Change-point model for a sequence of random variables with independent increments.

Dariush Ghorbanzadeh[1] Rachid Lounes [2] H. de Arazoza[3]

Abstract

In this paper, we consider a sequence of random variables with independent increments which can change their distribution at some unknown instant. We use this model with two different data sets.

Key words : change-point, independent increments, Brownian Bridge, Kolmogorov test.

1 Introduction

In the change-point literature, various authors have proposed different tests for the detection of a change-point and in most of them the detection concerns the mean. For the case of Poisson variables we can cite the papers of Raftery & Akman (1986), Simpkin & Downham (1988), Henderson & Matthews (1993), Ghorbanzadeh & Lounes (1996) and West & Ogden (1997).
Raftery & Akman (1986) proposed, a bayesian approach to the retrospective analysis of a Poisson process with a single change-point at an unknown time. They derived the posterior distribution for the parameters before and after the change time, and derived tests of the existence of a change-point. Simpkin & Downham (1988) have proposed a test of detection for Poisson random variables, based on maximum likelihood ratio statistic; assuming the change-point is fixed, they approximate the distribution of this statistic by a χ^2 distribution. Henderson & Matthews (1993) propose a classical model for

[1]CNAM-Département Mathématiques 292, Rue Saint-Martin
75141 PARIS cedex 03-FRANCE, ghorbanzadeh@vcnam.cnam.fr
[2]Université René Descartes, Laboratoire de Statistique Médicale, UPRES A CNRS 8071, 45 rue des Saints-Pères 75006 PARIS-FRANCE, lounes@citi2.fr
[3]Departemento Ecuaciones Diferenciales, Facultad Mat-Comp,
Universidad de la Habana, San Lázaro y L Habana 4 Cuba.

Poisson random variables to detect changes in the annual number of cases of haemolytic uraemic syndrome (HUS). They studied not only at most one change but also multi-change. West and Ogden (1997) used the HUS data for the estimation problem of change-points in a sequence of Poisson random variables, they approached the problem by allowing the change to range over a continuous time interval $(0, T)$, and derived a bayesian-based interval estimator. Ghorbanzadeh & Lounes (1996) proposed two methods for the detection of change-point for grouped Poisson random variables. The first method is based on penalized maximum likelihood ratio statistic. The second one uses a bayesian approach on the posterior distribution of change-point parameter.
In this work, we consider a sequence of random variables $X_1, \ldots, X_n$ with independent increments (i.e. $X_1, X_2 - X_1, \ldots, X_n - X_{n-1}$ independent) with Poisson distributions. The independent increments $X_1, \ldots, X_k - X_{k-1}$ follow a Poisson distribution $\mathcal{P}(\lambda)$ and the independent increments $X_{k+1} - X_k, \ldots, X_n - X_{n-1}$ follow a Poisson distribution $\mathcal{P}(\lambda + \frac{\delta}{\sqrt{n}})$ where the parameters λ and δ are unknow. We propose the following test problem:

$$\begin{cases} H_0 : & \delta = 0 \\ \\ H_1 : & \delta \neq 0 \end{cases} \tag{1}$$

2 Statistical Model

For testing the existence of the change-point we propose the statistic:

$$R_n(k) = \sqrt{n}\,(\frac{X_n - X_k}{n-k} - \frac{X_k}{k}) \tag{2}$$

This model is a particular case considered in Chapter 2 of [3] with a different methodology.

To study the behavior of R_n under the hypotheses H_0 and H_1 we have the folowing lemma:

Lemma 1 Under H_0, we have :

1. $X_i \sim \mathcal{P}(i\,\lambda)$
2. $\mathtt{Cov}(X_i, X_j) = \min(i, j)\,\lambda$

`Proof.` Let $Y_1 = X_1$ and for $i = 2, \ldots, n \quad Y_i = X_i - X_{i-1}$. From the fact that $X_1, \ldots, X_n$ have independent increments, then the variables $Y_1, \ldots, Y_n$

are independent and the variables $Y_1, \ldots, Y_k \sim \mathcal{P}(\lambda)$ and $Y_{k+1}, \ldots, Y_n \sim \mathcal{P}(\lambda + \frac{\delta}{\sqrt{n}})$. In the other hand, we have :

$$\forall i = 1, \ldots, n \qquad X_i = \sum_{j=1}^{i} Y_j \tag{3}$$

and from this we deduce that

$$\mathtt{Cov}(X_i, X_j) = \mathtt{Cov}\left(\sum_{r=1}^{\min(i,j)} Y_r \,, \sum_{r=1}^{\min(i,j)} Y_r \right) = \min(i,j)\,\lambda \tag{4}$$

From (4), we have the following lemma

Lemma 2 Under H_0, and for some k,

$$\mathbb{E}[R_n(k)] = 0$$

$$\forall k_1 \leq k_2 \quad \mathtt{Cov}(R_n(k_1)\,,\, R_n(k_2)) = \frac{\lambda}{\frac{k_2}{n}\,(1 - \frac{k_1}{n})} \tag{5}$$

Remark: From (4) we have: $\mathtt{Var}[R_n(k)] = \dfrac{\lambda}{\frac{k}{n}\,(1 - \frac{k}{n})}$ which is not bounded for k near n. In the rest of our paper, we will penalize the statistic R_n so that the variance remains bounded for all values k (see [6]).

Let φ be a stepwise continuous fonction on $[0,1]$ that satisfies $\displaystyle\int_0^1 \left(\frac{\varphi(t)}{\sqrt{t}}\right)^2 dt < \infty$ and

$$R^{\star}_{n,\varphi}(k) = \varphi(\frac{k}{n})\,\varphi(1 - \frac{k}{n})\,R_n(k) \tag{6}$$

From (5) we have : $\forall\, k_1 \leq k_2$

$$\mathtt{Cov}(R^{\star}_{n,\varphi}(k_1)\,,\, R^{\star}_{n,\varphi}(k_2)) = \frac{\varphi(\frac{k_1}{n})\,\varphi(1 - \frac{k_1}{n})\,\varphi(\frac{k_2}{n})\,\varphi(1 - \frac{k_2}{n})}{\frac{k_2}{n}\,(1 - \frac{k_1}{n})}\lambda \tag{7}$$

and

$$\mathtt{Var}(R^{\star}_{n,\varphi}(k)) = \left(\frac{\varphi(\frac{k}{n})\,\varphi(1 - \frac{k}{n})}{\sqrt{\frac{k}{n}\,(1 - \frac{k}{n})}} \right)^2 \lambda$$

We note that if we take $\varphi(t) = t$, (6) can be written in the form:

$$\mathbb{K}_n(k) = n^{-3/2}(k\,X_n - n\,X_k) \tag{8}$$

For any consistent estimator $\hat{\lambda}_n$ de λ, let

$$\mathbb{K}_n^\star(k) = \frac{1}{\sqrt{\hat{\lambda}_n}}\,\mathbb{K}_n(k)$$

We define $\{\mathbb{K}_n^\star(t)\,;\,t \in [0,1]\}$ as the linear interpolated process of the process $\{\mathbb{K}_n^\star(k)\,;\,1 < k < n\}$.

Theorem 1 Under H_0, $\mathbb{K}_n^\star(t)$ converges in the sense of weak convergence of the finite dimensional distributions to the Brownian Bridge $B(t) = W(t) - t\,W(1)$ where $\{W(t)\,;\,t \in [0,1]\}$ represent the Wiener process.

`Proof.` We can rewrite $\mathbb{K}_n(t)$ as difference of two independent asymptotically gaussian processes. In the fact we have

$$\mathbb{K}_n(t) = \frac{[nt](n-[nt])}{n^{3/2}}\left(\frac{1}{n-[nt]}\sum_{i=[nt]+1}^{n} Y_i - \frac{1}{[nt]}\sum_{i=1}^{[nt]} Y_i\right)$$

where $Y_1, \ldots, Y_n$ are defined in the proof of lemma 1. By using (7)we have

$$\forall t_1 \le t_2 \quad \mathtt{cov}(\mathbb{K}_n(t_1)\,,\,\mathbb{K}_n(t_2)) = \frac{[nt_1]}{n}\,(1 - \frac{[nt_2]}{n})\,\lambda \tag{9}$$

Then

$$\lim_{n\to\infty} \mathtt{Cov}(\mathbb{K}_n^\star(t_1)\,,\,\mathbb{K}_n^\star(t_2)) = t_1\,(1-t_2)$$

Now, for all $t_1 \le t_2$, we have $\mathtt{Cov}(B(t_1), B(t_2)) = t_1\,(1-t_2)$ and also, $\hat{\lambda}_n \underset{n\to\infty}{\xrightarrow{\mathrm{P}}} \lambda$ ■

We have then the following corollaries.

Corollary 1 Under H_0, and for all $\eta \ge 0$, we have :

$$\lim_{n\to\infty} P_{H_0}\,(\sup_{t\in[0,1]} |\mathbb{K}_n^\star(t)| \ge \eta) = P\,(\sup_{t\in[0,1]} |B(t)| \ge \eta) \tag{10}$$

`Proof of corollary 1.` Using the continuity of the function "*sup*", (cf.[2]), on space $\mathcal{C}([0,1])$.

Corollary 2 Let $\hat{\tau}_n = \text{Arg}\sup_{t\in[0,1]} |\mathbb{K}_n^\star(t)|$ and $\tau = \text{Arg}\sup_{t\in[0,1]} |B(t)|$. Under H_0, we have

$$\hat{\tau}_n \xrightarrow[n\to\infty]{\text{P}} \tau \tag{11}$$

`Proof of corollary 2.` The proof is the same as the Lemma 2.1 of [5].

For the behavior under the hypothesis H_1 of $\mathbb{K}_n^\star$, we have the following results.

Lemma 3 Under H_1, and for all $i = 1, \ldots, n$ we have:

$$\mathbb{K}[X_i] = \begin{cases} i\,\lambda & \text{if } 1 \le i \le k \\ \\ k\,\lambda + (i-k)\,(\lambda + \dfrac{\delta}{\sqrt{n}}) & \text{if } k+1 \le i \le n \end{cases} \tag{12}$$

`Proof.` Using (3), writing, X_i in the following form :

$$X_i = \begin{cases} \sum_{j=1}^{i} Y_j & \text{if } 1 \le i \le k \\ \\ \sum_{j=1}^{k} Y_j + \sum_{j=k+1}^{i} Y_j & \text{if } k+1 \le i \le n \end{cases}$$

Lemma 4 Under H_1, and for all k_1 et k_2 we have :

$$\texttt{Cov}(X_{k_1}, X_{k_2}) = \begin{cases} k\,\lambda + (\min(k_1,k_2) - k)\,(\lambda + \dfrac{\delta}{\sqrt{n}}) & \text{if } \min(k_1,k_2) \ge k \\ \\ \min(k_1,k_2)\,\lambda & \text{otherwise} \end{cases} \tag{13}$$

`Proof.` It is enough to note that from (3), for $k \le k_1 \le k_2$, we have :

$$X_{k_1} = \sum_{j=1}^{k} Y_j + \sum_{j=k+1}^{k_1} Y_j \text{ et } X_{k_2} = \sum_{j=1}^{k} Y_j + \sum_{j=k+1}^{k_1} Y_j + \sum_{j=k_1+1}^{k_2} Y_j.$$

Theorem 2 Under H_1, $\mathbb{K}_n^\star$ converges in the sense of weak convergence of the finite dimensional distributions to the process $B_\tau^\star$ defined by:

$$B_\tau^\star(t) = (t \wedge \tau)\,(1 - t \vee \tau)\,\frac{\delta}{\sqrt{\lambda}} + B(t) \tag{14}$$

where $t \wedge \tau = \inf\{t, \tau\}$ and $t \vee \tau = \sup\{t, \tau\}$.

`Proof.` The same argument as in the proof of theorem 1 shows that $\mathbb{K}_n(t)$ converge to gaussian process. From (12), for all $t \in [0,1]$ we have :

$$\mathbb{E}[\mathbb{K}_n(t)] = \begin{cases} \frac{[nt]}{n}\,(1-\frac{[n\tau]}{n})\,\delta & \text{if } t \le \tau \\ \frac{[n\tau]}{n}\,(1-\frac{[nt]}{n})\,\delta & \text{if } t \ge \tau \end{cases} \tag{15}$$

Let $t_1 \le t_2$. By (13), if $[nt_1] \le [nt_2] \le [n\tau]$

$$\mathtt{Cov}(\mathbb{K}_n(t_1), \mathbb{K}_n(t_2)) = \frac{[nt_1]}{n}\,(1-\frac{[nt_2]}{n})\lambda + \frac{[nt_1]}{n}\,\frac{[nt_2]}{n}\,(1-\frac{[n\tau]}{n})\,\frac{\delta}{\sqrt{n}}$$

If $[nt_1] \le [n\tau] \le [nt_2]$

$$\mathtt{Cov}(\mathbb{K}_n(t_1), \mathbb{K}_n(t_2)) = \frac{[nt_1]}{n}\,(1-\frac{[nt_2]}{n})\lambda + \frac{[nt_1]}{n}\,(1-\frac{[nt_2]}{n})\,\frac{[n\tau]}{n}\,\frac{\delta}{\sqrt{n}}$$

If $[n\tau] \le [nt_1] \le [nt_2]$

$$\begin{aligned} \mathtt{Cov}(\mathbb{K}_n(t_1), \mathbb{K}_n(t_2)) &= \frac{[nt_1]}{n}\,(1-\frac{[nt_2]}{n})\lambda \\ &\quad +(1-\frac{[nt_2]}{n})\,\left(\frac{[nt_1]}{n}\,\frac{[n\tau]}{n} + \frac{[nt_1]}{n} - \frac{[n\tau]}{n}\right)\,\frac{\delta}{\sqrt{n}} \end{aligned}$$

Hence

$$\lim_{n\to\infty} \mathtt{Cov}(\mathbb{K}_n(t_1), \mathbb{K}_n(t_2)) = t_1\,(1-t_2)\,\lambda \quad \blacksquare$$

3 Building the test

To build the test, we take as the rejection region

$$\{\sup_{t\in[0,1]} |\mathbb{K}_n^\star(t)| \ge \eta\}$$

For a fixed α in $[0,1]$, by (10), the level η_α is given by (see [9]) :

$$P\,(\sup_{t\in[0,1]} |B(t)| \ge \eta_\alpha) = 2\,\sum_{j=1}^{\infty} (-1)^{j+1}\,e^{-2\,j^2\,\eta_\alpha^2} \tag{16}$$

It is interesting to note that the proposed test has the same asymptotic critical value as the Kolmogorov test.

By (14) we obtain the power of the test:

$$\mathtt{p}(\tau) = P\,(\sup_{t\in[0,1]} |B^{\star}_{\tau}(t)| \geq \eta_{\alpha}) \tag{17}$$

In the following applications, we estimate the power of the test by building the percentiles table of (14) after replacing δ and τ by their estimate. This construction is based on Monte Carlo simulation of the brownian bridge, for this see [4].

4 Applications and Discussion

4.1 HSU case

In this section we apply the previous results to the data used by [7] and [11]. The data represent the number of haemolytic uraemic syndrome (HUS) cases in Birmingham and Newcastle from $1970-1989$.

In [7] the following model was used:

$$Y_i \sim \begin{cases} \mathcal{P}(\lambda_1) & \text{if } i = 1,\ldots,k \\ \\ \mathcal{P}(\lambda_2) & \text{if } i = k+1,\ldots,n \end{cases}$$

where Y_i denotes the number of HUS cases for the year i and $\mathcal{P}(\lambda)$ is the Poisson distribution with mean λ. By using the *Log-likelihood* statistic, for the at most one change point case they found that the change occurs at $k = 11$ for the Birmingham data and $k = 15$ for the Newcastle data. In [11] a Poisson process was used to suppress the hypothesis that the change is an integer. They applied this to the Newcastle data to obtain (maximum likelihood method), the value $\tau = 14.944$ which closely corresponds with the estimate of 15 given by [7].

We have applied the model (1) by taking $X_i = Y_1 + \ldots + Y_i$, where Y_i is the same as in [7].

For the Newcastle data, we have $\hat{\lambda}_n = \dfrac{x_n}{n} = 5$.

k	y_k	x_k	$\lvert\mathbb{K}_n^\star(k)\rvert$	k	y_k	x_k	$\lvert\mathbb{K}_n^\star(k)\rvert$
1	6	6	0.1	11	4	27	2.8
2	1	7	0.3	12	0	27	3.3
3	0	7	0.8	13	4	31	3.4
4	0	7	1.3	14	3	34	3.6
5	2	9	1.6	**15**	3	37	**3.8**
6	0	9	2.1	16	13	50	3
7	1	10	2.5	17	14	64	2.1
8	8	18	2.2	18	8	72	1.8
9	4	22	2.3	19	9	81	1.4
10	1	23	2.7	20	19	100	0

Table 1. Newcastle results

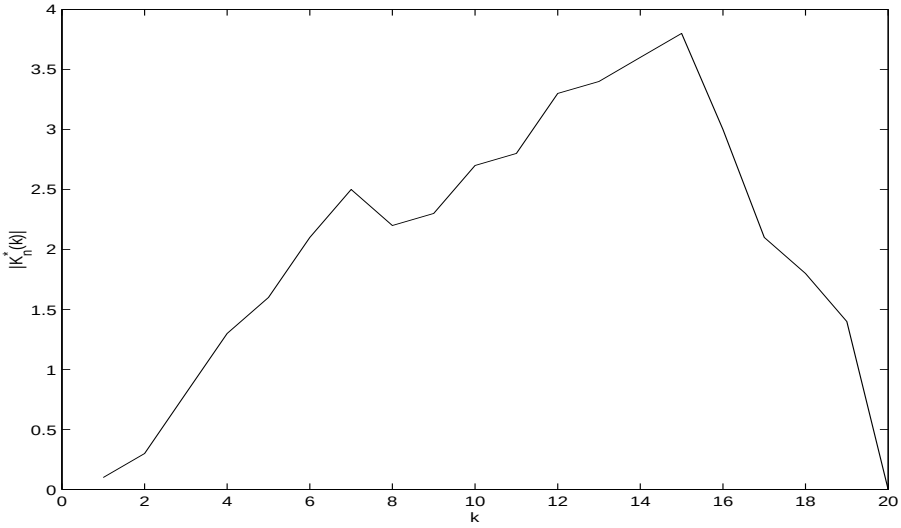

Figure 1: Graph of $\lvert\mathbb{K}_n^\star(k)\rvert$ for the Newcastle data .

The maximum is obtained at $k = 15$, corresponding to the value of the statistic $\lvert\mathbb{K}_n^\star(15)\rvert = 3.8$. For $\alpha = 5\%$, critical value is $\eta_\alpha = 1.36$, then we reject H_0. From (8), we obtain $\hat{\delta} = \lvert\mathbb{K}_n(15)\rvert = 8.50$,which gives an order of 21% for the estimation of the power.

For the Birmingham data, we have $\hat{\lambda}_n = \dfrac{x_n}{n} = 5.65$.

k	y_k	x_k	$\lvert\mathbb{K}_n^\star(k)\rvert$	k	y_k	x_k	$\lvert\mathbb{K}_n^\star(k)\rvert$
1	1	1	0.43744	**11**	4	18	**4.1533**
2	5	6	0.49858	12	0	25	4.0263
3	3	9	0.74787	13	4	36	3.523
4	2	11	1.0912	14	3	40	3.6782
5	2	13	1.4346	15	3	47	3.5512
6	1	14	1.872	16	13	57	3.142
7	0	14	2.4035	17	14	73	2.1684
8	0	14	2.935	18	8	89	1.1947
9	2	16	3.2784	19	9	98	0.87957
10	1	17	3.7158	20	19	113	0

Table 2. Birmingham results

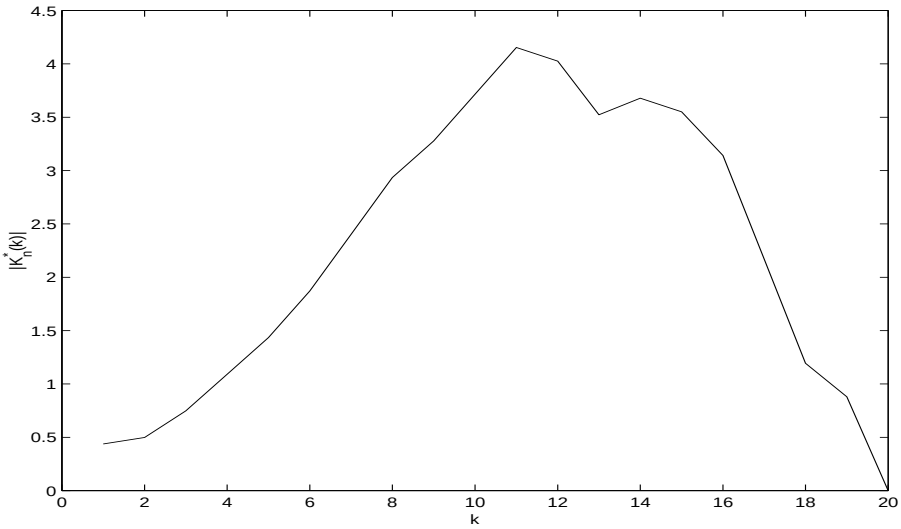

Figure 2: Graph of $\lvert\mathbb{K}_n^\star(k)\rvert$ for the Birmingham data .

The maximum is obtained at $k = 11$, corresponding to the value of the statistic $\lvert\mathbb{K}_n^\star(11)\rvert = 4.15$. For $\alpha = 5\%$, critical value is $\eta_\alpha = 1.36$, then we reject H_0. From (8), we obtain $\hat{\delta} = \lvert\mathbb{K}_n(11)\rvert = 9.87$,which gives an order of 41% for the estimation of the power.

We obtain the same point estimations for the AMOC case as in [7] and [11].

4.2 AIDS case

The Cuban case is an interesting one because the epidemiological information is reliable, due to the existence of a centralized system, with a National Reference Center that validates and confirms the new AIDS cases diagnosed, at regular intervals. Also not only AIDS but also HIV infection is under a mandatory report directive since 1985. This means that a high percentage of the HIV infected persons are known to the Health System, and have special medical care that makes the report of the dates of AIDS. In the case of the AIDS epidemic in Cuba, we will look at the number of new AIDS cases reported by semester since the start of the epidemic in 1986 [12]. If we take the new AIDS cases in Cuba by semester, following [1], the data can be regarded as independent Poisson random variables.

For this data, we have $\hat{\lambda}_n = \frac{x_n}{n} = 32.15$.

sem-year	k	y_k	x_k	$\lvert\mathbb{K}_n^\star(k)\rvert$	sem-year	k	y_k	x_k	$\lvert\mathbb{K}_n^\star(k)\rvert$
1st 1986	1	1	1	1.0775	2nd 1992	14	37	179	9.3781
2nd 1986	2	4	5	2.0512	1st 1993	15	43	222	9.0029
1st 1987	3	3	8	3.0595	2nd 1993	16	39	261	8.7662
2nd 1987	4	8	16	3.8949	1st 1994	17	37	298	8.5985
1st 1988	5	6	22	4.7994	2nd 1994	18	55	353	7.8084
2nd 1988	6	8	30	5.6348	1st 1995	19	50	403	7.1912
1st 1989	7	3	33	6.6431	2nd 1995	20	66	469	6.0206
2nd 1989	8	10	43	7.4093	1st 1996	21	48	517	5.4725
1st 1990	9	12	55	8.1064	2nd 1996	22	51	568	4.8207
2nd 1990	10	16	71	8.6651	1st 1997	23	55	623	4.0306
1st 1991	11	20	91	9.0854	2nd 1997	24	74	697	2.5833
2nd 1991	**12**	17	108	**9.6095**	1st 1998	25	55	752	1.7931
1st 1992	13	34	142	9.5457	2nd 1998	26	84	836	0

Table 3. Cuban AIDS results

The maximum is obtained at $k = 12$, corresponding to the value of the statistic $|\mathbb{K}_n^\star(12)| = 9.61$. For $\alpha = 5\%$, critical value is $\eta_\alpha = 1.36$, then we reject H_0. From (8), we obtain $\hat{\delta} = |\mathbb{K}_n(12)| = 54.49$,which gives an order of 98% for the estimation of the power.

We find that there is a change-point that corresponds to the second semester of 1991. This is consistent with what is known of the AIDS epidemic in Cuba. There were some persons infected before 1981-82, but the transmission of

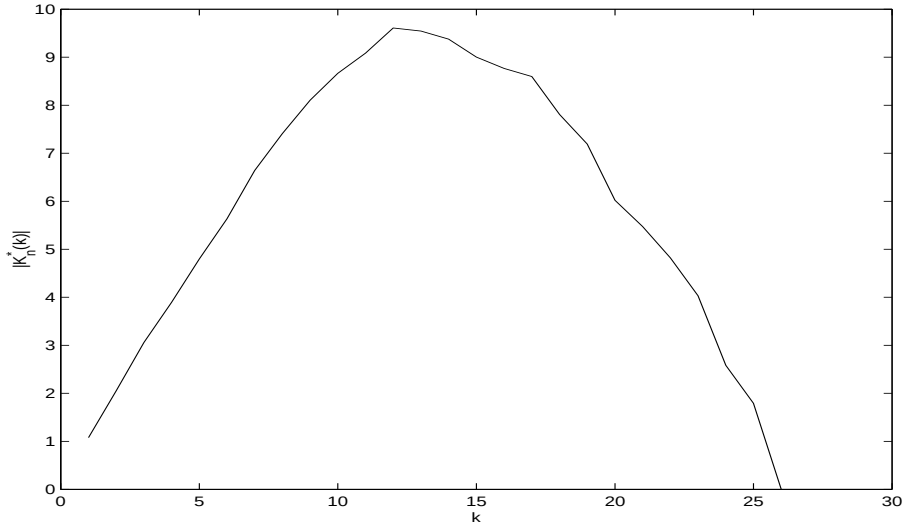

Figure 3: Graph of $|\mathbb{K}_n^\star(k)|$ for the Cuban AIDS data.

the HIV virus started to establish itself in the period 1982-86. Considering that the average incubation period was estimated at 9.2 years, an increase of the number of new AIDS cases was to be expected at some time in the period 1991-92. This is confirmed by our analysis of the data.

Taking into account the power, the model we propose in this paper, is best adapted to the AIDS data than to the HSU data in the sens that in the AIDS data case the jump is larger than in the two other cases.

References

[1] Bailey, N. Operational modelling of HIV/AIDS to assist public health control. *In Epidemic Models: Their Structure and Relation to Data (ed. D. Mollison 404-415* (1995). Cambridge University Press.

[2] Billingsley, P.Convergence of Probability Measures. (1968) *Wiley, New York.*

[3] Csörgö, M. and Horváth, L. Limit theorems in Change Point Analysis. (1997) J. Wiley, New York.

[4] Eastwood, B. J. and Eastwood, V. R. On the distribution of some gaussian functionals related to changepoint problems. Technical Report

146, Laboratory for Research in Statistics and Probability, Carleton University, Ottawa (1990).

[5] Ferger, D. Change-point estimators in case of small disorders. *J. Statist. Plann. Inference,* **40,**. *33-49* (1994).

[6] Ghorbanzadeh, D. and Lounes, R. Deux procédures de détection de rupture pour des observations poissonniennes groupées. *Revue de Statistique Appliqueé,* **XLIV(3),**. *47-61* (1996).

[7] Henderson, R. and Matthews, J. N. S. An Investigation of Changepoints in the Annual number of Cases of Haemolytic Uraemic Syndrome. *Appl. Statist* **42 3**. *461-471* (1993).

[8] Raftery, A.E. and Akman, V.E. Bayesian analysis of a Poisson process with a change-point. *Biometrika.***73** *85-89* (1986).

[9] Shorack, G.R. and Wellner, J.A. Empirical Processes With Applications to Statistics. *J. Wiley, New York* (1986).

[10] Simpkin, J.M. and Downham, D.Y. Testing for a change-point in registry data with an example on Hypospadias. *Statistics in Medicine.* **7** *387-393* (1988).

[11] West, R. W. and Ogden, R. T. Continuous-time estimation of a change-point in a Poisson process. *Journ. Statist. Comput. Simul.* **56** *293-302* (1997).

[12] The HIV-AIDS Epidemic in Cuba. Internal Report. Departamento de Epidemiología, Sanatorio Santiago de Las Vegas. Cuba 1999.