BAYESIAN ANALYSIS OF CHANGE-POINTS IN POISSON PROCESSES

K.D. Moloi and P.C.N. Groenewald

1. INTRODUCTION

Change-point analysis deals with the situation where an abrupt change has possibly taken place in the underlying mechanism that generates random variables. In a parametric setting, this means a change in the parameters of the underlying distribution. The interest is in whether such a change has actually taken place, and if it has, at which point in time. Also, there may have been more than one change during the period of interest. Application of change-point analysis is wide, but is particularly relevant in finance, the environment and medicine. The violability of markets may change abruptly, the rate and intensity of natural phenomena may change, or the effect of treatments in clinical trails may be studied.

The literature on change-point problems is, by now, enormous. In this study we consider only the so-called non-sequential or fixed sample size version, although an informal sequential procedure, which follows from Smith (1975), is a routine consequence. Still, literature is substantial and our focus is on a fully Bayesian parametric approach. Use of the Bayesian framework for inference with regard to the change-point dates to work by Chernoff and Zacks (1964). Smith (1975) presents the Bayesian formulation for a finite sequence of independent observations. See also Zacks (1983). In our study we will consider only Poisson sequences and will address four situations:

1) When it is assumed that there is exactly one change-point, and proper priors are used. This can be generalised to more than one change-point. If the number of change-points is fixed and known, improper priors are also valid as will be explained later.

2) When there is a fixed number of change-points, the Markov Chain Monte Carlo method of Chib (1998) is useful, especially for large samples and multiple change-points This approach will be described and applied.

3) When the number of change-points is unknown, and we want posterior probability distributions of the number of change-points, only proper priors are valid for calculating Bayes factors. In the case when no prior information is available, improper priors will cause the Bayes factor to have an indeterminate constant. In this case we apply the Fractional Bayes factor method of O'Hagan (1995).

4) When the data consists of multiple sequences, it is called multi-path change-point analysis, and the distribution from which the change-points are drawn is of interest. Here the posterior distributions of parameters are estimated by MCMC methods. All the techniques are illustrated using simulated and real data sets.

1.1 Bayes factors.

The Bayesian approach to hypothesis testing was developed by Jeffreys (1935, 1961) as major part of his program for scientific inference. The centrepiece was a number, now called the Bayes factor, which is the posterior odds of the null hypothesis when the prior probability on the null is one-half. Jeffreys was concerned with the comparison of predictions made by two competing scientific theories. In his approach, statistical models are introduced to represent the probability of the data according to each of the two theories and Bayes' theorem is used to compute the posterior probability that one of the theories is correct.

According to Kass and Raftery (1993), often lost from the controversy however, are the practical aspects of the Bayesian methods: how conclusions may be drawn from them, and how they can provide answers when non-Bayesian methods are hard to construct, what their strength and limitation are.

Kass and Raftery (1993) begin with data *D*, assumed to have arisen under one of the two hypotheses H_1 and H_2 according to a probability density $pr(D|H_1)$ or $pr(D|H_2)$. Given a priori probabilities $pr(H_1)$ and $pr(H_2)=1-pr(H_1)$, the data produce a posteriori probabilities $pr(H_1|D)$ and $pr(H_2|D) = 1-pr(H_1|D)$. From Bayes' theorem, we obtain

$$pr(H_k|D)=\frac{pr(D|H_k)pr(H_k)}{pr(D|H_1)pr(H_1)+pr(D|H_2)pr(H_2)} \quad , \quad k = 1,2,$$

so that

$$\frac{pr(H_1|D)}{pr(H_2|D)}=\frac{pr(D|H_1)pr(H_1)}{pr(D|H_2)pr(H_2)}$$

(1.1)

and the transformation is simply multiplication of the prior odds by

$$B_{12}=\frac{pr(D|H_1)}{pr(D|H_2)}, \qquad \text{which is the Bayes factor.}$$

(1.2)

Thus, in words, posterior $\propto$ Bayes factor $\times$ prior odds, and the Bayes factor is the ratio of the value of the posterior odds, regardless of the value of the prior odds. In the simple case, when the two hypotheses are single distributions with no free parameters (the case of "simple versus simple" testing), B_{12} is the likelihood ratio. In other cases, when there are unknown parameters under either or both of the hypotheses, the densities $pr(D|H_k)$, k = 1,2, are obtained by integrating over the parameter space, so that in equation (1.2),

$$pr(D|H_k) = \int pr(D|\boldsymbol{\theta}_k, H_k)\pi(\boldsymbol{\theta}_k|H_k)d\boldsymbol{\theta}_k,$$

(1.3)

where $\boldsymbol{\theta}_k$ is the parameter under H_k, $\pi(\boldsymbol{\theta}_k \mid H_k)$ is its prior density, $pr(D \mid \boldsymbol{\theta}_k, H_k)$ is the density of D given the value of $\boldsymbol{\theta}_k$, or the likelihood function of $\boldsymbol{\theta}_k$ ($\boldsymbol{\theta}_k$ may be a vector with dimension d_k). The prior distributions $\pi(\boldsymbol{\theta}_k \mid H_k)$, k = 1,2, are necessary to find posterior probabilities.

The quantity $pr(D|H_k)$ given by equation (1.3) is the marginal probability of the data , because it is obtained by integrating the joint density of $(D, \boldsymbol{\theta}_k)$ over $\boldsymbol{\theta}_k$. It is also the predictive probability of the data; that is , the probability of seeing the data that actually were observed, calculated before any data became available. It is also sometimes called a marginal likelihood, or an integrated likelihood. Note that, as in computing the likelihood ratio statistics but unlike in some other applications of likelihood, all constants appearing in the definition of the likelihood $pr(D|H_k, \boldsymbol{\theta}_k)$ must be retained when computing B_{12}. In fact, B_{12} is closely related to the likelihood ratio statistics, in which the parameters θ_k are eliminated by maximization rather than by integration.

Bayes factor calculations

The Bayesian framework is particularly attractive in the context of change-point analysis because these models are non-nested. In such settings, the marginal likelihood of the respective models, and Bayes factor are the preferred means for comparing models.(Kass and Raftery (1995), Berger and Perrichi,(1996)).

The computation of the marginal likelihood using the posterior simulation output has been an area of much current activity. A method developed by Chib (1995) is quite simple to implement. The key point is that the marginal likelihood of model M_r,

$$m(\boldsymbol{y} \mid M_r) = \int f(\boldsymbol{y} \mid M_r, \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid M_r) d\boldsymbol{\theta},$$

may be expressed as

$$m(\boldsymbol{y} \mid M_r) = \frac{f(\boldsymbol{y} \mid M_r, \boldsymbol{\theta}^*) \pi(\boldsymbol{\theta}^* \mid M_r)}{\pi(\boldsymbol{\theta}^* \mid \boldsymbol{y}, M_r)},$$

(1.4)

where $\boldsymbol{\theta}^*$ is any point in the parameter space. Given estimates of the marginal likelihood for two models M_r and M_s, the Bayes factor of r versus s is defined as

$$B_{rs} = \frac{m(\boldsymbol{y} \mid M_r)}{m(\boldsymbol{y} \mid M_s)}.$$

Large values of B_{rs} indicate that the data support M_r over M_s (Jeffrey, 1961).

1.2 The Change-point Model.

In general, when there is uncertainty about the existence of a change-point, the parametric models is described as follows:

Let $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$ be a sequence of observations from a distribution with pdf $f(.|.)$. Under model M_o (no change-point),

$$x_i \sim f(x_i \mid \boldsymbol{\theta}) \ , \quad i = 1,2,\ldots,n.$$

Under model M_k (a change after the k^{th} observation),

$$x_i \sim \begin{cases} f(x_i \mid \boldsymbol{\theta}_1), & i = 1,2,\ldots,k \\ f(x_i \mid \boldsymbol{\theta}_2), & i = k+1,\ldots,n \end{cases}, \qquad k = d, d+1, \ldots, n-d.$$

This is the parametric model and the assumption is that only the parameters, and not the distribution, can change at k. The dimension of the parameter space under M_o is d and under M_k it is $2d$, and the parameters under model M_k are only estimable for $d \le k \le n - d$. There are $n - 2d + 2$ possible models, and the Bayes factor in favour of M_o, when compared with M_k, is

$$B_{ok} = \frac{\int f(\boldsymbol{x} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid M_o) d\boldsymbol{\theta}}{\iint f(\boldsymbol{x}_1, \boldsymbol{x}_2 \mid \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \mid M_k) d\boldsymbol{\theta}_1 d\boldsymbol{\theta}_2}$$

$$= \frac{m(\boldsymbol{x} \mid M_o)}{m(\boldsymbol{x}_1, \boldsymbol{x}_2 \mid M_k)},$$

(1.5)

where $\boldsymbol{x}_1 = \{x_1, x_2, \ldots, x_k\}$ and $\boldsymbol{x}_2 = \{x_{k+1}, \ldots, x_n\}$. The following relations also hold for the Bayes factor:

$$B_{ij} = \frac{1}{B_{ji}} = \frac{B_{io}}{B_{jo}} \ .$$

(1.6)

The posterior probability of model M_k is then given by

$$pr(M_k \mid \boldsymbol{x}) = \left[\sum_j \frac{p_j}{p_k} B_{jk} \right]^{-1}, \quad \text{k} = 0,\text{d},\text{d}+1,\ldots,\text{n-d},$$

(1.7) where p_j is the prior probability for model M_j.

The prior distributions $\pi(\boldsymbol{\theta} \mid M_o)$ and $\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \mid M_k)$ should, in general, be proper, but in the next section some proposed methods for dealing with improper priors will be discussed.

2. THE POISSON MODEL

Raftery and Akman (1996) developed a Bayesian approach to estimate and test for a Poisson process with a change-point, assuming the change-point to be continuous. Carlin, Gelfand and Smith (1992) presented a general approach to hierarchical Bayes change-point models. In particular, desired marginal posterior densities are obtained utilising the Gibbs sampler. They included an application to changing Poisson processes, applied to the coal-mining disaster data of Jarrett (1979). Raftery and Akman (1996) also analysed the coal-mining disaster data.

There have been indications that the number of cases of diarrhoea-associated haemolytic uraemic syndrome increased abruptly, during the early part of the 1980's in England. Henderson and Matthews (1993) investigate this hypothesis

and applied change-point models for Poisson variables to two series of data from regional referral units in Newcastle-upon-Tyne and Birmingham. Using a direct re-sampling process, Broemeling and Gregurich (1996) developed a Bayesian approach for the analysis of the change-point problem. They illustrated this technique with examples involving one shift for the Poisson process.

First, let us consider a sequence of observations, $x_1, x_2, \ldots, x_n$, from a Poisson model with exactly one discrete change at an unknown point k:

$$x_i \sim Poisson(\lambda_1), \quad i = 1, \ldots, k$$

$$x_i \sim Poisson(\lambda_2), \quad i = k+1, \ldots, n.$$

The likelihood function is

$$L(\lambda_1, \lambda_2, k \mid \mathbf{x}) = \frac{1}{\prod_{i=1}^{n} x_i!} \lambda_1^{y_1} \lambda_2^{y_2} e^{-\lambda_1 k} e^{-(n-k)\lambda_2} \quad , 1 \le k \le n-1,$$

(2.1)

where $y_1 = \sum_{i=1}^{k} x_i$ *and* $y_2 = \sum_{i=k+1}^{n} x_i$.

Assuming that λ_1, λ_2 and k are independent a priori and that the prior densities have the conjugate form

$$\pi(\lambda_1, \lambda_2 \mid \alpha, \beta) = \frac{\beta^{2\alpha}}{\Gamma^2(\alpha)} \lambda_1^{\alpha-1} \lambda_2^{\alpha-1} e^{-\beta(\lambda_1+\lambda_2)}$$

(2.2)

and we have a discrete uniform prior on k so that

$$f(y_1, y_2 | k, \alpha, \beta) = \frac{\Gamma(\alpha + y_1)\Gamma(\alpha + y_2)}{(k+\beta)^{\alpha+y_1}(n-k+\beta)^{\alpha+y_2}}$$

(2.3)

and

$$\pi(k | y_1, y_2, \alpha, \beta) = \frac{f(y_1, y_2 \mid k, \alpha, \beta)}{\sum_{k=1}^{n-1} f(y_1, y_2 \mid k, \alpha, \beta)} .$$

(2.4)

If we let $\alpha \to 0$ and $\beta \to 0$ to represent non-informative priors, it follows that

$$\pi(k \mid y) \propto \Gamma(y_1)\Gamma(y_2) k^{-y_1} (n-k)^{-y_2}$$

(2.5)

Alternatively, if we let $\alpha \to \frac{1}{2}$ and $\beta \to 0$, we have the Jeffreys prior.

Furthermore,

$$\lambda_i \mid y_i, k, \alpha, \beta \sim \Gamma(\alpha + y_i, k_i + \beta) \quad , \quad k_2 = n - k,$$

(2.6)

and, unconditionally,

$$\pi(\lambda_i \mid y, \alpha, \beta) = \sum_k \pi(\lambda_i \mid y_i, k, \alpha, \beta)\pi(k \mid y, \alpha, \beta), \qquad i=1,2.$$

(2.7)

For the ratio $\tau = \frac{\lambda_1}{\lambda_2}$ it follows that

$$\pi(\tau \mid y, k, \alpha, \beta) \propto \left[1 + \frac{k}{n-k+\beta}\tau\right]^{-(2\alpha+y)} \tau^{\alpha+y-1} \quad ,$$

(2.8)

so that

$$\frac{2(\alpha + y_2)k}{2(\alpha + y_1)(n-k+\beta)}\tau \mid y, k \sim F_{v_1, v_2}$$

(2.9)

where $v_1 = 2(\alpha + y_1)$, $v_2 = 2(\alpha + y_2)$.

The posterior of τ, unconditional of k, is then given by

$$\pi(\tau \mid y) = \sum_k \pi(\tau \mid y, k, \alpha, \beta)\pi(k \mid y, \alpha, \beta) .$$

(2.10)

In the above analysis we assumed exactly one change-point. Considering the possibility of no change, let

$$\pi(k) = \begin{cases} q & , \quad k = 0 \\ \dfrac{1-q}{n-1} & , \quad k = 1,\ldots,n-1 \end{cases} ,$$

(2.11)

where k = 0 means no change in the sequence, and M_k denotes the model with a change-point at k, k = 0,1,2,…,n-1. This means a prior probability of q for the model of no change, while the rest of the probability is uniformly distributed over all possible change-point positions. In general we will divide the prior probabilities uniformly between the number of possible change-points, so that q = 0.5 if there is only one possible change-point. Then

$$f(y \mid k=0, \alpha, \beta) = \frac{\beta^{\alpha}\Gamma(\alpha + y)}{\Gamma(\alpha)\prod_{i=1}^{n} x_i!(n+\beta)^{\alpha+y}}$$

(2.12)

where $y = \sum_{i=1}^{n} x_i$, and the posterior probability of no change follows as

$$\pi(k=0 \mid y) = \frac{qf(y \mid k=n)}{\sum \frac{1-q}{n-1} f(y \mid k) + qf(y \mid k=n)}$$

$$= \left[1 + \frac{1-q}{q(n-1)\sum_{k=1}^{n-1} B_{ko}} \right]^{-1},$$

(2.13)

and

$$\pi(k \mid y) = B_{ko}\left[\frac{q(n-1)}{1-q} + \sum_{j=1}^{n-1} B_{jo} \right]^{-1}, \qquad k = 1,2,..,n-1,$$

(2.14)

.where

$$B_{ko} = \frac{f(y_1, y_2 \mid k, \alpha, \beta)}{f(y \mid k=0, \alpha, \beta)} = \frac{\beta^{\alpha}\Gamma(\alpha+y_1)\Gamma(\alpha+y_2)(n+\beta)^{\alpha+y}}{\Gamma(\alpha)(k+\beta)^{\alpha+y_1}(n-k+\beta)^{\alpha+y_2}\Gamma(\alpha+y)}$$

(2.15)

is the Bayes factor in favour of model M_k when compared with the model M_o.

2.1 Fractional Bayes Factors

As can be seen from the above equation, we cannot let $\alpha, \beta \to 0$ (using vague priors) since we will get an indeterminate result. In this case we will use partial Bayes factors.

O'Hagan (1995) advocated the fractional Bayes factor (FBF), a new variant of a partial Bayes factor, which uses the device of dividing the data into two parts, **x = (y,z)** .The first set **y** is be used as a training sample to provide "prior" information about the parameters. The second part, **z**, is then used for model comparison.
To avoid the arbitrariness of choosing a particular **y** or having to consider all possible subsets of a given size, O'Hagan uses a fraction of the likelihood function, instead of a fraction of the sample, to provide information about the parameters and thereby turning improper priors into proper ones. He defines a simplified form of the partial Bayes factor as

$$B_{12}^{F} = \frac{m_1^b(\boldsymbol{x})}{m_2^b(\boldsymbol{x})},$$

(2.16)

where

$$m_i^b(\boldsymbol{x}) = \frac{m_i(\boldsymbol{x})}{mb_i(\boldsymbol{x})} = \frac{\int f_i(\boldsymbol{x}\mid\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i}{\int[f_i(\boldsymbol{x}\mid\boldsymbol{\theta}_i)]^b\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i}$$

(2.17)

If $\pi_i(\boldsymbol{\theta}_i)$=c$_i h_i(\boldsymbol{\theta}_i)$, h_i a function whose integral over the $\boldsymbol{\theta}_i$-space converges, the indeterminate constant c$_i$ cancel out, leaving

$$m_i^b(\boldsymbol{x}) = \frac{\int h_i(\boldsymbol{\theta}_i) f_i(\boldsymbol{x}\mid\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i}{\int h_i(\boldsymbol{\theta}_i)[f_i(\boldsymbol{x}\mid\boldsymbol{\theta}_i)]^b d\boldsymbol{\theta}_i}.$$

(2.18)

So O'Hagan (1995) proposes using a fractional part of the entire likelihood, $[f(\boldsymbol{x}\mid\boldsymbol{\theta})]^b$, instead of a training sample. This tends to produce a more stable answer than the use of a particular training sample, but will fail the asymptotic criterion, unless *b* $\propto \frac{1}{n}$ as the sample size n increase. The behaviour of the fractional Bayes factor for such a *b* is well worth study, although it appears to be quite difficult to decide on a specific choice of *b*. O'Hagan suggested *b* = $\frac{m}{n}$, where *m* is the minimal sample size (when it is unique). Other suggestions are $\frac{1}{\sqrt{n}}$ and $\frac{\log(n)}{n}$.

With vague priors, $\pi(\lambda_i) \propto \lambda_i^{-\frac{1}{2}}$, i = 0,1 or 2, for the Poisson model with one possible change-point, for the fractional BF it follows that the marginal likelihood with the vague prior is

$$m_o(y) = \frac{\Gamma(y+\frac{1}{2})}{\Pi x_i!\, n^{y+\frac{1}{2}}},$$

while the fractional marginal likelihood is

$$mb_o(y) = \frac{\Gamma(by+\frac{1}{2})}{(nb)^{by+\frac{1}{2}}(\Pi x_i!)^b},$$

so that

$$m_o^b(y) = \frac{\Gamma(y+\frac{1}{2})b^{by+\frac{1}{2}}n^{-y(1-b)}b^{by+\frac{1}{2}}}{\Gamma(by+\frac{1}{2})(\Pi x_i!)^{(1-b)}},$$

(2.19)

and

$$m_k^b(y_1, y_2) = \frac{\Gamma(y_1 + \frac{1}{2})\Gamma(y_2 + \frac{1}{2}) b^{by} k^{y_1(b-1)} (n-k)^{y_2(b-1)} b^{by+1}}{\Gamma(by_1 + \frac{1}{2})\Gamma(by_2 + \frac{1}{2})(\Pi x_i!)^{(1-b)}}.$$

(2.20)

The fractional Bayes factor in favour of no change against a change after the k^{th} observation is then given by

$$B_{ok}^F = \frac{m_o(y)}{m_k(y_1, y_2)}$$

$$= \frac{\Gamma(y + \frac{1}{2})\Gamma(by_1 + \frac{1}{2})\Gamma(by_2 + \frac{1}{2})}{\Gamma(by + \frac{1}{2})\Gamma(y_1 + \frac{1}{2})\Gamma(y_2 + \frac{1}{2})} \left(\frac{k}{n}\right)^{y_1(1-b)} \left(\frac{n-k}{n}\right)^{y_2(1-b)} b^{-\frac{1}{2}}.$$

(2.21)

If we use the prior $\pi(\lambda_i) \propto \lambda_i^{-1}$, and $b = \frac{2}{n}$, since $m = 2$ is the minimal sample size to estimate the parameters under model M_k, it follows that

$$\mathrm{B}_{ok}^F = \frac{n^{-\frac{(y-2)}{n}} B(\frac{2y_1}{n}, \frac{2y_2}{n})}{k^{-y_1(n-2)}(n-k)^{-y_2(n-2)} B(y_1, y_2)}.$$

(2.22)

Posterior probabilities follow from equations (2.21) and (1.7) where $B_{ko} = B_{ok}^{-1}$.

2.3 Sensitivity of the Fractional Bayes Factor.

To examine the sensitivity of the Fractional Bayes factor to the sample size and the value of the fraction *b*, consider a data set that supports the model with no change-point perfectly, that is, all observations are equal. The posterior probability for no change, as opposed to one change-point, is calculated when the prior probabilities are uniformly distributed as in (2.13) with $q = 0.5$.

Let $\pi(\lambda) \propto \lambda^{-\frac{1}{2}}$ under M_o, and $\pi(\lambda_1, \lambda_2) \propto \lambda_1^{-\frac{1}{2}} \lambda_2^{-\frac{1}{2}}$ under model M_k be the Jeffreys priors. The sample size is n and let *y* be the common value of the observations. Then the Fractional Bayes factor in favour of M_o is given by

$$B_{ok}^F = \frac{\Gamma(ny + \frac{1}{2})\Gamma(bky + \frac{1}{2})\Gamma(b(n-k)y + \frac{1}{2}) n^{-(1-b)ny}}{\Gamma(bny + \frac{1}{2})\Gamma(ky + \frac{1}{2})\Gamma((n-k)y + \frac{1}{2}) k^{-(1-b)ky} (n-k)^{-(1-b)(n-k)y} b^{\frac{1}{2}}}.$$

(2.23)

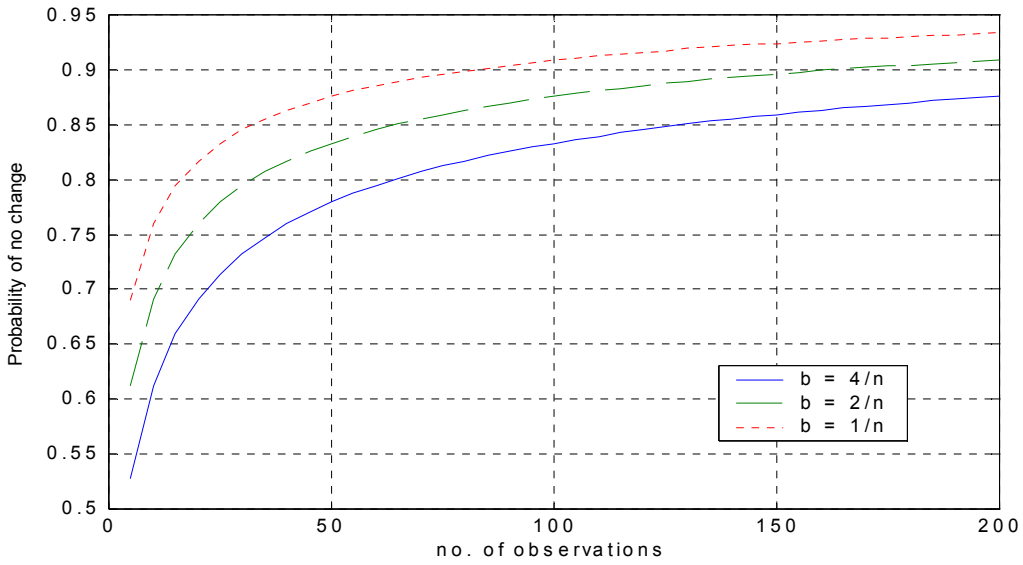
The posterior probability for no change is then given by

$$P_o = 1 - \sum_{k=1}^{n-1} \left[B_{ko}^F [n-1 + \sum_{k=1}^{n-1} B_{ko}^F]^{-1} \right] .$$

(2.24)

Figure 1.1 shows the posterior probability of no change as a function of sample size and for three values of the fraction *b* when the data supports the null model perfectly. The probability increases with sample size, but there remains a high degree of uncertainty for small and moderate samples.
Also, the Fractional Bayes factor discriminate better between models when the fraction b gets smaller, leaving more likelihood information free for model comparison. The actual value of the observation has very little effect on the posterior probabilities in Figure 1.1.

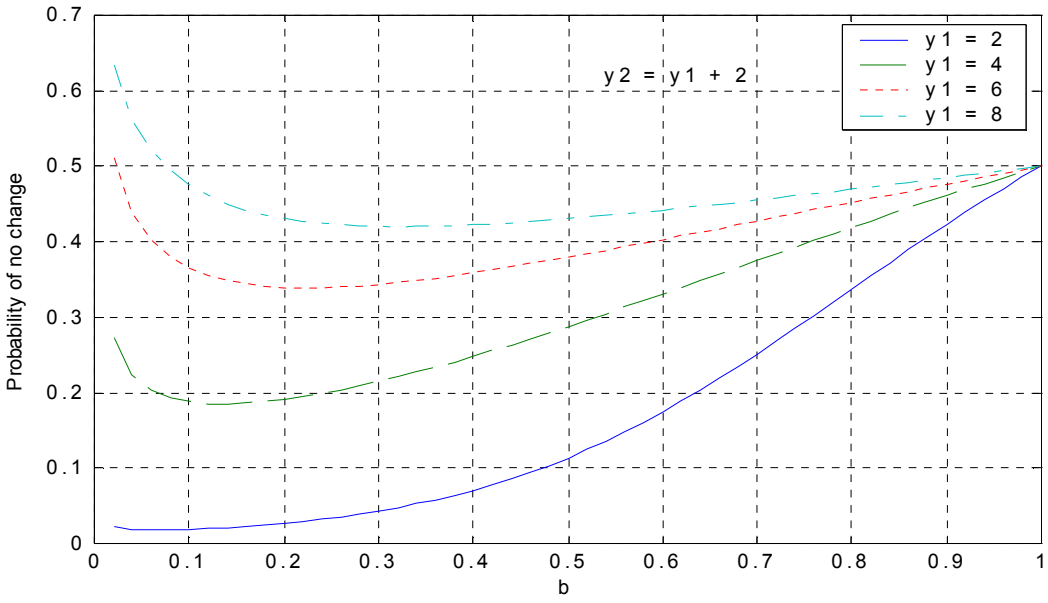
Figure 1.1: *Posterior Probability for Model M_o when all observations are equal.*



In summary, you can never achieve 100% certainty of no change, even when the data supports it perfectly and the sample size gets large, but if a change-point exists, it quickly becomes apparent when sample size and parameter value increase.

Also, the probabilities are sensitive to the value of *b*, as can be seen in Figure 1.2. There posterior probabilities are plotted as a function of *b* when *n* = 50 with a change at *k* = $\frac{n}{2}$, and the actual change in the data is an increase of 2 in the common value of the observations.

Figure 1.2: *Probability of no change as a function of b when n=50, k=*$\frac{n}{2}$*, $y_2 = y_1 + 2$*



As *b* approaches one, the probability approaches its prior value, 0.5, since there is no likelihood left for model comparison. As the observed values increase, it is naturally more difficult to discriminate when the difference is only 2. As the probability is always a convex function of *b*, it may be useful to report the lower bound, which does not seem to be overly biased against the probability of no change. However, the value of *b* remains a contentious issue when using the Fractional Bayes factor.

3. MULTIPLE CHANGE-POINTS

For a fixed known number of change-points, say r, we have a generalisation of equations (2.2) to (2.7). Let $\boldsymbol{k} = \{k_1, k_2,\ldots,k_r\}$ be the positions of the change-points where $k_1 < k_2 < \ldots < k_r$, and assume that $\lambda_i \sim Gamma(\alpha,\beta)$, $i = 1, 2,\ldots,r+1$, independently. We also assume that $\boldsymbol{k}$ is uniformly distributed over all possible partitions, so that

$$\pi(\boldsymbol{k}) = \left[\binom{n-1}{r}\right]^{-1}.$$

Let $y_i = \sum_{j=k_{i-1}+1}^{k_i} x_j$, where $k_o = 0$ and $k_{r+1} = n$. The marginal likelihood under a particular partition $\boldsymbol{k}$, model $M_{\boldsymbol{k}}$, is then

$$f(\boldsymbol{y}|\boldsymbol{k},\alpha,\beta) = \prod_{i=1}^{r+1} \frac{\Gamma(\alpha+y_i)}{(k_i+\beta)^{\alpha+y_i}} \qquad (3.1)$$

and

$$\pi(\boldsymbol{k}|\boldsymbol{y},\alpha,\beta) = \frac{f(\boldsymbol{y}\mid\boldsymbol{k},\alpha,\beta)}{\sum_{\boldsymbol{k}} f(\boldsymbol{y}\mid\boldsymbol{k},\alpha,\beta)}. \qquad (3.2)$$

The Bayes factor when comparing models $M_{\boldsymbol{k}}$ and $M_{\boldsymbol{s}}$ is just

$$B_{\boldsymbol{ks}} = \frac{f(\boldsymbol{y}|\boldsymbol{k},\alpha,\beta)}{f(\boldsymbol{y}|\boldsymbol{s},\alpha,\beta)} .$$

If $\alpha \to 0$ and $\beta \to 0$ it follows that

$$\pi(\boldsymbol{k} \mid \boldsymbol{y}) \propto \prod_{i=1}^{r+1} \Gamma(y_i) k_i^{-y_i} . \tag{3.3}$$

Notice that (3.3) only holds for partitions for which $y_i > 0$ for all *i*. With the Jeffreys prior ($\alpha = ½$) all partitions are valid.
Furthermore,

$$\lambda_i \mid y_i, k, \alpha, \beta \sim \Gamma(\alpha + y_i, k_i - k_{i-1} + \beta) \quad , \quad \text{i} = 1, 2, \ldots, \text{r+1} , \tag{3.4}$$

and the unconditional distribution of λ_i follows as in (2.7).

In the case of an **unknown** number of change-points maximum, but with a maximum of *R*, let h_r be the number of possible partitions given *r* change-points. Let

$$\pi(\boldsymbol{k} \mid r) = h_r^{-1} \quad \text{and} \quad \pi(\text{r}) = \frac{1}{R+1} \ , r=0,1, \ldots, R .$$

Define $B_{\boldsymbol{k}o}^r$ as the Bayes factor in favour of model $M_{\boldsymbol{k}}^r$, the model with r change-points, partitioned according to $\boldsymbol{k}$, when compared with the model M_o with no change-point. Then

$$B_{\boldsymbol{k}o}^r = \frac{f(\boldsymbol{y} \mid \boldsymbol{k},\alpha,\beta)}{f(\boldsymbol{y} \mid r=0,\alpha,\beta)} = \frac{\beta^{r\alpha}(n+\beta)^{\alpha+y}}{\Gamma^r(\alpha)\Gamma(\alpha+y)} \prod_{i=1}^{r+1} \frac{\Gamma(\alpha+y_i)}{(k_i - k_{i-1} + \beta)^{\alpha+y_i}} . \tag{3.5}$$

With the Jeffreys prior the Fractional Bayes factor is given by

$$B_{\boldsymbol{k}o}^{rF} = \frac{\Gamma(by+\frac{1}{2}) b^{\frac{r}{2}}}{\Gamma(y+\frac{1}{2}) n^{-y(1-b)}} \prod_{i=1}^{r+1} \frac{\Gamma(y_i+\frac{1}{2}) k_i^{y_i+\frac{1}{2}}}{\Gamma(by_i+\frac{1}{2})} , \tag{3.6}$$

where $b = \frac{r+1}{n}$, $\boldsymbol{k} = \{k_1, \ldots, k_{r+1}\}$.

The posterior distribution of the number of change-points *r* is given by

$$pr(r=0|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|r=0)pr(r=0)}{\sum_{j=0}^{R} \sum_{\boldsymbol{k}} f(\boldsymbol{y}|r=j,\boldsymbol{k})pr(r=j,\boldsymbol{k})}$$

$$= \frac{f(\boldsymbol{y} \mid r=0)}{\sum_{j=0}^{R} h_j^{-1} \sum_{\boldsymbol{k}} f(\boldsymbol{y} \mid r=j,\boldsymbol{k})}$$

$$= \left[1 + \sum_{j=1}^{R} h_j^{-1} \sum_{\boldsymbol{k}} B_{\boldsymbol{k}o}^j \right]^{-1} . \tag{3.7}$$

Also,

$$pr(r=t \mid \boldsymbol{y}) = \left[\frac{\sum_{j=0}^{R} h_j^{-1} \sum_{k} f(\boldsymbol{y} \mid \boldsymbol{k}, r=j)}{h_t^{-1} \sum_{k} f(\boldsymbol{y} \mid \boldsymbol{k}, r=t)} \right]^{-1}$$

$$= h_t^{-1} \sum_{\boldsymbol{k}} B_{\boldsymbol{k}o}^{t} \left[\sum_{j=0}^{R} h_j^{-1} \sum_{\boldsymbol{k}} B_{\boldsymbol{k}o}^{j} \right]^{-1}$$

$$= h_t^{-1} \sum_{\boldsymbol{k}} B_{\boldsymbol{k}o}^{t} pr(r=0 \mid \boldsymbol{y}), \qquad t = 1, 2, \ldots, R. \tag{3.8}$$

3.2 Alternative approach to multiple Change-points.

Chib (1998) proposed a new Bayesian approach for models with multiple change-points. The change-point model is formulated in terms of a latent discrete state variable that indicates the regime from with a particular observation has been drawn. This state variable is specified to evolve according to a discrete-time discrete-state Markov process with the transition probabilities constrained so that the state variable can either stay at the current value or jump to the next higher value. The model is estimated by Markov chain Monte Carlo methods using an approach that is based on Chib (1996). This approach is for a **known** number of change-points, but is useful since the computational effort does not increase exponentially with the sample size and the number of change-points, as is the case with the exact evaluation from the previous section. Also, proper priors are required but since there is a fixed number of change-points, vague proper priors ensure that the influence of the priors is minimal. In this section we will give a description of Chib's method as applicable to the Poisson model.

Assuming r change-points, the formulation is based on the introduction of the discrete variable s_t in each time period, the state of the system at time t, that takes the values of the integers $\{1, 2, \ldots, r+1\}$ and indicates the regime from which a particular observation x_t has been drawn. Specifically, $s_t = k$ indicates that x_t is drawn from $f(x_t \mid X_{t-1}, \lambda_k)$, where $X_{t-1} = \{x_1, x_2, \ldots, x_{t-1}\}$. The variable s_t is a Markov process with transition matrix

$$P = \begin{bmatrix} p_{11} & p_{12} & 0 & \ldots & 0 \\ 0 & p_{22} & p_{23} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & p_{rr} & p_{r,r+1} \\ 0 & 0 & \ldots & 0 & 1 \end{bmatrix}, \tag{3.9}$$

where $p_{ij} = pr(s_{t=j} \mid s_{t-1} = i)$. The chain begins in state 1 at time t = 1 and terminates in state r + 1. So s_t can either stay in the current state or move to the next higher one. The transitions of the state identify the change-points $\mathrm{K}_r = \{k_1, k_2, \ldots, k_r\}$.

Chernoff and Zacks (1964) propose a special case of this general model in which there is a constant probability of change at each time point. Yao (1984) specified the same model for the change points but assumed that the joint distribution of the parameters $\{\theta_k\}$ is exchangeable and independent of the change-points. Similar exchangeable models for the parameters have been studied by Carlin et al. (1992) in the context of a single change point, and by Inclán and Tiao (1994) in the context of multiple change-points.

Suppose prior density $\pi(\Lambda, P)$, where $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_{r+1}\}$, and data X_n, then the Monte Carlo sampling scheme is applied to obtain the posterior density $\pi(S_n, \Lambda, P \mid X_n)$, $S_n = \{s_1, s_2, \ldots, s_n\}$. The sampling method works recursively. First the states S_n are simulated conditional on the data and the other parameters, and second, the parameters are simulated conditional on the data and S_n. The MCMC algorithm is implemented by simulating as follows.

Simulation of $\{s_t\}$

Let $S^{t+1} = \{s_{t+1}, \ldots, s_n\}$,then the simulation consists of sampling, in turn,

- s_{n-1} from $f(s_{n-1} | X_n, s_n = r+1, \Lambda, P)$,
- s_{n-2} from $f(s_{n-2} | X_n, S^{n-1}, \Lambda, P)$,
- ……
- s_2 from $f(s_2 | X_n, S^3, \Lambda, P)$,

where $s_1 = 1$. Chib (1996) showed that

$$f(s_t \mid X_n, S^{t+1}, \Lambda, P) \propto f(s_t \mid X_n, P) f(s_{t+1} \mid s_t, P), \tag{3.10}$$

where s_t can take on only one of two possible values, conditional on s_{t+1}. The last term is just the probabilities from the transition matrix P, i.e.

$$f(s_{t+1} \mid s_t = i, P) = \begin{cases} i & \text{with probability } p_{ii} \\ i+1 & \text{with probability } 1 - p_{ii} \end{cases}.$$

To obtain the mass function $f(s_t \mid X_n, P)$, $t = 1, 2, \ldots, n$, a recursive calculation is required.

Starting with $t = 1$, where $f(s_1 = 1 \mid X_o, \Lambda) = 1$, the update is given by

$$f(s_t = j \mid X_t, \Lambda, P) = \frac{f(s_t = j \mid X_{t-1}, \Lambda, P) f(x_t \mid X_{t-1}, \lambda_j)}{\sum_{l=j-1}^{j} f(s_t = l \mid X_{t-1}, \Lambda, P) f(x_t \mid X_{t-1}, \lambda_l)}, \tag{3.11}$$

where

$$f(s_t = j \mid X_{t-1}, \Lambda, P) = \sum_{l=j-1}^{j} p_{lj} \times f(s_{t-1} = l \mid X_{t-1}, \Lambda, P) \tag{3.12}$$

and

$$f(x_t \mid X_{t-1}, \lambda_j) = \frac{\lambda_j^{x_t} e^{-\lambda_j}}{x_t!} \tag{3.13}$$

for $j = 1, 2,\ldots,r+1$ and p_{ij} is the Markov transition probabilities. With these mass functions at hand, the states are simulated from time n and working backwards according to the scheme described in (3.10).

Simulation of *P*

The full conditional distribution of P is independent of (X_n, Λ) given S_n, and the elements p_{ii} of P may be simulated from $f(P|S_n)$. We shall assume that $p_{ii} \sim Beta(a, b)$, independently, $i = 1, 2,\ldots,r$, where $a >> b$. The joint prior density of P is then

$$\pi(P) = \frac{1}{B^r(a,b)} \prod_{i=1}^{r} p_{ii}^{a-1}(1-p_{ii})^{b-1} . \tag{3.14}$$

The parameters a and b can be chosen so that $E(p_{ii}) = \frac{a}{a+b} \approx \frac{r+1}{n}$, with large variance. This means that apriori the mean lengths of all regimes are the same. Let n_{ii} denote the number of periods the process stays in state i, then the conditional distribution of p_{ii} is

$$p_{ii} \sim Beta(a + n_{ii}, b+1), \quad i = 1, 2,\ldots,r, \tag{3.15}$$

since $n_{i,i+1} = 1$. The p_{ii}'s can be simulated by letting $p_{ii} = \frac{x_1}{x_1+x_2}$, where

$x_1 \sim Gamma(a+n_{ii}, 1)$ and $x_2 \sim Gamma(b+1, 1)$.

Simulation of $\lambda_j, j = 1, 2,\ldots,r+1$.

Let $\lambda_j \sim Gamma(c, d), i = 1, 2,\ldots r+1$, independently, then the conditional distribution, $\Lambda \mid S_n, X_n, P$, factors into independent terms,

$$\lambda_j \mid X_n, S_n, P \sim Gamma(c + U_j, d + N_j), \quad j = 1, 2,\ldots,r+1, \tag{3.16}$$

where $U_j = \sum_{t=1}^{n} I(s_t = j) x_t$ and $N_j = \sum_{t=1}^{n} I(s_t = j)$. $I(s_t = j)$ is the indicator function that is equal to 1 if $s_t = j$ and zero otherwise. So N_j is simply the number of time periods the process spends in regime j, while U_j is the sum of the observations recorded while in regime j.

The sample output of the states S_n can be used to determine the posterior distribution of the change-points. Alternatively, the Monte Carlo estimate of $\pi(s_t \mid X_n)$ cab be found by taking an average of $f(s_t = j \mid X_{t-1}, \Lambda, P)$, from (3.12),

over the MCMC iterations of Λ and *P*. This is called Rao-Blackwellization, and is more efficient than taking the empirical distribution of the simulated states.

Chib (1998) also gives a MCMC approach to the calculation of marginal likelihoods, used for Bayes factor calculations when comparing models with different number of change-points.

4. APPPLICATIONS 1.

Example 4.1

As an example of the Poisson model with one change-point, we will use the diarrhoea-associated haemolytic uraemic syndrome (HUS) data used by Henderson and Matthews (1992). Haemolytic uraemic syndrome is a severe, life threatening illness, which predominantly affects infants and young children (Levin and Barrett, (1984)). The aetiology of HUS is unknown but various bacterial and viral agents have been implicated, with particular speculation of a link with the level in the environment of E. coli. There has been concern that the incidence of HUS has apparently increased sharply during the 1980's (Tarr et al. (1989), Coad et al. (1991)). As an example, we consider the frequency of cases of HUS treated in two specialist centres in Newcastle upon Tyne and Birmingham from 1970 to 1989. The data is given in Table 4.1.

Table 4.1: *Annual number of cases of HUS at each referral centre.*

Year	Newcastle	Birmingham	Year	Newcastle	Birmingham
1970	6	1	1980	4	1
1971	1	5	1981	0	7
1972	0	3	1982	4	11
1973	0	2	1983	3	4
1974	2	2	1984	3	7
1975	0	1	1985	13	10
1976	1	0	1986	14	16
1977	8	0	1987	8	16
1978	4	2	1988	9	9
1979	1	1	1989	19	15

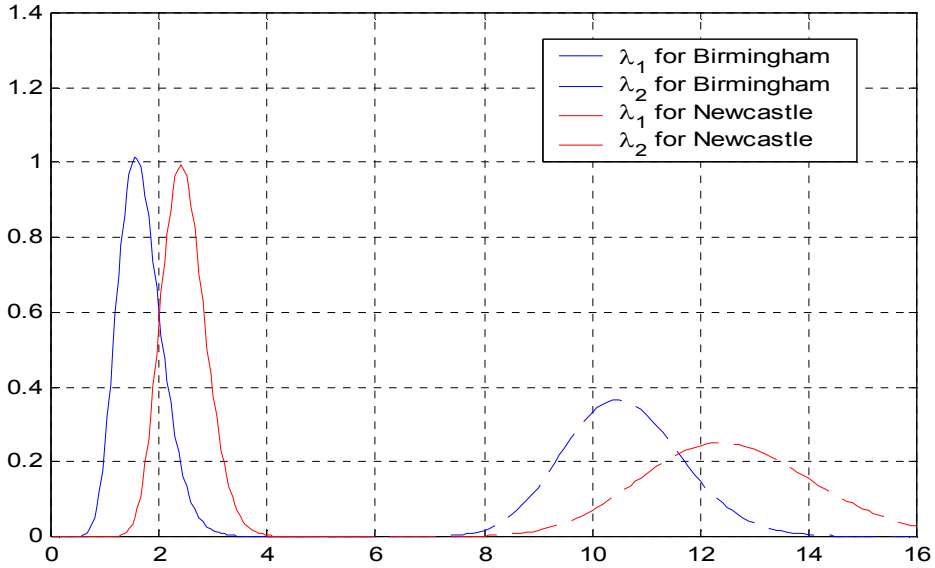
Assuming one change-point and using equation (2.5), the change at Newcastle occurred at k=15 (1984), and for Birmingham at k=11 (1980). Assuming at most one change-point and using the Fractional Bayes factor from (2.21) with (1.7) and $b = \frac{2}{n}$, we see that the probability for no change is virtually zero. The maximum probabilities and the probabilities for no change are given in Table 4.2 .

Table 4.2 *Posterior probabilities assuming at most one change-point.*

	Pr[No change|x]	Pr[k = 15|x]
Newcastle	1.680e-011	0.9834
	Pr[No change|x]	Pr[k = 11|x]
Birmingham	1.816e-013	0.9515

Figure 4.1 shows the posterior distributions of λ_1 and λ_2 for both cities, clearly showing the increase in cases.

Figure 4.1: *Posterior distribution of rate of incidences before and after change-point.*



Assuming two change-points, Figure 4.2 shows the distribution of the change-points for Newcastle, and figure 4.3 shows the distribution of the change-points for Birmingham.
For Newcastle the maximum probability for 2 change-points is 0.2712 at ***k*** = (7, 15), and for Birmingham it is 0.2507 at ***k*** = (11, 16).

Figure 4.2: *Posterior probability distribution: 2 Change-point for Newcastle*

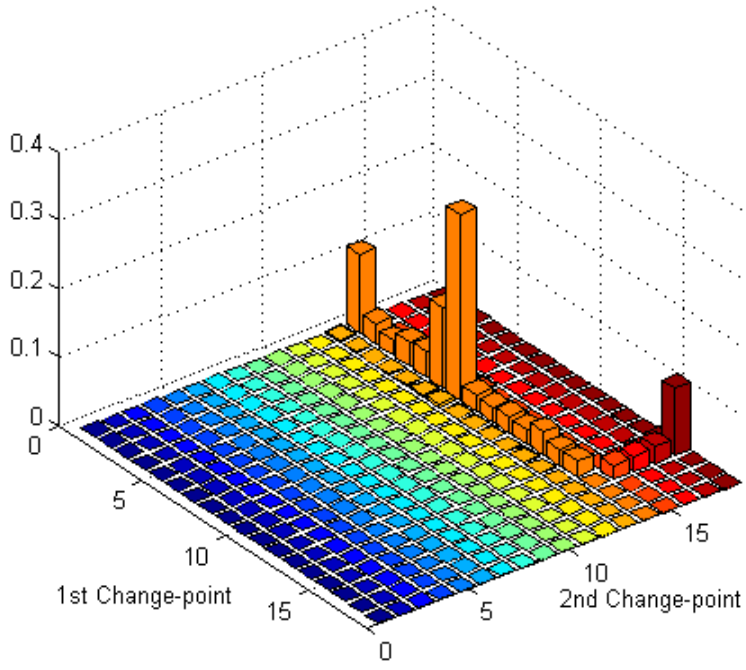
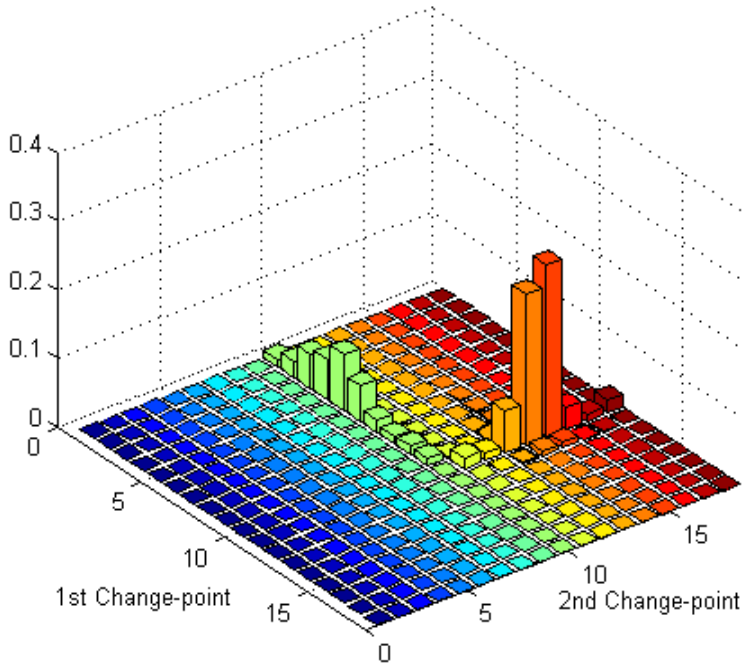


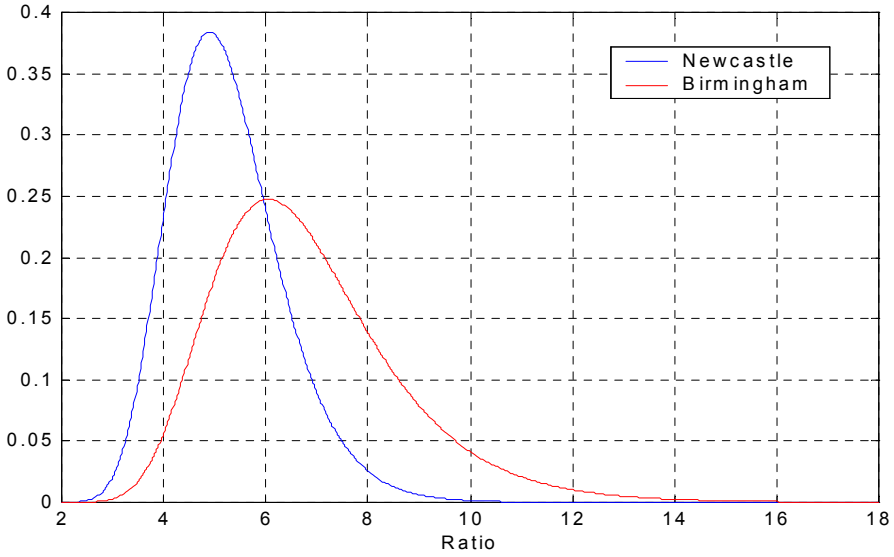
Figure 4.3: *Posterior probability distribution: 2 Change-point for Birmingham*



Assuming one change-point, Figure 4.4 shows the magnitude of the change $\frac{\lambda_2}{\lambda_1}$ (from equation (2.8)) for Birmingham and Newcastle. Clearly the change occurred later in Newcastle than in Birmingham and the magnitude of change is

greater in Birmingham with a mean increase of over 6 times compared to an increase of about 5 times in Newcastle.

Figure 4.4: *Posterior distribution of the ratio* $\tau = \frac{\lambda_2}{\lambda_1}$.



Considering models with up to four change-points, by using equations (3.6) to (3.8), the data seems to support a single change-point as seen in Table 4.3.

Table 4.3*: Posterior probabilities for multiple change-points, using the FBF.*

	No change	1 cp	2 cp's	3 cp's	4 cp's
Birmingham	0	0.4017	0.3825	0.1687	0.0471
Newcastle	0	0.3814	0.1921	0.2687	0.1577

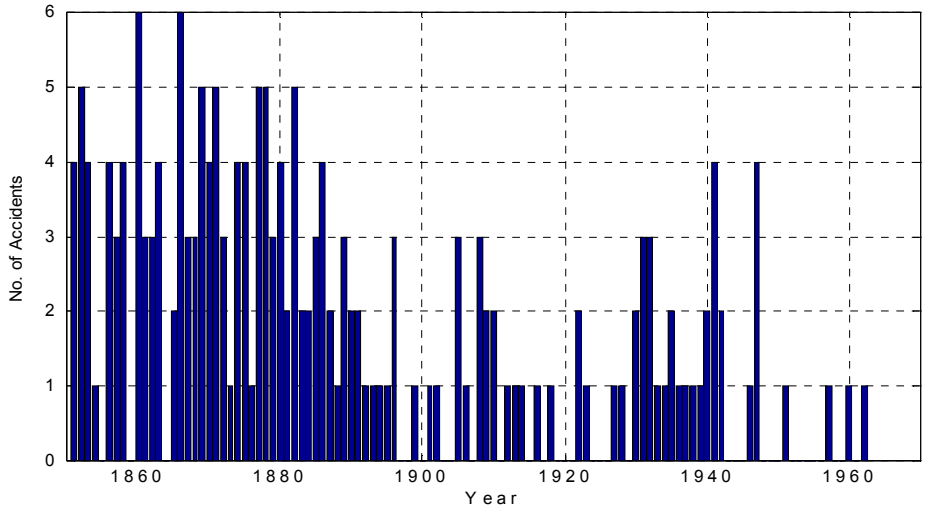
Henderson and Matthews (1992) compared the models from 0 to 3 possible change-points pairwise and concluded that there are 2 change-points for Birmingham at 11 and 16 (1980, 1985) and 3 change-points for Newcastle at 2, 7 and 15 (1970, 1976 and 1984). Our results from Table 4.3, however, do not strongly support this.

Example 2.2

As a second example of the Poisson model we will use the much analysed data set of yearly numbers of British coal-mining disasters during the 112-year period 1851-1962, gathered by Maquire, et al.(1952), extended and corrected by Jarrett (1979). Frequentist change-point investigations appear in Worsley (1986) and in Siegmund (1988), while Raftery and Akman (1986) apply their Bayesian model to investigate a continuous single change-point. Broemeling and Grequrich (1996) investigated a discrete single change-point, while Carlin, Gelfand and Smith (1992)

used Gibbs sampling in examining for a single change-point. Green (1995) considered multiple change-points with the reversible jump algorithm.

Figure 4.5: *Number of British coal-mining disasters during 1851 – 1962.*



Assuming one change-point, Carlin, Gelfand and Smith (1992) found a maximum probability of 0.2750 at 1891 (k = 41) with $\alpha = \frac{1}{2}$ and $\beta = 0$. The same result is obtained from equation (2.4). Equation (2.5) gives a maximum probability of 0.2421 (see Figure 4.6), while the fractional Bayes factor from (2.21) gives a probability of 0.2372 at 1891 with the probability for no change virtually zero. Allowing for at most 3 change-points, the posterior probabilities from equations (3.6) to (3.8) are given in Table 4.4, together with the results of Green (1995) who used the reversible jump algorithm with a Poisson prior on k with mean 3.

Figure 4.6: *Posterior probability for position of a single of change-point: Coalmine data*

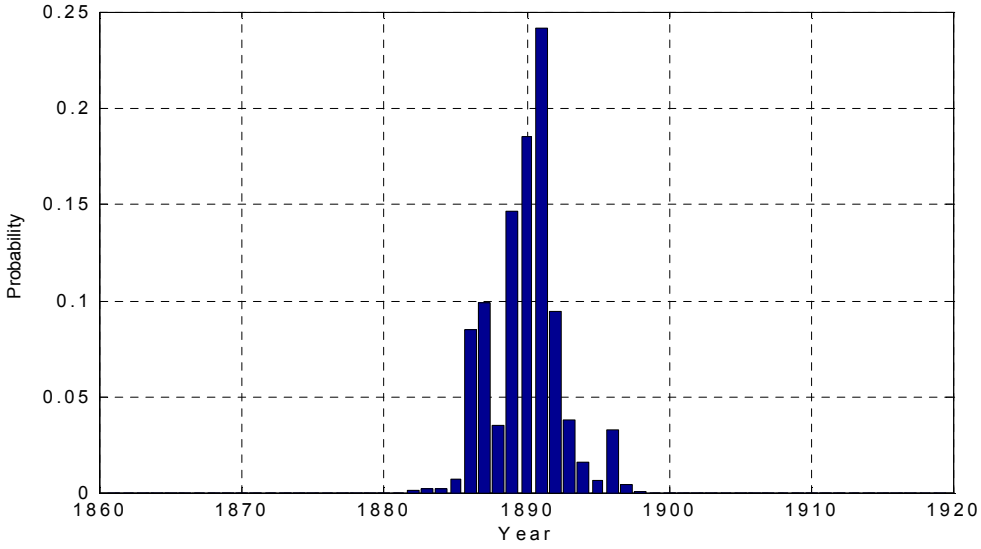


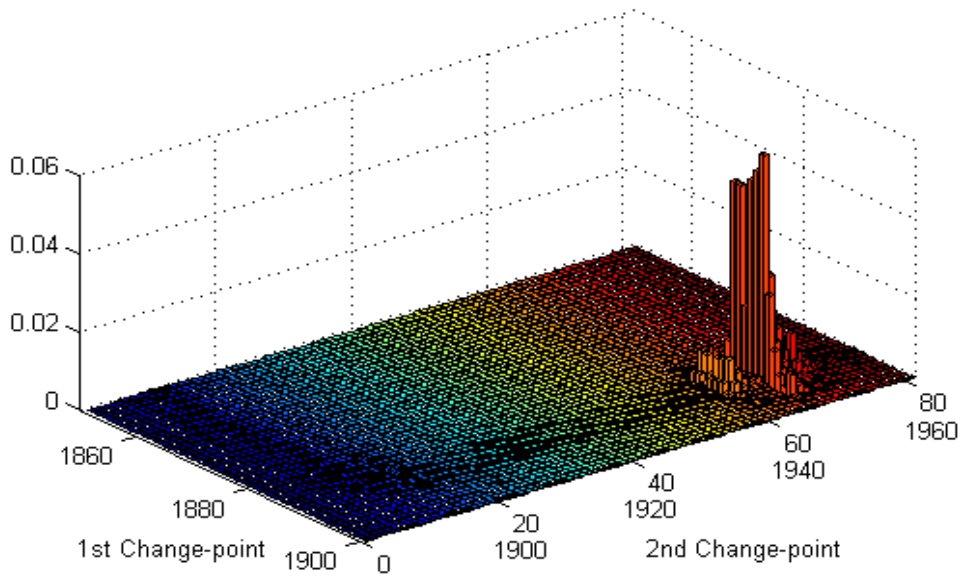
Table 4.4: *Posterior probabilities for the number of change-points: Coal-mine data*

No. of change-points	r = 0	r = 1	r = 2	r = 3	r ≥ 4
Change-point(s)	---	k=41	$\boldsymbol{k}$ = (41,97)	$\boldsymbol{k}$ = (41,79,97)	---
Posterior probability	3.9e-014	0.1763	0.4716	0.3521	---
Green (1995)	0	0.157	0.348	0.266	0.229

The evidence points to 2 change-points with maximum probability at $\boldsymbol{k}$ = (41, 97) which are 1891 and 1947. Worsley (1986) and Raftery and Akman (1986) give some possible historic reasons for the possible change-points. According to Worsley changes in the coal-mining regulations during 1896 may have reduced the probability of accidents. According to Raftery and Akman a fairly abrupt decrease around 1887-1895 may be associated with changes in the coal industry around that time, namely a severe decline in labour productivity starting at the end of the 1980's, an the emergence of the Miner's federation at the end of 1889. The change in 1947 may be due to changes in labour practice just after the war.

Under model M_2, the joint posterior of k_1 and k_2 is shown in Figure 4.7. The posterior mass is clearly concentrated around $\boldsymbol{k}$ given above. The posterior distributions of λ_1 , λ_2 and λ_3 are virtually the same as shown in Figure 4.9 below, and so the number of disasters has been significantly reduced each time.

Fig 4.7*: Joint distribution of k_1 and k_2 given 2 change-points for Example 2.2.*



Chib's approach

Chib's approach (1998) will be illustrated using the coal mining-disaster data from Britain used above. Let the count x_t in the year t be modelled via a hierarchical Poisson model, and consider determining the change-points for each of the two models M_1 and M_2. Under M_1 the data is subject to a single break with

$$\lambda_t = \begin{cases} \lambda_1 & for\ t \le \tau_1, \\ \lambda_2 & for \tau_1 + 1 \le t \le 112 \end{cases}$$

where $\lambda_1, \lambda_2 \sim Gamma\,(2,1)$, as assumed by Chib.

First $S = \{s_1, s_2, \ldots, s_n\}$ is simulated according to equations (3.10) to (3.13) with a starting value of 0.99 for p_{11}, after which p_{11} is simulated from (3.15) with a = 10 and b = 0.1. Finally λ_j, j = 1,2, follows from (3.16) with c = 2 and d = 1. This cycle was run 10 000 times, and the results are represented in the following figure. Figure 4.8(a) shows a different way of representing the position of the change-point, which follows naturally from Chib's approach. It shows the probability of being in the 1st regime (before the change), or being in the second regime (after the change), as a function of the time. The point were the lines cross is where the probability of being in the 2nd regime exceeds 0.5. The result corresponds with that obtained earlier, namely a change around 1891.

Figure 4.8: *Posterior results:1 change-point for mining accidents, Chib's method.*

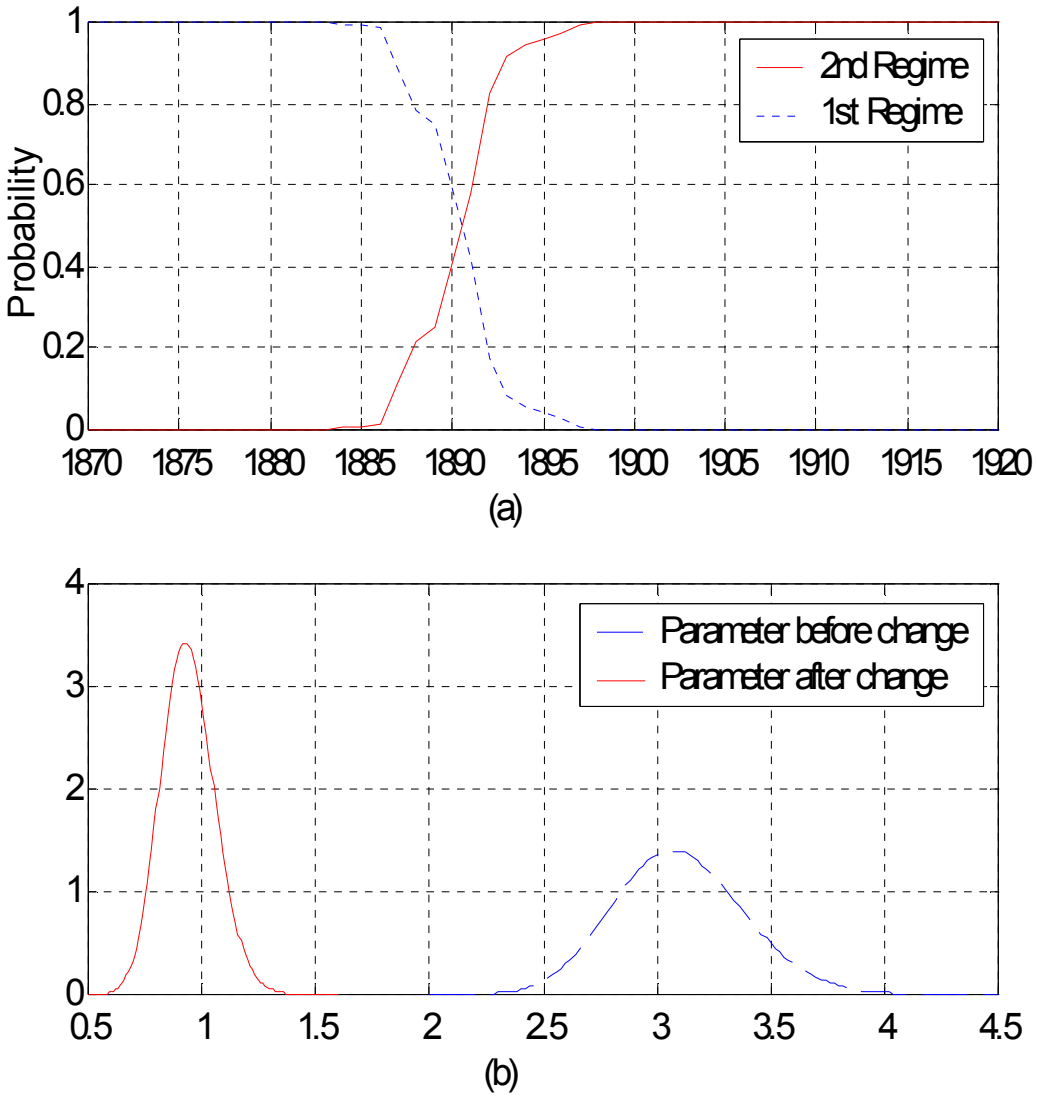
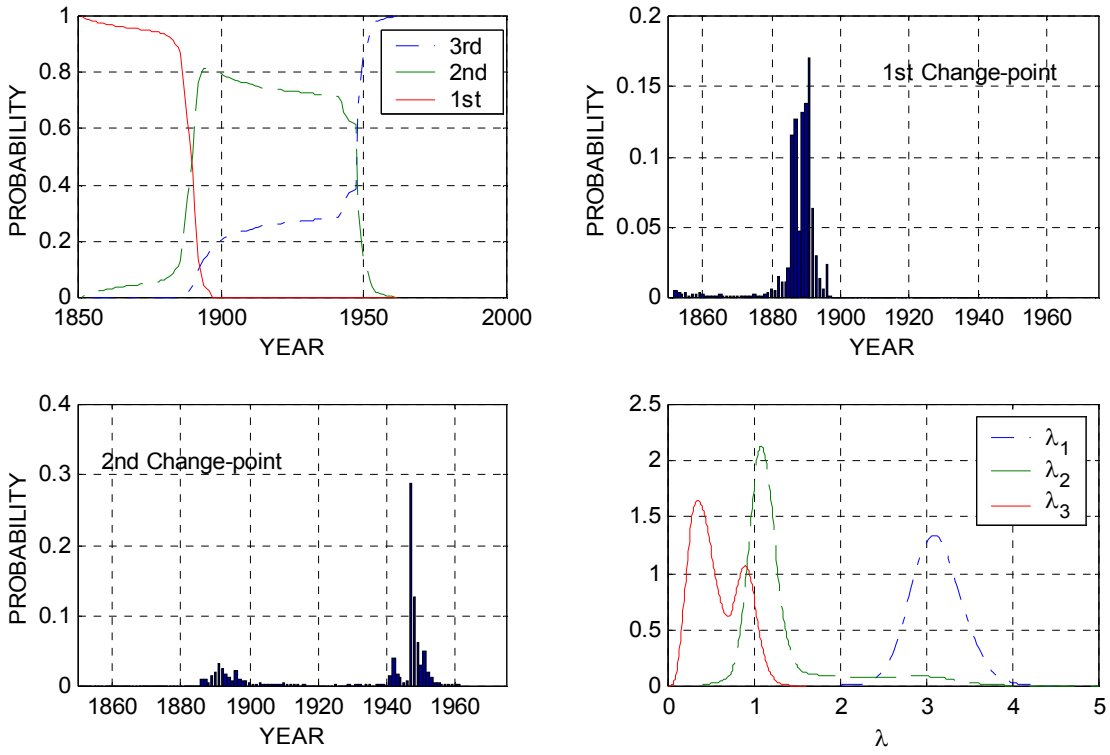


Figure 4.8(b) shows the posterior distributions of λ_1 and λ_2, clearly a significant decrease in the number of accidents, with the posterior mean of λ_1 equal to 3.099 and that of λ_2 equal to 0.938.

Under model M_2, λ_t is subject to two breaks with priors, $\lambda_1, \lambda_2, \lambda_3$ ~ Gamma (3,1). Results are shown in Figure 4.9, where the posterior means are:
$E[\lambda_1 \mid \boldsymbol{x}] = 3.118,\ E[\lambda_2 \mid \boldsymbol{x}] = 1.297,\ E[\lambda_3 \mid \boldsymbol{x}] = 0.506.$

Figure 4.9: *Posterior results: 2 change-points for mining accidents,Chib's method.*



When three change-points are assumed there is much more uncertainty as to the positions of the change-points, and the two change-point model seems the best, as confirmed in Table 4.4.

Example 2.3

Prussian military personnel killed by horse-kicks (1875-1894).
The "Horse-kicks" data of Bortkewitsch are amongst the most well-known collections of Poisson data. They summarise the number of Prussian military personnel killed by kicks of a horse for each of 14 corps in each of 20 successive years 1875-1894. The full data-table can be found in Hand et al. (1994) and is analysed by Preece et al. (1988).

The total number of deaths over all corps for all 280 years was 196. If each of the 280 years could reasonably be thought to be independent of all others, and the number of cavalry officers and their susceptibility to death from horse-kicks could be reasonably thought to be the same for each of the 280 units of observations, then a simple Poisson model for the observed frequencies would be reasonable. The expected frequencies for a Poisson distribution with mean $\frac{196}{280} = 0.700$ were given by Bortkewitsch and show a good agreement with the observed frequencies. Table 4.5 shows the posterior probability for no change for each of the 14 corps as well as the position and probability of the change-point with maximum probability.

Table 4.5: *Probability of no change and position of change-point for the horse-kicks data.*

Corps	Pr[k = 0|x]	Position of cp t	Pr[k = t|x]
G	0.6773	13	0.3130
I	0.5990	3	0.1035
II	0.6510	3	0.1034
III	0.6085	17	0.0586
IV	0.6669	7	0.0382
V	0.5908	4	0.0895
VI	0.5831	2 and 8	0.0487
VII	0.6787	12	0.0256
VIII	0.6892	16	0.0301
IX	0.4381	5	0.2112
X	0.6127	14	0.0534
XI	0.2002	4	0.5050
XIV	0.5732	18	0.0929
XV	0.6242	16	0.0518

It seems that there is little evidence of a change in the number of deaths caused by horse-kicks for many corps for the twenty years, with an exception of corps IX and XI, where there seems to be an abrupt change in the number of deaths at k = 5 (1879) and k = 4 (1878) respectively.

Analysing the totals over the 20 years and assuming the possibility of up to 5 change-points, using the fractional Bayes factor from equation (2.21) and b=4/n, we see that the highest probability is for 3 change-points from Table 4.6. However, any number from 1 to 5 has a reasonable probability, only the probability for no change is negligible.

Table 4.6: *Probabilities for the number of change-points in horse-kicks data*

No. of Change-point	0	1	2	3	4	5
Probability	0.030	0.143	0.212	0.227	0.209	0.179

The probability of no change-point for the totals is 0.173 when compared with the model with a single change-point, so, while it is unclear as to how many change-points there are, it seems that some changes did occur. This data set will be examined further in section 6.

5. MULTI-PATH CHANGE -POINT ANALYSIS.

Suppose we are given *M* sequences of random variables, each of length *N*, and we want to make inferences about a change-point τ_i, *i* = 1, 2,…,*M*, in each sequence. There are two main subdivisions of the multi-path change-point problem. If the change-point occurs at the same position in each sequence, (i.e. $\tau_1 = \tau_2 = ... = \tau_M = \tau$) or if the $\tau_i's$ occur at random positions in each sequence, $1 \le \tau_i \le N-1$.

The multi-path analysis was described by Bélisle et al. (1998) as follows:
Assume that there are data in the form of an *M* x *N* array

$$\mathrm{X} = \begin{Bmatrix} x_{11} & x_{12} & \cdots & x_{1,\tau_1} & x_{1,\tau_1+1} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2,\tau_2} & x_{2,\tau_2+1} & \cdots & x_{2N} \\ \cdot & & & & & & \\ \cdot & & & & & & \\ \cdot & & & & & & \\ x_{M1} & x_{M2} & \cdots & x_{M,\tau_M} & x_{M,\tau_M+1} & \cdots & x_{MN} \end{Bmatrix} \qquad (5.1)$$

Each sequence, $\mathrm{X}_{i1},\ldots,\mathrm{X}_{iN}$, represents observations over time from the *i*-th subject, *i*=1,…,*M*. A change-point is said to have occurred at τ_i in sequence or row *i*, $1 \le \tau_i \le N-1$, if $\mathrm{X}_{i1},\ldots,\mathrm{X}_{i\tau_i}$ are identically distributed with common distribution F_{i1}, which is different from the common distribution , F_{i2} of $\mathrm{X}_{i\tau_{i+1}},\ldots,\mathrm{X}_{iN}$. If $\tau_i = N$, then no change has occurred in row *i*. The distribution of the point of change, τ_i, an unknown parameters of the distributions F_{ik}; *i* = 1, …, *M*, *k* = 1,2 is to be estimated from the data matrix (5.1).

We assume that the times of change, τ_i, in each sequence are themselves independent and identically distributed from a given population, following a distribution *g(t) = pr(τ_i = t), i = 1,…,M; t=1,…N* which is to be estimated. If *g(N)* > 0, then it is possible that there is no change in some rows. Here *g(.)* represent the probability for the location of the change point for a randomly selected individual in the population.

It has been shown by Hinkley (1970) that the single-path maximum likelihood estimator of the change-point is not consistent, but the non-parametric estimator of *g*(.) has been shown in Joseph and Wolfson (1992) and Joseph,Vandal and Wolfson (1996) to be consistent under certain conditions in the multi-path case. In Joseph and Wolfson (1992) both bootstrap and empirical Bayes methods have been utilised in the multi-path context, and in Joseph et al. (1997) Bayesian analysis was employed.

5.1 Estimation of the parameters via the Gibbs sampler.

The likelihood for the model described in equation (5.1) is given by

$$f(x \mid \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \tau_1, ..., \tau_M) = \prod_{i=1}^{M} \left\{ \prod_{j=1}^{\tau_i} f(x_{ij} \mid \lambda_{i1}) \right\} \left\{ \prod_{i=\tau_{i+1}}^{N} f_2(x_{ij} \mid \lambda_{i2}) \right\} \qquad (5.2)$$

where x_{ij} follows a Poisson distribution with parameter λ_{ij} and $\boldsymbol{\pi} = (\pi_1,...,\pi_N)$, where $\pi_k = pr(\tau_i = k)$, i = 1,…,M; k = 1, 2,..., N. The parameters in the model are;

1) $\boldsymbol{\lambda}_1 = (\lambda_{11},...,\lambda_{M1})$ and $\boldsymbol{\lambda}_2 = (\lambda_{12},...,\lambda_{M2})$, vectors of the means of the Poisson distributions before and after the change-point in each row.

2) $\boldsymbol{\pi} = (\pi_1,...,\pi_N)$, the multinomial probabilities that a change occurs at position k in each row, k = 1,…,N.

3) $\boldsymbol{\tau} = (\tau_1,...,\tau_M)$, the unobserved latent data representing the change-points in each row.

4) In addition we have the parameters, θ_1 and θ_2, of the exchangeable priors of λ_1 and λ_2.

Prior distributions

Let $(\pi_1,...,\pi_{N-1})$ ~ Dirichlet$(\alpha_1,...,\alpha_{N-1})$ so that the joint distribution of $\pi_1,...,\pi_{N-1}$ is given by

$$f(\pi_1,...,\pi_{N-1}) = \frac{\Gamma(\sum_{j=1}^{N-1}\alpha_j)}{\prod_{j=1}^{N-1}\Gamma(\alpha_j)}\prod_{j=1}^{N-1}\pi_j^{\alpha_j-1}. \tag{5.3}$$

Next, let

$$\lambda_{i1} \sim \exp(\theta_1) \text{ and } \lambda_{i2} \sim \exp(\theta_2), \tag{5.4}$$

where the hyperparameters θ_1 and θ_2 have independent vague Jeffreys priors, $\pi(\theta_1,\theta_2) \propto \theta_1^{-\frac{1}{2}}\theta_2^{-\frac{1}{2}}$.

We are mainly interested in the posterior distributions of θ_1, θ_2 and $\boldsymbol{\pi}$. Implementation of the Gibbs sampler to find the marginal posterior distributions requires the specification of the full conditional distribution of all the parameters, i.e. the conditional distribution of each parameter given the values of all of the other parameters. These are derived as follows:

Conditional distributions.

The likelihood is

$$\begin{aligned} f(\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2,\tau_1,...,\tau_M \mid \boldsymbol{x}) &\propto \prod_{i=1}^{M}\left\{\prod_{j=1}^{\tau_i}e^{-\lambda_{i1}}\lambda_{i1}^{x_{ij}}\right\}\left\{\prod_{j=\tau_i+1}^{N}e^{-\lambda_{i2}}\lambda_{i2}^{x_{ij}}\right\} \\ &\propto \prod_{i=1}^{M}e^{-\tau_i\lambda_{i1}}e^{-(N-\tau_i)\lambda_{i2}}\lambda_{i1}^{y_{i1}}\lambda_{i2}^{y_{i2}}, \end{aligned} \tag{5.5}$$

where $y_{i1} = \sum_{j=1}^{\tau_i}x_{ij}$ and $y_{i2} = \sum_{j=\tau_i+1}^{N}x_{ij}$, so that with prior (5.4),

$$\lambda_{ij} \mid x_{ij},\tau_i,\theta_1 \sim Gamma(y_{ij}+1, \theta_j+\tau_i), \quad i = 1, 2,\ldots, M, \quad j = 1,2. \tag{5.6}$$

For the hyperparameters we have

$$\theta_j \mid \boldsymbol{\lambda}_j \sim Gamma(M+1/2, \sum \lambda_{ij}) \ , \quad j = 1,2. \tag{5.7}$$

For $\boldsymbol{\tau}$ we have

$$\Pr[\tau_i = k \mid \lambda_1, \lambda_2, \boldsymbol{\pi}, \boldsymbol{x}] \propto f(\lambda_1, \lambda_2, \tau_1, ..., \tau_M \mid \boldsymbol{x}) \Pr[\tau_i = k]$$

$$\propto \frac{\lambda_{i1}^{y_i(k_1)} \lambda_{i2}^{y_i(k_2)} e^{-k\lambda_{i1}} e^{-(N-k)\lambda_{i2}}}{\prod_{i=1}^{M} \Gamma(y_i(k_1)+1)\Gamma(y_i(k_2)+1)} \pi_k , \tag{5.8}$$

where $y_i(k_1) = \sum_{j=1}^{k} x_{ij}$ and $y_i(k_2) = \sum_{j=k+1}^{N} x_{ij}$.

Lastly, the conditional distribution of the elements of $\boldsymbol{\pi}$ follow a Dirichlet distribution,

$$\pi_1, ..., \pi_{N-1} \mid \boldsymbol{\tau} \sim Dirichlet\ (\boldsymbol{\beta}) \ , \tag{5.9}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_{N-1})$, and $\beta_j = \alpha_j + \sum_{i=1}^{M} I(\tau_i = k)$. The indicator function I is one if $\tau_i = k$ and zero otherwise.

The Gibbs sampler algorithm proceeds by drawing a random sample from each full conditional distribution (5.5) to (5.9) in turn. The parameters sampled from the immediately preceding random draw are used in the conditional distribution for subsequent draws. A large number of iterations are run, and after discarding iterates from an initial burn-in period to allow for the convergence of the algorithm, the remaining random vectors can be regarded as samples from the joint posterior distribution of the parameters, from which inference can be made. Marginal posterior density estimates can be obtained by what has become known as the Rao-Blackwell method (Gelfand and Smith (1990)). For example, the marginal posterior distribution of λ_{i1} can be obtained as

$$\pi(\lambda_{i1} \mid \boldsymbol{x}) = \frac{1}{L} \sum_{l=1}^{L} f(\lambda_{is} \mid \boldsymbol{x}, \tau_i^{(l)}, \theta_1^{(l)}) , \tag{5.10}$$

where L is the total number of cycles and $(\tau_i^{(l)}, \theta_1^{(l)})$ are the generated values during the *l*-th cycle. So we are averaging the conditional distribution of λ_{i1} over the simulated values of the conditioning parameters.

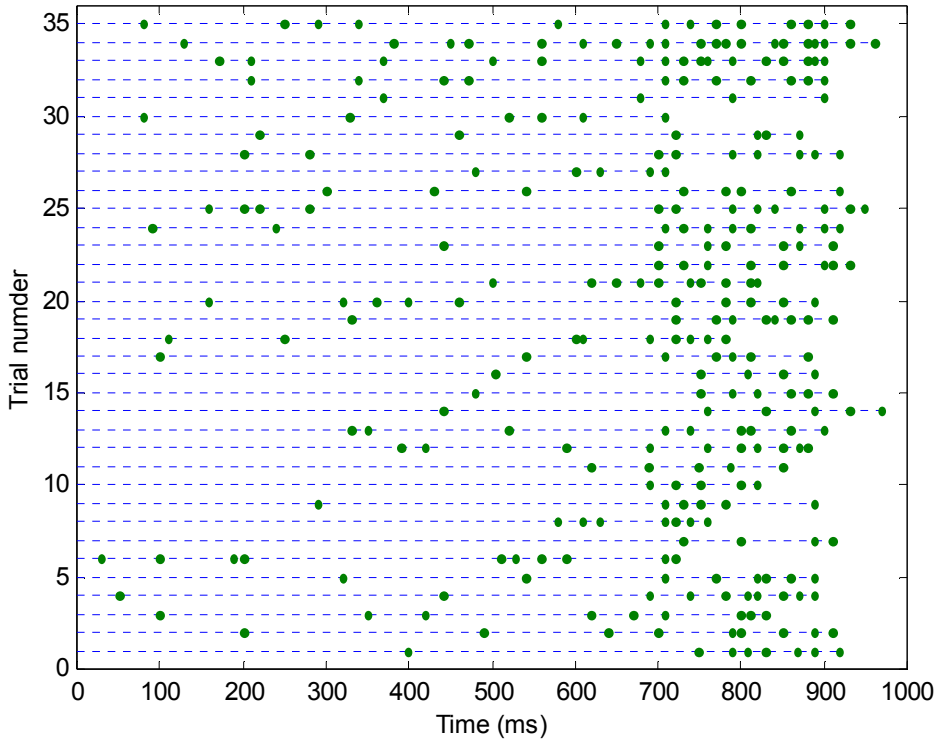
6. APPLICATIONS 2

Example 6.1: *Neuron spike train analysis.*

As an example of the Poisson model with multi-path change-points, we will use data from Bélisle et al. (1998), consisting of counts of electrical discharges in 20 milliseconds (ms) intervals, approximately one-half second before and after a stimulus was applied to the neuron at *t* = 500 ms. The counts of electrical discharges were observed on *M* = 35 data sequences. Each time the neuron was allowed to the resting state before the experiment was resumed.

All sequences had 25 observations before the stimulus was applied, but the number of observations after the stimulus varied between 11 and 24. The variation should not cause substantial bias in estimating π unless there is evidence that the change point occurred after approximately 220 ms post-stimulus, which was not the case in this data set.

Figure 6.1 *Data from the experiment with M=35 trials. A stimulus was applied at 500ms. Each line represents a spike train at the indicated time.*

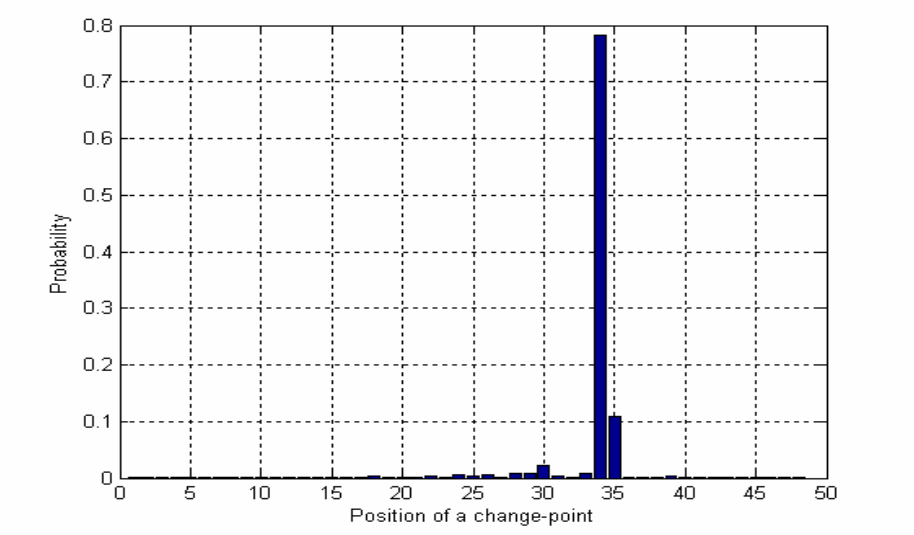


The output produced by the Gibbs sampler for π (from (5.9)) is a sample from a Dirichlet distribution in *N* = 49 dimensions. Summary statistics marginal Dirichlet posterior distributions can be calculated, and posterior marginal densities for selected change-point probabilities may be plotted. within each iteration, each sequence may have τ_i < *N* or τ_i = *N*, *i* = 1,...,*M*. A useful statistics is then $\{\#times\ \tau_i < N\}$/number of iterations. This approximates the sequence or trial-specific probability of a change-point.

To obtain relatively flat prior densities, so that the data themselves would contribute most of the information in the posterior densities, a Dirichlet prior density (from (5.3)) with α_i = 0.05 for all *i* was used. Bélisle et al. (1998) used $\alpha_1 = \alpha_2 = ... = \alpha_{24}$ = 0 and $\alpha_{25} = \alpha_{26} = ... = \alpha_{49} = 0.04$. The sample size equivalent of this prior density is two and a half observations ($\sum \alpha_i = 2.5$), so that 35/37.5 = 93% of the information in the marginal posterior density on π would come from the data. Bélisle et al. (1998) used Gamma(4, 0.03) and Gamma(8, 0.03) prior distributions for the before and after Poisson parameters, where we use Exponential priors with vague hyperpriors on its parameters, as given in equation (5.4).

Using the MATLAB software, 25000 sets of parameter values were generated. The mean marginal posterior change-point probability at τ_{34} was 0.855, indicating that there is indeed a change in electrical activity following the application of the stimulus, occurring roughly 180 ms after the stimulus. None of the other change-point mean marginal probabilities was greater than 0.025, and in particular, there was a negligible estimated probability of no change. The estimated posterior mean of $\boldsymbol{\pi}$ is depicted in Figure 6.2, and shows the high probability of 0.78 for a change after the 34th interval.

Figure 6.2: *Posterior means of probabilities* $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$*, neuron data.*



The marginal posterior distribution for π_{34} is given in figure 6.3. This figure indicates a 95% Highest Posterior Density interval of (0.60 – 0.99).

Figure 6.3: *Histogram of posterior probability distribution of* π_{34}

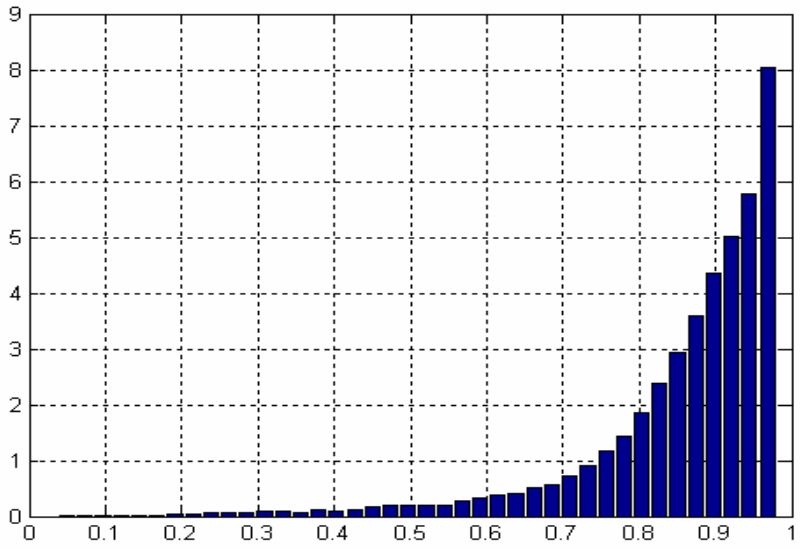
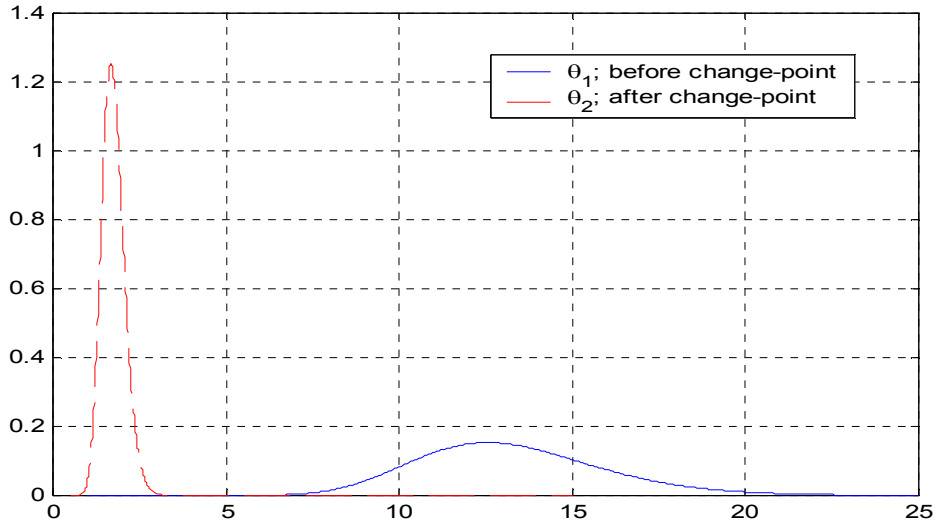


Figure 6.4 shows the posterior distributions of θ_1 and θ_2, derived by averaging over the conditional Gamma distributions given in (5.7). It shows clearly the difference between the mean rates before and after the change-point.

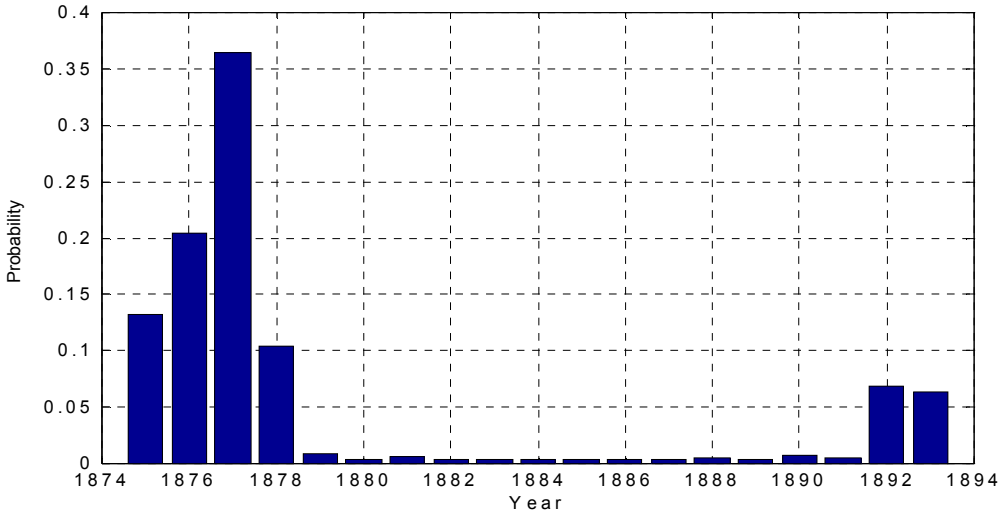
Figure 6.4: *Posterior distributions of θ_1 and θ_2 for neuron spike train data*



Example 6.2: *Horse-kicks*

The "horse-kicks" data is described in section 4, and here we will analyse it as multi-path data with 14 sequences, each of length 20. Assuming one change-point in each sequence, we want to derive the posterior distribution of $\boldsymbol{\pi}=(\pi_1, \pi_2, \ldots, \pi_{19})$, the probabilities for the position of the change-point. This is depicted in Figure 6.5.

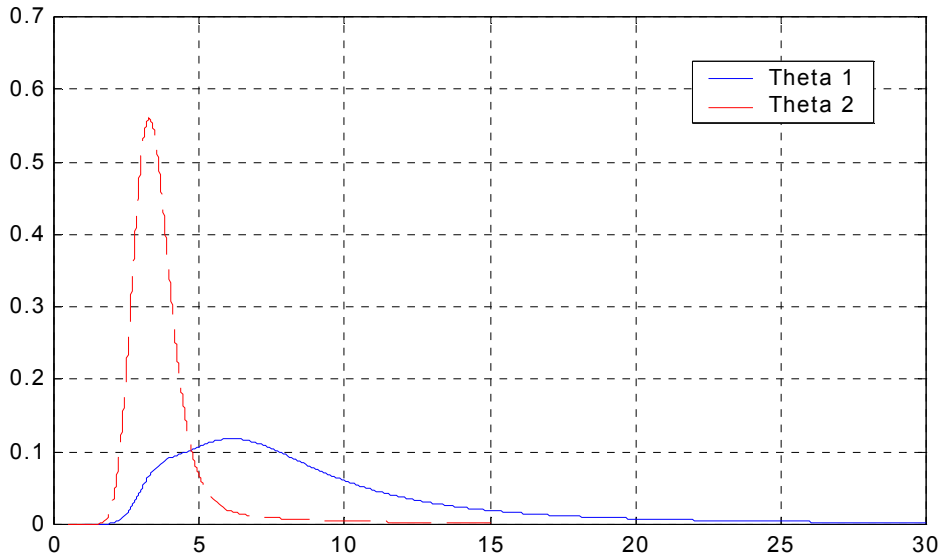
Figure 6.5: *Posterior means of probabilities $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$, horse-kick data*



It is clear that if there is a change, it is most likely to be early in the sequence, with a probability of 0.364 for 1877 and a probability of 0.805 that it is during the first

four years from 1875 to 1878. The marginal posteriors of θ_1 and θ_2 is given in Figure 6.6 and show the wide range of possible values, especially for θ_1. This is because of the uncertainty about the change-points in the sequences. From the analysis in section 4 it appears that most sequences exhibit more than one change-point.

Figure 6.6: *Marginal posterior distributions of θ_1 and θ_2, horse-kicks data.*



The years do not show an obvious trend except for the first 6 years, when there was a consistent increase in deaths. Corps G, I, VI and XI, which were noted as having a numerical composition particularly far from the average, have four of the five highest counts of deaths, the other corps with a high count being XIV.

7. REFERENCES

[1] Bélisle, P., Josheph, L., MacGibbon, B., Wolfson, D.B. and du Berger, R. (1998): Change-point analysis of neuron spike train data. *Biometrika*, **54** , 113 - 123.

[2] Berger, J. and Pericchi (1996): The intrinsic Bayes Factor for model selection and prediction. *J. Amer. Stat. Ass.*, **91**, 109 - 122.

[3] Bloemeling, L.D. and Gregurich, M.A. (1996): On a Bayesian approach for the shift point problem. *Comm. Stat-Theory Meth.*, **25**, 2267 - 2279.

[4] Carlin, B.P., Gelfand, A.E. and Smith, A.F.M. (1992): Hierarchical Bayesian analysis of changepoint problems. *Appl. Statist.*,**41**, 389 – 405.

[5] Chernoff, H and Zacks, S.(1964): Estimating the current mean of normal distribution which is subject to changes in time. *Ann. Math. Stat*. **35**, 999 - 1018.

[6] Chibb, S. (1995): Marginal likelihood from the Gibbs output. *J. Amer. Stat. Ass.*, **90**, 1313 – 1321.

[7] Chibb, S. (1996): Calculating posterior distributions and model estimates in Markov mixture models. *J. Econ*., **75**, 79 – 98.
[8] Chibb, S.(1998): Estimation and comparison of multiple change-point models. *J. Econ*., **86**, 221 – 241.
[9] Coad, A.N.G.,Marshall, T., Rowe, B. and Taylor, C.M.(1991): Changes in the postentro pathic form of the haemolytic uremic syndrome in children. *Clin. Nephrol*., **35**, 10-16.
[10] Green, P.J (1995): Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika,* **82**, 711 - 732.
[11] Hand, D.J., Daly, F., Lunn, A.D., McConway, K. J. and Ostrowski, E. (1994): A handbook of small data sets. Chapman & Hall, London.
[12] Henderson, R. and Matthews, J.N.S (1983) : An investigation of change-points in the annual number of cases of haemolytic uraemic syndrome. *App. Statist*., **42**, 461-471.
[13] Hinkley, D.V. (1970): Inference about the change-point in a sequence of random variables. *Biomerika*, **57**, 1 - 17.
[14] Jarrett, R.G.(1979): A note on the intervals between coal-mining disasters. *Biometrika*, **66**, 191 - 193.
[15] Jeffreys, H.(1961): Theory of probability, 3rd edition. Oxford University Press.
[16] Joseph, L. and Wolfson, D.B.(1992): Estimation in the multi-path change-point problems. *Comm. Stats-Theory Meth*., **21**, 897 - 913.
[17] Joseph, L., Vandal, A. and Wolfson, D. B. (1996): Estimation in the multi-path change-point problem for correlated data. *Can. J. Statist.* **24**, 37 – 53.
[18] Joseph, L.,Wolfson, D.B., du Berger, R. and Lyle, R.M. (1997): Analysis of panel data with change-points. *Statistica Sinica*, **7**, 687 - 703.
[19] Kass, R.E and Raftery, A.E. (1993): Bayes Factors. *J.Amer. Stat. Ass*., **90**, 773 -795.
[20] Levin, M and Barrett, T.M(1984): Haemolytic uraemic syndrome. *Arch. Dis.Chldhd*., **59**, 397-400.
[21] Maquire, B.A.,Pearson, E.S. and Wynn,A.H.A.(1952):The time intervals between industrial accidents. *Biometrika*, **38**, 168 - 180.
[22] O'Hagan, A. (1995): Fractional Bayes factors for model comparison. *J. Royal Stat. Soc. B.*, **57**, 99 - 138.
[23] Preece, D.A., Ross, G.J.S. and Kirby, S.P.J. (1988): Bortkewitsch's horse-kicks and the generalized linear model. *The Statistician*, **37**, 313 - 318.
[24] Raftery, A.E. and Akman, V.E. (1986): Bayesian analysis of a Poisson process with a change-point. *Biometrika*, **73**, 85 - 89.
[25] Siegmund, D.(1986): Boundary cross probabilities and statistical applications. *Ann. Stat*.,**14**, 361-404.
[26] Smith, A.F.M. (1975): A Bayesian approach to inference about a change-point in a sequence of random variables. *Biometrika*, **62**, 407 - 416.
[27] Tarr,P.I., Neill, A.M.,Allen, J., Siccardi, C.J. Watkins, S.L. and Hickman, R.O.(1989): The increasing incidence of the haemolytic uremic syndrome in King country, Washington: Lack of evidence for entertainment bias. *Am. J. Epedem*.,**129**,582-586.
[28] Worsley,K.J.(1986): Confidence Regions and tests for a change-point in a sequence of exponential family random variables. *Biometrika*., **73(1**),91-104 .

[29] Yao, Y.-C. (1987): Approximating the distribution of the maximum likelihood estimate of the change-point in a sequence of independent random variables. *Ann. Statist.*, **15**, 1321 – 1328.
[30] Zacks, S (1983): Survey of classical and Bayesian approaches to the change-point problem : Fixed sample and sequential procedures of testing and estimation. In Recent advances in statistics: Herman Chernoff Festschrift, 245 - 269, Academic Press, New York/ London.