Available online at www.sciencedirect.com
ScienceDirect
C. R. Acad. Sci. Paris, Ser. I 344 (2007) 697–701
COMPTES RENDUS MATHEMATIQUE
http://france.elsevier.com/direct/CRASS1/

Statistics

P-value calculations for multiple temporal cluster detection

Christophe Dematteï, Nicolas Molinari

Laboratoire de biostatistique, institut universitaire de recherche clinique, 641, avenue du Doyen Gaston-Giraud, 34093 Montpellier, France

Received 12 January 2006; accepted after revision 4 April 2007

Presented by Paul Deheuvels

Abstract

The aim of this Note is to propose a new approach to test multiple temporal cluster significance. Our method is based on a data transformation and on multiple structural change models, and it completes a former method (Molinari et al., 2001). Instead of using bootstrap replicates, we compute upper bounds for *p*-values using the Bernstein inequality. The inequalities on which the new detection method is based are detailed. ***To cite this article: C. Dematteï, N. Molinari, C. R. Acad. Sci. Paris, Ser. I 344 (2007).***

Résumé

Calcul de la p-valeur d'un test de détection de clusters temporaux multiples. L'objectif de cette Note est de proposer une nouvelle approche afin de tester la significativité de clusters temporaux multiples. Notre approche, qui est basée sur une transformation des données et sur des modèles de changements structurels multiples, complète une méthode existante (Molinari et al., 2001). Au lieu d'utiliser des simulations par bootstrap, nous calculons des bornes supérieures pour les *p*-valeurs en utilisant l'inégalité de Bernstein. Les inégalités servant de base à la nouvelle méthode de détection sont détaillées. ***Pour citer cet article : C. Dematteï, N. Molinari, C. R. Acad. Sci. Paris, Ser. I 344 (2007).***

Version française abrégée

La plupart des méthodes de détection de clusters temporaux ont comme point commun de faire appel à des simulations afin de tester la significativité des modèles envisagés (méthodes de Monte Carlo ou bootstrap). C'est notamment le cas d'une méthode par régression après transformation des données permettant de tester la présence de plusieurs clusters décrite dans Molinari et al. [6].

Dans cette Note, nous contournons l'utilisation de simulations en proposant une inégalité exponentielle.

Notons $X_1, \ldots, X_n$ des variables aléatoires i.i.d. uniformes $U(0, 1)$. Leur réalisations représentent les temps d'occurrence de n évènements dans l'intervalle $(0, 1)$. L'idée générale de la méthode est basée sur l'hypothèse que les évènements formant un agrégat sont consécutifs et très proches les uns des autres, et donc que la distance moyenne entre 2 évènements est plus faible à l'intérieur d'un cluster qu'à l'extérieur. Cette distance est notée $Y_i = X_{(i)} - X_{(i-1)}$ avec la convention $X_{(0)} = 0$. Afin de modéliser la présence de m points de cassures $(n_1, \ldots, n_m)$, et ainsi pouvoir re-

E-mail addresses: dematteï@iurc.montp.inserm.fr (C. Dematteï), molinari@iurc.montp.inserm.fr (N. Molinari).

1631-073X/$ – see front matter
doi:10.1016/j.crma.2007.04.004

pérer les clusters potentiels (portion entre deux points de cassure ayant une distance moyenne faible), les auteurs envisagent la fonction de régression

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

avec la convention $n_0 = 0$ et $n_{m+1} = n$. La notation $\bar{d}_{[i;j]}$ $(1 \leqslant i < j \leqslant n)$ désigne la moyenne des Y_k pour k dans $[i; j]$, et $I_{[i;j]}(t) = 1$ si $t \in [i; j]$ et 0 sinon. Notons que pour $m = 0$, la fonction $f(t)$ est constante et égale à la moyenne de toutes les distances. Les n_i sont estimés en résolvant le problème des moindres carrés sous contrainte défini par l'équation (1). Finalement, le modèle avec $m > 0$ points de cassure peut être testé contre l'hypothèse H_0 de non-agrégation (aucun point de cassure) en utilisant des simulations par bootstrap (Molinari et al. [6]).

Dans cette Note, nous exploitons notamment le fait que pour $i = 1, \ldots, n$, Y_i suit une distribution beta $\beta(1, n)$. Si on note N le nombre d'éléments compris dans une portion donnée, et en définissant $Z_i = (n + 1)Y_i$, $T = \frac{1}{N}\sum_{i=1}^{N} Z_i$ et $t > 0$, nous pouvons appliquer le théorème de Bernstein [2] pour obtenir l'inégalité (2). Ce résultat permet par la suite d'obtenir, sous H_0, un seuil pour T tout en contrôlant le risque de première espèce α, ou encore de calculer un majorant de la p-valeur p_u correspondant à la valeur observée u de T (voir (3) et (4)). Ainsi, si la moyenne des distances sur une portion se situe sous le seuil $1 - t_\alpha/N$, nous pouvons rejeter H_0 avec un risque α. Cette procédure peut être employée sur chacune des portions définies par les m cassures.

Un exemple d'application de cette approche est illustré par la Fig. 1. L'échantillon est simulé suivant un mélange de lois uniformes afin de former deux clusters, le premier avec une densité faible, le second avec une densité deux fois supérieure. Les seuils sont représentés par les lignes en pointillé et nous observons que le premier cluster potentiel présente une moyenne qui est au dessus du seuil, tandis que la moyenne du second est inférieure au seuil. Le second cluster est significatif.

1. Introduction

The term 'cluster' is an unusual aggregation, real or perceived, of events that are grouped together in time and/or space. Clusters of health events, such as chronic disease, injuries, and birth defects, 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 or spatial clustering and to establish an etiologic link with exposure. Temporal cluster detection affects severals fields: medicine, social sciences, agronomy and more. The question of whether events are clustered in time has received considerable attention in the literature since the 1960s (see Bonaldi [3] for a survey).

In a seminal paper, Kulldorff and Nagarwalla [5] propose an efficient method for detecting both temporal and spatial clusters. The authors use a scan statistic with variable window that allows one to the cluster window size (interval length) not to be chosen a priori. Their test is the generalized likelihood ratio test for a uniform null distribution against an alternative of non-random clustering. The test significance is provided by Monte Carlo simulations. The method was extended by Kulldorff [4] for detecting disease clusters in heterogeneous populations.

More recently, Molinari et al. [6] proposed a method which allows one to detect several temporal clusters. This algorithm is based on a simple data transformation. It determines a time window with excess events and scans continuously over the study period for any position of the window. Moreover, this approach is effective with changes in the population at risk. Its main weakness is that the presence of one or more clusters is determined by using bootstrapped simulations.

In this Note, we propose to overcome this difficulty by testing cluster significativity without using simulated samples. Our approach is based on an application of the inequality established by Bernstein [2] for the sum of independent random variables. We adapt this inequality to the multiple temporal cluster detection case and we obtain an upper bound for the p-value in the test for cluster significativity.

2. Potential cluster location

In this section we recall the method of Molinari et al. [6]. The method is based on a data transformation in order to obtain values corresponding to the time (the distance) between two successive events. Under the no-clustering

hypothesis H_0 (uniform distribution), these values can be estimated by a constant, the mean distance. Under the alternative, a piecewise constant model improves the fitting.

Let $X_1, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables which represent the times of occurrence of n events in an interval $(0, T)$. Without loss of generality, set $T = 1$. Suppose that $x_1, \ldots, x_n$ are dropped at random in the unit interval $(0, 1)$. Then, let $x_{(1)}, \ldots, x_{(n)}$ be the ordered distances of these points from the origin, and set for $i = 1, \ldots, n$, $y_i = x_{(i)} - x_{(i-1)}$ (by convention $x_{(0)} = 0$).

Consider now the data set $(k, y_k)_{k=1,\ldots,n}$. Under the no cluster hypothesis, an appropriate regression function to fit this data set would be the constant one

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

On the other hand, to determine the presence of m breaks ($m + 1$ regimes), the regression function taken into consideration is

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

where $n_1, \ldots, n_m$ are the m breaks, with the convention $n_0 = 0$ and $n_{m+1} = n$. The notation $\bar{d}_{[i;j]}$ $(1 \leqslant i < j \leqslant n)$ indicates the mean of y_t for t in $[i; j]$, and $I_{[i;j]}(t) = 1$ if $t \in [i; j]$ and 0 otherwise.

Breaks are estimated by resolving the constrained least square problem

$$\min_{(0<n_1<\cdots<n_m<n)} \quad \sum_{k=1}^{n}\bigl(y_k - f(k)\bigr)^2, \tag{1}$$

and we denote by $(\hat{n}_1, \ldots, \hat{n}_m)$ the solution.

A method to compute these estimates efficiently is presented by Bai and Perron [1]. It is based on a dynamic algorithm programming.

The general idea of the method of Molinari et al. [6] is based on the assumption that points included in a cluster are consecutive and that their associated distances are lower than those of points outside the cluster (because the density of points is higher within the cluster). Therefore, potential clusters are located by portions with a low mean distance $\bar{d}_{[\hat{n}_{j-1}+1;\hat{n}_j]}$. In the original approach, the model with m breaks is tested versus the no-clustering hypothesis H_0 using bootstrapped replicates. In the next section we propose a new approach to overcome this limitation.

3. Inequalities and test

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

N is a random variable which depends on m and more generally on the sample $X_1, \ldots, X_n$. In a general way, we cannot use N directly, but we have to compute all the probabilities conditionally to N. This difficulty is overcomed when another realization $\tilde{X}_1, \ldots, \tilde{X}_n$ is known. For example, we can divide the original data into two parts by taking one ordered time out of two. Hence $\tilde{X}_{(1)}, X_{(1)}, \ldots, \tilde{X}_{(n)}, X_{(n)}$ is the ordered original time series. The breaks and the number N of events falling into a given portion are then computed on the first half-sample, called training sample. What follows is applied to the second half-sample, called test sample, and N is not random.

Assuming that the X_i's are i.i.d. uniform $U(0, 1)$, then $X_{(1)}, \ldots, X_{(n)}$ are distributed as n-order statistics from a uniform $U(0, 1)$ parent. In this case, $X_{(i)}$ follows a beta distribution $\beta(i, n-i+1)$ and $Y_i = X_{(i)} - X_{(i-1)}$, the distance (time) between the successive events $X_{(i-1)}$ and $X_{(i)}$, 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. The hypothesis 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 $n \geqslant N$. For all $i \in \{1,\dots,N\}$, define $Z_i = (n+1)Y_i$. Let $T = \frac{1}{N}\sum_{i=1}^N Z_i$ and $t > 0$. Then we have*

$$\mathbb{P}\left(T \leqslant 1 - \frac{t}{N}\right) \leqslant \exp\left(-\frac{t^2}{2nN/(n+2) + 2t/3}\right). \tag{2}$$

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

Since Z_i is non-negative and $\mathbb{E}[Z_i] = 1$, we are in a position to apply the Bernstein inequality [2] to the random variables $1 - Z_i$:

$$\mathbb{P}\left(\sum_{i=1}^N (1-Z_i) - \mathbb{E}\left[\sum_{i=1}^N (1-Z_i)\right] \geqslant t\right) \leqslant \exp\left(-\frac{t^2}{2\sum_{i=1}^N \mathrm{Var}(Z_i) + 2t/3}\right). \tag{3}$$

Moreover, we clearly have $\sum_{i=1}^N (1-Z_i) - \mathbb{E}[\sum_{i=1}^N (1-Z_i)] = N(1-T)$. Therefore, (3) becomes

$$\mathbb{P}\big(N(1-T) \geqslant t\big) \leqslant \exp\left(-\frac{t^2}{2nN/(n+2) + 2t/3}\right),$$

as desired. □

Proposition 3.1 provides an upper bound on the probability that the mean of the Y_i is less than a given threshold. The following corollary allows one to make this result effective by controlling the type I error rate α:

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

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

Proof. The proof is clear. □

Setting the type I error rate to α, Corollary 3.2 allows one to specify the threshold $1 - t_\alpha/N$ 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, and this provides us a conservative procedure to test the significance of a given portion. Applying this procedure to each potential cluster allows one to detect several clusters. It is worth pointing out that it allows one to avoid using bootstrapped or Monte Carlo methods for inference.

As a by-product of Proposition 3.1, we give the p-value corresponding to the observed mean distance of a portion, say u:

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

$$\mathbb{P}(T \leqslant u) \leqslant p_u \quad \textit{with } p_u = \exp\left(-\frac{N(1-u)^2}{2n/(n+2) + 2(1-u)/3}\right). \tag{5}$$

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

Another way to use Proposition 3.1 is to set the threshold u in Corollary 3.3 to the observed mean distance of a given portion, which provides a p-value p_u. If $p_u \leqslant \alpha$, we can reject H_0 with a type I error rate α and the portion represents a significant temporal cluster.

To illustrate our results, we applied this method with $\alpha = 0.05$ to a sample of 200 times of occurrence, simulated by the mixture $0.5 \times \mathcal{U}(0,100) + 0.25 \times \mathcal{U}(20,40) + 0.25 \times \mathcal{U}(70,80)$. This mixture contains two clusters. The first

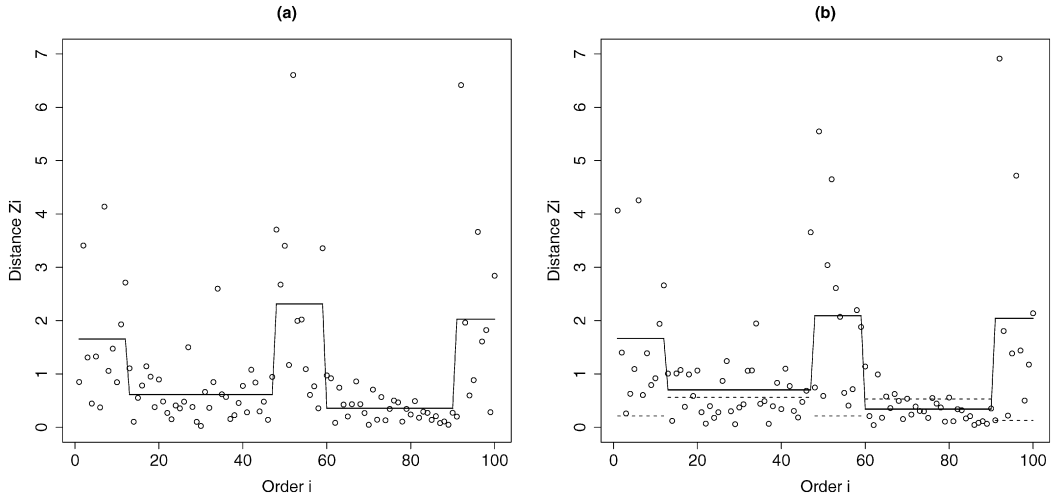

Fig. 1. Results obtained (a) on the training sample (to determine N_1 and N_2) and (b) on the test sample on which the inequality is applied. C_1 corresponds to the second portion (which includes orders 20 and 40), and contains N_1 events with a mean distance u_1. C_2 corresponds to the fourth portion (which includes orders 60 and 80), and contains N_2 events with a mean distance u_2. The dotted lines represent the thresholds $1 - t_\alpha/N$ computed for each portion. For C_1, the threshold is below u_1, which means that we cannot conclude that C_1 is significant. For C_2, the threshold is higher than u_2, which means that $p_{u_2} < 0.05$ and that C_2 is a significant cluster.

Fig. 1. Résultats obtenus (a) sur l'échantillon d'apprentissage (pour déterminer N_1 et N_2) et (b) sur l'échantillon de test sur lequel l'inégalité a été appliquée. C_1 correspond à la seconde portion (incluant les ordres 20 et 40), et contient N_1 évènements distants de u_1 en moyenne. C_2 correspond à la quatrième portion (incluant les ordres 60 et 80), et contient N_2 évènements distants de u_2 en moyenne. Les lignes en pointillés représentent les seuils $1 - t_\alpha/N$ calculés pour chaque portion. Pour C_1, le seuil est inférieur à u_1 ce qui signifie que l'on ne peut pas conclure à la significativité de C_1. Pour C_2, le seuil est supérieur à u_2 ce qui signifie que $p_{u_2} < 0.05$ et donc que C_2 est significatif.

one (C_1) has a low density. The other one, denoted by C_2, has a density twice higher. The breaks and the number of events falling in the two clusters, N_1 and N_2, were determined on the training half-sample (Fig. 1(a)). The regression plot for the model with $m = 4$ breaks, applied to the test half-sample, is presented in Fig. 1(b). In C_1, the mean of the distances Z_i is higher than the threshold and $p_{u_1} = 0.26 > 0.05$. In C_2, the mean of the distances Z_i is less than the threshold and $p_{u_2} = 0.0037 < 0.05$. We obtain one significant cluster (C_2).

The method presented here 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 present example since the best model selected contains two potential clusters: the method of Molinari et al. [6] can only test the model in its whole and would detect two clusters or no cluster, whereas the present method allows one to affirm that only one of the two potential clusters is significant. However, this approach has the disadvantage to consider only half of the data, which decreases the power of the test.

Acknowledgements

The authors wish to express their gratitude to Gérard Biau for his helpful comments.

References

[1] J. Bai, P. Perron, Computation and analysis of multiple structural change models, J. Applied Econometrics 18 (2003) 1–22.
[2] S. Bernstein, The Theory of Probabilities, Gostehizdat Publishing House, Moscow, 1946.
[3] C. Bonaldi, Analyse de clusters sur le temps, Thesis, University of Montpellier I, Montpellier, 2003.
[4] M. Kulldorff, A spatial scan statistic, Communications in Statistics—Theory and Methods 26 (1997) 1481–1496.
[5] M. Kulldorff, N. Nagarwalla, Spatial disease clusters: detection and inference, Statistics in Medicine 14 (1995) 799–810.
[6] N. Molinari, C. Bonaldi, J.P. Daurès, Multiple temporal cluster detection, Biometrics 57 (2001) 577–583.