

Multiple temporal cluster detection test using exponential inequalities

Christophe DEMATTEI¹ and Nicolas MOLINARI^{1,2}

¹ *Laboratoire de biostatistique, Institut Universitaire de Recherche Clinique, UFR Medecine Site Nord UPM/IURC, 640 avenue du Doyen Giraud, 34295 Montpellier Cedex 5, France*

² *UMR Analyse des Systemes et Biometrie, 2 place Pierre Viala, Campus ENSAM-INRA, 34060 Montpellier Cedex 1, France*

Abstract

The method of Molinari et al. (2001) is a multiple temporal cluster detection approach, which is based on a data transformation. The model selection procedure and the test of the cluster significance are achieved by bootstrap. The use of simulations is a common point between existing temporal cluster detection methods. The aim of this paper is to propose a new approach to avoid the use of such simulations in the cluster significance test stage. A direct application of the Bernstein inequality allows to compute upper bounds for p -values for each potential cluster. We also propose another model selection procedure based on multiple structural changes developed by Bai and Perron (1998). The new detection approach based on inequalities is detailed. Those inequalities are applied on simulated data and on two real data set. A discussion concludes the paper.

Keywords : Cluster detection, Bernstein inequality, Multiple structural changes, Temporal cluster, Break dates, Double maximum test

Email address: demattei@iurc.montp.inserm.fr (Christophe DEMATTEI¹).

Introduction

A temporal cluster is an unusual aggregation of events that are grouped together in time. Clusters of health events are often reported to health agencies. When the etiology of a disease has not yet been established, it is sometimes required to examine data for obtaining evidence of temporal clustering and to establish an etiologic link with exposure. Several fields are affected by temporal cluster detection such as medicine, social sciences, agronomy and more. The question of whether events are clustered in time has received considerable attention in the literature. A survey is presented in Bonaldi [6].

The scan statistic proposed by Kulldorff and Nagarwalla [11] is an efficient method for detecting temporal clusters. The size of scanning windows is variable, which allows to the cluster size (interval length) not to be chosen *a priori*. This test is the generalized likelihood ratio test for a uniform null distribution against an alternative of non random clustering. The significance of the test is provided by Monte Carlo simulations. The method was extended for detecting disease clusters in heterogeneous populations by Kulldorff [10].

Molinari et al. [12] proposed an original method that allows to detect several temporal clusters. This approach is based on a simple data transformation. The serie of dates of occurrence is replaced by the serie of the time between two successive events. This method determines a time window with excess events and scans continuously over the study period for any position of the window. A window is considered to have a high density of events when it groups together events that are near to each other. The method is effective with changes in the population at risk. Presence of one or more clusters is determined by using bootstrapped simulations and a classical model selection procedure.

The common point between existing methods to detect temporal clusters (multiple or not) is the need, to carry out the inference, to use bootstrap or Monte Carlo methods. In this paper, we propose to test the cluster significance without using simulated samples. This approach is based on the inequality established by Bernstein [5] for the sum of independant random variables. This inequality has been adapted to the temporal case. This allows to obtain upper bound for the p -value in the test for cluster significance. We propose also another model selection procedure, the double maximum test, developed by Bai and Perron [1] which is based on multiple structural change models. This procedure was previously used in the spatial case in Demattei et al. [7].

The first section presents briefly the method of Molinari et al. [12]. The model selection via multiple structural changes models is detailed in section 2. In the third section, the Bernstein inequality is recalled and then adapted to the detection of multiple temporal clusters. The method is applied both on simulated and real data in a fourth section. The paper is concluded by a discussion.

1. Standard method

The method of Molinari et al. [12] is based on a data transformation. Initial data are constituted by the times of occurrence of events in the interval of observation. The transformation consists in obtaining values corresponding to the time (the distance) between two successive events. Under the uniform distribution hypothesis (no cluster), these values can be estimated by a constant, the mean distance. Under the alternative, a piecewise constant model improves the fitting.

In this section, we recall the data transformation and the cluster location procedure. The model selection and the detection stages will be treated separately in following sections.

Let n be the number of events occurring in the interval of observation wich can be set to $[0; 1]$ without loss of generality. Times of occurrence of those n events are i.i.d random variables denoted X_1, \dots, X_n . Then, let x_1, \dots, x_n be a realization of X_1, \dots, X_n , and $x_{(1)}, \dots, x_{(n)}$ be the ordered serie of these times

from the origin. Finally, set for $k = 1, \dots, n$, $y_k = x_{(k)} - x_{(k-1)}$ (by convention $x_{(0)} = 0$).

In order to determine potential cluster bounds, we took the distance regression on the selection order k . Consider the data set $(k, y_k)_{k=1, \dots, n}$. Under the no-cluster hypothesis, an appropriate regression function to use would be the constant one

$$f(t) = \bar{y} = \frac{1}{n} \sum_{k=1}^n y_k .$$

Let $\bar{y}_{[i;j]}$ ($1 \leq i < j \leq n$) be the mean of y_t for t in $[i; j]$, and $I_{[i;j]}(t) = 1$ if $t \in [i; j]$ and 0 otherwise. To determine the presence of m breaks ($m + 1$ regimes), the regression function taken into consideration is :

$$f(t) = \sum_{j=1}^{m+1} \bar{y}_{[n_{j-1}+1; n_j]} \times I_{[n_{j-1}+1; n_j]}(t)$$

with the convention $n_0 = 0$ and $n_{m+1} = n$.

In order to carry out an asymptotic analysis in the detection stage, it is necessary to impose a minimum size for each portion between two breaks. The set of possible partitions is defined as follows : for some arbitrary positive number $\epsilon \in [0; 1]$, $\Delta_\epsilon = \{(n_1, \dots, n_m) ; \forall i = 1, \dots, m + 1, \text{card}([n_{i-1} + 1; n_i]) \geq \lfloor n\epsilon \rfloor\}$. For example, an ϵ of 0.2 means that the number of dates between two breaks is imposed as being at least 20% of the total number of dates. See Demattei et al. [7] for more explanations on the parameter ϵ .

Breaks (cluster bounds) are estimated by resolving the constrained least square problem

$$\min_{(n_1, \dots, n_m) \in \Delta_\epsilon} \sum_{t=1}^n (y_t - f(t))^2 .$$

We note $(\hat{n}_1, \dots, \hat{n}_m)$ the solution.

For computing these estimates efficiently, we can use the method of Bai and Perron [3] which is based on dynamic algorithm programming.

2. Model selection

In this section, we present the model selection procedure. This procedure is based on the double maximum test developed by Bai and Perron. We must first consider the no break test versus a fixed number $m = k$ of breaks. The test statistic proposed by Bai and Perron [1] is

$$F_n(\hat{n}_1, \dots, \hat{n}_k) = \left(\frac{n - (k + 1)}{k} \right) \hat{A} D' (D \hat{V}(\hat{A}) D')^{-1} D \hat{A} = \sup_{(n_1, \dots, n_k) \in \Delta_\epsilon} F_n(n_1, \dots, n_k)$$

in which $\hat{A} = (\hat{a}_1, \dots, \hat{a}_{k+1})'$ is the vector of the mean distance on each portion and D is a matrix so that $D \hat{A} = (\hat{a}_1 - \hat{a}_2, \dots, \hat{a}_k - \hat{a}_{k+1})'$. $\hat{V}(\hat{A})$ is an estimate of the variance covariance matrix of \hat{A} .

F_n is the statistic for testing $\hat{a}_1 = \dots = \hat{a}_{k+1}$ against $\hat{a}_i \neq \hat{a}_{i+1}$ for a certain i . A high value of F_n means a shift away from no break hypothesis. Critical values for this test statistic are given by Bai and Perron [2] for values of ϵ between 0.05 and 0.25.

The double maximum test, defined in Bai and Perron [1], allows us to test the null hypothesis of no break against an unknown number of breaks given a certain upper bound M . Let $c(\alpha, m)$ denote the asymptotic critical value of the test $F_n(\hat{n}_1, \dots, \hat{n}_m)$ for a significance level α . The test is denoted :

$$WD \max F_n(M) = \max_{1 \leq m \leq M} \frac{c(\alpha, 1)}{c(\alpha, m)} F_n(n_1, \dots, n_m) .$$

The best model can now be selected. The number of breaks is chosen as the *argmax* of the *WD max* statistic. Critical values for this corrected test statistic are given by Bai and Perron [2] for values of ϵ between 0.05 and 0.25 and $M \leq 9$.

3. Cluster detection using inequalities

The Bernstein inequality [5] is a particular case of Hoeffding [8] and Bennett [4] inequalities. Those three inequalities are based on the Chernoff's bounding method.

Theorem 1 (Bernstein inequality) *Let $(Z_i)_{i=1}^n$ be independant random variables with $Z_i - E[Z_i] \leq d$ for all $i \in \{1, \dots, n\}$. Let $S = \sum_{i=1}^n Z_i$ and $u > 0$. Then, with $\sigma_i^2 = E[Z_i^2] - E[Z_i]^2$ we have*

$$P(S - E[S] \geq u) \leq \exp\left(-\frac{u^2}{2 \sum_{i=1}^n \sigma_i^2 + \frac{2ud}{3}}\right). \quad (1)$$

□

We propose to test, for a given number of breaks m , the significance for each portion between two breaks, say $\hat{n}_k + 1$ and \hat{n}_{k+1} . In order to simplify the notation, we note $N = \hat{n}_{k+1} - \hat{n}_k$ and we rename $(Y_i)_{i=1}^N$ the distance series $(Y_i)_{i=\hat{n}_k+1}^{\hat{n}_{k+1}}$.

N cannot be used directly. All the probabilities have to be computed conditionnally to N . Indeed, N is a random variable which depends on m and more generally on the sample X_1, \dots, X_n . This difficulty is overcome when another realization $\tilde{X}_1, \dots, \tilde{X}_n$ is known. The number N of events falling into a given portion is then computed on this other sample, in order to suppress the dependency between N and X_1, \dots, X_n . We assume this in what follows.

Under the assumption that X_1, \dots, X_n are i.i.d. uniform $U(0, 1)$, $X_{(1)}, \dots, X_{(n)}$ are distributed as n -order statistics from a uniform $U(0, 1)$ parent. In this case, $X_{(i)}$ follows a beta distribution $\beta(i, n - i + 1)$. Let $Y_i = X_{(i)} - X_{(i-1)}$ be the distance (time) between the successive events $X_{(i-1)}$ and $X_{(i)}$. Y_i has a beta distribution $\beta(1, n)$. Thus, the null hypothesis of uniform distribution can be written H_0 : "the mean of Y_i on a portion is equal to the mean of a $\beta(1, n)$ distributed variable". For each portion, we propose to test H_0 versus H_1 : "the mean of Y_i is less than the mean of a $\beta(1, n)$ " by using the following inequalities. H_1 denotes the presence of a cluster on the considered portion.

Proposition 3.1 *Let $(Y_i)_{i=1}^N$ be independent random variables following a $\beta(1, n)$ distribution with $n \geq N$. For all $i \in \{1, \dots, N\}$, define $\tilde{Y}_i = (n+1)Y_i$. Let $T = \frac{1}{N} \sum_{i=1}^N \tilde{Y}_i$ and $u > 0$. Assume N is independent of $(Y_i)_{i=1}^{n-1}$. Then we have*

$$\mathbb{P}\left(T \leq 1 - \frac{u}{N}\right) \leq \exp\left(-\frac{u^2}{\frac{2nN}{n+2} + \frac{2u}{3}}\right). \quad (2)$$

Proof. Since $Y_i \sim \beta(1, n)$, Y_i is non-negative with $\mathbb{E}[Y_i] = \frac{1}{n+1}$ and $\text{Var}(Y_i) = \frac{n}{(n+1)^2(n+2)}$. Thus, for $i = 1, \dots, N$, the random variables $\tilde{Y}_i = (n+1)Y_i$ are independent, non-negative, with $\mathbb{E}[\tilde{Y}_i] = 1$ and $\text{Var}(\tilde{Y}_i) = (n+1)^2 \text{Var}(Y_i) = \frac{n}{n+2}$.

Set $Z_i = 1 - \tilde{Y}_i$. Since \tilde{Y}_i is non negative and $\mathbb{E}[Z_i] = 1 - \mathbb{E}[\tilde{Y}_i] = 0$, we have $Z_i - \mathbb{E}[Z_i] = Z_i = 1 - \tilde{Y}_i \leq 1$. We are now in a position to apply the Bernstein inequality to the random variables Z_i :

$$\mathbb{P}\left(\sum_{i=1}^N Z_i - \mathbb{E}\left[\sum_{i=1}^N Z_i\right] \geq u\right) \leq \exp\left(-\frac{u^2}{2 \sum_{i=1}^N \text{Var}(Z_i) + \frac{2u}{3}}\right). \quad (3)$$

Moreover, we clearly have $\sum_{i=1}^N Z_i - \mathbb{E} \left[\sum_{i=1}^N Z_i \right] = \sum_{i=1}^N (1 - \tilde{Y}_i) = N(1 - T)$, and $\text{Var}(Z_i) = \text{Var}(\tilde{Y}_i) = \frac{n}{n+2}$. Therefore, (3) becomes

$$\mathbb{P}(N(1 - T) \geq u) \leq \exp \left(-\frac{u^2}{\frac{2nN}{n+2} + \frac{2u}{3}} \right),$$

as desired. \square

This proposition provides an upper bound on the probability that the mean of the Y_i is less than a given threshold.

To make this result effective, we propose to control the type I error rate α . This is achieved by the following corollary.

Corollary 3.2 *Under the assumptions of Proposition 3.1, we have for all $\alpha \in (0, 1)$*

$$\mathbb{P} \left(T \leq 1 - \frac{u_\alpha}{N} \right) \leq \alpha \text{ with } u_\alpha = -\frac{\ln(\alpha)}{3} + \sqrt{\left(\frac{\ln(\alpha)}{3} \right)^2 - \frac{2nN \ln(\alpha)}{n+2}}. \quad (4)$$

Proof. The proof is clear. \square

If we set the type I error rate to α , this corollary makes it possible to specify the threshold $1 - u_\alpha/N$ associated to this value for α . Hence, under H_0 , the probability that the mean of a portion of size N is under the threshold is less than α . If the mean is effectively under the threshold, we can reject H_0 with a type I error rate less than α . This provides us a conservative procedure to test the significance of a given portion. Several clusters can be detected by applying this procedure to each potential cluster. It is worth pointing out that it allows to avoid using bootstrapped or Monte Carlo methods for inference.

The corollary below gives, as a by-product of Proposition 3.1, the p -value corresponding to the observed mean distance of a portion, say t .

Corollary 3.3 *Under the assumptions of Proposition 3.1, we have for all $t < 1$*

$$\mathbb{P}(T \leq t) \leq p_t \text{ with } p_t = \exp \left(-\frac{N(1-t)^2}{\frac{2n}{n+2} + \frac{2(1-t)}{3}} \right). \quad (5)$$

Proof. Just apply (2) with $u = N(1 - t)$ and note that $u > 0$ since $t < 1$. \square

Thus, another way to use Proposition 3.1 is to set the threshold t in Corollary 3.3 to the observed mean distance of a given portion. This provides a p -value p_t . If $p_t \leq \alpha$, H_0 can be rejected with a type I error rate α and the portion represents a significant temporal cluster.

We have to note that the threshold $1 - u_\alpha/N$ is negative when

$$\frac{1}{3N} + \frac{n}{N(n+2)} > -\frac{1}{2\ln(\alpha)}.$$

In this case, the threshold cannot be reached by the mean distance.

4. Applications

4.1. Sample run

We applied this method with $\alpha = 0.05$ to a sample of 100 times of occurrence. The times were simulated by a mixture $0.5 \times \mathcal{U}(0, 100) + 0.25 \times \mathcal{U}(25, 35) + 0.25 \times \mathcal{U}(60, 80)$. This mixture contains two potential clusters. The first one (C_1) has a high density and contains N_1 events. The other one, denoted by C_2 , has a density twice lower and contains N_2 events. N_1 and N_2 were determined on a replicate of the sample in order to suppress the dependency between the number of events in a given portion and the sample analyzed. The regression plot for the model with $m = 4$ breaks is presented in Figure 1. In C_1 , the mean of the distances \hat{Y}_i is less than the threshold and $p_{t_1} = 0.008 < 0.05$. In C_2 , the mean of the distances \hat{Y}_i is higher than the threshold and $p_{t_2} = 0.11$. We obtain one significant cluster (C_1) which corresponds to the high density portion.

[FIG. 1 about here.]

4.2. Knox data set

The Knox data set is a classical data set used to compare cluster detection methods. This data set consists of 35 cases of the birth defects esophageal atresia and tracheoesophageal fistula observed in a hospital in Birmingham, United Kingdom, between 1950 and 1955. The total time period of the study was 2191 days. Knox [9], Weinstock [14] with the scan statistic and Nagarwalla [13] with a scan statistic with variable window used this data set to illustrate their approaches. Molinari et al. [12] also used this data set on the regression method for multiple cluster detection. All these methods found the same cluster of 15 cases in the 258 days time interval from day 1233 to day 1491. The existence of a second cluster of 7 cases in the 125 days from day 2049 to day 2174 is controversial.

Due to the low number of cases, we choose an $\epsilon = 0.15$. Indeed, an $\epsilon = 0.1$ would have potentially found clusters of 4 cases, which is very few. The model with 3 breaks was selected, with two potential clusters. The first is the well-known cluster [1233-1491] with 15 cases in 258 days. The mean of the distance is below the threshold and $p = 0.045$. The second potential cluster is [2049-2174]. This cluster is not significant ($p = 0.33$). Those results are concordant with those obtained in Molinari et al. [12] and with the scan statistic ($p = 0.019$).

[FIG. 2 about here.]

4.3. Hospital hemoptysis admission data set

This data set consists of 62 days of hemoptysis admission at Nice University Hospital from January 1 to December 31, 1995. As proposed in Molinari et al. [12], the population at risk has been modified to take into account the of 0.72% per year and the crowd of 55000 tourists in summer.

Since there is more cases than in the previous example, we can choose $\epsilon = 0.1$. The model with 2 breaks was selected. The potential cluster is [58;87] as in Molinari et al. [12]. However, the p -value associated to this portion is 0.13. We cannot conclude that this portion is significant. The scan statistic led to the same result ($p = 0.29$). Those two results are in contradiction with the one obtained by Molinari et al. [12] ($p = 0.02$).

[FIG. 3 about here.]

Discussion

In this paper, we have modified the original method by avoiding the use of bootstrap simulations.

The first way to overcome those simulations is to use the *WD max* statistic, initially proposed in the econometric field by Bai and Perron [1]. This procedure, conceived for multiple structural changes, is particularly useful to select the best model in multiple cluster detection problems.

Our approach is a direct application of the Bernstein inequality for the sum of bounded random variables. This method provides an upper bound for the p -value. In that sense, the inequality allows to detect clusters in a conservative way. This method also has the advantage of being very flexible. Firstly, it can locate several potential clusters. Moreover, it makes possible to test the uniform distribution hypothesis for each cluster separately.

This last point is well illustrated in the simulated exemple since the best model selected contains two potential clusters : the method of Molinari et al. [12] can only test the model in its whole and would detect two clusters or no cluster, whereas the present method allows to affirm that only one of the two potential clusters is significant.

A perspective of this work is to adapt the Bernstein inequality to the multiple spatial cluster detection method of Dematteï et al. [7]. However, this extension is not immediate since the variables (corresponding to the distance from a point to its nearest neighbour) are weakly dependant in the spatial case (the distance depends on the trajectory already done until this point). Some inequalities exist for the sum of weakly dependent variables. To apply them to the spatial field, the work will be to characterize the dependance between the distance variables.

Références

- [1] J. Bai and P. Perron, Estimating and Testing Linear Models with Multiple Structural Changes, *Econometrica* 66 (1998), 47–78.
- [2] J. Bai and P. Perron, Critical Values for Multiple Structural Change tests, *Econometrics Journal* 6 (2003), 72–78.
- [3] J. Bai, P. Perron, Computation and analysis of multiple structural change models, *J. Appl. Econom.* 18 (2003) 1–22.
- [4] G. Bennett, Probability inequalities for the sum of independant random variables, *Journal of the American Statistical Association* 57 (1962), 33–45.
- [5] S. Bernstein, *The Theory of Probabilities*, Gastehizdat Publishing House, Moscow, 1946.
- [6] C. Bonaldi, *Analyse de clusters sur le temps*, Thesis, University of Montpellier I, Montpellier, 2003.
- [7] C. Dematteï, N. Molinari and J.P. Daurès, Arbitrarily Shaped Multiple Spatial Cluster Detection for Case Event Data, To appear in *Computational Statistics and Data Analysis* (2006).
- [8] W. Hoeffding, Probability inequalities for sums of bounded random variables, *Journal of the American Statistical Association* 58 (1963), 13–30.
- [9] G. Knox, Secular pattern of congenital oesophageal atresia, *British Journal of Preventive Social Medicine* 13 (1959), 222–226.
- [10] M. Kulldorff, A spatial scan statistic, *Communications in Statistics - Theory and Methods* 26 (1997) 1481–1496.
- [11] M. Kulldorff, N. Nagarwalla, Spatial disease clusters : detection and inference, *Statistics in Medicine* 14 (1995) 799–810.
- [12] N. Molinari, C. Bonaldi, J.P. Daurès, Multiple temporal cluster detection, *Biometrics* 57 (2001) 577–583.
- [13] N. Nagarwalla, A scan statistic with variable window, *Statistics in Medicine* 15 (1996), 845–850.
- [14] M.A. Weinstock, A generalized scan statistic test for the detection of clusters, *International Journal of Epidemiology* 10 (1981), 289–293.

Table des figures

- 1 Simulation results. C_1 corresponds to the second portion (which includes orders 20 and 40), and contains N_1 events with a mean distance t_1 . C_2 corresponds to the fourth portion (which includes orders 60 and 80), and contains N_2 events with a mean distance t_2 . The dotted lines represent the thresholds $1 - \frac{u_\alpha}{N}$ computed for each portion. For C_1 , the threshold is higher than t_1 , which means that $p_{t_1} < 0.05$ and that C_1 is a significant cluster. For C_2 , the threshold is below t_2 , which means that we cannot conclude that C_2 is significant. 9
- 2 Knox data set : regression plot. The dotted lines represent the thresholds $1 - \frac{u_\alpha}{N}$ computed for each portion. The first potential cluster is significant 10
- 3 Hemoptysis data set : regression plot. The dotted lines represent the thresholds $1 - \frac{u_\alpha}{N}$ computed for each portion. The potential cluster is not significant. 11

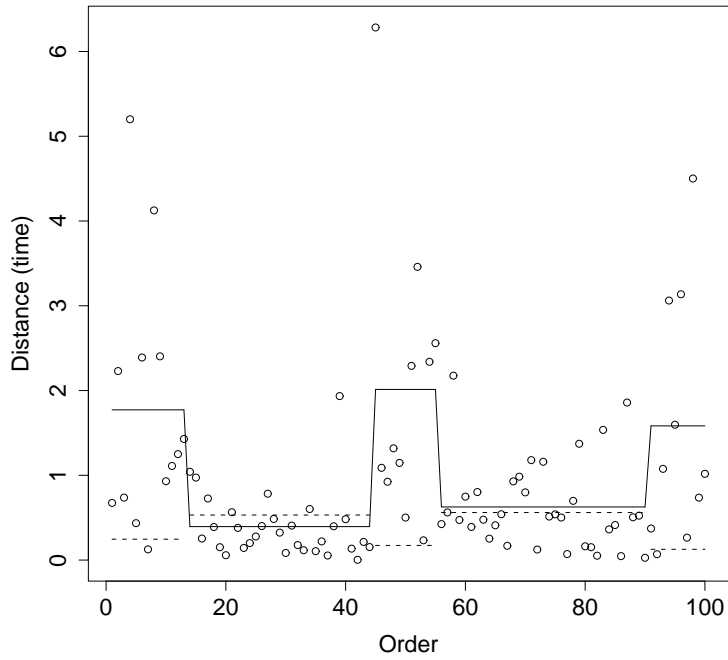


FIG. 1. Simulation results. C_1 corresponds to the second portion (which includes orders 20 and 40), and contains N_1 events with a mean distance t_1 . C_2 corresponds to the fourth portion (which includes orders 60 and 80), and contains N_2 events with a mean distance t_2 . The dotted lines represent the thresholds $1 - \frac{u_\alpha}{N}$ computed for each portion. For C_1 , the threshold is higher than t_1 , which means that $p_{t_1} < 0.05$ and that C_1 is a significant cluster. For C_2 , the threshold is below t_2 , which means that we cannot conclude that C_2 is significant.

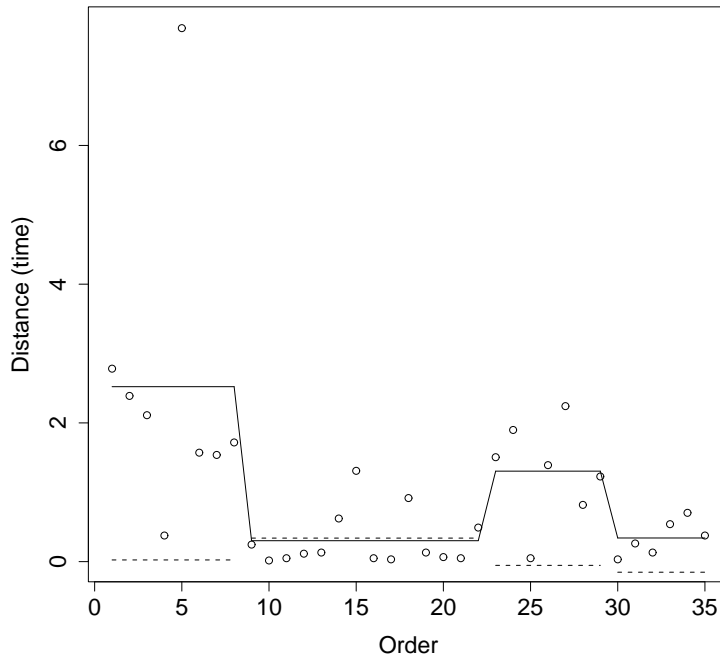


FIG. 2. Knox data set : regression plot. The dotted lines represent the thresholds $1 - \frac{u_{\alpha}}{N}$ computed for each portion. The first potential cluster is significant

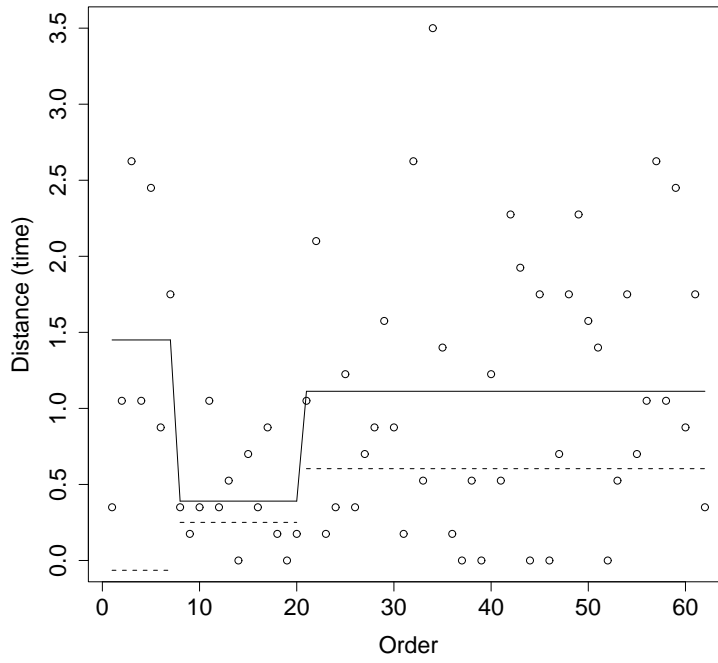


FIG. 3. Hemoptysis data set : regression plot. The dotted lines represent the thresholds $1 - \frac{u_{\alpha}}{N}$ computed for each portion. The potential cluster is not significant.