Optimal Generalized Sequential Monte Carlo Test

I. R. Silva*, R. M. Assunção

Departamento de Estatística, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

Abstract

Conventional Monte Carlo tests require the simulation of m independent copies from the test statistic U under the null hypothesis H_0. The execution time of these procedures can be substantially reduced by a sequential monitoring of the simulations. The sequential Monte Carlo test power and its expected time are analytically intractable. The literature has evaluated the properties of sequential Monte Carlo tests implementations by using some restrictions on the probability distribution of the p-value statistic. Such restrictions are used to bound the resampling risk, the probability that the accept/reject decision is different from the decision from the exact test. This paper develops a generalized sequential Monte Carlo test that includes the main previous proposals and that allows an analytical treatment of the power and the expected execution time. This results are valid for any test statistic. We also bound the resampling risk and obtain optimal schemes minimizing the expected execution time within a large class of sequential design.

Keywords: Sequential Monte Carlo test, Power loss, p-value density, Resampling risk, Sequential design, Sequential probability ratio test.

1. Introduction

In the conventional Monte Carlo (MC) tests, the user selects the number m of simulations of the test statistic U under H_0. A Monte Carlo p-value is calculated based on the proportion of the simulated values that are larger or equal than the observed value of U, assuming that large values of U lead to the null hypothesis rejection. This procedure can take a long time to run if the test statistic requires a complicated calculation as, for example, those involved in complex models. These situations are exactly those where the MC tests are likely to be most useful, as analytical exact or asymptotic results concerning the test statistic U is hard to obtain. The adoption of sequential procedures to carry out MC tests is a way to reach a faster decision. In contrast with the fixed size MC procedure, in the sequential MC test the number of simulated statistics is a random variable. The basic idea is to stop simulating as soon as there is enough evidence either to reject or to accept the null hypothesis. For example, it is intuitively clear that, if the observed value of

*Corresponding author
Email addresses: irs@ufmg.br (I. R. Silva), assuncao@est.ufmg.br (R. M. Assunção)

U is close to the median of the first 100 simulated values, the null hypothesis is not likely to be rejected even if we perform another 950 simulations. If a valid p-value could be provided, most researchers would be confident to stop at this point. Sequential Monte Carlo tests are procedures that provide valid p-values in these situations.

Let X_t be the number of simulated statistics under H_0 exceeding the observed value u_0 at t-th simulation. In general, sequential MC procedures track the X_t evolution by checking if it crosses an upper or a lower boundary. When it does, the test is halted and a decision is reached. Typically, crossing the lower boundary leads to the rejection of the null hypothesis while the upper boundary crossing leads to the acceptance of the null hypothesis.

There are different proposals for a sequential Monte Carlo test in the statistical literature. Besag and Clifford (1991) proposed a very simple scheme that provides valid p-value for a sequential test with an upper bound $n-1$ in the number of simulations of U. It depends on a single tuning parameter h, making it extremely simple to use. We stop the simulations when $X_t = h$ for the first time and $t < n$. If $X_{n-1} < h$, the simulations are halted. If $h \leq l \leq n-1$ is the number of simulations carried out and if we stop at time t, the sequential p-value is given by

$$p_{BC} = \begin{cases} X_t/t, & \text{if } X_t = h, \\ (X_t+1)/n, & \text{if } X_t < h. \end{cases} \tag{1}$$

The support set of p_{BC} is

$$S = \{1/n, 2/n, \ldots, h/n, h/(n-1), \ldots, h/(h+2), h/(h+1), 1\} \ .$$

and we have $\mathbb{P}(P_s \leq a) = a$ under the null hypothesis if $a \in S$. This is a valid p-value estimator, because, a p-value estimator P_e is valid if $\mathbb{P}(P_e \leq b) \leq b$, where b is an element from the support set of P_e. Additional randomization can provide a continuous p-value with uniform distribution in the interval $(0, 1)$, rather than distributed on the discrete set S.

Therefore, the boundaries of Besag and Clifford (1991) are given by the horizontal line $X_t = h$ and the vertical line $t = n-1$. There is no lower boundary but only a predetermined maximum number of simulations, typically called a truncated sequential Monte Carlo test. The Besag and Clifford sequential MC test brings a reduction in execution time only when the null hypothesis is true. When it is false, one will often run the Monte Carlo simulation up to its upper bound $n-1$. Therefore, additional gains could be obtained by adopting a stopping criterium based on large values of X_t. For any fixed type I error probability α, Silva et al. (2009) showed that one can design a Besag and Clifford sequential MC test with the same power as a conventional Monte Carlo test and with shorter running time. Silva et al. (2009) showed also the puzzling result that this sequential Monte Carlo should have a maximum sample size equal to $h/\alpha + 1$, because, for $n \geq h/\alpha + 1$, the power is constant.

In addition to Besag and Clifford (1991), alternative sequential Monte Carlo tests have been suggested recently. These other procedures are mainly concerned with the resampling risk, defined by Fay and Follmann (2002) as the probability that the test decision of a realized MC test will be different from a theoretical MC test with an infinite number of replications. Fay and Follmann (2002) proposed the curtailed sampling design, where, if $X_t \geq \lfloor \alpha(n+1) \rfloor$, the procedure is interrupted and H_0 is not rejected, and, if $t - X_t \geq \lceil (1-\alpha)(n+1) \rceil$ or the number of simulations reaches n, the procedure is interrupted and H_0 is rejected, where n is the maximum number of simulations. They also introduced the interactive push out (IPO) procedure that requires a sequential algorithm to define the boundaries of the sequential procedure. This procedure is not proven to be optimal but simply to decrease the sample size with respect to a curtailed sampling design. For all their results, Fay and Follmann (2002) assumed a specific class of distribution for the p-value statistic, that distribution implied by a test statistic U that follows the standard normal distribution under the null hypothesis and follows a $N(\mu, 1)$ under the alternative hypothesis. Conditional to this class of distributions, they found numerically the worst distribution to bound the resampling risk. IPO has a smaller expected execution time than the curtailed sampling design but its implementation is not practical for bounding the resampling risk in arbitrarily low values such as 0.01, for example. Also, we think that the assumption on the p-value distribution is too restrictive and, in fact, we show that it is not necessary to obtain optimal procedures.

Fay et al. (2007) proposed an algorithm (and an R package) to implement a truncated Sequential Probability Ratio Test (tSPRT) to bound the resampling risk and studied its behaviour as a function of the p-value. The algorithm, denoted here as the FKH algorithm, calculates a valid p-value, which depends on the calculation of the number of ways to reach each point on the stopping boundary of the MC test.

Gandy (2009) proposed an algorithm to build a sequential MC test that uniformly bounds the resampling risk in arbitrarily small values and provides lower bounds to the expected number of simulations. His algorithm is not truncated and the expected number of simulations can be infinite for p-values close to α. Therefore, the simulations may go on indefinitely. One missing issue in his paper is the lack of results concerning the type I error probability when the number of simulations is truncated.

Kim (2010) explored the approach from Fay and Follmann (2002) to bound the resampling risk using their same restrictive class of p-value distributions. She used the B-value boundaries proposed by Lan and Wittes (1988) and applied the algorithm of Fay et al. (2007) to obtain valid p-values estimates. She was able to obtain arbitrarily low bounds to the resampling risk and showed empirically that the B-value boundaries produces a smaller expected number of simulations than the IPO designs. In this paper, she also defined an approximated B-value procedure, which is easy to calculate and has analytical formulas that give insights on the choice of parameter values of the exact B-value design.

These B-value boundaries have the main advantages from the other procedures cited and, in our opinion,

is the best alternative for a sequential MC test at the moment. However, its main results, concerning the resampling risk and the expected number of simulations, depend on the same restrictive class of p-value distributions of Fay and Follmann (2002). Moreover, important topics were not explored for the B-value boundaries such as, for example, its power with respect to the conventional MC test or the establishment of lower bounds for the expected number of simulations for any test statistic.

In this paper, we introduce a generalized sequential Monte Carlo allowing any monotonic shapes for the boundaries. For example, it is possible to construct boundaries which are close to each other in the beginning of the simulations, departing from each other as the simulations proceed and approaching each other again in the end of the simulations. We have been able to obtain bounds for the power loss of the sequential MC test. In fact, we establish boundaries shapes such that the sequential MC test has the same power as the conventional MC test for any α level. These boundaries are simple to calculate and they are valid in the general case of any p-value distribution. Moreover, we are able to provide an algorithm to find the truncated boundaries that lead to a design with minimum expected sampling size. Concerning the resampling risk, we consider a larger class of distributions for the p-value than Fay and Follmann (2002) and we show that it is suitable to explicit algebraic manipulation allowing simple bounding of the resampling risk for any sequential MC test design.

This paper is organized in the following way. In the next section, we describe the B-value boundaries. Section 3 defines our sequential MC test and develops its properties. In Section 4 we discuss a general class for the p-value distribution and provide some analytical results for the sequential tests. Section 5 presents a numeric routine for the preliminary choice of our boundaries and some specific suggestions for practical use. Section 6 offers a comparison between the B-value procedure and our procedure. Section 7 closes the paper with some discussion.

2. The B-value Procedure

Consider a hypothesis test of a null hypothesis H_0 against an alternative hypothesis H_a by means of a test statistic U. The MC test can be seen as an estimation procedure to the unknown decision from the exact test based on the null hypothesis distribution of U. Kim (2010) has adopted this point of view by seeing the MC test as a decision procedure concerning in which $(0, 1)$ interval, either $(0, \alpha]$ or $(\alpha, 1)$, does belong the exact p-value associated with the test statistic U. The parameter α is the significance level of the exact test. This interpretation leads to the following pair of hypotheses:

$$\begin{aligned} H_0^* &: p \leq \alpha \\ H_A^* &: p > \alpha \end{aligned} \tag{2}$$

where p is the observed and unknown p-value generated from the random variable p-value. Viewed as a random variable, we denote the p-value by P. Clearly, the decision in favor of any hypotheses above leads to a decision concerning the original hypotheses H_0 and H_A.

Let U be the test statistic, u_0 be its observed value for a fixed sample and $u_i, i = 1, ...,$ be the independently simulated values from U under H_0. Let

$$X_t = \sum_{i=1}^{t} 1_{\{[u_0,\infty)\}}(u_i) \,,$$

where $1_{\{[u_0,\infty)\}}(u_i)$ is the indicator function that $u_i \geq u_0$.

Kim (2010) used the B-value introduced by Lan and Wittes (1988) to propose a sequential procedure to test H_0^* versus H_A^*. Define:

$$V(t) = min\left\{s \geq 0 : x - tx \geq c_1\sqrt{n\alpha(1-\alpha)}\right\}$$

and

$$L(t) = max\left\{s \geq 0 : x - t\alpha \leq c_2\sqrt{n\alpha(1-\alpha)}\right\} \,.$$

Define also:

$$B_{\text{Sup}} = \left\{(t,x) = (t, \min\{V(t), r_1\}) : t = t_0^+, t_0^+ + 1, ..., n)\right\} \,,$$

the upper boundary, and

$$B_{\text{Inf}} = \left\{(t,x) = (t, \max\{L(t), t - r_0\}) : t = t_0^-, t_0^- + 1, ..., n)\right\} \,,$$

the lower boundary of a sequential Monte Carlo test, where t_0^+ is the smaller value of t such that $V(t) \leq t$ and t_0^- is the smaller value of t such that $L(t) \geq 0$. Similarly, let t_1^+ be the smaller value of t such that $V(t) \geq r_1$ and t_1^- the smaller value of t such that $L(t) \leq t - r_0$. The stopping boundaries from Kim (2010) are given by $B = B_{\text{Inf}} \cup B_{\text{Sup}}$. The B boundaries are formed by the union of linear functions in t. Figure 1 illustrate the B-boundaries B_{Sup} and B_{Inf} using $c_1 = -c_2 = 1.282$, $n = 600$ and $\alpha = 0.05$.

The upper boundary B_{Sup} is formed by the union of the line $V(t) = c_1\sqrt{n\alpha(1-\alpha)} + t\alpha$ until $t = t_1^+$, when the upper boundary becomes the horizontal line with height $r_1 = \lfloor \alpha(n+1) \rfloor$. The lower boundary B_{Inf} is formed by the line $L(t) = c_2\sqrt{n\alpha(1-\alpha)} + t\alpha$ up to $t = t_1^-$ when it becomes the vertical line $r_0 = t - \lceil (1-\alpha)(n+1) \rceil$.

Kim (2010) uses ϕ_{FKH}, the test criterium based on the valid p-value presented in Fay et al. (2007). The valid p-value is defined as $\hat{p}_v(X_t, t) = F_{\hat{p}_{MLE}}(X_t/t)$, where $\hat{p}_{MLE}$ is the maximum likelihood estimator of p and $F_{\hat{p}}$ is defined in (5.2) from Fay et al. (2007). The estimate $\hat{p}_v(X_t, t)$ of the p-value can be computed using the FKH algorithm. The test adopted by Kim (2010) for the B boundaries is given by:

$$\phi_{FKH}(t,x) = \begin{cases} 1, & \text{if } \hat{p}_v(x,t) \leq \alpha \\ 0, & \text{if } \hat{p}_v(x,t) > \alpha. \end{cases}$$

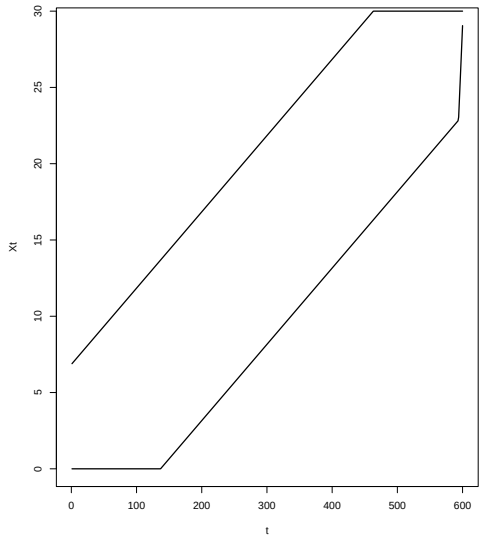

Figure 1: Example of the B boundaries with $\alpha = 0.05$ and maximum number of simulations equal to $n = 600$.

When $\phi_{FKH}(t, x) = 0$, H_0 is not rejected (because $H_0^* : p \leq \alpha$ is rejected). For $\phi_{FKH}(t, x) = 1$, H_0 is rejected (that is, $H_0^* : p \leq \alpha$ is not rejected). Henceforth, this procedure is called MC_B.

It is very important to remark that there is no need to check the value of X_t at every moment t. To see this, noticed that the boundaries B_{Sup} and B_{Inf} are composed by non-integer numbers while X_t is a count. As a consequence, there will be times t for which the simulations can not be interrupted by B_{Inf} and therefore there is no need to check against the lower boundary at these times. To illustrate this, consider the example from Kim (2010) illustrated in Figure 1. Table 1 shows the values of B_{Sup} and B_{Inf} between the times 134 and 179. The lower boundary is equal to zero until $t = 136$ and it is formed by numbers smaller than 1 until $t = 156$. Therefore, X_t reach the lower boundary during this period if $X_{137} = 0$ and there is no need to check against it for $t \leq 136$. Likewise, if X_t is not interrupted by B_{Inf} at $t = 156$ (that is, $X_{156} \geq 2$), it will not reach it at least until $t = 176$. Therefore, in practice, there is no need to check against the lower boundary for every simulated value. One needs to check only on those times t such that

$$B_{\text{Inf}}(t-1) < B_{\text{Inf}}(t)$$

for $t = 2, \ldots, m$ where $B_{\text{Inf}}(t)$ is the value of the lower boundary at time t. This will be explored by our generalized sequential Monte Carlo method described in Section 3.

Since B_{Sup} will typically be non-integer, it is always possible to define step functions equivalent to the upper boundary. To see this, consider again the Table 1. From $t = 134$ to $t = 143$, it is clear that the values $B_{\text{Sup}}(t)$ could be all substituted by 14 and the procedure would remain the same.

2.1. Bounding the resampling risk of MC_B

Fay and Follmann (2002) considered the IPO procedure that, interactively with the current simulations, adjusts the initial boundaries. This method allows the bounding of the resampling risk. The IPO procedure is not described in details here, but it should be noted that it is a computationally intensive procedure, and its implementation is intractable for bounding the resampling risk in arbitrarily small values (see (Kim, 2010)). Fay and Follmann (2002) considered a rather restrictive class of p-value distributions, with cumulative distribution function given by:

$$H_{\alpha,1-\beta}(p) = 1 - \Phi\left\{\Phi^{-1}(1-p) - \Phi^{-1}(1-\alpha) + \Phi^{-1}(\beta)\right\} \tag{3}$$

where $\Phi(.)$ is the cumulative distribution function of a standard Normal distribution, α is the desired significance level and β is the type II error probability. When $\alpha = 1 - \beta$, the cumulative distribution $H_{\alpha,1-\beta}(p)$ has a uniform distribution on $(0, 1)$, as is expected when H_0 is true.

The p-value distribution defined in (3) assumes a variety of shapes, but the analytical manipulation of the resampling risk or of the expected number of simulations is intractable. To circumvent this problem, Fay and Follmann (2002) used a Beta(a, b) distribution to approximate $H_{\alpha,1-\beta}(p)$, and this approximation is denoted by $\tilde{H}_{\alpha,1-\beta}(p)$. This approximation is chosen such that the expected value of P coincides with that from $H_{\alpha,1-\beta}(p)$ and such that $\tilde{H}_{\alpha,1-\beta}(\alpha) = H_{\alpha,1-\beta}(\alpha) = 1 - \beta$. Numerical studies were performed by Fay and Follmann (2002) to obtain the worst case $\tilde{F}$ within the class (3) in the sense of having the largest resampling risk. Let $\tilde{F}^*$ be the correspondent Beta distribution approximation to $\tilde{F}$.

Although MC_B is simpler and present a smaller expected time execution than the IPO procedure, it depends on the FKH algorithm which requires rather complex modifications for each type of sequential design. Kim (2010) proposes an approximation for the MC_B procedure. With this approximation, if B_{Sup} is reached before B_{Inf}, H_0^* is rejected, while H_A^* is accepted if B_{Inf} is reached first. The approximation may be used to gain analytic insights on the properties of the MC_B procedure or to help on choosing the parameters c_1, c_2, and n, as well as providing an approximation for the expected number of simulations. An undesirable characteristic of the approximated MC_B is that it is not truncated and the expected number of simulations must be calculated letting the maximum number of simulations go to infinity. Moreover, the approximation to MC_B does not offer guarantee that the type I error probability is under control for any choice of c_1 and c_2. For this reason, the approximation MC_B will not be explored here.

3. Our proposed generalized sequential Monte Carlo test

The analytical treatment of the MC test power function, when it is based on two interruption boundaries, is a cumbersome task. The reason is that it involves the calculation of the large number of possible trajectories of the random variable X_t responsible for H_0 rejection. Fay et al. (2007) present an algorithm to calculate

the terms associated with such number, and they used this algorithm to obtain both, the expected number of simulations and the resampling risk, for each fixed p-value. Fay et al. (2007) emphasize that such algorithm is valid only for the specific sequential procedure treated in that article, and adjustments are needed to use it with other sequential designs. Kim (2010) also used that algorithm for her calculations, and the approximate MC_B is an attempt to escape from the dependence on special algorithms.

Aiming to overcome this limitation, we propose a truncated sequential procedure with two boundaries that have the shape of step functions. The values of $X(t)$ are checked against the upper boundary for every t while they are checked against the lower boundary in an arbitrary set of predetermined discrete moments, possibly a smaller set than all integers between 1 and m. As we showed in Section 2, the B-boundaries can also be expressed by step functions with jumps equal to positive integer numbers. Therefore, the boundaries of MC_B and of our sequential procedure can be expressed in the same way. To express the boundaries by means of step functions is more cumbersome in terms of notation. The motivation for this design, where the lower boundary monitoring is not carried out for every time t, is mainly to allow for the analytical treatment of the power function, the expected number of simulations of the sequential MC test for any test statistic. We also bound the resampling risk of our sequential MC test.

Let $\eta^I = \{n_1^I, n_2^I, ..., n_{k_1}^I\}$, with $n_j^I < n_{j+1}^I$, be a set containing the moments when X_t must be checked against the lower boundary given by the values $I = \{I_1, I_2, ..., I_{k_1}\}$. If $X_{n_j^I} < I_j$, the simulations are interrupted and H_0 is rejected.

The monitoring of X_t with respect to the upper boundary crossing is carried out at all moments $t = 1, \ldots, m$ and this upper boundary is a step function. Let $\eta^S = \{n_1^S, n_2^S, ..., n_{k_2}^S\}$, with $n_j^S < n_{j+1}^S$ be the jump moments for the upper boundary. For $n_{j-1}^S \leq t < n_j^S$, the upper boundary is given by S_j where $n_0^S = 0$ and $S_1 < S_2 \ldots < S_{k_2}$. Let $S = \{S_1, S_2, ..., S_{k_2}\}$. Therefore, the simulations are interrupted if $D_t = 1$, where:

$$D_t = \begin{cases} 1, & \text{if } (t \in \eta^I \text{ and } X_t < I_j, \text{ for } t = n_j^I) \text{ or } (X_t = S_j, \text{ for } n_{j-1}^S < t \leq n_j^S) \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

or if the number of simulations reach a predetermined maximum equal to m.

Let x_t be the observed value of the random variable X_t. The p-value can be estimated by:

$$p_I = \begin{cases} x_t/t, & \text{if } x_t = S_j, n_{j-1}^S < t \leq n_j^S \\ (x_t + 1)/(t + 1), & \text{if } x_t < I_j, t = n_j^I. \end{cases}$$

We define the test decision function for this sequential test:

$$\phi_I(t, x) = \begin{cases} 1, & \text{if the lower boundary } I \text{ is reached before the upper } S \text{ or the simulations reach } m \\ 0, & \text{if the upper boundary } S \text{ is reached before the lower } I. \end{cases}$$

The hypothesis H_0 is rejected if $\phi_I = 1$ and it is not rejected if $\phi_I = 0$. This sequential MC test will be denoted by MC_G.

As an example, take $k_1 = k_2 = 10$, $m = 600$, and consider $I = \{0, 1, 2, 3, 9, 15, 20, 24, 27, 29\}$ for the lower boundary values, $S = \{5, 7, 9, 13, 17, 23, 26, 29, 29, 30\}$ for the upper boundary values, and $\eta^I = \eta^S = \{20, 50, 79, 119, 239, 359, 459, 539, 569, 600\}$. Figure 2 shows these boundaries as dashed lines.

The choice of the boundaries is closely linked to the desired α_{mc}, which is equal to 0.05 in this example. In Section 5, we present an algorithm to obtain the appropriate boundaries for any α_{mc} and m in an easy and fast way. The solid lines are the B boundaries calculated by Kim (2010) using $c_1 = -c_2 = 1.282$, $n = 600$, and $\alpha = 0.05$.

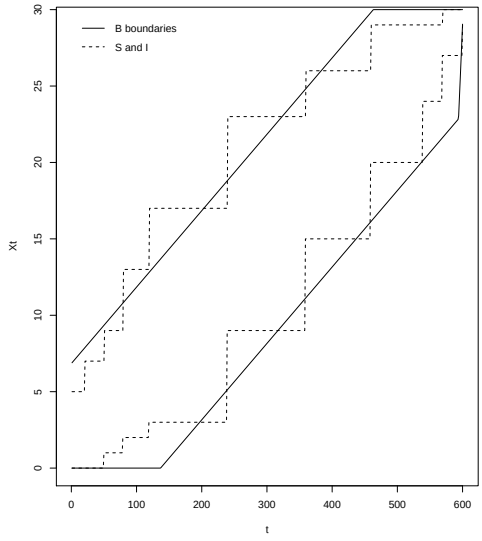

Figure 2: Example of the MC_G (S and I) and B boundaries with $\alpha = 0.05$ and a maximum number of simulations equal to $m = 600$.

3.1. Power and Size of the MC_G

In the MC_G procedure, the rejection of H_0 occurs in the first moment $t = n_j^I$ such that $x_t < I_j$. The power calculation is simpler if we merge the two sets η^I and η^S. Define $\eta = \eta^I \cup \eta^S = \{n_1, n_2, ..., n_k\}$ with $k = \#\eta$. Let $S' = \left\{S_1', S_2', ..., S_k'\right\}$ be the upper boundary adjusted for each $n_i \in \eta$ in the following way. If $n_i = n_j^S \in \eta^S$ for some j, then $S_i' = S_j$. If $n_i \in \eta^I \cap \left(\eta^S\right)^c$, then $S_i' = S_j$ where j is such that $n_j^S = \max\left\{n_r^S < n_i\right\}$. Thus, if n_i matches with some jump time in the set η^S, then S_i' is equal to the value in S for the time n_i. If n_i is not an element in η^S, then S_i' is the jump value of the time immediately preceding n_i.

Similarly, let $I' = \left\{I_1', I_2', ..., I_k'\right\}$ be the adjusted lower boundary. That is, when $n_i = n_j^I \in \eta^I$ or some j, then $I_i' = I_j$. If $n_i \in \eta^S$ but $n_i \notin \eta^I$, then $I_i' = I_j$ where j is such that $n_j^I = \max\left\{n_j^I < n_i\right\}$.

Thus, for a given value of $p \in (0,1)$, the power function of the MC_G procedure, is given by:

$$\begin{aligned}
\pi_G(p) &= \sum_{x_1=0}^{I_1'-1} \binom{n_1}{x_1} p^{x_1}(1-p)^{n_1-x_1} + \\
&+ \sum_{x_1=I_1'}^{\min\{S_1'-1, I_2'-1\}} \sum_{y=0}^{\min\{I_2'-x_1-1, n_2-n_1\}} \binom{n_2-n_1}{y} p^y (1-p)^{n_2-n_1-y} \times \\
&\times \left[\binom{n_1}{x_1} p^{x_1}(1-p)^{n_1-x_1} \right] + \\
&+ \sum_{j=2}^{k-1} \sum_{x_j=I_j'}^{\min\{S_j'-1, I_{j+1}'-1\}} \sum_{y=0}^{\min\{I_{j+1}'-x_j-1, n_{j+1}-n_j\}} \sum_{x_{j-1}=I_{j-1}'}^{\min\{S_{j-1}'-1, x_j\}} \cdots \\
&\cdots \sum_{x_1=I_1'}^{\min\{S_1'-1, x_2\}} \binom{n_{j+1}-n_j}{y} \times \\
&\times \binom{n_1}{x_1} p^{y+x_j}(1-p)^{n_{j+1}-y-x_j} \prod_{i=2}^{j} \binom{n_i-n_{i-1}}{x_i-x_{i-1}}. \qquad (5)
\end{aligned}$$

This expression is composed by k summands. If k is not too large, the direct application of this expression produces results quickly and easily. The calculation would be computationally hard if we used a similar expression for sequential procedure where $k = m$. Note that, in the MC_B procedure, the number of summands in (5) can reach up to $2\alpha_{mc}m$.

Under the null hypothesis, p follows the $U(0,1)$ distribution. Hence, by integrating out (5) with respect to p with a $U(0,1)$ density, we obtain the type I error probability for MC_G:

$$\begin{aligned}
\mathbb{P}\,(\text{type I error}) &= \int_0^1 \pi_G(p)dp = \frac{I_1'}{n_1+1} + \\
&+ \sum_{x_1=I_1'}^{\min\{S_1'-1, I_2'-1\}} \sum_{y=0}^{\min\{I_2'-x_1-1, n_2-n_1\}} \frac{\binom{n_2-n_1}{y}\binom{n_1}{x_1}}{(n_2+1)\binom{n_2}{y+x_1}} + \\
&+ \sum_{j=2}^{k-1} \sum_{x_j=I_j'}^{\min\{S_j'-1, I_{j+1}'-1\}} \sum_{y=0}^{\min\{I_{j+1}'-x_j-1, n_{j+1}-n_j\}} \sum_{x_{j-1}=I_{j-1}'}^{\min\{S_{j-1}'-1, x_j\}} \cdots \\
&\cdots \sum_{x_1=I_1'}^{\min\{S_1'-1, x_2\}} \frac{\binom{n_{j+1}-n_j}{y}\binom{n_1}{x_1}\prod_{i=2}^{j}\binom{n_i-n_{i-1}}{x_i-x_{i-1}}}{(n_{j+1}+1)\binom{n_{j+1}}{y+x_j}}.
\end{aligned}$$

Similarly to Silva and Assunção (2011), an upper bound for the power difference between MC_G and the exact test can be obtained by:

$$b_G = \max_{p\in(0,1)} \left\{1_{(0,\alpha]} - \pi_G(p)\right\} \tag{6}$$

where α is the significance level of the exact test.

The power function of $\pi_G(p)$ evaluated for a fixed p is equal to the probability of X_t reaching I before reaching S, and this probability is decreasing with p. In this way, the largest power loss of MC_G as compared to the exact test is given by:

$$b_G = \max_{p\in(0,\alpha]} \{1 - \pi_G(\alpha)\} = 1 - \pi_G(\alpha). \tag{7}$$

Let MC_m be the conventional MC test performed with a fixed number m of simulations. An upper bound for the power difference between MC_m and MC_G is given by:

$$b_{m,G} = \max_{p\in(0,1)} \{\pi_m(p) - \pi_G(p)\} \tag{8}$$

where $\pi_m(p) = \mathbb{P}(G \leq \lfloor m\alpha_{mc} \rfloor - 1)$ is the power function of MC_m for a given p, and G is distributed according to a binomial distribution with parameters $m-1$ and p.

3.2. Expected Number of Simulations for MC_G

Let L be the random variable that represents the number of simulations carried out until the halting moment. To perform the computation of the expectation of L, obtained by $\mathbb{E}(L|P=p) = \sum_{l=1}^{n_k} l\, \mathbb{P}(L = l|P=p)$, for each fixed p. The probability $\mathbb{P}(L=l|P=p)$ is given by:

$$\mathbb{P}(L=l|P=p) = \begin{cases} \begin{pmatrix} l-1 \\ l-S_1^{'} \end{pmatrix} p^{l-S_1^{'}}(1-p)^{S_1^{'}} & \text{if } l < n_1 \\ \begin{pmatrix} l-1 \\ l-S_1^{'} \end{pmatrix} p^{l-S_1^{'}}(1-p)^{S_1^{'}} + \sum_{x=0}^{I_1^{'}-1} \begin{pmatrix} n_1 \\ x \end{pmatrix} p^x(1-p)^{n_1-x} & \text{if } l = n_1 \\ \sum_{x=0}^{I_1^{'}-1} \begin{pmatrix} n_1 \\ x_1 \end{pmatrix} \begin{pmatrix} l-n_1-1 \\ l-n_1-(S_2^{'}-x) \end{pmatrix} p^{S_2^{'}}(1-p)^{l-S_2^{'}} & \text{if } n_1 < l < n_2 \\ \sum_{x=0}^{I_1^{'}-1} \begin{pmatrix} n_1 \\ x_1 \end{pmatrix} \begin{pmatrix} l-n_1-1 \\ l-n_1-(S_2^{'}-x) \end{pmatrix} p^{S_2^{'}}(1-p)^{l-S_2^{'}} + \\ \sum_{x=I_1^{'}}^{min\left\{S_1^{'}-1, I_2^{'}-1\right\}} \sum_{y=0}^{min\left\{I_2^{'}-x-1, n_2-n_1\right\}} \begin{pmatrix} n_2-n_1 \\ y \end{pmatrix} p^y(1-p)^{n_2-n_1-y} \times \\ \times \left[\begin{pmatrix} n_1 \\ x \end{pmatrix} p^x(1-p)^{n_1-x} \right] & \text{if } l = n_2. \end{cases}$$

We need to consider this calculation depending on l being equal to one of the n_j or not. For $l = n_j, j =$

$3, ..., k-1$, we have:

$$\begin{aligned}
\mathbb{P}(L=l|P=p) \quad = \quad & \sum_{x_{j-1}=I'_{j-1}}^{min\left\{S'_{j-1}-1, I'_j-1\right\}} \sum_{y=0}^{min\left\{I'_j-x_{j-1}-1, n_j-n_{j-1}\right\}} \sum_{x_{j-2}=I'_{j-2}}^{min\left\{S'_{j-2}-1, x_{j-1}\right\}} \cdots \\
\cdots \quad & \sum_{x_1=I'_1}^{min\left\{S'_1-1, x_2\right\}} \left(\begin{array}{c} n_j - n_{j-1} \\ y \end{array} \right) \times \\
\times \quad & \left(\begin{array}{c} n_1 \\ x_1 \end{array} \right) p^{y+x_{j-1}} (1-p)^{n_j-y-x_{j-1}} \prod_{i=2}^{j-1} \left(\begin{array}{c} n_i - n_{i-1} \\ x_i - x_{i-1} \end{array} \right) + \\
+ \quad & \sum_{x_{j-1}=I'_{j-1}}^{min\left\{S'_{j-1}-1, I'_j-1\right\}} \sum_{x_{j-2}=I'_{j-2}}^{min\left\{S'_{j-2}-1, x_{j-1}\right\}} \cdots \\
\cdots \quad & \sum_{x_1=I'_1}^{min\left\{S'_1-1, x_2\right\}} \left(\begin{array}{c} n_j - n_{j-1} - 1 \\ n_j - n_{j-1} - (S'_j - x_{j-1}) \end{array} \right) \times \\
\times \quad & \left(\begin{array}{c} n_1 \\ x_1 \end{array} \right) p^{l-S'_j} (1-p)^{S'_j} \prod_{i=2}^{j-1} \left(\begin{array}{c} n_i - n_{i-1} \\ x_i - x_{i-1} \end{array} \right).
\end{aligned}$$

For $n_{j-1} < l < n_j, j = 3, ..., k$:

$$\begin{aligned}
\mathbb{P}(L=l|P=p) \quad = \quad & \sum_{x_{j-1}=I'_{j-1}}^{min\left\{S'_{j-1}-1, I'_j-1\right\}} \sum_{x_{j-2}=I'_{j-2}}^{min\left\{S'_{j-2}-1, x_{j-1}\right\}} \cdots \\
\cdots \quad & \sum_{x_1=I'_1}^{min\left\{S'_1-1, x_2\right\}} \left(\begin{array}{c} n_j - n_{j-1} - 1 \\ n_j - n_{j-1} - (S'_j - x_{j-1}) \end{array} \right) \times \\
\times \quad & \left(\begin{array}{c} n_1 \\ x_1 \end{array} \right) p^{l-S'_j} (1-p)^{S'_j} \prod_{i=2}^{j-1} \left(\begin{array}{c} n_i - n_{i-1} \\ x_i - x_{i-1} \end{array} \right).
\end{aligned}$$

Finally, for $l = n_k$:

$$\begin{aligned}
\mathbb{P}(L=l|P=p) \quad = \quad & \sum_{x_{j-1}=I'_{j-1}}^{min\left\{S'_{j-1}-1, I'_j-1\right\}} \sum_{y=0}^{min\left\{I'_j-x_{j-1}-1, n_j-n_{j-1}\right\}} \sum_{x_{j-2}=I'_{j-2}}^{min\left\{S'_{j-2}-1, x_{j-1}\right\}} \cdots \\
\cdots \quad & \sum_{x_1=I'_1}^{min\left\{S'_1-1, x_2\right\}} \left(\begin{array}{c} n_j - n_{j-1} \\ y \end{array} \right) \times \\
\times \quad & \left(\begin{array}{c} n_1 \\ x_1 \end{array} \right) p^{y+x_{j-1}} (1-p)^{n_j-y-x_{j-1}} \prod_{i=2}^{j-1} \left(\begin{array}{c} n_i - n_{i-1} \\ x_i - x_{i-1} \end{array} \right). \qquad (9)
\end{aligned}$$

Using that p has a $U(0,1)$ distribution under the null hypothesis, we have

$$\mathbb{E}\left(L|H_0 \text{ is true}\right) = \int_0^1 \mathbb{E}(L|P=p)dp\,. \tag{10}$$

To calculate $\mathbb{E}(L)$ under H_A it is necessary to know the p-value distribution. However, a bound is easier to calculate as

$$b_{E(L)} = max_{p\in(0,1)}\left\{\mathbb{E}(L|P=p)\right\}. \tag{11}$$

$b_{E(L)}$ is a very conservative upper bound for $E(L)$. However, as we will illustrate in Section 5, this bound is useful to bound the expectation of L in values less than 65% of m.

4. A class of distributions for the p-value

Kim (2010) showed that, for $p=\alpha$, the resampling risk is at least 0.5. Hence, it is not possible to bound the resampling risk in relevant values if we allow all distributions of p-values. This is the reason to define a class for the p-value distribution, taken as the set $\Im$ of all continuous probability distributions in $(0,1)$ with differentiable densities that are non-increasing (that is, $f'_P(p) \leq 0$, for all $p \in (0,1)$, with f'_P representing the first derivative with respect to p of the p-value density function f.

From the p-value definition, its probability distribution function can be written in the following way:

$$\mathbb{P}(P \leq p) = 1 - F_A\left\{F_0^{-1}(1-p)\right\} \tag{12}$$

where F_A denotes the probability distribution function of the test statistic U under H_A and F_0 is the distribution of U under H_0.

Assuming the existence of densities functions $f_A(u)$ and $f_0(u)$ of U under H_A and H_0, respectively, the p-value density can be written as:

$$f_P(p) = \frac{f_A\left\{F_0^{-1}(1-p)\right\}}{f_0\left\{F_0^{-1}(1-p)\right\}}. \tag{13}$$

Hence, we can study the behavior of the p-value distribution by studying the behavior of the ratio between $f_A(u)$ and $f_0(u)$.

In the majority of the real applications, the ratio (13) is non-increasing with p and this is the motivation to restrict the analysis of the resampling risk to the set $\Im$. Let $\Im_B$ be the class of p-value distributions defined in Fay and Follmann (2002) with cumulative distribution $H_{\alpha,1-\beta}(p)$, as described in Section 2. Let π be the power of the exact test. We will show now that, for $\pi \geq \alpha$, $\Im$ is more general than $\Im_B$.

From the expression (3), the densities $h(p) \in \Im_B$ can be indexed by α and β and they are given by:

$$h_{\alpha,1-\beta}(p) = \exp\left\{-\frac{1}{2}\left[\Phi^{-1}(\beta) - \Phi^{-1}(1-\alpha)\right]\left[\Phi^{-1}(\beta) - \Phi^{-1}(1-\alpha) + 2\Phi^{-1}(1-p)\right]\right\} \tag{14}$$

where Φ^{-1} is the inverse function of the standard normal cumulative distribution function $\Phi(\cdot)$. The first derivative of $h_{\alpha,1-\beta}(p)$ with respect to p is equal to:

$$h'_{\alpha,1-\beta}(p) = \frac{\left[\Phi^{-1}(\beta) - \Phi^{-1}(1-\alpha)\right]}{\phi(\Phi^{-1}(1-p))} h_{\alpha,1-\beta}(p) \tag{15}$$

where $\phi(\cdot)$ is the density function of the standard normal distribution. For $1-\beta \geq \alpha$, we have $h'_{\alpha,1-\beta}(p) \leq 0$ for all $p \in (0,1)$.

Consider the subset of densities $\Im_B^* = \{f_P(p) \in \Im_B : 1-\beta \geq \alpha\}$. That is, $\Im_B^*$ is a subset from $\Im_B$ formed only by densities that implies an exact test power greater or equal to α. Therefore, $\Im_B^* \subset \Im$. Thus, at least for useful test statistic ($\mathbb{P}(P \leq p) \geq \alpha$), the class $\Im_B$ is a particular case from $\Im$.

The formulation of the class $\Im_B$ in Fay and Follmann (2002) was inspired on the behavior of the p-value distribution for the cases where $U_0 \sim N(0,1)$ and $U_A \sim N(\mu,1)$, with $\mu = \Phi^{-1}(1-\alpha) - \Phi^{-1}(\beta)$, which results in a distribution with shape $H_{\alpha,1-\beta}(p)$. Fay and Follmann (2002) have explained that this same distribution can be derived from the cases where $U_0 \sim \chi^2_1(0)$ and $U_A \sim \chi^2_1(\mu^2)$, where $\chi^2_1(\mu^2)$ is the random variable with non-central Chi-square distribution with 1 degree of freedom and non-centrality parameter equal to μ^2. They argued that, for the cases in which $U \sim F_{1,d}(\mu^2)$ the p-value distribution converges in distribution to $H_{\alpha,1-\beta}(p)$ when $d \to \infty$, where $F_{1,d}(\mu^2)$ is the random variable with F distribution with 1 and d degrees of freedom and non-centrality parameter equal to μ^2.

The class $\Im_B$ is smaller than $\Im$ and does not cover all cases of interest. For example, the spatial scan statistic developed by Kulldorff (2001) to detect spatial clusters follows very closely a Gumbel distribution under the null hypothesis and a chi-square distribution under H_A (see Abrams et al. (2010)). Therefore, even in interesting applied situations, there is not guarantee that $f_P(p) \in \Im_B$ and a larger class such as our $\Im$ may be useful.

It is worth mentioning that $h_{\alpha,1-\beta}(p)$ is a convex function when $1-\beta \geq \alpha$ and $p \leq 0.5$. Indeed, the second derivative of $h_{\alpha,1-\beta}(p)$ with respect to p is given by:

$$h''_{\alpha,1-\beta}(p) = \frac{\left[\Phi^{-1}(\beta) - \Phi^{-1}(1-\alpha) - \phi'(\Phi^{-1}(1-p))\right]}{\phi(\Phi^{-1}(1-p))} h'_{\alpha,1-\beta}(p) \tag{16}$$

and we have that

$$\phi'(\Phi^{-1}(1-p)) = \frac{\Phi^{-1}(1-p)}{\sqrt{2\pi}\phi(\Phi^{-1}(1-p))} \exp\left\{-1/2\left[\Phi^{-1}(1-p)\right]^2\right\} \geq 0$$

if $p \leq 0.5$.

Cases where the real situation of the data presents a small distance from H_0 are examples of applications in that the density of the p-value could escape from $\Im_B$. When the p-values tend to small values, in direction to α, that is the situation where the p-value density is deforming, from an uniform density, to an asymmetric

curve to the left hand, the convexity could not be verified for $p \leq 0.5$. For example, suppose $U_0 \sim \chi^2_1(0)$ e $U_A \sim \chi^2_{1,01}(0)$. The corresponding p-value density from this conjecture is not concave for $p > 0.32$.

The family $\Im$ for bounding the resampling risk is not restricted to families such as the normal, chi-square or F distributions. It also contains p-value densities with mixed shapes, with concave and convex parts. As an additional benefit, $\Im$ allows the bounding of the resampling risk in a very simple way.

In the next subsections, we analyze the power, the expected number of simulations and the resampling risk of our generalized Monte Carlo test procedure when the p-value distribution belongs to the class $\Im$. It is important to remember that, when using the MC_G, the class $\Im$ is not needed neither to calculate a bound for the power loss with respect to the MC_m or to the exact test nor to establish the bound for the expected number of simulations under H_A. Indeed, the results in the Sub-sections 3.1 and 3.2 are valid for any test statistic. However, when the additional assumption that the p-value density $f_P(p)$ belongs to $\Im$ holds, stronger results can be obtained.

4.1. Upper bound for the power difference between the exact test and MC_G

The power of the generalized Monte Carlo test is given by integrating out the probability $\pi_G(p)$ of rejecting the null hypothesis conditioned on the p-value p with respect to the p-value density:

$$\pi_G = \int_0^1 \pi_G(p) f_P(p) dp \, .$$

The power difference between the exact test and MC_G is given by:

$$\delta^*_G = \int_0^1 \left(1_{(0,\alpha_{mc}]}(p) - \pi_G(p)\right) f_P(p) dp. \tag{17}$$

An upper bound for δ^*_I can be obtained if we use $f_{P,w}(p) = 1/\alpha_{mc}$ if $p \in (0, \alpha_{mc}]$, and $f_{P,w}(p) = 0$, otherwise:

$$\begin{aligned} \delta^*_G \leq b^*_I &= \int_0^1 \left(1_{(0,\alpha_{mc}]}(p) - \pi_G(p)\right) f_{P,w}(p) dp = \int_0^{\alpha_{mc}} \frac{1}{\alpha_{mc}} dp - \int_0^{\alpha_{mc}} \pi_G(p) \frac{1}{\alpha_{mc}} dp \\ &= 1 - \frac{1}{\alpha_{mc}} \int_0^{\alpha_{mc}} \pi_G(p) dp. \end{aligned} \tag{18}$$

Because the function (5) is a sum of Beta(a, b) density kernels, the integral (18) can be rewritten as a function of incomplete Beta(a, b) functions, all of them evaluated at $p = \alpha_{mc}$, with a and b depending only of the parameters I, S e η. In the same way, an upper bound for the power difference between MC_m and MC_G is given by:

$$b^*_{m,G} = \int_0^{\alpha_{mc}} \left(\pi_m(p) - \pi_G(p)\right) \frac{1}{\alpha_{mc}} dp. \tag{19}$$

As before, (19) can also be expressed using incomplete beta functions.

4.2. An upper bound for the expected number of simulations

For values of p near 0, the simulation time is around n_1, the first checking point of the lower boundary. For values of p near 1, the simulation time is around $S_1^{'}$, the smallest height of the upper boundary. Numerically, we find that $\mathbb{E}(L|P=p)$ is maximized for p around α_{mc}. Let

$$p_{\max} = \arg\max_p \mathbb{E}(L|P=p)$$

and define $f_{P,max}(p) = 1/p_{max}$, for $p \in (0, p_{max}]$, and $\bar{f}_{P,max}(p) = 0$, otherwise. Thus, it follows that

$$\mathbb{E}(L) = \int_0^1 \mathbb{E}(L|P=p)\bar{f}_P(p)dp \leq \int_0^1 \mathbb{E}(L|P=p)\bar{f}_{P,max}(p)dp = \int_0^{1/p_{max}} \mathbb{E}(L|P=p)\frac{1}{p_{max}}dp. \tag{20}$$

The right hand side of the inequality (20) defines an upper bound $b^*_{E(L)}$ for $\mathbb{E}(L)$.

4.3. An upper bound for the resampling risk

Let RR be the resampling risk in a MC test defined as:

$$RR = \mathbb{P}_{mc}(H_0 \text{ is not rejected}|P \leq \alpha)\mathbb{P}(P \leq \alpha) + \mathbb{P}_{mc}(H_0 \text{ is rejected}|P \geq \alpha)\mathbb{P}(P \geq \alpha) \tag{21}$$

where $\mathbb{P}_{mc}$ is the probability measure associated with the events generated by MC simulations. For the MC_G test, denote its resampling risk by RR_G, which is computed as:

$$RR_G = \int_0^\alpha [1 - \pi_G(p)]\, f_P(p)dp + \int_\alpha^1 \pi_G(p) f_P(p)dp. \tag{22}$$

As $\pi_G(p)$ is a decreasing function, the function $\left[1_{p\in(0,\alpha]}(p) - \pi_G(p)\right]$ is maximum at $p = \alpha$. Thus, RR_G is maximum when $f_P(p)$ puts the largest possible mass at α, which is the worst case $f_{P,w}(p)$. Substituting $f_P(p)$ in (22) by $f_{P,w}(p)$ and setting $\alpha = \alpha_{mc}$, we have :

$$RR_G \leq 1 - \frac{1}{\alpha_{mc}} \int_0^{\alpha_{mc}} \pi_G(p)dp. \tag{23}$$

Therefore, an upper bound for RR_G is equal to the upper bound (18) for the power loss with respect to the exact test. That is, $b^*_{RR_G} = b^*_G$.

The expression (22) can be rewritten in a way that emphasizes another property. The situation where $\pi \geq \pi_G$ is that where the control of RR_G is important. If $\pi \geq \pi_G$, then $RR_G \geq \delta_I$, where δ_I is the power difference between the exact test and MC_G. Therefore, equal power of the exact test and the MC_G test does not imply a null resampling risk. To see this:

$$\begin{aligned} RR_G &= \int_0^{\alpha_{mc}} f_P(p)dp - \int_0^{\alpha_{mc}} \pi_G(p) f_P(p)dp + \int_{\alpha_{mc}}^1 \pi_G(p) f_P(p)dp \\ &= \pi - \pi_G + 2\int_{\alpha_{mc}}^1 \pi_G(p) f_P(p)dp = \delta_I + 2\int_{\alpha_{mc}}^1 \pi_G(p) f_P(p)dp. \end{aligned} \tag{24}$$

5. Choosing Parameters to Operate MC_G

This section aims to provide the reader with a useful set of choices for the parameters I, S and η to run the MC_G test. The choices we suggest produce a MC_G test with power equal to a MC_m test for any test statistic with small expected number of simulations.

Optimizing $\mathbb{E}(L)$ analytically is undoubtedly a complex task. In contrast, a numeric approach is feasible and simple to operate, and this is the approach adopted here. Define the class M, the set of MC_G procedures that, under H_0, leads to the same decision about rejecting H_0 than the MC_m. Conditioned on this class M, the three next steps were developed to estimate the parameters of the MC_G with minimum $\mathbb{E}(L)$. Let $MC_{I_{op}}$ be such scheme with minimum $\mathbb{E}(L)$.

1. This step is intended to emulate the X_t path under H_0. Generate N observations from an $U(0,1)$ distribution, and label them as p_i, $i = 1, ..., N$. For each p_i, generate m values x_{ij}, with $j = 1, \ldots, m$ following a Bernoulli distribution with success probability p_i. Define the partial sum processes

$$S_i = \left\{ S_{it}, \text{ such that } S_{it} = \sum_{l=1}^{t} x_{il}, t = 1, ..., m \right\} .$$

2. We build envelopes for the path X_t based on the simulated ones. For that, select those S_i sequences leading to the rejection of H_0 by MC_m. That is, to be selected the sequence S_i must satisfy $\max_t \{S_{it}\} < m\alpha_{mc}$. Suppose there are s of those sequences and they form the set $\mathcal{R}$. If N is large, we expect $s/N \approx \alpha_{mc}$. Define the sequence $\hat{S}_t = \{\max_i \{S_{it}\} + 1, i \in \mathcal{R}\}$. The curve $\hat{S}_t$ is an estimator for the upper boundary of $MC_{G_{op}}$.

   Next, take the r sequences S_i such that $\max_t \{S_{it}\} \geq m\alpha_{mc}$ and collect them in the set $\mathcal{A}$. These are the sequences $S_i's$ that do not reject H_0. Define the sequence $\hat{I}_t = \{\min_i \{S_{it}\}, i \in \mathcal{A}\}$. The curve $\hat{I}_t$ is the estimator for the lower boundary of $MC_{G_{op}}$.

3. Take the set $\hat{\eta^S}$ containing the jumping moments of $\hat{S}$. $\hat{\eta^S}$ is an estimator for η^S associated to $MC_{G_{op}}$. Take also the set $\hat{\eta^I}$ formed by the jumping moments of $\hat{I}$. $\hat{\eta^I}$ is an estimator of η^I associated to $MC_{G_{op}}$. Formally:

$$\hat{\eta^I} = \left\{ \hat{n}_t^I = \hat{n}_{t-1}^I \; if \; \left\lceil \hat{I}_t \right\rceil = \left\lceil \hat{I}_{t-1} \right\rceil, \text{ or } \hat{n}_t^I = t \text{ if } \left\lceil \hat{I}_t \right\rceil > \left\lceil \hat{I}_{t-1} \right\rceil \right\} \tag{25}$$

   with, $t = 2, ..., m$ and $\hat{n}_1^I = \min \left\{ l : \hat{I}_l > 0 \right\}, l = 1, ..., m$. Also,

$$\hat{\eta^S} = \left\{ \hat{n}_t^S = \hat{n}_{t-1}^S \; if \; \left\lfloor \hat{S}_t \right\rfloor = \left\lfloor \hat{S}_{t-1} \right\rfloor, \; or \; \hat{n}_t^S = t \; if \; \left\lfloor \hat{S}_t \right\rfloor > \left\lfloor \hat{S}_{t-1} \right\rfloor \right. \tag{26}$$

   with, $t = 2, ..., m$ and $\hat{n}_1^S = min \left\{ \hat{S}_t \right\}$. The estimation procedure ends here.

As an heuristic argument to show that these boundaries estimated using this algorithm are indeed estimates of the $MC_{G_{op}}$ boundaries, consider that, for N sufficiently large, $\hat{S}_t$ and $\hat{I}_t$ are constructed to ensure

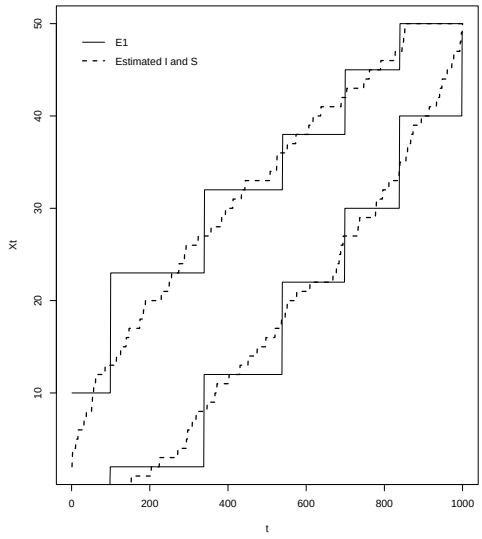

Figure 3: Scheme E_1 from Table 5 versus estimates for I and S considering $m = 1000$, $N = 100000$ and $\alpha_{mc} = 0.05$

that, under H_0, the decision of the MC_G will be always the same as that reached with MC_m. In addition, if $f_P(p) \in \Im$, we have

$$\mathbb{P}(X_t \text{ reach } I \text{ before reach } S \mid H_0) \leq \mathbb{P}(X_t \text{ reach } I \text{ before reach } S \mid H_A),$$

and then the power of these estimated boundaries is, at least, equal to that from MC_m. Concerning the expected number of simulations, $\mathbb{E}(L)$ decreases by increasing the I elements and by decreasing the S elements. The construction of $\hat{S}_t$ and $\hat{I}_t$ follows this logic. The reasoning is to scan each t taking the maximum value for I which does not restrict the X_t trajectories which would not reject H_0 by using MC_m. Simultaneously, it takes the minimum value for S which does not restrict the X_t trajectories which would reject H_0 by using MC_m.

The resulting estimators of η^I and η^S are quite sparse using this algorithm, while they could be computationally costly if calculated by means of the expressions developed in Section 3. An alternative and satisfactory way to construct η^I and η^S is based on the identification of the moments with high incidence of impact of X_t with the estimated boundaries. For that, let $n_q^* = \min\left\{t \in [1, ..., m], t : S_{it} \leq \hat{I}, i \in \mathcal{R}\right\}$ be an element of the sequence formed by the impact moments of each sequence S_{it} with $\hat{I}$ (considering only those sequences $i \in \mathcal{R}$). The most frequent impact moments of these sequences S_{it} with $\hat{I}$ are appropriate candidates for composing η^I. Apply the same reasoning for constructing η^S, and denote the correspondent sequence by n_r^*. Thus, as an alternative way to construct η^I and η^S, by arbitrary and conveniently low values k_1 and k_2, choose the most frequent elements in n_q^* and n_r^* to compose η^I and η^S, respectively. Extensive simulated examples indicate that, for $k_1 = k_2 \geq 5$, the exact computation of power loss and expected number

of simulations have a small computational cost, and the results are close to that using $\hat{\eta^I}$ and $\hat{\eta^S}$.

Figure 3 shows the estimates $\hat{I}$ and $\hat{S}$ obtained according to the steps 1, 2, and 3, of our algorithm, using $N = 100000$, $m = 1000$ and $\alpha_{mc} = 0.05$. The estimated boundaries are not parallel, but they are characterized by a funnel in the extremities. This behavior was verified in all of the simulations performed by us. For these specific estimates, if we take $\eta = \eta^S = \eta^I$, we obtain the times $\hat{\eta} = \{99, 339, 539, 699, 839, 999\}$. From the estimates plotted in Figure 3, we obtain $\hat{I} = \{2, 12, 22, 30, 40, 49\}$, and $\hat{S} = \{10, 23, 32, 38, 45, 50\}$. This specific scheme for our generalized sequential MC test is available in Table 5 and it is labeled as E_1. As we can see in the Table 2, column $b_{m,I}$, this scheme is efficient, presenting practically the same power than MC_m for $m = 1000$, with size equal to 0.049864. From Table 3, columns $\mathbb{E}(L|H_0)$ and $b_{E(L)}$, we see that this scheme have a small expected number of simulations, equal to 58.606 under H_0, and with an upper bound under H_A for any statistic, that is approximately 65% of the maximum 999. By using the class $\Im_B$ and the larger class $\Im$ for the p-value distribution, the bounds are expressively low, equal to 172.612 and 246.354, respectively. We consider that this scheme is a good option to replace MC_m. We must enphasize that, although the boundaries presented in Table 5 were guided by the algorithm above, all results in tables 2 and 3 are exact, because they were obtained by applying the expressions from Sections 3 and 4. Such algorithm is useful to construct preliminary choices of boundaries. The validation of an arbitrary design to practical use must be based on such exact calculations.

We provide other interesting schemes in Table 5. For each scheme, Table 2 offers the type I error probability, the upper bound for the power loss comparatively to MC_m and to the exact test, and the upper bound for the resampling risk. Table 3 gives the expected number of simulations under H_0 and the upper bounds under H_A. We adopt $b_{m,G}$ to denote the general upper bound for the power loss comparatively to MC_m, b^*_G and $b^*_{RR_G}$, the upper bounds for the power loss, with respect to the exact test, and for the resampling risk, respectively, where the super index $*$ indicates that the calculations are restricted to the p-value distribution on the class $\Im$. The same symbol was used to indicate the use of this class for the bounds in Table 3. Upper bounds using the class $\Im_B$ are also available in Tables 2 and 3, and they are indicated by a tilde accent. Concerning the use of $\Im_B$, the numerical explorations of Fay and Follmann (2002) were not used here to define the worst case of a p-value distribution with shape $H_{\alpha;1-\beta}(p)$. As discussed in Section 4.3, $h_{\alpha,1-\beta}(p)$ (for $1 - \beta \geq \alpha$) and π_G are decreasing with p. Therefore, the worst case within the class $\Im_B$, in the sense of bounding RR_I, occurs at the point of maximum of the function $H_{\alpha;1-\beta}(\alpha)$ with respect to β. For $1 - \beta \geq \alpha$, the point of maximum in β for the function $H_{\alpha;1-\beta}(\alpha)$ is 0.5. Then, the analytical worst case is given by $H_{\alpha;0.5}(\alpha)$. We used this result to compute $\tilde{b}_I$, $\tilde{b}_{RR_G}$ and $\tilde{b}_{E(L)}$ here.

6. MC_G versus MC_B

In this Section we offer a comparison between the MC_B and MC_G sequential test procedures. We use an example of the MC_B test given by Kim (2010). In this comparison, we focus on resampling risk bound and on the expected number of simulations. We assume that the p-value distribution $f_P(p)$ belongs to the class $\Im_B$. We did not consider other important characteristics of a test, such as the power loss with respect to the exact test and expected number of simulations for an arbitrary $f_P(p)$, because they were not treated by Kim (2010). We built our MC_G boundaries using the algorithm from Section 5. After securing an upper bound for the resampling risk to MC_G equal to that presented by the MC_B scheme developed in Kim (2010), we compared the average simulation time of the two procedures.

An obvious fact is that the B boundaries are particular cases of I and S, because MC_G was designed to be a generalized sequential with two stopping boundaries. We can rewrite the B boundaries using the MC_G notation, based on the sets I, S, η^I and η^S. In this way, for an MC_B test, the user can apply the general expressions for the power and the expected number of simulations developed in Section 3.

Define:

$$T_1^* = \left\{ t > 2 : \left\lceil B_{\text{Inf}}(t-1) \right\rceil < \left\lceil B_{\text{Inf}}(t) \right\rceil \right\}$$

and

$$T_2^* = \left\{ t > 2 : \left\lfloor B_{\text{Sup}}(t-1) \right\rfloor < \left\lfloor B_{\text{Sup}}(t) \right\rfloor \right\} .$$

Let $t_{11}^* < t_{12}^* < ... < t_{1k_1}^*$ be the ordered elements of T_1^*, and $t_{21}^* < t_{22}^* < ... < t_{2k_2}^*$ be the ordered elements of T_2^*. Rewritten in terms of I and S, the B boundaries are denoted by I^*, S^*, η^{I*} and η^{S*}, and they are built as follows:

$$I^* = \left\{ \left\lceil B_{Inf}(t_{11}^*) \right\rceil, \left\lceil B_{Inf}(t_{12}^*) \right\rceil, ..., \left\lceil B_{Inf}(t_{1k_1}^*) \right\rceil \right\}$$

$$S^* = \left\{ \left\lfloor B_{Inf}(t_{21}^*) \right\rfloor, \left\lfloor B_{Inf}(t_{22}^*) \right\rfloor, ..., \left\lfloor B_{Inf}(t_{2k_2}^*) \right\rfloor \right\}$$

$$\eta^{I*} = \left\{ t_{11}^*, t_{12}^*, ..., t_{1k_1}^* \right\}$$

$$\eta^{S*} = \left\{ t_{21}^*, t_{22}^*, ..., t_{2k_2}^* \right\}.$$

It should be noted that some important shapes for I and S, as the funnel behavior estimated in Section 5, can not be represented by the MC_B boundaries.

Table 4 shows the upper bounds for the resampling risk and for the expected number of simulations presented in Kim (2010) for $n = 600$, $\alpha = 0.05$ and $c_1 = -c_2 = 1.282$, as well the bounds associated to MC_G

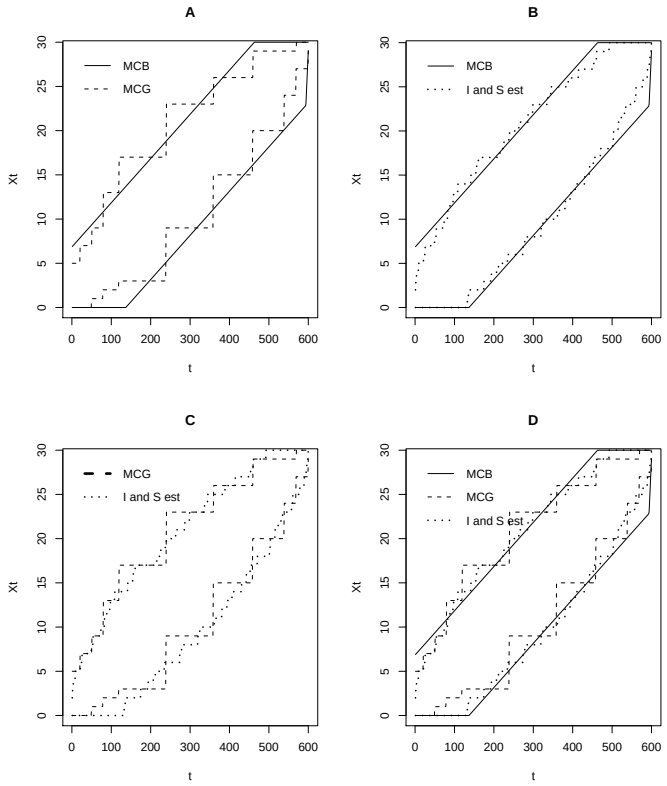

Figure 4: Boundaries for MC_G and MC_B sequential test procedures.

using the scheme E_{12} detailed in Table 5. Concerning the worst case of distribution within the class $\Im_B$, Kim (2010) adopted the numerical studies from Fay and Follmann (2002) and she found $\tilde{F}^* = H_{0.05;0.47}(p)$ for $\alpha = 0.05$, with approximation $\tilde{H}_{0.05;0.47}(p) := Beta(0.389; 2.523)$.

These bounds are also computed here for the sequential procedure proposed by Besag and Clifford (1991), which will be denoted by MC_h. This MC_h procedure is a very simple way to perform sequential tests, because it is based just in an upper boundary fixed in a value denoted by h and truncated in a maximum number of simulations n. Silva et al. (2009) had showed that MC_h has the same power than MC_m if $h = \alpha_{mc}m$ and its power is constant for $n \geq h/\alpha_{mc} + 1$, noting that the combination of this two last rules implies that MC_m must be replaced by MC_h, because they have the same power for a maximum number of simulations practically equal (that is, for $n = m + 1$).

Under either hypothesis, MC_G is substantially faster than the MC_B procedure with their expected time ratio being around 60%. This illustrates the gain provided by our MC_G algorithm. The estimated boundaries for $m = 600$ and $\alpha_{mc} = 0.05$ for the MC_G procedure are available in Figure 4, where we can also see the B boundaries and the boundaries I and S (built according to the scheme E_{12}). It is clear the greater flexibility

given by the MC_G boundaries. While the B boundaries are parallel almost up to the end of the experiment, the I and S boundaries are tapered when t gets close to the maximum number of simulations. This can be intuitively thought as if the boundaries were using the information that X_t not touching the boundaries after a long time was inducing a narrower vigilance.

7. Discussion

The generalized sequential Monte Carlo test has properties that recommend it in substitution to the conventional Monte Carlo test for any test statistic. In this paper, we gave simple expressions for the calculation of the test size, the expected number of simulations under the null hypothesis, and upper bounds for the expected number of simulations under the alternative hypothesis and for the power loss with respect to the fixed length conventional Monte Carlo test.

Exact calculations for some specific design indicates that our generalized sequential test has a substantially smaller execution time than other sequential methods proposed in the literature. Under a wide class of distributions for the p-value statistic, the bounds for the execution time under the alternative hypothesis is even more substantial. Under this class, the generalized sequential test has power virtually identical to the exact test, even for such intermediate maximum number of simulations as 4999. The use of this class allows the construction of optimal boundaries from a simple algorithm and they have a surprising funnel-type shape. These optimal boundaries, or any other generalized design, for any test statistic, can be evaluated by calculating the size, expected number of simulations under H_0, upper bound for the power loss and for the expected number of simulations under H_A, by using the expressions developed in section 3.

t	B_{Sup}	B_{Inf}	t	B_{Sup}	B_{Inf}
⋮	⋮	⋮	⋮	⋮	⋮
134	13.54	0.00	157	14.69	1.01
135	13.59	0.00	158	14.74	1.06
136	13.64	0.00	159	14.79	1.11
137	13.69	0.01	160	14.84	1.16
138	13.74	0.06	161	14.89	1.21
139	13.79	0.11	162	14.94	1.26
140	13.84	0.16	163	14.99	1.31
141	13.89	0.21	164	15.04	1.36
142	13.94	0.26	165	15.09	1.41
143	13.99	0.31	166	15.14	1.46
144	14.04	0.36	167	15.19	1.51
145	14.09	0.41	168	15.24	1.56
146	14.14	0.46	169	15.29	1.61
147	14.19	0.51	170	15.34	1.66
148	14.24	0.56	171	15.39	1.71
149	14.29	0.61	172	15.44	1.76
150	14.34	0.66	173	15.49	1.81
151	14.39	0.71	174	15.54	1.86
152	14.44	0.76	175	15.59	1.91
153	14.49	0.81	176	15.64	1.96
154	14.54	0.86	177	15.69	2.01
155	14.59	0.91	178	15.74	2.06
156	14.64	0.96	179	15.79	2.11
⋮	⋮	⋮	⋮	⋮	⋮

Table 1: Boundary values $B_Sup(t)$ and $B_{Inf}(t)$ for $134 \leq t \leq 179$.

α	n_k	Scheme	$\mathbb{P}$(erro tipo I)	$b_{m,I}$	$\tilde{b}_I$	$b_I^* = b_{RR_G}^*$	$\tilde{b}_{RR_G}$
		E_1	0.049864	0.031032	0.000000	0.060000	0.023000
		E_2	0.049681	0.031748	0.000000	0.060895	0.023337
0.05	999	E_3	0.049920	0.027828	0.000000	0.060118	0.022348
		E_4	0.049982	0.024751	0.000000	0.053977	0.022826
		E_5	0.049998	0.020890	0.000218	0.058518	0.021326
0.01	999	E_6	0.009999	0.030136	0.002921	0.136997	0.028048
		E_7	0.009993	0.000000	0.003344	0.140063	0.028443
0.05	4999	E_8	0.050401	0.000000	0.014908	0.021841	0.016746
0,01		E_9	0.010003	0.027643	0.000577	0.061952	0.012354
0.05	9999	E_{10}	0.050036	0.008435	0.000000	0.017947	0.001838
0.01		E_{11}	0.009992	0.023415	0.000106	0.043559	0.008821

Table 2: Effective test size, upper bound for the power loss, and for the resampling risk associated to the schemes in the Table 5.

α	n_k	Scheme	$\mathbb{E}(L\|H_0)$	$b_{E(L)}$	$\tilde{b}_{E(L)}$	$b_{E(L)}^*$
		E_1	58.606	644.654	172.612	246.354
		E_2	96.739	807.572	235.119	402.704
0.05	999	E_3	113.597	792.104	264.587	407.943
		E_4	108.472	670.761	276.496	396.895
		E_5	128.276	698.242	335.969	474.473
0.01	999	E_6	42.717	677.409	225.362	550.574
		E_7	41.728	625.841	225.334	523.601
0.05	4999	E_8	384.591	4283.575	1016.884	1716.704
0.01		E_9	121.992	3580.891	1816.592	687.744
0.05	9999	E_{10}	731.131	9477.250	1951.502	3502.344
0.01		E_{11}	190.355	8706.908	1182.462	3968.661

Table 3: Expected Number of Simulations for Schemes from Table 5.

	MC_G E_{12}	MC_B with $n = 600$ $c_1 = -c_2 = 1.282$	IPO with $n = 576$	MC_h with $h = 30$ and $m = 600$
P(erro tipo I)	0.050000	0.050000	0.050000	0.050000
$\tilde{b}$	0.001888	≤ 0.025000	≤ 0.025000	0.001888
$\tilde{b}_{RR}$	0.024562	0.025000	0.025000	0.024562
$\mathbb{E}(L\|H_0)$	33.720	51.169	62.850	119.351
$\tilde{b}_{E(L)}$	91.877	163.118	213.508	390.507

Table 4: Upper bounds for the resampling risk and expected number of simulations for comparison among MC_G using E_{12}, MC_B, with $n = 600$, $\alpha = 0.05$ and $c_1 = -c_2 = 1.282$, and IPO.

E_1	$I = \{2, 12, 22, 30, 40, 49\}$ $S = \{10, 23, 32, 38, 45, 50\}$ $\eta = \{99, 339, 539, 699, 839, 999\}$
E_2	$I = \{2, 12, 20, 35, 49\}$ $S = \{19, 30, 37, 45, 49\}$ $\eta = \{99, 379, 539, 779, 999\}$
E_3	$I = \{7, 16, 24, 35, 49\}$ $S = \{28, 35, 41, 49, 49\}$ $\eta = \{199, 499, 699, 899, 999\}$
E_4	$I = \{13, 22, 30, 35, 49\}$ $S = \{28, 35, 42, 45, 49\}$ $\eta = \{299, 559, 719, 799, 999\}$
E_5	$I = \{17, 26, 35, 42, 49\}$ $S = \{34, 39, 43, 46, 49\}$ $\eta = \{399, 639, 799, 899, 999\}$
E_6	$I = \{1, 5, 7, 8, 9\}$ $S = \{8, 9, 9, 9, 9\}$ $\eta = \{299, 599, 799, 899, 999\}$
E_7	$I = \{2, 5, 7, 8, 9\}$ $S = \{8, 8, 8, 9, 9\}$ $\eta = \{399, 599, 769, 899, 999\}$
E_8	$I = \{27, 100, 199, 249\}$ $S = \{80, 150, 219, 249\}$ $\eta = \{799, 2499, 3999, 4999\}$
E_9	$I = \{5, 20, 35, 49\}$ $S = \{20, 31, 43, 49\}$ $\eta = \{799, 2499, 3999, 4999\}$
E_{10}	$I = \{17, 79, 499\}$ $S = \{79, 249, 499\}$ $\eta = \{499, 2999, 9999\}$
E_{11}	$I = \{79, 199, 499\}$ $S = \{199, 499, 499\}$ $\eta = \{2399, 4999, 9999\}$
E_{12}	$I = \{0, 1, 2, 3, 9, 15, 20, 24, 27, 29\}$ $S = \{5, 7, 9, 13, 17, 23, 26, 29, 29, 30\}$ $\eta = \{20, 50, 79, 119, 239, 359, 459, 539, 569, 600\}$

Table 5: Appropriate Schemes for Replacing MC_m by MC_G.

References

Abrams, A., Kleinman, K., Kulldorff, M., 2010. Gumbel based p-value approximations for spatial scan statistics. International Journal of Health Geographics 9 (61).

Besag, J., Clifford, P., 1991. Sequential monte carlo p-value. Biometrika 78, 301–304.

Fay, M., Follmann, D., 2002. Designing monte carlo implementations of permutation or bootstrap hypothesis tests. The American Statistician 56 (1), 63–70.

Fay, M., Kim, H.-J., Hachey, M., 2007. On using truncated sequential probability ratio test boundaries for monte carlo implementation of hypothesis tests. Journal of Computational and Graphical Statistics 16, 946–967.

Gandy, A., 2009. Sequential implementation of monte carlo tests with uniformly bounded resampling risk. Journal of the American Statistical Association 104 (488), 1504–1511.

Kim, H.-J., 2010. Bounding the resampling risk for sequential monte carlo implementation of hypothesis tests. Journal of Statistical Planning and Inference 140, 1834–1843.

Kulldorff, M., 2001. Prospective time periodic geographical disease surveillance using a scan statistic. Journal of Royal Statistical Society 164A, 61–72.

Lan, K., Wittes, J., 1988. The b-value: a tool for monitoring data. Biometrics 44, 579–585.

Silva, I., Assunção, R., 2011. Monte carlo test under general conditions: Power and number of simulations. Paper submitted to Journal of Statistical Planning and Inference.

Silva, I., Assunção, R., Costa, M., 2009. Power of the sequential monte carlo test. Sequential Analysis 28 (2), 163–174.

Acknowledgements

We are grateful to Martin Kulldorff for very useful comments and suggestions in the manuscript.