journal of
statistical planning
and inference

ELSEVIER

Journal of Statistical Planning and
Inference 59 (1997) 337–353

Bayesian criteria for discriminating among regression models with one possible change point

Faouzi Lyazrhi*

Laboratoire de Statistique et Probabilités, Université Paul Sabatier, 118, Rte de Narbonne, 31062 Toulouse Cedex, France

Received 25 April 1994

Abstract

The change-point problem for normal regression models is considered here as the problem of choosing the hypothesis H_0 of no change or one of the hypotheses H_i that one or more parameters change after the ith observation. The observations are often associated with a known increasing sequence τ_i (for example, τ_i is the date of the ith observation). It then seems natural to introduce a quadratic loss function involving $(\tau_i - \tau_j)^2$ for selecting H_i instead of the true hypothesis H_j. A Bayes optimal invariant procedure is derived within such a framework and compared to previous proposals. When H_0 is rejected, large errors may arise in the estimation of the change point. To get around this difficulty another procedure is introduced whose main feature is to select one of the H_i's when H_0 is rejected only if there is sufficient evidence in favour of this choice.

AMS subject classification: 62J05; 62F15

Keywords: Bayes procedure; Change point; Invariant procedure; Linear regression model; Quadratic loss

1. Introduction

Let us consider a sequence of n independent observations, that is independent real-valued random variables Y_i $(i = 1,\dots,n)$. The set of indices is assumed to be naturally ordered, for example, the i's are associated with a given sequence of increasing times τ_i $(i = 1,\dots,n)$. In many cases, but not all, τ_i can be simply taken as i. The probability distribution of Y_i is known up to a parameter θ. If θ takes two values, one for $\tau_i \leqslant \tau_{i_0}$ and the other for $\tau_i > \tau_{i_0}$, one says that i_0 (or τ_{i_0}) is a change point for the observed sequence.

Two kinds of change-point problem have been dealt with in the literature. The first one is that of testing for the null hypothesis of no change versus the existence

* Tel.: +33 61556772; fax: +33 61556089; e-mail: lyazrhi@cix.cict.fr.

PII S0378-3758(96)00105-X

of a change occurring at some unknown time in a sequence of i.i.d. normal random variables (Page, 1955; Chernoff and Zacks, 1964; Gardner, 1969; Hawkins, 1975; Worsley, 1979), or in a simple linear model (Quandt, 1958; Farley and Hinich, 1970; Maronna and Yohai, 1978), or in a general linear model (Worsley, 1983; Jandhayala and MacNeill, 1991). The second problem is that of estimating the point at which the change occurs (Hinkley, 1969, 1971; Holbert and Broemeling, 1977; Schulze, 1982; Smith and Cook, 1980; Zacks, 1982). In practice, these two problems are strongly linked: when the null hypothesis is rejected it seems natural to wonder where the change occurs and, in fact, several usual test statistics provide implicitly a natural estimate of the change point.

In this paper we consider the normal linear model and give a formulation of the change-point problem dealing simultaneously with test and estimation by presenting the question as a choice between an hypothesis H_0 of no change and one of the hypotheses H_i's corresponding to a change at time τ_i.

Due to the previous assumptions, $|\tau_i - \tau_j|$ can be considered as a distance between H_i and H_j. This is exploited here by introducing a quadratic loss function for choosing H_i instead of H_j, as it is generally done in estimation problems (note, however, that we consider only a discrete set of possible changes as it is the case in most the literature, with the notable exceptions of Hinkley, 1969, 1971; Ferreira, 1975; Smith and Cook, 1980). This loss function leads to a new optimal procedure (Section 3), different from the likelihood ratio one (Worsley, 1983) which is Bayes optimal under a different set of assumptions, including in particular a simple (0/1) loss function (Lyazrhi, 1992). By some of its aspects, namely the choice between H_0 and $\bigcup_i \mathrm{H}_i$, our procedure is, however, close to Bayesian procedures previously proposed in the literature (Chernoff and Zacks, (1964) for the i.i.d. case, Farley and Hinich (1970) or Jandhayala and MacNeill (1991) for the linear regression).

It may be considered that the choice of one hypothesis H_i is only relevant if the probability of choosing the change point close to the true one is large enough. On the contrary, deciding $\bigcup_i \mathrm{H}_i$ without making a further choice between the H_i's may be wiser. In Section 4, we state such a decision problem and derive an optimal procedure for it.

Finally, Section 5 gives a concrete example and compares the behaviour of various procedures on simulated data for samples as well as linear models. It is worth noticing that our framework includes all the situations where the null hypothesis H_0 is contrasted with hypotheses H_i's stipulating that 'something happens at observation i'. Our presentation emphasizes the change-point problem, but the derived procedures may be easily adapted to several other problems, for example, to the detection of a possible outlier.

2. Notation and framework

The n-dimensional normal distribution with mean μ and variance-covariance matrix V will be denoted by $\mathrm{N}_n(\mu, V)$. We consider n random variables $Y_i\,(i = 1, \ldots, n)$ and a $n \times q$ matrix X of independent variables $(q < n)$. Without loss of generality, it is assumed that $\operatorname{rank}(X) = q$, so that the columns of X span a p-dimensional linear subspace of R^n, say Q. If Y denotes the column vector of the Y_i's and $\mathscr{L}(Y)$ the probability distribution of Y, it is assumed henceforth that $\mathscr{L}(Y) = \mathrm{N}_n(\mu, \sigma^2 I_n)$, where I_n is the unit matrix of order n and σ is a positive unknown parameter. The various models differ by the μ space.

The basic model, that is the null hypothesis H_0, is defined by $\mathrm{H}_0 : \mu \in Q$ or, equivalently, $\mu = X\beta$, where β is an unknown vector of q real parameters and X is a known $n \times q$ matrix.

Let Q_i be a given linear subspace of $\mathbb{R}^n$ contained in $Q^\perp$ (the linear subspace of $\mathbb{R}^n$ orthogonal to Q). The hypothesis H_i is defined by

$$\mathrm{H}_i : \quad \mu \in Q \oplus Q_i, \quad \mu \notin Q.$$

If the dimension of $Q \oplus Q_i$ is $q_i\,(> q)$, that is the dimension of Q_i is $k_i = q_i - q$, $Q \oplus Q_i$ can be spanned by the columns of an $n \times q_i$ matrix X_i and μ can be written as $\mu = X_i \beta_i$.

All the problems we deal with can be formulated as follows: a set of hypotheses H_i being given (including H_0 as the special case where Q_i reduces to zero), choose one of the models H_i.

In many problems the observations are ordered, for example, by the time. More precisely, we suppose that they are linked to a given increasing process $\tau_i\,(i = 1, \ldots, n)$ and, in practice, this allows us to introduce a distance $|\tau_i - \tau_j|$ between the indices i and j, when discriminating between H_i and H_j. Along the paper, $\pi_Q\,(\pi_{Q^\perp}$, etc.) denotes the orthogonal projector onto $Q\,(Q^\perp$, etc.) and we shall simplify the writing by putting π_i instead of π_{Q_i}.

Let us first consider some examples. Note that, in most cases, τ_i may be taken as i or as one of the explanatory variables according to their concrete meaning.

Examples

(i) *One change point in a sequence of i.i.d. random variables.* Consider a sequence of i.i.d. random variables with a change point in mean after the ith observation. Then we have

$$\mathrm{H}_i : \quad \mathbb{E}(Y) = \mu \mathbb{1} + \mu^* \mathbb{1}_i, \quad \mu^* \neq 0$$

with $\mathbb{1}_i = (0, \ldots, 0, 1, \ldots, 1)'$ where the first 1 occupies the $(i+1)$st rank. Here Q is spanned by $\mathbb{1} = \mathbb{1}_0$ and Q_i is spanned by $\pi_{Q^\perp}(\mathbb{1}_i)$, $q = 1$, $q_i = 1$ for any i.

(ii) *One change point in a multiple regression model.* Note $X = [x_1 | \ldots | x_n]'$ the $n \times q$ matrix of q independent variables, and $X_i^* = [0 | \ldots | 0 \ldots | x_{i+1} | \ldots | x_n]'$. If one change point occurs after the ith observation, the hypothesis H_i is

$$\mathrm{H}_i : \quad \mathbb{E}(Y) = X\beta + X_i^* \beta^*, \quad \beta^* \neq 0, \ q \leqslant i \leqslant n - q.$$

In this case Q_i is spanned by $[0 | \ldots | 0 \ldots | \pi_{Q^\perp}(x_{i+1}) | \ldots | \pi_{Q^\perp}(x_n)]$ and, in general, $k_i = q$ for all i.

(iii) *One outlier in a multiple regression model.* Let $e_i = \mathbb{1}_{i-1} - \mathbb{1}_i$ be the column matrix whose elements are 0 but the ith one which is 1. We have

$$\mathrm{H}_i: \ \mathbb{E}(Y) = X\beta + \beta^* e_i, \quad \beta^* \neq 0.$$

If $e_i \notin Q$, then Q_i is a one-dimensional space ($k_i = 1$) spanned by $\mathbb{1}_{Q^\perp}(e_i)$.

(iv) *One change point in simple linear regression constrained to continuity.* Suppose that a simple regression function can change while staying continous at x_i. Then X is a $n \times 2$ matrix $[x|\mathbb{1}]$, Q_i is spanned by $\Pi_{Q^\perp}(a_i)$ with $a_i = [0, \ldots, 0, x_{i+1} - x_i, \ldots, x_n - x_i]'$, $q_i = 1$, and the several hypotheses are

$$\mathrm{H}_0: \ \mathbb{E}(Y) = \beta_1 x + \beta_2 \mathbb{1},$$

$$\mathrm{H}_i: \ \mathbb{E}(Y) = \beta_1 x + \beta_2 \mathbb{1} + \beta^* a_i, \quad \beta^* \neq 0, \ 2 \leqslant i \leqslant n-2.$$

Invariance

These hypotheses are invariant under the group of transformations $\{y \to ay + b, a > 0, \ b \in Q\}$ and a maximal invariant is the normed vector of residuals $T = \pi_{Q^\perp}(Y)/\|\pi_{Q^\perp}(Y)\|$. In the sequel only invariant procedures are considered. The restriction to such procedures leads to performing the analysis through T. Under H_0 the distribution of T is the uniform probability on the unit sphere $S_{Q^\perp}$ of $Q^\perp$, that is $\mathscr{L}(T) = U_{Q^\perp}$. Under one of the alternatives, i.e. $\mu \in Q \oplus Q_i$, the distribution of T has a density $g_i(S_{Q^\perp} \to \mathbb{R}_+)$ with respect to $U_{Q^\perp}$ given by (see Caussinus and Vaillant, 1985)

$$g_i(t, \mu, \sigma) = \frac{1}{2^{m/2-1}\Gamma(\frac{m}{2})} \mathrm{e}^{-\|\pi_i(\mu)\|^2/2\sigma^2} h_m\left(\frac{\langle t, \pi_i(\mu)\rangle}{\sigma}\right), \ t \in S_{Q^\perp}, \tag{1}$$

where:

$$h_m(u) = \int_0^\infty \mathrm{e}^{uv}\mathrm{e}^{-v^2/2}\, v^{m-1}\,\mathrm{d}v, m = n - q.$$

Note that $g_i(t, \mu, \sigma)$ depends only on $\theta = \pi_i(\mu)/\sigma$. Therefore, it will be further denoted by $g_i(t, \theta)$.

3. Bayes optimal procedure

Let us denote by $\mathrm{H}_1, \ldots, \mathrm{H}_J$ the set of alternative hypotheses.

In this section we shall give a Bayes optimal (invariant) procedure $(d_0, d_1, \ldots, d_J)$ minimising the Bayes risk

$$R(d, P, l) = \sum_{i=0}^{J} \int_{S_{Q^\perp}} r_i(t)\,\mathrm{d}_i(t)\,\mathrm{d}U_{Q^\perp}(t)$$

with

$$r_i(t) = \sum_{j=0}^{J} p_j \int_{Q_J} g_j(t, \theta)\,\ell(i, j, \theta)\,\mathrm{d}P_j(\theta),$$

where g_j is given above ($g_0 = 1$) and p_j, ℓ and P_j are defined as follows:

(a) *Prior distributions*

- The prior probability of H_j is p_j ($\sum_{j=0}^{J} p_j = 1$).
- The conditional prior distribution of θ given the jth model is P_j. For $j \neq 0$, it will be assumed that P_j has a k_j-dimensional density f_j with respect to the Lebesgue measure of Q_j. For $j = 0$, P_0 is a Dirac measure on 0.

(b) *Loss function*

We define the following loss function for selecting H_i when H_j holds with the value θ of the parameter:

$$\ell(i,j,\theta) = \begin{cases} 0 & \text{if } i = j, \\ \ell_1 & \text{if } i \neq 0,\ j = 0, \\ \ell_2(\theta) & \text{if } i = 0,\ j \neq 0, \\ \ell_3(\theta)(\tau_i - \tau_j)^2 & \text{if } i \neq 0,\ j \neq 0, \end{cases}$$

where $\ell_1 \in \mathbb{R}_+$, ℓ_2 and ℓ_3 are positive functions.

When H_0 is rejected while it is true, it seems natural that the loss neither depends on i nor, of course, on θ. When H_0 is accepted when it is false, we assess that the loss is the same for any true hypothesis H_j. Finally, when H_0 is false and rejected, we assume that the loss is quadratic with respect to the difference between the true and estimated change points.

Let

$$A_j(t) = \int_{Q_j} \ell_3(\theta) g_j(t,\theta) f_j(\theta)\,\mathrm{d}\theta \quad (j \neq 0),$$

$$a_j(t) = \frac{p_j A_j(t)}{\sum_{h=1}^{J} p_h A_h(t)},$$

$$\bar{\tau}(t) = \sum_{j=1}^{J} a_j(t)\tau_j,$$

and let $i^*(t)$ be such that $\tau_{i^*(t)}$ is the closest τ_i to $\bar{\tau}(t)$.

From the foregoing assumptions (a) and (b), we have

$$r_0(t) = \sum_{j=1}^{J} p_j \int_{Q_j} g_j(t,\theta)\,\ell_2(\theta)\, f_j(t)\,\mathrm{d}\theta,$$

$$r_i(t) = \ell_1\, p_0 + \sum_{j=1}^{J} p_j A_j(t)(\tau_i - \tau_j)^2,$$

$r_i(t)$ is minimum over $i (i \neq 0)$ for $i = i^*(t)$, and

$$\min_{i \neq 0} r_i(t) = \ell_1\, p_0 + \sum_{j=1}^{J} p_j A_j(t)(\tau_{i^*(t)} - \tau_j)^2.$$

Hence we have the following.

Proposition 3.1. *Given the previous assumptions* (a) *and* (b), *the following rule is Bayes optimal*:

$$\begin{cases} \textit{decide in favour of } \mathrm{H}_0 \textit{ if} \\ r_0(T) < \ell_1 p_0 + \sum_{j=1}^{J} p_j A_j(T)(\tau_j - \tau_{i^*(T)})^2 \\ \textit{decide in favour of } \mathrm{H}_j\,(j \neq 0) \textit{ if } j = i^*(T) \textit{ and} \\ r_0(T) > \ell_1 p_0 + \sum_{j=1}^{J} p_j A_j(T)(\tau_j - \tau_{i^*(T)})^2, \end{cases}$$

Proposition 3.1 can be used in two steps : first decide whether or not there is a change and, if it this is the case, estimate its location by $\tau_{i^*(T)}$.

The decision rule provided by Proposition 3.1 depends heavily on ℓ_2, ℓ_3 and the prior densities f_j's. To operationalize the proposition, we first have to make some assumptions about these quantities. Since the hypotheses H_i express similar changes for any i, we assume that the Q_i's have the same dimension and we set

$$k_i = k \quad \text{for all } i = 1,\ldots,J. \tag{2}$$

We then suppose that the losses ℓ_2 and ℓ_3 verify

$$f_j(\theta)\, l_2(\theta) = a, \tag{3}$$

$$f_j(\theta)\, l_3(\theta) = b \tag{4}$$

for any $\theta \in Q_j \setminus \{0\}$ and any $j \neq 0$.

The conditions above seem fairly realistic since they express that the cost of deciding H_0 for a given $\theta\,(\theta \neq 0)$ increases as this value of θ becomes less probable. For example, for a proper prior, $f_j(\theta)$ is small when $\|\theta\|$ is large and $\ell_2(\theta)$ or $\ell_3(\theta)$ are accordingly large. Note that (3) and (4) hold in the special case where f_j is constant (P_j is a vague prior) while ℓ_2 and ℓ_3 are constant ('simple' loss function). This set of assumptions leads to explicit formulas for the risks $r_i(t)$. Actually, we have from (1), (2) and (4):

$$A_j(t) = \frac{b}{2^{m/2-1}\Gamma(\frac{m}{2})} \int_{v=0}^{\infty} \mathrm{e}^{-v^2/2}\, v^{m-1} \left[\int_{\theta \in Q_j} \mathrm{e}^{\langle v\,\pi_j(t), \theta \rangle}\, \mathrm{e}^{-\frac{\|\theta\|^2}{2}}\, \mathrm{d}\theta \right] \mathrm{d}v.$$

The integral between brackets is equal to $(2\pi)^{k/2}\, \mathrm{e}^{v^2\, \|\pi_j(t)\|^2/2}$ (moment generating function of the normal distribution up to a multiplicative factor) and the integral over v is then easily computed. A similar result is obtained with ℓ_2 instead of ℓ_3, and we have:

Lemma 3.1. *If* (a), (b) *and* (2)–(4) *hold*:
- $A_j(t) = b(2\pi)^{k/2}(1 - \|\Pi_j(t)\|^2)^{-m/2}, \; j = 1,\dots,J,$
- $r_0(t) = a(2\pi)^{k/2}\sum_{j=1}^{J} p_j(1 - \|\Pi_j(t)\|^2)^{-m/2}$

We can now prove the following proposition.

Proposition 3.2. *Under* (a), (b) *and* (2)–(4), *the following rule is Bayes optimal*:

$$\begin{cases} \text{decide in favour of } \mathrm{H}_0 \text{ if} \\ \displaystyle\sum_{j=1}^{J} p_j(1 - \|\Pi_j(T)\|^2)^{-m/2}\left(1 - \frac{b}{a}(\tau_j - \tau_{i^*(T)})^2\right) < w_0. \\ \text{otherwise decide in favour of } \mathrm{H}_{i^*(T)} \text{ so that } \tau_{i^*(T)} \text{ is the nearest } \tau_i \text{ to} \\ \dfrac{\sum_{j=1}^{J} p_j(1 - \|\Pi_j(T)\|^2)^{-m/2}\,\tau_j}{\sum_{j=1}^{J} p_j(1 - \|\Pi_j(t)\|^2)^{-m/2}}. \end{cases} \tag{5}$$

Proof. Using the results provided by Lemma 3.1, Proposition 3.1 leads to decide in favour of H_0 if

$$a(2\pi)^{k/2}\sum_{j=1}^{J} p_j(1 - \|\Pi_j(T)\|^2)^{-m/2} < \ell_1 p_0 + b(2\pi)^{k/2}\sum_{j=1}^{J} p_j(1 - \|\Pi_j(T)\|^2)^{-m/2} \times(\tau_i - \tau_{i^*(T)})^2.$$

Hence, we get the first step of the proposition with $w_0 = (\ell_1 p_0/a)(2\pi)^{-k/2}$. If H_0 is rejected, Proposition 3.1 leads to decide in favour of H_{i^*} so that $|\tau_i - \bar{\tau}(T)|$ reaches its minimum value over $i \neq 0$ for τ_{i^*}. The second step of the proposition comes out by using the actual value of $\bar{\tau}(T)$, viz. (5).

Remark 3.1. In practical situations the p_j's ($j \neq 0$) can be naturally taken as equal, but other choices are possible: see Ferreira (1975). A proper choice of b/a may be easily discussed. From (3), (4) and the definition of $\ell(i,j,\theta)$, $(b/a)(\tau_i - \tau_j)^2$ is the ratio of the loss of selecting H_i to the loss of selecting H_0 when the true hypothesis is H_j. If we assume that it is better to select H_0 than H_i when $|\tau_i - \tau_j|$ is greater than $\Delta\tau$, while it is better to select H_0 if the difference $|\tau_i - \tau_j|$ is smaller than $\Delta\tau$, we are led to put $(b/a)\Delta^2\tau = 1$ which provides a suitable value for b/a. On the other hand, w_0 is more difficult to assess from prior assumptions in practice. We propose to get round the difficulty by setting a given probability, say α, for wrongly rejecting H_0. For given b/a, this is theoretically possible since the probability distribution of T under H_0 depends neither on Q nor on σ. Let $w(\Delta\tau, \alpha)$ be the critical value corresponding to the risk α; therefore, the procedure given by Proposition 3.1 can be rewritten as:

Corollary 3.1. *Under* (a), (b) *and* (2)–(4), *and using the expression of* $\Delta\tau$, *the following rule is Bayes optimal:*

$$\begin{cases} \textit{accept } \mathrm{H}_0 \textit{ if } \sum_{j=1}^{J} p_j (1 - \|\Pi_j(T)\|^2)^{-m/2} \left(1 - \left(\frac{\tau_j - \tau_{i^*(T)}}{\Delta\tau}\right)^2\right) < w(\Delta\tau, \alpha) \\ \textit{otherwise, decide in favour of } \mathrm{H}_{i^*} \textit{ so that } \tau_{i^*(T)} \textit{ is the nearest } \tau_i \textit{ to} \\ \dfrac{\sum_{j=1}^{J} p_j (1 - \|\Pi_j(T)\|^2)^{-m/2}\, \tau_j}{\sum_{j=1}^{J} p_j (1 - \|\Pi_j(T)\|^2)^{-m/2}}. \end{cases}$$

Remark 3.2. (1) If $\Delta\tau$ is large, that is emphasis is put on the detection of a change point rather than the estimation of its precise location, then $\left((\tau_j - \tau_{i^*})/\Delta\tau\right)^2$ can be neglected and the first step of the rule becomes

$$\text{decide in favour of } \mathrm{H}_0 \text{ if } \sum_{j=1}^{J} p_j (1 - \|\Pi_j(T)\|^2)^{-m/2} < w(\infty, \alpha). \tag{6}$$

One can show that (6) is exactly the Bayes optimal rule that we obtain for testing H_0 against $\bigcup_{i=1}^{J} \mathrm{H}_i$ when (3) holds and the simple loss function (0/1) is considered (i.e. $\ell_2(\theta) = 1$).

(2) On the other hand, the Bayes factor for testing H_0 against $\bigcup_{i=1}^{J} \mathrm{H}_i$ (see Smith, 1975; Booth and Smith, 1982) is

$$B_0 = \sum_{j=1}^{J} p_j B_{j0}, \text{ with } B_{j0} = \frac{\mathrm{pr}(\mathrm{H}_j|T)}{\mathrm{pr}(\mathrm{H}_0|T)} \Big/ \frac{p_j}{p_0}.$$

(The notation pr(.|.) is a generic symbol for a conditional probability). Using Bayes' theorem we can reexpress this in the form: $B_{j0} = \mathrm{pr}(T|\mathrm{H}_j)/\mathrm{pr}(T|\mathrm{H}_0)$, with $\mathrm{pr}(T|\mathrm{H}_j) = \int_{Q_j} g_j(t,\theta)\,\mathrm{d}P_j(\theta)$ for $j = 0, 1$.

Then, if the P_j's are vague prior distributions, we get by using (3),

$$B_{j0} = a(2\pi)^{k/2} (1 - \|\Pi_j(T)\|^2)^{-m/2} \text{ and } B_0 = a(2\pi)^{k/2} \sum_{j=1}^{J} p_j (1 - \|\Pi_j(T)\|^2)^{-m/2}.$$

Of course, this Bayes factor depends on the undefined constant a. This leads us to propose the P value solution to get around this. For another solution, see, for example, Speigelhalter and Smith (1982). Note that the problem does not arise if we test H_i against H_j ($i \neq 0, j \neq 0$) since $B_{0i} = B_{i0}^{-1}$ and $B_{ij} = B_{i0} B_{0j}$.

(3) In practice, the change point will be the integer closest to $\bar{\tau}(T)$ given in (5) which is the Bayes' estimator $\mathbb{E}(\tau|T)$ arising from the quadratic term in the loss function (see loss function).

(4) Finally, when replacing m by n, the left-hand side of (6) is the average of the different likelihood ratios $(1 - \|\Pi_j(T)\|)^{-n/2}$ under the different alternatives. Hence, (6) is similar to the procedures considered by Chernoff and Zacks (1964), Jandhayala and MacNeill (1991), Farley and Hinich (1975), which rest on such averages.

4. Another Bayes optimal rule in a more realistic framework

The most current situation in the literature consists in testing H_0 against $\bigcup_{i=1}^{J} \mathrm{H}_i$ (that is no change against any change). Section 3 deals with the more specific problem where, when H_0 is rejected, an estimate of the change point is simultaneously provided. But we can imagine an intermediate situation with three kinds of decisions: (i) decide that there is no change, (ii) decide that there is a change and estimate its position, and (iii) decide that there is a change without specifying its position.

With respect to our previous framework, the latter element has now to be added to the set of all possible decisions. It will be indexed by C (for 'change'). This leads to the new loss function where j refers to the true hypothesis ($j = 0, 1, \ldots, J$) while i refers to the decision ($i = 0, C, 1, \ldots, J$):

$$\ell(i,j,\theta) = \begin{cases} 0 & \text{if } i = j, \\ \ell_1 & \text{if } i \neq 0,\ i \neq C,\ j = 0, \\ \ell_1' & \text{if } i = C,\ j = 0, \\ \ell_2(\theta) & \text{if } i = 0,\ j \neq 0, \\ \ell_3(\theta)(\tau_i - \tau_j)^2 & \text{if } i \neq 0,\ i \neq C,\ j \neq 0, \\ \ell_4(\theta) & \text{if } i = C,\ j \neq 0, \end{cases}$$

where ℓ_1, $\ell_2(.)$ and $\ell_3(.)$ are defined as in the previous section, $\ell_1' > 0$ and $\ell_4(.)$ is a positive function such that

$$\ell_4(\theta)\, f_j(\theta) = c \quad \text{for all } \theta \in Q_j \backslash \{0\},\ j \neq 0. \tag{7}$$

If H_j holds for some $j \neq 0$, then choosing H_0 is clearly worse than choosing H_C, since H_C includes H_j: hence, we must have

$$c < a \tag{8}$$

(even, in practice, c should be much less than a). ℓ_1' (resp. ℓ_1) is the cost of choosing H_C (resp. H_i, $i \neq 0$) when H_0 holds. In practice, it is clear that ℓ_1' should be equal to or perhaps slightly less than ℓ_1. A simple optimal procedure will be obtained by letting

$$\ell_1' = \ell_1 \frac{a-c}{a}. \tag{9}$$

Finally, let $\Delta^2\tau = c/b$, where c and b are defined by (7) and (4). Now, $\Delta\tau$ is a distance ($|\tau_i - \tau_j|$) between two hypotheses H_i and H_j such that the cost of choosing H_i for H_j is the same as the cost for choosing H_C for H_j. We get

$$r_C(t) = \ell_1' \, p_0 + c(2\pi)^{k/2} \sum_{j=1}^{J} p_j (1 - \|\Pi_j(t)\|^2)^{-m/2} = \ell_1' \, p_0 + c\,(2\pi)^{k/2} \sum_{j=1}^{J} C_j(t),$$

with $C_j(t) = p_j(1 - \|\Pi_j(t)\|^2)^{-m/2}$, while $r_0(t)$ and $r_i(t), i \neq 0, i \neq C$, remain unchanged. We have immediately

$$r_0(t) < r_C(t) \quad \text{if} \sum_{j=1}^{J} C_j(t) < \frac{\ell_1' p_0}{a-c}(2\pi)^{-k/2}$$

and, with $i^*(t)$ defined as above,

$$r_0(t) < r_{i^*}(t) \quad \text{if} \sum_{j=1}^{J} C_j(t) < \frac{\ell_1 p_0}{a}(2\pi)^{-k/2} + \frac{b}{a}\sum_{j=1}^{J} C_j(t)(\tau_i - \tau_{i^*(t)})^2.$$

From (9), it is clear that the former inequality implies the latter since $\ell_1' p_0/(a-c) = \ell_1 p_0/a$. Therefore, H_0 will be accepted if $\sum_{j=1}^{J} C_j < (\ell_1 p_0/a)(2\pi)^{-k/2}$.
Further, it is easy to verify

$$r_C(t) < r_{i^*}(t) \quad \text{if} \sum_{j=1}^{J} C_j(t)\left(1 - \left(\frac{\tau_j - \tau_{i^*(t)}}{\Delta\tau}\right)^2\right) < \frac{\ell_1 p_0}{a}(2\pi)^{-k/2}.$$

Let $\bar{\tau}(T)$ be still defined by (5). We have thus proved the following.

Proposition 4.1. *Within the framework defined above, with assumptions* (a), (b), (2)–(4), (7)–(9), *the following procedure is Bayes optimal:*

$$\begin{cases} \textit{accept } \mathrm{H}_0 \textit{ if } \displaystyle\sum_{j=1}^{J} p_j(1 - \|\Pi_j(T)\|^2)^{-m/2} < w_0 \\ \textit{accept } \mathrm{H}_C \textit{ if } \displaystyle\sum_{j=1}^{J} p_j(1 - \|\Pi_j(T)\|^2)^{-m/2} > w_0 \\ \qquad\qquad \textit{and } \displaystyle\sum_{j=1}^{J} p_j(1 - \|\Pi_j(T)\|^2)^{-m/2}\left(1 - \left(\frac{\tau_j - \tau_{i^*(T)}}{\Delta\tau}\right)^2\right) < w_0 \\ \textit{accept } \mathrm{H}_{i^*} \textit{ so that } \tau_{i^*(T)} \textit{ is the nearest } \tau_i \textit{ to } \bar{\tau}(T) \textit{ if} \\ \displaystyle\sum_{j=1}^{J} p_j(1 - \|\Pi_j(T)\|^2)^{-m/2}\left(1 - \left(\frac{\tau_j - \tau_{i^*(T)}}{\Delta\tau}\right)^2\right) > w_0. \end{cases}$$

The procedure given by Proposition 4.1 depends on two constants w_0 and $\Delta\tau$. The critical value w_0 can be determined by fixing a level α (it is then the same value as in (6)) and can be replaced by $w(\infty, \alpha)$, while $\Delta\tau$ is to be chosen by the user in the light of the discussion above, according to the importance given to the misspecification vs. the non-specification of a change point.

5. Applications

In this section we examine a numerical example to illustrate the behaviour of the procedures given in Corollary 3.1 and in Proposition 4.1 with several values of $\Delta\tau$. Then we present simulated experiments to compare the performances of these procedures to others procedures proposed in the literature, respectively, by Farley and Hinich (1970) and Worsley (1983). Two questions have been given attention: (i) The comparison of the powers, that is the probabilities of rejecting H_0 when it does not hold with no account of the actual choice of some H_i (this allows comparisons with procedures which do not include such a choice).

(ii) The simultaneous comparison of the probabilities of the various choices, with emphasis on finding a good estimate of the change point.

From the many cases which have been processed, a few typical ones have been selected to present the main features of the comparisons.

For these procedures the computations of critical values have been discussed in the literature from a number of viewpoints (Worsley, 1983; Kim and Siegmund, 1989; Jandhayala and MacNeill, 1991). Analytical determination is difficult for most of them as well as for the procedures proposed in this paper. It is very easy, however, and less costly to get good approximate values by simulation. Therefore, the critical values which could not be found in the literature have merely been obtained from 100 000 simulated experiments under H_0.

5.1. Human fertility in Iran

The data described in Raftery (1993) and Raftery et al. (1995) concern the human fertility in Iran for the years 1949–1977, the period leading up to the islamic revolution.

The fertility period effect by year Y_i is assumed to be a linear function of the time $x_i = i = \tau_i$, $i = 1, \ldots, n = 29$, that is

$$H_0: \quad \mathbb{E}(Y) = \beta_0 \mathbb{1} + \beta_1 x.$$

It is suspected, however, that the parameters of the regression may have changed after an unknown time i with constraint of continuity (see Section 1, example (iv)), hence the change-point regression model is

$$H_i: \quad \mathbb{E}(Y) = \beta_0 \mathbb{1} + \beta_1 x + \beta^* a_i, \quad i = 2, \ldots, n-2.$$

In this example we suppose that the p_i's ($i \neq 0$) are equal. The procedure given in Corollary 3.1 has been used first for the 0.01 level and various values of $\Delta\tau$. For $\Delta\tau$ infinite (procedure (6)), H_0 is rejected with $n^{-1}\sum_{j=1}^{J}(1 - \|\Pi_j(T)\|^2)^{-m/2} = 455.40$ and a critical value equal to 34.23. When H_0 is rejected, $\bar{\tau}(T) = 11.85$ indicates that a change point occurs at 1959.

This agrees with the results obtained by Raftery who uses the Bayes factor. In particular, under some hypotheses he shows that to test H_0 against $\bigcup_{i=2}^{n-2} H_i$ it is asymptotically equivalent to use $n^{-2}\sum_{j=1}^{J}(1 - \|\Pi_j(T)\|^2)^{-n/2}$ which is equal to 26.18

in this example. To estimate the change point, he uses the posterior probability of τ rather than $\bar{\tau}(T)$.

The fact that the acceptance or the rejection of H_0 depends on the choice of $\Delta\tau$ is an inconvenience of this procedure, so that it seems better to use the procedure given in Proposition 4.1 which gets around this. The results obtained are that we reject H_0 at the same 0.01 level and

- for $\Delta\tau \leqslant 4$ we accept H_C (i.e. do not try to estimate the change point).
- for $\Delta\tau > 4$ we estimate the change point as above.

5.2. *Power comparisons*

We give empirical comparisons of the procedure derived in Section 4 (denoted by Q3) and the procedures proposed, respectively, by Farley and Hinich (1970) and Worsley (1983). The discussion is limited here to the context of simple linear regression when the change occurs in slope, constrained to continuity (see Section 1 example (iv), from where the notation is borrowed). The values of the explanatory variable are $x_i = i/n$ and $\tau_i = i (i = 1, \ldots, n)$.

Worsley procedure is the likelihood ratio test of H_0 against $\bigcup_{i=2}^{n-2} \mathrm{H}_i$, which rests on $\max_{i \neq 0} \|\pi_i(T)\|$. Farley and Hinich (1970) derived a locally most powerful procedure for testing H_0 against $\bigcup_{i=2}^{n-2} \mathrm{H}_i$. Their statistic is: $z' \, T / \|\pi_{Q^\perp}(z)\|$, with

$$z = \left(0, \frac{\sum_{j=1}^{2}(x_2 - x_j)}{n}, \ldots, \frac{\sum_{j=1}^{n}(x_n - x_j)}{n}\right).$$

The critical value is the $(1 - \alpha/2)$th quantile of the N(0,1) distribution.

Under both the null and the alternative hypotheses, all the involved statistics do not depend on β_1 and β_2, hence we need not specify any values for these unknown parameters.

The sample size is $n = 50$. For selected values of $|\beta^*|$ (viz. $|\beta^*| = 1, 2, 3$), and of the change point ($j_0 = 5(5)45$), the various statistics have been computed 1000 times. Based on these values, the empirical power for each test statistic has then been evaluated for $\alpha = 0.05$ and presented in Figs. 1–3.

Roughly speaking, no procedure is uniformly most powerful, the efficiency depending on the amount of change and the location of the change point (see also James et al. (1987) or Sen and Srivastava (1975) for similar results concerning the comparison of Bayesian and likelihood approaches). The Farley and Hinich procedure performs well compared to that of Worsley for change points around the middle of the sample, which confirms the result of Jandhayala and MacNeill (1991) in the same context. On the other hand, our procedure is the most powerful when β^* is large and the change occurs far enough from the middle of the sample, but it is less powerful than the Farley and Hinich procedure (although better than the Worsley procedure) when the change occurs around the middle of the sample. Thus, our procedure may be the most powerful and is never the least powerful of the three.

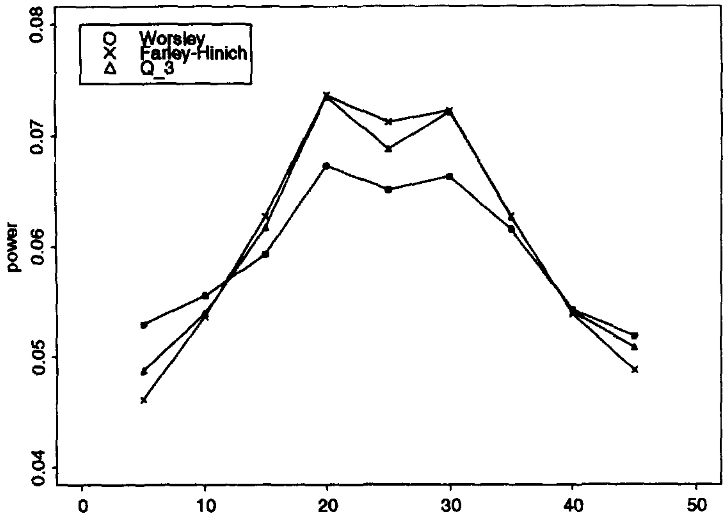

Fig. 1. Change in intercept and slope (constrained to continuity). $n = 50$, $\beta^* = 1$, $\alpha = 0.05$.

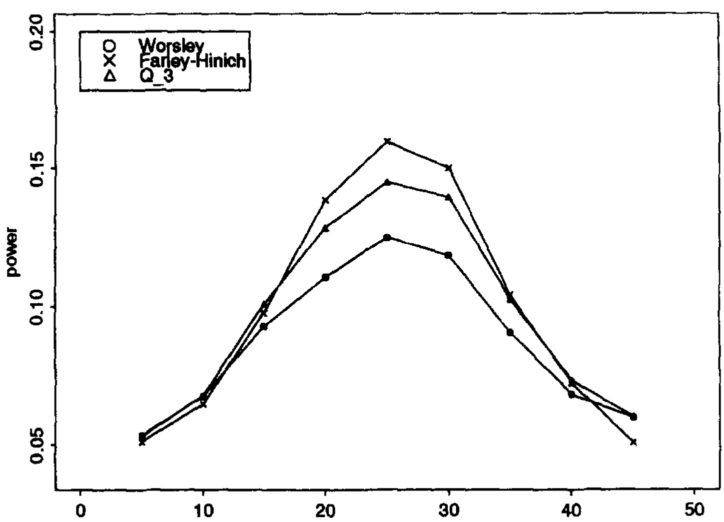

Fig. 2. Change in intercept and slope (constrained to continuity). $n = 50$, $\beta^* = 2$, $\alpha = 0.05$.

Simulations have been done in other contexts, in particular in comparing our procedure to the Bayes-type procedure (Jandhayala and MacNeill, 1991) and we arrived at similar conclusions.

5.3. Probabilities of good estimation

In this section simulation results are given for two cases: (i) a sequence of i.i.d. normal random variables with a possible shift in mean, and (ii) simple linear regression

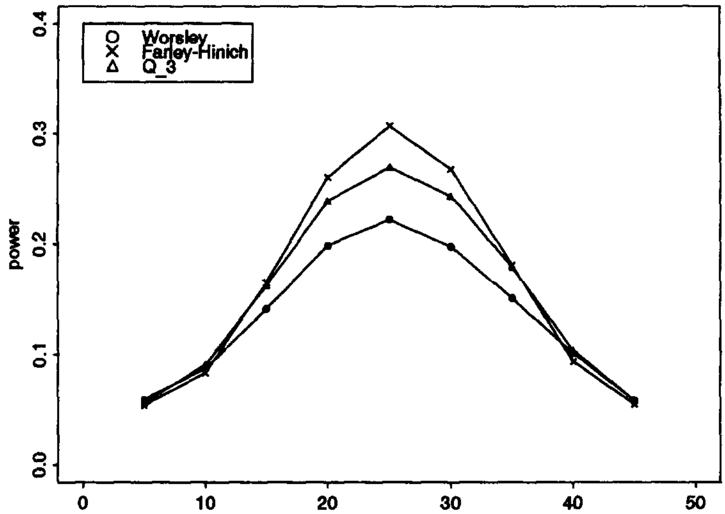

Fig. 3. Change in intercept and slope (constrained to continuity). $n = 50$, $\beta^* = 3$, $\alpha = 0.05$.

Table 1
Sequence of i.i.d. random variables. $n = 50$, $j_0 = 5$, $\mu^* = 1$, $\alpha = 0.05$

	Worsley	Q1	Q2	Q3
Reject H_0	0.342	0.335	0.270	0.335
H_5	0.104	0.031	0.031	0.021
$i \in [2,8]$	0.251	0.163	0.153	0.064
$i \in [2,11]$	0.275	0.231	0.202	0.074
$i < 2$, $i > 39$	0.021	0.009	0.006	0.001
H_C				0.746

with possible change in intercept and slope. In all cases $n = 50$ and $\sigma = 1$. For the first model, the values of the different parameters are $\mu^* = 1$ (with the notation of example (i) in Section 2) and $j_0 = 5, 15, 25$. In the second case (regression), the explanatory variable is $x_i = i/n$, $\tau_i = x_i\,(i = 1, \ldots, n)$, the change point is $j_0 = 5, 15, 25, 35$ and the amount of change is $\beta^* = [1, 1]'$ (see example (ii) in Section 2). The frequencies of the different decisions have been computed for the procedure proposed by Worsley (1983), the procedure given in Corollary 3.1 with $\Delta\tau$ infinite and $\Delta\tau = 10$ (denoted, respectively, by Q1 and Q2) and the procedure given in Proposition 4.1 with $\Delta\tau = 5$ (denoted by Q3).

The frequencies obtained over 1000 replications for $\alpha = 0.05$ are displayed in Tables 1–5. To improve the readability, the results are grouped in the following way:

row 1 : frequency of H_0 rejection
row 2 : frequency of exact choice of H_{j_0}
row 3 : frequency of 'good' estimates of H_{j_0}
row 4 : frequency of 'fair' estimates of H_{j_0}

Table 2
Sequence of i.i.d. random variables. $n = 50$, $j_0 = 25$, $\mu^* = 1$, $\alpha = 0.05$

	Worsley	Q1	Q2	Q3
Reject H_0	0.800	0.836	0.830	0.836
H_{25}	0.210	0.132	0.141	0.090
$i \in [22,28]$	0.547	0.596	0.611	0.363
$i \in [17,33]$	0.701	0.793	0.796	0.442
$i < 11$, $i > 39$	0.039	0.003	0.003	0.001
H_C				0.384

Table 3
Simple linear regression. $n = 50$, $j_0 = 15$, $\beta^* = [1,1]'$, $\alpha = 0.05$

	Worsley	Q1	Q2	Q3
Reject H_0	0.296	0.316	0.310	0.316
H_{15}	0.108	0.029	0.031	0.018
$i \in [12,18]$	0.196	0.132	0.150	0.059
$i \in [9,24]$	0.229	0.240	0.248	0.073
$i < 6$, $i > 35$	0.031	0.014	0.012	0.003
H_C				0.236

Table 4
Simple linear regression. $n = 50$, $j_0 = 25$, $\beta^* = [1,1]'$, $\alpha = 0.05$

	Worsley	Q1	Q2	Q3
Reject H_0	0.345	0.334	0.370	0.334
H_{25}	0.140	0.095	0.107	0.068
$i \in [22,28]$	0.285	0.237	0.147	0.086
$i \in [17,33]$	0.313	0.301	0.342	0.155
$i < 10$, $i > 35$	0.025	0.013	0.003	0.002
H_C				0.175

Table 5
Simple linear regression. $n = 50$, $j_0 = 35$, $\beta^* = [1,1]'$, $\alpha = 0.05$

	Worsley	Q1	Q2	Q3
Reject H_0	0.670	0.707	0.647	0.707
H_{35}	0.300	0.082	0.082	0.076
$i \in [32,38]$	0.498	0.327	0.342	0.201
$i \in [26,41]$	0.556	0.585	0.561	0.224
$i < 12$, $i > 44$	0.039	0.004	0.005	0.000
H_C				0.469

row 5 : frequency of 'bad' estimates of H_{j_0}
row 6 : frequency of decisions in favour of H_C

From the third to the fifth row, a statement such as $i \in [2,8]$ means that one of the hypotheses H_i for $i \in [2,8]$ is selected. The length of the intervals for which we say that the estimate of H_{j_0} is good, fair or bad is somewhat arbitrary. In fact these intervals have been chosen to be as illustrative as possible. Note they are not necessarily symmetric when the change point is far from the middle of the sequence of observations.

In case (i), the Worsley procedure seems more powerful than the others when the change occurs early (or late) in the sequence, while, when the change occurs around the middle of the sequence, our procedures (especially Q1 and Q3) become more powerful. Concerning the estimation of the change point, the Worsley procedure gives exactly the correct position more often than the other procedures. But it can also be quite misleading, since it gives bad estimates in too many cases. These unfortunate events are almost entirely avoided with our procedures, as this would be expected, in particular of course with procedure Q3.

If we consider cumulative frequencies of good or fairly good estimates, our procedures perform well except if the change point occurs at the end of the sequence. In case (ii), if the change occurs around the beginning (or the end) of a series, Q1 and Q3 are more powerful than the other procedures. But, around the middle, Q2 becomes the most powerful, and Worsley procedure is slightly better than Q1 and Q3. Concerning the estimation of the change point, the Worsley procedure is often the best to find it exactly. On the whole, it is equivalent to Q1 and Q3 in finding a fairly good estimates but more often provides bad estimates. Concerning the very bad estimates, we get the same conclusion as in case (i).

References

Booth, N.B and A.F.M. Smith (1982). A Bayesian approach to retrospective identification of change-point. *J. Econometrics* **19**, 7–22.

Caussinus, H. and J. Vaillant (1985). Some geometric tools for the Gaussian linear model. In: *Linear Statistical Inference*, Lecture Notes in Statistics. Vol. 35. Springer Berlin, 1–19.

Chernoff, H. and S. Zacks (1964). Estimating the current mean of a normal distribution which is subject to changes in time. *Ann. Math. Statist.* **35**, 999–1018.

Farley, J.U. and M.J. Hinich (1970). A test for a shifting slope coefficient in a linear model. *J. Amer. Statist. Assoc.* **65**, 1320–1399.

Ferreira, P. (1975). A Bayesian Analysis of a switching regression model: known number of regimes. *J. Amer. Statist. Assoc.* **10**, 370–375.

Gardner, L.A. (1969). On detecting changes in the mean of normal variates. *Ann. Statist.* **40**, 116–126.

Hawkins, D.M. (1977). Testing a sequence of observation for a shift in location. *J. Amer. Statist. Assoc.* **72**, 180–186.

Hinkley, D.J. (1969). Inference about the intersection in two-phase regression. *Biometrika* **56**, 495–504.

Hinkley, D.J. (1971). Inference in two-phase regression. *J. Amer. Statist. Assoc.* **66**, 736–745.

Holbert, D. and L.D. Broemeling (1977). Bayesian inference related to shifting sequences and two-phase regression. *Commun. Statist. Theoret. Methods* **6**, 265–275.

James, B., K.L. James and D. Siegmund (1987). Tests for a change-point. *Biometrika* **74**, 71–83.

Jandhyala, V.K and I.B. MacNeill (1991). Tests for parameter changes at unknown times in linear regression models. *J. Statist. Plann. Inference* **27**, 291–316.

Kim, H. and D. Siegmund (1989). The likelihood ratio test for a change point in simple linear regression. *Biometrika* **76**, 409–423.

Lyazrhi, F. (1991). Détection de ruptures dans un modèle linéaire gaussien: optimalité de procédures de décision multiple basées sur le rapport de vraisemblance. *Statistique Analyse Données* **16**, 107–125.

Maronna, R. and U. Yohai (1978). A Bivariate test for the detection of a systematic change in means. *J. Amer. Statist. Assoc.* **73**, 640–645.

Page, E.S. (1955). A test for a change in a parameter occurring at an unknown time point. *Biometrika* **42**, 523–526.

Quandt, R.E. (1958). The estimation of the parameter of a linear regression system obeing two separate regimes. *J. Amer. Statist. Assoc.* **67**, 306–310.

Raftery, A.E. (1993). Change-point and change curve modeling in stochastic processes and spatial statistics. Technical Report no. 253, Department of Statistics, University of Washington.

Schulze, U. (1982). Estimation in segmented regression. *Statistics* **13**, 295–317.

Sen, A.K. and M.S. Srivastava (1975). On tests for detecting change in means. *Ann. Statist.* **3**, 98–108.

Smith, A.F.M. (1975). A Bayesian approach to inference about a change point in a sequence of random variables. *Biometrika* **62**, 407–416.

Smith, A.F.M. and S. Cook (1980). Straight lines with a change-point: a Bayesian analysis of some renal transplant data. *Appl. Statist.* **29**, 180–189.

Speigelhalter, D.J. and A.F.M. Smith (1982). Bayes factor for linear and log-linear models with vague prior information. *J. Roy. Statist. Soc. B* **44**, 377–387.

Worsley, K.J. (1979). On the likelihood ratio test for shift in location of normal population. *J. Amer. Statist. Assoc.* **74**, 36–57.

Worsley, K.J. (1983). Testing for a two-phase multiple regression. *Technometrics* **25**, 35–42.

Zacks, S. (1982). Classical and Bayesian approaches to the change-point problem: fixed samples and sequential procedures. *Statistique Analyse Données* **7**, 48–68.