
ELSEVIER

Computational Statistics & Data Analysis 37 (2001) 323–341

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Fitting multiple change-point models to data ☆

Douglas M. Hawkins

School of Statistics, University of Minnesota, 313 Ford Hall, 224 Church Street NE, Minneapolis, MN 55455-0493, USA

Received 1 November 1999; received in revised form 1 November 2000; accepted 1 November 2000

Abstract

Change-point problems arise when different subsequences of a data series follow different statistical distributions – commonly of the same functional form but having different parameters. This paper develops an exact approach for finding maximum likelihood estimates of the change points and within-segment parameters when the functional form is within the general exponential family. The algorithm, a dynamic program, has execution time only linear in the number of segments and quadratic in the number of potential change points. The details are worked out for the normal, gamma, Poisson and binomial distributions. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Segmented regressions; Quality improvement; Regression trees; Time series

1. Introduction

The change-point model is appropriate for some data sets with a natural ordering. This model is that the sequence of data can be broken down into segments with the observations following the same statistical model within each segment, but different models in different segments. One example of a change-point model is that in which the data follow a common distributional form (for example normal) whose parameters

E-mail address: dong@stat.umn.edu (D.M. Hawkins).

☆ Work supported by the National Science Foundation under grant DMS 9803622.

0167-9473/01/$ - see front matter © 2001 Elsevier Science B.V. All rights reserved.
PII: S 0167-9473(00)00068-2

(mean, variance or both) change from one segment to another. Another more complex model is the discontinuous segmented regression model in which the observations in each segment follow a linear regression, but the parameter(s) of this regression (slopes and/or intercept) change from one segment to the next.

Change-point models involve three issues – the choice of suitable parametric forms for the within-segment models; the choice of segment boundaries, or change-points, and the determination of the appropriate number of change-points to use in modeling the specific data set. Our discussion focuses on the second of these questions. The third question is outside the scope of this article but will be commented upon.

The best-known application of change-point modeling in data analysis is that of regression trees. In the most widely used implementation (Breiman et al., 1984), the data set is ordered by a continuous or ordinal predictor and then split into two subsequences – those cases whose predictor value falls below some change-point and those whose predictor value is above the change-point. The change-point is chosen to maximize the separation between the two subsequences. The same binary splitting algorithm is then applied to each of the subsequences, and repeated recursively until the subsequences can no longer be usefully subdivided. This is a "greedy" algorithm – it seeks to select each change-point to maximize an immediate return. As is generally the case with greedy algorithms (and as we shall see later by example) this hierarchic binary splitting, though fast, usually fails to give the optimum splits if there are two or more of them.

In this paper, we provide an exact and reasonably fast algorithm for performing a multiway split. We will do this, not only for the case of a normal mean (as used in regression trees) but for an arbitrary parameter in an exponential family model.

In the following sections, we will derive the likelihood equations for optimal multiway splitting of data following an exponential-family distribution. Showing that this satisfies Bellman's 'Principle of Optimality' it then follows that the optimal splits can be found with a dynamic programming algorithm. Finally, we will work out the details for a number of common data modeling distributions and illustrate them with actual data sets.

1.1. The exponential family

The exponential family provides a rich set of models for data. Familiar members of the family are the normal distribution, the exponential, the gamma, the binomial and the Poisson. The family also includes normal-error linear regression and some generalized linear models. Starting with the simpler (non-regression) models, the canonical form of the exponential family distribution or density function is

$$f(\boldsymbol{x},\boldsymbol{\theta}) = \exp[-\boldsymbol{\theta}'\boldsymbol{x} + c(\boldsymbol{x}) + d(\boldsymbol{\theta})]. \tag{1}$$

The parameter $\boldsymbol{\theta}$ and data $\boldsymbol{X}$ may be either scalar or vector-valued. If vectors, they must be of the same dimension. Given a random sample of size n, $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$,

all mutually independent, the sufficient statistic for $\boldsymbol{\theta}$ is

$$\boldsymbol{S} = \sum_{i=1}^{n} \boldsymbol{X}_i. \tag{2}$$

This statistic is the maximum likelihood estimator (MLE) of the parametric function $nd'(\boldsymbol{\theta})$, for which it is unbiased. Solving the equation $d'(\hat{\boldsymbol{\theta}}) = \boldsymbol{S}/n$ gives the MLE of $\boldsymbol{\theta}$. Substituting this back into the likelihood gives the maximized likelihood.

2. The change-point model

Now extend the formulation to the change-point model. In this model, there are a number of change points, $\tau_1, \tau_2, \ldots, \tau_{k-1}$ such that the observations $\boldsymbol{X}_i$ with $\tau_{j-1} < i \leq \tau_j$ follow the particular exponential family model with parameter $\boldsymbol{\theta}_j$. In other words, the distributional form remains the same for all segments, but the parameter changes whenever one crosses over one of the change points τ_j.

As there are $k-1$ change-points, there are a total of k segments in this model. To simplify notation, we will augment these change-points with $\tau_0 = 0$, a notional changepoint to the left of the entire sequence, and $\tau_k = n$, a notional changepoint to the right of the sequence. The log likelihood of the data is then given by

$$L(\boldsymbol{X}, \boldsymbol{\theta}, \tau) = \sum_{j=1}^{k} \sum_{i=\tau_{j-1}+1}^{\tau_j} [-\boldsymbol{\theta}_j' \boldsymbol{X}_i + c(\boldsymbol{X}_i) + d(\boldsymbol{\theta}_j)]. \tag{3}$$

2.1. Maximum likelihood estimation of the parameters

The maximum likelihood estimators (MLEs) of the parameters $\tau_j, \boldsymbol{\theta}_j$ are found by maximizing the log likelihood (3). This can logically be separated into two stages; given the changepoints τ_j, the $\boldsymbol{\theta}_j$ maximize the within-segment likelihood. Thus the maximization consists of the "outer" problem of finding the MLE's $\hat{\tau}_j$ of the τ_j with the "inner" problem of finding the $\hat{\boldsymbol{\theta}}_j$ given the $\hat{\tau}_j$.

For arbitrary $0 < h < m \leq n$, write $\boldsymbol{S}(h,m) = \sum_{i=h+1}^{m} \boldsymbol{X}_i$. If the sequence $\boldsymbol{X}_{h+1}, \ldots, \boldsymbol{X}_m$ follows the model distribution with a parameter value $\boldsymbol{\theta}$, $\boldsymbol{S}(h,m)$ is the sufficient statistic for $\boldsymbol{\theta}$ using $\boldsymbol{X}_{h+1}, \ldots, \boldsymbol{X}_m$. Write $Q(h,m)$ for -2 times the maximized log likelihood obtained by substituting $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ in the log likelihood of this subsequence of the data. This will be

$$Q(h,m) = -2[\hat{\boldsymbol{\theta}}' \boldsymbol{S}(h,m) - (m-h)d(\hat{\boldsymbol{\theta}})] - 2 \sum_{i=h+1}^{m} c(\boldsymbol{X}_i). \tag{4}$$

When looking at the likelihood of the whole sequence, the $\sum c(\boldsymbol{X}_i)$ term is a constant, and can be dropped from the maximization, leaving us with the "stage return"

$$Q(h,m) = -2[\hat{\boldsymbol{\theta}}' \boldsymbol{S}(h,m) - (m-h)d(\hat{\boldsymbol{\theta}})]. \tag{5}$$

Omitting the $c(\boldsymbol{X}_i)$ terms, the overall maximized likelihood of the data set can then be written

$$-2\max_{\{\tau_j\},\{\boldsymbol{\theta}_j\}} L(\boldsymbol{X},\boldsymbol{\theta},\tau) = -2\max_{\{\tau_j\}}\left[\max_{\{\boldsymbol{\theta}_j\}} L(\boldsymbol{X},\boldsymbol{\theta},\tau)\right]$$
$$= -2\max_{\{\tau_j\}}\sum_{m=1}^{k} Q(\tau_{m-1},\tau_m).$$

Notice an important property of this likelihood – it is separable. The optimum for splitting cases $1,\ldots,n$ into k segments consists conceptually of first finding the rightmost changepoint $\hat{\tau}_k$. Once this is done, the remaining changepoints are found from the fact that they constitute the optimum for splitting cases $1,\ldots,\hat{\tau}_k$ into $k-1$ segments. This separability is Bellman's "principle of optimality" (see, for example, Bellman and Dreyfus, 1962). Because of it, the likelihood may be maximized by dynamic programming (DP).

Theorem. *Write $F(r,m)$ for -2 times the maximized log likelihood resulting from fitting an r-segment model to the sequence $\boldsymbol{X}_1,\boldsymbol{X}_2,\ldots,\boldsymbol{X}_m$ (omitting constants and the $c(\boldsymbol{X})$ terms). Then, $F(r,m)$ satisfies the recursion*

$$F(1,m) = Q(0,m), \tag{6}$$

$$F(r,m) = \min_{0<h<m} F(r-1,h) + Q(h,m). \tag{7}$$

Proof. The result follows immediately by contradiction. □

The dynamic programming algorithm: The DP algorithm follows these equations exactly. For each $m=1,\ldots,n$, calculate $F(1,m)=Q(0,m)$. Then for each $r=2,\ldots,k$, calculate $F(r,m), m=1,\ldots,n$. We calculate each table entry $F(r,m)$, using (7) to find the h value minimizing the sum of the previously calculated $F(r-1,h)$ and $Q(h,m)$. Along with each value $F(r,m)$, we keep a record of $H(r,m)$, the h value yielding the minimum.

Once these calculations have been done, the maximized log likelihood of the k segment model fitted to the full data set is $-\frac{1}{2}F(k,n)$.

The estimates $\hat{\tau}_j$ are given by the DP back-tracing operation $\hat{\tau}_k = n$, and for $r = k-1,k-2,\ldots,1$, $\hat{\tau}_r = H(r+1,\hat{\tau}_{r+1})$.

Finally, the within-segment $\hat{\boldsymbol{\theta}}_j$ are found as the values defined by the $Q(\hat{\tau}_{r-1},\hat{\tau}_r)$.

2.2. Computational complexity of the algorithm

At first glance, fitting the change-point problem appears to be very hard for k, the number of changepoints greater than 1 (see, for example, the comments by Chen and Gupta (1997)). The DP formulation though has a computational complexity just *linear* in k. To see this, we sketch a computation count. First, we can build up a single once-and-for-all table of the $Q(h,m), 0<h<m\le n$. Since there are $n(n-1)/2$ of these, this is an $O(n^2)$ calculation, which yields the $F(1,m)$ *en passant.*

Then, for each subsequent r value, we need to compute the values of $F(r,m), m = 1,\ldots,n$. As finding $F(r,m)$ involves a search over $m-1$ values of $F(r-1,h) + Q(h,m)$, this is an $O(m)$ calculation, so the total computation of $F(r,m)$, $m=1,\ldots,n$ is an $O(\sum_{m=1}^{n} m)=O(n^2)$ calculation. Since this needs to be done for each $r=2,\ldots,k$, the total computational complexity is $O(kn^2)$.

Note that the core computation that defines the computational complexity is very simple, requiring some indexing, one add, and a compare. Contrary therefore to the initial impression that the computation increases dramatically with the highly nonlinear parameters τ_r, computation actually increases only linearly, and with quite a small multiplier. Fitting say 10 segments takes a little less than twice as long as fitting 5. This means that fitting large numbers of changepoints is not a computational concern.

The quadratic complexity in n does mean that optimal segmentation of very long sequences may be computationally burdensome. It is possible to reduce the computational load for large samples substantially at the cost of introducing some approximation by evaluating the $F(r,m)$ not for all m, but for a grid of m values. Correspondingly, the search is only over m values on the grid. If, for example, we restricted the grid (and the possible changepoints) to be every fifth point of the data sequence, this would cut the computation down by a factor of nearly 25. This reduction occurs naturally in problems where there are blocks of data whose order is tied. In this case, it does not make sense to break a block across two segments, and the only natural positions for the changepoints are between blocks. Under these circumstances, the computation is reduced without introducing any approximation.

Note that in the process of finding the k segment fit to the full series of data, we obtain as a free by-product the optimal changepoints for all subsequences $1,\ldots,m$ for $m \leq n$, and all $r \leq k$. This greatly facilitates exploring the fitting of different numbers of segments – simply carry out the DP for the maximum number of segments in the range of interest, and all optimal segmentations using fewer changepoints are available for the miniscule effort of backtracing them with the $H(r,m)$ table.

The formulation also makes it easy to explore fitting different starting subsequences.

This dynamic programming algorithm is a development of Hawkins (1972), where a similar formulation is used for the numeric analysis problem of producing piecewise approximations of functions.

3. Particular applications

Changepoint in normal mean: We will start with the familiar example of scalar normal data with constant variance, where the mean may change from one segment to the next. This problem and the DP solution are discussed in more detail in Hawkins and Merriam (1973, 1975). As this is the problem addressed by regression trees (Breiman et al., 1984), it is particularly interesting to compare their implementation with exact optimization.

Turning the normal density into canonical exponential family form gives

$$f(x,\theta,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\, e^{-(x-\theta)^2/(2\sigma^2)}$$
$$= \exp\left[\frac{x\theta}{\sigma^2} - \frac{x^2}{2\sigma^2} - \frac{\theta^2}{2\sigma^2} - \log(\sqrt{2\pi}\sigma)\right].$$

The sufficient statistic for the mean θ is the mean of the data. Write $\bar{X}_{h,m} = S(h,m)/(m-h)$ for the mean of the subsequence $X_{h+1},\ldots,X_m$, Then, the maximized log likelihood for this subsequence is

$$Q(h,m) = \frac{\sum_{h+1}^{m} X_i^2}{\sigma^2} - \frac{(m-h)\bar{X}_{h,m}^2}{\sigma^2} + 2(m-h)\log(\sqrt{2\pi}\sigma)$$
$$= \sum_{h+1}^{m} (X_i - \bar{X}_{h,m})^2/\sigma^2 + 2(m-h)\log(\sqrt{2\pi}\sigma).$$

This is the within-segment sum of squared deviations from the mean, divided by the nuisance constant σ^2, plus the nuisance constant $2(m-h)\log(\sqrt{2\pi}\sigma)$. Note that, when considering the log likelihood of the entire sequence, the $2(m-h)\log(\sqrt{2\pi}\sigma)$ terms will sum to the constant $2n\log(\sqrt{2\pi}\sigma)$, regardless of the within-segment parameters, and so this term can be ignored. Similarly, the nuisance divisor σ^2, being constant for the whole data set, can be omitted.

The change-point problem for normal means therefore comes down to a one-way analysis of variance with the change points chosen to minimize the pooled within segment sum of squared deviations, thereby maximizing the between-segment sum of squared deviations. Writing $W(h,m)$ for the sum of squared deviations of the observations $X_{h+1}, X_{h+2},\ldots,X_m$ from their mean

$$W(h,m) = \sum_{i=h+1}^{m} (X_i - \bar{X}_{h,m})^2,$$

we can find the optimal changepoints from the DP

$$F(1,m) = W(1,m),$$

$$F(r,m) = \min_{0<h<m} [F(r-1,h) + Q(h,m)].$$

An algorithm mathematically equivalent to this was proposed independently by Venter and Steel (1996). Rather than minimize the within-segment sum of squares, they maximize the between-segment sum of squares. In computational terms, the resulting algorithm is somewhat faster than ours since the update for calculating our $Q(h,m)$ requires 6 floating-point operations while theirs requires only 4.

The Breiman et al. regression tree approach uses successive hierarchic binary splits. The best single changepoint in the sequence is defined by

$$F(2,n) = \min_{0<h<n} [W(0,h) + W(h,n)].$$

This leads to the same $\hat{\tau}_1$ as the exact DP for this two-segment solution. Subsequently, the hierarchic binary approach fixes this changepoint and then applies a

binary search to the left half $1,\ldots,\hat{\tau}_1$ and the right half $\hat{\tau}_1+1,\ldots,n$, adding the splitpoint of whichever of these two subsequences gives the smaller pooled residual sum of squares. The method continues in this way by attempting to split each of the subsequences it uncovers.

The binary hierarchic method is very fast – it is only linear in n. It suffers though from the fact that the true optimal changepoints are not necessarily hierarchic, and so it does not in general yield the correct optimum. The approach therefore is a tradeoff – a much faster computation that yields only an approximate solution to the MLE. Since the exact computation using the DP is itself quite fast except for huge data sets, we suggest this is a poor trade.

Changepoint in a gamma sequence: Consider next the gamma distribution with known shape parameter α.

$$f(x,\beta,\alpha)=\frac{x^{\alpha-1}\mathrm{e}^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}}.$$

If we fit this model to the sequence $X_{h+1},\ldots,X_m$, the MLE of β is

$$\hat{\beta}=S(h,m)/(m-h)=\bar{X}_{h,m}$$

and the maximized log likelihood gives

$$\begin{aligned}Q(h,m)&=2(m-h)[1+\log\Gamma(\alpha)-\alpha\log\bar{X}_{h,m}]-2(\alpha-1)\sum_{h+1}^{m}\log X_i\\&=2(m-h)\alpha\log\bar{X}_{h,m}-2(m-h)[1+\log\Gamma(\alpha)]+2(\alpha-1)\sum_{h+1}^{m}\log X_i.\end{aligned}$$

When finding the likelihood of the entire sequence, all but the first term in this expression will sum to a constant, and so may be ignored. For optimization purposes therefore it is sufficient to define the stage return

$$Q(h,m)=2(m-h)\alpha\log\bar{X}_{h,m}.$$

Normal data with fixed known mean and changing variance: If

$$Y_i\sim N(\mu,\beta_j),\quad \tau_{j-1}<i\le\tau_j,$$

where μ is known, and we define $X_i=(Y_i-\mu)^2$, then the sequence X_i follows this gamma change-point model with $\alpha=0.5$. The gamma formulation therefore also addresses the case of normal data with a constant known mean but a variance that changes between segments.

Normal data with change in mean and/or variance: Consider next the model

$$X_i\sim N(\mu_j,\sigma_j^2),\quad \tau_{j-1}<i\le\tau_j$$

The log likelihood of the sequence $X_{h+1},\ldots,X_m$ for a generic μ,σ is

$$\sum_{i=h+1}^{m}\left[-\frac{(X_i-\mu)^2}{2\sigma^2}-\log(\sqrt{2\pi}\sigma)\right].$$

This is maximized by the estimators $\hat{\mu}=\bar{X}_{h,m}$, $\hat{\sigma}^2=W(h,m)/(m-h)$, and the maximized likelihood gives

$$Q(h,m)=(m-h)[1+\log\{W(h,m)/(m-h)\}].$$

We may redefine $Q(h,m) = (m-h)\log[W(h,m)/(m-h)]$, since the omitted term sums to the constant n over the entire sample. A further modest refinement is a "degrees of freedom" correction to $\hat{\sigma}^2$, replacing the divisor $(m-h)$ by $(m-h-1)$ and leading to the expression

$$Q(h,m) = (m-h)\log[W(h,m)/(m-h-1)].$$

This model is an example of a vector-valued parameter $\boldsymbol{\theta} = (\mu, \sigma)$ with its vector-valued sufficient statistic $\sum X_i, \sum X_i^2$.

The Poisson distribution: The formulation works equally well with discrete members of the exponential family. For the Poisson distribution

$$f(x,\lambda) = \frac{\lambda^x e^{-\lambda}}{x!},$$

the MLE of λ over the data sequence $X_{h+1} \ldots X_m$ is $\bar{X}_{h,m}$, and the maximized log likelihood gives

$$Q(h,m) = -2(m-h)\bar{X}_{h,m}[\log \bar{X}_{h,m} - 1] - \sum_{h+1}^{m} \log X_i!$$

As the last term sums to a constant, there is no loss in defining

$$Q(h,m) = -2(m-h)\bar{X}_{h,m}[\log \bar{X}_{h,m} - 1].$$

The binomial distribution: For this discrete distribution,

$$f(x,\psi) = \binom{\alpha}{x} \psi^x (1-\psi)^{\alpha - x},$$

where α is the number of trials and is assumed known, and the probability of success is ψ.

For the sequence $X_{h+1}, \ldots, X_m$, the MLE of ψ is $\bar{X}_{h,m}/\alpha$, and substituting this in the expression for the log likelihood gives

$$Q(h,m) = -2\left[\sum_{h+1}^{m} \binom{\alpha}{X_i} + X_i \log \frac{\bar{X}_{h,m}}{\alpha} + (\alpha - X_i)\log\left(1 - \frac{\bar{X}_{h,m}}{\alpha}\right)\right].$$

As the first term is a constant when summed over the data, it may be omitted from the definition of $Q(h,m)$. The α divisor can also be simplified out, leading to the form

$$Q(h,m) = -2(m-h)[\bar{X}_{h,m} \log \bar{X}_{h,m} + (\alpha - \bar{X}_{h,m})\log(\alpha - \bar{X}_{h,m})].$$

Halpern (2000) gave a dynamic programming algorithm for the optimal segmentation of Bernoulli sequences. This algorithm however was based, not on likelihood maximization, but on maximizing the differences in proportions of "successes" between adjacent segments. This leads to the Venter–Steel algorithm.

3.1. Hierarchic binary splitting

We commented on the hierarchic binary segmentation in the context of regression trees. The other members of the exponential family can also be analyzed using the same approach. Chen and Gupta (1997), for example, use it for gamma sequences.

A modest refinement of hierarchic binary splitting is sketched in Hawkins (1976). This consists of adding a merging step in which similar subsequences are re-merged. While this does improve on the basic hierarchic splitting approach, it still does not produce the optimal changepoints reliably.

Other applications of the dynamic programming formulation include discontinuous piecewise regression models. These extend the "normal mean" formulation by allowing the segment mean $\boldsymbol{\theta}$ to be a regression on a set of covariates (Hawkins, 1976). This problem can be solved with a computational complexity of $O(n^2 p^2) + O(kn^2)$. The corresponding case of continuous piecewise regression models is set up as a DP in Bellman and Roth (1969). Some multivariate segmentation problems (Hawkins and Ten Krooden, 1979) can also be solved with the DP.

4. Formal testing for the number of segments

$F(k,n)$ is the negative doubled maximized likelihood of the model fitting k segments to the full sequence of data. It therefore gives rise to generalized likelihood ratio tests:

To test the null hypothesis of a single segment versus the alternative of k segments, the GLR statistic is $F(1,n) - F(k,n)$.

To test the null hypothesis of at most $(k-1)$ segments against the alternative of k, the GLR statistic is $F(k-1,n) - F(k,n)$.

On the face of it, the incremental change $F(k-1,n) - F(k,n)$ should follow an asymptotic null chi-squared distribution with degrees of freedom $1 + dim(\boldsymbol{\theta})$ given by parameter counting, and similarly for $F(1,n) - F(k,n)$. However, despite being GLR test statistics, neither of these quantities follows an asymptotic chi-squared distribution. This is a consequence of the failure of the Cramér regularity conditions. Little is known about the asymptotics of the tests in general – the only situation that has been well studied is that of normal data with constant variance in the test of whether a two segment model fits better than a single segment. For this problem, if the nuisance parameter σ is known, then the generalized likelihood ratio test for the null hypothesis of a single segment against the alternative that there are two is

$$[F(1,n) - F(2,n)]/\sigma^2.$$

If the nuisance parameter σ is unknown, then the generalized likelihood ratio test statistic (with a degrees of freedom correction) is

$$(n-2)[F(1,n) - F(2,n)]/F(2,n).$$

Far from having an asymptotic distribution as n increases, these statistics increase without bound (see, for example, Hinkley (1970), Hawkins (1979), Irvine (1982), Yao (1987) and the review by Bhattacharya (1994)). The failure of conventional asymptotics in even this easiest case is an indication of the technical difficulty of the more general situation.

In the absence of proper inferential bases for tests on the number of segments, there are some intuitive methods that may be useful. One is the "scree" test of

plotting $F(k-1,n)-F(k,n)$ against k and looking for an "elbow" in the plot. The rationale for this test is that, while real segment boundaries are being fitted there will be large reductions in $F(k,n)$, but once all real segment boundaries have been found there will be a more-or-less-constant decrease in $F(k,n)$ from fitting additional segments.

This heuristic does not rest on very solid foundations (in particular, it is not true that the null distribution of $F(k-1,n)-F(k,n)$ varies smoothly with k), but may suffice as a rough and ready practical tool. See also Venter and Steel (1996) and Halpern (2000).

5. Examples

5.1. A regression-tree-type example

We start with a data set showing a non-linear relationship between a predictor and a dependent variable. In the absence of a parametric model, this data set might be subjected to analysis with a regression tree. The data set is shown as Fig. 1a, and the optimal segmentation into $2,3,\ldots,6$ segments is Fig. 1b. Table 1 shows the optimal segment boundaries, the pooled residual sum of squares $F(r,n)$ and the change in residual sum of squares as we go from one value of r to the next. Note that the optimal changepoints are not hierarchic. In particular, the first changepoint found, 136, does not re-appear in any of the subsequent fits.

The 5-segment optimal fit seems to capture most of the structure in the data set, with a residual sum of squares of 137.6. The 5-segment fit returned by the hierarchic binary segmentation approach has a much larger residual sum of squares, 175.

The binary hierarchic approach therefore does not do very well here. It does almost catch up with the optimal fit at 10 segments, by which time the latter is overfitting. This illustrates the problem with hierarchic binary splitting; while it will uncover the broad structure in the end, it may do so using some unnecessary changepoints. These violate parsimony and can cloud interpretation.

We could formulate the changing-mean model as a non-linear regression

$$X_i = \beta_0 + \sum_{r=1}^{k} \beta_r I[i > \tau(r)],$$

where I is the usual indicator function, and try fitting using general-purpose non-linear regression software. As it stands the lack of differentiability with respect to the τ makes this formulation unworkable. We can however "smooth the corners" by approximating the staircase model with

$$X_i = \beta_0 + \sum_{r=1}^{k} \beta_r H[i - \tau(r)],$$

where $H(\cdot)$ is any convenient cumulative distribution function centered on zero – for example, that of the Cauchy. This non-linear regression formulation can be made to

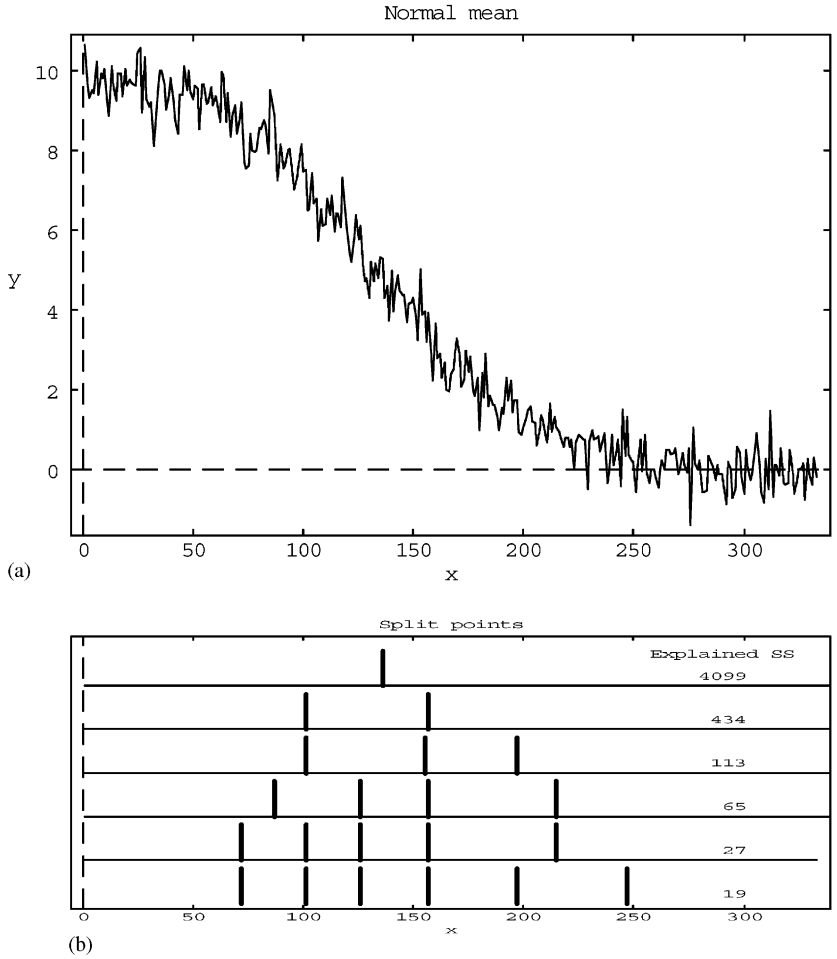


Fig. 1. (a) A general regression data set. (b) Optimal k-way splits on mean.

converge, but usually to poor local optima. This non-linear regression formulation is far inferior to even the hierarchic binary segmentation approach which at least is guaranteed to work for two segments even if for no more than two.

5.2. Stock market data

We now illustrate the analysis with three real data sets. The first data set (taken from Hsu, 1979) is the weekly log price relative (LPR) of the Dow Jones Industrial Average for the period July 1, 1971 to August 2, 1974. The series is plotted in Fig. 2a. A baseline model of this sequence would be that X_i, the LPR in week i,follows a $N(\mu, \sigma^2)$ distribution. The value of μ should be practically zero under modest "efficient market" assumptions. However, there can be doubts about the

Table 1
Changepoints for regression tree data

k	$F(k,n)$	Change	τ_1	τ_2	τ_3	τ_4	τ_5
2	749.22	4099.43	136				
3	315.77	433.45	101	157			
4	202.45	113.32	101	155	197		
5	137.64	64.81	87	126	157	215	
6	110.73	26.91	72	101	126	157	215

constancy of σ – visually the sequence seems to be more variable in the later periods than the earlier.

To check this possibility, we fitted the change-point model for a constant $\mu = 0$, but with σ changing from one segment to another. The $F(k,n)$ values for k values from 2 to 6 and the corresponding change points are given in Table 2a and plotted in Fig. 2b along with the changes $F(r-1,n)-F(r,n)$, the "explained deviance" due to going from $r-1$ to r segments.

Once again, we see that the optimal change points are not hierarchical. The best single changepoint, 89, is not chosen as one of the two best changepoints and this lack of hierarchy continues in the table. Visually, the change $[F(k-1,n)-F(k,n)]$ seems to stabilize around the value 5.5 (a pattern that is continued in fitting up to 10 segments), suggesting that the data set has no more than four "real" segments.

While a boundary at point 89 is visually reasonable from Fig. 2a, the boundaries at 28 and 31 seem to do no more than isolate a trio of LPR values that are unusually close to zero.

5.3. A binomial formulation

The log of the variance of short subsequences (as used in the likelihood function) may be thought to pay unnaturally close attention to fortuitously small LPR's in a few successive days. A different formulation that looks instead at the larger differences (and illustrates a different statistical distribution) may be interesting. We created a Bernoulli variable, set to 1 if the absolute value of the LPR is above the median absolute LPR of 0.0133, and 0 if it is below 0.0133. This transformation may be thought of as a non-parametric way to isolating subsequences of higher than average variability.

This random variable is binomially distributed with $\alpha=1$. Differing variances in the original data would translate into differing binomial probabilities in these 'yes/no' variables. Segmenting this sequence using the DP gives the split points shown in Fig. 2c, and tabulated in Table 2b.

The $F(k,n)$ values here more clearly suggests three segments, with boundaries at 89 and 133. Visually, this seems reasonable, with the segment to the right of 133 appearing to have fewer extreme LPR values than does the segment from 89 to 133.

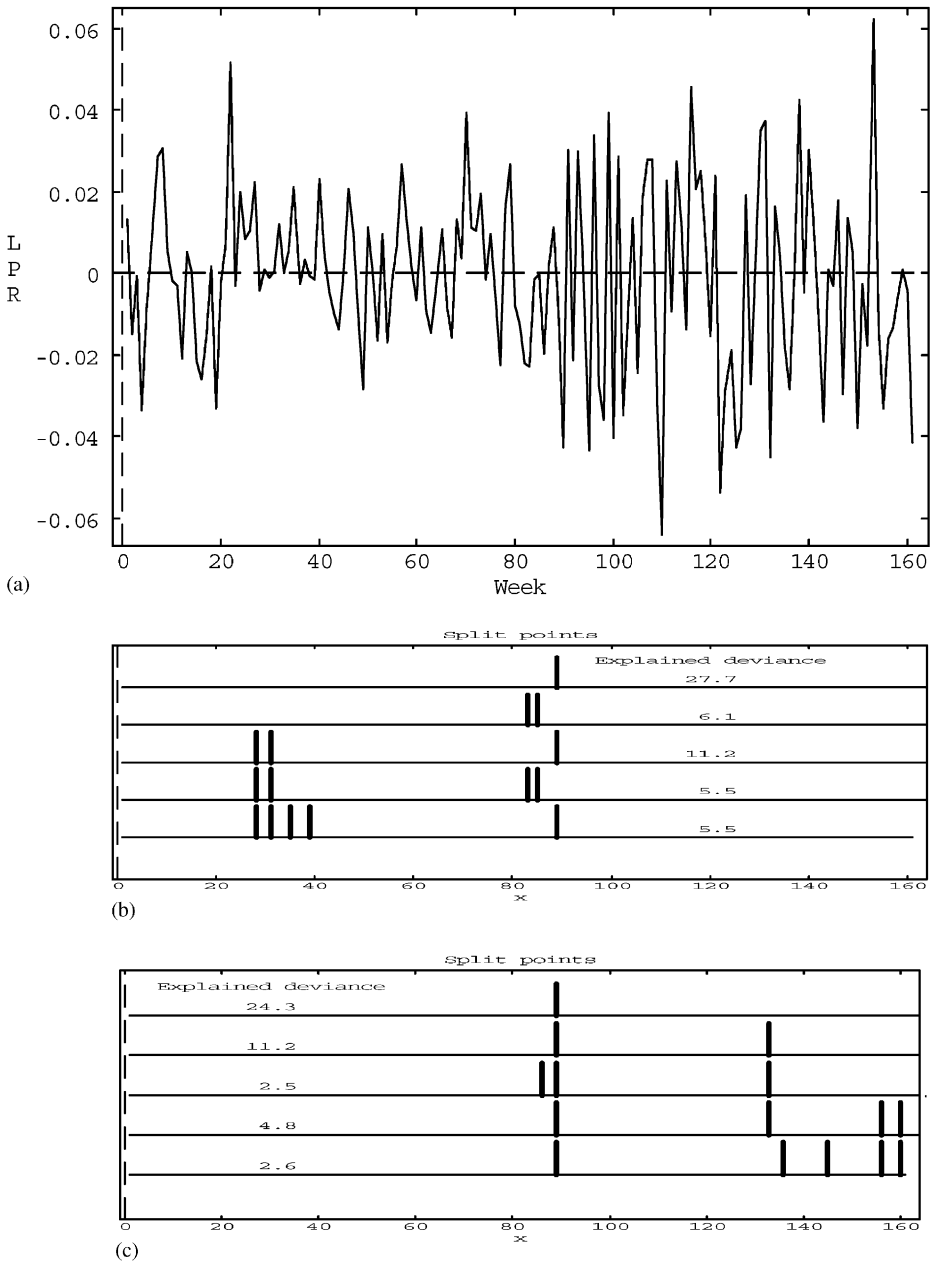


Fig. 2. (a) Log price relative of DJIA. (b) Cutpoints by variance. (c) Cutpoints by binomial splitting.

5.4. Aircraft arrivals

A second sequence by Hsu is of the times between arrivals of airplanes over Newark NJ for an 8-h period in April 1969. Under a Poisson process model, these

Table 2
Change points (a) for Dow Jones variance

k	$F(k,n)$	Change	τ_1	τ_2	τ_3	τ_4	τ_5
(a) for Dow Jones variance							
2	−1254.69	27.72	89				
3	−1260.83	6.14	83	85			
4	−1272.06	11.22	28	31	89		
5	−1277.53	5.48	28	31	83	85	
6	−1283.00	5.46	28	31	35	39	89
(b) for binomial Dow Jones scale measure							
2	198.84	24.30	89				
3	187.62	11.21	89	133			
4	185.10	2.51	86	89	133		
5	180.30	4.80	89	133	156	160	
6	177.66	2.63	89	136	145	156	160

inter-arrival times should follow an exponential distribution. Traffic intensities changing by time of day, however, would lead to shifts in the exponential parameter.

The data sequence is shown in Fig. 3a. No changepoints are visually obvious, though the values seem higher on the right than on the left. Modeling the sequence by the gamma distribution with $\alpha = 1$ (the exponential distribution) gives the fits plotted in Fig. 3b and tabulated in Table 3. The apparent oscillations in the size of the changes in $F(k,n)$ and their similar size suggest that there is not much structure in this sequence.

Following the approach of Worsley (1986), we find that the upper 5% point of the likelihood ratio statistic for a binary split is 4.43. This means that the initial binary split is not significant. The significance or otherwise of subsequent splits is not clear.

5.5. A Poisson formulation

Data modeled by a Poisson process can be tracked using either the exponential inter-arrival times, or the Poisson-distributed numbers of events within non-overlapping time windows. Poisson monitoring of the number of arrivals in successive time windows can be more effective than monitoring the inter-arrival times for the detection of sharp increases in mean inter-arrival time. The inter-arrival time data only produce an observation after there has been an arrival. If arrivals slow down or stop, there will be few or no inter-arrival data to which to react, but the Poisson monitor will produce a long stream of zero counts and detect the decrease in arrival rate.

To explore this alternative way of looking at the data, we transformed the data into the number of events in successive intervals of width 155 (the mean inter-arrival time) and analyzed this sequence of counts per standard interval using the Poisson

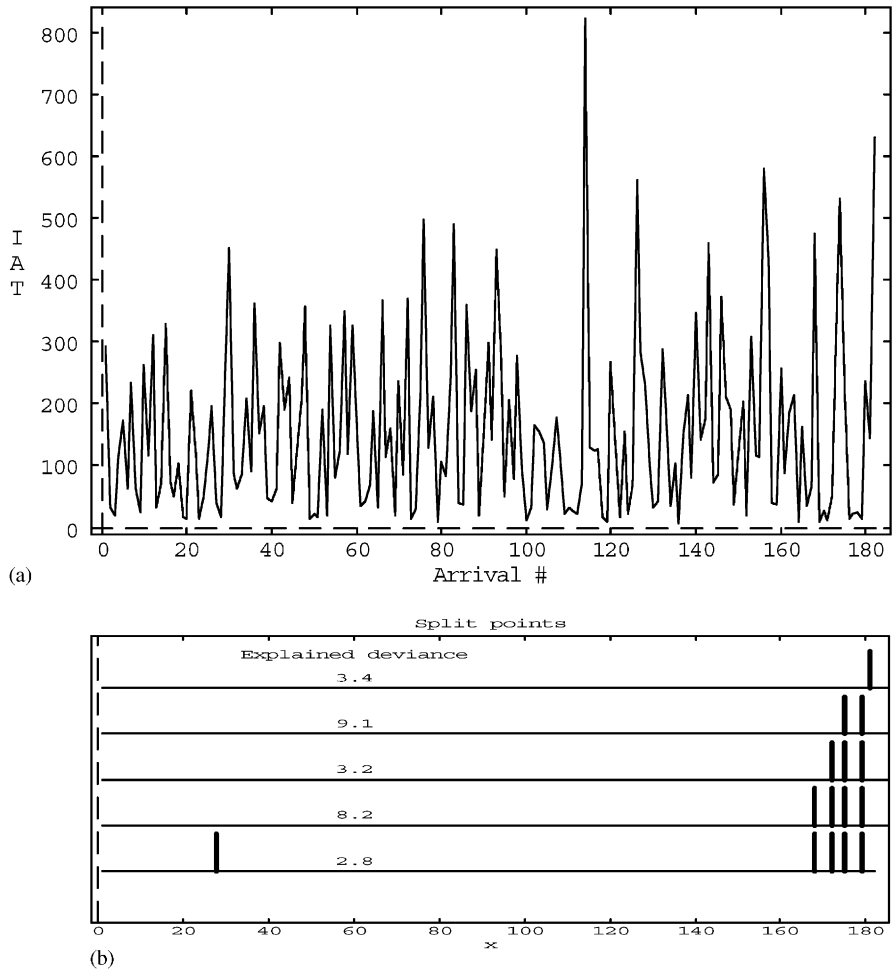


Fig. 3. (a) Aircraft inter-arrival times. (b) Cutpoints by gamma.

Table 3
Change points for aircraft inter-arrival times

k	$F(k,n)$	Change	τ_1	τ_2	τ_3	τ_4	τ_5
2	1833.33	3.36	181				
3	1824.25	9.09	175	179			
4	1821.09	3.16	172	175	179		
5	1812.94	8.15	168	172	175	179	
6	1810.15	2.79	28	168	172	175	179

model. Segmenting the sequence gives

k	$F(k,n)$	Change	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}_3$	$\hat{\tau}_4$	$\hat{\tau}_5$
2	5.95	5.95	3				
3	11.01	5.06	178	179			
4	16.97	5.94	3	178	179		
5	21.60	4.63	3	106	108	113	
6	27.97	6.37	3	108	113	178	179

This table is strikingly featureless. The explained deviance is steady across the whole range of values, and visually looks small. It appears that, whether we look at this data set in terms of the inter-arrival times or the number of events per fixed time interval, there is no segment structure in it and therefore no detectable non-constancy in the aircraft arrival process.

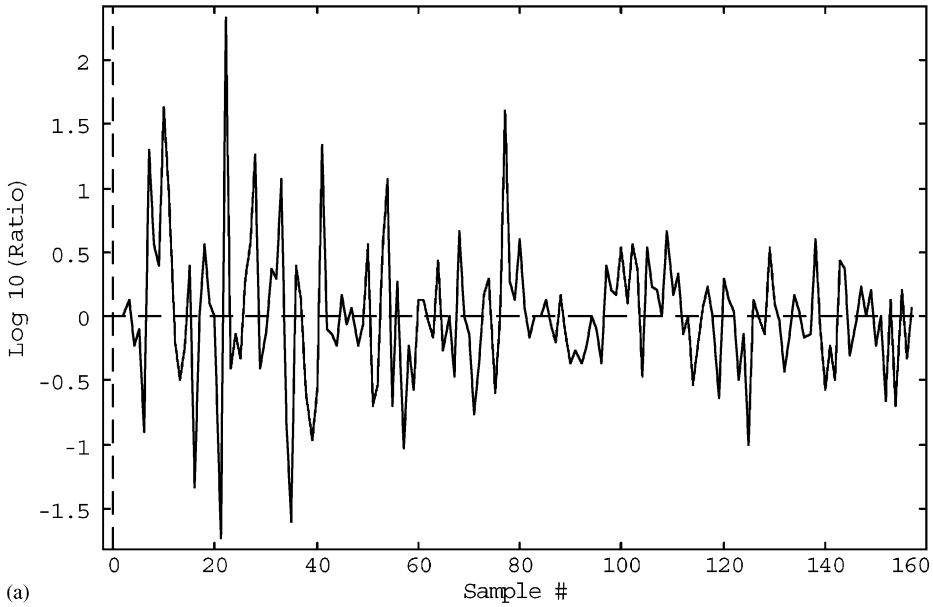
5.6. Gold mine sampling quality control

A third real data set is taken from mine quality control. Samplers in gold mines extract samples of the face at regular spacings and submit them for chemical assay for their gold content. As a quality check, supervisors cut out fresh samples at some of the locations already sampled. This gives rise to pairs of samples and of gold contents – one by the original sampler and one by the supervisor. The log of the ratio of these two gold contents has an approximately normal distribution. This distribution should have a zero mean (else the sampler is biased) and a small variance (else the sampler is erratic). Fig. 4a (taken from Rowland and Sichel, 1961) shows a sequence of such ratios for a junior sampler. It is of interest to see whether there have been changes in either mean or variance over the sequence, as either of these would have important implications for the valuation of the mine. We therefore apply the segmentation procedure for normal data shifting in mean, variance, or both.

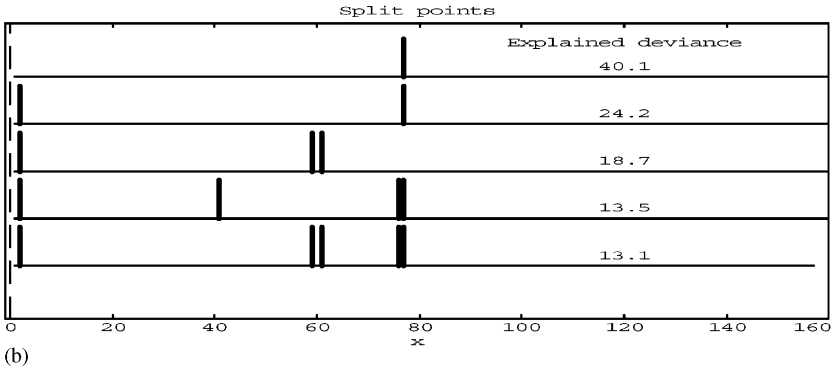
The split points are shown in Fig. 4b, and the numeric results in Table 4. The GLR test statistics for adding the second and subsequent changepoints have values around 15. Against this benchmark, the single change point at time 42 seems statistically as well as visually real. Looking at the summary statistics of the log ratio in the two segments we get

Segment $1-42$ mean $= 0.0698$, sd $= 0.8294$,
Segment $43-157$, mean $= -0.0177$, sd $= 0.3994$.

It appears that the difference is mainly in variance (initially poor but improving consistency) and not of a non-zero mean, so there is no cause for concern about bias. That the later variance is smaller than the earlier suggests some learning by the sampler, and consequent quality improvement. This halving of the standard deviation corresponds to a dramatic improvement in the quality of gold estimates over his early work.



(a)



(b)

Fig. 4. Gold mine quality control data. (b) Normal mean and/or variance.

Table 4
Change points for gold mine quality control data

k	$F(k,n)$	Change	τ_1	τ_2	τ_3	τ_4	τ_5
2	−217.55	39.77	42				
3	−231.71	14.16	75	78			
4	−249.03	17.32	42	75	78		
5	−263.71	14.67	42	75	78	79	
6	−277.88	14.16	37	38	75	78	79

6. Conclusion

The change-point model for the general exponential family can be thought of as a generalized non-linear model. As such it would seem to be computationally intensive in the number of non-linear parameters – the changepoints. On the contrary however, the model can be fitted in a time linear in the number of change-points using a dynamic programming formulation making it quite a small task with moderate size data sets.

We have discussed the single-parameter exponential family in some detail. The approach applies equally well though to discontinuous segmented regression models, and some multivariate models. It can also be applied to distributions outside the exponential family, but with the complication that the explicit estimates of within-segment parameters are replaced with much slower iterative calculations of maximum likelihood estimates.

A by-product of fitting k segments to n data points is the full set of optimal segmentations using k or fewer segments to points 1 through m for $1 < m < n$. We have not explored this use of the algorithm, but it has obvious implications. For example, in quality improvement, by ordering a data series from most to least recent, it can be used to help identify the duration of the current stable conditions.

The dynamic programming algorithm provides an alternative to the hierarchic binary splitting that dominates current work on multiple change points. It gives an exact optimum with a generally modest amount of computing and therefore seems preferable except for data sets so large that computation starts to be a concern.

Acknowledgements

The author is grateful to the referees for several suggestions for improving the paper.

References

Bellman, R.E., Dreyfus, S.E., 1962. Applied Dynamic Programming. Princeton University Press, Princeton.

Bellman, R., Roth, R., 1969. Curve fitting by segmented straight lines. J. Amer. Statist. Assoc. 64, 1079–1084.

Bhattacharya, P.K., 1994. Some aspects of change-point analysis. In: Carlstein, E., Muller, H.G., Siegmund, D. (Eds.), Change-Point Problems. IMS Monograph, Institute for Mathematical Statistics, Hayward, pp. 28–56.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth, Belmont, CA.

Chen, J., Gupta, A.K., 1997. Testing and locating variance change-points with applications to stock prices. J. Amer. Statist. Assoc. 92, 739–747.

Halpern, A.L., 2000. Multiple-changepoint testing for an alternating segments model of a binary sequence. Biometrics 56, 903–908.

Hawkins, D.M., Ten Krooden, J.A., 1979. Zonation of sequences of heteroscedastic multivariate data. Comput. Geosci. 5, 189–194.

Hawkins, D.M., 1972. On the choice of segments in piecewise approximation. J. Inst. Math. Appl. 9, 250–256.

Hawkins, D.M., 1976. Point estimation of the parameters of a piecewise regression model. Appl. Statist. 25, 51–57.

Hawkins, D.M., 1979. Testing a sequence of observations for a shift in location. J. Amer. Statist. Assoc. 72, 180–186.

Hawkins, D.M., Merriam, D.F., 1973. Optimal zonation of digitized sequential data. Math. Geo. 5, 389–396.

Hawkins, D.M., Merriam, D.F., 1975. Segmentation of discrete sequences of geologic data. In: Whitten, E.H.T. (Ed.), Quantitative Studies in the Geological Sciences. Geological Society of America, Washington, DC, pp. 311–316.

Hinkley, D.V., 1970. Inference about the change-point in a sequence of random variables. Biometrika 57, 1–17.

Hsu, D.A., 1979. Detecting shifts of parameter in gamma sequences with applications to stock price and air traffic flow analyses. J. Amer. Statist. Assoc. 74, 31–40.

Irvine, J.M., 1982. Changes in regime in regression models. Ph.D. Thesis, Yale University.

Rowland, R.St.J., Sichel, H.S., 1961. Statistical quality control of routine underground sampling. J. South African Inst. Mining Metall. 60, 251–284.

Venter, J.H., Steel, S.J., 1996. Finding multiple abrupt change points. Comput. Statist. Data Anal. 22, 481–504.

Worsley, K.J., 1986. Confidence regions and tests for a change-point in a sequence of exponential family random variables. Biometrika 73, 91–104.

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.