Cox process goodness-of-fit test. A Matlab file.

Bouzas, P.R.*
paula@ugr.es
Department of Statistics and Operations Research
University of Granada, 18071-Granada, Spain

Ruiz-Fuentes, N.
nfuentes@ujaen.es
Department of Statistics and Operations Research
University of Jaén, 23071-Jaén, Spain

October 7, 2011

Abstract

The Cox Process (CP) models many real phenomena dealing with counting data. Having observed sample paths of a counting process in a discrete set of time points and assuming that the phenomenon can be modeled by a Cox process or compound Cox process, an important task is to decide if those paths fit a given model. A goodnes-of-fit test to assess the coherence of the new observed data with the given Cox process has been proposed by the authors, taking into account if the process is parametrically known or it has to be estimated. This paper deals with a computational tool to support the test.

Keywords: Cox process, compound Cox process, goodness-of-fit test, simultaneous inference.

*This work was partially supported by project MTM2010-20502 of Dirección General de Investigación y Gestión del Plan Nacional I+D+I, Ministerio de Ciencia e Innovación, Spain.

1 Introduction

CpGFT (Bouzas et al. [5]) is a MATLAB function to test if a sample of observed paths of a counting process M(t) fits a given Cox Process, $N(t)$. The hypothesis test assesses the coherence of new observed data with a proposed CP, with known or estimated moments. For the second case, the authors suggest as an accurate estimation of the intensity the one proposed in [3].

A goodness-of-fit test for functional data was derived (see Bouzas et al. [4]), it is based on a simple discrete estimator of the intensity process of the CP at each observation point. The null hypothesis states that the intensity mean is the proposed one at each observation point. Afterwards, a multiple testing criterion (Benjamini Hochberg, [1]) yields a decision on the coherence of the observed complete sample path with the proposed CP.

Having observed s sample paths of $M(t)$ of dimension $p*m+1$, the discrete estimator of the intensity process $\lambda(t)$ for each sample path is calculated by splitting it up in m subpaths of $p+1$ points.The subpath i contains the values of $M(t)$ at time points $t_{(i-1)*p}$ to t_{i*p} for $i : 1, ..., m$. Then, the point estimator proposed in Bouzas et al. [2] is applied. With the s estimated sample paths of the intensity at each time point, an experimental value of the mean of the intensity is calculated and the conclusion of the hypothesis test is achieved at each t_j. Taking into account the information of all isolated tests, the simultaneous inference leads us to a final decision.

2 Function CpGFT, syntax and parameters

In this section we outline a description of the parameters to evaluate the `CpGFT` function. The syntax of the function is given by:

```
[status]=CpGFT(type,paths,timepoints,mean_sd,alpha)
```

The function is evaluated in the following set of parameters:

type Refers to features of the intensity process assumed as pattern.

- Set **type=0** if the intensity process is unknown. Its properties, the mean $\mu(t)$ and the standard deviation $\sigma(t)$, need to be estimated by the user in an external procedure.

– Set **type=1** if the intensity process is already known and so its mean and standard deviation, that is, the proposed underlying model is parametrically specified.

paths Put here an array $s \times (p*m+1)$, s paths of dimension $p*m+1$, these are the observed paths of the counting process $M(t)$. They are used to calculate a discrete estimation of the intensity process and then a point estimator of the intensity mean. As usual, $s \geq 30$ if the intensity cannot be assumed as gaussian.

timepoints Introduce here a $p+1$ vector $(t_0, t_1, ..., t_p)$ of equally or unequally spaced time points. If they are unequally spaced, at least delays between observations must be the same for all m subpaths in which all the s paths will be splitted up.

mean_sd It is a $2 \times (p+1)$ vector. The first row contains the values of the estimated mean of the intensity process at each time point (type=0) or the values of the actual mean of the intensity process at each time point (type=1). The second row contains the values of the estimated standard deviation of the intensity process at each time point (type=0) or the values of the actual standard deviation of the intensity process at each time point (type=1).

alpha Significance level (default = 0.05).

3 MATLAB function code

```
function [status]=CpGFT(type,paths,timepoints,mean_sd,alpha)
%Setting the parameters
if nargin < 5,
   alpha = 0.05;
end
if nargin < 4,
   error('Requires at least four input arguments.');
end
  T=timepoints;
  h=diff(timepoints);
  M=paths;
```

```
  s=size(M,1);
  p=length(h);
  m=(size(M,2)-1)/p;
  Mu=mean_sd(1,:);
  SD=mean_sd(2,:);
%Discrete estimation of the intensity process
%at the observation points
L=[];
for i=1:s
    L_aux=[];
    for j=1:m
        L_aux(j,:)=diff(M(i,(j-1)*p+1:j*p+1))./h;
    end
    L(i,:)=mean(L_aux);
end
Mu_exp=mean(L,1);
%Application of the multiple testing criterion
pv=[];
if type==1
   for j=1:p
       pv(j)=2*(1-normcdf(abs((Mu_exp(j)-Mu(j))/(SD(j)/sqrt(s)))));
   end
elseif type==0
   for j=1:p
       pv(j)=2*(1-tcdf(abs((Mu_exp(j)-Mu(j))/(SD(j)/sqrt(s))),s));
   end
end
PV=sort(pv');
vale=[];
for i=1:p
    if PV(i)>alpha*i/p
       vale(i)=1;
    else
       vale(i)=0;
    end
end
if sum(vale)==p
   status=1;
```

```
else
   status=0;
end
```

4 Results

The function returns status=1 if data are coherent with the proposal, estimated or real process, and status=0 otherwise.

As the Cox Process is quite flexible and suitable to model many counting phenomena, this seems to be an appropriate tool to assess if new observed data of such counting processes come from a given one. The function is based on a point estimator of the intensity process, that have been reported as well-behaved, and in simultaneous inference to achieve a final decision on the main task.

One example of application of this test with real counting data can be found in Bouzas et al. [4], it is applied to data of emitted particles of two different isotopes. For each new path, a point estimator and an estimated intensity mean and variance (type=0) were taken into account. The aim was to evaluate the proper working of a counting device and the test produced fairly good outcomes.

The function has been also used to detect deviations from the pattern in simulated contaminated paths of parametrically defined Cox processes with quite good results, detecting even slight deviations from the original patterns.

5 Acknowledgments

This work was partially supported by project MTM2010-20502 of Dirección General de Investigación and Gestión del Plan Nacional I+D+I and grants FQM-307 and FQM-246 of Conserjería de Innovación de la Junta de Andalucía, all in Spain.

References

[1] Y. Benjamini and Y. Hochberg, Controling the false discovery rate: a practical and powerful approach to multiple testing, *J. Roy. Statist. Soc. Ser. B*, 57, 289–300, 1995.

[2] P.R. Bouzas, A.M. Aguilera, M.J. Valderrama and N. Ruiz-Fuentes, On the structure of the stochastic process of mortgages in Spain, *Computation. Stat.*, 21, 73–89, 2006b.

[3] P.R. Bouzas, A.M. Aguilera and N. Ruiz-Fuentes, Functional estimation of the random rate of a Cox process, *Methodol. Comput. Appl. Probab.*, 2010a, 1-13, DOI: 10.1007/s11009-010-9173-z.

[4] P.R. Bouzas, N. Ruiz-Fuentes, A. Matilla, A.M. Aguilera and M.J. Valderrama, A Cox model for radioactive counting measure: inference on the intensity process, *Chemometr. Intell. Lab.*, 103, 116–121, 2010b.

[5] P.R. Bouzas, N. Ruiz-Fuentes, CpGFT Matlab file, v.1.0.2011, *Free access DIGIBUG.*