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

Faouzi Lyazrhi*

Laboratoire de Statistique et Probabilités, Université Paul Sabatier, 118, Rue 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 $i$th observation. The observations are often associated with a known increasing sequence $r_i$ (for example, $r_i$ is the date of the $i$th observation). It then seems natural to introduce a quadratic loss function involving $(r_i - r_j)^2$ 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 $Y_i$ ($i = 1, \ldots, n$). The set of indices is assumed to be naturally ordered, for example, the $i$’s are associated with a given sequence of increasing times $r_i$ ($i = 1, \ldots, n$). In many cases, but not all, $r_i$ can be simply taken as $i$. The probability distribution of $Y_i$ is known up to a parameter $\theta$. If $\theta$ takes two values, one for $r_i \leq r_{i_0}$ and the other for $r_i > r_{i_0}$, one says that $i_0$ (or $r_{i_0}$) 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@cict.fr.
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 $\tau_i$.

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 $\bigcup_i H_i$, 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 $\bigcup_i H_i$ 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 $\mu$ and variance-covariance matrix $V$ will be denoted by $N_n(\mu, V)$. 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 rank(X) = q, so that the columns of X span a p-dimensional linear subspace of \( \mathbb{R}^n \), say \( Q \). If \( Y \) denotes the column vector of the \( Y_i \)'s and \( \mathcal{L}(Y) \) the probability distribution of \( Y \), it is assumed henceforth that \( \mathcal{L}(Y) = \mathcal{N}(\mu, \sigma^2 I_n) \), where \( I_n \) is the unit matrix of order \( n \) and \( \sigma \) is a positive unknown parameter. The various models differ by the \( \mu \) space.

The basic model, that is the null hypothesis \( H_0 \), is defined by \( H_0 : \mu \in Q \) or, equivalently, \( \mu = X\beta \), where \( \beta \) 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

\[
H_i : \quad \mu \in Q + Q_i, \quad \mu \notin Q.
\]

If the dimension of \( Q + Q_i \) is \( q_i > q \), that is the dimension of \( Q_i \) is \( k_i = q_i - q \), \( Q \perp Q_i \) can be spanned by the columns of an \( n \times q_i \) matrix \( X_i \) and \( \mu \) 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, \( \pi_Q(\pi_{Q_i}, \text{etc.}) \) denotes the orthogonal projector onto \( Q(Q_i, \text{etc.}) \) and we shall simplify the writing by putting \( \pi_i \) instead of \( \pi_Q \).

Let us first consider some examples. Note that, in most cases, \( \tau_i \) 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 \( i \)th observation. Then we have

\[
H_i : \quad \mathbb{E}(Y) = \mu \mathbb{1} + \mu^* \mathbb{1}_i, \quad \mu^* \neq 0
\]

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 \( \pi_{Q_i}(\mathbb{1}_i) \), \( q = 1, q_i = 1 \) 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 \( X_i^* = [0| \ldots |0| x_{i+1}| \ldots |x_n]' \). If one change point occurs after the \( i \)th observation, the hypothesis \( H_i \) is

\[
H_i : \quad \mathbb{E}(Y) = X\beta + X_i^* \beta^*, \quad \beta^* \neq 0, \quad q_i = i \leq n - q.
\]

In this case \( Q_i \) is spanned by \([0| \ldots |0| \pi_{Q_i}(x_{i+1})| \ldots |\pi_{Q_i}(x_n)]\) and, in general, \( k_i = q \) for all \( i \).
(iii) **One outlier in a multiple regression model.** Let $e_i = \|_{i-1} - \|_i$ be the column matrix whose elements are 0 but the $i$th one which is 1. We have

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

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

(iv) **One change point in simple linear regression constrained to continuity.** Suppose that a simple regression function can change while staying continuous at $x_i$. Then $X$ is a $n \times 2$ matrix $[x_1 \ldots x_{i+1} - x_i, \ldots, x_n - x_i]'$, $q_i = 1$, and the several hypotheses are

$$H_0 : \mathbb{E}(Y) = \beta_1 x + \beta_2 1,$$

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

**Invariance**

These hypotheses are invariant under the group of transformations $\{y \rightarrow ay + b, \quad a > 0, \quad b \in Q\}$ 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 $\mathcal{L}(T) = U_{Q^\perp}$. Under one of the alternatives, i.e. $\mu \in Q \oplus Q$, the distribution of $T$ has a density $g_i(T \in Q^\perp)$ with respect to $U_{Q^\perp}$ given by (see Caussinus and Vaillant, 1985)

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

where:

$$h_m(u) = \int_0^\infty e^{uw} e^{-v/2} v^{m-1} dv, \quad m = n - q.$$  

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 $H_1, \ldots, H_J$ 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

$$R(d, \mu, l) = \sum_{j=0}^J \int_{S_0} r_j(t) d_j(t) dU_{Q^\perp}(t)$$

with

$$r_j(t) = \sum_{j=0}^J p_j \int_{Q_j} g_j(t, \theta) \zeta(i, j, \theta) dP_j(\theta),$$

where

$$\zeta(i, j, \theta) = \theta_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t - \langle t, \pi_i(\mu) \rangle / \sigma)^2}{2\sigma^2}}.$$
where $g_j$ is given above ($g_0 = 1$) and $p_j$, $\ell$ 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 $\theta$ given the $j$th 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 $\theta$ 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 $\ell_1 \in \mathbb{R}_+$, $\ell_2$ and $\ell_3$ 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 $\theta$. 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

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

$$
a_j(t) = \frac{p_jA_j(t)}{\sum_{h=1}^{J} p_hA_h(t)},
$$

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

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

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

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

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

$r_i(t)$ 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:

\[
\begin{align*}
& \text{decide in favour of } H_0 \text{ if } \\
& r_0(T) < \ell_1 p_0 + \sum_{j=1}^{J} p_j A_j(T)(\tau_j - \tau_{i^*(T)})^2 \\
& \text{decide in favour of } H_j (j \neq 0) \text{ if } j = i^*(T) \text{ and } \\
& r_0(T) > \ell_1 p_0 + \sum_{j=1}^{J} p_j A_j(T)(\tau_j - \tau_{i^*(T)})^2,
\end{align*}
\]

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 $\ell_2, \ell_3$ 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 $\ell_2$ and $\ell_3$ verify

\[ f_j(\theta) \ell_2(\theta) = a, \tag{3} \]
\[ f_j(\theta) \ell_3(\theta) = b \tag{4} \]

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 $\theta$ 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 $\ell_2$ and $\ell_3$ 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):

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

The integral between brackets is equal to $(2\pi)^{k/2} e^{\ell^2 \|\pi_j(t)\|^2/2}$ (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 $\ell_2$ instead of $\ell_3$, and we have:
Lemma 3.1. If (a), (b) and (2)–(4) hold:
- \( A_j(t) = b(2\pi)^{k/2}(1 - \|P_j(t)\|)^{-m/2}, \quad j = 1, \ldots, J \)
- \( r_0(t) = a(2\pi)^{k/2} \sum_{j=1}^{J} p_j (1 - \|P_j(t)\|)^{-m/2} \)

We can now prove the following proposition.

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

\[
\begin{cases}
\text{decide in favour of } H_0 \text{ if } \\
\sum_{j=1}^{J} p_j (1 - \|P_j(T)\|^2)^{-m/2} \left( 1 - \frac{b}{a} (\tau_j - \tau_{i^*}(T))^2 \right) < w_0.
\end{cases}
\]

otherwise decide in favour of \( H_i^* \) so that \( \tau_{i^*}(T) \) is the nearest \( \tau_i \) to

\[
\sum_{j=1}^{J} p_j (1 - \|P_j(t)\|^2)^{-m/2} \tau_j \\
\sum_{j=1}^{J} p_j (1 - \|P_j(t)\|^2)^{-m/2}.
\]

(5)

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

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

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 \( |\tau_j - \bar{\tau}(T)| \) reaches its minimum value over \( i \neq 0 \) for \( \tau_{i^*} \). 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 \( H_j \) 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 \( A \tau \), while it is better to select \( H_0 \) if the difference \( |\tau_i - \tau_j| \) is smaller than \( A \tau \), we are led to put \( (b/a) A^2 \tau = 1 \) 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 \( \alpha \), 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 \( \sigma \). Let \( w(A \tau, \alpha) \) be the critical value corresponding to the risk \( \alpha \); 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:

\[
\begin{align*}
\text{accept } H_0 & \text{ if } \sum_{j=1}^{J} p_j (1 - \|H_j(T)\|^2)^{-m/2} \left( 1 - \left( \frac{\tau_j - \tau_{\star}(T)}{\Delta \tau} \right)^2 \right) < w(\Delta \tau, \alpha) \\
\text{otherwise, decide in favour of } H_i^* \text{ so that } \tau_{\star}(T) \text{ is the nearest } \tau_i \\
\sum_{j=1}^{J} p_j (1 - \|H_j(T)\|^2)^{-m/2} \tau_j \\
\sum_{j=1}^{J} p_j (1 - \|H_j(T)\|^2)^{-m/2}.
\end{align*}
\]

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 $((\tau_j - \tau_{\star})/\Delta \tau)^2$ can be neglected and the first step of the rule becomes

\[
\text{decide in favour of } H_0 \text{ if } \sum_{j=1}^{J} p_j (1 - \|H_j(T)\|^2)^{-m/2} < w(\infty, \alpha).
\]

One can show that (6) is exactly the Bayes optimal rule that we obtain for testing $H_0$ against $\bigcup_{j=1}^{J} H_i$ 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 $\bigcup_{j=1}^{J} H_i$ (see Smith, 1975; Booth and Smith, 1982) is

\[
B_0 = \sum_{j=1}^{J} p_j B_{j0}, \text{ with } B_{j0} = \frac{\text{pr}(H_j|T)}{\text{pr}(H_0|T)} \left/ \frac{p_j}{p_0} \right..
\]

(The notation $\text{pr}(\cdot|\cdot)$ is a generic symbol for a conditional probability). Using Bayes’ theorem we can reexpress this in the form: $B_{j0} = \text{pr}(T|H_j)/\text{pr}(T|H_0)$, with $\text{pr}(T|H_j) = \int_{\Theta_j} g_j(t, \theta) \, dP_j(\theta)$ for $j = 0, 1$.

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

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

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 $B_{ij} = B_{0i} B_{0j}$.

(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 $(1 - \|H_j(T)\|^2)^{-m/2}$ 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 $\bigcup_{i=1}^{j} H_i$ (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 $\ell_1$, $\ell_2(.)$ and $\ell_3(.)$ are defined as in the previous section, $\ell'_1 > 0$ and $\ell_4(.)$ is a positive function such that

$$
\ell_4(\theta) f_j(\theta) = c \quad \text{for all } \theta \in \mathcal{O}_j \setminus \{0\}, j \neq 0. \quad (7)
$$

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

$$
c < a \quad (8)
$$

(even, in practice, $c$ should be much less than $a$). $\ell'_1$ (resp. $\ell_1$) is the cost of choosing $H_C$ (resp. $H_i, i \neq 0$) when $H_0$ holds. In practice, it is clear that $\ell'_1$ should be equal to or perhaps slightly less than $\ell_1$. A simple optimal procedure will be obtained by letting

$$
\ell'_1 = \ell_1 \frac{a - c}{a}. \quad (9)
$$

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

$$
r_C(t) = \ell'_1 p_0 + c(2\pi)^{j/2} \sum_{j-1}^{j} p_j (1 - \|H_j(t)\|^2)^{-m/2} = \ell'_1 p_0 + c(2\pi)^{j/2} \sum_{j-1}^{j} C_j(t),
$$
with \( C_j(t) = p_j(1 - \|P_I(t)\|^2)^{-m/2} \), while \( r_0(t) \) and \( r_i(t), i \neq 0, i \neq C, \) remain unchanged. We have immediately

\[
r_0(t) < r_C(t) \quad \text{if} \quad \sum_{j=1}^{J} C_j(t) < \frac{\ell'_1 p_0}{a - c} (2\pi)^{-k/2}
\]

and, with \( \tau^*(t) \) defined as above,

\[
r_0(t) < r_{\tau^*}(t) \quad \text{if} \quad \sum_{j=1}^{J} C_j(t) < \frac{\ell'_1 p_0}{a} (2\pi)^{-k/2} + \frac{b}{a} \sum_{j=1}^{J} C_j(t) (\tau_i - \tau_{\tau^*(t)})^2.
\]

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

\[
r_C(t) < r_{\tau^*}(t) \quad \text{if} \quad \sum_{j=1}^{J} C_j(t) \left(1 - \left(\frac{\tau_j - \tau_{\tau^*(t)}}{\Delta \tau}\right)^2\right) < \frac{\ell'_1 p_0}{a} (2\pi)^{-k/2}.
\]

Let \( \tau(T) \) 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:

\[
\begin{align*}
\text{accept } H_0 & \text{ if } \sum_{j=1}^{J} p_j (1 - \|P_I(T)\|^2)^{-m/2} < w_0 \\
\text{accept } H_C & \text{ if } \sum_{j=1}^{J} p_j (1 - \|P_I(T)\|^2)^{-m/2} > w_0 \\
& \quad \text{and } \sum_{j=1}^{J} p_j (1 - \|P_I(T)\|^2)^{-m/2} \left(1 - \left(\frac{\tau_j - \tau_{\tau^*(T)}}{\Delta \tau}\right)^2\right) < w_0 \\
\text{accept } H_{\tau^*} & \text{ so that } \tau_{\tau^*(T)} \text{ is the nearest } \tau_i \text{ to } \tau(T) \text{ if } \\
& \sum_{j=1}^{J} p_j (1 - \|P_I(T)\|^2)^{-m/2} \left(1 - \left(\frac{\tau_j - \tau_{\tau^*(T)}}{\Delta \tau}\right)^2\right) > w_0.
\end{align*}
\]

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 \( \alpha \) (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

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

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

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

In this example we suppose that the $\beta_i$'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 $n^{-1} \sum_{i=1}^{n} (1 - \|T_i(T)\|^2)^{-m/2} = 455.40$ and a critical value equal to 34.23. When $H_0$ is rejected, $\overline{T}(T) = 11.85$ 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 $n^{-2} \sum_{i=1}^{n} (1 - \|T_i(T)\|^2)^{-n/2}$ which is equal to 26.18
in this example. To estimate the change point, he uses the posterior probability of \( \tau \) rather than \( \tilde{\tau}(T) \).

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 \( \Delta \tau \leq 4 \) 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 \( \bigcup_{i=2}^{n-2} H_i \), which rests on maximal \( \pi(T) \). Farley and Hinich (1970) derived a locally most powerful procedure for testing \( H_0 \) against \( \bigcup_{i=2}^{n-2} H_i \). Their statistic is: \( z' T / \| \pi Q^{-1}(z) \| \), with

\[
\begin{align*}
0, \quad & \frac{\sum_{j=1}^{n-2} (x_2 - x_j)}{n}, \quad \ldots, \quad \frac{\sum_{j=1}^{n-2} (x_n - x_j)}{n}.
\end{align*}
\]

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 \( \beta_1 \) and \( \beta_2 \), 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 \( \beta^* \) 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.
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. \( n = 50, j_0 = 5, \mu^* = 1, \alpha = 0.05 \)

<table>
<thead>
<tr>
<th>Worsley</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject ( H_0 )</td>
<td>0.342</td>
<td>0.335</td>
<td>0.270</td>
</tr>
<tr>
<td>( H_5 )</td>
<td>0.104</td>
<td>0.031</td>
<td>0.031</td>
</tr>
<tr>
<td>( i \in [2,8] )</td>
<td>0.251</td>
<td>0.163</td>
<td>0.153</td>
</tr>
<tr>
<td>( i \in [2,11] )</td>
<td>0.275</td>
<td>0.231</td>
<td>0.202</td>
</tr>
<tr>
<td>( i &lt; 2, i &gt; 39 )</td>
<td>0.021</td>
<td>0.009</td>
<td>0.006</td>
</tr>
<tr>
<td>( H_C )</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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, $n = 50$, $j_0 = 25$, $\mu^* = 1$, $\alpha = 0.05$  

<table>
<thead>
<tr>
<th>Worsley</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject $H_0$</td>
<td>0.800</td>
<td>0.836</td>
<td>0.830</td>
</tr>
<tr>
<td>$H_{25}$</td>
<td>0.210</td>
<td>0.132</td>
<td>0.141</td>
</tr>
<tr>
<td>$i \in [22, 28]$</td>
<td>0.547</td>
<td>0.596</td>
<td>0.611</td>
</tr>
<tr>
<td>$i \in [17, 33]$</td>
<td>0.701</td>
<td>0.793</td>
<td>0.796</td>
</tr>
<tr>
<td>$i &lt; 11$, $i &gt; 39$</td>
<td>0.039</td>
<td>0.003</td>
<td>0.003</td>
</tr>
<tr>
<td>$H_C$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Worsley</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject $H_0$</td>
<td>0.296</td>
<td>0.316</td>
<td>0.310</td>
</tr>
<tr>
<td>$H_{15}$</td>
<td>0.108</td>
<td>0.029</td>
<td>0.031</td>
</tr>
<tr>
<td>$i \in [12, 18]$</td>
<td>0.196</td>
<td>0.132</td>
<td>0.150</td>
</tr>
<tr>
<td>$i \in [9, 24]$</td>
<td>0.229</td>
<td>0.240</td>
<td>0.248</td>
</tr>
<tr>
<td>$i &lt; 6$, $i &gt; 35$</td>
<td>0.031</td>
<td>0.014</td>
<td>0.012</td>
</tr>
<tr>
<td>$H_C$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Worsley</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject $H_0$</td>
<td>0.345</td>
<td>0.334</td>
<td>0.370</td>
</tr>
<tr>
<td>$H_{25}$</td>
<td>0.140</td>
<td>0.095</td>
<td>0.107</td>
</tr>
<tr>
<td>$i \in [22, 28]$</td>
<td>0.285</td>
<td>0.237</td>
<td>0.147</td>
</tr>
<tr>
<td>$i \in [17, 33]$</td>
<td>0.313</td>
<td>0.301</td>
<td>0.342</td>
</tr>
<tr>
<td>$i &lt; 10$, $i &gt; 35$</td>
<td>0.025</td>
<td>0.013</td>
<td>0.003</td>
</tr>
<tr>
<td>$H_C$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Worsley</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject $H_0$</td>
<td>0.670</td>
<td>0.707</td>
<td>0.647</td>
</tr>
<tr>
<td>$H_{35}$</td>
<td>0.360</td>
<td>0.082</td>
<td>0.082</td>
</tr>
<tr>
<td>$i \in [32, 38]$</td>
<td>0.498</td>
<td>0.327</td>
<td>0.342</td>
</tr>
<tr>
<td>$i \in [26, 41]$</td>
<td>0.556</td>
<td>0.585</td>
<td>0.561</td>
</tr>
<tr>
<td>$i &lt; 12$, $i &gt; 44$</td>
<td>0.039</td>
<td>0.004</td>
<td>0.005</td>
</tr>
<tr>
<td>$H_C$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 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


