Ann. Inst. Statist. Math.
Vol. 49, No. 4, 761–775 (1997)

CHOOSING A LINEAR MODEL WITH A RANDOM NUMBER OF CHANGE-POINTS AND OUTLIERS

HENRI CAUSSINUS AND FAOUZI LYAZRHI

Laboratoire de Statistique et Probabilités, UMR-CNRS 5583, Université Paul Sabatier, 118, Route de Narbonne, 31062 Toulouse Cedex, France

(Received July 27, 1995; revised July 30, 1996)

Abstract. The problem of determining a normal linear model with possible perturbations, viz. change-points and outliers, is formulated as a problem of testing multiple hypotheses, and a Bayes invariant optimal multi-decision procedure is provided for detecting at most k ($k > 1$) such perturbations. The asymptotic form of the procedure is a penalized log-likelihood procedure which does not depend on the loss function nor on the prior distribution of the shifts under fairly mild assumptions. The term which penalizes too large a number of changes (or outliers) arises mainly from realistic assumptions about their occurrence. It is different from the term which appears in Akaike's or Schwarz' criteria, although it is of the same order as the latter. Some concrete numerical examples are analyzed.

Key words and phrases: Akaike's criterion, Bayes decision procedure, change-point, invariance, maximal invariant, outliers, regression analysis, Schwarz' criterion.

1. Introduction

Let us consider n observations indexed by an ordered set, for example time, and depending on some given explanatory variables. They follow a linear model with a change-point at time τ if they can be described by a linear model up to time τ and another linear model after this time. There are two change-points if three different linear models are necessary to fit the data, the first from time 1 to τ_1, the second from τ_1 to τ_2 and the third from τ_2 to n. A similar definition holds for more change-points. In practice the positions of possible changes in the model are unknown. Determining if there are change-points and, in this case, where they are, can be formulated as a problem of testing multiple hypotheses, or a problem of model choice.

When it is assumed that there is only one change point, frequentist analyses are based on the likelihood ratio test. Page (1955), Gardner (1969), Hawkins (1977), Worsley (1979), consider a sequence of independent normal random variables and they test for no change in mean versus global existence of change.

Quandt (1958) was the first to propose a likelihood ratio procedure to test for separate regression lines. There is a substantial literature dealing with this case, for example Maronna and Yohai (1978), Kim and Siegmund (1989). Worsley (1983) extends the discussion to the multiple regression model.

Within the Bayesian framework, Chernoff and Zacks (1964) introduce the quasi Bayesian statistic to test for no change in a sequence of independent normal random variables versus a class of alternatives. The technique was extended by Farley and Hinich (1970) to the case of two separate lines, and by Jandhayala and MacNeill (1991) to the case of multiple regression model.

When the number of possible change-points is known in advance, say k, all the previous procedures can be straightforwardly generalized to the choice between no change and k changes. It is however more difficult to choose between models assuming, say, no or one or two changes, and assigning the location(s) of the change(s) in the latter cases: Smith (1980) proposes a stepwise procedure, Barry and Hartigan (1993) adopt product partition models, Kashiwagi (1991) applies the predictive log-likelihood to evaluate posterior distributions and Yao (1988) uses a version of the Schwarz (1978) criterion to estimate the number of change-points.

The detection of outliers in a linear model is very similar to the change-point problem from a practical as well as a theoretical viewpoint since an outlier is basically defined by its location. As in the previous problem, finding outliers requires a procedure to choose between no outlier and outliers at given locations. Here again, most of the previous papers do not address the question of detecting at most k outliers and estimating their locations (or they merely propose a heuristic stepwise procedure: Freeman (1980), Pettit (1992) and Alexander (1993)), but only the question of detecting exactly k outliers. A global heuristic procedure for at most k outliers can however be found in Caussinus and Vaillant (1985).

In this paper, we consider the problem of modelling the data by means of a linear model with at most k change-points and outliers by a global (non stepwise) multi-decision procedure. An invariant Bayes optimal solution is provided within a decision theoretic framework for some loss function and a prior distribution for (i) the changes in the parameters of the linear models, (ii) the number and the location of the change-points and outliers. Realistic assumptions concerning the latter point are the main feature of our way of dealing with the problem and the key aspect of the proposed procedure (for similar assumptions in a simpler context, see Yao (1984)). In fact, the problem is a special case of selection of variables in a linear model, but the actual specific situation leads to a particular choice of the term which penalizes too large a number of parameters. In particular, in our problem the number of competing models increases with n (such a situation has seldom been considered in the literature, with some notable exceptions, for example Hannan and Quinn (1979) and Shibata (1981)). Although the frameworks are fairly different, it turns out that our penalty term is somewhat similar to the term which appears in the Schwarz' criterion. They are however different and both criteria may lead to different decisions.

A similar derivation can be used for the cases where only outliers or only change-points are taken into consideration. However, looking at the same time for both kinds of perturbations of the null model does not result in much more

complication theoretically, due to the similarities of both situations, and seems realistic in many cases, especially when the change-point problem is the leading one (see Smith and West (1983), Taplin and Raftery (1994), the discussion of his example by Worsley (1983), and the results of our own examples in Section 5).

The general problem is formulated in Section 2 and an optimal invariant Bayes multi-decision procedure is derived in Section 3. This procedure depends heavily on the prior probabilities of occurrence of an outlier or a change-point at each possible location. A realistic model for the occurrence of such events is introduced in Section 4. The resulting optimal procedure is derived and a simple explicit approximation of this procedure is provided under fairly general and realistic assumptions concerning the prior distribution of the shift in the regression parameters. Finally, Section 5 is devoted to some examples.

2. Notation and framework

2.1 *Generalities*

We consider n random variables Y_i $(i = 1, \ldots, n)$, we denote by Y the column vector of the Y_i's and we assume that the probability distribution of Y is n-dimensional normal $N_n(\mu, \sigma^2 I_n)$, where I_n denotes the $n \times n$ unit matrix and σ is a positive unknown parameter. The scalar product on $\mathbb{R}^n$ is denoted by $\langle \cdot, \cdot \rangle$ and the corresponding norm by $\| \cdot \|$. The set of vectors $(e_1, \ldots, e_n)$ is the canonical basis of $\mathbb{R}^n$, that is e_j is the $n \times 1$ matrix whose all elements are zero but the j-th which is equal to one.

The various models differ from one another in the μ space. Let Q be a given q-dimensional linear subspace of $\mathbb{R}^n$, the basic model, that is the null hypothesis $H_\emptyset$, is defined by:

$$H_\emptyset : \mu \in Q$$

or, equivalently, $\mu = X\beta$, where β is an unknown vector of q parameters and X is a full rank $n \times q$ matrix whose columns span Q.

Let Q_J 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_J is defined by:

$$H_J : \mu \in Q \oplus Q_J$$

where $\oplus$ denotes the direct sum of two linear subspaces. If the dimension of Q_J is q_J, Q_J is usually spanned by the columns of an $n \times q_J$ matrix $\Pi_{Q^\perp} X_J$, where $\Pi_{Q^\perp}$ denotes the orthogonal projector on $Q^\perp$. Then μ can be written as $\mu = X\beta + \Pi_{Q^\perp} X_J \beta_J$.

All the problems we deal with can be formulated as the choice of a model from a set of hypothetical models H_J (including $H_\emptyset$ as the special case where Q_J reduces to $\{0\}$).

2.2 *Change-points and outliers*

The previous framework is well adapted to the problem of outlier(s) in the mean or change-point(s) provided the changes are assumed to occur at observed points (if this is not the case the problems are quite different: see, for example Hinkley (1971)). The hypothesis $H_\emptyset$ is still associated with the basic (non perturbated) model. If we consider one kind of perturbation of the basic model (e.g. change-points), J is then a subset of $\{1, 2, \ldots, n\}$ corresponding to their locations. If two kinds of perturbations are considered (change-points and outliers) J is a pair of subsets of $\{1, 2, \ldots, n\}$, say $J = (J_1, J_2)$, where J_1 corresponds to the location of the change-points and J_2 to the location of the outliers. One or both of these subsets can be empty: for instance, $J_2 = \emptyset$ means that there is no outlier and $J_1 = J_2 = \emptyset$ corresponds to the null hypothesis $H_\emptyset$. Let us consider some examples.

Example 1. One outlier in a linear model

We can set that Y_j is an outlier by writing:

$$H_j : \mathrm{E}(Y) = X\beta + \alpha e_j, \qquad \alpha \neq 0$$

or $\mathrm{E}(Y) \in Q \oplus Q_j$, where Q_j is spanned by $\Pi_{Q^\perp}(e_j)$. If any observation may be an outlier of the foregoing kind, we are led to consider the set of hypotheses $H_\emptyset$ and H_j $(j = 1, \ldots, n)$. Here $J_1 = \emptyset$ and $J_2 = \{j\}$, hence the notation $H_J = H_j$.

Example 2. One change-point in a multiple regression model

Let $X = [x_1, \ldots, x_n]'$ where x_j is a $q \times 1$ vector and $X_j = [0, \ldots, 0, x_{j+1}, \ldots, x_n]'$. The hypothesis H_j that one change-point occurs after the j-th observation is:

$$H_j : \mathrm{E}(Y) = X\beta + X_j\beta^*, \qquad \beta^* \neq 0$$

In this case Q_j is spanned by the columns of $\Pi_{Q^\perp} X_j$, $J_1 = \{j\}$ and $J_2 = \emptyset$.

Example 3. Outlier and change-point in a simple regression

Let $X = [x, \mathbb{1}]$, where $x = [x_1, \ldots, x_n]'$ is a column vector and $\mathbb{1} = [1, \ldots, 1]'$. A switch in regression of Y on x after the k-th observation and an outlier at the j-th observation $(j > k)$ may be written as:

$$\begin{aligned}
\mathrm{E}(Y_i) &= \alpha_1 + \beta_1 x_i && \text{for} \quad i = 1, \ldots, k, \\
&= \alpha_1 + \beta_1 x_i + \alpha_2 + \beta_2 x_i && \text{for} \quad i = k+1, \ldots, j-1, j+1, \ldots, n, \\
&= \alpha_1 + \beta_1 x_i + \alpha_2 + \beta_2 x_i + \lambda && \text{for} \quad i = j.
\end{aligned}$$

In this case Q is spanned by x and $\mathbb{1}$, $J_1 = \{k\}$, $J_2 = \{j\}$, $Q_J = Q_{(\{k\},\{j\})}$ is spanned by the three vectors $\Pi_{Q^\perp}(e_j)$, $\Pi_{Q^\perp}(\mathbb{1}_k)$ and $\Pi_{Q^\perp}(x_k^*)$ where $x_k^* = [0, \ldots, 0, x_{k+1}, \ldots, x_n]'$, and the first k elements of $\mathbb{1}_k$ are 0 while the others are 1. In general $q = 2$ and $q_J = 3$.

Example 4. One change-point in a simple linear regression constrained by continuity.

Suppose that a simple regression function can change but stays continuous at x_j. Then X is a $n \times 2$ matrix $[x, \mathbb{1}]$, Q_j is spanned by $\Pi_{Q^\perp}(a_j)$ with $a_j = [0, \ldots, 0, x_{j+1} - x_j, \ldots, x_n - x_j]'$, $J_1 = \{j\}$, $J_2 = \emptyset$, $q_J = 1$, and the several hypotheses are

$$H_\emptyset : \mathrm{E}(Y) = \beta_1 x + \beta_2 \mathbb{1}$$
$$H_j : \mathrm{E}(Y) = \beta_1 x + \beta_2 \mathbb{1} + \beta^* a_j, \qquad \beta^* \neq 0, \ 2 \leq j \leq n-2.$$

In general, if the number of change-points (resp. outliers) is denoted by $|J_1|$ (resp. $|J_2|$), then

$$q_J = |J_1| q + |J_2|. \tag{2.1}$$

If the changes, however, do not concern all the parameters (e.g. in Example 4 owing to the continuity assumption) then $q_J = |J_1| q^* + |J_2|$ with $q^* < q$.

2.3 *Invariance*

The statistical model as well as the different hypotheses are invariant under the group of transformations $\{y \to ay + b, a > 0, b \in Q\}$ and a maximal invariant is the vector of normed residuals $T = \frac{\Pi_{Q^\perp}(Y)}{\|\Pi_{Q^\perp}(Y)\|}$. We shall therefore restrict attention to invariant procedures which leads to performing the analysis through T. The distribution of T in the canonical basis does not have an easily handled distribution even under $H_\emptyset$. Actually, T belongs to the unit sphere $S_{Q^\perp}$ of $Q^\perp$ and its distribution is the uniform distribution $U_{Q^\perp}$ on $S_{Q^\perp}$ if $H_\emptyset$ holds. Under the alternative H_J, i.e. $\mu \in Q \oplus Q_J$, the distribution of T has a density g_J with respect to $U_{Q^\perp}$ given by (see Caussinus and Vaillant (1985)):

$$q_J(t, \theta) = \frac{1}{2^{m/2-1}\Gamma\left(\frac{m}{2}\right)} e^{-\|\theta\|^2/2} h_m(\langle t, \theta \rangle), \qquad t \in S_{Q^\perp} \tag{2.2}$$

where: $m = n - q$, $h_m(u) = \int_0^\infty e^{uv} e^{-v^2/2} v^{m-1} dv$ and $\theta = \frac{\Pi_{Q_J}(\mu)}{\sigma}$. Note that g_J depends only on μ and σ through the parameter θ.

3. Optimal multi-decision rule: general results

Let us consider a set of hypotheses H_J, where $J \in \mathcal{J}$, within the general framework of Subsection 2.1. The hypothesis $H_\emptyset$ is assumed to be an element of this set, that is $\emptyset \in \mathcal{J}$. The number of elements of the set $\mathcal{J}$ will be denoted by $|\mathcal{J}|$.

Prior distribution

- The prior probability of H_J is p_J ($\sum_{J \in \mathcal{J}} p_J = 1$).
- The conditional prior probability of θ given H_J is P_J, a probability distribution on Q_J which will be assumed to have a density function f_J for $J \neq \emptyset$. (The probability $P_\emptyset$ is the Dirac measure on $\{0\}$.)

Loss function

The loss function for selecting H_I when the true hypothesis is H_J with the value θ of the parameter is:

$$\ell(I,J,\theta) = \begin{cases} 0, & \text{if } I = J \\ \ell, & \text{if } I \neq J,\ J = \emptyset \\ \ell_J(\theta), & \text{if } I \neq J,\ J \neq \emptyset \end{cases}$$

where ℓ and ℓ_J are positive.

We shall consider below one of the following assumptions:

(3.1) $\quad \ell_J(\theta) f_J(\theta) = 1 \quad$ for all $J \in \mathcal{J} \setminus \{\emptyset\} \quad$ and for all $\theta \in Q_J \setminus \{0\}$

and

(3.2) $\quad$ there exists $(a,b) \in \mathbb{R}^2$ such that, $\quad 0 < a \leq \ell_J(\theta) f_J(\theta) \leq b$
for all $J \in \mathcal{J} \setminus \{\emptyset\} \quad$ and $\quad \theta \in Q_J \setminus \{0\}$.

Note that (3.1) holds in the special case where f_J is constant (P_J is a vague prior) while ℓ_J is constant ("simple" loss function). The latter assumptions are not however necessary to derive Proposition 3.2 below. Moreover, it is worth noticing that (3.1) is only introduced to make the presentation easier since the main practical result (Proposition 4.1) rests basically on the more general assumption (3.2).

Let $d(T)$ be an invariant multi-decision procedure, that is $d = (d_J)_{J \in \mathcal{J}}$ is a measurable function from $S_{Q^\perp}$ to $[0,1]^{|\mathcal{J}|}$ with $\sum_{J \in \mathcal{J}} d_J(T) = 1$, where $d_J(T)$ is the probability of selecting H_J for given T. Using the previous notation and results, the Bayes risk for this multi-decision rule is given by

$$R = \sum_{I \in \mathcal{J}} \int d_I(t) r_I(t) dU_{Q^\perp}(t),$$

where $r_I(t) = \sum_{J \in \mathcal{J}} p_J \int_{Q_J} g_J(t,\theta) \ell(I,J,\theta) dP_J(\theta)$.

PROPOSITION 3.1. *For $\ell(\cdot,\cdot,\cdot)$ and P_J defined above, an invariant Bayes optimal multi-decision rule is: select H_{J^*} if $a_{J^*}(T)$ is the maximum value of $a_J(T)$ for $J \in \mathcal{J}$, with $a_\emptyset(T) = p_\emptyset \ell$ and $a_J(T) = p_J \int_{Q_J} \ell_J(\theta) g_J(T,\theta) dP_J(\theta)$ for $J \neq \emptyset$.*

PROOF. It is well known (Ferguson (1967)) that a multi-decision rule d^* minimizing R is defined as follows:

$$\text{select } H_J \quad \text{if} \quad r_J < \min_{I \neq J} r_I.$$

Define $g_\emptyset$ by $g_\emptyset(t,0) = 1$ for all $t \in S_{Q^\perp}$ and $\ell_\emptyset(\theta) = \ell$ for all $\theta \in Q_J$. By using the expression of $\ell(\cdot,\cdot,\cdot)$, we get:

$$r_\emptyset(T) = \sum_{J \in \mathcal{J}} p_J \int_{Q_J} \ell_J(\theta) g_J(T,\theta) dP_J(\theta) - p_\emptyset \ell, \quad \text{and}$$

$$r_I(T) = \sum_{J \in \mathcal{J}} p_J \int_{Q_J} \ell_J(\theta) g_J(T,\theta) dP_J(\theta) - p_I \int_{Q_I} \ell_I(\theta) g_I(T,\theta) dP_I(\theta), \quad I \neq \emptyset.$$

Since the first term on the right side of $r_I(T)$ is independent of I, the proposition follows easily.

Remark 1. A less clear but more rigorous statement of Proposition 3.1 would be: any multi-decision rule d^* for which R reaches its minimum value is such that

$$d_I^*(T) = 0 \quad \text{if there exists } J \in \mathcal{J} \quad \text{such that } a_J(T) > a_I(T).$$

The formulation in Proposition 3.1 and analogous formulations below are actually valid only if tied values of the $a_J(T)$ are neglected, which is possible here since tied values arise with null probability.

PROPOSITION 3.2. *If assumption* (3.1) *holds and* $\ell(\cdot,\cdot,\cdot)$ *is the loss function defined above, an invariant Bayes optimal multi-decision rule is: select* H_{J^*} *such that*:

$$J^* = \arg\max_J (p_\emptyset \ell; p_J(1 - \|\Pi_{Q_J}(T)\|^2)^{-(n-q)/2}(2\pi)^{q_J/2}, \text{ for } J \neq \emptyset)$$

PROOF. By taking (3.1) into account, $a_J(T)$ (as defined in Proposition 3.1) becomes for $J \neq \emptyset$:

$$a_J(T) = p_J \int_{Q_J} g_J(T,\theta)\ell_J(\theta)f_J(\theta)d\theta = p_J \int_{Q_J} g_J(T,\theta)d\theta.$$

Now, for $v > 0$, $t \in S_{Q^\perp}$, $\theta \in Q_J \subset Q^\perp$, we have $\langle \theta, t\rangle = \langle \theta, \Pi_{Q_J}(t)\rangle$ and $\|\theta\|^2 - 2v\langle\theta,t\rangle = \|\theta - v\Pi_{Q_J}(t)\|^2 - v^2\|\Pi_{Q_J}(t)\|^2$. Hence, we get by using (2.2)

$$\int_{Q_J} g_J(T,\theta)d\theta = \frac{1}{2^{m/2-1}\Gamma\left(\frac{m}{2}\right)} \int_0^\infty \left[\int_{Q_J} e^{-\|\theta - v\Pi_{Q_J}(T)\|^2/2} d\theta\right] e^{-v^2(1-\|\Pi_{Q_J}(T)\|^2)/2} v^{m-1} dv$$

with $m = n - q$.

The integral between brackets is equal to $(2\pi)^{q_J/2}$. By integrating then over v, we obtain:

$$\int_{Q_J} g_J(T,\theta)d\theta = (2\pi)^{q_J/2}(1 - \|\Pi_{Q_J}(T)\|^2)^{-m/2}, \quad \text{for} \quad J \neq \emptyset. \tag{3.3}$$

Proposition 3.2 is then deduced from Proposition 3.1.

4. Optimal multi-decision rule: change-points and outliers

From the practical point of view, Proposition 3.2 suffers from (at least) two shortcomings:

(i) the prior probabilities p_J are generally unknown,

(ii) assumption (3.1) is quite restrictive.

This section is devoted to the derivation of an asymptotically optimal workable procedure under a more general assumption than (3.1), with a suitable choice of the prior probabilities p_J for the special case of interest described in Subsection 2.2. A convenient model for the p_J's arises from the specific meaning of the hypotheses H_J in the problem under consideration: an hypothesis corresponds to the place(s) where some events (change-points or outliers) happen. Assume first that there is one kind of perturbation (e.g. change-point) and that they may happen independently with probability p, which seems to be a sensible model in most practical cases. Then $p_J = p^{|J|}(1 \quad p)^{n-|J|}$.

Furthermore, assume that the maximum value of $|J|$ is k, a fixed positive integer. The probability of H_J given $|J| \leq k$ is then:

$$p_J = \lambda p^{|J|}(1-p)^{n-|J|} \quad \text{for} \quad 0 \leq |J| \leq k \tag{4.1}$$

$$\text{with } \lambda \quad \text{such that} \quad \sum_{|J|=0}^{k} p_J = 1.$$

If there are two possible kinds of perturbations, with probabilities p_1 and p_2 respectively, then (4.1) becomes by using the same arguments as above:

$$p_J = p_{(J_1,J_2)} = \lambda p_1^{|J_1|} p_2^{|J_2|}(1-p_1-p_2)^{n-|J_1|-|J_2|} \quad \text{for} \quad 0 \leq |J| \leq k, \tag{4.2}$$

where $|J| = |J_1| + |J_2|$ and λ is such that $\sum_{|J|=0}^{k} p_J = 1$.

Finally, we assume that, when $n \to \infty$:

$$np_1 \to r_1, \qquad np_2 \to r_2, \qquad r_1 > 0, \qquad r_2 > 0. \tag{4.3}$$

For large n, the number of change-points and the number of outliers are two independent Poisson random variables whose sum is truncated at k. Assuming that the total number of perturbations is bounded by k may be considered as unrealistic. In practice, it seems however that, if the number of perturbations may be very large, the model we are fitting will not be considered.

As for assumption (3.1), it will be replaced by the more general assumption (3.2) to obtain the following result.

PROPOSITION 4.1. *If assumptions* (3.2), (4.2) *and* (4.3) *hold, an invariant Bayes optimal multi-decision rule is: select* H_{J^*} *such that:*

$$J^* = \arg\min_J \left(0; \log(1 - \|\Pi_{Q_J}(T)\|^2) + \frac{2|J|}{m}\log n + \mathcal{O}\left(\frac{1}{n}\right), J \neq \emptyset\right).$$

(In the expression above, 0 is associated to $J = \emptyset$ which means that $H_\emptyset$ is selected if all the terms associated with $J \neq \emptyset$ are positive.)

PROOF. From condition (3.2) and result (3.3) we have:

$$ap_J(2\pi)^{q_J/2}(1 - \|\Pi_{Q_J}(T)\|^2)^{-m/2} \leq a_J(T) \leq bp_J(2\pi)^{q_J/2}(1 - \|\Pi_{Q_J}(T)\|^2)^{-m/2}.$$

From (4.2) and (4.3), it is clear that, for large n:

$$\frac{2}{m}\log p_J = -\frac{2|J|\log n}{m} + \mathcal{O}\left(\frac{1}{n}\right), \qquad \text{for} \quad 0 < |J| \leq k.$$

Thus, for large n and $J \neq \emptyset$

$$\begin{aligned} &-\log(1 - \|\Pi_{Q_J}(T)\|^2) - \frac{2|J|}{m}\log n + \mathcal{O}\left(\frac{1}{n}\right) \\ &\quad \leq \frac{2}{m}\log a_J(T) \\ &\quad < -\log(1 - \|\Pi_{Q_J}(T)\|^2) - \frac{2|J|}{m}\log n + \mathcal{O}\left(\frac{1}{n}\right). \end{aligned}$$

Moreover, since $a_\emptyset(T) = p_\emptyset \ell$, we have $\frac{2}{m}\log a_\emptyset(T) = \mathcal{O}(\frac{1}{n})$.

Proposition 4.1 is then derived from Proposition 3.1.

It is worth noticing that the unknown values a, b, ℓ, r_1 and r_2 occur only in the rest $\mathcal{O}(\frac{1}{n})$. The rest depends also on J via $|J_1|$, $|J_2|$ and q_J (note that q_J does not depend on n) but it is independent of T, being thus non stochastic.

A feasible multi-decision procedure

For a practical use, we propose the following procedure: select H_{J^*} such that:

$$J^* = \arg\min_J \left(0; \log(1 - \|\Pi_{Q_J}(T)\|^2) + \frac{2|J|}{m}\log n, J \neq \emptyset\right). \tag{4.4}$$

Procedure (4.4) can be easily carried out while being close, for large n, to the procedure given by Proposition 4.1 which is optimal under fairly general assumptions.

Remark 2. Procedure (4.4) turns out to be a log-likelihood procedure (based on the original data) with the penalty term $2|J|\frac{\log n}{m}$. A similar term appears in Schwarz' criterion, viz. $q_J\frac{\log n}{n}$, to penalize too large a number of perturbations (in general, q_J is given by (2.1)). Note however that both terms are not equal even if they are of the same order with respect to n. It is also worth noticing that our derivation is similar to the one based on Bayes factors, although it is a little more general: in fact, if $\ell_J(\theta)$ is assumed to be constant, then $a_J(T)$ is proportional to the probability of J given T. The corresponding optimal procedure is therefore equivalent to the one based on Bayes factors. The latter procedure

has been investigated by Smith and Spiegelhalter (1980) who find various penalty terms according to the prior for θ. On the contrary, under assumption (3.2), our criterion does not depend on the prior for θ, but it depends heavily on assumption (3.2) concerning the p_J's, which arises from the specific problem we are dealing with. (Another characteristic of our framework is that the hypotheses are not nested, but this does not give rise to serious differences: see, e.g. Leonard (1982), for a discussion with respect to the Schwarz' criterion.)

Remark 3. In the derivation of Proposition 4.1 (and thus the derivation of the procedure (4.4)), we assume implicitly that all the hypotheses H_J such that $0 \leq |J| \leq k$ are taken into consideration, that is all subsets $J_1 \cup J_2$ of $\{1, 2, \ldots, n\}$ with $|J_1 \cup J_2| < k$ are possible perturbation points. In practice, this is not always the case, if only since the detection of two change-points in a regression model is not possible if these points are too close. This reduces the number of hypotheses H_J under consideration, but it is easy to show that the asymptotic results above are still valid. On the other hand, the prior probabilities (4.1) and (4.2) are obtained by assuming that a change-point (or outlier) may happen with constant probability p_1 (or p_2) at each observation. This assumption can be unrealistic chiefly for the change-points, in particular if the Y_i's are not observed at regular "distances" when $i = 1, 2, \ldots, n$ (an example is provided by the men's olympic performances in Section 5 on account of the missing war years). The asymptotic results remain however valid if the constant probabilities p_1 and p_2 are replaced by p_{1i} and p_{2i} $(i = 1, \ldots, n)$, $\frac{p_{1i}}{p_1}$ and $\frac{p_{2i}}{p_2}$ belong to an interval $[\alpha, \beta]$ $(0 < \alpha < \beta < +\infty)$ for any i and (4.3) holds.

5. Applications

In this section, the behaviour of procedure (4.4) is appraised on several examples. In all the examples, the observations are ordered in time. The hypotheses under consideration are all those H_J such that $0 \leq |J| \leq k$ with $|J| = |J_1| + |J_2|$, except the hypotheses for which the regression parameters could not be identified: for example, two change-points are assumed to be separated by at least q observations. The value of k has to be chosen somewhat arbitrarily: in practice, we have carried on the computation up to the value of k such that the minimum of $C(J)$ over $|J| = k$ started increasing. The results obtained by procedure (4.4) are then compared to the ones given by Akaike's (1973) and Schwarz' (1978) criteria based on the original data, that is our procedure where the penalty term $2|J|\frac{\log n}{m}$ is changed into $2\frac{q_J}{n}$ (Akaike) or $q_J\frac{\log n}{n}$ (Schwarz).

For computation, the various models are defined in matrix form by $\mu = X\beta + \Pi_{Q^\perp} X_J \beta_J$ (see Subsection 2.1) and $\|\Pi_{Q_J}(T)\|^2$ is then:

$$\|\Pi_{Q_J}(T)\|^2 = \frac{R' X_J (X_J'(I_n - X(X'X)^{-1}X')X_J)^{-1} X_J' R}{\|R\|^2}$$

where R is the vector of residuals, that is $R = Y - X(X'X)^{-1}X'Y$ and $\|R\|^2 = R'R$.

We denote $C_n(J) = \log(1 - \|\Pi_{Q_J}(T)\|^2) + \frac{2|J|}{m}\log n$, with $C_n(\emptyset) = 0$.

5.1 *Gross domestic product in U.S.A.*

The data described in Maddala ((1977), Table 10.3, p. 196) concern the gross domestic product and labor and capital input in the United States for the years 1929–1967. The logarithm Y_i of the gross domestic product of year $1928+i$ is first modelled as a linear function of the logarithms of the labor input, x_{1i}, and the capital input, x_{2i}. The basic model is therefore

$$H_\emptyset : \mathrm{E}(Y_i) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i}, \qquad i = 1, \ldots, 39.$$

It is suspected, however, that the parameters of the regression may have changed after an unknown time.

These data were reanalyzed by Worsley (1983), who used the likelihood ratio statistic $\max_J \|\Pi_{Q_J}(T)\|^2$ to look for one change-point. He pointed out that one change-point occurs after 1942. He looked next for further changes in the two subsequences that are formed by the first split. He found that the first subsequence 1929–1942 contained no significant change, while the second subsequence 1943–1967 contained one change-point after 1946. He suspected also that the data might contain outliers.

The procedure (4.4) has been used to look for change-points (with constraint of continuity which seems natural in this example) and outliers. The constraint of continuity generalizes the one introduced in Example 4, Subsection 2.2. It means that if j is a change-point, $\mathbb{E}(Y_j)$ takes the same value whether it is computed with the coefficients of the regression before j or after j.

In this case $n = 39$, $q = 3$ and we have set $|J| \leq k - 4$. We give some numerical values of $C(J)$ including the minimum value obtained for each integer $|J|$ up to five. For better readability, the indices $i = 1, 2, \ldots$ have been replaced by the years $(1929, \ldots, 1967)$ in the definition of J_1 and J_2.

$$\begin{aligned}
&|J| = 1 : C(J) = -0.336 \quad \text{for } J_1 = \{1945\} \quad \text{and} \quad J_2 = \emptyset,\\
&|J| = 2 : C(J) = -0.711 \quad \text{for } J_1 = \{1944, 1948\} \quad \text{and} \quad J_2 = \emptyset,\\
&|J| = 3 : C(J) = -1.136 \quad \text{for } J_1 = \{1938, 1944, 1948\} \quad \text{and} \quad J_2 = \emptyset,\\
&|J| = 3 : C(J) = -0.928 \quad \text{for } J_1 = \{1938\} \quad \text{and} \quad J_2 = \{1945, 1946\},\\
&|J| = 4 : C(J) = -1.107 \quad \text{for } J_1 = \{1938, 1944, 1948, 1952\} \quad \text{and} \quad J_2 = \emptyset,\\
&|J| = 4 : C(J) = -1.076 \quad \text{for } J_1 = \{1938, 1944, 1948\} \quad \text{and} \quad J_2 = \{1951\}.
\end{aligned}$$

It is thus decided that there are three change-points, one after 1938, one after 1944, one after 1948 and no outlier.

Akaike's and Schwarz' criteria both decide the same change-points.

5.2 *Men's olympic performances*

We consider the data given in Hand *et al.* ((1994), tables 300, 311) of men's Olympic performances in the pole vault and long jump from 1896 to 1988 (completed up to 1992) and in 200 metres finals from 1900 to 1988 (completed up to 1992). There were no Olympic games in 1916, 1940 and 1944. The basic model is:

$$H_\emptyset : \mathrm{E}(Y_i) = \alpha + \beta x_i$$

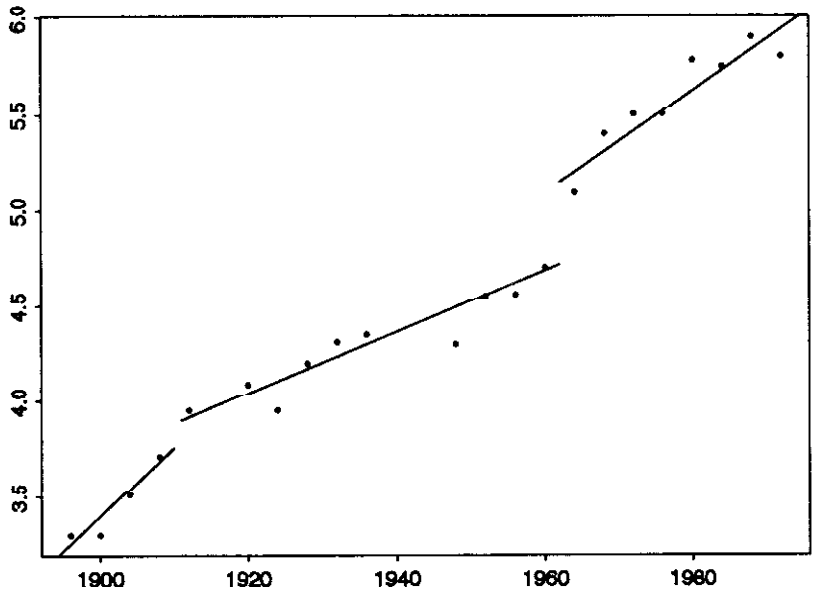

Fig. 1. Men's olympic performances in the pole vault.

where Y_i is the performance on year x_i ($i = 1, \ldots, 21$ or 22). This is obviously a naive model for several reasons (for example, it cannot be expected that the performances increase indefinitely). However, representing the data by means of a set of linear regressions may be adequate within short periods especially if the change-points may be easily interpreted. Hence, the procedure proposed in this paper may turn out to be of interest.

The data and the selected model are displayed in Figs. 1, 2 and 3.

5.2.1 *Pole vault*

In this case, Y is the height (in metres) jumped by the successive winners of the Olympic pole vault, $n = 22$, $|J| < 3$.

Our procedure decides that two change-points have occurred, one after 1908 and another after 1960, with $C(J) = -0.868$ for $J_1 = \{1908, 1960\}$ and $J_2 = \emptyset$ (the second change-point corresponds to a sudden, or at least very rapid, improvement in the equipment).

The smallest value of $C(J)$ for $|J| > 2$ is -0.850 obtained for $J_1 = \{1908, 1956, 1964\}$ and $J_2 = \emptyset$ (a more complicated and less interpretable model than the previous one), while the smallest value of $C(J)$ for $|J| = 1$ is -0.505 for $J_1 = \{1960\}$ and $J_2 = \emptyset$.

In this example, Akaike's and Schwarz' criteria decide that there are two change-points (1908, 1960) and one outlier (1992) (this outlier seems to have little practical support).

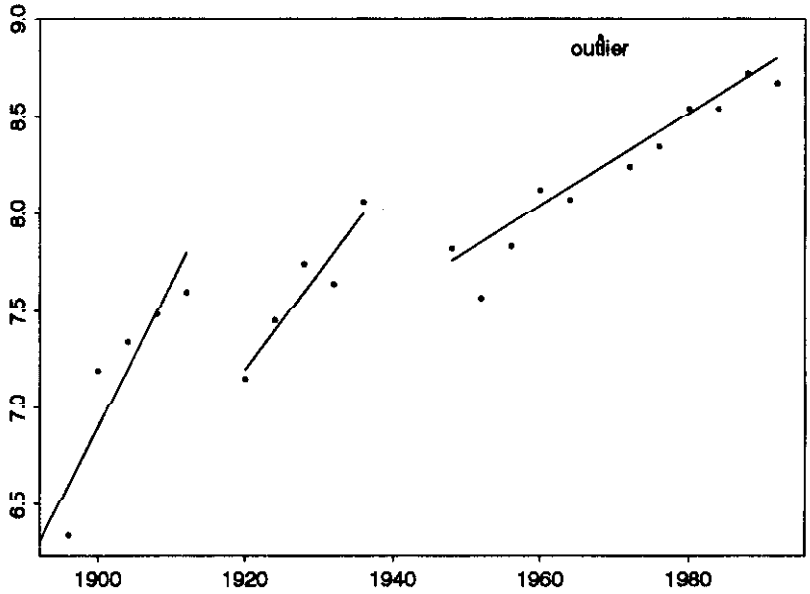

Fig. 2. Men's olympic performances in the long jump.

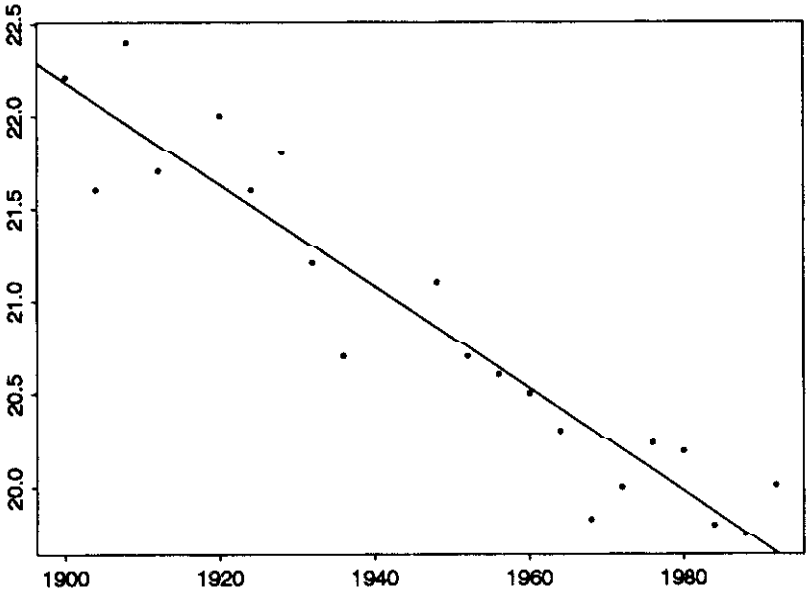

Fig. 3. Men's olympic performances in the 200 m finals.

5.2.2 *Long jump*

Here, Y is the distance (in metres) jumped by the successive winners of the Olympic long jump.

In this case $n = 22$, $|J| \leq 4$. The procedure decides that there are two change-points (1912 and 1936) and one outlier (1968), with $C(J) = -0.526$. The two change-points correspond to the two war periods, while the outlying performance in 1968 (Mexico) is well known.

The other smallest values of $C(J)$ for $|J| \leq 4$ are:

$$|J| = 1 : C(J) = -0.085 \quad \text{for } J_1 = \emptyset \quad \text{and} \quad J_2 = \{1896\},$$
$$|J| = 2 : C(J) = -0.280 \quad \text{for } J_1 = \{1900\} \quad \text{and} \quad J_2 = \{1968\},$$
$$|J| = 4 : C(J) = -0.489 \quad \text{for } J_1 = \{1908, 1936\} \quad \text{and} \quad J_2 = \{1896, 1968\}.$$

Note that a stepwise procedure would have been misleading.

In this example, Akaike's criterion provides the same result as ours and Schwarz' criterion decides that there are three outliers (1896, 1952, 1968) and no change-point.

5.2.3 *200 m*

Y is the time in seconds of the men's Olympic 200 m finals, $n = 21$, $|J| \leq 3$. In this case the least value of $C(J)$ over $|J| > 0$ is 0.10. The basic model without any perturbation is thus selected.

On the contrary, Akaike's and Schwarz' criteria both decide that there are three outliers (1904, 1936, 1968).

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *2nd International Symposium on Information Theory*, 267–281, Akademiai Kiado, Budapest.

Alexander, W. P. (1993). Testing the means of independent normal random variables, *Comput. Statist. Data Anal.*, **16**, 1–10.

Barry, D. and Hartigan, J. A. (1993). A Bayesian analysis for change point problems, *J. Amer. Statist. Assoc.*, **88**, 421, 309–319.

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

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

Ferguson, T. S. (1967). *Mathematical Statistics: a Decision Theoretic Approach*, Academic Press, New York and London.

Freeman, P. R. (1980). On the number of outliers in data from a linear model (with discussion), *Bayesian Statistics* (eds. J. M. Bernardo *et al.*), 349–365, University Press, Valencia.

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

Hand, D. J., Daly, F., Lunn, A. D., Mc Conway, K. J. and Ostrowski, E. (1994). *A Handbook of Small Data Sets*, Chapman & Hall, London.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression, *J. Roy. Statist. Soc. Ser. B*, **41**, 190–195.

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

Hinkley, D. J. (1971). Inference in two-phase regression, *J. Amer. Statist. Assoc.*, **66**, 736–743.

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

Kashiwagi, N. (1991). Bayesian detection of structural changes, *Ann. Inst. Statist. Math.*, **43**, 77–93.

Kim, H. and Siegmund, D. (1989). The likelihood ratio test for a change-point in simple linear regression, *Biometrika*, **76** (3), 409–423.

Leonard, T. (1982). Comment on M. Lejeune and G. D. Faulkenberry, "A simple predictive density function", *J. Amer. Statist. Assoc.*, **77**, 657–658.

Maddala, G. S. (1977). *Econometrics*, McGraw-Hill, Singapore.

Maronna, R. and Yohai, V. (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.

Pettit, L. I. (1992). Bayes factors for outlier models using the device of imaginary observations, *J. Amer. Statist. Assoc.*, **87**, 541–545.

Quandt, R. E. (1958). The estimation of the parameter of a linear regression system obeying two separate regimes, *J. Amer. Statist. Assoc.*, **53**, 873–880.

Schwarz, G. (1978). Estimating the dimension of a model, *Ann. Statist.*, **6**, 461–464.

Shibata, R. (1981). An optimal selection of regression variables, *Biometrika*, **68**, 45–54.

Smith, A. F. M. (1980). Change-point problems: approaches and applications, *Bayes Statistics* (eds. J. M. Bernardo *et al.*), 83–98, University Press, Valencia.

Smith, A. F. M. and Spiegelhalter, D. J. (1980). Bayes factors and choice criteria for linear models, *J. Roy. Statist. Soc. Ser. B*, **42**, 213–220.

Smith, A. F. M. and West, M. (1983). Monitoring renal transplants: an application of the multiprocess Kalman filter, *Biometrics*, **39**, 867–878.

Taplin, R. H. and Raftery, A. E. (1994). Analysis of agricultural fields trials in the presence of outliers and fertility jumps, *Biometrics*, **50**, 764–781.

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.

Yao, Y. C. (1984). Estimation of a noisy discrete-time step function: Bayes and empirical Bayes approaches, *Ann. Statist.*, **12**, 1434–1447.

Yao, Y. C. (1988). Estimating the number of change points by Schwarz's criterion, *Statist. Probab. Lett.*, **6**, 181–189.