Analyse non paramétrique de séquences de potentiels d'action. Construction de modèles et de tests de qualité d'ajustement.

Christophe Pouzat

Mathématiques Appliquées à Paris 5 (MAP5)

Université Paris-Descartes and CNRS UMR 8145

christophe.pouzat@parisdescartes.fr

Premières Rencontres R

Bordeaux, 3 juillet 2012

Où en est-on ?

Les données

Intensité conditionnelle

Transformation du temps

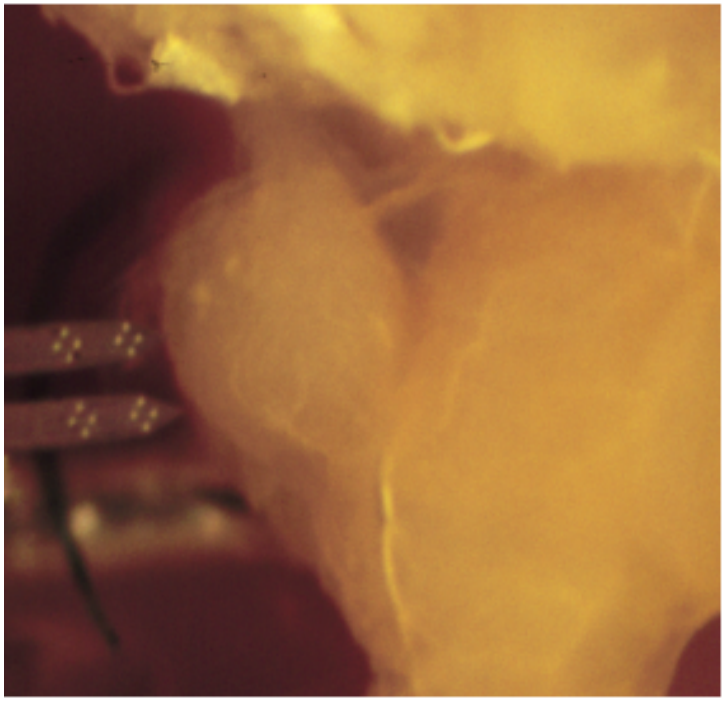
Un test basé sur le théorème de Donsker

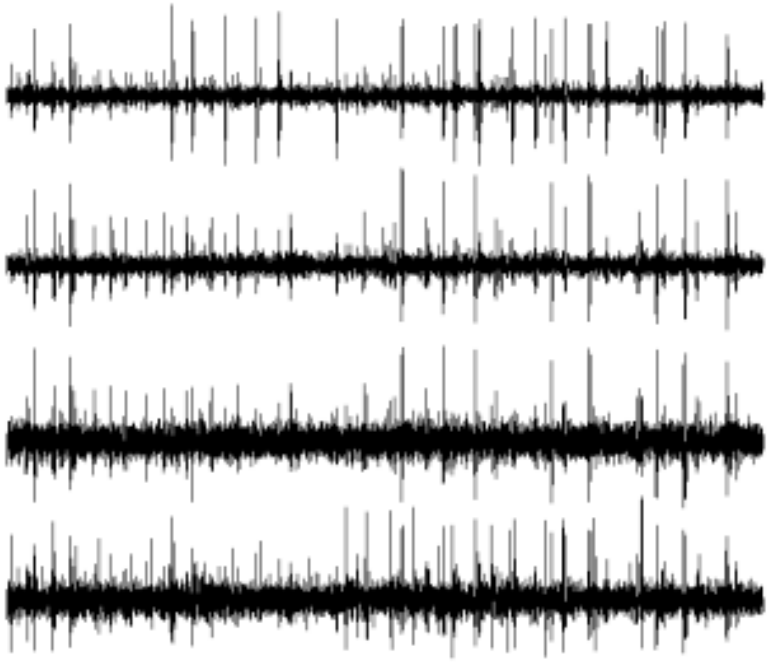
Estimation de l'intensité conditionnelle

Ajustement et diagnostiques

L'origine des données

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

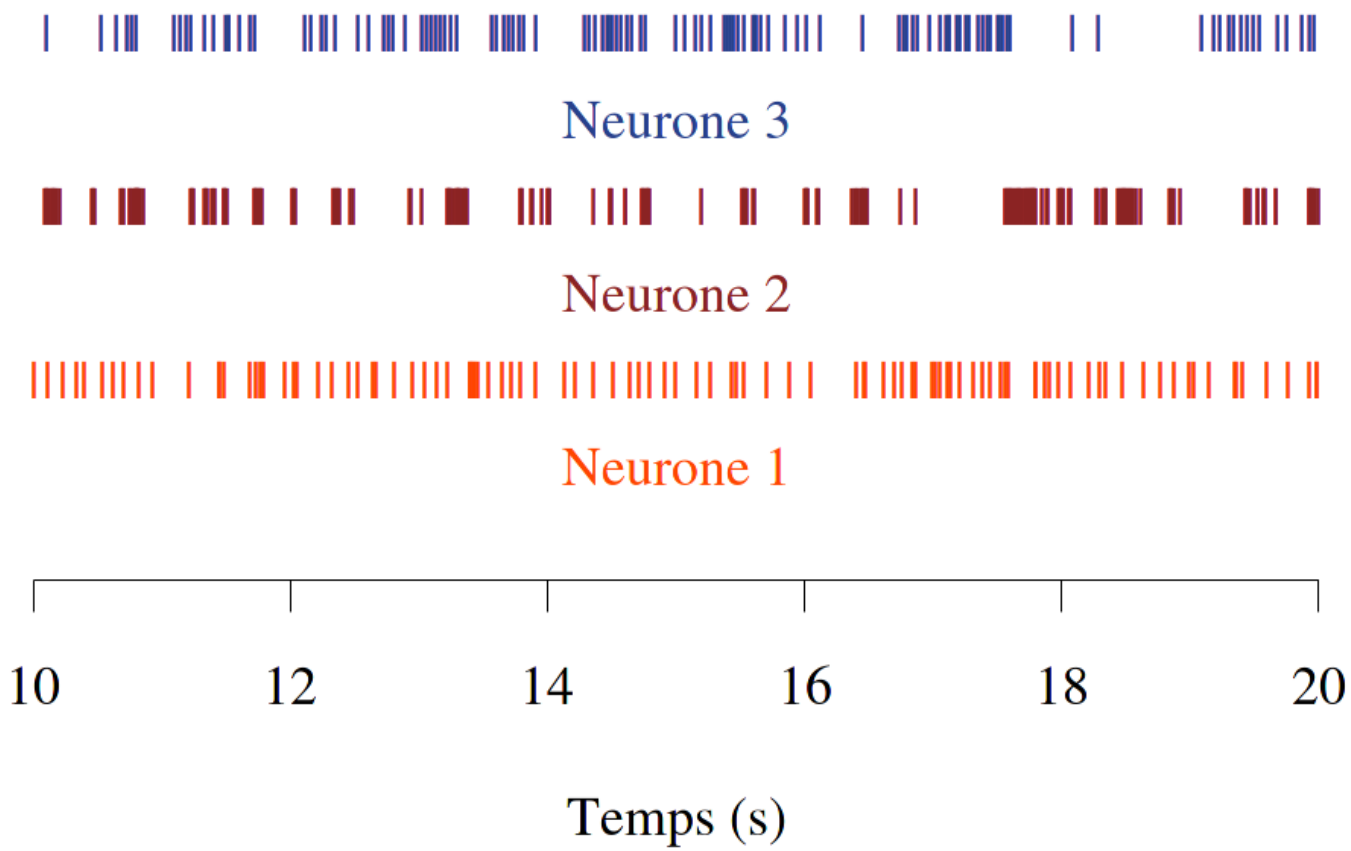




À gauche, le cerveau – d'un insecte – et la sonde d'enregistrement qui comporte 16 électrodes (les points brillants). La largeur d'une branche de la sonde est de 80 μm. A droite, 1 sec d'enregistrement sur 4 électrodes. Les pics sont des potentiels d'action.

Exemple de séquences de potentiels d'action

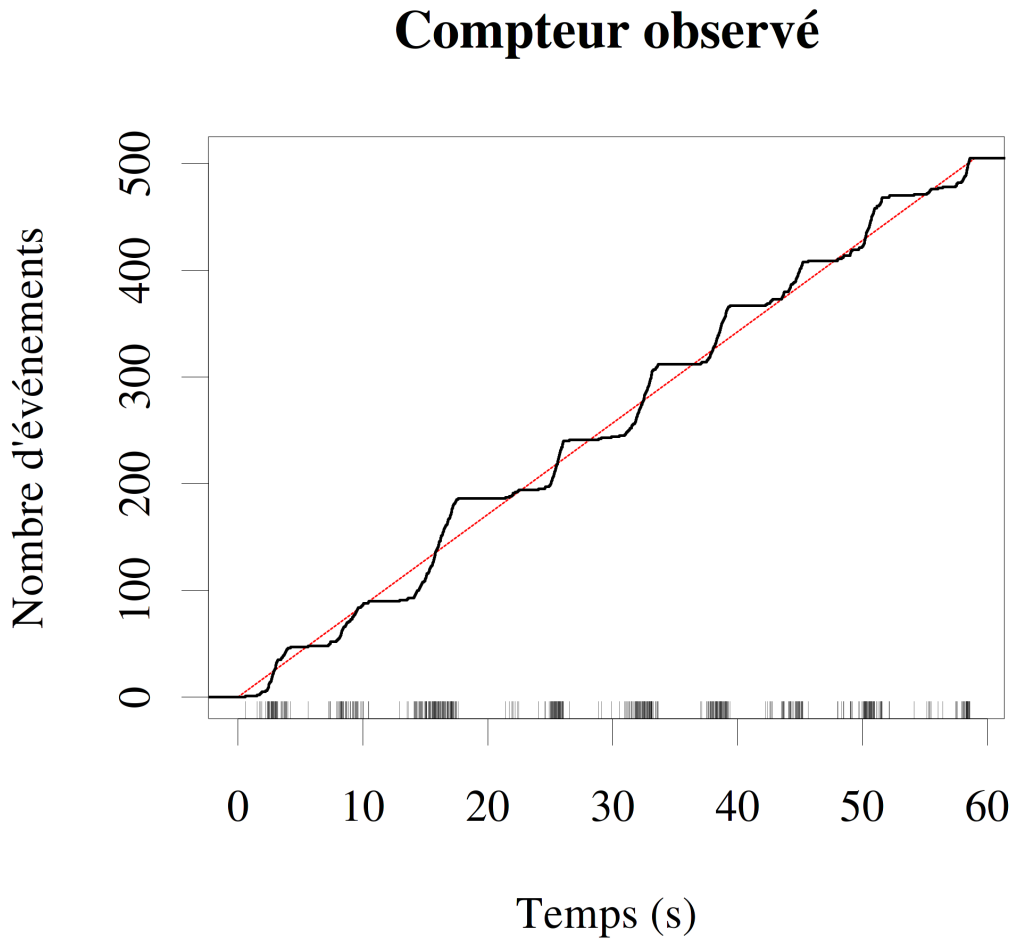
Après une étape de pré-traitement « assez lourde » appelée **tri des potentiels d'action** (PAs), on obtient le **graphe en raster** représentant les séquences ou trains de PAs :



Pourquoi et comment modéliser les trains de PAs ?

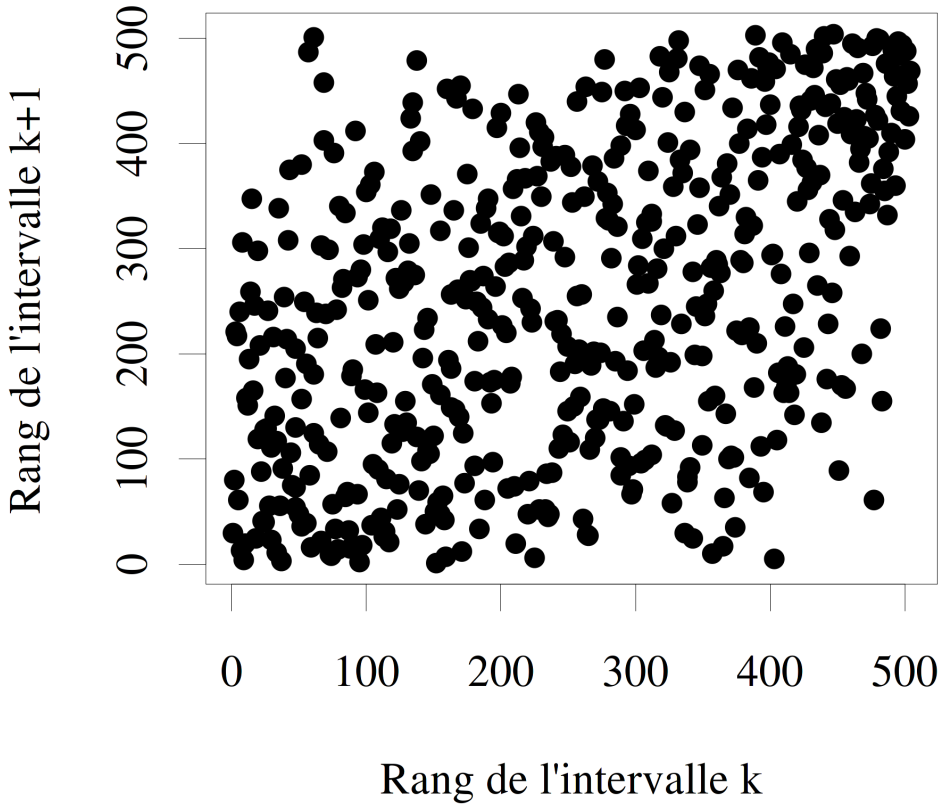
- une hypothèse de travail centrale en neurosciences est que les temps d'apparition des PAs, par opposition à leurs formes, sont le seul support de transmission de l'information entre régions du cerveau ;
- cette hypothèse légitime l'étude des trains de PAs en tant que séquences de points sur la demi-droite réelle (représentant le temps) sans nécessairement tenir compte des mécanismes biophysiques qui les génèrent ;
- Dans ce qui suit nous allons assimiler un train de PAs à un **processus ponctuel** auquel nous allons associer un **compteur** (ou **processus de comptage**).

Un cas compliqué (1)



L'espérance du compteur d'un processus de Poisson homogène – de même fréquence moyenne – est figurée en rouge (pointillés). Elle est utilisée comme « test » de stationnarité.

Un cas compliqué (2)



Un processus de renouvellement est ici inadéquat : le rang des intervalles inter-PA successifs **sont corrélés**.

Où en est-on ?

Les données

Intensité conditionnelle

Transformation du temps

Un test basé sur le théorème de Donsker

Estimation de l'intensité conditionnelle

Ajustement et diagnostiques

Cahier des charges du modèle

Notre modèle devrait nous permettre de travailler avec :

- le temps écoulé depuis le dernier PA du neurone (suffisant pour un processus de renouvellement homogène) ;
- des variables correspondant à l'histoire de la décharge du neurone, comme la durée de l'intervalle entre les deux derniers PAs ;
- des variables associées au temps écoulé depuis le dernier PA d'un neurone « fonctionellement associé » ;
- le temps écoulé par-rapport au début d'une stimulation.

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 :

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

- avec λ, nous obtenons une **description exhaustive** de notre processus / séquence de PAs.

Les deux problèmes à résoudre

Maintenant que nous avons adopté un formalisme basé sur l'intensité conditionnelle, nous devons :

- trouver un estimateur $\hat{\lambda}$ de λ ;
- trouver des tests de qualité d'ajustement de nos modèles à nos données.

Où en est-on ?

Les données

Intensité conditionnelle

Transformation du temps

Un test basé sur le théorème de Donsker

Estimation de l'intensité conditionnelle

Ajustement et diagnostiques

Que faire de λ : un résumé

Après avoir associé à λ, l'**intensité cumulée** :

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

il est facile – mais un peu long dans ce type de présentation – de prouver les résultats suivants :

- **si notre modèle est bon**, la densité des intervalles entre PAs successifs après transformation du temps :

  $$\{t_1, \dots, t_n\} \to \{\Lambda(t_1) \equiv \Lambda_1, \dots, \Lambda(t_n) \equiv \Lambda_n\}$$

  est **une densité exponentielle de paramètre 1** ;
- le processus ponctuel observé $\{\Lambda_1,\ldots,\Lambda_n\}$ est donc l'observation d'un **processus de Poisson homogène de paramètre 1**.

Les quelques diapos suivantes vont illustrer ces résultats.

Illustration de la transformation du temps sur données simulées (1)

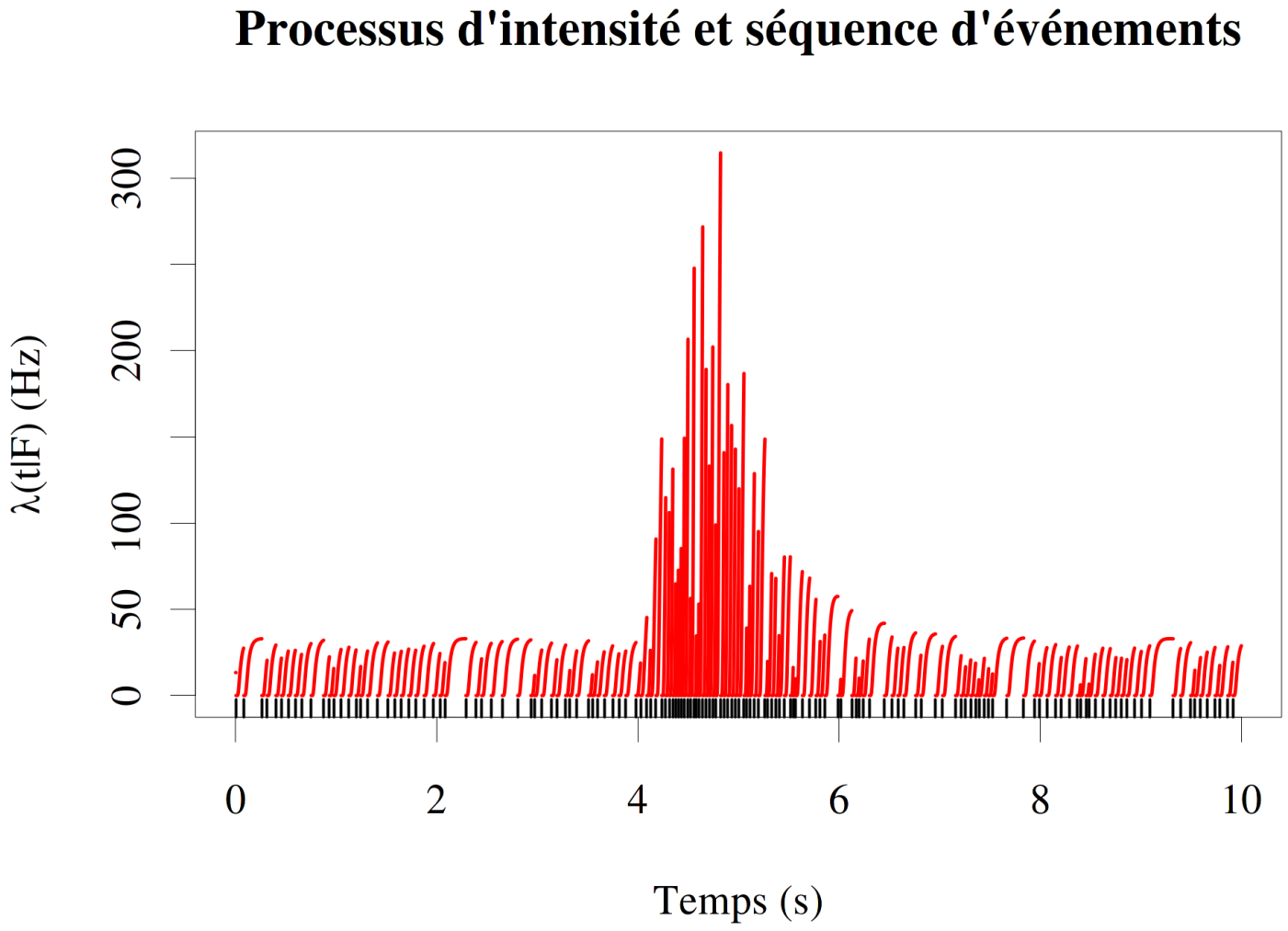


Illustration de la transformation du temps sur données simulées (2)

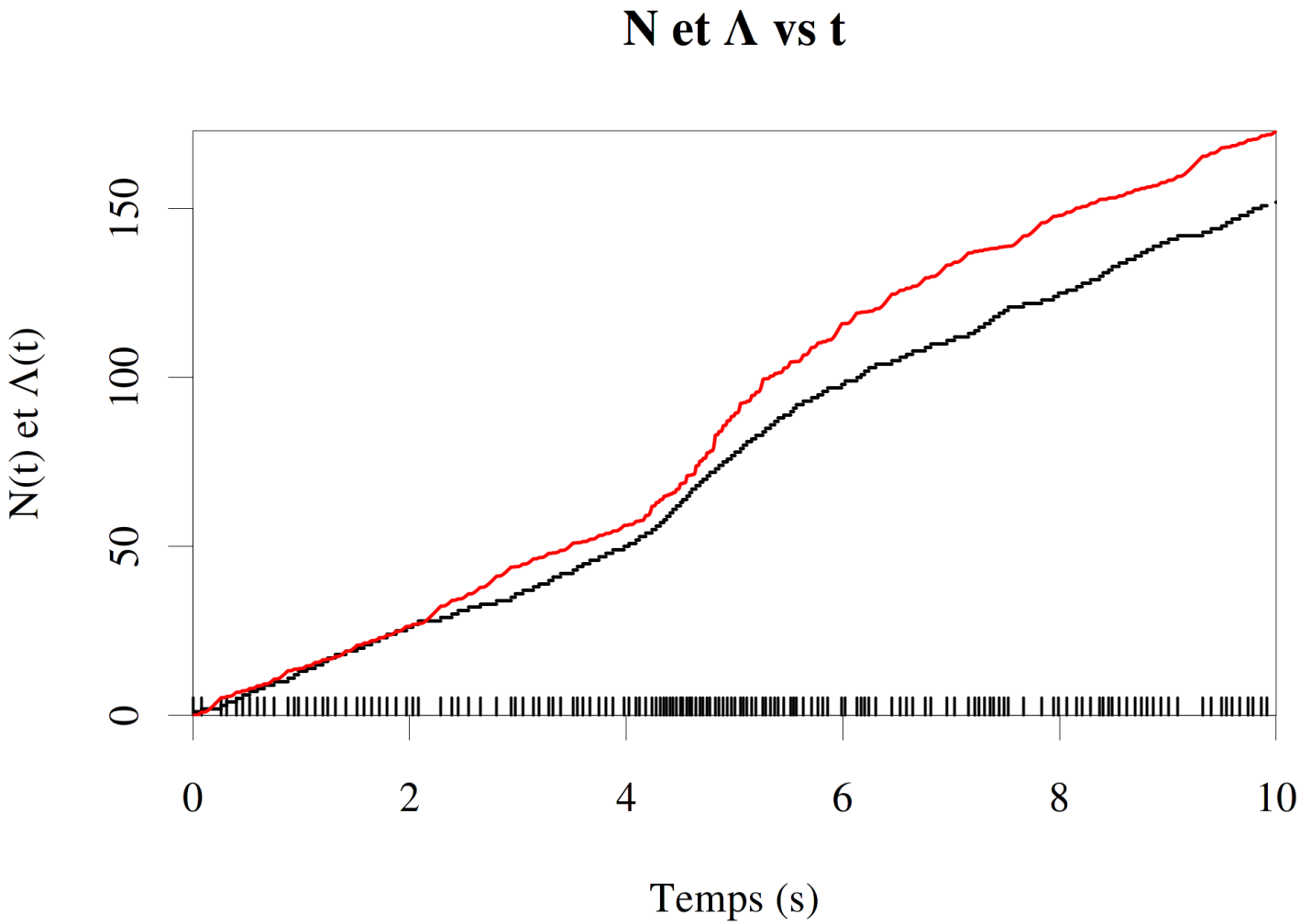
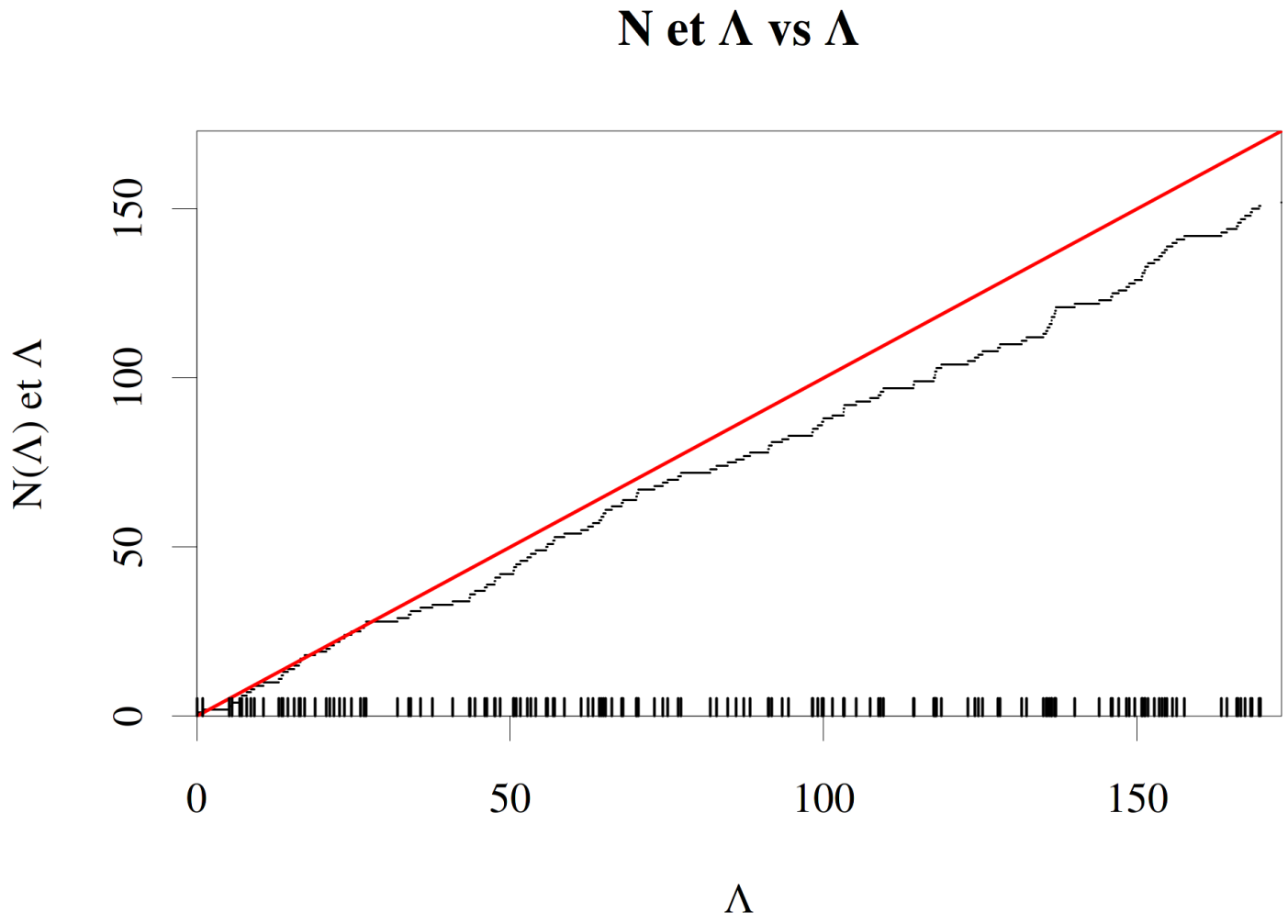


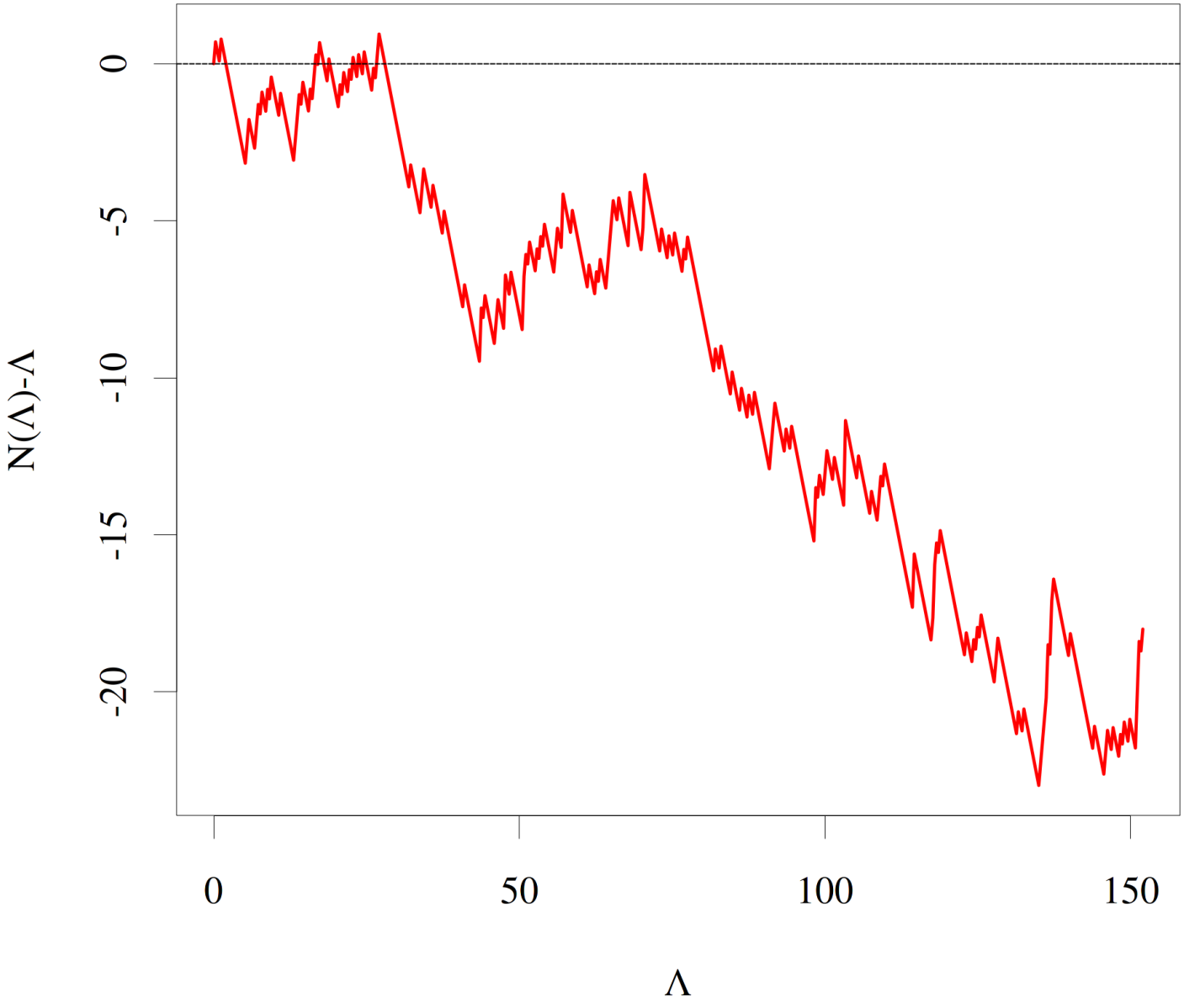
Illustration de la transformation du temps sur données simulées (3)



Les tests d'Ogata

- si, pour un bon modèle, la séquence « transformée » de PAs, $\{\Lambda_1,\ldots,\Lambda_n\}$, est une réalisation d'un processus de Poisson homogène de paramètre 1, il est possible de **tester** l'adéquation d'un modèle ajuster en comparant $\{\Lambda_1,\ldots,\Lambda_n\}$ à un Poisson homogène ;
- c'est ce qu'a proposé Yosihiko Ogata en 1988 (Statistical models for earthquake occurrences and residual analysis for point processes, Journal of the American Statistical Association, **83** : 9-27) ;
- une observation suggère, néanmoins, qu'un autre type de test pourrait s'appliquer à notre problème…

Un mouvement brownien ?



Où en est-on ?

Les données

Intensité conditionnelle

Transformation du temps

Un test basé sur le théorème de Donsker

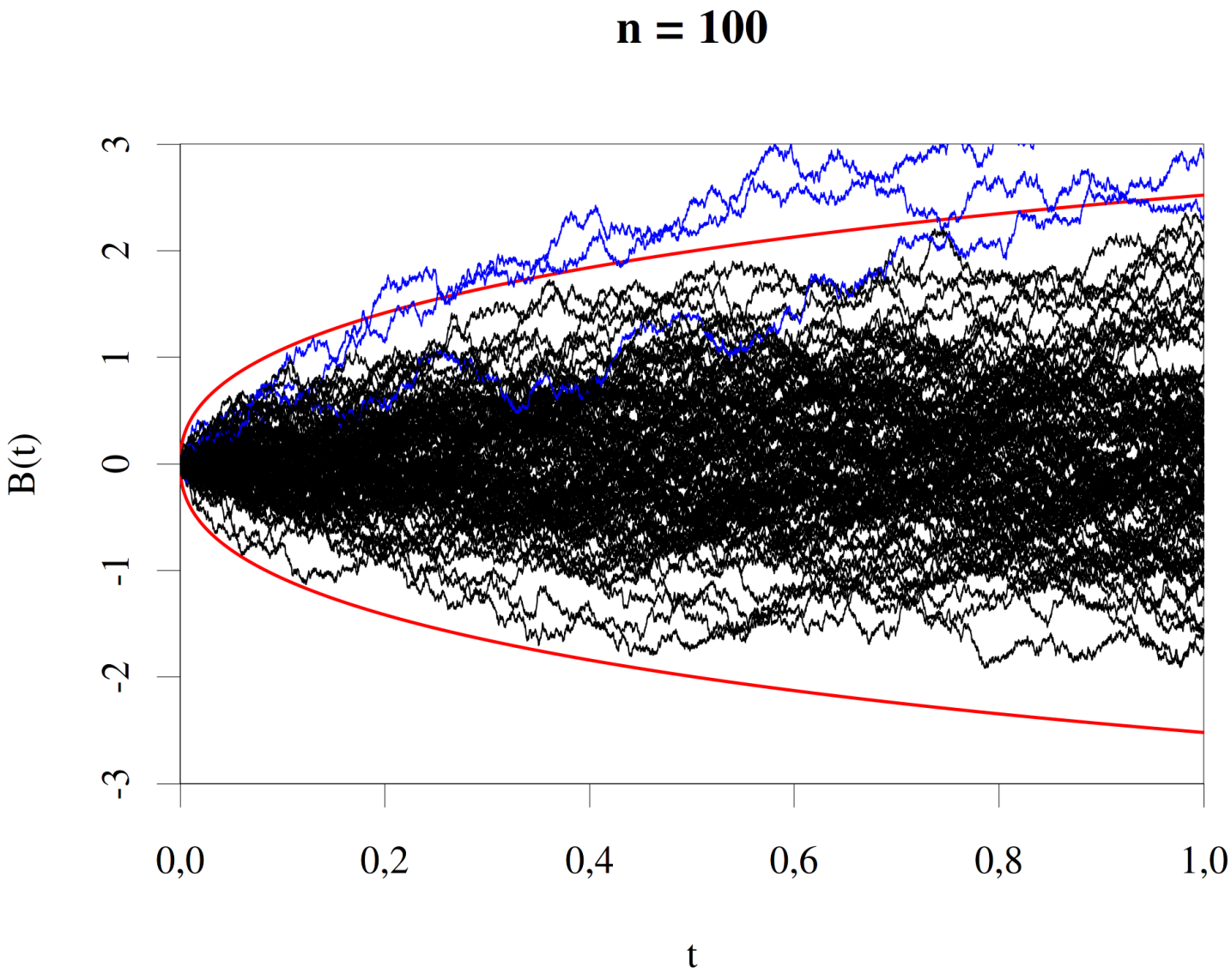
Estimation de l'intensité conditionnelle

Ajustement et diagnostiques

Théorème de Donsker et région de surface minimale

- l'intuition d'un convergence – d'une version correctement normalisée – du processus N(Λ)-Λ vers un mouvement brownien est correcte ;
- c'est le théorème de Donsker, comme me l'a indiqué Vilmos Prokaj sur la mailing list de `R` ;
- il est de plus possible de définir des régions de surfaces minimales ayant des probabilités données de contenir toutes les réalisations d'un mouvement brownien canonique (Kendall, Marin et Robert, 2007 ; Loader et Deely, 1987) ;
- nous obtenons ainsi un nouveau test de qualité d'ajustement.

Région de prédiction à 95 % de surface minimale



Où en est-on ?

Les données

Intensité conditionnelle

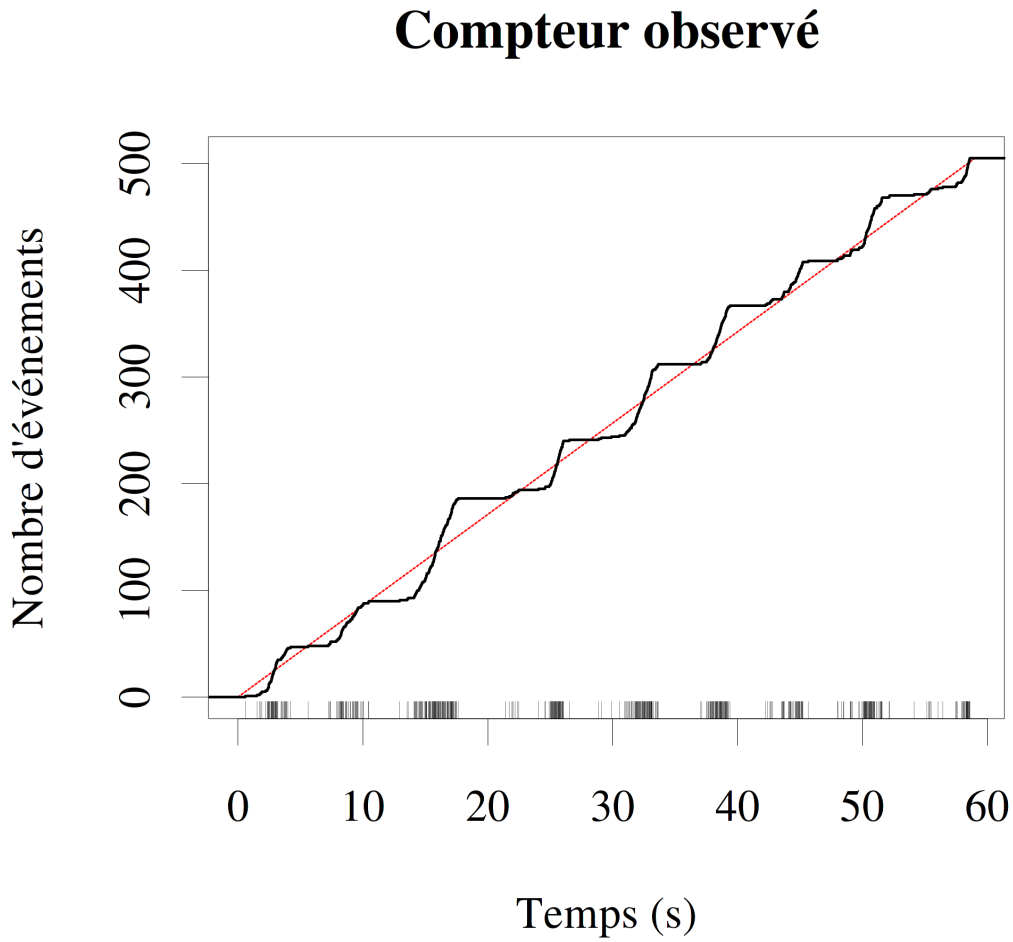
Transformation du temps

Un test basé sur le théorème de Donsker

Estimation de l'intensité conditionnelle

Ajustement et diagnostiques

Retour sur le cas « compliqué »



L'analyse exploratoire précédente nous impose de considérer un modèle minimal du type :

$$\lambda(t|\mathcal{F}_t) = f(t-t_d, t_d-t_{ad})$$

où t_d est le temps du dernier PA précédent t et t_{ad}, celui de l'avant dernier.

Approche de David Brillinger

- nous suivons Brillinger (1988) qui commence par discrétiser l'axe des temps en blocs de durée h ; h étant suffisamment petit pour avoir au plus un PA par bloc ;
- nous sommes ainsi ramenés à un problème de **régression binomiale** ;
- les « données discrétisées » sont alors considérées comme une observation d'une collection de variables aléatoires de Bernoulli $\{Y_1,\ldots,Y_k\}$ de paramètres : $f(t\text{-}t_d,t_d\text{-}t_{ad})\,h$;
- nous allons directement estimer :

$$\log\left(\frac{f(t-t_d,t_d-t_{ad})\,h}{1-f(t-t_d,t_d-t_{ad})\,h}\right)=\eta(t-t_d,t_d-t_{ad})\,.$$

Les données discrétisées

	event	time	neuron	lN.1	i1
14604	0	58.412	1	0.012	0.016
14605	1	58.416	1	0.016	0.016
14606	0	58.420	1	0.004	0.016
14607	1	58.424	1	0.008	0.016
14608	0	58.428	1	0.004	0.008
14609	0	58.432	1	0.008	0.008
14610	1	58.436	1	0.012	0.008
14611	0	58.440	1	0.004	0.012

event est la séquence de PAs discrétisée ; **time** est le temps du centre des blocs ; **neuron** est le neurone auquel **event** « appartient » ; **lN.1** est $t-t_d$; **i1** est t_d-t_{ad}. Ici, h est égal à 4 ms.

Splines de lissage

- comme la biophysique cellulaire ne nous donne que peu d'informations sur la forme fonctionnelle de η, nous allons employer des fonctions splines et nous allons pénaliser la vraisemblance (Wahba, 1990 ; Green et Silverman, 1994 ; Eubank, 1999 ; Gu, 2002) ;
- nous effectuons les calculs avec le paquet `gss` de Chong Gu ;
- $\eta(t\text{-}t_d, t_d\text{-}t_{ad})$ est décomposée de façon unique en :

$$\eta(t-t_d, t_d-t_{ad}) = \eta_\emptyset + \eta_1(t-t_d) + \eta_2(t_d-t_{ad}) + \eta_{1,2}(t-t_d, t_d-t_{ad}),$$

où les variables : $t\text{-}t_d$ et $t_d\text{-}t_{ad}$ ont été transformées (linéairement) pour avoir leurs domaines définitions égaux à [0,1] ;

- la décomposition est unique parce-qu'on impose des conditions du type : $\int_0^1 \eta_i = 0$.

Où en est-on ?

Les données

Intensité conditionnelle

Transformation du temps

Un test basé sur le théorème de Donsker

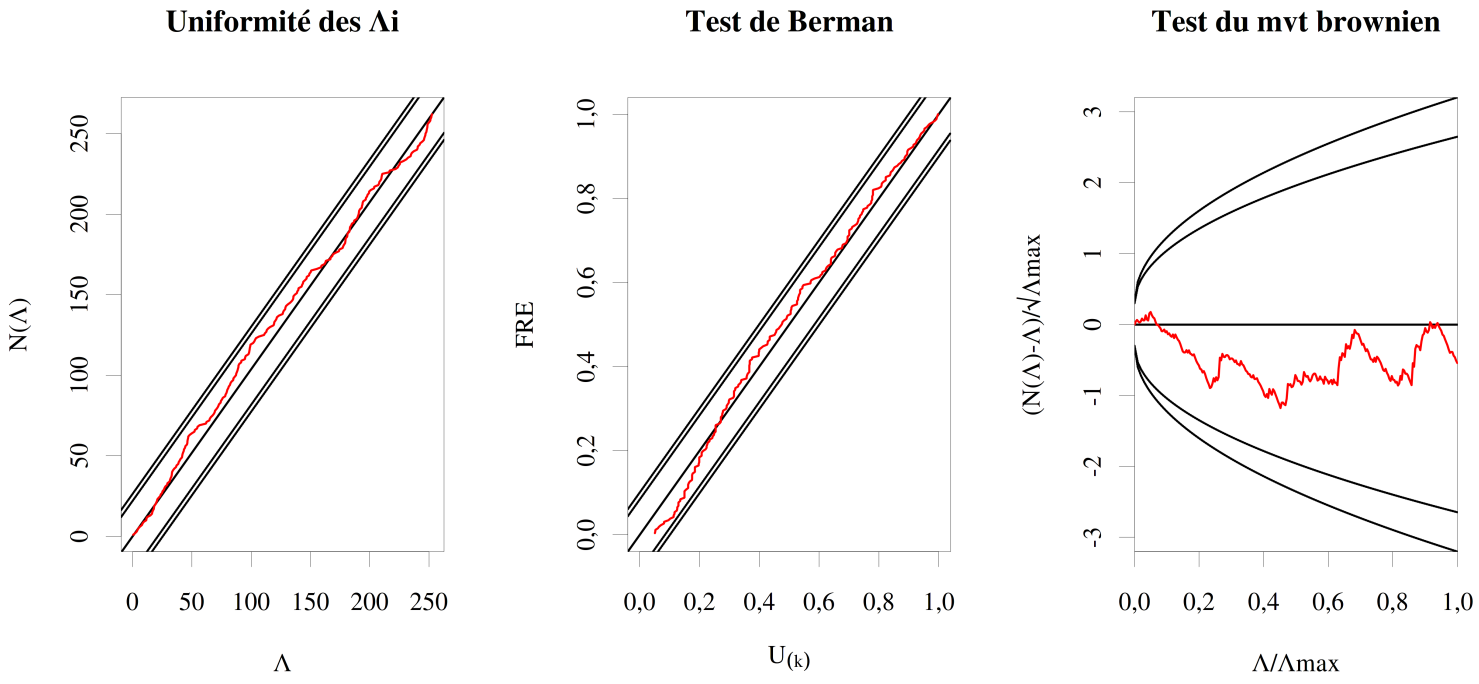
Estimation de l'intensité conditionnelle

Ajustement et diagnostiques

Remarque sur les tests

- les tests d'Ogata, tout comme le test « du mouvement brownien » que nous proposons, supposent qu'**aucun paramètre n'est dépendant des données** ;
- en général, quand la taille de l'échantillon est grande et que le nombre de paramètres est « petit », cela ne pose pas de problème ;
- mais nous sommes ici dans un contexte non-paramétrique : le nombre de paramètres augmente avec la taille de l'échantillon ;
- nous adoptons donc la stratégie suivante : nous ajustons le (ou les) modèle(s) sur une moitié des données et nous testons sur l'autre ; nous échangeons les rôles des deux moitiés.

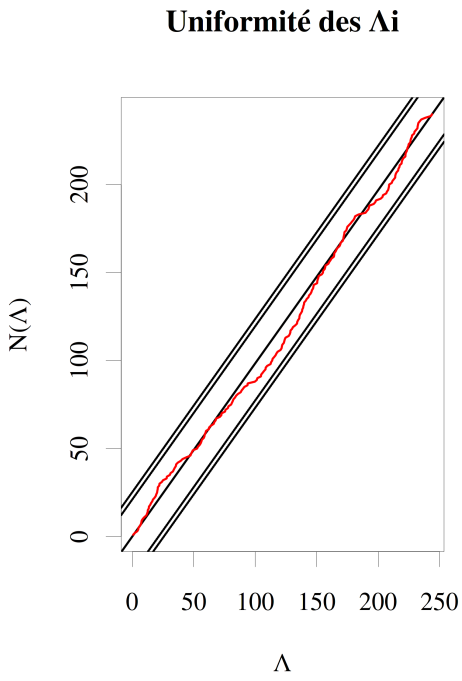
Modèle avec interaction : ajustement début / test fin

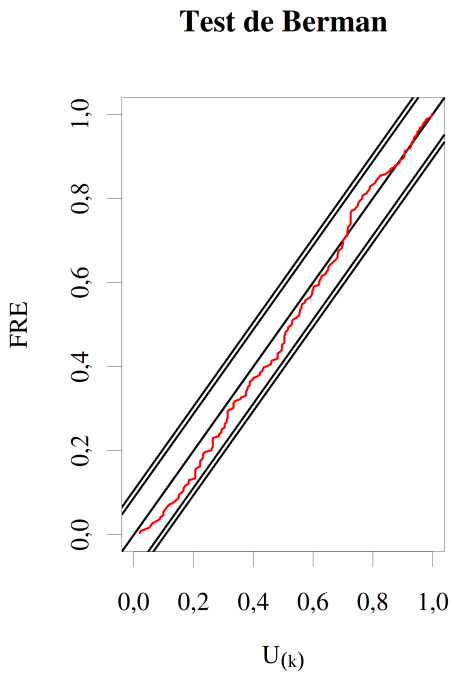


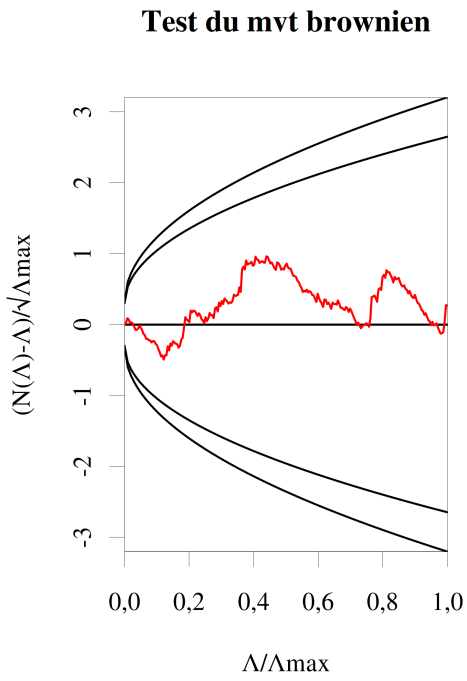
Le modèle est ici :

$$\eta(t-t_d, t_d-t_{ad}) = \eta_\emptyset + \eta_1(t-t_d) + \eta_2(t_d-t_{ad}) + \eta_{1,2}(t-t_d, t_d-t_{ad}).$$

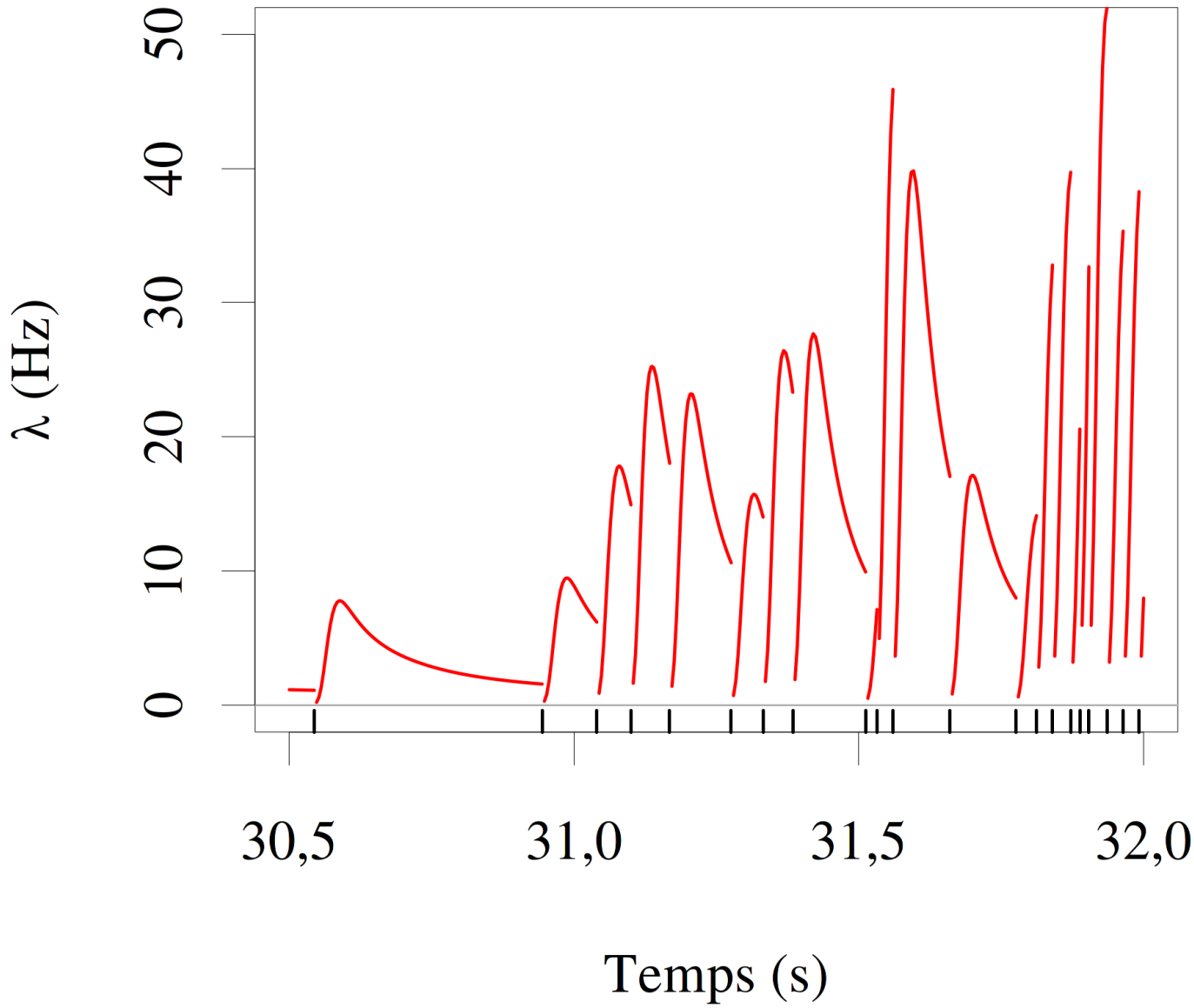
Modèle avec interaction : ajustement fin / test début







Données et $\hat{\lambda}$



Conclusions

- nous pouvons désormais estimer, « en routine », l'intensité conditionnelle de nos séquences de PAs ;
- nous pouvons inclure des interactions entre neurones et des réponses à des stimulations ;
- nous passons systématiquement des tests d'ajustement plus nombreux et plus drastiques que nos concurrents ;
- la question délicate du choix de modèles n'a pas était discutée ici, mais nous disposons d'une solution – un peu chère en temps de calcul ;
- vous pouvez en faire de même avec le paquet `STAR` disponible sur `CRAN` depuis 2009.

Remerciements

Je remercie :

- Antoine Chaffiol, pour les données et le tri des PAs ;
- Chong Gu, développeur de `gss` : mon principal collaborateur sur ce projet ;
- Vilmos Prokaj, Jonathan Touboul et Olivier Faugeras, pour m'avoir « signalé » le théorème de Donsker ;
- Carl van Vreeswijk et Avner Bar-Hen, pour les discussions ;
- Clément Léna et Yann-Suhan Senova, les courageux testeurs de `STAR` ;
- Ken Knoblauch, pour l'invitation à parler dans cette session ;
- Vous, pour m'avoir écouté.