\right\}(1-\eta_{k+1}-\ldots-\eta_{n-1})^{\alpha_n}. \tag{26}$$

The recommended carrying density is to approximate the density 26 by a suitably chosen Dirichlet as discussed above and the density 25 by independent Gammas with shape α_i and scale $(d_i+\alpha_i)^{-1}$. It is reasonable to restrict the range of the Gamma random variables and generate another Gamma if this does not hold. It may also be tempting to generate another suite of Gammas if their sum is either greater than one or even η_n say. However this has the effect of repeatedly generating Gammas for unfavourable $\eta_{k+1},\ldots,\eta_n$ and so it is necessary to generate a complete new η vector in this case. A potential η should be accepted if

$$U \leq \prod_{i=1}^{k} exp(\frac{\alpha_n\eta_i}{1-\sum_{j=k+1}^{n-1}\eta_j})$$

$$\left\{\prod_{i=k+1}^{n-1}\left(\frac{\eta_i}{v_i^*}\right)^{\alpha_i-c_0\alpha_n}\exp\{-d_i(\eta_i-v_i^*)\}\right\}\times$$

$$\left(1-\frac{\eta_1+\ldots+\eta_k}{1-\eta_{k+1}-\ldots-\eta_{n-1}}\right)^{\alpha_n}, \tag{27}$$

where $U \sim U(0,1)$ and v^* is calculated from $d_{k+1},\ldots,d_n$ only using equations 22 and 23.

There are two reasons for a vector η to be rejected:

- The generated deviate may not lie in the simplex which automatically means that it cannot be a candidate for a tilted Dirichlet variable.
- The vector η may not satisfy the rejection sampling inequality given in equation 27.

For large n, the combination of these two reasons can possibly lead to high probabilities of rejection and different methods are required. The method that we propose generates the vector η by Gibbs sampling where we generate the components of η in pieces of length k by rejection sampling based on equations 25, 26 and 27. It is well known (Smith & Roberts, 1993) that by using multivariate sections of the vector

rather than univariate sections, convergence can be greatly increased in situations where there is correlation between the components.

The comparison of the rejection sampling approach to the Gibbs sampling approximation considered previously is complicated by the dependance of the results on the particular situation considered. The convergence of the Gibbs subchain was examined by running parallel chains and examining the behaviour of the output over iterations, comparing this via Q-Q plots to a sample from the true distribution obtained via the rejection sampling algorithm. The results of this for $n = 30$ and a plausible set of λ and $P(\beta, x_i)$ showed that approximate convergence was obtained after n iterations, provided that η_n was taken to have $d_n = 0$. In this case the expected value of η_n is larger than that of the other η's, and the sampler can traverse the simplex more freely. In addition, the Gibbs subchain sampling of η is only a component of the full sampler, and approximate convergence is adequate for the convergence of the chain.

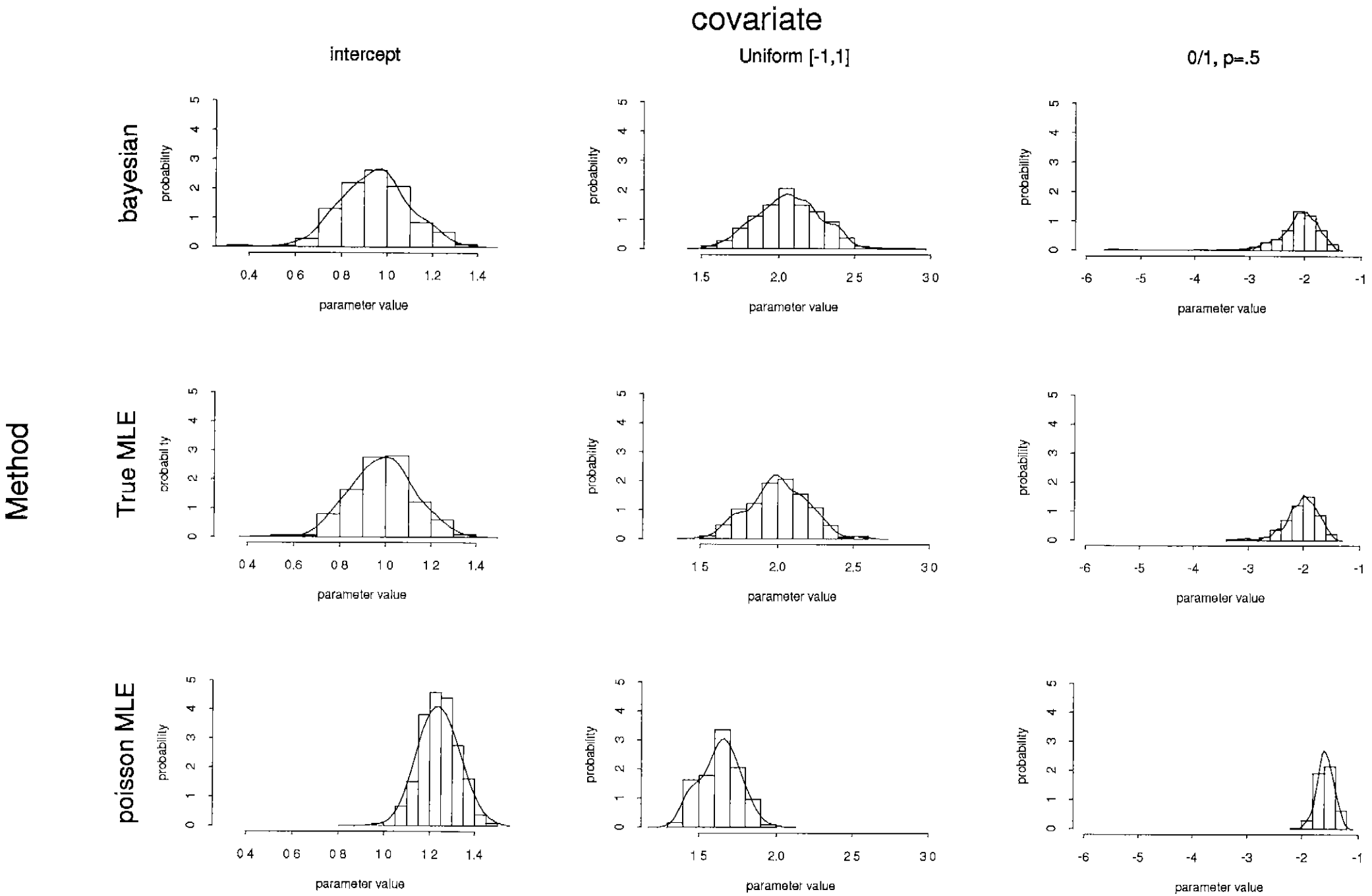

Figure 1. Distribution of modes over simulations

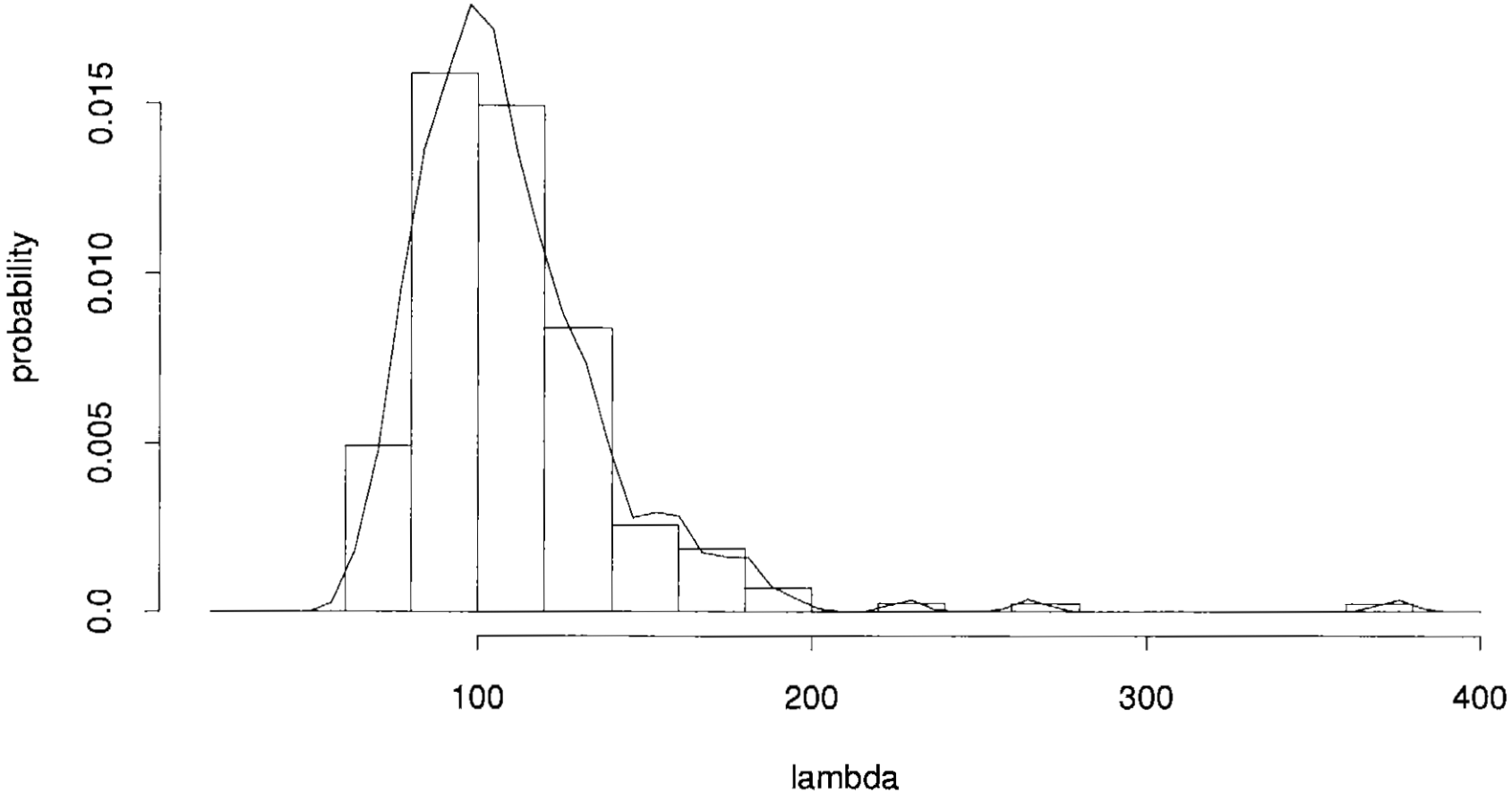

Figure 2: Distribution of modes of marginal posterier for lambda

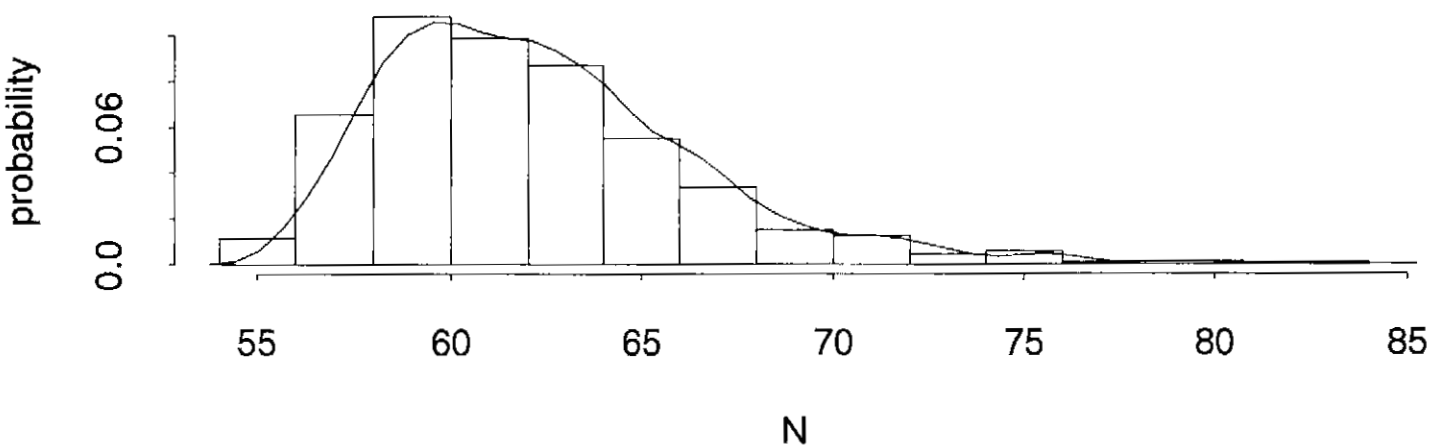

Figure 3: Approximate marginal posterier for N