Biometrika (2018), **105**, 4, *pp.* 931–944
Printed in Great Britain
doi: 10.1093/biomet/asy044
Advance Access publication 18 September 2018

Continuous testing for Poisson process intensities: a new perspective on scanning statistics

BY FRANCK PICARD

Centre National de la Recherche Scientifique, Laboratoire de Biométrie et Biologie Evolutive, 43 Boulevard du 11 Novembre 1918, 69622 Villeurbanne, France

franck.picard@univ-lyon1.fr

PATRICIA REYNAUD-BOURET

Centre National de la Recherche Scientifique, Laboratoire Jean Alexandre Dieudonné, Parc Valrose, 06108 Nice, France

Patricia.Reynaud-Bouret@unice.fr

AND ETIENNE ROQUAIN

Sorbonne Université, Laboratoire de Probabilités, Statistique et Modélisation, 4 Place Jussieu, 75005 Paris, France

etienne.roquain@upmc.fr

SUMMARY

We propose a continuous testing framework to test the intensities of Poisson processes that allows a rigorous definition of the complete testing procedure, from an infinite number of hypotheses to joint error rates. Our work extends procedures based on scanning windows by controlling the familywise error rate and the false discovery rate in a non-asymptotic manner and in a continuous way. We introduce the *p*-value process on which the decision rule is based. Our method is applied in neuroscience via the standard homogeneity and two-sample tests.

Some key words: False discovery rate; Familywise error rate; Multiple testing; Poisson process.

1. INTRODUCTION

Continuous testing has recently emerged as a suitable framework for testing a possibly infinite number of hypotheses. It is especially suited to situations where the data come from an underlying continuously indexed random process, such as white noise (Dümbgen & Spokoiny, 2001), or a point process (Blanchard et al., 2014). Point processes are discrete in essence, which allows a drastic reduction of the continuum number of hypotheses to a finite but random number of tests. We elaborate on this idea to propose a continuous framework that allows rigorous definition of the complete testing procedure, from multiple hypotheses to error rates. Our procedure is very general and can be fully implemented thanks to the use of sliding windows, which allows us to extend scanning statistics by providing theoretical controls on the familywise error and false discovery rates. We focus on Poisson processes, but the main results are valid for general point processes.

Scanning windows constitute possibly the most common method of performing tests on Poisson processes to detect unexpected clusters of observations (Kulldorf, 1997; Chan & Zhang, 2007; Perone Pacifico et al., 2007; Siegmund et al., 2011) or to compare samples (Walther, 2010). Test statistics are computed on overlapping windows of a given length that are shifted by all possible delays, with the advantage of avoiding discretization of the data (Tuleau-Malot et al., 2014). The decision rule is based on the distribution of the scan statistic, that is, the maximum of all test statistics, under the complete null hypothesis. This distribution can be either approximated (Naus, 1982; Chan & Zhang, 2007; Walther, 2010; Fu et al., 2012; Rivera & Walther, 2013) or learned by a randomization procedure (Kulldorff et al., 2005, 2009; Arias-Castro et al., 2018), which is common in multiple testing (Westfall & Young, 1993; Romano & Wolf, 2005). We adopt the conditional testing approach (Loader, 1991; Rivera & Walther, 2013), which has the advantage of eliminating nuisance parameters. This approach consists in deriving tests that are conditional on the number of occurrences for the homogeneity test or on their positions for the two-sample test. A major limitation of current window-based strategies is that they are global procedures that focus on a global null hypothesis only, which inevitably results in a global yes/no answer. However, there is a need for flexible approaches in order to improve sensitivity for local structures. Continuous testing allows us to revisit the scanning windows setting by defining an uncountably infinite set of local hypotheses and providing a rigorous framework for defining joint error rates.

A statistical challenge that arises when dealing with sliding windows is the appropriate control of Type I errors, which is made difficult by the induced dependencies. Current strategies mostly rely on asymptotic approximation of the distribution of the scan statistics (Kulldorf, 1997; Chan & Zhang, 2007), which ignore the dependence structure but are computationally efficient. Here, we adopt a non-asymptotic point of view and consider the whole p-value family, which becomes a stochastic process. This is easily computed for count statistics, but it can also be estimated by simulation. We propose a min-p procedure in continuous time to control the familywise error rate, based on a Monte Carlo scheme. Control of the false discovery rate, i.e., the expected proportion of errors among rejections, is more intricate, and attempts to achieve control are more recent. Perone Pacifico et al. (2004) proposed a continuous analogue of the false discovery rate in the context of random fields, where false positives are quantified via the Lebesgue measure. However, their procedure is based on a familywise error rate guarantee. Their strategy has also been considered in a setting close to our framework (Perone Pacifico et al., 2007). Siegmund et al. (2011) investigated the false discovery rate for scanning statistics, but under two strong conditions, namely a Poisson distribution for the number of false discoveries and independence between null and alternative statistics, which are not satisfied in standard models. Moreover, their procedure is based on discretized p-values. Thanks to the continuous testing framework, we propose a weighted Benjamini–Hochberg procedure in continuous time for Poisson process testing, which outperforms the procedure of Siegmund et al. (2011) in simulations. We generalize the classical step-up procedure of Benjamini & Hochberg (1995) to the continuous setting by extending the method of Blanchard et al. (2014). An important improvement is that we consider local hypothesis testing, which allows us to remove any smoothness assumptions while deriving a fully computable p-value process.

The present work is inspired by the analysis of spike train data in neuroscience; a spike train is a succession of action potentials of a given neuron (Pipa et al., 2013) and is commonly modelled by a Poisson process. In particular, inhomogeneous Poisson processes can be used as a powerful model to understand rapid changes in firing rates induced by certain stimuli (Kass et al., 2003). The estimation of the Poisson process intensity is well studied (Shimazaki & Shinomoto, 2010), but estimation is insufficient in practice. Indeed, the main practical problems are to determine whether the spike apparition process is influenced by the stimulus and, if so, when modifications

take place. Hence, estimation needs to at least be combined with confidence bands before one can take a decision (Pouzat & Chaffiol, 2009). But confidence bands are usually not corrected for multiplicity, and it might be more precise to view these questions directly as a multiple and continuous testing problem. Our method tackles these problems directly, by considering classical count or kernel statistics (Fromont et al., 2011; Gretton et al., 2012; Fromont et al., 2013). It is illustrated on simulated and experimental data, and is available from the first author in the form of the R (R Development Core Team, 2018) package contest.

2. Continuous testing framework

2·1. *Notation*

A point process is a random countable set of points, usually denoted by N; we shall also use N to denote the associated point measure. For any interval I, $N(I)$ is a random variable that corresponds to the number of points of N in I, and $N \cap I$ denotes the point process N restricted to I, i.e., the points that are in both N and I. For simplicity, we focus on processes in $[0, 1]$. Two different kinds of hypotheses are tested, either homogeneity of a given process or equality of intensities between two different processes. To propose a unified framework, we use the following notation: we observe a random set of points Y with distribution $P_{\theta,\lambda}$, where λ is a nuisance parameter and θ represents the signal of interest. The aim is to test whether θ is zero on some intervals.

2·2. *Homogeneity test*

We observe $Y = N$, a point process modelled by a heterogeneous Poisson process on $[0, 1]$ with unknown intensity ν with respect to Lebesgue measure. We would like to detect regions with an unexpected number of occurrences with respect to the stationary behaviour of the process. To do so, we compare the intensity ν of the observed process with the mean intensity $\lambda = \int_0^1 \nu(s)\,\mathrm{d}s$. Then we rewrite the distribution of Y in terms of the parameters λ and θ, with

$$\theta(t) = \frac{\nu(t)}{\lambda} - 1, \quad 0 \leqslant t \leqslant 1.$$

The signal of interest is θ, while λ may be viewed as a nuisance parameter because its value is unknown under the null hypothesis. The null hypothesis therefore reduces to $\theta(t) = 0$; it is composite because the distribution $P_{\theta,\lambda}$ can still vary with λ. We could also test $\nu = \lambda_0$, with λ_0 a known constant baseline rate. The corresponding baseline test is different from a homogeneity test in general, but if $\lambda_0 = \int_0^1 \nu(t)\,\mathrm{d}t$ then it can be viewed as a particular, easier case of the homogeneity test.

2·3. *Two-sample test*

We observe $Y = (N_A, N_B)$, where N_A and N_B are heterogeneous independent Poisson processes on $[0, 1]$ with intensities ν_A and ν_B with respect to Lebesgue measure. In the following, for any interval I, we denote by $Y(I)$ the pair of counts formed by $N_A(I)$ and $N_B(I)$, and $Y \cap I = (N_A \cap I, N_B \cap I)$ represents the composition of interval I in terms of points of N_A and N_B, including their precise positions inside the interval. Our aim here is to detect the regions where $\nu_A \neq \nu_B$ or $\nu_A > \nu_B$.

To reparameterize the distribution of Y, one can provide a one-to-one correspondence between Y and the pair (N, ε), where $N = N_A \cup N_B$ is the joint process and $\varepsilon = (\varepsilon_T)_{T \in N}$ is a set of marks

such that $\varepsilon_T = 1$ if T belongs to N_A and $\varepsilon_T = -1$ if T belongs to N_B (Kingman, 1993, pp. 16 and 53). Note that N is a Poisson process with intensity $\lambda = \nu_A + \nu_B$ and that, conditionally on N, the ε_T are independent variables with distribution

$$\varepsilon_T \mid N \sim 2\mathcal{B}\left\{\frac{\theta(T)+1}{2}\right\} - 1,$$

where $\mathcal{B}(p)$ represents the Bernoulli distribution with parameter p and

$$\theta(t) = \frac{\nu_A(t) - \nu_B(t)}{\nu_A(t) + \nu_B(t)}, \quad 0 \leqslant t \leqslant 1.$$

The two-sample null hypothesis is then once again that $\theta(t) = 0$. Since λ is unknown, this hypothesis is composite.

2·4. *Conditional distributions*

Our strategy is based on conditional testing (Loader, 1991; Rivera & Walther, 2013). For the homogeneity test, conditionally on the total number of points $N([0,1]) = n$, the observed process is actually a sample of density $1+\theta$ that does not depend on λ (Loader, 1991; Rivera & Walther, 2013). Similarly, in the two-sample case, given the positions of the joint process N, the remaining variability under the null hypothesis lies in the distribution of the marks $(\varepsilon_T)_{T\in N}$, which depends only on θ. Let us introduce the variable $\mathfrak{N}$ such that $\mathfrak{N} = N([0,1])$ for the homogeneity test and $\mathfrak{N} = N$ for the two-sample test. The nature of $\mathfrak{N}$ is different in the two cases: it is only a single variable for the homogeneity test, while it is an entire process in the two-sample case. Our procedure relies on the distribution of Y conditionally on $\mathfrak{N}$, denoted by P_θ below, which does not depend on λ; E_λ will refer to the expectation with respect to $\mathfrak{N}$.

3. INFINITE SET OF HYPOTHESES AND INTRODUCTION OF A p-VALUE PROCESS

3·1. *An infinite set of local null hypotheses*

Our continuous testing procedure aims to determine where the signal θ is nonzero. The classical strategy consists in testing the full null hypothesis $\{\theta = \theta_0\}$, where θ_0 is the null function on $[0,1]$. Therefore a unique hypothesis is tested and the test produces a binary answer. Here we wish to test $H_{0,t} : \{\theta(t) = 0\}$ versus $H_{1,t} : \{\theta(t) \neq 0\}$ or $H_{1,t} : \{\theta(t) > 0\}$, for $0 \leqslant t \leqslant 1$. Since finding information at the single-point level is impossible without smoothness assumptions on θ, we focus on finding intervals or unions of intervals on which the null hypothesis holds. We denote by $\mathcal{H}_0(I)$ the null hypothesis on interval I, that is, the nullity of θ on the whole interval I. As a special case, $\mathcal{H}_0([0,1])$ is the full null hypothesis.

3·2. *A continuum of scanning windows*

Our procedure is based on scanning windows of fixed width $\eta \in (0,1)$. Here, we envision the entire continuum of windows of length η. To avoid confusion, in what follows x will always denote a window centre whereas t will always denote a point, i.e., a possible value for the observations Y. Thus we have $x \in \mathcal{X} = [\eta/2, 1-\eta/2]$. The scanning windows are denoted by $I_x = (x-\eta/2, x+\eta/2]$ for any $x \in \mathcal{X}$. Therefore, a typical relationship between x and t is $t \in I_x$. All the proposed multiple testing procedures are based on single tests of the null hypothesis $\mathcal{H}_0(I_x)$, for all possible window centres $x \in \mathcal{X}$. The continuous testing procedure aims to accept

the set of true windows, which is indexed by their centres: $J_0(\theta) = \{x \in \mathcal{X} : \theta \text{ null on } I_x\}$; we will write simply J_0 when it is unambiguous.

3·3. *Construction of the p-value process with the count statistics*

For the homogeneity test, consider the test statistic $S(x) = N(I_x)$, the number of points in the window with centre x. Then, classically, a one-sided test based on this statistic rejects $\mathcal{H}_0(I_x)$ if $S(x)$ is large enough. Recall that for the homogeneity test, the conditioning variable is $\mathfrak{N} = N([0,1])$. Under $\mathcal{H}_0(I_x)$, the conditional distribution of $S(x)$ given $\mathfrak{N} = n$ is binomial with parameters n and η, denoted by $\mathcal{B}(n,\eta)$. For all z, let $F_{\mathcal{B}(n,\eta)}(z) = \text{pr}(Z \geqslant z)$ with $Z \sim \mathcal{B}(n,\eta)$. The p-value process associated with the classical one-sided single test of $\mathcal{H}_0(I_x)$ is then defined by

$$p(x) = F_{\mathcal{B}(\mathfrak{N},\eta)}\{S(x)\}, \quad x \in \mathcal{X}.$$

Since $F_{\mathcal{B}(\mathfrak{N},\eta)}$ is explicit, $p(x)$ can be computed at very low computational cost. If λ is known and equals λ_0, the conditional distribution is no longer needed; the unconditional distribution of $S(x)$ under $\mathcal{H}_0(I_x)$ is Poisson with mean $\eta\lambda_0$, denoted by $\mathcal{P}(\eta\lambda_0)$, with associated tail distribution function $F_{\mathcal{P}(\eta\lambda_0)}$. This can be used directly to define the p-value process

$$p(x) = F_{\mathcal{P}(\eta\lambda_0)}\{S(x)\}, \quad x \in \mathcal{X}. \tag{1}$$

This easier set-up is quite close to that of Chan & Zhang (2007).

For the two-sample test, the individual test statistic is $S(x) = N_A(I_x)$. A classical one-sided test based on this statistic also rejects $\mathcal{H}_0(I_x)$ if $S(x)$ is large enough. In this set-up, the conditioning variable $\mathfrak{N}$ is the whole joint process N. Under $\mathcal{H}_0(I_x)$, the conditional distribution of $S(x)$ given $\mathfrak{N}$ is the same as the conditional distribution of $S(x)$ given $N(I_x)$, i.e., binomial with parameters $N(I_x)$ and $1/2$. Indeed, under $\mathcal{H}_0(I_x)$, a point of N has equal chance of belonging to N_A and to N_B. The corresponding p-value process is then

$$p(x) = F_{\mathcal{B}\{N(I_x),1/2\}}\{S(x)\}, \quad x \in \mathcal{X}. \tag{2}$$

3·4. *Kernel-based statistics*

Following Gretton et al. (2012) and Fromont et al. (2013), who showed the power improvement of kernel-based tests over count-based tests, we consider a Gaussian kernel of bandwidth h, denoted by K_h, to estimate $\|1 + \theta\|_2^2$ in the homogeneity case and $\|\nu_A - \nu_B\|_2^2$ in the two-sample case. For the homogeneity test, given $N([0,1]) = n$, the test is based on the U-statistic

$$\frac{1}{n(n-1)} \sum_{T,T' \in N \cap I_x, T' \neq T} K_h(T - T').$$

For the two-sample case, the test statistic is

$$\sum_{T,T' \in N \cap I_x, T' \neq T} K_h(T - T')\, \varepsilon_T\, \varepsilon_{T'},$$

where N is the joint process and the ε_T are the marks. The derivations of our kernel-based statistics in the different cases and the Monte Carlo approximation of the p-values are presented in the Supplementary Material.

3·5. *Making continuous testing computationally tractable*

Even if the method involves a continuum of tests, our procedure is computationally tractable because it is based on $Y(I_x)$, the number of points within window I_x, which can be computed very efficiently and which is piecewise constant in x. In full generality, single test statistics may depend only on the composition of each window, $Y \cap I_x$, which records the random positions of points in Y that lie in I_x. In particular, this is the case for the kernel-based statistics. Since the composition is piecewise constant, with changes only at times when a point leaves or enters the scanning window, the single test statistic that depends only on the composition is piecewise constant on a random finite partition $\tau = (\tau_m)_{0 \leqslant m \leqslant M}$. Therefore, the continuum of tests reduces to a random finite number of tests, one for each segment defined by τ. An example is provided in the Supplementary Material.

3·6. *General definition of the p-value process*

We assume that a test statistic $S(x)$ is given for each window I_x and depends only on the composition of the window. Using the independence property of Poisson processes between disjoint sets, it is easy to show the following subset pivotality property (Westfall & Young, 1993; Romano & Wolf, 2005).

Property 1. For all θ, the conditional distribution of $\{S(x)\}_{x \in J_0(\theta)}$ given $\mathfrak{N}$ is the same whether $Y \sim P_\theta$ or $Y \sim P_{\theta_0}$.

Thus, the conditional distribution of $S(x)$ given $\mathfrak{N}$ is not modified by the true or false character of the remaining hypotheses. In particular, the conditional distribution of $S(x)$ given $\mathfrak{N}$ under $\mathcal{H}_0(I_x)$ is the same as under the full null hypothesis $\mathcal{H}_0([0,1])$. In general, the conditional distribution of $S(x)$ given $\mathfrak{N}$ depends not only on $\mathfrak{N}$ but also on the composition of the window centred at x. Therefore we define

$$F_{\mathfrak{N},x}(s) = P_{\theta_0}\{S(x) \geqslant s \mid \mathfrak{N}\}, \quad x \in \mathcal{X},$$

and the associated p-value process

$$p(x) = F_{\mathfrak{N},x}\{S(x)\}, \quad x \in \mathcal{X}, \tag{3}$$

which satisfies the following property.

Property 2. For all (θ, λ) and all $\alpha \in [0,1]$, if $\mathcal{H}_0(I_x)$ is true, then $P_{\theta,\lambda}\{p(x) \leqslant \alpha \mid \mathfrak{N}\} \leqslant \alpha$ almost surely.

Moreover, since $p(x)$ depends only on $S(x)$ conditionally on $\mathfrak{N}$, Property 1 also holds for $p(x)$, not just for $S(x)$.

4. FALSE POSITIVE CONTROL IN CONTINUOUS TIME

4·1. *Definition of the multiplicity error rates*

The result of a continuous testing procedure is a rejection set $\mathcal{R} \subset \mathcal{X}$ that is a function of the p-value process $\{p(x)\}_{x \in \mathcal{X}}$. A typical example is the rejection set with fixed threshold u, defined by $\mathcal{R}_u = \{x \in \mathcal{X} : p(x) \leqslant u\}$. The set of accepted windows is defined by the complementary set $\mathcal{A} = \mathcal{R}^c$. If the procedure behaves properly, the set $\mathcal{A}$ should be a good approximation to J_0;

or, similarly, $\mathcal{R}$ should be a good approximation to J_0^c. To evaluate the quality of a procedure $\mathcal{R}$, we focus on false positive windows that correspond to elements of $J_0 \cap \mathcal{R}$, whose size can be gauged in many ways. From a familywise point of view, one wants to avoid any false positive with a large probability, which corresponds to control of

$$\text{FWER}_{\theta,\lambda}(\mathcal{R}) = P_{\theta,\lambda}(J_0 \cap \mathcal{R} \neq \emptyset). \tag{4}$$

A procedure with controlled familywise error rate avoids false positive windows whatever their number and is generally conservative. A more balanced attitude is to allow the presence of a prespecified fraction of false positives among the positives. This can be achieved by controlling the false discovery rate

$$\text{FDR}_{\theta,\lambda}(\mathcal{R}) = E_{\theta,\lambda}\left\{\frac{\Lambda\,(J_0 \cap \mathcal{R})}{\Lambda\,(\mathcal{R})}\right\}, \tag{5}$$

with the convention that $0/0 = 0$. Since the setting here is continuous, the quantity of false positives is evaluated above by the Lebesgue measure Λ on $[0, 1]$.

4·2. *A continuous* min-*p procedure to control the familywise error rate*

We would like to find a threshold $u_\alpha \in [0, 1]$ such that for all (θ, λ), $\text{FWER}_{\theta,\lambda}(\mathcal{R}_{u_\alpha}) \leqslant \alpha$. For all $u \in [0, 1]$, the event $\{J_0 \cap \mathcal{R}_u \neq \emptyset\}$ means that there exists $x \in J_0$ such that $p(x) \leqslant u$. Therefore

$$\{J_0 \cap \mathcal{R}_u \neq \emptyset\} = \left\{\inf_{x \in J_0} p(x) \leqslant u\right\},$$

with $p(x)$ as defined in (3). The set J_0 being unknown, one can use some rough upper bound. Thanks to Property 1, for any $u > 0$, possibly depending on the conditioning variable $\mathfrak{N}$,

$$\text{FWER}_{\theta,\lambda}(\mathcal{R}_u) = E_\lambda\left[P_\theta\left\{\inf_{x \in J_0} p(x) \leqslant u \;\middle|\; \mathfrak{N}\right\}\right] = E_\lambda\left[P_{\theta_0}\left\{\inf_{x \in J_0} p(x) \leqslant u \;\middle|\; \mathfrak{N}\right\}\right].$$

Since $J_0 \subset \mathcal{X}$, we obtain

$$\text{FWER}_{\theta,\lambda}(\mathcal{R}_u) \leqslant E_\lambda\left[P_{\theta_0}\left\{\inf_{x \in \mathcal{X}} p(x) \leqslant u \;\middle|\; \mathfrak{N}\right\}\right].$$

Conditionally on $\mathfrak{N}$, the distribution of $\inf_{x \in \mathcal{X}} p(x)$ under θ_0 can be determined, and one can choose a suitable threshold u_α. However, due to the subtleties of discrete distributions, it is more powerful to use

$$q(x) = F^{\inf}_{\mathfrak{N}}\{p(x)\}, \quad x \in \mathcal{X}, \tag{6}$$

where $F^{\inf}_{\mathfrak{N}}$ is the conditional cumulative distribution function of $\inf_{x \in \mathcal{X}} p(x)$ under the full null hypothesis, i.e., $\theta = \theta_0$.

THEOREM 1. *Let $\alpha \in (0, 1)$ and let q be given by (6). Then for all (θ, λ) the procedure defined by $\mathcal{R}^{\inf}_\alpha = \{x \in \mathcal{X} : q(x) \leqslant \alpha\}$ satisfies* $\text{FWER}_{\theta,\lambda}(\mathcal{R}^{\inf}_\alpha) \leqslant \alpha$.

The $q(x)$ are often called adjusted p-values, as they are already adjusted to the multiplicity of the test (Dudoit & van der Laan, 2008, § 4.3.3).

Remark 1. The adjusted p-values given by (6) depend on $F^{\inf}_{\mathfrak{N}}$, which needs to be estimated, even for count-based test statistics. A Monte Carlo approximation yields the following substitute for $q(x)$:

$$\hat{q}(x) = \frac{1}{B+1}\left[1 + \sum_{b=1}^{B} 1_{\{\inf_{x\in\mathcal{X}} p^b(x) \leqslant p(x)\}}\right], \tag{7}$$

where each $p^b(x)$ $(b = 1, \ldots, B)$ is the p-value computed on the bth resampled process. The resampling process can be derived as follows. For the homogeneity test with unknown λ, conditionally on the event $N([0,1]) = n$, the bth resampled observation is obtained by simulating n independent and identically distributed random points uniformly on $[0,1]$. For the two-sample test, conditionally on the joint process N, the bth resampled observation is generated by simulating a set of independent uniform signs that mark the points of N. The familywise error rate control given in Theorem 1 is maintained upon replacing $q(x)$ by $\hat{q}(x)$ even if the underlying p-value process $p(x)$ in (7) is also the outcome of a Monte Carlo procedure; see the Supplementary Material.

Remark 2. The procedures proposed above are single-step, which means that the thresholds are calibrated by taking a minimum over the whole of $\mathcal{X}$. This can be refined by using a step-down procedure; see the Supplementary Material.

4·3. *Link with the scan statistic framework*

The p-value process for the homogeneity test does not depend on x. Moreover, using the notation $\mathfrak{N} = N([0,1])$, the function $F_{\mathcal{B}(\mathfrak{N},\eta)}$ is nonincreasing. Therefore

$$\inf_{x\in\mathcal{X}} p(x) = F_{\mathcal{B}(\mathfrak{N},\eta)}\Big\{\sup_{x\in\mathcal{X}} S(x)\Big\}. \tag{8}$$

Hence, rejecting for small values of the infimum of the p-value process and rejecting for large values of the supremum of the test statistics are equivalent (Dudoit et al., 2003). Procedures using the scan statistics $\sup_{x\in\mathcal{X}} S(x)$ have also been investigated by Chan & Zhang (2007) under the full null hypothesis $\mathcal{H}_0([0,1])$. The main difference from our work on homogeneity tests is that our p-values are conditional and exact and do not depend on the unknown λ; in contrast, the p-values of Chan & Zhang (2007) rely on an asymptotic approximation of the unconditional full null distribution, which depends on the unknown λ. More generally, scan statistic procedures only provide weak familywise error rate control, that is, control under the full null distribution. By contrast, our approach offers strong control that is valid for any configuration of true/false null hypotheses.

Nevertheless, this analogy with scan statistic procedures does not hold if the p-value distribution depends on the composition of the window under the full null hypothesis. This is typically the case for the two-sample test, where the function $F_{\mathcal{B}\{N(I_x),1/2\}}$ also depends on x. In this case, (8) fails and scan procedures are substantially different from min-p procedures. The min-p procedure is more appropriate because the p-value transform of each statistic before applying the minimum puts all local tests on the same significance scale (Dudoit et al., 2003).

4·4. *A continuous weighted Benjamini–Hochberg procedure*

To increase power, we propose a second procedure based on control of the false discovery rate (5). We would like to find a threshold $v_\alpha \in [0,1]$ such that for all (θ,λ), $\mathrm{FDR}_{\theta,\lambda}(v_\alpha) \leqslant \alpha$.

Following Blanchard et al. (2014), let us first explain heuristically how one can achieve this goal in a continuous framework. By Fubini's theorem, for all $v \in [0,1]$,

$$\mathrm{FDR}_{\theta,\lambda}(v) = E_{\theta,\lambda}\left\{\frac{\Lambda\left(J_0 \cap \mathcal{R}_v\right)}{\Lambda(\mathcal{R}_v)}\right\} = \int_{J_0} E_{\theta,\lambda}\left[\frac{1_{\{p(x)\leqslant v\}}}{\Lambda(\mathcal{R}_v)}\right] \mathrm{d}\Lambda(x).$$

Then, choosing v such that

$$\frac{\Lambda(\mathcal{R}_v)}{\Lambda(\mathcal{X})} \geqslant \frac{v}{\alpha}, \tag{9}$$

we obtain

$$\begin{aligned}\mathrm{FDR}_{\theta,\lambda}(v) &\leqslant \alpha \int_{J_0} v^{-1} E_{\theta,\lambda}\left[\frac{1_{\{p(x)\leqslant v\}}}{\Lambda(\mathcal{X})}\right] \mathrm{d}\Lambda(x) \\ &\leqslant \frac{\alpha}{\Lambda(\mathcal{X})} \int_{J_0} \frac{P_{\theta,\lambda}\left\{p(x) \leqslant v\right\}}{v} \mathrm{d}\Lambda(x) \leqslant \alpha \frac{\Lambda(J_0)}{\Lambda(\mathcal{X})} \leqslant \alpha\end{aligned} \tag{10}$$

using Property 2. However, this reasoning is heuristic because (9) entails that v is data-driven and hence nondeterministic, which makes (10) an approximation. This nonetheless suggests choosing V_α as

$$V_\alpha = \max\left\{v \geqslant 0 : \frac{\Lambda(\mathcal{R}_v)}{\Lambda(\mathcal{X})} \geqslant \frac{v}{\alpha}\right\}. \tag{11}$$

The quantile V^α is random, even conditionally on $\mathfrak{N}$, because it depends on the observed p-value process. Nevertheless, since the p-value process is piecewise constant in our framework, finding V_α is very simple. We propose a step-up algorithm that is inspired by the weighted procedure of Benjamini & Hochberg (1997), except that our algorithm is based on random weights. Recall that the processes $\{S(x)\}_{x\in\mathcal{X}_n}$ and hence $\{p(x)\}_{x\in\mathcal{X}_n}$ are piecewise constant on the random partition $\tau = (\tau_m)_{0\leqslant m\leqslant M}$, with the convention that $\tau_0 = 0$ and $\tau_M = 1$. Therefore, for all thresholds v,

$$\Lambda(\mathcal{R}_v) = \sum_{m=1}^{M} (\tau_m - \tau_{m-1}) 1_{\{p(\tau_{m-1})\leqslant v\}}.$$

This implies that the threshold V_α defined in (11) can be found as in the following algorithm.

Algorithm 1.
Compute the weights $w_m = (\tau_m - \tau_{m-1})/\Lambda(\mathcal{X})$ $(1 \leqslant m \leqslant M)$.
Compute $p_m = p(\tau_{m-1})$ $(1 \leqslant m \leqslant M)$.
Order these p-values as $p_{\sigma(1)} \leqslant \cdots \leqslant p_{\sigma(M)}$ for an appropriate permutation σ of $\{1,\ldots,M\}$.
Compute $\hat{k} = \max[k \in \{1,\ldots,M\} : p_{\sigma(k)} \leqslant \alpha \sum_{\ell=1}^{k} w_{\sigma(\ell)}]$, with $\hat{k} = 0$ if the set is empty.
Compute V_α as $\alpha \sum_{\ell=1}^{\hat{k}} w_{\sigma(\ell)}$.
Set $\mathcal{R}_\alpha^{w\mathrm{BH}} = \{x \in \mathcal{X} : p(x) \leqslant V_\alpha\}$.

In contrast to the classical Benjamini–Hochberg step-up procedure, each p-value has its own weight, which is not uniform but rather adaptive to the data. This weight means that the quantity

of interest is not the number of tests but the proportion of time that the p-value process spends at a particular value.

Remark 3. Note that $\mathcal{R}_\alpha^{w\mathrm{BH}} = \{x \in \mathcal{X} : q(x) \leqslant \alpha\}$, where the adjusted p-value process is

$$q(x) = \min_{k: p_{\sigma(k)} \geqslant p(x)} \left\{ \frac{p_{\sigma(k)}}{\sum_{\ell=1}^{k} w_{\sigma(\ell)}} \right\}.$$

Blanchard et al. (2014) proved that such a procedure is valid under the assumptions of measurability and positive dependence, the latter being an extension of the standard assumption of positive regression dependent on each member of a subset introduced by Benjamini & Yekutieli (2001). The difficulty arises from the interplay between the numerator and denominator within the expectation in (10) via v, because $v = V_\alpha$ is random. While the measurability condition is satisfied here by virtue of the piecewise-constant property, the positive dependence seems difficult to check in general. For instance, it is known to fail for two-sided tests (Yekutieli, 2008). We prove here that such a positive dependence condition holds in the one-sided test setting, both for the homogeneity test with a known λ_0 and for the two-sample test, by using properties of Poisson processes.

THEOREM 2. *For all $\alpha \in (0, 1)$, for a p-value process given by either* (1) *or* (2)*, for one-sided null hypotheses, and for all (θ, λ), the procedure $\mathcal{R}_\alpha^{w\mathrm{BH}}$ satisfies*

$$\mathrm{FDR}_{\theta,\lambda}(\mathcal{R}_\alpha^{w\mathrm{BH}}) \leqslant \alpha. \tag{12}$$

This result is proved in the Supplementary Material. Theorem 2 justifies the use of a weighted step-up procedure to control the continuous false discovery rate. It is valid for the one-sided version of our test statistics. However, our simulation results show that the false discovery rate control (12) is achieved in every situation we considered, even for more complex p-value processes than those in Theorem 2, e.g., using kernel-based statistics and/or two-sided testing. While control for the two-sample test is new, control for the homogeneity test with known intensity in Theorem 2 is similar to Corollary 4.5 in Blanchard et al. (2014). However, an important distinction concerns how the null hypotheses are defined: windows in our case versus points in their case. A consequence is that our approach does not rely on regularity of the underlying intensity λ. This makes the p-value process here computable in practice without prior knowledge of such regularity.

5. SIMULATIONS

5·1. *Simulations under the full null hypothesis*

We consider testing homogeneity under the full null hypothesis, $\mathcal{H}_0([0, 1])$, which corresponds to $\{\theta = \theta_0\}$; see § 3·1. We simulate a homogeneous Poisson process with intensity λ. We fix the window size of our procedure at $\eta = 0{\cdot}05$. The asymptotic methods proposed by Chan & Zhang (2007) and Siegmund et al. (2011) are only valid when there are enough observations in the testing windows, as shown in Table 1 with $\lambda = 5000$; otherwise they do not ensure the control of joint error rates. Moreover, the method of Siegmund et al. (2011) requires an additional discretization step that strongly influences the results.

Our min-p procedure strictly controls the familywise error rate but appears overly conservative for the count statistic, whereas the nominal level is attained with the kernel statistic. Since this

Table 1. *Error rate* (%) *control under the full null hypothesis for the one-sided homogeneity test; the level of the test is* $\alpha = 10\%$ *and the window size is* $\eta = 0{\cdot}05$

	FWER			FDR			
	CZ	min-p		SZY.5	SZY.50	wBH	
λ	Count	Count	Kernel	Count	Count	Count	Kernel
500	51	1·8	8	49	64	8	4
1000	45	1·0	10	42	59	5	4
5000	8	0·2	11	6	34	6	7

FWER, familywise error rate; FDR, false discovery rate; CZ: method of Chan & Zhang (2007); SZY, method of Siegmund et al. (2011) with 5 or 50 windows; wBH, weighted Benjamini–Hochberg procedure.

difference is specific to the min-p procedure, it can be explained by the discreteness of the count statistic, which induces discrete p-values with an infimum distribution of small finite support. Consequently, in this case it is impossible to produce p-values close to a prespecified level after adjustment. A classical way to circumvent this problem is to use randomization, but this would provide a nonreproducible p-value process, which may not be desirable in practice. Here we instead propose using the kernel-based test to derive a less conservative procedure. Our continuous Benjamini–Hochberg procedure controls the false discovery rate in every situation, even if it may be too conservative in some cases. The control is overly conservative, which may be related to the positive dependency hypothesis on which the procedure is constructed. Unfortunately, this hypothesis seems difficult to circumvent. Conclusions for two-sample tests are similar; see the Supplementary Material.

5·2. *Simulations with an alternative*

We consider a piecewise-constant signal function θ such that $\theta(t) = +\theta^*$ for $t \in \mathcal{I}_1^+$, $\theta(t) = -\theta^*$ for $t \in \mathcal{I}_1^-$, and $\theta(t) = 0$ otherwise. The full description of this function and the definitions of the intervals $\mathcal{I}_1^+$ and $\mathcal{I}_1^-$ are provided in the Supplementary Material. The parameter θ^* is used to measure the distance to homogeneity, from the nearly homogeneous regime to high separability. For a fixed background intensity parameter λ, we draw 1000 Poisson processes of intensity $\nu(t) = \lambda\{1 + \theta(t)\}$. To compare with other scanning procedures, we use one-sided tests. In the following, η is set to 0·05.

The results are similar to those for the full null configuration. Asymptotic methods control error rates when the number of observations per window is sufficient, as shown in Fig. 1. As our methods are non-asymptotic, they control the familywise error rate and the false discovery rate whatever the intensity of the signal. The min-p procedure is still overly conservative, and the kernel statistic does not allow the procedure to reach the nominal level, because the distribution of the infimum of the p-value process is considered over $\mathcal{X}$ rather than J_0, which is unknown. Step-down procedures, as defined in the Supplementary Material, can be used to overcome this difficulty. We also present Type II errors in the Supplementary Material, but they should be compared with caution since the different procedures do not have the same Type I errors.

6. Do neurons respond differently to two stimulii ?

The succession of action potentials of a neuron, called a spike train, constitutes a key biological signal that carries functional information. In particular, understanding the rapid changes in firing

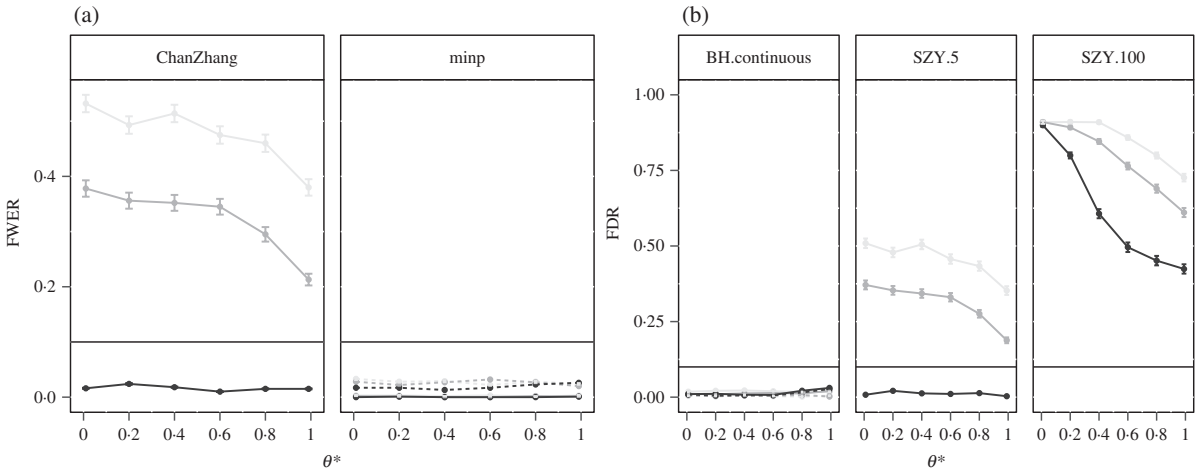

Fig. 1. Comparison of methods for the control of Type I error rates ($\alpha = 10\%$): (a) method of Chan & Zhang (2007), ChanZhang, and our min-p procedure, minp; (b) Benjamini–Hochberg procedure, BH.continuous, and the method of Siegmund et al. (2011) with 5 and 100 windows of identical length, SYZ.5 and SYZ.100. In each panel the lines correspond to $\lambda = 5000$ (black), $\lambda = 1000$ (dark grey), $\lambda = 500$ (light grey), count statistic (solid), and kernel statistic (dashed). FWER and FDR refer to the measures in (4) and (5); the familywise error rate is controlled using a single-step min-p procedure.

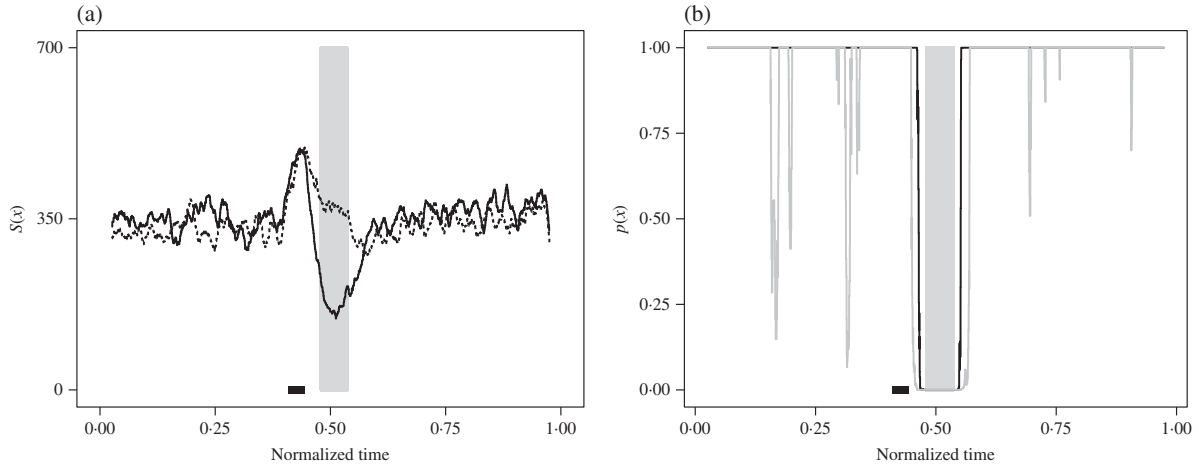

Fig. 2. Two-sample test for the spike trains of a cockroach neuron (neuron 2) excited at $[0{\cdot}42, 0{\cdot}45]$ normalized time (black band): (a) count statistic $S(x)$ with $\eta = 1/20$ for citronellal (solid) and terpineol (dotted), with the grey strip representing the time interval rejected by the weighted Benjamini–Hochberg procedure ($\alpha = 5\%$); (b) p-value process $p(x)$ adjusted with the min-p (black) or weighted Benjamini–Hochberg (grey) procedures.

rate due to a stimulus (Kass et al., 2003) is challenging, and inhomogeneous Poisson processes have often been used to model such data (Shimazaki & Shinomoto, 2010). To illustrate our procedure we analysed public experimental data (Pouzat & Chaffiol, 2009) that consist of spike trains of a cockroach neuron stimulated by odour puffs of citronellal or terpineol, two different chemical molecules.

We focus on the differential response of the neuron to different stimuli using a two-sided two-sample test; see the Supplementary Material. The data were scaled to be in $[0, 1]$, and the time variable after such pre-processing is called normalized time; see Fig. 2. Classical confidence bands obtained without multiplicity correction, as done by the STAR package (Pouzat & Chaffiol, 2009), would detect differences even before the stimulus, which does not seem to agree with the experimental set-up. Our min-p and weighted Benjamini–Hochberg procedures identify a region after the stimulus. Despite differences in the counts, nothing is detected before the stimulus, and

the differential response is detected between 0·47 and 0·53 in normalized time, which corresponds to a 0·625 ms delay after stimulation.

7. Conclusion

The application to experimental data motivates extensions of our framework to the testing of more than two conditions or to other tests, such as local independence tests, that are essential in neuroscience (Albert et al., 2015). Calibrating the window size remains an open question. Although the subject has been extensively investigated in the context of model estimation, only a few articles have considered the problem in the testing framework (Gao & Gijbels, 2008). One challenge will be to determine if the continuous testing framework would be appropriate for deriving a calibration method to automatically select the resolution of the test from the data.

Acknowledgement

We are grateful to Anne-Laure Fougères and Matthieu Lerasle for discussions, and to Ivan Bardet for his work on the Chan & Zhang (2007) paper. This research was partially supported by the Agence Nationale de la Recherche of France. Franck Picard is also affiliated with the Université de Lyon. Patricia Reynaud-Bouret is also affiliated with the Université Côte d'Azur.

Supplementary material

Supplementary material available at *Biometrika* online includes further test procedures and the proofs of Theorems 1 and 2.

References

Albert, M., Bouret, Y., Fromont, M. & Reynaud-Bouret, P. (2015). Bootstrap and permutation tests of independence for point processes. *Ann. Statist.* **43**, 2537–64.

Arias-Castro, E., Castro, R. M., Tánczos, E. & Wang, M. (2018). Distribution-free detection of structured anomalies: Permutation and rank-based scans. *J. Am. Statist. Assoc.* **113**, 789–801.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B **57**, 289–300.

Benjamini, Y. & Hochberg, Y. (1997). Multiple hypotheses testing with weights. *Scand. J. Statist.* **24**, 407–18.

Benjamini, Y. & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29**, 1165–88.

Blanchard, G., Delattre, S. & Roquain, E. (2014). Testing over a continuum of null hypotheses with false discovery rate control. *Bernoulli* **20**, 304–33.

Chan, H. & Zhang, N. (2007). Scan statistics with weighted observations. *J. Am. Statist. Assoc.* **102**, 595–602.

Dudoit, S., Shaffer, J. P. & Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statist. Sci.* **18**, 71–103.

Dudoit, S. & van der Laan, M. J. (2008). *Multiple Testing Procedures with Applications to Genomics*. New York: Springer.

Dümbgen, L. & Spokoiny, V. G. (2001). Multiscale testing of qualitative hypotheses. *Ann. Statist.* **29**, 124–52.

Fromont, M., Laurent, B. & Reynaud-Bouret, P. (2011). Adaptive tests of homogeneity for a Poisson process. *Ann. Inst. Henri Poincaré Prob. Statist.* **47**, 176–213.

Fromont, M., Laurent, B. & Reynaud-Bouret, P. (2013). The two-sample problem for Poisson processes: Adaptive tests with a non-asymptotic wild bootstrap approach. *Ann. Statist.* **41**, 1431–61.

Fu, J. C., Wu, T.-L. & Lou, W. Y. W. (2012). Continuous, discrete, and conditional scan statistics. *J. Appl. Prob.* **49**, 199–209.

Gao, J. & Gijbels, I. (2008). Bandwidth selection in nonparametric kernel testing. *J. Am. Statist. Assoc.* **103**, 1584–94.

Gretton, A., Borgwardt, K., Rasch, M., Schoelkopf, B. & Smola, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res.* **13**, 723–73.

KASS, R. E., VENTURA, V. & CAI, C. (2003). Statistical smoothing of neuronal data. *Network Comp. Neural Syst.* **14**, 5–15.

KINGMAN, J. F. C. (1993). *Poisson Processes*. Oxford: Clarendon Press.

KULLDORF, M. (1997). A spatial scan statistic. *Commun. Statist.* A **26**, 1481–92.

KULLDORFF, M., HEFFERNAN, R., HARTMAN, J., ASSUNÇÃO, R. & MOSTASHARI, F. (2005). A space–time permutation scan statistic for disease outbreak detection. *PLoS Med.* **2**, e59.

KULLDORFF, M., HUANG, L. & KONTY, K. (2009). A scan statistic for continuous data based on the normal probability model. *Int. J. Health Geogr.* **8**, 58. DOI: 10.1186/1476-072X-8-58.

LOADER, C. R. (1991). Large-deviation approximations to the distribution of scan statistics. *Adv. Appl. Prob.* **23**, 751–71.

NAUS, J. I. (1982). Approximations for distributions of scan statistics. *J. Am. Statist. Assoc.* **77**, 177–83.

PERONE PACIFICO, M., GENOVESE, C., VERDINELLI, I. & WASSERMAN, L. (2004). False discovery control for random fields. *J. Am. Statist. Assoc.* **99**, 1002–14.

PERONE PACIFICO, M., GENOVESE, C., VERDINELLI, I. & WASSERMAN, L. (2007). Scan clustering: A false discovery approach. *J. Mult. Anal.* **98**, 1441–69.

PIPA, G., GRÜN, S. & VAN VREESWIJK, C. (2013). Impact of spike train autostructure on probability distribution of joint spike events. *Neural Comp.* **25**, 1123–63.

POUZAT, C. & CHAFFIOL, C. (2009). Automatic spike train analysis and report generation: An implementation with R, R2HTML and STAR. *J. Neurosci. Meth.* **181**, 119–44.

R DEVELOPMENT CORE TEAM (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.

RIVERA, C. & WALTHER, G. (2013). Optimal detection of a jump in the intensity of a Poisson process or in a density with likelihood ratio statistics. *Scand. J. Statist.* **40**, 752–69.

ROMANO, J. P. & WOLF, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *J. Am. Statist. Assoc.* **100**, 94–108.

SHIMAZAKI, H. & SHINOMOTO, S. (2010). Kernel bandwidth optimization in spike rate estimation. *J. Comp. Neurosci.* **29**, 171–82.

SIEGMUND, D. O., ZHANG, N. R. & YAKIR, B. (2011). False discovery rate for scanning statistics. *Biometrika* **98**, 979–85.

TULEAU-MALOT, C., ROUIS, A., GRAMMONT, F. & REYNAUD-BOURET, P. (2014). Multiple tests based on a Gaussian approximation of the unitary events method. *Neural Comp.* **26**, 1408–54.

WALTHER, G. (2010). Optimal and fast detection of spatial clusters with scan statistics. *Ann. Statist.* **38**, 1010–33.

WESTFALL, P. H. & YOUNG, S. S. (1993). *Resampling-Based Multiple Testing*. New York: John Wiley & Sons.

YEKUTIELI, D. (2008). False discovery rate control for non-positively regression dependent test statistics. *J. Statist. Plan. Infer.* **138**, 405–15.

[*Received on* 30 *June* 2017. *Editorial decision on* 1 *June* 2018]