Symétrisation dans les problèmes à deux échantillons : le cas des processus de Poisson

Magalie Fromont [1] & Béatrice Laurent [2] & Patricia Reynaud-Bouret [3]

[1] *IRMAR - Université européenne de Bretagne - Campus de Villejean - 35043 Rennes Cedex (France) - magalie.fromont@univ-rennes2.fr*
[2] *IMT, INSA Toulouse - 135 Avenue de Rangueil - 31077 Toulouse Cedex 4 (France) - Beatrice.Laurent@insa-toulouse.fr*
[3] *CNRS Université de Nice Sophia-Antipolis - 06108 Nice Cedex 2 (France) - Patricia.Reynaud-Bouret@unice.fr*

Résumé. Nous considérons ici le problème dit "à deux échantillons" pour des processus de Poisson, qui consiste à tester l'hypothèse nulle d'égalité des intensités de deux processus de Poisson indépendants. Plus précisément, nous nous intéressons à l'utilisation d'une astuce de symétrisation pour construire des tests non paramétriques et non asymptotiques, partant de statistiques de test dont la loi n'est pas nécessairement libre de la loi - inconnue - des processus sous l'hypothèse nulle. Cette astuce est appliquée en particulier à des statistiques de test basées sur des noyaux généraux. Les tests ainsi construits sont alors du niveau voulu et sont optimaux au sens du minimax sur certaines classes d'alternatives.

Mots-clés. Problème à deux échantillons, processus de Poisson, symétrisation, bootstrap, test de permutation.

Abstract. We here consider the two-sample problem for Poisson processes, which consists in testing the null hypothesis of equality of the intensities of two independent Poisson processes. We more precisely focus on the use of a symmetrization trick to construct nonparametric and nonasymptotic tests, from testing statistics whose distribution is not necessarily free from the - unknown - underlying distribution of the processes under the null hypothesis. This symmetrization principle is in particular applied on testing statistics based on general kernels. The obtained tests are then of the prescribed level and optimal in the minimax sense over various classes of alternatives.

Keywords. Two-sample problem, Poisson process, symmetrization, bootstrap, permutation test.

Ce travail est présenté pour la session spéciale du Groupe de Statistique Mathématique portant sur la technique de symétrisation.

1 Introduction

Nous considérons dans ce travail le problème dit "à deux échantillons" pour des processus de Poisson, qui consiste à tester l'égalité des intensités (non nécessairement constantes) de deux processus de Poisson indépendants observés sur un espace mesurable $\mathbb{X}$. En particulier, nous nous intéressons à l'utilisation d'une astuce de symétrisation pour construire des tests non asymptotiques de niveau voulu.
Après un point sur l'utilisation de la symétrisation dans les tests statistiques et son lien avec le wild bootstrap dans des cadres plus classiques (*c.f.* [3] ou [1] par exemple), nous présentons le principe de symétrisation spécifique au cadre des problèmes à deux échantillons pour des processus de Poisson. Nous expliquons comment il peut être utilisé pour construire les valeurs critiques associées à des statistiques de test dont la loi n'est pas libre de la loi des processus sous l'hypothèse nulle.
Nous appliquons en particulier ce principe à la construction de tests non paramétriques et non asymptotiques dont les statistiques de test, basées sur des noyaux généraux, peuvent être vues comme des extensions de celles proposées par Li [6] ou Gretton et al. [5] dans le modèle de densité. Nous exhibons pour ces tests des conditions suffisantes de contrôle du risque de seconde espèce, et montrons qu'ils sont minimax sur diverses classes d'alternatives.
Notons enfin que les tests construits ici peuvent être agrégés dans une procédure de test dite "à noyaux multiples", de niveau voulu, et adaptative au sens du minimax.

2 Problème à deux échantillons pour des processus de Poisson

Soit $X^{(1)}$ et $X^{(2)}$ deux processus de Poisson observés sur un espace mesurable $\mathbb{X}$, d'intensités respectives s_1 et s_2 par rapport à une mesure $\mu = nd\nu$, où ν est une mesure non atomique, σ-finie sur $\mathbb{X}$. Supposant que $s_1, s_2 \in \mathbb{L}^1(\mathbb{X}, d\nu) \cap \mathbb{L}^\infty(\mathbb{X}) \subset \mathbb{L}^2(\mathbb{X}, d\nu)$, on souhaite tester l'hypothèse nulle (H_0) "$s_1 = s_2$" contre l'alternative (H_1) "$s_1 \neq s_2$".
Ce problème de test, bien qu'étroitement lié au problème à deux échantillons classique dans le modèle de densité, a été moins étudié dans la littérature statistique. Il présente néanmoins de nombreuses applications, notamment en biologie et en économie.

On note $\mathbb{P}_{s_1,s_2}$ la loi jointe de $(X^{(1)}, X^{(2)})$. Pour tout événement $\mathcal{A}$ construit sur $(X^{(1)}, X^{(2)})$, $\mathbb{P}_{(H_0)}(\mathcal{A})$ désigne $\sup_{(s_1,s_2),s_1=s_2} \mathbb{P}_{s_1,s_2}(\mathcal{A})$.

On introduit le processus agrégé X dont la mesure ponctuelle est définie par $dX = dX^{(1)} + dX^{(2)}$, et dont les points sont notés $X_1, \ldots, X_N$.

3 Astuce de symétrisation

Considérons une statistique de test $T(X^{(1)}, X^{(2)})$ quelconque, dont la loi n'est pas libre de $s_1 = s_2$ sous (H_0). Choisir une valeur critique associée à $T(X^{(1)}, X^{(2)})$ de telle sorte que le test correspondant soit de niveau voulu s'avère être dans ce cas une question fondamentale. Dans un modèle de densité, ce choix se fait généralement par des approches de bootstrap ou de permutation. Les tests de permutation sont en fait basés sur l'idée que la statistique de test calculée sur les échantillons d'origine et la statistique de test calculée sur des échantillons obtenus après permutation aléatoire des éléments de l'échantillon agrégé suivent la même loi sous l'hypothèse nulle.
De la même façon, dans un modèle de régression, ce choix peut se faire par des approches de bootstrap. Lorsque la loi des bruits est symétrique, il peut aussi se faire à l'aide d'une approche de symétrisation, liée au wild bootstrap (*c.f.* [3] ou [1]). Cette approche repose sur l'égalité des lois des statistiques de test calculées sur les échantillons d'origine et sur les échantillons "symétrisés" ou pondérés à l'aide de variables de loi de Rademacher.
Ici, comme dans un modèle de régression, on introduit une suite $(\varepsilon_i)_{i\in\mathbb{N}}$ de variables i.i.d. de loi de Rademacher, indépendante de X. Soit $X^{\varepsilon(1)}$ et $X^{\varepsilon(2)}$ les deux processus ponctuels dont les points sont respectivement définis par $\{X_i,\ i = 1 \dots N,\ \varepsilon_i = 1\}$ et $\{X_i,\ i = 1 \dots N,\ \varepsilon_i = -1\}$. On a alors le résultat suivant.

Proposition 1. *Sous (H_0), conditionnellement à X, $(X^{(1)}, X^{(2)})$ est de même loi que $\left(X^{\varepsilon(1)}, X^{\varepsilon(2)}\right)$.*

En conséquence, sous (H_0), conditionnellement à X, $T(X^{(1)}, X^{(2)})$ est de même loi que $T\left(X^{\varepsilon(1)}, X^{\varepsilon(2)}\right)$. Ainsi, étant donné $\alpha \in (0, 1)$, si $q_{1-\alpha}(X)$ désigne le $(1-\alpha)$ quantile de la loi conditionnelle de $T\left(X^{\varepsilon(1)}, X^{\varepsilon(2)}\right)$ sachant X, sous (H_0),

$$\mathbb{P}_{s_1,s_2}\Big(T(X^{(1)}, X^{(2)}) > q_{1-\alpha}(X)\Big|X\Big) \leq \alpha. \tag{1}$$

En particulier, on en déduit que

$$\mathbb{P}_{(H_0)}\Big(T(X^{(1)}, X^{(2)}) > q_{1-\alpha}(X)\Big) \leq \alpha. \tag{2}$$

Le test rejetant (H_0) lorsque $T(X^{(1)}, X^{(2)}) > q_{1-\alpha}(X)$ est donc de niveau α.
Mais notons que la propriété (1) est en réalité plus forte que le contrôle usuel du risque de première espèce maximal (sans conditionnement), comme dans (2).

4 Procédures de test basées sur des méthodes à noyaux

Soit $K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ une fonction noyau générale telle que

$$\int_{\mathbb{X}^2} K^2(x, x')(s_1 + s_2)(x)(s_1 + s_2)(x')d\nu_x d\nu_{x'} < +\infty,$$

et $K(x,x') = K(x',x)$ pour tout $(x,x') \in \mathbb{X}^2$.

On introduit alors la statistique de test $T_K(X^{(1)}, X^{(2)})$ définie par :

$$T_K(X^{(1)}, X^{(2)}) = \sum_{i,j \in \{1,\dots,N\}, i \neq j} K(X_i, X_j) \varepsilon_i^0 \varepsilon_j^0,$$

où les ε_i^0 sont des marques sur X, telles que pour tout $i \in \{1,\dots,N\}$, $\varepsilon_i^0 = 1$ si $X_i \in X^{(1)}$, $\varepsilon_i^0 = -1$ si $X_i \in X^{(2)}$.

Nous avons choisi de considérer trois exemples particuliers de noyaux.

Exemple 1. K est un noyau de projection sur une base orthonormée $\{\varphi_\lambda, \lambda \in \Lambda\}$ pour la norme $\mathbb{L}^2$ associée à ν, notée $\|.\|$, défini par $K(x,x') = \sum_{\lambda \in \Lambda} \varphi_\lambda(x)\varphi_\lambda(x')$. Dans ce cas, on montre que $T_K(X^{(1)}, X^{(2)})$ est un estimateur sans biais de $n^2 \|\Pi_S(s_1 - s_2)\|^2$, où Π_S désigne la projection orthogonale sur l'espace vectoriel S engendré par $\{\varphi_\lambda, \lambda \in \Lambda\}$.

Exemple 2 pour $\mathbb{X} = \mathbb{R}^d$ *et* ν *la mesure de Lebesgue.* K est basé sur un noyau d'approximation $k \in \mathbb{L}^2(\mathbb{R}^d)$, tel que $k(-z) = k(z)$: pour $x = (x_1,\dots,x_d)$, $x' = (x'_1,\dots,x'_d)$ dans $\mathbb{X}$, $K(x,x') = \frac{1}{\prod_{i=1}^d h_i} k\left(\frac{x_1 - x'_1}{h_1}, \dots, \frac{x_d - x'_d}{h_d}\right)$, où $h = (h_1,\dots,h_d)$ est un vecteur de d fenêtres positives. Dans ce cas, $T_K(X^{(1)}, X^{(2)})$ est un estimateur sans biais de $n^2 \int_{\mathbb{X}} k_h * (s_1 - s_2)(x)(s_1 - s_2)(x) d\nu_x$, où $k_h(u_1,\dots,u_d) = \frac{1}{\prod_{i=1}^d h_i} k\left(\frac{u_1}{h_1}, \dots, \frac{u_d}{h_d}\right)$ et $*$ est l'opérateur de convolution usuel par rapport à ν.

Exemple 3. K est un noyau reproduisant tel que $K(x,x') = \langle \theta(x), \theta(x') \rangle_{\mathcal{H}_K}$, où θ une fonction de représentation et $\mathcal{H}_K$ un RKHS associé à K. Ici, $T_K(X^{(1)}, X^{(2)})$ est un estimateur sans biais de $n^2 \left\|m_{s_1} - m_{s_2}\right\|^2_{\mathcal{H}_K}$, où $\|.\|_{\mathcal{H}_K}$ est la norme du RKHS, et $m_s = \int_{\mathbb{X}} K(.,x)s(x)d\nu_x$. Si $\int_{\mathbb{X}} s_1(x)d\nu_x = \int_{\mathbb{X}} s_2(x)d\nu_x = 1$, m_{s_1} et m_{s_2} sont les plongements moyens des lois $s_1 d\nu$ et $s_2 d\nu$ dans le RKHS $\mathcal{H}_K$, et si K est un noyau dit "caractéristique" (*c.f.* [7]), alors $m_{s_1} = m_{s_2}$ si et seulement si $s_1 = s_2$.

L'approche de symétrisation décrite ci-dessus, appliquée à la statistique $T_K(X^{(1)}, X^{(2)})$ conduit à considérer une suite $(\varepsilon_i)_{i \in \mathbb{N}}$ de variables de loi de Rademacher i.i.d. indépendantes de X, et les processus $X^{\varepsilon(1)}$ et $X^{\varepsilon(2)}$ dont les points sont respectivement définis par $\{X_i,\ i = 1 \dots N,\ \varepsilon_i = 1\}$ et $\{X_i,\ i = 1 \dots N,\ \varepsilon_i = -1\}$. Sous (H_0), conditionnellement à X, on a vu que $T_K(X^{(1)}, X^{(2)})$ est de même loi que $T_K\left(X^{\varepsilon(1)}, X^{\varepsilon(2)}\right)$, qui elle-même est de même loi que

$$T_K^\varepsilon = \sum_{i,j \in \{1,\dots,N\}, i \neq j} K(X_i, X_j) \varepsilon_i \varepsilon_j.$$

Ainsi, si $q_{K,1-\alpha}(X)$ désigne le $(1-\alpha)$ quantile de la loi conditionnelle de T_K^ε sachant X, le test rejetant (H_0) lorsque $T_K(X^{(1)}, X^{(2)}) > q_{K,1-\alpha}(X)$ est de niveau α.

Étudiant le risque de seconde espèce du test, on obtient des conditions suffisantes sur l'alternative (s_1, s_2) garantissant que

$$\mathbb{P}_{s_1,s_2}\left(T_K(X^{(1)}, X^{(2)}) > q_{K,1-\alpha}(X)\right) \geq 1 - \beta,$$

pour $\beta \in (0,1)$.
Dans le cas particulier de noyaux de projection, ces conditions s'expriment de la façon suivante.

Theorem 1. *Soit $\alpha, \beta \in (0,1)$. Si K est un noyau de projection construit à partir d'une base orthonormée $\{\varphi_\lambda, \lambda \in \Lambda\}$ d'un sous-espace vectoriel de dimension D de $\mathbb{L}^2(\mathbb{X}, d\nu)$, il existe $\kappa > 0$ telle que si*

$$\|s_1 - s_2\|_2^2 \geq \|(s_1 - s_2) - \Pi_S(s_1 - s_2)\|_2^2 + \frac{(4 + 2\sqrt{2}\kappa \ln(2/\alpha))\|s_1 + s_2\|_\infty \sqrt{D}}{n\sqrt{\beta}} + \frac{8\|s_1 + s_2\|_\infty}{\beta n},$$

alors

$$\mathbb{P}_{s_1,s_2}\left(T_K(X^{(1)}, X^{(2)}) > q_{K,1-\alpha}(X)\right) \geq 1 - \beta.$$

La preuve de ce théorème est basée sur le fait que la statistique T_K^ε est un chaos de Rademacher, que l'on peut contrôler précisément, à l'aide d'une inégalité de De La Peña et Giné [2]. De ce résultat, on déduit en particulier que les tests considérés sont minimax sur des classes d'alternatives liées à des espaces de Besov, et qu'ils peuvent être intégrés dans une procédure de test agrégé, à noyaux multiples, qui sera adaptative au sens du minimax sur les mêmes classes d'alternatives (*c.f.* [4] pour plus de détails).

Références

[1] Arlot, S., Blanchard, G., and Roquain, E. (2010). *Some nonasymptotic results on resampling in high dimension, I : Confidence regions and II : Multiple tests*, *Ann. Statist.* **38**, 51–82, 83–99.

[2] de la Peña, V. H. and Giné, E. (1999). *Decoupling : From dependence to independence. Randomly stopped processes. U-statistics and processes. Martingales and beyond*, Springer, New York.

[3] Durot, C. and Rozenholc, Y. (2006). *An adaptive test for zero mean*, *Math. Meth. Statist.* **15 (1)**, 26–60.

[4] Fromont, M. and Laurent, B. and Reynaud-Bouret, P. (2013). *The two-sample problem for Poisson processes : Adaptive tests with a nonasymptotic wild bootstrap approach*, *Ann. Statist.* **41(3)**, 1431–1461.

[5] Gretton, A. and Borgwardt, K. M. and Rasch, M. J. and Schölkopf, B. and Smola, A. (2012). *A kernel two-sample test*, *J. Mach. Learn. Res.* **13**, 723–773.

[6] Li, Q. (1999). *Nonparametric testing the similarity of two unknown density functions : local power and bootstrap analysis*, *J. Nonparametr. Statist.* **11(1-3)**, 189–213.

[7] Sriperumbudur, B. K. and Fukumizu, K. and Lanckriet, G. R. G. (2011). *Universality, characteristic kernels and RKHS embeddings of measures*, *J. Mach. Learn. Res.* **12**, 2389–2410.