Tester si un processus de Poisson est modifié en admettant que certains événements ponctuels se regroupent en classes

Franz Streit

Tester si un processus de Poisson est modifié en admettant que certains événements ponctuels se regroupent en classes.

Streit Franz

Section de Mathématiques de l' Université de Genève,
Case postale 64, CH-1211 Genève 4, Suisse.

Résumé.
En observant la réalisation d'un processus ponctuel on est amené à se poser la question si cette réalisation a été engendrée par un processus homogène de Poisson ou par un modèle stochastique plus compliqué. Quand on envisage comme alternative un processus de Poisson à événements ponctuels regroupés en classes, on trouve peu d'indications utiles et explicites dans la littérature pour aider à faire le choix entre ces hypothèses. Cela s'explique probablement par le fait que la fonction de vraisemblance du processus de Poisson à événements ponctuels regroupés en classes ('Poisson cluster processes') est une expression assez complexe parce que le nombre des regroupements possibles augmente rapidement avec le nombre des événements ponctuels. Il s'avère cependant que d'habitude seulement relativement peu de termes de cette fonction de vraisemblance fournissent une contribution non-nulle à la statistique du score efficace. Cette constatation permet dans certains cas d'indiquer d'une manière explicite des tests localement les plus puissants pour distinguer un processus homogène de Poisson des alternatives susmentionnées.

Mots clés: Processus, Statistique mathématique.

Summary
When observing the realisation of a point process it is natural to try to find out whether this realisation has been generated by a homogeneous Poisson process or by a more complicated stochastic model. When a Poisson cluster process is taken into consideration as alternative one finds little useful and explicite indications which would help to choose between these hypotheses. This is probably due to the fact that the likelihood function for Poisson cluster processes is quite a complex expression since the number of ways to form clusters increases rapidly with the number of point events. However it turns out that only relatively few terms of this likelihood function yield a non-zero contribution to the efficient score statistic. This fact permits in certain cases to derive a locally most powerful test of a homogeneous Poisson point process against cluster alternatives.

Texte de la communication

En observant la réalisation d'un processus ponctuel on est dans bien des cas intéressé à connaître le modèle stochastique sous-jacent. La supposition la plus standard consiste à adopter l'hypothèse que la réalisation soit engendrée par un processus ponctuel de Poisson. Quand on choisit comme hypothèse alternative que la réalisation provienne d'un processus de Poisson à événements ponctuels regroupés en classes on trouve peu d'indications utiles et explicites dans la littérature statistique pour élaborer des tests pour cette situation. Comme on sait, la construction de tests statistiques optimaux se fait en déterminant et analysant la fonction de vraisemblance. Or, il s'avère que la fonction de vraisemblance du processus de Poisson à événements ponctuels regroupés en classes (en anglais 'Poisson cluster process') est d'une complexité décourageante [voir Daley et Vere-Jones (1988,p.501)], ce qui explique probablement le manque de méthodes inférentielles relatives à ce problème.

Nous proposons dans ce contexte de se servir des tests localement les plus puissants pour faire le choix entre les hypothèses. On désire donc à discriminer les configurations des événements ponctuels sans regroupement des configurations où la formation de classes est prévue selon le modèle stochastique, mais se réalise avec une faible probabilité. Il s'avère que cette formulation du problème conduit à une simplification considérable. En effet, les regroupements en plusieures classes se font avec petites probabilités et donnent des contributions zéro à la statistique du test.

Les processus ponctuels considérés sont tous supposés simples et à espace des états $\mathcal{R}^+ = \{t : t > 0\}$. On désignera les variables aléatoires et leurs valeures réalisées par la même lettre, mais en utilisant la forme majuscule respectivement minuscule. Soit $t_i[i = 1, 2, \ldots]$ le temps d'occurrence du i-ème événement ponctuel après le début de l'expérience à $t = 0 = t_0$ et $u_i = t_i - t_{i-1}.\, n(t)$ désignera le nombre des événements ponctuels se présentant dans la période $(0, t]$ et $G[h]$ la fonction génératrice du processus ponctuel qui spécifie en résumé succinct le modèle stochastique [voir Daley et Vere-Jones (1988, p.220)]

$G[h] = E[\exp[\int_{\mathcal{R}^+} \log(h(x)dN(x)]]$

(où $0 \leq h \leq 1$ et $h = 1$ à l'extérieur d'un sous-ensemble borné de $\mathcal{R}^+$). $\sim$ signifie 'suit la distribution', χ^2_ν désigne une variable aléatoire de chi-carré à ν degrés de liberté, $R_{(0,b]}$ la distribution uniforme à support $(0, b]\,[b > 0]$, $Exp[\lambda]$ la distribution exponentielle de paramètre $\lambda > 0$ et $N^*(a, g)$ la distribution normale d'espérance mathématique a et de variance $g > 0$.

Nous déterminons la version de la formule pour la fonction de vraisemblance dont nous avons besoin en choisissant d'abord l'intervalle fixe (0,t] comme période d'observation du processus. Les événements ponctuels observés sont soit des événements ponctuels primaires engendrés par un processus de Poisson homogène à taux d'intensité global μ, soit des événements ponctuels secondaires associés à un événement ponctuel primaire. Un événement ponctuel primaire se présentant à $t = x$ donne lieu avec probabilité δ à un (seul) événement ponctuel secondaire se trouvant à $t = y$, où y est choisi à l'aide de la

fonction de densité $f(y|x)$. Il s'agit de tester les hypothèses $H_0 : \delta = 0$ (représentant le processus de Poisson pur) versus $H_1 : \delta > 0$ (représentant le processus contaminé par des regroupements en classes). Sous l'hypothèse générale $H_0 \cup H_1$ la fonction génératrice du processus observé est donnée par
$G[\tilde{h} : \delta] = \exp[\int_0^t (1-\delta)(\tilde{h}(x)-1)\mu dx + \int_0^t \int_0^\infty \delta(\tilde{h}(x)\tilde{h}(y)-1) f y|x) dy \mu dx$
$= \exp[-\mu t] \cdot \exp[\int_0^t h(x)(1-\delta)\mu dx + \int_0^t \int_0^t \delta h(x)h(y)f(y|x)dy\mu dx + \int_0^t \delta h(x)(1-F(t|x))\mu dx]$,
où $\tilde{h}(z) = h(z)\,[0 < z \leq t]$ et $\tilde{h}(z) = 1[z > t]$ et $F(t|x) = \int_0^t f(y|x)dy$ et où on a tenu compte qu'un événement ponctuel est soit primaire soit secondaire et qu'un événement ponctuel secondaire à position $y > t$ n'est pas observé.
La valeur de la densité de Janossi locale de la réalisation observée vaut donc
$j_{n(t)}(t_1, \ldots, t_{n(t)}|(0,t] : \delta) =$
$(1-\delta)^{n(t)}\mu^{n(t)}e^{-\mu t} + \delta(1-\delta)^{n(t)-2}\mu^{n(t)}e^{-\mu t}\sum_{i_1=1}^{n(t)}\sum_{i_2=1, i_2\neq i_1}^{n(t)}(f(t_{i_2}|t_{i_1})/\mu) +$
$\delta(1-\delta)^{n(t)-1}\mu^{n(t)}e^{-\mu t}\sum_{i=1}^{n(t)}(1-F(t|t_i)) + o(\delta)$
et donne selon la théorie générale des processus ponctuels une expression valide pour $\delta \to 0$ de la fonction de vraisemblance [voir Daley et Vere-Jones (1988,p.497)].
On trouve
$\partial(\log(j_{n(t)}(t_1, \ldots, t_{n(t)}|(0,t] : \delta)/\partial\delta|_{\delta=0} = -\sum_{i=1}^{n(t)} F(t|t_i) +$
$\sum_{i_1=1}^{n(t)}\sum_{i_2=1, i_1\neq i_2}^{n(t)}(f(t_{i_2}|t_{i_1})/\mu)$,
et on obtient pour la statistique du score efficace [voir Cox et Hinkley(1979, p.107 et p.113)]
$T_{Y\sim F(|x),\mu}(0) = -\sum_{i=1}^{N(t)} F(t|T(i)) + \sum_{i_1=1}^{N(t)}\sum_{i_2=1, i_1\neq i_2}^{N(t)}(f(T(i_2)|T(i_1))/\mu)$.
En remplaçant la période d'observation du processus $(0,t]$ par $(0,t_{n_0}]$, donc en se basant sur la réalisation du processus ponctuel jusqu'à l'apparition du n_0-ième événement ponctuel où n_0 est un entier fixe > 1, la déduction analogue aboutit à la statistique
$T_{Y\sim F(|x),\mu,n_0}(0) = -\sum_{i=1}^{n_0} F(T(n_0)|T(i)) + \sum_{i_1=1}^{n_0}\sum_{i_2=1, i_1\neq i_2}^{n_0}(f(T(i_2)|T(i_1))/\mu)$.
Dans la suite nous discutons quelques cas spéciaux permettant de décrire les tests d'une manière explicite.

Cas I
Choix de la période d'observation: $(0, t_{n_0}]$.
Choix de la fonction de densité de Y : $f(y|x) = f(y) = 1_{(0,t_{n_0}]}(y)/t_{n_0}$,
où $1_A(z) = 1\,[z \in A]$ et $1_A(z) = 0\,[z \notin A]$. La statistique du test prend la forme
$T_{Y\sim R_{(0,T(n_0)]},\mu,n_0}(0) = n_0(((n_0-1)/\mu T(n_0)) - 1)$,
la région critique consistant des valeures élevées de $t_{Y\sim R_{(0,T(n_0)]},\mu,n_0}(0)$. On constate que $T_{Y\sim R_{(0,T(n_0)]},\mu,n_0}(0)$ dépend du hazard uniquement en termes de la variable aléatoire $T(n_0)$. Sous H_0 $\mu T(n_0)$ suit la loi gamma à fonction de densité $f_{\mu T(n_0)}(x) = x^{n_0-1}e^{-x}/(n_0 - 1)!\,[x > 0]$ [voir Cox et Lewis (1968, p.24)]. Grâce à la relation liant les distributions gamma aux distributions de chi-carré, on obtient pour la région critique du test localement le plus puissant

$2\mu t_{n_0} < \chi^2_{2n_0}(\alpha)$ où $P(\chi^2_{2n_0} < \chi^2_{2n_0}(\alpha)) = \alpha$.
Comme pour des valeurs élevées de n_0 on a $T(n_0) \sim N^*(n_0/\mu, n_0/\mu^2)$, on peut se servir pour n_0 grand aussi de la région critique asymptotique
$n_0^{-1/2}(\mu t_{n_0} - n_0) < z(\alpha)$ où $Z \sim N^*(0,1)$ et $P(Z < z(\alpha)) = \alpha$.

Cas II
Choix de la période d'observation: $(0, t]$.
Choix de la fonction de densité de Y : $f(y|x) = f(y) = 1_{(0,t]}(y)/t$.
Pour ce choix la statistique de score efficace se réduit à
$T_{Y\sim R_{(0,t]},\mu}(0) = (N(t)(N(t)-1)/(\mu t)) - N(t)$.
On a $E[T_{Y\sim R_{(0,t]},\mu}(0) : \delta = 0] = (\mu t + (\mu t)^2 - \mu t)/(\mu t) - \mu t = 0$
et un petit calcul à l'aide de formules pour les moments non-centrés de la distribution de Poisson [voir Johnson, Kotz et Kemp(1992, p.156)] donne le résultat
$Var[T_{Y\sim R_{(0,t]},\mu}(0) : \delta = 0] = 2 + \mu t$.
Asymptotiquement pour $t \to \infty$ et en tenant μ fixe, on trouve
$T_{Y\sim R_{(0,t]},\mu}(0) \sim (N(t)/(\mu t))(N(t) - \mu t - 1) \sim N(t) - \mu t - 1 \sim N^*(-1, \mu t)$ et donc
$(T_{Y\sim R_{(0,t]},\mu}(0) + 1)/(\mu t)^{1/2} \sim (T_{Y\sim R_{(0,t]},\mu}(0))/\sqrt{2 + \mu t} \sim N^*(0,1)$.
La région critique asymptotique d'un test localement le plus puissant de H_0 versus H_1 est donc spécifiée par
$[2 + \mu t]^{-1/2} t_{Y\sim R_{(0,t]},\mu} > z(1-\alpha)$ où $P(Z > z(1-\alpha)) = \alpha$.

Cas III
Choix de la période d'observation: $(0, t]$.
Choix de la fonction de densité de Y : $f(y|x) = 1_{(x,x+t]}(y)/t$.
Sous ces conditions on obtient pour la statistique du score efficace
$T_{Y-X\sim R_{(0,t]},\mu}(0) = -\sum_{i=1}^{N(t)}(1 - (T(i)/t)) + (N(t)(N(t)-1)/(2\mu t)$.
Sous H_0 on a conditionellement à $N(t) = n$
$\sum_{i=1}^{N(t)}(1 - (T(i)/t))|N(t) = n \sim \sum_{i=1}^{n} R_{(0,1]}(i)$
où $R_{(0,1]}(1), \ldots, R_{0,1]}(n)$ désigne une suite de variables aléatoires indépendantes et identiquement distribuées selon $R_{(0,1]}$ [voir (Albrecht,1968)].
Il en découle que
$E[T_{Y-X\sim R_{(0,t]},\mu}(0)|N(t) : \delta = 0] = -(N(t)/2) + (N(t)(N(t)-1))/(2\mu t)$ et
$Var[T_{Y-X\sim R_{(0,t]},\mu}(0)|N(t) : \delta = 0] = N(t)/12$.
Pour $t \to \infty$ et sous H_0
$T_{Y-X\sim R_{(0,t]},\mu}(0)/\sqrt{N(t)} = -(\sum_{i=1}^{N(t)}(1 - (T(i)/t) - 1/2)/\sqrt{N(t)}) + ((N(t)/(\mu t)) \cdot ((N(t) - \mu t - 1)/(2\sqrt{N(t)}))) \sim N^*(0, 1/3)$
selon le théorème de limite centrale de Anscombe-Rényi [voir DasGupta(2008 ,p.71)].
La région critique du test asymptotique de niveau de signification α pour $t \to \infty$ est donc

donnée par
$(\sqrt{3/\mu t})\, t_{Y-X\sim R_{(0,t]},\mu}(0) > z(1-\alpha)$.

Cas IV
Choix de la période d'observation: $(0, t_{n_0}]$.
Choix de la fonction de densité de Y : $f(y|x) = 1_{\{y>x\}}(y)\tilde{f}(y-x)/(\mu e^{-\mu(y-x)})$
où $\tilde{f}$ est une fonction de densité concentrée sur $\mathcal{R}^+$. Ce choix de la loi de la position de Y correspond à la situation, où l'événement ponctuel secondaire ne précède jamais l'événement ponctuel primaire correspondant et où l'action du processus homogène de Poisson est interrompue pendant le temps d'attente sur l'arrivée de l'événement ponctuel secondaire associé. La statistique du test égale à
$T^-_{Y-X\sim\tilde{F}(|x),\mu,n_0}(0) \sim \sum_{l=2}^{n_0}((f(T(l)|T(l-1))-1)) = \sum_{l=2}^{n_0}((\tilde{f}(U(l))/(\mu e^{-\mu U(l)}))-1)$.
Sous H_0 les variables aléatoires sont indépendantes et identiquement distribuées selon $Exp[\mu]$ et on peut donc appliquer le théorème limite centrale standard, ce qui conduit à
$T^-_{Y-X\sim\tilde{F}(|x),\mu,n_0}(0) \sim N^*(0, Var[T^-_{Y-X\sim\tilde{F}(|x),\mu,n_0}(0) : \delta = 0])$
pour $n_0 \to \infty$ à condition que $0 < Var[T^-_{Y-X\sim\tilde{F}(|x),\mu,n_0}(0) : \delta = 0] < \infty$.
Par exemple, si $\tilde{f}(\tau) = \lambda e^{-\lambda\tau}\,[\tau > 0]$ on trouve
$Var[T^-_{Y-X\sim Exp[\lambda],\mu,n_0}(0) : \delta = 0] = (n_0-1)[\lambda^2/(\mu(2\lambda-\mu))-1]$
pour $\lambda > \mu/2$ et cette variance est positive à l'exception du cas $\lambda = \mu$ (qui est sans intérêt).
Quand $\tilde{f} = 1_{(0,c]}(\tau)/c$ on obtient
$Var[T^-_{Y-X\sim R_{(0,c]},\mu,n_0}(0) : \delta = 0] = (n_0-1)(e^{c\mu}-1-c^2\mu^2)/(c\mu)^2 > 0$.
Basé sur ces informations les régions critiques des tests se déterminent selon la façon usuelle.

Bibliographie

[1] Albrecht,P. (1982) Testing the goodness of fit of a mixed Poisson process. *Insurance: Mathematics and Economics*, 1,27–33.
[2] Cox,D.R. et Lewis,P.A.W. (1968) *The statistical analysis of series of events*, Methuen, London.
[3] Cox,D.R. et Hinkley,D.V.(1979) *Theoretical statistics*, Chapman and Hall, London.
[4] Daley,D.J. et Vere-Jones,D. (1988) *An introduction to the theory of point processes*, Springer, New York.
[5] DasGupta,A. (2008) *Asymptotic theory of statistics and probability*, Springer, New York.
[6] Johnson,N.L., Kotz,S. et Kemp,A. (1992) *Univariate discrete distributions*, J.Wiley, New York.