UNIVERSITY OF SOUTHAMPTON

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS
School of Mathematics

Model-Based Adaptive Cluster Sampling

by

Veronica Elizabeth Rapley

Thesis submitted for the degree of Doctor of Philosophy
August 2004

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS
SCHOOL OF MATHEMATICS

Doctor of Philosophy

MODEL-BASED ADAPTIVE CLUSTER SAMPLING

by Veronica Elizabeth Rapley

Adaptive cluster sampling was introduced as a refined method for estimating the size of sparse clustered populations of plants or animals. Thompson (1990) formalised the strategy of increasing survey effort around where a plant or animal of interest is found and developed a design-based analysis of the resulting sampling scheme. We wish to view this sampling scheme from a model-based perspective. The general theory for making maximum likelihood (model-based) inference from sample survey data is presented in Breckling et al. (1994). We present an overview of this theory and use this to create a model-based approach to analysing sparse, clustered data.

The ideas contained within Breckling et al. (1994) are expanded by creating a model not dissimilar to the final model proposed, but simple enough to give an idea of the possibilities of modelling this situation in a frequentist framework and also an indication of where some of the complexities underlying the problem arise. In itself this section presents an interesting and stand alone extension to Breckling et al. (1994). We then explore some of the literature on modelling oil-pools, a situation involving continuous measurements which poses similar problems to those needing to be addressed in the discrete clustered case. This gives us a critical insight into how to create a likelihood which includes a sampling proportional to size strategy.

The main work of the thesis is in the synthesis of these ideas into a model which gives a more efficient estimate of overall population totals than the design-based estimates proposed by Thompson (1990). This model is predictably complex and despite critical insights into simplifying the problem, such as finessing the spatial component of the clusters, we necessarily use Bayesian methodology to make inference from the sample. The estimates produced prove to be more efficient than the design-based estimates and the model is a success.

Contents

List of Figures

List of Tables

Acknowledgments

A special thank you to my supervisor Professor Alan Welsh for his continued help, guidance, support and patience. Without his encouragement and confidence I would have struggled to complete this work.

I would like to acknowledge the EPSRC and University of Southampton for funding me throughout my three years and making my continued study possible.

I would like to thank Dr Jon Forster for all of his help and guidance, Dr David Woods and Mr David Gurnell for their time and help in programming and computing in general.

Thank you to my family, office mates and friends for their unfailing support and to my partner for his patience and understanding.

Chapter 1

Preface

Adaptive cluster sampling was proposed as a refined method for estimating the size of sparse clustered populations of plants and animals. In the field if we are dealing with a rare species of bird which is normally found in groups, it is tempting once a specimen is found to examine the surrounding area for more. Thus the scientist gains more data with little extra effort and more can be learned about the species. The motivation for a sampling scheme which allows for this type of sampling is clear. Thompson (1990) formalised the strategy of increasing survey effort around where a plant or animal of interest is found and developed a design-based analysis of the resulting sampling scheme. The methods proposed in Thompson (1990) are used not only for animal and plant populations, but also in epidemiological surveys of rare diseases and social surveys. In chapter 2 we examine the approach proposed in Thompson (1990) and look at some of the applications and adaptations in more detail.

The aim of this thesis is to view the problem of analysing sparse, clustered data sampled by adaptive cluster sampling, from a model-based perspective. Chapter 2 continues with a brief discussion of design-based versus model-based approaches to inference. We then proceed to describe a method for making maximum likelihood inference from sample survey data presented

in Breckling et al. (1994). This paper develops a method for making model-based inference about an entire population from a sample. The idea behind it is fundamental to this research.

In chapter 3 we examine Nair and Wang (1989) in which an oil play is modelled. This situation involves continuous measurements such as surface area, volume, net pay and depth it also poses similar problems to those we address in the discrete clustered population case. The paper gives an insight into how we create a likelihood which incorporates a sampling proportional to size strategy. The maximum likelihood estimate is found from the model they present by using the EM algorithm due to the difficulty in working with the integrals involved in the model. A brief outline of the EM algorithm is given at the beginning of this chapter as a reference.

The work by Nair and Wang (1989) was limited by the types of distributions which could be used in the model due to the difficulties arising in performing the necessary integrals. An obvious extension of this work is to take the model presented in this paper and apply Bayesian analysis. This is done in West (1996). Chapter 4 begins with an overview of Bayesian methodology and goes on to summarise West (1996). While this paper only examines this particular model, this work generalises the analysis in a way which allows more complicated models to be constructed and analysed using the same methods.

In chapter 5 we begin to expand upon the ideas contained in Breckling et al. (1994) by constructing a model similar to the final model proposed, but simple enough that while it can be analysed within a frequentist framework it also gives an indication of where some of the complexities underlying the problem arise. We analyse the model in both a frequentist and Bayesian framework and present a comparison of these approaches. This section presents an interesting and stand alone extension to Breckling et al. (1994), while also giving us the chance to ensure that we can produce similar results using both methods and explore some of the difficulties this problem presents.

In chapter 6 we synthesise all of the ideas examined to this point into a model for a sparse, clustered population. This model is predictably complex and despite the critical insights made into finessing the spatial aspect of this problem we find that the model is analytically intractable. We therefore use a Gibbs sampler similar to that proposed in West (1996) and described earlier in chapter 4 to perform our analysis.

Chapter 7 is the culmination of the work in this thesis. It takes the model developed in the previous chapter and extends it to model the population counts, giving us a method for predicting population totals from our model. The work in this chapter is also presented in a preprint by the author and Prof. Alan Welsh. The estimates produced using this model are compared to those obtained from the design based methods presented in Thompson (1990). As hoped the estimates produced are more efficient than the design based estimates.

Finally in chapter 8 we propose an alternative method for modelling sparse, clustered data. We then return to the continuous case and examine the possibility of using our model in the continuous case of oil pools.

Chapter 2

Literature Review for Sampling and Modeling

2.1 Adaptive Cluster Sampling

2.1.1 Introduction

In conventional sampling plans, the entire set of units which we wish to observe can be selected prior to the survey. By contrast, in adaptive cluster sampling the procedure for selecting units to include in the sample depends on the variable of interest, which is only observed during the survey (Thompson and Seber, 1996). We sample units in order to achieve a given aim, for instance to make inference about the population as a whole. In the instance of clustering, adaptive cluster sampling improves on simple random sampling by allowing us to increase survey effort around where a plant or animal of interest is found.

2.1.2 Methodology

Adaptive sampling is frequently used when populations are sparse and clustered since it can give predictions of population totals which in general

reflect the true values more closely. To demonstrate this, first consider a biological population spread homogeneously over a region with a grid of a given size superimposed upon it. If a simple random sample of units (grid cells) is taken, we can estimate the population total by using the 'expansion estimator' (Thompson, 1992) which scales up directly from our sample. In this homogeneous case a relatively accurate estimate can be produced. Now consider a non-homogeneous population, for instance one in which there are clusters. The expansion estimator can give us an inefficient estimator of the population total. If the sample includes several cluster cells the population total will be over estimated. If the sample includes very few cluster cells it will underestimate. In this situation using an adaptive sampling strategy can give more efficient estimators and is therefore to be preferred in most cases (Thompson, 1990).

In its simplest form adaptive cluster sampling requires that if a selected unit contains a member of the biological population, then the surrounding 'nearest neighbour' units are also sampled. This will continue until a group of units each containing at least one member of the biological population is completely surrounded by units which do not contain any of the biological population. (The 'nearest neighbour' units can be defined in many ways: the simplest way, and the method applied throughout this thesis, is to define them as the units sharing a common edge with the current unit.) The set of contiguous units containing members of the population make up a network, while the set of contiguous units sampled, both the network and the 'empty' units, is termed a 'cluster'. It is convenient to define all singular 'empty' cells as networks in their own right, so an edge unit is in fact a network of size one. These definitions are illustrated in Figure 2.1.

When the data are analysed, the networks become the analysis units and the boundary units of the networks are ignored if they do not already appear in the original sample (Thompson and Seber, 1996). Networks are used as analysis units because it is possible to calculate inclusion probabilities for each network, allowing the size of the networks to be accounted for. Networks are used as analysis units in preference to clusters because they

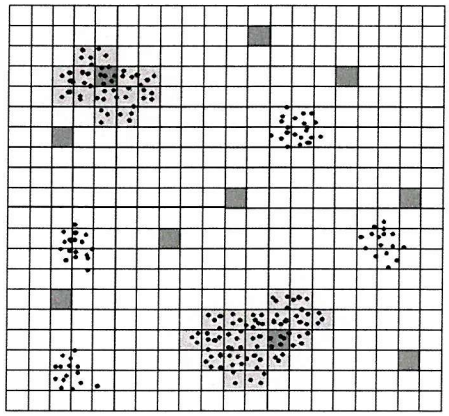


Figure 2.1: Illustration of adaptive cluster sampling, showing a number of point objects in a study region of 400 units. The dark grey units are the original sampled units, the medium grey units highlight the sampled networks and the light grey units are the boundary cells.

do not contain boundary units so there can be no overlap of networks. The networks are disjoint and form a partition over the region. Networks are used as analysis units in preference to the grid cells because the grid cells within networks have a dependence structure: working at the network level allows us to avoid making this dependence structure explicit.

2.1.3 Extensions of adaptive cluster sampling

Several of the extensions to simple random sampling, for instance stratified simple random sampling, can also be applied to adaptive sampling. Methods for stratified adaptive cluster sampling were first proposed in Thompson (1991b). Another extension was proposed in Thompson (1991a). In this approach primary sampling units, for example groups of units arranged in strips or rectangles, are defined and then sampled randomly. If a member of the population is encountered within a primary unit, secondary units outside of the primary unit are added to the sample in the same way as in normal

adaptive cluster sampling (Thompson, 1991a). This approach is illustrated in figure 2.2.

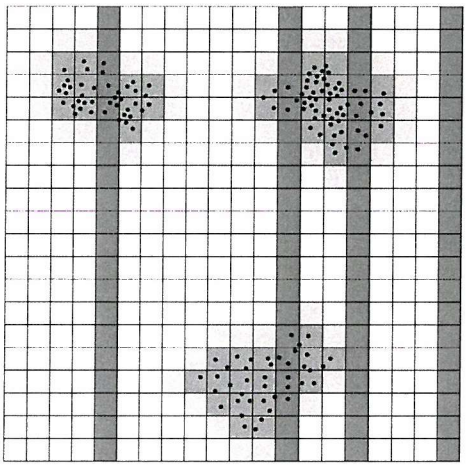


Figure 2.2: Illustration of stratified adaptive cluster sampling. Here the darkest shaded squares are the primary units (Thompson, 1991a). The networks are then formed from the secondary units

The methodology for adaptive cluster sampling is very flexible and has therefore been applied in various situations such as epidemiological surveys of rare diseases and in observing clustering of lichens.

Work has been undertaken to produce improvements on the unbiased estimators of Thompson (1990). Discussion of this work can be found in Sarndal (1996) and Felix-Medina (2000).

2.2 Model-based versus Design-based Inference

Adaptive cluster sampling is usually treated from a design-based as opposed to a model-based perspective. There have been numerous discussions of the relative advantages of the model-based and design-based approaches, see Royall (1976); Hansen et al. (1983); Brus and de Gruijter (1997); Thompson (2002). In general terms, design-based sampling and inference treats

population values as *fixed*. The key points underlying design-based inference are:

- We deal with a fixed finite population.
- Selection of sampling units depends on randomisation.
- Each member of the population has a known chance of being in the sample and inference depends on the randomisation and repeated sampling behaviour.
- Valid inference statements can only be constructed about finite population parameters (enumerative inference).

In model based sampling and inference population values are treated as realisations of random variables drawn from some superpopulation distribution. This means that for model-based inference:

- randomisation is not required (although this is still a useful concept).
- Inference depends on the assumed model.

The main benefit of model-based inference is its efficiency in comparison to design-based inference when the model holds. If the model fails then design-based inference may or may not be more efficient, depending on the robustness of the procedures.

Adaptive cluster sampling gives a way of dealing with clustered populations in a design-based framework. However, there is as yet no comparable technique using the model-based approach.

2.3 Maximum Likelihood Inference from Sample Survey Data

2.3.1 Introduction

When sampling from a population we aim to estimate an aspect of that population, be it some quantitative aspect of a finite population or the parameters of the distribution assumed to have generated the population values (Skinner et al., 1989). This thesis is initially concerned with the goal of estimating the parameters characterising the underlying distribution of a sparse clustered population. To achieve this we introduce some techniques developed for survey sampling in Breckling et al. (1994).

2.3.2 Maximum Likelihood Estimation in Survey Sampling

One method of estimating the parameters of a distribution is to calculate the maximum likelihood estimators of the parameters. When the likelihood is sufficiently smooth these are found by setting the score function for the parameter to zero, solving for the parameter and ascertaining whether the information function is positive definite.

Using the notation and development of Breckling et al. (1994), consider a finite population $\mathcal{P}$ made up of N elements. We consider D_j to be a vector containing all random variables associated with the jth element of the population, $j \in \mathcal{P}$. So, for example in a population $\mathcal{P}$ of lizards, D_j could be made up of several random variables associated with the jth lizard, for instance its length and weight. We denote the number of random variables associated with each member of the population $j \in \mathcal{P}$ by p, so that D_j is a p-vector. We let D be a $N \times p$ matrix with jth row D_j^T

Let the joint distribution of these variables over $\mathcal{P}$ be $f(D;\theta)$. We wish to estimate θ. If D is fully observed this can usually be achieved by solving the equation $\text{sc}(\theta) = 0$, subject to $\text{info}(\theta)$ being positive definite. We

define $\partial_\theta \log f(D;\theta)$ to be the vector of partial derivatives of the density function with respect to the set of parameters and $\partial_{\theta\theta} \log f(D;\theta)$ to be the Hessian matrix of the the density function with respect to the set of parameters. Then $\text{sc}(\theta)$ $=\text{sc}(\theta; D) = \partial_\theta \log f(D;\theta)$ is the score function for θ and $\text{info}(\theta)$ $=\text{info}(\theta; D) = -\partial_\theta \text{sc}(\theta) = -\partial_{\theta\theta} \log f(D;\theta)$ is the observed information function for θ.

In sampling situations, we only have a sample (s) of size $n < N$. We can also have non-response within the sample itself, so in actuality we only observe part of D, together with information about which members of the population have been sampled and which members of the sample have responded. We denote the observed part of the sample by $\mathcal{D}_{s1}$ and the sample design by Z (an $N \times q$ matrix of values for q known auxiliary variables (Scott, 1977)). Let I be a vector of inclusion indicators (1 if an element is included and 0 otherwise) and R_s be a response indicator for the sampled elements (1 if a response has been recorded and 0 otherwise). The available sample data set is $\jmath_s = (\mathcal{D}_{s1}, R_s, I, Z)$.

We wish to find the maximum likelihood estimator of the parameters as a function only of the variables in the set $\jmath_s$. However, it should be noted that the model is now subtly different. We no longer have simply a parameter θ to estimate because the joint distribution of all the variables in $\jmath_s$ involves additional parameters and the different ways to parameterise the model must be taken into account. If a different parameterisation of the model is made and the relationship between the new and original parameters is known then maximum likelihood estimators for the original parameters can still be found by transforming the score function for the new parameters into a score function for the original parameters. Hence reparameterisation does not always pose a significant problem.

The sample score for a generic parameterisation is calculated (Breckling et al., 1994) as follows:

$$sc_s(\aleph) = E_s[sc(\aleph)],$$

where $\aleph$ is a generic parameter representing one of many possible parameterisations of the data and $E_s(\cdot) \equiv E(\cdot \mid \jmath_s)$. The sample score function is therefore the expected value of the score function given the sample (Breckling et al., 1994; Fisher, 1925). To find the sample information function it is possible to differentiate the sample score directly. Alternatively, we can also find the sample information from the original score and information function as follows (Breckling et al., 1994):

$$info_s(\aleph) = E[info(\aleph) \mid \jmath_s] - var[sc(\aleph) \mid \jmath_s].$$

This result is both satisfying and intuitive. It shows the loss of information incurred when we only have sample values as opposed to the values for the whole population.

The actual calculation of the sample score function begins with the joint density of the variables D, R_s, I and Z. We express this as a function of $\aleph$ and then differentiate with respect to $\aleph$ to obtain the population score function. We then take the expectation of the score function given $\jmath_s$ to get the sample score function. This last step is accomplished by taking the marginal distribution of $\jmath_s$ and then finding the conditional distribution of $D_s \mid \jmath_s$. These last two steps can be avoided in some cases as the formulae will simplify. We include the steps for completeness.

Chapter 3

Successive Sampling Discovery Models

3.1 The EM Algorithm

The EM algorithm can be used to find maximum likelihood estimators in incomplete data problems, although the algorithm itself is versatile and can be applied to many different problems. It was first proposed in Dempster et al. (1977) and has since been widely used in a number of areas.

The algorithm essentially consists of an E-step and an M-step, giving the algorithm its name.

E-step: This finds the conditional expectation of the log-likelihood with respect to the distribution of the 'missing data' given the already observed data and current estimation parameters, these are then substituted for the missing data values.

M-step: This is the maximum likelihood estimation of the parameters using the new values in place of the missing data.

More precisely, if we let θ^t be the current estimate of the parameter θ and Y be the observed data, the E-step finds the expected log-likelihood Q as though $\theta = \theta^t$,

$$Q(\theta \mid \theta^t) = \int l(\theta \mid Y) f(Y_{missing} \mid Y_{observed}, \theta = \theta^t) dY_{missing} \ .$$

The M-step then determines $\theta^{(t+1)}$ by maximising the expected likelihood

$$Q(\theta^{(t+1)} \mid \theta^t) \geq Q(\theta \mid \theta^t) \qquad \forall\, \theta \ .$$

We then iterate through these two steps successively.

3.2 The construction of a Successive Sampling Discovery Model

This section is a summary of a paper by Nair and Wang (1989) and some of the key ideas developed within it. The paper applies some of the methods to overcome size bias in survey sampling to oil pools and is a generalization of Kaufman et al. (1975). The general idea is to model the discovery process as sampling successively without replacement and with *probability proportional to size.*

Let $\mathbf{Y}_1, ..., \mathbf{Y}_N$ represent the variables of interest. Each $\mathbf{Y}_i$ is a vector of measurements related to the ith unit of the population, the population has N members where N is assumed known. In the case of oil pools the measurements are depth, net pay surface area and volume. We assume that the $\mathbf{Y}$'s are all independent random variables and that the density function of the $\mathbf{Y}$'s with respect to the parameter vector $\boldsymbol{\theta}$ ($f_{\boldsymbol{\theta}}$) is continuous.

We let $w(\cdot)$ denote a positive weight function, so the probability of observing an *ordered* sample $(i_1, ..i_n)$ is

$$Pr\{(i_1, ..., i_n) | \mathbf{Y}_i = \mathbf{y}_i, i = 1, ..., N\} = \prod_{j=1}^{n} \frac{w(\mathbf{y}_{i_j})}{\sum_{i=1}^{N} w(\mathbf{y}_i) - \sum_{k=0}^{j-1} w(\mathbf{y}_{i_k})}$$

$$where \ \ w(y_{i_0}) \equiv 0 \ .$$

That is, the sample is obtained by selection without replacement and with probability proportional to $w(\mathbf{y})$ from a population of size N. This is different from the usual probability-proportional-to-size sampling because the size measures are unknown a priori; in more familiar situations the size would be known a priori.

The sample taken is necessarily ordered. Let $\mathbf{X}_j$ be the value associated with the jth discovery. The *observed ordered sample* can be denoted $(\mathbf{x}_1, ..., \mathbf{x}_n)$, so the density function of the $\mathbf{X}$'s is also continuous with respect to $\boldsymbol{\theta}$. If the labeling is irrelevant, we can relabel the first n elements of the population so $\mathbf{X}_j = \mathbf{Y}_j (j = 1, ..., n)$ and the probability of observing the ordered sample $(\mathbf{x}_1, ..., \mathbf{x}_n)$ given the population is

$$
\begin{aligned}
&[\mathbf{x}_1, ..., \mathbf{x}_n | \mathbf{y}_1, ..., \mathbf{y}_N] \\
&\qquad = \prod_{j=1}^{n} \frac{w(\mathbf{y}_{i_j})}{\sum_{i=1}^{N} w(\mathbf{y}_i) - \sum_{k=0}^{j-1} w(\mathbf{y}_{i_k})} \\
&\qquad = \prod_{j=1}^{n} \frac{w(\mathbf{y}_{i_j})}{\sum_{i=1}^{n} w(\mathbf{y}_i) - \sum_{k=0}^{j-1} w(\mathbf{y}_{i_k}) + w(\mathbf{y}_{n+1}) + \cdots + w(\mathbf{y}_N)} \\
&\qquad = \prod_{i=1}^{n} \frac{w(\mathbf{x}_i)}{w(\mathbf{x}_i) + \cdots + w(\mathbf{x}_n) + w(\mathbf{y}_{n+1}) + \cdots + w(\mathbf{y}_N)} \\
&\qquad = \prod_{i=1}^{n} \frac{w(\mathbf{x}_i)}{b_i + R} \,,
\end{aligned}
$$

where $b_i = w(\mathbf{x}_i) + \cdots + w(\mathbf{x}_n)$ and $R = w(\mathbf{y}_{n+1}) + \cdots + w(\mathbf{y}_N)$.

The ultimate aim of the paper is to predict the values of the unobserved data from the sample. This is accomplished by treating the unobserved data as missing values and using the EM algorithm exactly as described in the previous section to estimate θ. We will reproduce these calculations here as they will be relevant later in the thesis.

First find $[\mathbf{x}_1, ..., \mathbf{x}_n, \mathbf{y}_1, ..., \mathbf{y}_N]$ and then use this to find $[\mathbf{x}_1, ..., \mathbf{x}_n]$. We

have already found that

$$[\mathbf{y}_1, ..., \mathbf{y}_N] = \prod_{i=1}^{N} f_{\boldsymbol{\theta}}(\mathbf{y}_i)$$

and

$$\begin{aligned}[\mathbf{x}_1, ..., \mathbf{x}_n | \mathbf{y}_1, ..., \mathbf{y}_N] &= \prod_{i=1}^{n} \frac{w(\mathbf{x}_i)}{b_i + R} \\ &= \prod_{i=1}^{n} \left\{ \frac{w(\mathbf{x}_i)}{b_i} \right\} c_i\end{aligned}$$

where $c_i = \frac{b_i}{b_i+R}$, which means

$$\begin{aligned}[\mathbf{x}_1, ..., \mathbf{x}_n, \mathbf{y}_1, ..., \mathbf{y}_N] &= [\mathbf{x}_1, ..., \mathbf{x}_n | \mathbf{y}_1, ..., \mathbf{y}_N][\mathbf{y}_1, ..., \mathbf{y}_N] \\ &= \prod_{i=1}^{n} \frac{w(\mathbf{x}_i)}{b_i} c_i \prod_{i=1}^{N} f_{\boldsymbol{\theta}}(\mathbf{y}_i). \qquad (3.1)\end{aligned}$$

All that remains in order to find the unconditional distribution of the sample is to integrate over the unknown values and sum over all of the possible ways of choosing $\mathbf{y}_1, ..., \mathbf{y}_n$.

We first manipulate (3.1) to make the calculation more tractable:

$$\begin{aligned}[\mathbf{x}_1, ..., \mathbf{x}_n, \mathbf{y}_1, ..., \mathbf{y}_N] &= \prod_{i=1}^{n} \frac{w(\mathbf{x}_i)}{b_i} c_i \prod_{i=1}^{N} f_{\boldsymbol{\theta}}(\mathbf{y}_i) \\ &= \prod_{i=1}^{n} \frac{w(\mathbf{x}_i)}{b_i} c_i \prod_{i=1}^{n} f_{\boldsymbol{\theta}}(\mathbf{y}_i) \prod_{i=n+1}^{N} f_{\boldsymbol{\theta}}(\mathbf{y}_i) \\ &= \prod_{i=1}^{n} \frac{w(\mathbf{x}_i) f_{\boldsymbol{\theta}}(\mathbf{x}_i) c_i}{b_i} \prod_{i=n+1}^{N} f_{\boldsymbol{\theta}}(\mathbf{y}_i) \\ &= \prod_{i=1}^{n} \frac{w(\mathbf{x}_i) f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{b_i} \prod_{k=1}^{n} c_k \prod_{i=n+1}^{N} f_{\boldsymbol{\theta}}(\mathbf{y}_i) \, .\end{aligned}$$

The values $\mathbf{x}_1, ..., \mathbf{x}_n$ form an ordered sample of the population $\mathbf{y}_1, ..., \mathbf{y}_N$. We wish to find the density of $(\mathbf{x}_1, ..., \mathbf{x}_n)$ so we must re-label the values $\mathbf{y}_1, ..., \mathbf{y}_N$ introducing a normalization constant $\frac{N!}{(N-n)!}$ to the density function (Andreatta and Kaufman, 1986; Kaufman et al., 1975). The calculation therefore is as follows,

$$\begin{aligned}
[\mathbf{x}_1, ..., \mathbf{x}_n] &= \frac{N!}{(N-n)!} \prod_{i=1}^{n} \frac{w(\mathbf{x}_i) f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{b_i} \\
&\qquad \int ... \int \prod_{k=1}^{n} c_k \prod_{i=n+1}^{N} f_{\boldsymbol{\theta}}(\mathbf{y}_i)\ d\mathbf{y}_{n+1} ... d\mathbf{y}_N \\
&= \frac{N!}{(N-n)!} \prod_{i=1}^{n} \frac{w(\mathbf{x}_i) f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{b_i}\ E_{\boldsymbol{\theta}} \left(\prod_{k=1}^{n} c_k \right) . \qquad (3.2)
\end{aligned}$$

We can express the expectation in (3.2) in various ways.

Let $T = \Sigma_{i=1}^{n} \frac{\epsilon_i}{b_i}$ where $\epsilon_i \sim$ independent identically distributed standard exponential variates independent of the $\mathbf{y}$'s. Then the moment generating function for ϵ_i is

$$\begin{aligned}
E_{\epsilon_i}(e^{-t\epsilon_i}) &= \int_0^\infty \lambda e^{-\lambda\epsilon} e^{-t\epsilon}\, d\epsilon \\
&= \int_0^\infty \lambda e^{-(\lambda+t)\epsilon}\, d\epsilon \\
&= \frac{\lambda e^{-(\lambda+t)\epsilon}}{\lambda + t} \Bigg|_0^\infty \\
&= \frac{\lambda}{\lambda + t} .
\end{aligned}$$

In our case $\lambda = 1$ and $t = \frac{R}{b_i}$, so

$$\begin{aligned} E_T e^{-RT} &= E_{\epsilon_1 \ldots \epsilon_n} \left\{ \prod_{i=1}^{n} e^{\frac{-R\epsilon_i}{b_i}} \right\} \\ &= \prod_{i=1}^{n} E_{\epsilon_i} \left\{ e^{\frac{-R\epsilon_i}{b_i}} \right\} \\ &= \prod_{i=1}^{n} \frac{1}{1 + \frac{R}{b_i}} \\ &= \prod_{i=1}^{n} \frac{b_i}{b_i + R} \, . \end{aligned} \tag{3.3}$$

Since the $\mathbf{y}_i$'s are independent, if we define $\phi(t; \boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(e^{-w(\mathbf{Y}_1)t})$ and let $g_n(t)$ denote the density of T then (3.3) allows us to write

$$\begin{aligned} E_{\boldsymbol{\theta}} \left\{ \prod_{i=1}^{n} \frac{b_i}{b_i + R} \right\} &= E_R E_T[exp(-RT)] \\ &= E_T E[e^{-(w(\mathbf{y}_{n+1}) + \cdots + w(\mathbf{y}_N))T}] \\ &= E_T E \prod_{i=n+1}^{N} e^{-w(\mathbf{y}_i)T} \\ &= E_T \prod_{i=n+1}^{N} E(e^{-w(\mathbf{y}_i)T}) \\ &= E_T \prod_{i=n+1}^{N} \phi(T; \boldsymbol{\theta}) \\ &= E_T \left\{ \phi(t; \boldsymbol{\theta})^{(N-n)} \right\} \\ &= \int_0^\infty \phi(t; \boldsymbol{\theta})^{(N-n)} g_n(t) \, dt \\ &= S(\boldsymbol{\theta}) \, , \end{aligned}$$

say, so finally we can write (3.2) as

$$\begin{aligned} f(\mathbf{x}_1, ..., \mathbf{x}_n) &= \frac{N!}{(N-n)!} \prod_{i=1}^{n} \frac{w(\mathbf{x}_i) f_{\boldsymbol{\theta}}(\mathbf{x})}{b_i} \int_0^\infty \phi(t;\boldsymbol{\theta})^{(N-n)} g_n(t)\, dt \\ &= \frac{N!}{(N-n)!} \prod_{i=1}^{n} \frac{w(\mathbf{x}_i) f_{\boldsymbol{\theta}}(\mathbf{x})}{b_i} \; S(\boldsymbol{\theta}) \; . \end{aligned} \tag{3.4}$$

We now have an expression for the likelihood

$$\exp\left(l(\boldsymbol{\theta})\right) = \frac{N!}{(N-n)!} \prod_{i=1}^{n} \frac{w(\mathbf{x}_i) f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{b_i} \; S(\boldsymbol{\theta}) \; ,$$

so, we take logs to find

$$l(\boldsymbol{\theta}) = constant + \sum_{i=1}^{n} \log f_{\boldsymbol{\theta}}(\mathbf{x}_i) + \log S(\boldsymbol{\theta}) \; .$$

By differentiating $l(\theta)$, we get the likelihood equations, which are equivalent to the score function and can be used to find the parameter estimates. Differentiating $\log S(\theta)$ first for clarity, we obtain

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\theta}_r} \log S(\boldsymbol{\theta}) &= \frac{\frac{\partial}{\partial \boldsymbol{\theta}_r} S(\boldsymbol{\theta})}{S(\boldsymbol{\theta})} \\ &= \frac{\frac{\partial}{\partial \boldsymbol{\theta}_r} \left\{ \int_0^\infty \phi(t;\boldsymbol{\theta})^{(N-n)} g_n(t)\, dt \right\}}{S(\boldsymbol{\theta})} \\ &= \frac{(N-n) \int_0^\infty \phi(t;\boldsymbol{\theta})^{(N-n-1)} \frac{\partial}{\partial \boldsymbol{\theta}_r} \phi(t;\boldsymbol{\theta}) g_n(t)\, dt}{S(\boldsymbol{\theta})} \\ &= \frac{(N-n) \int_0^\infty \phi(t;\boldsymbol{\theta})^{(N-n)} \frac{\frac{\partial}{\partial \boldsymbol{\theta}_r} \phi(t;\boldsymbol{\theta})}{\phi(t;\boldsymbol{\theta})} g_n(t)\, dt}{S(\boldsymbol{\theta})} \\ &= \frac{(N-n) \int_0^\infty \left[\frac{\partial}{\partial \boldsymbol{\theta}_r} \log \phi(t;\boldsymbol{\theta}) \right] \phi(t;\boldsymbol{\theta})^{(N-n)} g_n(t)\, dt}{S(\boldsymbol{\theta})} \\ &= (N-n) \int_0^\infty \left[\frac{\partial}{\partial \boldsymbol{\theta}_r} \log \phi(t;\boldsymbol{\theta}) \right] h_{\boldsymbol{\theta}}(t)\, dt \end{aligned}$$

where $r = 1, .., m$ and m is the dimension of $\boldsymbol{\theta}$. We define the density function h_θ as

$$h_{\boldsymbol{\theta}}(t) = \frac{[\phi(t;\boldsymbol{\theta})]^{N-n} g_n(t)}{S(\boldsymbol{\theta})} .$$

In full then,

$$\begin{aligned} \frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_r} &= \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}_r} \log f_{\boldsymbol{\theta}}(\mathbf{x}_i) \\ &\quad + (N-n) \int_0^\infty \left[\frac{\partial}{\partial \boldsymbol{\theta}_r} \log \phi(t;\boldsymbol{\theta}) \right] h_{\boldsymbol{\theta}}(t)\, dt. \end{aligned} \tag{3.5}$$

We can manipulate the score function further using the fact that

$$\begin{aligned} \frac{\partial \log \phi(t;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_r} &= \frac{\frac{\partial \phi(t;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_r}}{\phi(t;\boldsymbol{\theta})} \\ &= \frac{1}{\phi(t;\boldsymbol{\theta})} \frac{\partial}{\partial \boldsymbol{\theta}_r} \int e^{-tw(y)} f_{\boldsymbol{\theta}}(\mathbf{y})\, d\mathbf{y} \\ &= \frac{1}{\phi(t;\boldsymbol{\theta})} \int e^{-tw(\mathbf{y})} \frac{\partial}{\partial \boldsymbol{\theta}_r} f_{\boldsymbol{\theta}}(\mathbf{y})\, d\mathbf{y} \\ &= \frac{1}{\phi(t;\boldsymbol{\theta})} \int e^{-tw(\mathbf{y})} \left\{ \frac{\frac{\partial}{\partial \boldsymbol{\theta}_r} f_{\boldsymbol{\theta}}(\mathbf{y})}{f_{\boldsymbol{\theta}}(\mathbf{y})} \right\} f_{\boldsymbol{\theta}}(\mathbf{y})\, d\mathbf{y} \\ &= \frac{1}{\phi(t;\boldsymbol{\theta})} \int e^{-tw(\mathbf{y})} \frac{\partial}{\partial \boldsymbol{\theta}_r} \log f_{\boldsymbol{\theta}}(\mathbf{y}) f_{\boldsymbol{\theta}}(\mathbf{y})\, d\mathbf{y} \\ &= \int \left[\frac{\partial}{\partial \boldsymbol{\theta}_r} \log f_{\boldsymbol{\theta}}(\mathbf{y}) \right] k(\mathbf{y}|t;\boldsymbol{\theta})\, d\mathbf{y} , \end{aligned} \tag{3.6}$$

where we define another density function,

$$k(\mathbf{y}|t;\boldsymbol{\theta}) = \frac{e^{-tw(\mathbf{y})} f_{\boldsymbol{\theta}}(\mathbf{y})}{\phi(t;\boldsymbol{\theta})} .$$

Substituting (3.6) into the score function (3.5) we have,

$$\begin{aligned}\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_r} &= \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}_r} \log f_{\boldsymbol{\theta}}(\mathbf{x}_i) \\ &\quad +(N-n)\int_0^\infty \left\{ \int \left[\frac{\partial}{\partial \boldsymbol{\theta}_r} \log f_{\boldsymbol{\theta}}(\mathbf{y}) \right] k(\mathbf{y}|t;\boldsymbol{\theta}) h_{\boldsymbol{\theta}}(t)\, d\mathbf{y} \right\} dt \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}_r} \log f_{\boldsymbol{\theta}}(\mathbf{x}_i) \qquad (3.7) \\ &\quad +(N-n)\int \left[\frac{\partial}{\partial \boldsymbol{\theta}_r} \log f_{\boldsymbol{\theta}}(\mathbf{y}) \right] \left(\int_0^\infty k(\mathbf{y}|t;\boldsymbol{\theta}) h_{\boldsymbol{\theta}}(t)\, dt \right) d\mathbf{y}\ .\end{aligned}$$

Now, in the same way as in Breckling et al. (1994), the likelihood is split into the observed and unobserved components, so the integral above is actually the distribution of the unobserved data given the sample; we can show this using (3.1) and (3.4).

We first use (3.1) to find $[\mathbf{x}_1, ..., \mathbf{x}_n, \mathbf{y}_{n+1}, ..., \mathbf{y}_N]$ by integrating out the values $\mathbf{y}_1, ..., \mathbf{y}_n$. We still have an ordered sample, so to find the unordered density function we again multiply the density by $\frac{N!}{(N-n)!}$, just as we did to find (3.4). So,

$$\begin{aligned}&[\mathbf{x}_1, ..., \mathbf{x}_n, \mathbf{y}_{n+1}, ..., \mathbf{y}_N] \\ &\qquad = \frac{N!}{(N-n)!} \int ... \int [\mathbf{x}_1, ..., \mathbf{x}_n, \mathbf{y}_1, ..., \mathbf{y}_N] d\mathbf{y}_1 ... d\mathbf{y}_n \\ &\qquad = \frac{N!}{(N-n)!} \prod_{i=1}^{n} \frac{w(\mathbf{x}_i) f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{b_i} \prod_{k=1}^{n} c_k \prod_{i=n+1}^{N} f_{\boldsymbol{\theta}}(\mathbf{y}_i).\end{aligned}$$

Then,

$$[\mathbf{y}_{n+1},...,\mathbf{y}_N|\mathbf{x}_1,...,\mathbf{x}_n] = \frac{[\mathbf{x}_1,...,\mathbf{x}_n,\mathbf{y}_{n+1},...,\mathbf{y}_N]}{[\mathbf{x}_1,...,\mathbf{x}_n]}$$

$$= \frac{\frac{N!}{(N-n)!}\prod_{i=1}^n \frac{w(\mathbf{x}_i)f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{b_i}\prod_{k=1}^n c_k \prod_{i=n+1}^N f_{\boldsymbol{\theta}}(\mathbf{y}_i)}{\frac{N!}{(N-n)!}\prod_{i=1}^n \frac{w(\mathbf{x}_i)f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{b_i}\ S(\boldsymbol{\theta})}$$

$$= \frac{\prod_{i=1}^n c_i \prod_{i=n+1}^N f_{\boldsymbol{\theta}}(\mathbf{y}_i)}{S(\boldsymbol{\theta})}$$

$$= \prod_{i=1}^n \left\{\frac{b_i}{b_i+R}\right\} \prod_{i=n+1}^N \frac{f_{\boldsymbol{\theta}}(\mathbf{y}_i)}{S(\boldsymbol{\theta})}\,.$$

Now, using (3.3), we have

$$\begin{aligned}
\prod_{i=1}^n \frac{b_i}{b_i+R} &= E_T e^{-RT} \\
&= E_T e^{-(w(\mathbf{y}_{n+1})+\cdots+w(\mathbf{y}_N))T} \\
&= E\prod_{i=n+1}^N e^{-w(\mathbf{y}_i)T} \\
&= \prod_{i=n+1}^N E_T(e^{-w(\mathbf{y}_i)T}) \\
&= \prod_{i=n+1}^N \int_0^\infty e^{-w(\mathbf{y}_i)t} g_n(t)\,dt\ . \qquad (3.8)
\end{aligned}$$

If we now substitute (3.8) we obtain

$$\begin{aligned}
[\mathbf{y}_{n+1},...,\mathbf{y}_N|\mathbf{x}_1,...,\mathbf{x}_n] &= \frac{\int_0^\infty \prod_{i=n+1}^N e^{-tw(\mathbf{y}_i)} g_n(t) f_{\boldsymbol{\theta}}(y_i)\,dt}{S(\boldsymbol{\theta})} \\
&= \frac{\int_0^\infty \prod_{i=n+1}^N k(\mathbf{y}_i|t;\boldsymbol{\theta}) g_n(t)\,dt}{[\phi(t;\boldsymbol{\theta})]^{N-n} S(\boldsymbol{\theta})} \\
&= \int_0^\infty \prod_{i=n+1}^N k(\mathbf{y}_i|t;\boldsymbol{\theta}) h_{\boldsymbol{\theta}}(t)\,dt.
\end{aligned}$$

This density is symmetric in its arguments, therefore the conditional distribution of any one of the unknown variables $\mathbf{Y}_{n+1}$ given the data is

$$f_{\boldsymbol{\theta}}(\mathbf{y} \mid data) = \int_0^\infty k(\mathbf{y}|t;\boldsymbol{\theta}) h_{\boldsymbol{\theta}}(t)\, dt. \tag{3.9}$$

We can use (3.9) to solve (3.7) as

$$\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_r} = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}_r} \log f_{\boldsymbol{\theta}}(\mathbf{x}_i) + E_{\boldsymbol{\theta}} \left(\left[\frac{\partial}{\partial \boldsymbol{\theta}_r} \log f_{\boldsymbol{\theta}}(\mathbf{Y}) \right] \mid \mathbf{x}_1, ..., \mathbf{x}_n \right) .$$

Now for the M-step we must maximise the log-likelihood, so we solve,

$$0 = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}_r} \log f_{\boldsymbol{\theta}}(\mathbf{x}_i) + (N-n) E_{\boldsymbol{\theta}^*} \left(\left[\frac{\partial}{\partial \boldsymbol{\theta}_r} \log f_{\boldsymbol{\theta}}(\mathbf{Z}) \right] \mid \mathbf{x}_1, ..., \mathbf{x}_n \right),$$

where $\mathbf{Z}$ denotes any one of the $(N-n)$ missing variables because they are independent. The conditional distribution of $\mathbf{Z}$, given the data, is given by (3.9) so the EM algorithm can be iterated as described above.

However, even with the EM algorithm several of the integrals outlined above still need to be calculated. These integrals are intractable when not special cases which leads to the use of numerical methods. A paper extending these ideas (West, 1996) uses Bayesian inference and Markov Chain Monte Carlo simulation solve this problem.

Chapter 4

Bayesian Methodology and Applications to a Successive Sampling Discovery Model

4.1 Introduction

Bayesian theory provides an approach to statistical inference which is different from the classical approach. The basic philosophy underlying Bayesian inference is that probability is the only sensible measure of uncertainty. The Bayesian approach treats parameters not as fixed and unknown, but as random variables in their own right. Treating parameters in this way can mean that complex integrals must be evaluated, which in turn has led to using numerical forms of integration to find parameter estimates, for example Markov chain Monte Carlo.

Markov chain Monte Carlo or MCMC as it is more widely known has become popular in the last fifteen years or so because it can provide a useful way to analyse problems that were previously considered intractable. It provides a method for simulating complex, non-standard multivariate distributions (Chib and Greenberg, 1995). The original method was developed

in Metropolis et al. (1953) and was then generalised by Hastings (1970). The Gibbs sampler arises as a special case of the Metropolis-Hastings algorithm (Gelman, 1992; Chib and Greenberg, 1995).

4.2 Bayesian Statistics and The Metropolis Hastings Algorithm

4.2.1 Background

In Bayesian inference the parameters are random variables and therefore have both prior and posterior distributions. Let θ represent the unknown parameters, then our prior beliefs about θ are represented by the prior $\pi(\theta)$, a probability density function. Let $\mathbf{x}$ be the data which can be written as $x_1, x_2, ..., x_n$ and let $f(x_1, ..., x_n \mid \theta)$ be the likelihood of the data given the unknown parameters. The posterior, $\pi(\theta \mid x_1, ...x_n)$, is then our modified belief about θ in light of the observed data.

Bayes' Theorem states that the posterior distribution is proportional to the likelihood multiplied by the prior:

$$\pi(\theta \mid x_1, ..., x_n) \propto f(x_1, ..., x_n \mid \theta)\pi(\theta).$$

This is the underlying theorem which underpins all of Bayesian statistics.

4.2.2 Prior Choice

It is clear that Bayesian inference depends upon the prior. A prior represents the knowledge we already have about the parameter we are hoping to estimate, before the sample is taken. So if we know something about what the parameter estimate should be this information is included within the prior. In order to follow the Bayesian method, we are adopting a subjective

interpretation of probability. In this interpretation probability represents a degree of belief in a proposition, based on all of the available information.

Prior choice in Bayesian statistics has produced much discussion and is different in each situation, so a general method for choosing a prior is unlikely. A good overview of the subject is provided by Kass and Wasserman (1996); Jeffreys (1961); Besag and Green (1993).

Conjugate Priors

For a given class of likelihood functions $[\mathbf{X}|\theta]$, the class Ω of priors $\pi(\theta)$ is called a conjugate if the posterior $\pi(\theta|\mathbf{X})$ is of the same class as Ω. If the prior is tractable this has the desirable outcome of making the posterior tractable (Bernado and Smith, 1998). For example, a binomial likelihood and a Beta prior distribution will give a Beta posterior distribution, which is tractable and easily sampled from. Conjugate priors were widely used before numerical methods allowed posteriors to be estimated because they allowed relatively simple analytical calculation of the posterior distributions.

There are many known distributions which have conjugate priors and these priors are given hyperparameters to make the shape of the prior fit the prior knowledge. So, if we knew that our likelihood was binomial but that the parameter we were sampling for was likely to be small, we would choose a beta distribution which was weighted towards the lower tail.

Non-Informative Priors

If a prior distribution does not contain any information about θ, it is called a non-informative (vague, diffuse or flat) prior.

Ideally, from a subjective viewpoint we would elicit a prior on the basis of available information, expert opinion or past experience. However, there are many situations where we may have little or no prior information about the

parameter we wish to estimate. We are therefore interested in a prior under partial or complete ignorance about a likely prior value, particularly in high dimensional problems.

Although proponents of subjective probability question whether a state of complete ignorance exists, many will assume total ignorance as an approximation if prior information is weak. This approach allows some reconciliation between frequentist and Bayesian approaches to inference because less prior knowledge is assumed.

The most widely used non-informative priors are Jeffreys (1961) priors:

$$\pi(\theta) = \{I(\theta)\}^{\frac{1}{2}},$$

where $I(\theta)$ is the Fisher information. We cannot simply use a uniform distribution as a non-informative prior in all cases because complete ignorance about θ implies knowledge about $\phi = g(\theta)$. A Jeffreys prior gives a solution to this problem. The Jeffreys prior is often improper: this is not a problem providing that the resulting posterior distribution is proper, something which should always be checked before inference is made.

Conclusion

There are many possible choices of prior depending on the particular situation. There are no actual rules for which prior to choose and problems can arise if the wrong choice is made. In fact the wrong choice of prior can outweigh or swamp the data. Obviously this is a problem and one which many believe to be one of the main difficulties in the Bayesian approach to statistics.

4.2.3 Computation and Evaluation of Integrals

Let $g(x)$ be a normalised density function, and $f(x)$ be the non-normalised density function, so $g(x) = \frac{f(x)}{\int f(x)dx}$. Then let $b(x)$ be a function of interest.

We often want to compute expressions of the form,

$$E_g(b) = \int b(x)g(x)dx = \frac{\int b(x)f(x)dx}{\int f(x)dx}.$$

However, these integrals are often analytically intractible.

Various methods for numerical integration (trapezoid rule or Simpson's rule for example) are widely known and can be found in any good mathematics textbook. However, these approaches are only feasible in low dimensions. Many distributions of interest are in higher dimensions, so new methods for performing integrals become necessary. Monte Carlo integration is just such a method (Evans and Swartz, 2000).

Suppose we can draw a sample $x^{(1)}, x^{(2)}, ..., x^{(N)}$ from $g(x)$, then we can estimate $E_g(b)$ by

$$E_g[b(x)] \approx \bar{b}_N = \frac{1}{N}\sum_{i=1}^{N} b(x^{(i)}).$$

It can be shown that $\bar{b}_N$ approaches $E_g(b)$ for large N. This method is called Monte Carlo integration.

4.2.4 Markov Chain Monte Carlo

In high dimensions, the previous integration methods become difficult to use or fail completely. This is the point where iterative Monte Carlo integration becomes necessary and this method is called Markov Chain Monte Carlo. There are a number of good papers which explain these ideas more fully and in more general terms, for example Geyer (1992); Tierney (1994); Chib and Greenberg (1995).

The most general Markov Chain Monte Carlo algorithm is known as the Metropolis-Hastings algorithm, which is implemented as follows.

1. Start at any point, say $x^{(t)} = x$.

2. Generate y from $q(y \mid x)$ where y is a candidate point and q is the proposal or candidate distribution.

3. Calculate $\alpha(x, y) = min\left\{1, \frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}\right\}$.

4. $x^{t+1} = \begin{cases} y & \text{with probability } \alpha \\ x & \text{otherwise .} \end{cases}$

4.2.5 The Gibbs Sampler

Gibbs sampling is a special case of the Metropolis-Hastings algorithm where $q(y \mid x) = \prod \pi(y_i \mid x_{<i}, x_{>i}) \Rightarrow \alpha(x, y) = 1$, so the acceptance probability is always 1, candidates are always accepted. Thus, Gibbs sampling consists purely of sampling from full conditional distributions and there are many ways of sampling from the full conditional distributions including using the Metropolis-Hastings algorithm. The Gibbs sampler was given its name by Geman and Geman (1984) who first introduced MCMC into mainstream statistics. To date most statistical applications of MCMC have used Gibbs Sampling (W.R.Gilks et al., 1996).

For a discussion of the theory behind the Gibbs sampler see Casella and George (1992); Smith and Roberts (1993); Tierney (1994). The Gibbs sampler is implemented as follows.

Let $x = (x_1, ..., x_k)^T$ and $\pi(x)$ be a k-dimensional distribution. The algorithm is started at any point in the support of π, say x^0, then given $x^{(t)}$,

$$\begin{aligned}
x_1^{(t+1)} &\sim \pi(x_1 \mid x_2^{(t)}, ..., x_k^{(t)}) \\
x_2^{(t+1)} &\sim \pi(x_2 \mid x_1^{(t+1)}, x_3^{(t)}, ..., x_k^{(t)}) \\
&\vdots \qquad \vdots \qquad\qquad \vdots \\
x_k^{(t+1)} &\sim \pi(x_k \mid x_1^{(t+1)}, ..., x_{k-1}^{(t+1)}).
\end{aligned}$$

The densities on the right are the full conditional distributions and it should be noted that the most up-to-date version of x is always used.

This simple formula has a surprisingly rich set of applications. The idea behind the Gibbs sampler is that while we may not be able to find the posterior, if we treat each parameter separately and find the conditionals for each parameter, these conditional distributions may well be recognisable and we can then sample from them to approximate the posterior. Each conditional is sampled from in turn and the updated values are then used to sample from the next conditional.

4.2.6 Implementation and Convergence

Convergence of a Gibbs sampler to a stationary distribution can be slow, which is why it is customary to have a 'burn-in', where the first values generated by the sampler are ignored while it converges. There is a theoretical method for diagnosing convergence in some special cases, however, this is not possible in general and many different 'convergence diagnostic' methods are often used instead. Although these methods do not prove convergence, they are often felt to be an adequate check. For a further discussion of this subject see Gelman (1996); Sahu and Roberts (1997, 1999); Brooks and Roberts (1998).

While there are many convergence diagnostics available, in this thesis we restrict ourselves to those that are termed output diagnostics. In other words those techniques which monitor selected output from the Markov Chain itself and therefore require no extra calculation; this avoids the need to run more simulations than is necessary for the inference alone. In particular we examine convergence using trace plots and the CUSUM method (Yu and Mykland, 1998). We summarise the 'hairiness' of the CUSUM plot using the method of Brooks and Roberts (1998). We will present a full description of these particular techniques in Chapter 6 as it is convenient to discuss these methods at the same time as we present the examples of their use.

There is also debate on whether multiple short chains of the sampler should be used, or whether one long chain is more efficient, see Gelman and Rubin

(1992); Geyer (1992); Raferty and Lewis (1992). This is due to the fact that multiple short chains generated from a wide variety of initial starting values give an idea of how well the chain is 'mixing' ie. how distinguishable the chains are from one another. In the examples in this thesis we use one long chain because of the implementation issues. However, convergence occurs in a small number of iterations of the chain so this is not an issue.

4.2.7 Further Gibbs samplers

Gibbs samplers can be implemented only if a full conditional for all of the parameters can be found. However, with very complicated distributions this is not always possible. In these cases more advanced techniques must be used.

There are many different modifications of the Gibbs sampler. In this thesis we will be using a Gibbs sampler with a Metropolis-Hastings step incorporated within it to allow us to sample from some of the more complicated conditionals. An accept-reject Metropolis-Hastings step, performed in exactly the way outlined earlier in this chapter, is written in to the Gibbs sampler in place of the unknown conditionals. However, because we only need to generate one value per iteration of the Gibbs sampler, only one accept-reject step is performed in each loop of the Gibbs sampler.

For a detailed discussion of some of the theory behind this methodology see Gilks et al. (1995).

4.3 Using a Gibbs sampler in a Successive Sampling Discovery Model

This section is a summary of West (1994, 1996) which extends the methods described in Nair and Wang (1989) by using a Gibbs sampler to evaluate the integrals, thus making it unnecessary to evaluate the integrals analytically.

West (1996) starts with exactly the same density function as Nair and Wang (1989). However, in West's paper the data, $\mathbf{Y}$, is split into the observed part $\mathbf{D} = \mathbf{y}_1, .., \mathbf{y}_n$ and the unobserved part $\mathbf{U} = \mathbf{y}_{n+1}, .., \mathbf{y}_N$ and $R = \sum_{n+1}^{N} w(\mathbf{y}_i)$ is denoted by $t(U)$. This is an extension of the notation which will make the following derivations clearer because the known and unknown data can be separated. Adopting the notation of West (1996), we write the density function as,

$$[D \mid \boldsymbol{\theta}] \;=\; \frac{N!}{(N-n)!} \int \cdots \int \prod_{i=1}^{n} \frac{w(\mathbf{y}_i)}{t(U)+b_i} \prod_{i=1}^{N} [\mathbf{y}_i \mid \boldsymbol{\theta}] \; d\mathbf{y}_{n+1} \ldots d\mathbf{y}_N \;,$$

which is equivalent to equation (3.2). Unlike Nair and Wang (1989), West (1996) does not require the integrals to be calculated analytically therefore generalising the work of Nair and Wang (1989).

The Gibbs sampler is implemented as follows. The joint density function is written as,

$$[\mathbf{Y} \mid \boldsymbol{\theta}] \equiv [\mathbf{D}, \mathbf{U} \mid \boldsymbol{\theta}] = \left\{ \prod_{i=1}^{n} \frac{w(\mathbf{y}_i)}{(t(\mathbf{U})+b_i)} \right\} \prod_{i=1}^{N} f_{\boldsymbol{\theta}}(\mathbf{y}_i) \;. \tag{4.1}$$

The next step is to find the conditional densities for the random variables $\boldsymbol{\theta}$ and $\mathbf{U}$.

For known $\mathbf{U}$ (4.1) implies that,

$$[\boldsymbol{\theta} \mid \mathbf{D}, \mathbf{U}] \propto p(\boldsymbol{\theta}) \prod_{i=1}^{n} f_{\boldsymbol{\theta}}(\mathbf{y}_i), \tag{4.2}$$

where $p(\theta)$ denotes the prior for $\boldsymbol{\theta}$.

For known $\mathbf{D}$ and $\boldsymbol{\theta}$, (4.1) gives,

$$[\mathbf{U} \mid \mathbf{D}, \boldsymbol{\theta}] \propto \left\{ \prod_{i=1} (t(\mathbf{U})+b_i)^{-1} \right\} \prod_{i=n+1}^{N} f_{\boldsymbol{\theta}}(\mathbf{y}_i).$$

It is clear that $\mathbf{U}$ appears in complicated ways throughout this expression, so at this point West (1996) uses augmentation to simplify the conditional distributions so that it can be sampled from.

Let $(t(\mathbf{U}) + b_i)^{-1} = \int_0^\infty \exp(-(t(\mathbf{U}) + b_i)\phi_i) d\phi_i$ for each i. Then,

$$[\mathbf{U} \mid \boldsymbol{\theta}, \mathbf{D}] \propto \left\{ \prod_{i=1}^{n} \int_0^\infty \exp(-(t(\mathbf{U}) + b_i)\phi_i) d\phi_i \right\} \prod_{i=n+1}^{N} f_{\boldsymbol{\theta}}(\mathbf{y}_i). \tag{4.3}$$

Let $\boldsymbol{\phi} = (\phi_1, ..., \phi_n)$, then (4.3) is the marginal density for $\mathbf{U}$ from a joint density for $(\mathbf{U}, \boldsymbol{\phi} \mid \boldsymbol{\theta}, \mathbf{D})$ with the following conditional distributions:

$$[\boldsymbol{\phi} \mid \boldsymbol{\theta}, \mathbf{U}, \mathbf{D}] \propto \prod_{i=1}^{n} \exp(-(t(\mathbf{U}) + b_i)\phi_i), \tag{4.4}$$

so the ϕ_i are conditionally independent exponential random variables.

$[\mathbf{U} \mid \boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{D}] \propto \{\prod_{i=1}^{n} \exp(-t(u)\phi_i)\} \prod_{i=n+1}^{N} f_{\boldsymbol{\theta}}(\mathbf{y}_i)$.

Define $r = \sum_{i=1}^{n} \phi_i$ then with some rearrangement we have,

$$[\mathbf{U} \mid \boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{D}] \propto \prod_{i=n+1}^{N} \exp(-rw(\mathbf{y}_i)) f_{\boldsymbol{\theta}}(\mathbf{y}_i). \tag{4.5}$$

We now have full conditional distributions which can be used within a Gibbs sampler.

To sample for θ, first transform the $\mathbf{y}$ values, defining $\mathbf{x}$ to be the p-vector whose elements are the natural logs of the corresponding $\mathbf{y}$. In this case $p = 4$ as there are four measurements taken for each $\mathbf{y}$, the volume, area, net-pay and depth. We assume that $x \sim N(\mu, \Sigma)$ is a p-variate normal distribution whose mean and variance determine the parameter $\boldsymbol{\theta} = \{\mu, \Sigma\}$. So, as in Nair and Wang (1989), $f_{\boldsymbol{\theta}}(y)$ is a multivariate log-normal density. Also, following Nair and Wang (1989) assume the weighting of the sampled units is log-linear with $w(y) = e^{a'\mathbf{x}}$ for some p-vector a. In this case the vector

used is $a' = (0, 0.84, 0.82, -2.68)$. In this special case conjugate priors for θ lie in the normal-inverse Wishart class (see Press (1985) Sec.7.1.6), so let $\bar{\mathbf{x}} = \sum_{i=1}^{N} \frac{\mathbf{x}}{N}$ and $S = \sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$; then

$$\mu \mid \Sigma \sim N(\bar{\mathbf{x}}, \Sigma/N) \tag{4.6}$$

and

$$\Sigma \sim W^{-1}(S, p, N + p)\,. \tag{4.7}$$

To sample for ϕ, we sample from an exponential distribution with parameter $(t(U) + b_i)$.

Finally we need a method for sampling $\mathbf{U}$. The general method will be to use rejection sampling. We note that $e^{-rw(\mathbf{y})} = P(x > rw(y) \mid \mathbf{y})$, where $[\mathbf{x} \mid \mathbf{y}] \sim Ex(w(y))$; then we sample from $f_{\boldsymbol{\theta}}(\mathbf{y})$ and $u \sim U(0, 1)$. If $\log(1 - u) > -rw(\mathbf{y})$ we reject $\mathbf{y}$ otherwise we accept $\mathbf{y}$ as one of the sampled $\mathbf{y}$'s. In this case $\mathbf{y}$ is a vector, so for ease of computation, we first map $z = a'\mathbf{x}$ and find the distribution for z. We then generate z and transform back to the $\mathbf{y}$'s, vastly reducing the computation involved in the Gibbs sampler. (Finding the distributions needed here is unnecessary in this overview so the details are omitted. For full calculations see West (1996).)

The Gibbs sampler is constructed as follows:

(a) Choose an initial U and compute $t(U)$.

(b) Sample $\boldsymbol{\theta}$ from (4.6) and (4.7).

(c) Sample the exponentials from (4.4) conditional on the current $\mathbf{U}$ and calculate $r = \sum_{i=1}^{n} \phi_i$.

(d) Draw U as outlined above using the current values of r and $\boldsymbol{\theta}$.

(e) Iterate from (b).

Discussion

This methodology can be implemented to give estimates for population totals and to generate missing data. The Bayesian approach also gives us a method which is more versatile that that of Nair and Wang (1989) because the Gibbs sampler allows us to use density functions where the integrals cannot be determined analytically or by using the EM algorithm.

Chapter 5

Applications of Breckling et al. (1994) using both a Frequentist and Bayesian Approach

The aim of the following sections is to apply some of the theory discussed in Breckling et al. (1994) and chapter 2 to some new examples to gain insight into the way the method can be applied to more complex situations.

5.1 Bernoulli model with informative sampling

Let N be the total number of units and n be the number of units sampled. Let $X_1, ..., X_N$ be independent and identically distributed Bernoulli random variables with parameter α. Let $\mathbf{X}$ be a vector containing $X_1, ..., X_N$. We denote the realisations of these random variables by $\mathbf{x}$ and $x_1, ..., x_N$ respectively. As in the notation used in design-based sampling (Thompson, 1992) we allow the vector $\mathbf{X}$ to be ordered into the sampled and non-sampled sections respectively, so $\mathbf{X} = (\mathbf{X}_s, \mathbf{X}_r)^T$.

We will assume that the sample inclusion indicators are denoted by a vector $\mathbf{I}$, where given $\mathbf{X} = \mathbf{x}$, each component I_i is an independent Bernoulli random variable. So if $I_i = 1$ the ith unit is contained within the sample and if $I_i = 0$ the ith unit is not contained within the sample. (The realisations of these sample inclusion indicators are not denoted by lower case I's because i is generally associated with an index set for the realisations and dual use can be unclear.)

We allow the selection probabilities to depend on X_i as

$$P(I_i = 1|X_i = x_i) \quad = \quad \pi(x_i) \quad ,$$

for some $\pi(\cdot)$. In fact, as X_i is binary, there are only two values for $\pi(\cdot)$, namely

$$P(I_i = 1|X_i = 1) \quad = \quad \pi(1)$$

and

$$P(I_i = 1|X_i = 0) \quad = \quad \pi(0) \; .$$

We will assume initially that the sampling probabilities $\pi(0)$ and $\pi(1)$ are known.

5.1.1 Known Sampling Probabilities

We will summarise the model as follows, remembering that the $\mathbf{X}_i$'s are independent.

$$[X] \quad = \alpha^{\sum X_i}(1-\alpha)^{\sum(1-X_i)}$$

and

$$[I_i = I|X_i = x] \quad = \pi(x)^I(1-\pi(x))^{1-I} \; .$$

It is natural to construct the model in this order, with I dependent on $\mathbf{X}$, because the sampling probability is dependent on the variable. It is less

intuitive to formulate the model in terms of the variable $\mathbf{X}_i$ being dependent on the inclusion indicator I_i even though this makes mathematical sense. The likelihood is

$$\exp(l(\alpha)) = \alpha^{\sum X_i}(1-\alpha)^{\sum(1-X_i)} \prod\prod \pi(X_i)^{I_i}(1-\pi(X_i))^{(1-I_i)}, \quad (5.1)$$

where the sums and product are over $i = 1, ..., N$, and the log likelihood is

$$\begin{aligned} l(\alpha) &= \sum_{i=1}^{N} X_i \log(\alpha) + \sum_{i=1}^{N}(1-X_i)\log(1-\alpha) \\ &\quad + \sum_{i=1}^{N} I_i \log(\pi(X_i)) + \sum_{i=1}^{N}(1-I_i)\log(1-\pi(X_i)) \, . \end{aligned} \quad (5.2)$$

The first step is to find the maximum likelihood estimator for α. We will do this in exactly the same stages as explained earlier (see also Breckling et al. (1994)). The population score function is the derivative of the log-likelihood,

$$\frac{\partial l(\alpha)}{\partial \alpha} = \frac{\sum_{i=1}^{N} X_i}{\alpha} - \frac{\sum_{i=1}^{N}(1-X_i)}{1-\alpha} \, .$$

This is then split into the sample and non-sample contributions,

$$\begin{aligned} \frac{\partial l(\alpha)}{\partial \alpha} &= \frac{\sum_{i=1}^{n} X_i}{\alpha} - \frac{\sum_{i=1}^{n}(1-X_i)}{1-\alpha} \\ &\quad + \frac{\sum_{i=n+1}^{N} X_i}{\alpha} - \frac{\sum_{i=n+1}^{N}(1-X_i)}{1-\alpha} \, . \end{aligned}$$

We still have a function which involves the non-sampled, or unknown, elements so the next stage is to find the expectation of the unknown parts given the sample. To do this we must first find the distribution of $\mathbf{X}_r$. The X_i's are independent, therefore the distribution of $\mathbf{X}_r$ is equal to the product of

the individual distributions of the X_i's for all $i \in r$. It is therefore enough to find the distribution for a single $X_i \in \mathbf{X}_r$, which is equivalent to the distribution of $X_i \mid I_i = 0$ for some $i \in r$. The distribution of $X_i \mid I_i = 1$ is unnecessary at this stage because this is part of the sample, which is known.

The joint density function is constructed as follows:

$$[X_i, I_i] \quad = \alpha^{x_i}(1-\alpha)^{1-x_i}\pi(x_i)^{I_i}(1-\pi(x_i))^{1-I_i} \quad .$$

We can therefore find the density for I_i by summing over all possible values of X_i,

$$\begin{aligned}[I_i] \quad &= \sum_x \alpha^x(1-\alpha)^{1-x}\pi(x)^{I_i}(1-\pi(x))^{1-I_i} \\ &= (1-\alpha)\,\pi(0)^{I_i}(1-\pi(0))^{1-I_i} \\ &\qquad +\alpha\,\pi(1)^{I_i}(1-\pi(1))^{1-I_i} \quad .\end{aligned}$$

The density of $[X_i \mid I_i = 0]$ is obtained by simply substituting $I = 0$ into the relevant parts of the expression

$$[X_i|I_i] = \frac{[X_i, I_i]}{[I_i]} \ ,$$

so

$$[X_i|I_i = 0] \quad = \quad \frac{\alpha^{x_i}(1-\alpha)^{1-x_i}(1-\pi(x_i))}{(1-\alpha)(1-\pi(0)) + \alpha(1-\pi(1))} \ .$$

Equivalently we can write,

$$[X_i = 0|I_i = 0] \quad = \frac{(1-\alpha)(1-\pi(0))}{(1-\alpha)(1-\pi(0)) + \alpha(1-\pi(1))} \ ,$$

and

$$[X_i = 1|I_i = 0] \quad = \frac{\alpha(1-\pi(1))}{(1-\alpha)(1-\pi(0)) + \alpha(1-\pi(1))} \ .$$

Thus, $X_i \mid I_i = 0$ has a Bernoulli distribution,

$$X_i|I_i = 0 \sim Bernoulli \left(\frac{\alpha(1-\pi(1))}{(1-\alpha)(1-\pi(0)) + \alpha(1-\pi(1))} \right) .$$

The expected values we require are,

$$\begin{aligned} E(X_i|I=0) &= \sum_{0,1} X_i[X_i|I_i = 0] \\ &= \frac{\alpha(1-\pi(1))}{(1-\alpha)(1-\pi(0)) + \alpha(1-\pi(1))} , \end{aligned} \tag{5.3}$$

and

$$\begin{aligned} E(1-X_i|I=0) &= \sum_{0,1} (1-X_i)[X_i|I_i = 0] \\ &= \frac{(1-\alpha)(1-\pi(0))}{(1-\alpha)(1-\pi(0)) + \alpha(1-\pi(1))} , \end{aligned} \tag{5.4}$$

so we can write the sample score in full, as

$$\begin{aligned} sc_s(\alpha) &= \frac{\sum_{i=1}^n X_i}{\alpha} - \frac{\sum_{i=1}^n (1-X_i)}{1-\alpha} \\ &\quad + \frac{(N-n)(1-\pi(1))}{(1-\alpha)(1-\pi(0)) + \alpha(1-\pi(1))} \\ &\quad - \frac{(N-n)(1-\pi(0))}{(1-\alpha)(1-\pi(0)) + \alpha(1-\pi(1))} \\ &= \frac{\sum_{i=1}^n X_i}{\alpha} - \frac{\sum_{i=1}^n (1-X_i)}{1-\alpha} \\ &\quad + \frac{(N-n)(\pi(0)-\pi(1))}{(1-\alpha)(1-\pi(0)) + \alpha(1-\pi(1))} . \end{aligned}$$

The parameter α is estimated by setting the score function equal to zero and then solving for α.

The next step is to ensure that the sample information function is positive definite (Breckling et al., 1994).

We take the second derivative of the log-likelihood function

$$\frac{\partial^2 l(\alpha)}{\partial \alpha^2} = \frac{-\sum_{i=1}^{N} X_i}{\alpha^2} - \frac{\sum_{i=1}^{N}(1-X_i)}{(1-\alpha)^2} \quad ,$$

then the information function is the negative of the second derivative,

$$Info(\alpha) = \frac{\sum_{i=1}^{N} X_i}{\alpha^2} + \frac{\sum_{i=1}^{N}(1-X_i)}{(1-\alpha)^2} \quad .$$

The expectation of the information with respect to the unknown X_i's is,

$$\begin{aligned} E(\frac{\partial^2 l(\alpha)}{\partial \alpha^2}) &= \frac{\sum_{i=1}^{n} X_i}{\alpha^2} + \frac{\sum_{i=1}^{n}(1-X_i)}{(1-\alpha)^2} \\ &\quad + \frac{(N-n)(1-\pi(1))}{\alpha((1-\alpha)(1-\pi(0)) + \alpha(1-\pi(1)))} \\ &\quad + \frac{(N-n)(1-\pi(0))}{(1-\alpha)((1-\alpha)(1-\pi(0)) + \alpha(1-\pi(1)))} \quad . \end{aligned}$$

We also need to find the variance of the unknown part of the score function (the variance of the known part is of course zero), so

$$\begin{aligned} var(sc_s(\alpha)) &= \sum_{i=n+1}^{N} \left(var_s \left(\frac{X_i}{\alpha} - \frac{(1-X_i)}{(1-\alpha)} \right) \right) \\ &= \sum_{i=n+1}^{N} \left(E_s \left\{ \frac{X_i}{\alpha} - \frac{(1-X_i)}{(1-\alpha)} \right\}^2 \right) \\ &\quad - \left\{ E_s \left(\frac{X_i}{\alpha} - \frac{(1-X_i)}{(1-\alpha)} \right) \right\}^2 . \end{aligned}$$

The variance after the expectations have been taken is written,

$$\begin{aligned} var_s(sc) &= (N-n) \left[\frac{(1-\pi(1))}{\alpha((1-\alpha)(1-\pi(0)) + \alpha(1-\pi(1)))} \right. \\ &\quad + \frac{(1-\pi(0))}{(1-\alpha)((1-\alpha)(1-\pi(0)) + \alpha(1-\pi(1)))} \\ &\quad - \left\{ \frac{(1-\pi(1))}{(1-\alpha)(1-\pi(0)) + \alpha(1-\pi(1))} \right. \\ &\quad \left. \left. - \frac{(1-\pi(0))}{(1-\alpha)(1-\pi(0)) + \alpha(1-\pi(1))} \right\}^2 \right] . \end{aligned}$$

The sample information function is then formed directly from these equations,

$$\begin{aligned} Info_s(\alpha) &= \frac{\sum_{i=1}^{n} X_i}{\alpha^2} + \frac{\sum_{i=1}^{n}(1-X_i)}{(1-\alpha)^2} \\ &\quad +(N-n)\left\{\frac{(\pi(0)-\pi(1))^2}{\{(1-\alpha)(1-\pi(0))+\alpha(1-\pi(1))\}^2}\right\}, \end{aligned}$$

which is positive for all values of α.

The score equation can be solved using the usual formula to solve quadratic expressions to find a value for the parameter α, so

$$\hat{\alpha} = \frac{-b-\sqrt{b^2-4ac}}{2a}, \tag{5.5}$$

where

$$\begin{aligned} a &= N(\pi(1)-\pi(0)) , \\ b &= \sum_{i=1}^{n} X_i(\pi(0)-\pi(1)) - n(1-\pi(0)) + (N-n)(\pi(0)-\pi(1)) , \\ c &= (1-\pi(0))\sum_{i=1}^{n} X_i . \end{aligned}$$

The other possible solution to the quadratic gives a value of $\hat{\alpha}$ which lies outside the parameter space.

It is interesting to note, that if $\pi(0) = \pi(1)$ the score is not quadratic in α in this case it can be shown that the solution is

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} X_i}{n} .$$

In other words if $\pi(0) = \pi(1)$ the maximum likelihood estimate for α is the expansion estimator as we might expect.

5.1.2 Unknown Sampling Probabilities

Now suppose that the sampling probabilities are unknown. The variable $\mathbf{X}$ will be defined in the same way and the likelihood is the same as (5.2). However, the values of $\pi(x_i)$ are now unknown, therefore $\pi(0)$ and $\pi(1)$ will become parameters in the log-likelihood. We re-write (5.2) here for clarity.

$$\begin{aligned} l(\alpha, \pi_0, \pi_1) &= \sum_{i=1}^{N} X_i \log(\alpha) + \sum_{i=1}^{N} (1 - X_i) \log(1 - \alpha) \\ &\quad + \sum_{i=1}^{N} I_i \log(\pi(X_i)) + \sum_{i=1}^{N} (1 - I_i) \log(1 - \pi(X_i)) \, . \end{aligned}$$

The score function for α will remain as before but will now include the score function for π. The full score function is now a vector of length 3.

The first step, as before, is to differentiate the log-likelihood with respect to the parameters and then separate out the function into the sampled and non-sampled parts. We will leave out the calculations for α as these are identical to those given above. So,

$$\begin{aligned} \frac{\partial l}{\partial \pi(0)} &= \frac{\sum_{i=1}^{N} I_i (1 - X_i)}{\pi(0)} - \sum_{i=1}^{N} \left(\frac{(1 - I_i)(1 - X_i) \pi(1)^{X_i} \pi(0)^{-X_i}}{1 - \pi(1)^{X_i} \pi(0)^{(1 - X_i)}} \right) \quad \text{and} \\ \frac{\partial l}{\partial \pi(1)} &= \frac{\sum_{i=1}^{N} I_i X_i}{\pi(1)} - \sum_{i=1}^{N} \left(\frac{(1 - I_i) X_i \, \pi(0)^{(1 - X_i)} \pi(1)^{(X_i - 1)}}{1 - \pi(1)^{X_i} \pi(0)^{(1 - X_i)}} \right) , \end{aligned}$$

for $i = 1, ..., n$, $I_i = 0$ and for $I = n + 1, ..., N$, $I_i = 1$ by definition. So, written with the sample and non-sample parts separated, we have

$$\begin{aligned} \frac{\partial l}{\partial \pi(0)} &= \frac{\sum_{i=1}^{n} (1 - X_i)}{\pi(0)} - \sum_{i=n+1}^{N} \left(\frac{(1 - X_i) \pi(1)^{X_i} \pi(0)^{-X_i}}{1 - \pi(1)^{X_i} \pi(0)^{(1 - X_i)}} \right) \quad \text{and} \\ \frac{\partial l}{\partial \pi(1)} &= \frac{\sum_{i=1}^{n} X_i}{\pi(1)} - \sum_{i=n+1}^{N} \left(\frac{X_i \, \pi(0)^{(1 - X_i)} \pi(1)^{(X_i - 1)}}{1 - \pi(1)^{X_i} \pi(0)^{(1 - X_i)}} \right) . \end{aligned}$$

The sample score is then calculated by taking the expectation of the non-sampled part,

$$\begin{aligned}
\frac{\partial l_s}{\partial \pi(0)} &= \frac{\sum_{i=1}^{n}(1-X_i)}{\pi(0)} \\
&\quad - \sum_{i=n+1}^{N}\left(\frac{1}{1-\pi(0)} \times \frac{(1-\alpha)(1-\pi(0))}{(1-\alpha)(1-\pi(0))+\alpha(1-\pi(1))}\right) \text{ and} \\
\frac{\partial l_s}{\partial \pi(1)} &= \frac{\sum_{i=1}^{n} X_i}{\pi(1)} \\
&\quad - \sum_{i=n+1}^{N}\left(\frac{1}{1-\pi(1)} \times \frac{\alpha(1-\pi(1))}{(1-\alpha)(1-\pi(0))+\alpha(1-\pi(1))}\right).
\end{aligned}$$

The complete sample score function simplifies to

$$sc_s = \begin{pmatrix} \frac{\sum_{i=1}^{n} X_i}{\alpha} - \frac{\sum_{i=1}^{n}(1-X_i)}{1-\alpha} + \frac{(N-n)(\pi(0)-\pi(1))}{(1-\alpha)(1-\pi(0))+\alpha(1-\pi(1))} \\ \frac{\sum_{i=1}^{n}(1-X_i)}{\pi(0)} - \frac{(N-n)(1-\alpha)}{(1-\alpha)(1-\pi(0))+\alpha(1-\pi(1))} \\ \frac{\sum_{i=1}^{n} X_i}{\pi(1)} - \frac{\alpha(N-n)}{(1-\alpha)(1-\pi(0))+\alpha(1-\pi(1))} \end{pmatrix}.$$

This score equation cannot be solved for a unique solution for all three parameters because the equations are reducible. If we multiply the third equation by $\frac{\pi(1)}{\alpha}$ and the second equation by $\frac{\pi(0)}{1-\alpha}$ and then subtract the second equation from the third equation we obtain the first equation. This implies that there are an infinite set of solutions to this set of equations.

It is still possible find a solution for $\pi(0)$ and $\pi(1)$ in terms of α (the equations are not presented here due to their complexity) and α in terms of $\pi(0)$ and $\pi(1)$ as before. Also as before if $\pi(0) = \pi(1)$ we can find a solution which turns out to be the expansion estimator for α.

The fact that this set of equations cannot be solved for a unique solution can be explained if we consider what the parameters represent. If we obtain

a low number of ones in the sample there are two equally likely explanations for this; we have a low value of α in other words the probability in the Bernoulli distribution is low or we have a high value for α but a low value for $\pi(1)$. These two cases are indistinguishable.

The likelihood function can be re-parameterized as follows. Let $\omega_1 = \alpha\pi(1)$ and $\omega_2 = (1-\alpha)\pi(1)$. Then

$$\exp(l(\alpha)) \quad = \quad (\omega_1)^{\sum_{i=1}^n X_i}[\omega_2]^{\sum_{i=1}^n (1-X_i)}[1-\omega_1-\omega_2]$$

The log-likelihood is exactly the same as before, however, we can now solve for ω_1 and ω_2.

$$\omega_1 \quad = \quad \frac{\sum_{i=1}^n X_i}{N}$$

$$\omega_2 \quad = \quad \frac{\sum_{i=1}^n (1-X_i)}{N}$$

This reinforces the point that it is the fact that α and $\pi(1)$ are so closely related that causes the problem in solving for all three parameters.

The calculations for the information function are similar to those carried out in the previous example and are not given in full here. We obtain

$$Inf = -\begin{pmatrix} I_{11} & I_{12} & I_{13} \\ I_{21} & I_{22} & I_{23} \\ I_{31} & I_{32} & I_{33} \end{pmatrix},$$

where

$$I_{11} = -\frac{\sum_{i=1}^n X_i}{\alpha^2} - \frac{\sum_{i=1}^n (1-X_i)}{(1-\alpha)^2} - \frac{(N-n)(\pi(0)-\pi(1))^2}{\{(1-\alpha)(1-\pi(0)) + \alpha(1-\pi(1))\}^2},$$

$$I_{12} = \frac{(N-n)(1-\pi(1))}{\{(1-\alpha)(1-\pi(0)) + \alpha(1-\pi(1))\}^2},$$

$$I_{13} = -\frac{(N-n)(1-\pi(0))}{\{(1-\alpha)(1-\pi(0))+\alpha(1-\pi(1))\}^2},$$

$$I_{21} = \frac{(N-n)(1-\pi(1))}{\{(1-\alpha)(1-\pi(0))+\alpha(1-\pi(1))\}^2},$$

$$I_{22} = -\frac{\sum_{i=1}^{n} X_i}{\pi(0)^2} - \frac{(N-n)(1-\alpha)^2}{\{(1-\alpha)(1-\pi(0))+\alpha(1-\pi(1))\}^2},$$

$$I_{23} = -\frac{\alpha(N-n)(1-\alpha)}{\{(1-\alpha)(1-\pi(0))+\alpha(1-\pi(1))\}^2},$$

$$I_{31} = -\frac{(N-n)(1-\pi(0))}{\{(1-\alpha)(1-\pi(0))+\alpha(1-\pi(1))\}^2},$$

$$I_{32} = -\frac{\alpha(N-n)(1-\alpha)}{\{(1-\alpha)(1-\pi(0))+\alpha(1-\pi(1))\}^2} \quad \text{and}$$

$$I_{33} = -\frac{\sum_{i=1}^{n} X_i}{\pi(1)^2} - \frac{(N-n)\alpha^2}{\{(1-\alpha)(1-\pi(0))+\alpha(1-\pi(1))\}^2}.$$

At this point it is clear that it would be difficult to proceed with this model as the non-identifiability problem will remain. However, to understand exactly where the inference problem occurs we will return to the fully identifiable case and add in an independent variable to ascertain that it is the dependence structure between the parameters α and the π values which causes the problems.

5.1.3 Two independent variables

Assuming that π_0 and π_1 are known, we include a simple continuous exponential variable as this changes the calculations from summations to integrals which on the whole can be performed more easily.

We will label the second dependent variable $\mathbf{Y}$ and assume independence throughout, so

$$\begin{aligned}
[I_i|X_i = x_i] &= \pi(x_i)^{I_i}(1-\pi(x_i))^{1-I_i}\,, \\
[X_i] &= \alpha^{x_i}(1-\alpha)^{1-x_i}\,, \\
[Y_i] &= \frac{1}{\lambda}e^{\frac{-y_i}{\lambda}}\,.
\end{aligned}$$

The joint distribution of X_i, Y_i and I_i is

$$\begin{aligned}
[X_i, Y_i, I_i] &= \pi(x_i)^{I_i}(1-\pi(x_i))^{1-I_i}\alpha^{x_i}(1-\alpha)^{1-x_i}\frac{1}{\lambda}e^{\frac{-y_i}{\lambda}}\,, \\
[I_i] &= \sum_{x_i}\int_0^\infty \pi(x_i)^{I_i}(1-\pi(x_i))^{1-I_i}\alpha^{x_i}(1-\alpha)^{1-x_i}\frac{1}{\lambda}e^{\frac{-y_i}{\lambda}}dy_i \\
&= \int_0^\infty \pi(0)^{I_i}(1-\pi(0))^{1-I_i}(1-\alpha)\frac{1}{\lambda}e^{\frac{-y_i}{\lambda}}dy_i \\
&\quad + \int_0^\infty \pi(1)^{I_i}(1-\pi(1))^{1-I_i}\alpha\frac{1}{\lambda}e^{\frac{-y_i}{\lambda}}dy_i\,.
\end{aligned}$$

The conditional distribution of X_i and Y_i given $I_i = 0$ is

$$[X_i, Y_i|I_i = 0] = \frac{(1-\pi(x_i))\alpha^{x_i}(1-\alpha)^{(1-x_i)}\frac{1}{\lambda}e^{\frac{-y_i}{\lambda}}}{(1-\alpha)\{1-\pi(0)\}\int_0^\infty \frac{1}{\lambda}e^{\frac{-y_i}{\lambda}}dy_i + \alpha\{1-\pi(1)\}\int_0^\infty \frac{1}{\lambda}e^{\frac{-y_i}{\lambda}}dy_i}\,.$$

We can use the above density to find the expected values needed to evaluate the score function as follows,

$$\begin{aligned}
E(X_i|I_i = 0) &= \frac{\alpha(1-\pi(1))\int_0^\infty \frac{1}{\lambda}e^{\frac{-y_i}{\lambda}}dy_i}{(1-\alpha)\int_0^\infty\{1-\pi(0)\}\frac{1}{\lambda}e^{\frac{-y_i}{\lambda}}dy_i + \alpha\int_0^\infty\{1-\pi(1)\}\frac{1}{\lambda}e^{\frac{-y_i}{\lambda}}dy_i} \\
&= \frac{\alpha(1-\pi(1))}{(1-\alpha)\{1-\pi(0)\}+\alpha\{1-\pi(1)\}}\,.
\end{aligned}$$

This is identical to (5.3) as we would expect because of the independence in the model.

We now must also find the expectation for Y_i, so

$$
\begin{aligned}
&E(Y_i|I_i=0) \\
&= \int_0^\infty \left\{ \frac{y_i \frac{1}{\lambda} e^{\frac{-y_i}{\lambda}} (1-\alpha)(1-\pi(0))}{(1-\alpha)\{1-\pi(0)\}\int_0^\infty \frac{1}{\lambda} e^{\frac{-y_i}{\lambda}} dy_i + \alpha\{1-\pi(1)\}\int_0^\infty \frac{1}{\lambda} e^{\frac{-y_i}{\lambda}} dy_i} \right. \\
&\quad \left. + \frac{\alpha y_i \frac{1}{\lambda} e^{\frac{-y_i}{\lambda}} (1-\pi(1))}{(1-\alpha)\{1-\pi(0)\}\int_0^\infty \frac{1}{\lambda} e^{\frac{-y_i}{\lambda}} dy_i + \alpha\{1-\pi(1)\}\int_0^\infty \frac{1}{\lambda} e^{\frac{-y_i}{\lambda}} dy_i} \right\} dy_i \,,
\end{aligned}
$$

which can be evaluated using integration by parts to

$$
\begin{aligned}
E(Y_i|I_i=0) \quad = \qquad &= \frac{\lambda(1-\alpha)(1-\pi(0))}{(1-\alpha)\{1-\pi(0)\}+\alpha\{1-\pi(1)\}} \\
&\quad + \frac{\lambda\alpha(1-\pi(1))}{(1-\alpha)\{1-\pi(0)\}+\alpha\{1-\pi(1)\}} \,.
\end{aligned}
$$

The log-likelihood function is written as follows

$$
\begin{aligned}
l(\alpha,\lambda) \quad = \quad & \sum_{i=1}^{N} X_i \log(\alpha) + \sum_{i=1}^{N} (1-X_i)\log(1-\alpha) + \sum_{i=1}^{N} I_i \log \pi(X) \\
& + \sum_{i=1}^{N} (1-I_i)\log(1-\pi(X)) + N \, \log(\frac{1}{\lambda}) - \sum_{i=1}^{N} \frac{Y_i}{\lambda} \quad .
\end{aligned}
$$

This is used to find the population score function ,

$$\begin{aligned}
\frac{\partial \log(l)}{\partial \alpha} &= \frac{\sum_{i=1}^{N} X_i}{\alpha} - \frac{\sum_{i=1}^{N}(1-X_i)}{(1-\alpha)} \\
&= \frac{\sum_{i=1}^{n} X_i}{\alpha} - \frac{\sum_{i=1}^{n}(1-X_i)}{(1-\alpha)} + \frac{\sum_{i=n+1}^{N} X_i}{\alpha} \\
&\qquad - \frac{\sum_{i=n+1}^{N}(1-X_i)}{(1-\alpha)} \\
\frac{\partial \log(l)}{\partial \lambda} &= -\frac{N}{\lambda} + \frac{\sum_{i=1}^{N} Y_i}{\lambda^2} \\
&= -\frac{N}{\lambda} + \frac{\sum_{i=1}^{n} Y_i}{\lambda^2} + \frac{\sum_{i=n+1}^{N} Y_i}{\lambda^2} \quad .
\end{aligned}$$

The sample score function for these two parameters is then

$$sc_s = \begin{pmatrix} \frac{\sum_{i=1}^{n} X_i}{\alpha} - \frac{\sum_{i=1}^{n}(1-X_i)}{1-\alpha} \\ \qquad +(N-n)\left\{\frac{(\pi(0)-\pi(1))}{(1-\alpha)\{1-\pi(0)\}+\alpha\{1-\pi(1)\}}\right\} \\ \\ \frac{-N}{\lambda} + \frac{\sum_{i=1}^{n} Y_i}{\lambda^2} + \frac{(N-n)}{\lambda}\left\{\frac{(1-\alpha)(1-\pi(0))}{(1-\alpha)\{1-\pi(0)\}+\alpha\{1-\pi(1)\}}\right. \\ \qquad\qquad \left. + \frac{\alpha(1-\pi(1))}{(1-\alpha)\{1-\pi(0)\}+\alpha\{1-\pi(1)\}}\right\} \end{pmatrix} .$$

This system of equations can be solved for each parameter allowing us to conclude that it is indeed the dependence structure between the parameters α and the π values which causes the identifiability problems. The sample score can then be used to find the Information function in the usual way.

$$Inf = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix}$$

$$
\begin{aligned}
I_{11} &= \frac{\sum_{i=1}^{n} X_i}{\alpha^2} + \frac{\sum_{i=1}^{n}(1-X_i)}{(1-\alpha)^2} + \frac{(N-n)(\pi(0)-\pi(1))^2}{\{(1-\alpha)(1-\pi(0)) + \alpha(1-\pi(1))\}^2} \\
I_{12} &= 0 \\
I_{21} &= 0 \\
I_{22} &= -\frac{N}{\lambda^2} + \frac{2\sum_{i=1}^{n} Y_i}{\lambda^3} + \frac{(N-n)}{\lambda^2}\left\{\frac{(1-\alpha)(1-\pi(0))}{(1-\alpha)\{1-\pi(0)\} + \alpha\{1-\pi(1)\}}\right. \\
&\qquad \left. - \frac{\alpha(1-\pi(1))}{(1-\alpha)\{1-\pi(0)\} + \alpha\{1-\pi(1)\}}\right\}
\end{aligned}
$$

5.2 Analysis of the Bernoulli model with informative sampling, using a Gibbs sampler

We will go on to use a Gibbs sampler to further explore the models in section (5.1). The Gibbs sampler should produce similar estimates to $\hat{\alpha}$ for the Bernoulli model with known sampling probabilities and fail to produce good estimates for the model with unknown sampling probabilities because the maximum likelihood estimators of the parameters cannot be found.

5.2.1 Known Sampling Probabilities

We construct the Gibbs sampler using the likelihood (5.1) as follows.

$$
\begin{aligned}
&[X_1, ..., X_N, \alpha, \pi(0), \pi(1)] \\
&\qquad = \alpha^{\sum_{i=1}^{N} X_i}(1-\alpha)^{\sum_{i=1}^{N}(1-X_i)} \prod_{i=1}^{N} \pi(X_i)^{I_i}(1-\pi(X_i))^{(1-I_i)}.
\end{aligned}
$$

The likelihood for α is proportional to a Beta density, so we will use the Beta(γ, δ) distribution as a prior. The conditional distribution is constructed

as follows,

$$
\begin{aligned}
&[\alpha \mid X_1, ..., X_N, \pi(0), \pi(1)] \\
&\qquad \propto prior(\alpha) \times \alpha^{\sum_{i=1}^{N} X_i}(1-\alpha)^{\sum_{i=1}^{N}(1-X_i)} \\
&\qquad \propto \frac{\Gamma(\gamma+\delta)}{\Gamma(\gamma)\Gamma(\delta)}\alpha^{(\gamma-1)}(1-\alpha)^{(\delta-1)} \\
&\qquad\qquad \times \alpha^{\sum_{i=1}^{N} X_i}(1-\alpha)^{\sum_{i=1}^{N}(1-X_i)} \\
&\qquad \propto \alpha^{\sum_{i=1}^{N} X_i+\gamma-1}(1-\alpha)^{\sum_{i=1}^{N}(1-X_i)+\delta-1} \\
&\qquad \propto Beta\left(\sum_{i=1}^{N} X_i + \gamma - 1, \sum_{i=1}^{N}(1-X_i) + \delta - 1\right).
\end{aligned}
\tag{5.6}
$$

The distribution for the unknown X's has already been calculated in section (5.1),

$$
[X_i \mid \alpha, \pi(0), \pi(1), I = 0] \propto Bernoulli\left(\frac{\alpha(1-\pi(1))}{(1-\alpha)(1-\pi(0)) + \alpha(1-\pi(1))}\right). \tag{5.7}
$$

From these two conditionals we can construct a Gibbs sampler to predict an estimated value for α. To initialise the samplers we will generate populations with underlying parameter values from $\alpha = 0.1$ to $\alpha = 0.5$ with an increment of 0.1 and sample from these populations using given values of $\pi(0)$ and $\pi(1)$ again each ranging from 0.1 to 0.5. The Gibbs sampler will be initialised from the sample.

We let $\gamma = 2$ and $\delta = 5$ in the sampler because the mean of the prior distribution is then in the centre of the possible range of α and when compared to other possible priors it showed the best fit for the model. Several other possible parameter values were tried in the prior and the model proved to be relatively insensitive to prior choice. For a full discussion and comparison of the possible priors see appendix (A.2). All of the Gibbs samplera in this section are run for 9000 iterations with a burn-in of 2000. In all cases convergence was achieved in less than 500 iterations using the techniques discussed at the end of the following chapter.

Results

To see how well the Gibbs sampler is performing, we will first compare the estimated values of α returned from the Gibbs sampler to the values of $\hat{\alpha}$ calculated from equation (5.5).

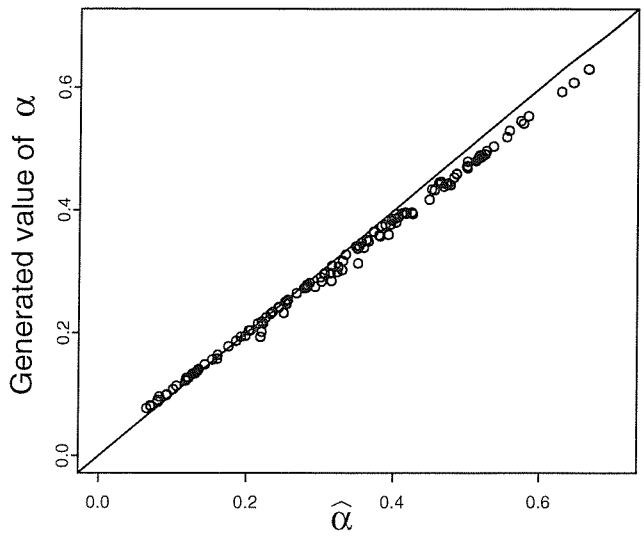


Figure 5.1: Plot of generated estimates for α (where the generated values are the mean values obtained from one Gibbs sampler) against the values obtained using the analytical solution for α, $\hat{\alpha}$, for all of the different values of $\pi(1)$ and $\pi(0)$.

The Gibbs sampler is performing well. We can see from the graph that the predictions are very similar to the corresponding values of $\hat{\alpha}$ although, as we might expect, the estimates improve at $\hat{\alpha} = 0.25$ because here the mean of the prior is equal to the estimator. The estimator $\hat{\alpha}$ generally overestimates for high values of α, the generated estimate is lower than $\hat{\alpha}$ for high underlying values of α possibly showing a better fit than $\hat{\alpha}$.

The generated values at $\alpha = 0.1$ are slightly higher then the predicted values. This needs further investigation although the discrepancy is not serious enough to indicate the Gibbs sampler is performing badly.

If we now look at how well the sample mean performs as a comparison,

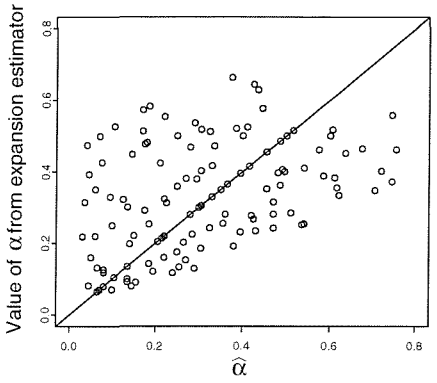


Figure 5.2: Plot of the estimates obtained using the expansion estimator for α against the values obtained using the analytical solution for α, $\hat{\alpha}$, for all of the different values of $\pi(1)$ and $\pi(0)$.

we can see that the sample mean performs badly against $\hat{\alpha}$. So the Gibbs sampler performs significantly better than the sample mean in this case.

To look in more detail at exactly how well the sampler is performing we can look at the difference between the actual values of α (those used to generate the original population) and the estimates for α generated from the sampler. We are only really interested in how big these differences are so we have taken the modulus of all of the differences.

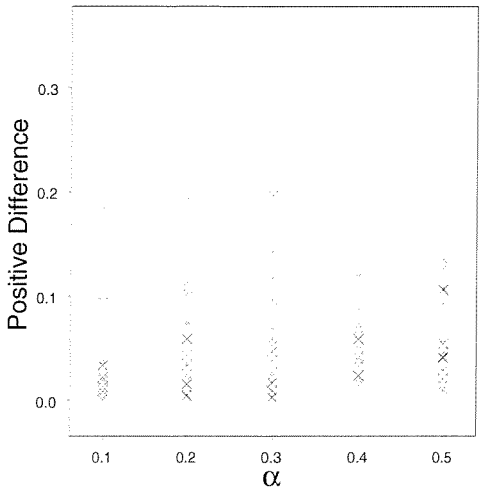


Figure 5.3: Plot of the positive differences between the actual values of α and the generated values of α (where the generated values are the mean values obtained from one Gibbs sampler) against the actual values of α.

We can see that as in figure (5.4) the predictions improve as we approach $\alpha = 0.5$. However, most of the estimates lie within 0.15 of the actual value and so are acceptable. The few outliers occur when α is small, suggesting that the estimator does not perform as well for very small values of α. To explore this further we can examine the plots for the different values of π.

If we look at how well the Gibbs sampler is performing for the different values of $\pi(0)$ and $\pi(1)$ we get the following plots.

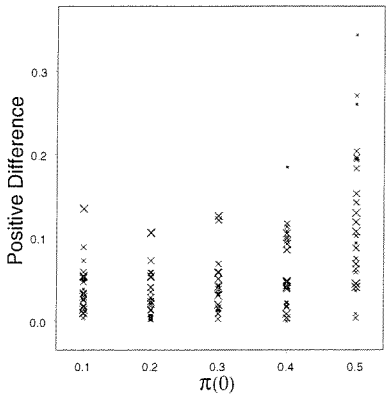


Figure 5.4: Plot of the positive actual differences between the true values of α and the generated values of α (where the generated values are the mean values obtained from one Gibbs sampler) against the true value of $\pi(0)$. Here the larger the cross, the bigger the true value of α.

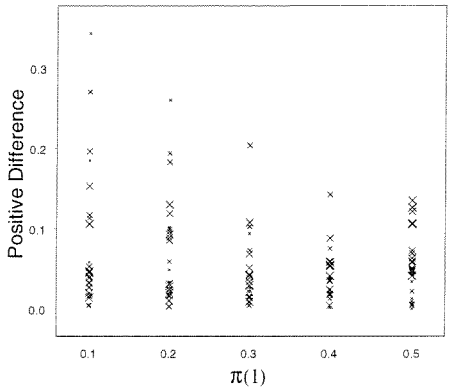


Figure 5.5: Plot of the positive actual differences between the true values of α and the generated values of α (where the generated values are the mean values obtained from one Gibbs sampler) against the true value of $\pi(1)$. Here the larger the cross, the bigger the true value of α.

It is clear that if $\pi(0)$ is large and $\pi(1)$ is small the predictions are poor for smaller values of alpha. If the values of π were to have any effect it is not surprising that they would have most effect at their extreme values. At these parameter settings we have the fewest number of 1's being sampled with a very low probability, which is clearly the most difficult case to make inference from. In most cases the sampler is actually producing over-estimates.

To see how the values of $\pi(0)$ and $\pi(1)$ interact we can plot them both against the positive difference for each level of α. We will only show the plots for $\alpha = 0.1$, $\alpha = 0.3$ and $\alpha = 0.5$ here for a full set of these graphs and a plot showing where the values are known, see (A.4).

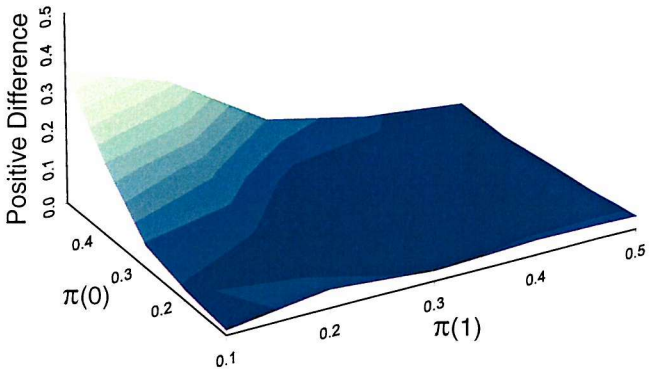


Figure 5.6: Plot of the positive actual differences between the true values of α and the generated values of α (where the generated values are the mean values obtained from one Gibbs sampler) for $\alpha = 0.1$.

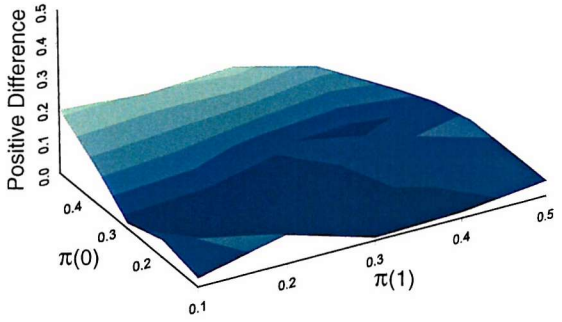


Figure 5.7: Plot of the positive actual differences between the true values of α and the generated values of α (where the generated values are the mean values obtained from one Gibbs sampler) for $\alpha = 0.3$.

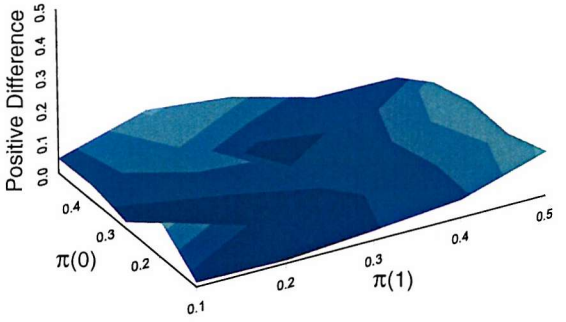


Figure 5.8: Plot of the positive actual differences between the true value of α and the generated values of α (where the generated values are the mean values obtained from one Gibbs sampler) for $\alpha = 0.5$.

These plots reaffirm that the estimated values of α are less affected by the values of π as α increases. All of the slopes on the plots decrease as α increases. However, it seems to be only the combination of a large value of $\pi(0)$ and a small value of $\pi(1)$ that causes a real problem.

Conclusion

The estimator generated by the Gibbs sampler performs well in general, significantly better than the sample mean and slightly better than $\hat{\alpha}$. However, the Gibbs sampler estimator begins to fail if high values of $\pi(1)$, low values of $\pi(0)$ and low values of α are combined. If we limit the parameter spaces to exclude these possibilities for π then the estimator will produce consistently good estimates for α.

5.2.2 Unknown Sampling Probabilities

It is possible to construct a Gibbs sampler for this model although we would expect it to give unreliable results because there is no unique solution to the score equations.

To sample for α and X we will use the same conditional distributions as before (5.6 and 5.7). Then the conditional distributions for $\pi(0)$ and $\pi(1)$ are calculated with their conjugate Beta distribution priors as follows,

$$\begin{aligned}[\pi(0) \mid \pi(1), X_1, ..., X_N, \alpha] &\propto prior(\pi(0)) \times \prod_{X_i\,'s=0} \pi(1)^{I_i}(1-\pi(1))^{1-I_i} \\ &\propto Beta(\sum_{X_i\,'s=0} I_i + \mu - 1, \sum_{X_i\,'s=0} (1-I_i) + \sigma - 1),\end{aligned} \tag{5.8}$$

$$\begin{aligned}[\pi(1) \mid \pi(0), X_1, ..., X_N, \alpha] &\propto prior(\pi(1)) \times \prod_{X_i\,'s=1} \pi(0)^{I_i}(1-\pi(0))^{1-I_i} \\ &\propto Beta(\sum_{X_i\,'s=1} I_i + \varepsilon - 1, \sum_{X_i\,'s=1} (1-I_i) + \omega - 1).\end{aligned} \tag{5.9}$$

The populations are generated in the same way as before and the sampler is run for 9000 iterations with a burnin of 2000. We will set the prior values at $\mu = 2$, $\sigma = 5$, $\varepsilon = 2$ and $\omega = 5$.

Results

We must first ascertain that the Gibbs sampler is producing sensible estimates for α from the data it is given. To do this we plot the estimated values against the values obtained for $\hat{\alpha}$ calculated using the generated values of $\pi(0)$ and $\pi(1)$.

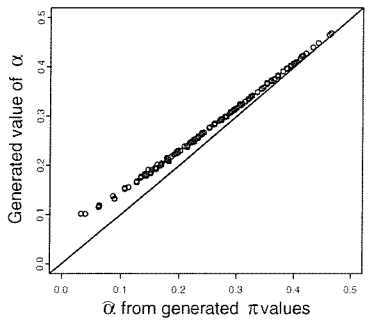


Figure 5.9: Plot of the generated values of α (where the generated values are the mean values obtained from one Gibbs sampler) against the estimated values, $\hat{\alpha}$, where $\hat{\alpha}$ is calculated using the values of π generated in the Gibbs sampler.

We can see from figure 5.9 that there is a clear correlation, although the sampler estimates are all slightly too high. This is to be expected because we are adding in more error by estimating the value of α using estimated values of π. There is enough correlation to feel confident that there is no flaw in the execution of the Gibbs sampler.

If we now look at a plot of the generated values of α against the value of $\hat{\alpha}$ obtained from the true values of π, we can see that there is little or no correlation at all.

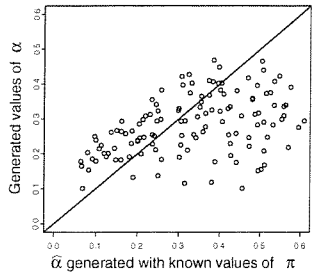


Figure 5.10: Plot of the generated values of α (where the generated values are the mean values obtained from one Gibbs sampler) against the estimated values of α for the actual values π.

The estimator is no improvement on what we would expect to see from the sample mean, in fact it is slightly less accurate because it fails to generate any high values of α.

The Gibbs sampler is failing to estimate values for α if we assume the sampling values are unknown. In fact the sampler fails to produce any values above $\alpha = 0.4$ at all.

If we look at where the generated values lie in relation to the actual values,

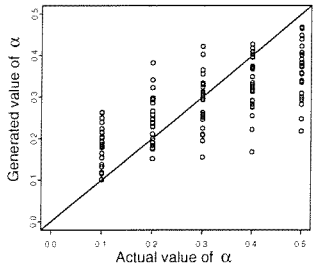


Figure 5.11: Plot of the generated values of α (where the generated values are the mean values obtained from one Gibbs sampler) against the actual values of α when the π values are unknown.

we can see that all of the points generated for $\alpha = 0.1$ are above the actual value so it is always predicting higher values, yet all of the values for $\alpha = 0.5$ are too low. So the model is inaccurate at both extremes and the average gradient does not reflect the 45° line.

This is in marked contrast to the previous example, where the points are evenly spread around the actual value and the generated gradient obviously follows the 45° line as we can see below.

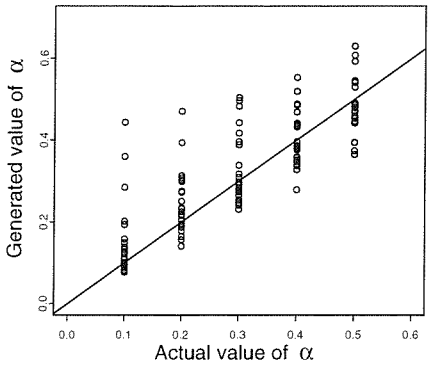


Figure 5.12: Plot of the generated values of α (where the generated values are the mean values obtained from one Gibbs sampler) against the actual values of α when the π values are known.

If we look at the fit of the π variables the reason for the second estimator performing so badly becomes apparent.

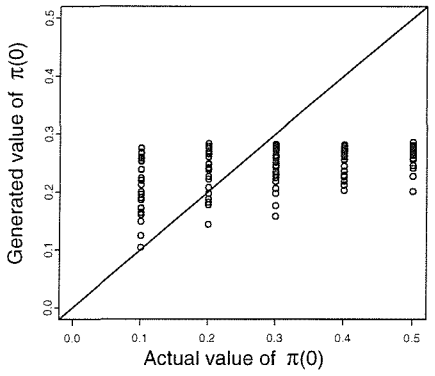


Figure 5.13: Plot of the generated values of $\pi(0)$ (where the generated values are the mean values obtained from one Gibbs sampler) against the actual values of $\pi(0)$

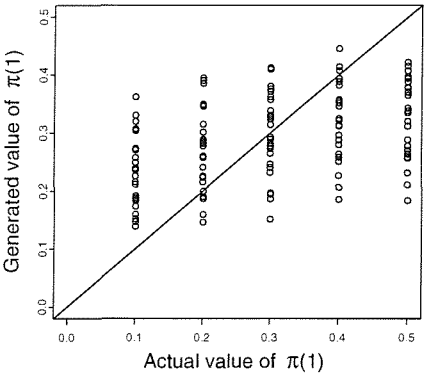


Figure 5.14: Plot of the generated values of $\pi(1)$ (where the generated values are the mean values obtained from one Gibbs sampler) against the actual values of $\pi(1)$

The sampler is completely failing to produce accurate values for the π values as we would expect, giving instead the same estimate each time, regardless of the true value. The estimates of π produced are those more likely to generate lower values of α, explaining the earlier reasonable estimates for these values.

Conclusion

If we assume that all of the underlying parameters are unknown in this model then we cannot make inference. There simply is not enough information in the sample data to estimate all of the required unknown values. The Gibbs sampler is reflecting this well by failing to give accurate estimates for any of the parameters. However, we can make inference if we know α or we know $\pi(0)$ and $\pi(1)$. We can also make inference about functions of these variables for instance $\alpha\pi(1)$ and $(1-\alpha)\pi(1)$, however, this does not help us to distinguish the situation where we have a low value of α from the situation where we have a low value of $\pi(1)$.

At this point it is clear that these simple models will not help us in our overall aim, although they are interesting in their own right. We will not know the sampling probabilities before hand and we need to be able to make inference in this case. We must therefore begin to examine more complex models which contain more information about the population.

Chapter 6

Modelling Network Sizes

6.1 Introduction

The aim of the following chapters is to apply a model-based Bayesian analysis to the Adaptive Cluster sampling methods described in chapter 2. It is in effect a synthesis of the work presented so far.

At present one of the main approaches to modelling spatially clustered data is by using point processes (Diggle, 1975; Baddeley and Turner, 2000; Brix and Diggle, 2001) where, although the spatial relationship between the clusters is considered, the clusters themselves are thought of as points and so have no internal spatial structure. In our methodology we super-impose a grid on a region containing a clustered population. The population is then modelled within this grid structure giving an interdependence within and spatial component to the clusters themselves. This locates the clusters within the grid and gives them a spatial size.

Our ultimate aim will be to take a sample of the grid cells from which we can make inference about the total number of members within the population. The sample is taken in the same way as in the design-based methodology of Thompson (1990).

6.2 A Spatial Model

Clustered data has an intrinsic spatial component, so constructing a model which explicitly uses this spatial aspect of the data would seem to be an obvious step. We would hope that by modelling the spatial component as well as the size of the clusters we can obtain more information for the sample. For instance there is knowledge about the size of a cluster contained in its distance from its nearest neighbour. If two cluster are close together they are more likely to be small.

In fact we will learn that the real elegance of the final model proposed is that it will finesse the spatial component of the model enabling many of the difficulties discussed in this section to be avoided, while still enabling us to capture enough information from the sample to make inference about the population total.

To reflect the structure of the thesis we present this model in a Bayesian framework, although much of the original work was attempted using frequentist methods which quickly proved analytically intractable.

We start with a grid of an arbitrary size in which there is some clustered population.

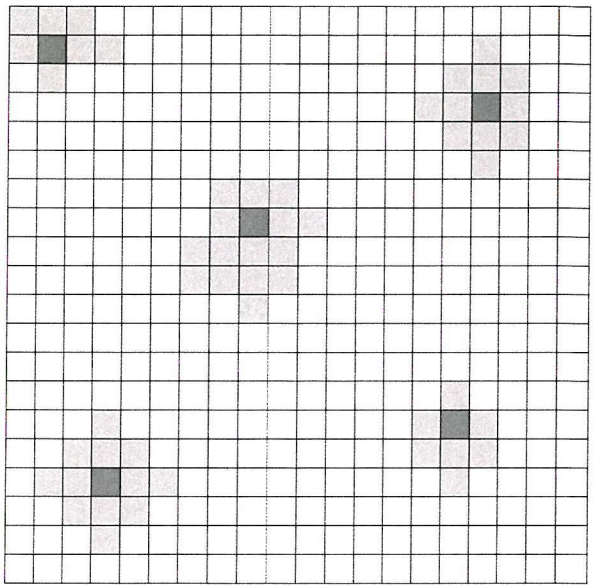


Figure 6.1: Illustration of a clustered population, dark grey cells indicate cluster centres, light grey indicate cluster cells.

Consider a region containing a sparse, clustered population and super-impose a regular grid of N cells over this region The following variables are defined: the total number of grid cells N, the width of the grid c, the height of the grid r and the number of cluster centers W. Then we let $S_1, ..., S_W$ be the coordinates of the cluster centers, $X_1, ..., X_W$ be the number of cells contained in each cluster, $Y_1, ..., Y_W$ be the population counts in each cluster, and finally $d_1, ..., d_W$ denote the distances between each cluster centre and the nearest centre to it.

We generate the number of cluster centers using a Poisson distribution dependent on the grid size and then generate coordinates for these centers using a uniform distribution, although neighbouring centres will necessarily be rejected. We can then generate $X_1, ..., X_W$ and $Y_1, ..., Y_W$ from Poisson distributions dependent on the distances between clusters and the size of the clusters respectively. The distances between clusters are not random, but are a function of $\mathbf{S}$ and W.

We can construct the following likelihood,

$$
\begin{aligned}
&[W, \mathbf{S}, \mathbf{X}, \mathbf{Y}, \mathbf{d}|\alpha, \beta, \gamma] \\
&\quad = [W][\mathbf{S}|W][\mathbf{X}|W, \mathbf{S}][\mathbf{Y}|\mathbf{X}, \mathbf{W}, \mathbf{S}] \\
&\quad = \frac{e^{-\alpha N}(\alpha N)^W}{W![1 - e^{-\alpha N}]} \times \prod_{i=1}^{W} \frac{1}{r}\frac{1}{c} \\
&\qquad \times \prod_{i=1}^{W} \frac{e^{-\beta f_i(W,S)}(\beta f_i(W,S))^{x_i}}{x_i![1 - \{e^{\beta f_i(W,S)}(1 + \beta f_i(W,S))\}]} \\
&\qquad\quad \times \prod_{i=1}^{W} \frac{e^{-\gamma x_i}(\gamma x_i)^{y_i}}{y_i![1 - \sum_{j=1}^{x_i} exp\{-\gamma x_i + j\log(\gamma x_i) - log(j!)\}]}.
\end{aligned}
$$

This model seems initially attractive, it has many of the same features as the models used later in this thesis and after the sampling stage has been added the likelihood can be analysed using a Gibbs sampler in a similar way to the previous chapter. However, there are several problems underlying this likelihood which need to be explored in more detail.

The first problem is how to determine the position of a centre within a cluster from the sample data. This is illustrated by the following examples.

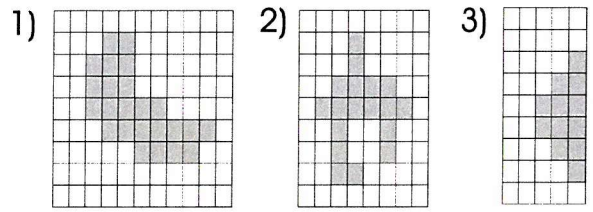


Figure 6.2: An illustration of some cluster shapes which pose difficulties when trying to determine a centre

In the first example there could conceivably be dispute over which cell is the 'true' cluster centre. While most people could probably agree on a center 'by eye', it is not clear what the 'true' center would be, yet this is exactly what is required as sample data in the model. This problem is not insurmountable,

there are several possible (although ad hoc) methods for finding a cluster centre.

One possible (although computationally expensive) method could be to define the centre as the cell which minimises the following sum,

$$\sum_b d(a,b)^2 \ .$$

Here a and b are cells within a particular cluster. This method could be seen as similar to minimising the sum of squares in regression analysis. We are minimising the distance between a given cell and the remaining cells in the cluster. The cell a which minimises this sum will be defined as the centre of the cluster. In the event of multiple cells minimising this sum the centre will be randomly chosen from these cells.

This method requires the summation to be performed for each cell in the cluster; in large clusters this will become time consuming. A second possible method which could be carried out quickly by hand is as follows:

- Find the longest row of horizontal squares and the longest column of vertical cells.
- Extend the row and column (outside of the cluster if necessary) to find the intersection of row and column.
- Count the number of cells within the cluster along both diagonals through the intersection.
- Take the largest of these values and find the mid-cell along this diagonal (in the case of multiple possibilities the center is chosen at random from these possibilities).
- The cell chosen will act as the cluster center.

This method fails if we have a puncture or 'hole' in the shape as in example 2. To overcome this we would have to use other conditions, for instance if

the center lies in the puncture take the cell contained in the cluster with the closest Euclidean distance to it. Otherwise, we could take all of the surrounding cells contained in the cluster and have the cluster center as the cell with the largest population count.

This leaves the third example, if we have a cluster on the edge of our area, the center could be a cell on the boundary of the area, yet most algorithms would put the center away from this boundary.

It is not clear whether these problems would cause enough error in the predictions to be of concern, however, it is important to acknowledge their existence.

We now move on the function $f(W, S)$, the distance between a cluster centre and it's nearest neighbour. This distance is necessary as we would like to model the size of clusters dependent upon it. We would expect that if two cluster centers are close together the clusters will be smaller so that they do not overlap, if they are 'far apart' we would expect they could be larger. We find the distance between centers by defining it as the Euclidean distance between the middle of the cluster centers cells.

Problems arise modelling the size of cluster in this way if we allow two clusters to be very close together, the cells would need to be arranged within the cluster so as to ensure that the two clusters did not overlap. This can be achieved in two ways:

- Collapse the clusters down to the cluster centres and do not worry about their spatial shape, in which case we are dealing with a point process;
- Allow the clusters to overlap in the hope that this does not affect the overall estimations.

It may well be possible to produce a good spatial model using these methods, although it is clearly not as straight forward as we initially conceived. This

is why the main model presented in this thesis is so appealing, it allows us to model the data spatially, while not forcing us to work with all of the problems that spatially clustered data can pose.

6.3 The Model

We now propose a model which finesses the difficulties described above.

Consider again a region containing a sparse, clustered population and superimpose a regular grid of N cells over this region. We let X be the number of nonempty grid cells, where a cell is defined to be nonempty if it contains at least one member of the population. Let P represent the number of nonempty networks where a network is defined in the same way as in Thompson (1990). Then let $\mathbf{Y} = (Y_1, ..., Y_P)$ denote the number of nonempty grid cells within each nonempty network so that $X = \sum_{i=1}^{P} Y_i$. The number of sampled units is denoted by n, where a sampled unit is either one whole network or one 'empty' cell. Therefore the total number of units which can be sampled is $(N - X + P)$.

At this first stage we assume that each cell contained within a nonempty network contains a member of the population. However, we will not actually model the number of members of the population within each cell until the next chapter.

We let be a truncated binomial $X \sim Binomial(N, \alpha), \quad X \neq 0$ and $P \sim Binomial(X, \beta), \quad p \neq 0$ to ensure that $P \leq X$. Then the size of each network $\mathbf{Y} = (Y_1, ..., Y_P)$ will be a partition of the X cells over the P cluster centers.

Modelling the population in this way enables us to finesse the problem in such a way that it is no longer necessary to know where the networks are in space, we simply need to know the size of each network. This crucial insight reduces the problem, yet because in the end we are only interested in the population total not the precise physical location of the units of the

population, we are not losing any information. Although we would obviously obtain less information in this case, this loss of information is not enough to warrant the added complications.

6.4 The model with a fixed number of grid cells contained within clusters

6.4.1 Introduction

We let the total number of grid squares be N. As in West (1996) we first assume that the number of cells contained within networks, $X = X_0$, is known. We then have a binomial distribution over X_0 determining the value of P and a multinomial distribution to create the Y values. We multiply these two distributions together to form the likelihood for this initial model.

The likelihood function of the initial model taking into account the fact that $P > 0$ is,

$$[Y, P|\beta] = \binom{X_0}{P} \frac{\beta^P (1-\beta)^{X_0 - P}}{1-(1-\beta)^{X_0}} \times (X_0 - P)! \frac{\left(\frac{1}{P}\right)^{(y_1-1)} \cdots \left(\frac{1}{P}\right)^{(y_P-1)}}{(y_1-1)! \cdots (y_P-1)!}.$$

6.4.2 The Notation within the Observation process

The networks will be split into the unobserved and observed networks using the subscript 0 for the observed values X_0, P_0 and $\mathbf{Y}_0$ and the subscript 1 for the unobserved values are then X_1, P_1 and $\mathbf{Y}_1$.

The observed and unobserved data is then denoted by $\mathbf{D}$ and $\mathbf{U}$. So $\mathbf{D} = \{X_0, P_0, \mathbf{Y}_0\}$ and $\mathbf{U} = \{X_1, P_1, \mathbf{Y}_1\}$.

In order to construct sampling probabilities we introduce the further notation $\mathbf{Z} = (Z_1, ..., Z_{N-X+P})$ a vector of the sizes of all possible units which can be sampled, both empty and nonempty. An empty unit is defined to be of size 1.

We sample using the informative sampling strategy of Thompson (1990). This proposes that if one cell within a network is sampled the entire network is sampled and therefore the probability of sampling a particular network is weighted according to its size. As we have seen a similar sampling scheme has been modelled in Nair and Wang (1989) by introducing a weight function. We now construct a parallel weight function for the discrete case using the same principles. The discrete case, however, is complicated by the possibility of sampling two units of the same size.

The sample is ordered and any unit once sampled is not replaced. We define $b_i = z_i + ... + z_n$ and $t(U) = \sum_{i=n+1}^{N-X+P}(z_i)$ in a similar way to Nair and Wang (1989).

Consider the probability of sampling the first unit z_1. If there is only one unit of size z_1 in the population, the probablilty of sampling it will be $\frac{z_1}{b_1+t(U)}$. If there are multiple units of size z_1 the probablilty of sampling this particular unit becomes $\frac{z_1 \times g_1}{b_1+t(U)}$. We define g_i to be the number of networks of size z_i remaining after $(i-1)$ previous units have been sampled.

We now see that the probability of taking a particular sample of size n in the discrete case is

$$\prod_{i=1}^{n} \frac{z_i \times g_i}{b_i + t(U)} .$$

This is simply and clearly illustrated in the following example.

Consider a population consisting of eight units of the following sizes,

$$5, 5, 3, 6, 1, 2, 2, 7$$

from which the following sample is taken

$$5, 3, 5, 2.$$

The probability of taking this sample is then

$$\left\{2 \times \left(\frac{5}{31}\right)\right\} \times \left\{1 \times \left(\frac{3}{26}\right)\right\} \times \left\{1 \times \left(\frac{5}{23}\right)\right\} \times \left\{2 \times \left(\frac{2}{18}\right)\right\}.$$

If we introduce this sampling probability to our model in the same way as Nair and Wang (1989) we have the following likelihood,

$$\begin{aligned}[\mathbf{Y}, P|\beta] &= \prod_{i=1}^{n} \frac{z_i \times g_i}{b_i + t(U)} \times \binom{X_0}{P} \frac{\beta^P (1-\beta)^{X_0 - P}}{1-(1-\beta)^{X_0}} \\ &\quad \times (X_0 - P)! \frac{\left(\frac{1}{P}\right)^{(y_1 - 1)} \cdots \left(\frac{1}{P}\right)^{(y_P - 1)}}{(y_1 - 1)! \cdots (y_P - 1)!}.\end{aligned}$$

6.4.3 Outline of the Gibbs Sampler

The model outlined above is too complicated to be fitted analytically, so we will use a Gibbs sampler to find estimates of the underlying parameter β. The accuracy of the parameter estimates will be tested by generating 30 populations using the same known value of β and then taking 5 samples of size 15 from these populations. Each sample will then be used to initialise the Gibbs sampler and the parameter estimate can be compared to the original parameter value. This method is equivalent to that used to test the Gibbs samplers in section (5.2). In this section all Gibbs sampler are run for 40000 iterations with a burn-in of 2000. This burn-in size is precautionary as all chains were shown to converge within 500 iterations using the techniques described at the end of the chapter.

We present a brief overview of the steps in the Gibbs sampler before going on to find the full conditionals and describing how these will be sampled.

1) Initialize. We will need to provide the initial values for the unsampled components, P_1 and $\mathbf{Y}_1$.

2) Generate β from the conditional $[\beta|D, U]$

3) Generate P_1 and $\mathbf{Y}_1$ from the conditional $[P_1, \mathbf{Y}_1|D, \beta]$.

Repeat from 2).

6.4.4 Conditional Posterior Distributions and Sampling Strategies

Posterior distribution for β

The posterior distribution for β is proportional to the product of all of the terms in which β appears in the likelihood multiplied by the prior distribution for β,

$$\begin{aligned}[\beta|D, U] &\propto prior(\beta) \times \binom{X_0}{P} \frac{\beta^P (1-\beta)^{X_0 - P}}{1-(1-\beta)^{X_0}} \\ &\propto prior(\beta) \times \frac{\beta^P (1-\beta)^{X_0 - P}}{1-(1-\beta)^{X_0}}.\end{aligned}$$

Sampling Strategy for β

The posterior distribution for β is not simply a beta distribution due to the truncation, so a Metropolis-Hastings accept-reject step is used to sample for β. Due to the close resemblance of the likelihood to the beta distribution, a $Beta(\eta + P, X_0 - P + \zeta)$ will be used as the proposal distribution. The prior for β will be a $Beta(\eta, \zeta)$ distribution.

We will make $\eta = 1$ and $\zeta = 1$ so there is a uniform prior distribution for β.

It is necessary to ensure that the sampler is moving well over the sample space. To do this we look at a series of plots of the values of β generated at each iteration of the Gibbs samplers. The following is a very small section of one of these plots, all of which show that the sampler is moving well over the parameter space. It gives a clear indication that the Markov chain is mixing well.

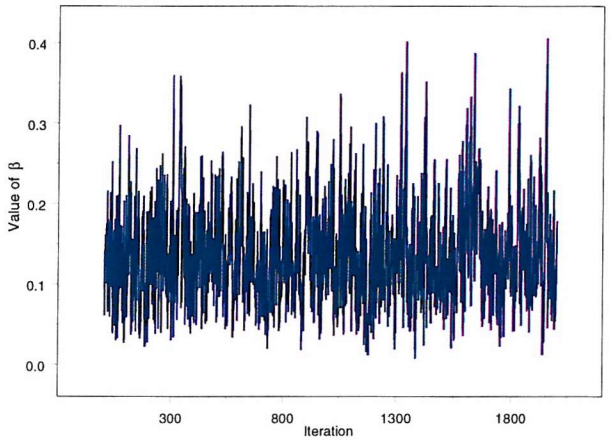


Figure 6.3: Plot showing the movement of β over the parameter space when the actual value of $\beta = 0.15$.

Conditional Posterior Distribution for Unobserved Clusters and Cluster Sizes

It is clear that $\mathbf{Y}$ is highly dependent on the values of X and P. In these circumstances it is usual to generate these parameters together in one step of the Gibbs sampler (Seewald, 1992; Liu et al., 1994; Sahu and Roberts, 1999). Obviously the conditional distribution for these parameters will not be easily sampled from so we will use a second Metropolis-Hastings step.

The conditional posterior distribution is formed from all of the terms in the likelihood which depend on either P_1 or Y_1.

$$
\begin{aligned}
[\mathbf{Y}_1, P_1 | D, \beta] &= \prod_{i=1}^{n} \frac{z_i \times g_i}{b_i + t(U)} \times \binom{X_0}{p_0 + p_1} \frac{\beta^{p_0+p_1}(1-\beta)^{X_0-p_0-p_1}}{1-(1-\beta)^{X_0}} \\
&\qquad \times (X_0 - p_0 - p_1)! \frac{\left(\frac{1}{p_0+p_1}\right)^{(y_1-1)} \cdots \left(\frac{1}{p_0+p_1}\right)^{(y_{p_0+p_1}-1)}}{(y_1-1)! \cdots (y_{p_0+p_1}-1)!} \\
&\propto \prod_{i=1}^{n} \frac{g_i}{b_i + t(U)} \times \frac{\beta^{p_1}(1-\beta)^{-p_1}}{(p_0+p_1)!} \\
&\qquad \times \left(\frac{1}{p_0+p_1}\right)^{\sum_{i=1}^{p_0+p_1}(y_i-1)} \times \prod_{i \notin s}^{p_1} \frac{1}{(y_i-1)!}.
\end{aligned}
$$

Sampling Strategy for Unobserved Clusters

We first construct a proposal distribution from which it is straightforward to sample (P_1, Y_1) jointly and accept or reject these values using the above distribution as the target.

It is convenient to generate P_1 using a discrete random walk; given the past value of P_1, P_1^*, we generate the new value of P_1 from the discrete distribution centered at P_1^* with support $\{P_1^* \pm k : k = 1\}$. Then given P_1 generate Y_1 from the distribution of $[Y_1|Y_0, X_0, P]$, which is a $\mathbf{1}_{P_1}$+multinomial$((X_0 - X_s - P_1), \frac{1}{P_1}\mathbf{1}_{P_1})$ distribution.

The proposal distribution is therefore

$$
[P_1, \mathbf{Y}_1]_{prop} = \frac{(X_0 - X_s - P_1)}{2} \times \left(\frac{1}{p_1}\right)^{\sum_{i \notin s}(y_i-1)} \prod_{i \notin s} \frac{1}{(y_i-1)!}.
$$

By construction it is now simple to generate from the proposal distribution. Letting C denote the constant of proportionality, the logged ratio of the target distribution over the proposal distribution is given by

$$\log\left(\frac{[P_1, \mathbf{Y}_1|D,\beta]}{[P_1, \mathbf{Y}_1]_{prop}}\right) = \log(2C) + \sum_{i=1}^{n} \log\left(\frac{g_{i,j}}{b_i + t(U)}\right)$$
$$- \log((P_0 + P_1)!(X_0 - X_s - P_1)) + p_1 \log\left(\frac{\beta}{1-\beta}\right)$$
$$- \sum_{i \in s}(y_i - 1)\log(p_0 + p_1) + \sum_{i \notin s}(y_i - 1)\log\left(\frac{p_1}{p_0 + p_1}\right).$$

If we define P_1' and $\mathbf{Y}_1'$ to be the new proposed variables, the Metropolis-Hastings test criterion is

$$\log\left(\frac{[P_1', \mathbf{Y}_1'|D,\beta][P_1, \mathbf{Y}_1]_{prop}}{[P_1, \mathbf{Y}_1|D,\beta][P_1', \mathbf{Y}_1']_{prop}}\right).$$

Note that $g_{i,j}$ must be recalculated at each step.

6.4.5 Results

The model was tested using two different values for the total number of non-empty cells, forty and eighty. These values seem appropriate given the number of cells in the region and the fact that we are trying to model sparse clusters.

We will show results for the parameter estimates as these will form the basis from which the rest of the model is constructed.

Several different prior distributions for β were used to measure prior sensitivity the results of which can be found in Appendix B. The estimates do seem somewhat sensitive to the prior chosen but the effect is slight. The best estimates were obtained with a uniform prior, suggesting that at this stage there is a lot of data in our sample and that the model is performing well.

We present parameter estimates which are a median over thirty populations, each of which has been sampled five times thereby giving us one hundred and fifty estimates for each level of the parameter β. We have taken the median of the estimates as this is the measure we will use further on in the thesis as there can be large outliers in the distribution of estimates. At this stage, however, mean and median produce very similar plots.

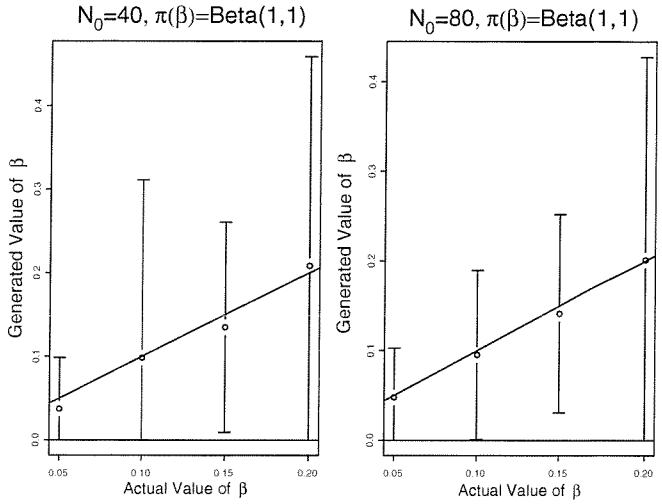


Figure 6.4: Plots of the actual underlying parameter values against the median of those predicted by the Gibbs sampler, here the intervals bars represent the entire range of values generates by the gibbs sampler for β.

The results are extremely promising at this stage allowing us to add further complexity to the model and begin to model the non-empty cells.

6.5 The Model with an Unknown Number of grid cells contained within clusters

6.5.1 Introduction

The previous section only models the situation where there are a known number of grid cells contained within networks. We wish ultimately to be able to model the number of grid cells contained within networks and make inference about this, so we include the number of grid cells contained within networks as another variable within the Gibbs sampler.

The new variable will determine the number of 'non-empty' cells contained within the networks. We place a distribution on this variable so both the underlying parameter value and the actual value of the variable must now be sampled. We call the new parameter α and model the number of cells in networks using a binomial distribution over the actual number of grid cells.

6.5.2 The Likelihood Function

The likelihood function, remembering that we are assuming that $X > 0$ and $P > 0$, is

$$\begin{aligned}[Y, X, P|\alpha, \beta] &= \binom{N}{X} \frac{\alpha^X (1-\alpha)^{N-X}}{1-(1-\alpha)^N} \times \binom{X}{P} \frac{\beta^P (1-\beta)^{X-P}}{1-(1-\beta)^X} \\ &\quad \times (X-P)! \frac{\left(\frac{1}{P}\right)^{(y_1-1)} \cdots \left(\frac{1}{P}\right)^{(y_P-1)}}{(y_1-1)! \cdots (y_P-1)!}.\end{aligned}$$

We sample from the population as before and construct the weight function in the same way so the sampled likelihood function becomes,

$$
\begin{aligned}
[Y, X, P|\alpha, \beta, D] \quad = \quad & \prod_{i=1}^{n} \frac{z_i \times g_i}{b_i + t(U)} \\
& \times \binom{N}{X} \frac{\alpha^X (1-\alpha)^{N-X}}{1-(1-\alpha)^N} \\
& \times \binom{X}{P} \frac{\beta^P (1-\beta)^{X-P}}{1-(1-\beta)^X} \\
& \times (X-P)! \frac{\left(\frac{1}{P}\right)^{(y_1-1)} \cdots \left(\frac{1}{P}\right)^{(y_P-1)}}{(y_1-1)! \cdots (y_P-1)!}.
\end{aligned}
$$

6.5.3 Outline of the Gibbs sampler

As before we present a brief overview of the steps in the Gibbs sampler before proceeding to find the new conditional posterior distributions and describing how these will be sampled. The conditional posterior distribution for β remains unchanged as does the sampling method. In this new model we may find it necessary to change the parameter values of the prior distribution. This will be discussed in the results section of this chapter.

1) Initialize. We will need to provide the initial values for the unsampled components, P_1, X_1 and $\mathbf{Y}_1$.

2) Generate α from the conditional $[\alpha|\beta, D, U]$

3) Generate β from the conditional $[\beta|\alpha, D, U]$

4) Generate X_1 from a random walk

5) Generate $[P_1|D, X_1, \beta] \sim TruncatedBinomial\,(X_1, \beta)$

6) Generate $[\mathbf{Y}_1|D, X_1, P_1] \sim Multinomial\,(X_1, P_1)$

7) Accept or reject $X_1, P_1, \mathbf{Y}_1$ using a Metropolis-Hastings step within the Gibbs sampler.

Repeat from 2).

Conditional Posterior distribution for α

The conditional posterior distribution for α is very similar to that for β and is again formed from the product of the prior distribution and all of the terms in the likelihood containing α;

$$[\alpha|D, U, X, P_0, P_1] \propto prior(\alpha) \times \frac{\alpha^X (1-\alpha)^{N-X}}{1-(1-\alpha)^N} .$$

Sampling Strategy for α

For the same reason we used a Beta prior distribution while sampling for β we take a $Beta(\delta, \epsilon)$ prior distribution and use this within a Metropolis Hastings algorithm. We will take the $Beta(\delta + X, N - X + \epsilon)$ distribution as the proposal. We vary δ and γ within the prior distribution and discuss the effects of this in the results section.

We demonstrate that the sample estimates are moving around the parameter space in the plot below.

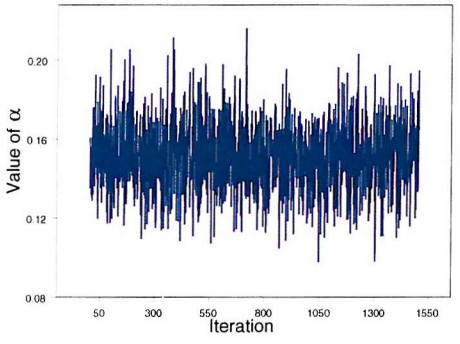


Figure 6.5: Series of generated values of α for one of the populations generated when the actual value of $\alpha = 0.15$.

As in the previous case it is clear that the sampler is mixing well even over such a small subset of the iterations.

Remaining Conditional Posterior Distributions

We will sample X_1, P_1 and $\mathbf{Y}_1$ in the same step because the variables are all very highly correlated and it therefore makes sense to generate them together as discussed in the previous model.

The conditional distribution for these parameters is,

$$\begin{aligned}
[X_1, P_1, \mathbf{Y}_1 | D, \alpha, \beta] &= \prod_{i=1}^{n} \frac{z_i \times g_i}{b_i + t(U)} \times \binom{N}{x_0 + x_1} \frac{\alpha^{x_0+x_1}(1-\alpha)^{N-x_0+x_1}}{1-(1-\alpha)^N} \\
&\quad \times \binom{x_0}{p_0 + p_1} \frac{\beta^{p_0+p_1}(1-\beta)^{x_0+x_1-p_0-p_1}}{1-(1-\beta)^{x_0}} \\
&\quad \times (x_0 + x_1 - p_0 - p_1)! \\
&\quad \frac{\left(\frac{1}{p_0+p_1}\right)^{(y_1-1)} \cdots \left(\frac{1}{p_0+p_1}\right)^{(y_{p_0+p_1}-1)}}{(y_1-1)! \cdots (y_{p_0+p_1}-1)!} \\
\propto \prod_{i=1}^{n} \frac{g_i}{b_i + t(U)} &\times \frac{\alpha^{x_1}(1-\alpha)^{-x_1}}{(N-x_0-x_1)!} \\
&\quad \times \frac{\beta^{p_1}(1-\beta)^{x_1-p_1}}{(1-(1-\beta)^{x_0+x_1})(p_0+p_1)!} \\
&\quad \times \left(\frac{1}{p_0+p_1}\right)^{\sum_{i=1}^{p_0+p_1}(y_i-1)} \times \prod_{i \notin s}^{p_1} \frac{1}{(y_i-1)!}.
\end{aligned}$$

Sampling Strategy

We generate X_1 using a discrete random walk; given the past value of X_1, X_1^*, we generate the new value of X_1 from the discrete distribution centered

at X_1^* with support $\{X_1^* \pm k : k = 1\}$. Then given X_1 and β we generate P_1 from the truncated binomial(X_1, β) distribution. Given X_1 and P_1 we generate $\mathbf{Y}_1$ from the distribution of $[Y_1|Y_0, X, P]$ which is a $\mathbf{1}_{P_1}$+multinomial$((X_1 - P_1), \frac{1}{P_1}\mathbf{1}_{P_1})$ distribution. We sample for X_1, P_1 and $\mathbf{Y}_1$ using a Metropolis Hastings accept-reject step.

The proposal distribution is

$$[X_1, P_1, \mathbf{Y}_1]_{prop} = \frac{1}{2} \times \frac{x_1}{p_1} \frac{\beta^{p_1}(1-\beta)^{x_1-p_1}}{1-(1-\beta)^{x_1}} \left(\frac{1}{p_1}\right)^{\sum_{i \notin s}(y_i-1)} \prod_{i \notin s} \frac{1}{(y_i-1)!} .$$

By construction it is simple to generate from these proposal distributions. If B is now the constant of proportionality, the new logged ratio of the target distribution divided by the proposal distribution is given by

$$\begin{aligned}
\log\left(\frac{[X_1, P_1, \mathbf{Y}_1|D, \alpha, \beta]}{[X_1, P_1, \mathbf{Y}_1]_{prop}}\right) &= \log(2B) + \sum_{i=1}^{n} \log\left(\frac{g_{i,j}}{b_i + t(U)}\right) \\
&\quad - \log\left(\frac{(N - x_0 - x_1)!(p_0 + p_1)! x_1!}{p_1!}\right) \\
&\quad + x_1 \log\left(\frac{\alpha}{1-\alpha}\right) + \log\left(\frac{1-(1-\beta)^{x_1}}{1-(1-\beta)^{x_0+x_1}}\right) \\
&\quad - \sum_{i \in s}(y_i - 1)\log(p_0 + p_1) \\
&\quad + \sum_{i \notin s}(y_i - 1)\log\left(\frac{p_1}{p_0 + p_1}\right) .
\end{aligned}$$

If we define X_1', P_1' and $\mathbf{Y}_1'$ to be the new proposed variables, the Metropolis-Hastings test criterion is

$$\log\left(\frac{[X_1', P_1', \mathbf{Y}_1'|D, \beta][X_1 P_1, \mathbf{Y}_1]_{prop}}{[X_1, P_1, \mathbf{Y}_1|D, \beta][X_1', P_1', \mathbf{Y}_1']_{prop}}\right) .$$

Note that again $g_{i,j}$ must be recalculated at each step.

6.5.4 Results

We generate thirty different populations from several different combinations of underlying parameters and take five samples from these populations using the sampling scheme described in Thompson (1990). We use these sample values within the Gibbs sampler and generate parameter estimates which are as close as possible to the original values. We again run the sampler for 40000 iterations with a burn-in of 2000.

We have very little sample information and are not expecting our estimates to be very accurate. However, with a choice of prior which is sufficiently robust to changes in underlying parameter values we can in fact get very good estimates at all of our chosen parameter values. The next four charts show the best prior combination for each initial value of the parameter α. The confidence interval increases as the priors change due to the shape of the prior for α. The Beta$(1, 25)$ has a much narrower peak that the Beta$(2.5, 9)$ and so the confidence interval will also be narrower. A full set of plots for each prior is given in appendix C.

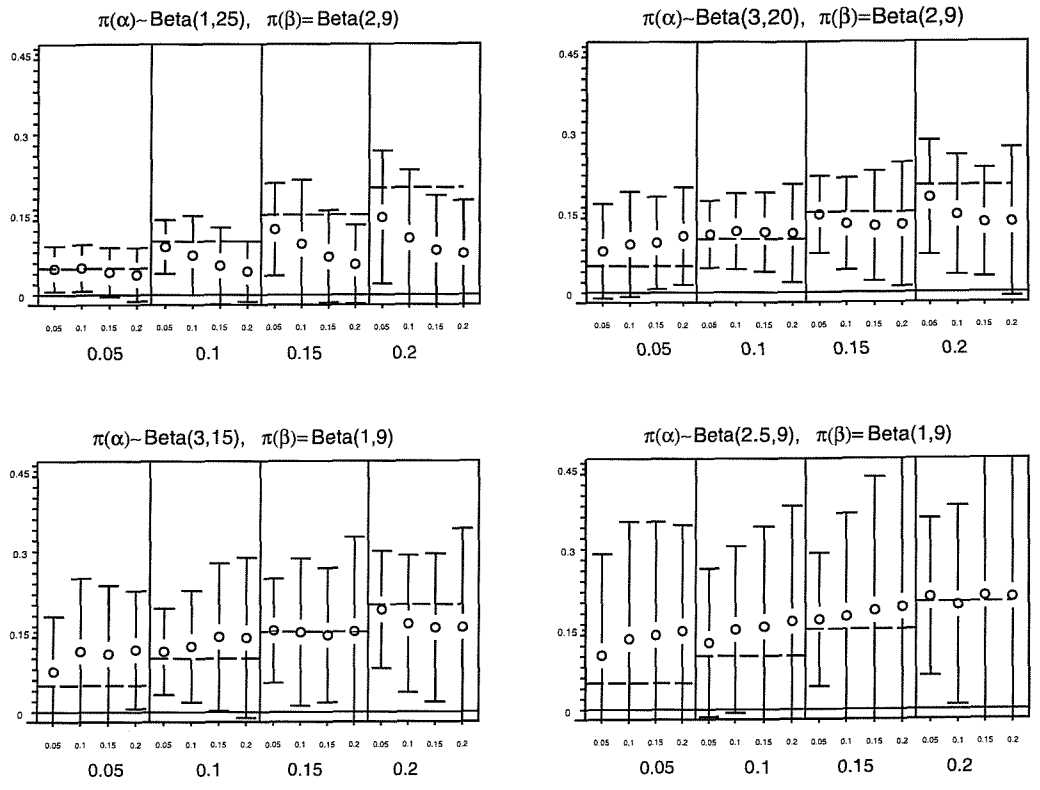


Figure 6.6: Plot showing the median predicted values of the parameter α. The first row under each graph gives the initial value of β and the second gives the initial value of α. Interval width is three standard deviations (calculated over the different population values) in each direction so negative values should be read as zero.

Interestingly, to obtain the best estimates of the parameter β slightly different prior parameter values are needed. For the best fit of α it seems that β must be a very slight underestimate. However, this difference is negligible as can be seen in the following plots.

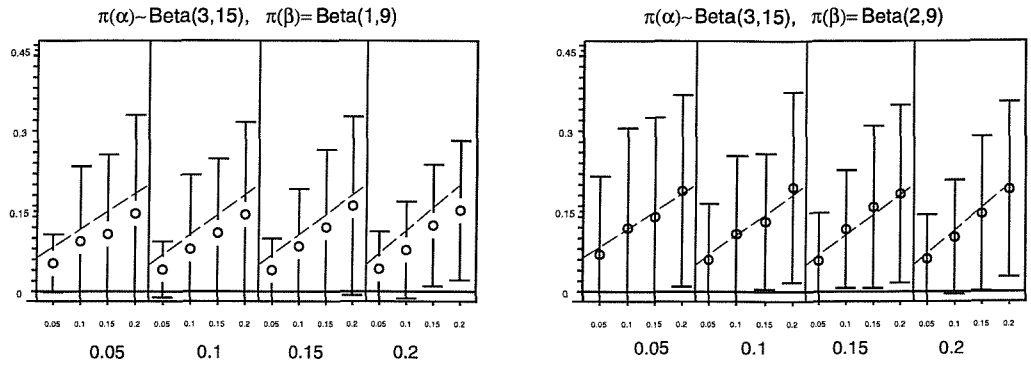


Figure 6.7: Plot showing the median predicted values of the parameter β. The first row under each graph gives the initial value of β and the second gives the initial value of α. Interval width is three standard deviations (calculated over the different population values) in each direction so negative values should be read as zero.

Plots averaging the medians over the initial parameters α and β can also be found in appendix C. These reaffirm the conclusions already drawn but do not shed any new insight so we have not included them here.

In our sampler we are generating the value of P from X so when we add in the population total it is more important to have a good fit for X. For this reason we conclude that the $\pi(\alpha)$ =Beta$(3, 15)$ and $\pi(\beta)$ =Beta$(1, 9)$ should be used in all situations where no further information is available because these priors are most robust to changes of underlying parameter values and they predict reasonably well for all of the initial values of α and β explored, which is our main aim. However, when applying the model, much better results can be obtained by using some of the other prior combinations provided we have some knowledge of the structure of the population, for instance if we know that α is likely to be very small we recommend that a $Beta(1, 25)$ prior is used for α and a $Beta(1, 9)$ prior for β.

6.6 Convergence

It is convenient at this point to discuss the convergence of the Gibbs sampler. We will not examine convergence again in the following chapter as the likelihood functions will not change.

As discussed in chapter 4 we now assess convergence using CUSUM plots (Yu and Mykland, 1998). We use this method because it can be run using the outputs from the sample runs, therefore keeping additional calculations to a minimum and limiting computational expense. A more detailed discussion of which method to use to assess convergence in any particular case can be found in Brooks and Roberts (1998).

The 'burn-in' length was taken to be 2000, however, in all cases convergence was obtained in less than 500 iterations. As an example of when convergence occurs we present in figure 6.8 the following trace plot showing the initial convergence of one of the plots.

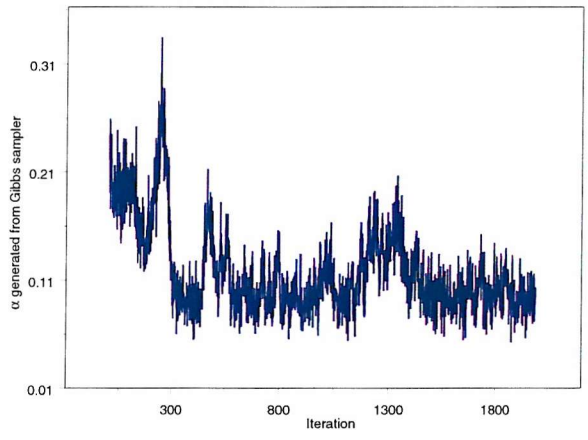


Figure 6.8: A trace plot showing how the paramter estimates for alpha converge.

The initial value was set to be deliberately high to accentuate the burn-in and as we can see convergence occurs in a few hundred iterations.

To construct a CUSUM path plot, we first discard the first n_0 iterations of the sample runs or the 'burn-in'. We assume $\{\mathbf{x}_1, ..., \mathbf{x}^n\}$ is the output from a single run of the sampler. Let θ be the parameter we wish to monitor, so in this case either α or β. We then calculate

$$\hat{\mu} = \frac{1}{(n - n_0)} \sum_{i=n_0+1}^{n} \theta(\mathbf{x}^i) \, .$$

The CUSUM is then calculated as follows,

$$\hat{S}_T = \sum_{i=n_0+1}^{T} [\theta(\mathbf{x}^i) - \hat{\mu}] \qquad \text{for} \quad T = n_0 + 1, ..., n \, .$$

We then plot $\{\hat{S}_T\}$ against T for $T = n_0 + 1, ..., n$. If the plot produced is smooth then it indicates that the chain is mixing slowly in θ, a "hairy" plot indicates fast mixing in θ. To gain an indication of the smoothness the plot it is compared to a benchmark plot of *iid* random normal variates with a mean and variance of the estimated mean and variance of θ.

We present an example of two of the plots obtained, one for α and one for β in figures 6.9 and 6.10. We show plots only 20000 iterations after the burn-in period for clarity, although the plots remain similar if extended.

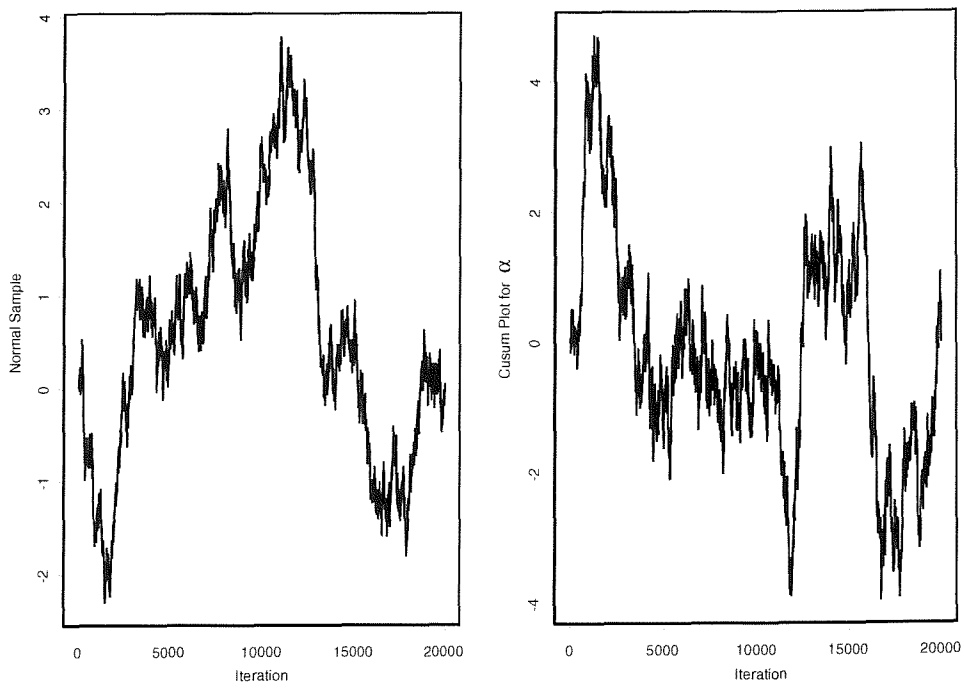


Figure 6.9: An example of a CUSUM plot for α compared with a plot generated from *iid* normal variates with the mean and variance estimated from α

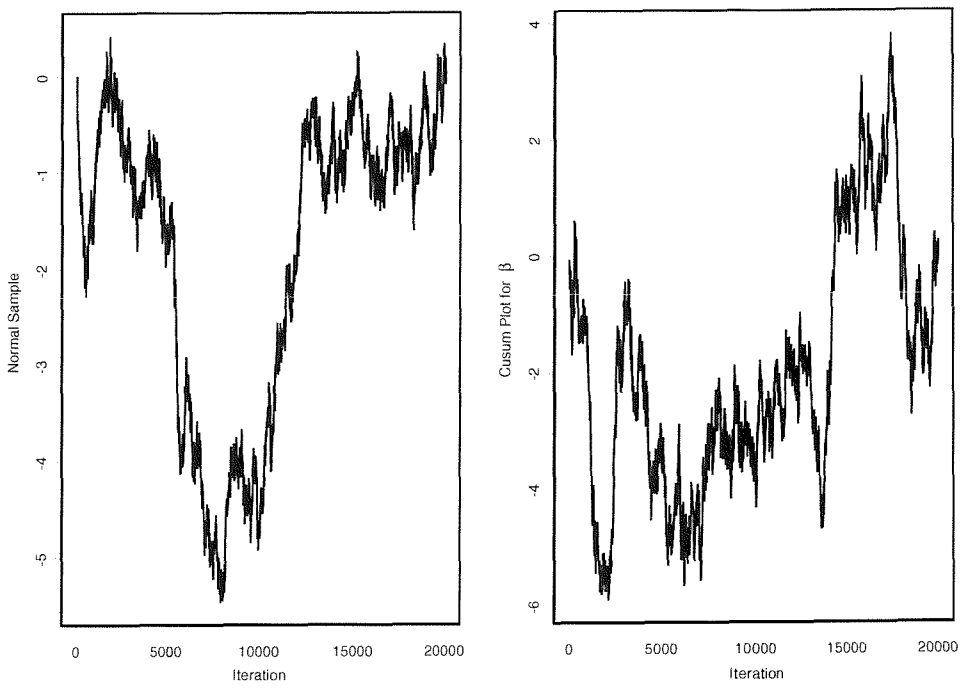


Figure 6.10: An example of a CUSUM plot for β compared with a plot generated from *iid* normal variates with the mean and variance estimated from α

It is clear from figures 6.9 and 6.10 that the chains have converged, as the comparison between the normal variate and the θ variate is favourable. However, Brooks (1996) argues that this is subjective and suggests a quantitative measure. He defines

$$d_T = \begin{cases} 1 & \text{if } S_{T-1} > S_T \text{ and } S_T < S_{T+1} \\ & \text{or } S_{T-1} < S_T \text{ and } S_T > S_{T+1} \\ 0 & \text{else} \end{cases} .$$

Then as a measure of 'hairiness'

$$D_{n_0,n} = \frac{1}{n - n_0} \sum_{i=n_0+1}^{n-1} d_T \ .$$

In this measure if $D_{n_0,n} = 0$ we have maximum smoothness, if $D_{n_0,n} = 1$ we have maximum hairiness. The 'perfect score' then would be $D_{n_0,n} = 0.5$. In our example $D_{n_0,n} = 0.501$ for α and $D_{n_0,n} = 0.536$ for β. This again shows that our Gibbs sampler is converging as we would hope.

Chapter 7

Modelling Population totals in clustered populations

7.1 Introduction

We are now in a situation where it is a relatively simple extra step to model the actual population total. We do this by defining a distribution for the number of members of the population found in each network. Let $\mathbf{M} = M_1, ..., M_P$ be the population totals in each network and conditionally model $M_1, ..., M_P$ to be truncated Poisson distributed with parameters $y_1\gamma, ...y_P\gamma$. Here $y_1, ..., y_P$ are the network sizes as in the previous chapter. The truncation takes into account the fact that each grid cell within the network is nonempty and so must contain at least one member of the population, i.e. $M_i \geq Y_i$ for $i = 1, ..., P$. The posterior conditional distribution for $\mathbf{M}$ is then,

$$[\mathbf{M}|D, U, \gamma] = \prod_{i=1}^{P} \frac{e^{-\gamma y_i}(\gamma y_i)^{m_i}}{m_i![1 - \sum_{j=1}^{y_i} exp\{-\gamma y_i + j\log(\gamma y_i) - log(j!)\}]}.$$

It is interesting to note that this model does not reflect the loglinear model which frequentists might expect to see used at this point. It requires more

numeric approximation to work with a loglinear model where the mean would be taken to be $e^{\gamma y_i}$ as opposed to the mean of γy_i above. We would need to adopt numerical techniques to generate γ as it would no longer resemble a gamma distribution. While techniques have been constructed to fit these models (see Wild and Gilks (1992, 1993)), they are computationally complicated.

We initially fitted a loglinear model using basic numerical techniques but in a structure such as this adding in a further fairly inaccurate numerical step simply made the estimates unacceptable. While we could theoretically have used the techniques in Wild and Gilks (1993) rather than our crude numerical ones the length of time needed to run this simulation proved prohibitive (taking run times from weeks to months on a behemoth cluster of pentium fours). This seemed unnecessary when a simple Poisson distribution could be used much more simply and with less numerical error.

7.2 The Model

The complete density function for our full model of population counts now becomes,

$$\begin{aligned}[\mathbf{M}, \mathbf{Y}, X, P|\alpha, \beta, \gamma] \quad = \quad & \binom{N}{X} \frac{\alpha^X (1-\alpha)^{N-X}}{1-(1-\alpha)^N} \times \binom{X}{P} \frac{\beta^P (1-\beta)^{X-P}}{1-(1-\beta)^X} \\ & \times (X-P)! \frac{\left(\frac{1}{P}\right)^{(y_1-1)} \cdots \left(\frac{1}{P}\right)^{(y_P-1)}}{(y_1-1)! \cdots (y_P-1)!} \\ & \times \prod_{i=1}^{P} \frac{e^{-\gamma y_i} (\gamma y_i)^{m_i}}{m_i! [1 - \sum_{j=1}^{y_i} exp\{-\gamma y_i + j \log(\gamma y_i) - log(j!)\}]}.\end{aligned}$$

We sample from the population and construct the weights in the same way so,

$$
\begin{aligned}
[\mathbf{M}, \mathbf{Y}, X, P|\alpha, \beta, \gamma, D] \quad = \quad & \prod_{i=1}^{n} \frac{z_i \times g_i}{b_i + t(U)} \\
& \times \binom{N}{X} \frac{\alpha^X (1-\alpha)^{N-X}}{1-(1-\alpha)^N} \\
& \times \binom{X}{P} \frac{\beta^P (1-\beta)^{X-P}}{1-(1-\beta)^X} \\
& \times (X-P)! \frac{\left(\frac{1}{P}\right)^{(y_1-1)} \cdots \left(\frac{1}{P}\right)^{(y_P-1)}}{(y_1-1)! \cdots (y_P-1)!} \\
& \prod_{i=1}^{P} \frac{e^{-\gamma y_i} (\gamma y_i)^{m_i}}{m_i! [1 - \sum_{j=1}^{y_i} exp\{-\gamma y_i + j \log(\gamma y_i) - log(j!)\}]}.
\end{aligned}
$$

7.3 The Gibbs sampler

The steps in the Gibbs sampler remain the same as in the previous chapter, we simply add a further step to sample for γ. The value of $\mathbf{M}$ will be sampled with X_1, P_1 and $\mathbf{Y}$.

7.3.1 Conditional Posterior Distribution for γ

The conditional distribution for γ is formed as follows

$$
[\gamma|D, U, \mathbf{M}] \propto prior(\gamma) \times \prod_{i=1}^{P} \frac{e^{-\gamma y_i} (\gamma y_i)^{m_i}}{m_i! [1 - \sum_{j=1}^{y_i} exp\{-\gamma y_i + j \log(\gamma y_i) - log(j!)\}]}.
$$

Sampling Strategy for γ

The numerator in the likelihood for γ is proportional to a gamma density, so to make the posterior distribution tractable we have the prior distribution $Gamma(\zeta, \tau)$ and the proposal distribution for γ a $Gamma\left(\sum_{i=1}^{P} m_i + \zeta, \frac{\tau}{\tau X + P}\right)$ distribution.

We will then sample using a Metropolis-Hastings accept-reject step in the same way as previously.

7.3.2 Remaining Conditional Distribution

We will sample $\mathbf{M}_1, X_1, P_1$ and $\mathbf{Y}_1$ together. This is again necessary because of the dependence between $\mathbf{M}$ and $\mathbf{Y}$.

An interesting point to note at this stage is that we envisaged adding the population total to the model as a simple addition which would not affect the underlying structure of the model. Clearly this is impossible due to the dependence on $\mathbf{Y}$. However, in some senses it can still be separated and this is reflected in the target distribution given below. To illustrate these points we note that the sampling design satisfies the condition

$$[\{i_1, ..., i_n\}|X, P, \mathbf{Y}, \mathbf{M}] = [\{i_1, ..., i_n\}|X, P, \mathbf{Y}],$$

so we can factorise the model as follows,

$$\begin{aligned}
[\{i_1, ..., i_n\}, X, P, \mathbf{Y}, \mathbf{M}|\alpha, \beta, \gamma] &= [\{i_1, ..., i_n\}|X, P, \mathbf{Y}, \mathbf{M}][X, P, \mathbf{Y}, \mathbf{M}|\alpha, \beta, \gamma] \\
&= [\{i_1, ..., i_n\}|X, P, \mathbf{Y}][X, P, \mathbf{Y}, \mathbf{M}|\alpha, \beta, \gamma] \\
&= [\{i_1, ..., i_n\}|X, P, \mathbf{Y}][X, P, \mathbf{Y}|\alpha, \beta][\mathbf{M}|X, P, \mathbf{Y}, \gamma] \\
&= [\{i_1, ..., i_n\}X, P, \mathbf{Y}|\alpha, \beta][\mathbf{M}|P, \mathbf{Y}, \gamma].
\end{aligned}$$

Thus the model can be factorised into two terms, at the population level (α, β) and γ are orthogonal and the model can be interpreted as a spatial version of the count data models in Mullahy (1986); Heilbron (1994); Welsh

et al. (1996). However, at the sample level the likelihood is obtained as we have seen in chapter 2 by summing over the unknown variables, so even if there is no dependence between $\mathbf{M}$ and $\mathbf{Y}$ the factors are still linked by the common unobserved P_1.

The full conditional distribution for the remaining unknown parameters is now,

$$[X_1, P_1, \mathbf{Y}_1, \mathbf{M}_1 | D, \mathbf{M}_0, \alpha, \beta, \gamma]$$

$$\propto \prod_{i=1}^{n} \frac{g_i}{b_i + t(U)} \times \frac{\alpha^{x_1}(1-\alpha)^{-x_1}}{(N - x_0 - x_1)!} \times \frac{\beta^{p_1}(1-\beta)^{x_1 - p_1}}{(1-(1-\beta)^{x_0+x_1})(p_0+p_1)!} \times \left(\frac{1}{p_0+p_1}\right)^{\sum_{i=1}^{p_0+p_1}(y_i-1)} \times \prod_{i \notin s}^{p_1} \frac{1}{(y_i-1)!} \times \prod_{i=1}^{P} \frac{e^{-\gamma y_i}(\gamma y_i)^{m_i}}{m_i![1 - \sum_{j=1}^{y_i} exp\{-\gamma y_i + j\log(\gamma y_i) - log(j!)\}]}.$$

Sampling Strategy

We will perform a Metropolis Hastings accept-reject step in exactly the same way as in the previous chapter, generating $\mathbf{M}_1$ as independent truncated Poisson($exp\{\gamma y_i\}$) random variables. The proposal is therefore,

$$[X_1, P_1, \mathbf{Y}_1]_{prop} = \frac{1}{2} \times \frac{x_1}{p_1} \frac{\beta^{p_1}(1-\beta)^{x_1-p_1}}{1-(1-\beta)^{x_1}} \left(\frac{1}{p_1}\right)^{\sum_{i \notin s}(y_i-1)} \prod_{i \notin s} \frac{1}{(y_i-1)!} \times \prod_{i \notin s} \frac{e^{-\gamma y_i}(\gamma y_i)^{m_i}}{m_i![1 - \sum_{j=1}^{y_i} exp\{-\gamma y_i + j\log(\gamma y_i) - log(j!)\}]}.$$

Constructing the proposal in this way means that all of the terms depending on $\mathbf{M}$ cancel with the target as noted above, so the logged ratio of the target to the proposal remains identical to the previous chapter,

$$\log\left(\frac{[X_1, P_1, \mathbf{Y}_1\mathbf{M}_1|D, \mathbf{M}_0, \alpha, \beta, \gamma]}{[X_1, P_1, \mathbf{Y}_1, \mathbf{M}_1]_{prop}}\right) = \log(2B) + \sum_{i=1}^{n} \log\left(\frac{g_{i,j}}{b_i + t(U)}\right)$$
$$- \log\left(\frac{(N - x_0 - x_1)!(p_0 + p_1)!x_1!}{p_1!}\right)$$
$$+ x_1 \log\left(\frac{\alpha}{1-\alpha}\right) + \log\left(\frac{1-(1-\beta)^{x_1}}{1-(1-\beta)^{x_0+x_1}}\right)$$
$$- \sum_{i \in s}(y_i - 1)\log(p_0 + p_1)$$
$$+ \sum_{i \notin s}(y_i - 1)\log\left(\frac{p_1}{p_0 + p_1}\right).$$

7.4 Results

The choice of prior for γ has little affect so we will summarise the results obtained for $\pi(\gamma)$ =Gamma$(2, 7)$ only. The results for $\pi(\gamma)$ =Gamma$(2, 5)$ and $\pi(\gamma)$ =Gamma$(5, 5)$ are very similar and a full set of plots for these priors can be found in Appendix D. We note that all interval bars in the appendix are for the inter-quartile range as this gives a much clearer indication of the fit for each different prior. In the following graph, however, the interval bars represent the 95% confidence interval. The setting for the Gibbs sampler remain as before we generate thirty populations which we sample five times and run a Gibbs sampler of 40000 iterations with a burn-in of 2000.

As expected, the best overall results were obtained when $\pi(\alpha)$ =Beta$(3, 15)$ and $\pi(\beta)$ =Gamma$(1, 9)$ as we see overleaf.

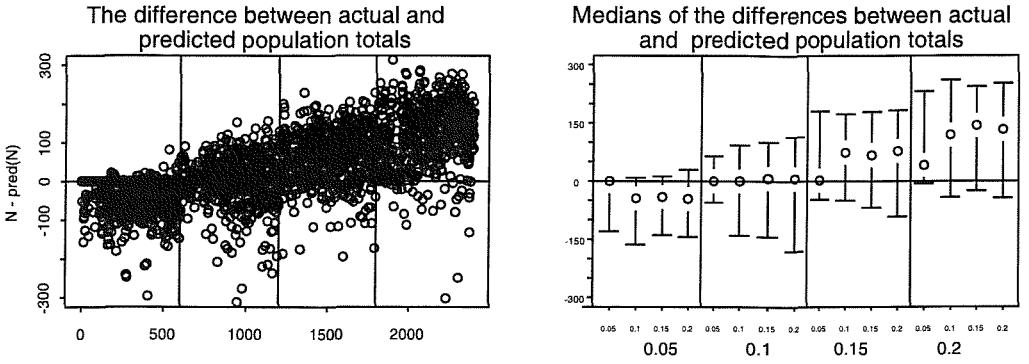


Figure 7.1: Plots showing the best overall fit generated from a single set of priors ($\pi(\alpha)$ =Beta(3, 15), $\pi(\beta)$ =Gamma(1, 9)). Interval bars represent the 95% confidence interval.

The positive gradient in these plots is not a cause for concern because the initial values of α and β are increasing from left to right and the fit is dependent on these parameters. We could change the ordering of the underlying variables and the positive gradient would no longer be apparent in the plots.

This trend does emphasise the fact that how well the estimator does is highly dependent on the underlying parameter values. The larger α and β, the more the population total is underestimated. This makes intuitive sense; we are assuming that we are modelling sparse clustered data and if α and β are large then we have many small clusters which does not hold with this assumption. In some senses in this case we have a much more evenly distributed population; with larger grid cells the same population could be thought of as randomly scattered single units rather than clusters as the clusters will be of a similar size.

In general, the model seems to be working well. If we have an idea of the size and structure of the population which we can incorporate into our prior beliefs then we can predict very well. If we do not have this information then we can seriously underestimate the population total. Even in this worst case our estimates are still as good as those used previously to estimate population totals from this kind of data.

7.5 A comparison with the estimator in Thompson (1990)

We will compare our estimator to the design-based estimator suggested in Thompson (1990) to assess how well it performs in comparison. Two estimators are applied in this paper, we concern ourselves with the estimator which produced the most accurate estimates, the modified Horvitz-Thompson type estimator.

The Horvitz-Thompson estimator, in which each sampled unit is divided by its inclusion probability, is an unbiased estimator in sampling designs where the probability that a unit is included in the sample is known for every unit. The Horvitz-Thompson estimator is modified in Thompson (1990) so that it only makes use of the inclusion probabilities if the unit is included in the initial sample, it can therefore be applied in this case.

We first look at the comparison when we use the combination of priors which which are most robust to parameter changes and would therefore be used in general. The Gibbs sampler is run as before, however, in these comparisons only three samples are taken from each population due to the similarity of the sample generated.

$\pi(\alpha)$ =Beta$(3, 15)$, $\pi(\beta)$ =Beta$(1, 9)$

Sample Size	$\widehat{var}(t_{HT*})$	$\widehat{var}(t_B)$	$\overline{(r_{HT*})}$	$\overline{(r_B)}$	$med(r_B)$	$var(r_{HT*})$	$var(r_B)$
10	16884.48	8617.66	-86.53	44.79	35.00	40510.35	7801.35
15	13750.54	7264.52	-47.39	29.86	17.50	21442.58	8231.73

Table 7.1: Table showing a comparison between the design-based and model-based estimates. B denotes the Bayesian estimator, r denotes the average difference between the true and predicted population totals (the bias), t_{HT*} is the modified Horvitz-Thompson estimator.

We clearly see that the model-based method produces estimates with smaller actual variances than the design-based method of Thompson (1990). The estimator is also on average closer to the true values in these particular cases. We have not presented the efficiency statistic given in Thompson (1990) as clearly this is not appropriate, the variances are too large for comparison with the very small variance of the simple random sampling method. The simple random sampling method may give small variances in all cases but the population estimates are so biased as to make the method useless in this case. The average mean square error of the population total obtained using simple random sampling is 48866, while the average variance is 1.458 for a sample size of 15.

We also present below a comparison of our estimator to the Horvitz-Thompson estimator when a prior which is not robust to parameter changes is used. In other words a prior which gives poor results for some parameter combinations.

$\pi(\alpha)$ =Beta$(1, 25)$, $\pi(\beta)$ =Beta$(1, 9)$

Sample Size	$\widehat{var}(t_{HT*})$	$\widehat{var}(t_B)$	$\overline{(r_{HT*})}$	$\overline{(r_B)}$	$med(r_B)$	$var(r_{HT*})$	$var(r_B)$
15	21480.4	1314.2	-44.58	89.61	77.50	14080.44	8257.69

Table 7.2: Table showing a comparison between the design-based and model-based estimates when an inappropriate prior is used. B denotes the Bayesian estimator, r denotes the average difference between the true and predicted population totals (the bias), t_{HT*} is the modified Horvitz-Thompson estimator.

We see that while the Horvitz-Thompson estimator on average produces slightly better estimates of the population it still has a larger variance.

If we look at a breakdown of the different starting values of the parameters generating the population we see the following results.

$\pi(\alpha)$ =Beta(3, 15), $\pi(\beta)$ =Beta(1, 9), sample size of 15.

Value α	$\widehat{var}(t_{HT*})$	$\widehat{var}(t_B)$	$\overline{(r_{HT*})}$	$\overline{(r_B)}$	$var(r_{HT*})$	$var(r_B)$
0.05	14320.6	7901.73	-71.16	-62.10	2304.06	1393.84
0.1	19736.42	7332.59	-60.05	4.425	9654.03	1920.51
0.15	23436.47	8398.16	-33.64	57.17	16700.88	3057.24
0.2	28276.79	9294.46	-24.70	120.0	24193.69	4718.04

Table 7.3: Table showing a comparison between the design-based and model-based estimates when the estimates are broken down over the initial parameter values. B denotes the Bayesian estimator, r denotes the average difference between the true and predicted population totals (the bias), t_{HT*} is the modified Horvitz-Thompson estimator.

Interestingly the Horvitz-Thompson estimator improves if used with populations in which there are more small networks. This makes sense as it is based on simple random sampling so the closer the population comes to being randomly spaced rather than clustered the better it will do. This is in contrast to our model which assumes few large networks and so does better in the clustered cases.

In conclusion our estimator out-performs the Horvitz-Thompson estimator when the prior is appropriate, but does not perform as well in comparison when the prior is not appropriate. However, our variances are always much smaller irrespective of the prior choice.

Chapter 8

A Spatial Model and Possible Extensions of this Work

8.1 Applying the Model to the Continuous Case

The model constructed in this thesis has some of its origins in the model for oil pools discussed in Nair and Wang (1989) and West (1996). It seems a logical step to attempt to extend our model to the case where we are no longer dealing with discrete populations but with continuous pools.

Our model enables us to consider more data than the model of Nair and Wang (1989) because we use the information about where clusters, or pools in this case, have not been discovered as well as information about pools we have seen. In this section we will consider some of the added complexities and look at how we can adapt our model to this situation.

8.1.1 Constructing a framework

The first step in looking at this new situation is to realise that working with grid cells is no longer appropriate. If we work with cells we will have the situation where a pool covers only part of a cell. To overcome this we

propose changing from grid cells to sampling points, where the sampling points will form a lattice over the area. This allows us to remain in the case of finite sampling units, while still modelling continuous pools.

Our area would then be constructed as follows.

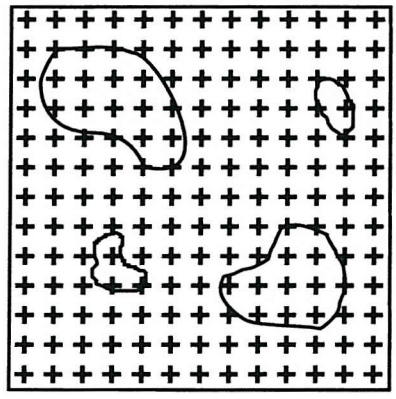


Figure 8.1: Illustration of a continuous population. Here grey shapes represent pools and crosses are possible sampling points

We can initially sample any of the crosses. Then, upon finding a pool, we sample all neighbouring crosses lying within the pool. Thus our sampled units would consist of 'empty' crosses and groups of crosses within a pool.

The next step would depend on the inference we wished to make about the pools. Clearly, if we wished to make inference on the total volume of the lakes from the sample it would be possible to replace the Poisson distribution of our earlier model and model the Y values using a lognormal distribution. We could run the Gibbs sampler in a similar way.

This would be in keeping with the work done by West (1996) who modelled the logs of the volume, depth, net pay and area as a multivariate normal distribution.

If we wished to make inference on surface area this would cause more difficulty. To see this we present the following diagram,

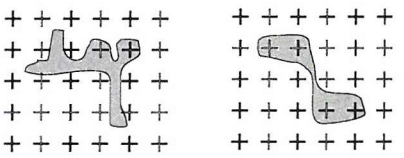


Figure 8.2: Examples of irregular continuous clusters. Here grey shapes represent pools and crosses are possible sampling points

As we can see in the first picture, the surface area is not easy to estimate simply from the number of crosses within the cluster. One could imagine cases where the surface area could be considerably bigger than expected, or conversely, cases where the surface area is considerably smaller. It would still seem intuitive that the number of crosses should lay some sort of restriction on the surface area.

The second problem, illustrated by the second diagram, is that the method of sampling nearest neighbour units may not give a clear indication of the size of pools. In this case it could be assumed that this pool was in fact two disjoint pools, or in a worse case scenario if both sets of crosses were not sampled we would only see half of a pool.

In order to avoid this difficulty we could adapt the sampling scheme presented by Thompson (1990) slightly. In this case if a point is found to lie within a pool, we would simply sample all of the points lying in the pool. This would be more difficult in the case of oil pools, as these tend to lie below ground and the surface area is not clearly visible.

Alternatively, we could think of approximating the surface area by making the log of the surface area follow a normal distribution with the 90% confidence interval being defined as lying between the area within the points and the area made up of the next set of points out from this, ie.

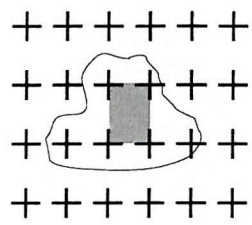


Figure 8.3: Illustration of a possible surface area approximation. Here the light grey area will form the 90% quantile and the dark grey area will form the 10% quantile

This distribution would have to be truncated at zero.

By modelling the surface are in this way we could simply exchange the Poisson for this truncated normal distribution and proceed as before.

It would be possible to model the oil pool data using this 'add on' method, exchanging the Poisson distribution in chapter 7 for the multivariate normal distribution and modelling the logged values in the same way as West (1996). The proposal distribution should always be chosen carefully, however, so as not to affect the ratio in the Metropolis-Hastings step.

8.2 Applying the Model to The Snowy Mountain Reservoir System

We cannot obtain the spatial data necessary to apply the model to the Rimbey-Meadowbrook Reef chain due to the sensitive nature of the oil pool data. We will therefore examine a series of reservoirs located in New South Wales, Australia. The data is similar, but a map of the locations is available which enables us to analyse the spatial aspect of the data.

We present a map of the system below.

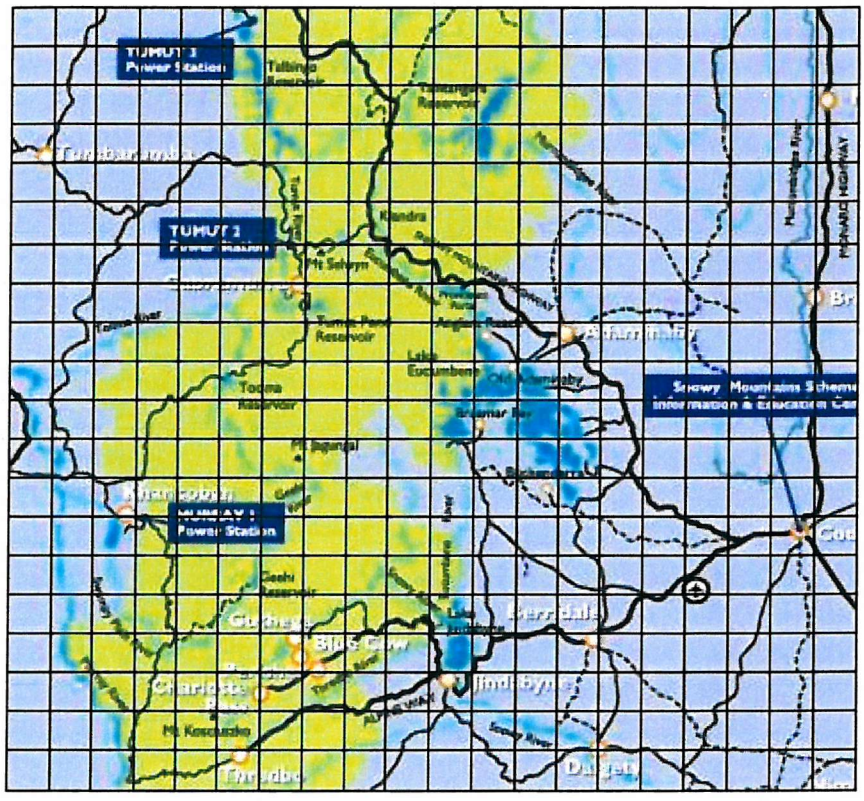


Figure 8.4: Snowy Moutain System of Reservoirs

It is an interesting aspect of this particular region that the reservoirs are very narrow, due to their location in the valleys of a mountain range. This

narrowness means that the point sampling method described above is inappropriate. It is clear that we would have to make the grid very fine to have a point within a reservoir. In this particular case we will revert to a grid sampling scheme. We assume that a grid cell satisfies the adaptive sampling condition if any part of the reservoir lies within it. This change is irrelevant as far as the implementation of the model is concerned. It could be considered as a design choice to be made at the time of sampling.

We will model the total volume of water contained in the system because this data is readily available. In this model the m_i's represent the log of the volume of water contained in reservoir i. We construct the likelihood function including the sampling distributions as follows,

$$\begin{aligned}
[\mathbf{M}, \mathbf{Y}, X, P|\alpha, \beta, \gamma, D] \quad = \quad & \prod_{i=1}^{n} \frac{z_i \times g_i}{b_i + t(U)} \\
& \times \binom{N}{X} \frac{\alpha^X (1-\alpha)^{N-X}}{1-(1-\alpha)^N} \\
& \times \binom{X}{P} \frac{\beta^P (1-\beta)^{X-P}}{1-(1-\beta)^X} \\
& \times (X-P)! \frac{\left(\frac{1}{P}\right)^{(y_1-1)} \cdots \left(\frac{1}{P}\right)^{(y_P-1)}}{(y_1-1)! \cdots (y_P-1)!} \\
& \times \prod_{i=1}^{P} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(m_i-\gamma y_i)^2}{2\sigma^2}} .
\end{aligned}$$

We allow the volume of water in each reservoir to depend on the size of the particular reservoir in the same way as for the discrete case.

8.3 The Gibbs sampler

The steps in the Gibbs sampler remain the same as in the previous chapters. The value of $\mathbf{M}$ will again be sampled with X_1, P_1 and $\mathbf{Y}$.

8.3.1 Conditional Posterior Distribution for γ

The conditional distribution for γ is formed as follows

$$\begin{aligned}
[\gamma|D,U,\mathbf{M}] &\propto prior(\gamma) \times \prod_{i=1}^{P} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(m_i-\gamma y_i)^2}{2\sigma^2}} \\
&\propto prior(\gamma) \times \frac{\sum_{i=1}^{P} y_i}{\sqrt{2\pi\sigma^2}} e^{\frac{-\sum_{i=1}^{P} y_i^2 \left(\gamma - \left(\frac{\sum_{i=1}^{P} y_i m_i}{\sum_{i=1}^{P} y_i^2}\right)\right)^2}{2\sigma^2}}.
\end{aligned}$$

Sampling Strategy for γ

We can sample γ and σ^2 from known distributions without the use of a Metropolis-Hastings step due to the form of the posterior.

As the prior for $\mathbf{M}$ did not affect the posterior in the previous example, we will assume a flat prior for γ and σ^2.

We can therefore sample γ from a Normal$\left(\frac{\sum_{i=1}^{P} y_i m_i}{\sum_{i=1}^{P} y_i^2}, \frac{\sigma^2}{\sum_{i=1}^{P} y_i^2}\right)$ and $\frac{1}{\sigma^2}$ from a Gamma$\left(\frac{P}{2}, \frac{1}{2}\sum(m_i - \gamma y_i)^2\right)$ distribution .

8.3.2 Remaining Conditional Distribution

We will again sample $\mathbf{M}_1, X_1, P_1$ and $\mathbf{Y}_1$ together. This is necessary because of the dependence between $\mathbf{M}$ and $\mathbf{Y}$.

The full conditional distribution for the remaining unknown parameters is

now,

$$[X_1, P_1, \mathbf{Y}_1, \mathbf{M}_1 | D, \mathbf{M}_0, \alpha, \beta, \gamma]$$

$$\propto \prod_{i=1}^{n} \frac{g_i}{b_i + t(U)} \times \frac{\alpha^{x_1}(1-\alpha)^{-x_1}}{(N - x_0 - x_1)!}$$
$$\times \frac{\beta^{p_1}(1-\beta)^{x_1 - p_1}}{(1-(1-\beta)^{x_0+x_1})(p_0+p_1)!}$$
$$\times \left(\frac{1}{p_0+p_1}\right)^{\sum_{i=1}^{p_0+p_1}(y_i-1)} \times \prod_{i \notin s}^{p_1} \frac{1}{(y_i-1)!}$$
$$\times \prod_{i=1}^{P} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(m_i - \gamma y_i)^2}{2\sigma^2}} .$$

Sampling Strategy

We will perform a Metropolis-Hastings accept-reject step in exactly the same way as in the previous chapter, generating $\mathbf{M}_1$ as independent Normal$(\varphi\bar{y}, \sigma^2)$ random variables. The proposal is therefore,

$$[X_1, P_1, \mathbf{Y}_1]_{prop} = \frac{1}{2} \times \frac{x_1}{p_1} \frac{\beta^{p_1}(1-\beta)^{x_1-p_1}}{1-(1-\beta)^{x_1}} \left(\frac{1}{p_1}\right)^{\sum_{i \notin s}(y_i-1)} \prod_{i \notin s} \frac{1}{(y_i-1)!}$$
$$\times \prod_{i \notin s} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(m_i - \gamma y_i)^2}{2\sigma^2}} .$$

Constructing the proposal in this way means that all of the terms depending on $\mathbf{M}$ cancel with the target as noted above, so the logged ratio of the target to the proposal remains identical to the previous chapter,

$$\log\left(\frac{[X_1, P_1, \mathbf{Y}_1\mathbf{M}_1|D, \mathbf{M}_0, \alpha, \beta, \gamma]}{[X_1, P_1, \mathbf{Y}_1, \mathbf{M}_1]_{prop}}\right) \propto \log(2B) + \sum_{i=1}^{n} \log\left(\frac{g_{i,j}}{b_i + t(U)}\right)$$
$$- \log\left(\frac{(N - x_0 - x_1)!(p_0 + p_1)!x_1!}{p_1!}\right)$$
$$+ x_1 \log\left(\frac{\alpha}{1-\alpha}\right) + \log\left(\frac{1-(1-\beta)^{x_1}}{1-(1-\beta)^{x_0+x_1}}\right)$$
$$- \sum_{i \in s} (y_i - 1) \log(p_0 + p_1)$$
$$+ \sum_{i \notin s} (y_i - 1) \log\left(\frac{p_1}{p_0 + p_1}\right).$$

8.4 Results

We ran a Gibbs sampler for 40000 iterations with a burn-in of 2000. We sample as described earlier from the population of reservoirs in the snowy mountain reservoir system. The following data and the earlier map are taken from the Snowy Hydro website (Snowy-Hydro, 2005).

Reservoir	Size (Grid cells)	Volume ($10^3 m^3$)
Talbingo	4	920600
Eucumbene	13	4798400
Geehi	1	21100
Tumut Pond	1	52800
Jindabyne	4	689900
Tooma	1	28100
Tumut 2	1	2700
Tantangara	4	254100
Jounama	1	43500
Murray 2	1	1760
Guthega	1	1550
Khancoban	1	21500

Table 8.1: Table giving the sizes and volumes of the reservoirs in the snowy mountain chain. Note the Blowering reservoir is not included due to it's location outside of the square grid.

The gross volume of the reservoirs is 6836000 $10^3 m^3$. We took three separate samples of size 25 from the population of grid cells. In the first sample we sampled Eucumbene, Jindabyne and Tooma , in the second we sampled Eucumbene, Jindabyne and Murray 2 and in the third we sampled Eucumbene and Tantangara. The results were as follows. The first predicted total was 5084049 with a interquartile range of $(4685468, 73612846)$, the second 5011639 with a interquartile range of $(54657482, 41034987)$ and the third 66335674 with a interquartile range of $(5565467, 76606300788)$. We note that the distributions are heavily skewed, as we might expect from a log-normal distribution.

It is encouraging that we obtain an estimate of the right order of magnitude considering the data we have. If we look at the data we see that there are a large number of lakes in the area. In this example we have an X value of 33 which lies well within the advised range. However, there are 12 separate lakes in this chain, giving us a value for β of 0.36, which is well outside of the advised range. We would not therefore expect an accurate prediction. As in the previous example the risk is that we over estimate the total volume. This happens in the case where we have the least data, when we only sample two lakes and both of these contain a much higher volume of water than most of the other lakes. In this case the effect is exacerbated due to the fact that any one lake can contain a very large volume of water. Therefore overestimating the number of lakes by just 1 can raise the estimated total significantly. This effect is also exaggerated by the fact that a small overestimate on the log scale can create a marked difference on the true scale. In the cases where we have more data, however, the true answer is contained in the interquartile range and the estimate is of the right order of magnitude showing that the model is working well even in this case which is slightly outside of the recommended range.

The estimate we receive in this case while not accurate is of the correct magnitude and given that we are working slightly outside of the advised parameter range the model is behaving as we well as we would hope.

Chapter 9

Discussion and further work

We have successfully produced a model-based approach for modelling sparse clustered data which is more efficient than the design-based approach adopted in Thompson (1990).

The model can always produces results equivalent in accuracy and efficiency to the design-based methods. However, with a good choice of prior we can produce results which are closer to the true value and more efficient.

Due to the assumptions made in constructing the model, it performs better when there are a small number of cells containing members of the population. This is in contrast to the design-based estimator.

The beauty of the model developed in this thesis is its flexibility. It could feasibly be applied in any situation where adaptive cluster sampling is used and we have a known value for N, the number of units with a possibility of containing a member of the population.

It would be interesting to extend this work by implementing a Markov Chain Monte Carlo algorithm where all components are proposed simultaneously instead of the Gibbs sampler we use currently. Due to the very high dependence between the variables, the Gibbs sampler steps must be performed in groups. It would be interesting to see if sampling from the full distribution

in one step produced more accurate results, as the chain may move more freely.

Another extension could be to adapt this model for use in an epidemiological problem. Here we would not specify a grid; our sampling units would become the population members. A network would be formed from links between the population members, for instance by using family ties. In this case it would be conceivable that we would want to examine the case where N is not known. A distribution could be placed on N and could be sampled relatively simply within the Gibbs sampler. However, this adding this step would create more uncertainty within the model and it is unclear whether there would be enough data contained within the sample to make clear inference.

Appendix A

Bayesian Analysis of extension of Breckling et al. (1994)

A.1 Full set of Graphs for Bernoulli Model with known sampling probabitities

These graphs are created from values at twenty five points shown below.

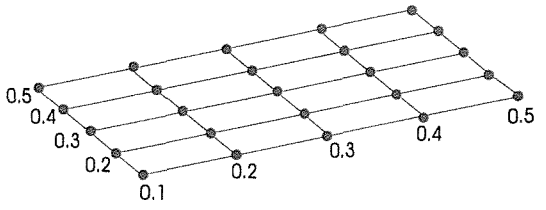


Figure A.1: Plot of the positions for which we have values. Each position is denoted by a filled circle.

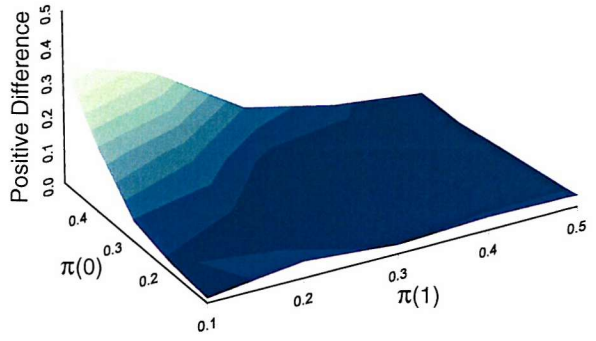


Figure A.2: Plot of the positive actual differences between the true values of α and the generated values of α for $\alpha = 0.5$.

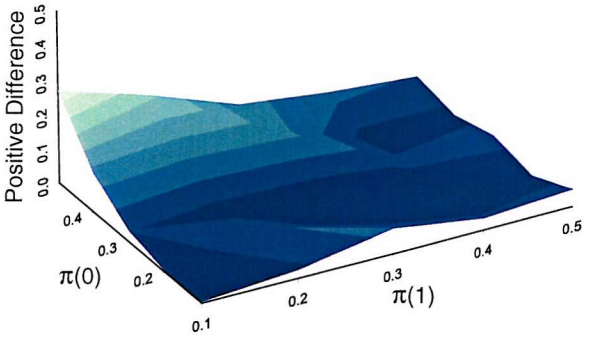


Figure A.3: Plot of the positive actual differences between the true values of α and the generated values of α for $\alpha = 0.5$.

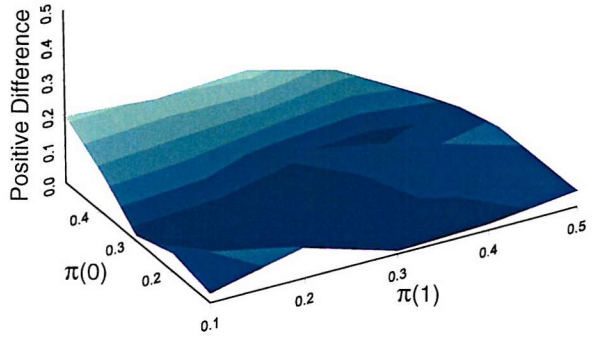


Figure A.4: Plot of the positive actual differences between the true values of α and the generated values of α for $\alpha = 0.5$.

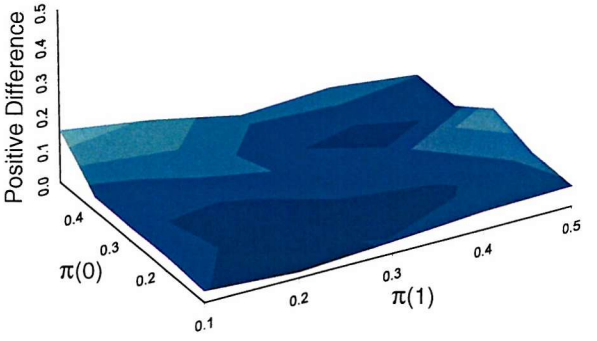


Figure A.5: Plot of the positive actual differences between the true values of α and the generated values of α for $\alpha = 0.5$.

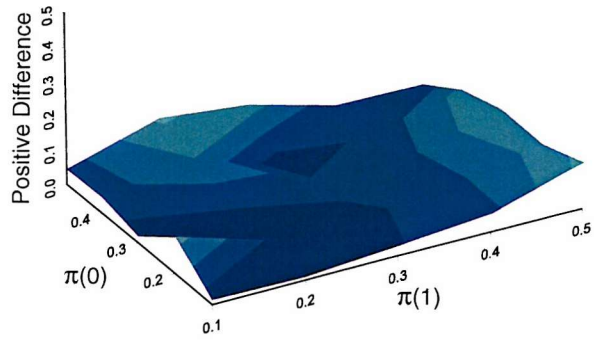


Figure A.6: Plot of the positive actual differences between the true values of α and the generated values of α for $\alpha = 0.5$.

A.2 Discussion of the Priors

In general we hope that a conjugate prior will not affect the parameter estimates for a model. We want only our sample data to determine the parameter estimates.

We therefore need to check that the parameter estimates remain relatively unchanged if the hyperparameters within the prior distribution are changed.

To determine this for the Bernoulli model with known sampling probabilities we put in several different hyperparameter values and compare the estimates returned for α.

The least informative prior is a uniform prior. This is achieved in a Beta prior by using the values $\gamma = 1$, $\delta = 1$. To see what estimates were produced we plot them against the value for $\hat{\alpha}$.

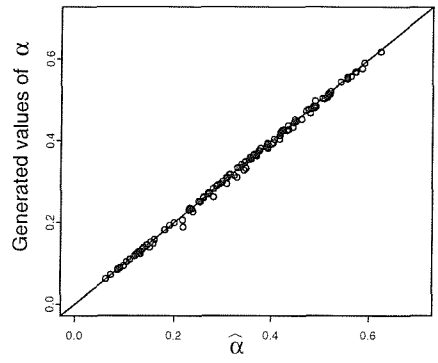


Figure A.7: Plot of the generated values of α against $\hat{\alpha}$ for uniform prior.

It is clear that the uniform prior is producing good estimates for α. We will compare this to two other possible priors.

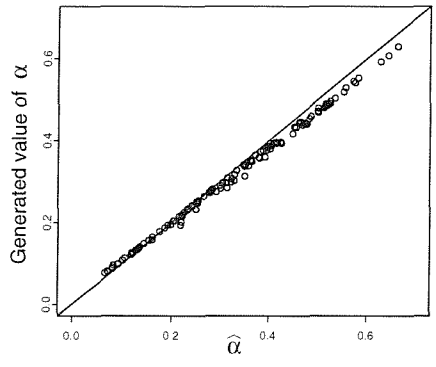


Figure A.8: Plot of generated values of α against $\hat{\alpha}$ for Beta$(2, 5)$.

All three priors give similar results. Although the second two are slightly affected by the prior values it is not a significant enough effect to cause concern. As we would expect, the non-uniform priors produce slightly more

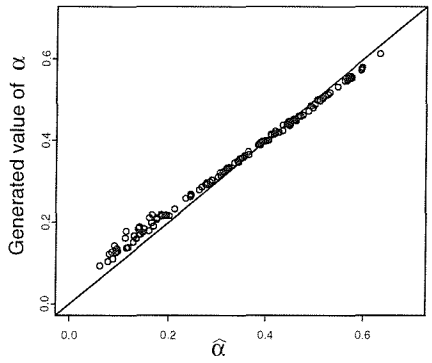


Figure A.9: Plot of the generated values of α against $\hat{\alpha}$ for uniform prior for Beta$(4, 5)$.

accurate estimates where their mean value is similar to that of $\hat{\alpha}$ because the data does not have to 'move' the mean of the prior.

It is clear that the model is well defined and that the priors do not swamp the data. However, a choice of prior must still be made. To do this we will compare the parameter estimates to their true values (the values of $\hat{\alpha}$ are biased). If we sum the positive differences between the actual value of alpha and the estimated values of alpha we can obtain an idea of how well each set of estimates fits the real data. As we might expect, when the mean of the prior is high the sum is highest (at 8.2). The uniform prior gives a sum of 7.8 which slightly better. The prior which allows the data to best fit the true values, however, is the Beta$(2, 5)$ prior distribution, with a summed value of 7.3. This can be explained if we consider that the mean of this data is lower than the higher values of α we are trying to estimate. So it is not surprising that it will pull down the parameter estimates for the higher values slightly. This means that the model will fit the true data better because the value of $\hat{\alpha}$ is slightly too large for higher values of α.

A.2.1 Conclusion

The choice of prior for this model is arbitrary, changing the prior does not make a substantial difference to the fit of the model. We know the actual population values so it would make sense to use the prior which best fits these values. However, it should be noted that exactly the same conclusions can be drawn if either of the other priors were used. To show this the full set of graphs for the uniform and Beta$(4, 5)$ priors have been included in the next two sections for comparison.

A.3 Uniform prior

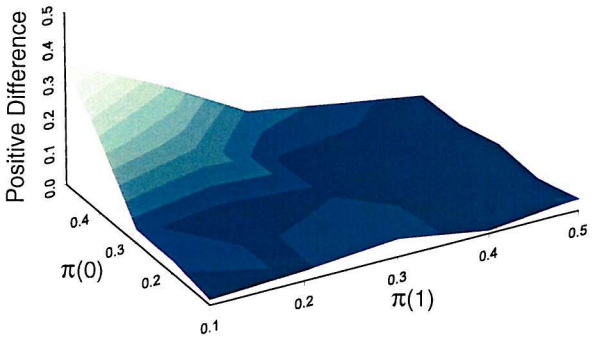


Figure A.10: Plot the of positive actual differences between the true values of α and the generated values of α for $\alpha = 0.5$.

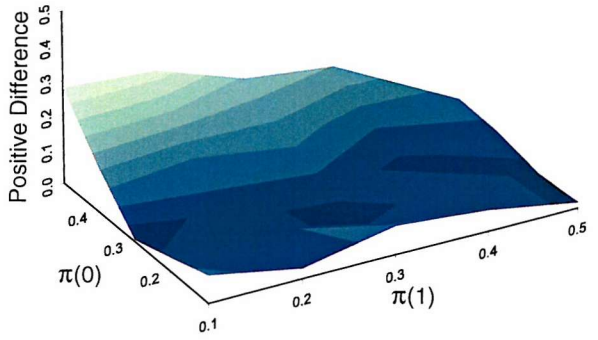


Figure A.11: Plot the of positive actual differences between the true values of α and the generated values of α for $\alpha = 0.5$.

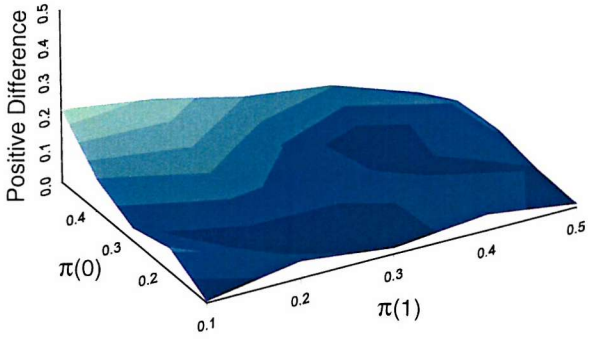


Figure A.12: Plot of the positive actual differences between the true values of α and the generated values of α for $\alpha = 0.5$.

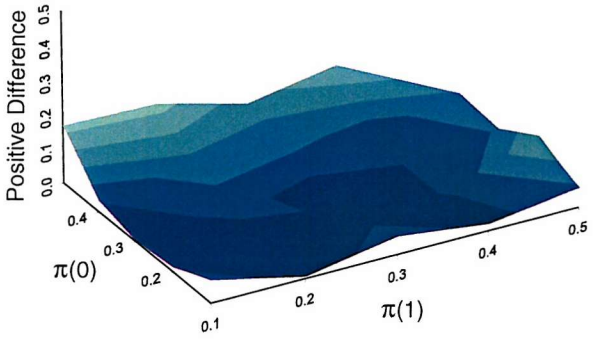


Figure A.13: Plot of the positive actual differences between the true values of α and the generated values of α for $\alpha = 0.5$.

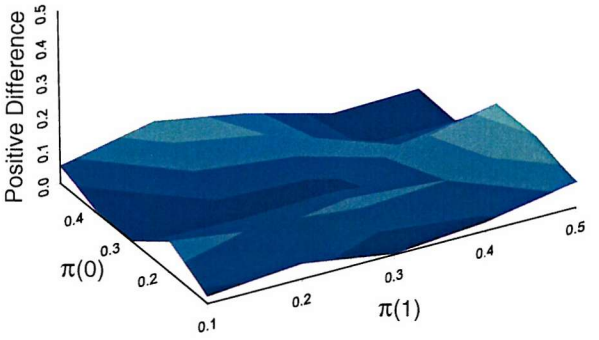


Figure A.14: Plot of the positive actual differences between the true values of α and the generated values of α for $\alpha = 0.5$.

A.4 Beta$(4, 5)$ prior

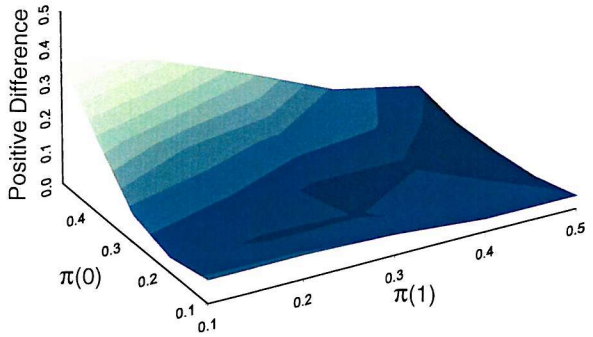


Figure A.15: Plot of the positive actual differences between the true values of α and the generated values of α for $\alpha = 0.5$.

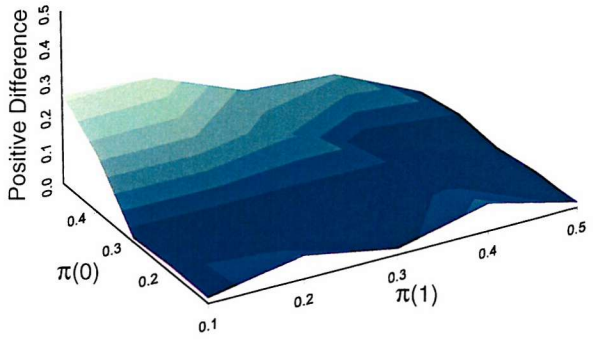


Figure A.16: Plot of the positive actual differences between the true values of α and the generated values of α for $\alpha = 0.5$.

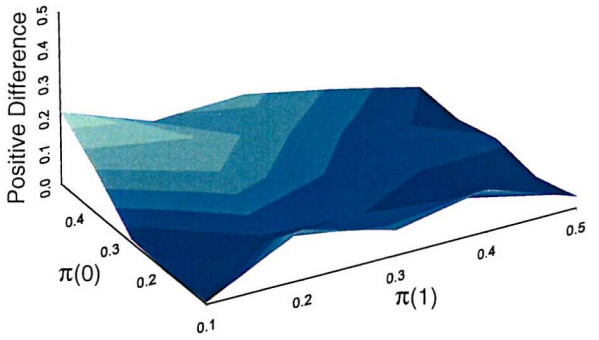


Figure A.17: Plot of the positive actual differences between the true values of α and the generated values of α for $\alpha = 0.5$.

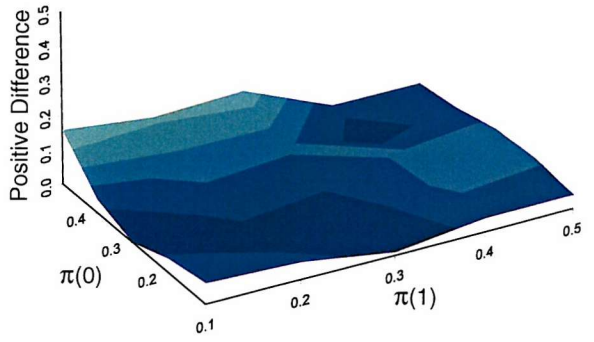


Figure A.18: Plot of the positive actual differences between the true values of α and the generated values of α for $\alpha = 0.5$.

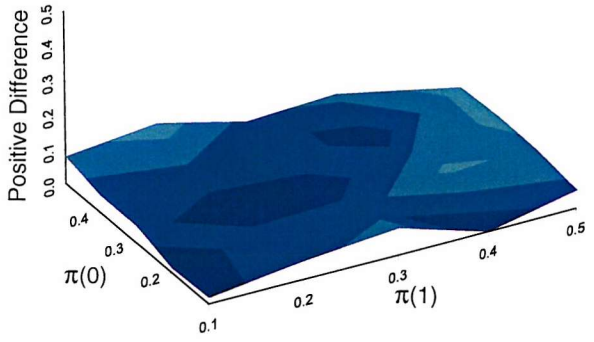


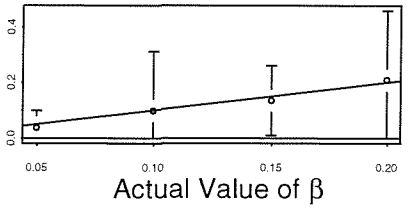
Figure A.19: Plot of the positive actual differences between the true values of α and the generated values of α for $\alpha = 0.5$.

Appendix B

Bayesian Analysis of model with Fixed number of nonempty cells

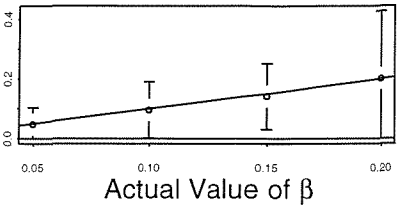
We present the full set of plots of actual parameter values against the medians of those generated from the Gibbs sampler for all priors. It is clear that this model is fairly robust to changes in prior.

N_0=40, $\pi(\beta)$=Beta(1,1)



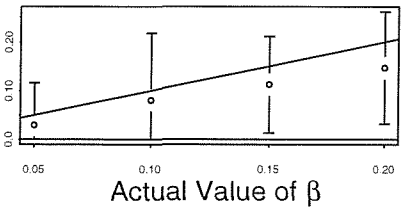
Actual Value of β

N_0=80, $\pi(\beta)$=Beta(1,1)



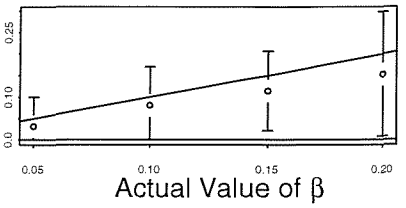
Actual Value of β

N_0=40, $\pi(\beta)$=Beta(1,9)



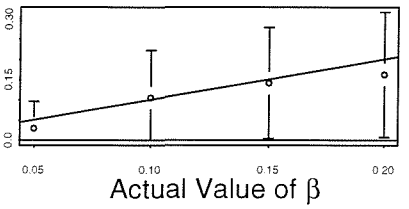
Actual Value of β

N_0=80, $\pi(\beta)$=Beta(1,9)



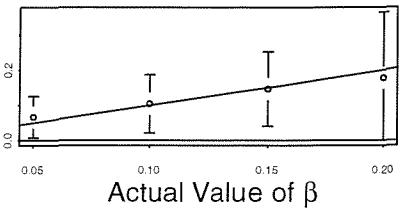
Actual Value of β

N_0=40, $\pi(\beta)$=Beta(2,9)



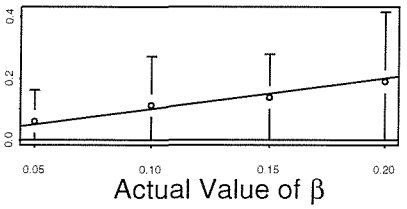
Actual Value of β

N_0=80, $\pi(\beta)$=Beta(2,9)



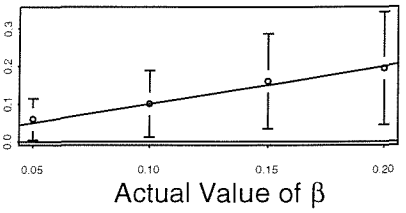
Actual Value of β

N_0=40, $\pi(\beta)$=Beta(2,5)



Actual Value of β

N_0=80, $\pi(\beta)$=Beta(2,5)

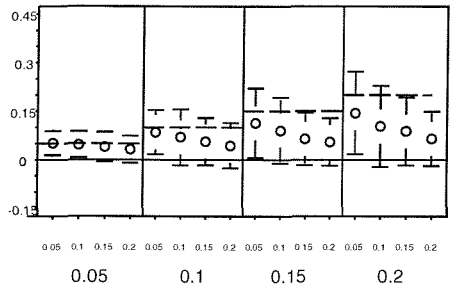


Actual Value of β

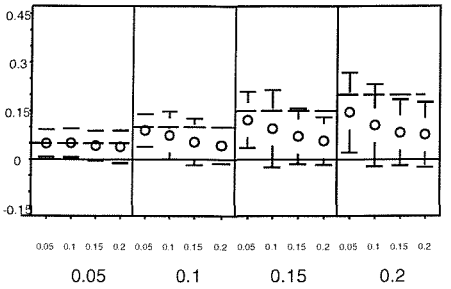
Appendix C

Full set of Graphs for model with a random number of non-empty cells

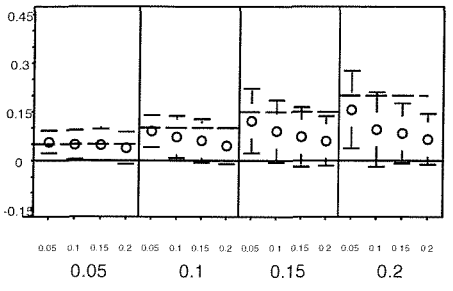
π(α)~ Beta(1,25), π(β)= Beta(1,9)



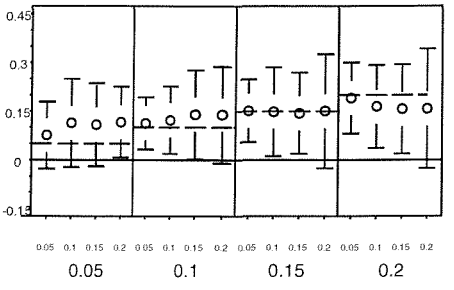
π(α)~ Beta(1,25), π(β)= Beta(2,9)



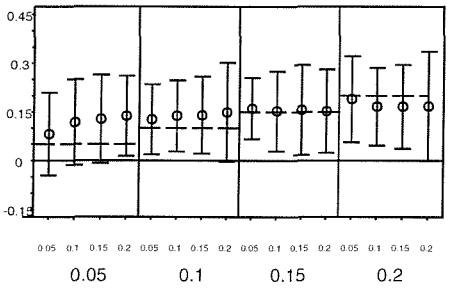
π(α)~ Beta(1,25), π(β)= Beta(2,5)



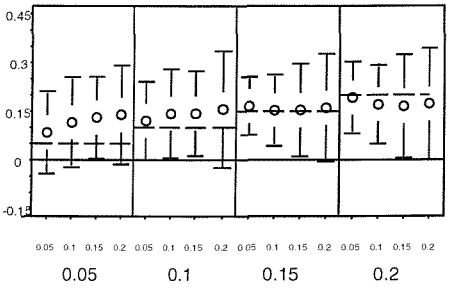
π(α)~ Beta(3,15), π(β)= Beta(1,9)



π(α)~ Beta(3,15), π(β)= Beta(2,9)



π(α)~ Beta(3,15), π(β)= Beta(2,5)



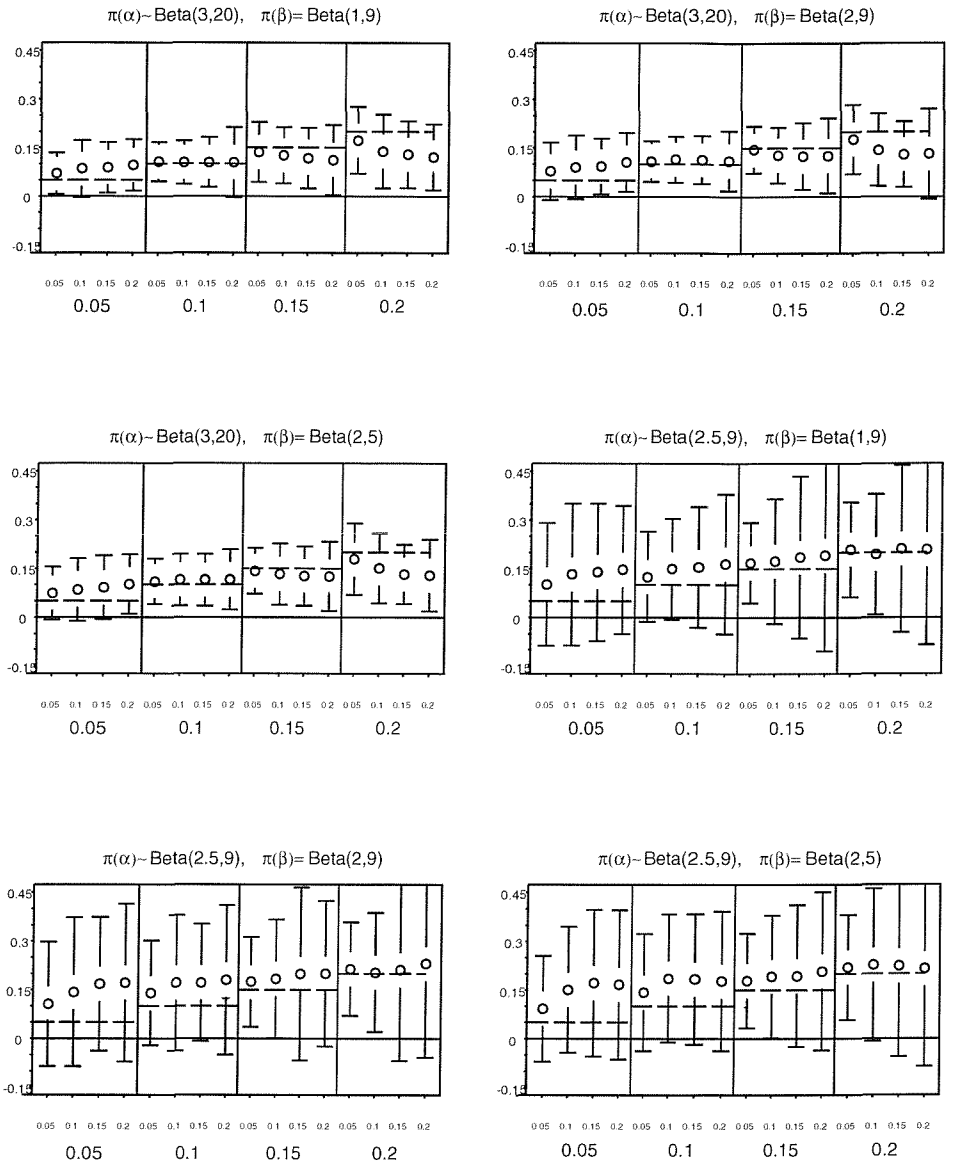
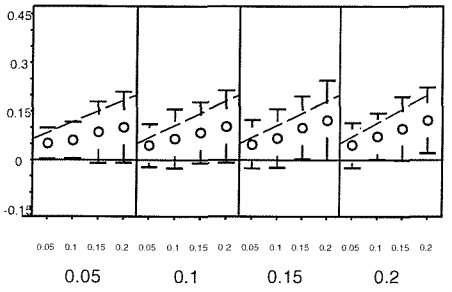
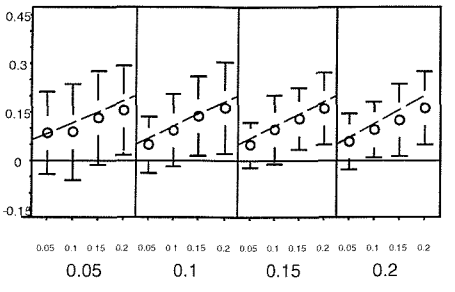


Figure C.1: Plot showing the median predicted values of the parameter α. The first row under each graph gives the initial value of β and the second gives the initial value of α. Interval width is three standard deviations in each direction so negative values should be read as zero.

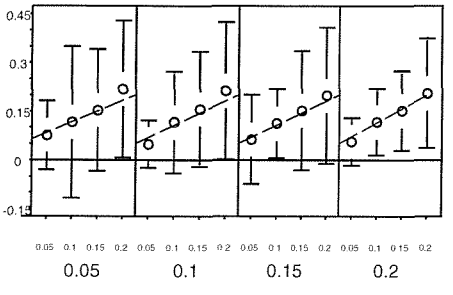
π(α)~ Beta(1,25), π(β)= Beta(1,9)



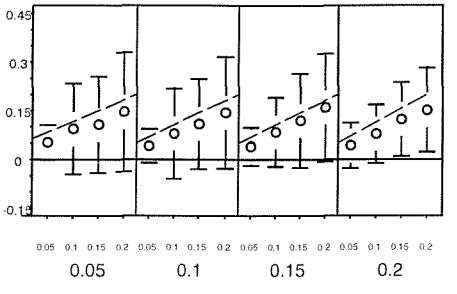
π(α)~ Beta(1,25), π(β)= Beta(2,9)



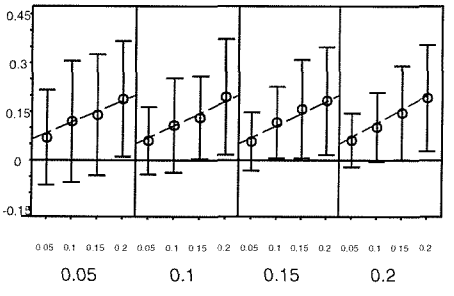
π(α)~ Beta(1,25), π(β)= Beta(2,5)



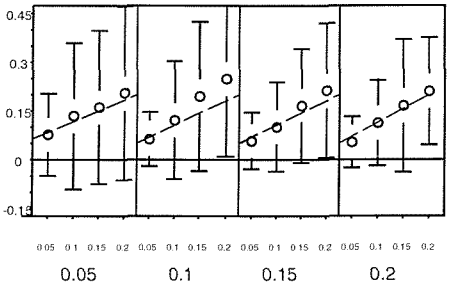
π(α)~ Beta(3,15), π(β)= Beta(1,9)



π(α)~ Beta(3,15), π(β)= Beta(2,9)



π(α)~ Beta(3,15), π(β)= Beta(2,5)



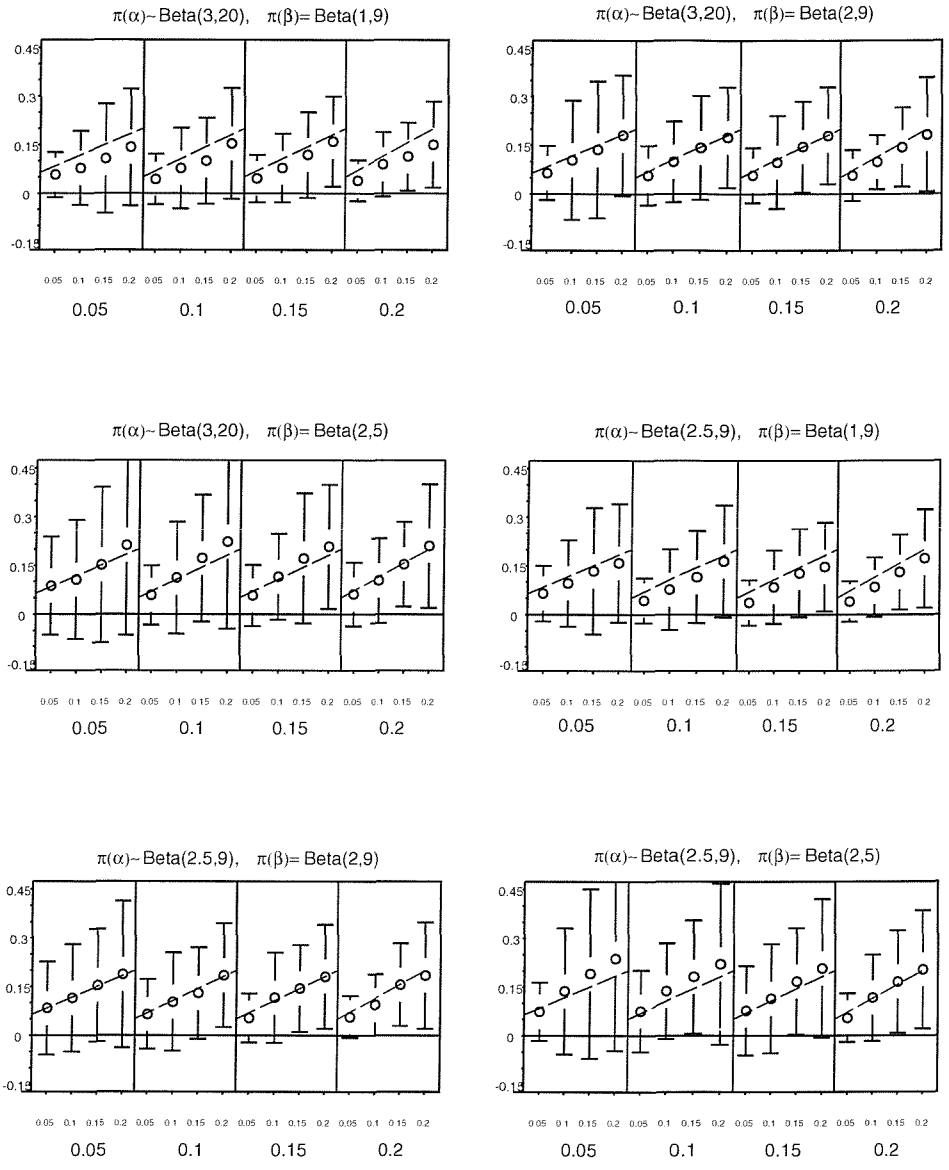


Figure C.2: Plot showing the median predicted values of the parameter β. The first row under each graph gives the initial value of β and the second gives the initial value of α. Interval width is three standard deviations in each direction so negative values should be read as zero.

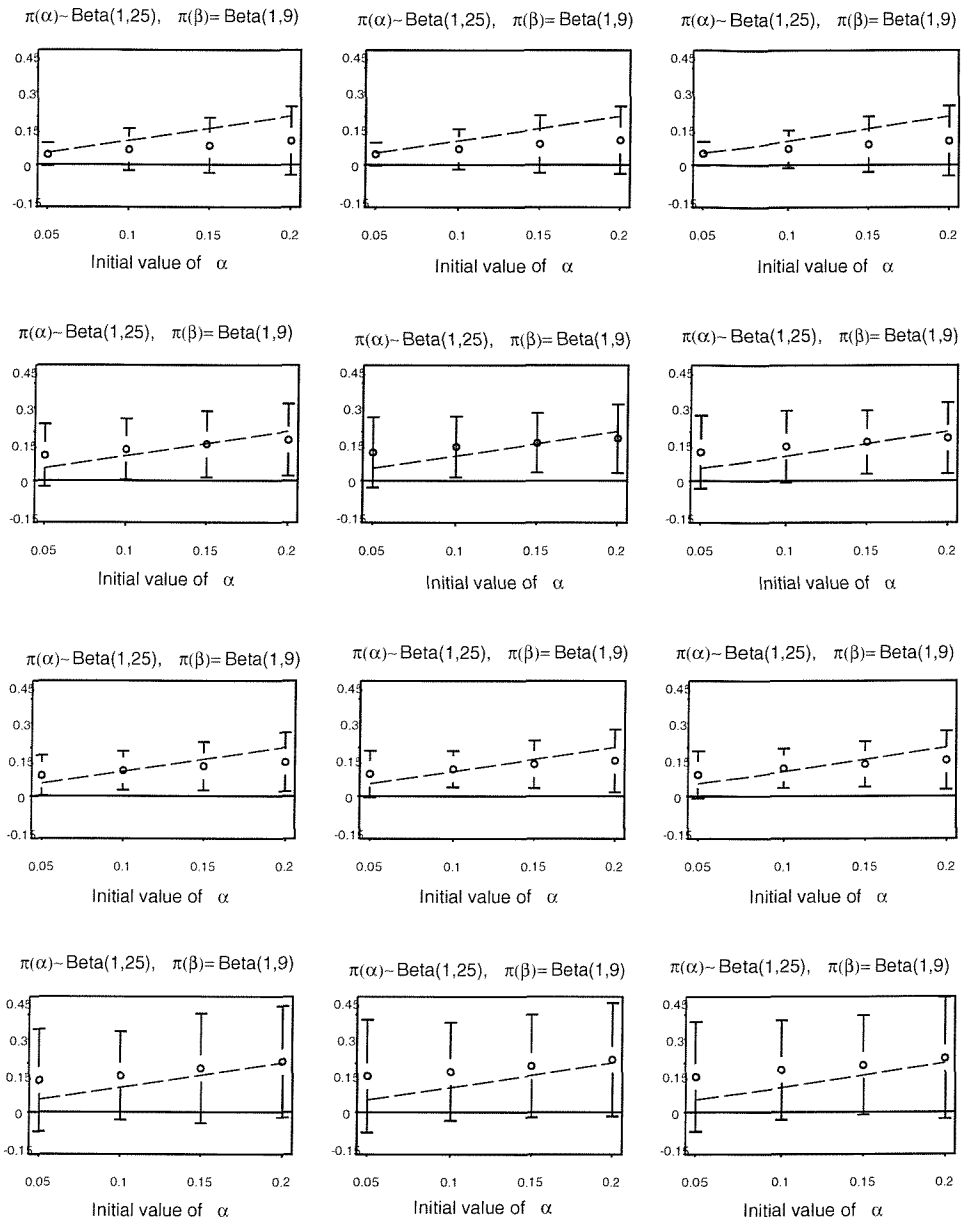


Figure C.3: Plot of the median predicted value of the parameter α against the initial parameter values of α.

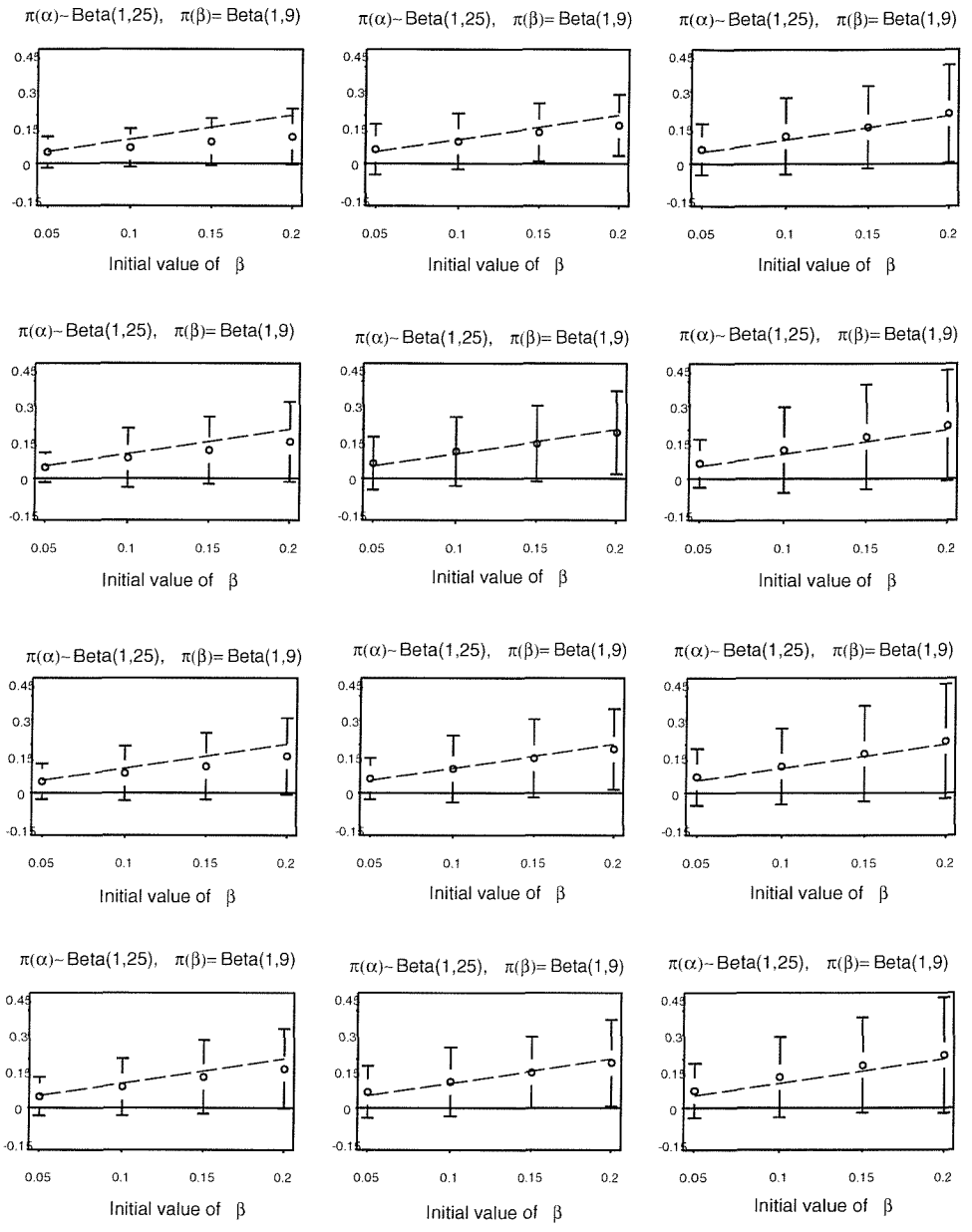
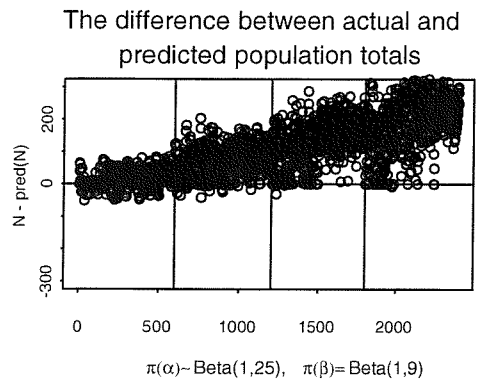


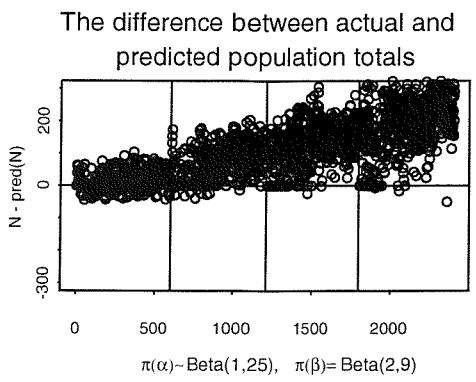
Figure C.4: Plot of the median predicted value of the parameter β against the initial parameter values of β.

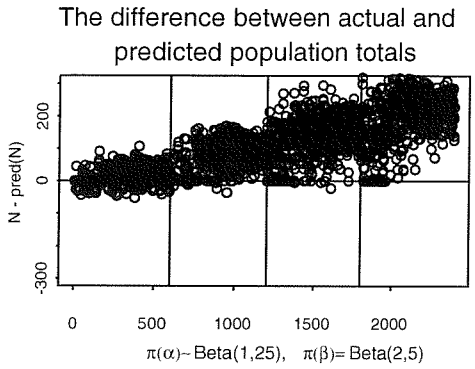
Appendix D

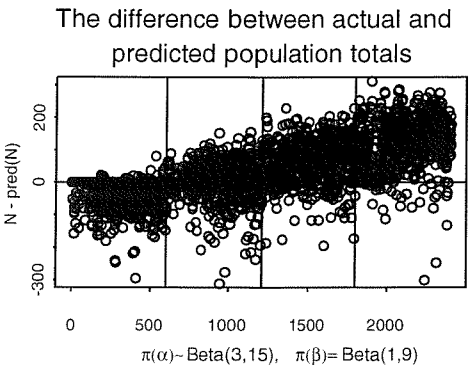
Full set of Graphs with an unknown population total

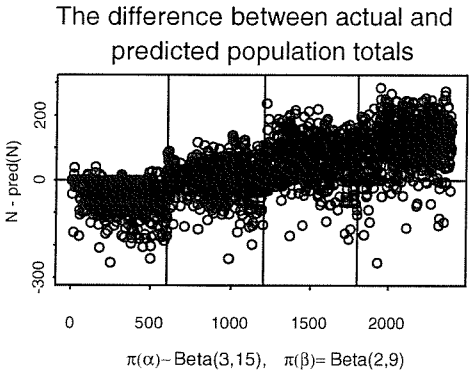
D.0.1 $\pi(\gamma) =$**Gamma**$(2, 7)$

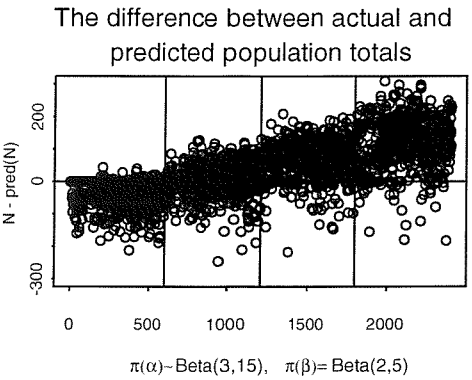

The difference between actual and predicted population totals
N - pred(N)
200
0
-300
0
500
1000
1500
2000
π(α)~Beta(1,25), π(β)= Beta(1,9)


The difference between actual and predicted population totals
N - pred(N)
200
0
-300
0
500
1000
1500
2000
π(α)~Beta(1,25), π(β)= Beta(2,9)


The difference between actual and predicted population totals
N - pred(N)
200
0
-300
0
500
1000
1500
2000
π(α)~Beta(1,25), π(β)= Beta(2,5)


The difference between actual and predicted population totals
N - pred(N)
200
0
-300
0
500
1000
1500
2000
π(α)~Beta(3,15), π(β)= Beta(1,9)


The difference between actual and predicted population totals
N - pred(N)
200
0
-300
0
500
1000
1500
2000
π(α)~Beta(3,15), π(β)= Beta(2,9)


The difference between actual and predicted population totals
N - pred(N)
200
0
-300
0
500
1000
1500
2000
π(α)~Beta(3,15), π(β)= Beta(2,5)

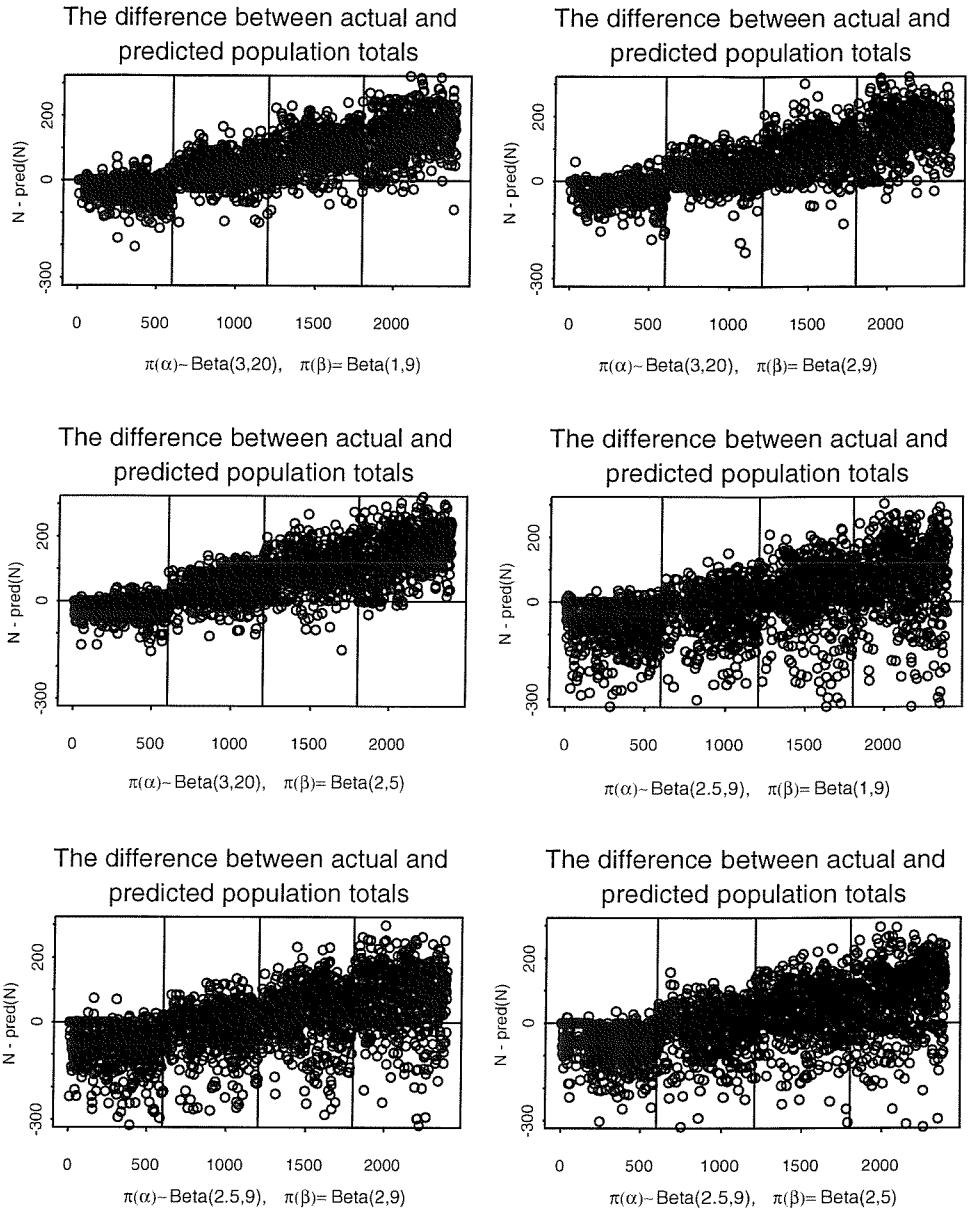
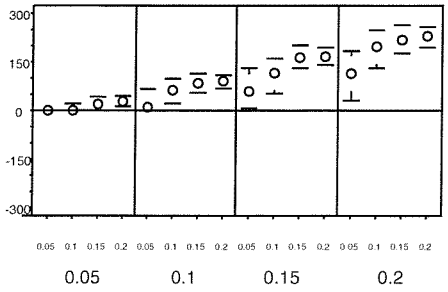
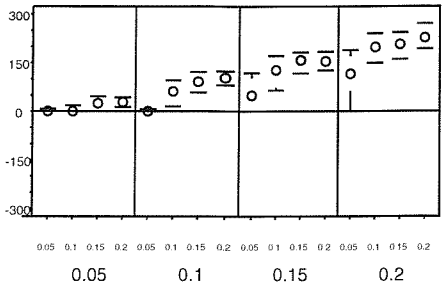


Figure D.1: Plot of the difference between the actual values of M and the generated values of M over the different initial values of α. In the first section of each plot $\alpha = 0.05$, the second $\alpha = 0.1$, the third $\alpha = 0.15$ and the fourth $\alpha = 0.2$. In each section β increases over the same intervals from left to right.

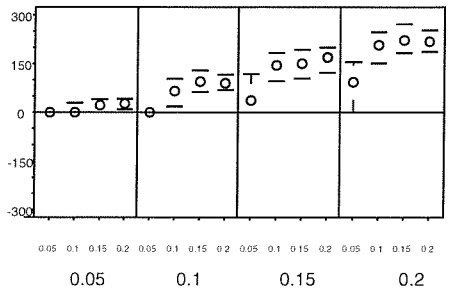
π(α)~Beta(1,25), π(β)= Beta(1,9)



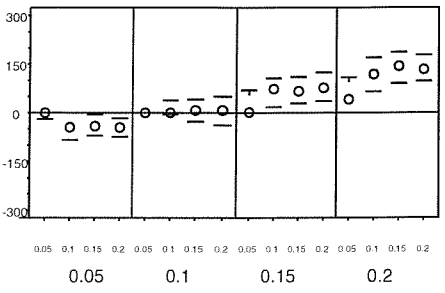
π(α)~Beta(1,25), π(β)= Beta(2,9)



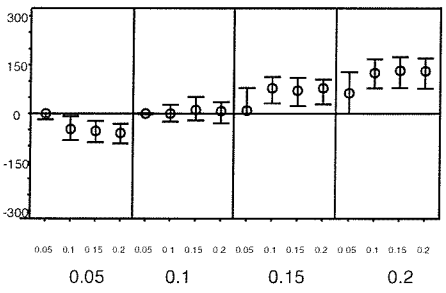
π(α)~Beta(1,25), π(β)= Beta(2,5)



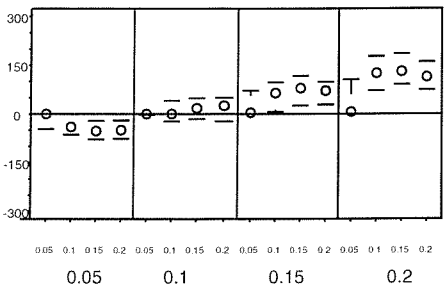
π(α)~Beta(3,15), π(β)= Beta(1,9)



π(α)~Beta(3,15), π(β)= Beta(2,9)



π(α)~Beta(3,15), π(β)= Beta(2,5)



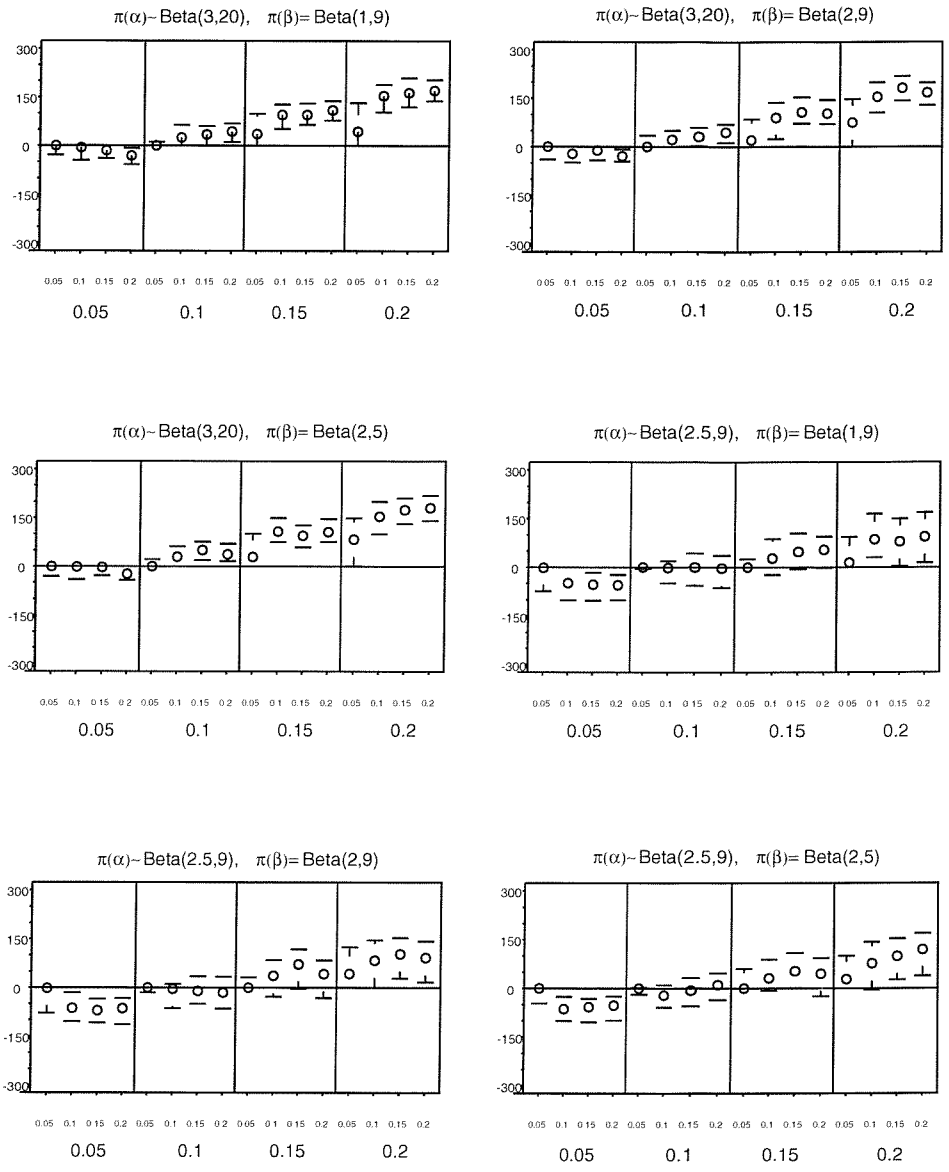


Figure D.2: Plot of the median differences between the actual values of M and the generated values of M.

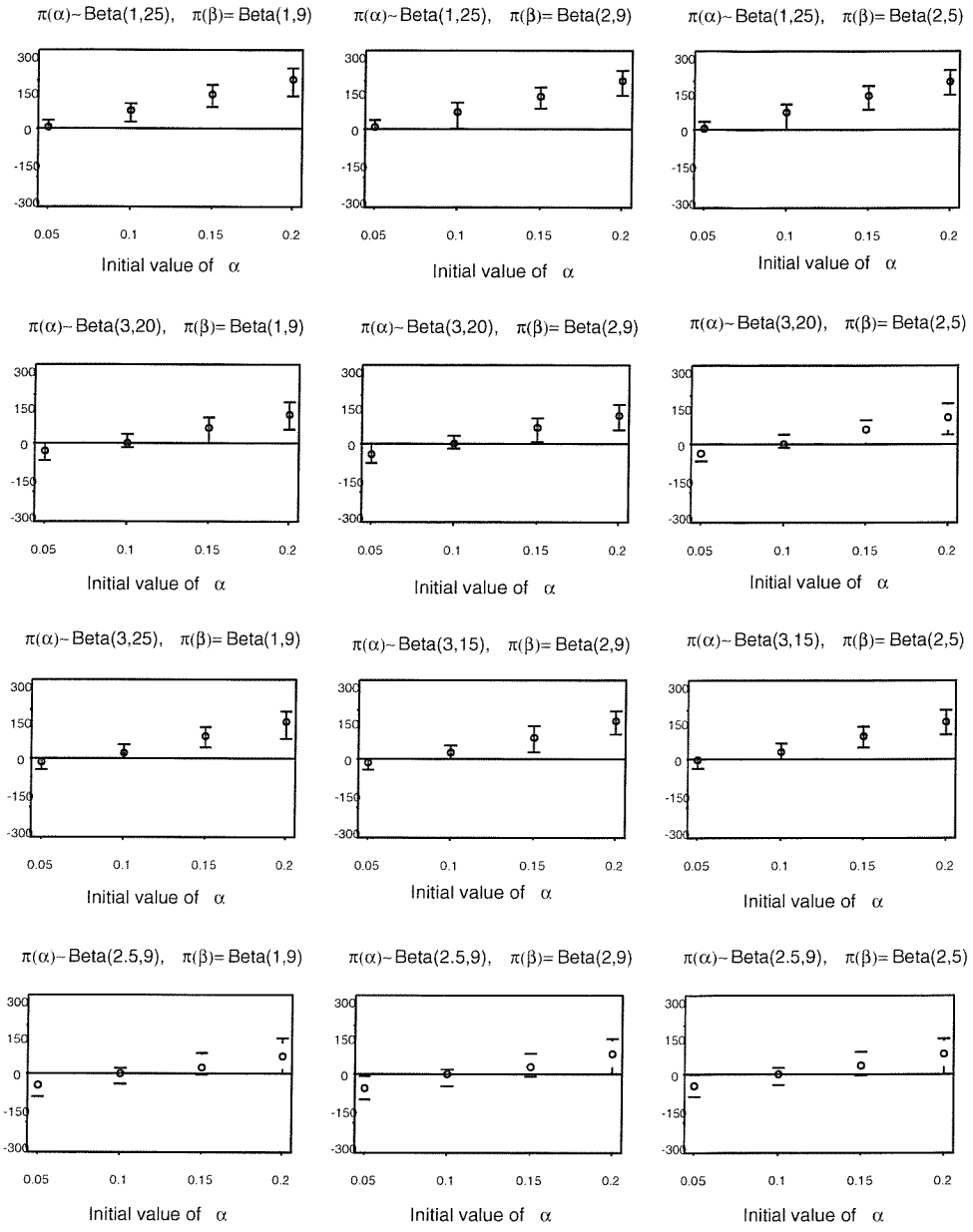


Figure D.3: Plot of the the median values of the difference between M and the predicted value of M over the initial parameter values of α.

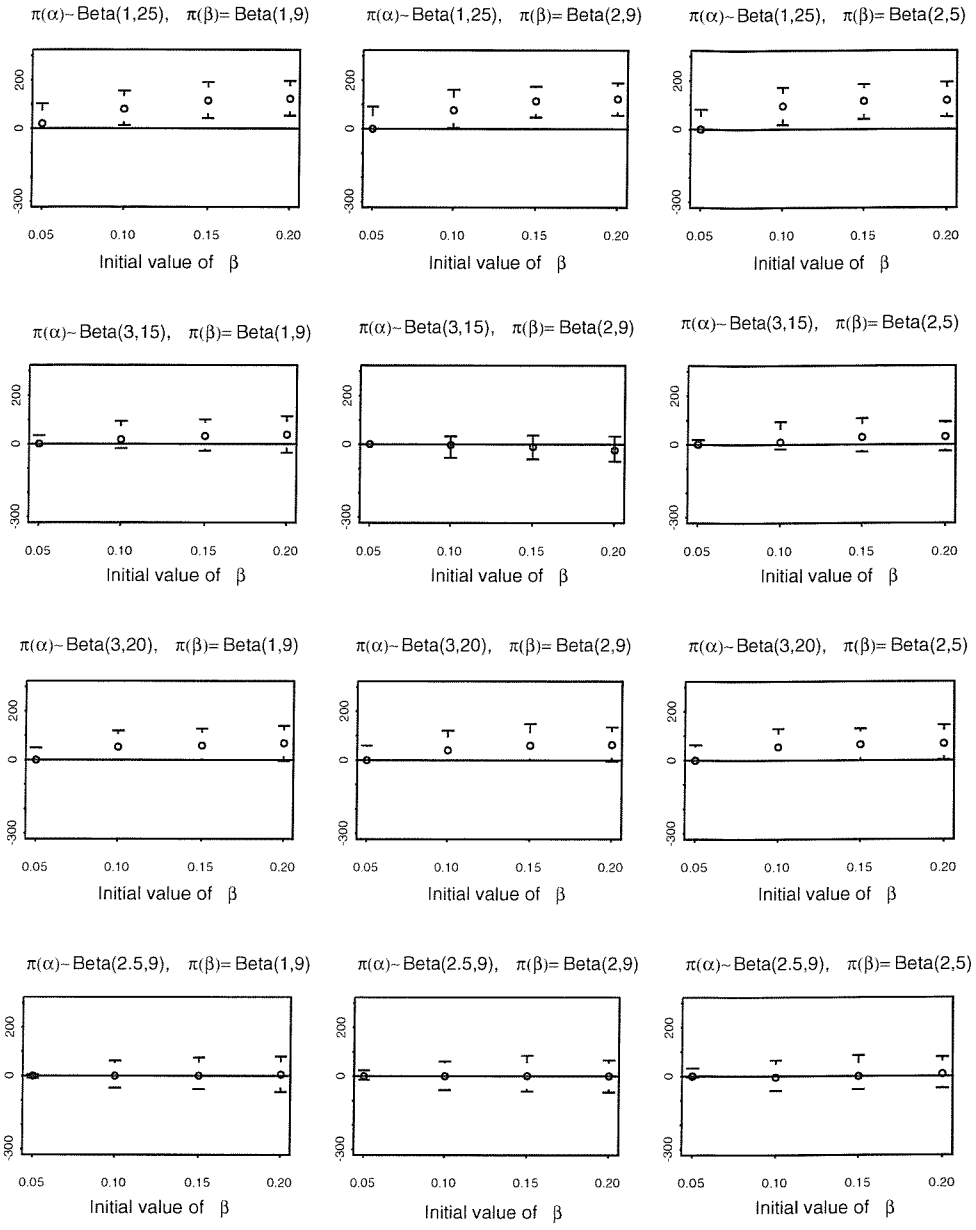
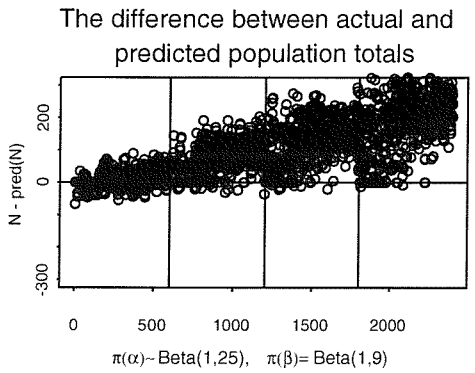
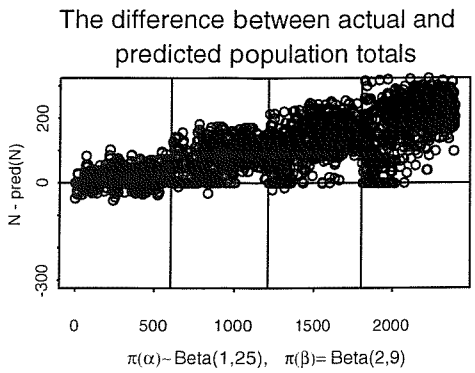
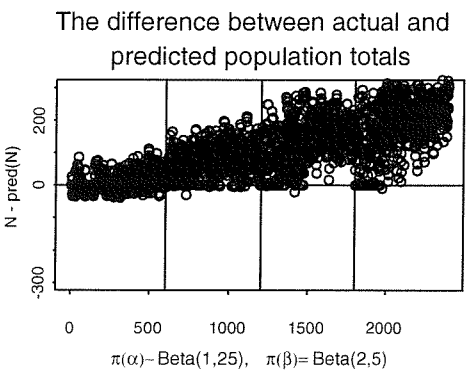


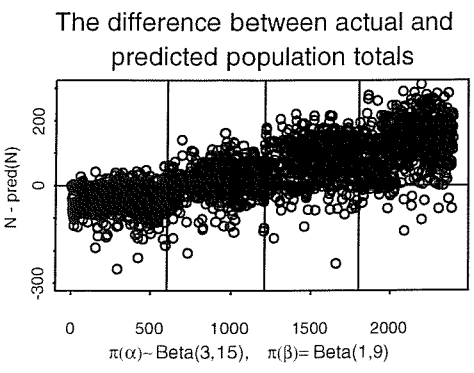
Figure D.4: Plot of the the median values of the difference between M and the predicted value of M over the initial parameter values of β.

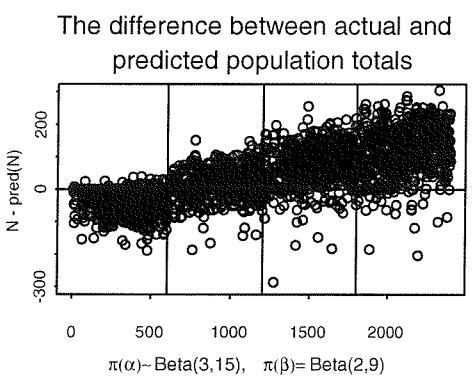
D.0.2 $\pi(\gamma)$ =Gamma$(5, 2)$

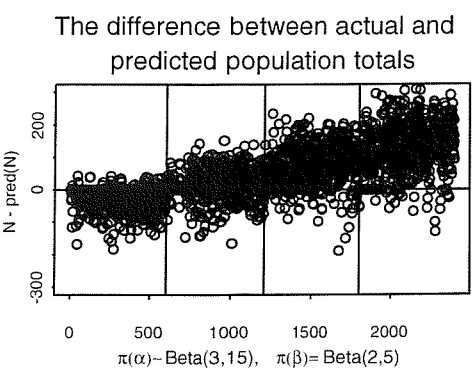












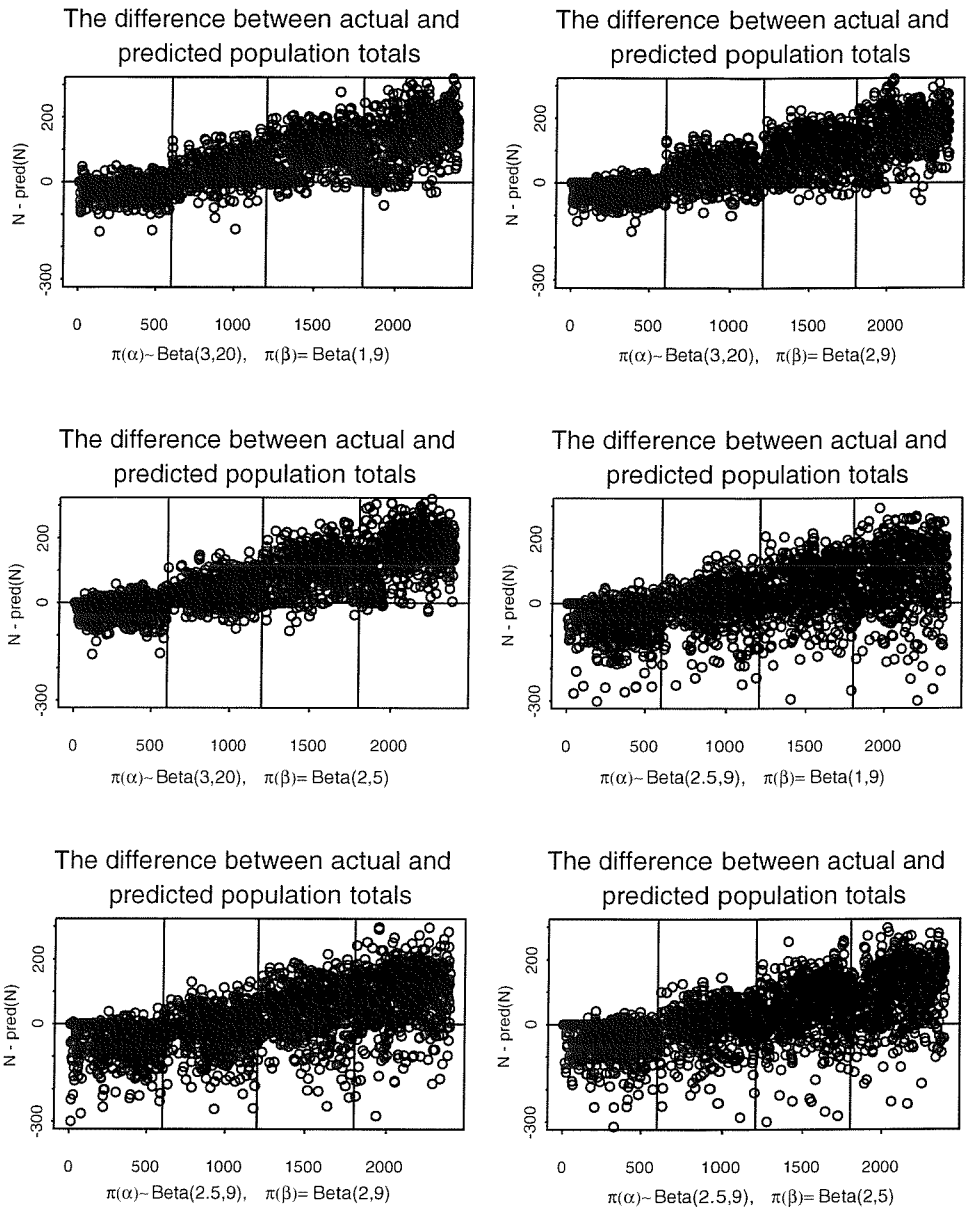
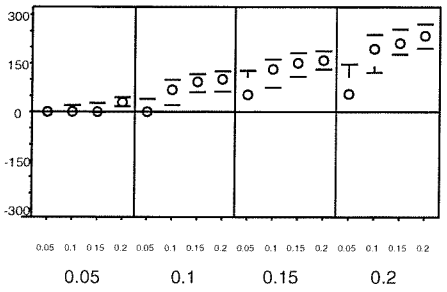
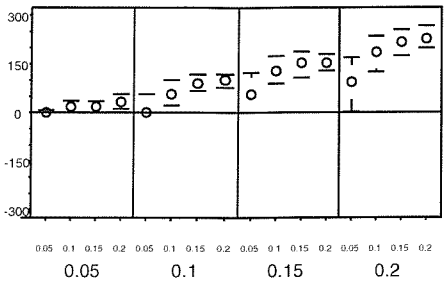


Figure D.5: Plot of the difference between the actual values of M and the generated values of M over the different initial values of α. In the first section of each plot $\alpha = 0.05$, the second $\alpha = 0.1$, the third $\alpha = 0.15$ and the fourth $\alpha = 0.2$. In each section β increases over the same intervals from left to right.

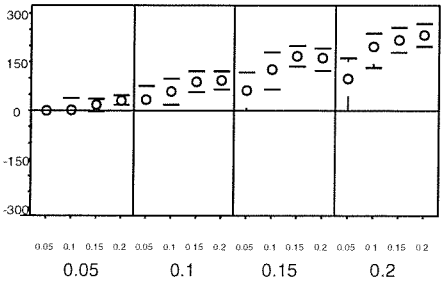
π(α)~Beta(1,25), π(β)= Beta(1,9)



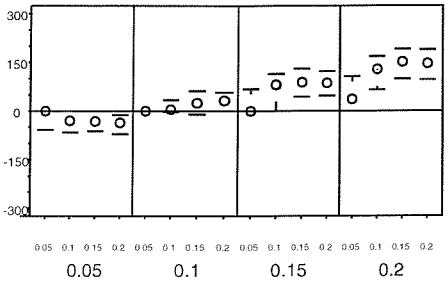
π(α)~Beta(1,25), π(β)= Beta(2,9)



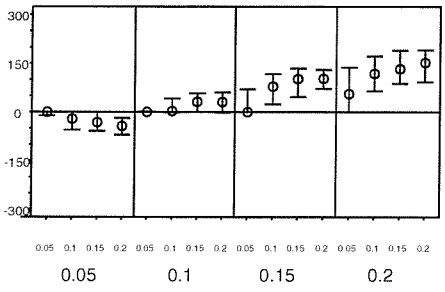
π(α)~Beta(1,25), π(β)= Beta(2,5)



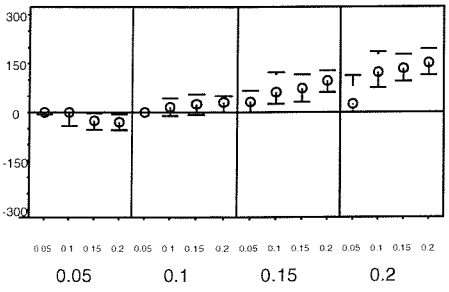
π(α)~Beta(3,15), π(β)= Beta(1,9)



π(α)~Beta(3,15), π(β)= Beta(2,9)



π(α)~Beta(3,15), π(β)= Beta(2,5)



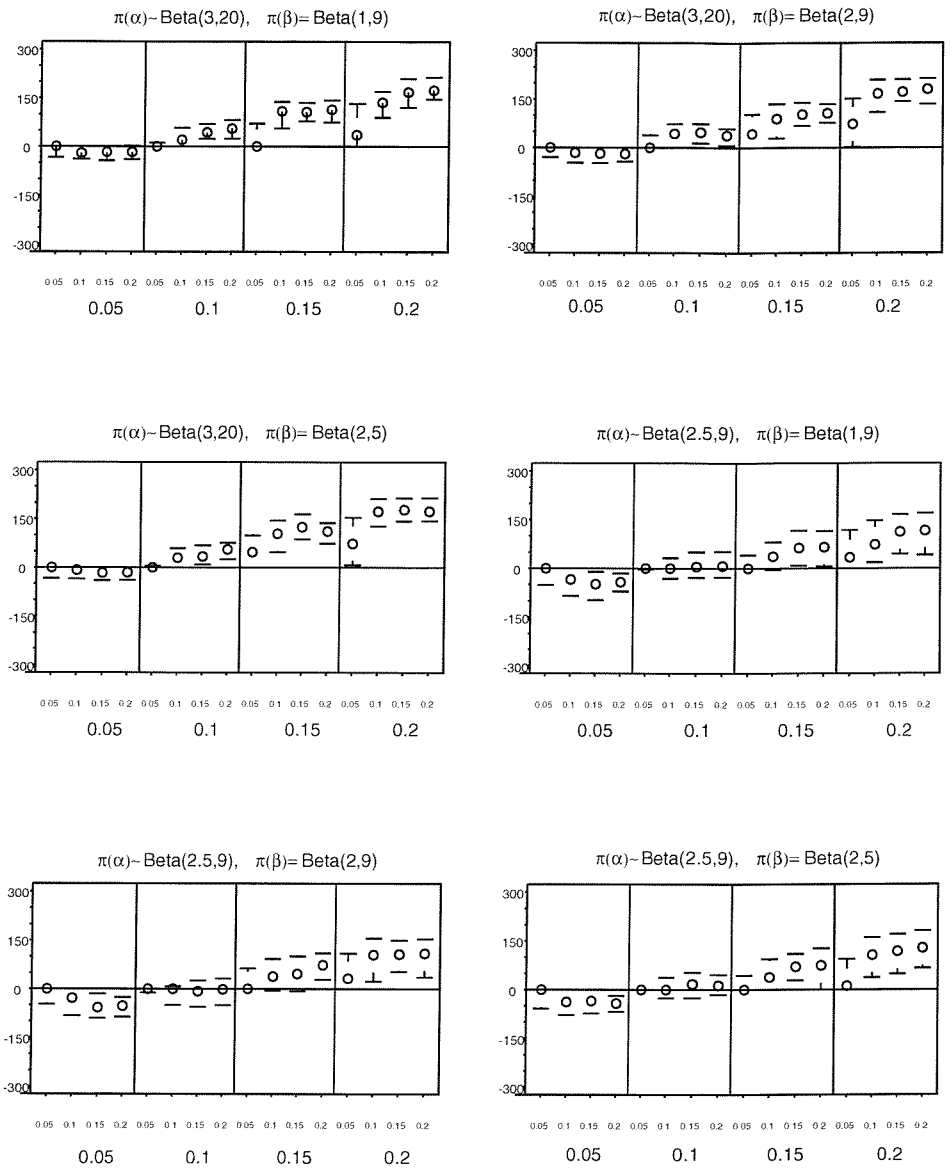


Figure D.6: Plot of the median differences between the actual values of M and the generated values of M.

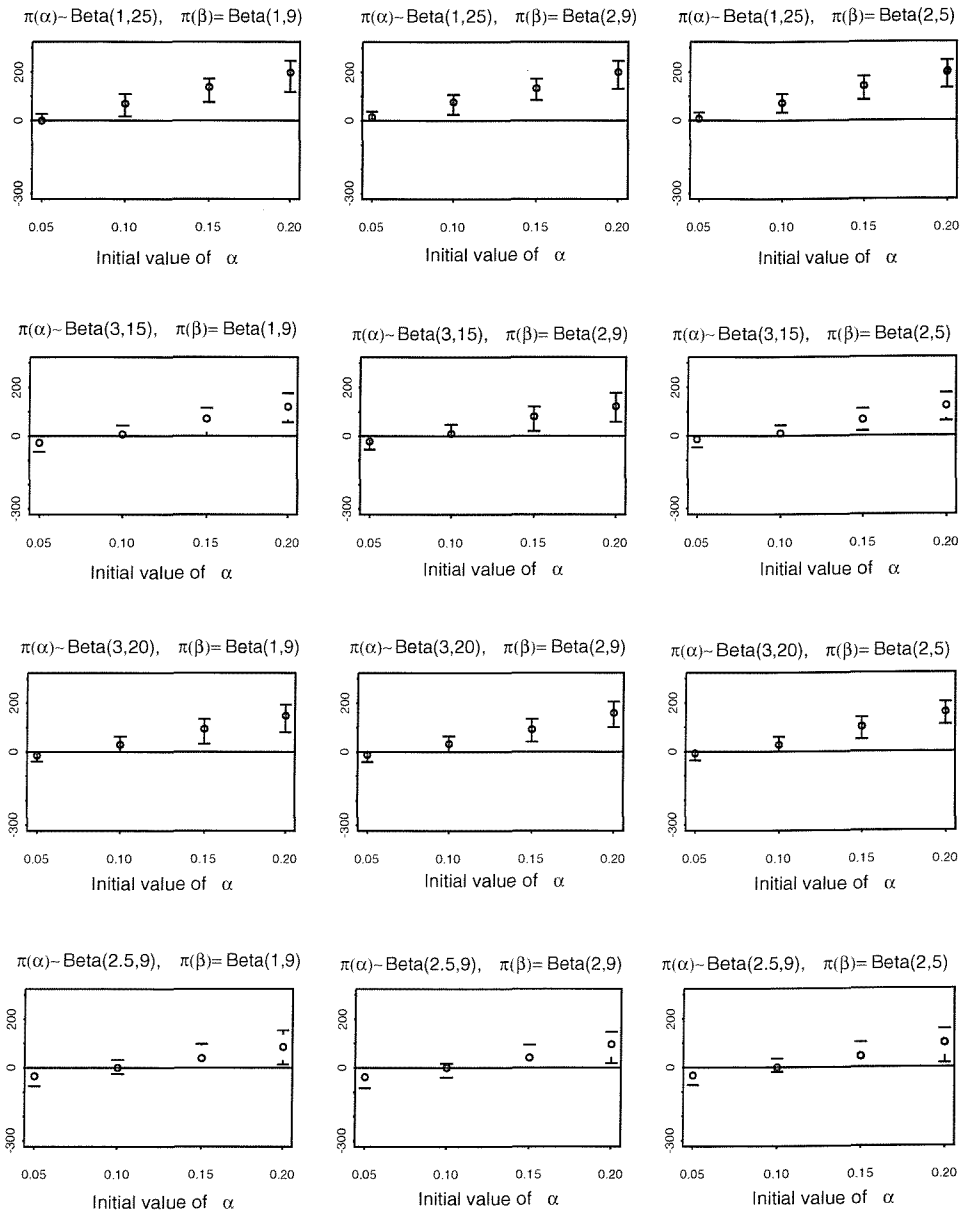


Figure D.7: Plot of the the median values of the difference between M and the predicted value of M over the initial parameter values of α.

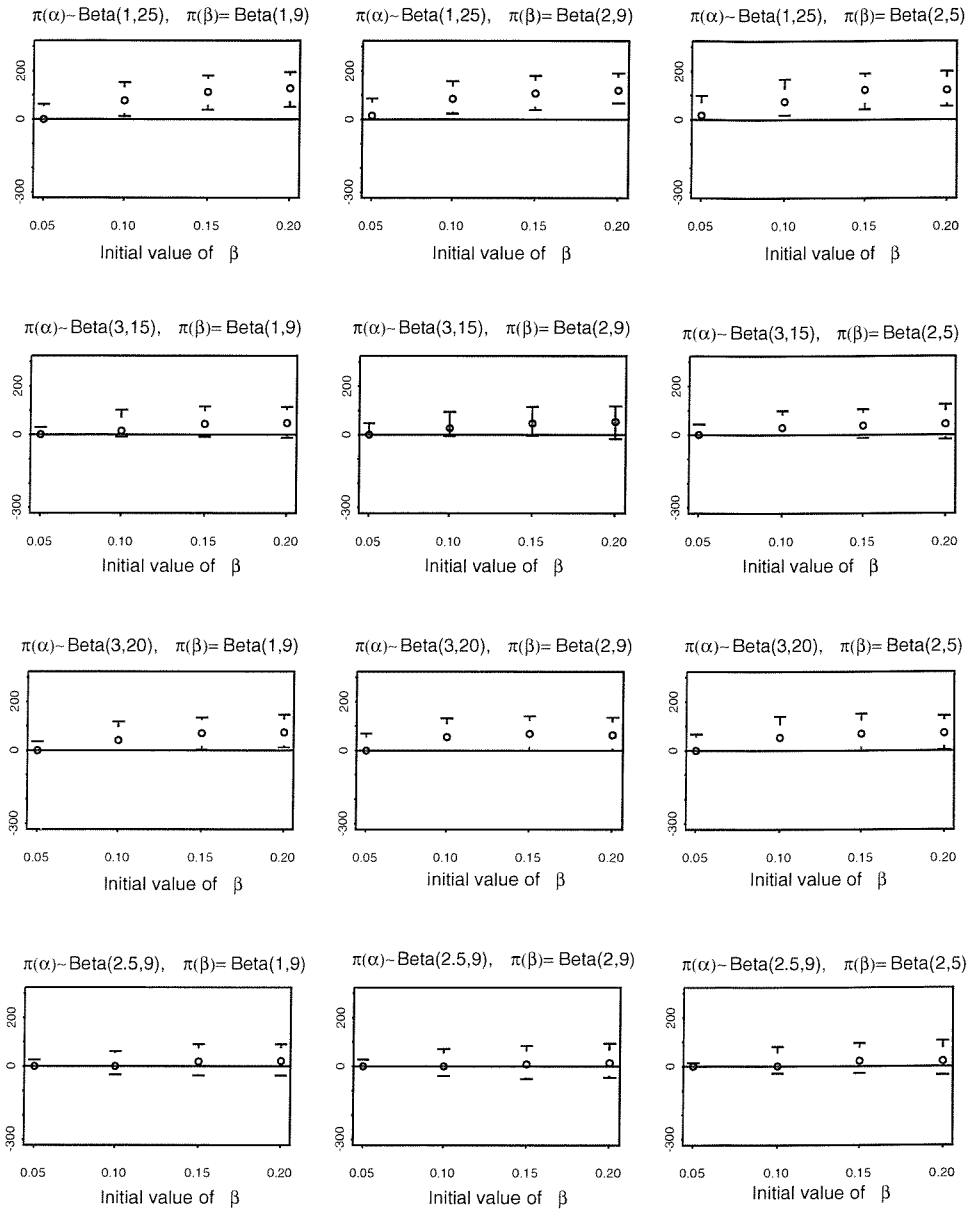
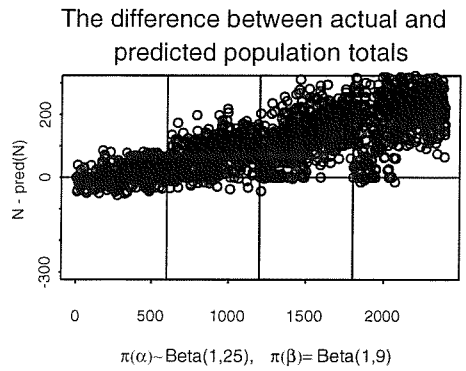
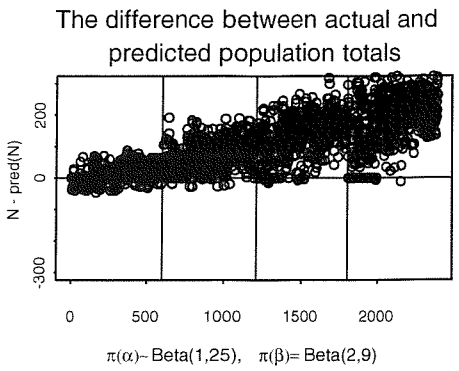
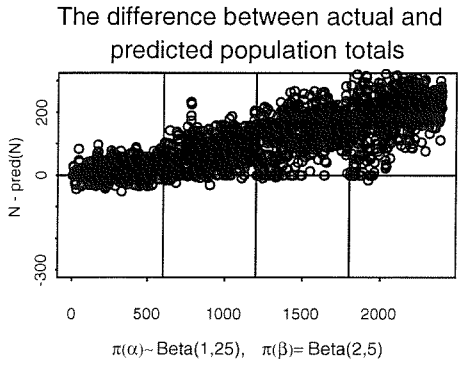


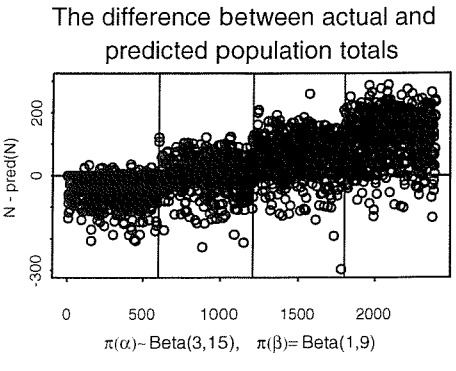
Figure D.8: Plot of the the median values of the difference between M and the predicted value

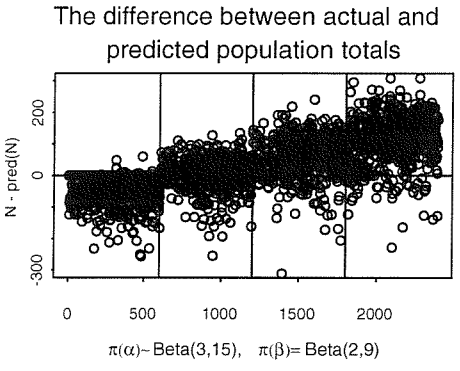
D.0.3 $\pi(\gamma)$ =Gamma$(5,5)$

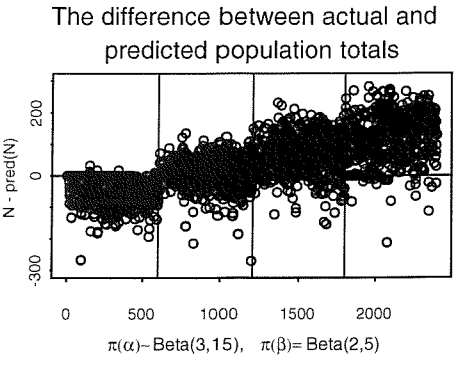












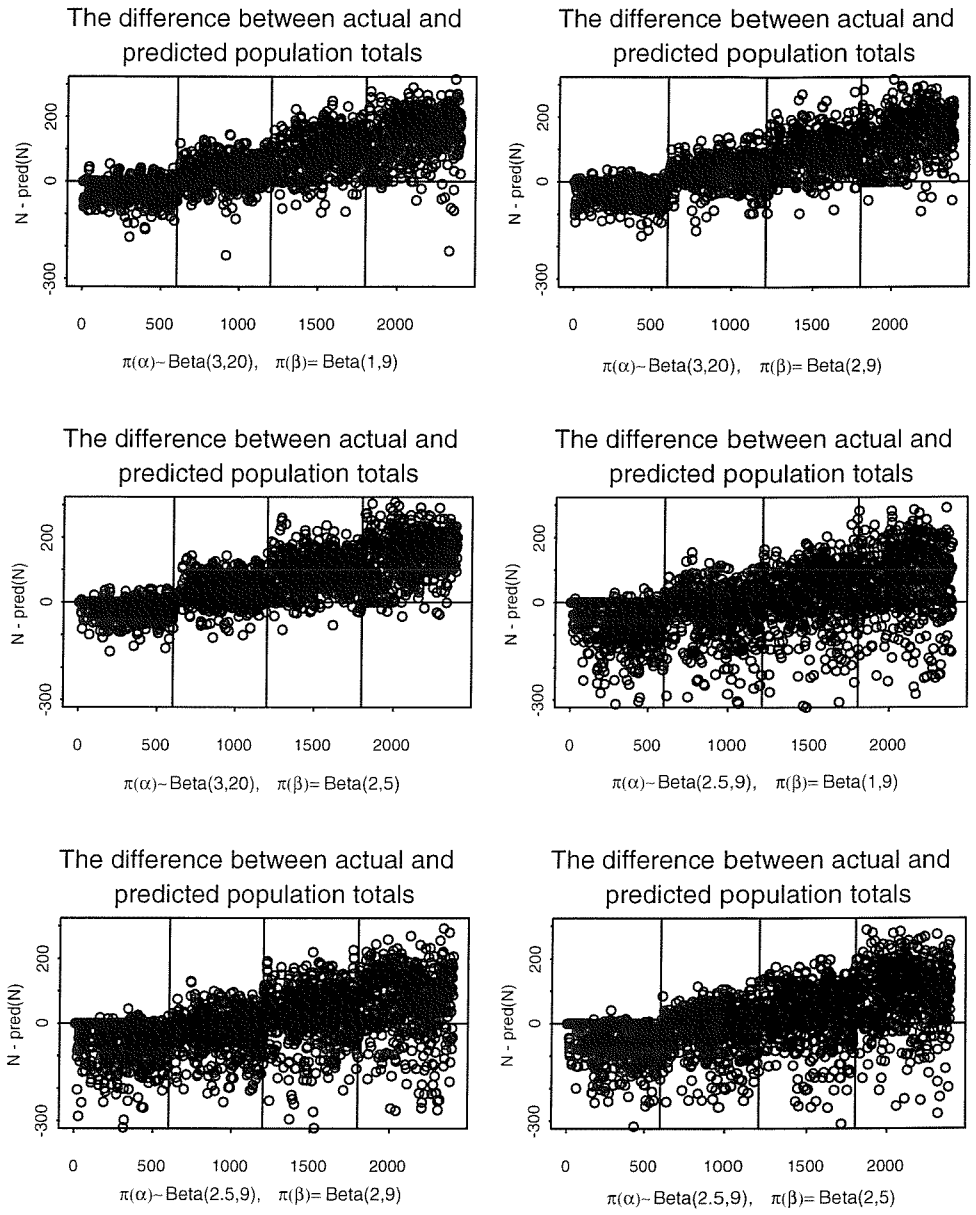
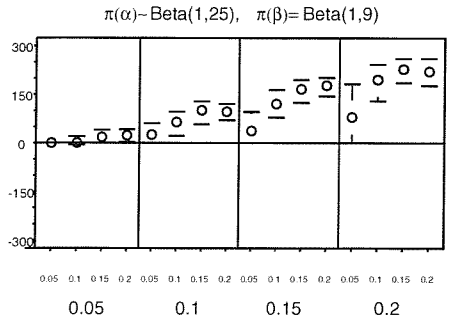
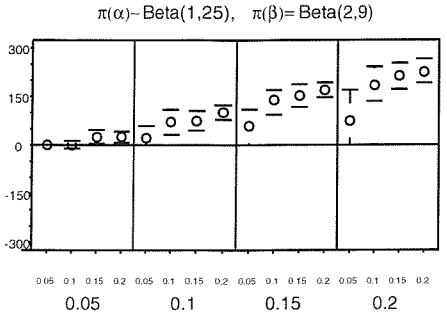
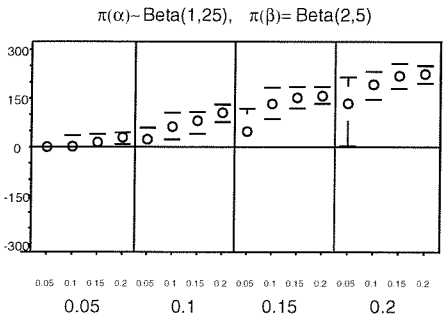
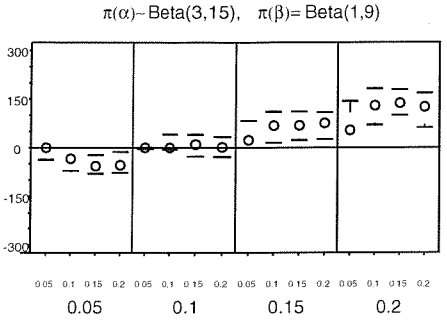


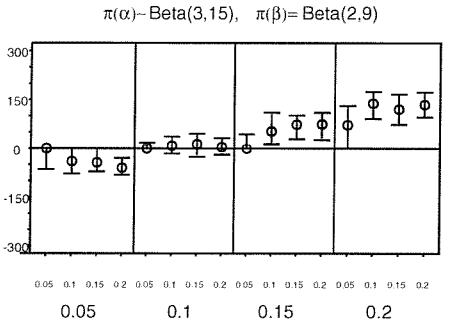
Figure D.9: Plot of the difference between the actual values of M and the generated values of M over the different initial values of α. In the first section of each plot $\alpha = 0.05$, the second $\alpha = 0.1$, the third $\alpha = 0.15$ and the fourth $\alpha = 0.2$. In each section β increases over the same intervals from left to right.

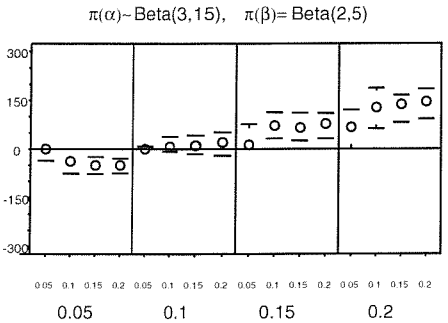

π(α)~Beta(1,25), π(β)= Beta(1,9)
300
150
0
-150
-300
0.05 0.1 0.15 0.2 0.05 0.1 0.15 0.2 0.05 0.1 0.15 0.2 0.05 0.1 0.15 0.2
0.05
0.1
0.15
0.2


π(α)~Beta(1,25), π(β)= Beta(2,9)
300
150
0
-150
-300
0.05 0.1 0.15 0.2 0.05 0.1 0.15 0.2 0.05 0.1 0.15 0.2 0.05 0.1 0.15 0.2
0.05
0.1
0.15
0.2


π(α)~Beta(1,25), π(β)= Beta(2,5)
300
150
0
-150
-300
0.05 0.1 0.15 0.2 0.05 0.1 0.15 0.2 0.05 0.1 0.15 0.2 0.05 0.1 0.15 0.2
0.05
0.1
0.15
0.2


π(α)~Beta(3,15), π(β)= Beta(1,9)
300
150
0
-150
-300
0.05 0.1 0.15 0.2 0.05 0.1 0.15 0.2 0.05 0.1 0.15 0.2 0.05 0.1 0.15 0.2
0.05
0.1
0.15
0.2


π(α)~Beta(3,15), π(β)= Beta(2,9)
300
150
0
-150
-300
0.05 0.1 0.15 0.2 0.05 0.1 0.15 0.2 0.05 0.1 0.15 0.2 0.05 0.1 0.15 0.2
0.05
0.1
0.15
0.2


π(α)~Beta(3,15), π(β)= Beta(2,5)
300
150
0
-150
-300
0.05 0.1 0.15 0.2 0.05 0.1 0.15 0.2 0.05 0.1 0.15 0.2 0.05 0.1 0.15 0.2
0.05
0.1
0.15
0.2

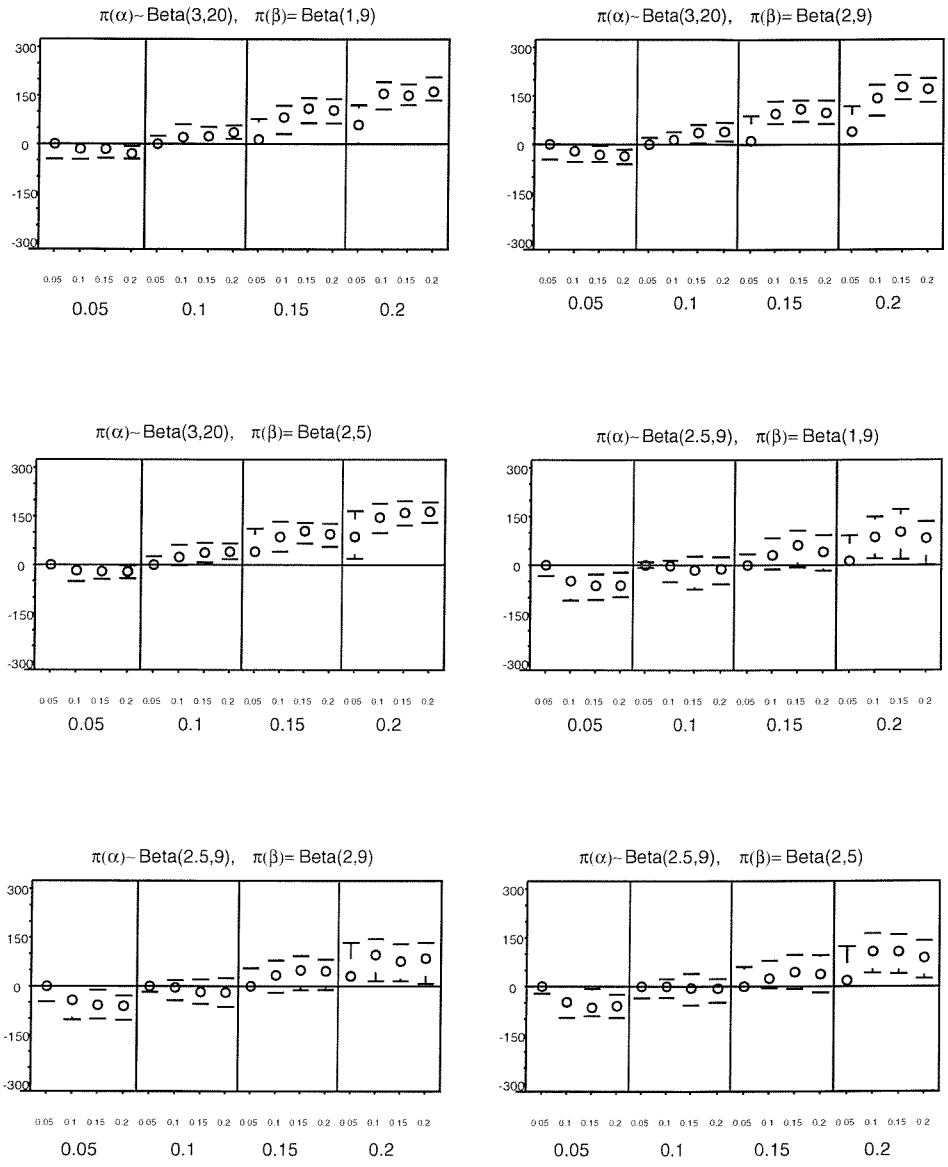


Figure D.10: Plot of the median differences between the actual values of M and the generated values of M.

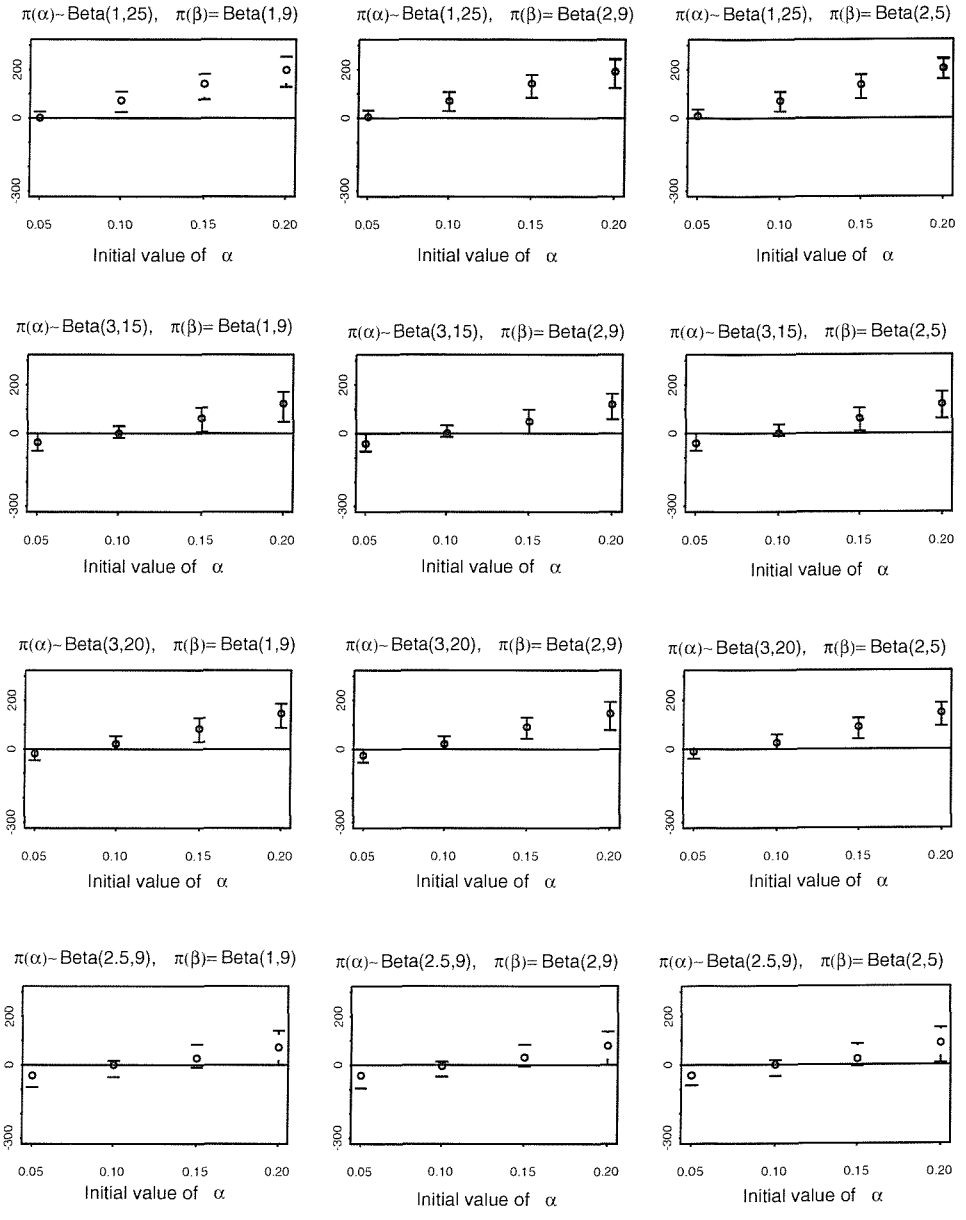


Figure D.11: Plot of the the median values of the difference between M and the predicted value of M over the initial parameter values of α.

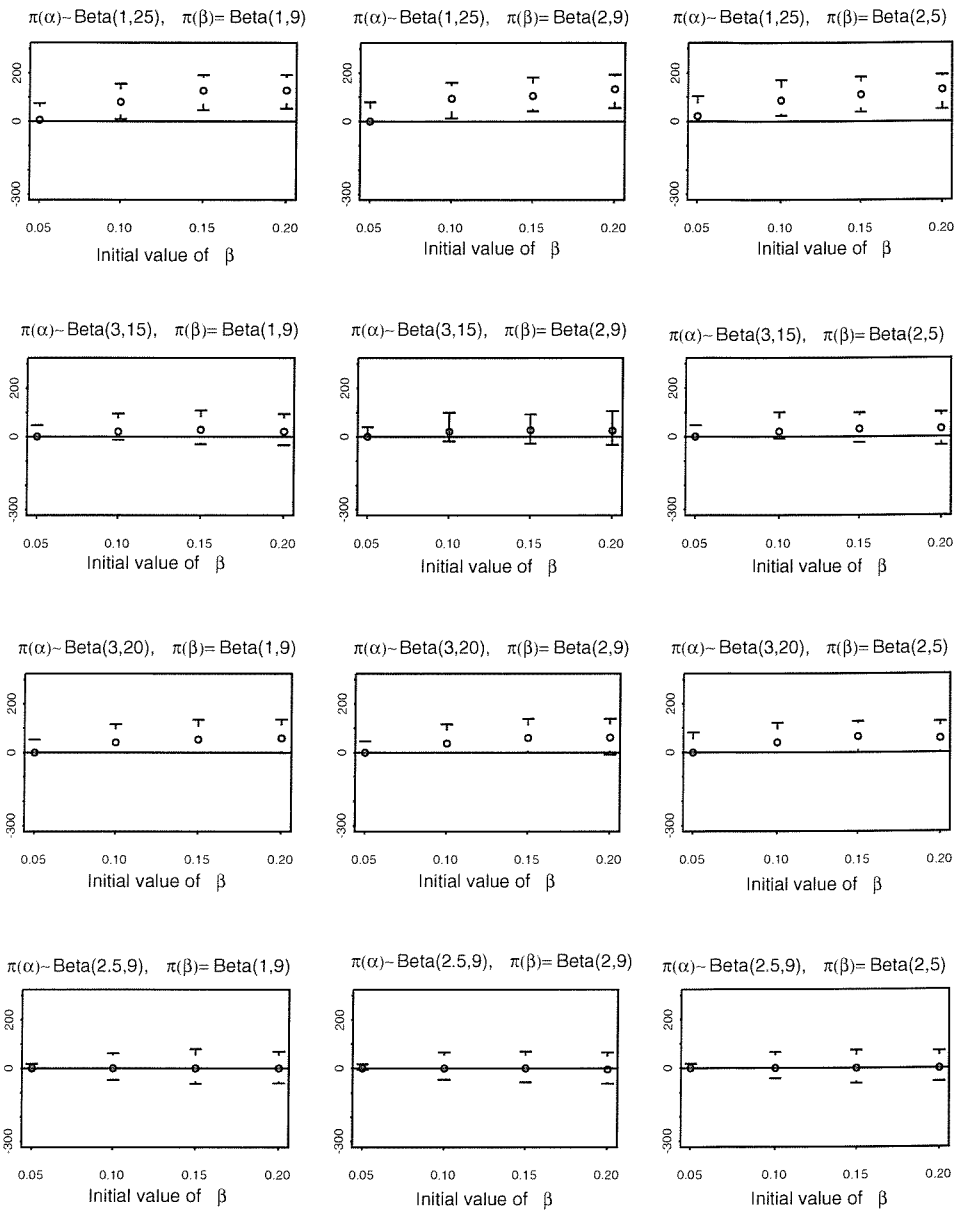


Figure D.12: Plot of the the median values of the difference between M and the predicted value of M over the initial parameter values of β.

Bibliography

Andreatta, G. and Kaufman, G. M. (1986) Estimation of finits population properties when sampling is without replacement and proportional to magnitude. *Journal of the American Statistical Association*, **81**, 657–666.

Baddeley, A. and Turner, R. (2000) Practical maximum pseudolikelihood for spatial point patterns. *Australian Statistical Publishing Association Inc.*, **42**, 283–322.

Bernado, J. M. and Smith, A. F. (1998) *Bayesian Theory.* Wiley InterScience.

Besag, J. and Green, P. J. (1993) Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society: Series B (Methodology)*, **55**, 25–37.

Breckling, J. U., Chambers, R. L., Dorfman, A. H., Tam, S. and H.Welsh, A. (1994) Maximum likelihood inference from sample survey data. *International Statistical Review*, **62**, 349–363.

Brix, A. and Diggle, P. J. (2001) Spatiotemporal prediction for log-Gaussian Cox processes. *Journal of the Royal Statistical Society B.*, **63**, 823–841.

Brooks, S. (1996) Quantitive convergence diagnosis for MCMC vis CUSUMS. *Technical report, University of Bristol.*

Brooks, S. and Roberts, G. (1998) Assessing convergence of Markov Chain Monte Carlo algorithms. *Statistics and Computing*, **8**, 319–335.

Brus, D. and de Gruijter, J. (1997) Random sampling or Geostatistical Modelling? choosing between design-based and model-based sampling strategies for soil (with discussion). *Geoderma*, **80**, 1–44.

Casella, G. and George, E. I. (1992) Explaining the Gibbs sampler. *The American Statistitian*, **46**, 167–174.

Chib, S. and Greenberg, E. (1995) Understanding the Metropolis Hastings algorithm. *The American Statistitian*, **49**, 327–335.

Dempster, A., Laird, N. and Rubin, D. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Seried B (Methodological)*, **39**, 1–38.

Diggle, P. J. (1975) Robust density estimation using distance methods. *Biometrika*, **62**, 39–48.

Evans, M. and Swartz, T. (2000) *Approximating integrals via Monte-Carlo and Deterministic methods*. Oxford University Press: Oxford.

Felix-Medina, M. H. (2000) Analytical expressions for Rao-Blackwell estimators in adaptive cluster sampling. *Journal of Statistical Planning and Inference*, **84**, 221–236.

Fisher, R. A. (1925) Theory of statistical estimation. *Proc. Camb. Phil. Soc.*, **22**, 700–725.

Gelman, A. (1992) Iterative and non-iterative simulation algorithms. *Computing Science and Statistics (Interface Proceedings)*, **24**, 433–438.

Gelman, A. (1996) Inference and monitoring convergence. In *Markov Chain Monte Carlo in Practice* (eds. W. R. Gilks, S. Richardson and D. J. Spiegelhalter), 131–143. Chapman and Hall.

Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, **7**, 457–511.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Geyer, C. J. (1992) Practical Markov chain Monte Carlo. *Statistical Science*, **7**, 473–483.

Gilks, W., Best, N. G. and Tan, K. (1995) Adaptive rejection Metropolis sampling within Gibbs sampling. *Applied Statistics*, **44**, 455–472.

Hansen, M. H., Madow, W. G. and Tepping, B. J. (1983) An evaluation of model-dependant and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, **78**, 776–793.

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

Heilbron, D. (1994) Zero-altered and other models for count data with added zeros. *Biometrical Journal*, **36**, 531–547.

Jeffreys, H. (1961) *Theory of Probability (3rd Edition)*. London: Oxford University Press.

Kass, R. E. and Wasserman, L. (1996) The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, **91**, 1343–1370.

Kaufman, G., Balcer, Y. and Kruyt, D. (1975) A probabilistic model of oil and gas discovery. In *Estimating the Volume of Undiscovered oil and gas resources*, 109–117. The American Association of Petroleum Geologists: Tulsa, Oklahoma.

Liu, J. S., Wong, W. H. and Kong, A. (1994) Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, **81**, 27–40.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953) Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1092.

Mullahy, J. (1986) Specification and testing of some modified count data models. *Journal of Econometrics*, **33**, 341–365.

Nair, V. N. and Wang, P. C. C. (1989) Maximum likelihood estimation under a successive sampling discovery model. *Technometrics*, **31**, 423–436.

Press, S. (1985) *Applied Multivariate analysis: Using Bayesian and Frequentist methods of inference*. Krieger, CA.

Raferty, A. E. and Lewis, S. M. (1992) How many iterations in the Gibbs sampler? In *Bayesian Statistics* (eds. J. M. Bernado, J. Berger, A. Dawid and A. Smith), vol. 4, 763–773. Oxford University Press: oxford.

Royall, R. M. (1976) Current advances in sampling theory: Implications for human observational studies. *American Journal of Epidemiology*, **104**, 463–477.

Sahu, S. K. and Roberts, G. O. (1997) On convergence of the EM algorithm and the Gibbs sampler. *Statistics and Computing*, **59**, 291–317.

Sahu, S. K. and Roberts, G. O. (1999) Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society, Series B*, **9**, 55–64.

Sarndal, C.-E. (1996) Efficient estimators with simple variance in unequal probability sampling. *Journal of the American Statistical Association*, **91**, 1289–1300.

Scott, A. J. (1977) On the problem of randomization in survey sampling. *Sankhya C*, **39**, 1–9.

Seewald, W. (1992) Discussion on parameterisation issues in bayesian inference. In *Bayesian Statistics 4* (eds. J. Bernado, J. Berger, A. Dawid and A. Smith), pp. 241–243. Oxford:Oxford University Press.

Skinner, C. J., Holt, D. and Smith, T. M. F. (1989) *Analysis of Complex Surveys*. Wiley Inter-Science.

Smith, A. F. M. and Roberts, G. O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistics Society, Series B (Methodological)*, **55**, 3–23.

Snowy-Hydro (2005) www.snowyhydro.com.au. *Last Accessed 5/2/05.*

Thompson, S. K. (1990) Adaptive cluster sampling. *Journal of the American Statistical Association*, **85**, 1050–1059.

Thompson, S. K. (1991a) Adaptive cluster sampling: Designs with primary and secondary units. *Biometrics*, **47**, 1103–1115.

Thompson, S. K. (1991b) Stratified adaptive cluster sampling. *Biometrika*, **78**, 389–397.

Thompson, S. K. (1992) *Sampling.* Wiley Inter-Science.

Thompson, S. K. (2002) On sampling and experiments. *Environmetrics*, **13**, 429–436.

Thompson, S. K. and Seber, G. A. F. (1996) *Adaptive Sampling.* Wiley Inter-Science.

Tierney, L. (1994) Markov Chains for exploring posterior distributions. *The Annals of Statistics*, **22**, 1701–1728.

Welsh, A. H., Cunningham, R. B., Donnelly, C. F. and Lindenmayer, D. (1996) Modelling the abundance of rare species - statistical models with extra zeros. *Ecological Modelling*, **88**, 297–308.

West, M. (1994) Discovery sampling and selection models. In *Decision Theory and Related Topics* (eds. J. Berger and S. Gupta), vol. IV, 221–235. Springer Verlag: New York.

West, M. (1996) Inference in seccussive sampling discovery models. *Journal of Econometrics*, **75**, 217–238.

Wild, P. and Gilks, W. R. (1992) Adaptive rejection sampling for gibbs sampling. *Applied Statistics*, **41**, 337–348.

Wild, P. and Gilks, W. R. (1993) Algorithm as 287: Adaptive rejection sampling from log-concave density functions. *Applied Statistics*, **42**, 701–709.

W.R.Gilks, Richardson, S. and Spiegelhalter, D. J. (1996) *Markov Chain Monte Carlo*. Chapman and Hall.

Yu, B. and Mykland, P. (1998) Looking at Markov samplers through cusum path plots: a simple diagnostic idea. *Statistics and Computing*, **8**, 275–286.