

Analytical approach to changepoint detection in Laplacian noise

M. Wu
W.J. Fitzgerald

Indexing terms: Bayesian inference, Changepoint detection, Laplace distribution, Markov chain Monte Carlo method

Abstract: The paper presents an analytical method using the Bayesian inference framework for the identification of time-series discontinuities, i.e. changepoints, in impulsive Laplacian noise. Exact expressions for the posterior density of the changepoint positions and the associated Bayesian model evidence are given for DC step changes. The performance of the analytical approach is compared to that predicted by a Gaussian assumption to the noise statistics and Markov chain Monte Carlo methods.

1 Introduction

The detection and location estimate of changepoints (parametric discontinuities) in data are of considerable interest to a variety of data analysis areas. One example is the analysis of well log data, which conveys geophysics information about rock structure. A changepoint in such data normally indicates a boundary between two different rock strata. Other applications can be found in areas such as medicine, edge detection in image processing and so on.

Recent work on changepoint detection has been largely dependent on the assumption that noise can be modelled as normally (Gaussian) and independently distributed. One can justifiably expect the distribution of the observations to be approximately Gaussian on the grounds that the noise arises from a multitude of independent sources which yields a central limit tendency. However, there are cases where the overall noise distribution is determined by a dominant non-Gaussian source. In these circumstances the central limit theorem, which describes an asymptotic property, no longer provides justifications for the use of a Gaussian assumption. One category of non-Gaussian noise frequently encountered in practice is impulsive noise, which can be characterised by a distribution that tends to have a less pronounced 'shoulder' and a heavier 'tail'. The Laplace probability density function (PDF) can be adopted to model the impulsive or burst-type of noise, namely,

$$p(x) = \frac{\exp(-|x|/\lambda)}{\lambda} \quad (1)$$

The simple exponential behaviour of the Laplacian PDF (rather than the quadratic exponential of the Gaussian PDF) means high amplitudes having correspondingly higher probabilities, a characteristic of the impulse type of noise. The Laplace distribution can be viewed as a special member of a family of symmetric distributions clustered about the normal [1]. They can be written in the general form

$$p(y|\theta, \phi, \beta) = k\phi^{-1} \exp\left(-\frac{1}{2}\left|\frac{y-\theta}{\phi}\right|^{2/(1+\beta)}\right) \\ -\infty < y < +\infty \quad (2)$$

where it can be easily seen that when $\beta = 0$, the distribution is normal. When $\beta = 1$, the distribution is Laplacian. When $\beta \rightarrow -1$, it can be shown that the distribution tends to the rectangular distribution. If β is employed as a measure of nonnormality, the Laplace distribution will thus appear to be as discrepant from the Gaussian distribution as the rectangular distribution.

Practical cases abound in which the additive noise is more appropriately modelled as Laplacian. The Laplace distribution can be used as a model for the distributions of clicks that degrade audio signals and noises such as scratches or dropouts that corrupt images. The well log data mentioned earlier can also be very impulsive and consequently sometimes it is more appropriate to assume the corrupting noise distribution to be Laplacian. In all those cases, the use of Laplacian statistics is more robust than the Gaussian. One difficulty with the analysis of those data contaminated by Laplace distributed noise, as opposed to Gaussian noise, is that posterior distributions are generally not expressible in terms of simple mathematically tractable functions of the observations.

2 Single changepoint detection

Consider the case of single changepoint detection. Suppose that there are N samples of data \mathbf{d} corrupted by Laplace noise \mathbf{e} , i.e.

$$d_i = \begin{cases} \mu_1 + e_i & 1 \leq i \leq m \\ \mu_2 + e_i & m+1 \leq i \leq N \end{cases} \quad (3)$$

where m denotes the position of the changepoint and μ_1 and μ_2 are signal levels prior to and after the changepoint, respectively. Bayes' theorem states that the probability distribution for m posterior to the data \mathbf{d} is proportional to the product of the distribution for m prior to the data and the likelihood for \mathbf{d} given m . Assuming that these N observations are independently

© IEE, 1995

Paper 1919K (E5), received 19th September 1994

The authors are with the Signal Processing & Communications Laboratory, Cambridge University Engineering Department, Trumpington Street, Cambridge CB2 1PZ, United Kingdom

drawn, the likelihood function is

$$p(\mathbf{d} | m, \lambda, \mu_1, \mu_2) \propto \prod_{i=1}^N \frac{\exp(-|e_i|/\lambda)}{\lambda} \propto \exp\left[\frac{\left(\sum_{i=1}^m |d_i - \mu_1|\right)}{-\lambda}\right] \times \frac{\exp\left[\frac{\left(\sum_{i=m+1}^N |d_i - \mu_2|\right)}{-\lambda}\right]}{\lambda^N} \quad (4)$$

As little is known *a priori* in relation to the parameters, a noninformative prior has to be adopted to characterise this state of ignorance. For m , which, barring it occurring at two ends, can only take on values between 2 and $N - 1$, a locally uniform prior is assigned. μ_1 and μ_2 are supposed *a priori* to take on any arbitrary real values and hence they should be assigned a uniform prior over the entire range. λ , without loss of generality, is assumed to be positive only, in which case, following Jeffreys [2], its noninformative prior is represented by its logarithm being uniformly distributed. Further assume that all these four parameters are judged *a priori* to be distributed independently with each other, then

$$p(m, \lambda, \mu_1, \mu_2) \propto \frac{u(m-2) - u(m-(N-1))}{N-2} \frac{u(\lambda)}{\lambda} \quad (5)$$

Here $u(x)$ denotes the unit step function. Using Bayes' theorem, one can therefore express the joint posterior distribution as

$$p(m, \lambda, \mu_1, \mu_2 | \mathbf{d}) \propto u(\lambda)[u(m-2) - u(m-(N-1))] \exp\left[\frac{\left(\sum_{i=1}^m |d_i - \mu_1|\right)}{-\lambda}\right] \exp\left[\frac{\left(\sum_{i=m+1}^N |d_i - \mu_2|\right)}{-\lambda}\right] \times \frac{1}{\lambda^{N+1}} \quad (6)$$

From the Bayesian perspective, the overall inferences which can be made about m from the observation data \mathbf{d} are summarised in its marginal posterior distribution, which results from integrating out the nuisance parameters λ, μ_1, μ_2 from the joint distribution. That is

$$p(m | \mathbf{d}) \propto [u(m-2) - u(m-(N-1))] \int_0^\infty \left\{ \int_{-\infty}^\infty \exp\left[\frac{\left(\sum_{i=1}^m |d_i - \mu_1|\right)}{-\lambda}\right] d\mu_1 \times \int_{-\infty}^\infty \exp\left[\frac{\left(\sum_{i=m+1}^N |d_i - \mu_2|\right)}{-\lambda}\right] d\mu_2 \right\} d\lambda \quad (7)$$

Consider the following integral

$$\int_{-\infty}^\infty \exp\left[\frac{\left(\sum_{i=1}^m |d_i - \mu_1|\right)}{-\lambda}\right] d\mu_1 \quad (8)$$

The presence of the absolute value sign in the formula obviously has made an otherwise easy integration very difficult. To circumvent this a novel solution is devised which divides the integration interval of μ_1 into a series of subintervals in accordance with the magnitude of the data samples $\{d_i\}$. The integral is to be re-evaluated on

these subintervals separately and subsequently summed up. First, let $\mathbf{d1} \triangleq \{d_i\}$, for $1 \leq i \leq m$ and $\mathbf{d2} \triangleq \{d_{m+j}\}$, for $1 \leq j \leq N - m$. Moreover, $\mathbf{d1}$ and $\mathbf{d2}$ are subject to the constraint such that their respective entries are arranged in an ascending fashion,

$$d1_i \leq d1_{i+1} \quad \text{for any } i \in [1, m-1]$$

and

$$d2_j \leq d2_{j+1} \quad \text{for any } j \in [1, n-1] \quad (9)$$

where

$$n \triangleq N - m \quad (10)$$

It is readily observed from the definitions that data vectors $\mathbf{d1}$ and $\mathbf{d2}$ are functions of m and the combination of the two represents just the re-ordered version of the original data vector \mathbf{d} . Although the value of m is yet to be estimated, a value can be tentatively assigned to it which then yields a corresponding pair of $\mathbf{d1}$ and $\mathbf{d2}$. The posterior density for m can thus be evaluated at this particular value. The algorithm works by exhaustively evaluating the posterior density at all possible values of m . The optimal estimate of m is the one that results in the maximum value of the posterior density.

Following the definitions of $\mathbf{d1}$, the integral of μ_1 can be rewritten as

$$\int_{-\infty}^\infty \exp\left[\frac{\left(\sum_{i=1}^m |d_i - \mu_1|\right)}{-\lambda}\right] d\mu_1 = \int_{-\infty}^{d1_1} \cdots d\mu_1 + \int_{d1_1}^{d1_2} \cdots d\mu_1 + \cdots + \int_{d1_m}^\infty \cdots d\mu_1 = \sum_{i=0}^m g_i(\lambda) \quad (11)$$

where

$$g_0(\lambda) \triangleq \int_{-\infty}^{d1_1} \exp\left[\frac{\left(\sum_{i=1}^m |d_i - \mu_1|\right)}{-\lambda}\right] d\mu_1 = \int_{-\infty}^{d1_1} \exp\left[\frac{\left(\sum_{i=1}^m |d1_i - \mu_1|\right)}{-\lambda}\right] d\mu_1 \quad (12)$$

$$g_i(\lambda) \triangleq \int_{d1_i}^{d1_{i+1}} \exp\left[\frac{\left(\sum_{i=1}^m |d_i - \mu_1|\right)}{-\lambda}\right] d\mu_1 = \int_{d1_i}^{d1_{i+1}} \exp\left[\frac{\left(\sum_{i=1}^m |d1_i - \mu_1|\right)}{-\lambda}\right] d\mu_1 \quad (13)$$

$$g_m(\lambda) \triangleq \int_{d1_m}^\infty \exp\left[\frac{\left(\sum_{i=1}^m |d_i - \mu_1|\right)}{-\lambda}\right] d\mu_1 = \int_{d1_m}^\infty \exp\left[\frac{\left(\sum_{i=1}^m |d_i - \mu_1|\right)}{-\lambda}\right] d\mu_1 \quad (14)$$

It can be readily shown that

$$g_0(\lambda) = \int_{-\infty}^{d1_1} \exp\left[\frac{\left(\sum_{i=1}^m (d1_i - \mu_1)\right)}{-\lambda}\right] d\mu_1 = \frac{\exp\left[\frac{\left(\sum_{i=1}^m d1_i\right)}{-\lambda}\right] \exp(-m d1_1 / -\lambda)}{(m/\lambda)} \quad (15)$$

$$g_m(\lambda) = \int_{d1_m}^\infty \exp\left[\frac{\left(\sum_{i=1}^m - (d1_i - \mu_1)\right)}{-\lambda}\right] d\mu_1 = \frac{\exp\left[-\frac{\left(\sum_{i=1}^m d1_i\right)}{-\lambda}\right] \exp(m d1_m / -\lambda)}{(m/\lambda)} \quad (16)$$

When μ_1 falls into any of the intermediate regions $[d1_i, d1_{i+1}]$,

$$\sum_{i=1}^m |d1_i - \mu_1| = \sum_{r=1}^i -(d1_r - \mu_1) + \sum_{r=i+1}^m (d1_r - \mu_1) \quad (17)$$

Hence,

$$g_i(\lambda) = \int_{d1_i}^{d1_{i+1}} \exp \left[\left(\sum_{r=1}^i -(d1_r - \mu_1) + \sum_{r=i+1}^m (d1_r - \mu_1) \right) / -\lambda \right] d\mu_1$$

$$= \left\{ \frac{\exp \left[\left(\sum_{r=i+1}^m d1_r - \sum_{r=1}^i d1_r \right) / -\lambda \right] \times [\exp \{(2i - m)d1_{i+1} / -\lambda\} - \exp \{(2i - m)d1_i / -\lambda\}]}{(2i - m)/\lambda} \right\}$$

$$i \in [1, m-1] \quad (18)$$

One important point worthy of note is that for m even, the value of the expression $\sum_{i=1}^m |d1_i - \mu_1|$ is constant inside $[d1_{m/2}, d1_{m/2+1}]$. In consequence, the above formula for g_i is invalid when $i = m/2$, since the denominator vanishes. It can however be modified to the following expression

$$g_{m/2} = \exp \left[\left(\sum_{r=m/2+1}^m d1_r - \sum_{r=1}^{m/2} d1_r \right) / -\lambda \right] \times (d1_{m/2+1} - d1_{m/2}) \quad (19)$$

Similarly, the integral with respect to μ_2 can be expressed as

$$\int_{-\infty}^{\infty} \exp \left[\left(\sum_{i=m+1}^N |d_i - \mu_2| \right) / -\lambda \right] d\mu_2 = \sum_{j=1}^n h_j(\lambda) \quad (20)$$

where

$$h_0(\lambda) \triangleq \int_{-\infty}^{d2_1} \exp \left[\left(\sum_{i=1}^m |d_i - \mu_1| \right) / -\lambda \right] d\mu_2$$

$$= \frac{\exp \left[\left(\sum_{j=1}^n d2_j \right) / -\lambda \right] \exp(-n d2_1 / -\lambda)}{n/\lambda} \quad (21)$$

$$h_j(\lambda) \triangleq \int_{d2_j}^{d2_{j+1}} \exp \left[\left(\sum_{i=1}^m |d_i - \mu_2| \right) / -\lambda \right] d\mu_2$$

$$= \left\{ \frac{\exp \left[\left(\sum_{s=j+1}^n d2_s - \sum_{s=1}^j d2_s \right) / -\lambda \right] \times [\exp \{(2j - n)d2_{j+1} / -\lambda\} - \exp \{(2j - n)d2_j / -\lambda\}]}{(2j - n)/\lambda} \right\}$$

$$j \in [1, n-1] \quad (22)$$

$$h_n(\lambda) \triangleq \int_{d2_n}^{\infty} \exp \left[\left(\sum_{j=1}^n |d_j - \mu_2| \right) / -\lambda \right] d\mu_2$$

$$= \frac{\exp \left[- \left(\sum_{j=1}^n d2_j \right) / -\lambda \right] \exp(n d2_n / -\lambda)}{n/\lambda} \quad (23)$$

With μ_1 and μ_2 having been integrated out, the marginal posterior distribution for m is shown to be

$$p(m | \mathbf{d}) \propto [u(m-2) - u(m-(N-1))] \times \int_0^{\infty} \frac{\sum_{i=0}^m h_i(\lambda) \sum_{j=0}^n g_j(\lambda)}{\lambda^{N+1}} d\lambda$$

$$\propto [u(m-2) - u(m-(N-1))] \times \sum_{i=0}^m \sum_{j=0}^n \int_0^{\infty} \frac{h_i(\lambda) g_j(\lambda)}{\lambda^{N+1}} d\lambda \quad (24)$$

Consider the integration term inside the double summation. Rearranging the integrand gives rise to a gamma integral form, i.e. if

$$c_{ij}^{(1)} \triangleq \sum_{r=i+1}^m d1_r - \sum_{r=1}^i d1_r + \sum_{s=j+1}^n d2_s - \sum_{s=1}^j d2_s + (2i - m)d1_{i+1}(2j - n)d2_{j+1} \quad (25)$$

$$c_{ij}^{(2)} \triangleq \sum_{r=i+1}^m d1_r - \sum_{r=1}^i d1_r + \sum_{s=j+1}^n d2_s - \sum_{s=1}^j d2_s + (2i - m)d1_{i+1}(2j - n)d2_j \quad (26)$$

$$c_{ij}^{(3)} \triangleq \sum_{r=i+1}^m d1_r - \sum_{r=1}^i d1_r + \sum_{s=j+1}^n d2_s - \sum_{s=1}^j d2_s + (2i - m)d1_i(2j - n)d2_{j+1} \quad (27)$$

$$c_{ij}^{(4)} \triangleq \sum_{r=i+1}^m d1_r - \sum_{r=1}^i d1_r + \sum_{s=j+1}^n d2_s - \sum_{s=1}^j d2_s + (2i - m)d1_i(2j - n)d2_j \quad (28)$$

then

$$\int_0^{\infty} \frac{h_i(\lambda) g_j(\lambda)}{\lambda^{N+1}} d\lambda = \int_0^{\infty} \frac{(1/\lambda)^{N-1}}{(2i - m)(2j - n)} \exp \{[c_{ij}^{(1)}] / -\lambda\} d\lambda$$

$$- \int_0^{\infty} \frac{(1/\lambda)^{N-1}}{(2i - m)(2j - n)} \exp \{[c_{ij}^{(2)}] / -\lambda\} d\lambda$$

$$- \int_0^{\infty} \frac{(1/\lambda)^{N-1}}{(2i - m)(2j - n)} \exp \{[c_{ij}^{(3)}] / -\lambda\} d\lambda$$

$$+ \int_0^{\infty} \frac{(1/\lambda)^{N-1}}{(2i - m)(2j - n)} \exp \{[c_{ij}^{(4)}] / -\lambda\} d\lambda$$

$$\Gamma(N-2) \{ [c_{ij}^{(1)}]^{N-2} - [c_{ij}^{(2)}]^{N-2} - [c_{ij}^{(3)}]^{N-2} + [c_{ij}^{(4)}]^{N-2} \}$$

$$= \frac{\Gamma(N-2) \{ [c_{ij}^{(1)}]^{N-2} - [c_{ij}^{(2)}]^{N-2} - [c_{ij}^{(3)}]^{N-2} + [c_{ij}^{(4)}]^{N-2} \}}{(2i - m)(2j - n)} \quad (29)$$

Finally, one arrives at the following expression for the posterior density of m , which is a function of m and \mathbf{d} only

$$p(m | \mathbf{d}) \propto [u(m-2) - u(m-(N-1))] \Gamma(N-2) \{ [c_{ij}^{(1)}]^{N-2} - [c_{ij}^{(2)}]^{N-2} - [c_{ij}^{(3)}]^{N-2} + [c_{ij}^{(4)}]^{N-2} \}$$

$$\times \sum_{i=0}^m \sum_{j=0}^n \frac{\Gamma(N-2) \{ [c_{ij}^{(1)}]^{N-2} - [c_{ij}^{(2)}]^{N-2} - [c_{ij}^{(3)}]^{N-2} + [c_{ij}^{(4)}]^{N-2} \}}{(2i - m)(2j - n)} \quad (30)$$

3 Multiple changepoint detection

The method developed in the previous Section can be easily extended to the problem of multiple changepoint detection. Suppose that it is known *a priori* that there are altogether T changepoints in a data vector \mathbf{d} of length N , denoted by m_1, m_2, \dots, m_T . The problem can be formulated by the following model:

$$d_i = \begin{cases} \mu_1 + e_i & 1 \leq i \leq m_1 \\ \mu_2 + e_i & m_1 + 1 \leq i \leq m_2 \\ \vdots & \\ \mu_{T+1} + e_i & m_T + 1 \leq i \leq N \end{cases} \quad (31)$$

Applying the same definition of $\mathbf{d1}$ and $\mathbf{d2}$, one can segment \mathbf{d} into $T + 1$ subvectors: $\mathbf{d}^1, \mathbf{d}^2, \dots, \mathbf{d}^T, \mathbf{d}^{T+1}$. And the entries of each subvectors are sorted in ascending order. Although the precise values of the set of changepoints are not available, a provisional set of changepoint values can be assigned. The resultant optimal estimate is the set that leads to MAP. The posterior distribution for the multiple changepoint is given by

$$p(m_1, m_2, \dots, m_T | \mathbf{d}) \propto V(\{m_i\}) \times \sum_{i_1=1}^{m_1} \sum_{i_2=1}^{m_2-m_1} \dots \sum_{i_{T+1}=1}^{N-m_T} \int_0^\infty \frac{g_{i_1}^{(1)}(\lambda) g_{i_2}^{(2)}(\lambda) \dots g_{i_{T+1}}^{(T+1)}(\lambda)}{\lambda^{N+1}} d\lambda \quad (32)$$

where $V(\{m_i\})$ is a function that determines the range of the values on which a set of T changepoints can possibly take. For example, when $T = 2$, $V(\{m_i\})$ takes the following form

$$V(\{m_i\}) = \{u(m_1 - 2) - u[m_1 - (N - 2)]\} \times \{u[m_2 - (m_1 + 1)] - u[m_2 - (N - 3)]\} \quad (33)$$

Following the definitions of the last Section, one can define

$$g_i^{(j)}(\lambda) \triangleq \int_{d_{ji}}^{d_{j+1}} \exp\left[-\left(\sum_{i=1}^{m_j-m_{j-1}} |d_{ji} - \mu_j|\right) / -\lambda\right] d\mu_j \quad (34)$$

Likewise, the integrand term inside the multisummation can be manipulated to a form with which use can be made of the gamma integral formulas. Summing all the integration results then amounts to the posterior density for a particular set of changepoints. Since the number of possible changepoint configurations are finite given N , the optimal set is the one among all the candidate changepoints that gives the maximum posterior density value.

If the corrupting noise is Gaussian, it is useful to formulate the changepoint problem by adopting a general matrix form. For instance, one can use the following to formulate the double-changepoint problem.

$$\begin{bmatrix} d_1 \\ \vdots \\ d_{m_1} \\ d_{m_1+1} \\ \vdots \\ d_{m_2} \\ d_{m_2+1} \\ \vdots \\ d_N \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_{m_1} \\ e_{m_1+1} \\ \vdots \\ e_{m_2} \\ e_{m_2+1} \\ \vdots \\ e_N \end{bmatrix} \quad (35)$$

Define \mathbf{G} to be the matrix whose elements are chosen to be either 0 or 1 according to the changepoint positions. If the matrix \mathbf{G} is of size $N \times M$ and the noise is assumed to be zero-mean Gaussian, O'Ruanaidh and Fitzgerald [4] show that

$$p(\{m_i\} | \mathbf{d}) \propto \frac{[\mathbf{d}^T \mathbf{d} - \mathbf{d}^T \mathbf{G} (\mathbf{G}^T \mathbf{G})^{-1} \mathbf{G}^T \mathbf{d}]^{-(N+M)/2}}{\sqrt{[\det(\mathbf{G}^T \mathbf{G})]}} \quad (36)$$

In some cases, the precise number of changepoints present in the data may not be available *a priori*. It has to be inferred as well as the associated locations. For the sake of simplicity, consider the scenario where there are conceivably either T_1 or T_2 of changepoints in the data. Let M_1 and M_2 denote the two models which assume T_1 and T_2 number of changepoints, respectively. Bayesian inference favours whichever model has a greater posterior probability, i.e.

$$p(M_1 | \mathbf{d}) \stackrel{T_1}{\cong} p(M_2 | \mathbf{d}) \quad (37)$$

From Bayes' rule, this can be rewritten as

$$p(\mathbf{d} | M_1) p(M_1) \stackrel{T_1}{\cong} p(\mathbf{d} | M_2) p(M_2) \quad (38)$$

The term $p(\mathbf{d} | M_i)$ is frequently referred to as the Bayesian evidence of the i th model, and it can be calculated by

$$p(\mathbf{d} | M_i) = \int_{\Omega_i} p(\omega_{i_1}, \dots, \omega_{i_n} | \mathbf{d}, M_i) \times p(\omega_{i_1}, \dots, \omega_{i_n} | M_i) d\omega_{i_1}, \dots, d\omega_{i_n} \quad (39)$$

where Ω_i denotes the parameter space of the i th model. The product term under the multi-integration sign is exactly the form which the previous treatment took as the joint posterior density. Nevertheless, the posterior densities obtained previously cannot be applied directly in this context without some necessary modifications. The reason is that in the foregoing analysis, improper unnormalised priors were used for the parameters. It has been widely appreciated that in any model comparison, the Bayesian evidence evaluation depends significantly on the prior distributions specified for parameters of each model. If competing linear models are of different dimensionality, the use of improper uninformative priors can mislead one to support the model with highest dimension. The fact that such a dependence can be rather striking is well illustrated by the phenomenon referred to as the Lindley Paradox [6].

To overcome this difficulty, instead of assuming the signal levels $\{\mu_i\}$ to be uniformly distributed over the entire real range, one can assume their distributions to be uniform over some finite ranges. There is however no need for λ 's prior to be normalised. Since it is used in all the model formulations, the impact of its prior on the different model evidences is equal. Thus as far as the comparison of different model evidences is concerned, the prior specifications of common parameters are irrelevant.

In the light of the discrete nature of the changepoint location parameters, the evidence expression, after integrating out all the other continuous parameters, is given by

$$p(\mathbf{d} | M_i) = \sum_{m_1} \dots \sum_{m_{T_i}} \frac{1}{S(M_i)} p(m_1, \dots, m_{T_i} | \mathbf{d}, M_i) \quad (40)$$

where $S(M_i)$ represents the size of the i th model's discrete changepoint parameter space, which is the number of total distinct candidate changepoint positions.

4 Simulation results

The performance of the proposed algorithm has been examined in comparison with methods based on a Gaussian assumption and Markov chain Monte Carlo algorithms, namely, the Gibbs sampler and the Metropolis algorithm [3, 4]. When the noise level is low, the differences between these methods are not appreciable, all of which can correctly estimate changepoint positions on simulated data. The advantage of the analytical approach manifests itself when the impulsiveness of the underlying noise gets severe. Consider the data shown in Fig. 1,

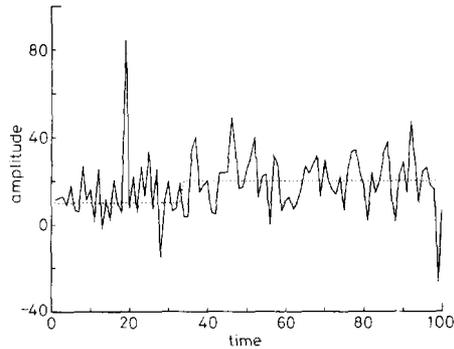


Fig. 1 Single changepoint data 1

where the dotted line represents the signal component and the corrupting noise has a Laplacian distribution with λ set at 10. It can be seen from Fig. 2 that under

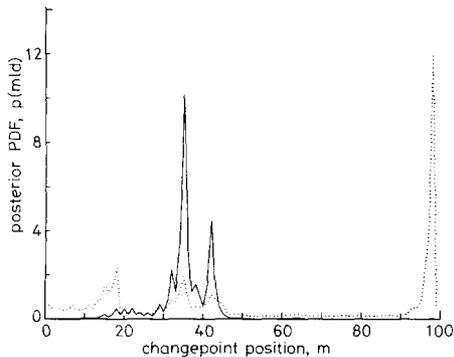


Fig. 2 Data 1: Posterior density for single changepoint in data 1 using analytical Laplacian approach compared with that using Gaussian assumption

— analytical Laplacian
 Gaussian assumption

such an intensively impulsive noisy environment, a Gaussian approximation no longer serves any use. It is particularly vulnerable to outliers occurring towards the two ends of a data record, as can be indicated by the resultant posterior density which shows that the Gaussian assumption misinterprets a noise outlier close to the end of the data as a signal step change.

Fig. 3 shows the histogram approximation to the posterior density $p(m)$ using the Gibbs sampler and Metropolis algorithm. Given the starting point $(\omega_1^{(0)}, \omega_2^{(2)}, \dots, \omega_i^{(0)})$, the Gibbs algorithm iterates the following loop:

- (a) Sample $\omega_1^{(i+1)}$ from $p(\omega_1 | \omega_2^{(i)}, \dots, \omega_i^{(i)}, \mathbf{d})$
- (b) Sample $\omega_2^{(i+1)}$ from $p(\omega_2 | \omega_1^{(i+1)}, \omega_3^{(i)}, \dots, \omega_i^{(i)}, \mathbf{d})$
- ...
- (i) Sample $\omega_i^{(i+1)}$ from $p(\omega_i | \omega_1^{(i+1)}, \omega_2^{(i+1)}, \dots, \omega_{i-1}^{(i+1)}, \mathbf{d})$

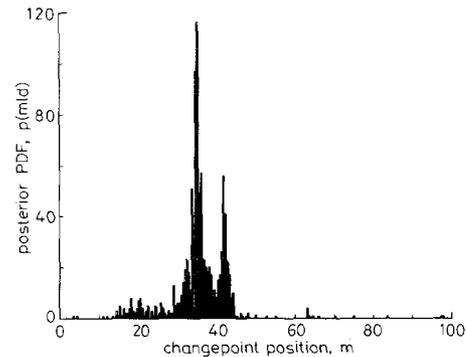


Fig. 3 Data 1: histogram plot of posterior density using Markov chain Monte Carlo methods

The vectors $\omega^0, \omega^1, \dots, \omega^n, \dots$ are a realisation of a Markov chain and converge geometrically to $p(\omega_1, \dots, \omega_i | \mathbf{d})$, as $n \rightarrow \infty$ [5]. Furthermore, $\omega_1^{(0)}, \omega_1^{(1)}, \dots, \omega_1^{(m)}, \dots$ converge to $p(\omega_i | \mathbf{d})$ as $m \rightarrow \infty$. So far as the changepoint problem is concerned, four simulations are performed per iteration corresponding to the four individual parameters λ, μ_1, μ_2, m . The Metropolis algorithm, which constructs a Markov chain whose equilibrium distribution is the one from which samples are intended to be drawn, is employed to generate samples from four conditional distributions. As the constructed chain is typically driven by a random walk process, its rate of convergence tends to be slow. The resultant histogram, to a striking extent, matches almost identically with the PDF profile resulting from the analytical approach. The amount of computing power required for the Monte Carlo algorithms to attain such a degree of accuracy, however exceeds that for the analytical approach by at least a factor of 50.

The data shown in Fig. 4 is noisier than that in Fig. 1. Whereas the signal element remains the same in both

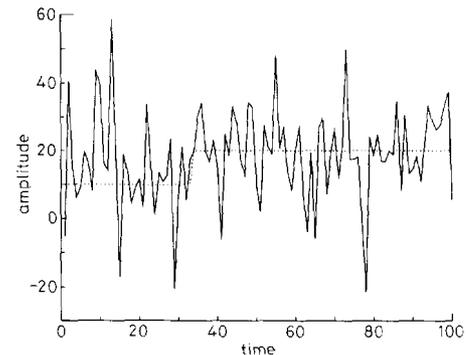


Fig. 4 Single changepoint data 2

cases, the noise distribution in Fig. 4 has become more Leptokurtic and disagrees more with a Gaussian assumption than that in Fig. 1. With reference to Fig. 5, it can be seen that the Gaussian approximation assigns maximum weights to the two end points, implicitly conveying the message of the absence of changepoints in the given data record. The analytical approach, despite failing to locate

the exact changepoint position, nonetheless still assigns considerable weight to the true changepoint. And its PDF profile matches well, in terms of both the multiple

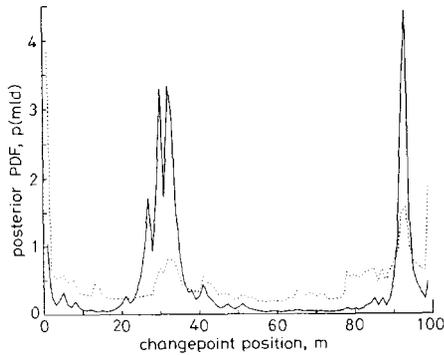


Fig. 5 Data 2: posterior density for single changepoint using analytical Laplacian approach compared with that using Gaussian assumption
 — analytical Laplacian
 Gaussian assumption

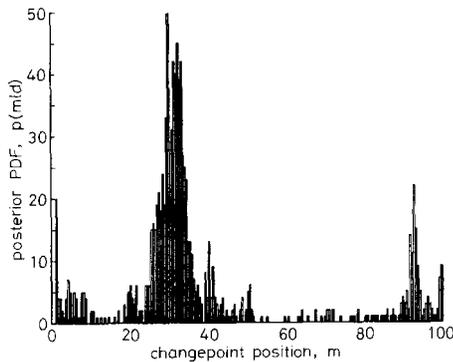


Fig. 6 Data 2: histogram plot of posterior density using Markov chain Monte Carlo methods

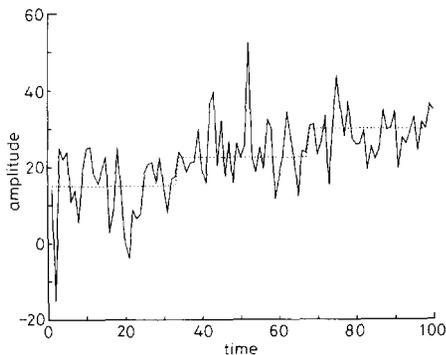


Fig. 7 Double changepoint data

modes' locations and curvatures, with the histogram obtained using the Monte Carlo method in Fig. 6. The Markov chain Monte Carlo method outperforms the other two in this extremely noisy case, albeit at the expense of computing power. The performance advantage achieved by it, however, is believed attributable to nothing fundamental other than the favourable randomness occasionally resulted from the stochastic relaxation type of algorithm. Given a sufficiently long period of

time, the histogram is expected to ultimately converge to the same result as is given by analytical approach.

Double-changepoint data is plotted in Fig. 7. The corresponding 3D posterior densities using a Gaussian

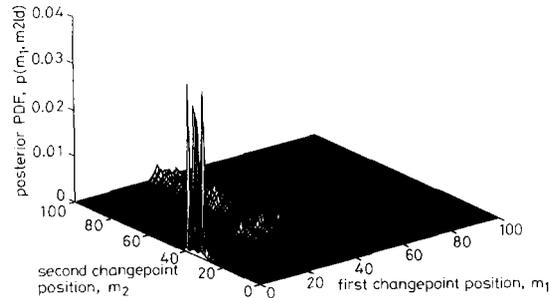


Fig. 8 Posterior density using Gaussian assumption

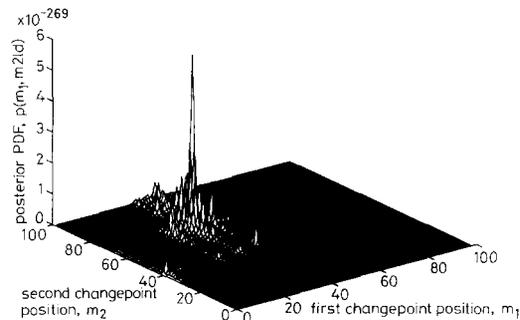


Fig. 9 Posterior density using analytical Laplacian approach

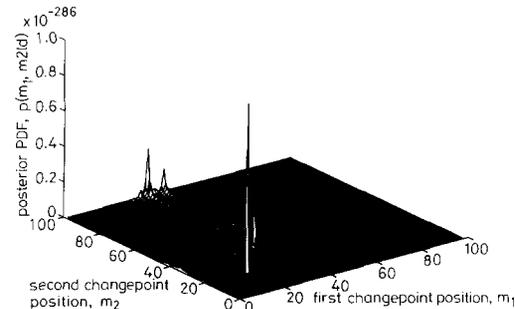


Fig. 10 Data 1: posterior density under assumption of two changepoints

approximation and the analytical Laplacian approach are shown in Figs. 8 and 9, respectively, where once again the robustness of the analytical method against noise impulses at the endpoints is evidently demonstrated.

The formulas derived in the last Section to compute the Bayesian model evidence has been applied to the single-changepoint data shown in Fig. 1. Here it is supposed that there are just two possible models which fit the data, model A having a single changepoint and model B having two changepoints. Differentiation between these two cases is made by comparing their associated model evidences. Fig. 10 plots the posterior density under the assumption of two changepoints.

Assume that all the signal-level parameters are uniformly distributed in $[-1000, 1000]$. Their normalised

prior density can then be expressed as

$$p(\mu_i) = \frac{u(\mu_i + 1000) - u(\mu_i - 1000)}{2000} \quad (41)$$

The ratio of the two evidences is calculated to be

$$\begin{aligned} \frac{E_B}{E_A} &= \frac{\sum_{m_1} \sum_{m_2} \frac{p(m_1, m_2 | \mathbf{d})}{99.98}}{\sum_m \frac{p(m | \mathbf{d})}{99}} \\ &= 0.0103 \end{aligned} \quad (42)$$

One can then deduce from the above calculations that model A is more plausible, which is in accordance with the fact.

5 Conclusion

This paper addresses the issue of changepoint detection in an impulsive noisy environment, where the presence of a large number of outliers renders the Gaussian assumption for the noise distribution inadequate of providing robust estimates. Markov chain Monte Carlo methods, such as the Metropolis algorithm and Gibbs sampler, are often employed in those cases to accommodate the mathematical intractability posed by the use of non-Gaussian statistical descriptions of the additive noise. Nevertheless, due to the inherent random-walk nature of those methods, tremendous computing power is required and moreover convergence is difficult to monitor in the discrete changepoint parameter space. Sampling from nonGaussian densities and parameter initialisations are also issues that are difficult to tackle.

To find an accurate and yet relatively parsimonious solution, an analytical approach to detecting signal changepoints in the presence of additive Laplacian noise has been proposed. It is demonstrated that at least in the context of DC step changes, by rearranging data samples with regard to their magnitude, purely analytical expressions for posterior density of changepoint locations and Bayesian model evidences are obtainable. This method can be extended to other symmetrical exponential distributions whose mathematical expressions involve absolute value signs and whose departure from normality is not as great as the Laplacian. Simulation results have displayed superior performance of this method over conventional methods, such as one which assumes that the noise distribution is Gaussian.

6 References

- 1 BOX, G.E.P., and TIAO, G.C.: 'Bayesian inference in statistical analysis' (Addison-Wesley, Reading, MA, 1973)
- 2 JEFFREYS, H.: 'Theory of probability' (Oxford University Press, 1961, 3rd edn.)
- 3 NEAL, R.: 'Probabilistic inference using Markov chain Monte Carlo method', Technical report CRG-TR-93-1, Department of Computer Science, University of Toronto, 1993
- 4 O'RUANAIDH, J.J.K., and FITZGERALD, W.J.: 'The identification of discontinuities in time series using the Bayesian piecewise linear model'. Technical report CUED/F-INFENG/TR 134, Department of Engineering, University of Cambridge, 1993
- 5 TANNER, M.A.: 'Tools for statistical inference: methods for the exploration of posterior distributions and likelihood functions' (Springer-Verlag, New York, 1993)
- 6 BERNARDO, J.M., and SMITH, A.F.M.: 'Bayesian theory' (Wiley, Chichester, 1994)
- 7 MARKS, R.J., WISE, G.L., HALDEMAN, D.G., and WHITED, J.L.: 'Detection in Laplace noise', *IEEE Trans.*, 1978, AES-6, pp. 866-871