Efficient simulation methods for scan statistics: a comparison study.

Alexandru Amărioarei

Laboratoire de Mathématiques Paul Painlevé
Département de Probabilités et Statistique
Université de Lille 1, INRIA Modal Team

The 17^{th} Conference of the Romanian Society of Statistics and Probability
25-26 April, 2014, București, România

Outline

Outline

The d-dimensional discrete scan statistics

Let $T_1, T_2, \ldots, T_d$ be positive integers, with $d \geq 1$

- The rectangular region, $\mathcal{R}_d = [0, T_1] \times [0, T_2] \times \cdots \times [0, T_d]$
- The r.v.'s $X_{s_1,s_2,\ldots,s_d}$, $1 \leq s_j \leq T_j$, $j \in \{1, 2, \ldots, d\}$

Let $2 \leq m_j \leq T_j$, $1 \leq j \leq d$, be positive integers

- Define for $1 \leq i_l \leq T_l - m_l + 1$, $1 \leq l \leq d$,

$$Y_{i_1,i_2,\ldots,i_d} = \sum_{s_1=i_1}^{i_1+m_1-1} \sum_{s_2=i_2}^{i_2+m_2-1} \cdots \sum_{s_d=i_d}^{i_d+m_d-1} X_{s_1,s_2,\ldots,s_d}$$

- The d-dimensional discrete scan statistic,

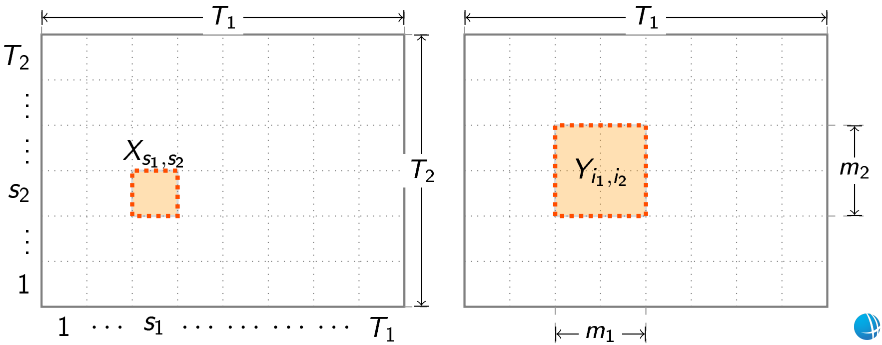
$$S_{\mathbf{m}}(\mathbf{T}) = \max_{\substack{1 \leq i_j \leq T_j - m_j + 1 \\ j \in \{1,2,\ldots,d\}}} Y_{i_1,i_2,\ldots,i_d}$$

with $\mathbf{m} = (m_1, m_2, \ldots, m_d)$ and $\mathbf{T} = (T_1, T_2, \ldots, T_d)$

Example: two dimensional scan statistics ($d = 2$)

We have for $d = 2$

$$Y_{i_1,i_2} = \sum_{s_1=i_1}^{i_1+m_1-1} \sum_{s_2=i_2}^{i_2+m_2-1} X_{s_1,s_2}, \quad S_{m_1,m_2}(T_1, T_2) = \max_{\substack{1\leq i_1\leq T_1-m_1+1 \\ 1\leq i_2\leq T_2-m_2+1}} Y_{i_1,i_2}$$



Outline

1. Introduction
   - Framework
   - Problem
2. Simulation Methods
   - Classical Monte Carlo Approach (MC)
   - Importance Sampling Approach (IS)
   - Computational Aspects
3. Comparison Study
   - One Dimensional Setting
   - Other Simulation Methods
   - Numerical Examples
4. References

Problem

The distribution of $S_{\mathbf{m}}(\mathbf{T})$ is used for testing the null hypotheses of randomness against the alternative hypothesis of clustering.

Example: Bernoulli model

H_0: The r.v.'s $X_{s_1,s_2,\ldots,s_d}$ are i.i.d. $\mathcal{B}(p)$

H_1: There exists
$\mathcal{R}(i_1,i_2,\ldots,i_d) = [i_1-1,i_1+m_1-1]\times\cdots\times[i_d-1,i_d+m_d-1]\subset\mathcal{R}_d$
where the r.v.'s $X_{s_1,s_2,\ldots,s_d}\sim\mathcal{B}(p')$, $p'>p$ and $X_{s_1,s_2,\ldots,s_d}\sim\mathcal{B}(p)$ outside $\mathcal{R}(i_1,i_2,\ldots,i_d)$

Goal

Find a good estimate for the distribution of d-dimensional discrete scan statistic

$$Q_{\mathbf{m}}(\mathbf{T}) = \mathbb{P}\left(S_{\mathbf{m}}(\mathbf{T})\leq\tau\right)$$

Outline

Naive Hit-or-Miss MC

Fix a threshold value τ.

For each $1 \leq k \leq ITER$ (iterations number)

- Generate $\mathbf{X}^{(\mathbf{i})} = \left\{X_{s_1,s_2,\dots,s_d}^{(i)}, 1 \leq s_j \leq T_j, 1 \leq j \leq d\right\}$ under H_0
- Compute the d-dimensional scan statistics $S_{\mathbf{m}}^{(i)}(\mathbf{T})$

Return

$$\widehat{p_{MC}} = \frac{1}{ITER}\sum_{i=1}^{ITER}\mathbf{1}_{\left\{S_{\mathbf{m}}^{(i)}(\mathbf{T})\geq\tau\right\}}, \quad \widehat{s.e._{MC}} = \sqrt{\frac{\widehat{p_{MC}}(1-\widehat{p_{MC}})}{ITER}}$$

the unbiased direct Monte Carlo estimate of $p = \mathbb{P}\left(S_{\mathbf{m}}(\mathbf{T}) \geq \tau\right)$ and its consistent standard error estimate.

- computationally intensive since just a fraction of the generated observations will cause a rejection
- needs a large number of replications in order to reduce the standard error estimate to an acceptable level (especially for $d \geq 2$)

Outline

Generalities on IS

Variance reduction technique employed especially when dealing with rare events ([Fishman, 1996], [Rubino and Tuffin, 2009]).

Problem

Let W be a random vector with joint density f. Estimate the expectation

$$\theta = \mathbb{E}_f\left[G(W)\right] = \int G(\mathbf{x})f(\mathbf{x})d\mathbf{x}$$

Possible solution

Introduce another probability density g such that Gf is dominated by g and use

$$\theta = \int \left[\frac{G(\mathbf{x})f(\mathbf{x})}{g(\mathbf{x})}\right] g(\mathbf{x})d\mathbf{x} = \mathbb{E}_g\left[\frac{G(W)f(W)}{g(W)}\right]$$

Finding a suitable change of measure g is a difficult problem ([Rubino and Tuffin, 2009]).

IS for scan statistics: $d = 2$

The method was used for solving the problem of:

- union count ([Frigessi and Vercellis, 1984], [Fishman, 1996])
- exceeding probabilities ([Naiman and Wynn, 1997])
- scan statistics ([Naiman and Priebe, 2001], [Malley et al., 2002])

We are interested in evaluating the probability

$$\mathbb{P}_{H_0}\left(S_{\mathbf{m}}(\mathbf{T}) \geq \tau\right) = \mathbb{P}\left(\bigcup_{i_1=1}^{T_1-m_1+1} \bigcup_{i_2=1}^{T_2-m_2+1} E_{i_1,i_2}\right) = \int G(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}$$

where $E_{i_1,i_2} = \{Y_{i_1,i_2} \geq \tau\}$, $G(\mathbf{x}) = \mathbf{1}_E(\mathbf{x})$, $E = \bigcup_{i_1=1}^{T_1-m_1+1} \bigcup_{i_2=1}^{T_2-m_2+1} E_{i_1,i_2}$ and f is the joint density of Y_{i_1,i_2} under H_0.

IS for scan statistics: $d = 2$

We introduce the change of measure

$$g(\mathbf{x}) = \sum_{j_1=1}^{T_1-m_1+1} \sum_{j_2=1}^{T_2-m_2+1} \left\{ \frac{\mathbb{P}\left(E_{j_1,j_2}\right)}{B(2)} \right\} \left\{ \frac{\mathbf{1}_{E_{j_1,j_2}} f(\mathbf{x})}{\mathbb{P}\left(E_{j_1,j_2}\right)} \right\}$$

and we observe that $\mathbb{P}_{H_0}\left(S_{\mathbf{m}}(\mathbf{T}) \geq \tau\right) = B(2)\rho(2)$

- the Bonferroni upper bound $B(2)$ and the correction factor $\rho(2)$

$$B(2) = \sum_{i_1=1}^{T_1-m_1+1} \sum_{i_2=1}^{T_2-m_2+1} \mathbb{P}\left(E_{i_1,i_2}\right), \quad \rho(2) = \sum_{j_1=1}^{T_1-m_1+1} \sum_{j_2=1}^{T_2-m_2+1} p_{j_1,j_2} \int \frac{1}{C(\mathbf{Y})} d\mathbb{P}_{H_0}(\cdot|E_{j_1,j_2})$$

where

$$p_{j_1,j_2} = \frac{1}{(T_1-m_1+1)(T_2-m_2+1)}, \quad C(\mathbf{Y}) = \sum_{i_1=1}^{T_1-m_1+1} \sum_{i_2=1}^{T_2-m_2+1} \mathbf{1}_{E_{i_1,i_2}}$$

IS for scan statistics: $d = 2$ – Algorithm

Algorithm 1 Importance Sampling Algorithm for Scan Statistics

Begin
Repeat for each k from 1 to $ITER$ (iterations number)

1: Generate uniformly the point $(i_1^{(k)}, i_2^{(k)})$ from the set $\{1, \ldots, T_1 - m_1 + 1\} \times \{1, \ldots, T_2 - m_2 + 1\}$.

2: Given the point $(i_1^{(k)}, i_2^{(k)})$, generate a sample of the random field $\tilde{\mathbf{X}}^{(k)} = \left\{\tilde{X}_{s_1,s_2}^{(k)}\right\}$, with $s_j \in \{1, \ldots, T_j\}$ and $j \in \{1, 2\}$, from the conditional distribution of $\mathbf{X}$ given $\left\{Y_{i_1^{(k)}, i_2^{(k)}} \geq \tau\right\}$.

3: Take $c_k = C(\tilde{\mathbf{X}}^{(k)})$ the number of all couples (i_1, i_2) for which $\tilde{Y}_{i_1,i_2} \geq \tau$ and put $\widehat{\rho}_k(2) = \frac{1}{c_k}$.

End Repeat
Return

$$\widehat{\rho}(2) = \frac{1}{ITER}\sum_{k=1}^{ITER}\widehat{\rho}_k(2), \quad Var\left[\widehat{\rho}(2)\right] \approx \frac{1}{ITER-1}\sum_{k=1}^{ITER}\left(\widehat{\rho}_k(2) - \frac{1}{ITER}\sum_{k=1}^{ITER}\widehat{\rho}_k(2)\right)^2$$

End

Example

We evaluate the simulation error corresponding to $\mathbb{P}\left(S_{5,5,5}(60,60,60) \leq 2\right)$ in the Bernoulli model with $p = 0.0001$.

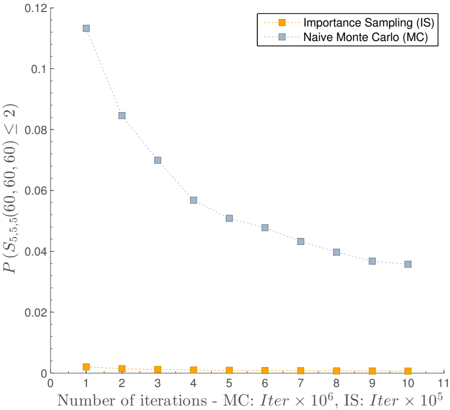


Figure 1 : The evolution of simulation error in MC and IS methods

Outline

Implementation issues

Algorithm 1 presents two main difficulties:

a) being able to sample from the conditional distribution of $\mathbf{X}$ given $\left\{Y_{i_1^{(k)}, i_2^{(k)}} \geq \tau\right\}$ in **Step** 2

b) the number of locality statistics that exceed the predetermined threshold is supposed to be found in a *reasonable* time

Partial solutions were found for:

a) binomial, Poisson and Gaussian model

b) *cumulative counts* or *fast spatial scan* techniques (see [Neil, 2006], [Neil, 2012])

Outline

Discrete scan statistics for normal data

Consider $d = 1$ and let $2 \leq m_1 \leq T_1$, m_1 and T_1 be positive integers

- $X_{s_1} \sim \mathcal{N}(\mu, \sigma^2)$ are i.i.d., $1 \leq s_1 \leq T_1$

The variables $Y_{i_1} = \sum_{s_1=i_1}^{i_1+m_1-1} X_{s_1}$ follow a multivariate normal distribution with mean $\bar{\mu} = m_1\mu$ and covariance matrix $\Sigma = (\Sigma_{i_1,j_1})$

$$\Sigma_{i_1,j_1} = Cov\left[Y_{i_1}, Y_{j_1}\right] = \begin{cases} (m_1 - |i_1 - j_1|)\,\sigma^2 & , |i_1 - j_1| < m_1 \\ 0 & , \text{otherwise.} \end{cases}$$

Step 2 in Algorithm 1

Step 2 requires to sample:

- $Y_{i_1^{(k)}}$ from the tail distribution $\mathbb{P}\left(Y_{i_1^{(k)}} \geq \tau\right)$ ([Devroye, 1986])
- for the other indices, from the conditional distribution given $\left\{Y_{i_1^{(k)}} \geq \tau\right\}$

For $\mathbf{W}_1 = \left(Y_1, \dots, Y_{i_1^{(k)}-1}\right)$ and $\mathbf{W}_2 = \left(Y_{i_1^{(k)}+1}, \dots, Y_{T_1-m_1+1}\right)$

$$\overline{\mathbf{W}}_1 = \mathbf{W}_1 | (Y_{i_1^{(k)}} = t) \sim \mathcal{N}\left(\mu_{w_1|t}, \Sigma_{w_1|t}\right) \text{ and } \overline{\mathbf{W}}_2 = \mathbf{W}_2 | (Y_{i_1^{(k)}} = t) \sim \mathcal{N}\left(\mu_{w_2|t}, \Sigma_{w_2|t}\right)$$

where for $i \in \{1,2\}$,

$$\mu_{w_i|t} = \mathbb{E}[\mathbf{W}_i] + \frac{1}{Var[Y_{i_1^{(k)}}]} Cov[\mathbf{W}_i, Y_{i_1^{(k)}}](t - \mathbb{E}[Y_{i_1^{(k)}}]),$$

$$\Sigma_{w_i|t} = Cov(\mathbf{W}_i) - \frac{1}{Var[Y_{i_1^{(k)}}]} Cov[\mathbf{W}_i, Y_{i_1^{(k)}}] Cov^T[\mathbf{W}_i, Y_{i_1^{(k)}}].$$

Outline

Alternative approaches

Several other methods were proposed:

i) [Genz and Bretz, 2009] developed a quasi Monte Carlo algorithm for numerically approximate the distribution of a multivariate normal, the algorithm was implemented in R and Matlab ([Wang and Glaz, 2013])

ii) [Shi et al., 2007] introduced another IS algorithm (Algo 2)
   - idea: imbed the probability measure under H_0 into an exponential family

▸ Details Algo 2

To measure the efficiency of the methods we evaluate the *relative efficiency* introduced by [Malley et al., 2002]

$$Rel\ Eff = \frac{\sigma^2_{method\ 1} \times CPU\ \ Time_{method\ 1}}{\sigma^2_{method\ 2} \times CPU\ \ Time_{method\ 2}}$$

Outline

Numerical results

All the results are compared with respect to Algo 1 for $ITER = 10000$

Table 1 : Algorithm [Genz and Bretz, 2009], IS (Algo 1) and the relative efficiency (Rel Eff)

T_1	m_1	τ	Genz	Err Genz	IS Algo 1	Err Algo 1	Rel Eff
200	15	12	0.932483	0.000732	0.933215	0.000743	7
500	25	18	0.976117	0.000460	0.975797	0.000425	518
750	30	24	0.998454	0.000125	0.998493	0.000024	688
800	40	30	0.999752	0.000029	0.999742	0.000004	617

Table 2 : Naive Monte Carlo (MC), IS (Algo 1) and the relative efficiency (Rel Eff)

T_1	m_1	τ	MC	Err MC	IS Algo 1	Err Algo 1	Rel Eff
200	15	12	0.932624	0.000694	0.933215	0.000743	15
500	25	18	0.975880	0.000425	0.975797	0.000425	33
750	30	24	0.998515	0.000061	0.998493	0.000024	101
800	40	30	0.999741	0.000009	0.999742	0.000004	602

Numerical results

Table 3 : IS algorithms (Algo 2 and Algo 1) and the relative efficiency (Rel Eff)

T_1	m_1	τ	IS Algo 2	Err Algo 2	IS Algo 1	Err Algo 1	Rel Eff
200	15	12	0.932744	0.000839	0.933215	0.000743	3
500	25	18	0.976105	0.000448	0.975797	0.000425	3.5
750	30	24	0.998508	0.000032	0.998493	0.000024	3.5
800	40	30	0.999740	0.000006	0.999742	0.000004	3.6

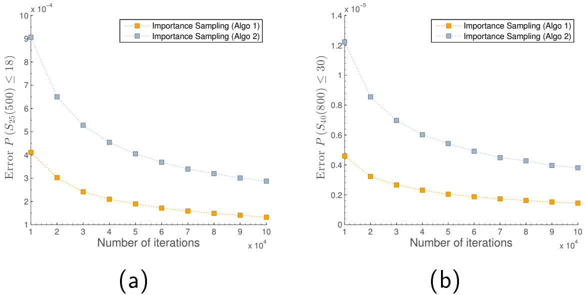


Figure 2 : The evolution of simulation error in IS Algorithm 1 and IS Algorithm 2

Devroye, L. (1986).
Non uniform random variate generation.
Springer-Verlag, New York.

Fishman, G. (1996).
Monte Carlo: Concepts, Algorithms and Applications.
Springer Series in Operations Research. Springer-Verlag, New York.

Frigessi, A. and Vercellis, C. (1984).
An analysis of Monte Carlo algorithms for counting problems.
Department of Mathematics, University of Milan.

Genz, A. and Bretz, F. (2009).
Computation of Multivariate Normal and T Probabilities.
Springer-Verlag, New York.

Malley, J., Naiman, D. Q., and Bailey-Wilson, J. (2002).
A compresive method for genome scans.
Human Heredity, 54:174–185.

Naiman, D. Q. and Priebe, C. E. (2001).
Computing scan statistic p values using importance sampling, with applications to genetics and medical image analysis.
J. Comput. Graph. Statist., 10:296–328.

Naiman, D. Q. and Wynn, P. (1997).
Abstract tubes, improved inclusion exclusion identities and inequalities and importance sampling.
The Annals of Statistics, 25:1954–1983.

Neil, D. (2006).
Detection of spatial and spatio-temporal clusters.
PhD thesis, School of Computer Science, Carnegie Mellon University.

Neil, D. (2012).
Fast subset scan for spatial pattern detection.
Journal of the Royal Statistical Society, 74(2):337–360.

Rubino, G. and Tuffin, B. (2009).
Rare event simlation using Monte Carlo methods.

Wiley-Interscience [John Wiley & Sons], New York.

Shi, J., Siegmund, D., and Yakir, B. (2007).
Importance sampling for estimating p values in linkage analysis.
Journal of American Statistical Association, 102:929–937.

Wang, X. and Glaz, J. (2013).
A variable window scan statistic for $MA(1)$ process.
In *Proceedings, 15th Applied Stochastic Models and Data Analysis (ASMDA 2013)*, pages 905–912.

Importance sampling algorithm [Shi et al., 2007]

Algorithm 2 Second Importance Sampling Algorithm for Scan Statistics

Take $d\mathbb{P}_{\xi,r_1} = \frac{e^{\xi Y_{r_1}}}{\mathbb{E}_{H_0}\left[e^{\xi Y_{r_1}}\right]} d\mathbb{P}_{H_0}$ and compute

$$\xi \approx \frac{\tau}{m_1\sigma^2} - \frac{\mu}{\sigma^2}, \quad \mathbb{E}_{\xi,r_1}\left[Y_{i_1}\right] = \xi Cov_{H_0}\left[Y_{i_1}, Y_{r_1}\right] + m_1\mu, \quad Cov_{\xi,r_1}\left[Y_{i_1}, Y_{j_1}\right] = Cov_{H_0}\left[Y_{i_1}, Y_{j_1}\right]$$

Repeat for each k from 1 to $ITER$ (iterations number)

1: Generate uniformly $i_1^{(k)}$ from the set $\{1, \ldots, T_1 - m_1 + 1\}$.

2: Given $i_1^{(k)}$, generate the Gaussian process Y_{i_1} according to the new measure $d\mathbb{P}_{\xi, i_1^{(k)}}$.

3: Compute $\widehat{\rho}_k(1)$ based on

$$\widehat{\rho}_k(1) = \frac{T_1 - m_1 + 1}{\sum_{j_1=1}^{T_1-m_1+1} e^{\xi Y_{j_1} - m_1\left(\mu\xi + \frac{\sigma^2\xi^2}{2}\right)}} 1\{S_{m_1}(T_1) \geq \tau\}$$

End Repeat

Return

$$\widehat{\rho}(1) = \frac{1}{ITER}\sum_{k=1}^{ITER}\widehat{\rho}_k(1), \quad Var\left[\widehat{\rho}(1)\right] \approx \frac{1}{ITER - 1}\sum_{k=1}^{ITER}\left(\widehat{\rho}_k(1) - \frac{1}{ITER}\sum_{k=1}^{ITER}\widehat{\rho}_k(1)\right)^2$$

Return

Université Lille1 Sciences et Technologies