

Une application simple du théorème de Donsker pour la construction de tests de qualité d'ajustement de modèles de l'intensité conditionnelle d'un processus de comptage

Christophe Pouzat

Laboratoire de Physiologie Cérébrale, CNRS UMR 8118, Université
Paris-Descartes

e-mail: christophe.pouzat@gmail.com

web:

http://www.biomedicale.univ-paris5.fr/phycerv/C_Pouzat.html

SELECT : 26 Octobre 2009

Sommaire

Introduction

Définitions

Interprétation de λ et transformation du temps

Transformation du temps sur données réelles

Mouvement brownien canonique et théorème de Donsker

Comment reconnaît-on un ivrogne à son trajet dans la rue ?

Approche de Fierro

Approche de la "région de prédiction"

Conclusions et remerciements

Sommaire

Introduction

Définitions

Interprétation de λ et transformation du temps

Transformation du temps sur données réelles

Mouvement brownien canonique et théorème de Donsker

Comment reconnaît-on un ivrogne à son trajet dans la rue ?

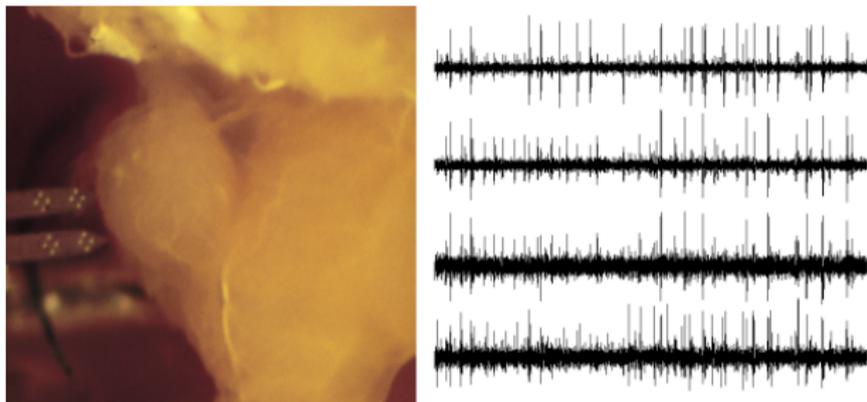
Approche de Fierro

Approche de la "région de prédiction"

Conclusions et remerciements

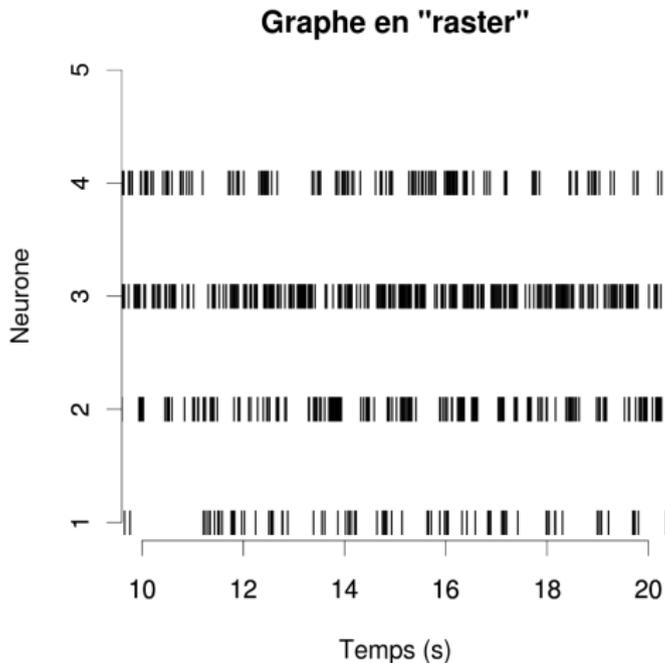
Un contexte concret : les trains de potentiels d'action

Vue de l'extérieur, l'activité des neurones se manifeste par l'émission d'impulsions électriques très brèves : **les potentiels d'actions**.



A gauche, le cerveau et la sonde d'enregistrement qui comporte 16 électrodes (les points brillants). La largeur d'une branche de la sonde est de $80 \mu m$. A droite, 1 sec d'enregistrement sur 4 électrodes. Les pics sont des potentiels d'action.

Exemple de trains de potentiels d'actions



Après une étape de pré-traitement "assez lourde" appelée **tri des potentiels d'action**, on obtient les trains de PAs.

Pourquoi modéliser les trains de PAs ?

- ▶ Une hypothèse de travail centrale en Neurosciences est que les temps d'apparition des PAs, par opposition à leurs formes, sont le seul support de transmission de l'information entre régions du cerveau.
- ▶ Cette hypothèse légitime l'étude des trains de PAs en tant que séquence de points sur la demi-droite réelle (représentant le temps) sans nécessairement tenir compte des mécanismes biophysiques les générant.

Train de PAs et processus de comptage

- ▶ Dans ce qui suit nous allons assimiler un train de PAs à un **processus ponctuel** auquel nous allons associer un **processus de comptage** (ou compteur).

Processus de comptage ou compteur

- ▶ A des points $\{t_i\}_{i \in \mathbb{N}^*} \subset \mathbb{R}^{+*}$, tels que $t_j > t_i$ pour $j > i$, on associe le **processus de comptage** ou **compteur** défini par :

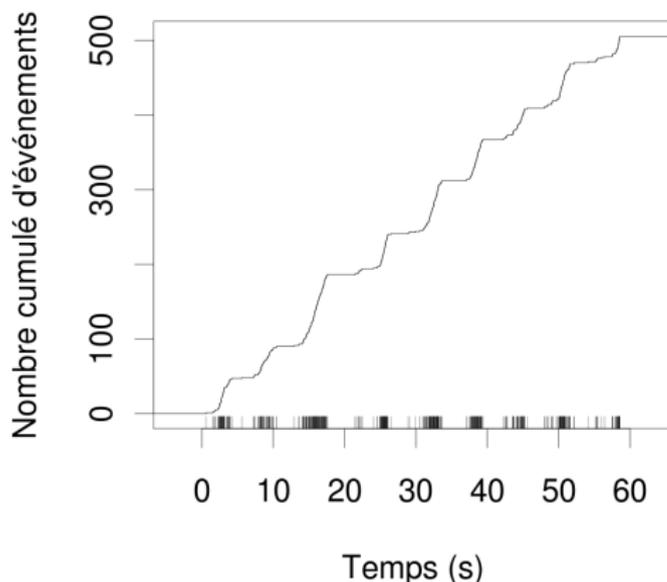
$$N(t) \equiv \#\{t_i \mid t_i \leq t\}$$

- ▶ Lorsque plusieurs neurones sont enregistrés simultanément on introduit un compteur, $N^{(k)}(t)$, par neurone. Ici l'exposant (k) est un indice de neurone. Alors pour que le processus multidimensionnel, $(N^{(1)}(t), \dots, N^{(K)}(t))$, reste "simple", on rajoute la condition :

$$t_i^{(k)} \neq t_j^{(l)}, \quad \forall i, j \in \mathbb{N}^* \quad \text{et} \quad \forall k \neq l, \quad k, l \in \{1, \dots, K\}$$

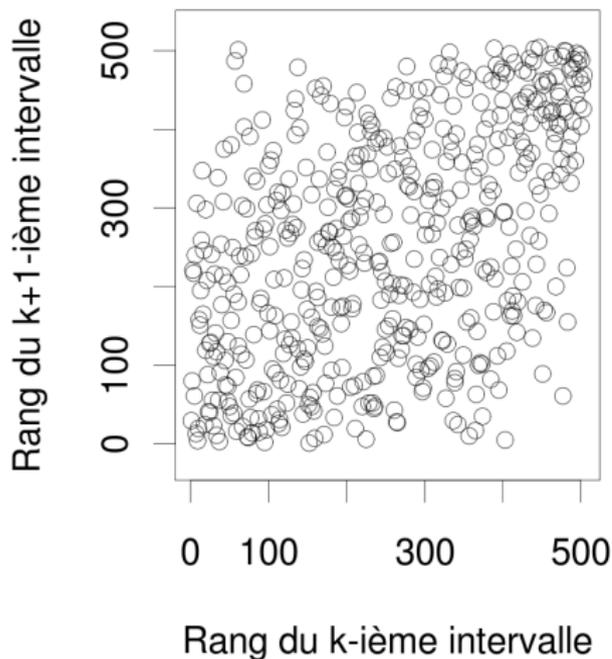
où $t_i^{(k)}$ est le temps du i -ème PA du neurone k .

Exemple de processus de comptage observé



Ici la répartition des temps des PAs sur $[0, 59]$ est bien trop régulière pour un processus de Poisson homogène.

Les trains de PAs stationnaires correspondent rarement à des processus de renouvellement



Corrélation des rangs
des intervalles inter PAs
successifs pour le train
précédent.

Conséquences de l'inadéquation des processus de renouvellement

- ▶ Si une modélisation basée sur un processus de renouvellement été adéquate, notre problème “statistique” se réduirait au choix d'une densité d'intervalles entre PAs appropriée et serait relativement simple.
- ▶ Comme le processus de renouvellement n'est que rarement pertinent, nous allons avoir recours au formalisme plus général basé sur la notion d'**intensité conditionnelle**.

Conséquences de l'inadéquation des processus de renouvellement

- ▶ Si une modélisation basée sur un processus de renouvellement été adéquate, notre problème “statistique” se réduirait au choix d'une densité d'intervalles entre PAs appropriée et serait relativement simple.
- ▶ Comme le processus de renouvellement n'est que rarement pertinent, nous allons avoir recours au formalisme plus général basé sur la notion d'**intensité conditionnelle**.
- ▶ Trois problèmes doivent alors être résolus :
 - ▶ Des modèles de l'intensité conditionnelle (ou fonction de risque instantané) doivent être construits et ajustés aux données.
 - ▶ **L'adéquation des modèles aux données doit être évaluée.**
 - ▶ Quand plusieurs modèles sont “satisfaisant” il faut faire un choix.

Sommaire

Introduction

Définitions

Interprétation de λ et transformation du temps

Transformation du temps sur données réelles

Mouvement brownien canonique et théorème de Donsker

Comment reconnaît-on un ivrogne à son trajet dans la rue ?

Approche de Fierro

Approche de la "région de prédiction"

Conclusions et remerciements

Quelques définitions

- ▶ Nous allons continuer en introduisant quelques définitions (un peu) plus “formelles”, essentiellement pour clarifier nos notations.
- ▶ Pour des définitions correctes voir par exemple :
 - ▶ Rebolledo (1979) *La méthode des martingales appliquée à l'étude de la convergence en loi des processus*. Mémoires de la Société Mathématiques de France, **62**, pp 102–114.
 - ▶ Fleming & Harrington (2005) *Counting Processes and Survival Analysis*. Wiley

Filtration, histoire et intensité conditionnelle

- ▶ Les probabilistes des processus introduisent la **filtration** ou l'**histoire** : une famille de tribus croissantes, $(\mathcal{F}_t)_{0 \leq t \leq \infty}$, telle que toute l'information relative au processus au temps t , peut-être représentée par un élément de \mathcal{F}_t .
- ▶ L'**intensité conditionnelle** du compteur $N(t)$ est alors définie par :

$$\lambda(t | \mathcal{F}_t) \equiv \lim_{h \downarrow 0} \frac{\text{Prob}\{N(t+h) - N(t) = 1 | \mathcal{F}_t\}}{h}$$

Remarques

- ▶ En principe on s'attend à ce que \mathcal{F}_t fasse intervenir les trains de PAs de **tous** les neurones qui influent sur le neurone étudié.
- ▶ En pratique on a toujours enregistré au mieux qu'une toute petite fraction des ces neurones...
- ▶ On est alors "forcé" de considérer des approximations de $\lambda(t | \mathcal{F}_t)$ du type :

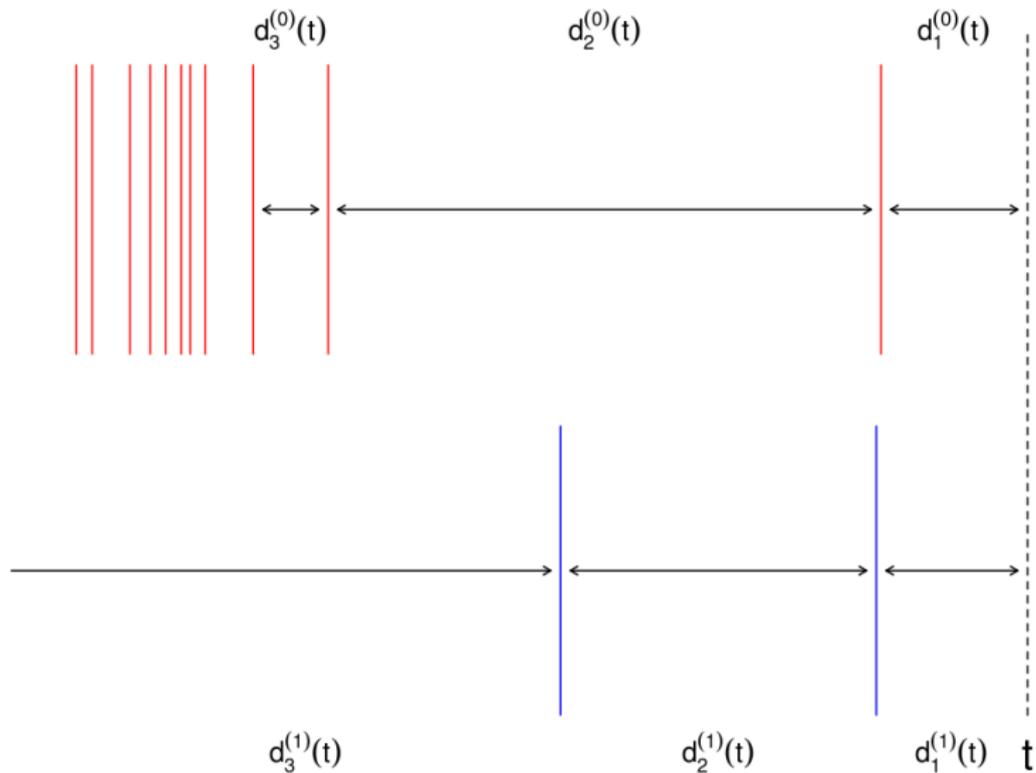
$$f \left(d_1^{(0)}(t), \dots, d_{m_0}^{(0)}(t), d_1^{(1)}(t), \dots, d_{m_1}^{(1)}(t), \dots, d_1^{(K)}(t), \dots, d_{m_K}^{(K)}(t) \right)$$

où, (0) est l'indice du neurone "étudié" et (1), ..., (K), sont les indices de K autres neurones enregistrés, où :

$$d_i^{(k)}(t) \equiv d_{i-1}^{(k)}(t) - \max \left\{ t_j^{(k)} \mid t_j^{(k)} < d_{i-1}^{(k)}(t) \right\}, \quad \text{et} \quad d_0^{(k)}(t) = t.$$

On a de plus supposé implicitement que le processus est stationnaire.

Les $d_i^{(k)}(t)$ illustrés



Remarques sur notre approximation de λ

- ▶ Lorsque nous employons pour $\lambda(t | \mathcal{F}_t)$ l'approximation :

$$f \left(d_1^{(0)}(t), \dots, d_{m_0}^{(0)}(t), \dots, d_1^{(K)}(t), \dots, d_{m_K}^{(K)}(t) \right)$$

nous devons décider **quels neurones inclure** parmi ceux enregistrés et, **pour chaque neurone $k \in \{0, \dots, K\}$, quelle valeur de m_k** utiliser. Nous sommes donc immédiatement confronté à un problème de choix de “modèle”.

- ▶ Une fois ces choix effectués, il nous faut trouver la fonction f .
- ▶ La seule information directement exploitable fournie par des années de recherche en biophysique des neurones est que :

$$\lim_{d_1^{(0)}(t) \downarrow 0} f \left(d_1^{(0)}(t), \dots, d_{m_0}^{(0)}(t), \dots, d_1^{(K)}(t), \dots, d_{m_K}^{(K)}(t) \right) = 0$$

Sommaire

Introduction

Définitions

Interprétation de λ et transformation du temps

Transformation du temps sur données réelles

Mouvement brownien canonique et théorème de Donsker

Comment reconnaît-on un ivrogne à son trajet dans la rue ?

Approche de Fierro

Approche de la "région de prédiction"

Conclusions et remerciements

Interprétation de λ

- ▶ Nous avons défini l'intensité du compteur $N(t)$ par :

$$\lambda(t | \mathcal{F}_t) \equiv \lim_{h \downarrow 0} \frac{\text{Prob}\{N(t+h) - N(t) = 1 | \mathcal{F}_t\}}{h}$$

- ▶ Ce que nous interprétons comme suit :

$$\begin{aligned} \text{Prob}\{N(t+h) - N(t) = 0 | \mathcal{F}_t\} &= 1 - \lambda(t | \mathcal{F}_t) h + o(h) \\ \text{Prob}\{N(t+h) - N(t) = 1 | \mathcal{F}_t\} &= \lambda(t | \mathcal{F}_t) h + o(h) \\ \text{Prob}\{N(t+h) - N(t) > 1 | \mathcal{F}_t\} &= o(h) \end{aligned}$$

où $o(h)$ est telle que :

$$\lim_{h \rightarrow 0} \frac{o(h)}{h} = 0$$

Probabilité de "pas d'événement" sur $]t, t + s]$

- ▶ On cherche la probabilité, $P_0(t, s)$, de ne pas avoir de PA sur l'intervalle $]t, t + s]$ lorsque s n'est pas infiniment petit.
- ▶ Nous divisons $]t, t + s]$ en sous intervalles disjoints :

$$\mathcal{I}_j =]t + j h, t + (j + 1) h], \quad j = 0, \dots, k - 1 \quad \text{et} \quad h = s/k$$

- ▶ Alors :

$$P_0(t, s) = \prod_{j=0}^{k-1} (1 - \lambda(t + j h | \mathcal{F}_{t+jh}) h + o(h))$$

$$P_0(t, s) = \exp \left[\sum_{j=0}^{k-1} \log (1 - \lambda(t + j h | \mathcal{F}_{t+jh}) h + o(h)) \right]$$

$$P_0(t, s) = \exp \left[- \sum_{j=0}^{k-1} \lambda(t + j h | \mathcal{F}_{t+jh}) h + o(k h) \right]$$

$$P_0(t, s) \xrightarrow{k \rightarrow \infty} \exp \left(- \int_t^{t+s} \lambda(u | \mathcal{F}_u) du \right)$$

Densité des intervalles et intensité cumulée

- ▶ La fonction $P_0(t, s)$ dont nous venons d'obtenir l'expression est une **fonction de survie** à partir de laquelle on obtient la densité des intervalles entre PAs :

$$\begin{aligned} p(t, s) &= -\frac{dP_0(t, s)}{ds} \\ &= \lambda(t + s | \mathcal{F}_{t+s}) \exp\left(-\int_t^{t+s} \lambda(u | \mathcal{F}_u) du\right) \end{aligned}$$

- ▶ Si maintenant nous introduisons l'**intensité cumulée** :

$$\Lambda(t) \equiv \int_0^t \lambda(u | \mathcal{F}_u) du$$

- ▶ On voit que :

$$p(t, s) ds = d\Lambda(t + s) \exp - (\Lambda(t + s) - \Lambda(t))$$

Transformation du temps

- ▶ Si $\lambda(u | \mathcal{F}_u) > 0$ alors le processus Λ est strictement croissant et la **transformation du temps** :
 $t \in \mathbb{R}^{+*} \mapsto \Lambda(t) \in \mathbb{R}^{+*}$ admet une transformation inverse.
- ▶ La densité des intervalles entre PAs successifs après transformation du temps :
 $\{t_1, \dots, t_n\} \mapsto \{\Lambda(t_1) \equiv \Lambda_1, \dots, \Lambda(t_n) \equiv \Lambda_n\}$ est alors donnée, d'après la dernière équation de la diapo précédente, par :

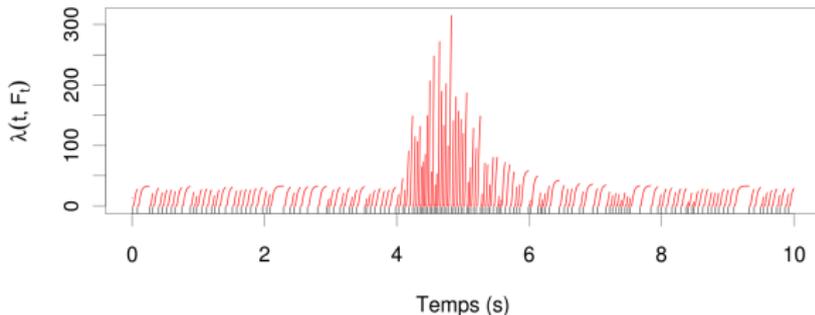
$$p(\Lambda_{i+1} - \Lambda_i) = \exp -(\Lambda_{i+1} - \Lambda_i)$$

C'est une densité exponentielle de paramètre 1.

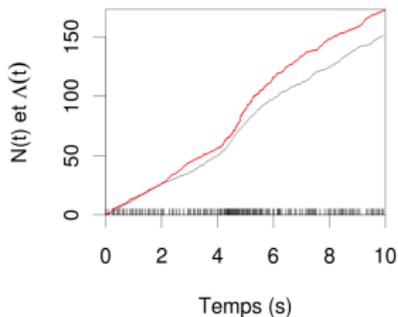
- ▶ Le processus ponctuel observé $\{\Lambda_1, \dots, \Lambda_n\}$ est donc une observation d'un **processus de Poisson de paramètre 1**.

Illustration de la transformation du temps sur données simulées

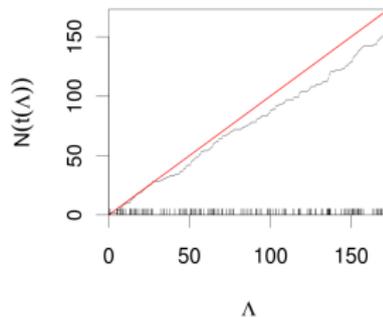
Processus d'intensité et séquence d'événements



N et Λ vs t



N et Λ vs Λ



Sommaire

Introduction

Définitions

Interprétation de λ et transformation du temps

Transformation du temps sur données réelles

Mouvement brownien canonique et théorème de Donsker

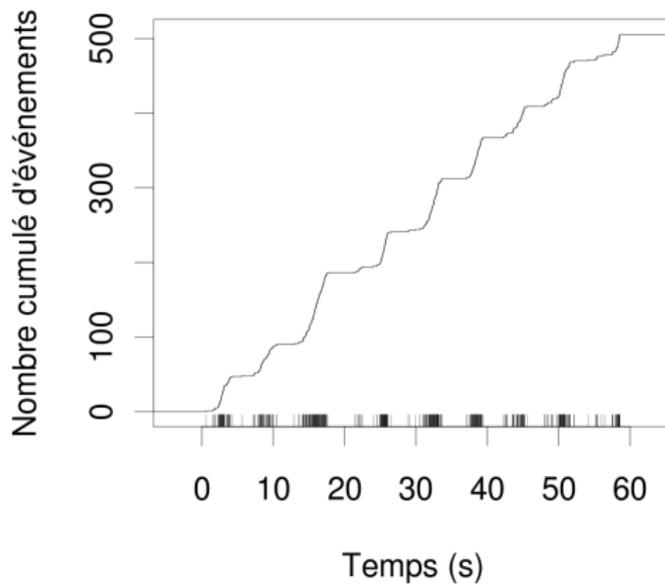
Comment reconnaît-on un ivrogne à son trajet dans la rue ?

Approche de Fierro

Approche de la "région de prédiction"

Conclusions et remerciements

Les données



Remarques

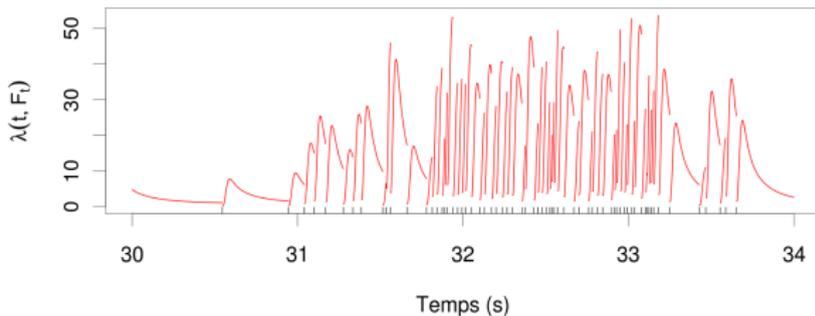
- ▶ Nous n'avons pas le temps aujourd'hui de donner suffisamment de détails sur la manière dont l'estimation de l'intensité est effectuée.
- ▶ Succinctement, nous employons une discrétisation du temps introduite par D Brillinger (1992, JASA, 87 : 270) qui ramène le problème dans le cadre des modèles linéaires généralisés (régression logistique).
- ▶ L'estimation non paramétrique du modèle est alors obtenue avec des splines de lissage (Gu, 2002, *Smoothing Spline ANOVA Models*, Springer).
- ▶ Pour les données précédentes, nous ajustons sur les 29,5 premières seconde un modèle du type :

$$f \left(d_1^{(0)}(t), d_2^{(0)}(t) \right)$$

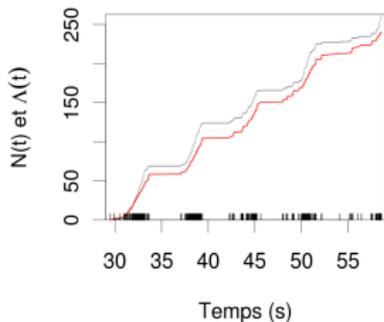
- ▶ Nous transformons ensuite le temps des PAs de la seconde partie du jeu de données avec ce modèle.

Illustration de la transformation du temps sur données réelles

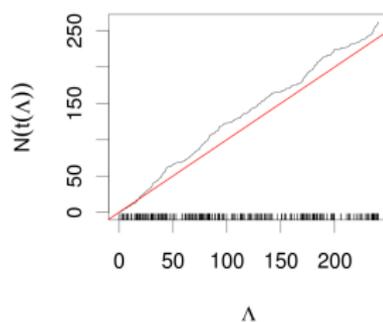
Processus d'intensité et séquence d'événements



N et Λ vs t



N et Λ vs Λ



Sommaire

Introduction

Définitions

Interprétation de λ et transformation du temps

Transformation du temps sur données réelles

Mouvement brownien canonique et théorème de Donsker

Comment reconnaît-on un ivrogne à son trajet dans la rue ?

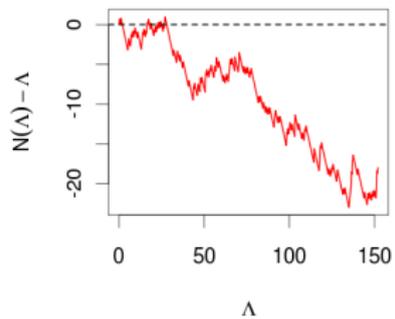
Approche de Fierro

Approche de la "région de prédiction"

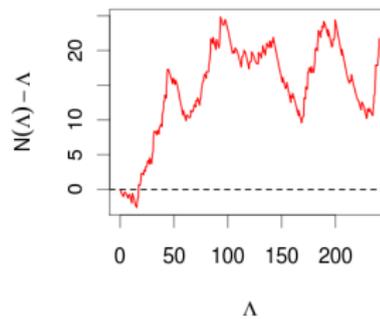
Conclusions et remerciements

$N(\Lambda) - \Lambda$ vs Λ

Données simulées

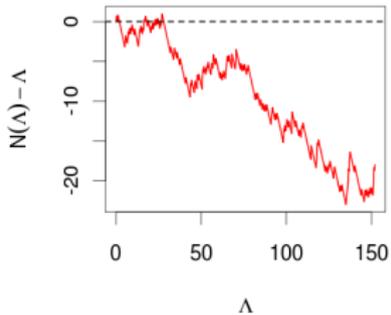


Données réelles

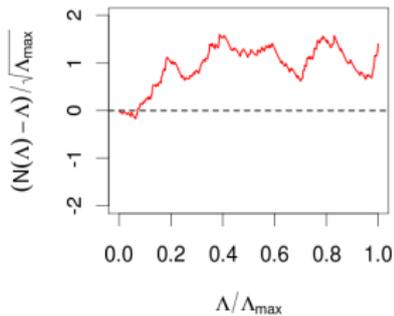
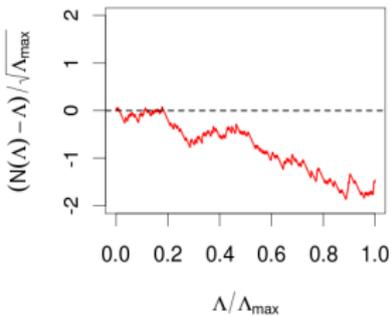
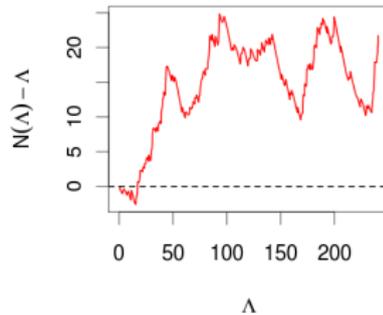


$$(N(\Lambda) - \Lambda) / \sqrt{\Lambda_{\max}} \text{ VS } \Lambda / \Lambda_{\max}$$

Données simulées



Données réelles



Théorème de Donsker

- ▶ Soit X_1, X_2, \dots une suite de variables aléatoires IID telles que $EX_i = 0$ et $EX_i^2 = 1$, alors la suite de processus :

$$S_n(t) = \frac{1}{\sqrt{n}} \sum_{i=0}^{\lfloor nt \rfloor} X_i, \quad 0 \leq t \leq 1, \quad X_0 = 0$$

converge en loi vers le mouvement brownien canonique.

- ▶ Pour une preuve du théorème de Donsker voir :
 - ▶ R Durrett (2009) *Probability : Theory and Examples*. CUP. Sec. 7.6, pp 323-329.
 - ▶ P Billingsley (1999) *Convergence of Probability Measures*. Wiley. p 121.

Application du théorème de Donsker. 1

- ▶ Si notre modèle est bon, alors les $\{\Lambda_1, \Lambda_2, \dots, \Lambda_n\}$ constituent une observation d'un processus de Poisson de paramètre 1.
- ▶ Les X_i de notre énoncé du théorème peuvent alors correspondre aux $\Lambda_{i+1} - \Lambda_i - 1$ (du fait de la loi exponentielle de paramètre 1 des intervalles).
- ▶ Alors le graphe de :

$$\frac{(\Lambda(t) - \Lambda(t_1)) - (N(t) - 1)}{\sqrt{N(t_{max}) - 1}} \text{ vs } \frac{\Lambda(t) - \Lambda(t_1)}{\Lambda(t_{max}) - \Lambda(t_1)}, \text{ pour } t_1 \leq t \leq t_{max}$$

va “ressembler de plus en plus” au chemin du mouvement brownien canonique sur $[0, 1]$.

Application du théorème de Donsker. 2

- ▶ Les X_i de l'énoncé du théorème peuvent encore être donnés par : $N(k+1) - N(k) - 1$, où $k \in \mathbb{N}$ et où le processus de comptage est défini sur l'échelle de temps transformé (du fait de la loi de Poisson de paramètre 1 suivie par le nombre d'événements générés par un processus de Poisson de paramètre 1 sur une durée de 1).
- ▶ Alors le graphe de :

$$\frac{N(t) - \Lambda(t)}{\sqrt{[\Lambda(t_{max})]}} \text{ vs } \frac{\Lambda(t)}{[\Lambda(t_{max})]} \text{ pour } 0 \leq t \leq \max\{t \mid \Lambda(t) \leq [\Lambda(t_{max})]\}$$

va “ressembler de plus en plus” au chemin du mouvement brownien canonique sur $[0, 1]$.

Remarques

- ▶ Si pour $0 \leq t \leq \max\{t \mid \Lambda(t) \leq \lfloor \Lambda(t_{max}) \rfloor\}$ nous définissons :

$$n_{max} \equiv \lfloor \Lambda(t_{max}) \rfloor, \quad x \equiv \frac{\Lambda(t)}{n_{max}} \quad \text{et} \quad \tilde{N}(u) \equiv N\left(\Lambda^{-1}(u)\right)$$

- ▶ Alors notre dernier résultat devient :

$$\frac{\tilde{N}(n_{max}x) - n_{max}x}{\sqrt{n_{max}}} \xrightarrow{\text{Loi}} B(t)$$

où B désigne le mouvement brownien canonique.

- ▶ C'est un résultat classique qu'on trouve par exemple dans :
 - ▶ Rebolledo (1979) p 107.
 - ▶ Andersson & Britton (2000) *Stochastic Epidemic Models and Their Statistical Analysis*. Springer. p 39.

Alors quoi de neuf ?

- ▶ Notre proposition, qui semble être originale, est d'utiliser la convergence en loi du processus de comptage "compensé" et exprimé sur une nouvelle échelle de temps vers un mouvement brownien canonique **quand le modèle est bon** pour tester si notre modèle est effectivement bon.
- ▶ Pour cela il faut que nous devenions capables de reconnaître un mouvement brownien canonique sur $[0, 1]$ lorsque nous en voyons un...

Sommaire

Introduction

Définitions

Interprétation de λ et transformation du temps

Transformation du temps sur données réelles

Mouvement brownien canonique et théorème de Donsker

Comment reconnaît-on un ivrogne à son trajet dans la rue ?

Approche de Fierro

Approche de la "région de prédiction"

Conclusions et remerciements

Comment reconnaître un mouvement brownien : L'approche de Fierro

- ▶ Raúl Fierro dans son article “A test of goodness of fit testing for stochastic intensities associated to counting processes” (2003) *Statistics & Probability Letters* **64** : 287-294, suggère d'utiliser comme Statistique :

$$S^2 \equiv \int_0^1 B(t)^2 dt$$

- ▶ Il donne l'expression de la transformée de Laplace de la loi de S^2 :

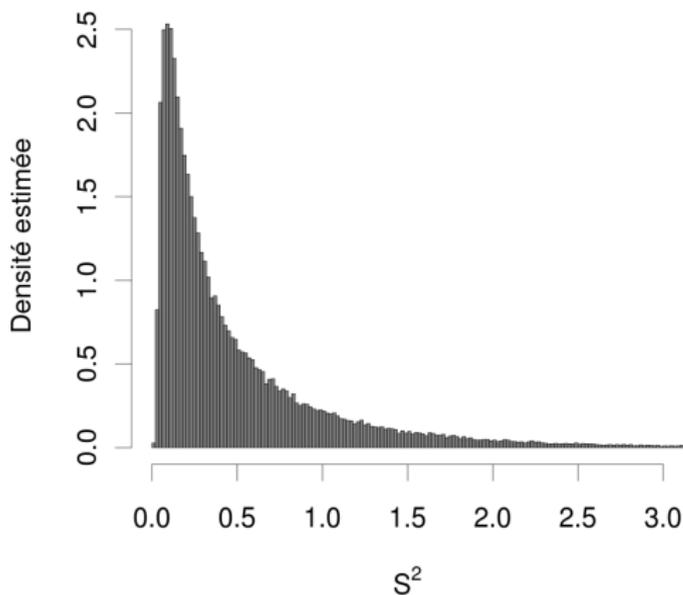
$$E \exp(-\nu S^2) = \left(\sqrt{\cosh(\sqrt{2\nu})} \right)^{-1/2}$$

- ▶ D'où nous déduisons que :

$$E S^2 = 1/2 \quad \text{et} \quad \text{Var } S^2 = 1/3$$

Estimation de la densité de S^2

$\widehat{\text{pdf}}(S^2)$ ($n = 10^5$)



Il est risqué de résumer une telle loi par son approximation gaussienne.

Approche de la "région de prédiction"

- ▶ Soit $B(t)$ un mouvement brownien canonique et soit $f(t)$ une "frontière". On définit le **temps de premier passage** $T(f)$ par :

$$T(f) \equiv \inf\{t \mid B(t) \geq f(t)\}$$

- ▶ Si on arrive à exprimer (ou estimer) la loi de $T(f)$ "en fonction de" f , on peut chercher $f_{0.975}$ telle que :
 $\text{Prob}(T(f_{0.975}) > 1) = 0.975$.
- ▶ Il suffit alors de définir :

$$\tilde{T}(f_{0.975}) \equiv T(f_{0.975}) \wedge T(-f_{0.975})$$

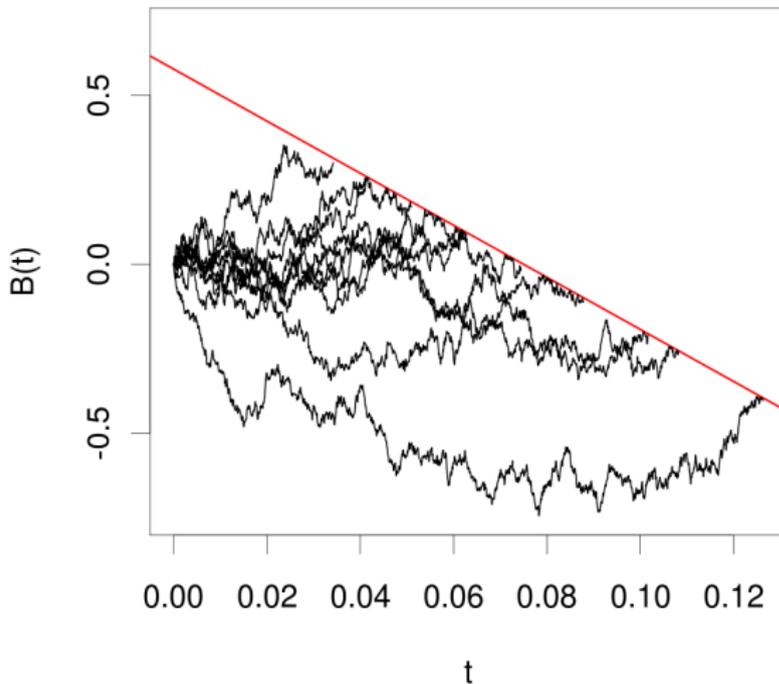
et la propriété de symétrie du mouvement brownien nous donne :

$$\text{Prob}\left(\tilde{T}(f_{0.975}) > 1\right) = 0.95$$

- ▶ Alors le domaine $[0, 1] \times [-f_{0.975}, f_{0.975}]$ contiendra **dans leur totalité** 95% des mouvements browniens observés sur $[0, 1]$.

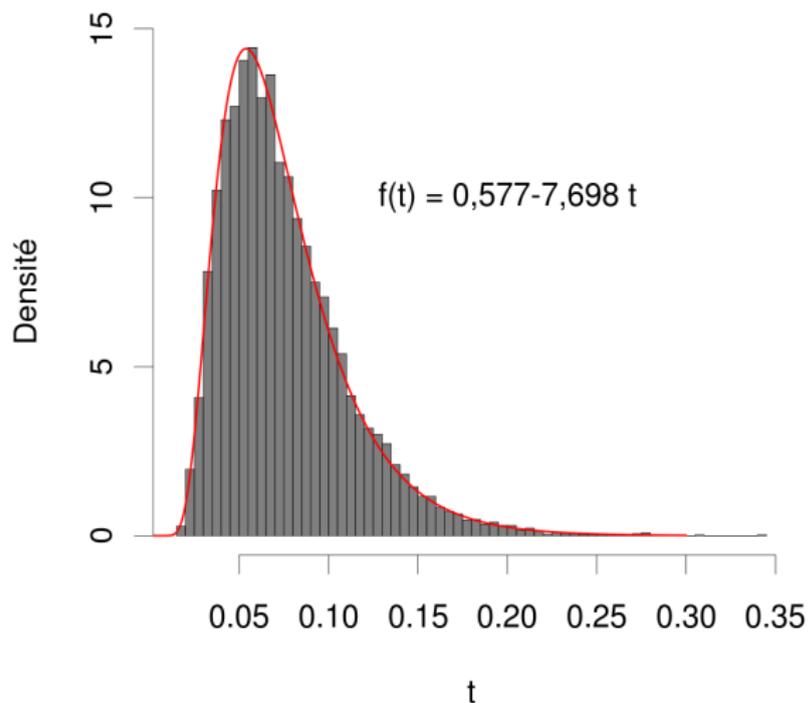
Illustration du temps de premier passage

$$f(t) = 0,577 - 7,698 t$$



Loi d'un temps de premier passage

Densité du temps de premier passage



Loi gaussienne inverse d'un temps de premier passage

- ▶ Whitmore et Sheshadri montrent dans un élégant papier de 1987 (*The American Statistician* **41** : 280-281) que la loi du premier temps de passage à travers une frontière :

$$f(t) = a + b t, \quad \text{avec } a > 0 \quad \text{et} \quad b < 0$$

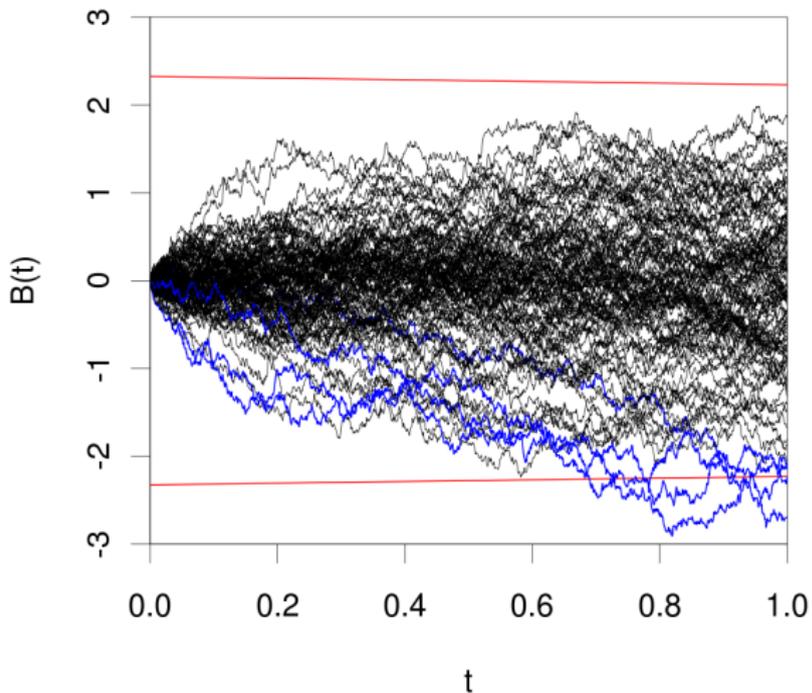
est une loi gaussienne inverse de densité :

$$g(t) = \frac{a}{\sqrt{2 \pi t^3}} \exp\left(-\frac{(a + b t)^2}{2 t}\right)$$

- ▶ Une optimisation numérique montre alors que $G(1) = 0.025$ pour $a = 2.324$ et $b = -0.095$, G désignant ici la fonction de répartition de la loi gaussienne inverse.

Région de prédiction à frontières linéaires

n = 100



Région de prédiction de surface minimale

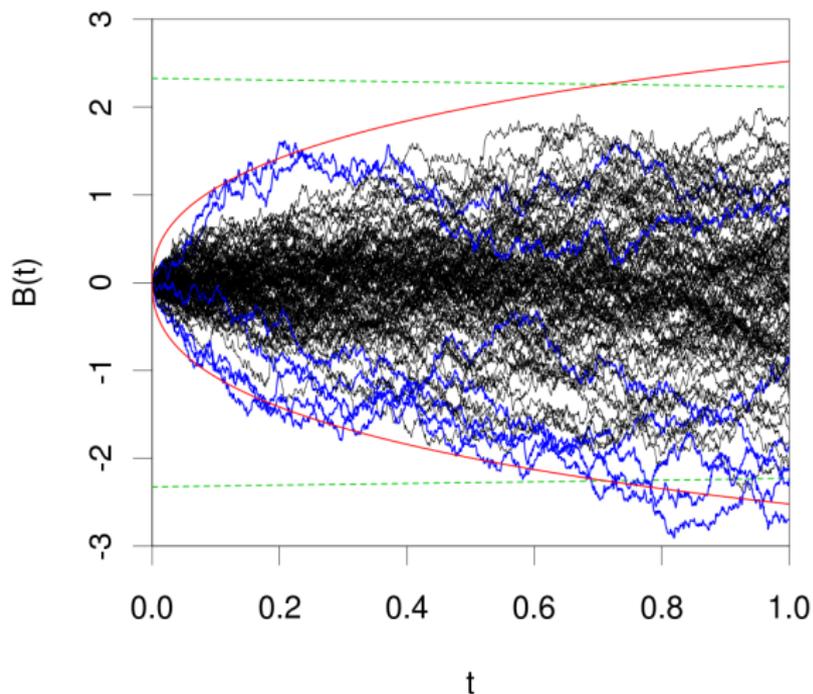
- ▶ Dès lors qu'une approche basée sur une région de prédiction est adoptée, il est raisonnable pour ne pas dire souhaitable d'utiliser une **région de surface minimale**.
- ▶ Dans un article (pas franchement trivial) de 2007, Kendall, Marin et Robert (Brownian Confidence Bands on Monte Carlo Output, *Statistics and Computing*, **17** : 1-10) ont montré que la frontière "supérieure" de cette surface minimale est donnée par :

$$u^*(t) \equiv \sqrt{-W_{-1}(-(\kappa t)^2)} \sqrt{t}, \quad \text{pour } \lambda t \leq 1/\sqrt{e}.$$

où W_{-1} est la "branche réelle secondaire" de la fonction W de Lambert définie comme la solution de : $W \exp W = x$ et où κ est un paramètre à ajuster pour obtenir la probabilité de recouvrement désirée.

Région de prédiction minimale à 95%

n = 100



Comment trouver κ pour avoir une probabilité de recouvrement donnée ?

- ▶ La mise en oeuvre pratique de tests de qualité d'ajustement basés sur une région de prédiction requiert que nous soyons capable de construire une région dont nous contrôlons la probabilité de recouvrement.
- ▶ Dit autrement, il faut que je vous explique comment j'ai pu trouver le κ de la dernière figure qui donne une probabilité de 95%.
- ▶ La solution (numérique) à ce problème est donnée dans un article de Loader et Deely (1987, Computations of boundary crossing probabilities for the Wiener process. *J. Statist. Comput. Simulation* **27** : 95-105).

Une équation intégrale de Volterra

Dans leur article Loader et Deely montrent que la fonction de répartition $G(T)$ du temps de premier passage à travers une frontière $f(t)$ est solution d'une équation intégrale de Volterra de première espèce :

$$H(t) = \int_0^t K(t, u) dG(u)$$

où :

$$H(t) = \Phi\left(-\frac{f(t)}{\sqrt{t}}\right) + \exp\left(-2(f(t) - t)\right) \Phi\left(\frac{-f(t) + 2t}{\sqrt{t}}\right)$$

Φ désignant la fonction de répartition d'une loi normale réduite, et :

$$K(t, u) = \Phi\left(\frac{f(u) - f(t)}{\sqrt{t-u}}\right) + \exp\left(-2(f(t) - f(u) - (t - u))\right) \Phi\left(\frac{f(u) - f(t) + 2(t-u)}{\sqrt{t-u}}\right)$$

Solution numérique par la méthode du point milieu

- ▶ Cette équation intégrale de Volterra peut-être résolue numériquement par la méthode du point milieu.
- ▶ On procède en choisissant : $0 = t_0 < t_1 < \dots < t_n = t$, $t_j = j h$ pour un $h > 0$.
- ▶ On définit : $t_{j-1/2} = (t_j + t_{j-1})/2$.
- ▶ Alors :

$$H(t_j) = \sum_{i=1}^j K(t_j, t_{i-\frac{1}{2}}) \Delta_j, \quad j = 1, \dots, n$$

où $\Delta_j = G(t_j) - G(t_{j-1})$.

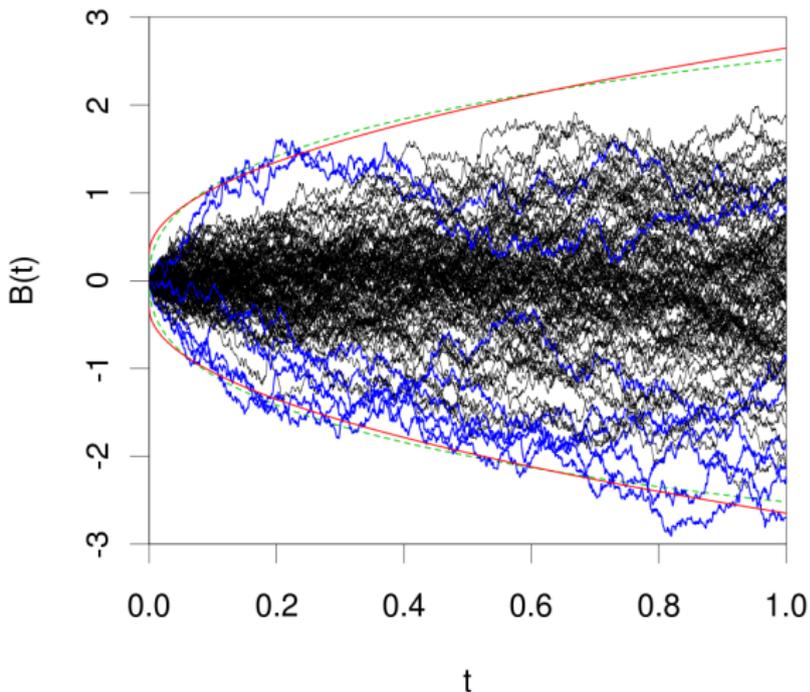
- ▶ On a alors un **système triangulaire inférieur** à résoudre ce qui est facile et rapide.
- ▶ Il est de plus possible, dans les cas qui nous intéressent, de borner l'erreur absolue de cette solution numérique.

Une région de prédiction presque minimale

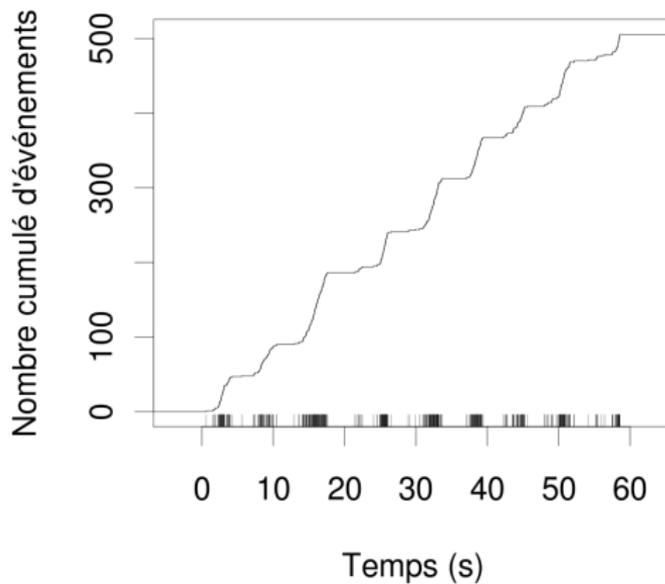
- ▶ Un problème potentiel de l'approche de la région de prédiction de surface minimale est que votre logiciel favori ne comporte pas nécessairement la fonction W_{-1} (le paquet `gsl` de \mathbb{R} la contient).
- ▶ Dans ce cas on peut utiliser un résultat supplémentaire de Kendall, Marin et Robert : une frontière du type $f(t) = a + b\sqrt{t}$ est presque de surface minimale.

Région de prédiction presque minimale à 95%

n = 100

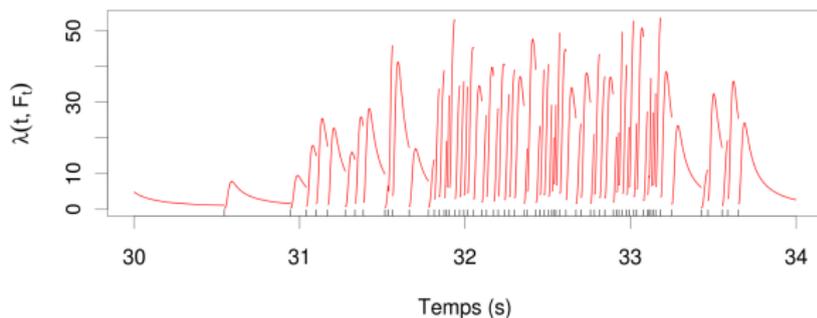


Retour sur les données

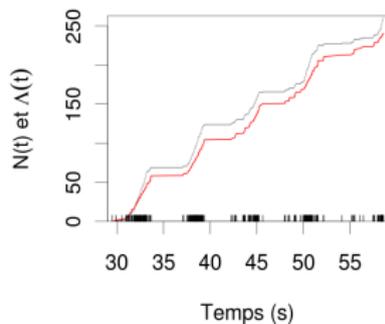


La transformation du temps sur données réelles

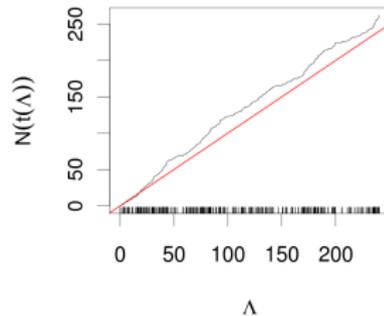
Processus d'intensité et séquence d'événements



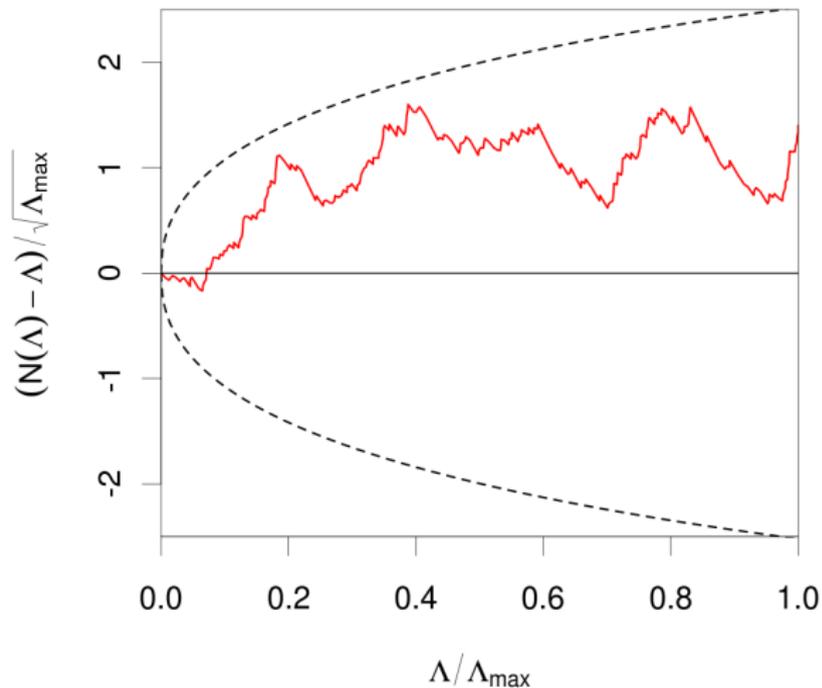
N et Λ vs t



N et Λ vs Λ



Région de prédiction minimale à 95% sur données réelles



Sommaire

Introduction

Définitions

Interprétation de λ et transformation du temps

Transformation du temps sur données réelles

Mouvement brownien canonique et théorème de Donsker

Comment reconnaît-on un ivrogne à son trajet dans la rue ?

Approche de Fierro

Approche de la "région de prédiction"

Conclusions et remerciements

Conclusions

- ▶ Quand des données sont modélisées par un processus de comptage et qu'une estimation de l'intensité conditionnelle est obtenue, la qualité de l'ajustement du modèle aux données peut être évaluée comme suit :

1. Construire le graphe de :

$$\frac{N(t) - \Lambda(t)}{\sqrt{[\Lambda(t_{max})]}} \text{ vs } \frac{\Lambda(t)}{[\Lambda(t_{max})]} \text{ pour } 0 \leq t \leq \max\{t \mid \Lambda(t) \leq [\Lambda(t_{max})]\}$$

2. Rajouter les frontière "de Lambert" ou "racine carrée" qui donnent le probabilité de recouvrement (asymptotique) désirée.
 3. Regarder si le graphe est entièrement dans la région de prédiction.
- ▶ Le test de Fierro et / ou la batterie de tests d'Ogata (1988, *JASA* **83** : 9-27) peuvent aussi être utilisés.

Remerciements

Je remercie :

- ▶ Antoine Chaffiol pour les données,
- ▶ Chong Gu pour les splines de lissage,
- ▶ Vilmos Prokaj, Olivier Faugeras et Jonathan Touboul, pour le théorème de Donsker,
- ▶ Vous pour m'avoir écouté.