UNIVERSITY OF SOUTHAMPTON

Faculty of Engineering, Science and Mathematics

School of Mathematics

Bayesian Sampling Methods in Epidemic and Finite Mixture Models

by

Christine Susan Mary Currie

submitted for the degree of Doctor of Philosophy

November 2004

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS
SCHOOL OF MATHEMATICS

Doctor of Philosophy

BAYESIAN SAMPLING METHODS IN EPIDEMIC AND FINITE MIXTURE MODELS

Christine Susan Mary Currie

This thesis describes the use of sampling methods in two applications: an epidemic model of tuberculosis (TB) and HIV, and the estimation of the number of components in finite normal mixture models. We use Bayesian statistics for the analysis, which enables us to take into account prior information about parameter values in the case of the epidemic modelling, and smooths the likelihood function when considering finite mixture models. The convergence properties of importance sampling are investigated and methods for diagnosing non-convergence of importance sampling are discussed. We use importance sampling to analyse finite normal mixture models and Markov Chain Monte Carlo sampling to fit the epidemic model. Results for effectiveness and cost-effectiveness of different interventions against TB and HIV are presented.

Contents

List of Figures

List of Tables

Acknowledgements

I would first like to thank my supervisor Professor Russell Cheng for the assistance, advice and support he has given freely throughout the course of my PhD study.

I am also grateful to the World Health Organization and the EPSRC who funded me throughout this research.

Many people at the World Health Organization have helped me through the course of my PhD, especially Dr Brian Williams and Dr Chris Dye. Thanks also to Dr Katherine Floyd for her invaluable assistance with health economics.

I would also like to thank Adam Brentnall, Ian Woodward, Ebert Brea, Jenni Sykes, Naomi Powell and Lai Soon Lee with whom I have had many interesting conversations during the times I shared offices with them. Thanks also to Katy Hoad for making me think more broadly about TB, and Emily Webb and Veronica Rapley for their wise words on statistics.

Last but not least, I would very much like to thank my husband David Currie, whose continued support, patience and good humour throughout my time of study has been invaluable to me.

Chapter 1

Introduction

This thesis describes the application of a Bayesian methodology to the analysis of two very different problems: model selection for mixtures of normal distributions and uncertainty analysis of a compartmental disease model. The methodology uses efficient sampling of parameter space to obtain posterior distributions for outputs of interest.

With the ongoing improvements in computing power, it is becoming less time-consuming to integrate numerically over a large number of dimensions. This makes uncertainty analysis possible for statistically non-standard problems and very complex models. The first example given in this thesis is of model selection for finite mixture models, and is statistically non-standard. The second involves the uncertainty analysis of a compartmental model of tuberculosis (TB) and HIV, a complex model with a large number of parameters.

As the number of dimensions of the integration increases, more samples are required to evaluate an integral accurately. Using variance reduction methods can improve the convergence, as discussed in many books on Monte Carlo sampling (e.g. [58], [47]). We consider two methods for variance reduction here: importance sampling and Markov Chain Monte Carlo (MCMC). In both of these techniques, the variance is reduced by concentrating the sample points in areas where the integrand is significant.

The methodology that we use for the analysis of the two examples considered in this thesis is based on Bayesian statistics. With the mixture model application, we have only vague prior information, and the benefit of using Bayesian statistics for this example is that the posterior distribution function is smoother than the likelihood function and has no discontinuities. Using Bayesian statistics to analyse the TB-HIV model allows us to take into account prior information about parameter values in the uncertainty analysis, as well as the fit of the model to available TB incidence and HIV prevalence data.

1.1 Bayesian Statistics

In Bayesian statistics, a parameter θ of a statistical model is regarded as the realised value of a random variable Θ with probability distribution function $\pi(\theta)$. We call $\pi(\theta)$ the prior distribution. Information about the value of θ comes from both the prior distribution and observations of the data D that the statistical model is describing. All of this information can be summarised in the conditional distribution of θ, conditioned on D. We use Bayes' theorem [8], [9] to form this conditional distribution, such that

$$P(\theta|D) = P(D|\theta)P(\theta)/P(D). \tag{1.1}$$

The conditional probability $P(\theta|D)$ is called the posterior distribution, and $P(D|\theta)$ the likelihood. The function $P(D)$ is a normalising factor.

We can evaluate the product of the likelihood and the prior distribution at any point in parameter space Θ and so obtain the shape of the posterior probability distribution, but to obtain a proper probability distribution for the posterior distribution, we must evaluate the constant of proportionality $P(D)$. It is given by

$$P(D) = \int_{\Theta} P(D|\theta)P(\theta)d\theta, \tag{1.2}$$

the product of the likelihood and the prior probability distribution integrated over parameter space. In the examples that we consider in this thesis, the integral cannot be computed analytically, and we use Monte Carlo sampling to evaluate it, in the form of importance sampling or MCMC.

1.2 Sampling Parameter Space

As discussed in the previous section, the purpose of the sampling in this thesis is to evaluate the normalising constant $P(D)$ by integrating the product of the likelihood and the prior probability distribution over parameter space. The variance of the sampling can be reduced if we make use of the available information about the shape of the posterior distribution when devising our sampling methodology. We consider two sampling methods here: importance sampling and MCMC sampling. Both make use of a candidate distribution to focus the sampling in more important areas of parameter space, where importance is measured by the size of the posterior probability. By allowing for the fact that we are sampling from a candidate distribution, these methods enable us to produce a sample that is effectively drawn from $P(\theta|D)$, the posterior distribution.

The sampling convergence is improved if the candidate distribution is similar to the posterior probability distribution. We obtain knowledge about the shape of the posterior distribution prior to sampling by optimizing it to find its mode, and using the inverse Hessian of the negative log of the posterior distribution at the mode to estimate its covariance matrix. This knowledge is then used to define the candidate distribution for the sampling.

1.2.1 Importance Sampling

Importance sampling is a numerical method for evaluating a general integral $\int_\Theta h(\theta)d\theta$ and was first used in the late 1940s and early 1950s. Early discussions of its use are given by, among others, Kahn and Marshall [63]. It was popularised by Hammersley [58] in the 1960s, helped by an expository paper by Clarke [25], which discusses its use within operational research. In importance sampling, samples are drawn from a candidate distribution and weighted by the ratio of the integrand $h(\theta)$ evaluated at the sample point to the value of the candidate distribution at that point. If the candidate distribution is chosen correctly, this results in the sampling being concentrated in parts of parameter space at which the integrand is large, i.e. more important parts of parameter space. In its application to the normalisation of the posterior probability distribution, $h(\theta)$ is the product of the prior and likelihood distributions.

The improvements in convergence of the sampling are dependent on the quality of the candidate distribution. We investigate the choice of candidate distribution in Chapter 2, and find that the best candidate distribution is one that is proportional to the modulus of the integrand, as shown by Marshall [66] in 1954, and more recently by Rubinstein [89] and Evans and Swartz [46]. Using this as the candidate distribution is not practical as it requires knowledge of the integral that we are trying to calculate. We therefore investigate practical solutions to the choice of candidate distribution in the normalisation of the posterior distribution in Section 2.3, giving general results for functions of the exponential family that add detail to the usual rule of thumb that the tails of the candidate distribution should be fatter than those of the integrand.

As the number of dimensions increases, knowledge of the integrand becomes more important. We show in Section 2.4.1 that when the posterior distribution is multivariate normal, and we use a multivariate normal as the candidate distribution, the variance of the sampling increases exponentially with the dimension when there are discrepancies in the mean. Discrepancies in the mean have a greater effect on convergence than discrepancies in the covariance structure.

One of the main advantages of importance sampling is its simplicity. It is easy to implement and easy to understand. In addition, samples output by importance sampling are independent making them easier to work with than those output by adaptive algorithms such as MCMC or adaptive importance sampling. Many authors also comment on the ease of assessing the convergence of importance sampling [46], but few seem to perform a formal analysis of convergence [65]. We discuss methods for analysing the convergence of importance sampling in Chapter 3, and show how extreme value theory can be used to help diagnose a lack of convergence.

1.2.2 Markov Chain Monte Carlo Sampling

Markov Chain Monte Carlo sampling (MCMC) was first used in statistical physics by Metropolis et al in the 1950s [67], who introduced the Metropolis algorithm. This was generalised by Hastings in 1970 [59] to give the Metropolis-Hastings algorithm. An MCMC algorithm for the problem of finding the posterior distribution is designed so that, after a steady state has been reached, the points generated by the algorithm will come from a Markov chain with stationary distribution given by the posterior distribution. In the Metropolis-Hastings algorithm, a point Y_i are generated from a candidate distribution $q(Y_i, X_i)$, which may depend on X_i, the current position of the algorithm. The algorithm will move to Y_i with probability α, where

$$\alpha = \min\left\{1, \frac{f(Y_i)q(Y_i, X_i)}{f(X_i)q(X_i, Y_i)}\right\}, \tag{1.3}$$

where $f(.)$ is the product of the prior and likelihood distributions.

Most MCMC algorithms are adaptive, in that the parameters of the candidate distribution depend on the current position of the algorithm. For example, in determining the posterior distribution of the parameters in the TB-HIV model, we use a random-walk Metropolis algorithm with a multivariate t-distribution as the candidate. We set the mean of the t-distribution to be equal to the last accepted point, but use the same covariance structure throughout the sampling. This adaptability reduces the importance of knowledge of the posterior distribution prior to the sampling as information obtained during the sampling is used to improve the convergence. Adaptive algorithms do present problems when analysing output however, because the individual observations are not independent. A further difficulty with MCMC algorithms is the difficulty in assessing convergence, although many methods have been devised to do this [51].

1.3 Finite Mixture Models

Mixture models are used where a statistical dataset is not homogeneous but is composed of a number of distinct component distributions. An example is the galaxy dataset introduced by Roeder [87], where it is believed that there are a number of different groups of galaxies present. The number of components then relates directly to the number of galaxy groups. A further use is in semiparametric density estimation, such as modelling input data for simulation models [21]. A number of datasets that arise in the mixture models context are examined in Chapter 4.

We shall discuss only continuous finite normal mixture models, where the probability density function can be written as

$$f(x) = \sum_{i=1}^{k} w_i g_i(x|\theta_i), \tag{1.4}$$

where $g_i(.)$ is a normal distribution and w_i are weights such that $\sum_{i=1}^{k} w_i = 1$ and $w_i \geq 0$.

We wish to determine the number of components k in a finite mixture model for different datasets. The problem is statistically non-standard as it is possible for components to be present in the mixture that are not represented within the data.

We focus on examples for which there is no prior information available, but use a Bayesian framework because of the inherent problems with maximum likelihood methods. The likelihood surface is often multimodal and has discontinuities near the boundaries, e.g. when the component variances tend to zero. This can occur, for example, when a component is centred on just one data point and tends to a delta function at that point. In addition, the likelihood increases as more components are added, even if these components contribute very little to the model, making determination of the optimal number of components difficult. Using Bayesian statistics, the prior distribution smooths out the discontinuities in the likelihood function, though the posterior distribution can still be multimodal. The posterior distribution for the number of components k also tends to have a peak at $k < n$, where n is the number of data points. The posterior distribution for the number of components is thus usually a more meaningful measure of the number of components in the mixture than the likelihood function.

We use importance sampling to determine the posterior distribution for the number of components, contrasting with much of the established literature in which MCMC methods dominate [84], [97]. MCMC algorithms for this problem tend to be complicated as they must include some mechanism for jumping between different models (different values of k). By contrast, the application of importance sampling is relatively simple, with the candidate distribution including a function describing the probability of selecting a model with a particular k, and a function for sampling parameter values that is dependent on the model with the chosen k.

1.4 Model of Tuberculosis and HIV

The second part of the thesis discusses an application of the Bayesian analysis methodology in epidemiological modelling. We use a compartmental model to describe disease progression through tuberculosis (TB), and the effects of HIV on that progression. In this example, we have good prior information about model parameters, which comes from medical studies within the literature (see the supplementary material of [36] for full details). The extent of the prior knowledge makes a Bayesian methodology particularly attractive. With 23 parameters, the model is relatively complex and determining the posterior distribution is time-consuming. We use a Metropolis MCMC algorithm to determine the posterior distribution of the model parameters, using the output of this algorithm to estimate the uncertainty on our estimates of the costs, effects and cost-effectiveness of the different intervention strategies, as well as our predictions of TB incidence and deaths. We chose to use MCMC sampling for this example because importance sampling performed relatively poorly due to the shape of the posterior distribution, which is skewed and so not very similar to a normal distribution.

The modelling study focuses on high burden countries in Sub-Saharan Africa, where HIV prevalence is greater than 10% and there has been a marked increase in TB incidence as a result of the HIV epidemic. We use the model to predict TB incidence in the future and the effects that different interventions against the two diseases will have on the future course of the TB epidemic. Further work, discussed in Chapter 6, evaluates the cost-effectiveness of different intervention strategies, measured in terms of the costs per disability adjusted life year (DALY) averted.

1.5 Outline of the Thesis

The optimal choice of sampling function to use in importance sampling is discussed in Chapter 2, including an evaluation of the convergence of importance sampling in many dimensions. We then present diagnostic and statistical methods for assessing the convergence in Chapter 3. In Chapter 4 we apply the methodology to the statistically non-standard problem of determining the number of components in a finite normal mixture model. The TB-HIV model and the methodology that we used for the uncertainty analysis is described in Chapter 5. Predictions for the TB incidence in Kenya, Uganda and South Africa, and estimates of the effects that different interventions will have on reducing TB incidence and TB deaths in these countries are shown here. Further work analysed the costs, effects and cost-effectiveness of different intervention strategies against TB and HIV in terms of costs per disability adjusted life year (DALY) averted, and this is discussed in Chapter 6. We conclude in Chapter 7.

Chapter 2

Choosing the Candidate Function in Importance Sampling

Importance sampling is a numerical method of evaluating an integral

$$I = \int h(\theta) d\theta, \qquad \theta \in D \subset R^n. \tag{2.1}$$

In standard Monte Carlo sampling, I would be evaluated by taking K samples distributed uniformly over the region D, giving $\hat{I} = \sum_{i=1}^{K} h(\theta_i)/K$. In importance sampling, we concentrate the sample points in areas of "importance" within D by sampling from a candidate distribution $w(\theta, \beta)$. We then allow for the fact that we are sampling from $w(\theta, \beta)$, rather than a uniform distribution, by weighting each of the observations of $h(\theta)$ and so finding the expectation of $h(\theta)/w(\theta, \beta)$. The integral I can therefore be approximated by

$$\hat{I}_w = \frac{1}{K} \sum_{i=1}^{K} \frac{h(\theta_i)}{w(\theta_i, \beta)}. \tag{2.2}$$

The choice of $w(\theta, \beta)$ affects the convergence rate of $\hat{I}_w$ to I and the optimal choice for $w(\theta, \beta)$ is $\frac{|h(\theta)|}{\int h(\theta) d\theta}$, as we show for the particular example of statistical estimation in Section 2.2. This requires full knowledge of the integral that we are trying to evaluate and so it is not a practical solution to the problem.

This chapter focuses on the use of importance sampling in statistical estimation, and in Section 2.3, we consider the problem of normalising univariate functions of the general exponential family using importance sampling. We present conditions on the parameters of the candidate distribution for sampling to converge and demonstrate these results by considering three examples of common probability distributions.

The convergence of importance sampling is dependent on the number of dimensions of the integration, and we discuss the dependence in Section 2.4. We begin by considering how the number of dimensions and discrepancies between the candidate distribution $w(\theta, \beta)$ and the integrand affect the convergence of the sampling, where the integrand and the candidate distribution are both assumed to be multivariate normal. We then go on to present four numerical examples of importance sampling in two dimensions, which demonstrate some of the limitations of importance sampling.

2.1 Importance Sampling for Statistical Estimation

Often in problems of statistical estimation, we need to evaluate the expectation of a statistic $m(\theta)$. In this case, I can be written as

$$I(m) = \int_{\Theta} m(\theta) f(\theta) d\theta, \tag{2.3}$$

where $f(\theta)$ is a probability density function. A typical example is where $m(\theta) = \theta$. This kind of integral is very common in statistical calculations but frequently cannot be calculated analytically. The integral can be estimated by sampling from $f(\theta)$ and calculating $m(\theta)$ at each sample point θ_i, such that $\hat{I}(m) = \sum_{i=1}^{K} m(\theta_i)/K$. Convergence can be improved by using importance sampling and sampling from a candidate distribution $w(\theta, \beta)$ instead of $f(\theta)$. As in Equation 2.2, we must weight each of the observations by the probability of having selected that sample point, and so the integral $I(m)$ is approximated by

$$\hat{I}_w(m) = \frac{1}{K}\sum_{i=1}^{K}\frac{m(\theta_i)f(\theta_i)}{w(\theta_i,\beta)}. \tag{2.4}$$

Another integral of interest in statistics is the normalisation of the posterior probability distribution in Bayesian statistics. We say that the posterior probability distribution of the parameters θ, given available data x is $p(\theta|x)$, which is proportional to the product of the likelihood $L(x|\theta)$ and the prior probability distribution for θ, $\pi(\theta)$. To obtain a proper probability distribution for the posterior distribution, it is necessary to find the normalising factor by integrating $L(x|\theta)\pi(\theta)$ over parameter space. Setting $m(\theta) = L(x|\theta)$ and $f(\theta) = \pi(\theta)$ we can see that this integral has the same general form as $I(m)$ and the results of subsequent sections will therefore apply to this problem.

2.2 Theoretical Results

Theorem

The variance of importance sampling is minimized if

$$w(\theta,\beta) = \frac{|m(\theta)|f(\theta)}{\int_\Theta f(\theta)|m(\theta)|d\theta} \tag{2.5}$$

Proof

That this is the optimal form for $w(\theta,\beta)$ has been shown by Evans and Swartz [46] and Rubinstein [89]. Evans and Swartz use the law of the iterated logarithm attributable to Durrett, and Rubinstein uses the Cauchy Schwarz inequality. These proofs are more complicated, and will not be discussed here. Instead we present a simpler more direct derivation using the calculus of variations.

We assume that our candidate distribution, $w(\theta, \beta)$ is a true probability distribution function such that

1. $\int_\Theta w(\theta, \beta) d\theta = 1$
2. $w(\theta, \beta) \geq 0$

The variance of the sampling can be written as

$$Var_w \left[\frac{m(\theta) f(\theta)}{w(\theta, \beta)} \right] = \int_\Theta \left(\frac{m(\theta) f(\theta)}{w(\theta, \beta)} \right)^2 w(\theta, \beta) d\theta - \left[\int_\Theta m(\theta) f(\theta) d\theta \right]^2 . \quad (2.6)$$

The second term in Equation 2.6 is independent of $w(\theta, \beta)$, therefore can be ignored when choosing the best form for $w(\theta, \beta)$. Therefore, we are left with the problem of minimising $\int_\Theta \left(\frac{m(\theta) f(\theta)}{w(\theta, \beta)} \right)^2 w(\theta, \beta) d\theta$ subject to the conditions given above. Using Lagrangian multipliers to take account of the first constraint on $w(\theta, \beta)$, the objective function becomes

$$\int_\Theta \left(\frac{m(\theta) f(\theta)}{w(\theta, \beta)} \right)^2 w(\theta, \beta) d\theta + \lambda \left(\int_\Theta w(\theta, \beta) d\theta - 1 \right) , \quad (2.7)$$

where λ is a Lagrangian multiplier. Using Euler's equation, the optimal form for $w(\theta, \beta)$ must obey

$$-\frac{m^2(\theta) f^2(\theta)}{w^2(\theta, \beta)} + \lambda = 0, \quad (2.8)$$

therefore

$$w^2(\theta, \beta) = \frac{m^2(\theta) f^2(\theta)}{\lambda}. \quad (2.9)$$

To find λ, we substitute the expression for $w(\theta)$ into the normalisation constraint. Assuming that $f(\theta)$ is a proper probability distribution function such that $f(\theta) \geq 0$, and remembering that $w(\theta, \beta) \geq 0$, the normalisation constraint becomes

$$\int_\Theta \frac{|m(\theta)| f(\theta)}{\sqrt{\lambda}} d\theta = 1. \quad (2.10)$$

Solving for λ and substituting this back into Equation 2.9, we find that

$$w(\theta, \beta) = \frac{|m(\theta)| f(\theta)}{\int_\Theta |m(\theta)| f(\theta) d\theta}, \quad (2.11)$$

and so the theorem is proved.

Geweke [53] uses a similar method to Evans and Swartz [46] to find the optimum sampling density but obtains a different expression,

$$w(\theta, \beta) \propto |m(\theta) - \overline{m}| f(\theta) \tag{2.12}$$

In arriving at this expression Geweke uses a central limit theorem to make the assumption that in the limit of a large number of samples $n^{1/2}(\overline{m}_n - \overline{m})$ is described by a normal distribution with zero mean and variance

$$\begin{aligned} \sigma^2 &= E\{[m(\theta) - \overline{m}]^2 f(\theta)/w(\theta, \beta)\} \\ &= \int_\Theta [m(\theta) - \overline{m}]^2 f(\theta)/w(\theta, \beta).f(\theta) d\theta, \end{aligned} \tag{2.13}$$

where the expectation is taken over the distribution $f(\theta)$. Minimising σ^2 is equivalent to maximising the rate of convergence, and a minimum is achieved when $w(\theta, \beta)$ is given by the expression in Equation 2.12.

The expression Geweke uses for σ^2 is derived making the assumption that we are sampling from $f(\theta)$ and calculating $m(\theta)f(\theta)/w(\theta, \beta)$ at every sample point, whereas we are actually sampling from $w(\theta, \beta)$. Therefore, σ^2 should be given by

$$\begin{aligned} \sigma^2 &= E_w\{[m(\theta)f(\theta)/w(\theta, \beta) - \overline{mf/w}]^2\} \\ &= \int_\Theta [m(\theta)f(\theta)/w(\theta, \beta) - \overline{mf/w}]^2 w(\theta, \beta) d\theta \\ &= \int_\Theta m^2(\theta) f^2(\theta)/w(\theta, \beta) d\theta - \left[\int_\Theta m(\theta) f(\theta) d\theta\right]^2, \end{aligned} \tag{2.14}$$

where the subscript w implies that we are calculating the expectation assuming sampling from $w(\theta, \beta)$. The correct expression for σ^2, given in Equation 2.14, is minimised for $w(\theta, \beta) = \frac{|m(\theta)| f(\theta)}{\int |m(\theta)| f(\theta) d\theta}$. This is an identical result to that obtained above.

This result makes more sense intuitively than the result presented by Geweke as it suggests that more points should be sampled in regions close to the mean, where $m(\theta)f(\theta)$ is highest. The result presented by Geweke suggests that more points should be sampled at points distant from the mean, where $|m(\theta) - \overline{m}|$ is large. Geweke's result would minimise the sampling variance of the expectation over $f(\theta)$ of $m(\theta) - \overline{m}$.

2.3 Choosing the Candidate Distribution in Practice: Results for Functions in the Exponential Family

Although the results of Section 2.2 show that the optimal choice of candidate distribution is $w(\theta, \beta) = |m(\theta)|f(\theta)/\int |m(\theta)|f(\theta)d\theta$, this is not a practical solution as it requires knowledge of the integral that we are trying to evaluate. In practice we must choose a candidate distribution that is simple to sample from, with parameters that can be estimated without excessive preliminary investigations of $m(\theta)f(\theta)$. Often $w(\theta, \beta)$ has a parametric form that can be adjusted to change its shape and the choice of functional for $w(\theta, \beta)$ must take into account the ease of adjusting the parameters β to obtain a good fit to $|m(\theta)|f(\theta)$. Although computing time may be saved by using a candidate distribution that is very close to the function being estimated, if estimating or sampling from this distribution requires a large amount of computing time, any gains in efficiency due to good convergence will be lost.

In this section, we give the form of the candidate distribution that should be used to ensure convergence for a general function from the exponential family. We then go on to find expressions for the variance associated with the sample for some standard probability distributions. We find analytical expressions for the variance associated with sampling the normal, gamma and student-t distributions in Sections 2.3.1, 2.3.2 and 2.3.3 respectively.

A general discussion of the choice of sampling candidate distribution is given by Robert and Cassella [85], while Geweke [53] gives a set of conditions that a sampling distribution must obey to ensure convergence. Although his estimate of the variance of the sampling is different from ours, his conditions still hold. Using the notation of Equation 2.3, they are equivalent to

1. $f(\theta)m(\theta)/w(\theta,\beta) < c < \infty$

2. θ is compact and $f(\theta)m(\theta) < k < \infty$

where k and c are arbitrary constants greater than zero. The first condition ensures that the ratios calculated during the sampling are always finite and the second that the integral being evaluated is always finite.

We reduce the problem of Equation 2.3 to one of obtaining a sample from $f(\theta)$, by sampling from the importance sampling candidate distribution $w(\theta,\beta)$. This allows us to draw conclusions about the efficiency of importance sampling for the generation of samples from the posterior distribution, where $f(\theta)$ is now considered to be the posterior distribution. The variance associated with the sample, $Var_w[f(\theta)/w(\theta,\beta)]$, is then given by

$$\begin{aligned} Var_w[f(\theta)/w(\theta,\beta)] &= E_w[(f(\theta)/w(\theta,\beta))^2] - (E_w[f(\theta)/w(\theta,\beta)])^2 \\ &= \int f^2(\theta)/w(\theta,\beta)dx - \left[\int f(\theta)d\theta\right]^2. \qquad (2.15) \end{aligned}$$

The second term in the expression is unaffected by the choice of candidate distribution. Therefore, assuming that $\int f(\theta)d\theta$ is not divergent, to obtain a finite variance, $w(\theta,\beta)$ must be chosen such that $\int f^2(\theta)/w(\theta,\beta)d\theta = E_f[f(\theta)/w(\theta,\beta)]$ is finite. However, as we calculate $f(\theta)/w(\theta,\beta)$ at each step of the sampling, we must also impose the condition that $f(\theta)/w(\theta,\beta)$ is finite.

Let the function that we are trying to sample be $f(\theta;\alpha_1,\ldots,\alpha_m)$, where $f(\theta;\alpha_1,\ldots,\alpha_m)$ is from the exponential family and of the form

$$f(\theta;\alpha) = \exp\left(\sum_{j=1}^{m} p_j(\alpha_1,\ldots,\alpha_m)k_j(\theta) + s(\theta) + q(\alpha_1,\ldots,\alpha_m)\right). \qquad (2.16)$$

We assume the candidate distribution $w(\theta;\beta_1,\ldots,\beta_l)$ to also be from the exponential family such that

$$w(\theta;\beta) = \exp\left(\sum_{i=1}^{l}\pi_i(\beta_1,\ldots,\beta_l)\gamma_i(\theta) + \sigma(\theta) + \phi(\beta_1,\ldots,\beta_l)\right). \qquad (2.17)$$

In this context, the conditions for the importance sampling to converge are

1. $f(\theta)/w(\theta,\beta)$ is finite for all θ

2. $\int_{-\infty}^{\infty} f^2(\theta)/w(\theta,\beta)d\theta$ is finite

The first condition will hold if $w(\theta;\beta) > 0, \forall\theta$; $f(\theta)$ is not divergent over the range $[-\infty,+\infty]$; and

$$\frac{d}{d\theta}(f(\theta)/w(\theta))\begin{cases}\leq 0 & \theta\rightarrow\infty\\ \geq 0 & \theta\rightarrow-\infty\end{cases}. \qquad (2.18)$$

This last condition implies that the expression

$$\sum_{j=1}^{m} p_j(\alpha_1,\ldots,\alpha_m)\frac{dk_j(\theta)}{d\theta} + \frac{ds(\theta)}{d\theta} - \sum_{i=1}^{l}\pi_i(\beta_1,\ldots,\beta_l)\frac{d\gamma_i(\theta)}{d\theta} - \frac{d\sigma(\theta)}{d\theta} \qquad (2.19)$$

must be less than or equal to zero as $\theta\rightarrow\infty$ and greater than or equal to zero as θ tends to $-\infty$.

Using these results, we can determine whether importance sampling will converge when sampling a probability density function $f(\theta)$ with a given candidate distribution $w(\theta,\beta)$. We consider the sampling of three specific probability distribution functions from the general exponential family: normal, gamma and student t distributions.

2.3.1 Sampling a Normal with a Normal

In this section, we use the conditions of Section 2.3 to determine the limits on the parameters of the candidate distribution $w(\theta,\beta)$, when sampling a normal distribution of mean α_1, variance α_2^2, under the assumption that the candidate distribution is also normal, with mean β_1 and variance β_2^2.

Using the notation of the general exponential family,

$$\begin{aligned}
q(\alpha_1,\alpha_2) &= -\frac{1}{2}\left[\ln(2\pi\alpha_2^2)+\frac{\alpha_1^2}{\alpha_2^2}\right] \\
p_1(\alpha_1,\alpha_2) &= \frac{\alpha_1}{\alpha_2^2} \\
p_2(\alpha_1,\alpha_2) &= -\frac{1}{2\alpha_2^2} \\
k_1(\theta) &= \theta \\
k_2(\theta) &= \theta^2 \\
\phi(\beta_1,\beta_2) &= -\frac{1}{2}\left[\ln(2\pi\beta_2^2)+\frac{\beta_1^2}{\beta_2^2}\right] \\
\pi_1(\beta_1,\beta_2) &= \frac{\beta_1}{\beta_2^2} \\
\pi_2(\beta_1,\beta_2) &= -\frac{1}{2\beta_2^2} \\
\gamma_1(\theta) &= \theta \\
\gamma_2(\theta) &= \theta^2. \qquad (2.20)
\end{aligned}$$

Ignoring the constant terms in $f^2(\theta)/w(\theta)$, we can write this as

$$f^2(\theta)/w(\theta) \propto \exp\left[\left(\frac{2\alpha_1}{\alpha_2^2}-\frac{\beta_1}{\beta_2^2}\right)\theta-\left(\frac{1}{\alpha_2^2}-\frac{1}{2\beta_2^2}\right)\theta^2\right]. \qquad (2.21)$$

In the limit that $\theta \to \pm\infty$, the θ^2 term will dominate and so for the variance to be finite,

$$\beta_2^2 > \alpha_2^2/2. \qquad (2.22)$$

Writing out the expression for the variance of the sampling with $f(\theta,\alpha)$ and $w(\theta,\beta)$ defined as in Equation 2.20, we can show how the choice of parameters for a normal importance sampling distribution affects the convergence to the actual function. The integrand, $f^2(\theta;\alpha)/w(\theta;\beta)$ is given by

$$\begin{aligned}\frac{f^2(\theta;\alpha)}{w(\theta;\beta)} &= \frac{1}{\alpha_2^2}\sqrt{\frac{\beta_2^2}{2\pi}}\exp\left[-\theta^2\left(\frac{1}{\alpha_2^2}-\frac{1}{2\beta_2^2}\right)+2\theta\left(\frac{\alpha_1}{\alpha_2^2}-\frac{\beta_1}{2\beta_2^2}\right)-\frac{\alpha_1^2}{\alpha_2^2}+\frac{\beta_1^2}{2\beta_2^2}\right]\\ &= \epsilon\exp\left[-\left(\frac{1}{\alpha_2^2}-\frac{1}{2\beta_2^2}\right)(\theta-\phi)^2\right], \qquad (2.23)\end{aligned}$$

where ϕ and ϵ are constants,

$$\begin{aligned}\phi &= \frac{\frac{\alpha_1}{\alpha_2^2}-\frac{\beta_1}{2\beta_2^2}}{\frac{1}{\alpha_2^2}-\frac{1}{2\beta_2^2}}\\ \epsilon &= \frac{1}{\alpha_2^2}\sqrt{\frac{\beta_2^2}{2\pi}}\exp\left[-\frac{\alpha_1^2}{\alpha_2^2}+\frac{\beta_1^2}{2\beta_2^2}+\left(\frac{1}{\alpha_2^2}-\frac{1}{2\beta_2^2}\right)\phi^2\right]. \qquad (2.24)\end{aligned}$$

We can integrate $f^2(\theta;\alpha)/w(\theta;\beta)$ to obtain the expected value,

$$\begin{aligned}E_w[f^2(\theta;\alpha)/w^2(\theta;\beta)] &= \epsilon\int_{-\infty}^{\infty}\exp\left[-\frac{(\theta-\phi)^2}{(1/\alpha_2^2-1/2\beta_2^2)^{-1}}\right]d\theta\\ &= \epsilon\sqrt{\frac{\pi}{(1/\alpha_2^2-1/2\beta_2^2)}} \qquad (2.25)\end{aligned}$$

and writing out the expression for ϵ in full,

$$E_w[f^2(\theta;\alpha)/w^2(\theta;\beta)] =$$

$$\frac{1}{\alpha_2^2}\sqrt{\frac{\beta_2^2}{2\left(\frac{1}{\alpha_2^2}-\frac{1}{2\beta_2^2}\right)}}\exp\left[-\frac{\alpha_1^2}{\alpha_2^2}+\frac{\beta_1^2}{2\beta_2^2}+\left(\frac{1}{\alpha_2^2}-\frac{1}{2\beta_2^2}\right)\left(\frac{\frac{\alpha_1}{\alpha_2^2}-\frac{\beta_1}{2\beta_2^2}}{\frac{1}{\alpha_2^2}-\frac{1}{2\beta_2^2}}\right)^2\right]. \qquad (2.26)$$

Using Equation 2.26, we find expressions for the β_2^2 and β_1 that minimise and maximise $E_w[f^2(\theta;\alpha)/w^2(\theta;\beta)]$, by differentiating with respect to β_2^2 and β_1. Differentiating with respect to β_1 initially, we find that there is an extreme value at $\tilde{\beta}_1$, where

$$\tilde{\beta}_1 - \frac{\alpha_1/\alpha_2^2-\tilde{\beta}_1/2\beta_2^2}{1/\alpha_2^2-1/2\beta_2^2} = 0, \qquad (2.27)$$

therefore,

$$\tilde{\beta}_1 = \alpha_1, \qquad (2.28)$$

i.e. the means of the sampling distribution $w(\theta;\alpha)$ and the distribution being sampled from $f(\theta;\alpha)$ are equal.

Using the same principle, we can find $\tilde{\beta}_2^2$, the value of β_2^2 at which $E_w[f^2(\theta;\alpha)/w^2(\theta;\beta)]$ takes on an extreme value. After much algebra, it is found that $\tilde{\beta}_2^2$ obeys the expression

$$0 = \left(\tilde{\beta}_2^2 - \alpha_2^2\right)\left(\tilde{\beta}_2^2 - \alpha_2^2/2\right) - (\beta_1 - \alpha_1)^2 \tag{2.29}$$

and solving the quadratic equation,

$$\tilde{\beta}_2^2 = 3\alpha_2^2/4 \pm \frac{\alpha_2^2}{2}\sqrt{1/4 - 4(\beta_1 - \alpha_1)^2/\alpha_2^2} \tag{2.30}$$

with the smaller solution a maximum and the larger solution a minimum.

If we now set $\beta_1 = \alpha_1$,

$$\begin{aligned} \tilde{\beta}_2^2 &= 3\alpha_2^2/4 \pm \alpha_2^2/4 && (2.31)\\ &= \begin{cases} \alpha_2^2 \\ \alpha_2^2/2, \end{cases} && (2.32) \end{aligned}$$

and we can see that the variance of the sampling has a minimum where the variance of the candidate distribution equals the variance of $f(\theta)$ and a maximum where its variance is equal to half this value.

We now investigate the behaviour of the optimal value of β_2^2 as the discrepancy $(\beta_1 - \alpha_1) = \delta$ is increased. We assume δ is small relative to the variance α_2^2, i.e. that we have a good estimate of the mean of the normal distribution that we are trying to find,

$$\tilde{\beta}_2^2 \simeq \begin{cases} \alpha_2^2 + 8\delta^2/2\alpha_2^2 \\ \alpha_2^2/2 - 8\delta^2/2\alpha_2^2 \end{cases} \tag{2.33}$$

Therefore, as we move further from the mean, the value of β_2^2 that minimises $E_w[f^2(\theta;\alpha)/w^2(\theta;\beta)]$ increases and the worst value decreases.

We can write expressions for $E_w[f^2(\theta;\alpha)/w^2(\theta;\beta)]$ for known mean and known variance to determine how knowledge of the mean and variance influences the variance of the sampling. Where the mean is known,

$$E_w[f^2(\theta;\alpha)/w^2(\theta;\beta)](\beta_1=\alpha_1)=\frac{1}{\alpha_2^2}\left(\frac{\beta_2^2}{2\left(\frac{1}{\alpha_2^2}-\frac{1}{2\beta_2^2}\right)}\right)^{1/2}; \qquad (2.34)$$

and where the variance is known

$$E_w[f^2(\theta;\alpha)/w^2(\theta;\beta)](\beta_2^2=\alpha_2^2)=\exp\left[\frac{(\beta_1-\alpha_1)^2}{\alpha_2^2}\right]. \qquad (2.35)$$

Equations 2.34 and 2.35 suggest that knowledge of α_1 is more important than knowledge of α_2, providing the variance of the candidate distribution is greater than the variance of the function being sampled, with a polynomial increase in the variance for worsening estimates of α_2^2 and an exponential increase in the variance for worsening estimates of α_1. Oh [76] also discusses this, showing graphically that knowledge of the position, in this case α_1, improves the convergence more than knowledge of the scale, here given by α_2^2.

For β_2^2 less than α_2^2 there is a sharp increase in the variance as β_2^2 is decreased to $\alpha_2^2/2$, where the variance is infinite. Below $\alpha_2^2/2$, the variance expression given in Equation 2.34 is imaginary.

2.3.2 Sampling a Gamma Distribution

The gamma distribution,

$$f(\theta)=\frac{\theta^{\gamma-1}e^{-\theta/\delta}}{\Gamma(\gamma)\delta^\gamma} \qquad \theta>0 \qquad (2.36)$$

can be written as a function of the exponential family, with

$$\begin{aligned}
q(\alpha_1,\alpha_2) &= -\ln[\Gamma(\alpha_1)\alpha_2^{\alpha_1}] \\
p_1(\alpha_1,\alpha_2) &= \alpha_1-1 \\
p_2(\alpha_1,\alpha_2) &= -1/\alpha_2 \\
k_1(\theta) &= \ln\theta \\
k_2(\theta) &= \theta, \qquad (2.37)
\end{aligned}$$

where $\alpha_1 = \gamma$ and $\alpha_2 = \delta$. We begin by investigating sampling a gamma with a gamma. Considering the θ dependent terms in Equation 2.37, and using β_1 and β_2 to describe the parameters of the gamma distribution we are using as the candidate distribution, we can write $f^2(\theta)/w(\theta, \beta)$ as

$$f^2(\theta)/w(\theta,\beta) \propto \exp\left[(2\alpha_1 - \beta_1 - 1)\ln\theta + \left(\frac{1}{\beta_2} - \frac{2}{\alpha_2}\right)\theta\right]. \tag{2.38}$$

The θ term will dominate as $\theta \rightarrow \infty$. Therefore, for the integral to be finite and the sampling to converge,

$$\beta_2 > \alpha_2/2. \tag{2.39}$$

If we instead use a normal distribution as a sampling distribution for the gamma distribution, we can write $f^2(\theta)/w(\theta, \beta)$ as

$$f^2(\theta)/w(\theta,\beta) \propto \exp\left[2(\alpha_1 - 1)\ln\theta - \left(\frac{2}{\alpha_2} - \frac{\beta_1}{\beta_2^2}\right)\theta + \frac{\theta^2}{2\beta_2^2}\right]. \tag{2.40}$$

As $\theta \rightarrow \infty$ the θ^2 term will dominate. This term is greater than zero for $\theta > 0$, therefore the variance of the sampling is never finite, and using a normal distribution as a sampling function for a gamma distribution will never result in the sampling converging.

2.3.3 Sampling a Student t-Distribution

The t-distribution,

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma(\nu/2)}\frac{1}{\sqrt{\nu\pi}}\frac{1}{(1+t^2/\nu)^{(\nu+1)/2}} \tag{2.41}$$

can be written as a function of the exponential family if we make the transformation $\theta = t/\sqrt{\alpha}$, where $\alpha \equiv \nu$. Using the notation of the exponential family,

$$\begin{aligned} q(\alpha) &= \ln\left(\frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{\Gamma(\alpha/2)}\frac{1}{\sqrt{\pi}}\right) \\ p(\alpha) &= -(\alpha+1)/2 \\ k(\theta) &= \ln(1+\theta^2). \end{aligned} \tag{2.42}$$

We begin by investigating the sampling of a t-distribution with another t-distribution, with β degrees of freedom and a similar transformation of t. Considering just the θ dependent terms of $f^2(\theta)/w(\theta)$,

$$f^2(\theta)/w(\theta) \propto \exp\left[-\frac{1}{2}(2\alpha - \beta + 1)\ln(1+\theta^2)\right]. \tag{2.43}$$

Therefore, for the sampling to converge,

$$\beta < 2\alpha + 1. \tag{2.44}$$

If we now consider sampling a t-distribution using a normal distribution as a candidate distribution,

$$f^2(\theta)/w(\theta) \propto \exp\left[-(\alpha+1)\ln(1+\theta^2) - \frac{\beta_1}{\beta_2^2}\theta + \frac{\theta^2}{2\beta_2^2}\right]. \tag{2.45}$$

As $\theta \longrightarrow \infty$, the θ^2 term dominates and the variance will tend to infinity for all real β_2. Therefore, a normal distribution could not be used as a candidate distribution for a t-distribution. However, importance sampling will always converge when using a t-distribution as a candidate distribution when sampling a normal distribution, because as $\theta \longrightarrow \infty$, the θ^2 term wil again dominate, but will now be negative. Therefore, the sampling variance will always tend to zero.

2.4 Sampling Multivariate Distributions

We begin this section by considering how the convergence of importance sampling of a multivariate normal distribution varies with the number of dimensions n, and draw some general conclusions from this analysis. We then go on to consider three different specific candidate distributions: the multivariate normal distribution; a multivariate generalisation of the t-distribution and a non-standard adaptation of a multivariate normal distribution. The adaptation allows the axes of symmetry of the normal distribution to curve. These distributions are described in Sections 2.4.2, 2.4.3 and 2.4.4.

In Sections 2.4.5 to 2.4.8 we consider a two dimensional example and describe the performance of these three different candidate distributions. We show the fit of the candidate distributions to the actual function and the results of the importance sampling. Conclusions are then drawn about the suitability of each of these functions as importance sampling candidate distributions.

2.4.1 Dependence of Convergence Rate on the Number of Dimensions

The problem of how the convergence of importance sampling varies as the number of dimensions n is increased has been considered before by Au and Beck [3] and Oh [76]. Au and Beck introduce a function describing the relative entropy of the candidate distribution and the function being sampled and examine its variation to determine how convergence of the importance sampler will change with the number of dimensions. We find the variance a more useful measure of the convergence and use this and the unit coefficient of variation (unit c.o.v.) introduced by Oh [76], which is the standard deviation divided by the mean, to describe the convergence.

We assume that the function being sampled $f(\theta)$ is a standard multivariate normal distribution of n dimensions with mean vector $(\mu_1, \mu_2, \ldots, \mu_n)^T$ and covariance matrix σ, and that we use an importance sampler $w(\theta, \beta)$ that is a multivariate normal distribution with mean vector $(m_1, m_2, \ldots, m_n)^T$ and covariance matrix s. The squared unit c.o.v. for importance sampling can be written as

$$\Delta_{IS}^2 = \frac{\int_\Theta \frac{f^2(\theta)}{w(\theta,\beta)} d\theta}{\left[\int_\Theta f(\theta) d\theta\right]^2} - 1, \tag{2.46}$$

where $f(\theta)$ is the function being sampled over the range Θ and $w(\theta, \beta)$ is the importance sampler.

We find an expression for Δ^2_{IS} for general multivariate normal distributions $f(\theta)$ and $w(\theta, \beta)$. Oh [76] considers only the situation where $\mu = \mathbf{0}$ and $\sigma = \mathbf{I}$. Having obtained the general result, we go on to consider two situations; in the first we assume that the mean is known and in the second that the covariance structure is known. This allows us to determine how the variance of importance sampling is affected by the number of dimensions for discrepancies in the mean and in the covariance structure.

We consider the integral

$$\int_\Theta \frac{f^2(\theta)}{w(\theta, \beta)} d\theta. \tag{2.47}$$

As $f(\theta)$ is a normal distribution, the denominator of the first term of Equation 2.46 is one and so the integral in Equation 2.47 determines the behaviour of the unit covariance. Writing the expression out in full

$$\begin{aligned}\int_\Theta \frac{f^2(\theta)}{w(\theta)} d\theta = \frac{|s|^{1/2}}{|\sigma|(2\pi)^{n/2}} \int_\Theta \exp\Bigg[&- (\theta - \mu)^T \sigma^{-1} (\theta - \mu) \\ &+ \frac{1}{2}(\theta - m)^T s^{-1} (\theta - m)\Bigg] d\theta,\end{aligned} \tag{2.48}$$

where n is the number of dimensions. This can be written as

$$\int_\Theta \frac{f^2(\theta)}{w(\theta)} d\theta = \frac{1}{(2\pi)^{n/2}} \frac{|s|^{1/2}}{|\sigma|} \int_\Theta \exp\left[-\frac{1}{2}(\theta - \xi)\chi^{-1}(\theta - \xi) - p/2 \right] d\theta, \tag{2.49}$$

where

$$\begin{aligned} \chi &= (2\sigma^{-1} - s^{-1})^{-1} \\ \xi &= (2\sigma^{-1} - s^{-1})^{-1}(2\sigma^{-1}\mu - s^{-1}m) \\ p &= 2\mu^T \sigma^{-1} \mu - m^T s^{-1} m - \xi^T \chi^{-1} \xi. \end{aligned} \tag{2.50}$$

Evaluating the integral,

$$\begin{aligned}\int_\Theta \frac{f^2(\theta)}{w(\theta)} d\theta = &\sqrt{\frac{|s|}{|\sigma|^2 |2\sigma^{-1} - s^{-1}|}} \\ &\exp\left[-\frac{1}{2}\left(2\mu^T \sigma^{-1} \mu - m^T s^{-1} m - \xi^T \chi^{-1} \xi\right)\right].\end{aligned} \tag{2.51}$$

We consider two special cases: unknown covariance, known mean; known covariance, unknown mean.

When the covariance is unknown but the mean is known, $m = \mu$, and Equation 2.51 can be simplified to

$$\sqrt{\frac{|s|}{|\sigma|^2|2\sigma^{-1} - s^{-1}|}}. \tag{2.52}$$

To demonstrate the effect of the dimension, we consider a specific example in which the covariance matrix σ of $f(\theta)$ is diagonal and the covariance matrix s of the candidate distribution $w(\theta, \beta)$ is also diagonal, such that $s_{ii} = \sigma_{ii}(1 + \gamma)$. Under these assumptions, in n dimensions, Equation 2.51 reduces to

$$\int_{\Theta} \frac{f^2(\theta)}{w(\theta, \beta)} d\theta = \frac{(1 + \gamma)^n}{(1 + 2\gamma)^{n/2}}. \tag{2.53}$$

When the covariance is known, but the mean is unknown, $s = \sigma$ and Equation 2.51 can be simplified to

$$\exp\left[(\mu - m)^T \sigma^{-1} (\mu - m)\right]. \tag{2.54}$$

If we assume that $\mu_i - m_i = \delta$ for $i = 1, \ldots, n$ then Equation 2.54 can be written as

$$\exp\left[(\mu - m)^T \sigma^{-1} (\mu - m)\right] = \exp\left[\delta^2 \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_{ij}\right], \tag{2.55}$$

in n dimensions. Further assuming that σ is the identity matrix, we find that

$$\int_{\Theta} \frac{f^2(\theta)}{w(\theta)} d\theta = \exp(n\delta^2). \tag{2.56}$$

Therefore, as the number of dimensions increases, the variance of the sampling will increase exponentially with n, the number of dimensions for unknown mean.

Considering Equations 2.53 and 2.56, we can see that errors in the estimate of the mean will have a greater effect on the variance of the sampling for large n than a lack of knowledge about the covariance structure. For small n, the relative effects of knowledge about the mean and knowledge about the variance would depend more on the relative sizes of the discrepancies between the actual and estimated values. These two results agree with the findings of Oh [76], who considered the effect of varying ϵ and δ when sampling $f(\theta) \sim N(0, I)$ with candidate distribution $w(\theta, \beta) \sim N(\epsilon\mathbf{1}, \delta I)$.

Although we have focused only on the multivariate normal in this section, it is suspected that similar results will hold for other distribution functions, i.e. that knowledge of the mean is more important than knowledge of the shape, assuming that the candidate distribution has fat enough tails for convergence to be possible.

2.4.2 Multivariate Normal Distribution

The multivariate normal distribution has the probability density function

$$f(\mathrm{x}) = (2\pi)^{-n/2}|\sigma^{-1}| \exp\left\{-\frac{1}{2}(\mathrm{x}-\mu)^T\sigma^{-1}(\mathrm{x}-\mu)\right\}, \qquad (2.57)$$

in n dimensions, where μ is the vector of means and σ is the covariance matrix.

Samples from the multivariate normal distribution can be generated in a number of ways. We describe here the method attributed to Box and Muller for generation of standard normal variates [15] and extend this to n dimensions, using the preferred method of Barr and Slezak [7].

The proof of Box and Muller's method is given in most simulation text books (e.g. [16]) and will not be reproduced here. If U_1 and U_2 are independent random variates from a uniform distribution between 0 and 1, then Z_1 and Z_2 will be independent standard normal variates where

$$\begin{aligned} Z_1 &= (-2\ln U_1)^{\frac{1}{2}}\cos(2\pi U_2) \\ Z_2 &= (-2\ln U_1)^{\frac{1}{2}}\sin(2\pi U_2). \qquad (2.58) \end{aligned}$$

This can be extended to n dimensions using the Cholesky factorization C of the covariance matrix σ. The variable $\mathbf{W}$ will be a multivariate normal variate with covariance matrix σ and mean μ, where

$$\mathrm{W} = \mu + \mathrm{CZ}, \tag{2.59}$$

and Z is a vector of standard normal variates.

2.4.3 A Multivariate Generalisation of the Student t-Distribution

The student t-distribution has fatter tails than the normal distribution and as was shown in Section 2.3.3, is a good choice of candidate distribution when sampling a normal distribution. In this section we introduce a multivariate generalisation of the student t-distribution. This is not identical to the multivariate t-distribution introduced by Dunnett and Sobel [43], but has similar characteristics, as will be discussed below.

We use the same method as in Equation 2.59 to generate a variate from the multivariate t-distribution,

$$\mathbf{W} = \mu + \mathbf{CT}/\sqrt{\nu/(\nu - 2)}, \tag{2.60}$$

where $\mathbf{T}$ is a vector of independent t-distributed variates with ν degrees of freedom, C is the Cholesky factorisation of the covariance matrix and $\sqrt{\nu/(\nu - 2)}$ is the standard deviation of a t-distribution with ν degrees of freedom.

In generating the vector of t-distributed variates we make use of the relationship between the student t-distribution and the chi-squared and normal distributions. The random variate X will have a t-distribution with ν degrees of freedom when X is given by

$$X = \frac{Z}{\sqrt{Y/\nu}}, \tag{2.61}$$

where Z is a standard normal variate and Y is a random variate generated from a chi-squared distribution with ν degrees of freedom.

The chi-squared distribution with ν degrees of freedom is equivalent to a gamma distribution with $\alpha = \nu/2$ and $\beta = 2$. We therefore use Cheng's gamma generator [19] to generate the chi-squared variates. Combining a chi-squared variate with a standard normal variate according to Equation 2.61, we can generate a t-distributed variate. This process is followed n times to generate the n t-distributed variates $\mathbf{T}$ used to calculate $\mathbf{W}$, as described in Equation 2.60.

This multivariate generalisation differs from the multivariate t-distribution described for example in [43]. In the alternative formulation, variates are generated using the transform

$$\mathbf{W}' = \mu + \frac{(\mathbf{CZ})}{S/\sqrt{\nu}}, \tag{2.62}$$

where $\mathbf{Z}$ is a vector of standard normals, $\mathbf{C}$ is the Cholesky factorisation of the covariance matrix, S^2 is a variate from the chi-squared distribution with ν degrees of freedom and μ is the mean vector. The only difference between this method of generation and the method that we use is that only one variate is generated from the chi-squared distribution, rather than one for each component of $\mathbf{Z}$.

We can derive the probability density function for our multivariate generalisation of a t-distribution. The probability density function for a vector of independent t-distributed variates is given by

$$f(\mathbf{t}) = \left(\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma(\nu/2)}\frac{1}{\sqrt{\nu\pi}}\right)^n \prod_{i=1}^{n}(1 + t_i^2/\nu)^{-(\nu+1)/2}. \tag{2.63}$$

Using this expression and Equation 2.60, we can write the probability density function of $\mathbf{W}$, a vector of correlated t-distributed variates, with covariance structure σ and mean μ as

$$\begin{aligned} f(\mathbf{w}) = \frac{1}{|\sigma|^{1/2}} &\left(\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma(\nu/2)}\frac{1}{\sqrt{\nu\pi}}\right)^n \\ &\prod_{i=1}^{n}\left[1 + \frac{1}{\nu-2}\left(\sum_{j=1}^{n}\mathrm{C}_{ij}^{-1}(w_j - \mu_j)\right)^2\right]^{-(\nu+1)/2}. \end{aligned} \tag{2.64}$$

Working through the transformation of Equation 2.60, we can see that

$$\begin{aligned} Var\mathbf{W} &= E((\mathbf{W}-\mu)(\mathbf{W}-\mu)^T) \\ &= E(\mathbf{C}\mathbf{T}\mathbf{T}^T\mathbf{C}^T) \\ &= \mathbf{C}E(\mathbf{T}\mathbf{T}^{\mathrm{T}})\mathbf{C}^T \\ &= \mathbf{C}\mathbf{C}^T, \end{aligned} \tag{2.65}$$

as $\mathbf{T}$ is a vector of standardised student t-variates. Therefore, the covariance structure of $\mathbf{W}$ is $\sigma = \mathbf{C}\mathbf{C}^T$.

Using the alternative formulation, with only one chi-squared variable, the probability density function for the multivariate t-distribution can be written as

$$f(\mathrm{w}') = \frac{\Gamma\left[(\nu+n)/2\right]}{(\nu\pi)^{n/2}\Gamma(\nu/2)|\sigma|^{1/2}}\left[1+\frac{1}{\nu}(\mathrm{w}'-\mu)^T\sigma^{-1}(\mathrm{w}'-\mu)\right]^{-(n+\nu)/2}, \tag{2.66}$$

a tidier expression than Equation 2.64. However, the expression for the covariance matrix is more complicated in this case.

2.4.4 Bent Multivariate Normal Distribution

The bent multivariate normal distribution is not a standard distribution and was devised as part of the thesis for the purpose of testing importance sampling on a multivariate probability distribution with non-elliptical contours. Only a two-dimensional example has been considered so far. We wish to deform a multivariate normal distribution so that instead of having elliptical contours with axes of symmetry that are straight lines, we instead have contours that have curved axes of symmetry. To create a function which is not symmetric about the x-axis, i.e. has a bend in the y-direction, we make the transformations

$$\begin{aligned} x_1 &= z_1\sigma_1+\mu_1, \\ x_2 &= \mu_2+(1+d^2\sigma_1^2z_1^2)(\sigma_2z_2+a), \end{aligned} \tag{2.67}$$

where z_1 and z_2 are standard normal variates and d dictates the angle of the bend. To make this function more general, we can rotate the coordinates through an angle ϕ, allowing the function to be oriented in any direction. Incorporating this rotation, the transformations given in Equation 2.67 become

$$\begin{aligned} y_1 &= (\sigma_1 z_1 + \mu_1)\cos\phi - [\mu_2 + (\sigma_2 z_2 + a)(1 + d^2\sigma_1^2 z_1^2)]\sin\phi, \\ y_2 &= (\sigma_1 z_1 + \mu_1)\sin\phi + [\mu_2 + (\sigma_2 z_2 + a)(1 + d^2\sigma_1^2 z_1^2)]\cos\phi. \end{aligned} \tag{2.68}$$

The probability density function of this bent bivariate normal can then be written as

$$\begin{aligned} f(y_1, y_2) =& \frac{1}{2\pi\sigma_1\sigma_2(1 + d^2(u - \mu_1)^2)} \\ & \exp\left[-\frac{(u-\mu_1)^2}{2\sigma_1^2} - \frac{1}{2\sigma_2^2}\left(\frac{v - \mu_2}{1 + d^2(u-\mu_1)^2} - a\right)^2\right], \end{aligned} \tag{2.69}$$

where we use

$$\begin{aligned} u &= y_1\cos\phi + y_2\sin\phi \\ v &= -y_1\sin\phi + y_2\cos\phi \end{aligned} \tag{2.70}$$

for conciseness.

In generating a sample from this distribution, we make use of the transformation equations (Equation 2.68) describing the relationship between (y_1, y_2) and (z_1, z_2) to transform standard normal variates to random variates of the bent normal distribution. A contour plot of one instance of the bent normal distribution is given in Figure 2.1.

2.4.5 Using a Normal as Candidate Distribution for a Bent Normal Distribution

In this example, $f(\theta)$ is the bent normal distribution with $\sigma_1 = 2$, $\sigma_2 = 0.1$, $\mu_1 = -0.1$, $\mu_2 = -1.2$, $d = 2$, $a = 0.5$ and $\phi = 0.2$, as shown in Figure 2.1. The

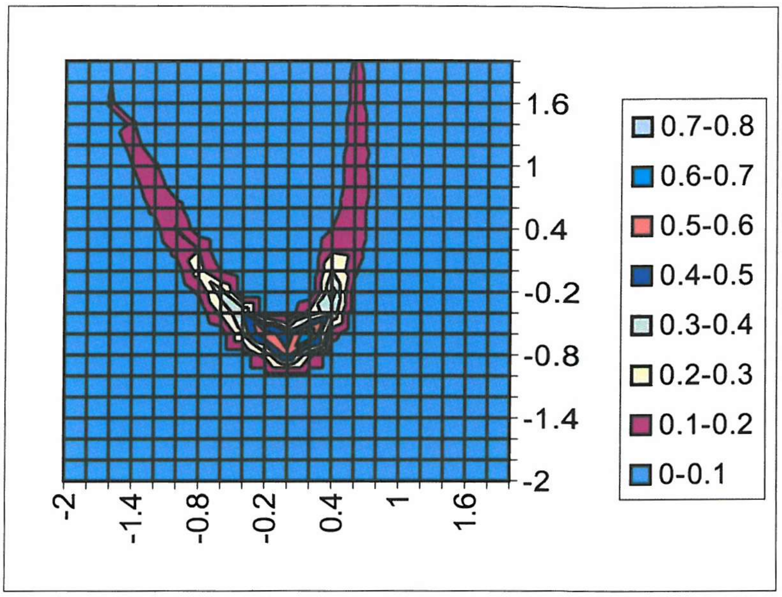

Figure 2.1: Contour plot of a bent normal distribution.

mean of the normal distribution used as the candidate distribution, $w(\theta, \beta)$ was set to be the position of the mode of the bent normal distribution, as found using the Nelder Mead optimization routine [74]. The Hessian was then calculated at the mode, and the inverse of this was used as the covariance matrix.

Sampling with the normal distribution provides a good estimate of the function close to the mode. Few points are sampled in the arms of the function however, and so a large part of the function is ignored. The reason for the difference in range of the normal candidate distribution and the bent normal is a result of the method used to estimate the covariance matrix.

Assuming for simplicity that $\phi = 0$, the matrix of second derivatives of the negative logarithm of the bent normal distribution is

$$\mathbf{H} = \begin{pmatrix} 1/\sigma_1^2 + 2d^2 & 0 \\ 0 & 1/\sigma_2^2 \end{pmatrix}. \tag{2.71}$$

Assuming normality, the covariance matrix is given by the inverse of the Hessian at the mode,

$$\sigma = \begin{pmatrix} (1/\sigma_1^2 + 2d^2)^{-1} & 0 \\ 0 & \sigma_2^2 \end{pmatrix}. \tag{2.72}$$

Therefore, $Var[x_1]$ is estimated to be $(1/\sigma_1^2 + 2d^2)^{-1}$, and $Var[x_2]$ is estimated to be σ_2^2. Using the definition of variance,

$$Var[x_i] = \int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2 (x_i - \mu_i)^2 f(x_1, x_2), \; i = 1, 2, \tag{2.73}$$

the true variances of the bent normal distribution are found to be

$$\begin{aligned} Var[x_1] &= \sigma_1^2, \\ Var[x_2] &= (\sigma_2^2 + a^2)(1 + 3d^4\sigma_1^4 + 2d^2\sigma_1^2). \end{aligned} \tag{2.74}$$

Comparison with the expressions for the variances derived assuming normality shows a significant difference. Inputting the parameter values used in the trial function, we find that the actual variance of x_1 is 4, whereas the estimated variance of x_1 is 0.121, and the actual and estimated variances of x_2 are 208 and 0.01 respectfully. Hence the normal sampling distribution will miss a significant part of the actual function when practical sample sizes are used.

2.4.6 Using a Student t-Distribution as a Candidate Distribution for a Bent Normal Distribution

We use the same bent normal distribution as in Section 2.4.5 for $f(\theta)$. The mean and the covariance structure of the t-distribution were calculated in the same way as for the normal distribution, using the mode of the function as the mean and the inverse of the Hessian as the covariance matrix. The t-distribution, therefore suffers from the same problem as the normal distribution in that the area sampled covers only a small fraction of the significant part of the actual function. This is shown very clearly in Figure 2.2 which shows the points sampled from the candidate distribution during the importance sampling and points sampled from the

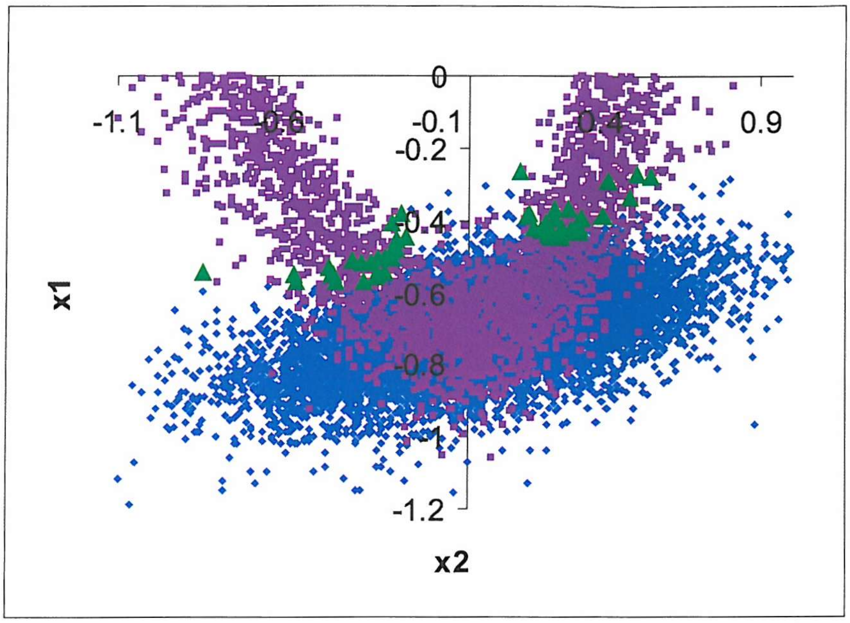

Figure 2.2: Samples from the t-distributed candidate distribution (blue) and the bent normal distribution (pink) highlighting the points with very high ratios (green).

actual function. Highlighted on the graph are the points with the highest ratios of bent normal distribution to candidate distribution. These are all in the tails of the t-distribution, but where the bent normal distribution still has a relatively high density. High ratios cause problems with convergence in importance sampling and generally also suggest that the candidate distribution being used is unsuitable, as is seen to be the case here.

The t-distribution appears to perform worse than the normal distribution when only a small number of degrees of freedom are used. The standardized t-distribution has much longer tails than the normal distribution, but the standardization means that the peak is much narrower. This means that a significant proportion of the points sampled from a t-distribution will correspond to fairly low values of the probability density function. Therefore, if the function being sampled has a significant probability in the tails of the t-distribution, as is the case in this example, points sampled in the tails will correspond to very high ratios of actual distribution to candidate distribution. This will result in very high peaks in the tails of the candidate distribution, which are artefacts of the sampling rather than true features of the function being estimated. The t-distribution suffers more from this problem in this example than the normal distribution because it samples more points in its tails.

2.4.7 Using a Bent Normal Distribution as a Candidate Distribution for a Bent Normal Distribution

Results given at the start of this chapter suggest that ideally the sampling function should be identical to the function being sampled. Therefore, we consider using a bent normal distribution as a candidate distribution. We estimate the mode and second derivatives of the bent normal distribution being sampled and use these to determine the optimal values for most of the parameters of the candidate distribution, $w(\theta, \beta)$. The mode of the bent normal distribution occurs at

$$\begin{aligned} x_1 \cos\phi + x_2 \sin\phi &= \mu_1 \\ x_2 \cos\phi - x_1 \sin\phi &= \mu_2 + a. \end{aligned} \tag{2.75}$$

The second derivatives of the bent normal distribution at the mode are given by

$$\left.\frac{\partial^2 f}{\partial x_1^2}\right|_{\text{mode}} = -N\left[(1/\sigma_1^2 + 2d^2)\cos^2\phi + 1/\sigma_2^2 \sin^2\phi\right] \tag{2.76}$$

$$\left.\frac{\partial^2 f}{\partial x_1 x_2}\right|_{\text{mode}} = -N\left(1/\sigma_1^2 + 2d^2 - 1/\sigma_2^2\right)\sin\phi\cos\phi \tag{2.77}$$

$$\left.\frac{\partial^2 f}{\partial x_2^2}\right|_{\text{mode}} = -N\left[(1/\sigma_1^2 + 2d^2)\sin^2\phi + 1/\sigma_2^2 \cos^2\phi\right], \tag{2.78}$$

where N is the value of the function at the mode. For the normalized function, $N = \frac{1}{2\pi\sigma_1\sigma_2}$ but in general the posterior distribution will not be normalized, and we assume throughout that N can take any (unknown) value. Solving Equations 2.76 - 2.78, gives us expressions for σ_2, $2d^2 + 1/\sigma_1^2$ and ϕ in terms of known variables. This leaves four variables undetermined a, μ_2, σ_1 and d, with expressions linking a and μ_2, and d and σ_1. We used the Nelder Mead optimization routine [74] to find the best values for these parameters, with the best values defined as being those that minimize the sum of the squared error between the actual function and the candidate distribution for a set of one thousand randomly generated samples. Table 2.4.7 shows the parameter values for the actual function and the sampling function, demonstrating how good a fit can be obtained using this method.

	Actual	Sampling
σ_1	2	2.01
σ_2	0.1	0.100
μ_1	-0.1	-0.0966
μ_2	-1.2	-1.19
d	2	2.01
a	0.5	0.492
θ	0.2	0.193

Table 2.1: Comparison between the parameters used in the sampling function and those of the actual function

Using this function as a candidate distribution gives very disappointing results. We obtain very high values for the ratio of actual function to candidate distribution in the tails of the function. These values far exceed those obtained in regions where the actual and candidate distributions both have higher values. In fact, where the candidate distribution is a good fit, the ratio of actual function to candidate distribution tends to be of order one. Elsewhere, the ratios rise to the order of hundreds and thousands, with a continuum up to the very high values. The parameter values corresponding to the large ratios have a very small probability of being sampled (of the order of 10^{-5} or less of the maximum), but the actual function is often a factor of 10^8 or more different from the sampling probability. The discrepancy in the parameter values makes very little difference to the fit of the candidate distribution to the actual function, but a huge difference to the results of the importance sampling.

These results suggest that finding a good fit to the function being estimated is not always good enough to ensure that the importance sampling works well. It is also essential to find a sampling function with fatter tails than the function being estimated.

2.4.8 Using a Bent t-Distribution as a Candidate Distribution for a Bent Normal Distribution

In this section, we introduce a bent t-distribution and present some of the results obtained when using this function as a sampling distribution for the bent normal distribution, shown in Figure 2.1. By constructing a bent t-distribution, we hope to construct a suitable candidate distribution for a bent normal distribution, that has a similar shape but fatter tails.

We use the transformations described in Equation 2.68, using standard t-variates in place of the standard normal variates z_1 and z_2 to generate variates of the bent t-distribution. The probability density function for the bent t-distribution with ν degrees of freedom can be written as

$$\begin{aligned} f(x_1, x_2) =& \frac{K}{1 + d^2(x_1 - \mu_1)^2} \\ & \left\{ \left[1 + \frac{(x_1 - \mu_1)^2}{\nu \sigma_1^2} \right] \left[1 + \frac{1}{\sigma_2^2 \nu} \left(\frac{x_2 - \mu_2}{1 + d^2(x_1 - \mu_1)^2} - a \right)^2 \right] \right\}^{-(\nu+1)/2} \end{aligned} \tag{2.79}$$

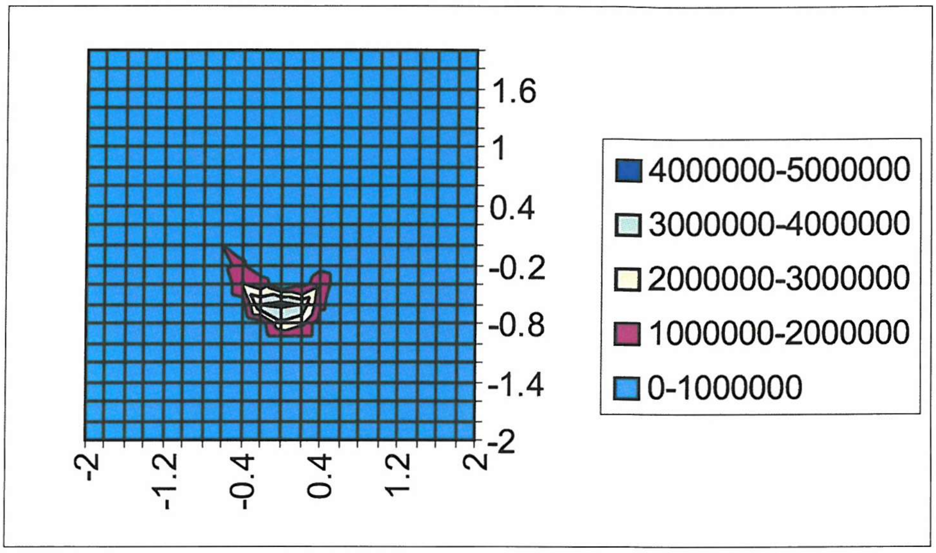

Figure 2.3: Contour plot of the bent t-distribution with four degrees of freedom.

in unrotated coordinates, where

$$K = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \left(\frac{\nu-2}{\nu}\right)^{1/2} \frac{1}{\nu\pi\sigma_1\sigma_2}. \tag{2.80}$$

In rotated coordinates, x_1 would be replaced by $y_1 \cos\phi + y_2 \sin\phi$ and x_2 by $y_2 \cos\phi - y_1 \sin\phi$.

We obtain best-fit parameters using the procedure detailed in Section 2.4.7, for finding the parameters of a bent normal distribution, and then experiment with different values for the degrees of freedom ν to obtain the best results for the estimated function.

Figure 2.3 shows the bent t-distribution for $\nu = 4$. Comparison with Figure 2.1 (the bent normal distribution) shows that the bent normal distribution has arms of approximately equal length, whereas the bent t-distribution does not. The estimate to the function obtained using this candidate distribution is given in Figure 2.4 and shows the effect of this discrepancy. In addition to the peak corresponding to the maximum of the function, there is a second peak in the region where the difference

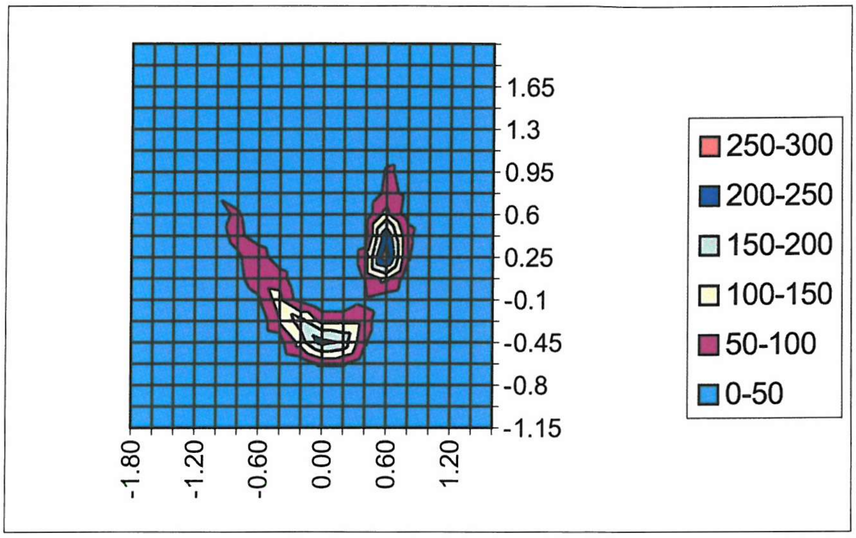

Figure 2.4: Contour plot of the estimated bent normal distribution.

between the bent normal and the bent t-distribution is largest. Similar results are obtained with higher values of ν.

With $\nu = 3$, the estimated function is similar to that obtained using the standard bent normal distribution as candidate distribution, with very high peaks in the tails of the function.

These results suggest that the bent t-distribution would not be suitable as a candidate distribution for a bent normal distribution as the shapes of the two functions are different.

2.5 Discussion

The optimal candidate distribution to use in importance sampling is a normalised version of the function whose integral we are trying to estimate, as we showed in Section 2.2. However, finding this function involves evaluating the integral and so this is not a practical solution to the problem. The results suggest that we should use a candidate distribution that is as close as possible in form to the function whose integral we are evaluating.

The wrong choice of candidate distribution in importance sampling can result in a situation where the sampling does not converge, and the variance of the importance sampling is infinite. In general, this can be overcome by ensuring that the candidate distribution used has thicker tails than the function being integrated. We discussed this in Section 2.3 for one-dimensional functions, and gave some general results for functions from the exponential family.

As the number of dimensions increases, knowledge about the function becomes more critical. Expressions were derived for the variance of importance sampling when both the sampler function and the function being integrated are multivariate normal, and are given in Section 2.4.1. These showed that knowledge of the mean is more important than knowledge of the covariance structure when defining the candidate distribution in importance sampling.

The practical examples introduced in Section 2.4 show that obtaining convergence of importance sampling in multi-dimensional space is difficult when the function being integrated has a different shape from a normal distribution, with non-elliptical contours. None of the candidate distributions tried worked well in this situation, but the lack of convergence was easily diagnosed by the extreme values of the ratios of the function being sampled to the candidate distribution for a few of the observations in each of the runs of the sampling.

The results of this chapter suggest that some time should be spent learning about the function being sampled prior to defining the candidate distribution. They also demonstrate that importance sampling does not perform well in all situations and that some kind of robustness test is required to check that the sampling has converged. Chapter 3 discusses a number of convergence tests for importance sampling, all of which work by examining the distribution of the ratios output for each observation. The ratios should be approximately equal if importance sampling is performing well, with extreme ratios suggesting that importance sampling is not converging.

Chapter 3

Techniques for Measuring the Convergence of Importance Sampling

3.1 Introduction

In this chapter, we describe diagnostic and statistical methods for assessing the convergence of importance sampling. As discussed at the end of Chapter 2, the distribution of the importance sampling weights, which are the ratio of the function being sampled to the candidate distribution at each sampling point, give a good indication of whether the sampling has converged. All of the methods we consider for assessing convergence in this chapter use only the values of these weights in their assessment.

The diagnostic tests that we consider mainly involve graphical indicators of convergence, such as plotting the variation in the variance over the sampling. We also consider statistical tests based on extreme value theory that test whether the variance of the sampling is finite or not. To compare the performance of the different methods and demonstrate their use, we apply them to two simple examples, one of well-behaved and the other of non-convergent importance sampling.

3.2 Diagnostic Tests

3.2.1 Plot of Top One Hundred Weights and Variance of Weights

In this test, the highest one hundred weights are plotted as they occur in the sample, with the running estimate of the variance of the weights plotted on the same graph. From this we can tell if the sample is being biased by any very large weights. With perfect convergence, all the weights would be equal, and if importance sampling is converging well this plot should show weights to be of a similar order of magnitude, and the variance of the sample should not be affected significantly by any individual weight.

For example, Figure 3.1 shows the top one hundred weights in a sample of ten thousand when a student t-distribution is being used to sample a normal distribution, with the variance of the weights over the run superimposed. This is an example where importance sampling does work well. Figure 3.2 on the other hand shows the top one hundred weights in a sample of ten thousand when using a normal distribution as a candidate distribution for a t-distribution. This is an example where importance sampling does not perform well and we can see that there are two very high-valued weights that have a great effect on the variance of the sampling.

We can relate the variance of the weights to the variance of the sampling as follows. An expression for the variance of the sampling is given in Equation 2.6. The second term of this expression is independent of the candidate distribution; therefore assuming that the quantity we are trying to estimate has a finite variance, this term can be ignored. In assessing whether convergence will occur, we can

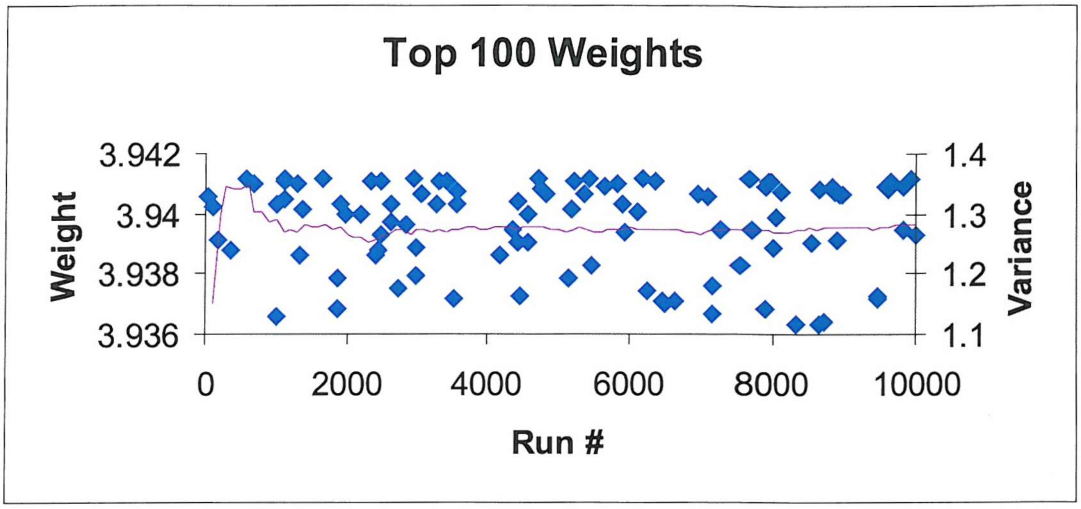

Figure 3.1: Distribution of the highest one hundred importance sampling weights when sampling a normal distribution with mean zero and variance one with a t-distribution with three degrees of freedom. The variance of the weights over the run is superimposed.

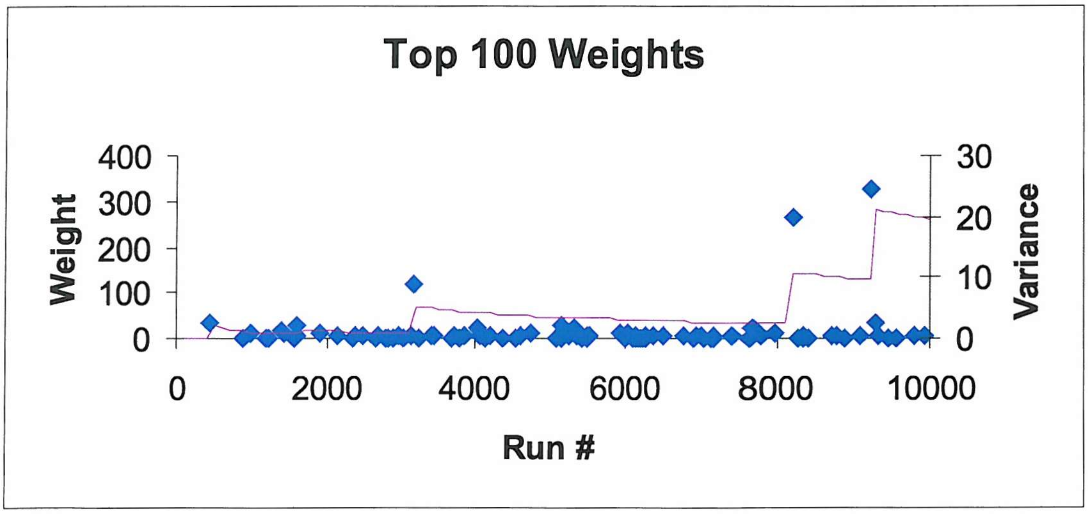

Figure 3.2: Distribution of the highest one hundred importance sampling weights when sampling a t-distribution with ten degrees of freedom with a normal distribution with mean zero and variance one. The variance of the weights over the run is superimposed.

therefore concentrate on the first term, which can be approximated by

$$\int_{\Theta} \left(\frac{m(\theta)f(\theta)}{w(\theta,\beta)} \right)^2 w(\theta,\beta) d\theta \approx \frac{1}{n} \sum_{i=1}^{n} \left(\frac{m(\theta_i)f(\theta_i)}{w(\theta_i,\beta)} \right)^2, \tag{3.1}$$

where the n samples $\{\theta_i\}$ are drawn from the candidate distribution $w(\theta, \beta)$. The variance of the weights can be expressed as

$$Var_w = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{m(\theta_i)f(\theta_i)}{w(\theta_i,\beta)} \right)^2 - \left(\frac{1}{n} \sum_{i=1}^{n} \frac{m(\theta_i)f(\theta_i)}{w(\theta_i,\beta)} \right)^2, \tag{3.2}$$

which is equal to the expression in Equation 3.1 minus the square of the mean of the weights. Therefore, if the variance of the weights is divergent, the variance of the sampling will also be divergent.

We can judge whether the variance is tending to some finite value or is divergent by evaluating the variance of the weights over the run and observing any trends. Necessarily, this judgment must be subjective, but as Figures 3.2 and 3.1 show, observing the evolution of the variance of the weights can give a good indication of sampling convergence.

3.2.2 Distribution of Weights Between Observations

This is a diagnostic test that we have developed to give an indication of the proportion of the sum of the weights that is being assigned to individual observations and to groups of observations. We calculate the maximum normalised weight initially to determine how extreme this weight is. The proportion of the sample making up different proportions of the sum of the weights can also be useful in assessing convergence.

This test can find examples where the importance sampling is definitely not performing well but can occasionally suggest excellent convergence when other tests indicate that this is not the case. For example, Figures 3.3 and 3.4 show that

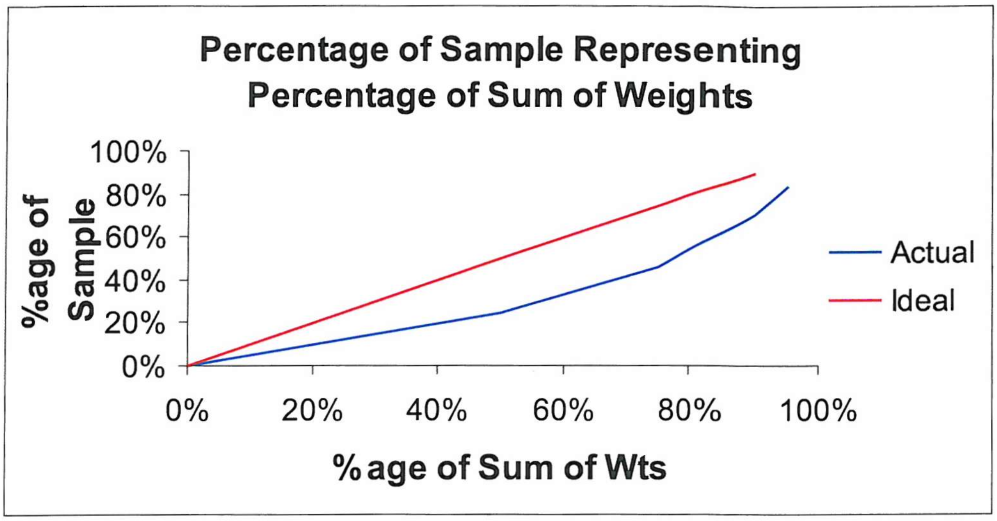

Figure 3.3: The percentage of the sample points making up percentages of the sum of the weights when sampling a normal distribution with mean zero and variance one with a t-distribution with three degrees of freedom.

for the case in which importance sampling is less appropriate (3.4), the results appear better than for the case in which importance sampling should definitely converge (3.3). Considering the highest-valued weights for these examples, however, we find that the example in which a t-distribution is used as a candidate distribution for a normal (a good use of importance sampling), has a maximum normalised weight of 0.02% and in the example where a normal distribution is used as a candidate distribution for a t-distribution (a poor use of importance sampling) the maximum normalised weight is 3.0%. In an ideal situation, all normalised weights should have an equal value of one over the number of samples; in this case a value of 0.01%.

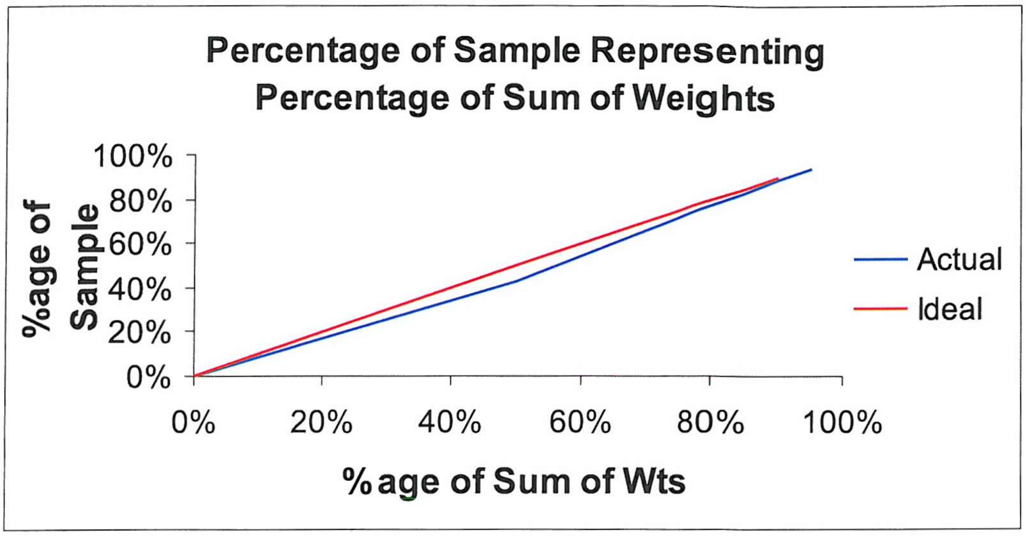

Figure 3.4: The percentage of the sample points making up percentages of the sum of the weights when sampling a t-distribution with ten degrees of freedom with a normal distribution with mean zero and variance one.

3.3 Statistical Tests

Importance sampling is only valid if the variance of the sampling is finite, as was shown in Chapter 2 and has also been discussed by Geweke [53]. Proving that the variance is finite can be very difficult for high-dimensional complex integrals. We consider below a method based on extreme value theory that was recently proposed by Koopman and Shephard [65].

3.3.1 Tests Based on Extreme Value Theory

To determine whether the variance is finite, we need to investigate the distribution of the weights. We make use of results from extreme value theory and fit a generalised pareto distribution (GPD) to the highest valued weights. The shape parameter ξ of the GPD determines the number of moments that exist. The best-fit value for ξ for a set of weights can therefore be used to determine whether the variance of the weights is finite. This work follows that of Koopman and Shephard [65].

The generalised pareto distribution (GPD) describes the distribution of excesses over a threshold and has probability density function

$$f(z) = \frac{1}{\beta}\left(1 + \xi\frac{z}{\beta}\right)^{-\frac{1}{\xi}-1}, \qquad (3.3)$$

where z are the exceedances over the threshold u, such that $z > 0$. For $\xi < 0$, we have the additional constraint that $z < -\beta/\xi$.

According to Smith [96], if we have a set of independent, identically distributed weights $\{y_i\}$, then as the threshold u increases, the limit distributions of the random variables over this threshold $z_i = (y_i - u)$ will be generalised pareto. The threshold u is defined by the user, and the choice must be made carefully to ensure that u is low enough for there to be sufficient data available to use for inference, but high enough for the excesses to follow a GPD distribution.

As only $1/\xi$ moments exist, the variance is finite only if $\xi \leq 0.5$. Following Koopman and Shephard [65], we test the hypothesis

$$H_0 : \xi = \tfrac{1}{2} \quad \text{and} \quad H_1 : \xi > \tfrac{1}{2}, \qquad (3.4)$$

where equality is used in the expression for H_0 to simplify the statistical analysis.

The score vector s of the parameters $\lambda = (\xi, \beta)$ for a sample of n exceedances z_i is given by

$$\begin{aligned} \begin{pmatrix} s_\xi \\ s_\beta \end{pmatrix} &= \frac{\partial \log f(z;\lambda)}{\partial \lambda} \\ &= \begin{pmatrix} \xi^{-2}\sum_{i=1}^{n}\log x_i - (1+\xi^{-1})\beta^{-1}\sum_{i=1}^{n} z_i/x_i \\ -n\beta^{-1} + (1+\xi)\beta^{-2}\sum_{i=1}^{n} z_i/x_i \end{pmatrix}, \end{aligned} \qquad (3.5)$$

where we use the shorthand $x_i = 1 + \xi\beta^{-1}z_i$. The expected information matrix of λ is then nI, where

$$I = \frac{1}{(1+2\xi)(1+\xi)} \begin{pmatrix} 2 & \beta^{-1} \\ \beta^{-1} & \beta^{-2}(1+\xi) \end{pmatrix}. \tag{3.6}$$

This is a different expression from that of [65]. Using this expression for the information matrix, the asymptotic distribution of the maximum likelihood estimator $\hat{\lambda}$ is given by

$$\sqrt{n}(\hat{\lambda} - \lambda) \rightarrow^{d} N(0, I^{-1}), \tag{3.7}$$

where

$$I^{-1} = (1+\xi) \begin{pmatrix} 1+\xi & -\beta \\ -\beta & 2\beta^2 \end{pmatrix}. \tag{3.8}$$

We know from [96] that likelihood inference is regular for $\xi > -1/2$.

We use three different hypothesis tests:

1. The Wald test is based directly on the result of Equation 3.7 and involves computation of an asymptotic signed t-test

$$t = \frac{\left(\hat{\xi} - \frac{1}{2}\right)}{\sqrt{\sigma_{\xi_0\xi_0}}}, \tag{3.9}$$

where $\sigma_{\xi_0\xi_0}$ is the diagonal component of I^{-1} corresponding to ξ, evaluated at $\xi = 1/2$. This gives

$$t = \frac{2\sqrt{n}}{3}\left(\hat{\xi} - \frac{1}{2}\right), \tag{3.10}$$

where the null hypothesis is rejected if t takes a large positive value compared with the standard normal.

2. In the score test we consider the score value for the null hypothesis

$$s_{\xi}^{0} = 4\sum_{i=1}^{n} \log x_i - 3\beta^{-1}\sum_{i=1}^{n} \frac{z_i}{1+\beta^{-1}z_i/2}. \tag{3.11}$$

Its standardised value

$$s_\xi^* = \frac{s_\xi^0}{\sqrt{I_{\xi_0\xi_0}}}, \tag{3.12}$$

is asymptotically N(0,1) under H_0. Substituting $\xi = 1/2$ into Equation 3.6, we can evaluate $I_{\xi_0\xi_0}$ and write s_ξ^* as

$$s_\xi^* = \sqrt{\frac{2}{3n}} s_\xi^0. \tag{3.13}$$

In the score test we therefore reject the null if s_ξ^* is significantly positive compared with a standard normal.

3. The likelihood ratio test compares the log likelihoods of the best fit parameters $(\hat{\xi}, \hat{\beta})$ under the restriction that $\hat{\xi} \geq 1/2$, and the best fit parameters where ξ is restricted to be equal to $1/2$. The maximum likelihood estimator for β is then given by $\hat{\beta}_0$. We evaluate

$$LR = 2\left[\log f(z; \hat{\xi}, \hat{\beta}) - \log f(z; \hat{\beta}_0, \xi = 1/2)\right]. \tag{3.14}$$

Using Equations 3.3 and 3.14,

$$\begin{aligned} LR =& 2\left[n(\ln\hat{\beta}_0 - \ln\beta) + 3\sum_{i=1}^{n} \ln(1 + z_i/(2\hat{\beta}_0))\right. \\ & \left. -(1 + \frac{1}{\xi})\sum_{i=1}^{n} \ln(1 + \beta\xi z_i/\beta)\right]. \end{aligned} \tag{3.15}$$

The null hypothesis is rejected if LR is high compared with $(\chi_0^2 + \chi_1^2)/2$, where χ_0^2 is a unit point mass at the origin. The χ_0^2 term arises because H_0 is on a boundary [23].

Asymptotically these tests should give the same results, with the likelihood ratio considering the differences in log likelihood between the maximum and the hypothesis point, the Wald test considering the difference in the position of the maximum likelihood estimator and hypothesis point, and the score test considering the difference between the gradient of the log likelihood surface at the hypothesis point and at the actual maximum. With a finite number of data points, the results of the three tests may be different, and it is useful to take all of them into account when assessing the convergence of the importance sampling.

We use Davidon's method of conjugate gradients [37] (a good description is given in [17]) to fit the unrestricted maximum likelihood estimates $(\hat{\xi}, \hat{\beta})$, and a Fibonacci line search to fit $\hat{\beta}_0$. We found that the method of Fisher Scoring suggested by Koopman and Shephard [65] frequently entered infeasible regions of parameter space. We compared the results of our maximum likelihood fitting with those produced using the ExtRemes toolkit [54] and found that our estimates of ξ and β matched their results in all cases of interest.

When the weights are very small, problems are encountered fitting the GPD. The parameter β, which acts as a scale parameter is very small in these situations, but tends to have very high derivatives, making the optimization routine suggest infeasible values. Rescaling the data before performing the fitting routine seems to help, but more investigations are required into the sensitivity of the final values to the scale parameter used. Earlier work on fitting a GPD to data [18] has tended to concentrate on datasets for which $\xi < 1/2$, considering the case where $\xi > 1/2$ to be less practically useful. The problem of very small data values also does not appear to have been considered within the literature.

Results for the two examples are given in Table 3.3.1. We find that the results corroborate the theory of Chapter 2, that using a t-distribution as the candidate distribution when sampling a normal distribution results in a finite sampling variance, whereas using a normal distribution as a sampling function for a t-distribution leads to the sampling having a non-finite variance.

3.4 Evaluation of Convergence Tests

We find the diagnostic tests presented here to be very useful tools for assessing convergence of importance sampling and indicating situations where the candidate distribution is not sufficiently close to the actual function for importance sampling to work efficiently. We have introduced one new diagnostic test to those of Koopman and Shepherd [65], evaluating the proportion of the sample accounting for different proportions of the sum of weights. This did not prove useful for assessing convergence for the two examples presented here, suggesting good convergence for the example where convergence was poor and worse convergence in the example where convergence was in fact good. In other examples, this test has been found to be useful, where there are a few very extreme weights that absorb most of the probability. We have not included an histogram of all but the top one hundred weights, as suggested by Koopman and Shephard because we did not find this to be a useful tool. It told us very little about the extreme weights, which seem to be principally responsible for non-convergence.

Fitting the generalised pareto distribution to the weights can be time consuming as this needs to be done for a number of different thresholds to check that the

Example	Test	Statistic	Result
Good IS	Wald test	-1300	Accept H_0
	Score test	-178	Accept H_0
	Likelihood ratio	-43600	Accept H_0
Poor IS	Wald test	3.72	Reject H_0
	Score test	253	Reject H_0
	Likelihood ratio	2650	Reject H_0

Table 3.1: Results of statistical tests of the hypothesis given in Equation 3.4 to assess convergence of importance sampling.

weights included in the fitting process come from a distribution of this type. If one threshold only could be chosen, this would reduce CPU time, but perhaps at the expense of accuracy. The statistical tests appear to be useful in assessing the convergence and possibly provide a more concrete measure than the diagnostic tests.

The recommendation based on this investigation is to use both diagnostic and statistical tests. Non-convergence can generally be determined from the diagnostic tests, with the statistical tests confirming the user's beliefs.

Chapter 4

Bayesian Model Selection

4.1 Introduction

In this chapter, we describe the application of Bayesian methods to model selection in normal mixture models. These are models of the form

$$f(x) = \sum_{i=1}^{k} w_i g(x|\theta_i), \tag{4.1}$$

where $g(.)$ is a normal distribution. The problem that we consider is the statistically non-standard one of finding the probability distribution for k, the number of components in the normal mixture. The focus of the work is on examples in which there is no prior information available. We use importance sampling to find the posterior distribution for the number of components in the mixture and apply our methodology to a number of standard datasets. There are two main applications: semiparametric density estimation, such as that used in input modelling for simulation models [21] and determination of the number of distinct groups present in a dataset, when it is known prior to the investigation that such groups do exist.

We use a Bayesian framework to analyse the problem, finding the posterior distribution of the number of components in the mixture. Maximum likelihood methods have some drawbacks in this application as the likelihood surface has discontinuities near its boundaries, e.g. as the component variances tend to zero. In addition, a maximum likelihood methodology will tend to prefer the model with the largest number of components, as this is always the model that best fits the data. Including the prior distribution acts to smooth out the discontinuities in the likelihood.

A review of the available literature on the analysis of finite mixture models is given in Section 4.2. There are a number of issues concerned with model selection in finite mixture models, and these are discussed in Section 4.3. We then go on to describe the methodology that we have used to solve this problem in Sections 4.4 and 4.5. Results are presented in Section 4.6 and are followed by a discussion in Section 4.7.

4.2 Literature Review

The main issues in designing a Bayesian methodology for the solution of this problem are the choice of prior distribution and the sampling methodology used to find the posterior distribution. We begin by discussing the choice of prior distribution in Section 4.2.1 and then go on to describe the different sampling methodologies used to find the posterior distribution in Section 4.2.2.

4.2.1 Prior Distribution

It is not possible for the prior distribution used to be fully non-informative and still to obtain proper posterior distributions for mixture models. The choice of prior distribution in a mixture model setting that is proper and yet sufficiently non-informative is an important part of the methodology.

Both Richardson and Green [84] and Roeder and Wasserman [88] use a Dirichlet distribution with parameters set equal to 1 as a prior for the component weights, and a discrete uniform distribution to describe the number of components in the model such that $Pr\{K = k\} = 1/k_{max}$, $k \leq k_{max}$ and zero for all other values of k. Providing k_{max} is chosen to be sufficiently large, these priors impart no influence on the posterior distribution.

Phillips and Smith [80] use a modified Poisson distribution as a prior distribution for the number of components, in which the probability of obtaining no components is zero, and a uniform distribution as a prior distribution for the weights. The use of a Poisson distribution places a greater probability mass on values closer to the input parameter of the distribution, which in this example is a hyperparameter that is chosen by the user. Therefore, this prior distribution is less flat than the discrete uniform prior described in the previous paragraph.

The prior distributions for the number of components and the weights used by Escobar and West [45] are fundamentally different from those described above. They use a Dirichlet process as a prior distribution for the mixture, with the distribution of the $(n+1)^{th}$ sample, conditional on the previous n estimates, equal to

$$\pi_{n+1}|\pi \sim \alpha a_n G_0(\pi_{n+1}) + a_n \sum_{j=1}^{k} \pi_j \delta_{\pi_j}(\pi_{n+1}). \tag{4.2}$$

Therefore there is a positive probability that the n^{th} sample from the distribution comes from the same component as one of the previous samples. Using this distribution, the expected number of components in the mixture for a sample of size n is proportional to $\ln(1 + n/\alpha)$. Therefore, as the sample size increases, the expected number of components also increases. Although to a certain extent this is logical, it does lead to some influence in the prior distribution. Priors of this form, with Dirichlet mixtures and Poisson-like priors for the number of components are more geared toward density estimation than to determining the number of components, with the number of components being treated more as a nuisance parameter [99].

Most authors use a normal prior for the means of the components and a gamma distribution for the inverse variances. This choice of distributions gives some advantages of conjugacy. Richardson and Green [84] extend this to include a hyperprior structure for the shape parameter in the gamma distribution for the inverse variance. They argue that although it is possible to define a sufficiently vague prior distribution for the mean based on the range of the data available, little information can be gleaned from the data about the variances of the components. In their model,

$$\begin{aligned} \mu_i &\sim N(\xi, \kappa^{-1}) \\ \sigma_i^{-2} &\sim \Gamma(\alpha, \beta) \\ \beta &\sim \Gamma(g, h), \end{aligned} \tag{4.3}$$

where g, h ξ, κ and α are hyperparameters to be determined by the user. With this approach, the prior probability distributions for the means and variances of the components are independent of each other.

Hierarchical priors are also advocated by Berkhoff, van Mechelen and Gelman [11] who investigated the sensitivity of the prior structure for a latent class model. They argue that by using a hierarchical model for the prior distribution, they are selecting prior distributions that are not contradicted by the data. Applying this methodology to a model of psychiatric symptoms, they find that the hierarchical prior distribution produces more sensible posterior distributions than the non-hierarchical distributions.

Roeder and Wasserman [88] use what they describe as partially proper priors for the means and standard deviations of the component parameters. These are partially proper in the sense that the overall scale and location of the parameters require no subjective input but the parameters for different components are linked. The means are loosely linked through a Markov Chain, which means that the prior distribution for the position of an individual component mean in parameter space is flat but the distribution describing the distance between two component means is not. The joint prior distribution for the component variances is a product of scaled inverse-chi distributions with a common scale parameter and common degrees of freedom. This has the effect of pushing all of the component standard deviations toward some common, unspecified value. The prior requires two hyperparameters, one influencing the distance between the component means and the other affecting the difference in the scale of the component variances.

Although this choice of prior distribution could be used in many different applications without adaptation, it does impose some structure on the problem through having non-flat distributions describing the distance between the component means and the difference in scale of the component variances. A prior distribution that imposes some scale on the component means and variances but treats them independently may actually impart less information. Further problems arrive with Roeder and Wasserman's approach if the data being modelled comes from a mixture of components when two or more of those components have the same mean. The prior that they use has zero probability of this occurring and so prevents the correct posterior probability distribution being obtained.

Stephens [97] suggests however, that choosing a vague prior distribution for this problem is more difficult than it might first appear, and this point is also picked up by Jennison in the discussion of Richardson and Green's paper [61]. Both show the dependence of the posterior distribution for the number of components k, on the prior distributions used for the component means and variances. Stephens discusses how for a very small variance in the prior distribution for the component means or variances, models with a low number of components are favoured, then as the variance is increased, the prior distribution favours models with high numbers of components. As the variance is increased even further, to very high levels, the prior distribution again begins to favour models with few components. We discuss this sensitivity of the posterior distribution of the number of components to the prior distributions of the other model parameters further in Section 4.3.3.

4.2.2 Sampling Methodologies

Previous work in this area has mainly concentrated on the use of Markov Chain Monte Carlo (MCMC) methods for the solution of the problem. Richardson and Green [84] and Phillips and Smith [80] describe reversible jump methodologies for model selection. Stephens [97] again uses MCMC sampling but considers an alternative to the reversible jump methodologies. Independence sampling has been considered by Cheng [20] and importance sampling by Raftery [82]. Raftery's approach is based on the estimation of marginal likelihoods for each of the possible models, with several methods of determining the marginal likelihoods proposed, importance sampling being just one. His favoured approach is the Laplace-Metropolis estimator, which is based on the Laplace method, but uses posterior simulation to estimate the quantities that the Laplace method needs. In the example he considers of one-dimensional mixing, Gibbs sampling is used to perform the posterior simulation.

In jump-diffusion sampling [80], [84], the Markov chain can make discrete transitions between different models (jumps) and can sample model-specific parameters between these transitions (diffusion). Two different jump dynamics are described by Phillips and Smith [80]: Gibbs and Metropolis-Hastings. In Gibbs jump dynamics, the jump intensity, i.e. the probability of jumping from the current model to a new model, is proportional to the full conditional distribution. With the Metropolis-Hastings jump dynamics, jump times are calculated using a modified jump intensity dependent on the prior distribution, and are accepted with a probability of $\min\{1, \exp\left[L_j(\phi) - L_k(\phi)\right]\}$, where $L_i(\theta)$ is the log likelihood of a parameter set θ for model i. The diffusion step then consists of a Langevin diffusion in the subspace corresponding to the current model.

Phillips and Smith [80] have applied this methodology to normal mixture models, and here they restricted the jump space of the sampler, allowing it only to jump from its current state k to models with either $k-1$ or $k+1$ components. Therefore, the number of components can only increase or decrease by 1 in each move of the sampler. Metropolis-Hastings jump dynamics were used, as obtaining full conditional distributions for the parameters is difficult for mixture models.

A similar methodology is used by Richardson and Green [84], who also use Metropolis-Hastings jump dynamics in a reversible jump MCMC sampler. Instead of performing the random sampling of jump times, as used by Phillips and Smith [80], Richardson and Green use a systematic approach, in which the parameters in the current model are sampled from their full conditional distributions, and the sampler then goes on to either split one component or combine two components, resulting in a model with either one more or one less component.

Stephens's MCMC sampler [97] appears simpler than the reversible jump samplers described above. The method is based on the construction of a continuous time Markov birth-death process which has the posterior as its stationary distribution. The number of components is varied in the model by allowing new components to be born and old components to die. Births occur at a constant rate from the prior while deaths occur at a rate dependent on the quality of the component.

A further example of the use of MCMC methods is given by Escobar and West [45], who use Gibbs sampling to find the full posterior distribution. This is made possible by their choice of prior distribution for the number of components and the weightings associated with these components, as it means that a discrete move between different models is not required.

Other methods proposed in the literature determine the posterior distribution using the marginal likelihoods of the different models. The posterior distribution for a model M_k and the Bayes factor for comparing that model with others are dependent on the marginal likelihood $P(M_k|D)$, where D is the available data. The marginal likelihood for a model is defined to be the integral over that model's parameter space of the prior distribution multiplied by the likelihood. Raftery [82] proposes using importance sampling or maximum likelihood methods for finding the marginal likelihood, while Roeder and Wasserman [88] and Chib [24] use Gibbs sampling. As a result of using partially proper priors, Roeder and Wasserman can only estimate the marginal likelihoods up to an unknown factor, therefore, the Schwarz criterion [94] is used to find the optimal model. Under the Schwarz criterion, the optimal model is that for which $\ln[p(D|\hat{\theta}_k)] - \frac{1}{2}d_k \ln[n]$ is largest, where model k has d_k parameters θ_k, and n is the number of observations, denoted by D.

Berkhof et al [11] also use marginal likelihoods to calculate Bayes factors for model selection, using a variant of Chib's estimator [24] for the computation. The variant involves implementing a relabelling transition for the mixture components, as suggested by Neal [73], to enhance mixing between different modes of the mixture distribution. Due to non-identifiability of the components in a mixture model, the posterior for a model with k components will have $k!$ symmetrical modes. Neal argues that marginal likelihoods cannot be used to compare models if the sampling does not allow sufficient mixing between these different modes, as the estimates of the marginal likelihoods will be incorrect. Further discussion of identifiability in mixture models is given by Crawford [31] and in Section 4.3.1.

Cheng [20] describes a different method again of determining the posterior distribution of a mixture model. He assumes during the sampling that all components are present in the model, up to a maximum number k_{max}. Markov Chain Monte Carlo sampling is then used to produce m samples from the posterior distribution of this model. For each of the samples, if the weighting assigned to a component is below some predetermined value δ, the component is ignored. No rule for choosing δ is given and this must be determined subjectively by the user. In his paper, Cheng uses an independence sampler to sample from the posterior distribution. Gibbs sampling can also be used for most problems and this has been found to give better results.

4.3 Issues

4.3.1 Label-Switching

The components in a mixture model are non-identifiable, which means that the posterior distribution will have $k!$ symmetric modes for a model with k components. If the separation between component means is small, there could be interference from one or more of the other $k! - 1$ symmetric modes. We assume that we identify only one of these modes in the optimization, and that the importance sampling only samples from close to this mode. For the datasets analysed in this thesis, we assume in addition that the $k!$ modes are well separated in the best-fit models.

The problem of label-switching, as the phenomenon described above is referred to, is not necessarily important in mixture analysis. In situations where the data are known to be made up of finite mixtures, it is important to identify and label the components correctly, as the component means and variances have a physical meaning. On the other hand, where finite mixtures are being used simply as semi-parametric density estimates, the component parameters are of less interest and the analysis should focus on quantities such as the probability density, which will be invariant to label-switches. The aim of our analysis is somewhere between these two extremes, and is probably best described as "investigating heterogeneity", a term used by Richardson and Green to describe their own work [83]. We wish to determine the posterior distribution for the number of components in the mixture model, and we have less interest in the accurate determination of component means and variances. The label-switching problem is therefore not of paramount interest to us, as the results of the importance sampling for the number of components will be invariant to label-switching.

4.3.2 Bayesian versus Frequentist Argument

The likelihood of the parameters of mixtures of normal distributions with different variances has several problems. As a component variance tends to the boundary level of zero, the likelihood tends to an infinite spike. This corresponds to the situation where one component is fitting to just one data point, and the component tends to a delta function centred on that point. The likelihood function also suffers from the existence of local maxima, which can create some computational difficulties when trying to estimate the number of components. The effect of the prior can alleviate these difficulties in the Bayesian analysis, although the posterior distribution can also suffer from local maxima. For example, the prior distribution for the variance of the sampling will usually associate a very low or zero prior probability with the variance of the component tending to zero. The prior probability can be seen as introducing a smoothing effect, resulting in a posterior distribution that is easier to deal with than the corresponding likelihood distribution.

There are ways of overcoming the problem of zero variance in maximum likelihood estimation. These usually rely on restricting the variance, and it could be argued that these use much the same methods as incorporating a prior distribution but in a less transparent way.

4.3.3 Sensitivity to the Prior Distribution

Both Richardson and Green [84] and Stephens [97] comment on the sensitivity of the posterior distribution for the number of components in a finite mixture model

(k) to the prior distributions for other parameters in the mixture model.

Richardson and Green found that when the variance of the prior distribution is small, representing a strong belief that the means are at the mean of the prior distribution, models with a small number of components are favoured. As the variance is increased, to represent vaguer prior knowledge of the position of the component means, initially more components are fitted with means spread across the range of the data, but continuing to increase the variance will eventually favour fitting fewer components. In the limit of the variance tending to ∞, the distribution of k becomes independent of the data (according to Stephens [97]) and this heavily favours a one component model.

We investigate this dependence further by considering a very simple prior distribution for the means, variances and weights of a normal mixture model. We assume that the means follow a uniform distribution with minimum at $\chi - R/2$ and maximum at $\chi + R/2$ and that the variances also follow a uniform with lower and upper values at 0 and T, where χ, R and T are hyperparameters to be set by the user. The values of the parameters R and T will determine how vague the prior distributions for the component means and variances are, and χ sets the location of the mean. The weights are assumed to follow a Dirichlet distribution with parameter δ set equal to one. The prior probability of choosing a model with k components is assumed to be $1/k_{max}$ for k up to k_{max}. Therefore, the prior probability of a model with k components with parameters θ is

$$\pi(k,\theta) \propto \frac{\Gamma(k)}{(RT)^k}, \tag{4.4}$$

where $0 < \sigma < T, -R/2 < \mu - \chi < R/2, 0 < k \leq k_{max}$. Writing $RT = S$, and expanding the gamma function we obtain an expression for the prior probability in terms of k and S,

$$\pi(k,\theta) = \frac{(k-1)!}{S^k}. \tag{4.5}$$

We use Maple to plot this function for different values of S in Figure 4.1. Higher values of S imply vaguer prior knowledge. As can be seen in the figure, the prior distribution has a minimum for some value of k. As S increases and the prior knowledge becomes vaguer, the minimum occurs at higher values of k. Thus the scale parameters used in the prior distributions for the component means and variances impact on the prior distribution for k. In general, intermediate values of k seem to have a lower probability than very low or very high values of k.

A 3-d plot showing the variation of $\pi(k, \theta)$ with k and S is shown in Figure 4.2. This is not very clear, which is why the set of 6 graphs for different values of S were produced. The graphs in Figure 4.1 all use an integer value for S and it is interesting to note, that $\pi(k, \theta)$ is equal at $k = S$ and $k = S + 1$, with the minimum of $\pi(k, \theta)$ with respect to k always lying between S and $S + 1$. In fact, this is also true for non-integer S as is easily shown by considering the definition of $\pi(k, \theta)$ in Equation 4.4.

So what does it mean geometrically? As we are using uniform distributions, the prior distribution $\pi(k, \theta)$ gives an indication of the volume of our available parameter space. The weights are restricted to always sum to one, therefore the volume of our parameter space is defined by a simplex in $k - 1$ dimensions (parameter space of the weights) multiplied by a cuboid of side S in k dimensions (parameter space of combined means and variances). As k increases, the volume of the cuboid increases for $S > 1$, decreases for $S < 1$ and remains constant for $S = 1$; the volume of the simplex always decreases. The maximum volume of parameter space, and so the minimum value for $\pi(k, \theta)$, will therefore occur at different values for different S.

We can obtain an approximate expression for the value of k at which the prior distribution has a minimum by using Stirling's approximation to the factorial function,

$$n! \simeq n^n e^{-n} \sqrt{2\pi n}, n \to \infty. \tag{4.6}$$

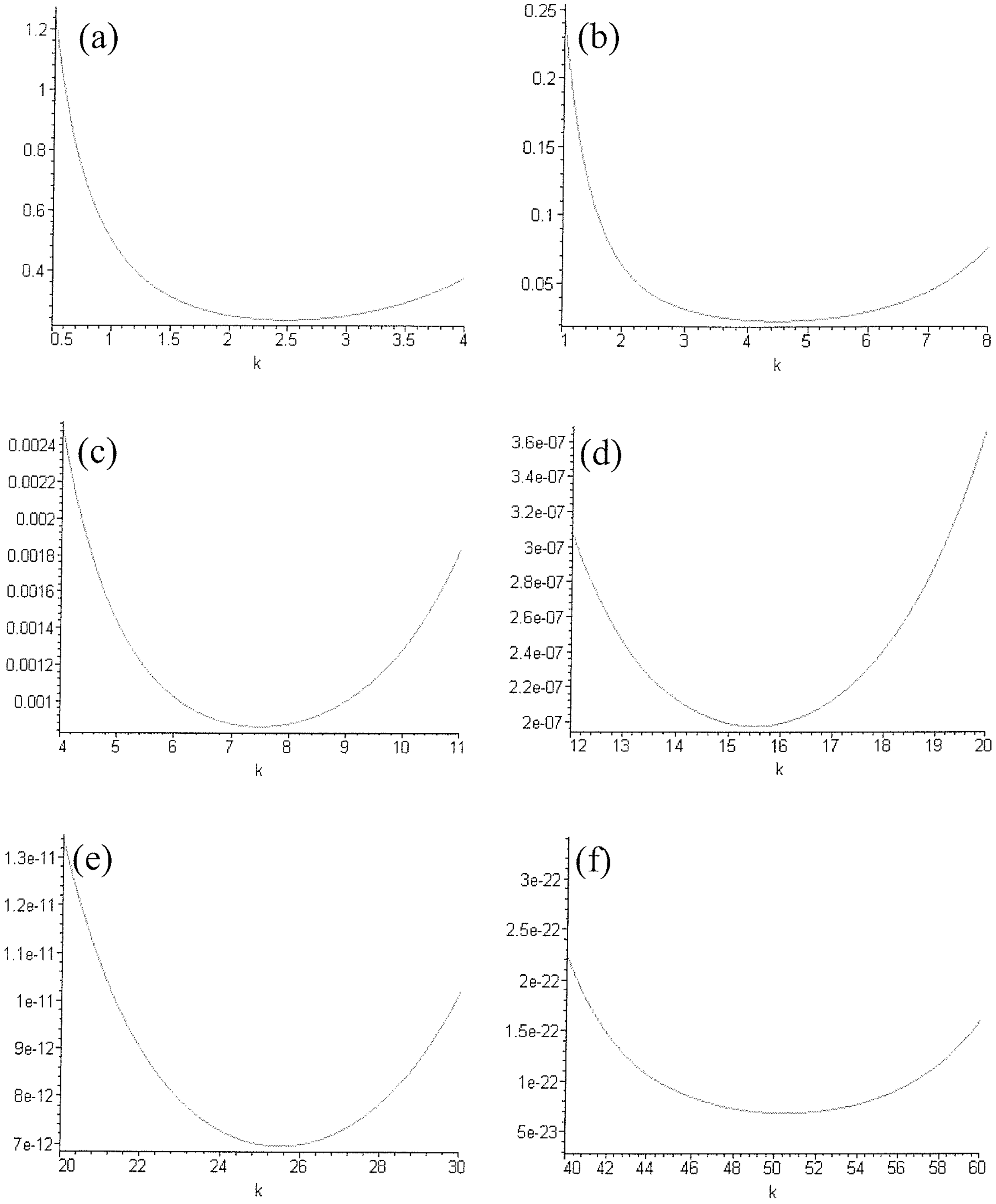

Figure 4.1: The variation of the prior distribution with k for (a)$S = 2$, (b) $S = 4$, (c) $S = 7$, (d)$S = 15$, (e)$S = 25$, (f) $S = 50$.

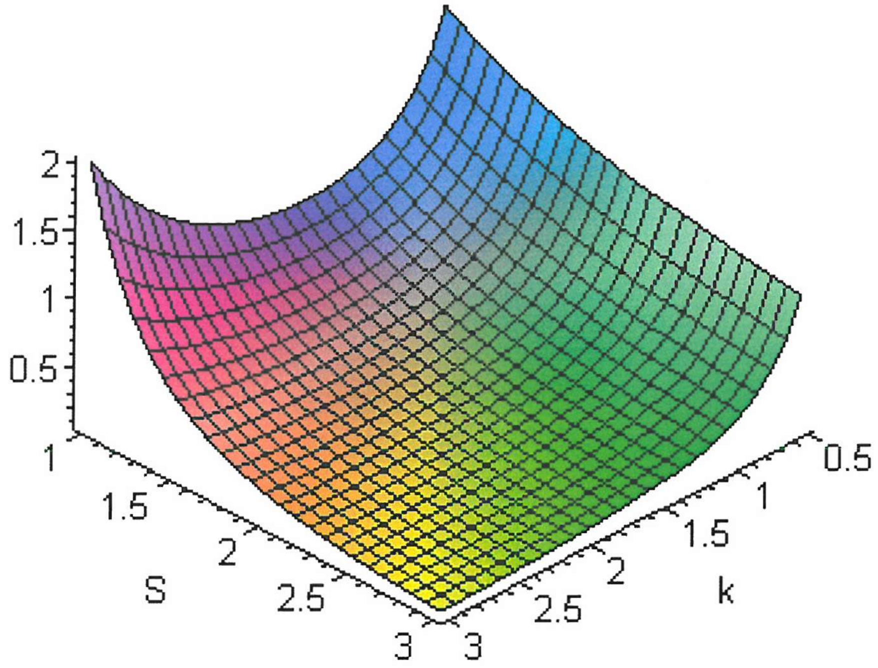

Figure 4.2: The variation of the prior distribution with k and S.

Substituting this into Equation 4.5,

$$\pi(k,\theta) = \frac{(k-1)^{k-1}e^{-(k-1)}\sqrt{2\pi(k-1)}}{S^k}. \tag{4.7}$$

Differentiating and setting the differential equal to zero, we find that k has a minimum at

$$k_{min} = \frac{-1 + LambertW\left(-\frac{1}{2S}\right)}{LambertW\left(-\frac{1}{2S}\right)}, \tag{4.8}$$

where $LambertW(x)$ is defined such that

$$LambertW(x)\exp(LambertW(x)) = x. \tag{4.9}$$

Therefore, the joint prior distribution for k, the number of components in a normal mixture model, is dependent on the prior distribution for the component means and variances.

Rescaling the data could result in a change in the prior distribution, and so an increase or a reduction in S, with no increase in the vagueness of the priors for the means and variances. This could then change the joint prior distribution for the number of components. For example, if the data, and corresponding prior distributions, were rescaled so that $S < 1$, the prior distribution would have a minimum at $k = 1$, rather than at the integer value of k between unscaled values of S and $S + 1$. This results from the interplay between the volume of the simplex defining the parameter space of the weights and the cuboid defining the parameter space of the means and variances.

Although this analysis has been conducted for uniform distributions for the means and variances, it is suspected that the results will be similar for other distributions as the cause of the variation is the interplay between the increase in the volume of the parameter space of the component means and variances and the decrease in the volume of the parameter space of the component weights.

4.4 Prior Distribution

We use as a prior distribution

$$\begin{array}{lcll} y_i & \sim & N(\mu_{z_i}, \sigma^2_{z_i}) & \\ P(z_i = j) & = & a_j & \\ \mathbf{a} & \sim & Di(\delta) & \\ \mu_j & \sim & N(m, s^2) & \\ \tau_j^2 & \sim & Ga(\alpha, \beta) & \\ Pr(k) & = & 1/k_{max} & k = 1, 2, \ldots, k_{max}, \end{array} \tag{4.10}$$

where $Di(\delta)$ denotes a Dirichlet distribution with parameter vector δ and $\tau_j^2 = 1/\sigma_j^2$. The parameters $m, s^2, \alpha, \beta, \delta$ are chosen in advance to give an uninformative prior. We set m equal to the mean of the data, and s^2 equal to the sample variance multiplied by a stretch factor, that we set equal to one thousand. We set α, the shape parameter in the prior distribution for the inverse variance, equal to one and

β such that the mean of the distribution is equal to one over the variance of the data multiplied by the stretch factor. The parameter in the Dirichlet distribution, δ is set to one, placing equal prior probability on values of the component weights, and k_{max} is chosen to be 10 in all of the examples considered.

The prior distribution given in Equation 4.10 is of a simpler form than that used by Richardson and Green [84] and described in Section 4.2.1 as it includes one less hierarchical layer. This has no effect on the prior distributions for the component means and weightings nor on that for the number of components in the mixture. It will result in a slightly more restrictive prior distribution for the component variances.

We investigated the effect of changing the scale parameters in the prior distributions for the component means and variances, considering three different datasets. Results suggested that for this form of the prior distribution, changing the scale parameters (and so altering some of the hyperparameters) had little effect on the posterior distribution for k, with the optimal number of components changing by at most one, but in the main staying the same.

Choosing a higher value for δ could have a more significant effect on the optimal number of components, as this parameter affects the the size of the component weights. Setting δ to one allows the model to choose zero values for some of the weights. This means that a model that appears to have many components could actually be a model with only a few components, as some of the component weights may be zero or very close to zero. As δ is increased beyond one, the dirichlet distribution favours larger weights. It could be argued that the best value for this parameter is slightly higher than one, and this should be the subject of further investigation.

4.5 Importance Sampling

Obtaining good convergence using importance sampling requires good knowledge about the function being sampled and the methodology we propose here incorporates an initial step in which we investigate the form of the posterior distribution by finding the modes of the posterior distributions of each of the models in terms of the component means, variances and weights, and the covariance structure at the modes. We then use this information to set our candidate distribution for the importance sampling.

4.5.1 Optimization to Find the Mode of the Posterior Distribution

Three methods were evaluated for finding the maxima of the posterior distributions: the Nelder Mead optimization routine [74], conjugate gradient optimization [37] and the EM algorithm [39]. Nelder Mead was chosen for the final methodology because it was found be more robust than the EM algorithm and to give better results than the conjugate gradient optimization. We discuss the implementations of the different optimization routines further below.

With each of the optimization methods tried, starting parameters for the models with $k \leq k_{max} - 1$ are determined from the best estimates for the model with $k + 1$ components, usually by combining two of the original components. Unless otherwise stated, two components are combined to give a new component that has a mean equal to the weighted average of the means of the original components, a variance equal to a weighted average of the variances of the original components and a weight equal to the sum of the two original weights. All other parameter values remain the same. Five different methods of choosing which components to combine are considered.

1. Combine the two components that give the highest posterior probability, before optimization.

2. The collapsing method introduced by Sahu and Cheng [91], in which the closest two components, measured in terms of the difference in the means, are combined.

3. Any components with very large variances have their weights added to the component with the closest mean. If no large variance components exist, the optimization is started from two points: combining the components with the closest means, and combining the component with the smallest central weight with its nearest component, measured in terms of the difference in the mean. The solution that has the higher posterior probability following optimization is retained.

4. As the previous method but instead of trying just two combinations of components in the case of there being no large variance components, we try all combinations of adjacent components, where components are adjacent if their means are adjacent. Select the solution that has the highest posterior probability following optimization.

5. Run the optimization for all possible combinations of adjacent components, where components are combined as described in method two. Select the solution that has the highest posterior probability following optimization.

The five different methods of combining components have been tried on a number of different data sets. We find that method three works well for most examples, performing a much smaller number of iterations than methods four and five and finding optima that are either similar or better than these more thorough methods. Methods one and two involve the smallest number of iterations but the optima that they find are generally not as good as those found by the other three methods.

All methods have problems fitting skewed data, with the more thorough search methods performing better.

Nelder Mead

In the final methodology we used the Nelder Mead optimization routine [74]. Moves were chosen by the Nelder Mead routine ignoring the positivity constraints on the τ_i and w_i and the constraint that the weights summed to one. We dealt with the positivity constraints by imposing a very large penalty on transgressions into infeasible areas. We ensured the sum of the weights remained equal to one by renormalising the weights at each new point.

The Nelder Mead is a local optimization routine. We tested how local the optimization was by running it from a number of different starting points. Only the start point for the model with k_{max} components was changed. In the test, we used method five for combining components for $k < k_{max}$, as described above. Four different sets of initial conditions were considered:

1. Standard initial conditions: put the data into non-decreasing order of means, and split into k_{max} groups. These are assumed to be a very rough approximation to the k_{max} components, and we take the initial component means to be the group means and the initial component weights to be $1/k_{max}$. The initial group variances are assumed to be equal and are set to be the data variance divided by k_{max}^2.

2. Initial component variances and weights are set using the standard initial conditions; the initial component means are set to be equal, at the mean of the whole dataset.

3. Initial component means and variances are set using the standard initial conditions; the initial component weights are set so that there is one very low

weight and $k_{max} - 1$ high weights.

4. Initial component means and weights set using the standard initial condition; the initial component variances are set to be equal to the data variance.

The optimal solutions found by the Nelder Mead routine were very similar for each of the start points, with the main differences being between the models with k_{max} components. When applying the routine to the galaxy dataset [87], we found that the sum of the squared differences between the component means of the different solutions and the solution found using the standard initial conditions was at most 1.37 (for scenario 4), when considering all possible models and 4.63 x 10^{-5} when considering only the first $k_{max} - 1$ models. Similarly, when looking at the variances, the sum of the squared difference between the $\tau_i = 1/\sigma_i$ was 2150 (for scenario 4) when considering all models and 0.0311 (for scenario 2) when considering only the first $k_{max} - 1$ models. For the weights, the sum of the squared difference was at most 2.18 x 10^{-3} (for scenario 4) when considering all models and 2.34 x 10^{-6} (for scenario 4) when looking at only the first $k_{max} - 1$ models. This suggests that the Nelder Mead is performing a sufficiently wide-ranging search for the first $k_{max} - 1$ models but that the initial conditions have a greater effect on the optimum reached for the model with k_{max} components. The optimum is likely to not be as well-defined with higher numbers of components and so any optimization routine would have problems finding the global optimum.

We calculated Anderson-Darling statistics for each of the optimal sets of parameters found by the Nelder Mead routine for the four different starting points and for each value of k. These are given in Table 4.1. Critical values are not available but the magnitude of the statistics allows an informal comparison of fits across scenarios. We find that there is only a big difference in these statistics for the model with k_{max} components, for which scenario three has an Anderson-Darling statistic that is double that found for the other four scenarios.

We find that a total of approximately 25,000 runs (and no more than 30,000) are required to find the optimal solutions for all of the different models considered using the Nelder Mead routine.

Conjugate Gradient Optimization

The BFGS (Broyden-Fletcher-Goldfarb-Shanno) method of gradient-based optimization introduced by Davidon [37], was tried initially, as the gradient of the posterior distribution can be calculated. A good review of conjugate gradient methods is given in Chapter Two of [17]. Transformed parameters were used to ensure that positivity constraints on the inverse variances τ_i and the weights a_i were always satisfied, with the weights also always summing to one, such that

$$\begin{aligned} \tau_i &= e^{d_i} \qquad i = 1, \ldots, k \\ a_i &= \frac{e^{b_i}}{\sum_{j=1}^{k} e^{b_i}} \qquad i = 1, \ldots, k \end{aligned} \tag{4.11}$$

where d and b are optimized and can vary between $-\infty$ and ∞ with the constraints always being met.

We used the algorithm to minimise minus the posterior and minus the log of the posterior. When minimising the posterior, we found that the algorithm did not

k	Scenario 1	Scenario 2	Scenario 3	Scenario 4
1	3.86	3.85	3.85	3.85
2	1.98	1.99	1.98	1.98
3	0.521	0.521	0.519	0.520
4	0.136	0.136	0.135	0.136
5	0.101	0.101	0.101	0.101
6	0.0750	0.0749	0.135	0.0786

Table 4.1: Anderson-Darling statistics for optimal models found using the Nelder Mead routine, with different initial solutions.

move far from its starting point, as the gradients calculated at the initial points were very small. When instead minimising the negative log of the posterior, the algorithm frequently moved to areas of parameter space associated with a very low posterior probability. The errors causing this originated in the routine updating H, the estimate of the covariance matrix and we suspect were due to the surface being a long way from quadratic.

EM Algorithm

We also considered using the EM algorithm, introduced by Dempster et al [39] to find the mode of the posterior. Traditionally, the EM algorithm has been used to find the maximum likelihood solution, but it can be easily adapted to instead find the maximum of the posterior distribution. The basic idea behind the EM algorithm is to augment the original data with latent data in order to obtain a more tractable expression for the likelihood. When applied to mixture models, the latent variables are assumed to be the components that data points have been generated from. A good introduction to the EM algorithm and its application to mixture models is given in [12]. We describe its application to this problem below.

Using the prior distributions given in Section 4.4, the log of the prior probability can be incorporated into the expression for Q given in Section 3 of [12] to give

$$Q = \sum_{\mathbf{y}\in\Upsilon} \{\log[\pi(\theta)] + \log[L(\theta|\Xi, \mathbf{y})]\}\, p(\mathbf{y}|\Xi, \Theta), \tag{4.12}$$

where Ξ is the data, θ is the vector of parameters and $\mathbf{y}$ is the unknown data, which tells us which component each of the data points was generated from. The first term is the prior distribution and the second the likelihood.

Using this expression for Q, we find that the updated μ_l, τ_l and a_l, $l = 1, \dots, k$

in the g^{th} iteration should be

$$\begin{aligned} \mu_l^g &= \frac{\xi\kappa + \sum_{i=1}^N \tau_l x_i p(l|x_i, \Theta^{g-1})}{\kappa + \sum_{i=1}^N \tau_l^{g-1} p(l|x_i, \Theta^{g-1})} \\ \tau_l^g &= \frac{2(\alpha - 1) + \sum_{i=1}^N p(l|x_i, \Theta^{g-1})}{2/\beta + \sum_{i=1}^N (x_i - \mu_l^{g-1})^2 p(l|x_i, \Theta^{g-1})} \\ a_l^g &= \frac{\delta_l - 1 + \sum_{i=1}^N p(l|x_i, \Theta^{g-1})}{N - k + \sum_{l=1}^k \delta_l}. \end{aligned} \tag{4.13}$$

The algorithm is run until convergence is reached, where convergence is measured by the similarity in the Q values between subsequent iterations.

We found that the EM algorithm did not converge to as good an optimum and was more sensitive to the starting solution than the Nelder Mead. It also did not converge for some initial solutions. Often this occurred when a large number of components were being fitted to a dataset for which only a small number of components might be required, and took the form of one of the τ_l tending to zero for a component with a very small weight a_l. The sensitivity of the limiting solution to the initial solution and the convergence to local maxima or saddle points are drawbacks that have been discussed elsewhere in the literature, e.g. in [41].

The EM algorithm is much quicker than the Nelder Mead algorithm, performing about 100 iterations per model compared with a few thousand for the Nelder Mead. We have not pursued this method further but there is scope for more research in this area, possibly considering an adaptation of a more sophisticated version of the EM algorithm, such as that put forward in [2], or the use of a stochastic EM algorithm, which has previously been applied to mixture models by Diebolt and Robert [42]. If a sufficiently good optimum could be obtained without a significant increase in the number of runs required, this method could out-perform the Nelder Mead.

4.5.2 Estimation of the Covariance Matrix

Having found the modes of the posterior distributions for each model, we can then estimate the covariance matrix for each of the models using the information matrix at the maxima of the model posteriors.

Estimating the covariance matrices for a mixture model of k components is non-trivial because the weights, w_i must sum to one. Let the vector of weights be

$$\mathbf{w} = (w_1, ..., w_k)^T \tag{4.14}$$

then

$$\sum_{i=1}^{k} w_i = 1 \tag{4.15}$$

and

$$w_i \geq 0, \;\; \text{all } i. \tag{4.16}$$

Let all the other parameters be written as $\alpha = (\alpha_1, \alpha_2, ..., \alpha_m)^T$ and let

$$L = L(\alpha, \mathbf{w}) \tag{4.17}$$

be the log posterior probability . Suppose its maximum occurs at $(\hat{\alpha}, \hat{\mathbf{w}})$, where this optimum has been obtained subject to $\sum_{i=1}^{k} w_i = 1$. Let the negative Hessian be

$$\mathbf{H} = \begin{pmatrix} \mathbf{H}_{\alpha,\alpha} & \mathbf{H}_{\alpha,\mathbf{w}} \\ \mathbf{H}^T_{\alpha,\mathbf{w}} & \mathbf{H}_{\mathbf{w},\mathbf{w}} \end{pmatrix} \tag{4.18}$$

with, in particular,

$$\mathbf{H}_{\mathbf{w},\mathbf{w}}(\hat{\mathbf{w}}) = -\left.\frac{\partial^2 L(\mathbf{w})}{\partial \mathbf{w}^2}\right|_{\mathbf{w}=\hat{\mathbf{w}}} . \tag{4.19}$$

Here, the partial derivatives of $\mathbf{H}$ are obtained ignoring the restriction on the weights that $\sum_{i=1}^{k} w_i = 1$.

Suppose that we now replace w_i by the parameters θ_i where

$$w_i = \theta_i + k^{-1}\left(1 - \sum_{j=1}^{k} \theta_j\right), \; i = 1, ..., k. \tag{4.20}$$

This ensures that

$$\sum_{i=1}^{k} w_i = 1. \tag{4.21}$$

The Jacobian matrix of the transformation is

$$\mathbf{J} = \frac{\partial \mathbf{w}}{\partial \theta} = (\mathbf{I}_k - k^{-1}\mathbf{1}_k\mathbf{1}_k^T), \tag{4.22}$$

where $\mathbf{I}_k$ is the $k-$component identity matrix and $\mathbf{1}_k= (1, 1, ..., 1)^T$ is the k-component vector with unit entries. The log posterior density in terms of this parameterization, $L = L(\alpha, \theta)$, where $\theta = (\theta_1, ..., \theta_k)^T$, has negative Hessian

$$\mathbf{A}(\alpha, \theta) = \mathbf{A}(\alpha, \mathbf{w}) = \begin{pmatrix} \mathbf{H}_{\alpha,\alpha} & \mathbf{H}_{\alpha,\mathbf{w}}\mathbf{J}^T \\ \mathbf{J}\mathbf{H}_{\alpha,\mathbf{w}}^T & \mathbf{J}\mathbf{H}_{\mathbf{w},\mathbf{w}}\mathbf{J}^T \end{pmatrix}, \tag{4.23}$$

which we write as $\mathbf{A}$ from now on, and gives the joint distributional behaviour of $\hat{\alpha}$, $\hat{\mathbf{w}}$, subject to $\sum_{i=1}^{k} w_i = 1$,.

More precisely, the inverse of $\mathbf{A}$ gives the covariance of $(\hat{\alpha}, \hat{\mathbf{w}})$ subject to $\sum_{i=1}^{k} w_i = 1$. Clearly therefore $\mathbf{A}$ must be singular, and indeed the sub matrix $\mathbf{J}\mathbf{H}_{\mathbf{w},\mathbf{w}}\mathbf{J}^T$ is singular, as $\det(\mathbf{J}) = 0$.

Thus $\mathbf{A}$ does not have a full inverse. However it does have a generalised inverse, $\mathbf{G}$, which by definition will satisfy

$$\mathbf{AGA} = \mathbf{A}. \tag{4.24}$$

To find the generalised inverse, we begin by assuming that $\mathbf{H}$ and $\mathbf{A}$ are evaluated at $\alpha = \hat{\alpha}$ and $\mathbf{w} = \hat{\mathbf{w}}$. We then let $\mathbf{P}$ be the orthogonal matrix formed from the eigenvectors of $\mathbf{A}$ (so that $\mathbf{P}\mathbf{P}^T= \mathbf{P}^T\mathbf{P} = \mathbf{I}_{m+k}$). Then

$$\mathbf{P}^T\mathbf{A}\mathbf{P} = \mathbf{D}, \tag{4.25}$$

where $\mathbf{D}$ is the diagonal matrix of eigenvalues corresponding to the eigenvectors forming $\mathbf{P}$. Then it is known (see for example [95]) that a generalised inverse of $\mathbf{A}$ is

$$\mathbf{G} = \mathbf{P}\mathbf{M}\mathbf{P}^T \tag{4.26}$$

where $\mathbf{M}$ is the diagonal matrix whose non-zero diagonal elements are the reciprocals of the non-zero diagonal elements of $\mathbf{D}$. For example, if

$$\mathbf{D} = \begin{array}{c} \left(\begin{array}{cccc} \mathbf{D}_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \mathbf{D}_2 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right) \end{array} \begin{array}{l} \}m_1 \\ \}m - m_1 \\ \}k_1 \\ \}k - k_1 \end{array} \tag{4.27}$$

where $\mathbf{D}_1$ and $\mathbf{D}_2$ are non-singular diagonal matrices, then

$$\mathbf{M} = \left(\begin{array}{cccc} \mathbf{D}_1^{-1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \mathbf{D}_2^{-1} & 0 \\ 0 & 0 & 0 & 0 \end{array}\right) \begin{array}{l} \}m_1 \\ \}m - m_1 \\ \}k_1 \\ \}k - k_1 \end{array} \tag{4.28}$$

Note that $m_1 \leq m$ with equality allowed, but the last row and column of $\mathbf{M}$ are zeros, as will be shown in the lemma below, therefore

$$k - k_1 \geq 1. \tag{4.29}$$

From the definition of $\mathbf{M}$ it easily follows that

$$\mathbf{DMD} = \mathbf{D} \tag{4.30}$$

and rearranging Equation (4.25),

$$\mathbf{A} = \mathbf{PDP}^T. \tag{4.31}$$

Using these two expressions, we can show that $\mathbf{G}$ satisfies Equation 4.24 and is thus the generalised inverse of $\mathbf{A}$

$$\mathbf{AGA} = \mathbf{PDP}^T\mathbf{PMP}^T\mathbf{PDP}^T = \mathbf{PDMDP}^T = \mathbf{PDP}^T = \mathbf{A}. \tag{4.32}$$

4.5.3 Candidate Distribution

When the modes and covariance matrices for the posterior distributions have been found, we use a multivariate generalisation of a student t-distribution (defined in Section 2.4.3) as a sampler for component weights, means and variances. We assume that we are sampling within only one of the $k!$ identical simlexes of parameter space, and so the probability of sampling each of the points must be scaled by a factor of $1/k!$. A uniform distribution is used as a sampler for k, the number of components in the model.

The algorithm for this method is then

1. Sample k with probability $1/k_{max}$ of sampling each of the $k = 1, 2, \ldots, k_{max}$.
2. Sample the mixture model parameters from a multivariate t-distribution with mean given by the mode of the posterior for the model with k components and covariance matrix $\mathbf{A}$, using the method described in Section 4.5.2.
3. Calculate the posterior probability divided by sampler probability (sampling ratio).
4. Output parameters and the sampling ratio to the worksheet.
5. Repeat N times, where N is the number of samples required.

A weighted frequency plot using the sampling ratios as the weights will then give the posterior probability density function, if the sampling has converged.

4.5.4 Generating Parameters Using the Generalised Inverse

If $\mathbf{H}$ is positive semidefinite then so is $\mathbf{A}$ and all its eigenvalues are non-negative, meaning that the diagonal elements of $\mathbf{M}$ are also non-negative. We can therefore write $\mathbf{M} = \mathbf{M}^{\frac{1}{2}}\mathbf{M}^{\frac{1}{2}}$ and define a matrix $\mathbf{L}$

$$\mathbf{L} = \mathbf{P}\mathbf{M}^{\frac{1}{2}}. \tag{4.33}$$

such that

$$\mathbf{G} = \mathbf{P}\mathbf{M}^{\frac{1}{2}}\mathbf{M}^{\frac{1}{2}}\mathbf{P}^{T}= \mathbf{L}\mathbf{L}^{T}. \tag{4.34}$$

Lemma

(i) The vector

$$\mathbf{p}_0 = \begin{pmatrix} \mathbf{0}_m \\ \mathbf{1}_k \end{pmatrix} \begin{matrix} \}m \\ \}k \end{matrix} \tag{4.35}$$

is an eigenvector of $\mathbf{A}$ with eigenvalue 0.

(ii) All other eigenvectors of $\mathbf{A}$, which we write as

$$\mathbf{p}_j = \begin{pmatrix} \mathbf{p}_j^{\alpha} \\ \mathbf{p}_j^{\mathrm{w}} \end{pmatrix} \begin{matrix} \}m \\ \}k \end{matrix} \quad , \ j = 1, 2, ..., \nu \tag{4.36}$$

where $\nu = m + k - 1$, satisfy

$$\mathbf{1}_k^T \mathbf{p}_j^{\mathrm{w}} = 0. \tag{4.37}$$

Proof The matrix $\mathbf{A}$ is singular and so, by definition, has at least one eigenvalue that is equal to zero. Therefore, in order to prove that part (i) of the lemma is true, we simply need to show that

$$\mathbf{A}\mathbf{p}_0 = \mathbf{0}. \tag{4.38}$$

Using the expansion given in Equation 4.23, we can rewrite this condition as

$$\mathbf{A}\mathbf{p}_0 = \begin{pmatrix} \mathbf{H}_{\alpha,w}\mathbf{J}^T\mathbf{1}_k \\ \mathbf{J}\mathbf{H}_{w,w}\mathbf{J}^T\mathbf{1}_k \end{pmatrix}. \tag{4.39}$$

The expression for the Jacobian $\mathbf{J}$ is given in Equation 4.22 and it is easy to show that $\mathbf{J}^T\mathbf{1}_k$ is equal to $\mathbf{0}_k$, the k-dimensional column vector of zeros. Hence, Equation 4.38 holds and part (i) of the lemma is proved.

To prove part (ii) we simply note that the matrix $\mathbf{A}$ is symmetric and therefore has orthogonal eigenvectors. Thus $\mathbf{p}_j\mathbf{p}_0 = \mathbf{0}_{\nu+1}$, $j = 1, 2, \ldots, \nu$. As the upper m components of p_0 are zero, the orthogonality condition reduces to $\mathbf{p}_j\mathbf{1}_k = \mathbf{0}_k$, hence proving part (ii) of the lemma. □

We now put $\mathbf{p}_0$ in the last column of $\mathbf{P}$, and write

$$\mathbf{P} = \begin{pmatrix} \mathbf{P}_1 & \mathbf{p}_0 \end{pmatrix} \tag{4.40}$$

with

$$\mathbf{P}_1 = \begin{pmatrix} \mathbf{p}_1 & \mathbf{p}_2 & \ldots & \mathbf{p}_\nu \end{pmatrix} \tag{4.41}$$

the matrix comprising the other eigenvectors. From the Lemma we can write

$$\mathbf{M}^{\frac{1}{2}} = \begin{pmatrix} \Lambda & \mathbf{0}_\nu \\ \mathbf{0}_\nu^T & 0 \end{pmatrix}, \tag{4.42}$$

where $\mathbf{0}_d$ is the $d-$dimensional column vector of zeros and

$$\Lambda = \begin{pmatrix} \mathbf{D}_\alpha^{-\frac{1}{2}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \mathbf{D}_\mathbf{w}^{-\frac{1}{2}} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{array}{l} \}m_1 \\ \}m - m_1 \\ \}k_1 \\ \}k - k_1 - 1 \end{array} \tag{4.43}$$

These new expressions for $\mathbf{P}$ and $\mathbf{M}^{1/2}$ can then be substituted into Equation 4.33 to yield

$$\begin{aligned} L &= \begin{pmatrix} \mathbf{P}_1 & \mathbf{p}_0 \end{pmatrix} \begin{pmatrix} \Lambda & \mathbf{0}_\nu \\ \mathbf{0}_\nu^T & 0 \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{P}_1\Lambda & \mathbf{0}_{m+k} \end{pmatrix}. \end{aligned} \tag{4.44}$$

We now consider how the above results can be applied to the generation of variates $\mathbf{x}$ from the singular multivariate t-distribution

$$\mathbf{x} = \begin{pmatrix} \alpha \\ \mathbf{w} \end{pmatrix} \sim ST\left(\begin{pmatrix} \hat{\alpha} \\ \hat{\mathbf{w}} \end{pmatrix}, \mathbf{LL}^T \right), \tag{4.45}$$

where ST indicates a singular multivariate generalisation of the student t-distribution as defined in Section 2.4.3. A variate from this distribution can be generated using

$$\mathrm{x} = \begin{pmatrix} \hat{\alpha} \\ \hat{\mathbf{w}} \end{pmatrix} + \mathbf{P}_1 \Lambda \mathbf{z}_\nu, \tag{4.46}$$

where $\mathbf{z}_\nu$ is a vector of standard t-variates. These can have arbitrary degrees of freedom δ, and are derived from non-standard t-variates by dividing through by $\sqrt{\nu/(\nu-2)}$, the variance of the student t-distribution. The covariance of the x generated in this way is then

$$\mathrm{Var}(\mathrm{x}) = \mathrm{E}(\mathbf{P}_1 \Lambda \mathbf{z}_\nu \mathbf{z}_\nu^T \Lambda \mathbf{P}_1^T) = \mathbf{P}_1 \Lambda \Lambda \mathbf{P}_1^T = \mathrm{G}. \tag{4.47}$$

Moreover, using the result of Equation 4.37,

$$(\mathbf{0}_m^T, \mathbf{1}_k^T)\mathbf{P}_1 = \mathbf{0}_\nu^T, \tag{4.48}$$

and the sum of the component weights is given by

$$\begin{aligned} \sum_{i=1}^{k} w_i &= (\mathbf{0}_m^T, \mathbf{1}_k^T) \begin{pmatrix} \alpha \\ \mathbf{w} \end{pmatrix} \\ &= (\mathbf{0}_m^T, \mathbf{1}_k^T) \begin{pmatrix} \hat{\alpha} \\ \hat{\mathbf{w}} \end{pmatrix} + (\mathbf{0}_m^T, \mathbf{1}_k^T)\mathbf{P}_1 \Lambda \mathbf{z}_\nu \\ &= (\mathbf{0}_m^T, \mathbf{1}_k^T) \begin{pmatrix} \hat{\alpha} \\ \hat{\mathbf{w}} \end{pmatrix} + \mathbf{0}_\nu^T \Lambda \mathbf{z}_\nu \\ &= \sum_{i=1}^{k} \hat{w}_i = 1. \end{aligned} \tag{4.49}$$

Thus under this sampling we are restricted to the simplex $\sum_{i=1}^{k} w_i = 1$.

4.5.5 Restricting the Range of the Weights to (0,1)

The above transformation needs an adjustment to ensure that in addition to summing to unity, the final weights each lie in the unit interval. We write

$$\begin{aligned} \mathbf{x} &= \begin{pmatrix} \hat{\alpha} \\ \hat{\mathbf{w}} \end{pmatrix} + \mathbf{P}_1\Lambda\mathbf{z}_\nu \\ &= \begin{pmatrix} \hat{\alpha} \\ \hat{\mathbf{w}} \end{pmatrix} + \begin{pmatrix} \mathbf{Q}_\alpha \\ \mathbf{Q}_\mathbf{w} \end{pmatrix} \mathbf{z}_\nu, \end{aligned} \tag{4.50}$$

where $\mathbf{Q}_\alpha$ is the first m rows of $\mathbf{P}_1\Lambda$ and $\mathbf{Q}_\mathbf{w}$ are the remaining k rows, and let

$$\xi = \xi_{k\times 1} = \mathbf{Q}_\mathbf{w}\mathbf{z}_\nu. \tag{4.51}$$

We define a new vector $\mathbf{y}$ such that

$$y_i = \frac{\hat{w}_i \exp(a_i\xi_i)}{\hat{w}_i \exp(a_i\xi_i) + 1 - \hat{w}_i} \quad i = 1, 2, ..., k, \tag{4.52}$$

where

$$a_i = \frac{1}{\hat{w}_i(1-\hat{w}_i)}. \tag{4.53}$$

The effect of this transform is to ensure that the y_i are all positive and less than one. For the final weights w_i we use the transform

$$w_i = \frac{y_i}{y_1 + y_2 + ... + y_k} \quad i = 1, 2, ..., k \tag{4.54}$$

to ensure that the y_i also sum to one.

The vector of weights, $\mathbf{w} = (w_1, w_2, ..., w_k)$ clearly has a singular distribution. Let ω be the $(k-1)$ dimensional vector formed from the first $(k-1)$ components of $\mathbf{w}$ and write

$$\phi = \begin{pmatrix} \alpha \\ \omega \end{pmatrix}. \tag{4.55}$$

In terms of the importance sampling we can think of the probabilistic distribution as being completely determined just by ϕ for the parameter set. Therefore, when determining the probability of sampling a particular set of parameters, we need only determine the candidate probability of ϕ, which is non-degenerate and given by

$$\begin{aligned} f_{\phi}(\alpha,\omega) &= f_{\mathbf{z}_\nu}(\mathbf{z}_\nu)\left|\frac{\partial \mathbf{z}_\nu}{\partial(\alpha,\omega)}\right| \\ &= f_{\mathbf{z}_\nu}(\mathbf{z}_\nu)\left|\frac{\partial(\alpha,\omega)}{\partial(\mathbf{z}_\nu)}\right|^{-1}, \end{aligned} \qquad (4.56)$$

where $f_{\mathbf{z}_\nu}(\mathbf{z}_\nu)$ is the joint probability of sampling the ν standard t-variates. The $(m+k-1)\times(m+k-1)$ matrix, $\frac{\partial(\alpha,\omega)}{\partial(\mathbf{z}_\nu)}$, is made up of two parts:

$$\frac{\partial(\alpha)}{\partial(\mathbf{z}_\nu)} = \mathbf{Q}_\alpha \qquad (4.57)$$

and

$$\begin{aligned} \frac{\partial(\omega)}{\partial(\mathbf{z}_\nu)} &= \frac{\partial(\omega)}{\partial(\mathbf{y})}\frac{\partial(\mathbf{y})}{\partial(\xi)}\frac{\partial(\xi)}{\partial(\mathbf{z}_\nu)} \\ &= \frac{\partial(\omega)}{\partial(\mathbf{y})}\frac{\partial(\mathbf{y})}{\partial(\xi)}\mathbf{Q}_\omega, \end{aligned} \qquad (4.58)$$

where

$$\left.\begin{aligned} \frac{\partial(\omega_i)}{\partial(y_j)} &= \frac{(1-\omega_i)}{\sum_{l=1}^{k} y_l} \quad j=i \\ &= \frac{-\omega_i}{\sum_{l=1}^{k} y_l} \quad j\neq i \end{aligned}\right\} \quad \begin{aligned} 1 \le j \le k, \\ 1 \le i \le k-1 \end{aligned} \qquad (4.59)$$

and

$$\left.\begin{aligned} \frac{\partial(y_i)}{\partial(\xi_j)} &= \frac{y_i(1-y_i)}{\hat{w}_i(1-\hat{w}_i)} \quad j=i \\ &= 0 \quad j\neq i \end{aligned}\right\} \quad 1 \le i,\ j \le k, \qquad (4.60)$$

where the matrix $\mathbf{Q}_\omega$ is the k by $k-1$ matrix that forms the bottom right hand corner of G.

We have thus shown how parameters for a mixture model can be generated using a multivariate t-distribution, going into some detail over how the component weights are generated to ensure that they both sum to one and are in the range (0,1). Expressions have also been given for the probability of sampling parameter sets under this sampling procedure, which are vital for importance sampling.

4.5.6 Convergence Statistics

We make use of the methods presented in Chapter 3 and in addition, evaluate the variance of p_k, the probability that the number of components in the model is equal to k,

$$Var[p_k] = \sum_{i=1}^{N} \left(\frac{f(\theta_i)}{w(\theta_i)} \right)^2 \delta_k(\theta_i) - \frac{1}{N} \left[\sum_{i=1}^{N} \frac{f(\theta_i)}{w(\theta_i)} \delta_k(\theta_i) \right]^2 , \qquad (4.61)$$

where $f(\theta_i)$ is the posterior probability of the i^{th} sample θ_i, $w(\theta_i)$ is the probability of sampling the parameter set θ_i, and

$$\delta_k(\theta_i) = \begin{cases} 1 & \text{number of components is } k \\ 0 & \text{otherwise} \end{cases} \qquad (4.62)$$

The variance of p_k can be calculated for each value of k, $k = 1, \ldots, k_{max}$ to give a measure of the quality of the solution. We also consider the unit coefficient of variance Δ_k for each of the p_k, which is defined to be the estimated standard deviation divided by the estimated mean.

The maximum of the normalised importance sampling weights is the percentage of the probability distribution included in just one point of the sample. This gives a further indication of convergence and is presented for all examples. In addition, we fit an extreme value distribution to the importance sampling weights, and use the results of the fitting to assess the convergence, as described in Section 3.3.

4.6 Examples

There are a number of standard datasets for the mixture model problem, most of which are discussed by Richardson and Green [84]. We here present results for four examples: a generated distribution of three normal distributions; a dataset

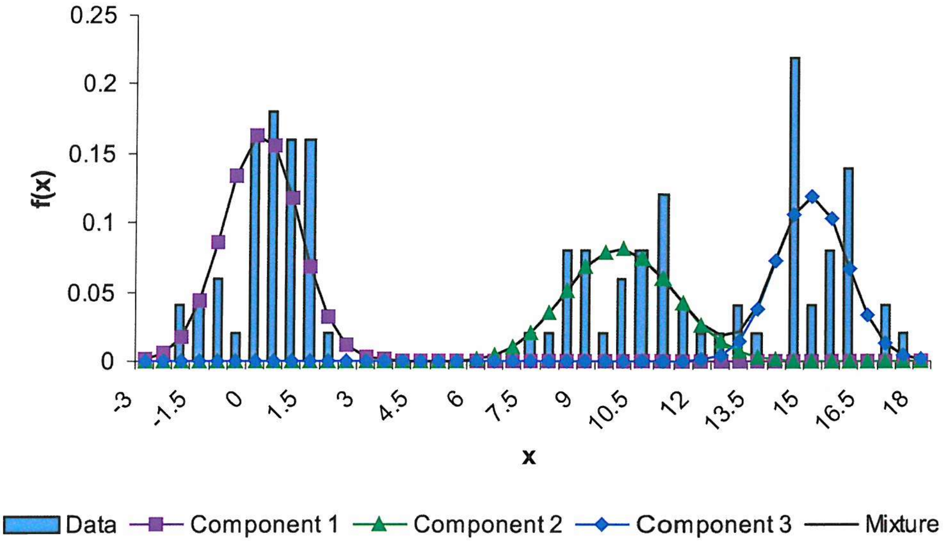

Figure 4.3: Probability density function for the three normals data.

describing the speeds of galaxies [87]; one of the more difficult standard datasets, the enzyme data [10]; and finally a dataset describing the acidity of lakes in north-central Wisconsin, [32].

4.6.1 Example 1: Three Normals

This is a test dataset of 100 data points sampled from a mixture of three normals with component means at 0, 10 and 15, component variances of 1, 2 and 1, and with equal weight applied to each component of the mixture. The sampling assigns a posterior probability of 0.87 to there being 3 components and 0.13 to there being 2 components. The best-fit distribution with three components (as found using the optimization routine), the data and the actual distribution are given in Figure 4.3. It has been argued [1] that it is easier to assess the fit of a distribution to data using an EDF and this is given in Figure 4.4

We assess the convergence of the importance sampling using the techniques

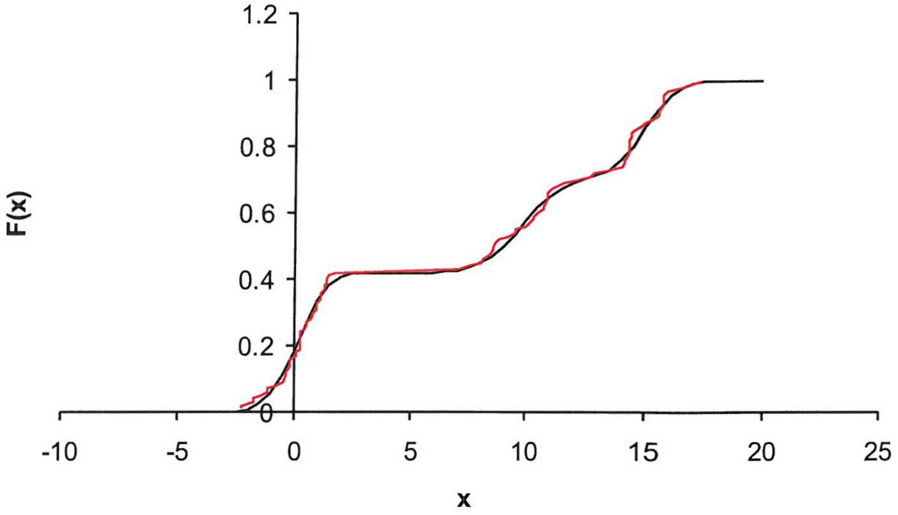

Figure 4.4: Empirical distribution function for the three normals data.

described in Chapter 3 and Section 4.5.6. Convergence statistics for the sampling are given in Table 4.2. These include the results of fitting an extreme-value distribution to the importance sampling weights. The tests work by fitting a generalised pareto distribution to the exceedances over a threshold. Weights that are smaller than the threshold are not included in the fitting. Choosing the threshold is a matter of judgment, and we used a number of different thresholds for each example. Only the results for the most representative thresholds for each of the three tests described in Section 3.3 are included.

The convergence statistics suggest that the importance sampling has converged for this example, with a relatively small sampling variance and maximum weight. The unit coefficients of variance are also small for the models of interest, becoming larger for models with high numbers of components.

One difficulty with the methodology that we have used for the importance sampling here is that we do not take full account of our knowledge of the posterior distribution. Although by looking at the data for this example, we can be reasonably confident that the number of components will be less than five, we still assign

Measure	Result
Δ_1	0.0366
Δ_2	0.0659
Δ_3	0.0585
Δ_4	0.138
Δ_5	0.424
Δ_6	0.388
Δ_7	0.597
Δ_8	0.386
Δ_9	0.976
Δ_{10}	0.951
$Var(f/w)$	2.02 x10^{-7}
Max(f/w)	0.00818
Wald Test	-4 (accept)
Score Test	-2.2 (accept)
Likelihood Ratio	-700 (accept)

Table 4.2: Unit coefficients of variance and convergence statistics for the importance sampling in the three normals example.

equal probability to sampling models with numbers of components between one and k_{max} (in this case k_{max} is 10). Models with numbers of components very different from three will have a relatively small posterior probability associated with them, and so relatively small sampling weights. If there were zero posterior probability associated with the other models, the weights associated with the correct model should be approximately k_{max}/N, where N is the number of runs performed during the importance sampling, k_{max} times larger than the weights generated by a model in which the sampling function is a good description of the actual function. In reality the posterior probabilities for the other models will be non-zero, but will contribute less than $1-1/k_{max}$ to the posterior distribution. This necessarily worsens the convergence of the importance sampling.

4.6.2 Example 2: Enzyme Data

The enzyme data comes from [10] and is made up of 245 data points. The results suggest that there is a 70% probability of the model being made up of four components and a 30% chance that it has only three components. The estimated probability distribution, with parameters set at the mode of the posterior distribution, and a histogram of the data are given in Figure 4.5. We also include the EDF of the data and the estimated cumulative distribution function in Figure 4.6

Convergence statistics for the enzyme data are given in Table 4.3. They suggest relatively high unit coefficients of variance and a relatively high maximum weight. Two out of three of the EVT statistics suggest that the variance does exist.

4.6.3 Example 3: Acidity Data

The acidity data comes from [32] and is made up of 155 data points. The results suggest that the mixture distribution is made up of two components, with a 99.8%

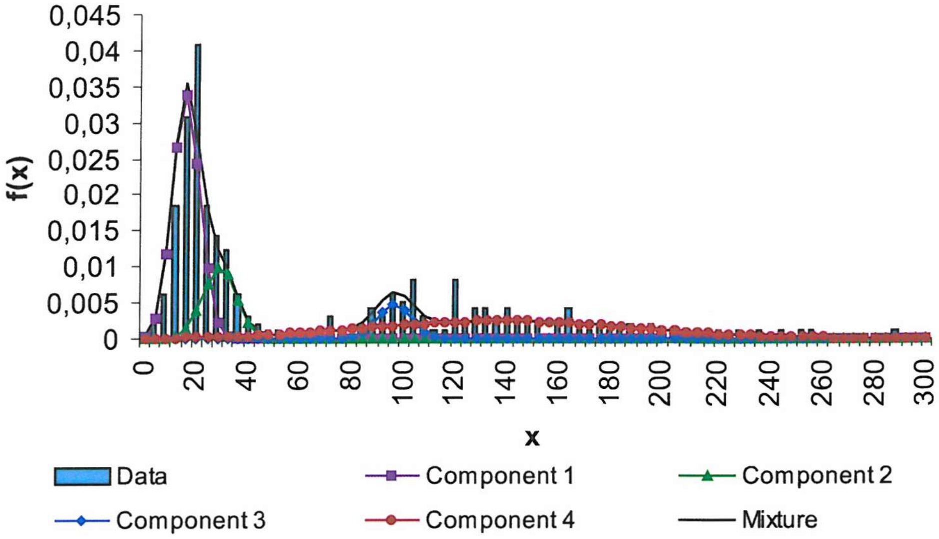

Figure 4.5: Probability density function for the enzyme data.

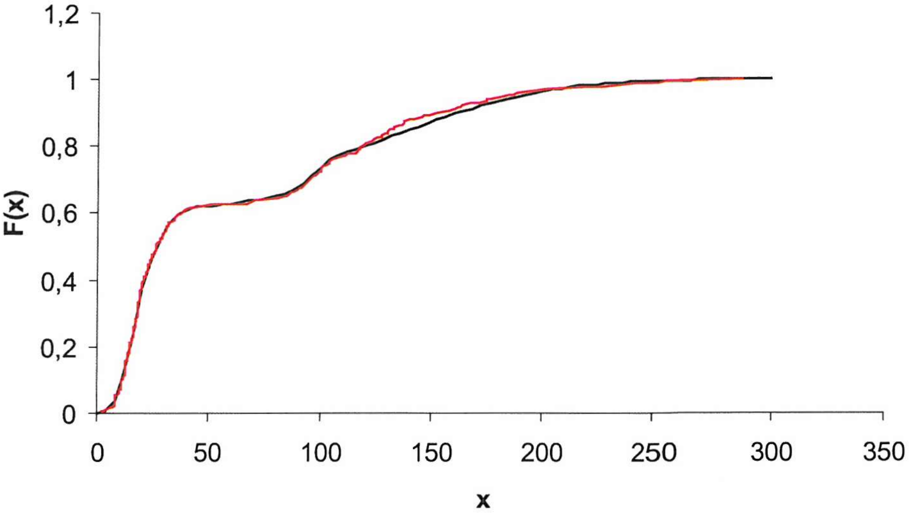

Figure 4.6: Empirical distribution function for the enzyme data.

Measure	Result
Δ_1	0.0367
Δ_2	0.991
Δ_3	0.423
Δ_4	0.198
Δ_5	0.194
Δ_6	0.265
Δ_7	0.777
Δ_8	0.838
Δ_9	0.825
Δ_{10}	0.496
$Var(f/w)$	7.75 x10^{-6}
Max(f/w)	0.254
Wald Test	-0.1 (accept)
Score Test	-0.1 (accept)
Likelihood Ratio	2900 (reject)

Table 4.3: Unit coefficients of variance and convergence statistics for the importance sampling in the enzyme example.

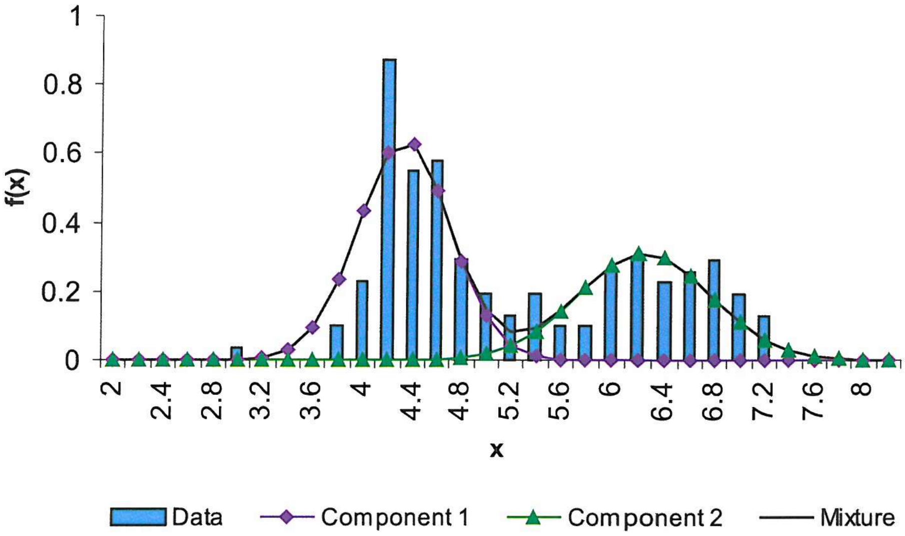

Figure 4.7: Probability density function for the acidity data.

posterior probability of this being the correct model. Figure 4.7 shows the probability distribution function for the model with two components, using the modal parameter values found in the optimization, with a histogram of the data. We also give the empirical distribution function of the data in 4.8

As Table 4.4 shows, the variance of the sampling and the maximum weight are small for this example, suggesting that the importance sampling has converged, which is confirmed by the extreme value statistics. Unit coefficients of variance are also small for the models of interest, again increasing for models with higher numbers of components.

4.6.4 Example 4: Galaxy Data

The galaxy data comes from [87] and is made up of 82 data points. The posterior distribution has a maximum for $k = 3$, with a 99% chance that the model has three components, and a 1% chance that it has only two. We present the probability

Measure	Result
Δ_1	0.0368
Δ_2	0.0774
Δ_3	0.132
Δ_4	0.157
Δ_5	0.254
Δ_6	0.284
Δ_7	0.627
Δ_8	0.712
Δ_9	0.554
Δ_{10}	0.759
$Var(f/w)$	5.99 x10^{-7}
Max(f/w)	0.0485
Wald Test	-2.7 (accept)
Score Test	-0.5 (accept)
Likelihood Ratio	1900 (reject)

Table 4.4: Unit coefficients of variance and convergence statistics for the importance sampling in the acidity example.

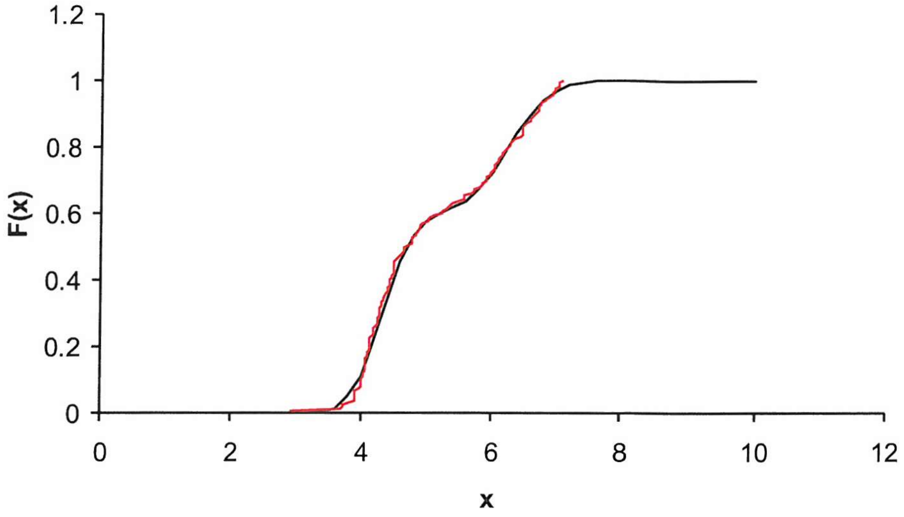

Figure 4.8: Empirical distribution function for the acidity data.

distribution function for a model with three components, using the modal parameters estimated using the optimization routine, alongside a histogram of the data in Figure 4.9. The empirical distribution function for the three component model is given in Figure 4.10.

The convergence statistics shown in Table 4.5 are slightly puzzling. The unit coefficients of variance, the variance of the sampling and the size of the maximum weight suggest that the sampling has converged. However the extreme value statistics suggest the opposite. We can be relatively confident that the model with two components is the most likely but can probably be less confident about the value given for the posterior probability of it being the true model.

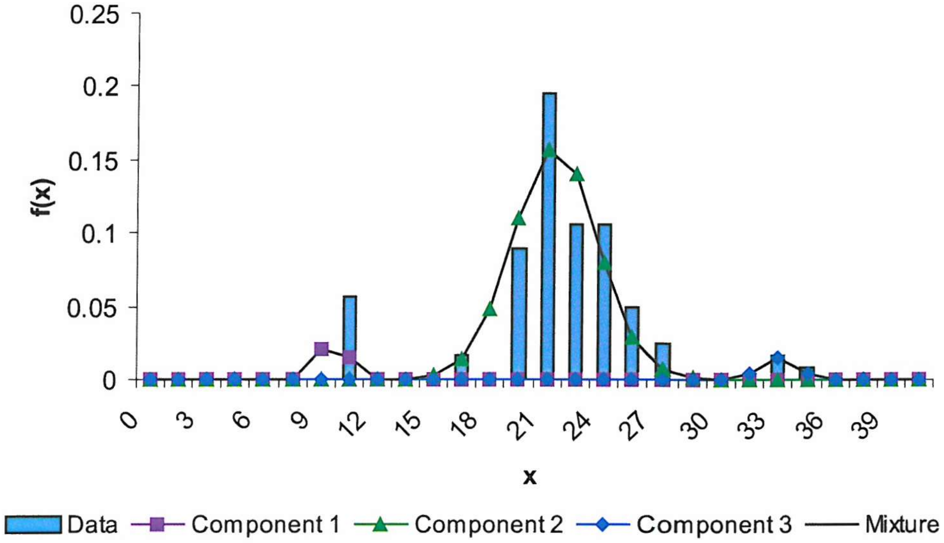

Figure 4.9: Probability density function for the galaxy data.

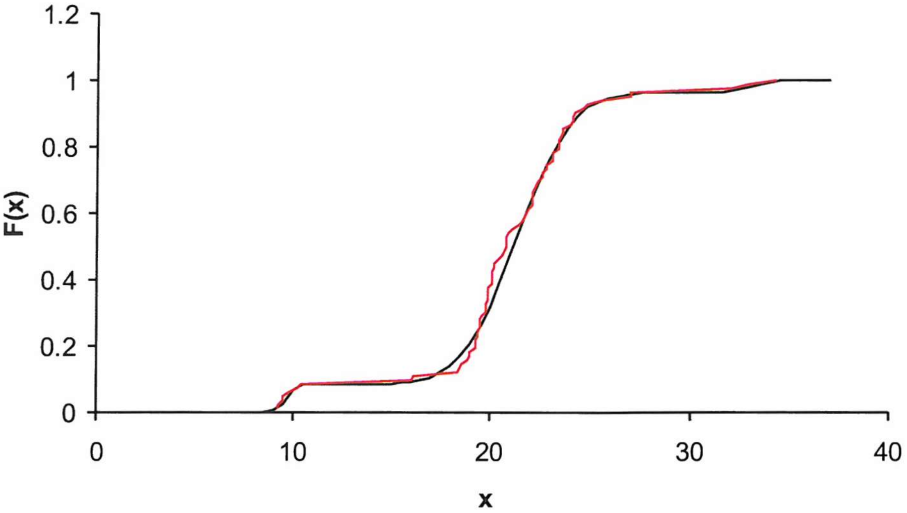

Figure 4.10: Empirical distribution function for the galaxy data.

Measure	Result
Δ_1	0.0378
Δ_2	0.0915
Δ_3	0.128
Δ_4	0.720
Δ_5	0.667
Δ_6	0.442
Δ_7	0.559
Δ_8	0.964
Δ_9	0.651
Δ_{10}	0.734
$Var(f/w)$	1.52 x10^{-6}
Max(f/w)	0.0555
Wald Test	8 (reject)
Score Test	-0.4 (accept)
Likelihood Ratio	1000 (reject)

Table 4.5: Unit coefficients of variance and convergence statistics for the importance sampling in the galaxy example.

4.7 Discussion

We have successfully used importance sampling to determine the posterior probability distributions of normal mixture models for a number of standard datasets, as shown in Section 4.6. The convergence statistics that we have presented suggest that importance sampling can be an efficient method of model selection, when combined with a prior investigation of parameter space to determine the optimal sampler function.

4.7.1 Comparison of the Results with the Literature

Comparison of our results with those of Richardson and Green [84], show that we suggest more definite posterior probability distributions for the number of components in the model, which generally predict a smaller number of components in the mixture. This may be due to the choice of prior distributions, as discussed in Section 4.3.3. Alternatively, it could reflect differences in the methodology. A comment by Cheng and Liu in the discussion of the Richardson and Green paper [22] suggests the possibility that models in which the number of components are greater than the true number of components could have a finite posterior probability incorrectly associated with them. It is possible to generate a model in which two or more of the components are very similar or one or more components have a very high variance or a very low weight. In such models, one or more of the components could be combined with other components, or removed, without significantly altering the probability density, and so these are effectively models with a smaller number of components than are actually used. These models are probably only rarely generated using our methodology because the importance sampling focuses on areas of parameter space relatively close to the mode found by the optimization. It may be more likely that such models are generated using the reversible jump MCMC and this may be an explanation for the difference in the results.

It is possible that a small change to the prior distributions might make it less likely that models are generated with very small component weights. This would involve increasing the parameter in the Dirichlet distribution from one to some higher value. The Dirichlet distribution describes the prior probability function for the component weights and increasing its parameter beyond one assigns a zero probability to component weights of zero, which is desirable, but also introduces some bias towards more uniform component weights in the prior distribution. How best to balance these two effects could be the subject of future research.

Looking at this problem from the frequentist point of view, one suggestion for determining the optimal number of components is to make use of the fact that the Fisher information matrix of a model becomes close to singular when the model is being overfitted. The fit of the model will always improve with an increase in the number of components, therefore from the frequentist perspective the optimal number of components will be the smallest number that still produces a reasonable fit.

4.7.2 Discussion of the Sampling Methodology

Importance sampling has several advantages over MCMC including the lack of serial correlation between samples and the better measures of convergence. These are discussed by Evans and Swartz [46] in their review paper. There are additional advantages in this particular example because of the difficulty of designing a MCMC routine that can jump between different models. With importance sampling, the choice of model can be made in an identical manner to the choice of parameters in the models. However, for importance sampling to be efficient, time must be spent investigating the distribution being sampled from and this time must be combined with the run length of the sampling itself to give the total computing time expended on the problem. For the examples considered here, the total number of function evaluations was approximately 25,000 for the optimization plus 10,000 for the importance sampling, giving a total of 35,000.

For some examples, especially those with a large number of data points, it can be difficult to find the modes of the posterior for models with a large number of components. The estimates found can often be such that the Hessian matrices calculated at the modes are not positive-definite. In these cases, we transform the Hessian matrix to a matrix of eigenvalues, swap the sign of any negative eigenvalues and then use this in our calculation of the generalised inverse. We suspect that these problems arise mainly in situations where the posterior distribution is very flat at the mode, generally where the model has too many components for the data.

The methodology that we use for the importance sampling results in imperfect convergence, as we take no account of our knowledge of which is the correct model when setting the candidate function in the importance sampling. Instead, an equal probability is assigned to the sampling of each of the k_{max} models. This allows us to argue that we introduce no bias on the choice of model into the importance sampling, but will result in the sampler wasting time sampling parameters for models with a very low posterior probability. One small extension to the methodology that could be investigated in the future is to use different probabilities for sampling different models.

Chapter 5

Investigation of the Effectiveness of Interventions Against Tuberculosis and HIV Using a Compartmental Model

5.1 Introduction

In this chapter we use use Bayesian methods to fit a compartmental difference equation model of tuberculosis (TB) driven by HIV The model is then used to compare the effectiveness of preventive and curative methods for the control of TB in high HIV prevalence settings. This is a slight adaptation of a model described previously [36] and will be described in Section 5.3. The methodology used for the uncertainty analysis is very similar to that used in the analysis of finite mixture models, although the application is very different.

We fit the model using literature estimates for the model parameter values as prior information, and time series of HIV prevalence and TB incidence to estimate the likelihood. The methodology used for the fitting process includes an initial optimization routine to find the maximum of the posterior distribution, followed by calculation of the Hessian matrix at that optimum to define a good candidate distribution. We tried using both importance sampling and Markov Chain Monte Carlo sampling (MCMC) to find the posterior distribution. MCMC was chosen as the final sampling methodology because it converged much better than the importance sampling. The results of the importance sampling were often biased by very large weights, corresponding to points in parameter space at which the candidate probability is low but the posterior probability is high. The output of the MCMC is used to determine the expected TB incidence and HIV prevalence and projections of the effectiveness of interventions. Sampling from the output, equivalent to sampling from the posterior distribution, enables the estimation of confidence limits that incorporate the knowledge coming from prior information on parameter values and the fit of the model to the available data.

We describe the aims of the study and some of the context to the problem in Section 5.2. The model is described in Section 5.3, and the modelling of the interventions in Section 5.4. A full description of the Bayesian methodology used to fit the model is given in Section 5.5, and the results of the study are given in Section 5.6. We conclude in Section 5.7.

5.2 Background to the Problem

Mycobacterium tuberculosis (TB) and the human immunodeficiency virus (HIV) are the leading causes of death due to infectious diseases among adults [30], [77]. The spread of HIV infection has already led to a dramatic increase in TB cases in eastern and southern Africa [79], where up to 60% of TB patients are co-infected with HIV [101], and threatens to do so elsewhere. The World Health Organization's DOTS strategy for TB control [79], [78], based on the provision of adequate resources, accurate diagnosis, good treatment, the use of the correct drugs and good monitoring to ensure that active cases of disease are rapidly found and cured, forms the basis of most national TB control programmes. In recent years however even good DOTS programmes have failed to check the rapid increase in TB cases in countries with a high prevalence of HIV, and this has stimulated the search for new ways to manage TB epidemics [38].

Since HIV is a potent risk factor for the development of TB, it should be possible to avert new TB cases by reducing HIV transmission through behavioural interventions (promoting condoms, changing sexual behaviour etc.), boosting patients' immunity by treating them with highly-active anti-retroviral therapy (ART) [93], or by administering TB preventive therapy (IPT), usually through 6-9 months' treatment with isoniazid [26]. Previous studies have attempted to calculate the number of TB cases and deaths that can be averted by finding and treating active TB cases during the course of HIV epidemics [44], [71], [81], but none have evaluated the curative approach against the three principal means of prevention.

The analysis of the effectiveness of the different interventions focused on Kenya. There are reasonably good data available both for HIV and TB in Kenya, and the epidemic is more advanced than in some African countries (such as South Africa), but less advanced than in others (such as Uganda). It is impossible to determine from the available data whether the prevalence of HIV will continue to rise, remain steady or fall, so we consider three different underlying HIV epidemics in which HIV incidence, in the absence of any further intervention, levels off at its current value, increases by half, or falls by half (Figures 5.6 to 5.8). To explore the generality of the findings for Kenya, we also fitted the model to data from Uganda and South Africa [75], where the HIV epidemics are, respectively, more and less advanced.

5.3 Mathematical Model of TB-HIV

We reduced an earlier compartmental model of TB-HIV epidemiology [44] to a single age class (adults 15-49 years), and extended the modelling of interventions to include TB preventive therapy, the administration of ART and the effect of HIV prevention methods, as well as TB case detection and cure. The model was written in Visual Basic and combines a compartmental model of TB progression with a statistical model of HIV prevalence. Figure 5.1 illustrates the general structure of the TB model. For clarity, non-infectious TB states have been omitted from this diagram and in the full model active TB may be infectious or non-infectious, with movement allowed from active non-infectious disease to active infectious disease. An identical sub-model, with different parameter values, describes the progression of those in the later stages of HIV (Stages 3 and above of the WHO staging system [4], [68]). Movement between the two sub-models is governed by the statistical model of HIV prevalence, described in Section 5.3.1. Death can occur in any state, but death rates are higher for patients with active disease. An early version of this model was presented in [34].

Active TB can arise through any of three mechanisms. Those who acquire a new TB infection either develop progressive primary disease within 1 year, or enter a latent state from which TB can arise by reactivation or re-infection. The same proportion of individuals who are latently infected can also develop TB within one year of re-infection or reactivation. We use a time step of three months in the model and assume that those developing primary disease move straight to the active disease state. Active TB may be infectious or non-infectious.

During the later stages of HIV, co-infection leads to a greatly increased risk of developing TB, though a smaller fraction of active TB cases becomes infectious. Individuals with late-stage HIV infections (WHO stages three and above) also have higher death rates, with and without active TB.

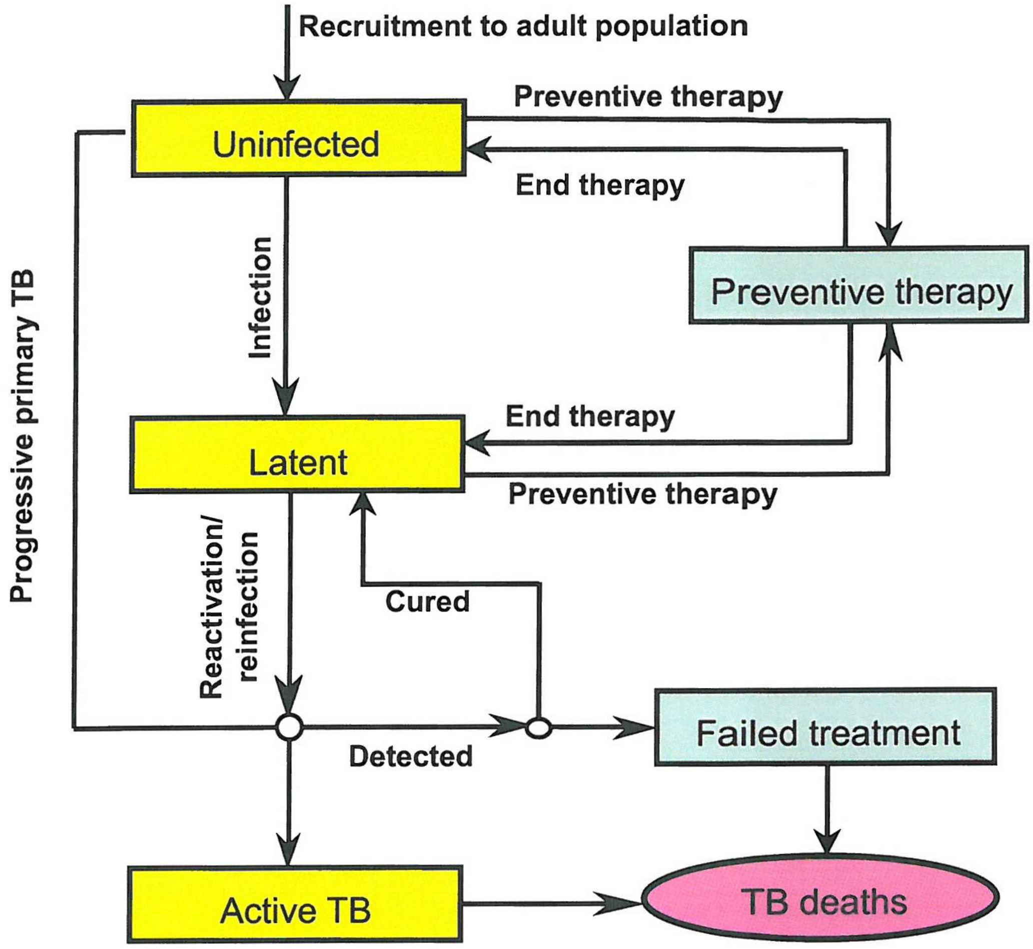

Figure 5.1: Outline of the TB sub-model.

5.3.1 HIV Model

The purpose of the separate HIV model is to determine the incidence of HIV in each time step and from this, estimate the rate at which people move from one TB sub-model to another. We assume that approximately four years after infection with HIV, individuals will enter late-stage HIV, where this time lag is described by the model parameter tLs. Therefore, in a given time period, the number of individuals moved from the first sub-model to the corresponding state in the second sub-model (TB uninfected or latently infected), is equal to the HIV incidence tLs time periods previous, corrected for deaths.

Reasonably good data are available for the prevalence of HIV infection over time in the countries of interest but the future course of the epidemic is much less certain. It is therefore desirable to use a flexible model to describe the HIV epidemic, allowing the HIV prevalence to increase, decrease or remain constant in the future.

The model that we developed originally in [36] fits a logistic function to HIV prevalence data and then uses a given relationship between prevalence and incidence to derive the HIV incidence from this functional form. This method is attractive as HIV epidemics are traditionally measured in terms of prevalence (number of cases per unit population) rather than incidence (number of cases per unit time), and it allows the long-term behaviour of the HIV epidemic to be set in terms of long-term HIV prevalence. The incidence can then be derived from this. Of the other authors who have considered this problem Salomon and Murray [92] have approached it from the opposite perspective, selecting a functional form for the incidence and using the relationship between incidence and prevalence to find the HIV prevalence. The others [29], [102] approach it from a similar perspective to [36].

Investigations of the different models showed that the method used originally in [36] and the method of Williams et al [102] result in the incidence of HIV being very dependent on the value of the time step ts used in the calculations. The method of Salomon and Murray [92] is also dependent on ts but the dependence is much weaker, and insignificant with respect to the results. The model described by Colvin et al [29] would also be stable to changes in the time step as earlier estimates of incidence have no effect on future values of the incidence, but restricts the choice of survival function. The Salomon and Murray method seems to be the most attractive and is the method used in this study to derive the HIV incidence. Although the HIV prevalence is now derived from an estimate of the HIV incidence, it is still possible to define long-term scenarios for the HIV epidemic, but expressed in terms of the long-term HIV-incidence rather than the long-term HIV prevalence.

Using Salomon and Murray's model [92], the prevalence $p(t)$ is given by

$$p(t) = \sum_{i=0}^{t-1} Inc(i)F(t-i)ts, \tag{5.1}$$

where $F(\tau)$ is the probability of surviving τ time periods and $Inc(i)$ is the HIV incidence in time period i. Salomon and Murray use a Weibull function to describe the time from infection to death. We instead use a survival function which mirrors that used in the TB compartmental model, such that

$$F(t) = \begin{cases} \exp(-\mu t) & t \leq tLs \\ \exp(-\mu_{HIV} t) & t > tLs \end{cases} \tag{5.2}$$

where μ is the background death rate and μ_{HIV} is the death rate for those in late-stage HIV. We use the function suggested by Salomon and Murray for the incidence of HIV,

$$Inc(t) = \frac{\gamma\beta^{-\alpha}(t-t_0)^{\alpha-1}e^{-(t-t_0)/\beta}}{\Gamma(\alpha)} \tag{5.3}$$

for $t \leq t_0 + \beta(\alpha-1)$ and

$$\begin{aligned} Inc(t) = \frac{\gamma\beta^{-\alpha}}{\Gamma(\alpha)} & [(1-\theta)(t-t_0)^{\alpha-1}e^{-(t-t_0)/\beta} \\ & +\theta(\beta(\alpha-1))^{\alpha-1}e^{-\beta(\alpha-1)/\beta}] \end{aligned} \tag{5.4}$$

for $t > t_0 + \beta(\alpha - 1)$. This is a gamma distribution with a multiplicative factor γ to allow for differences in scale, and an additive term that is used to describe the long-term incidence. The variables α and β set the shape of the incidence curve, γ sets the scale of the curve and θ sets the level of the long-term incidence, which is equal to θ multiplied by the peak incidence.

We use the parameter θ to define the scenario for long-term behaviour that we are considering (incidence falls to 50% of current value $-$ $\theta = 0.5$, incidence remains at its current level $-$ $\theta = 1$, incidence increases to 150% of its current value $-$ $\theta = 1.5$) and fit the model to HIV prevalence data by varying the parameters α, β and γ. We assume that the HIV prevalence data points have normal errors and therefore find the optimal set of parameters by minimising the sum of the squared difference between the model's estimate for HIV prevalence and the HIV prevalence data.

5.4 Interventions

The TB case detection rate is the proportion of new, active cases that are found and begin treatment during a given time period. The cure rate is the proportion of those who are treated that become non-infectious and are at no additional risk of dying from TB. We assume that cured TB patients uninfected with HIV, or in the early stages of HIV infection, remain infected with TB; those that have late-stage HIV infections return to the uninfected state, which gives them some immunity against developing active TB. Among patients that fail treatment, a proportion remains infectious; the remainders do not transmit TB, but have a high probability of relapsing to active disease, compared with patients that were deemed to have been cured at first treatment.

The main effect of preventive therapy for TB is to eliminate the chance of developing active TB for 70% of infected people who receive it; the other 30%

are assumed to receive no benefit [55], [100]. Ideally, TB preventive therapy is given only to those who are already infected with TB (and never to active cases); however, TB infections cannot always be identified by tuberculin skin-testing, especially in anergic subjects co-infected with HIV [26]. We therefore use HIV infection as the criterion for the administration of TB preventive therapy, and coverage is measured as the fraction of patients that receive one course of treatment between initial HIV infection and death. We assume that those given preventive therapy for TB are protected from TB infection for the duration of treatment [62]. Treatment is either for six months or for life; patients treated for six months return to their previous state, either latent or uninfected.

By reducing TB prevalence among HIV-positives we effectively reduce the death rate of those in late-stage HIV, thereby increasing the late-stage HIV population. We assume that this has a negligible effect on HIV transmission and do not include a corresponding rise in HIV incidence.

In our model ART returns patients to their corresponding TB state in early-stage HIV infection, and prevents their HIV infection from progressing for as long as they continue to take the appropriate combination of drugs. Since the increase in life expectancy of patients on ART (as currently formulated) has been measured at 5-7 years [40], [27], [28], or less [50], this is an optimistic view of the effectiveness of ART. As yet, there are few data on compliance with ART. We consider an optimistic scenario and a more realistic scenario for dropout from ART, with dropout rates of 5% and 20% per year [98]. We have not explicitly allowed for the emergence of drug resistance under ART, and we assume that ART has no impact on HIV transmission.

The coverage of interventions that do reduce HIV transmission (condoms, change of sexual behaviour, etc) is expressed in terms of the effects on HIV incidence, e.g. reducing the annual HIV incidence rate by 1% from the point of intervention onwards.

We measure the impact of interventions over and above present levels of coverage. For the TB treatment measures, we assume that currently 50% of the new infectious TB cases that arise each year are detected, and 70% of these are cured, which is thought to be typical for sub-Saharan Africa [79]. For ART, coverage is measured as the fraction of HIV-infected persons progressing to AIDS that receive antiretroviral drugs. Similarly, the coverage of TB preventive therapy is measured as the fraction of HIV-infected persons (including all TB and HIV co-infected persons) given one course of treatment between HIV infection and death. For condoms and other measures designed to prevent infection, we express coverage in terms of its effect on HIV incidence, applying a fixed percentage reduction in annual HIV incidence from the point of intervention onwards. For all three of the preventive measures, we assume that coverage was negligible prior to the modelled interventions. Thus, there is great potential to improve on prevention, much less to improve on cure.

5.5 Fitting the Model to Data

We use a Bayesian methodology [64], [46] to fit the model to the available data. Prior estimates of the distribution of each parameter are combined with the likelihood function to give the posterior distribution. The likelihood is estimated by fitting the model output to estimates of TB incidence and HIV prevalence from each country [79], [75], assuming normal errors. Prior distributions for the parameters describing transitions between TB states were obtained from published studies (Further information on these studies is given in the supplementary material of [36]). These prior distributions are all assumed to be normal. Little prior knowledge about the parameters describing the HIV epidemic was available and vague priors were used (uniform distributions with lower limits of zero and very high upper limits). As all of the parameters must be greater than zero and the rate parameters should be between zero and one, gamma or beta distributions may be more appropriate prior distributions for most of the model parameters. However, none of the parameters have priors with a significant probability outside the permissible region and as we assume normal errors for the data, with normal priors the posterior distribution is also likely to be close to a normal distribution. We assume that the posterior is normally distributed in defining the candidate functions for both the importance sampling and the Markov Chain Monte Carlo sampling (MCMC).

The advantage of using a Bayesian approach in this situation is that prior information on most parameter values is good and using a likelihood approach, this prior information would simply have been ignored. The scarcity of TB incidence data and the large number of parameters used in the model means that it is especially useful to use all of the prior information. With the Bayesian approach, if more information is available from the prior distribution than from the new data for a particular parameter the posterior distribution will depend mainly on the prior information. Conversely, if the prior information on a particular parameter is weak and the data constrain the parameter to a relatively small range of values, the posterior distribution will depend mainly on the likelihood function.

Two methods were tried for finding the posterior probability distributions of the parameters: the importance sampling methodology described earlier in this thesis and MCMC sampling. In both cases, before conducting the sampling, we investigated the form of the posterior probability distribution, using the Nelder Mead optimization routine [74] to find the mode of the posterior distribution. We then estimated the Hessian matrix at the mode and used this to find the covariance matrix of the posterior distribution.

5.5.1 Importance Sampling

Given the form of the prior distributions and the likelihood function, we assume that the posterior distribution will be approximately multivariate normal. Therefore, based on the results of Chapter 2 a multivariate t-distribution with 4 degrees of freedom is used as the candidate distribution in the importance sampling with mean given by the mode of the posterior distribution and covariance matrix as calculated at the mode.

We found that importance sampling converges relatively poorly for this example. This is principally due to the presence of very large weights, which produce high, narrow peaks in the the resulting posterior distribution, and introduce a bias into statistics such as the mean. In fact the top five normalised weights were 1, 3.19 x 10^{-16}, 3.46 x 10^{-40}, 1.03 x 10^{-44} and 1.10 x 10^{-52}, compared with the ideal weight size of 1/80,000 (1.25 x 10^{-5}). Obtaining very high weights is generally regarded as a symptom of poor convergence, and suggests that the sampling function used was not sufficiently close to the function being integrated over (in this case the posterior distribution) or that the number of runs is insufficient. In previous work [36] we used importance sampling to determine the posterior distribution and convergence was reasonable. In that study, we assumed a multivariate normal prior distribution for the HIV prevalence parameters, based on the fit of the HIV prevalence model to the HIV prevalence data and used a different model for estimating HIV incidence. We then only incorporated the fit of the model to the TB incidence data in calculating the likelihood. This meant that the posterior distribution for the HIV parameters was close to multivariate normal. In this study, we use a uniform as the prior distribution for the HIV parameters, and incorporate the fit of the HIV model into the calculation of the likelihood. This means that the posterior distribution for the HIV parameters is further from a normal distribution, as the results of the MCMC sampling show. This probably explains why the

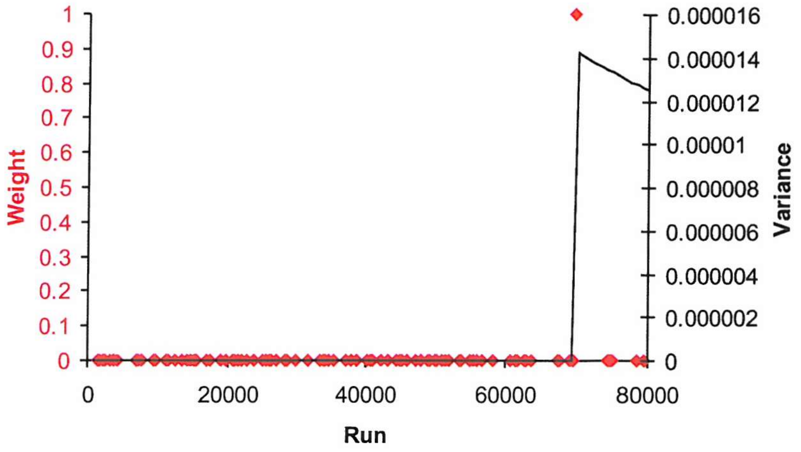

Figure 5.2: Highest one hundred normalised weights and sampling variance using importance sampling to find the posterior distribution of the parameters of the TB-HIV model.

importance sampling convergence was worse in this study than in [36].

We make use of the results of Chapter 3 and perform diagnostic and statistical tests of convergence. Both suggest that the importance sampling is not performing well. Figure 5.2 shows how the variance of the sampling varied during the 80,000 runs, with the top one hundred weights superimposed on this. It clearly shows how one high weight affects the sampling variance. The statistical tests of convergence based on extreme value theory, as described in Section 3.3 unequivocally state that the sampling has not converged.

The estimates of the posterior distribution obtained by importance sampling for the parameter w is shown in Figure 5.3. This parameter describes the rate at which non-infectious active TB becomes infectious active TB for those who are HIV-negative/early HIV-positive. There is one large spike in the graph, which corresponds to the set of parameters with the largest weight, and demonstrates the non-convergence.

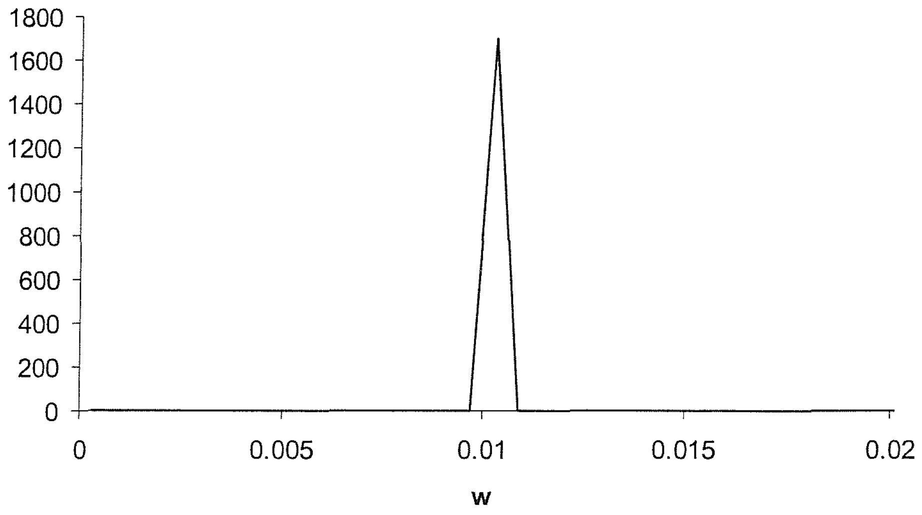

Figure 5.3: Estimated posterior distribution for w (rate at which non-infectious active TB becomes infectious active TB), determined using importance sampling.

5.5.2 Markov Chain Monte Carlo Sampling

We use the Metropolis-Hastings algorithm for the MCMC sampling, with a multivariate t-distribution with four degrees of freedom as the candidate distribution. We set the mean to be the current position of the chain in parameter space and the covariance structure of the candidate distribution to be the inverse of the Hessian matrix at the mode of the distribution, multiplied by a scaling factor, where the scaling factor is chosen based on observations of the mixing of the chains. Ideally, the mixing should be such that the probability of the chain moving to a new position should be between about 15 and 50% [86]. To achieve a level of mixing within this range, the best scaling factor appears to be 0.23, for which the probability of the chain moving to a new position is approximately 16%.

The warm up is set to be 3000 runs based on observations of the trace. To check that the simulation is covering the full range of parameter space, the algorithm is started from five different points: the mode, all parameter values below their mode values, all parameter values above their mode values and parameter values set at a mixture of low and high values in the other two chains. Following the warm up, we run each of the chains for 5000 iterations.

The MCMC is run separately for two different groups of parameters, with the first group containing all of the TB model parameters and the second group containing the parameters used in the estimate of the HIV incidence. The parameter determining the long-term HIV incidence θ is held constant during the sampling. It is set by the user to one of three different values corresponding to three scenarios for the HIV epidemic: incidence decreases by 50% ($\theta = 0.5$), remains the same ($\theta = 1$), or increases by 50% ($\theta = 1.5$), in the long-term. The parameters are split to improve convergence of the MCMC to the posterior distribution. The covariance matrix estimated following the optimization suggests little correlation between the HIV model parameters and the TB model parameters, and so the split seems reasonable.

Various methods exist for checking the convergence of MCMC. Comparing the traces of the different chains can be used as an initial check. If, after discarding the initial warm up runs, the traces seem to overlap and appear to have been produced by the same process, then they have probably reached a stage where the starting position of a chain is no longer influencing its current position in parameter space. The analytical methods described in [51] give a more quantitative method of measuring the same thing, i.e. whether the chains have reached a stationary state. These techniques compare the between-chain-variance

$$B = \frac{n}{m}\sum_{i=1}^{m}(\bar{\psi}_{i.} - \bar{\psi}_{..})^2 \tag{5.5}$$

with the within-chain-variance

$$W = \frac{1}{m}\sum_{i=1}^{m} s_i^2, \tag{5.6}$$

where m is the number of chains, n is the number of runs made after the warm up for each chain and

$$\begin{aligned}\bar{\psi}_{i.} &= \frac{1}{n}\sum_{j=1}^{n}\psi_{ij} \\ \bar{\psi}_{..} &= \frac{1}{m}\sum_{i=1}^{m}\bar{\psi}_{i.} \\ s_i^2 &= = \frac{1}{n-1}\sum_{j=1}^{n}(\psi_{ij} - \bar{\psi}_{i.})^2. \qquad (5.7)\end{aligned}$$

The within-sequence-variance, W should be an underestimate of the variance of ψ because each of the individual sequences will not have moved over the whole range of ψ. We can also calculate an overestimate of the variance of ψ,

$$V = \frac{n-1}{n}W + \frac{1}{n}B. \qquad (5.8)$$

This estimate will be unbiased if the starting points were drawn from the target distribution, but an overestimate under the more realistic assumption that the starting points are over-dispersed. By measuring the ratio of these two quantities we can estimate the factor by which W, the conservative estimate of the range of ψ, might be reduced. Gelman terms this the "estimated potential scale reduction", given by

$$\sqrt{\hat{R}} = \sqrt{\frac{V}{W}}, \qquad (5.9)$$

the ratio between the estimated upper and lower bounds for the standard deviation of ψ.

We here estimate $\sqrt{\hat{R}}$ for each of the parameters, with a value close to 1 suggesting good convergence. Results shown in Table 5.5.2 show that $\sqrt{\hat{R}}$ is less than 1.10 for all of the parameters, which suggests that a stationary distribution has been reached. We use this as our threshold for convergence: if the model has converged, $\sqrt{\hat{R}} \leq 1.10$.

Parameter	$\sqrt{\hat{R}}$
λ_0	1.00
p	1.03
p_{HIV}	1.04
v	1.01
v_{HIV}	1.00
x	1.03
x_{HIV}	1.02
f	1.02
f_{HIV}	1.03
ϕ	1.01
w	1.02
w_{HIV}	1.02
μ	1.03
μ^{inf}	1.02
μ^{inf}_{HIV}	1.02
$\mu^{non-inf}$	1.01
$\mu^{non-inf}_{HIV}$	1.02
e	1.01
rf	1.00
tLs	1.00
tD	1.01
α	1.01
β	1.01
γ	1.01

Table 5.1: Convergence results for the MCMC sampling, showing $\sqrt{\hat{R}}$, the estimated potential scale reduction, for the mean of each of the parameters, where $\theta = 1$.

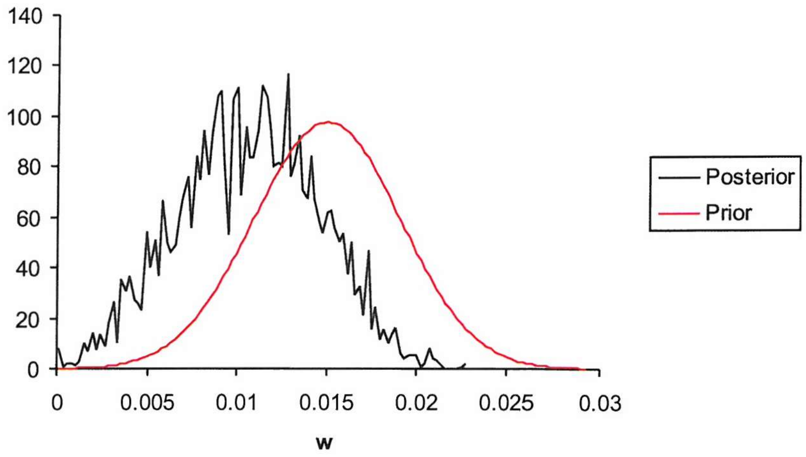

Figure 5.4: Estimated posterior distribution for w (rate at which non-infectious active TB becomes infectious active TB), determined using MCMC sampling.

5.5.3 Posterior Distributions of Parameters

The posterior means and variances of the parameters are given in Table 5.2, alongside the prior means and prior variances for the parameters, for $\theta = 1$. These show how our beliefs about the model parameters change based on the fit of the model to the data. Figures 5.4 and 5.5 show the prior and posterior distributions for parameters w and w_{HIV}. These parameters describe the rate of movement from non-infectious active tuberculosis to infectious active tuberculosis, among HIV-negatives/early-stage HIV patients, and late-stage HIV patients, respectively. As the figure shows, the posterior distribution for w is very different from the prior distribution, showing that the likelihood function has strongly influenced the posterior distribution. The posterior distribution for w_{HIV} on the other hand is almost identical to the prior distribution.

The estimated posterior correlation matrix for the TB parameters is given in Appendix A. The estimated posterior correlation matrix for the HIV parameters is

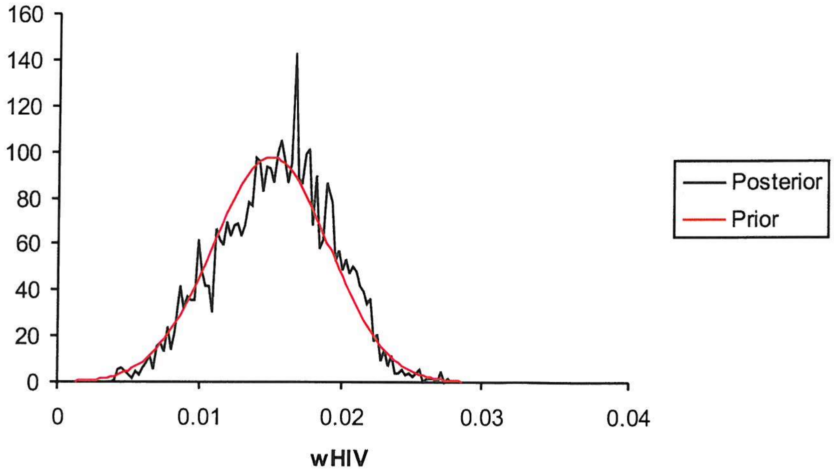

Figure 5.5: Estimated posterior distribution for w_{HIV} (rate at which non-infectious active TB becomes infectious active TB among late-stage HIV-positives), determined using MCMC sampling.

more interesting and is given in Table 5.3. This shows significant correlations between the parameters of the HIV model. As the prior distribution includes no dependence between parameters, these correlations must be coming from the likelihood function. This helps to demonstrate one of the advantages of using a Bayesian approach in this situation. Determining the confidence intervals around the predictions from models such as this has often been done by Monte Carlo sampling from the prior distributions. No account can then be taken of these correlations, and confidence intervals are therefore often predicted to be wider than they should be.

Parameter	Prior Mean	Prior Variance	Posterior Mean	Posterior Variance
λ_0	0.0095	1.00 x 10^{-6}	0.0116	6.74 x 10^{-6}
p	0.14	3.15 x 10^{-3}	0.244	4.47 x 10^{-4}
p_{HIV}	0.67	0.0250	0.892	0.0233
v	1.13 x 10^{-4}	9.10 x 10^{-9}	1.24 x 10^{-4}	5.84 x 10^{-9}
v_{HIV}	0.17	4.40 x 10^{-3}	0.107	3.78 x 10^{-4}
x	0.35	0.0163	0.313	0.0105
x_{HIV}	0.75	0.0163	0.787	0.0171
f	0.45	2.03 x 10^{-3}	0.383	2.13 x 10^{-3}
f_{HIV}	0.3	3.15 x 10^{-3}	0.435	2.15 x 10^{-3}
ϕ	0.5	0.0163	0.451	0.0159
w	0.015	1.67 x 10^{-5}	0.0107	1.53 x 10^{-5}
w_{HIV}	0.015	1.67 x 10^{-5}	0.0154	1.59 x 10^{-5}
μ	0.0185	7.33 x 10^{-7}	0.0185	6.80 x 10^{-7}
μ^{inf}	0.3	2.60 x 10^{-3}	0.430	1.43 x 10^{-3}
μ^{inf}_{HIV}	1	0.01	0.808	0.0106
$\mu^{non-inf}$	0.1	4.16 x 10^{-4}	0.109	3.43 x 10^{-4}
$\mu^{non-inf}_{HIV}$	1	0.01	0.997	9.92 x 10^{-3}
e	0.5	2.50 x 10^{-3}	0.512	2.73 x 10^{-3}
rf	0.3	0.0104	0.435	7.30 x 10^{-3}
tLS	3.79	0.156	3.51	0.0269
tD	8.72	0.0268	8.78	0.0262
α	n/a	n/a	78.1	178
β	n/a	n/a	0.128	5.13 x 10^{-4}
γ	n/a	n/a	0.0499	2.87 x 10^{-5}

Table 5.2: Prior and posterior distributions for the parameters of the TB-HIV model, estimated using MCMC sampling.

5.5.4 Estimating Confidence Intervals

Confidence intervals around each series of projected TB incidence and death rates are obtained by carrying out 1000 simulations using parameter values randomly chosen from the posterior distributions. Samples are drawn independently from the two groups of parameters (TB parameters and HIV parameters). This will result in a slight over-estimate of the confidence intervals due to the small interactions between HIV and TB parameters.

We use sensitivity analysis to identify the model parameters that most influenced our results, judging their influence from partial rank correlation coefficients calculated between each outcome measure and each of the parameters in the model [13]. These suggest that the parameters responsible for most of the uncertainty in model outputs are also those for which there is least information, i.e. those describing the effect of HIV infection on the course of TB. They are, for those with HIV, the rate of progression from co-infection to active TB, the proportion of active TB cases that is infectious, the death rate of TB cases, and the relapse rate to active TB among those who have failed treatment. The accuracy of the results depends on the structure of the TB-HIV model as well as the parameter values. Although a simpler model may still have captured the main features of the data, the model structure is the simplest that could be used to explore all of the interventions considered in this study.

α	1	-0.980	-0.903
β	-0.980	1	0.942
γ	-0.903	0.942	1

Table 5.3: Posterior correlation matrix for the HIV parameters in the TB-HIV model, estimated using MCMC sampling.

5.6 Results

Figures 5.6 to 5.8 show the projected TB incidence and HIV prevalence for each of the three scenarios for the HIV epidemic with 95% confidence intervals. If HIV prevalence declines in Kenya, we expect TB incidence to fall, even without additional interventions (Figure 5.7). The time lag between the start of the HIV epidemic and the increase in TB incidence is approximately four years. The delay is due to the time lag between becoming infected with HIV and becoming more susceptible to TB, as the TB epidemic is fuelled by those in late-stage HIV. If HIV prevalence stabilises or continues to increase, then the number of TB cases is also expected to increase (Figures 5.6 and 5.8), by about 60% for constant HIV incidence, and by approximately 70% for a 50% increase in HIV incidence.

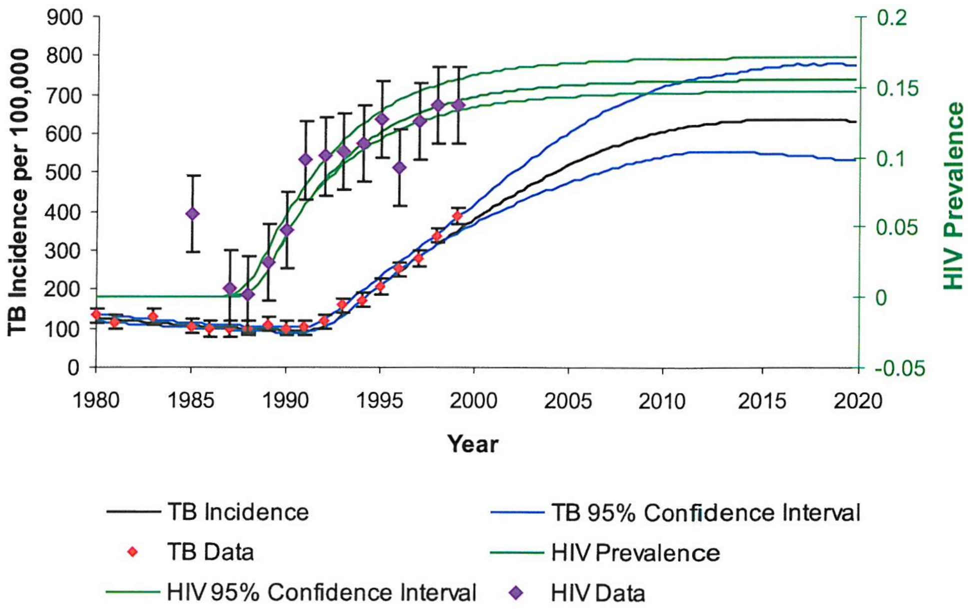

Figure 5.6: Estimated TB incidence and HIV prevalence in Kenya assuming HIV incidence remains approximately constant ($\theta = 1$). Confidence intervals are 95%.

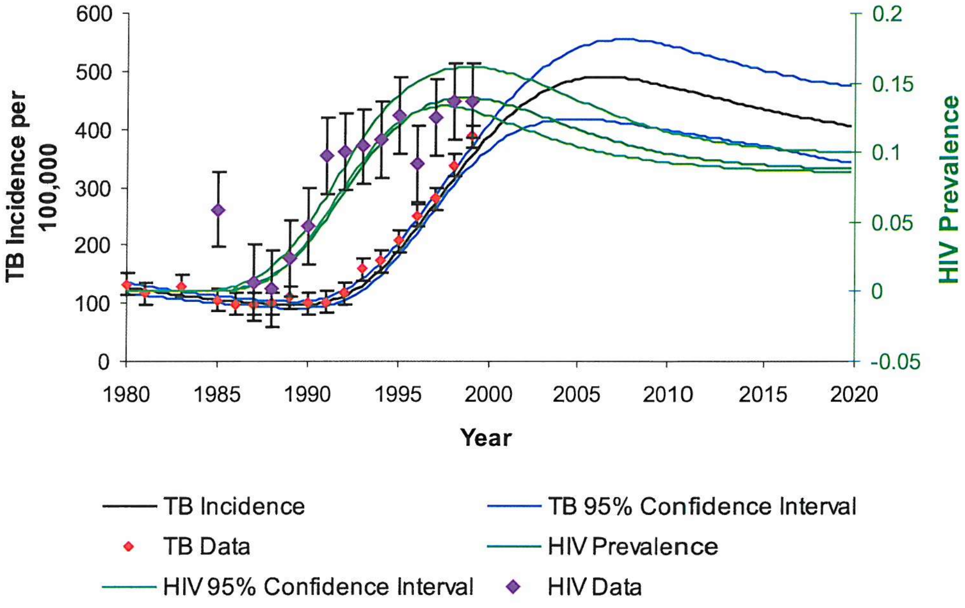

Figure 5.7: Estimated TB incidence and HIV prevalence in Kenya assuming HIV incidence declines by approximately 50% ($\theta = 0.5$). Confidence intervals are 95%.

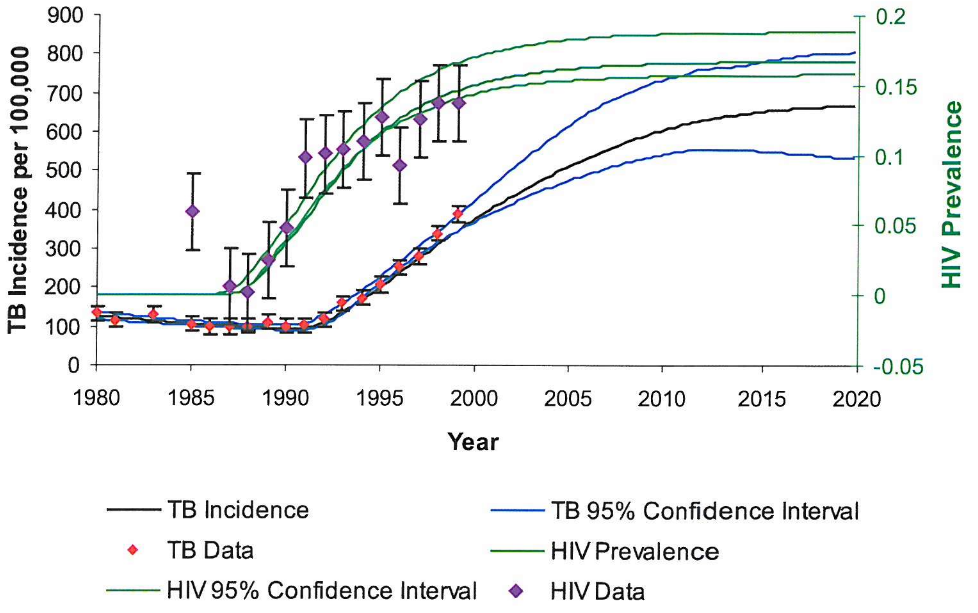

Figure 5.8: Estimated TB incidence and HIV prevalence in Kenya assuming HIV incidence increases by approximately 50% ($\theta = 1.5$). Confidence intervals are 95%.

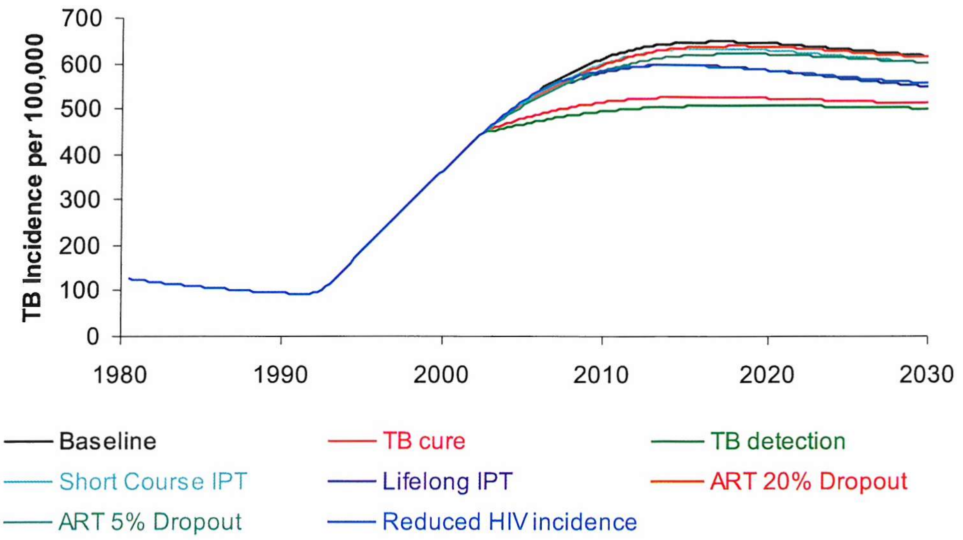

Figure 5.9: Effect of increasing intervention levels by 10% on TB incidence in Kenya, assuming constant HIV incidence in the long-term.

Figure 5.9 shows the impact on TB incidence of a 10% increase in coverage of each intervention in 2001 for the epidemic in which HIV prevalence stabilises at its current level. This demonstrates the advantage that increasing TB cure or detection rates has over implementing the other interventions, in that the impact is immediate. Reducing HIV incidence or administering TB preventive therapy have a delayed effect on TB incidence rates. Although ART has a high initial impact, when drop out from the therapy is incorporated into the modelling, its effects diminish with time.

The relative effectiveness of the different interventions is judged first by applying and maintaining the same, small improvements in coverage, and recording consequent reductions, over 10 years, in the numbers of new TB cases and TB deaths. Figure 5.10 shows the number of TB cases averted over 10 years when

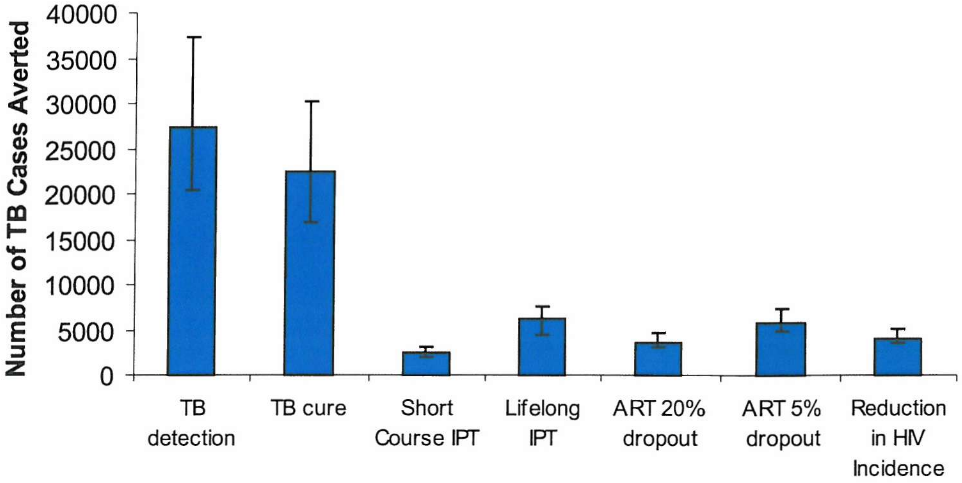

Figure 5.10: Effect of increasing intervention levels by 1% on the number of cases of active TB over ten years, assuming constant HIV incidence in the long-term. Confidence intervals are 95%.

HIV incidence is assumed to remain approximately constant. Figure 5.11 shows the number of deaths averted by the different interventions over 10 years, assuming constant HIV incidence in the long-term.

Larger HIV epidemics generate larger burdens of TB, and so more cases (Figure 5.10) and deaths (Figure 5.11) are averted by each intervention. In all of the scenarios considered, the most effective way to reduce TB incidence is by increasing TB case detection and cure rates (Figure 5.10). Reducing HIV incidence or administering ART or preventive therapy for TB never appear to be highly effective interventions.

The most effective way to avert TB deaths (Figure 5.11) is by improving case detection. For both TB cases and deaths, TB preventive therapy is relatively ineffective, although the effectiveness improves for lifelong treatment.

Since unit changes in the coverage of very different interventions are unlikely

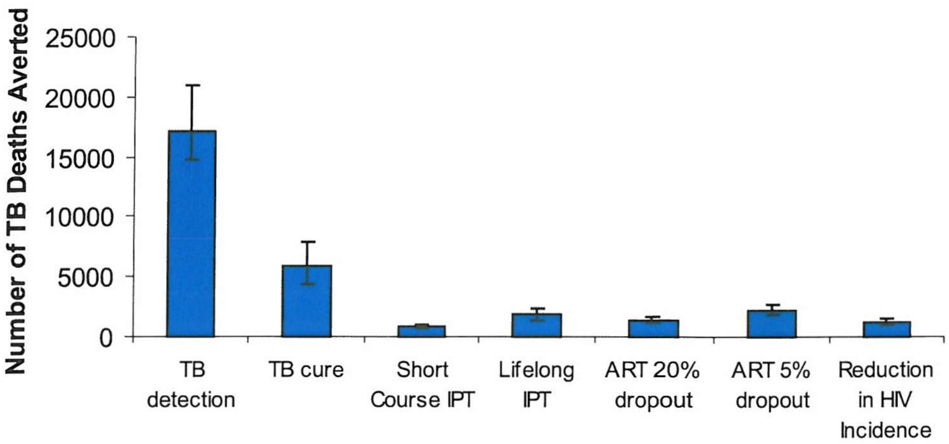

Figure 5.11: Effect of increasing intervention levels by 1% on the number of TB deaths over ten years, assuming constant HIV incidence in the long-term. Confidence intervals are 95%.

to be equally feasible or equally costly, these results are more of interest than of use to policy-makers. In Chapter 6 we extend this analysis to incorporate the costs of implementing the different interventions and evaluate the cost-effectiveness of a number of intervention strategies.

A third way of comparing interventions is to ask what improvement in coverage would be needed to match the impact of a 5% increase in the case detection rate, over the baseline level of 50%. We calculate that the same reduction in the number of TB cases over 10 years could be obtained by any of the following means: reduce HIV incidence by 50%; increase the coverage of ART from 0% to 90%, assuming 20% dropout each year; provide six months TB preventive therapy to 90% of all HIV-infected persons; or increase the TB cure rate from 70% to 79%. Thus, all interventions, except augmenting the TB cure rate, require relatively large increases in coverage to compete with a 5% improvement in case detection.

In order to assess how generally our results apply, we conducted a similar

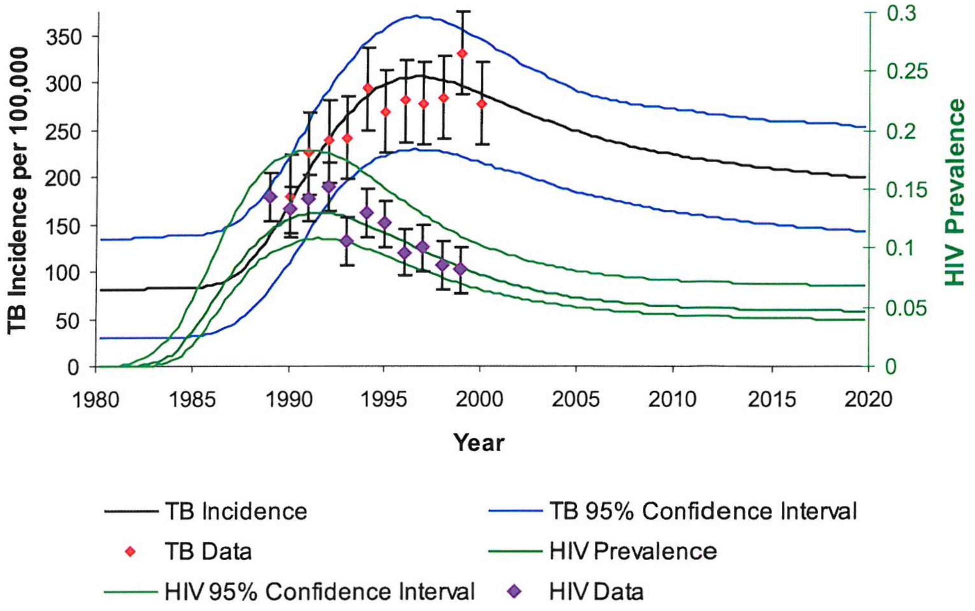

Figure 5.12: Projected TB incidence and HIV prevalence in Uganda assuming HIV incidence declines to 25%. Confidence intervals are 95%.

exercise for Uganda, which has an earlier epidemic than Kenya, and for South Africa, which has a later epidemic than Kenya. We assume that HIV incidence in Uganda will decline to 25% of its peak and in South Africa to 50% of its peak. The projected HIV and TB epidemics along with predicted reductions in the number of TB cases and deaths with increases in intervention levels are shown in Figures 5.12 and 5.13. We expect the number of TB cases in Uganda to be declining, irrespective of any change in control efforts, because the prevalence of HIV peaked in the early 1990s and has fallen by about 50% since then.

As HIV incidence is falling in Uganda [60], the measures aimed at curbing the effect of HIV on TB, such as ART or reducing HIV incidence, are much less effective than improving TB case detection and cure at averting TB cases and deaths (Figure 5.14). By contrast, South Africa appears to be on the threshold of a very large TB epidemic, driven by HIV. With no additional interventions, we

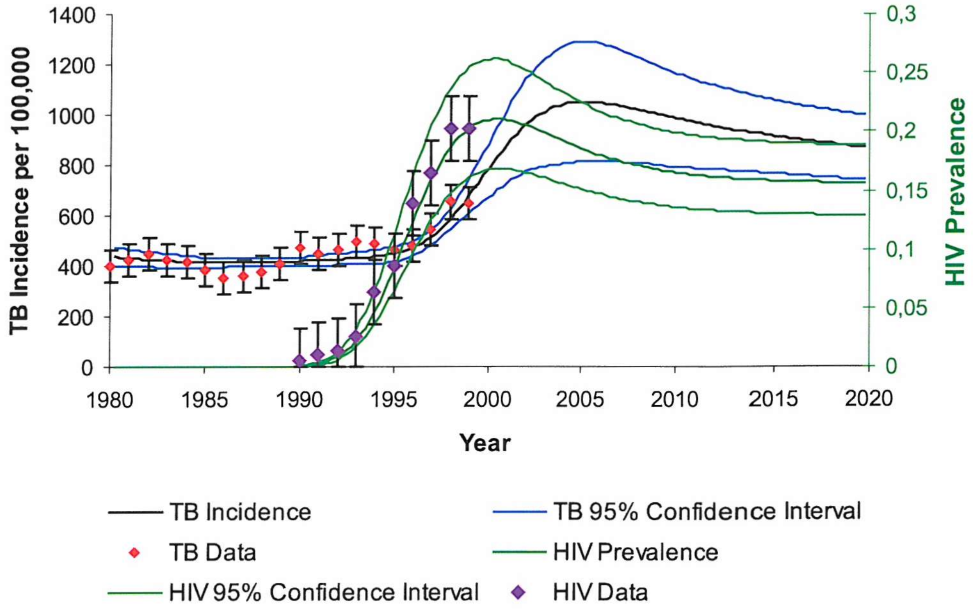

Figure 5.13: Projected TB incidence and HIV prevalence in South Africa assuming HIV incidence declines to 50%. Confidence intervals are 95%.

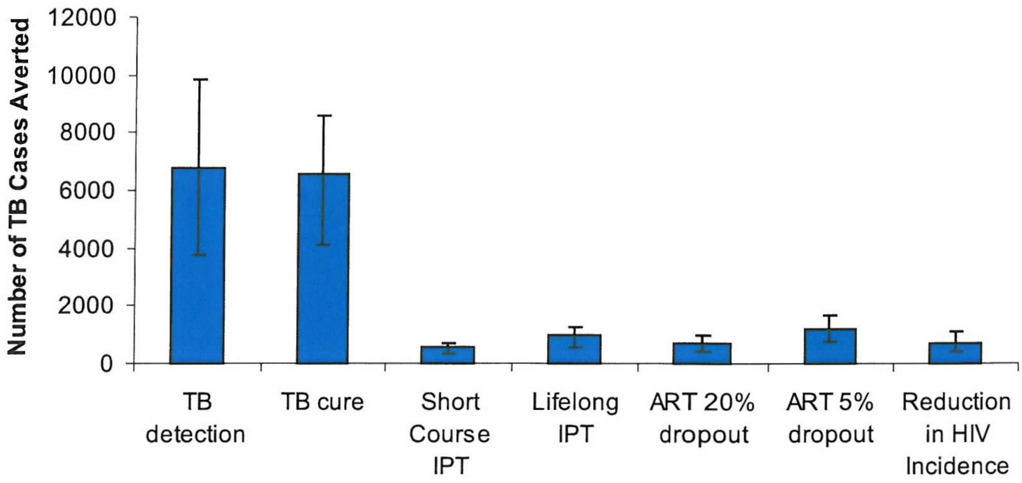

Figure 5.14: Number of TB cases averted in Uganda by increasing intervention levels by 1% from base. Confidence intervals are 95%.

forecast a 60% increase in TB incidence from 1999 levels, before 2010, whatever the future course of the HIV epidemic. Despite the different characteristics of the South Africa epidemic curative measures are, per unit improvement in coverage, still the best way to diminish TB incidence, as for Uganda (Figure 5.15).

5.7 Discussion

5.7.1 Study Results

The results of this study suggest that the best way to manage TB epidemics driven by HIV over the next five to ten years is to find and treat TB cases, rather than to prevent or mitigate the effects of HIV infection. These results are robust to uncertainties in the values of model parameters and are similar for early (Uganda), intermediate (Kenya) and late (South Africa) epidemics. The principal explanation for this finding is that curative measures reduce deaths and decrease transmission

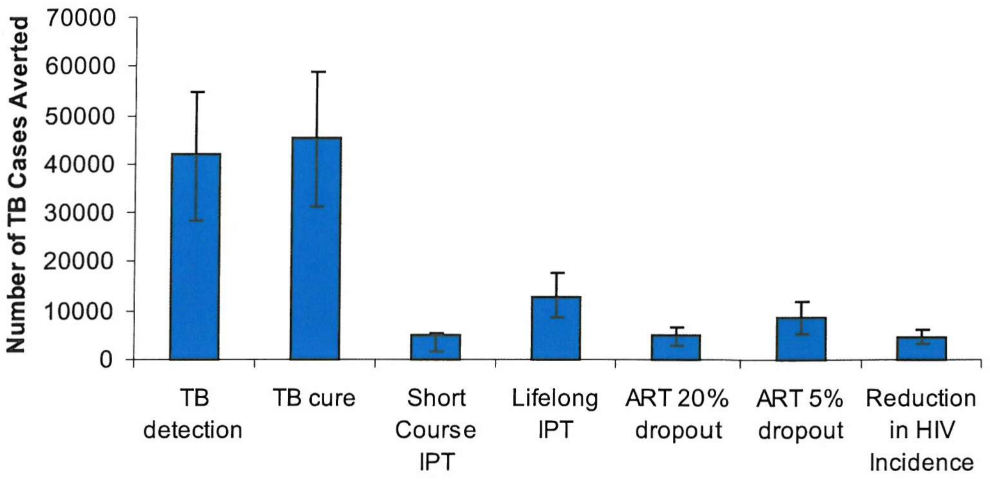

Figure 5.15: Number of TB cases averted in South Africa by increasing intervention levels by 1% from base. Confidence intervals are 95%.

immediately in all TB patients, irrespective of whether patients are infected with HIV. By contrast, the preventive methods are directed at people co-infected with TB and HIV, who typically represent only one third to one half of the sources of new TB cases in eastern and southern Africa [30]. In addition, whilst preventing HIV infection removes the underlying cause of rising TB incidence, the benefits only begin to appear after approximately four years [81], [4], the time lag between HIV infection and late-stage HIV (WHO stage three).

National TB control programmes in many African countries are already implementing the WHO DOTS strategy [79], which gives curative measures an additional practical advantage, because coverage can be improved by strengthening existing programmes.

Even if DOTS is necessary to contain the HIV-related epidemics of TB, it may not be sufficient to bring such epidemics under control for two reasons. First, although curing TB cases is relatively effective, the results of this analysis suggest that curative programmes on their own will stabilise, but not reverse, TB incidence and deaths. Second, methods for preventing and ameliorating the effects of HIV infection will be essential for tackling AIDS in general, as distinct from HIV-related TB in particular. The principal recommendation from this initial study is that national TB programmes in areas of high HIV prevalence should continue to strengthen their curative services, using preventive measures in addition to, but not as a substitute for, finding and treating active TB cases.

5.7.2 Methodology

With both the Markov Chain Monte Carlo sampling (MCMC) and the importance sampling, knowledge of the posterior function, especially of its mode, is necessary to ensure convergence, and the computing time used by the optimization routine to find the maximum of the posterior distribution seems to be better spent in this manner than it would be performing additional MCMC or importance sampling runs.

The MCMC works well for most scenarios, obtaining good convergence for approximately 80,000 model runs (3000 runs warm up and 5000 runs for each of 5 chains with separate runs for the TB and the HIV parameters). Some scenarios are more troublesome, most notably Uganda, where the data is worse than in Kenya and South Africa and the epidemic characteristics are very different. In Uganda the HIV prevalence data show a decline from a maximum in the early 1990s and no data are available for the years in which HIV prevalence was increasing. This may introduce some ambiguity into the fitting of the HIV parameters.

Importance sampling shows poor convergence, diagnosed by the existence of very high-valued weights, corresponding to points in parameter space at which the candidate distribution is very low and the posterior distribution relatively high. This suggests that the candidate distribution is not a good enough approximation to the posterior distribution. Importance sampling did work reasonably well in a previous study [36], and we suspect that the reason why convergence was so poor in this situation was due to the non-normal shape of the posterior distribution for the HIV parameters. High-valued weights skew the estimate of the posterior distribution of the model parameters, as Figure 5.3 shows, and in so doing lead to incorrect estimates of results such as TB incidence and intervention effectiveness, and the confidence intervals around these results.

We found the convergence of MCMC to be less dependent on the knowledge of the posterior distribution prior to sampling than importance sampling. As MCMC is an adaptive sampling procedure, this is to be expected. An additional advantage of MCMC is that the output, when the warm up has been removed, is a sample from the posterior distribution. This makes the sampling of parameters from their posterior distributions, e.g. for the estimation of uncertainty intervals on model results, easier than with importance sampling, where each of the sets of parameters must be weighted by the ratio of the posterior probability to candidate probability.

Chapter 6

Cost-Effectiveness Analysis of TB and HIV Interventions

6.1 Introduction

This chapter describes the cost-effectiveness analysis of interventions against tuberculosis (TB) and HIV, extending the analysis of Chapter 5. The work described here is not directly related to the main academic thread of the thesis; however the successful application of the Bayesian methodology to the initial study of the effects of different interventions led to a fuller requirement for an economic analysis. Introducing costs to the analysis of Chapter 5 allows us to measure the effort involved with increasing intervention levels on the same scale: that of money. Using more generic measures of effectiveness such as disability adjusted life years (DALYs) gained also enables a fairer comparison of interventions against HIV and interventions against TB. The work described in this chapter is therefore of great practical use to policy-makers.

As discussed in Chapter 5, TB remains the most common opportunistic infection associated with HIV in Sub-Saharan Africa, and interventions aimed at either disease must be considered in the context of a joint epidemic. Where budgets are limited, decisions must be made as to which interventions should be prioritized and implemented first. Cost and cost-effectiveness analysis can play an important role in this decision process, because they allow an assessment of which of many competing interventions are affordable, and which provide the best value for money.

A recent systematic review [33] identified 24 cost-effectiveness studies of 31 different HIV prevention, treatment and care interventions in sub-Saharan Africa that allowed cost-effectiveness to be assessed using a generic indicator of effectiveness (DALYs averted). Several studies of the cost-effectiveness of TB treatment in sub-Saharan Africa have also been undertaken, and two recent reviews are available [14], [48]. However, the existing studies have three important limitations. First, almost all studies consider only one intervention rather than comparing a range of interventions in the same setting. This limits the extent to which fair comparisons among interventions can be made. None of the published studies consider a range of TB/HIV interventions simultaneously. Second, they employ different approaches to transmission of both HIV and TB; in some studies, transmission is not considered and where it is, the methods for estimating the costs and effects associated with an intervention's impact on transmission vary. The only cost-effectiveness study of TB treatment in Africa that has incorporated transmission in the analysis focused on the treatment of HIV-negative patients, and cost per DALY averted figures that applied in high HIV prevalence settings were not reported [69]. Third, few studies consider the total number of people that would need to receive an intervention if existing policy was implemented and control targets met, and few analyse the related total costs, effects, affordability and cost-effectiveness of interventions. For example, all cost-effectiveness studies of TB treatment relate to existing levels of case detection and cure. They do not assess the cost-effectiveness of improving case detection and cure rates beyond their existing levels, even though this is needed if global TB control targets are to be achieved. None of the limited number of cost-effectiveness studies of antiretroviral treatment (ART) relate to the coverage levels needed to achieve the World Health Organization's recently announced goal of enrolling three million people on ART by 2005 (the "3 by 5 initiative").

In Chapter 5 and [36], we compared the effects of several strategies to reduce the burden of TB and HIV in high HIV prevalence countries in Africa. This analysis used a mathematical model that allowed impacts on transmission to be considered in a consistent way. Here, we extend this work to assess the costs, effects, affordability and cost-effectiveness of six strategies for reducing the burden of TB and HIV, using data for Kenya. Each strategy relates to existing targets or policy for TB control and ART enrolment, and in each case we include assessment of the total number of people that would need to be reached. The analysis follows that described in [35], but uses the model described in Chapter 5 rather than that described in [36].

6.2 Methods

6.2.1 Country and Strategies Considered

Our analysis focuses on Kenya. Kenya has an HIV epidemic that is typical of the region, good data on the prevalence and incidence of HIV and TB are available, and detailed costing studies of TB treatment have recently been undertaken.

We considered six strategies for reducing the burden of TB and HIV in Kenya. These were

1. Improving TB case detection rates so that the WHO target of 70% is reached in 2005 and then sustained

2. Improving TB cure rates so that the WHO target of 85% is achieved in 2005 and then sustained

3. Simultaneously improving both TB case detection and cure rates so that both WHO targets are met in 2005 and then sustained (DOTS)

4. Providing ART so that the targets for enrolment included in the recent WHO "3 by 5" initiative are met i.e. 50% of the estimated population in need receives treatment

5. Providing isoniazid preventive therapy (IPT) to HIV-positive individuals without TB for six months

6. Providing IPT for life to HIV positive individuals without TB

All six strategies were assessed for the ten year period 2005-2014, and compared with a scenario (which we term the baseline scenario) in which interventions continue at their existing levels. This means a 50% TB case detection rate, a 70% TB cure rate, and no implementation of either ART or preventive therapy (we acknowledge that there is some provision of ART and preventive therapy in Kenya, but this is very limited).

We also analyse the cost-effectiveness of reducing HIV incidence, but this was assessed differently from the other interventions for reasons that are discussed in Section 6.2.3

6.2.2 Analysis of Numbers to be Treated, Costs and Effects

The numbers to be treated in each strategy, and the associated costs and effects, were estimated using the mathematical model described in detail in Chapter 5. We focus on the scenario in which the HIV epidemic stabilises at a prevalence of 14% in adults, the value observed in ante-natal clinic surveys in Kenya in 1999 [75].

The model was extended to include the annual numbers of patients detected and treated (for strategies to improve TB case detection and cure rates, and preventive therapy for six months) and the annual person years of treatment (ART, lifelong preventive therapy) as model outputs. Unit costs of TB detection and treatment, one year of ART, and a six month course of preventive therapy were also incorporated into the model and were used, in combination with the model estimates of the numbers detected, numbers treated or the person years of treatment, to produce the total annual costs of each strategy as model outputs.

Costs incurred in future years were discounted at 3%, in line with recent international guidelines [90], [56]. Costs were assessed from the perspective of the health system only (i.e. costs incurred by patients themselves were not included) in year 2003 US$. It is important to highlight that because ART will defer costs associated with treatment of AIDS-related opportunistic infections (OIs) and palliative care, our analysis allowed for treatment savings arising from the provision of ART. For each year, costs for the treatment of OIs and palliative care were estimated as the total people years of treatment multiplied by the average annual cost of such treatment. The total people years of treatment were based on the numbers with AIDS (estimated as a fixed proportion - 40% - of the numbers in late stage HIV) and the fraction assumed to access care (assumed to be 50%). As ART reduces the numbers of people with AIDS, the total annual cost of OI treatment and palliative care is lowered when the strategy of providing ART is implemented. The cost parameters used, and the related assumptions and sources of data, are given in Appendix B.

The measure of effectiveness used in this analysis is the number of disability adjusted life years (DALYs) gained by each of the interventions. Our previous analysis focused on TB deaths and TB cases averted, but a fair comparison of the cost-effectiveness of interventions requires that the analysis captures a) differences in the years of life gained from averting deaths in HIV-positive and HIV-negative individuals, and b) the prevention of deaths from causes other than TB. If this is not done, the analysis will be biased against interventions that prevent relatively higher numbers of deaths in HIV-negative individuals and/or deaths unrelated to TB. The mathematical model does not include age structure; therefore, we estimate the average number of DALYs gained by averting a death among HIV-negative TB patients, HIV-infected TB patients and HIV-positives to be

$$\begin{aligned} DALY = &\sum_{i=1}^{n} p_i \left[\frac{KCe^{ra_i}}{(r+\beta)^2} \left[e^{-(r+\beta)(L_i+a_i)} \left(-(r+\beta)(L_i+a_i)-1\right) \right.\right. \\ &\left.\left. -e^{-a_i(r+\beta)} \left(-(r+\beta)a_i - 1\right) \right] + \frac{1-K}{r}\left(1-e^{-rL_i}\right) \right], \end{aligned} \tag{6.1}$$

based on the standard equation for a DALY averted as given in [70]. Here a_i is the average age in age group i, p_i is the proportion of deaths in age group i for the population under consideration, L_i is the life expectancy for someone in the given population at age a_i, K is the age weighting modulation factor, C is a constant, r is the discount rate and β is the parameter from the age-weighting function. The values of K, C and β come from [49] and are given in Table 6.1.

Parameter	Value
K	1
C	0.1658
r	0.03
β	0.04

Table 6.1: Parameter values used in the calculation of disability adjusted life years (DALYs) gained.

Using Equation 6.1, life expectancy data for Kenya [103], evidence that life expectancy among HIV-positive TB patients is approximately three years [72], [5], [6], and the assumption that the death rate among TB patients is the same in each age group, we estimate that the DALYs gained by averting a TB death in an

HIV-positive TB patient would be 4 years, and that the gain in an HIV-negative individual would be 24 years. To capture the effect of ART on non-TB related mortality, we further assume that 1 DALY is averted for each person year of ART. To avoid double counting of deaths, we assume that ART can only affect TB deaths among HIV-negatives and not those among HIV-positives. For consistency with the analysis of total costs, DALYs averted in future years were discounted at 3%.

The model was run for the baseline scenario and the addition of each of the six intervention strategies to the baseline. When considering the baseline, cost-effectiveness is estimated as the net change in costs from a situation where no interventions are applied, divided by the net increase in DALYs averted from a situation in which no interventions are applied. The six intervention strategies were applied individually, allowing comparison between strategies, which would not have been possible if the model had been run with all strategies applied simultaneously. Cost-effectiveness was calculated as the net change in costs from the baseline scenario divided by the net increase in the number of DALYs averted compared with the baseline scenario. Uncertainty intervals were obtained by sampling 1000 sets of model parameters from the output of the Markov Chain Monte Carlo sampling (MCMC), while costs were sampled from the distributions given in Table B.1. The uncertainty intervals therefore simultaneously incorporate uncertainty about unit costs, the numbers given interventions in each scenario, and effects.

In order to assess the how generally applicable our results are, we obtained results for a number of possible scenarios for Kenya:

1. The implementation of the six intervention strategies occurring at 50% and 25% of the rate required to meet the targets specified above

2. Assessment of the results over five and twenty years

Results for these scenarios will not be presented here but will be discussed in Section 6.3.

6.2.3 Reducing HIV Incidence

In Chapter 5, we considered the effect of reducing HIV incidence. The model cannot be used to estimate the total costs of implementing interventions aimed at reducing HIV incidence because the effect of reducing HIV incidence is explored simply by changing the assumed trajectory of the HIV epidemic, with no consideration of the specific interventions that would be required to achieve this and how many people they would need to reach. We therefore estimated the threshold costs per HIV infection averted at which reducing HIV incidence would have the same cost-effectiveness as the other six strategies, and compared these with existing published data. For HIV prevention, we estimate that 22 DALYs are averted for each HIV infection averted, based on the the standard DALY formula (Equation 6.1) and demographic data for Kenya.

6.3 Results

In the baseline scenario, the only interventions offered are treatment for active TB, with a cure rate of 70% and a case detection rate of 50%, and treatment for AIDS-related OIs and palliative care. The model estimated that 96,000 (95% confidence intervals [81,000, 120,000]) people are treated per year for TB, at a total cost of US$ 17 million [US$ 15 million, US$ 21 million], and 185,000 [175,000, 195,000] people receive treatment and care for AIDS, at a cost of US$ 37 million [US$ 33 million, US$ 41 million]. This results in 2 million [1.5 million, 2.8 million] DALYs being averted per year compared with a situation in which no interventions are offered. The cost per DALY averted is US$ 8.70 [US$ 6.60, US$ 11.10].

The numbers of people treated for TB per year for each of the six strategies to reduce the burden of TB and HIV are shown in Figure 6.1. Numbers increase for most interventions because of population growth. Some other trends are worth noting. Increasing TB case detection or implementing the DOTS strategy result in an initial increase in the number of TB cases being treated. After several years, however the impact of this improved control strategy on TB transmission becomes obvious as the number of TB cases needing treatment drops and the number being treated for TB under the DOTS strategy is lower than for any other intervention strategy. Increasing TB cure rates also results in smaller numbers being treated, because those who are treated are more likely to recover. Administration of ART also reduces the number of people given TB treatment. Those taking ART are assumed to have the same risk of developing TB as someone who is HIV-negative, thus reducing the expected number of TB cases. Other interventions have little effect on the numbers being treated for TB.

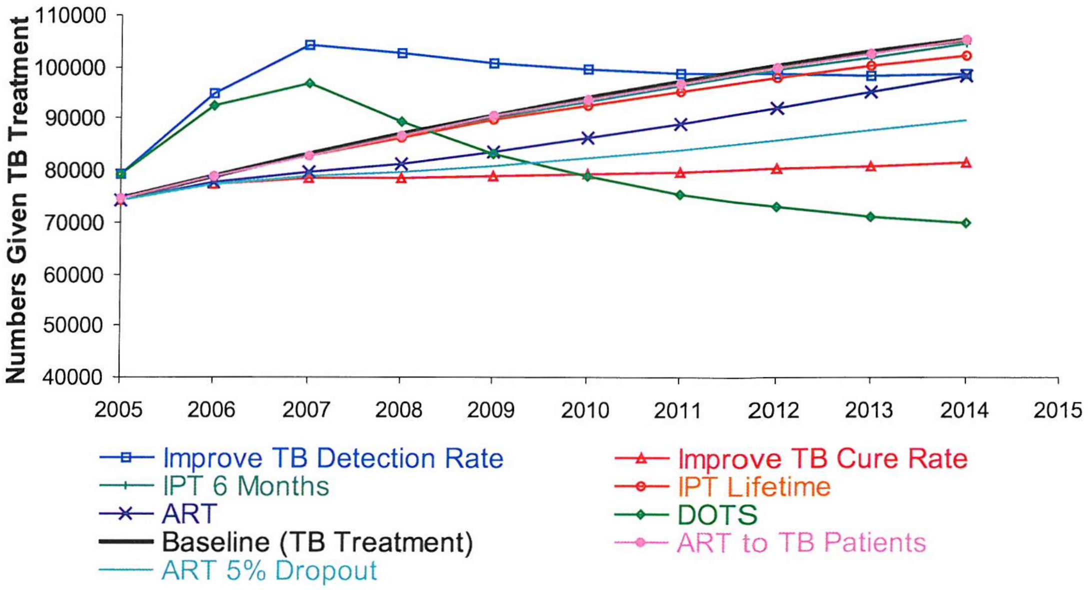

Figure 6.1: Numbers of people given TB treatment under each of the six intervention strategies.

As shown in Figure 6.2, when preventive therapy is provided for six months, the average number on treatment each year is stable at around 20,000 per year. When provided for life, there is a steady increase in the numbers on treatment from zero to 180,000 after 10 years.

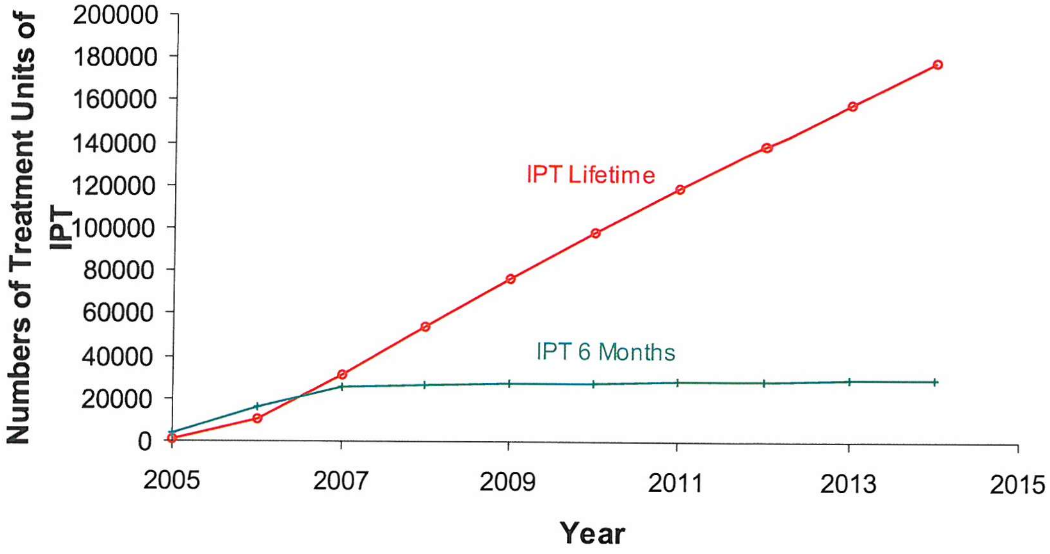

Figure 6.2: Numbers of people given TB preventive therapy under the two TB preventive therapy intervention strategies.

Provision of ART so that the "3 by 5" target for Kenya is met and then followed by enrolment of 50% of those in need of treatment, is associated with an increase from less than 10,000 on treatment in 2005 to 490,000 after 10 years, as shown in Figure 6.3. If the annual drop out rate from ART were 5% rather than 20%, numbers taking ART would increase to 880,000 after 10 years.

The numbers of people with AIDS receiving OI treatment and palliative care for the different intervention strategies are shown in Figure 6.4. Only administration of ART results in a substantial reduction (10% on average over the ten years) in the numbers of people with AIDS receiving treatment. As more people drop

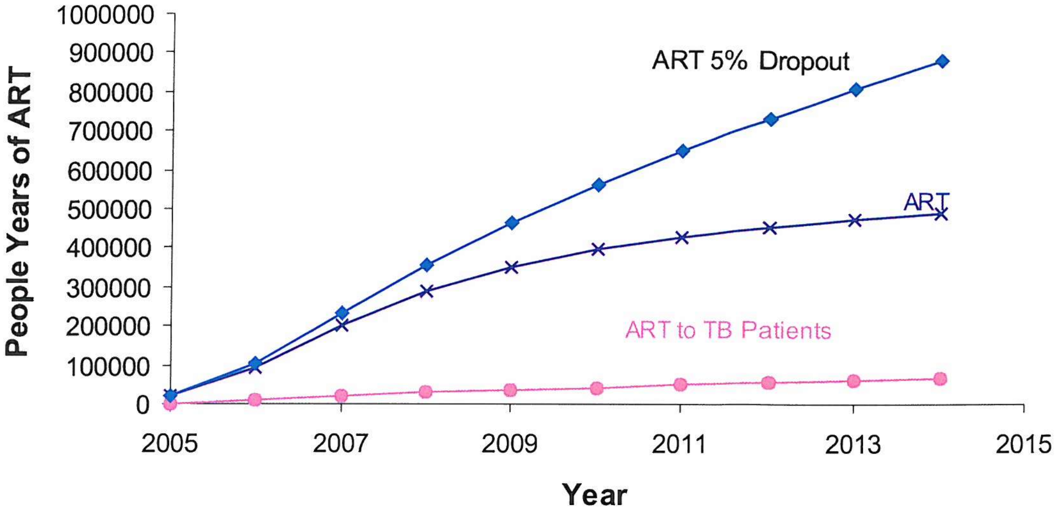

Figure 6.3: Numbers of people given antiretroviral therapy under different scenarios for dropout and administration.

out of antiretroviral treatment, the numbers given treatment for OIs and palliative care start to increase again, a trend that is not so pronounced with the lower annual dropout rate of only 5%.

The change in total annual costs (including cost-savings associated with reductions in the number treated for TB and AIDS-related OIs and palliative care) compared with the baseline situation is shown in Figure 6.5. Improving TB case detection results in a slight increase in costs (average US$ 2.6 million per year), while improving cure rates reduces costs (average of US$ 1.9 million per year). Provision of preventive therapy increases costs by between US$ 0.9 million (for 6 months of treatment) and US$ 5.3 million (lifetime treatment) per year, both of which are a small percentage of existing total health care expenditure. The most dramatic change in costs is for provision of ART: in 2014, ART will cost just over US$ 200 million per year more than the baseline strategy — greater than total gov-

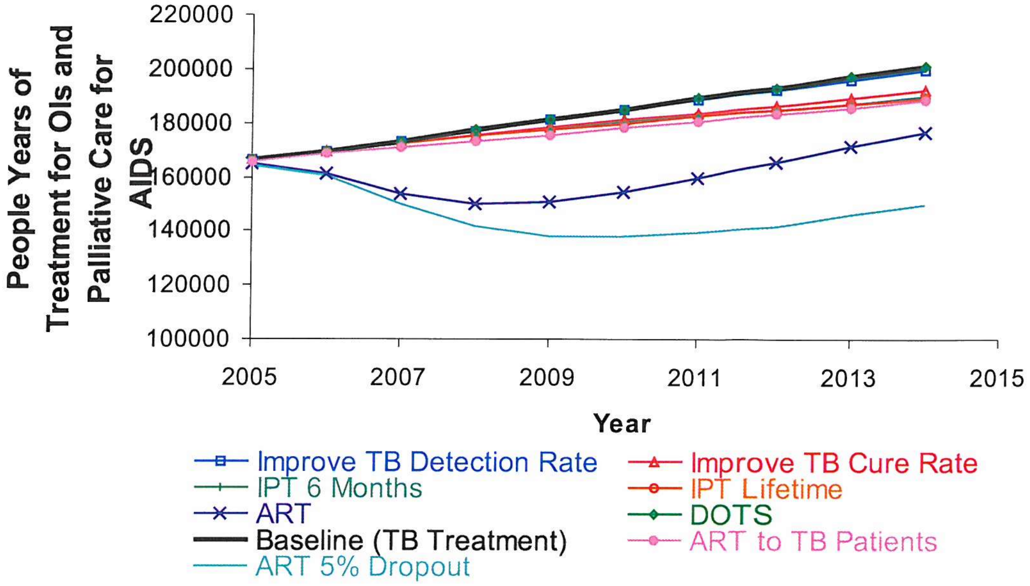

Figure 6.4: Numbers of people with AIDS receiving treatment for opportunistic infections and palliative care.

ernment health expenditure in 2000. This equates to an average annual cost of administering ART of US$ 163 million over the ten years.

We measure the effectiveness of the different intervention strategies by the number of disability adjusted life years (DALYs) that they avert, and the expected annual numbers averted over the next ten years are given in Figure 6.6. Provision of ART averts the most DALYs over the ten years, followed by simultaneously improving TB case detection and cure rates (DOTS) and increasing TB case detection rates. Other strategies avert far fewer DALYs. As people drop out of ART, it becomes less effective than other interventions, and it averts fewer DALYs than DOTS in the period 2012 to 2014.

The cost per DALY averted varies widely (Figure 6.7). Improving TB cure rates saves DALYs and lowers costs, and thus has a negative cost per DALY averted. Improving case detection has a very low cost per DALY averted

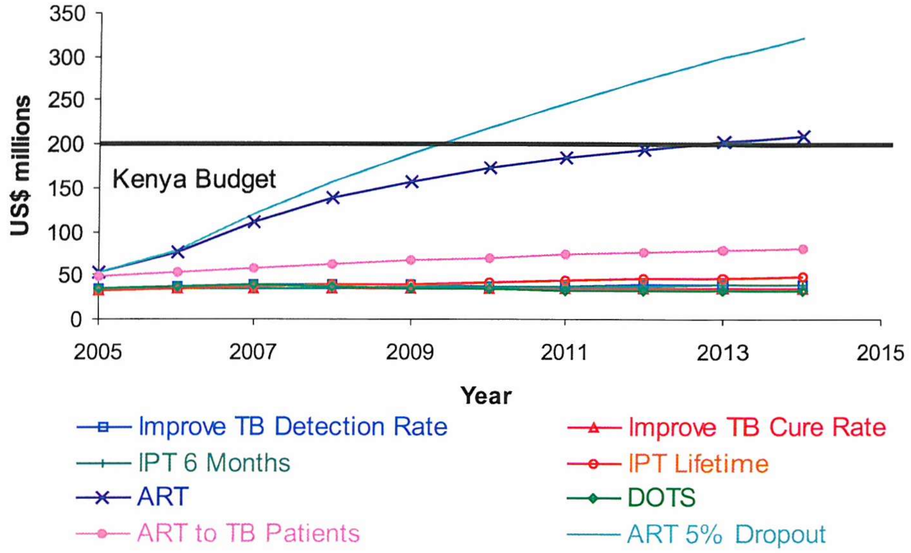

Figure 6.5: Additional annual costs over baseline for the six intervention strategies and government health expenditure for Kenya in 2000.

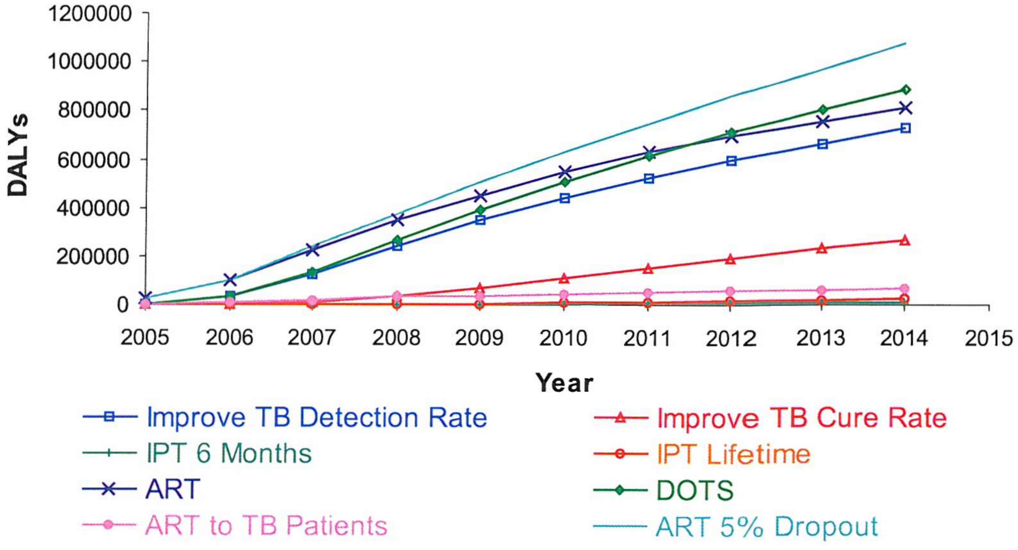

Figure 6.6: Annual DALYs averted for the six intervention strategies.

(US$ 7 [US$ 3, US$ 12). Other interventions cost more – the mean costs per DALY averted for TB preventive therapy for 6 months and for ART are approximately US$ 225, although the 95% confidence interval for the cost-effectiveness of short-course TB preventive therapy is wide, between US$170 and US$290. The strategy with the highest cost per DALY averted is provision of lifetime preventive therapy (mean US$ 690 [US$500, US$910]). Decreasing the dropout rate from ART to 5% will result in a higher cost per DALY averted of US$ 268 [US$265, US$271].

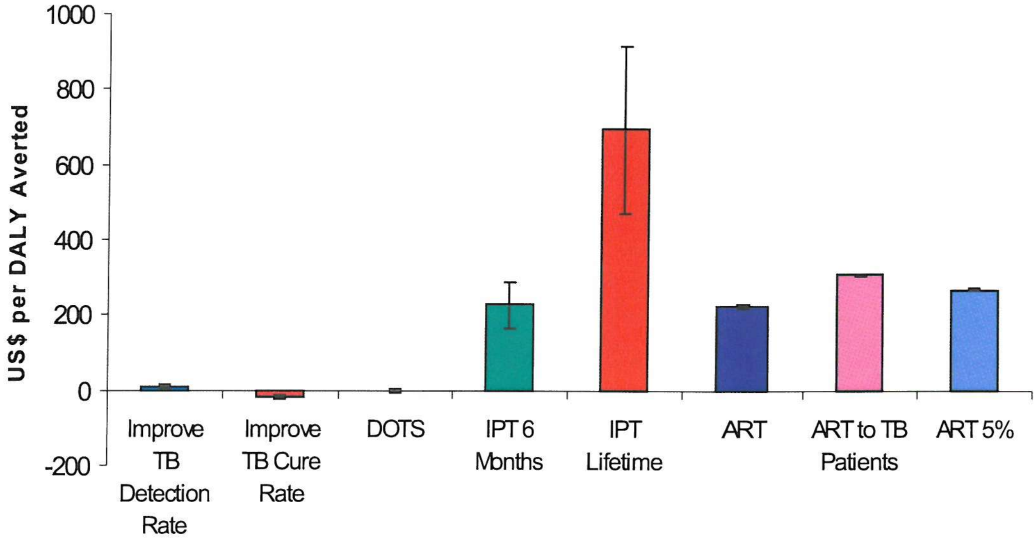

Figure 6.7: Average cost per DALY averted for the six intervention strategies.

Improving TB case detection and cure rates simultaneously could reduce TB incidence to 284 per 100,000 by the end of 2014, 60% of the estimated incidence rate for 2004 (Figure 6.8). Increasing either the case detection or cure rates independently gives a smaller effect, with a reduction of 5% by the end of 2014 for increases to the cure rate and of 20% for increases to the detection rate. Provision of ART results in an increase in the TB incidence, with TB incidence just 7%

below that of the baseline scenario in 2014. Preventive therapy has only a small effect on TB incidence.

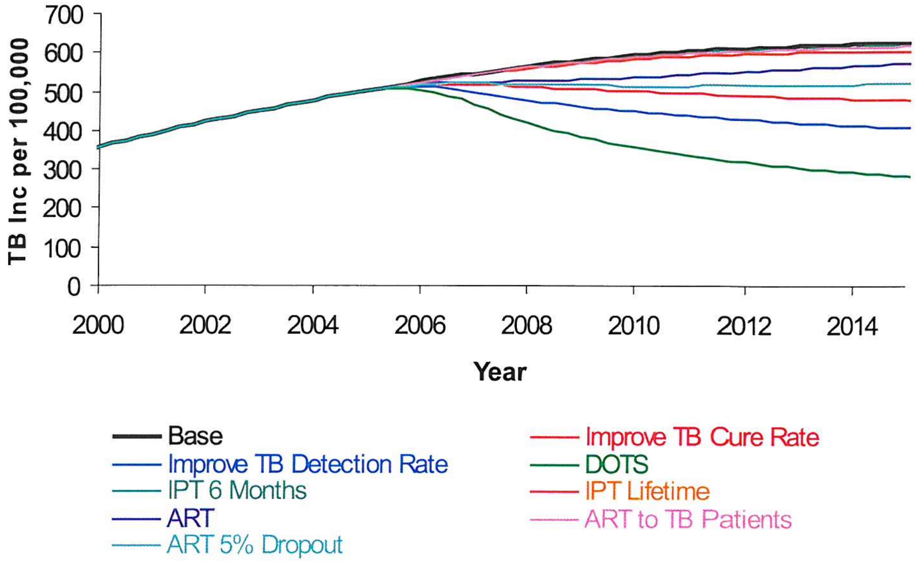

Figure 6.8: TB incidence over time for the six intervention strategies.

The picture is similar for the effect of interventions on deaths from TB, shown in Figure 6.9, with DOTS cutting the rate of TB deaths by just under 50% by the end of 2014, in line with the Millennium Development Goals, which state that TB deaths should be reduced by 50% by 2015. Improvements to the TB case detection rate have a proportionally greater effect on TB deaths than TB incidence, reducing the number of TB deaths to 142 per hundred thousand by the end of 2014. ART has a slightly greater effect on TB deaths than on TB incidence, with TB deaths per year 9% lower than baseline at the end of 2014, but still higher than the number of TB deaths per year in 2004.

The threshold costs per HIV infection averted for HIV prevention strategies compared with the other intervention strategies are given in Table 6.3. We can see that the threshold costs are all relatively high. The threshold cost compared with increasing the TB cure rate is negative because, even taking into account the

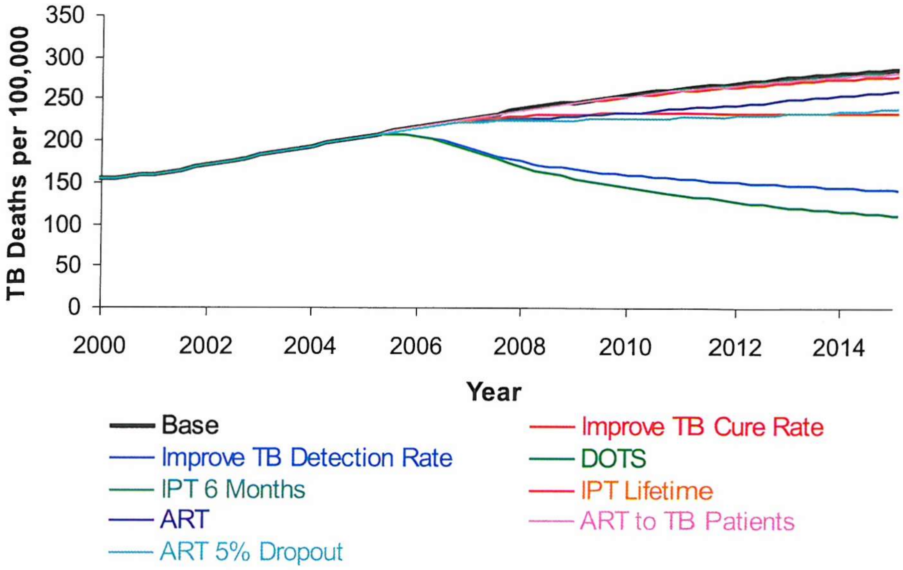

Figure 6.9: TB deaths per year for the six intervention strategies.

additional cost of improved treatment, increasing the TB cure rate is still more cost-effective than spending no money on reducing HIV incidence.

Intervention	Threshold Cost per HIV Infection Averted
Increasing TB Detection Rate	$208
Increasing TB Cure Rate	-$215
DOTS	$76.90
Administering TB Preventive Therapy (6 Months)	$3 600
Administering TB Preventive Therapy (Lifetime)	$10 800
Administering ART	$3 560

Table 6.2: Threshold costs per HIV infection averted to be as cost-effective as the other intervention strategies.

We can compare the threshold costs of reducing HIV incidence with the results presented in [33] for the costs per HIV infection averted of some standard HIV prevention strategies. Based on these, we can conclude that condom distribution of provision of blood safety measures could be more cost-effective than increasing TB treatment to DOTS levels or increasing the TB detection rate. Giving preventive therapy for TB or administering ART are both less cost-effective than all of the strategies for reducing HIV incidence discussed in [33].

The cost-effectiveness of the different interventions measured over periods of five and twenty years, and with different rates of progress toward targets, was very similar to those for the scenario presented here, where results were measured over ten years. We did not consider the effect of the HIV epidemic only being half that predicted by the antenatal clinic data from Kenya. In a previous study [35], this was found to reduce the numbers given ART and preventive therapy for TB (as this is given only to HIV-positives), and to reduce the numbers of DALYs averted by these interventions. Consequently, simultaneous improvements to TB case detection and cure rates were found to be the most effective interventions at gaining DALYs.

6.4 Discussion

6.4.1 Main Findings

The results suggest that the priority for TB programmes in high HIV settings should be to concentrate on doing better what they already do, i.e. improving TB treatment by increasing the TB cure and detection rates. This has been shown to be more cost-effective than the other interventions considered and is affordable with existing national health budgets.

Providing ART at the levels suggested in the "3 by 5" initiative has the potential to avert the most DALYs, 15% more than implementation of DOTS over the period 2005 to 2014. However realising this potential will require significant new funding, equivalent to a doubling of annual health spending in Kenya by 2013. Even if the money is made available from other external sources, the problem of absorbing a doubling in annual health expenditure over such a short space of time will remain.

Although low cost, the cost-effectiveness of IPT for 6 months is approximately equal to that of ART. The higher cost of lifelong IPT, and the small additional benefit associated with extending treatment beyond 6 months, make it a much less cost-effective strategy.

Condom distribution or improvements to blood safety could be more cost-effective than all of the interventions considered, with the exception of increasing the TB cure rate. Administering ART or TB preventive therapy is less cost-effective than all of the HIV prevention strategies considered in [33].

6.4.2 Limitations of Analysis

The nature of the mathematical model used prevents full account being taken of any reduction in HIV transmission caused by these interventions. Even allowing for this however, we would recommend that further efforts be concentrated on improvements to TB treatment programmes and implementation of HIV prevention strategies, with increasing coverage of ART as a secondary aim.

No data exist on the costs of improving case detection rates, and so it was necessary to make assumptions. However, even taking the cost of finding additional cases to be double the existing level, improving case detection is very cost-effective. It could be 30 times more costly before it would be less cost-effective than ART or short-course IPT. Assumptions were also made about the cost of improving cure rates due to a lack of data for Kenya.

We encountered further problems with limited data when estimating the costs of providing ART in practice. Costs may fall over time due to economies of scale; alternatively, some costs may have been underestimated in the existing analysis due to the limited experience of administration to such large numbers of people.

Little data exists in the literature about the life expectancy of patients who default from ART. We have assumed that, following default from ART, a person is at the same position in the natural history of HIV as someone entering late-stage HIV (WHO stage 3). This means that those given ART effectively pass through stage 3 twice: first before being given ART; second following default from ART. During stage 3, they are more susceptible to TB, and therefore HIV-positives given ART have an increased risk of TB for longer than HIV-positives not given ART. This partly explains why the numbers on TB treatment under the ART strategy increase to the same level as under the baseline strategy when those first given ART start to drop out of treatment.

6.4.3 Verification of Results

The results that we present here are based on a model of the situation in Kenya. To verify these results and compare them with results obtained in other similar situations, we compare them with those presented in Creese et al [33] for the cost per DALY gained of TB treatment, ART and TB preventive therapy, which are reproduced in Table 6.4.3. The comparison shows that the results given here are not dissimilar to previously published estimates with the exception of costs for ART. The costs estimated here tend to be on the low side of the literature estimates, which is to be expected, as the epidemiological model takes account of reductions in TB transmission, whereas this has not been possible in previous studies. Costs per DALY gained for ART are also significantly lower because of the low costs per person year of treatment used (based on the "3 by 5" analysis), and due to the allowance made for the reduced AIDS costs under the ART strategy.

Intervention	Cost per DALY (2003 US$): Creese et al [33]	Cost per DALY (2003 US$): this Analysis
TB Treatment	$2-$75	-$20 [-$24, -$15] (Increase cure rate) $7 [$3, $12] (Increase detection rate) -$1.20 [-$4.60, $2.70] (Increase cure and detection rate)
ART	$1200-$2000	$224 [$220, $227] $268 [$265, $271] for 5% dropout
TB Preventive Therapy (6 months)	$185-$320	$226 [$173, $290]
TB Preventive Therapy (Lifetime)	None available	$692 [$503, $913]

Table 6.3: Comparison of our estimates of the cost per DALY with current literature estimates. Confidence intervals are 95%

6.4.4 Conclusions

We have shown in this chapter, that the most cost-effective intervention strategy, measured in terms of cost per DALY gained, is to increase the cure rate for active TB, with improvements to the TB case detection rate also being highly cost-effective. Increasing the TB case detection and cure rates to DOTS levels of 70% and 85% will result in a 40% reduction in TB incidence by the end of 2014 and a just under 50% reduction in TB deaths, as compared with the 2004 estimate, suggesting that implementation of DOTS alone will be sufficient to meet the Millennium Development Goal of reducing TB deaths by 50% by 2015. Some HIV prevention strategies are also very cost-effective and all are estimated to be more cost-effective than administering ART or TB preventive therapy. Increasing the coverage of ART has the greatest effect on reducing DALYs but suffers from very high costs. Provision of ART and HIV prevention interventions will however be necessary to reduce the burden of HIV.

Chapter 7

Conclusion

In this thesis we have described a Bayesian methodology to analyse complex statistical models. The methodology uses Monte Carlo sampling to integrate over the posterior distribution. We concentrated initially on importance sampling; in Chapter 2 discussing how the candidate distribution should be chosen to improve convergence. We then went on to show how the convergence of importance sampling can be measured in Chapter 3. The methodology has been applied to two examples: in Chapter 4 we considered the non-standard statistical problem of determining the number of components in a finite normal mixture model; and we described the Bayesian uncertainty analysis of a compartmental model of tuberculosis (TB) and HIV in Chapter 5. The ease with which the results of the sampling can be used was demonstrated in Chapter 6, where we used the parameter values output by the sampling to evaluate the cost-effectiveness of different interventions against TB and HIV and the uncertainty around these results.

7.1 Bayesian Statistics

In Bayesian statistics we work with the posterior probability distributions of model parameters. The posterior distribution is proportional to the product of the prior distribution of the parameters and the likelihood function. It therefore combines any prior knowledge of parameter values with the fit of the model to the data. No prior knowledge was available in the finite mixture models example and a Bayesian methodology was used because of the smoothing effect of the prior distribution on the likelihood distribution. The posterior distribution was therefore better behaved than the likelihood distribution; in particular not suffering from discontinuities. Good prior knowledge of the model parameter values for the TB-HIV model was available in the medical literature. Using Bayesian statistics in this example meant that we could give an estimate of the uncertainty on the results that took into account the fit of the model to the data and our prior knowledge of the parameter values.

7.2 Sampling Methodology

Normalising the posterior probability distribution involves integrating the product of the prior distribution and the likelihood over parameter space. The integral can sometimes be calculated analytically, but in the two examples that we considered in this thesis it was necessary to integrate numerically. We used Monte Carlo methods to perform this integration, using importance sampling in the analysis of finite mixture models and Markov Chain Monte Carlo sampling (MCMC) when analysing the model of TB and HIV.

Both importance sampling and MCMC require some information about the posterior distribution for them to be more efficient than simple Monte Carlo sampling. In importance sampling, all of the knowledge must be obtained before starting the sampling as the candidate distribution used is fixed. Most MCMC algorithms are adaptive however, meaning that the candidate distribution changes over the course of the sampling. The advantages of the adaptive approach were observed in the analysis of the TB-HIV model where the normal candidate distribution was very different from the skewed posterior distribution. Here, importance sampling performed badly but MCMC converged well.

The examples we considered were assumed to have approximately multivariate normal posterior distributions. We therefore restricted our investigations of the posterior distribution prior to the sampling to finding its mode and estimating the covariance matrix.

7.2.1 Importance Sampling

In Chapter 2 we investigated the choice of candidate distribution in importance sampling, showing that the optimal sampling function is the posterior distribution. When sampling to find the normalising factor for the posterior distribution, this is not a practical solution, and the chapter went on to discuss some of the practicalities of importance sampling in the context of statistical estimation. As the number of dimensions increases, discrepancies between the candidate distribution and the posterior distribution become more costly in importance sampling. We showed in Section 2.4.1 that the variance of the sampling increases exponentially with the dimension for any discrepancies in the mean, when both the candidate and the posterior distributions are multivariate normal. Knowledge of the mean was found to be more important than knowledge of the covariance.

Assessing the convergence of importance sampling focuses on the distribution of the weights, the ratios of the posterior distribution to candidate distribution at each of the sampling points. Very high weights are generally indicative of a lack of convergence. We described a number of diagnostic and statistical tests of importance sampling convergence in Chapter 3. The diagnostic tests relied heavily on graphical indicators of convergence and the statistical tests made use of results from extreme value theory to determine whether the weights had a finite variance. We found both to be useful but the diagnostic tests much more straightforward to use and interpret. We would recommend the use of both diagnostic and statistical tests. Non-convergence can generally be identified using the diagnostic tests, and confirmed by the statistical tests.

7.2.2 Markov Chain Monte Carlo

In MCMC, a Markov chain is constructed that has as its stationary distribution the distribution being integrated over; in our case this is the posterior distribution. We used the Metropolis-Hastings algorithm to determine the posterior distribution for the TB-HIV model, updating the mean after every acceptance. The adaptability of this algorithm means that it can cope better with the posterior distribution being different from the initial candidate distribution.

Individual samples generated by MCMC are not independent, and when analysing the TB-HIV model we ran several chains from different starting points to avoid problems of autocorrelation. Output from the different chains was then compared to determine whether the chains had been run for long enough for the output to be unaffected by their start points.

7.3 Model Selection for Finite Mixture Models

We used importance sampling to find the posterior distribution of the number of components in a finite normal mixture model. Comparison of our results with the literature, especially those of Richardson and Green [84], shows that we suggest more definite posterior distributions for the number of components. We suspect that this is due to the choice of prior distribution, but may be due to the presence of "nuisance" components in the results of the reversible jump sampling used in [84]. These components may have a very large variance, or alternatively may have a very similar mean and variance to another component in the mixture, effectively doubling up on one component.

Importance sampling has advantages over MCMC in this example. Constructing a Markov chain that moves between the different possible models is very difficult and requires rather complicated methodology, such as jump diffusion sampling [84]. In comparison, the importance sampling methodology is relatively simple, as movement between models with different numbers of components is dictated by the candidate distribution, which is made up of the probability of sampling each model, in addition to the probabilities of sampling each of the parameters of the model. The posterior distribution was found to be approximately multivariate normal, ensuring that the importance sampling also converged.

7.4 Model of Tuberculosis and HIV

The Bayesian analysis of a compartmental disease model of TB and HIV was described in Chapter 5. We found that MCMC worked well in determining the posterior distribution of the model parameters. Importance sampling demonstrated poor convergence, possibly due to the posterior distribution being skewed and therefore not a close enough match to the candidate distribution, a multivariate t-distribution.

The model results suggested that, in countries with high HIV prevalence and high TB incidence, the best method of reducing TB incidence over the next five to ten years is to improve treatment of TB, by detecting more cases and curing them more effectively. Interventions aimed at reducing HIV incidence or mitigating the effects of HIV infection will have a smaller effect on TB incidence and TB deaths. Improvements to TB treatment are also relatively cheap, and we showed in Chapter 6 that the cost per disability adjusted life year (DALY) gained for implementing the World Health Organization targets of 70% TB case detection and 85% TB cure was -\$1.20 [-\$4.60, \$2.70]. This made it much more cost-effective than antiretroviral therapy for which the cost per DALY averted was \$224 [\$220, \$227]. Antiretroviral therapy is very effective at reducing DALYs but the large costs involved in implementing it reduce its cost-effectiveness.

7.5 Further Work

In many areas of research, stochastic models are used more widely than deterministic models and are considered to be a better description of reality. One possible extension of this methodology would be to the analysis of stochastic models. The additional uncertainty in the model output may make determining convergence more difficult.

One simple extension of the model selection work would be to consider mixtures of distributions other than the normal distribution. For example mixtures of skewed distributions could provide better descriptions of skewed data sets, and use fewer components. The basic methodology would not need to change substantially to make this extension, with most of the work being involved in choosing appropriate prior distributions and refining the optimization routine. The methodology could also be extended to model selection in regression analysis, which is also a statistically non-standard problem. Our approach would probably be closest to work by George and McCulloch [52] and Cheng [20].

We have assessed the cost-effectiveness of interventions against TB and HIV when they are applied individually. In practice, several interventions will be implemented together, and decision-makers are interested in the cost-effectiveness of mixes of interventions. A resource allocation study may be useful in determining the best mix of interventions in different resource settings.

7.6 Discussion

We have demonstrated the use of a Bayesian methodology involving Monte Carlo sampling on two examples: determining the posterior distribution of the number of components in a finite normal mixture model and estimating the uncertainty in the output of a compartmental model of TB and HIV. We found that importance sampling worked well in the mixture models example, providing a relatively simple mechanism for jumping between different models. MCMC worked better in the TB-HIV model, where the posterior distribution had a very different shape from the chosen candidate distribution. The adaptability of the MCMC algorithm was an advantage in this case.

Implementing importance sampling is relatively straightforward and the algorithm is easily understood. In addition, as the samples output are independent, assessing convergence is much easier than for MCMC. Therefore, if it is possible to obtain reasonable convergence with importance sampling, without too much additional effort learning about the distribution, we would recommend its use in preference to MCMC. The main advantage of MCMC is its adaptability, but this contributes to the problems assessing its convergence because it means that the samples output will be correlated.

In conclusion, with the right choice of Monte Carlo sampling algorithm, the Bayesian methodology described in this thesis can be used to determine the posterior distribution of a complex system. We have demonstrated its use on model selection for finite normal mixture models and uncertainty analysis for a compartmental model of disease.

Appendix A

Posterior Correlation Matrix of the Tuberculosis Parameters

The estimated posterior correlation matrix of the TB parameters in the TB-HIV model described in Chapter 5 is given in Tables A.1 and A.2. The matrix has been split across the two tables for presentation purposes.

	λ_0	p	p_{HIV}	v	v_{HIV}	x	x_{HIV}	f	f_{HIV}	ϕ
λ_0	1	-0.003	0.020	-0.021	-0.400	-0.308	-0.033	0.411	0.043	0.317
p	-0.003	1	-0.071	-0.250	0.195	-0.550	0.077	-0.111	-0.044	0.103
p_{HIV}	0.020	-0.071	1	0.005	-0.251	0.114	-0.029	0.117	-0.069	0.025
v	-0.021	-0.250	0.005	1	-0.020	0.086	0.097	0.004	0.054	0.064
v_{HIV}	-0.400	0.195	-0.251	-0.020	1	0.184	0.010	0.048	-0.235	-0.121
x	-0.308	-0.550	0.114	0.086	0.184	1	-0.006	-0.055	0.082	-0.003
x_{HIV}	-0.033	0.077	-0.029	0.097	0.010	-0.006	1	-0.001	-0.036	0.049
f	0.411	-0.111	0.117	0.004	0.048	-0.055	-0.001	1	0.152	0.019
f_{HIV}	0.043	-0.044	-0.069	0.054	-0.235	0.082	-0.036	0.152	1	0.069
ϕ	0.317	0.103	0.025	0.064	-0.121	-0.003	0.049	0.019	0.069	1
w	-0.487	-0.175	0.068	0.022	0.045	-0.002	-0.066	0.178	0.028	0.033
w_{HIV}	0.004	0.100	-0.170	-0.073	0.066	-0.087	0.037	-0.023	-0.068	0.037
μ	-0.133	-0.133	-0.037	0.016	0.070	0.064	0.066	0.080	-0.040	0.102
μ^{inf}	0.221	0.221	-0.110	0.044	-0.222	-0.133	-0.003	0.226	-0.128	-0.065
μ^{inf}_{HIV}	0.262	0.097	-0.031	-0.079	0.004	-0.145	0.005	-0.021	0.251	0.035
$\mu^{non-inf}$	0.111	-0.023	0.081	0.017	-0.073	-0.025	-0.152	-0.050	-0.013	-0.004
$\mu^{non-inf}_{HIV}$	0.033	-0.079	-0.044	-0.008	-0.019	-0.032	0.008	0.058	0.048	-0.048
e	0.099	0.026	0.041	0.024	0.083	-0.045	-0.011	0.126	-0.010	-0.074

	λ_0	p	p_{HIV}	v	v_{HIV}	x	x_{HIV}	f	f_{HIV}	ϕ
rf	0.064	-0.102	0.028	0.009	-0.198	0.074	0.002	0.201	-0.037	0.207
tLs	-0.020	0.082	-0.058	0.035	0.454	-0.103	0.006	0.030	-0.092	-0.033

Table A.1: Posterior correlation matrix for the first ten TB parameters in the TB-HIV model, estimated using MCMC sampling.

	w	w_{HIV}	μ	μ^{inf}	μ^{inf}_{HIV}	$\mu^{non-inf}$	$\mu^{non-inf}_{HIV}$	e	rf	tLs	tD
λ_0	-0.487	0.004	-0.133	0.221	0.262	0.111	0.033	0.099	0.064	-0.020	-0.006
p	-0.175	0.100	-0.133	0.221	0.097	-0.023	-0.079	0.026	-0.102	0.082	0.036
p_{HIV}	0.068	-0.170	-0.037	-0.110	-0.031	0.081	-0.044	0.041	0.028	-0.058	0.081
v	0.022	-0.073	0.016	0.044	-0.079	0.017	-0.008	0.024	0.009	0.035	0.061
v_{HIV}	0.045	0.066	0.070	-0.222	0.004	-0.073	-0.019	0.083	-0.198	0.454	0.046
x	-0.002	-0.087	0.064	-0.133	-0.145	-0.025	-0.032	-0.045	0.074	-0.103	0.017
x_{HIV}	-0.066	0.037	0.066	-0.003	0.005	-0.152	0.008	-0.011	0.002	0.006	0.016
f	0.178	-0.023	0.080	0.226	-0.021	-0.050	0.058	0.126	0.201	0.030	0.054
f_{HIV}	0.028	-0.068	-0.040	-0.128	0.251	-0.013	0.048	-0.010	-0.037	-0.092	0.006
ϕ	0.033	0.037	0.102	-0.065	0.035	-0.004	-0.048	-0.074	0.207	-0.033	0.025
w	1	-0.011	0.011	0.049	-0.177	0.189	0.067	-0.055	0.130	0.010	0.007
w_{HIV}	-0.011	1	0.096	0.118	0.066	-0.078	-0.018	-0.055	-0.065	0.053	0.001
μ	0.011	0.096	1	-0.002	-0.048	0.004	-0.021	-0.034	0.098	-0.020	-0.008
μ^{inf}	0.049	0.118	-0.002	1	0.187	-0.020	-0.011	-0.025	-0.007	-0.100	-0.026
μ^{inf}_{HIV}	-0.177	0.066	-0.048	0.187	1	0.060	0.036	-0.032	0.080	0.189	0.072
$\mu^{non-inf}$	0.189	-0.078	0.004	-0.020	0.060	1	0.060	-0.002	-0.019	-0.051	-0.010
$\mu^{non-inf}_{HIV}$	0.067	-0.018	-0.021	-0.011	0.036	0.060	1	0.015	-0.057	0.049	-0.019
e	-0.055	-0.055	-0.034	-0.025	-0.032	-0.002	0.015	1	-0.184	0.040	0.035

	w	w_{HIV}	μ	μ^{inf}	μ^{inf}_{HIV}	$\mu^{non-inf}$	$\mu^{non-inf}_{HIV}$	e	rf	tLs	tD
rf	0.130	-0.065	0.098	-0.007	0.080	-0.019	-0.057	-0.184	1	-0.025	-0.076
tLs	0.010	0.053	-0.020	-0.100	0.189	-0.051	0.049	0.040	-0.025	1	0.121

Table A.2: Posterior correlation matrix for the last eleven TB parameters in the TB-HIV model, estimated using MCMC sampling.

Appendix B

Unit Costs of Treatments for Tuberculosis and HIV

The table gives unit costs of treatments for tuberculosis (TB) and HIV used in the cost-effectiveness analysis in Chapter 6. For normal distributions, the first figure gives the mean and the second the variance. For uniform distributions, the two figures give the lower and upper limits. Confidence intervals are 95%. (Although data are available on the costs of treatment and care for people with AIDS in Kenya, these data were not used because they are out-of-date and currently implausible - the cost per person multiplied by the number of people with AIDS gives a total cost in excess of the country's total government health care expenditure.

Treatment	Unit	Unit Cost in US$ (year 2003 prices)	Uncertainty Distribution	Reference/Assumptions
TB diagnosis costs, existing level of case detection	SS+ case detected	101	Normal(101, 25)	Nganda et al [10] For every SS+ case detected, assume 10 suspects are seen. For each suspect, assume 3 sputum smears and 1 chest X-ray are done.
	SS+ case detected	152	Uniform(101, 202)	Detecting additional cases is likely to be more costly on a per case basis than treatment at existing level of case detection. No data are available to suggest what these costs would be so we allow them to vary between 1 and 2 times the existing cost.
Short course treatment for TB (SS+), existing level of case detection	Person treated	140	Normal(140, 49)	Nganda et al [10]

Treatment	Unit	Unit Cost in US$ (year 2003 prices)	Uncertainty Distribution	Reference/Assumptions
Short course treatment for TB (SS-), existing level of case detection	Person treated	130	Normal(130, 43)	Nganda et al [10]
Short course treatment for TB (SS+), any additional case above existing case detection levels	Person treated	210	Uniform(140, 280)	Treating additional cases is likely to be more costly on a per patient basis than treatment at existing level of case detection. No data are available to suggest what these costs would be, so we allow them to vary between 1 and 2 times the existing cost.
Short course treatment for TB (SS-), any additional case above existing case detection level	Person treated	195	Uniform(130, 260)	As above for treatment of SS+ cases

Treatment	Unit	Unit Cost in US$ (year 2003 prices)	Uncertainty Distribution	Reference/Assumptions
Isoniazid preventive therapy (6 months)	Person treated	32	Uniform(27, 37)	Bell et al [11] and evidence from ProTEST pilot projects (need ref). Assume 13% adult population accesses VCT each year, 36% are HIV+, 100% are screened for IPT, 43% start treatment of whom 38% complete treatment (give refs)
Isoniazid preventive therapy (lifetime)	Person year of treatment	64	Uniform(54, 74)	As above for isoniazid preventive therapy for six months, plus assumption that treatment for one year is double the cost of treatment for six months
Treatment for AIDS-related opportunistic infections and palliative care in the absence of ART	Person year of treatment	199	Normal(199, 99)	Cost and access to care assumptions used in recent cost analysis of "3 by 5". 56 - 78 % of people with AIDS assumed to access treatment (WHO/UNAIDS working group, unpublished report)

Treatment	Unit	Unit Cost in US$ (year 2003 prices)	Uncertainty Distribution	Reference/Assumptions
ART	Person year of treatment	308 for all except people with TB , 548 for patients with TB	None specified	Cost assumptions used in recent cost analysis of "3 by 5" [57]

Table B.1: Unit costs of treatments for tuberculosis (TB) and HIV used in the cost-effectiveness analysis in Chapter 6

References

[1] M. Aitkin. Discussion of paper by Richardson and Green. *Journal of the Royal Statistical Society Series B*, 59:764, 1997.

[2] P. Arcidiacono and J. Bailey Jones. Finite mixture distributions, sequential likelihood and the EM algorithm. *Econometrica*, 71:933–946, 2003.

[3] S.K. Au and J.L. Beck. Important sampling in high dimensions. *Structural Safety*, 25:139–163, 2003.

[4] M. Badri, R. Ehrlich, T. Pulerwitz, R. Wood, and G. Maartens. Tuberculosis should not be considered an AIDS-defining illness in areas with a high tuberculosis prevalence. *International Journal of Tuberculosis and Lung Disease*, 6:231–237, 2002.

[5] World Bank. *World Development Report 1993: Investing in health*. Oxford University Press, 1993.

[6] World Bank. *Malawi AIDS Assessment Study*. World Bank, Washington DC, 1998.

[7] D.R. Barr and N.L. Slezak. A comparison of multivariate normal generators. *Communications of the ACM*, 15:1048–1049, 1972.

[8] T. Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society London*, 53:370 – 418, 1763.

[9] T. Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society London*, 54:296 – 325, 1763. Reprinted in Biometrika **45** (1958), 293-315, with a biographical note by G.A. Barnard.

[10] Y.C. Bechtel, C. Bonaiti-Pellie, M. Poisson, J. Magnette, and P.R. Bechtel. A population and family study of n-acetyltransferase using caffeine urinary metabolites. *Clinical Pharmacology and Therapeutics*, 54:134–141, 1993.

[11] J. Berkhof, I. van Mechelen, and A. Gelman. A Bayesian approach to the selection and testing of mixture models. *Statistica Sinica*, 13:423–442, 2003.

[12] J.A. Bilmes. A gentle tutorial of the EM algorithm and its applications to parameter estimation for gaussian mixture and hidden Markov models. TR-97-021, U.C. Berkeley, 1998.

[13] S.M. Blower and H. Dowlatabadi. Sensitivity and uncertainty analysis of complex models of disease transmission: an HIV model, as an example. *International Statistical Review*, 62:229–243, 1994.

[14] M.W. Borgdorff, K. Floyd, and J.F. Broekmans. Interventions to reduce tuberculosis mortality and transmission in low- and middle-income countries. *Bulletin of the World Health Organization*, 80:217–227, 2002.

[15] G.E.P. Box and M.E. Muller. A note on the generation of random normal deviates. *Annals of Mathematical Statistics*, 29:610–611, 1958.

[16] P. Bratley, B.L. Fox, and L.E. Schrage. *A Guide to Simulation*. Springer-Verlag, 1983.

[17] D.M. Burley. *Studies in Optimization.* International Textbook Co. Ltd., 1974.

[18] E. Castillo and A.S. Hadi. Fitting the generalised pareto distribution to data. *Journal of the American Statistical Association*, 92:1609–1620, 1997.

[19] R.C.H. Cheng. The generation of gamma variates with non-integral shape parameter. *Applied Statistics*, 26:71–75, 1977.

[20] R.C.H. Cheng. Bayesian model selection when the number of components is unknown. In D.J. Medeiros, E.R. Watson, J.S. Carson, and M.S. Manivannan, editors, *Proceedings of the Winter Simulation Conference*, pages 653–659, 1998.

[21] R.C.H. Cheng and C.S.M. Currie. Prior and candidate models in the Bayesian analysis of finite mixtures. In S. Chick, P.J. Sanchez, D. Ferrin, and D.J. Maurice, editors, *Proceedings of the Winter Simulation Conference 2003*, pages 392 – 398, 2003.

[22] R.C.H. Cheng and W.B. Liu. Discussion of paper by Richardson and Green. *Journal of the Royal Statistical Society Series B*, 59:776, 1997.

[23] H. Chernoff. On the distribution of the likelihood ratio. *Annals of Mathematical Statistics*, 25:573–578, 1954.

[24] S. Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90:1313–1321, 1995.

[25] C.E. Clark. Importance sampling in Monte Carlo analysis. *Operations Research*, 9:603–620, 1961.

[26] D.L. Cohn and W.M. El-Sadr. Treatment of latent tuberculosis infection. In L.B. Reichman and E.S. Herchfield, editors, *Tuberculosis: a Comprehensive International Approach*, pages 471–501. Marcel Dekker, 2000.

[27] Cascade Collaboration. Survival after introduction of HAART in people with known duration of HIV-1 infection. *The Lancet*, 355:1158–1159, 2000.

[28] Cascade Collaboration. Time from HIV-1 seroconversion to AIDS and death before widespread use of highly-active anti-retroviral therapy: a collaborative re-analysis. *The Lancet*, 355:1131–1137, 2001.

[29] C. Colvin, M. Connolly and E. Gouws. HIV prevalence and projections of HIV prevalence and incidence in Daimler Chrysler employees: results of an HIV surveillance study. *Industrial Report*, 2002.

[30] E.L. Corbett, C.J. Watt, N. Walker, D. Maher, B.G. Williams, M.S. R'aviglione, and C. Dye. The growing burden of tuberculosis. global trends and interactions. *Archives of Internal Medicine*, 163:1009–1021, 2003.

[31] S. Crawford. An application of the Laplace method to finite mixture distributions. *Journal of the American Statistical Association*, 89:259–267, 1994.

[32] S.L. Crawford, M.H. DeGroot, J.B. Kadane, and M.J. Small. Modelling lake chemistry distributions: approximate Bayesian methods for estimating a finite mixture model. *Technometrics*, 34:441–453, 1992.

[33] A. Creese, K. Floyd, A. Alban, and L. Guinness. Cost-effectiveness of HIV/AIDS interventions in Africa: a systematic review of the evidence. *The Lancet*, 359:1635–1642, 2002.

[34] C.S.M. Currie. A study of the effectiveness of different interventions in reducing the severity of tuberculosis epidemics in countries with a high prevalence of human immunodeficiency virus. Master's thesis, University of Southampton, 2001.

[35] C.S.M. Currie, K. Floyd, B.G. Williams, and C. Dye. Cost, affordability and cost-effectiveness of strategies to control tuberculosis in countries with high HIV prevalence. *AIDS*, Submitted September 2004.

[36] C.S.M. Currie, B.G. Williams, R.C.H. Cheng, and C. Dye. Tuberculosis epidemics driven by HIV: is prevention better than cure? *AIDS*, 17:2501–2508, 2003.

[37] W.C. Davidon. Variable metric method for minimization. Argonne Nat. Lab. report ANL-5990 Rev., 1959.

[38] K.M. De Cock and R.E. Chaisson. Will DOTS do it? A reappraisal of tuberculosis control in countries with high rates of HIV infection. *International Journal of Tuberculosis and Lung Disease*, 3:457–465, 1999.

[39] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, 39:1–38, 1977.

[40] R. Detels, A. Munoz, G. McFarlane, and et al. Effectiveness of potent antiretroviral therapy on time to AIDS and death in men with known HIV infection duration. *Journal of the Americal Medical Association*, 280:1497–1503, 1998.

[41] J. Diebolt and E.H.S. Ip. Stochastic EM: method and application. In W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, chapter 15. Chapman and Hall, 1996.

[42] J. Diebolt and C.P. Robert. Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society Series B*, 56:363–375, 1994.

[43] C.W. Dunnett and M. Sobel. A bivariate generalization of student's t-distribution, with tables for certain special cases. *Biometrika*, 41:153–169, 1954.

[44] C. Dye, G.P. Garnett, K. Sleeman, and B.G. Williams. Prospects for worldwide tuberculosis control under the WHO DOTS strategy. *Lancet*, 352:1886–1891, 1998.

[45] M.D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588, 1995.

[46] M. Evans and T. Swartz. Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statistical Science*, 10:254–272, 1995.

[47] M. Evans and T. Swartz. *Approximating Integrals via Monte Carlo and Deterministic Methods*. Oxford University Press, 2000.

[48] K. Floyd. Costs and effectiveness - the impact of economic studies on TB control. *Tuberculosis*, 83:187–200, 2003.

[49] J.A. Fox-Rushby and K. Hanson. Calculating and presenting disability adjusted life years (DALYs) in cost-effectiveness analysis. *Health Policy and Planning*, 16:326–331, 2001.

[50] K.A. Freedberg, E. Losina, M.C. Weinstein, et al. The cost effectiveness of combination antiretroviral therapy for HIV disease. *New England Journal of Medicine*, 344:824–831, 2001.

[51] A. Gelman. Inference and monitoring convergence. In W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, chapter 8. Chapman and Hall, 1996.

[52] E.I. George and R.E. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88:881–889, 1993.

[53] J. Geweke. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 57:1317–1339, 1989.

[54] E. Gilleland, R. Katz, and G. Young. The extRemes toolkit: weather and climate applications of extreme value statistics. http://www.esig.ucar.edu/extremevalues/evtk.htm, 2004.

[55] P. Godfrey-Faussett. Policy statement on preventive therapy against tuberculosis in people living with HIV. Geneva: World Health Organisation WHO/TB/98.255, 1998.

[56] M.R. Gold, J.E. Siegel, L.B. Russell, and M.C. Weinstein. *Cost-Effectiveness in Health and Medicine*. Oxford University Press, 1996.

[57] J.P. Gutierrez, B. Johns, T. Adam, and et al. Achieving the WHO/UNAIDS antiretroviral treatment 3-by-5 goal: what will it cost? *Lancet*, 364:63–64, 2004.

[58] J.M. Hammersley and D.C. Handscomb. *Monte Carlo Methods*. Methuen, 1964.

[59] W.K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[60] J.A. Hogle. What happened in Uganda? declining HIV, behaviour change and the national response. USAID Lessons Learned Case Study, 2002.

[61] C. Jennison. Discussion of paper by Richardson and Green. *Journal of the Royal Statistical Society Series B*, 59:778, 1997.

[62] J.L. Johnson, A. Okwera, and D.L. Hom. Duration of efficacy of treatment of latent tuberculosis infection in HIV-infected adults. *AIDS*, 15:2137–2147, 2001.

[63] H. Kahn and A.W. Marshall. Methods of reducing sample size in Monte Carlo computations. *Journal of the Operational Research Society of America*, 1:263–278, 1953.

[64] T. Kloek and H.K. van Dijk. Bayesian estimates of equation system parameters: an application of integration by Monte Carlo. *Econometrica*, 46:1–19, 1978.

[65] S.J. Koopman and N. Shepherd. Testing the assumptions behind the use of importance sampling. http://www.nuff.ox.ac.uk/economics/papers/2002/w17/importance.pdf, 2003.

[66] A.W. Marshall. The use of multi-stage sampling schemes in Monte Carlo computations. In H.A. Meyer, editor, *Symposium on Monte Carlo Methods*, pages 123–140. John Wiley and Sons, 1956.

[67] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equations of state calculations by fast computing machine. *Journal of Chemical Physics*, 21:1087 – 1091, 1953.

[68] D. Morgan, C. Mahe, B. Mayanja, and J.A.G. Whitworth. Progression to symptomatic disease in people infected with HIV-1 in rural Uganda: prospective cohort study. *British Medical Journal*, 324:193–197, 2002.

[69] C.J.L. Murray and H.J. Chum. Cost-effectiveness of chemotherapy for pulmonary tuberculosis in 3 Sub-Saharan countries. *The Lancet*, 338:1305–1308, 1991.

[70] C.J.L. Murray and A.D. Lopez. *Global Burden of Disease and Injury, Volume 1*. Harvard University Press, 1996.

[71] C.J.L. Murray and J. Salomon. Modelling the impact of global tuberculosis control strategies. *Proceedings of the National Academy of Sciences USA*, 95:13881–13886, 1998.

[72] P. Musgrove. Public spending on healthcare: how are different criteria related? *Health Policy*, 47:207–223, 1993.

[73] R.M. Neal. Erroneous results in 'Marginal likelihoods from the Gibbs output'. http://www.cs.utoronto.ca/ radford, 1999.

[74] J.A. Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, 7:308–313, 1965.

[75] US Bureau of Cenus. HIV/AIDS surveillance database. Available at www.census.gov/ipc/www/hivaidsd.html, 2001.

[76] M.S. Oh. *Contemporary Mathematics*, pages 165–188. American Mathematical Society, 1991. N. Flourney and R. Tsutakawa editors.

[77] Geneva: World Health Organization. World Health Report. 1999.

[78] World Health Organization. The global plan to stop TB. WHO/CDS/STB/2001.16, 2001.

[79] World Health Organization. Global tuberculosis control: surveillance, planning, financing, who report 2002. WHO/TB/2002.295, 2002.

[80] D.B. Phillips and A.F.M. Smith. Bayesian model comparison via jump diffusions. In W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, chapter 13. Chapman and Hall, 1996.

[81] T.C. Porco, P.M. Small, and S.M. Blower. Amplification dynamics: predicting the effect of HIV on tuberculosis outbreaks. *Journal of Acquired Immune Deficiency Syndromes*, 28:437–444, 2001.

[82] A.E. Raftery. Hypothesis testing and model selection. In W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, chapter 10. Chapman and Hall, 1996.

[83] S. Richardson and P.J. Green. Discussion of paper by Richardson and Green. *Journal of the Royal Statistical Society Series B*, 59:785, 1997.

[84] S. Richardson and P.J. Green. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society B*, 59:731–792, 1997.

[85] C.P. Robert and G. Casella. *Monte Carlo Statistical Models.* Springer-Verlag, 1999.

[86] G.O. Roberts. Markov chain concepts related to sampling algorithms. In W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, chapter 3. Chapman and Hall, 1996.

[87] K. Roeder. Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, 85:617–624, 1992.

[88] K. Roeder and L. Wasserman. Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, 92:894–902, 1997.

[89] R.Y. Rubinstein. *Simulation and the Monte Carlo Method.* Wiley, 1981.

[90] L.B. Russell, M.R. Gold, J.E. Siegel, N. Daniels, and M.C. Weinstein. The role of cost-effectiveness analysis in health and medicine. panel on cost-effectiveness in health and medicine. *Journal of the American Medical Association*, 276:1172–1177, 1996.

[91] S.K. Sahu and R.C.H. Cheng. A fast distance-based approach for determining the number of components in mixtures. *The Canadian Journal of Statistics*, 31:3–22, 2003.

[92] J.A. Salomon and C.L. Murray. Modelling HIV/AIDS epidemics in Sub-Saharan Africa using seroprevalence data from antenatal clinics. *Bulletin of the World Health Organisation*, 79:596–607, 2001.

[93] G. Santoro-Lopez, A.M. Felix de Pinho, L.H. Harrison, and M. Schechter. Reduced risk of tuberculosis among Brazilian patients with Human Immunodeficiency Virus infection treated with highly active antiretroviral therapy. *Clinical Infectious Diseases*, 34:543–546, 2002.

[94] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[95] S.R. Searle. *Linear Models*. Wiley, New York, 1971.

[96] R.L. Smith. Estimating tails of probability distributions. *Annals of Statistics*, 15:1174–1207, 1987.

[97] M. Stephens. Bayesian analysis of mixture models with an unknown number of components - an alternative to reversible jump methods. *Annals of Statistics*, 28:40–74, 2000.

[98] P. Weidle and et al. Evaluation of patients accessing antiretroviral therapy in the UNAIDS HIV drug access initiative in Uganda. In *XIII International AIDS Conference Durban*, 2000. Abstract ThPeB5231.

[99] M. West. Discussion of paper by Richardson and Green. *Journal of the Royal Statistical Society Series B*, 59:783, 1997.

[100] D. Wilkinson, S.B. Squire, and P. Garner. Effect of preventive treatment for tuberculosis in adults infected with HIV: systematic review of randomised placebo controlled trials. *British Medical Journal*, 317:625, 1998.

[101] B.G. Williams and C. Dye. Antiretroviral drugs for tuberculosis control in the era of HIV/AIDS. *Science*, 301:1535–1537, 2003.

[102] B.G. Williams, E. Gouws, D. Wilkinson, and S. Abdool Karim. Estimating HIV incidence rates from age prevalence data in epidemic situations. *Statistics in Medicine*, 20:2003–2016, 2001.

[103] World Health Organization. WHO Life Tables. 2003.