Numéro d'ordre : 41498 Année 2014

THESE DE DOCTORAT

présentée par

Alexandru Amărioarei

en vue de l'obtention du grade de

Docteur en Sciences de l'Université de Lille 1

Discipline : Mathématiques Appliquées

Approximations for Multidimensional Discrete Scan Statistics

Soutenue publiquement le **15 Septembre, 2014** devant le jury composé de:

Jury:

Directeur de these:	Cristian PREDA	*Université de Lille 1, France*
Rapporteurs:	Joseph GLAZ	*University of Connecticut, USA*
	Claude LEFEVRE	*Université Libre de Bruxelles, Belgique*
Examinateurs:	Azzouz DERMOUNE	*Université de Lille 1, France*
	Stéphane ROBIN	*AgroParisTech/INRA, France*
	George HAIMAN	*Université de Lille 1, France*
	Manuela SIDOROFF	*National Institute of R&D for Biological Sciences, Roumanie*

To my parents

Acknowledgements

I am using this opportunity to express my gratitude to everyone who supported me throughout the course of this work.

I would like to express my special appreciation and thanks to my advisor, Prof. Cristian PREDA, who has been and still is a tremendous mentor for me. I would like to thank him for introducing me to the wonderful subject of *scan statistics*, for encouraging my research and for allowing me to grow as a research scientist. His advices on research, career as well as on life in general are priceless.

It is a pleasure to recognize my dissertation committee members, Prof. Joseph GLAZ, Prof. Claude LEFEVRE, Prof. Azzouz DERMOUNE, Dr. Stéphane ROBIN, Prof. George HAIMAN and Dr. Manuela SIDOROFF, for taking interest in my research and accepting to examine my work. Special thanks to Prof. Joseph GLAZ and Prof. Claude LEFEVRE for their thorough reviews, fruitfull discussions and insightful comments.

I am very grateful to Prof. George HAIMAN from whom I had the opportunity to learn many interesting aspects about the subject of the present work, his taste of details and mathematical rigour, which improved my knowledge in the field of *scan statistics*.

I want to emphasize the role of Dr. Manuela SIDOROFF, who guided me during my first years of research at the National Institute of R&D for Biological Sciences in Bucharest, and without whom I would never had the chance to meet my advisor, Prof. Cristian PREDA and complete this work.

I would also like to thank the members of INRIA/$M\Theta DAL$ team for welcoming me among them and for making the last years more enjoyable, for their support, discussions and encouragements. I've learned a lot from them.

Many of my thanks and my gratitude go to Acad. Ioan CUCULESCU, the first person who showed me the depths of research and guided me on this path.

I express my gratitude to all of my friends who supported me in writing, and encouraged me to strive towards my goal.

A special thanks goes to my family. Words cannot express how grateful I am to my mother and my father for all of the sacrifices that they've made on my behalf. I dedicate this thesis to them.

Last, but not least, I would like express appreciation to my beloved Raluca who spent sleepless nights with and was always my support in the moments when there was no one to answer my queries.

Résumé

Dans cette thèse nous obtenons des approximations et les erreurs associées pour la distribution de la statistique de scan discrète multi-dimensionnelle. La statistique de scan est vue comme le maximum d'une suite de variables aléatoires stationnaires 1-dépendante. Dans ce cadre, nous présentons un nouveau résultat pour l'approximation de la distribution de l'extremum d'une suite de variables aléatoire stationnaire 1-dépendante, avec des conditions d'application plus larges et des erreurs d'approximations plus petites par rapport aux résultats existants en littérature. Ce résultat est utilisé ensuite pour l'approximation de la distribution de la statistique de scan. L'intérêt de cette approche par rapport aux techniques existantes en littérature est du à la précision d'une erreur d'approximation, d'une part, et de son applicabilité qui ne dépend pas de la distribution du champ aléatoire sous-adjacent aux données, d'autre part.

Les modèles considérés dans ce travail sont le modèle i.i.d et le modèle de dépendance de type block-factor.

Pour la modélisation i.i.d. les résultats sont détaillés pour la statistique de scan uni, bi et tri-dimensionnelle. Un algorithme de simulation de type "importance sampling" a été introduit pour le calcul effectif des approximations et des erreurs associées. Des études de simulations démontrent l'efficacité des résultats obtenus. La comparaison avec d'autres méthodes existantes est réalisée.

La dépendance de type block-factor est introduite comme une alternative à la dépendance de type Markov. La méthodologie développée traditionnellement dans le cas i.i.d. est étendue à ce type de dépendance. L'application du résultat d'approximation pour la distribution de la statistique de scan pour ce modèle de dépendance est illustrée dans le cas uni et bi-dimensionnel.

Ces techniques, ainsi que celles existantes en littérature, ont été implémentées pour la première fois à l'aide des programmes Matlab® et une interface graphique.

Abstract

In this thesis, we derive accurate approximations and error bounds for the probability distribution of the multidimensional scan statistics.

We start by improving some existing results concerning the estimation of the distribution of extremes of 1-dependent stationary sequences of random variables, both in terms of range of applicability and sharpness of the error bound. These estimates play the key role in the approximation process of the multidimensional discrete scan statistics distribution.

The presented methodology has two main advantages over the existing ones found in the literature: first, beside the approximation formula, an error bound is also established and second, the approximation does not depend on the common distribution of the observations. For the underlying random field under which the scan process is evaluated, we consider two models: the classical model, of independent and identically distributed observations and a dependent framework, where the observations are generated by a block-factor.

In the i.i.d. case, in order to illustrate the accuracy of our results, we consider the particular settings of one, two and three dimensions. A simulation study is conducted where we compare our estimate with other approximations and inequalities derived in the literature. The numerical values are efficiently obtained via an importance sampling algorithm discussed in detail in the text.

Finally, we consider a block-factor model for the underlying random field, which consists of dependent data and we show how to extend the approximation methodology to this case. Several examples in one and two dimensions are investigated. The numerical applications accompanying these examples show the accuracy of our approximation.

All the methods presented in this thesis leaded to a Graphical User Interface (GUI) software, implemented in Matlab®.

Contents

List of Figures

List of Tables

Introduction

There are many fields of application where an observed cluster of events could have a great influence on the decision taken by an investigator. To know if such an agglomeration of events is due to hazard or not, plays an important role in the decision-making process. For example, an epidemiologist observes over a predefined period of time (a week, a month, etc.) an accumulation of cases of an infectious disease among the population of a certain region. Under some model for the distribution of events, if the probability to observe such an unexpected cluster is small, with respect to a given threshold value, then the investigator can conclude that an atypical situation occurred and can take the proper measures to avoid a pandemic crisis. The problem of identifying accumulations of events that are *unexpected* or *anomalous* with respect to the distribution of events belongs to the class of cluster detection problems. Depending on the application domain, these *anomalous* agglomeration of events can correspond to a diversity of phenomena: for example one may want to search for clusters of stars, deposits of precious metals, outbreaks of disease, batches of defective pieces, brain tumors and many other possibilities.

A general class of testing procedures, used by practitioners to evaluate the likelihood of such clusters of events, are the tests based on scan statistics. These statistics, considered for the first time in the work of Naus in the 60s, are random variables defined as the maximum number of observations in a scanning window of predefined size and shape that is moved in a continuous fashion over all possible locations of the region of study. The tests based on scan statistics are usually employed when one wants to detect a local change (a hot spot) in the distribution of the underlying random field via testing the null hypothesis of uniformity against an alternative hypothesis which favors clusters of events. The importance of the tests based on scan statistics have been noted in many scientific and technological fields, including: DNA sequence analysis, brain imaging, distributed target detection in sensors networks, astronomy, reliability theory and quality control among many other domains.

To implement these testing procedures, one needs to find the distribution of the scan statistics. The main difficulty in obtaining the distribution of the scan random variable, under the null hypothesis, resides in the high dependent structure of the observations over which the maximum is taken. As consequence, several approximations have been proposed, especially for the case of one, two and three dimensional scan statistics.

In this thesis, we consider the multidimensional discrete scan statistics into a general framework. Viewed as maximum of some 1-dependent sequences of random variables, we derive accurate approximations and error bounds for its distribution. Our methodology applies to a larger class of distributions and extends the i.i.d. case to some new dependent models based on block-factor constructions.

This manuscript is organized into four chapters as follows.

In **Chapter 1**, we review some of the existing approaches used to develop exact and approximate results for the distribution of the unconditional discrete scan statistics. We consider, separately, the cases of one, two and three dimensional scan statistics. In the one dimensional setting, we include, along various approximations and bounds, three general methods used for determining the exact distribution of the scan statistics over a sequence of i.i.d. binary trials: the combinatorial method, the finite Markov chain imbedding technique and the conditional probability generating function method. A new upper bound for the distribution of the two dimensional scan statistics is presented. We should mention that most of the results presented in this chapter are given in their general form, extending thus their corresponding formulas that appear in the literature.

Chapter 2 introduces a series of results concerning the approximation of the distribution of the extremes of 1-dependent stationary sequences of random variables. We improve some existing results in terms of error bounds and range of applicability. Our new approximations will constitute the *main tools* in the estimation of the distribution of the multidimensional discrete scan statistics derived in the subsequent chapters.

The general case of d dimensional discrete scan statistics, $d \geq 1$ for independent and identically distributed observations, is considered in **Chapter 3**. Employing the results derived in Chapter 2, we present the methodology used for obtaining the approximation of the probability distribution function of the multidimensional dimensional discrete scan statistics. The main advantage of the described approach is that, beside the approximation formula, we can also establish sharp error bounds. Since the quantities that appear in the approximation of the scan statistics formula are usually evaluated by simulation, two types of errors are considered: the theoretical error bounds and the simulation error bounds. We give detailed expressions, based on recursive formulas, for the computation of these bounds. Due to the simulation nature of the problem, we also include a general importance sampling procedure to increase the efficiency of the proposed estimation. We discuss different computational aspects of the procedure and we compare it with other existing algorithms. We conclude the chapter with a series of examples for the special cases of one, two and three dimensional scan statistics. In these frameworks, we explicit the general formulas obtained for the d dimensional setting and we investigate their accuracy via a simulation study.

In **Chapter 4**, we consider the multidimensional discrete scan statistics over a random field generated by a block-factor model. This dependent model generalizes the i.i.d. model presented in Chapter 3. We extend the approximation methodology developed for the i.i.d. case to this model. We provide recurrent formulas for the computation of the approximation, as well as for the associated error bounds. In the final section, we present several examples for the special cases of one and two dimensional scan statistics to illustrate the method. In particular, we give an estimate for the length of the longest increasing run in a sequence of i.i.d. random variables and we investigate the scan statistics for moving average of order q models. Numerical results are included in order to evaluate the efficiency of our results.

To illustrate the efficiency and the accuracy of the methods presented in this thesis, we developed a Graphical User Interface (GUI) software, implemented in Matlab®. This software application provides estimates for the distribution of the discrete scan statistics for different scenarios. In this GUI, the user can choose the dimension of the problem and the distribution of the random field under which the scan process is performed. We consider the cases of one, two and three dimensional scan statistics over a random field distributed according to a Bernoulli, binomial, Poisson or Gaussian model. In the particular situation of one dimensional scan statistics, we have also included a moving average of order q model. A more detailed description of this GUI application is given in Appendix C.

Chapter 1

Existing methods for finding the distribution of discrete scan statistics

In this chapter, we review some of the existing methods used in the study of the unconditional discrete scan statistics. In Section 1.1, we consider the one dimensional case and describe some of the approaches used to determine the exact distribution of scan statistics along with various approximations and bounds. In Section 1.2 and Section 1.3, we focus on the two and three dimensional scan statistics, respectively.

Contents

1.1 One dimensional scan statistics

There are many situations when an investigator observes an accumulation of events of interest and wants to decide if such a realisation is due to hazard or not. These type of problems belong to the class of cluster detection problems, where the basic idea is to identify regions that are *unexpected* or *anomalous* with respect to the distribution of events. Depending on the application domain, these *anomalous* agglomeration of events can correspond to a diversity of phenomena: for example one

may want to find clusters of stars, deposits of precious metals, outbreaks of disease, minefield detections, defectuous batches of pieces and many other possibilities.
If such an observed accumulation of events exceeds a preassigned threshold, usually determined from a specified significance level corresponding to a *normal* situation (the null hypothesis), then it is legitimate to say that we have an unexpected cluster and proper measures has to be taken accordingly.
Searching for unusual clusters of events is of great importance in many scientific and technological fields including: DNA sequence analysis ([Sheng and Naus, 1994], [Hoh and Ott, 2000]), brain imaging ([Naiman and Priebe, 2001]), target detection in sensors networks ([Guerriero et al., 2009], [Guerriero et al., 2010b]), astronomy ([Darling and Waterman, 1986], [Marcos and Marcos, 2008]), reliability theory and quality control ([Boutsikas and Koutras, 2000]) among many other domains. One of the tools used by practitioners to decide on the unusualness of such agglomeration of events is the *scan statistics*. Basically, the tests based on scan statistics are looking for events that are clustered amongst a background of those that are sporadic.
Let $2 \leq m_1 \leq T_1$ be two positive integers and $X_1, \ldots, X_{T_1}$ be a sequence of independent and identically distributed random variables with the common distribution F_0. The *one dimensional discrete scan statistics* is defined as

$$S_{m_1}(T_1) = \max_{1 \leq i_1 \leq T_1 - m_1 + 1} Y_{i_1}, \tag{1.1}$$

where the random variables Y_{i_1} are the moving sums of length m_1 given by

$$Y_{i_1} = \sum_{i=i_1}^{i_1+m_1-1} X_i. \tag{1.2}$$

Usually, the statistical tests based on the one dimensional discrete scan statistics are employed when one wants to detect a local change in the signal within a sequence of T_1 observations via testing the null hypothesis of uniformity, H_0, against a cluster alternative, H_1 (see [Glaz and Naus, 1991] and [Glaz et al., 2001]). Under H_0, the random observations $X_1, \ldots, X_{T_1}$ are i.i.d. distributed as F_0, while under the alternative hypothesis, there exists a location $1 \leq i_0 \leq T_1 - m_1 + 1$ where X_i, $i \in \{i_0, \ldots, i_0 + m_1 - 1\}$, are distributed according to $F_1 \neq F_0$ and outside this region X_i are distributed as F_0.
We observe that whenever $S_{m_1}(T_1)$ exceeds the threshold τ, where the value of τ is computed based on the relation $\mathbb{P}_{H_0}(S_{m_1}(T_1) \geq \tau) = \alpha$ and α is a preassigned significance level of the testing procedure, the generalized likelihood ration test rejects the null hypothesis in the favor of the clustering alternative (see [Glaz and Naus, 1991]).
It is interesting to note that most of the research has been done for F_0 being binomial, Poisson or normal distribution (see [Naus, 1982], [Glaz and Naus, 1991], [Glaz and Balakrishnan, 1999], [Glaz et al., 2001] or [Wang et al., 2012]).
In Chapter 3, we present a new approximation method for the distribution of the discrete scan statistics, following the work of [Haiman, 2000], that can be evaluated no matter what distribution we have under the null hypothesis.

In this section, we revisit some of the existing methods used to obtain the exact values or the approximation of the distribution of the one dimensional discrete scan statistics.

1.1.1 Exact results for binary sequences

In this section, we consider that the random variables X_1, X_2, ..., X_{T_1}, are i.i.d. binary trials (Bernoulli model). From the best of our knowledge, there are three main approaches used for investigating the exact distribution of the one dimensional discrete scan statistics: the combinatorial method, the Markov chain imbedding technique (MCIT) and the conditional probability generating function method. We will give a short description of each method in the subsequent sections.

1.1.1.1 The combinatorial approach

Let $X_1, \ldots, X_{T_1}$ be a sequence of i.i.d. $0-1$ Bernoulli random variables of parameter p. The combinatorial method is based on a consequence of a Markov process result of [Karlin and McGregor, 1959], used for solving the ballot problem (see [Barton and Mallows, 1965, page 243]) and is briefly described in the following.
The probability distribution function of the one dimensional discrete scan statistics can be obtained, using the law of total probability, from the relation

$$\mathbb{P}\left(S_{m_1}(T_1) \leq n\right) = \sum_{k=0}^{T_1} \binom{T_1}{k} p^k (1-p)^{T_1-k} \mathbb{P}\left(S_{m_1}(T_1) \leq n \middle| \sum_{i=1}^{T_1} X_i = k\right), \quad (1.3)$$

where in the last relation we used the fact that $X_1 \in \{0,1\}$ with $\mathbb{P}(X_1 = 1) = p$ and $\sum_{i=1}^{T_1} X_i$ follows a binomial distribution with parameters T_1 and p.
In [Naus, 1974, Theorem 1] (see also [Glaz et al., 2001, Chapter 12]), the author presented a combinatorial formula for the conditional distribution of $S_{m_1}(T_1)$ given the total number of realisations in T_1 trials, i.e. $\sum_{i=1}^{T_1} X_i = k$.
Assuming that $T_1 = m_1 L$ and partitioning the total number of trials into L disjoint groups of size m_1, we have

$$\mathbb{P}\left(S_{m_1}(T_1) \leq n \middle| \sum_{i=1}^{T_1} X_i = k\right) = \frac{(m_1!)^L}{\binom{T_1}{k}} \sum_{\sigma \in \Gamma_n} det|d_{i,j}|, \quad (1.4)$$

where σ denote a partition of the k realisations (successes) into L numbers $(n_1, \ldots, n_L)$ such that $n_i \geq 0$ represents the number of realisations in the ith group. The set Γ_n denote the collection of all the partitions σ such that for each $i \in \{1, \ldots, L\}$ we have $n_i \leq n$. The $L \times L$ matrices $d_{i,j}$ are determined based on the formulas

$$d_{i,j} = \begin{cases} 0, \text{ if } c_{i,j} < 0 \text{ or } c_{i,j} > m_1 \\ \frac{1}{c_{i,j}!(n+1-c_{i,j})!}, \text{ otherwise} \end{cases} \quad (1.5)$$

with

$$c_{i,j} = \begin{cases} (j-i)(n+1) - \sum_{\substack{l=1 \\ i}}^{j-1} n_k + n_i, \text{ for } i < j \\ (j-i)(n+1) + \sum_{l=j} n_k, \text{ for } i \geq j. \end{cases} \tag{1.6}$$

It is important to emphasize that the evaluation of $\mathbb{P}(S_{m_1}(T_1) \leq n)$ via the combinatorial method is complex and requires excessive computational time. The problem arises from the large number of terms in the set Γ_n and from the fact that for each element in such a set one needs to evaluate the determinant of a $L \times L$ matrix. As [Naus, 1982] noted, the expression in Eq.(1.3) can only be evaluated for small window sizes and moderate L.

We include, for completeness, the formulas for the particular cases of $T_1 = 2m_1$ and $T_1 = 3m_1$. [Naus, 1982] using the relations in Eq.(1.3) and Eq.(1.4), gave the following closed form expressions for the distribution of the discrete scan statistics:

$$\begin{aligned} \mathbb{P}(S_{m_1}(2m_1) \leq n) &= F^2(n; m_1, p) - nb(n+1; m_1, p)F(n-1; m_1, p) \\ &\quad + m_1 p b(n+1; m_1, p)F(n-2, m_1 - 1), \end{aligned} \tag{1.7}$$

$$\mathbb{P}(S_{m_1}(3m_1) \leq n) = F^3(n; m_1, p) - A_1 + A_2 + A_3 - A_4, \tag{1.8}$$

with

$$\begin{aligned} A_1 &= 2b(n+1; m_1, p)F(n; m_1, p)[nF(n-1; m_1, p) - m_1 pF(n-2; m_1 - 1, p)], \\ A_2 &= 0.5b^2(n+1; m_1, p)\left[n(n-1)F(n-2; m_1, p) - 2(n-1)m_1 F(n-3; m_1 - 1, p)\right. \\ &\quad \left. + m_1(m_1-1)p^2 F(n-4; m_1 - 2, p)\right], \\ A_3 &= \sum_{r=1}^{n} b(2(n+1) - r; m_1, p)F^2(r-1; m_1, p), \\ A_4 &= \sum_{r=2}^{n} b(2(n+1) - r; m_1, p)b(r+1; m_1, p)\left[rF(r-1; m_1, p)\right. \\ &\quad \left. - m_1 pF(r-2; m_1 - 1, p)\right], \end{aligned} \tag{1.9}$$

and where $b(s; t, p)$ and $F(s; t, p)$ are the probability mass function and cumulative distribution function of the binomial random variable $\mathcal{B}(t, p)$, that is

$$b(s; t, p) = \binom{t}{s} p^s (1-p)^{t-s} \tag{1.10}$$

$$F(s; t, p) = \sum_{i=0}^{s} b(i; t, p). \tag{1.11}$$

As we will see in Section 1.1.2, the foregoing relations are successfully used in the product type approximation of the scan statistics over a Bernoulli sequence.

1.1.1.2 Markov chain imbedding technique

In this section, we present succinctly the second approach used for finding the exact distribution of the one dimensional discrete scan statistics for binary trials, namely the Markov Chain Imbedding Technique (for short MCIT). The method was developed by [Fu, 1986] and [Fu and Koutras, 1994] and successfully employed to derive the exact distribution associated with several runs statistics in either independent and identically distributed or Markov dependent trials. The two monographs of [Fu and Lou, 2003] and [Balakrishnan and Koutras, 2002] give a full account of the development and the applications of the method and also provide a lot of references in this research area.

The main idea behind the MCIT method, as the name suggests, is to imbed the random variable of interest into a Markov chain with an appropriate state space. Using the Chapman-Kolmogorov equation, the desired probability can be computed via multiplication of the transition probability matrices of the imbedded Markov chain. To know if the studied random variable is Markov chain embeddable is not an easy task. For some particular classes of runs in multistate trials (number of occurrences of pattern, waiting time variables associated with a pattern, etc.) a general procedure, called the *forward-backward principle*, was introduced in [Fu, 1996] (see also [Fu and Lou, 2003, Chapter 4]). This procedure systematically shows how to imbed the random variable of interest into a Markov chain that carry all the necessary information.

In [Koutras and Alexandrou, 1995, Section 4c], the authors showed that the scan statistics $S_{m_1}(T_1)$ is Markov chain embeddable, thus, its distribution function, $\mathbb{P}(S_{m_1}(T_1) < n)$, can be evaluated in terms of the transition probability matrix of the imbedded chain. The imbedded Markov chain is defined on a state space that, at each step $t > m_1$, keeps track of the number of the occurrences of the event $\{S_{m_1}(t) < n\}$ in the first t trials and of the last m_1 realisations of the sequence $X_{t-m_1+1}, \ldots, X_t$. A similar methodology was adopted by [Wu, 2013], where the author defined the imbedded chain on the state space of all tuples containing the locations of the successes (1's) counting backward in a window of size m_1 at each time t. The drawback of both of their approaches is that the state space of the imbedded chain is rather large (is equal with 2^{m_1}), so the transition matrix becomes quickly intractable even for moderate window sizes.

[Fu, 2001] (see also [Fu and Lou, 2003, Section 5.9]) suggested a different approach for finding the distribution of the scan statistics which involves a smaller state space for the imbedded chain. The author expressed the distribution function $\mathbb{P}(S_{m_1}(T_1) < n)$ as the probability of the tail of a waiting time variable associated to a compound pattern. This type of random variable was extensively studied and has been shown to be Markov chain embeddable (see [Fu and Chang, 2002] or [Fu and Lou, 2003, Chapter 5]).

We will describe succinctly both the approach proposed by [Fu, 2001] and the one in [Wu, 2013], in order of publication.

For simplicity, we consider only the case of i.i.d. two state Bernoulli trials, the

case of homogeneous Markov trials being similar. Let $X_1, \ldots, X_{T_1}$ be i.i.d. $0-1$ Bernoulli random variables with success probability $\mathbb{P}(X_1 = 1) = p$.
For a given window size m_1 and n, with $0 \leq n \leq m_1$, we consider the following set of simple patterns:

$$\mathcal{F}_{m_1,n} = \{\Lambda_i | \Lambda_1 = \underbrace{1 \ldots 1}_{n}, \Lambda_2 = 10\underbrace{1 \ldots 1}_{n-1}, \ldots, \Lambda_l = \overbrace{\underbrace{1 \ldots 1}_{n-1}0 \ldots 01}^{m_1}\} \tag{1.12}$$

that is the set of all the simple patterns that start and end with a success (1), contain exactly n symbols of 1 and with length at most m_1. As [Fu, 2001] showed, the number of elements in $\mathcal{F}_{m_1,n}$ is given by the formula:

$$l = \sum_{j=0}^{m_1-n} \binom{n-2+j}{j}. \tag{1.13}$$

Considering the compound pattern $\Lambda_{m_1,n} = \bigcup_{i=1}^{l} \Lambda_i$, we observe that the scan statistics $S_{m_1}(T_1)$ and the waiting time $W(\Lambda_{m_1,n})$ to observe one of the simple patterns $\Lambda_1 \ldots, \Lambda_l$, are related by the following relation:

$$\mathbb{P}\left(S_{m_1}(T_1) < n\right) = \mathbb{P}\left(W(\Lambda_{m_1,n}) > T_1\right). \tag{1.14}$$

The state space of the imbedded chain can be written as

$$\Omega = \{0,1\} \bigcup_{i=1}^{l} \mathcal{S}(\Lambda_i), \tag{1.15}$$

where $\mathcal{S}(\Lambda_i)$ is the set of all sequential subpatterns of Λ_i.
[Chang, 2002] (see also [Fu and Chang, 2002]) showed that the transition matrix of the Markov chain that corresponds to the waiting time $W(\Lambda_{m_1,n})$ variable has the form

$$\boldsymbol{M}_{m_1,n} = \begin{matrix} \Omega \diagdown A \\ A \end{matrix} \begin{pmatrix} \boldsymbol{N}_{m_1,n} & \boldsymbol{C} \\ \boldsymbol{O} & \boldsymbol{I} \end{pmatrix}_{d \times d} \tag{1.16}$$

where Ω is the state space of the chain and has d elements, $A = \{\alpha_1, \ldots, \alpha_l\}$ denote the set of the absorbing states corresponding to the simple patterns $\Lambda_1 \ldots, \Lambda_l$ respectively, $\boldsymbol{N}_{m_1,n}$ is an $(d-l) \times (d-l)$ matrix called *essential matrix*, $\boldsymbol{O}$ is an $l \times (d-l)$ zero matrix and $\boldsymbol{I}$ is an $l \times l$ identity matrix. Given $u, v \in \Omega$, the transition probabilities of the imbedded chain $\{Z_t\}$ are computed via

$$p_{u,v} = \mathbb{P}\left(Z_t = v | Z_{t-1} = u\right) = \begin{cases} q, \text{ if } X_t = 0,\ u \in \Omega \diagdown A \text{ and } v = [u,0]_\Omega, \\ p, \text{ if } X_t = 1,\ u \in \Omega \diagdown A \text{ and } v = [u,1]_\Omega, \\ 1, \text{ if } u \in A \text{ and } v = u, \\ 0, \text{ otherwise}, \end{cases} \tag{1.17}$$

where the notation $v = [u, a]_\Omega$, $a \in \{0, 1\}$, means that v is the longest ending subsequence of observations that determines the status of forming the next pattern Λ_i from $X_{t-m_1+1}, \ldots, X_t$ (see [Fu and Lou, 2003, Theorem 5.2]).
It follows from [Fu, 2001, Theorem 2] that the probability of the tail of the waiting time $W(\Lambda_{m_1,n})$ is given by:

$$\mathbb{P}\left(W(\Lambda_{m_1,n}) > T_1\right) = \boldsymbol{\xi} \boldsymbol{N}_{m_1,n}^{T_1} \mathbf{1}^\intercal, \tag{1.18}$$

where $\boldsymbol{\xi} = (q, p, 0, \ldots, 0)$ is the initial distribution ($q = 1-p$), $\boldsymbol{N}_{m_1,n}$ is the essential matrix and $\mathbf{1}^\intercal$ is the transpose of the vector $(1, \ldots, 1)$.
The methodology proposed by [Wu, 2013] is different from the foregoing procedure and is worth mentioning since it leads to a recurrence relation for the distribution of the one dimensional scan statistics. As pointed out at the beginning of the section, the idea is to construct a Markov chain that keeps track of the locations of the successes counting backward in the window of size m_1 at each moment t. Hence, the state space of the imbedded chain $\{Z_t\}$ is

$$\Omega = \{0\} \cup \{w_1 \ldots w_j \,|\, w_j = 1, \ldots, m_1,\, w_1 < \cdots < w_j,\, j = 1, \ldots, m_1\}. \tag{1.19}$$

Assume that at time t the Markov chain takes the value $Z_t = w_1 \ldots w_j$, then the $m_1 - w_s + 1$, $s \in \{1, \ldots, j\}$, element of the realisation of $X_{t-m_1+1}, \ldots, X_t$ is a success. Employing a base-2 relabeling of the state space Ω,

$$v = w_1 \ldots w_j \to v_L = \sum_{i=1}^{j} 2^{w_i - 1} + 1, \tag{1.20}$$

the author showed ([Wu, 2013, Theorem 1]) that the distribution function of the scan statistics variable $S_{m_1}(T_1)$ can be evaluated via

$$\mathbb{P}\left(S_{m_1}(T_1) < n\right) = \sum_{\substack{v \to v_L \\ v_L \in A_n}} \mathbb{P}\left(Z_{T_1} = v\right) = \sum_{v_L \in A_n} a_{T_1}(v_L), \tag{1.21}$$

where A_n is the set of labels associated with the states that corresponds to less than n successes within the m_1 trailing observations and where $a_{T_1}(v_L)$ are computed by the following recurrence relations:

$$a_t(v_L) = \begin{cases} q a_{t-1}\left(\frac{v_L+1}{2}\right) + q a_{t-1}\left(2^{m_1-1} + \frac{v_L+1}{2}\right), & v_L \text{ odd}, \\ p a_{t-1}\left(\frac{v_L}{2}\right) + p a_{t-1}\left(2^{m_1-1} + \frac{v_L}{2}\right), & v_L \text{ even}, \end{cases} \tag{1.22}$$

with $t \in \{1, \ldots, T_1\}$, $v_L \in \{1, \ldots, 2^{m_1}\}$ and the initial conditions $a_0(1) = 1$, $a_0(v_L) = 0$ if $v_L \neq 1$, and $a_t(v_L) = 0$ if $v_L \notin A_n$.
We note that in the above relations both the size of the scanning window and the size of the sequence play an important role. In the approach presented by [Fu, 2001], only the size of the window influenced the transition matrix. Nevertheless, due to the recursive aspect, the second method seems to compute faster the distribution of the scan statistics for small-moderate values of m_1 ($m_1 \leq 30$) and moderate values of T_1 ($T_1 \leq 500$). For long sequences ($T_1 \geq 1000$) and small-moderate window sizes, the first method is desirable.

Remark 1.1.1. *Recently, [Nuel, 2008a] and [Nuel, 2008b] proposed a new approach for constructing the optimal state space of the imbedded Markov chain. The method, called Pattern Markov Chain, is based on the theory of formal language and automata (deterministic finite automata) and was successfully applied in the context of biological data.*

1.1.1.3 Probability generating function methodology

Another method used for determining the exact distribution of the one dimensional discrete scan statistics is the conditional probability generating function (*pgf*) approach. This method provides a way of deriving the probability generating function for the variable of interest and then, by differentiating it a number of times, yield the probability distribution function. The *pgf* method was introduced by [Feller, 1968, Chapter XIII] in the context of the theory of recurrent events and has been intensively used in research and education since then. This method is also one of the main tools for investigating the exact distribution of runs and scans statistics and, as in the case of the Markov chain imbedding technique, can be applied in both the case of independent and identically distributed or Markov dependent trials.
For example, [Ebneshahrashoob and Sobel, 1990] applied the method for determining the exact distribution of the sooner and later waiting time of some succession of events in Bernoulli trials, [Aki, 1992] (see also [Uchida, 1998]) used the same approach to generalize this problem to independent nonnegative integer value random variables and [Uchida and Aki, 1995] extended the results to the Markov dependent binary case. [Balakrishnan et al., 1997] successfully applied the method for start-up demonstration tests under Markov dependence.
The basic idea behind the *pgf* approach is to construct a system of recurrent relations of conditional probability generating functions of the desired random variable by considering the condition of one-step ahead from every condition. Then, by solving the resulted system of equations with respect to the conditional generating functions, we can get the (unconditional) probability generating function (see for example [Han and Hirano, 2003] or [Chaderjian et al., 2012]). It is interesting to note that [Feller, 1968, Chapter XIV] exploited this idea for finding the generating function for the first passage time in the classical ruin problem. Even if in some cases the system of recurrent relations cannot be solved symbolically, due to the large dimension of the problem, in many cases the conditional probability generating functions can be written explicitly (for example in [Aki, 1992] or [Uchida and Aki, 1995]).
Recently, [Ebneshahrashoob et al., 2004] and [Ebneshahrashoob et al., 2005], by utilizing sparse matrix computational tools, enlarged the range of applicability of the *pgf* method.
In [Ebneshahrashoob et al., 2005], the authors applied the method of conditional probability generating functions to derive the exact distribution of the one dimensional discrete scan statistics over a sequence of i.i.d. or Markov binary trials. In what follows, we give an outline of their results. We consider, for simplicity, the case of i.i.d. observations. As we will see, both the *pgf* approach of

[Ebneshahrashoob et al., 2005] and the MCIT given by [Fu, 2001] require roughly the same computational effort.
Let $2 \leq m_1 \leq T_1$ be positive integers and $X_1, \ldots, X_{T_1}$ be a sequence of i.i.d. $0-1$ Bernoulli of parameter p random variables. Let $W_{m_1,n}$, $n \leq m_1$, denote the waiting time until we first observe at least n successes in a window of size m_1. Clearly (see [Glaz et al., 2001, Chapter 13]), the random variables $S_{m_1}(T_1)$ and $W_{m_1,n}$ satisfy the relation

$$\mathbb{P}\left(S_{m_1}(T_1) \geq n\right) = \mathbb{P}\left(W_{m_1,n} \leq T_1\right). \tag{1.23}$$

Thus, finding the exact distribution of the waiting time $W_{m_1,n}$, automatically gives the exact distribution of the scan statistics $S_{m_1}(T_1)$ via Eq.(1.23). To derive the distribution of $W_{m_1,n}$, we employ the *pgf* methodology.
Let $G(t)$ be the probability generating function of the waiting time variable,

$$G(t) = \mathbb{E}\left[t^{W_{m_1,n}}\right] = \sum_{k=0}^{\infty} t^k \mathbb{P}\left(W_{m_1,n} = k\right) \tag{1.24}$$

and denote with $G_1(t)$ and $G_0(t)$,

$$G_1(t) = \sum_{k=0}^{\infty} t^k \mathbb{P}\left((W_{m_1,n}|X_1 = 1) = k\right), \tag{1.25}$$

$$G_0(t) = \sum_{k=0}^{\infty} t^k \mathbb{P}\left((W_{m_1,n}|X_1 = 0) = k\right), \tag{1.26}$$

the pgf's of the conditional distribution of the waiting time given that the first trial was a success or a failure, respectively. We immediately observe (see [Feller, 1968, Chapter XIV, Section 4]), that between the probability generating functions $G(t)$, $G_1(t)$ and $G_0(t)$ the following recurrence relation holds

$$G(t) = ptG_1(t) + qtG_0(t), \tag{1.27}$$

where $q = 1 - p$ is the failure probability. To take into account the realisations of $W_{m_1,n}$, we condition on the positions of the successes in the last m_1 trials. If $k \leq n$ and $1 \leq x_1 < x_2 < \cdots < x_k < m_1$, let $G_{x_1,\ldots,x_k}(t)$ denote the probability generating function of the waiting time given that there was a success x_1, x_2, $\ldots$, x_k steps back, respectively and no other in the last m_1 trials. As in the case of Eq.(1.27), the pgf's $G_{x_1,\ldots,x_k}(t)$ verify the following recurrences

$$G_{x_1,\ldots,x_k}(t) = ptG_{1,x_1+1,\ldots,x_k+1}(t) + qtG_{x_1+1,\ldots,x_k+1}(t). \tag{1.28}$$

We should note that the above approach resembles the procedure proposed by [Wu, 2013] in the Markov chain imbedding context.
Note that the Eqs.(1.27) and (1.28) lead to a system of linear equations consisting of conditional probability generating functions whose solution determines the pgf of

the waiting time $W_{m_1,n}$. If the pgf $G(t)$ can be computed, then the distribution of the waiting time can be evaluated by differentiation via

$$\mathbb{P}\left(W_{m_1,n}=k\right)=\frac{G^{(k)}(0)}{k!},\ k\in 1,2,\ldots. \tag{1.29}$$

We remark that among the conditional pgf's that satisfy Eq.(1.28) there are several which are redundant and need to be eliminated. For example, if $m_1-x_1<n-1$, then $G_{x_1}(t)=G_0(t)$, since the success occurred x_1 steps back can no longer influence the outcome of the waiting time variable (from the definition of $W_{m_1,n}$). Similarly we have

$$G_{x_1,\ldots,x_k}(t)=G_{x_1,\ldots,x_{k-1}}(t),\text{ if } m_1-x_k<n-k, \tag{1.30}$$
$$G_{x_1,\ldots,x_k}(t)=1,\text{ if } k=n. \tag{1.31}$$

Since for large values of the scanning window m_1 the resulted system of equations cannot be solved symbolically, [Ebneshahrashoob et al., 2005] proposed an alternative method to overcome this difficulty. The idea is to write the system in matrix form and to employ sparse matrix tools for the evaluation. Let $\mathbf{G}(t)$ be the $N\times 1$ vector of pgf's

$$\mathbf{G}(t)=\left(G(t),G_0(t),G_1(t),\ldots,G_{m_1-n+1,\ldots,m_1-1}(t)\right)^{\intercal}, \tag{1.32}$$

where N is the number of non constant pgf's $G_{x_1,\ldots,x_k}(t)$ after applying the reduction rules in Eqs.(1.30) and (1.31). It is not hard to verify (see for example [Ebneshahrashoob et al., 2005] or [Gao and Wu, 2006]), that this number is equal with

$$N=1+\binom{m_1}{n-1}. \tag{1.33}$$

Note that in the expression of $\mathbf{G}(t)$, the pgf $G_{x_1,\ldots,x_k}(t)$, $k\geq 1$, occupy the position given by the index

$$N_{x_1,\ldots,x_k}=1+\binom{m_1-n+k+1}{k}-\binom{m_1-n+k-x_1}{k}-\binom{m_1-n+k-x_2}{k-1}$$
$$-\cdots-\binom{m_1-n+k-x_k}{1}. \tag{1.34}$$

We observe that the system of recurrence relations given by Eqs.(1.27) and (1.28), can be written in matrix form

$$\mathbf{G}(t)=tA\mathbf{G}(t)+tb, \tag{1.35}$$

where A is a $N\times N$ matrix and b is a $N\times 1$ vector.
The entries in the matrix A and the vector b are determined based on the recurrence relations in Eq.(1.28) and the elimination rules from Eqs.(1.30) and (1.31), used in order to verify if a generated pgf is new, constant or equivalent with a previous one. For example, in Eq.(1.28), consider $N_{x_1,\ldots,x_k}$ the position of the pgf $G_{x_1,\ldots,x_k}(t)$ in

the vector $\mathbf{G}(t)$. If after applying the reduction rules $G_{1,x_1+1,\ldots,x_k+1}(t) = 1$, then the $N_{x_1,\ldots,x_k}$'s component of the vector b is equal with p. By contrary, if none of the pgf's $G_{1,x_1+1,\ldots,x_k+1}(t)$ and $G_{x_1+1,\ldots,x_k+1}(t)$ is constant and after the application of the reduction rules are transformed in $G_{y_1,\ldots,y_l}(t)$ and $G_{z_1,\ldots,z_r}(t)$, respectively, then

$$A\left(N_{x_1,\ldots,x_k}, N_{y_1,\ldots,y_l}\right) = p, \tag{1.36}$$

$$A\left(N_{x_1,\ldots,x_k}, N_{z_1,\ldots,z_r}\right) = q, \tag{1.37}$$

and $A(i,j) = 0$ otherwise. We note that the matrix A has at most two non zero values on each row thus its sparse character. It is interesting to observe that, as the authors pointed out, the matrix A and the matrix $\boldsymbol{N}_{m_1,n}$ obtained from the MCIT by [Fu, 2001] are similar.
Simple calculation show that

$$\frac{\mathbf{G}^{(k)}(0)}{k!} = A^{k-1}b, \tag{1.38}$$

hence $\mathbb{P}\left(W_{m_1,n} = k\right)$ is the first component of $A^{k-1}b$.

Remark 1.1.2. *[Shinde and Kotwal, 2008] derived the exact distribution of the one dimensional scan statistics in a sequence of binary trials in a different fashion. They employed the conditional probability generating functions method to determine the joint distribution of $(M_{T_1,m_1,1}, \ldots, M_{T_1,m_1,m_1})$, where $M_{T_1,m_1,k}$ is the number of overlapping m_1-tuples which contain at least k successes in a sequence of T_1 trials. Then the distribution of the scan statistic $S_{m_1}(T_1)$ can be obtained from the relation*

$$\mathbb{P}\left(S_{m_1}(T_1) = n\right) = \mathbb{P}\left(M_{T_1,m_1,n+1} = 0\right) - \mathbb{P}\left(M_{T_1,m_1,n} = 0\right). \tag{1.39}$$

Since their formulas are rather long we will not include them here.

1.1.2 Approximations

Due to the high complexity and the limited range of application of the exact formulas, a considerable number of approximations and bounds have been developed for the estimation of the distribution of the one dimensional discrete scan statistics, for example [Naus, 1982], [Glaz and Naus, 1991], [Chen and Glaz, 1997] or [Wang et al., 2012]. A full treatment of these results is presented in the reference books of [Glaz and Balakrishnan, 1999] and [Glaz et al., 2001]. In this section, we choose to describe the class of *product type* approximations since, in most cases, are the most accurate ones.
Let $X_1, \ldots, X_{T_1}$ be a sequence of independent and identically distributed random variables with the common distribution F_0. Assume that $T_1 = Lm_1 + l$, with $L \geq 3$, $m_1 \geq 2$ and $0 \leq l \leq m_1 - 1$, and denote the distribution function of the scan statistics by

$$Q_{m_1}(T_1) = \mathbb{P}\left(S_{m_1}(T_1) \leq n\right). \tag{1.40}$$

[Naus, 1982] showed that the following product type approximation

$$\mathbb{P}\left(S_{m_1}(T_1) \leq n\right) \approx Q_{m_1}(2m_1 + l) \left[\frac{Q_{m_1}(3m_1)}{Q_{m_1}(2m_1)}\right]^{L-2}, \tag{1.41}$$

is highly accurate for the entire range of the distribution. A slightly different product type approximation is given by

$$\mathbb{P}\left(S_{m_1}(T_1) \leq n\right) \approx Q_{m_1}(2m_1 - 1) \left[\frac{Q_{m_1}(2m_1)}{Q_{m_1}(2m_1 - 1)}\right]^{T_1 - 2m_1 + 1}. \tag{1.42}$$

The foregoing relation is especially useful when one can evaluate $Q_{m_1}(T)$ only for $T < 3m_1$. The intuition behind the approximation formula given by Eq.(1.41) arise from writing the distribution of the scan statistics as the intersection of L 1-dependent events. Applying the product rule, we have

$$\begin{aligned}\mathbb{P}\left(S_{m_1}(T_1) \leq n\right) &= \mathbb{P}\left(\bigcap_{i=1}^{L} E_i\right) = \mathbb{P}\left(E_1\right)\mathbb{P}\left(E_2|E_1\right)\cdots\mathbb{P}\left(E_L\middle|\bigcap_{i=1}^{L-1} E_i\right) \\ &= \mathbb{P}\left(E_1 \cap E_2\right)\prod_{t=3}^{L-1}\mathbb{P}\left(E_t\middle|\bigcap_{i=1}^{t-1} E_i\right)\mathbb{P}\left(E_L\middle|\bigcap_{i=1}^{L-1} E_i\right),\end{aligned} \tag{1.43}$$

where

$$E_t = \begin{cases} \left\{\max\limits_{(t-1)m_1+1 \leq i_1 \leq tm_1+1} Y_{i_1} \leq n\right\}, \text{ for } 1 \leq t \leq L-1 \\ \left\{\max\limits_{(L-1)m_1+1 \leq i_1 \leq (L-1)m_1+l+1} Y_{i_1} \leq n\right\}, \text{ for } t = L. \end{cases} \tag{1.44}$$

As in [Naus, 1982], we use the Markov like approximation for the conditional probabilities of the form

$$\mathbb{P}\left(E_t\middle|\bigcap_{i=1}^{t-1} E_i\right) \approx \mathbb{P}\left(E_t|E_{t-1}\right), \ 3 \leq t \leq L-1. \tag{1.45}$$

Since for $3 \leq t \leq L-1$ the intersections $E_t \cap E_{t-1}$ are stationary, due to exchangeability, we conclude that

$$\begin{aligned}\mathbb{P}\left(S_{m_1}(T_1) \leq n\right) &\approx \mathbb{P}(E_1 \cap E_2)\left[\frac{\mathbb{P}(E_1 \cap E_2)}{\mathbb{P}(E_1)}\right]^{L-3}\frac{\mathbb{P}(E_{L-1} \cap E_L)}{\mathbb{P}(E_{L-1})} \\ &= Q_{m_1}(3m_1)\left[\frac{Q_{m_1}(3m_1)}{Q_{m_1}(2m_1)}\right]^{L-3}\frac{Q_{m_1}(2m_1 + l)}{Q_{m_1}(2m_1)}.\end{aligned} \tag{1.46}$$

We observe that in order to use the product type approximations in Eqs.(1.41) and (1.42) we need to evaluate $Q_{m_1}(2m_1 - 1)$, $Q_{m_1}(2m_1)$ and $Q_{m_1}(3m_1)$. For the particular case when F_0 is a Bernoulli distribution, [Naus, 1982] derived exact formulas for $Q_{m_1}(2m_1)$ and $Q_{m_1}(3m_1)$ as we saw in Section 1.1.1.1 Eqs.(1.7)-(1.9). For the case of binomial and Poisson distribution, [Karwe and Naus, 1997], based on a generating function approach, gave recursive formulas for the computation of $Q_{m_1}(T)$ up to $T \leq 3m_1$. For completeness, we include in Section A.1 of Appendix A the algorithm of [Karwe and Naus, 1997] for the evaluation of $Q_{m_1}(2m_1 - 1)$ and $Q_{m_1}(2m_1)$.

Remark 1.1.3. *Other approximations for the distribution of the one dimensional discrete scan statistics have been developed by [Haiman, 2007]. These results are presented in detail in Chapter 3.*

1.1.3 Bounds

Several bounds have been proposed for the estimation of $\mathbb{P}(S_{m_1}(T_1) \leq n)$. Basically these inequalities can be divided into two classes, depending on the method employed in their derivation: product type bounds and Bonferroni type bounds. It is important to note that the inequalities in the first category are tighter than the ones provided by the Bonferroni approach, since the method of proof in this case takes into account the dependence structure of the moving sums Y_{i_1}. A detailed comparison between the two methodologies was given in [Glaz, 1990].
[Glaz and Naus, 1991], in the case of discrete variables, and later [Wang et al., 2012], for the continuous case, developed the following product type bounds for the distribution of the one dimensional scan statistics (see also [Glaz et al., 2001, Chapter 13]):

a) Lower bounds

$$\mathbb{P}(S_{m_1}(T_1) \leq n) \geq \frac{Q_{m_1}(2m_1)}{\left[1 + \frac{Q_{m_1}(2m_1-1) - Q_{m_1}(2m_1)}{Q_{m_1}(2m_1-1)Q_{m_1}(2m_1)}\right]^{T_1-2m_1}}, \; T_1 \geq 2m_1 \tag{1.47}$$

$$\geq \frac{Q_{m_1}(3m_1)}{\left[1 + \frac{Q_{m_1}(2m_1-1) - Q_{m_1}(2m_1)}{Q_{m_1}(3m_1-1)}\right]^{T_1-3m_1}}, \; T_1 \geq 3m_1. \tag{1.48}$$

b) Upper bounds

$$\mathbb{P}(S_{m_1}(T_1) \leq n) \leq Q_{m_1}(2m_1)\left[1 - Q_{m_1}(2m_1-1) + Q_{m_1}(2m_1)\right]^{T_1-2m_1}, \; T_1 \geq 2m_1 \tag{1.49}$$

$$\leq Q_{m_1}(3m_1)\left[1 - Q_{m_1}(2m_1-1) + Q_{m_1}(2m_1)\right]^{T_1-3m_1}, \; T_1 \geq 3m_1. \tag{1.50}$$

The quantities that appear in the above lower and upper bounds can be evaluated, in the case of discrete random variables, via the recurrence relations of [Karwe and Naus, 1997] (see Section A.1).
We now describe briefly the Bonferroni type bounds for $Q_{m_1}(T_1)$. Since, in general, Bonferroni inequalities deal with union of events, the first step is to express the event of interest $\{S_{m_1}(T_1) \leq n\}$ as a union of events. Employing the notations used in Section 1.1.2 we can write

$$\mathbb{P}(S_{m_1}(T_1) \leq n) = \mathbb{P}\left(\bigcap_{i=1}^{L} E_i\right) = 1 - \mathbb{P}\left(\bigcup_{i=1}^{L} E_i^c\right), \tag{1.51}$$

where $T_1 = Lm_1 + l$ and E_i are given in Eq.(1.44). [Hunter, 1976] (see also [Worsley, 1982]), using graph theory arguments (spanning tree structure), derived the following upper bound for the union of events:

$$\begin{aligned}\mathbb{P}\left(\bigcup_{i=1}^{L} E_i^c\right) &\leq \sum_{i=1}^{L} \mathbb{P}\left(E_i^c\right) - \sum_{i=1}^{L-1} \mathbb{P}\left(E_i^c \cap E_{i+1}^c\right) \\ &= (L-1)\left[1-\mathbb{P}(E_1)\right] - (L-2)\left[1-2\mathbb{P}(E_1)+\mathbb{P}(E_1\cap E_2)\right] \\ &\quad + \mathbb{P}(E_L^c) - \mathbb{P}\left(E_{L-1}^c \cap E_L^c\right), \end{aligned} \tag{1.52}$$

which substituted into Eq.(1.51) gives the following lower bound for the distribution of $S_{m_1}(T_1)$:

$$\mathbb{P}\left(S_{m_1}(T_1) \leq n\right) \leq 1 - (L-2)\left[Q_{m_1}(2m_1) - Q_{m_1}(3m_1)\right] + Q_{m_1}(2m_1+l). \tag{1.53}$$

In Eq.(1.52) and Eq.(1.53), based on the stationarity of the events E_i, we used the relations $\mathbb{P}(E_i) = \mathbb{P}(E_1) = Q_{m_1}(2m_1)$, $\mathbb{P}(E_i \cap E_{i+1}) = \mathbb{P}(E_1 \cap E_2) = Q_{m_1}(3m_1)$ for $1 \leq i \leq L-1$ and $\mathbb{P}(E_{L-1} \cap E_L) = Q_{m_1}(2m_1+l)$.
For the upper bound, we employ the inequality in [Dawson and Sankoff, 1967] to get

$$\mathbb{P}\left(S_{m_1}(T_1) \leq n\right) \leq 1 - \frac{2S_1}{u} + \frac{2S_2}{u(u-1)}, \tag{1.54}$$

where

$$S_1 = \sum_{i=1}^{L} \mathbb{P}(E_i^c) = 1 + (L-2)\left[1 - Q_{m_1}(2m_1)\right] - Q_{m_1}(m_1+l), \tag{1.55}$$

$$\begin{aligned} S_2 &= \sum_{1\leq i<j\leq L} \mathbb{P}(E_i^c \cap E_j^c) = \sum_{j=2}^{L}\sum_{i=1}^{j-1}\left[1-\mathbb{P}(E_i)-\mathbb{P}(E_j)+\mathbb{P}(E_i\cap E_j)\right] \\ &= \sum_{j=2}^{L}\sum_{i=1}^{j-1}\left[1-2\mathbb{P}(E_1)+\mathbb{P}(E_i\cap E_j)\right] + \sum_{i=1}^{L-1}\left[1-\mathbb{P}(E_1)-\mathbb{P}(E_L)+\mathbb{P}(E_i\cap E_L)\right] \\ &= \frac{1}{2}(L-1)(L-2)\left[1-2Q_{m_1}(2m_1)\right] + (L-1)\left[1-Q_{m_1}(2m_1)-Q_{m_1}(m_1+l)\right] \\ &\quad + \sum_{j=2}^{L-1}\sum_{i=1}^{j-1}\mathbb{P}(E_i\cap E_j) + \sum_{i=1}^{L-1}\mathbb{P}(E_i\cap E_L) \\ &= \frac{1}{2}(L-1)(L-2)\left[1-2Q_{m_1}(2m_1)\right] + (L-1)\left[1-Q_{m_1}(2m_1)-Q_{m_1}(m_1+l)\right] \\ &\quad + \frac{1}{2}(L-2)(L-3)\left[Q_{m_1}(2m_1)\right]^2 + (L-2)\left[Q_{m_1}(2m_1)Q_{m_1}(m_1+l) + Q_{m_1}(3m_1)\right] \\ &\quad + Q_{m_1}(2m_1+l) \end{aligned} \tag{1.56}$$

and where u is the integer part of $2 + 2\frac{S_2}{S_1}$. In the derivation of S_1 and S_2 we used the one dependence and stationarity property of the events E_i.
In [Kwerel, 1975] (see also [Galambos and Simonelli, 1996, Inequality I7]) it is shown that the bound in Eq.(1.54) is the best possible in the class of linear inequalities of

the form $a_1S_1 + a_2S_2$. A slightly better result can be attained if one employs the recent inequality of [Kuai et al., 2000]

$$\mathbb{P}\left(S_{m_1}\ (T_1) \leq n\right) = 1 - \mathbb{P}\left(\bigcup_{i=1}^{L} E_i^c\right)$$
$$\leq 1 - \sum_{i=1}^{L}\left[\frac{\theta_i \mathbb{P}(E_i^c)^2}{\sum_{j=1}^{L}\mathbb{P}\left(E_i^c \cap E_j^c\right) + (1-\theta_i)\mathbb{P}(E_i^c)} + \frac{(1-\theta_i)\mathbb{P}(E_i^c)^2}{\sum_{j=1}^{L}\mathbb{P}\left(E_i^c \cap E_j^c\right) - \theta_i\mathbb{P}(E_i^c)}\right], \tag{1.57}$$

where

$$\theta_i = \frac{\sum_{j\neq i}\mathbb{P}\left(E_i^c \cap E_j^c\right)}{\mathbb{P}(E_i^c)} - \left\lfloor \frac{\sum_{j\neq i}\mathbb{P}\left(E_i^c \cap E_j^c\right)}{\mathbb{P}(E_i^c)} \right\rfloor \tag{1.58}$$

and $\lfloor x \rfloor$ is the integer part of x.
[Kuai et al., 2000] showed that the bound in Eq.(1.57) is tighter than the one in Eq.(1.54) of [Dawson and Sankoff, 1967] and involves basically the same computational complexity.
As before, we have

$$\mathbb{P}(E_i^c) = \begin{cases} 1 - Q_{m_1}(2m_1), \text{ if } i \neq L \\ 1 - Q_{m_1}(m_1 + l), \text{ if } i = L \end{cases} \tag{1.59}$$

and

$$\sum_{j=1}^{L}\mathbb{P}\left(E_i^c \cap E_j^c\right) = \begin{cases} (L-3)\left[Q_{m_1}(2m_1)\right]^2 + Q_{m_1}(2m_1)\left[1 + Q_{m_1}(m_1+l)\right] \\ +Q_{m_1}(3m_1), \text{ for } i = 1 \\ \\ (L-4)\left[Q_{m_1}(2m_1)\right]^2 + Q_{m_1}(2m_1)\left[1 + Q_{m_1}(m_1+l)\right] \\ +2Q_{m_1}(3m_1), \text{ for } 2 \leq i \leq L-1 \\ \\ (L-2)Q_{m_1}(2m_1)Q_{m_1}(m_1+l) + Q_{m_1}(m_1+l) \\ +Q_{m_1}(2m_1+l), \text{ for } i = L. \end{cases} \tag{1.60}$$

All the unknown quantities that appear in Eqs.(1.55), (1.56), (1.59) and (1.60) can be evaluated via [Karwe and Naus, 1997] algorithm, for discrete variables, or by simulation.

1.2 Two dimensional scan statistics

The two dimensional discrete scan statistics was introduced in [Chen and Glaz, 1996] and extends, in a natural way, the one dimensional case.

Let T_1, T_2 be positive integers, $\mathcal{R} = [0, T_1] \times [0, T_2]$ be a rectangular region and $\{X_{i,j} \mid 1 \leq i \leq T_1, 1 \leq j \leq T_2\}$ be a family of independent and identically distributed random variables. In many applications, the practitioners have at their disposal only the counts of the observed events of interest in smaller subregions within the studied region. In such situations, the random variables X_{ij} take non-negative integer values and one can interpret them as representing the number of events observed in the elementary square sub-region $[i-1, i] \times [j-1, j]$.
Let m_1, m_2 be positive integers such that $2 \leq m_1 \leq T_1$, $2 \leq m_2 \leq T_2$. For $1 \leq i_1 \leq T_1 - m_1 + 1$, $1 \leq i_2 \leq T_2 - m_2 + 1$ define

$$Y_{i_1,i_2} = Y_{i_1,i_2}(m_1, m_2) = \sum_{i=i_1}^{i_1+m_1-1} \sum_{j=i_2}^{i_2+m_2-1} X_{i,j} \tag{1.61}$$

to be the random variables which counts the number of the observed events in the rectangular region $\mathcal{R}(i_1, i_2) = [i_1 - 1, i_1 + m_1 - 1] \times [i_2 - 1, i_2 + m_2 - 1]$, comprised of $m_1 m_2$ adjacent elementary sub-regions.
The *two dimensional discrete scan statistic* is defined as the largest number of events in any rectangular scanning window $\mathcal{R}(i_1, i_2)$, within the rectangular region $\mathcal{R}$, i.e.

$$S_{m_1,m_2}(T_1, T_2) = \max_{\substack{1 \leq i_1 \leq T_1 - m_1 + 1 \\ 1 \leq i_2 \leq T_2 - m_2 + 1}} Y_{i_1,i_2}. \tag{1.62}$$

We denote the distribution of two dimensional scan statistic over the region $[0, T_1] \times [0, T_2]$ by

$$Q_{m_1,m_2}(T_1, T_2) = \mathbb{P}\left(S_{m_1,m_2}(T_1, T_2) \leq n\right).$$

When the parameters m_1, m_2 and n are clearly understood, we abbreviate the notation to $Q(T_1, T_2)$.
The aim of this section is to review some of the existing formulas used in the estimation of $Q_{m_1,m_2}(T_1, T_2)$. We will consider the particular cases of i.i.d. Bernoulli, binomial and Poisson observations. For an overview of the methods and the potential application of the two dimensional scan statistics one can refer to the monographs of [Glaz et al., 2001, Chapter 16] and more recently the one of [Glaz et al., 2009, Chapter 6].

1.2.1 Approximations

In this section, we present a series of product-type approximations for the distribution of two dimensional scan statistics. These approximations are accurate and can be employed for all parameters T_1, T_2, m_1, m_2 and n. Although many of the formulas that appear bellow were given in the particular situation when $T_1 = T_2$ and $m_1 = m_2$, we include them here in their general form.
Consider the case when $X_{i,j}$ are i.i.d. $0-1$ Bernoulli random variables of parameter p. In [Boutsikas and Koutras, 2000], the authors derived, using the *Markov*

Chain Imbedding Technique, the following approximation for the distribution of two dimensional scan statistics:

$$\mathbb{P}\left(S_{m_1,m_2}(T_1,T_2)\leq n\right)\approx \frac{Q(m_1,m_2)^{(T_1-m_1-1)(T_2-m_2-1)}}{Q(m_1,m_2+1)^{(T_1-m_1-1)(T_2-m_2)}} \times \frac{Q(m_1+1,m_2+1)^{(T_1-m_1)(T_2-m_2)}}{Q(m_1+1,m_2)^{(T_1-m_1)(T_2-m_2-1)}}. \tag{1.63}$$

We employ the same notation as in Section 1.1.1 for the *pmf* and *cdf* of the binomial with parameters k, m, p, namely

$$b(s;n,p)=\binom{n}{s}p^s(1-p)^{n-s}$$
$$F(s;n,p)=\sum_{i=0}^{s}b(i;n,p).$$

The quantities that appear in the approximation given by Eq.(1.63) can be computed via

$$Q(m_1,m_2)=F\left(n;m_1m_2,p\right), \tag{1.64}$$
$$Q(m_1+1,m_2)=\sum_{s=0}^{n}F^2\left(n-s;m_2,p\right)b\left(s;(m_1-1)m_2,p\right), \tag{1.65}$$
$$\begin{aligned}Q(m_1+1,m_2+1)=&\sum_{s_1,s_2=0}^{n}\sum_{t_1,t_2=0}^{n}\sum_{i_1,i_2,i_3,i_4=0}^{1}b(s_1;m_1-1,p)b(s_2;m_1-1,p)\\&\times b(t_1;m_2-1,p)b(t_2;m_2-1,p)p^{\sum i_j}(1-p)^{4-\sum i_j}\\&\times F(u;(m_1-1)(m_2-1),p),\end{aligned} \tag{1.66}$$

where in the last relation u is given by

$$u=\min\left\{n-s_1-t_1-i_1,n-s_2-t_1-i_2,n-s_1-t_2-i_3,n-s_2-t_2-i_4\right\}.$$

In the case of independent and identically distributed nonnegative integer valued observations, based on [Boutsikas and Koutras, 2000] approach and a Markov type approximation (see [Glaz et al., 2001, Section 16.1.6] for the Bernoulli case), [Chen and Glaz, 2009] proposed the following product-type approximation:

$$\mathbb{P}\left(S_{m_1,m_2}(T_1,T_2)\leq n\right)\approx \frac{Q(m_1+1,m_2+1)^{(T_1-m_1)(T_2-m_2)}}{Q(m_1+1,m_2)^{(T_1-m_1)(T_2-m_2-1)}} \times \frac{Q(m_1,2m_2-1)^{(T_1-m_1-1)(T_2-2m_2)}}{Q(m_1,2m_2)^{(T_1-m_1-1)(T_2-2m_2+1)}}. \tag{1.67}$$

In the framework of binomial and Poisson model for the underlying random field, the probabilities $Q(m_1,2m_2-1)$ and $Q(m_1,2m_2)$ can be evaluated by adapting the algorithm developed by [Karwe and Naus, 1997] (see Section A.1). The other two

unknown quantities, $Q(m_1+1,m_2)$ and $Q(m_1+1,m_2+1)$, can be computed via a conditioning argument as the one presented in [Guerriero et al., 2009] for the Poisson distribution. For completeness we have included in Section A.2.1 of Appendix A these formulas.

Remark 1.2.1. *In [Chen and Glaz, 1996] and [Glaz et al., 2001, Section 16.1.4], the authors included two Poisson type approximations and a compound Poisson approximation. Their simulation study showed that the most accurate estimate, between the four compared, was the product type approximation.*
We should mention that despite the fact that these approximations are accurate, none of them give an order of this accuracy, that is there are no error bounds corresponding to these formulas. Expressing the scan statistics $S_{m_1,m_2}(T_1,T_2)$ as the maximum of a 1-dependent sequence of random variables, [Haiman and Preda, 2006] derived an accurate approximation formula as well as associated error bounds. We present these results in a larger context in Chapter 3.

1.2.2 Bounds

Bounds for the distribution of the two dimensional discrete scan statistics can be found only for particular situations. In [Boutsikas and Koutras, 2003], based on specific techniques used in reliability theory, the authors developed a series of bounds for the special case of Bernoulli observations. Their results are summarized in the following. If $X_{i,j}$ are i.i.d. Bernoulli random variables of parameter p, then

a) Lower bound

$$Q(T_1,T_2)\geq (1-Q_1)^{(T_1-m_1)(T_2-m2)}\,(1-Q_2)^{T_1-m_1}\,(1-Q_3)^{T_2-m_2}\,(1-Q_4) \tag{1.68}$$

b) Upper bound

$$\begin{aligned}Q(T_1,T_2)\leq{}&(1-A_1)\left[1-q^{(m_1-1)(3m_2-2)+(2m_1-1)(m_2-1)}A_1\right]^{(T_1-m_1-1)(T_2-m_2-1)}\\
&\times\left[1-q^{m_1(m_2-1)}A_1\right]^{T_2-m_2-1}\left[1-q^{(m_1-1)(2m_2-1)+(m_1-1)(m_2-1)}A_1\right]^{T_1-m_1-1}\\
&\times\left[1-q^{(m_1-1)(2m_2-1)+m_1(m_2-1)}A_2\right]^{T_1-m_1}\left[1-q^{(m_1-1)(2m_2-1)+m_1(m_2-1)}A_4\right]\\
&\times\left[1-q^{(m_1-1)(3m_2-2)+m_1(m_2-1)+(m_1-1)(m_2-1)}A_3\right]^{T_2-m_2},\end{aligned} \tag{1.69}$$

where $q=1-p$ and

$$\begin{aligned}A_1&=F^c_{k+1,m_1m_2}-q^{m_2}F^c_{k+1,(m_1-1)m_2}-q^{m_1}F^c_{k+1,m_1(m_2-1)} &(1.70)\\
&\quad+q^{m_1+m_2-1}F^c_{k+1,(m_1-1)(m_2-1)}, &(1.71)\\
A_2&=F^c_{k+1,m_1m_2}-q^{m_2}F^c_{k+1,(m_1-1)m_2}, &(1.72)\\
A_3&=F^c_{k+1,m_1m_2}-q^{m_1}F^c_{k+1,m_1(m_2-1)}, &(1.73)\\
A_4&=F^c_{k+1,m_1m_2}, &(1.74)\\
F^c_{i,m}&=1-F(i-1;m,p).\end{aligned}$$

We should note that similar bounds were obtained by [Akiba and Yamamoto, 2005]. In the general case, [Chen and Glaz, 1996] proposed a Bonferroni type inequality for the lower bound of the distribution of the scan statistics. As their simulations showed, these bounds are not as sharp as one expect. Using [Hoover, 1990] Bonferroni type inequality of order $r \geq 3$

$$\begin{aligned}\mathbb{P}\left(\bigcup_{i_1=1}^{T_1-m_1+1}\bigcup_{i_2=1}^{T_2-m_2+1} A_{i_1,i_2}\right) \leq & \sum_{i_1=1}^{T_1-m_1+1}\sum_{i_2=1}^{T_2-m_2+1}\mathbb{P}\left(A_{i_1,i_2}\right) \\ & - \sum_{i_1=1}^{T_1-m_1+1}\sum_{i_2=1}^{T_2-m_2}\mathbb{P}\left(A_{i_1,i_2}\cap A_{i_1,i_2+1}\right) - \sum_{i_1=1}^{T_1-m_1}\mathbb{P}\left(A_{i_1,1}\cap A_{i_1+1,1}\right) \\ & - \sum_{i_1=1}^{T_1-m_1+1}\sum_{l=2}^{r-1}\sum_{i_2=1}^{T_2-m_2+1-l}\mathbb{P}\left(A_{i_1,i_2}\cap A^c_{i_1,i_2+1}\ldots A^c_{i_1,i_2+l-1}\cap A_{i_1,i_2+l}\right)\end{aligned} \tag{1.75}$$

with $A_{i_1,i_2} = \{Y_{i_1,i_2} > n\}$ and $r = 4$, we have

$$\begin{aligned}\mathbb{P}\left(S_{m_1,m_2}(T_1,T_2)\leq n\right) \geq & \left(T_1-m_1\right)\left[Q(m_1+1,m_2)-2Q(m_1,m_2)\right] \\ & -(T_1-m_1+1)(T_2-m_2-3)Q(m_1,m_2+2) \\ & +(T_1-m_1+1)(T_2-m_2-2)Q(m_1,m_2+3).\end{aligned} \tag{1.76}$$

The unknown probabilities in Eq.(1.76): $Q(m_1,m_2)$, $Q(m_1+1,m_2)$, $Q(m_1,m_2+2)$, $Q(m_1,m_2+3)$ are evaluated via a conditional argument (see Section A.2.1) or an adaptation of the algorithm of [Karwe and Naus, 1997].
For the upper bound we propose to adapt the inequality of [Kuai et al., 2000] to the two dimensional framework. Using the events A_{i_1,i_2}, defined above, we can write

$$\begin{aligned}\mathbb{P}\left(S_{m_1,m_2}(T_1,T_2)\leq n\right) &= 1-\mathbb{P}\left(\bigcup_{i_1=1}^{T_1-m_1+1}\bigcup_{i_2=1}^{T_2-m_2+1}A_{i_1,i_2}\right) \\ &\leq 1-\sum_{i_1=1}^{T_1-m_1+1}\sum_{i_2=1}^{T_2-m_2+1}\left[\frac{\theta_{i_1,i_2}\mathbb{P}(A_{i_1,i_2})^2}{\Sigma(i_1,i_2)+(1-\theta_{i_1,i_2})\mathbb{P}(A_{i_1,i_2})}\right. \\ &\quad \left.+\frac{(1-\theta_{i_1,i_2})\mathbb{P}(A_{i_1,i_2})^2}{\Sigma(i_1,i_2)-\theta_{i_1,i_2}\mathbb{P}(A_{i_1,i_2})}\right],\end{aligned} \tag{1.77}$$

where

$$\Sigma(i_1,i_2) = \sum_{j_1=1}^{T_1-m_1+1}\sum_{j_2=1}^{T_2-m_2+1}\mathbb{P}\left(A_{i_1,i_2}\cap A_{j_1,j_2}\right) \tag{1.78}$$

and θ_{i_1,i_2} can be computed from

$$\theta_{i_1,i_2} = \frac{\Sigma(i_1,i_2)}{\mathbb{P}(A_{i_1,i_2})} - \left\lfloor\frac{\Sigma(i_1,i_2)}{\mathbb{P}(A_{i_1,i_2})}\right\rfloor. \tag{1.79}$$

We observe that if $|i_1 - j_1| \geq m_1$ or $|i_2 - j_2| \geq m_2$, then the events A_{i_1,i_2} and A_{j_1,j_2} are independent, thus

$$\mathbb{P}\left(A_{i_1,i_2} \cap A_{j_1,j_2}\right) = \left[1 - Q(m_1, m_2)\right]^2 . \tag{1.80}$$

On the other hand, if $|i_1 - j_1| < m_1$ and $|i_2 - j_2| < m_2$, then

$$\mathbb{P}\left(A_{i_1,i_2} \cap A_{j_1,j_2}\right) = 1 - 2Q(m_1, m_2) + \mathbb{P}\left(Y_{i_1,i_2} \leq n, Y_{j_1,j_2} \leq n\right). \tag{1.81}$$

The last term in Eq.(1.81) can be evaluated via a conditioning argument. For example, in the case of a binomial model with parameters r and p, denoting with Z the random variable

$$Z = \sum_{s=i_1 \vee j_1}^{(i_1+m_1-1)\wedge(j_1+m_1-1)} \sum_{t=i_2 \vee j_2}^{(i_2+m_2-1)\wedge(j_2+m_2-1)} X_{s,t}, \tag{1.82}$$

we have, due to the independence of $Y_{i_1,i_2} - Z$ and $Y_{j_1,j_2} - Z$, that

$$\mathbb{P}\left(Y_{i_1,i_2} \leq n, Y_{j_1,j_2} \leq n\right) = \sum_{k=0}^{n} \mathbb{P}(Z = k)\mathbb{P}(Y_{i_1,i_2} - Z \leq n - k)^2. \tag{1.83}$$

In Eq.(1.83), the random variables Z, $Y_{i_1,i_2} - Z$ are binomially distributed with parameters $r(m_1 - |i_1 - j_1|)(m_2 - |i_2 - j_2|)$, p and $r|i_1 - j_1||i_2 - j_2|$ and p, respectively. Hence, to calculate the upper bound in Eq.(1.77), it is enough to pre compute the $m_1 \times m_2$ matrix with the entries given by the probabilities $\mathbb{P}\left(A_{i_1,i_2} \cap A_{j_1,j_2}\right)$, found above, for $|i_1 - j_1| < m_1$ and $|i_2 - j_2| < m_2$.

Remark 1.2.2. *A different upper bound can be obtained from Eq.(1.77) if one consider the events E_{i_1,i_2} given by*

$$E_{i_1,i_2} = \left\{ \max_{\substack{(i_1-1)m_1+1 \leq s_1 \leq i_1 m_1+1 \\ (i_2-1)m_2+1 \leq s_2 \leq i_2 m_2+1}} Y_{s_1,s_2} \leq n \right\} \tag{1.84}$$

instead of the events A_{i_1,i_2}. This approach generalizes the upper bound in the one dimensional case presented in Section 1.1.3. Details about this alternative margin are given in Section A.2.2 of Appendix A.

1.3 Three dimensional scan statistics

In this section, we extend the notion of the two dimensional discrete scan statistics defined in Section 1.2, to the three dimensional case. The three dimensional discrete scan statistics was introduced in [Guerriero et al., 2010a], where the authors derived a product-type approximation and three Poisson approximations for the $0 - 1$ Bernoulli model. As the authors mention in the cited article, the most accurate estimate within the four is the product-type approximation. Based on

this observation, we include bellow the formula for the product type estimate in a somewhat general framework. Detailed expressions of the unknown quantities that appear in this formula are presented in Section A.3 of Appendix A.
Let T_1, T_2, T_3 be positive integers, $\mathcal{R} = [0, T_1] \times [0, T_2] \times [0, T_3]$ be a rectangular region and $\{X_{i,j,k} | 1 \leq i \leq T_1, 1 \leq j \leq T_2, 1 \leq k \leq T_3\}$ be a family of independent and identically distributed integer valued random variables from a specified distribution. For each $l \in \{1, 2, 3\}$, consider the positive integers m_l such that $1 \leq m_l \leq T_l$, and define the random variables

$$Y_{i_1,i_2,i_3} = \sum_{i=i_1}^{i_1+m_1-1} \sum_{j=i_2}^{i_2+m_2-1} \sum_{k=i_3}^{i_3+m_3-1} X_{i,j,k}, \quad 1 \leq i_l \leq T_l - m_l + 1. \tag{1.85}$$

If the random variables $X_{i,j,k}$ are interpreted as the number of events observed in the elementary square subregion $[i-1, i] \times [j-1, j] \times [k-1, k]$, then Y_{i_1,i_2,i_3} counts the events observed in the rectangular region

$$\mathcal{R}(i_1, i_2, i_3) = [i_1 - 1, i_1 + m_1 - 1] \times [i_2 - 1, i_2 + m_2 - 1] \times [i_3 - 1, i_3 + m_3 - 1],$$

comprised of $m_1 m_2 m_3$ adjacent elementary square subregions and with the south-west corner at the point $(i_1 - 1, i_2 - 1, i_3 - 1)$.
The three dimensional discrete scan statistic is defined as the maximum number of events in any rectangle $\mathcal{R}(i_1, i_2, i_3)$ within the region $\mathcal{R}$,

$$S_{m_1,m_2,m_3}(T_1, T_2, T_3) = \max_{\substack{1 \leq i_j \leq T_j - m_j + 1 \\ j \in \{1,2,3\}}} Y_{i_1,i_2,i_3}. \tag{1.86}$$

The distribution of the scan statistic,

$$Q_{m_1,m_2,m_3}(T_1, T_2, T_3) = \mathbb{P}\left(S_{m_1,m_2,m_3}(T_1, T_2, T_3) \leq n\right),$$

was successfully used in astronomy ([Darling and Waterman, 1986]), image analysis ([Naiman and Priebe, 2001]), reliability theory ([Boutsikas and Koutras, 2000]) and many other domains. An overview of the potential application of space-time scan statistics can be found in the monograph [Glaz et al., 2009, Chapter 6].
From a statistical point of view, the scan statistic $S_{m_1,m_2,m_3}(T_1, T_2, T_3)$ is used for testing the null hypothesis of randomness that X_{ijk}'s are independent and identically distributed according to some specified distribution. Under the alternative hypothesis there exists one cluster location where the X_{ijk}'s have a larger mean than outside the cluster. As an example, in the Poisson model, the null hypothesis, H_0, assumes that X_{ijk}'s are i.i.d. with $X_{ijk} \sim Pois(\lambda)$ whereas the alternative hypothesis of clustering, H_1, assumes the existence of a rectangular subregion $\mathcal{R}(i_0, j_0, k_0)$ such that for any $i_0 \leq i \leq i_0 + m_1 - 1$, $j_0 \leq j \leq j_0 + m_2 - 1$ and $k_0 \leq k \leq k_0 + m_3 - 1$, X_{ijk} are i.i.d. Poisson random variables with parameter $\lambda' > \lambda$. Outside the region $\mathcal{R}(i_0, j_0, k_0)$, X_{ijk} are i.i.d. distributed according to the distribution specified by the null hypothesis. The generalized likelihood ratio test rejects H_0 in favor of the local change alternative H_1, whenever $S_{m_1,m_2,m_3}(T_1, T_2, T_3)$ exceeds the threshold τ determined from $\mathbb{P}\left(S_{m_1,m_2,m_3}(T_1, T_2, T_3) \geq \tau | H_0\right) = \alpha$ and where α represents the significance level of the testing procedure ([Glaz et al., 2001, Chapter 13]).

1.3.1 Product-type approximation

Consider that $X_{i,j,k}$ are independent and identically distributed nonnegative integer valued random variables. When the parameters involved in the distribution of the three dimensional scan statistics are clearly understood we abbreviate the notation to $Q(T_1, T_2, T_3)$.
Following the approach proposed in [Guerriero et al., 2010a] for the i.i.d. $0-1$ Bernoulli model, we obtain the following product-type estimate:

$$\begin{aligned}\mathbb{P}\left(S_{m_1,m_2,m_3}(T_1,T_2,T_3) \leq n\right) \approx & \frac{Q(m_1+1,m_2+1,m_3+1)^{(T_1-m_1)(T_2-m_2)(T_3-m_3)}}{Q(m_1,m_2,m_3)^{(T_1-m_1-1)(T_2-m_2-1)(T_3-m_3-1)}} \\ & \times \frac{Q(m_1+1,m_2,m_3)^{(T_1-m_1-1)(T_2-m_2)(T_3-m_3)}}{Q(m_1,m_2+1,m_3+1)^{(T_1-m_1)(T_2-m_2-1)(T_3-m_3-1)}} \\ & \times \frac{Q(m_1,m_2+1,m_3)^{(T_1-m_1)(T_2-m_2-1)(T_3-m_3)}}{Q(m_1+1,m_2,m_3+1)^{(T_1-m_1-1)(T_2-m_2)(T_3-m_3-1)}} \\ & \times \frac{Q(m_1,m_2,m_3+1)^{(T_1-m_1)(T_2-m_2)(T_3-m_3-1)}}{Q(m_1+1,m_2+1,m_3)^{(T_1-m_1-1)(T_2-m_2-1)(T_3-m_3)}}.\end{aligned} \tag{1.87}$$

Particularizing, for $T_1 = T_2 = T_3 = N$ and $m_1 = m_2 = m_3 = m$ we get the same approximation formula as in [Guerriero et al., 2010a, Eq. 6]. Explicit expressions for the unknown quantities in Eq.(1.87) are given in Section A.3 of Appendix A for the binomial and Poisson model.

Remark 1.3.1. *As far as we know, the only result on bounds for the distribution of the three dimensional scan statistics is the one of [Akiba and Yamamoto, 2004]. The authors, using techniques from reliability theory, obtained closed form lower and upper bounds for the Bernoulli model. Their formulas are complex and, even for moderate scanning window sizes, requires excessive computational time. We should mention that alternative bounds can be obtained extending the approach presented in Section 1.2.2 to the three dimensional setting.*

Chapter 2

Extremes of 1-dependent stationary sequences of random variables

In this chapter, we present some results concerning the approximation of the distribution of the maximum and minimum of the first n terms of a 1-dependent stationary sequence of random variables. These approximations extend the original ones developed in [Haiman, 1999], both in terms of range of applicability and in sharpness of the error bounds. We begin in Section 2.1 with some definitions and remarks concerning the m-dependent sequences of random variables. Next, we give the formulation of the problem and the intuition behind the proposed method. The description of the main results and some numerical aspects are presented in Section 2.2. We conclude the chapter with the proofs of the results in Section 2.3. Parts of the work considered in this chapter appeared in [Amărioarei, 2012] and make the object of an article submitted for publication.

Contents

2.1 Introduction

We consider the problem of approximating the distribution of the extremes (maxima and minima) of 1-dependent stationary sequences of random variables.

2.1.1 Definitions and notations

We say that a sequence of random variables is m-dependent if observations separated by m units are stochastically independent and is stationary (in the strong sense) whenever its finite dimensional distributions are invariant under time shifts. To be more precise we have the following definitions:

Definition 2.1.1. *The sequence $(W_k)_{k\geq 1}$ of random variables is m-dependent, $m \geq 1$, if for any $h \geq 1$ the σ-fields generated by $\{W_1, \dots, W_h\}$ and $\{W_{h+m+1}, \dots\}$ are independent.*

Definition 2.1.2. *The sequence $(W_k)_{k\geq 1}$ of random variables is stationary (in the strong sense) if for all $n \geq 1$, for all $h \geq 0$ and for all $t_1, \dots, t_n$ the families*

$$\{W_{t_1}, W_{t_2}, \dots, W_{t_n}\} \quad \text{and} \quad \{W_{t_1+h}, W_{t_2+h}, \dots, W_{t_n+h}\}$$

have the same joint distribution.

A large class of m-dependent sequences is constructed based on the *block-factor* terminology. The following definition describes the notion of *l block-factor* (see also [Burton et al., 1993]):

Definition 2.1.3. *The sequence $(W_k)_{k\geq 1}$ of random variables with state space S_W is said to be l block-factor of the sequence $(Y_k)_{k\geq 1}$ with state space S_Y if there is a measurable function $f : S_Y^l \to S_W$ such that*

$$W_k = f\left(Y_k, Y_{k+1}, \dots, Y_{k+l-1}\right)$$

for all k.

2.1.2 Remarks about m-dependent sequences and block-factors

The study of m-dependent sequences can be regarded as a different model for dependence besides the well known Markov model. Even if the later has been investigated thoroughly for a long time, not so much can be said about the former model.
Maybe the most common examples of m-dependent sequences are obtained from $(m+1)$ block-factors of independent and identically distributed sequences of random variables. In [Ibragimov and Linnik, 1971], the authors conjectured that the converse is not true; they affirmed (without giving an example) that there are m-dependent sequences which are not $(m+1)$ block-factor of any i.i.d. sequence. Progress in this direction was made by [Aaronson et al., 1989], who showed that there exist two valued 1-dependent processes which cannot be expressed as a 2 block-factor of an i.i.d. sequence. In [Aaronson and Gilat D., 1992] it is shown that

a stationary 1-dependent Markov chain with no more than four states can be represented by a 2 block-factor. Their result is sharp in the sense that there is an example of such a sequence with five states which is not a 2 block-factor. In particular, [de Valk, 1988] proved that a two valued Markov chain becomes an i.i.d. sequence if in addition it is stationary and 1-dependent. This result was extended by [Matus, 1998, Lemma2] who gave a characterization of binary sequences which are Markov of order n and m-dependent. The author also presented a necessary and sufficient condition for the existence of a m-dependent Markov chain with a finite state space (see [Matus, 1998, Proposition 1]). In [Burton et al., 1993], the authors gave an example of a four state 1-dependent process which is not a l block-factor for any $l \geq 2$, thus confirming the conjecture. More details about 1-dependent processes can be found in [de Valk, 1988] and [Goulet, 1992].
To ask whether a sequence of m-dependent random variables is a $(m+1)$ block-factor or not, seems to be a natural question. From the best of our knowledge there is no general result concerning this problem. Nevertheless, some partial answers can be found in the literature. In [de Valk and Ruschendorf, 1993] was obtained a regression representation for a particular class of m-dependent sequences. In the same paper, the authors presented a constructive method to check if such a sequence can be viewed as a monotone $(m+1)$ block-factor. In [Broman, 2005] it is shown that 1-dependent trigonometric determinantal processes are 2 block-factors. Another class of processes that can be expressed as a $(m+1)$ block-factor is represented by stationary m-dependent Gaussian sequences. The following proposition justifies this affirmation:

Proposition 2.1.4. *Let $(W_k)_{k\in\mathbb{Z}}$ be a stationary m-dependent Gaussian sequence of random variables such that $\mathbb{E}W_0 = \mu$. Then there exists a sequence $(a_k)_{k=0}^m$ such that*

$$(W_k)_{k\in\mathbb{Z}} \stackrel{d}{=} \left(\mu + \sum_{i=0}^{m} a_i \eta_{k-i}\right)_{k\in\mathbb{Z}}, \tag{2.1}$$

where $(\eta_j)_{j\in\mathbb{Z}}$ are i.i.d. standard normal random variables.

Generalizations of the foregoing result, concerning m-dependent stationary infinite divisible sequences, can be found in [Harrelson and Houdre, 2003].

2.1.3 Formulation of the problem and discussion

The problem that we treat in this chapter can be formulated as follows. Let $(X_n)_{n\geq 1}$ be a strictly stationary sequence of 1-dependent random variables with marginal distribution function $F(x) = \mathbb{P}(X_1 \leq x)$ and take x such that

$$\inf\{u|F(u) > 0\} < x < \sup\{u|F(u) < 1\}.$$

For each $n \geq 1$, we define the sequences

$$p_n = p_n(x) = \mathbb{P}(\min\{X_1, X_2, \ldots, X_n\} > x), \tag{2.2}$$
$$q_n = q_n(x) = \mathbb{P}(\max\{X_1, X_2, \ldots, X_n\} \leq x). \tag{2.3}$$

Our aim in this work is to find good estimates for p_n and q_n together with the corresponding error margins. Asymptotic results for extremes of m-dependent sequences of random variables were obtained by many authors among which we can mention [Watson, 1954], [Newell, 1964], [Galambos, 1972] and [Flak and Schmid, 1995].
The approach we are using to attain these estimates differs from the ones presented in the cited papers and it has its origin in the paper of [Haiman, 1999]. To get the intuition behind the method presented in this work, consider $q_{-1} = q_0 = 1$ and take

$$D(z) = D_x(z) = \sum_{k=0}^{\infty} q_{k-1} z^k \tag{2.4}$$

the generating function of the sequence $(q_k)_{k\geq 0}$. We observe that this generating function depends on x, since the sequence $(q_k)_{k\geq 0}$ does. Suppose that, for example, the function $D(z)$ has the following expression:

$$D(z) = 1 + \frac{Az}{1 - \frac{z}{\lambda}},\ A \neq 0. \tag{2.5}$$

Then it is easy to see that

$$\sum_{k=0}^{\infty} q_k z^{k+1} = \sum_{k=0}^{\infty} \frac{A}{\lambda^k} z^{k+1} \tag{2.6}$$

and therefore $q_n = \frac{A}{\lambda^n}$. The following example shows that there are situations where the assumption made in Eq.(2.5) is valid.

Example 2.1.5. *Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d. random variables. In particular, the sequence is 1-dependent. Then one can easily show that*

$$D(z) = 1 + \frac{z}{1 - (1 - p_1)z}$$

which gives $\lambda = \frac{1}{1-p_1}$ and $q_n = (1 - p_1)^n = \frac{1}{\lambda^n}$.

Of course that the situation presented in Example 2.1.5 is an extreme one and such an assumption is not feasible in the general context. Perhaps, the next best thing that we can hope for is to have an asymptotic relation in the neighborhood of the singularity λ, namely

$$D(z) \sim 1 + \frac{Az}{1 - \frac{z}{\lambda}}. \tag{2.7}$$

In which case, one can wish to obtain a relation of the form

$$q_n \sim \frac{A}{\lambda^n}. \tag{2.8}$$

The above observations lead to the problem of showing that the generating function $D(z)$ has a *nice* singularity. It is obvious that the problem of studying the singularities of D is related to the analysis of the zeros of the function $\frac{1}{D}$. Now, if we consider the generating function associated to the sequence $\{(-1)^k p_{k-1}\}_{k\geq 0}$,

$$C(z) = C_x(z) = 1 + \sum_{k=1}^{\infty} (-1)^k p_{k-1} z^k \tag{2.9}$$

where $p_{-1} = p_0 = 1$, one can prove that (see Lemma 2.3.5)

$$D(z) = \frac{1}{C(z)}. \tag{2.10}$$

This last relation gives us the means to analyze the singularity of D by focusing our attention on the zeros of the generating function $C(z)$. As it turns out, to analyze the zeros of the function C is much easier because of two aspects: an elementary upper bound for the probabilities p_n (given in Eq.(2.25)) and the fact that C is an alternating series. We will finish this section with an example that illustrates the above ideas:

Example 2.1.6. *Let $(U_n)_{n\geq 1}$ be a sequence of i.i.d. $0-1$ Bernoulli random variables with $\mathbb{P}(U_n = 1) = p$. We define for each $n \geq 1$ the random variables $X_n = U_n U_{n+1}$. It is clear that the sequence $(X_n)_{n\geq 1}$ is stationary and 1-dependent, since is defined as a 2 block-factor of an i.i.d. sequence. Clearly*

$$\begin{aligned} p_n &= \mathbb{P}(X_1 = 1, X_2 = 1, \ldots, X_n = 1) \\ &= \mathbb{P}(U_1 = 1, U_2 = 1, \ldots, U_{n+1} = 1) = p^{n+1}, \end{aligned}$$

so the generating function in Eq.(2.9) becomes

$$C(z) = -z + pz + \frac{1}{1+pz}.$$

Solving the equation $C(z) = 0$ we obtain the value of the zero

$$\lambda = \frac{1}{2p}\left(-1 + \sqrt{1 + \frac{4p}{1-p}}\right).$$

Notice that λ belongs to an interval of the form $(1, 1+u)$, which will be in accordance with the result presented in Theorem 2.2.3. To get the value of q_n, we first notice that according to Lemma 2.3.5 we have

$$q_n = \sum_{k=0}^{n} (-1)^{n-k} p_{n-k} q_{k-1}, \quad n \geq 0, \quad p_0 = q_{-1} = q_0 = 1.$$

By writing $b_n = q_n p^{-n}$ and using the fact that $p_n = p^{n+1}$, we obtain the following second order recurrence:

$$b_{n+1} = \frac{1-p}{p}\left(b_n + b_{n-1}\right), \quad b_0 = 1, b_1 = \frac{1-p^2}{p}$$

which, after solving for q_n, leads to the solution

$$q_n = c\frac{1}{\lambda^n} + (1-c)\frac{1}{\eta^n},$$

where

$$\eta = \frac{-1}{2p}\left(1 + \sqrt{1 + \frac{4p}{1-p}}\right)$$

and $c = \frac{\eta+1}{\lambda+\eta}$. For p small enough, it can be seen that the value of η^{-1} decrease much faster than of λ^{-1}, so the main contribution to the value of q_n is made by the first term.

In the next section we will see that there exists a zero λ_x of C_x in an interval of the form $(1, 1+u)$ and the result in Theorem 2.2.3 will give an estimate of this zero. We will conclude the section with an approximation and the corresponding error bound for the value of q_n.

2.2 Main results

Starting from the results obtained by [Haiman, 1999] and presented in Section 2.2.1 bellow, we extend these results by enlarging their range of applicability and providing sharper error bounds. These new expressions will constitute the main tools in finding the approximation of the distribution of the scan statistics as it will be seen in the next chapters.

2.2.1 Haiman results

In [Haiman, 1999], the author presented a series of results concerning the approximation of the distribution of the maxima of a stationary 1-dependent sequence of random variables. He started by showing that if the tail probability p_1 is small enough, then the series given in Eq.(2.9) has an unique zero in the interval $(1, 1+2p_1)$ and gave explicit an estimate of its value. The next theorem illustrates this result:

Theorem 2.2.1. *For x such that $0 < p_1(x) \leq 0.025$, $C_x(z)$ has a unique zero $\lambda(x)$, of order of multiplicity 1, inside the interval $(1, 1+2p_1)$, such that*

$$|\lambda - (1 + p_1 - p_2 + p_3 - p_4 + 2p_1^2 + 3p_2^2 - 5p_1p_2)| \leq 87p_1^3. \tag{2.11}$$

The relation between the zero λ found in the foregoing result and the probability of the maximum of the first n elements of the sequence $(X_n)_{n\geq 1}$ is given by the following theorem:

Theorem 2.2.2. *We have*

$$q_1 = 1 - p_1, \qquad q_2 = 1 - 2p_1 + p_2, \qquad q_3 = 1 - 3p_1 + 2p_2 + p_1^2 - p_3$$

and for $n > 3$ if $p_1 \leq 0.025$,

$$|q_n\lambda^n - (1 - p_2 + 2p_3 - 3p_4 + p_1^2 + 6p_2^2 - 6p_1p_2)| \leq 561p_1^3. \tag{2.12}$$

These theorems were successfully employed in a series of applications: the distribution of the maximum of the increments of the Wiener process ([Haiman, 1999]), extremes of Markov sequences ([Haiman et al., 1995]), the distribution of scan statistics, both in one dimensional (see [Haiman, 2000, Haiman, 2007, Haiman and Preda, 2013]) and two dimensional case (see [Haiman and Preda, 2002, Haiman and Preda, 2006]).

2.2.2 New results

In this section, we extend the theorems presented in Section 2.2.1 but we keep their initial form. The reason behind this approach is that we want to have a reference for comparison. The results presented in subsequent will have parameterized error coefficients depending on the tail probability p_1, which will prove to be much smaller than the initial ones. The advantage of the parameterized error coefficients over the fixed ones presented above become clear when the value of p_1 decreases toward zero and lead to sharper bounds. The following statement gives a parametric form of the Theorem 2.2.1, improving both the range of $p_1(x)$, from 0.025 to 0.1, and the error coefficient:

Theorem 2.2.3. *For x such that $0 < p_1(x) \leq 0.1$, $C_x(z)$ has an unique zero $\lambda = \lambda(x)$, of order of multiplicity 1, inside an interval of the form $(1, 1 + lp_1)$, such that*

$$|\lambda - (1 + p_1 - p_2 + p_3 - p_4 + 2p_1^2 + 3p_2^2 - 5p_1p_2)| \leq K(p_1)p_1^3, \tag{2.13}$$

where $l = l(p_1) > t_2^3(p_1)$, $t_2(p_1)$ is the second root in magnitude of the equation $p_1t^3 - t + 1 = 0$ and $K(p_1)$ is given by

$$K(p_1) = \frac{\frac{11-3p_1}{(1-p_1)^2} + 2l(1+3p_1)\frac{2+3lp_1-p_1(2-lp_1)(1+lp_1)^2}{[1-p_1(1+lp_1)^2]^3}}{1 - \frac{2p_1(1+lp_1)}{[1-p_1(1+lp_1)^2]^2}}. \tag{2.14}$$

Using the stationarity and the one dependence of the sequence $(X_n)_{n\geq 1}$, we have the following:

Corollary 2.2.4. *Let λ be defined as in Theorem 2.2.3, then*

$$|\lambda - (1 + p_1 - p_2 + 2(p_1 - p_2)^2)| \leq (1 + p_1K(p_1))p_1^2. \tag{2.15}$$

The choice between the estimate of λ given in Theorem 2.2.3 and in Corollary 2.2.4 is made according to the information available about the random sequence. If one does not have enough information to compute (or simulate) the values of p_3 and p_4, then the obvious estimate is the one given in Corollary 2.2.4.
To get a better grasp of the bounds in Theorem 2.2.3 and Corollary 2.2.4, we present in Table 2.1, for selected values of p_1, the values taken by the coefficients in Eq.(2.13) and Eq.(2.15):
Notice that for $p_1 = 0.0025$ we are in the hypothesis of Theorem 2.2.2 and Theorem 2.2.3. The corresponding value for $K(0.0025)$ in Table 2.1 shows that the error bound in Eq.(2.13) is almost five times smaller then the one in Eq.(2.11). Figure 2.1 presents the behaviour of the error coefficient function $K(p_1)$ as p_1 varies between 0 and 0.1. The fixed value of the coefficient in Theorem 2.2.1 ($= 87$) is also plotted, but only between 0 and 0.025.
Before presenting the parametric extension of Theorem 2.2.2, we should stop for a little to illustrate the estimate of the zero λ in the context of Example 2.1.6.

p_1	l	$K(p_1)$	$1+p_1K(p_1)$
0.100	1.5347	38.6302	4.8630
0.050	1.1893	21.2853	2.0642
0.025	**1.0835**	**17.5663**	**1.4391**
0.010	1.0313	15.9265	1.1592

Table 2.1: Selected values for the error coefficients in Theorem 2.2.3 and Corollary 2.2.4

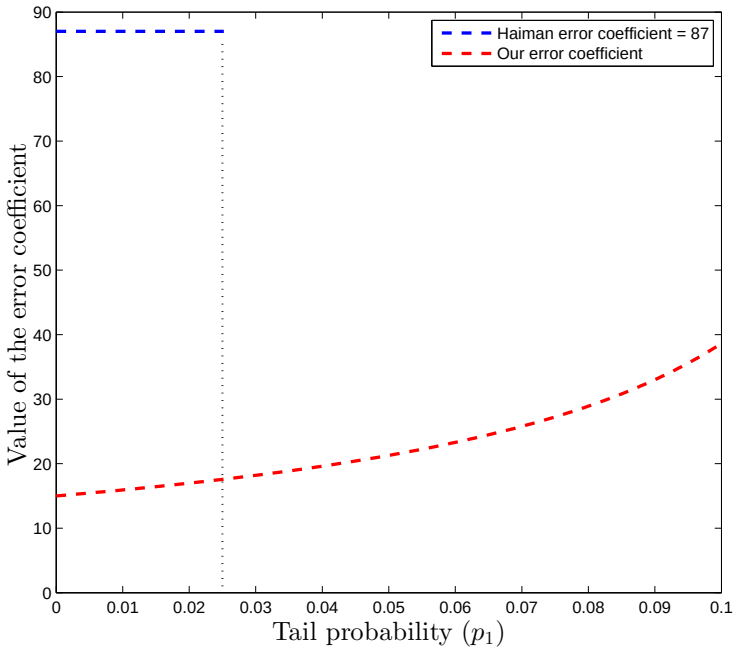


Figure 2.1: The behaviour of the function $K(p_1)$

Example 2.2.5 (Continuation of Example 2.1.6)**.** *As we have shown previously, the value of the zero λ is equal with*

$$\lambda = \frac{1}{2p}\left(-1+\sqrt{1+\frac{4p}{1-p}}\right)$$

and it belongs to an interval of the form $(1, 1+u)$. It is not hard to see that if one takes $u = tp$ with $t > \frac{p}{1-p^2}$, then $\lambda \in (1, 1+tp)$.
We will verify only the estimate in Corollary 2.2.4, as it involves less computations. First, observe that the approximation of λ in Eq.(2.15) can be rewritten as

$$\begin{aligned}\nu &= 1+p_1-p_2+2(p_1-p_2)^2\\ &= 1+p^2(1-p)+2p^4(1-p)^2,\end{aligned}$$

since $p_n = p^{n+1}$ for $n \geq 1$.

Using Taylor's formula with Lagrange form of the remainder for the function $\sqrt{1+x}$, *we have*

$$\sqrt{1+x} = 1 + \frac{1}{2}x - \frac{1}{8}x^2 + \frac{1}{16}x^3 - \frac{5}{128}x^4 + \frac{7}{256}(1+\xi)^{-9/2}x^5.$$

Substituting the above expansion in the expression of λ *for* $x = \frac{4p}{1-p}$, *we obtain*

$$\lambda = \frac{1}{1-p} - \frac{p}{(1-p)^2} + \frac{2p^2}{(1-p)^3} - \frac{5p^3}{(1-p)^4} + \frac{14p^4}{(1-p)^5}(1+\xi)^{-9/2},$$

where $\xi \in \left(0, \frac{4p}{1-p}\right)$. *Elementary calculations give us that*

$$\lambda - \nu = p^4\left[\frac{1}{(1-p)^5}(1+\xi)^{-9/2} - 2(1-p)^2 + \frac{-11+10p-5p^2+p^3}{(1-p)^4}\right].$$

We can see that $|\lambda - \nu|$ *is a multiple of* $p_1^2 = p^4$, *which is in accordance with our result. Notice that the difference between the estimate in Theorem 2.2.3 and the one in Corollary 2.2.4 is of the order* $p^4(1-p)^2$, *but the first result gives an accuracy of order* p^6, *while the second only of* p^4.

An analogue result to the one presented in Theorem 2.2.2 is given bellow:

Theorem 2.2.6. *Assume that* x *is such that* $0 < p_1(x) \leq 0.1$ *and define* $\eta = 1 + lp_1$ *with* $l = l(p_1) > t_2^3(p_1)$ *and* $t_2(p_1)$ *the second root in magnitude of the equation* $p_1t^3 - t + 1 = 0$. *If* $\lambda = \lambda(x)$ *is the zero obtained in Theorem 2.2.3, then the following relation holds:*

$$|q_n\lambda^n - (1 - p_2 + 2p_3 - 3p_4 + p_1^2 + 6p_2^2 - 6p_1p_2)| \leq \Gamma(p_1)p_1^3, \tag{2.16}$$

where $\Gamma(p_1) = 36.1 + (1-p_1)^2P(p_1) + E(p_1)$, $K(p_1)$ *is given by Eq.(2.14) and*

$$\begin{aligned} P(p_1) &= 3K(p_1)(1+p_1+3p_1^2)[1+p_1+3p_1^2+K(p_1)p_1^3] + p_1^6K^3(p_1) \\ &\quad + 9p_1(4+3p_1+3p_1^2) + 19, \end{aligned} \tag{2.17}$$

$$E(p_1) = \frac{\eta^5\left[1+(1-2p_1)\eta\right]^4\left[1+p_1(\eta-2)\right]\left[1+\eta+(1-3p_1)\eta^2\right]}{2(1-p_1\eta^2)^4\left[(1-p_1\eta^2)^2 - p_1\eta^2(1+\eta-2p_1\eta)^2\right]}. \tag{2.18}$$

As an immediate consequence of Theorem 2.2.6 we have the following corollary which involves only the values of p_1 and p_2:

Corollary 2.2.7. *In the conditions of Theorem 2.2.6 we have*

$$|q_n\lambda^n - (1-p_2)| \leq (3 + p_1\Gamma(p_1))p_1^2. \tag{2.19}$$

From Table 2.2 we can see that for $p_1 = 0.0025$, the coefficient in the error bound given by Eq.(2.16) is about three times smaller then the one from Eq.(2.12).
We continue the exposition with two theorems that answer the problem proposed in Section 2.1.3, namely finding an estimate for the maxima q_n. The first result

p_1	$\Gamma(p_1)$	$3+p_1\Gamma(p_1)$
0.100	480.696	51.0696
0.050	180.532	12.0266
0.025	**145.202**	**6.6300**
0.010	131.438	4.3143

Table 2.2: Selected values for the error coefficients in Theorem 2.2.6 and Corollary 2.2.7

presents an estimate for q_n in terms of q_1, q_2, q_3 and q_4, while for the second only the expressions of q_1 and q_2 are involved. These two statements, especially the second one, constitutes the cornerstone for the approximation developed in the following chapter. As we will see, the estimates are very sharp for those values of x for which the tail probability is small.
Combining the results obtained in Theorem 2.2.3 and Theorem 2.2.6, we get the following approximation:

Theorem 2.2.8. *Let x such that $q_1(x) \geq 1-p_1(x) \geq 0.9$. If $\Gamma(\cdot)$ and $K(\cdot)$ are the same as in Theorem 2.2.6, then*

$$\left| q_n - \frac{6(q_1-q_2)^2+4q_3-3q_4}{(1+q_1-q_2+q_3-q_4+2q_1^2+3q_2^2-5q_1q_2)^n} \right| \leq n\Delta_1(1-q_1)^3, \tag{2.20}$$

with

$$\Delta_1 = \Delta_1(q_1,n) = K(1-q_1) + \frac{\Gamma(1-q_1)}{n}. \tag{2.21}$$

In the same fashion, combining the results from Corollary 2.2.4 and Corollary 2.2.7, we get

Theorem 2.2.9. *If x is such that $q_1(x) \geq 1-p_1(x) \geq 0.9$, then*

$$\left| q_n - \frac{2q_1-q_2}{[1+q_1-q_2+2(q_1-q_2)^2]^n} \right| \leq n\Delta_2(1-q_1)^2, \tag{2.22}$$

with

$$\Delta_2 = \Delta_2(q_1,n) = 1+\frac{3}{n}+\left[K(1-q_1)+\frac{\Gamma(1-q_1)}{n}\right](1-q_1) \tag{2.23}$$

and $\Gamma(\cdot)$ and $K(\cdot)$ are as in Theorem 2.2.6.

We should mention that in [Haiman, 1999, Theorem 4], the author obtained a similar formula for the error bound in Eq.(2.22), the only difference being that Δ_2 is replaced by Δ_2^H, where

$$\Delta_2^H = \frac{9}{n}+\frac{561}{n}(1-q_1)+3.3\left[1+4.7n(1-q_1)^2\right]. \tag{2.24}$$

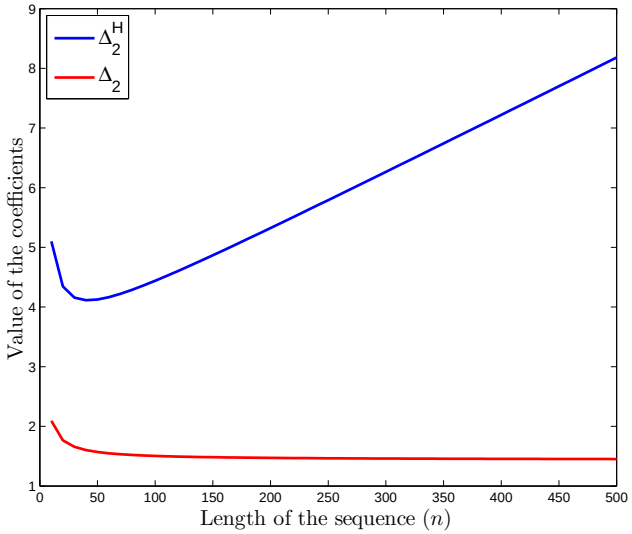
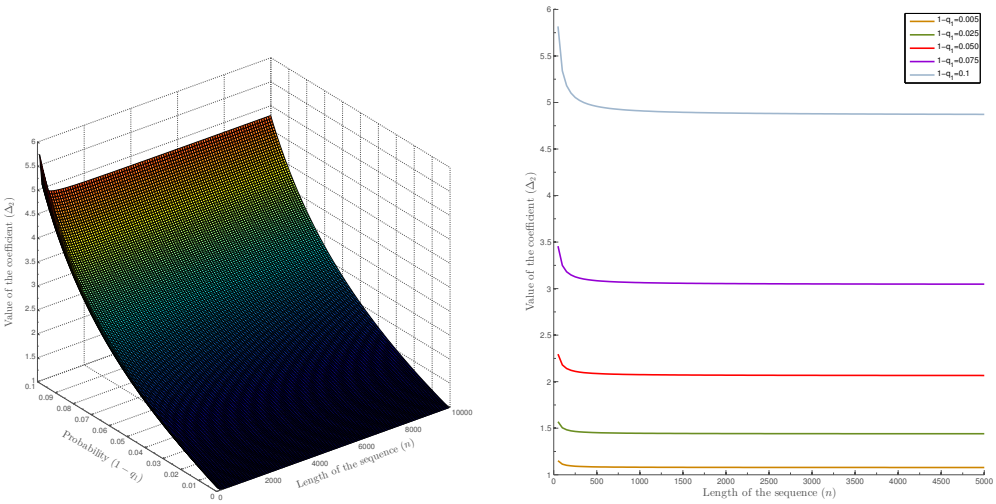


Figure 2.2: The relation between Δ_2^H and Δ_2 when $1 - q_1 = 0.025$ and n varies

As can be seen from Figure 2.2, when $1 - q_1$ is fixed and takes the value 0.025 (the upper limit in Haiman's results) and the length of the sequence (n) increases, the value of Δ_2^H increases with n, while the value of Δ_2 decreases. Also, one can note that the new error coefficient is much smaller than the old one. In Figure 2.3 we present the behaviour of the coefficient function Δ_2; in Figure 2.3(a) we illustrate the three dimensional plot and in Figure 2.3(b) we include the level curves for different values of $1 - q_1$.



(a) (b)

Figure 2.3: Illustration of the coefficient function Δ_2 for different values of p_1 and n

2.3 Proofs

In this section, we present the proofs for the results described in this chapter. The proofs for the main results described in Section 2.2.2 rely on a series of technical lemmas which will be stated in Section 2.3.1. We conclude the section by giving the proof of Proposition 2.1.4 presented at the end of Section 2.1.2.

2.3.1 Technical lemmas

We consider the framework described in Section 2.1.3. We begin by giving the following key upper bound for the probabilities of the minimum of the first n terms of the sequence $(X_n)_{n\geq 1}$. Using the stationarity and 1-dependence properties of the sequence we have

$$\begin{aligned} p_n = \mathbb{P}(X_1 > x, X_2 > x, \dots, X_n > x) &\leq \mathbb{P}(X_1 > x, X_3 > x, \dots, X_n > x) \\ &= \mathbb{P}(X_1 > x)\mathbb{P}(X_3 > x, \dots, X_n > x) \\ &= p_1 p_{n-2}, \end{aligned}$$

which gives the basic inequality

$$p_n \leq p_1^{[\frac{n+1}{2}]}. \tag{2.25}$$

This bound will be used over and over again in our results. Recall that in Eq.(2.9) we have defined the generating function

$$C(z) = C_x(z) = 1 + \sum_{k=1}^{\infty} (-1)^k p_{k-1} z^k. \tag{2.26}$$

The following lemmas will provide various estimates on the function $C(z)$.

Lemma 2.3.1. *The function $C_x(z)$ has a zero $\lambda = \lambda(x)$ in the interval $(1, 1 + lp_1)$, where $l = l(p_1) = t_2^3(p_1) + \varepsilon$, $\varepsilon > 0$ arbitrarily small and $t_2(p_1)$ is the second root in magnitude of the equation $p_1 t^3 - t + 1 = 0$.*

Proof. To show that $C(z)$ has a zero in the interval $(1, 1+lp_1)$ it is enough to verify that $C(1) > 0$ and $C(1 + lp_1) < 0$. It easy to see that $C(1) > 0$, since

$$C(1) = 1 + \sum_{k=1}^{\infty} (-1)^k p_{k-1} = \underbrace{(p_1 - p_2)}_{\geq 0} + \underbrace{(p_3 - p_4)}_{\geq 0} + \cdots \geq 0. \tag{2.27}$$

For $C(1 + lp_1)$ we have:

$$\begin{aligned} C(1 + lp_1) &= 1 + \sum_{k=1}^{\infty} (-1)^k p_{k-1}(1 + lp_1)^k \\ &= -lp_1 + \sum_{k=1}^{\infty} (1 + lp_1)^{2k} \underbrace{[p_{2k-1} - p_{2k}(1 + lp_1)]}_{\leq p_{2k-1} \leq p_1^k} \\ &\leq -lp_1 + \sum_{k=1}^{\infty} [(1 + lp_1)^2 p_1]^k. \end{aligned} \tag{2.28}$$

From the definition of l and the relation $1 < t_2(p_1) < t_2(p_1) + \varepsilon < \frac{1}{\sqrt{3p_1}}$, we obtain $(t_2 = t_2(p_1))$

$$t_2^2 + t_2\sqrt[3]{t_2^3+\varepsilon} + \sqrt[3]{\left(t_2^3+\varepsilon\right)^2} \leq t_2^2 + t_2(t_2+\varepsilon) + (t_2+\varepsilon)^2 < \frac{1}{3p_1} + \frac{1}{3p_1} + \frac{1}{3p_1} = \frac{1}{p_1},$$

which implies $t_2 - \sqrt[3]{t_2^3+\varepsilon} + p_1\varepsilon < 0$. Combining this last relation with the fact that t_2 is a root of the equation $p_1t^3 - t + 1 = 0$, we conclude that

$$p_1(t_2^3+\varepsilon) - \sqrt[3]{t_2^3+\varepsilon} + 1 < 0.$$

The foregoing relation can be rewritten as $(1+lp_1)^3 < l$ and since $p_1(1+lp_1)^2 < \frac{lp_1}{1+lp_1} < 1$, the series in Eq.(2.28) is convergent and we have $C(1+lp_1) < 0$.
Notice that the choice of l in the statement of the lemma was made such that the inequality $(1+lp_1)^3 < l$ holds. ■

Lemma 2.3.2. *Consider the series*

$$R = \sum_{k=2}^{\infty} 2kp_{2k-1} - \sum_{k=2}^{\infty}(2k+1)p_{2k}. \tag{2.29}$$

The following inequality holds:

$$|R| \leq 2p_1^2\left[\frac{1}{(1-p_1)^2} + \frac{1}{1-p_1}\right]. \tag{2.30}$$

Proof. To approximate R, we observe that since $p_{2k-1} - p_{2k} \geq 0$, the series can be bounded by

$$-\sum_{k=2}^{\infty} p_{2k} \leq R \leq \sum_{k=2}^{\infty} 2k\,(p_{2k-1} - p_{2k}).$$

With the help of Eq.(2.25), we obtain

$$-\frac{p_1^2}{1-p_1} = -\sum_{k=2}^{\infty} p_1^k \leq R \leq 2p_1\sum_{k=2}^{\infty} kp_1^{k-1} = 2p_1^2\left[\frac{1}{(1-p_1)^2} + \frac{1}{1-p_1}\right], \tag{2.31}$$

which implies the desired estimate for $|R|$. ■

Lemma 2.3.3. *Take $C''(z)$, the second derivative of the function $C(z)$. Then for all $z \in (1, 1+lp_1)$, we have*

$$-2lp_1^2\frac{1+3p_1(1+lp_1)^2}{[1-p_1(1+lp_1)^2]^3} - \frac{4p_1(1+lp_1)}{[1-p_1(1+lp_1)^2]^2} \leq C''(z) \leq 2p_1\frac{1+3p_1(1+lp_1)^2}{[1-p_1(1+lp_1)^2]^3}, \tag{2.32}$$

where $l = l(p_1)$ is defined in Lemma 2.3.1.

Proof. After derivation we have

$$\begin{aligned} C''(z) &= 2p_1 - 2\cdot 3p_2 z + 3\cdot 4p_3 z^2 - 4\cdot 5p_4 z^3 + 5\cdot 6p_5 z^4 - \dots \\ &= \sum_{k=0}^{\infty}(2k+1)(2k+2)p_{2k+1}z^{2k} - \sum_{k=0}^{\infty}(2k+2)(2k+3)p_{2k+2}z^{2k+1}. \end{aligned} \tag{2.33}$$

Based on the estimate in Eq.(2.25) and using the fact that $z \in (1, 1+lp_1)$, we find the upper bound

$$\begin{aligned} C''(z) &\le p_1 \sum_{k=0}^{\infty}(2k+1)(2k+2)(z\sqrt{p_1})^{2k} \\ &\le 2p_1 \frac{1+3p_1(1+lp_1)^2}{[1-p_1(1+lp_1)^2]^3}. \end{aligned} \tag{2.34}$$

Similarly, we obtain the following lower bound:

$$\begin{aligned} C''(z) &\ge -lp_1 \sum_{k=0}^{\infty}(2k+1)(2k+2)p_{2k+2}z^{2k} - 4\sum_{k=0}^{\infty}(k+1)p_{2k+2}z^{2k+1} \\ &\ge -lp_1^2 \sum_{k=0}^{\infty}(2k+1)(2k+2)(z\sqrt{p_1})^{2k} - 4p_1 z\sum_{k=0}^{\infty}(k+1)(p_1 z^2)^k \\ &\ge -2lp_1^2 \frac{1+3p_1(1+lp_1)^2}{[1-p_1(1+lp_1)^2]^3} - \frac{4p_1(1+lp_1)}{[1-p_1(1+lp_1)^2]^2}. \end{aligned} \tag{2.35}$$

From the upper and lower bounds in Eq.(2.34) and Eq.(2.35), respectively, we have the estimate

$$|C''(z)| \le 2lp_1^2 \frac{1+3p_1(1+lp_1)^2}{[1-p_1(1+lp_1)^2]^3} + \frac{4p_1(1+lp_1)}{[1-p_1(1+lp_1)^2]^2}. \quad \blacksquare \tag{2.36}$$

Lemma 2.3.4. *Let $C'(z)$ be the first derivative of the function $C(z)$. For all $z \in (1, 1+lp_1)$, we have the bounds*

$$C'(z) \le -1 + \frac{2p_1}{(1-p_1)^2} + 2lp_1^2 \frac{1+3p_1(1+lp_1)^2}{[1-p_1(1+lp_1)^2]^3}, \tag{2.37}$$

$$|C'(z)|^{-1} \le \left[1 - \frac{2p_1(1+lp_1)}{[1-p_1(1+lp_1)^2]^2}\right]^{-1}, \tag{2.38}$$

where $l = l(p_1)$ is defined in Lemma 2.3.1.

Proof. We will derive first the upper bound for $C'(z)$ given by (2.37). Using Lagrange theorem on the interval $[1, 1+a]$ with $a \le lp_1$, we get for $\theta \in (0,1)$:

$$C'(1+a) = C'(1) + aC''(1+\theta a). \tag{2.39}$$

We observe that

$$\begin{aligned} C'(1) &= -1 + 2p_1 - 3p_2 + 4p_3 - 5p_4 + 6p_5 - 7p_6 + \dots \\ &= -1 + 2p_1 - 3p_2 + R, \end{aligned} \tag{2.40}$$

where R is defined in Lemma 2.3.2.
Applying the estimate in Lemma 2.3.2 for R and the upper bound for $C''(z)$, derived in Lemma 2.3.3, we deduce

$$C'(z) \leq -1 + \frac{2p_1}{(1-p_1)^2} + 2lp_1^2 \frac{1+3p_1(1+lp_1)^2}{[1-p_1(1+lp_1)^2]^3}. \tag{2.41}$$

Notice that for the derivation of the foregoing expression we have also used the fact that $a \leq lp_1$.
To derive the second estimate, observe that

$$\begin{aligned} |C'(z)|^{-1} &= |1-(2p_1z-3p_2z^2+4p_3z^3-5p_4z^4+\dots)|^{-1} \\ &\leq \left|1-|2p_1z-3p_2z^2+4p_3z^3-5p_4z^4+\dots|\right|^{-1}, \end{aligned} \tag{2.42}$$

where we used the inequality $|1-x| \geq |1-|x||$. Taking into account that $z \in (1, 1+lp_1)$ and denoting the expression inside the second absolute value in the denominator of Eq.(2.42) with T, we have the upper margin

$$\begin{aligned} T &= \sum_{k=1}^{\infty} 2k(p_{2k-1}-p_{2k}z)z^{2k-1} - \sum_{k=1}^{\infty} p_{2k}z^{2k} \\ &\leq 2(p_1-p_2)z\sum_{k=1}^{\infty} k(p_1z^2)^{k-1} \leq \frac{2p_1z}{(1-p_1z^2)^2}. \end{aligned} \tag{2.43}$$

In the same way,

$$\begin{aligned} T &\geq -\sum_{k=1}^{\infty} p_{2k}z^{2k} - 2lp_1\sum_{k=1}^{\infty} kp_{2k}z^{2k-1} \\ &\geq -\sum_{k=1}^{\infty} p_1^k z^{2k} - 2lp_1^2 z\sum_{k=1}^{\infty} k(p_1z^2)^{k-1} \\ &\geq -p_1z\frac{2lp_1+z(1-p_1z^2)}{(1-p_1z^2)^2} > -\frac{2p_1z}{(1-p_1z^2)^2}. \end{aligned} \tag{2.44}$$

Combining the bounds in Eq.(2.43) and Eq.(2.44) along with $z \leq 1+lp_1$, leads to

$$|T| \leq \frac{2p_1(1+lp_1)}{[1-p_1(1+lp_1)^2]^2}. \tag{2.45}$$

Substitute the bound from Eq.(2.45) in Eq.(2.42) to obtain the estimate

$$|C'(z)|^{-1} \leq \frac{1}{1-\frac{2p_1(1+lp_1)}{[1-p_1(1+lp_1)^2]^2}}. \ \blacksquare \tag{2.46}$$

The following results will be used in the proof of Theorem 2.2.6. Recall, from Section 2.1.3, that

$$D(z) = D_x(z) = \sum_{k=0}^{\infty} q_{k-1}z^k. \tag{2.47}$$

The next lemma presents the relation between the generating functions $D(z)$ and $C(z)$. As we saw in Section 2.1.3, this formula will constitute the key idea behind our approach.

Lemma 2.3.5. *Let $C(z)$ and $D(z)$ be the generating functions defined by Eq.(2.26) and Eq.(2.47), respectively. The following relation holds:*

$$D(z) = \frac{1}{C(z)}. \tag{2.48}$$

Proof. We define the following series:

$$\tilde{D}(z) = \frac{1}{C(z)} = \sum_{k=0}^{\infty} d_k z^k, \tag{2.49}$$

which exists since the free term in the series C is equal to 1. Based on the relation $C(z)\tilde{D}(z) = 1$, we deduce that $d_0 = 1$ and

$$\sum_{j=0}^{n} (-1)^{n-j} p_{n-j-1} d_j = 0, \quad n \geq 1. \tag{2.50}$$

Define for each $k = \{0, 1, \ldots, n\}$ the events

$$A_k = \{X_1 \leq x, \ldots, X_k \leq x, X_{k+1} > x, \ldots, X_n > x\}.$$

We observe that $\mathbb{P}(A_0) = p_n$, $\mathbb{P}(A_n) = q_n$ and

$$\mathbb{P}(A_k) + \mathbb{P}(A_{k-1}) = p_{n-k} q_{k-1}, \quad k \geq 1. \tag{2.51}$$

Multiplying Eq.(2.51) with $(-1)^{k-1}$ and summing over k gives

$$q_n = \sum_{k=0}^{n} (-1)^{n-k} p_{n-k} q_{k-1}, \quad n \geq 0, \quad q_{-1} = q_0 = 1. \tag{2.52}$$

Notice that from the above relation and Eq.(2.50), after using mathematical induction, we obtain $d_{n+1} = q_n$. From the definition of the series $\tilde{D}$ we conclude that

$$\tilde{D}(z) = \sum_{k=0}^{\infty} q_{k-1} z^k, \quad q_{-1} = q_0 = 1, \tag{2.53}$$

which is exactly the generating function $D(z)$ and the proof is complete. ■

In the following lemma, we assume that the result in Theorem 2.2.3 is known.

Lemma 2.3.6. *We have*

$$q_3 \lambda^3 = 1 - p_2 + 2p_3 - 3p_4 + p_1^2 + 6p_2^2 - 6p_1 p_2 + \mathcal{O}(L(p_1) p_1^3), \tag{2.54}$$

where $\mathcal{O}(x)$ is a function such that $|\mathcal{O}(x)| \leq |x|$, $L(p_1) = 36 + (1-p_1)^2 P(p_1)$ and $P(p_1)$ has the form

$$\begin{aligned} P(p_1) = {} & 3K(p_1)(1 + p_1 + 3p_1^2)[1 + p_1 + 3p_1^2 + K(p_1)p_1^3] + p_1^6 K^3(p_1) \\ & + 9p_1(4 + 3p_1 + 3p_1^2) + 19. \end{aligned} \tag{2.55}$$

Proof. From Eq.(2.13) of Theorem 2.2.3 we can write

$$\lambda = 1 + p_1 - p_2 + p_3 - p_4 + 2p_1^2 + 3p_2^2 - 5p_1p_2 + \mathcal{O}(K(p_1)p_1^3) \tag{2.56}$$

and raising to the third power we get

$$\begin{aligned}\lambda^3 = (1+\zeta_1+\zeta_2)^3 + 3(1+\zeta_1+\zeta_2)^2\mathcal{O}(K(p_1)p_1^3) + \mathcal{O}(K^3(p_1)p_1^6p_1^3)+\\ + 3(1+\zeta_1+\zeta_2)\mathcal{O}(K^2(p_1)p_1^3p_1^3).\end{aligned} \tag{2.57}$$

In the above formulas we have used the notations

$$\zeta_1 = p_1 - p_2, \tag{2.58}$$
$$\zeta_2 = p_3 - p_4 + 2p_1^2 + 3p_2^2 - 5p_1p_2. \tag{2.59}$$

From the definitions of ζ_1 and ζ_2, it is not hard to see that $\zeta_1 = \mathcal{O}(p_1)$ and

$$\zeta_2 \leq p_1(p_1 - p_2) + 2(p_1 - p_2)^2 - p_2(p_1 - p_2) = \mathcal{O}(3p_1^2).$$

From these observations we deduce that

$$\lambda^3 - (1+\zeta_1+\zeta_2)^3 = \mathcal{O}(S(p_1)p_1^3), \tag{2.60}$$

with $S(p_1)$ given bellow by

$$S(p_1) = 3(1+p_1+3p_1^2)^2K(p_1) + 3p_1^3(1+p_1+3p_1^2)K^2(p_1) + p_1^6K^3(p_1). \tag{2.61}$$

Now observe that by expanding $(1+\zeta_1+\zeta_2)^3$, we can write

$$\begin{aligned}(1+\zeta_1+\zeta_2)^3 &= 1 + 3(\zeta_1+\zeta_2+\zeta_1^2) + 6\zeta_1\zeta_2 + 3\zeta_2^2 + 3\zeta_1\zeta_2^2 + 3\zeta_1^2\zeta_2\\ &\quad + \zeta_1^3 + \zeta_2^3\\ &= 1 + 3(\zeta_1+\zeta_2+\zeta_1^2) + \mathcal{O}(18p_1^3) + \mathcal{O}(27p_1p_1^3) + \mathcal{O}(27p_1^2p_1^3)\\ &\quad + \mathcal{O}(9p_1p_1^3) + \mathcal{O}(p_1^3) + \mathcal{O}(27p_1^3p_1^3),\end{aligned} \tag{2.62}$$

which along with Eq.(2.60) and Eq.(2.61) gives an expression for λ^3,

$$\lambda^3 = 1 + 3(\zeta_1+\zeta_2+\zeta_1^2) + \mathcal{O}(P(p_1)p_1^3). \tag{2.63}$$

The function $P(p_1)$, in Eq.2.63 above, is computed by the formula

$$\begin{aligned}P(p_1) &= 3K(p_1)(1+p_1+3p_1^2)[1+p_1+3p_1^2+K(p_1)p_1^3] + p_1^6K^3(p_1)\\ &\quad + 9p_1(4+3p_1+3p_1^2) + 19.\end{aligned} \tag{2.64}$$

Recall, from the previous lemma, that we have the recurrence

$$q_n = \sum_{k=0}^{n}(-1)^{n-k}p_{n-k}q_{k-1}, \quad n \geq 0, \quad q_{-1} = q_0 = 1,$$

which for $n = 3$ implies that

$$q_3 = 1 - p_1 - 2(p_1 - p_2) + p_1^2 - p_3 = \mathcal{O}((1-p_1)^2). \tag{2.65}$$

Now it is clear that from Eq.(2.63) and Eq.(2.65) we obtain

$$q_3\lambda^3 = q_3[1 + 3(\zeta_1 + \zeta_2 + \zeta_1^2)] + \mathcal{O}(P(p_1)(1-p_1)^2p_1^3). \tag{2.66}$$

The last step in our proof is to find an approximation for the first term on the right in Eq.(2.66). If we write

$$q_3[1 + 3(\zeta_1 + \zeta_2 + \zeta_1^2)] = 1 - p_2 + 2p_3 - 3p_4 + p_1^2 + 6p_2^2 - 6p_1p_2 + H, \tag{2.67}$$

then H verifies

$$\begin{aligned} H &= 3(p_1 - p_2)\left\{(p_1^2 - p_3)[3(p_1 - p_2) + 1 - p_2] + p_2^2 - 9(p_1 - p_2)^2\right\} \\ &\quad - 3(p_3 - p_4)[p_1 + 2(p_1 - p_2) - (p_1^2 - p_3)] \\ &= \mathcal{O}(36p_1^3). \end{aligned} \tag{2.68}$$

Finally, combining Eq.(2.66), Eq.(2.67) and Eq.(2.68) leads to

$$q_3\lambda^3 = 1 - p_2 + 2p_3 - 3p_4 + p_1^2 + 6p_2^2 - 6p_1p_2 + \mathcal{O}(L(p_1)p_1^3), \tag{2.69}$$

where we used the notation

$$L(p_1) = 36 + (1-p_1)^2P(p_1). \quad \blacksquare \tag{2.70}$$

2.3.2 Proof of Theorem 2.2.3

We saw from Lemma 2.3.1 that the series $C(z)$ has a zero $\lambda = \lambda(x)$ in the interval $(1, 1 + lp_1)$, where $l = l(p_1) = t_2^3(p_1) + \varepsilon$, $\varepsilon > 0$ arbitrarily small and $t_2(p_1)$ is the second root in magnitude of the equation $p_1t^3 - t + 1 = 0$. To show that this zero is unique, we will prove that $C(z)$ is strictly decreasing on the interval $(1, 1 + lp_1)$, i.e. $C'(z) < 0$. In Lemma 2.3.4, we found the following upper bound for $C'(z)$:

$$C'(z) \le -1 + \frac{2p_1}{(1-p_1)^2} + 2lp_1^2\frac{1 + 3p_1(1 + lp_1)^2}{[1 - p_1(1 + lp_1)^2]^3}.$$

We observe that the expression of the bound in the above relation is increasing in p_1 and since for $p_1 = 0.1$, $l \le 1.1535$, we get

$$C'(z) < -0.7,$$

which proves that C is strictly decreasing.
Now we try to approximate the zero λ. From Lagrange theorem applied on the interval $[1, \lambda]$, we have $C(\lambda) - C(1) = (\lambda - 1)C'(u)$, with $u \in (1, \lambda) \subset (1, 1 + lp_1)$. Since $C(\lambda) = 0$, we get

$$\lambda - 1 = -\frac{C(1)}{C'(u)} \tag{2.71}$$

and taking $\mu = p_1 - p_2 + p_3 - p_4 + 2p_1^2 + 3p_2^2 - 5p_1p_2$ as in [Haiman, 1999], we obtain

$$\lambda - (1+\mu) = -\frac{C(1) + \mu C'(u)}{C'(u)}. \tag{2.72}$$

Applying Lagrange theorem one more time for C', as in the proof of Lemma 2.3.4, on the interval $[1, 1+a]$ with $a \leq lp_1$, we get for $\theta \in (0,1)$

$$C'(1+a) = C'(1) + aC''(1+\theta a).$$

Recall that

$$C'(1) = -1 + 2p_1 - 3p_2 + R,$$

with R defined by Lemma 2.3.2. After the proper substitutions, the relation in Eq.(2.72) becomes

$$\lambda - (1+\mu) = -\frac{C(1) + \mu(-1 + 2p_1 - 3p_2) + \mu(R + aC''(1+\theta a))}{C'(u)}, \tag{2.73}$$

where $a = u - 1$ and $\theta \in (0,1)$.
If we denote $A = C(1) + \mu(-1 + 2p_1 - 3p_2)$, then

$$\begin{aligned} A &= (p_1 - p_2 + p_3 - p_4 + p_5 - p_6 + \dots) - \mu + (2p_1 - 3p_2)\mu \\ &= (p_5 - p_6 + \dots) + (p_1 - p_2)(2p_1 - 3p_2)^2 + 2(p_1 - p_2)(p_3 - p_4) \\ &\quad - p_2(p_3 - p_4). \end{aligned} \tag{2.74}$$

Applying the inequality from Eq.(2.25) in Eq.(2.74), we have

$$-p_1^3 \leq -p_2(p_3 - p_4) \leq A \leq \sum_{k=2}^{\infty} p_1^{k+1} + 4p_1^3 + 2p_1^3,$$

which gives us the estimate

$$|C(1) + \mu(-1 + 2p_1 - 3p_2)| \leq p_1^3 \left[6 + \frac{1}{1-p_1}\right]. \tag{2.75}$$

Notice that $\mu \geq 0$ and

$$\begin{aligned} \mu &= (1-p_2)(p_1 - p_2) + p_3 - p_4 + 2(p_1 - p_2)^2 \\ &\leq (1-p_2)(p_1 - p_2) + p_1(p_1 - p_2) + 2(p_1 - p_2)^2 \\ &= 3(p_1 - p_2)^2 + p_1 - p_2 \leq p_1(1 + 3p_1). \end{aligned} \tag{2.76}$$

Recall that, from Lemma 2.3.4, we have the following bound for $|C'(z)|^{-1}$:

$$|C'(z)|^{-1} \leq \left[1 - \frac{2p_1(1+lp_1)}{[1 - p_1(1+lp_1)^2]^2}\right]^{-1}.$$

Also, the bounds in Lemma 2.3.3 lead to the estimate

$$|C''(z)| \leq 2lp_1^2 \frac{1 + 3p_1(1+lp_1)^2}{[1 - p_1(1+lp_1)^2]^3} + \frac{4p_1(1+lp_1)}{[1 - p_1(1+lp_1)^2]^2}. \tag{2.77}$$

Substituting the estimate for $|R|$ derived in Lemma 2.3.2, along with the bounds in Eqs.(2.75), (2.76), (2.38) and (2.77) in Eq.(2.73) and using that $a \leq lp_1$, gives the approximation

$$\begin{aligned} |\lambda - (1+\mu)| &\leq \frac{|C(1) + \mu(-1+2p_1-3p_2)| + |\mu|(|R| + |a||C''(1+\theta a)|)}{|C'(u)|} \\ &\leq K(p_1)p_1^3 \end{aligned} \tag{2.78}$$

where

$$K(p_1) = \frac{\frac{11-3p_1}{(1-p_1)^2} + 2l(1+3p_1)\frac{2+3lp_1-p_1(2-lp_1)(1+lp_1)^2}{[1-p_1(1+lp_1)^2]^3}}{1 - \frac{2p_1(1+lp_1)}{[1-p_1(1+lp_1)^2]^2}}. \tag{2.79}$$

2.3.3 Proof of Corollary 2.2.4

From Theorem 2.2.3 we have

$$\left|\lambda - (1 + p_1 - p_2 + p_3 - p_4 + 2p_1^2 + 3p_2^2 - 5p_1p_2)\right| \leq K(p_1)p_1^3.$$

To prove the statement in Corollary 2.2.4, we first observe that

$$1+p_1-p_2+p_3-p_4+2p_1^2+3p_2^2-5p_1p_2 = 1+p_1-p_2+2(p_1-p_2)^2+[p_3-p_4-p_2(p_1-p_2)].$$

Using the 1-dependence of the sequence X_n, we notice that

$$\begin{aligned} p_3 - p_4 &= \mathbb{P}(X_1 > x, X_2 > x, X_3 > x, X_4 \leq x) \\ &\leq p_1\mathbb{P}(X_1 > x, X_2 \leq x) = p1(p_1 - p_2), \end{aligned} \tag{2.80}$$

which leads to the bound

$$|p_3 - p_4 - p_2(p_1 - p_2)| \leq p_3 - p_4 + p_2(p_1 - p_2) \leq p_1^2.$$

Combining the above relations, we conclude that

$$|\lambda - (1 + p_1 - p_2 + 2(p_1 - p_2)^2)| \leq p_1^2(1 + K(p_1)p_1) \leq (1 + p_1K(p_1))p_1^2. \quad \blacksquare$$

2.3.4 Proof of Theorem 2.2.6

From Lemma 2.3.5 we know that

$$D(z) = \frac{1}{C(z)}.$$

Taking λ to be zero defined in Theorem 2.2.3, we can write $C(z) = U(z)\left(1 - \frac{z}{\lambda}\right)$ and observe that if we put $U(z) = \sum_{k=0}^{n} u_k z^k$, then

$$\left(\sum_{k=0}^{n} u_k z^k\right)\left(1 - \frac{z}{\lambda}\right) = 1 + \sum_{k=1}^{\infty}(-1)^k p_{k-1} z^k, \tag{2.81}$$

which shows that

$$u_n - \frac{u_{n-1}}{\lambda} = (-1)^n p_{n-1},\ u_0 = 1,\ n \geq 1. \tag{2.82}$$

Multiplying Eq.(2.82) with λ^n and summing over n gives

$$u_n = \frac{1 + \sum_{k=1}^{n} (-1)^k p_{k-1} \lambda^k}{\lambda^n},\quad n \geq 1. \tag{2.83}$$

If we denote with $T(z) = \frac{1}{U(z)} = \sum_{k=0}^{\infty} t_k z^k$, then

$$D(z)\left(1 - \frac{z}{\lambda}\right) = T(z) \tag{2.84}$$

and using the same arguments as in Lemma 2.3.5 we get $t_0 = 1$ and

$$t_n = q_{n-1} - \frac{q_{n-2}}{\lambda},\quad n \geq 1, \tag{2.85}$$

so

$$q_n \lambda^{n+1} = t_0 + t_1 \lambda + \cdots + t_n \lambda^n + t_{n+1} \lambda^{n+1}. \tag{2.86}$$

To obtain the desired result, we begin by giving an approximation of u_n:

$$\begin{aligned} |u_n| &= \frac{\left|1 + \sum_{k=1}^{n} (-1)^k p_{k-1} \lambda^k\right|}{\lambda^n} \underset{C(\lambda)=0}{=} \frac{\left|\sum_{k=n+1}^{\infty} (-1)^k p_{k-1} \lambda^k\right|}{\lambda^n} \\ &\leq \frac{\lambda^{n+1}}{\lambda^n} \left|p_n - p_{n+1}\lambda + p_{n+2}\lambda^2 - p_{n+3}\lambda^3 + \dots\right| \\ &\leq \lambda(|p_n - p_{n+1}\lambda| + |p_{n+2} - p_{n+3}\lambda|\lambda^2 + \dots). \end{aligned} \tag{2.87}$$

Since $\lambda \in (1, 1 + lp_1)$, we have

$$p_n - p_{n+1}(1 + lp_1) \leq p_n - p_{n+1}\lambda \leq p_n - p_{n+1}, \tag{2.88}$$

which shows that

$$|p_n - p_{n+1}\lambda| \leq p_n - p_{n+1}(1 - lp_1). \tag{2.89}$$

Let $h = 1 - lp_1$. Using in Eq.(2.87) the fact that the sequence of probabilities $(p_n)_n$ is decreasing and the bound from Eq.(2.89), we obtain

$$\begin{aligned} \frac{|u_n|}{\lambda} &\leq p_n - p_{n+1}h + (p_{n+2} - p_{n+3}h)\lambda^2 + \dots \\ &= (p_n + p_{n+2}\lambda^2 + p_{n+4}\lambda^4 + \dots) - h(p_{n+1} + p_{n+3}\lambda^2 + p_{n+5}\lambda^4 + \dots) \\ &\leq W - h(p_{n+2} + p_{n+4}\lambda^2 + p_{n+6}\lambda^4 + \dots) \\ &= W - \frac{h}{\lambda^2}(W - p_n) = W\left(1 - \frac{h}{\lambda^2}\right) + \frac{h}{\lambda^2} p_n, \end{aligned} \tag{2.90}$$

where, based on the estimate from Eq.(2.25),

$$\begin{aligned} W &= p_n + p_{n+2}\lambda^2 + p_{n+4}\lambda^4 + \dots \\ &\le p_1^{\left[\frac{n+1}{2}\right]} + p_1^{\left[\frac{n+1}{2}\right]} p_1\lambda^2 + p_1^{\left[\frac{n+1}{2}\right]} p_1^2\lambda^4 + \dots \\ &= p_1^{\left[\frac{n+1}{2}\right]}(1 + p_1\lambda^2 + p_1^2\lambda^4 + \dots) = \frac{p_1^{\left[\frac{n+1}{2}\right]}}{1 - p_1\lambda^2}. \end{aligned} \tag{2.91}$$

From Eq.(2.90) and Eq.(2.91), we conclude that

$$\begin{aligned} \frac{|u_n|}{\lambda} &\le p_1^{\left[\frac{n+1}{2}\right]} \left[\frac{1}{1 - p_1\lambda^2} + \frac{h}{\lambda^2}\left(1 - \frac{1}{1 - p_1\lambda^2}\right)\right] \\ &= \frac{1 - p_1 h}{1 - p_1\lambda^2} p_1^{\left[\frac{n+1}{2}\right]}. \end{aligned} \tag{2.92}$$

Until now we have an approximation for u_n, but we still need one for t_n and to solve this aspect let us write

$$T(z) = \frac{1}{U(z)} = \frac{1}{1 - (1 - U(z))} = \sum_{n\ge 0} (-1)^n (U-1)^n, \tag{2.93}$$

which is true since the convergence of $C(z)$ implies $|z| < \frac{1}{\sqrt{p_1}}$ and in turn gives

$$\begin{aligned} |1 - U| &\le \frac{\lambda(1 - p_1 h)}{1 - p_1\lambda^2} \sum_{n\ge 1} p_1^{\frac{n}{2}} |z|^n \le \frac{|z|\lambda\sqrt{p_1}\,[1 - p_1(1 - lp_1)]}{(1 - p_1\lambda^2)(1 - |z|\sqrt{p_1})} \\ &\le \frac{\sqrt{p_1}(1 + lp_1^2)(1 + lp_1)^2}{[1 - p_1(1 + lp_1)^2]\,[1 - \sqrt{p_1}(1 + lp_1)]} < 0.8. \end{aligned} \tag{2.94}$$

Since $u_0 = 1$, we have $U - 1 = \sum_{n\ge 1} u_n z^n$ and

$$(U-1)^k = \sum_{l\ge 1} \sum_{\substack{i_1 + \dots + i_k = l \\ i_j \ge 1, j = \overline{1,k}}} u_{i_1} \dots u_{i_k} z^l. \tag{2.95}$$

Combining Eq.(2.93) with Eq.(2.95), we obtain

$$\sum_{n\ge 0} t_n z^n = \sum_{k=0}^{\infty} (-1)^k \sum_{l=1}^{\infty} b_{l,k} z^l, \tag{2.96}$$

where

$$b_{l,k} = \sum_{\substack{i_1 + \dots + i_k = l \\ i_j \ge 1, j = \overline{1,k}}} u_{i_1} \dots u_{i_k}. \tag{2.97}$$

Identifying the coefficients in Eq.(2.96) shows that $t_0 = 1$ and

$$t_n = \sum_{k=1}^{n} (-1)^k \sum_{\substack{i_1 + \dots + i_k = n \\ i_j \ge 1, j = \overline{1,k}}} u_{i_1} \dots u_{i_k}, \quad k \ge 1. \tag{2.98}$$

Notice that the coefficients $b_{n,k}$ can be bounded by

$$|b_{n,k}| \leq \delta^k \sum_{\substack{i_1+\cdots+i_k=n \\ i_j \geq 1, j=\overline{1,k}}} p_1^{\left[\frac{i_1+1}{2}\right]} \ldots p_1^{\left[\frac{i_k+1}{2}\right]}, \tag{2.99}$$

where $\delta = \frac{\lambda(1-p_1 h)}{1-p_1\lambda^2}$.
By induction, one can easily check the validity of the identity

$$\left[\frac{i_1+1}{2}\right] + \cdots + \left[\frac{i_k+1}{2}\right] \geq \left[\frac{i_1+\cdots+i_k+1}{2}\right] = \left[\frac{n+1}{2}\right]. \tag{2.100}$$

We observe that the number of terms in the sum of Eq.(2.99) is equal with the number of different positive integers solutions of the equation $i_1+\cdots+i_k=n$ and is given by $\binom{n-1}{k-1}$ (see for example [Yaglom and Yaglom, 1987, Problem 31]). This shows that

$$|b_{n,k}| \leq p_1^{\left[\frac{n+1}{2}\right]} \binom{n-1}{k-1} \delta^k. \tag{2.101}$$

Now, from Eq.(2.98) and Eq.(2.101), we have

$$-p_1^{\left[\frac{n+1}{2}\right]} \delta \sum_{k=0}^{\left[\frac{n-1}{2}\right]} \binom{n-1}{2k} \delta^{2k} \leq t_n \leq p_1^{\left[\frac{n+1}{2}\right]} \delta \sum_{k=0}^{\left[\frac{n}{2}\right]-1} \binom{n-1}{2k+1} \delta^{2k+1}, \tag{2.102}$$

which lead to the bound

$$|t_n| \leq \frac{\delta}{2} p_1^{\left[\frac{n+1}{2}\right]} \left[(1+\delta)^{n-1} + (1-\delta)^{n-1}\right]. \tag{2.103}$$

We observe that from Eq.(2.86) and Eq.(2.103), the difference $|q_n\lambda^n - q_3\lambda^3|$ can be bounded by

$$\begin{aligned} |q_n\lambda^n - q_3\lambda^3| &= \left| \frac{\sum_{s=0}^{n+1} t_s\lambda^s - \sum_{s=0}^{4} t_s\lambda^s}{\lambda} \right| = \left| \sum_{s=5}^{n+1} t_s\lambda^{s-1} \right| \\ &\leq \frac{\delta}{2} \sum_{s=5}^{\infty} p_1^{\left[\frac{s+1}{2}\right]} \left[(1+\delta)^{s-1} + (1-\delta)^{s-1}\right] \lambda^{s-1}. \end{aligned} \tag{2.104}$$

If we denote by $\sigma_1 = (1+\delta)\lambda$, $\sigma_2 = (1-\delta)\lambda$ and by V the upper bound in Eq.(2.104), it is not hard to see that

$$V = \frac{\delta p_1^3}{2} \left[\frac{\sigma_1^4(1+\sigma_1)}{1-p_1\sigma_1^2} + \frac{\sigma_2^4(1+\sigma_2)}{1-p_1\sigma_2^2}\right]. \tag{2.105}$$

Recalling that $h = 1 - lp_1$, $\lambda \in (1, 1+lp_1)$ and $p_1 \leq 0.1$, we observe that σ_2 is bounded by

$$-\frac{lp_1[1+2p_1(1+lp_1)]}{1-p_1(1+lp_1)^2} \leq \sigma_2 \leq -\frac{lp_1^2(1+lp_1)}{1-p_1},$$

which gives $|\sigma_2| < 0.5$ and $\frac{\delta\sigma_2^4(1+\sigma_2)}{2(1-p_1\sigma_2^2)} < 0.1$.
Substituting the last relation in Eq.(2.105), we can rewrite the bound in Eq.(2.104) as

$$|q_n\lambda^n - q_3\lambda^3| \leq E(p_1)p_1^3 \tag{2.106}$$

where, if we denote by $\eta = 1 + lp_1$,

$$E(p_1) = 0.1 + \frac{\eta^5\left[1+(1-2p_1)\eta\right]^4\left[1+p_1(\eta-2)\right]\left[1+\eta+(1-3p_1)\eta^2\right]}{2(1-p_1\eta^2)^4\left[(1-p_1\eta^2)^2 - p_1\eta^2(1+\eta-2p_1\eta)^2\right]}. \tag{2.107}$$

We saw in Lemma 2.3.6 that

$$q_3\lambda^3 = 1 - p_2 + 2p_3 - 3p_4 + p_1^2 + 6p_2^2 - 6p_1p_2 + \mathcal{O}(L(p_1)p_1^3),$$

where $L(p_1) = 36+(1-p_1)^2P(p_1)$ and $P(p_1)$ is given by Eq.(2.55). Using Eq.(2.106) and the above estimate, we conclude that

$$|q_n\lambda^n - (1 - p_2 + 2p_3 - 3p_4 + p_1^2 + 6p_2^2 - 6p_1p_2)| \leq \Gamma(p_1)p_1^3, \tag{2.108}$$

with $\Gamma(p_1) = 36 + (1-p_1)^2P(p_1) + E(p_1)$ and this ends the proof of the theorem. Remark that in the statement of Theorem 2.2.6 we gave $E(p_1)$ without the 0.1 term, but we have added it to the constant term 36. ■

2.3.5 Proof of Corollary 2.2.7

We know from Theorem 2.2.6 that the following relation holds:

$$|q_n\lambda^n - (1 - p_2 + 2p_3 - 3p_4 + p_1^2 + 6p_2^2 - 6p_1p_2)| \leq \Gamma(p_1)p_1^3.$$

Based on this bound, we can write

$$|q_n\lambda^n - (1-p_2)| \leq \Gamma(p_1)p_1^3 + |2p_3 - 3p_4 + p_1^2 + 6p_2^2 - 6p_1p_2|. \tag{2.109}$$

Observe that

$$0 \leq p_1^2 + 2p_3 - 3p_4 = \underbrace{p_1^2 - p_4}_{\leq p_1^2} + 2(\underbrace{p_3 - p_4}_{\leq p_1^2}) \leq 3p_1^2$$

and that

$$-3p_1^2 \leq 6p_2^2 - 6p_1p_2 \leq 0.$$

By adding these two set of inequalities, we have the bound

$$\left|p_1^2 + 2p_3 - 3p_4 + 6p_2^2 - 6p_1p_2\right| \leq 3p_1^2,$$

which substituted in Eq.(2.109) implies

$$|q_n\lambda^n - (1-p_2)| \leq (3 + p_1\Gamma(p_1))p_1^2. \quad ■ \tag{2.110}$$

2.3.6 Proof of Theorem 2.2.8

Denoting by

$$\mu_1 = 1 - p_2 + 2p_3 - 3p_4 + p_1^2 + 6p_2^2 - 6p_1p_2,$$
$$\mu_2 = 1 + p_1 - p_2 + p_3 - p_4 + 2p_1^2 + 3p_2^2 - 5p_1p_2$$

we observe that

$$0 \leq \frac{\mu_1}{\mu_2} < 1$$

and that

$$\mu_2 = 1 + \underbrace{(p_1 - p_2)(1 - p_2) + p_3 - p_4 + 2(p_1 - p_2)^2}_{\geq 0} \geq 1. \tag{2.111}$$

With the help of Eq.(2.13) and Eq.(2.16), we get

$$\begin{aligned}\left|q_n - \frac{\mu_1}{\mu_2^n}\right| &\leq \left|q_n - \frac{\mu_1}{\lambda^n}\right| + \left|\frac{\mu_1}{\lambda^n} - \frac{\mu_1}{\mu_2^n}\right| \\ &\leq \Gamma(\alpha)p_1^3 + |\mu_1|\left|\frac{1}{\lambda} - \frac{1}{\mu_2}\right|\left|\frac{1}{\lambda^{n-1}} + \cdots + \underbrace{\frac{1}{\lambda^{n-j}\mu_2^j}}_{\leq 1} + \cdots + \frac{1}{\mu_2^{n-1}}\right| \\ &\leq \left[\Gamma(p_1) + nK(p_1)\right]p_1^3. \end{aligned} \tag{2.112}$$

To express μ_1 and μ_2 in terms of q's, we have to observe first that

$$\begin{aligned} p_1 &= 1 - q_1 \\ p_2 &= 1 - 2q_1 + q_2 \\ p_3 &= 1 - 3q_1 + 2q_2 + q_1^2 - q_3 \\ p_4 &= 1 - 4q_1 + 3q_2 - 2q_1q_2 + 3q_1^2 - 2q_3 + q_4 \end{aligned}$$

and after the proper substitutions, we get

$$\mu_1 = 1 + q_1 - q_2 + q_3 - q_4 + 2q_1^2 + 3q_2^2 - 5q_1q_2 \tag{2.113}$$
$$\mu_2 = 6(q_1 - q_2)^2 + 4q_3 - 3q_4. \tag{2.114}$$

To conclude the proof of Theorem 2.2.8, it is enough to replace the above relations in Eq.(2.112). We obtain

$$\left|q_n - \frac{6(q_1 - q_2)^2 + 4q_3 - 3q_4}{(1 + q_1 - q_2 + q_3 - q_4 + 2q_1^2 + 3q_2^2 - 5q_1q_2)^n}\right| \leq n\Delta_1(1 - q_1)^3, \tag{2.115}$$

where Δ_1 is given by

$$\Delta_1 = \Delta_1(q_1, n) = K(1 - p_1) + \frac{\Gamma(1 - q_1)}{n}. \quad \blacksquare$$

2.3.7 Proof of Theorem 2.2.9

For the proof of Theorem 2.2.9, we use the same approach as in Theorem 2.2.8. From Eq.(2.15) and Eq.(2.19), we get

$$\begin{aligned}\left|q_n - \frac{\nu_1}{\nu_2^n}\right| &\leq \left|q_n - \frac{\nu_1}{\lambda^n}\right| + \left|\frac{\nu_1}{\lambda^n} - \frac{\nu_1}{\nu_2^n}\right| \\ &\leq (3 + p_1\Gamma(p_1))p_1^2 + n\nu_1\frac{|\lambda - \nu_2|}{\lambda\nu_2} \\ &\leq [3 + \Gamma(p_1)p_1 + n(1 + p_1K(p_1))]\, p_1^2, \end{aligned} \tag{2.116}$$

where $\nu_1 = 1 - p_2$ and $\nu_2 = 1 + p_1 - p_2 + 2(p_1 - p_2)^2$. We express ν_1 and ν_2 in terms of q's using the relations for p_1 to p_4 from the proof of Theorem 2.2.8, so

$$\nu_1 = 2q_1 - q_2 \tag{2.117}$$

$$\nu_2 = 1 + q_1 - q_2 + 2(q_1 - q_2)^2. \tag{2.118}$$

Substituting these formulas in Eq.(2.116), we obtain

$$\left|q_n - \frac{2q_1 - q_2}{[1 + q_1 - q_2 + 2(q_1 - q_2)^2]^n}\right| \leq n\Delta_2(1 - q_1)^2, \tag{2.119}$$

where

$$\Delta_2 = \Delta_2(q_1, n) = 1 + \frac{3}{n} + \left[K(1 - q_1) + \frac{\Gamma(1 - q_1)}{n}\right](1 - q_1). \quad \blacksquare$$

2.3.8 Proof of Proposition 2.1.4

From the stationarity of the sequence $(W_k)_{k\in\mathbb{Z}}$, we know that the autocovariance function

$$c(j) = Cov(W_0, W_j) \tag{2.120}$$

is nonnegative definite. Using that $\mathbb{E}[W_0] = 0$ and the m-dependence property, we obtain that $c(j) = 0$ for all $|j| > m$.
We associate to the sequence of autocovariances $\{c(j)\}_{j=-m}^{j=m}$ the trigonometric polynomial

$$t(\theta) = \sum_{j=-m}^{m} c(j)e^{ij\theta}. \tag{2.121}$$

From the above observations we have that $t(\theta)$ takes only positive real values for all θ. This remark will constitute the key behind the proof. The following lemma, due to Fejér and Riesz (see [Riesz and Nagy, 1990, pag 117] for a proof), will elucidate this statement:

Lemma 2.3.7 (Fejér-Riesz). *Let the trigonometric polynomial*

$$p(x) = \sum_{k=-n}^{n} c_k e^{ikx}.$$

If $p(x) \geq 0$ for all values of x, then there exists a sequence $\{b_j\}_{j=0}^{j=n}$ such that

$$p(x) = \left| \sum_{j=0}^{n} b_j e^{ijx} \right|^2 .$$

Since the trigonometric polynomial t defined in Eq.(2.121) verifies the hypothesis of Fejér-Riesz lemma, there is a sequence $\{a_j\}_{j=0}^{j=m}$ such that we can write

$$\sum_{j=-m}^{m} c(j) e^{ij\theta} = \left| \sum_{k=0}^{m} a_k e^{ik\theta} \right|^2 . \tag{2.122}$$

By expanding Eq.(2.122) and identifying the coefficients of $e^{ij\theta}$, we obtain for each $j \in \{0, 1, \ldots, m\}$

$$c(j) = \sum_{k=0}^{m-j} a_k a_{k+j}. \tag{2.123}$$

Now consider the sequence of moving averages of order m

$$Y_j = \sum_{k=0}^{m} a_k \eta_{j-k}, \tag{2.124}$$

where $\eta_j \sim \mathcal{N}(0,1)$ are i.i.d. standard normal random variables. Clearly $(Y_j)_{j\in\mathbb{Z}}$ is a stationary, m-dependent Gaussian sequence. Since $(W_k)_{k\in\mathbb{Z}}$ is in addition a Gaussian sequence, it remains to verify that

$$Cov(Y_0, Y_j) = \sum_{k=0}^{m-j} a_k a_{k+j}. \tag{2.125}$$

But it is an exercise to observe that

$$\begin{aligned} Cov(Y_0, Y_j) &= \mathbb{E}\left[\sum_{k,l} a_k a_l \eta_{-k} \eta_{j-l} \right] \\ &= \sum_{k,l} a_k a_l \mathbb{E}[\eta_{-k} \eta_{j-l}] \\ &= \sum_{k=0}^{m-j} a_k a_{k+j} \end{aligned} \tag{2.126}$$

and the proof is complete. ■

Chapter 3

Distribution of scan statistics viewed as maximum of 1-dependent stationary sequences

In this chapter we consider the general case of the discrete scan statistics in d dimensions, $d \geq 1$. After introducing the principal notions in Section 3.1, we present in Section 3.2 the methodology used for finding the approximation for the distribution of the d dimensional discrete scan statistics. The advantage of the described method is that we can also establish sharp error bounds for the estimation. Section 3.3 shows how to evaluate these errors. Since the simulation process plays an important role in our approach, in Section 3.4 we present a general importance sampling algorithm that increase the efficiency of the proposed approximation. To investigate the accuracy of our estimation, in Section 3.5 we explicit the formulas for the approximation and the error bounds for the special cases of one, two and three dimensional scan statistics and perform a series of numerical applications. We also compare our approximation with some of the existing ones presented in Chapter 1. The work presented in this chapter, for the particular case of three dimensional scan statistics, makes the object of the article [Amărioarei and Preda, 2013a].

Contents

3.1 Definitions and notations

Let T_1, T_2, ..., T_d be positive integers with $d \geq 1$ and consider the d-dimensional rectangular region, $\mathcal{R}_d$, defined by

$$\mathcal{R}_d = [0, T_1] \times [0, T_2] \times \cdots \times [0, T_d]. \tag{3.1}$$

For $1 \leq s_j \leq T_j$, $j \in \{1, 2, \ldots, d\}$, we can associate to each elementary rectangular region

$$r_d(s_1, s_2, \ldots, s_d) = [s_1 - 1, s_1] \times [s_2 - 1, s_2] \times \cdots \times [s_d - 1, s_d] \tag{3.2}$$

a real valued random variable $X_{s_1,s_2,\ldots,s_d}$. Notice that one can imagine the region $\mathcal{R}_d$ as a d-dimensional lattice (characterized by the centers of the elementary subregions), such that to each point of the lattice it corresponds a random variable.
Let $2 \leq m_j \leq T_j$, $1 \leq j \leq d$, be positive integers and define the random variables

$$Y_{i_1,i_2,\ldots,i_d} = \sum_{s_1=i_1}^{i_1+m_1-1} \sum_{s_2=i_2}^{i_2+m_2-1} \cdots \sum_{s_d=i_d}^{i_d+m_d-1} X_{s_1,s_2,\ldots,s_d}, \tag{3.3}$$

where $1 \leq i_l \leq T_l - m_l + 1$, $1 \leq l \leq d$. The random variables $Y_{i_1,i_2,\ldots,i_d}$ associate to each rectangular region comprised of $m_1 m_2 \ldots m_d$ adjacent elementary subregions,

$$\mathcal{R}(i_1, i_2, \ldots, i_d) = [i_1 - 1, i_1 + m_1 - 1] \times \cdots \times [i_d - 1, i_d + m_d - 1], \tag{3.4}$$

a numerical value equal with the sum of all the attributes in that region.
For example, consider the case when the random variables $X_{s_1,s_2,\ldots,s_d}$ are $0-1$ Bernoulli of parameter p. The value 1 can correspond to the situation when in the elementary rectangular region $r_d(s_1, s_2, \ldots, s_d)$, an event of interest has been observed and the value 0, otherwise. In this framework, the random variables $Y_{i_1,i_2,\ldots,i_d}$ give the number of events observed in the rectangular region $\mathcal{R}(i_1, i_2, \ldots, i_d)$.
We define the *d-dimensional discrete scan statistics* as the maximum number of events in any rectangular region $\mathcal{R}(i_1, i_2, \ldots, i_d)$ within the region $\mathcal{R}_d$,

$$S_{\mathbf{m}}(\mathbf{T}) = \max_{\substack{1 \leq i_j \leq T_j - m_j + 1 \\ j \in \{1,2,\ldots,d\}}} Y_{i_1,i_2,\ldots,i_d}, \tag{3.5}$$

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

Remark 3.1.1. *The random variable $S_{\mathbf{m}}(\mathbf{T})$ defined in Eq.(3.5) can be viewed as an extension of the one dimensional scan statistics presented in [Glaz et al., 2001] ($d = 1$), the two dimensional scan statistics introduced in [Chen and Glaz, 1996] ($d = 2$) and of the three dimensional scan statistics studied in [Guerriero et al., 2010a] and [Amărioarei and Preda, 2013a] ($d = 3$).*

The distribution of the scan statistics,

$$Q_{\mathbf{m}}(\mathbf{T}) = \mathbb{P}\left(S_{\mathbf{m}}(\mathbf{T}) \leq n\right), \tag{3.6}$$

has been used in [Chen and Glaz, 1996] and [Guerriero et al., 2009] for the two dimensional case and in [Guerriero et al., 2010a] and [Amărioarei and Preda, 2013a] for the three dimensional one, to test the null hypothesis of randomness against an alternative of clustering. Under the null hypotheses, H_0, it is assumed that the random variables $X_{s_1,s_2,\ldots,s_d}$ are independent and identically distributed according to some specified distribution. For the alternative hypothesis of clustering, one can specify a region $\mathcal{R}(i_1, i_2, \ldots, i_d)$ where the random variables $X_{s_1,s_2,\ldots,s_d}$ have a larger mean than outside this region. As an example, consider $d = 2$ and the binomial model. The null hypothesis, H_0, assumes that X_{s_1,s_2}'s are i.i.d. with $X_{s_1,s_2} \sim Bin(r, p)$ whereas the alternative hypothesis of clustering, H_1, assumes the existence of a rectangular subregion $\mathcal{R}(i_0, j_0)$ such that for any $i_0 \leq s_1 \leq i_0 + m_1 - 1$ and $j_0 \leq s_2 \leq j_0 + m_2 - 1$, X_{s_1,s_2} are i.i.d. binomial random variables with parameters r and $p' > p$. Outside the region $\mathcal{R}(i_0, j_0)$, X_{s_1,s_2} are i.i.d. distributed according to the distribution specified by the null hypothesis. The generalized likelihood ratio test rejects H_0 in favor of the local change alternative H_1, whenever $S_{\mathbf{m}}(\mathbf{T})$ exceeds the threshold τ, determined from $\mathbb{P}(S_{\mathbf{m}}(\mathbf{T}) \geq \tau | H_0) = \alpha$ and where α represents the significance level of the testing procedure ([Glaz et al., 2001, Chapter 13]).

3.2 Methodology

As noted in Chapter 1 (see also [Cressie, 1993, pag 313]), finding the exact distribution of the d-dimensional scan statistics for $d \geq 2$ has proved elusive. In this section, we present an alternative method, based on the results derived in Chapter 2, for finding accurate approximations for the distribution $Q_{\mathbf{m}}(\mathbf{T})$ of the d-dimensional scan statistics generated by an i.i.d. sequence. One of the main features of this approach is that, besides the approximation formula, it also provides sharp error bounds. It is also important to mention that the method can be applied for any distribution, discrete or continuous, of the random variables $X_{s_1,s_2,\ldots,s_d}$. In Section 3.5, we include numerical results to emphasize this remark.

This approach is not new and was successfully used to approximate the distribution of scan statistics, both in discrete and continuous cases, in a series of articles: for one-dimensional case in [Haiman, 2000] and [Haiman, 2007], for two-dimensional case in [Haiman and Preda, 2002] and [Haiman and Preda, 2006] and for the three-dimensional case in [Amărioarei and Preda, 2013a].

Consider the framework described in Section 3.1 and let the sequence of random variables $X_{s_1,s_2,\ldots,s_d}$ be independent and identically distributed according to a specified distribution. The key idea behind the approximation method is to express the scan statistic random variable $S_{\mathbf{m}}(\mathbf{T})$ as maximum of a 1-dependent stationary sequence of random variables and to employ the estimate from Theorem 2.2.9, Chapter 2. The approximation will be carried out in a series of steps.

Assume that $L_j = \frac{T_j}{m_j - 1}$, $j \in \{1, 2, \ldots, d\}$, are positive integers and define for each

$k_1 \in \{1, 2, \dots, L_1 - 1\}$ the random variables

$$Z_{k_1} = \max_{\substack{(k_1-1)(m_1-1)+1 \le i_1 \le k_1(m_1-1) \\ 1 \le i_j \le (L_j-1)(m_j-1) \\ j \in \{2,\dots,d\}}} Y_{i_1,i_2,\dots,i_d}. \tag{3.7}$$

We observe that the random variable Z_{k_1} defined in Eq.(3.7) corresponds, in fact, to the d-dimensional discrete scan statistics over the multidimensional rectangular strip

$$[(k_1 - 1)(m_1 - 1), (k_1 + 1)(m_1 - 1)] \times [0, T_2] \times \cdots \times [0, T_d].$$

We claim that the set of random variables $\{Z_1, \dots, Z_{L_1-1}\}$ forms a 1-dependent stationary sequence according to Definition 2.1.1 and Definition 2.1.2. Indeed, from Eq.(3.7) we notice that for $k_1 \ge 1$

$$\sigma(Z_1, \cdots, Z_{k_1}) \subset \sigma\left(\{X_{s_1,s_2,\dots,s_d} | 1 \le s_1 \le (k_1 + 1)(m_1 - 1), 1 \le s_j \le T_j, j \ge 2\}\right)$$

and

$$\sigma(Z_{k_1+2}, \cdots) \subset \sigma\left(\{X_{s_1,s_2,\dots,s_d} | (k_1 + 1)(m_1 - 1) + 1 \le s_1, 1 \le s_j \le T_j, j \ge 2\}\right).$$

The independence of the sequence $X_{s_1,s_2,\dots,s_d}$ implies that the σ-fields $\sigma(\cdots, Z_{k_1})$ and $\sigma(Z_{k_1+2}, \cdots)$ are independent, so the 1-dependence of the set of random variables $\{Z_1, \dots, Z_{L_1-1}\}$ is verified. The stationarity is immediate since the random variables $X_{s_1,s_2,\dots,s_d}$ are also identically distributed.
For a better understanding, we consider the three dimensional setting ($d = 3$) to exemplify our approach. In Figure 3.1 we illustrate the sequence $(Z_{k_1})_{k_1=1}^{L_1-1}$, emphasizing its 1-dependent structure.
Notice that from Eq.(3.7) and the definition of the d-dimensional scan statistics in Eq.(3.5) we have the following identity

$$\begin{aligned} S_{\mathbf{m}}(\mathbf{T}) &= \max_{\substack{1 \le i_j \le T_j - m_j + 1 \\ j \in \{1,2,\dots,d\}}} Y_{i_1,i_2,\dots,i_d} \\ &= \max_{1 \le k_1 \le L_1 - 1} \left(\max_{\substack{(k_1-1)(m_1-1)+1 \le i_1 \le k_1(m_1-1) \\ 1 \le i_j \le (L_j-1)(m_j-1) \\ j \in \{2,\dots,d\}}} Y_{i_1,i_2,\dots,i_d} \right) \\ &= \max_{1 \le k_1 \le L_1 - 1} Z_{k_1}. \end{aligned} \tag{3.8}$$

The above relation shows that the scan statistic random variable can be expressed as the maximum of a 1-dependent stationary sequence and gives us the proper setting for applying the estimates developed in the previous chapter.
Recall from Chapter 2 that given a 1-dependent stationary sequence of random variables $(W_k)_{k \ge 1}$, if the tail distribution $\mathbb{P}(W_1 > x)$ is small enough, then we have the following estimate for the distribution of the maximum of the first m terms:

$$\left| q_m - \frac{2q_1 - q_2}{[1 + q_1 - q_2 + 2(q_1 - q_2)^2]^m} \right| \le mF(q_1, m)(1 - q_1)^2, \tag{3.9}$$

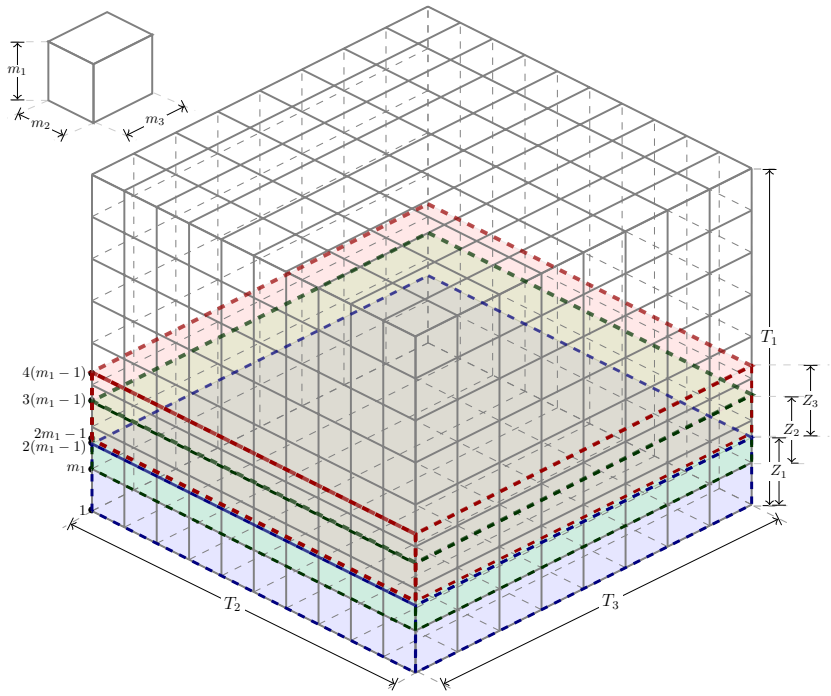


Figure 3.1: Illustration of Z_{k_1} in the case of $d = 3$, emphasizing the 1-dependence

where $q_m = \mathbb{P}(\max\{W_1, \ldots, W_m\} \leq x)$,

$$F(q_1, m) = 1 + \frac{3}{m} + \left[K(1-q_1) + \frac{\Gamma(1-q_1)}{m}\right](1-q_1), \tag{3.10}$$

and $\Gamma(\cdot)$ and $K(\cdot)$ are as in Theorem 2.2.6 (see Theorem 2.2.9). Notice that $F(x, n)$, described above, corresponds to the error coefficient $\Delta_2(x, n)$, defined by Eq.(2.23). Define for $t_1 \in \{2, 3\}$,

$$Q_{t_1} = Q_{t_1}(n) = \mathbb{P}\left(\bigcap_{k_1=1}^{t_1-1}\{Z_{k_1} \leq n\}\right) = \mathbb{P}\left(\max_{\substack{1\leq i_1\leq (t_1-1)(m_1-1)\\ 1\leq i_j\leq (L_j-1)(m_j-1)\\ j\in\{2,\ldots,d\}}} Y_{i_1,i_2,\ldots,i_d} \leq n\right). \tag{3.11}$$

It is clear that Q_{t_1} coincides with $Q_{\mathbf{m}}((t_1(m_1-1), T_2, \ldots, T_d))$, the distribution of the d-dimensional scan statistics over the rectangular region $[0, t_1(m_1-1)] \times [0, T_2] \times \cdots \times [0, T_d]$ (see also Figure 3.1 when $d = 3$).
For n such that $Q_2(n) \geq 0.9$, we can apply the result in Theorem 2.2.9 (enounced above) to obtain the first step approximation

$$\left|Q_{\mathbf{m}}(\mathbf{T}) - \frac{2Q_2 - Q_3}{[1 + Q_2 - Q_3 + 2(Q_2 - Q_3)^2]^{L_1-1}}\right| \leq (L_1 - 1)F(Q_2, L_1 - 1)(1 - Q_2)^2. \tag{3.12}$$

In order to evaluate the approximation in Eq.(3.12), one has to find suitable estimates for the quantities Q_2 and Q_3. To simplify the results of the presentation, in what follows we abbreviate the approximation formula by

$$H(x,y,m)=\frac{2x-y}{[1+x-y+2(x-y)^2]^{m-1}}. \tag{3.13}$$

The second step in our approximation consists in defining two new sequences of 1-dependent stationary random variables, such that each Q_{t_1} can be expressed as the distribution of the maximum of the variables in the corresponding sequence. We define, as in Eq.(3.7), for $t_1\in\{2,3\}$ and $k_2\in\{1,2,\dots,L_2-1\}$ the random variables

$$Z_{k_2}^{(t_1)}=\max_{\substack{1\le i_1\le (t_1-1)(m_1-1)\\ (k_2-1)(m_2-1)+1\le i_2\le k_2(m_2-1)\\ 1\le i_j\le (L_j-1)(m_j-1)\\ j\in\{3,\dots,d\}}} Y_{i_1,i_2,\dots,i_d}. \tag{3.14}$$

Observe that the random variables $Z_{k_2}^{(t_1)}$ coincide with the d-dimensional scan statistics across the overlapping strips of size $t_1(m_1-1)\times 2(m_2-1)\times T_3\times\cdots\times T_d$

$$[0,t_1(m_1-1)]\times[(k_2-1)(m_2-1),(k_2+1)(m_2-1)]\times\cdots\times[0,T_d].$$

Based on similar arguments as in the case of $(Z_{k_1})_{k_1=1}^{L_1-1}$, the random variables in the sets $\{Z_1^{(t_1)},Z_1^{(t_1)},\dots,Z_{L_2-1}^{(t_1)}\}$ are 1-dependent and stationary. Moreover, for each $t_1\in\{2,3\}$ we have

$$Q_{t_1}=\mathbb{P}\left(\max_{1\le k_2\le L_2-1} Z_{k_2}^{(t_1)}\le n\right). \tag{3.15}$$

Set for $t_1,t_2\in\{2,3\}$,

$$Q_{t_1,t_2}=Q_{t_1,t_2}(n)=\mathbb{P}\left(\bigcap_{k_2=1}^{t_2-1}\{Z_{k_2}^{(t_1)}\le n\}\right)=\mathbb{P}\left(\max_{\substack{1\le i_1\le (t_1-1)(m_1-1)\\ 1\le i_2\le (t_2-1)(m_2-1)\\ 1\le i_j\le (L_j-1)(m_j-1)\\ j\in\{3,\dots,d\}}} Y_{i_1,i_2,\dots,i_d}\le n\right) \tag{3.16}$$

and observe that $Q_{t_1,t_2}=Q_{\mathbf{m}}\left((t_1(m_1-1),t_2(m_2-1),\dots,T_d)\right)$.
Whenever n is such that $Q_{t_1,2}(n)\ge 0.9$, we are in the hypothesis of Theorem 2.2.9 and for each $t_1\in\{2,3\}$, we approximate Q_{t_1} with

$$|Q_{t_1}-H\left(Q_{t_1,2},Q_{t_1,3},L_2\right)|\le(L_2-1)F(Q_{t_1,2},L_2-1)(1-Q_{t_1,2})^2. \tag{3.17}$$

Combining the estimate in Eq.(3.12) with the ones in Eq.(3.17), we obtain an expression for the approximation of $Q_{\mathbf{m}}(\mathbf{T})$ depending on the four distribution functions $Q_{2,2}$, $Q_{3,2}$, $Q_{2,3}$ and $Q_{3,3}$.
Depending on the dimension of the problem, the above procedure can be repeated for a number of steps (at most d) to get simpler terms in the approximation formula. In

general, at the s step, $1 \leq s \leq d$, the problem is to approximate for each $t_j \in \{2,3\}$, $j \in \{1,\ldots,s-1\}$, the distribution function of the d-dimensional scan statistics over the rectangular region

$$[0, t_1(m_1-1)] \times \ldots [0, t_{s-1}(m_{s-1}-1)] \times [0, T_s] \cdots \times [0, T_d].$$

We adopt the following notation for these distribution functions:

$$Q_{t_1,t_2,\ldots,t_{s-1}} = Q_{t_1,t_2,\ldots,t_{s-1}}(n) = \mathbb{P}\left(\max_{\substack{1 \leq i_l \leq (t_l-1)(m_l-1) \\ l \in \{1,\ldots,s-1\} \\ 1 \leq i_j \leq (L_j-1)(m_j-1) \\ j \in \{s,\ldots,d\}}} Y_{i_1,i_2,\ldots,i_d} \leq n \right). \tag{3.18}$$

As described in the first two steps, the idea is to define for each point $(t_1,\ldots,t_{s-1}) \in \{2,3\}^{s-1}$ a set of stationary and 1-dependent random variables such that $Q_{t_1,t_2,\ldots,t_{s-1}}$ corresponds to the distribution of the maxima of these variables. We will focus on reducing the size of the s-th dimension.
Define for $t_l \in \{2,3\}$, $l \in \{1,\ldots,s-1\}$ and $k_s \in \{1,2,\ldots,L_s-1\}$ the random variables

$$Z_{k_s}^{(t_1,t_2,\ldots,t_{s-1})} = \max_{\substack{1 \leq i_l \leq (t_l-1)(m_l-1) \\ l \in \{1,2,\ldots,s-1\} \\ (k_s-1)(m_s-1)+1 \leq i_s \leq k_s(m_s-1) \\ 1 \leq i_j \leq (L_j-1)(m_j-1) \\ j \in \{s+1,\ldots,d\}}} Y_{i_1,i_2,\ldots,i_d}. \tag{3.19}$$

Based on the same arguments as for the first and the second step, the random variables $\left\{ Z_1^{(t_1,t_2,\ldots,t_{s-1})}, \ldots, Z_{L_s-1}^{(t_1,t_2,\ldots,t_{s-1})} \right\}$ form a 1-dependent stationary sequence and verify the following relation

$$Q_{t_1,t_2,\ldots,t_{s-1}} = \mathbb{P}\left(\max_{1 \leq k_s \leq L_s-1} Z_{k_s}^{(t_1,t_2,\ldots,t_{s-1})} \leq n \right). \tag{3.20}$$

Clearly, from Eq.(3.18) and Eq.(3.19) we have

$$Q_{t_1,t_2,\ldots,t_s} = Q_{t_1,t_2,\ldots,t_s}(n) = \mathbb{P}\left(\bigcap_{k_s=1}^{t_s-1} \{ Z_{k_s}^{(t_1,t_2,\ldots,t_{s-1})} \leq n \} \right). \tag{3.21}$$

If we take n such that $Q_{t_1,t_2,\ldots,t_{s-1},2}(n) \geq 0.9$, that is we can apply Theorem 2.2.9, then the s step approximation is given by

$$\begin{aligned} \left| Q_{t_1,\ldots,t_{s-1}} - H\left(Q_{t_1,\ldots,t_{s-1},2}, Q_{t_1,\ldots,t_{s-1},3}, L_s\right) \right| &\leq (L_s-1) F(Q_{t_1,\ldots,t_{s-1},2}, L_s-1) \\ &\quad \times (1 - Q_{t_1,\ldots,t_{s-1},2})^2. \end{aligned} \tag{3.22}$$

Substituting for each $s \in \{2,\ldots,d\}$, the Eqs. (3.21) in Eq. (3.12), we get an approximation formula for the distribution of the d-dimensional scan statistics depending on the 2^d quantities $Q_{t_1,\ldots,t_d}$, that we propose to be evaluated by simulation. To get a better filling of the approximation process described above, we include in Fig. 3.2 a diagram that illustrates the steps involved in the approximation of Q_2 for the three dimensional scan statistics.

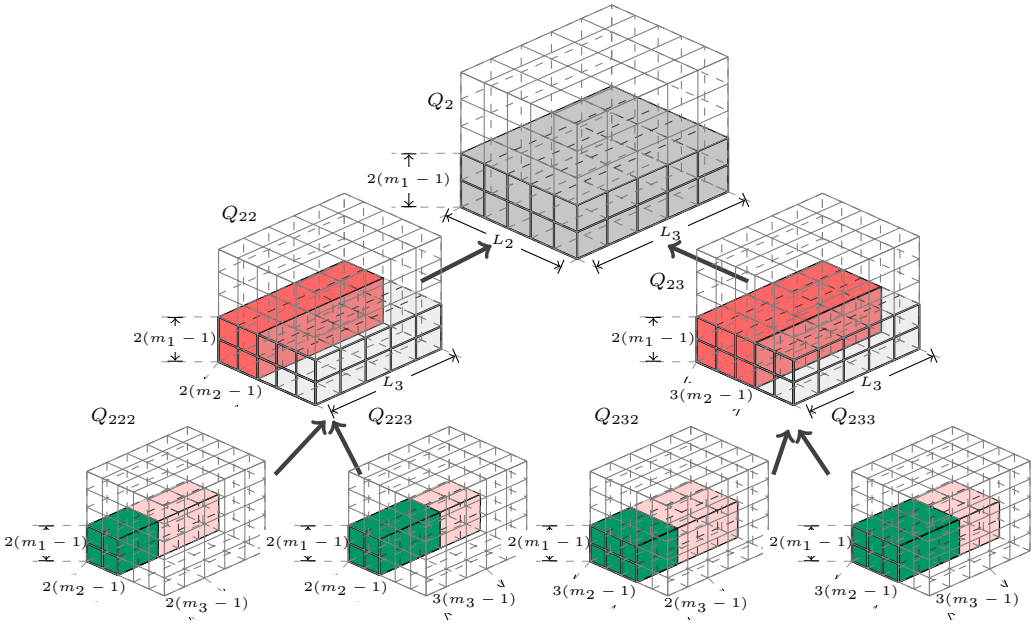


Figure 3.2: Illustration of the approximation of Q_2 in three dimensions

Remark 3.2.1. *If there are indices $j \in \{1, 2, \ldots, d\}$ such that T_j are not multiples of $m_j - 1$, then we take $L_j = \left\lfloor \frac{T_j}{m_j - 1} \right\rfloor$. Based on the inequalities*

$$\mathbb{P}(S_{\mathbf{m}}(\mathbf{M}_1) \leq n) \leq Q_{\mathbf{m}}(\mathbf{T}) \leq \mathbb{P}(S_{\mathbf{m}}(\mathbf{M}_2) \leq n), \tag{3.23}$$

with

$$\mathbf{M}_1 = ((L_1 + 1)(m_1 - 1), \ldots, (L_d + 1)(m_d - 1)),$$
$$\mathbf{M}_2 = (L_1(m_1 - 1), \ldots, L_d(m_d - 1)),$$

we can approximate the distribution of the scan statistics $Q_{\mathbf{m}}(\mathbf{T})$ by the following multi-linear interpolation procedure:
Suppose that we have a function $y = f(x_1, \ldots, x_d)$ and we are given 2^d points $x_{1,s} \leq x_{2,s}$, $s \in \{1, \ldots, d\}$, that are the vertices of a right rectangular polytope. We also assume that the values of the function in the vertices,

$$v_{i_1,\ldots,i_d} = f(x_{i_1,1}, \ldots, x_{i_d,d}),\ i_j \in \{1, 2\},\ j \in \{1, \ldots, d\},$$

are known. Given a point $(\bar{x}_1, \ldots, \bar{x}_d)$ in the interior of the polytope, we can approximate the value of $\bar{v} = f(\bar{x}_1, \ldots, \bar{x}_d)$ based on the following formula:

$$\bar{v} = \sum_{i_1,\ldots,i_d \in \{1,2\}} v_{i_1,\ldots,i_d} V_{i_1,\ldots,i_d},$$

where the normalized volume $V_{i_1,\ldots,i_d}$ is given by

$$V_{i_1,\ldots,i_d} = \frac{\alpha(i_1, 1) \cdots \alpha(i_d, d)}{(x_{2,1} - x_{1,1}) \cdots (x_{2,d} - x_{1,d})},$$

and

$$\alpha(i_s, s) = \begin{cases} x_{2,s} - \bar{x}_s & , \text{ if } i_s = 1 \\ \bar{x}_s - x_{1,s} & , \text{ if } i_s = 2. \end{cases}$$

Observe that in the particular case of one dimension ($d = 1$), if we are given $x_1 \leq \bar{x} \leq x_2$, $y_1 = f(x_1)$ and $y_2 = f(x_2)$ then, for $\bar{y} = f(\bar{x})$, the foregoing relations reduces to

$$\bar{y} = y_1 \frac{x_2 - \bar{x}}{x_2 - x_1} + y_2 \frac{\bar{x} - x_1}{x_2 - x_1},$$

which is exactly the linear interpolation formula.

3.3 Computation of the approximation and simulation errors

In this section, we present the derivation of the errors resulted from the approximation process described in Section 3.2. To see from where the errors appear, it is convenient to introduce some notations.
Let $\gamma_{t_1,\ldots,t_d} = Q_{t_1,\ldots,t_d}$, with $t_j \in \{2, 3\}$, $j \in \{1, \ldots, d\}$, and define for $2 \leq s \leq d$

$$\gamma_{t_1,\ldots,t_{s-1}} = H\left(\gamma_{t_1,\ldots,t_{s-1},2}, \gamma_{t_1,\ldots,t_{s-1},3}, L_s\right). \tag{3.24}$$

Since there are no exact formulas available for the computation of $Q_{t_1,\ldots,t_d}$ for $d \geq 2$, these quantities will be estimated by Monte Carlo simulation. Denote with $\hat{Q}_{t_1,\ldots,t_d}$ the estimated value of $Q_{t_1,\ldots,t_d}$ and define for $2 \leq s \leq d$

$$\hat{Q}_{t_1,\ldots,t_{s-1}} = H\left(\hat{Q}_{t_1,\ldots,t_{s-1},2}, \hat{Q}_{t_1,\ldots,t_{s-1},3}, L_s\right), \tag{3.25}$$

the estimated value of $Q_{t_1,\ldots,t_{s-1}}$.
Our goal is to approximate $Q_{\mathbf{m}}(\mathbf{T})$ with $H\left(\hat{Q}_2, \hat{Q}_3, L_1\right)$ and to find the corresponding error bounds. We observe that

$$\begin{aligned} \left|Q_{\mathbf{m}}(\mathbf{T}) - H\left(\hat{Q}_2, \hat{Q}_3, L_1\right)\right| &\leq \left|Q_{\mathbf{m}}(\mathbf{T}) - H\left(\gamma_2, \gamma_3, L_1\right)\right| \\ &\quad + \left|H\left(\gamma_2, \gamma_3, L_1\right) - H\left(\hat{Q}_2, \hat{Q}_3, L_1\right)\right|, \end{aligned} \tag{3.26}$$

which shows that the error bound has two components: an approximation error resulted from the first term in the right hand side of Eq.(3.26) and a simulation error associated with the second term (the simulation error corresponding to the approximation formula). Notice also that the approximation error is of a theoretical interest since it involves only the true values of the quantities $Q_{t_1,\ldots,t_d}$. Due to the simulation nature of the problem, this error is also bounded by what we call: the simulation error corresponding to the approximation error.

3.3.1 Computation of the approximation error

To simplify the presentation and the derivation of the approximation error formula, we define for $2 \leq s \leq d$,

$$F_{t_1,\ldots,t_{s-1}} = F\left(Q_{t_1,\ldots,t_{s-1},2}, L_s - 1\right) \tag{3.27}$$

and $F = F\left(Q_2, L_1 - 1\right)$.
In practice, since the function $F(q,m)$ defined by Eq.(3.10) is slowly decreasing in q, the values $F_{t_1,\ldots,t_{s-1}}$ will be computed by $F\left(\hat{Q}_{t_1,\ldots,t_{s-1},2}, L_s - 1\right)$, and we treat them as known values in the derivation process. Rigorously, these quantities can be bounded above by $F\left(\hat{Q}_{t_1,\ldots,t_{s-1},2} - \varepsilon_{sim}, L_s - 1\right)$, where ε_{sim} is the simulation error that corresponds to $\hat{Q}_{t_1,\ldots,t_{s-1},2}$.
The goal of this section is to find an error bound for the first term in the right hand side of Eq.(3.26), namely for the difference

$$\left|Q_{\mathbf{m}}(\mathbf{T}) - H\left(\gamma_2, \gamma_3, L_1\right)\right|.$$

We have the following lemma (a proof is given in Appendix B):

Lemma 3.3.1. *Let* $H(x,y,m) = \frac{2x-y}{[1+x-y+2(x-y)^2]^{m-1}}$. *If* $y_i \leq x_i$, $i \in \{1,2\}$, *then:*

$$|H(x_1,y_1,m) - H(x_2,y_2,m)| \leq \begin{cases} (m-1)\left[|x_1 - x_2| + |y_1 - y_2|\right], & 3 \leq m \leq 5 \\ (m-2)\left[|x_1 - x_2| + |y_1 - y_2|\right], & m \geq 6. \end{cases} \tag{3.28}$$

Hereinafter, we employ the result from Lemma 3.3.1 (first branch) without restrictions whenever is necessary. This is in agreement with the numerical values considered in Section 3.5. We begin by observing that

$$\begin{aligned} \left|Q_{\mathbf{m}}(\mathbf{T}) - H\left(\gamma_2,\gamma_3,L_1\right)\right| &\leq \left|Q_{\mathbf{m}}(\mathbf{T}) - H\left(Q_2,Q_3,L_1\right)\right| \\ &\quad + \left|H\left(Q_2,Q_3,L_1\right) - H\left(\gamma_2,\gamma_3,L_1\right)\right| \\ &\leq (L_1-1)F\left(1-Q_2\right)^2 + (L_1-1)\sum_{t_1\in\{2,3\}} |Q_{t_1} - \gamma_{t_1}|. \end{aligned} \tag{3.29}$$

Similarly, $|Q_{t_1} - \gamma_{t_1}|$ for $t_1 \in \{2,3\}$ is bounded by

$$\begin{aligned} |Q_{t_1} - \gamma_{t_1}| &\leq |Q_{t_1} - H\left(Q_{t_1,2}, Q_{t_1,3}, L_2\right)| + |H\left(Q_{t_1,2}, Q_{t_1,3}, L_2\right) - H\left(\gamma_{t_1,2}, \gamma_{t_1,3}, L_2\right)| \\ &\leq (L_2-1)F_{t_1}\left(1-Q_{t_1,2}\right)^2 + (L_2-1)\sum_{t_2\in\{2,3\}} |Q_{t_1,t_2} - \gamma_{t_1,t_2}|. \end{aligned} \tag{3.30}$$

Continuing this process, at the s step, with $2 \leq s \leq d$, we have

$$\begin{aligned} \left|Q_{t_1,\ldots,t_{s-1}} - \gamma_{t_1,\ldots,t_{s-1}}\right| &\leq \left|Q_{t_1,\ldots,t_{s-1}} - H\left(Q_{t_1,\ldots,t_{s-1},2}, Q_{t_1,\ldots,t_{s-1},3}, L_s\right)\right| \\ &\quad + \left|H\left(Q_{t_1,\ldots,t_{s-1},2}, Q_{t_1,\ldots,t_{s-1},3}, L_s\right) - H\left(\gamma_{t_1,\ldots,t_{s-1},2}, \gamma_{t_1,\ldots,t_{s-1},3}, L_s\right)\right| \\ &\leq (L_s-1)\left[F_{t_1,\ldots,t_{s-1}}\left(1-Q_{t_1,\ldots,t_{s-1},2}\right)^2 + \sum_{t_s\in\{2,3\}} |Q_{t_1,\ldots,t_s} - \gamma_{t_1,\ldots,t_s}|\right]. \end{aligned} \tag{3.31}$$

Substituting at each step $s \in \{2, \ldots, d\}$, the corresponding Eq.(3.31) into Eq.(3.29), we obtain

$$|Q_{\mathbf{m}}(\mathbf{T}) - H(\gamma_2, \gamma_3, L_1)| \leq \sum_{s=1}^{d} (L_1 - 1) \cdots (L_s - 1) \times \sum_{t_1,\ldots,t_{s-1} \in \{2,3\}} F_{t_1,\ldots,t_{s-1}} \left(1 - Q_{t_1,\ldots,t_{s-1},2}\right)^2, \quad (3.32)$$

where we adopt the convention that $\sum_{t_1,t_0 \in \{2,3\}} x = x$, $F_{t_1,t_0} = F$ and $Q_{t_1,t_0,2} = Q_2$, which implies that the first term in the sum is $(L_1 - 1)F(1 - Q_2)^2$.
To express the bound in Eq.(3.32) only in terms of $\gamma_{t_1,\ldots,t_s}$, with $1 \leq s \leq d$, we introduce the following variables.
Let $B_{t_1\ldots,t_d} = 0$,

$$\begin{aligned} B_{t_1\ldots,t_{d-1}} &= (L_d - 1)F_{t_1,\ldots,t_{d-1}} \left(1 - Q_{t_1,\ldots,t_{d-1},2}\right)^2 \\ &= (L_d - 1)F_{t_1,\ldots,t_{d-1}} \left(1 - \gamma_{t_1,\ldots,t_{d-1},2} + B_{t_1\ldots,t_{d-1},2}\right)^2 \end{aligned} \quad (3.33)$$

and for $2 \leq s \leq d-1$ define

$$B_{t_1\ldots,t_{s-1}} = (L_s - 1)\left[F_{t_1,\ldots,t_{s-1}} \left(1 - \gamma_{t_1,\ldots,t_{s-1},2} + B_{t_1\ldots,t_{s-1},2}\right)^2 + \sum_{t_s \in \{2,3\}} B_{t_1\ldots,t_s}\right]. \quad (3.34)$$

It can be easily verified that

$$|Q_{t_1,\ldots,t_d} - \gamma_{t_1,\ldots,t_d}| = 0 \leq B_{t_1\ldots,t_d} \quad (3.35)$$

and

$$\left|Q_{t_1,\ldots,t_{d-1}} - \gamma_{t_1,\ldots,t_{d-1}}\right| \leq B_{t_1\ldots,t_{d-1}}, \quad (3.36)$$

from the definition of $B_{t_1\ldots,t_d}$ and $B_{t_1\ldots,t_{d-1}}$.
Since

$$1 - Q_{t_1,\ldots,t_{d-1}} \leq 1 - \gamma_{t_1,\ldots,t_{d-1}} + \left|Q_{t_1,\ldots,t_{d-1}} - \gamma_{t_1,\ldots,t_{d-1}}\right|, \quad (3.37)$$

we deduce, by substituting Eq.(3.36) and Eq.(3.37) into Eq.(3.31) and from the recurrence given in Eq.(3.34), that

$$\left|Q_{t_1,\ldots,t_{d-2}} - \gamma_{t_1,\ldots,t_{d-2}}\right| \leq (L_{d-1} - 1)\left[F_{t_1,\ldots,t_{d-2}} \left(1 - \gamma_{t_1,\ldots,t_{d-2},2} + B_{t_1\ldots,t_{d-2},2}\right)^2 + \sum_{t_{d-1} \in \{2,3\}} B_{t_1,\ldots,t_{d-1}}\right] = B_{t_1,\ldots,t_{d-2}}. \quad (3.38)$$

Using mathematical induction and noticing that the relation in Eq.(3.37) remains valid for $2 \leq s \leq d$, we can verify that

$$|Q_{t_1,\ldots,t_s} - \gamma_{t_1,\ldots,t_s}| \leq B_{t_1,\ldots,t_s},\ s \in \{1, \ldots, d\}. \quad (3.39)$$

A simple computation, similar with the one used to obtain Eq.(3.32), lead us to

$$B_{t_1,\ldots,t_{s-1}} = \sum_{h=s}^{d} (L_s - 1)\cdots(L_h - 1) \times \sum_{t_s,\ldots,t_{h-1}\in\{2,3\}} F_{t_1,\ldots,t_{h-1}} \left(1 - \gamma_{t_1,\ldots,t_{h-1},2} + B_{t_1,\ldots,t_{h-1},2}\right)^2, \quad (3.40)$$

for $2 \leq s \leq d$ and with the convention: $\sum_{t_s,t_{s-1}\in\{2,3\}} x = x$.
Substituting Eqs.(3.37) and (3.39) in Eqs.(3.32), we derive the formula for the theoretical approximation error

$$E_{app}(d) = \sum_{s=1}^{d} (L_1 - 1)\cdots(L_s - 1) \times \sum_{t_1,\ldots,t_{s-1}\in\{2,3\}} F_{t_1,\ldots,t_{s-1}} \left(1 - \gamma_{t_1,\ldots,t_{s-1},2} + B_{t_1,\ldots,t_{s-1},2}\right)^2, \quad (3.41)$$

where for $s = 1$: $\sum_{t_1,t_0\in\{2,3\}} x = x$, $F_{t_1,t_0} = F$, $\gamma_{t_1,t_0,2} = \gamma_2$ and $B_{t_1,t_0,2} = B_2$.

3.3.2 Computation of the simulation errors

In this section, we deal with the computation of the simulation errors that appear in our approximation. We employ the notations introduced in Eq.(3.25) for the simulated values corresponding to $Q_{t_1,\ldots,t_s}$ for $s \in \{1,\ldots,d\}$ and $t_j \in \{2,3\}$, $j \in \{1,\ldots,d\}$. As we remarked before, there are two simulation errors resulting from the approximation process: the simulation error associated with the approximation formula (bounding the second term of the right hand side of Eq.(3.24)) and the simulation error due to the theoretical approximation error ($E_{app}(d)$). We treat these errors separately.
Suppose that we have a simulation method to estimate the values of $Q_{t_1,\ldots,t_d}$. Then, between the true and the estimated values, one can always find a relation of the form

$$\left|Q_{t_1,\ldots,t_d} - \hat{Q}_{t_1,\ldots,t_d}\right| \leq \beta_{t_1,\ldots,t_d}, \quad t_j \in \{2,3\},\ j \in \{1,\ldots,d\}. \quad (3.42)$$

If, for example, $ITER$ is the number of iterations used in the Monte Carlo simulation algorithm for the estimation of $Q_{t_1,\ldots,t_d}$, then one can consider the naive bound provided by the Central Limit Theorem with a 95% confidence level (see [Fishman, 1996])

$$\beta_{t_1,\ldots,t_d} = 1.96\sqrt{\frac{\hat{Q}_{t_1,\ldots,t_d}(1 - \hat{Q}_{t_1,\ldots,t_d})}{ITER}}. \quad (3.43)$$

We assume in subsequent that these values are known. To find the simulation error that corresponds to the approximation formula, we observe that, by applying

successively Lemma 3.3.1 to the second term of Eq. (3.30), we get

$$\begin{aligned}
\left|H\left(\gamma_2,\gamma_3,L_1\right)-H\left(\hat{Q}_2,\hat{Q}_3,L_1\right)\right| &\leq (L_1-1)\sum_{t_1\in\{2,3\}}\left|\gamma_{t_1}-\hat{Q}_{t_1}\right| \\
&= (L_1-1)\sum_{t_1\in\{2,3\}}\left|H\left(\gamma_{t_1,2},\gamma_{t_1,3},L_2\right)-H\left(\hat{Q}_{t_1,2},\hat{Q}_{t_1,3},L_2\right)\right| \\
&\leq (L_1-1)(L_2-1)\sum_{t_1,t_2\in\{2,3\}}\left|\gamma_{t_1,t_2}-\hat{Q}_{t_1,t_2}\right| \\
&\vdots \\
&\leq (L_1-1)\ldots(L_{d-1}-1)\sum_{t_1,\ldots,t_{d-1}\in\{2,3\}}\left|\gamma_{t_1,\ldots,t_{d-1}}-\hat{Q}_{t_1,\ldots,t_{d-1}}\right| \\
&\leq (L_1-1)\ldots(L_d-1)\sum_{t_1,\ldots,t_d\in\{2,3\}}\left|Q_{t_1,\ldots,t_d}-\hat{Q}_{t_1,\ldots,t_d}\right|. \qquad (3.44)
\end{aligned}$$

Combining Eq.(3.42) and Eq.(3.44), we obtain the simulation error associated with the approximation formula

$$E_{sf}(d)=(L_1-1)\ldots(L_d-1)\sum_{t_1,\ldots,t_d\in\{2,3\}}\beta_{t_1,\ldots,t_d}. \qquad (3.45)$$

As we will see in the numerical section, this simulation error has the largest contribution to the total error.
In order to find the simulation error corresponding to the approximation error bound given by Eq.(3.41), we need to introduce some notations. Set $A_{t_1,\ldots,t_d}=\beta_{t_1,\ldots,t_d}$ and take for $2\leq s\leq d$

$$A_{t_1,\ldots,t_{s-1}}=(L_s-1)\ldots(L_d-1)\sum_{t_s,\ldots,t_d\in\{2,3\}}\beta_{t_1,\ldots,t_d}. \qquad (3.46)$$

Based on similar arguments as for obtaining Eq.(3.44), we deduce that

$$\left|\hat{Q}_{t_1,\ldots,t_s}-\gamma_{t_1,\ldots,t_s}\right|\leq A_{t_1,\ldots,t_s},\ s\in\{1,\ldots,d\}. \qquad (3.47)$$

Let $C_{t_1\ldots,t_d}=0$ and for $2\leq s\leq d$, define

$$C_{t_1\ldots,t_{s-1}}=(L_s-1)\left[F_{t_1,\ldots,t_{s-1}}\left(1-\hat{Q}_{t_1,\ldots,t_{s-1},2}+A_{t_1\ldots,t_{s-1},2}+C_{t_1\ldots,t_{s-1},2}\right)^2+\sum_{t_s\in\{2,3\}}C_{t_1\ldots,t_s}\right]. \qquad (3.48)$$

We observe that replacing $s=d$ in Eq.(3.48)

$$C_{t_1\ldots,t_{d-1}}=(L_d-1)F_{t_1,\ldots,t_{d-1}}\left(1-\hat{Q}_{t_1,\ldots,t_{d-1},2}+\beta_{t_1\ldots,t_{d-1},2}\right)^2, \qquad (3.49)$$

since $C_{t_1\ldots,t_d} = 0$ and $A_{t_1\ldots,t_{d-1},2} = \beta_{t_1\ldots,t_{s-1},2}$.
As in Eq.(3.37),

$$1 - \gamma_{t_1,\ldots,t_{d-1}} \leq 1 - \hat{Q}_{t_1,\ldots,t_{d-1}} + \left|\hat{Q}_{t_1,\ldots,t_{d-1}} - \gamma_{t_1,\ldots,t_{d-1}}\right|, \tag{3.50}$$

so one can deduce from Eqs.(3.33) and (3.49), that

$$B_{t_1\ldots,t_{d-1}} \leq C_{t_1\ldots,t_{d-1}}. \tag{3.51}$$

From the definition of $B_{t_1\ldots,t_{s-1}}$ and $C_{t_1\ldots,t_{s-1}}$ in Eq.(3.34) and Eq.(3.48) respectively, we conclude, using mathematical induction, that

$$B_{t_1\ldots,t_{s-1}} \leq C_{t_1\ldots,t_{s-1}},\ s \in \{2,\ldots,d\}. \tag{3.52}$$

Clearly, from Eqs.(3.50), (3.47) and (3.52), we have

$$1 - \gamma_{t_1,\ldots,t_{s-1}} + B_{t_1,\ldots,t_{s-1}} \leq 1 - \hat{Q}_{t_1,\ldots,t_{s-1}} + A_{t_1,\ldots,t_{s-1}} + C_{t_1,\ldots,t_{s-1}} \tag{3.53}$$

and the simulation error corresponding to the approximation error follows from Eq.(3.51) and the foregoing equation

$$E_{sapp}(d) = \sum_{s=1}^{d} (L_1 - 1)\cdots(L_s - 1) \sum_{t_1,\ldots,t_{s-1}\in\{2,3\}} F_{t_1,\ldots,t_{s-1}} \Big(1 - \hat{Q}_{t_1,\ldots,t_{s-1},2} + A_{t_1,\ldots,t_{s-1},2} + C_{t_1,\ldots,t_{s-1},2}\Big)^2, \tag{3.54}$$

where for $s = 1$: $\sum_{t_1,t_0\in\{2,3\}} x = x$, $F_{t_1,t_0} = F$, $\hat{Q}_{t_1,t_0,2} = \hat{Q}_2$, $A_{t_1,t_0,2} = A_2$ and $C_{t_1,t_0,2} = C_2$.
The total error is obtained by adding the two simulation error terms from Eq.(3.45) and Eq.(3.54)

$$E_{total}(d) = E_{sf}(d) + E_{sapp}(d). \tag{3.55}$$

To efficiently evaluate Eq.(3.55), one needs to find suitable values for the error bounds $\beta_{t_1,\ldots,t_d}$. The bounds from Eq.(3.43), provided by the Central Limit Theorem, have been used in [Haiman and Preda, 2006] for the two dimensional case. As the authors pointed out, the main contribution to the total error is due to the simulation error $E_{sf}(d)$, especially for small sizes of the window scan with respect to the scanning region. Our numerical study shows that these error bounds are not feasible for the scan problem in more than three dimensions, the simulation error associated with the approximation formula $E_{sf}(d)$ being too large with respect to the other error $E_{sapp}(d)$. Thus, for the simulation of $\hat{Q}_{t_1,\ldots,t_d}$, we use an importance sampling technique introduced in [Naiman and Wynn, 1997]. Next section illustrates how to adapt this simulation method to our problem.

3.4 Simulation using Importance Sampling

Usually, the scan statistic-based tests are employed when dealing with the detection of an unusually large cluster of events (for example detection of a bioterrorist attack, brain tumor, minefield reconnaissance etc.). Normally, in such problems, the practitioner wants to find p values that involve small tail probabilities (smaller than 0.1). Therefore, the problem is to estimate the high order quantiles of the d-dimensional scan statistics. To solve this problem, we propose the approximation developed in Section 3.2. To check the accuracy of our proposed approximation, we need to find a suitable method to estimate the quantities involved in our formula, namely to estimate $Q_{t_1,\dots,t_d}$, $t_j \in \{2,3\}$ and $j \in \{1,\dots,d\}$. For comparison reasons, we also want to find an alternative method to evaluate the scan statistics over the whole region.

A direct approach to this problem is to use the naive hit-or-miss Monte Carlo, as described bellow. Let $\mathbf{X^{(i)}} = \left\{X^{(i)}_{s_1,s_2,\dots,s_d}, s_j \in \{1,\dots,T_j\}, j \in \{1,\dots,d\}\right\}$, with $1 \le i \le ITER$, be $ITER$ independent realizations of the underlying random field (under the null hypothesis). For each realization, we compute the d-dimensional scan statistics $S^{(i)}_{\mathbf{m}}(\mathbf{T})$ and we define

$$\widehat{p_{MC}} = \frac{1}{ITER}\sum_{i=1}^{ITER} \mathbf{1}_{\left\{S^{(i)}_{\mathbf{m}}(\mathbf{T}) \ge \tau\right\}} \tag{3.56}$$

and

$$\widehat{s.e._{MC}} = \sqrt{\frac{\widehat{p_{MC}}(1-\widehat{p_{MC}})}{ITER}}, \tag{3.57}$$

the unbiased direct Monte Carlo estimate of $p = \mathbb{P}\left(S_{\mathbf{m}}(\mathbf{T}) \ge \tau\right)$ and its consistent standard error estimate. As we will see from the numerical results, this approach is computationally intensive since just a fraction of the generated observations will cause a rejection and thus many replications are necessary in order to reduce the standard error estimate to an acceptable level (especially for $d \ge 2$).

In general, there are many variance reduction techniques (see for example [Ross, 2012, Chapters 9 and 10]) that can be used to improve the efficiency of the naive Monte Carlo approach. In the following, we present a variance reduction method based on importance sampling that will provide more accurate estimates.

3.4.1 Generalities on importance sampling

In this subsection, we introduce some basic theoretical aspects of the importance sampling technique. This simulation method is usually employed when dealing with rare event probabilities and often lead to substantial variance reduction (see [Rubino and Tuffin, 2009]). Expositions on importance sampling can be found in [Ross, 2012], [Rubinstein and Kroese, 2008] or [Fishman, 1996].

Suppose that we want to estimate, by Monte Carlo methods, the expectation of a

function $G(W)$ of a random vector W having a joint density function f. Let

$$\theta = \mathbb{E}_f\left[G(W)\right] = \int G(\mathbf{x})f(\mathbf{x})d\mathbf{x} \tag{3.58}$$

be the expectation to be determined. Observe that in Eq.(3.58) we used the notation $\mathbb{E}_f$ with the subscript f to emphasizes that the expectation is taken with respect to the density f.
The naive Monte Carlo approach suggests to draw N independent samples $W^{(1)}$, $W^{(2)}$, ..., $W^{(N)}$ from the density $f(\mathbf{x})$ and to take

$$\widehat{\theta_{MC}} = \frac{1}{N}\sum_{i=1}^{N} G\left(W^{(i)}\right), \tag{3.59}$$

as an estimate for θ. This approach may prove to be ineffective in many situations. One possible cause can be that we cannot simulate random vectors from the distribution of W, another is that the variance of $G(W)$ is too large (see [Ross, 2012] for further discussion).
To overcome such problems, it may be useful to introduce another probability density function g such that Gf is dominated by g, that is if $g(\mathbf{x}) = 0$ then $G(\mathbf{x})f(\mathbf{x}) = 0$ (to keep the estimator unbiased) and to give the following alternative expression for θ:

$$\theta = \int \left[\frac{G(\mathbf{x})f(\mathbf{x})}{g(\mathbf{x})}\right] g(\mathbf{x})d\mathbf{x} = \mathbb{E}_g\left[\frac{G(W)f(W)}{g(W)}\right]. \tag{3.60}$$

Consequently, if now we draw N i.i.d. samples $W^{(1)}$, $W^{(2)}$, ..., $W^{(N)}$ from the density $g(\mathbf{x})$, then

$$\widehat{\theta_{IS}} = \frac{1}{N}\sum_{i=1}^{N} \frac{G\left(W^{(i)}\right) f\left(W^{(i)}\right)}{g\left(W^{(i)}\right)} \tag{3.61}$$

is an unbiased estimator for θ. If the density function g can be chosen such that the random variable $\frac{G(W)f(W)}{g(W)}$ has a small variance (the likelihood ratio $f/g \ll 1$), then $\widehat{\theta}$ is an efficient estimator. In the particular case when one wants to estimate the probability $\theta = \mathbb{P}\left(W \in A\right)$, that is $G(W) = \mathbf{1}_{\{W\in A\}}$, a good choice of g is such that $g \gg f$ on the set A. To see this, we consider the comparison between the exact variances of the estimators in the direct ($\widehat{\theta_{MC}}$) and IS approach ($\widehat{\theta_{IS}}$):

$$\begin{aligned} Var\left[\widehat{\theta_{IS}}\right] &= \frac{1}{N}Var_g\left[\mathbf{1}_{\{W\in A\}}\frac{f(W)}{g(W)}\right] = \frac{1}{N}\left[\mathbb{E}_g\left[\mathbf{1}_{\{W\in A\}}\frac{f^2(W)}{g^2(W)}\right] - \theta^2\right] \\ &\ll \frac{1}{N}\left[\mathbb{E}_g\left[\mathbf{1}_{\{W\in A\}}\frac{f(W)}{g(W)}\right] - \theta^2\right] = \frac{1}{N}\left[\int \mathbf{1}_{\{W\in A\}}f(\mathbf{x})d\mathbf{x} - \theta^2\right] \\ &= \frac{1}{N}Var_f\left[\mathbf{1}_{\{W\in A\}}\right] = Var\left[\widehat{\theta_{MC}}\right]. \end{aligned}$$

In general, finding a good change of measure g that leads to an efficient sampling process can be difficult (see [Rubino and Tuffin, 2009]). Fortunately, for our problem at hand, there are efficient change of measures and we discuss one of them in the next subsection.

3.4.2 Importance sampling for scan statistics

As we saw in the foregoing section, the general idea behind the importance sampling technique is to change the distribution to be sampled from in such a way that the new estimator to remain unbiased. In this subsection, we present an importance sampling approach for the estimation of the significance level of hypothesis tests based on d-dimensional scan statistics. This method was introduced by [Naiman and Priebe, 2001] and was successfully used to solve the problem of exceeding probabilities, that is, the probability that one or more tests statistics exceeds some given threshold. Their methodology builds upon a procedure introduced by [Frigessi and Vercellis, 1984] to solve the union count problem (see also [Fishman, 1996, pag 261]).
Assume that we are in the framework of the d-dimensional scan statistics described in Section 3.1. As we previously saw, the scan statistics $S_{\mathbf{m}}(\mathbf{T})$ random variable is usually employed for testing the null hypothesis of randomness (H_0) against a clustering alternative (H_1). By the generalized likelihood ratio test, the null hypothesis H_0 is rejected in favor of the local change alternative H_1 whenever $S_{\mathbf{m}}(\mathbf{T})$ is sufficiently large (see [Glaz and Naus, 1991, Section 3] for an outline of the proof for $d = 1$ or [Glaz et al., 2001, Chapter 13]). If τ denotes the observed value of the test statistic from the actual data, then we want to find the p-value

$$p = \mathbb{P}_{H_0}\left(S_{\mathbf{m}}(\mathbf{T}) \geq \tau\right). \tag{3.62}$$

The main idea behind the importance sampling approach in the scan statistics setting is to sample only data such that the rejection of H_0 occurs, and then to determine the collection of all the locality statistics that generate a rejection. We describe this method in what follows.
Let $E_{i_1,\ldots,i_d}$, for $1 \leq i_j \leq T_j - m_j + 1$, $j \in \{1,\ldots,d\}$, denote the event that $Y_{i_1,\ldots,i_d}$ exceeds the threshold τ. We are interested in evaluating the probability

$$\mathbb{P}_{H_0}\left(S_{\mathbf{m}}(\mathbf{T}) \geq \tau\right) = \mathbb{P}\left(\bigcup_{i_1=1}^{T_1-m_1+1} \cdots \bigcup_{i_d=1}^{T_d-m_d+1} E_{i_1,\ldots,i_d}\right). \tag{3.63}$$

Under the notations made in the preceding section, the above equation can be rewritten as

$$\theta = \int G(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}, \tag{3.64}$$

where $\theta = \mathbb{P}_{H_0}\left(S_{\mathbf{m}}(\mathbf{T}) \geq \tau\right)$, $G(\mathbf{x}) = \mathbf{1}_E(\mathbf{x})$, $E = \bigcup_{i_1=1}^{T_1-m_1+1} \cdots \bigcup_{i_d=1}^{T_d-m_d+1} E_{i_1,\ldots,i_d}$ and f is the joint density of $Y_{i_1,\ldots,i_d}$ under the null hypothesis.
Let

$$B(d) = \sum_{i_1=1}^{T_1-m_1+1} \cdots \sum_{i_d=1}^{T_d-m_d+1} \mathbb{P}\left(E_{i_1,\ldots,i_d}\right) \tag{3.65}$$

denote the usual Bonferroni upper bound for $\mathbb{P}(E)$. Under the null hypothesis, due to stationarity, this bound becomes

$$B(d) = (T_1 - m_1 + 1) \cdots (T_d - m_d + 1)\mathbb{P}\left(E_{1,\ldots,1}\right). \tag{3.66}$$

The probability in Eq.(3.63) can be expressed as

$$\begin{aligned}
\mathbb{P}_{H_0}\left(S_{\mathbf{m}}\ (\mathbf{T}) \geq \tau\right) &= \mathbb{P}\left(\bigcup_{i_1=1}^{T_1-m_1+1} \cdots \bigcup_{i_d=1}^{T_d-m_d+1} E_{i_1,\ldots,i_d}\right) = \int \mathbf{1}_E d\mathbb{P}_{H_0} \\
&= \int \frac{\mathbf{1}_E}{\sum_{i_1=1}^{T_1-m_1+1} \cdots \sum_{i_d=1}^{T_d-m_d+1} \mathbf{1}_{E_{i_1,\ldots,i_d}}} \sum_{i_1=1}^{T_1-m_1+1} \cdots \sum_{i_d=1}^{T_d-m_d+1} \mathbf{1}_{E_{i_1,\ldots,i_d}} d\mathbb{P}_{H_0} \\
&= \sum_{j_1=1}^{T_1-m_1+1} \cdots \sum_{j_d=1}^{T_d-m_d+1} \int \frac{\mathbf{1}_E}{\sum_{i_1=1}^{T_1-m_1+1} \cdots \sum_{i_d=1}^{T_d-m_d+1} \mathbf{1}_{E_{i_1,\ldots,i_d}}} \mathbf{1}_{E_{j_1,\ldots,j_d}} d\mathbb{P}_{H_0} \\
&= \sum_{j_1=1}^{T_1-m_1+1} \cdots \sum_{j_d=1}^{T_d-m_d+1} \int \frac{1}{C(\mathbf{Y})} \mathbf{1}_{E \cap E_{j_1,\ldots,j_d}} d\mathbb{P}_{H_0} \\
&= \sum_{j_1=1}^{T_1-m_1+1} \cdots \sum_{j_d=1}^{T_d-m_d+1} \mathbb{P}\left(E_{j_1,\ldots,j_d}\right) \int \frac{1}{C(\mathbf{Y})} \frac{\mathbf{1}_{E_{j_1,\ldots,j_d}}}{\mathbb{P}\left(E_{j_1,\ldots,j_d}\right)} d\mathbb{P}_{H_0} \\
&= B(d) \sum_{j_1=1}^{T_1-m_1+1} \cdots \sum_{j_d=1}^{T_d-m_d+1} \frac{\mathbb{P}\left(E_{j_1,\ldots,j_d}\right)}{B(d)} \int \frac{1}{C(\mathbf{Y})} d\mathbb{P}_{H_0}(\cdot | E_{j_1,\ldots,j_d}) \\
&= B(d) \sum_{j_1=1}^{T_1-m_1+1} \cdots \sum_{j_d=1}^{T_d-m_d+1} p_{j_1,\ldots,j_d} \int \frac{1}{C(\mathbf{Y})} d\mathbb{P}_{H_0}(\cdot | E_{j_1,\ldots,j_d}),
\end{aligned} \tag{3.67}$$

where, by stationarity, $p_{j_1,\ldots,j_d}$ defines an uniform probability distribution over $\{1,\ldots,T_1 - m_1 + 1\} \times \cdots \times \{1,\ldots,T_d - m_d + 1\}$,

$$\begin{aligned}
p_{j_1,\ldots,j_d} &= \frac{\mathbb{P}\left(E_{j_1,\ldots,j_d}\right)}{\sum_{i_1=1}^{T_1-m_1+1} \cdots \sum_{i_d=1}^{T_d-m_d+1} \mathbb{P}\left(E_{i_1,\ldots,i_d}\right)} \\
&= \frac{1}{(T_1 - m_1 + 1) \cdots (T_d - m_d + 1)}
\end{aligned} \tag{3.68}$$

and where $C(\mathbf{Y})$ represents the number of d-tuples $(i_1,\ldots,i_d)$ for which exceedance $(Y_{i_1,\ldots,i_d} \geq \tau)$ occurs, that is

$$C(\mathbf{Y}) = \sum_{i_1=1}^{T_1-m_1+1} \cdots \sum_{i_d=1}^{T_d-m_d+1} \mathbf{1}_{E_{i_1,\ldots,i_d}}. \tag{3.69}$$

Remark that from the above identity in Eq.(3.67), the importance sampling change of measure g is a mixture of conditional distributions and is given by

$$g(\mathbf{x}) = \sum_{j_1=1}^{T_1-m_1+1} \cdots \sum_{j_d=1}^{T_d-m_d+1} \left\{ \frac{\mathbb{P}\left(E_{j_1,\ldots,j_d}\right)}{B(d)} \right\} \left\{ \frac{\mathbf{1}_{E_{j_1,\ldots,j_d}} f(\mathbf{x})}{\mathbb{P}\left(E_{j_1,\ldots,j_d}\right)} \right\}. \tag{3.70}$$

From Eq.(3.62) and Eq.(3.67), we observe that our p-value can be expressed as the Bonfferoni conservative bound $B(d)$ times a correction factor $\rho(d)$ between 0 and 1, with

$$\rho(d) = \sum_{j_1=1}^{T_1-m_1+1} \cdots \sum_{j_d=1}^{T_d-m_d+1} p_{j_1,\ldots,j_d} \int \frac{1}{C(\mathbf{Y})} d\mathbb{P}_{H_0}(\cdot | E_{j_1,\ldots,j_d}). \tag{3.71}$$

Notice that one can define a larger class of importance sampling algorithms based on higher order inclusion-exclusion identities, generalizing the foregoing procedure, as provided in [Naiman and Wynn, 1997, Section 4].
The correction factor that appears Eq.(3.71) can be estimated from the following algorithm:

Algorithm 1 Importance Sampling Algorithm for Scan Statistics

Begin

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

1: Generate uniformly the d-tuple $(i_1^{(k)}, \ldots, i_d^{(k)})$ from the set $\{1, \ldots, T_1 - m_1 + 1\} \times \cdots \times \{1, \ldots, T_d - m_d + 1\}$.
2: Given the d-tuple $(i_1^{(k)}, \ldots, i_d^{(k)})$, generate a sample of the random field $\tilde{\mathbf{X}}^{(\mathbf{k})} = \left\{ \tilde{X}^{(k)}_{s_1,s_2,\ldots,s_d} \right\}$, with $s_j \in \{1, \ldots, T_j\}$ and $j \in \{1, \ldots, d\}$, from the conditional distribution of $\mathbf{X}$ given $\left\{ Y_{i_1^{(k)},\ldots,i_d^{(k)}} \geq \tau \right\}$.
3: Take $c_k = C(\tilde{\mathbf{X}}^{(\mathbf{k})})$ the number of all d-tuple $(i_1, \ldots, i_d)$ for which $\tilde{Y}_{i_1,\ldots,i_d} \geq \tau$ and put $\widehat{\rho}_k(d) = \frac{1}{c_k}$.

End Repeat

Return $\widehat{\rho}(d) = \dfrac{1}{ITER} \displaystyle\sum_{k=1}^{ITER} \widehat{\rho}_k(d)$.

End

Clearly, $\widehat{\rho}(d)$ is an unbiased estimator for $\rho(d)$ with estimated variance

$$Var\left[\widehat{\rho}(d)\right] \approx \frac{1}{ITER - 1} \sum_{i=1}^{ITER} \left(\widehat{\rho}_i(d) - \frac{1}{ITER} \sum_{k=1}^{ITER} \widehat{\rho}_k(d) \right)^2. \tag{3.72}$$

For $ITER$ sufficiently large, as a consequence of Central Limit Theorem, the error between the true and the estimated value of the tail $\mathbb{P}\left(S_{\mathbf{m}}(\mathbf{T}) \geq \tau\right)$, corresponding

to a 95% confidence level, is given by

$$\beta = 1.96B(d)\sqrt{\frac{Var\left[\widehat{\rho}(d)\right]}{ITER}}. \tag{3.73}$$

Notice that for the simulation of $Q_{t_1,\ldots,t_d}$, we substitute T_1, $\ldots$, T_d in the above algorithm with $t_1(m_1-1)$, $\ldots$, $t_d(m_d-1)$, respectively. Therefore, we obtain the corresponding values for $\beta_{t_1,\ldots,t_d}$ as described by Eq.(3.73).
In Figure 3.3, we present the difference between the naive Monte Carlo and the Importance Sampling method in the particular case of a three dimensional scan statistics. 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 $p = 0.0001$. For the Monte Carlo approach, we used replications in the range $\{10^6,\ldots,10^7\}$, while for the Importance Sampling algorithm the range was $\{10^5,\ldots,10^6\}$. We observe that in the case of hit and miss Monte Carlo approach the simulation error is rather large, even for 10^7 iterations.

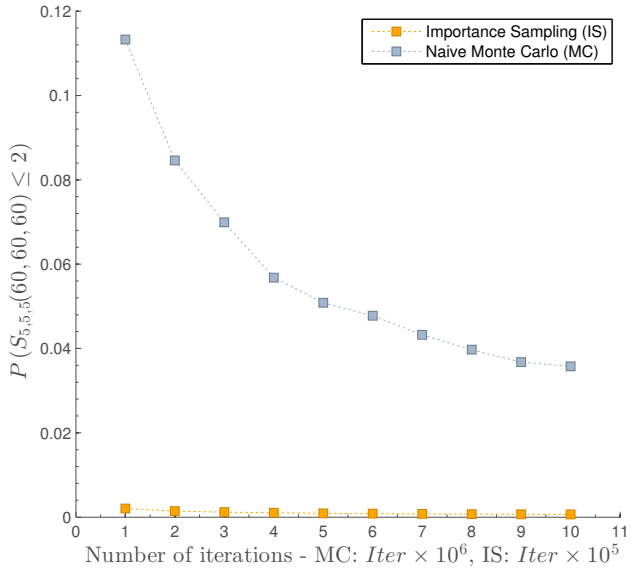


Figure 3.3: The evolution of simulation error in MC and IS methods

3.4.3 Computational aspects

The algorithm discussed in the previous section presents two implementation difficulties: first, it assumes that one is able to efficiently generate the underlying random field $\tilde{\mathbf{X}}$ from the required conditional distribution (see **Step 2**) and second, the number of locality statistics that exceed the predetermined threshold is supposed to be found in a *reasonable* time. We address these problems separately.
The problem of sampling from the conditional distribution in **Step 2** of the algorithm depends on the initial distribution of the i.i.d. random variables $X_{s_1,s_2,\ldots,s_d}$, $s_j \in \{1,\ldots,T_j\}$ and $j \in \{1,\ldots,d\}$. To illustrate the procedure we consider two

examples for the random field distribution: binomial of parameters ν and p and Gaussian with known mean μ and variance σ^2.

Example 3.4.1 (Binomial model). *In this example we consider that $X_{s_1,s_2,\ldots,s_d}$ are i.i.d. binomial $\mathcal{B}(\nu,p)$ random variables. Clearly, the random variables $Y_{i_1,\ldots,i_d}$ are also binomially distributed with parameters $w=\nu m_1\cdots m_d$ and p.*
*The main idea is to generate a sample not from the conditional distribution given $\left\{Y_{i_1^{(k)},\ldots,i_d^{(k)}}\geq\tau\right\}$ as described in **Step 2** of the algorithm, but from the conditional distribution given $\left\{Y_{i_1^{(k)},\ldots,i_d^{(k)}}=t\right\}$, for some value $t\geq\tau$. This can be achieved since, if we define the events $G_{j_1,\ldots,j_d}(t)=\{Y_{j_1,\ldots,j_d}=t\}$, for $t\in\{\tau,\ldots,w\}$ and $1\leq j_i\leq T_i-m_i+1$, $i\in\{1,\ldots,d\}$, then we observe that the events $E_{j_1,\ldots,j_d}$ can be expressed as*

$$E_{j_1,\ldots,j_d}=\bigcup_{t=\tau}^{w}G_{j_1,\ldots,j_d}(t)$$

and the Eq.(3.67) can be rewritten as

$$\begin{aligned}\mathbb{P}_{H_0}\left(\ S_{\mathbf{m}}(\mathbf{T})\geq\tau\right)&=B(d)\sum_{j_1=1}^{T_1-m_1+1}\cdots\sum_{j_d=1}^{T_d-m_d+1}p_{j_1,\ldots,j_d}\int\frac{1}{C(\mathbf{Y})}d\mathbb{P}_{H_0}(\cdot|E_{j_1,\ldots,j_d})\\&=B(d)\sum_{j_1=1}^{T_1-m_1+1}\cdots\sum_{j_d=1}^{T_d-m_d+1}p_{j_1,\ldots,j_d}\sum_{t=\tau}^{w}r_{j_1,\ldots,j_d}(t)\int\frac{1}{C(\mathbf{Y})}\frac{\mathbf{1}_{G_{j_1,\ldots,j_d}(t)}}{\mathbb{P}\left(G_{j_1,\ldots,j_d}(t)\right)}d\mathbb{P}_{H_0}\\&=B(d)\sum_{j_1=1}^{T_1-m_1+1}\cdots\sum_{j_d=1}^{T_d-m_d+1}p_{j_1,\ldots,j_d}\sum_{t=\tau}^{w}r_{j_1,\ldots,j_d}(t)\int\frac{1}{C(\mathbf{Y})}d\mathbb{P}_{H_0}(\cdot|G_{j_1,\ldots,j_d}(t)).\end{aligned}\tag{3.74}$$

The weights $r_{j_1,\ldots,j_d}(t)$ do not depend on the indices $(j_1,\ldots,j_d)$, due to the stationarity of the random field and are computed from the relation

$$r_{j_1,\ldots,j_d}(t)=\frac{\mathbb{P}\left(Y_{j_1,\ldots,j_d}=t\right)}{\mathbb{P}\left(Y_{j_1,\ldots,j_d}\geq\tau\right)}=\frac{\mathbb{P}\left(Y_{1,\ldots,1}=t\right)}{\mathbb{P}\left(Y_{1,\ldots,1}\geq\tau\right)}.\tag{3.75}$$

To generate a sample from the conditional distribution that appears in Eq.(3.74), reduces to show how one can sample uniformly a vector $\mathbf{u}=(u_1,\ldots,u_l)$ of size $l=m_1\cdots m_d$ satisfying $u_1+\cdots+u_l=t$ and $0\leq u_i\leq\nu$, from the set of all such vectors denoted by $\Gamma(l,t,\nu)$. This later aspect can be achieved by the use of urn models. Assume that we have l urns with ν balls each and a null vector $\mathbf{u}$ of length l and we take t (an integer) draws without replacement. At the i-th step ($i\leq t$), we choose uniformly an urn (from the remaining ones, that is the ones that are not empty at this step) and draw a ball without replacement. We add one to the component of the vector $\mathbf{u}$ that corresponds to the index of the chosen urn. We repeat the procedure t steps and we obtain a vector $\mathbf{u}$ whose components take values between 0 and ν and have the sum equal with t.
According to the above observations, the second step in Algorithm 1 can be rewritten in the following way

Step 2a *Generate a threshold value $t \geq \tau$ from the distribution*

$$r_{1,\dots,1}(t) = \frac{\mathbb{P}\left(Y_{1,\dots,1} = t\right)}{\mathbb{P}\left(Y_{1,\dots,1} \geq \tau\right)}.$$

Step 2b *Conditionally, given t and the d-tuple $(i_1^{(k)}, \dots, i_d^{(k)})$, generate $\tilde{X}^{(k)}_{s_1,s_2,\dots,s_d}$ for $i_j^{(k)} \leq s_j \leq i_j^{(k)} + m_j - 1$, $j \in \{1, \dots, d\}$, uniformly from the set $\Gamma(m_1 \cdots m_d, t, \nu)$ and take the remaining $\tilde{X}_{s_1,s_2,\dots,s_d}$ distributed according to the null distribution $\mathbb{P}_{H_0}$.*

It is interesting to remark that the foregoing procedure can be applied, with small modifications, to a larger class of discrete integer valued random variables: Poisson, geometric, negative binomial etc..

Example 3.4.2 (Gaussian model)**.** *In this example we consider that the underlying random field is generated by independent and identically distributed normal random variables with known mean μ and variance σ^2 ($X_{s_1,\dots,s_d} \sim \mathcal{N}(\mu, \sigma^2)$). Since the random variables $Y_{i_1,\dots,i_d}$ are sums of i.i.d. normals, clearly they follow a multivariate normal distribution with mean and covariance matrix, given by the next lemma (whose proof can be found in Appendix B):*

Lemma 3.4.3. *Let $X_{s_1,\dots,s_d} \sim \mathcal{N}(\mu, \sigma^2)$ be i.i.d. random variables for all $1 \leq s_j \leq T_j$, $j \in \{1, \dots, d\}$. The random variables $Y_{i_1,\dots,i_d}$ defined by Eq.(3.3) follow a multivariate normal distribution with mean $\bar{\mu} = m_1 \cdots m_d \mu$ and covariance matrix $\Sigma = (Cov\left[Y_{i_1,\dots,i_d}, Y_{j_1,\dots,j_d}\right])$, given by*

$$Cov\left[Y_{i_1,\dots,i_d}, Y_{j_1,\dots,j_d}\right] = \begin{cases} (m_1 - |i_1 - j_1|) \cdots (m_d - |i_d - j_d|)\, \sigma^2 & , |i_s - j_s| < m_s \\ & s \in \{1, \dots, d\}, \\ 0 & , \text{ otherwise.} \end{cases} \tag{3.76}$$

*Notice that **Step 2** of Algorithm 1 requires one to sample $Y_{i_1^{(k)},\dots,i_d^{(k)}}$ from the tail distribution $\mathbb{P}\left(Y_{i_1^{(k)},\dots,i_d^{(k)}} \geq \tau\right)$ and, for the other indices, from the conditional distribution given $\left\{Y_{i_1^{(k)},\dots,i_d^{(k)}} \geq \tau\right\}$.*
For generating a sample from the tail of a normal variable, we use the acceptance-rejection algorithm proposed in the classical paper of [Marsaglia, 1963] (see also [Devroye, 1986, pag. 380] for a faster alternative based on exponential random variables). For the second part, we propose two alternative methods (only the second one can be successfully applied for $d \geq 2$).
The first variant, which can be viewed as a direct approach, is to use the standard method of sampling from the posterior normal distribution of the unobserved components given the observed ones, of a multivariate Gaussian vector partitioned in observed and unobserved elements. To apply this method in our context, we need first to rewrite all the random variables $Y_{i_1,\dots,i_d}$ in an ordered sequence, that is to give a bijection between the set of all d-tuples $(i_1, \dots, i_d)$ and the set

$\{1,\dots,(T_1-m_1+1)\cdots(T_d-m_d+1)\}$.[1] *For such a bijection, we denote the formed sequence by* $\mathbf{Z}=(Z_1,\dots,Z_N)$, *where* $N=(T_1-m_1+1)\cdots(T_d-m_d+1)$. *If we consider that* $\left(i_1^{(k)},\dots,i_d^{(k)}\right)\to l$, *then to generate* $\mathbf{Z}$ *given* $Z_l\geq\tau$, *we partition it into* $\mathbf{Z}=(\mathbf{W}_1,Z_l,\mathbf{W}_2)$, *where the observed component is* Z_l *and the unobserved are* $\mathbf{W}_1=(Z_1,\dots,Z_{l-1})$ *and* $\mathbf{W}_2=(Z_{l+1},\dots,Z_N)$.
It is easy to establish (see for example [Tong, 1990, Chapter 3]) that given $Z_l=t$, *for some* $t\geq\tau$ *(obtained by sampling from the tail distribution of* Z_l*), we have*

$$\overline{\mathbf{W}}_1=\mathbf{W}_1|(Z_l=t)\sim\mathcal{N}\left(\mu_{w_1|t},\Sigma_{w_1|t}\right)\ \text{and}\ \overline{\mathbf{W}}_2=\mathbf{W}_2|(Z_l=t)\sim\mathcal{N}\left(\mu_{w_2|t},\Sigma_{w_2|t}\right) \tag{3.77}$$

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

$$\mu_{w_i|t}=\mathbb{E}[\mathbf{W}_i]+\frac{1}{Var[Z_l]}Cov[\mathbf{W}_i,Z_l](t-\mathbb{E}[Z_l]), \tag{3.78}$$

$$\Sigma_{w_i|t}=Cov(\mathbf{W}_i)-\frac{1}{Var[Z_l]}Cov[\mathbf{W}_i,Z_l]Cov^T[\mathbf{W}_i,Z_l]. \tag{3.79}$$

The covariance matrices that appear in the above equations can be computed using the result in Lemma 3.4.3. Notice that, in order to sample from $\mathcal{N}\left(\mu_{w_i|t},\Sigma_{w_i|t}\right)$, *one has to consider the Cholesky decomposition of* $\Sigma_{w_i|t}$ *and take*

$$\overline{\mathbf{W}}_i=\mu_{w_i|t}+Chol(\Sigma_{w_i|t})U_i, \tag{3.80}$$

with $U_i\sim\mathcal{N}(0,I)$ *vectors of independent standard normal random variables.*
Since computing the Cholesky decomposition at each iteration step can be time consuming, we propose an alternative method for sampling from the posterior of the Gaussian vector $\mathbf{Z}$. *This method was introduced by [Hoffman and Ribak, 1991] (see also the note of [Doucet, 2010]) and requires only to be able to simulate a random vector from the prior distribution. The algorithm can be summarized as follows*

- *Generate* $\mathbf{Z}\sim\mathcal{N}(\bar{\mu},\Sigma)$
- *Take* $\overline{\mathbf{W}}_i=\mathbf{W}_i+\frac{1}{Var[Z_l]}Cov[\mathbf{W}_i,Z_l](t-Z_l)$

The validity of the foregoing algorithm is presented in Appendix B. We remark that in the above procedure we need to compute the Cholesky decomposition of Σ *only once, thus reducing the execution time of Algorithm 1.*
The foregoing approach can be successfully applied in the case of one dimensional discrete scan statistics. Nevertheless, for higher dimensions ($d\geq 2$), *the covariance matrix is very large and to store it will require a large memory space. Take, for example, the two dimensional setting with* $T_1=T_2=200$ *and* $m_1=m_2=10$. *Clearly, the covariance matrix will be of size* 36481×36481 *and to store it in Matlab, for example, would be necessary almost* 10 *Gb of (RAM) space. Increasing the*

[1] Such a bijection is given by $f(i_1,\dots,i_d)=\sum_{s=1}^{d-1}(i_s-1)L_{s+1}\cdots L_d+i_d$, $L_0=1$.

dimensions of the problem and/or the sizes of the region $\mathcal{R}_d$ to be scanned, shows that this approach is unfeasible for $d \geq 2$.
To overcome the above difficulties, we present a second method for generating the random field $\tilde{\mathbf{X}}^{(\mathbf{k})} = \left\{\tilde{X}^{(k)}_{s_1,s_2,\ldots,s_d}\right\}$, with $s_j \in \{1,\ldots,T_j\}$ and $j \in \{1,\ldots,d\}$, from the conditional distribution given the d-tuple $(i_1^{(k)},\ldots,i_d^{(k)})$ and $\left\{Y_{i_1^{(k)},\ldots,i_d^{(k)}} \geq \tau\right\}$. The idea behind this method is to directly generate the underlying random field and not the random variables $Y_{i_1,\ldots,i_d}$, as in the previous approach and is based on the following result:

Lemma 3.4.4. *In the usual d dimensional setting, let $X_{s_1,s_2,\ldots,s_d}$, for all $s_j \in \{1,\ldots,T_j\}$ and $j \in \{1,\ldots,d\}$, be i.i.d. $\mathcal{N}(\mu,\sigma^2)$ random variables. If $w = m_1\cdots m_d$, then conditionally given the d-tuple $(i_1^{(k)},\ldots,i_d^{(k)})$ and $\left\{Y_{i_1^{(k)},\ldots,i_d^{(k)}} = t\right\}$, the random variables $X_{s_1,s_2,\ldots,s_d}$, $(s_1,\ldots,s_d) \neq \left(i_1^{(k)},\ldots,i_d^{(k)}\right)$, are jointly distributed as the random variables $\tilde{X}_{s_1,s_2,\ldots,s_d}$, $(s_1,\ldots,s_d) \neq \left(i_1^{(k)},\ldots,i_d^{(k)}\right)$, where*

$$\tilde{X}_{s_1,s_2,\ldots,s_d} = \frac{t-\mu\sqrt{w}}{w} - \frac{1}{w-1}\left(1-\frac{1}{w}\right)\left(Y_{i_1^{(k)},\ldots,i_d^{(k)}} - X_{i_1^{(k)},\ldots,i_d^{(k)}}\right) + X_{s_1,s_2,\ldots,s_d} \tag{3.81}$$

for $(s_1,\ldots,s_d) \in \Gamma_{i_1^{(k)},\ldots,i_d^{(k)}}$,

$$\tilde{X}_{s_1,s_2,\ldots,s_d} = X_{s_1,s_2,\ldots,s_d},\ \text{for } (s_1,\ldots,s_d) \notin \Gamma_{i_1^{(k)},\ldots,i_d^{(k)}} \tag{3.82}$$

and

$$\tilde{X}_{i_1^{(k)},\ldots,i_d^{(k)}} = t - \sum_{(s_1,\ldots,s_d)\in\Gamma_{i_1^{(k)},\ldots,i_d^{(k)}}} \tilde{X}_{s_1,s_2,\ldots,s_d} \tag{3.83}$$

and where

$$\Gamma_{i_1^{(k)},\ldots,i_d^{(k)}} = \left\{(r_1,\ldots,r_d) \neq (i_1^{(k)},\ldots,i_d^{(k)}) \,|\, i_j^{(k)} \leq r_j \leq i_j^{(k)} + m_j - 1,\ 1 \leq j \leq d\right\}.$$

The result in Lemma 3.4.4 leads to the following modification of the second step in Algorithm 1:

Step 2a *Given the d-tuple $(i_1^{(k)},\ldots,i_d^{(k)})$, generate a value $Y_{i_1^{(k)},\ldots,i_d^{(k)}} = t$ from the tail distribution $\mathbb{P}\left(Y_{i_1^{(k)},\ldots,i_d^{(k)}} \geq \tau\right)$.*

Step 2b *Conditionally, given the d-tuple $(i_1^{(k)},\ldots,i_d^{(k)})$ and $Y_{i_1^{(k)},\ldots,i_d^{(k)}} = t$, generate the random field $\tilde{\mathbf{X}} = \left\{\tilde{X}_{s_1,s_2,\ldots,s_d}\right\}$, with $s_j \in \{1,\ldots,T_j\}$ and $j \in \{1,\ldots,d\}$, based on the Eqs.(3.81)-(3.83) from Lemma 3.4.4*

The advantage of this method is that we don't need to compute any covariance matrix, the memory problem being solved. It is also important to notice the scalability of these formulas to higher dimensions.

We conclude by mentioning that there are other methods to sample from the discussed conditional distribution: for example in [Malley et al., 2002], the authors use a discrete fast Fourier transforms approach to solve this problem. However, their methodology make use of the covariance matrix, so it cannot be scaled to higher dimensions due to the memory problem described above.

Concerning the problem of efficiently searching for the locality statistics that exceed a given threshold τ, we adopt a technique based on *cumulative counts* (see [Neil, 2006]). This method searches the d-dimensional region $\mathcal{R}_d$ in constant time (see Figure 3.4(a)). We illustrate this method in the two dimensional setting, taking a square region and scanning the region with a square window ($d = 2$, $T_1 = T_2 = T$ and $m_1 = m_2 = m$). The idea is to precompute a matrix of cumulative counts using dynamic programming, step that takes about $\mathcal{O}(T^2)$ operations (such a function is implemented in almost all computational softwares: Matlab, R, Maple, Mathematica etc.). If we denote by $\boldsymbol{M}$ this matrix, then the element from the i-th line and j-th column is given by

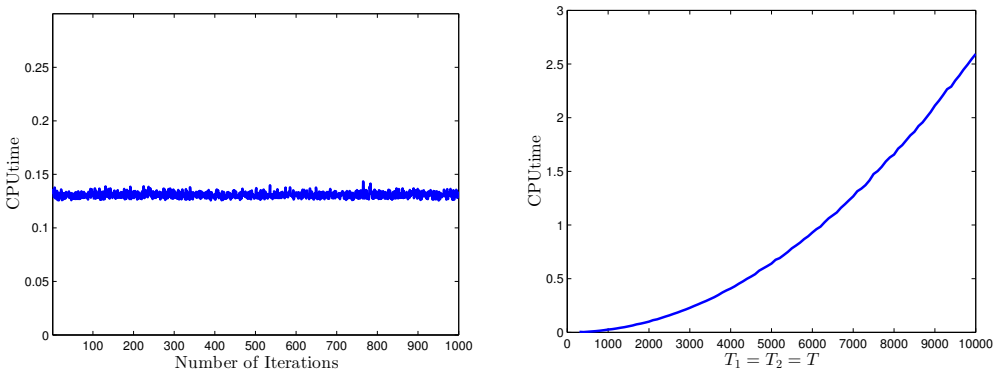
$$\boldsymbol{M}(i,j) = \sum_{k=1}^{i}\sum_{l=1}^{j} X_{k,l},$$

so the locality statistic, Y_{i_1,i_2} can be found by the relation

$$\begin{aligned} Y_{i_1,i_2} = \boldsymbol{M}(i_1+m-1, i_2+m-1) - \boldsymbol{M}(i_1+m-1, i_2-1) \\ - \boldsymbol{M}(i_1-1, i_2+m-1) + \boldsymbol{M}(i_1-1, i_2-1). \end{aligned}$$

The last computation can be done in $\mathcal{O}(1)$, which shows that if $ITER$ is the number of replicas made by our algorithm, then the required time for finding the number of locality statistics superior to a given τ is about $\mathcal{O}(ITER \times T^2)$.
In Figure 3.4, we considered two scenarios: on the left (see Figure 3.4(a)), we fixed the region and the scanning window sizes at $T_1 = T_2 = 2500$ and $m_1 = m_2 = 50$, respectively and for 10^3 replicas we plotted the run time necessary to find the number of locality statistics exceeding $\tau = 23$; on the right side (see Figure 3.4(b)), for the same values of $m_1 = m_2$ and τ, we illustrate the run time of the algorithm given that the size of the region increases from $T_1 = T_2 = 300$ to $T_1 = T_2 = 10000$. One advantage of this method is that can be easily scaled to d-dimensions[2]. In this situation, the computational time is about $\mathcal{O}(ITER \times T^d)$ for d-dimensional hypercubes. We should also mention that there are other methods, some of them faster (see for example [Neill et al., 2005] and [Neil, 2012]), that can be used to solve our problem. Nevertheless, their implementations are more difficult and involve advanced programming skills.

[2] In general, if $\boldsymbol{M}(y_1,\dots,y_d) = \sum_{s_1=1}^{y_1}\cdots\sum_{s_d=1}^{y_d} X_{s_1,\dots,s_d}$ and $W_{x_1,\dots,x_d}^{y_1,\dots,y_d} = \sum_{s_1=x_1}^{y_1}\cdots\sum_{s_d=x_d}^{y_d} X_{s_1,\dots,s_d}$ then $W_{x_1,\dots,x_d}^{y_1,\dots,y_d} = \boldsymbol{M}(y_1,\dots,y_d) + \sum_{k=1}^{d}\sum_{1\le j_1<\dots<j_k\le d} (-1)^k \boldsymbol{M}(y_1,\dots,x_{j_1}-1,\dots,x_{j_k}-1,\dots,y_d)$. To compute $Y_{i_1,\dots,i_d}$ it is enough to take $x_r = i_r$ and $y_r = i_r + m_r - 1$, $r \in \{1,\dots,d\}$ in the above formula.



(a) (b)

Figure 3.4: Illustration of the run time using the *cumulative counts* technique

3.4.4 Related algorithms: comparison for normal data

In this subsection, we present another importance sampling algorithm for the estimation of the significance level of hypothesis tests based on d-dimensional scan statistics. The algorithm was introduced by [Shi et al., 2007] and applied in the context of genetic linkage analysis. The idea behind the algorithm is to imbed the probability measure under the null hypothesis into an exponential family. We describe the algorithm in the context of an underlying random field generated by i.i.d. $\mathcal{N}(\mu, \sigma^2)$ random variables, but we should mention that it can be extended to any exponential families of distributions.

Consider the usual d-dimensional setting in which $X_{s_1,\dots,s_2} \sim \mathcal{N}(\mu, \sigma^2)$, $1 \leq s_j \leq T_j$, $j \in \{1, \dots, d\}$. As we saw in the Example 3.4.2, the random variables $Y_{i_1,\dots,i_d}$, $1 \leq i_j \leq T_j - m_j + 1$, $j \in \{1, \dots, d\}$, follow a multivariate normal distribution with mean and covariance matrix given by Lemma 3.4.3.

We define the new probability measure

$$d\mathbb{P}_{\xi,(r_1,\dots,r_d)} = \frac{e^{\xi Y_{r_1,\dots,r_d}}}{\mathbb{E}_{H_0}\left[e^{\xi Y_{r_1,\dots,r_d}}\right]} d\mathbb{P}_{H_0} \tag{3.84}$$

for a given ξ and d-tuple $(r_1, \dots, r_d)$. We observe that under the defined measure, the random field remains Gaussian and has the mean and the covariance matrix given by

$$\mathbb{E}_{\xi,(r_1,\dots,r_d)}\left[Y_{i_1,\dots,i_d}\right] = \xi Cov_{H_0}\left[Y_{i_1,\dots,i_d}, Y_{r_1,\dots,r_d}\right] + m_1 \cdots m_d \mu, \tag{3.85}$$

$$Cov_{\xi,(r_1,\dots,r_d)}\left[Y_{i_1,\dots,i_d}, Y_{j_1,\dots,j_d}\right] = Cov_{H_0}\left[Y_{i_1,\dots,i_d}, Y_{j_1,\dots,j_d}\right]. \tag{3.86}$$

The importance sampling algorithm is build based on the following identity:

$$\mathbb{P}_{H_0}\left(S_{\mathbf{m}}(\mathbf{T}) \geq \tau\right) = \frac{1}{N} \sum_{k_1,\ldots,k_d} \mathbb{E}_{\xi,(k_1,\ldots,k_d)} \left[\frac{N \mathbf{1}_{\{S_{\mathbf{m}}(\mathbf{T}) \geq \tau\}}}{\sum\limits_{j_1,\ldots,j_d} e^{\xi Y_{j_1,\ldots,j_d} - \log \mathbb{E}_{H_0}\left[e^{\xi Y_{j_1,\ldots,j_d}}\right]}} \right], \tag{3.87}$$

with $N = (T_1 - m_1 + 1) \cdots (T_d - m_d + 1)$.
This relation can be deduced from the following argument:

$$\begin{aligned}
\mathbb{P}_{H_0}\left(S_{\mathbf{m}}(\mathbf{T}) \geq \tau\right) &= \int \mathbf{1}_{\left\{ \max\limits_{\substack{1 \leq i_j \leq T_j - m_j + 1 \\ j \in \{1,2,\ldots,d\}}} Y_{i_1,i_2,\ldots,i_d} \geq \tau \right\}} d\mathbb{P}_{H_0} \\
&= \int \frac{\sum\limits_{j_1=1}^{T_1-m_1+1} \cdots \sum\limits_{j_d=1}^{T_d-m_d+1} e^{\xi Y_{j_1,\ldots,j_d}}}{\sum\limits_{j_1=1}^{T_1-m_1+1} \cdots \sum\limits_{j_d=1}^{T_d-m_d+1} e^{\xi Y_{j_1,\ldots,j_d}}} \mathbf{1}_{\{S_{\mathbf{m}}(\mathbf{T}) \geq \tau\}} d\mathbb{P}_{H_0} \\
&= \sum_{k_1,\ldots,k_d} \int \frac{e^{\xi Y_{k_1,\ldots,k_d}}}{\sum\limits_{j_1,\ldots,j_d} e^{\xi Y_{j_1,\ldots,j_d}}} \mathbf{1}_{\{S_{\mathbf{m}}(\mathbf{T}) \geq \tau\}} d\mathbb{P}_{H_0} \\
&= \sum_{k_1,\ldots,k_d} \int \frac{\mathbf{1}_{\{S_{\mathbf{m}}(\mathbf{T}) \geq \tau\}}}{\sum\limits_{j_1,\ldots,j_d} e^{\xi Y_{j_1,\ldots,j_d} - \log \mathbb{E}_{H_0}\left[e^{\xi Y_{j_1,\ldots,j_d}}\right]}} d\mathbb{P}_{\xi,(k_1,\ldots,k_d)} \\
&= \frac{1}{N} \sum_{k_1,\ldots,k_d} \int \frac{N}{\sum\limits_{j_1,\ldots,j_d} e^{\xi Y_{j_1,\ldots,j_d} - \log \mathbb{E}_{H_0}\left[e^{\xi Y_{j_1,\ldots,j_d}}\right]}} \mathbf{1}_{\{S_{\mathbf{m}}(\mathbf{T}) \geq \tau\}} d\mathbb{P}_{\xi,(k_1,\ldots,k_d)},
\end{aligned} \tag{3.88}$$

where the log moment generating function is equal with

$$\log \mathbb{E}_{H_0}\left[e^{\xi Y_{j_1,\ldots,j_d}}\right] = m_1 \cdots m_d \left(\mu\xi + \frac{\sigma^2 \xi^2}{2}\right), \tag{3.89}$$

and where we have adopted the shorthand notation $\sum\limits_{k_1,\ldots,k_d} = \sum\limits_{k_1=1}^{T_1-m_1+1} \cdots \sum\limits_{k_d=1}^{T_d-m_d+1}$.
The choice of the parameter ξ in the above relation should satisfy the following

equation (see [Shi et al., 2007, Appendix B] for an heuristic argument):

$$\mathbb{E}_{\xi,(r_1,\dots,r_d)}\left[Y_{r_1,\dots,r_d}\right] \approx \tau, \tag{3.90}$$

which leads to the solution $\xi \approx \frac{\tau}{m_1\cdots m_d\sigma^2} - \frac{\mu}{\sigma^2}$.
Based on the identity in Eq.(3.87), we have the following importance sampling algorithm:

Algorithm 2 Second Importance Sampling Algorithm for Scan Statistics

Begin

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

1: Generate uniformly the d-tuple $(i_1^{(k)},\dots,i_d^{(k)})$ from the set $\{1,\dots,T_1-m_1+1\}\times\cdots\times\{1,\dots,T_d-m_d+1\}$.
2: Given the d-tuple $(i_1^{(k)},\dots,i_d^{(k)})$, generate a sample of the Gaussian process $Y_{i_1,\dots,i_d}$ according to the new measure $d\mathbb{P}_{\xi,(i_1^{(k)},\dots,i_d^{(k)})}$.
3: Compute $\widehat{\rho}_k(d)$ based on

$$\widehat{\rho}_k(d) = \frac{N}{\sum\limits_{j_1,\dots,j_d} e^{\xi Y_{j_1,\dots,j_d} - \log \mathbb{E}_{H_0}\left[e^{\xi Y_{j_1,\dots,j_d}}\right]}} \mathbf{1}_{\{S_{\mathbf{m}}(\mathbf{T})\geq\tau\}}.$$

End Repeat

Return $\widehat{\rho}(d) = \dfrac{1}{ITER}\displaystyle\sum_{k=1}^{ITER}\widehat{\rho}_k(d)$.

End

In order to illustrate the efficiency of the foregoing algorithm and Algorithm 1 described in Section 3.4.2[3], we consider the problem of estimating the distribution of one dimensional scan statistics over a standard Gaussian random field. In our simulation study we include, for comparison, another two approaches for estimating the desired p value: the direct Monte Carlo method given at the beginning of Section 3.4 and a method based on the quasi Monte Carlo algorithm of [Genz and Bretz, 2009] for numerically approximate the distribution of a multivariate normal.
To evaluate the efficiency of the algorithms, we define the measure of *relative efficiency* between two algorithms, to be equal with the ratio between the computation time that is required by each algorithm to achieve a given error variance. This measure of efficiency has been used, in a slightly general form, in [Wu and Naiman, 2005] and [Priebe et al., 2001], where the authors considered the ratio between the computation time times the variance of the estimator.
In Tables 3.1-3.3, we present a comparison study between the four methods. We took as the reference point the importance sampling algorithm presented in Section 3.4.2.

[3] Actually, in the second step of the algorithm, we take the first approach presented in Example 3.4.2 for fair comparison, since both methods use the covariance matrix of the process.

We have evaluated the distribution of the one dimensional discrete scan statistics for selected values of the length of the sequence $T_1 \in \{200, 500, 750, 800\}$ and of the scanning window $m_1 \in \{15, 25, 30, 40\}$. The last column gives the value of the relative efficiency measure for each compared method (Genz, direct Monte Carlo and the new importance sampling algorithm) with respect to Algorithm 1. It can be observed that Algorithm 1 is the most efficient between the four, in some cases the differences being huge, especially in comparison with the direct Monte Carlo method and the algorithm proposed by [Genz and Bretz, 2009] (in some cases the importance sampling method is 600 times faster).

T_1	m_1	n	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 3.1: A comparison of the p values as evaluated by simulation using [Genz and Bretz, 2009] algorithm (Genz), importance sampling (Algo 1) and the relative efficiency between the methods (Rel Eff)

T_1	m_1	n	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

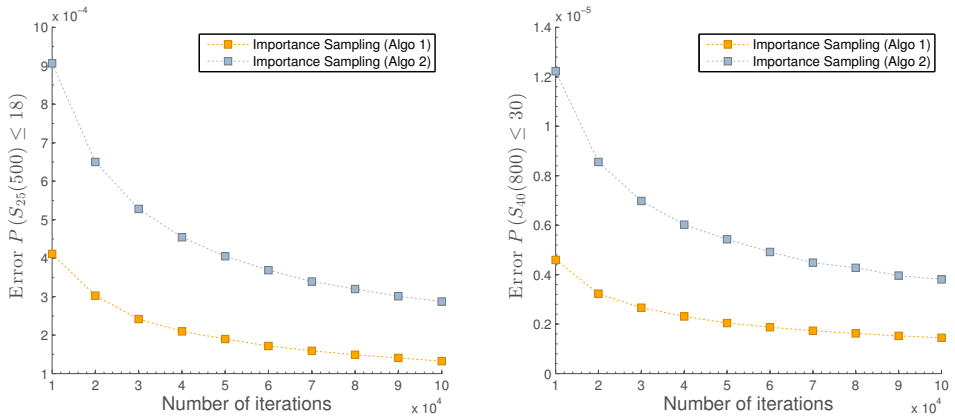
Table 3.2: A comparison of the p values as evaluated by naive Monte Carlo (MC), importance sampling (Algo 1) and the relative efficiency between the methods (Rel Eff)

T_1	m_1	n	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

Table 3.3: A comparison of the p values as evaluated by the two importance sampling algorithms (Algo 2 and Algo 1) and the relative efficiency between the methods (Rel Eff)

3.5 Examples and numerical results

To illustrate the accuracy of our approximations, we consider in this section the particular cases of one, two and three dimensional discrete scan statistics. We compare our results with the existing ones presented in Chapter 1.



(a) (b)

Figure 3.5: The evolution of simulation error in IS Algorithm 1 and IS Algorithm 2

3.5.1 One dimensional scan statistics

Based on the methodology presented in Section 3.2 and Section 3.3, we have the following approximation for the one dimensional scan statistics:

$$\mathbb{P}\left(S_{m_1}(T_1) \leq n\right) \approx \frac{2Q_2 - Q_3}{[1 + Q_2 - Q_3 + 2(Q_2 - Q_3)^2]^{L_1 - 1}}. \tag{3.91}$$

From Eq.(3.41), the theoretical error bound is

$$E_{app}(1) = (L_1 - 1)F\left(Q_2, L_1 - 1\right)\left(1 - Q_2\right)^2, \tag{3.92}$$

while the simulation errors, according to Eq.(3.45) and Eq.(3.54), are given by

$$E_{sapp}(1) = (L_1 - 1)F\left(\hat{Q}_2, L_1 - 1\right)\left(1 - \hat{Q}_2 + \beta_2\right)^2, \tag{3.93}$$

$$E_{sf}(1) = (L_1 - 1)(\beta_2 + \beta_3). \tag{3.94}$$

For selected values of m_1, T_1 and n and selected values of the parameters of the binomial, Poisson and normal distributions, we evaluate the accuracy of the approximation in Eq. (3.91), as well as the sharpness of the error bounds given above. In Table 3.4, we compare our approximation with the exact value (column Exact Value), the product type approximation and the lower and upper margins of the distribution of the scan statistics presented in Chapter 1. The exact values were obtained using the Markov chain imbedding methodology described in Section 1.1.1.2. Numerical values for the distribution of $S_{m_1}(T_1)$ in the case of binomial and Poisson models are illustrated in Table 3.5. To evaluate the two models, we considered a sequence of length $T_1 = 5000$, distributed according to a binomial distribution with parameters $r = 5$ and $p = 0.01$ and a Poisson of mean $\lambda = rp$, respectively and scanned with a window of size $m_1 = 50$. We have also included the simulated value, the product type approximation (Eq.(1.46)) and the lower and upper bounds

n	Exact Value	AppH Eq.(3.91)	$E_{app}(1)$ Eq.(3.92)	AppPT Eq.(1.46)	LowB Eq.(1.47)	UppB Eq.(1.49)
4	0.853857	0.853949	0.000673	0.853861	0.853583	0.853982
5	0.983090	0.983092	0.000007	0.983091	0.983087	0.983092
6	0.998628	0.998628	0.000000	0.998628	0.998628	0.998628
7	0.999916	0.999916	0.000000	0.999916	0.999916	0.999916

Table 3.4: Approximations for $\mathbb{P}(S_{m_1}(T_1) \leq n)$ in the Bernoulli $\mathcal{B}(0.05)$ case: $m_1 = 15$, $T_1 = 1000$, $ITER = 10^5$

(Eqs.(1.47) and (1.49)) for comparison reasons. For the simulated value (Sim) and our approximation (AppH), we used the importance sampling algorithm 1, as explained in Example 3.4.1, with $ITER_{sim} = 10^4$ and $ITER_{app} = 10^5$, respectively. The Gaussian model with mean $\mu = 0$ and variance $\sigma^2 = 1$ is presented in Table 3.6. For the evaluation of our approximation and of the simulation, we applied the method described in Example 3.4.2, where we used 10^5 replicas of the algorithm for the approximation and 10^4 replicas for the simulation.

n	AppH Eq.(3.91)	$E_{sapp}(1)$ Eq.(3.93)	$E_{sf}(1)$ Eq.(3.94)	Total Error	Sim	AppPT Eq.(1.46)	LowB Eq.(1.47)	UppB Eq.(1.49)
				$Bin(5, 0.01)$				
8	0.612778	0.004227	0.014486	0.018712	0.612987	0.605202	0.602589	0.605584
9	0.872277	0.000284	0.003811	0.004095	0.875072	0.870080	0.869625	0.870115
10	0.966629	0.000016	0.000904	0.000920	0.967064	0.966169	0.966095	0.966171
11	0.992442	0.000001	0.000193	0.000193	0.992507	0.992359	0.992346	0.992359
12	0.998485	0.000000	0.000037	0.000037	0.998465	0.998447	0.998445	0.998447
13	0.999716	0.000000	0.000007	0.000007	0.999718	0.999711	0.999711	0.999711
				$Poiss(0.05)$				
8	0.587242	0.004974	0.015845	0.020819	0.585250	0.587028	0.584203	0.587451
9	0.859921	0.000350	0.004239	0.004589	0.860858	0.859601	0.859087	0.859643
10	0.962599	0.000021	0.001029	0.001050	0.961894	0.962222	0.962137	0.962225
11	0.991108	0.000001	0.000228	0.000229	0.991120	0.991167	0.991152	0.991167
12	0.998140	0.000000	0.000046	0.000046	0.998126	0.998135	0.998132	0.998135
13	0.999642	0.000000	0.000008	0.000008	0.999635	0.999639	0.999638	0.999639

Table 3.5: Approximations for $\mathbb{P}(S_{m_1}(T_1) \leq n)$ for binomial and Poisson cases: $m_1 = 50$, $T_1 = 5000$, $ITER_{app} = 10^5$, $ITER_{sim} = 10^4$

We observe from the numerical results that our proposed approximation is quite accurate. Moreover, we see that the main contribution to the overall error is given by the simulation error associated with the approximation formula ($E_{sf}(1)$). Increasing the number of iterations of our algorithm can solve this aspect.

n	AppH Eq.(3.91)	$E_{sapp}(1)$ Eq.(3.93)	$E_{sf}(1)$ Eq.(3.93)	Total Error	Sim	AppPT Eq.(1.46)	LowB Eq.(1.47)	UppB Eq.(1.49)
17	0.657701	0.022872	0.011252	0.034124	0.665150	0.655664	0.630177	0.638129
18	0.759762	0.008325	0.007202	0.015527	0.754713	0.756074	0.755434	0.759405
19	0.838230	0.003112	0.004488	0.007600	0.836921	0.836993	0.846807	0.848582
20	0.893662	0.001119	0.002751	0.003871	0.893977	0.892007	0.882018	0.882768
21	0.933325	0.000390	0.001632	0.002022	0.933530	0.934386	0.929999	0.930284
22	0.959535	0.000132	0.000948	0.001080	0.959285	0.960504	0.962083	0.962185
23	0.976400	0.000044	0.000538	0.000582	0.976006	0.976686	0.976825	0.976860
24	0.986511	0.000014	0.000297	0.000311	0.986695	0.986749	0.985801	0.985812
25	0.992579	0.000004	0.000159	0.000163	0.992463	0.992358	0.992754	0.992757
26	0.996014	0.000001	0.000083	0.000084	0.995919	0.995792	0.996078	0.996079
27	0.997884	0.000000	0.000042	0.000043	0.997879	0.998012	0.997835	0.997835
28	0.998927	0.000000	0.000021	0.000021	0.998905	0.999349	0.998962	0.998962
29	0.999467	0.000000	0.000010	0.000010	0.999456	0.999428	0.999480	0.999480
30	0.999741	0.000000	0.000005	0.000005	0.999740	1.000017	0.999754	0.999754

Table 3.6: Approximations for $\mathbb{P}(S_{m_1}(T_1) \leq n)$ in the Gaussian $\mathcal{N}(0,1)$ case: $m_1 = 40$, $T_1 = 800$, $ITER_{app} = 10^5$, $ITER_{sim} = 10^4$

3.5.2 Two dimensional scan statistics

The distribution of the two dimensional discrete scan statistics can be approximated via Eq.(3.26) by

$$\mathbb{P}\left(S_{\mathbf{m}}(\mathbf{T}) \leq n\right) \approx \frac{2\hat{Q}_2 - \hat{Q}_3}{\left[1 + \hat{Q}_2 - \hat{Q}_3 + 2(\hat{Q}_2 - \hat{Q}_3)^2\right]^{L_1 - 1}}, \tag{3.95}$$

where

$$\hat{Q}_2 = H\left(\hat{Q}_{2,2}, \hat{Q}_{2,3}, L_2\right)$$
$$\hat{Q}_3 = H\left(\hat{Q}_{3,2}, \hat{Q}_{3,3}, L_2\right).$$

The simulation errors follow from Eq.(3.45) and Eq.(3.54) and can be expresses as

$$E_{sf}(2) = (L_1 - 1)(L_2 - 1)\left(\beta_{2,2} + \beta_{2,3} + \beta_{3,2} + \beta_{3,3}\right), \tag{3.96}$$

$$E_{sapp}(2) = (L_1 - 1)\left[F\left(1 - \hat{Q}_2 + A_2 + C_2\right)^2 + (L_2 - 1)F_2\left(1 - \hat{Q}_{2,2} + \beta_{2,2}\right)^2 + (L_2 - 1)F_3\left(1 - \hat{Q}_{3,2} + \beta_{3,2}\right)^2\right], \tag{3.97}$$

where $F = F\left(\hat{Q}_2, L_1 - 1\right)$, $F_2 = F\left(\hat{Q}_{2,2}, L_2 - 1\right)$ and $F_3 = F\left(\hat{Q}_{3,2}, L_2 - 1\right)$, based on Eq.(3.27). The value of A_2 and C_2 can be computed, according to Eqs.(3.46) and (3.48), by the following formulas:

$$A_2 = (L_2 - 1)\left(\beta_{2,2} + \beta_{2,3}\right),$$
$$C_2 = (L_2 - 1)F_2\left(1 - \hat{Q}_{2,2} + \beta_{2,2}\right)^2.$$

Notice that, from Eq.(3.41), the theoretical error becomes

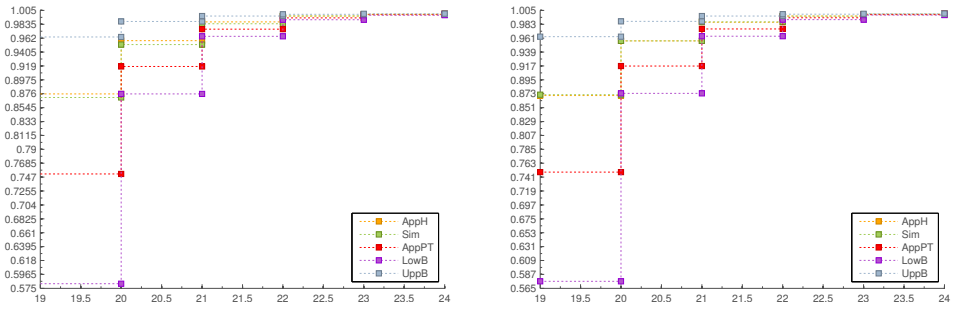
$$E_{app}(2) = (L_1 - 1)F\left(1 - \gamma_2 + B_2\right)^2 + (L_1 - 1)(L_2 - 1)\left[F_2\left(1 - Q_{2,2}\right)^2 + F_3\left(1 - Q_{3,2}\right)^2\right], \quad (3.98)$$

where $\gamma_2 = H\left(Q_{2,2}, Q_{2,3}, L_2\right)$ and $B_2 = (L_2 - 1)F_2\left(1 - Q_{2,2}\right)^2$.
To investigate the accuracy of the approximation and of the error bounds, described by Eqs.(3.95), (3.96) and (3.97), for the two dimensional discrete scan statistics, we present a series of numerical results. We evaluate the distribution of the scan statistics for selected values of the parameters of the binomial, Poisson and Gaussian models.
In Table 3.7, we include numerical values for two models: the binomial model with parameters $r = 10$ and $p = 0.001$ and the Poisson model with mean $\lambda = rp = 0.01$. To compare our results, we have included the product type approximation (AppPT) developed by [Chen and Glaz, 1996] (Eq.(1.67)) and the lower (LowB) and upper (UppB) bounds discussed in Section 1.2.2. The second column (AppH), that corresponds to our proposed approximation given by Eq.(3.95) and the sixth column (Sim), which corresponds to the simulated value, are evaluated via the importance sampling Algorithm 1 developed in Section 3.4, with $ITER_{app} = 10^4$ and $ITER_{sim} = 10^3$ replications, respectively. We observe that our approximation is quite accurate (see also Figure 3.6).

n	AppH Eq.(3.95)	$E_{sapp}(2)$ Eq.(3.96)	$E_{sf}(2)$ Eq.(3.97)	Total Error	Sim	AppPT Eq.(1.67)	LowB Eq.(1.76)	UppB Eq.(1.77)
				$Bin(10, 0.001)$				
19	0.875746	0.002681	0.049130	0.051811	0.870224	0.751860	0.582136	0.963836
20	0.958078	0.000281	0.015167	0.015448	0.952038	0.918137	0.875687	0.987744
21	0.987112	0.000022	0.004277	0.004299	0.984329	0.975919	0.964797	0.996133
22	0.996120	0.000002	0.001158	0.001159	0.996326	0.993379	0.990489	0.998849
23	0.998895	0.000000	0.000311	0.000311	0.998958	0.998269	0.997543	0.999675
24	0.999715	0.000000	0.000076	0.000076	0.999721	0.999567	0.999392	0.999912
				$Poiss(0.01)$				
19	0.870181	0.002964	0.051728	0.054692	0.871521	0.748912	0.576146	0.963406
20	0.956632	0.000290	0.016027	0.016317	0.956656	0.916879	0.873615	0.987569
21	0.986116	0.000024	0.004496	0.004520	0.986791	0.975478	0.964122	0.996069
22	0.995983	0.000002	0.001207	0.001209	0.995806	0.993239	0.990281	0.998827
23	0.998936	0.000000	0.000317	0.000317	0.998850	0.998228	0.997483	0.999668
24	0.999707	0.000000	0.000078	0.000078	0.999707	0.999556	0.999375	0.999910

Table 3.7: Approximations for $\mathbb{P}(S_{\mathbf{m}}(\mathbf{T}) \leq n)$ for binomial and Poisson models: $m_1 = 20$, $m_2 = 30$, $T_1 = 500$, $T_2 = 600$, $ITER_{app} = 10^4$, $ITER_{sim} = 10^3$

Numerical values for the Gaussian model of mean $\mu = 1$ and variance $\sigma^2 = 0.5$ are illustrated in Table 3.8. An extension to the two dimensional case of the importance sampling procedure presented in Example 3.4.2 was used to obtain the simulated values (Sim) and our approximation (AppH). We used $ITER_{app} = 10^4$ replications of the algorithm for approximation and $ITER_{sim} = 10^3$ for simulation.



(a) Binomial model (b) Poisson model

Figure 3.6: The empirical cumulative distribution function for the binomial and Poisson models in Table 3.7

n	AppH Eq.(3.95)	$E_{sapp}(2)$ Eq.(3.96)	$E_{sf}(2)$ Eq.(3.97)	Total Error	Sim	AppPT Eq.(1.67)	LowB Eq.(1.76)	UppB Eq.(1.77)
244	0.791513	0.011671	0.052791	0.064462	0.794643	0.777919	0.599404	0.918219
245	0.856678	0.004354	0.033828	0.038181	0.858459	0.847748	0.743415	0.943556
246	0.904917	0.001847	0.021621	0.023469	0.902215	0.895243	0.837316	0.961718
247	0.936329	0.000742	0.013350	0.014092	0.936598	0.932676	0.897894	0.974514
248	0.957904	0.000277	0.008174	0.008451	0.959433	0.957042	0.936562	0.983322
249	0.975042	0.000113	0.004976	0.005090	0.974738	0.972970	0.960985	0.989204
250	0.983983	0.000042	0.003014	0.003056	0.982867	0.982776	0.976247	0.993119
251	0.989632	0.000017	0.001805	0.001821	0.990106	0.989496	0.985685	0.995678
252	0.993801	0.000006	0.001067	0.001073	0.993643	0.993734	0.991460	0.997298
253	0.996329	0.000002	0.000631	0.000633	0.996193	0.996150	0.994956	0.998341
254	0.997863	0.000001	0.000359	0.000360	0.997922	0.997716	0.997051	0.998998
255	0.998689	0.000000	0.000205	0.000205	0.998708	0.998720	0.998294	0.999393
256	0.999264	0.000000	0.000118	0.000118	0.999259	0.999222	0.999022	0.999638

Table 3.8: Approximations for $\mathbb{P}(S_{\mathbf{m}}(\mathbf{T}) \leq n)$ in the Gaussian $\mathcal{N}(1, 0.5)$ model: $m_1 = 10$, $m_2 = 20$, $T_1 = 400$, $T_2 = 400$, $ITER_{app} = 10^4$, $ITER_{sim} = 10^3$

3.5.3 Three dimensional scan statistics

Following the methodology presented in Section 3.2 and the notations introduced in Section 3.3, we observe that in the three dimensional setting, the approximation formula for the distribution of the three dimensional discrete scan statistics is given by

$$\mathbb{P}\left(S_{\mathbf{m}}(\mathbf{T}) \leq n\right) \approx \frac{2\hat{Q}_2 - \hat{Q}_3}{\left[1 + \hat{Q}_2 - \hat{Q}_3 + 2(\hat{Q}_2 - \hat{Q}_3)^2\right]^{L_1 - 1}}, \tag{3.99}$$

where for $t_1, t_2 \in \{2, 3\}$ we have

$$\hat{Q}_{t_1} = H\left(\hat{Q}_{t_1,2}, \hat{Q}_{t_1,3}, L_2\right)$$
$$\hat{Q}_{t_1,t_2} = H\left(\hat{Q}_{t_1,t_2,2}, \hat{Q}_{t_1,t_2,3}, L_3\right)$$

with H given in Eq.(3.13).

It follows from Eq.(3.54) that the simulation error corresponding to the approximation formula is given by

$$E_{sf}(3) = (L_1-1)(L_2-1)(L_3-1)\left(\sum_{t_1,t_2,t_3\in\{2,3\}}\beta_{t_1,t_2,t_3}\right). \tag{3.100}$$

The second simulation error, which is associated with the approximation error, results from Eq.(3.45) and is expressed as

$$\begin{aligned}E_{sapp}(3) &= (L_1-1)F\left(1-\hat{Q}_2+A_2+C_2\right)^2+(L_1-1)(L_2-1)\left[F_2\left(1-\hat{Q}_{2,2}+\right.\right.\\ &\left.+A_{2,2}+C_{2,2}\right)^2+F_3\left(1-\hat{Q}_{3,2}+A_{3,2}+C_{3,2}\right)^2\Big]+(L_1-1)(L_2-1)\\ &\times(L_3-1)\left[F_{2,2}\left(1-\hat{Q}_{2,2,2}+\beta_{2,2,2}\right)^2+F_{2,3}\left(1-\hat{Q}_{2,3,2}+\beta_{2,3,2}\right)^2\right.\\ &\left.+F_{3,2}\left(1-\hat{Q}_{3,2,2}+\beta_{3,2,2}\right)^2+F_{3,3}\left(1-\hat{Q}_{3,3,2}+\beta_{3,3,2}\right)^2\right].\end{aligned} \tag{3.101}$$

Notice that, based on Eq.(3.27), for all $t_1,t_2\in\{2,3\}$, the coefficients F, F_{t_1} and F_{t_1,t_2} are computed as $F\left(\hat{Q}_2,L_1-1\right)$, $F\left(\hat{Q}_{t_1,2},L_2-1\right)$ and $F\left(\hat{Q}_{t_1,t_2,2},L_3-1\right)$, respectively.
From Eq.(3.46), we have

$$\begin{aligned}A_{2,2} &= (L_3-1)\left(\beta_{2,2,2}+\beta_{2,2,3}\right),\\ A_{3,2} &= (L_3-1)\left(\beta_{3,2,2}+\beta_{3,2,3}\right),\\ A_2 &= (L_2-1)(L_3-1)\left(\beta_{2,2,2}+\beta_{2,2,3}+\beta_{2,3,2}+\beta_{2,3,3}\right),\end{aligned}$$

while from Eq.(3.48)

$$\begin{aligned}C_{2,2} &= (L_3-1)F_{2,2}\left(1-\hat{Q}_{2,2,2}+\beta_{2,2,2}\right)^2,\\ C_{2,3} &= (L_3-1)F_{2,3}\left(1-\hat{Q}_{2,3,2}+\beta_{2,3,2}\right)^2,\\ C_{3,2} &= (L_3-1)F_{3,2}\left(1-\hat{Q}_{3,2,2}+\beta_{3,2,2}\right)^2,\\ C_2 &= (L_2-1)\left[F_2\left(1-\hat{Q}_{2,2}+A_{2,2}+C_{2,2}\right)^2+C_{2,2}+C_{2,3}\right].\end{aligned}$$

For selected values of the parameters of the binomial, Poisson and normal distributions, we evaluate the approximation introduced in Section 3.2 and provide the corresponding error bounds. We show the contributions of the simulation errors ($E_{sapp}(3)$ and $E_{sf}(3)$) in the overall error.
For all our simulations we used the importance sampling algorithm with $ITER = 10^5$ replications. We compare our results with those existing in literature, see [Guerriero et al., 2010a] for the Bernoulli model and with the simulated value of the scan statistics obtained by scanning the whole region $\mathcal{R}_3$, denoted by $\hat{\mathbb{P}}(S\leq n)$.

The scanning of $\mathcal{R}_3$ being more time consuming than the scanning of the subregions corresponding to Q_{t_1,t_2,t_3}, we used 10^3 repetitions of the algorithm.
In Table 3.9, we compare the results obtained by our approximation with the product type approximation presented by [Guerriero et al., 2010a] (see also Chapter 1, Eq.(1.87)). We observe that our approximation is very sharp. Table 3.10 presents

n	$\hat{\mathbb{P}}(S \leq n)$	AppPT Eq.(1.87)	AppH Eq.(3.99)	E_{sapp} Eq.(3.101)	E_{sf} Eq.(3.100)	Total Error
			$p = 0.00005$			
1	0.841806	0.841424	0.851076	0.011849	0.064889	0.076738
2	0.999119	0.999142	0.999192	0.000000	0.000170	0.000170
3	0.999997	0.999998	0.999997	0.000000	3×10^{-7}	3×10^{-7}
			$p = 0.0001$			
2	0.993294	0.993241	0.993192	0.000010	0.001367	0.001377
3	0.999963	0.999964	0.999963	0.000000	0.000005	0.000005
4	0.999999	0.999999	0.999999	0.000000	2×10^{-9}	2×10^{-9}

Table 3.9: Approximations for $\mathbb{P}(S_{\mathbf{m}}(\mathbf{T}) \leq n)$ in the Bernoulli model: $m_1 = m_2 = m_3 = 5$, $T_1 = T_2 = T_3 = 60$, $ITER_{app} = 10^5$, $ITER_{sim} = 10^3$

the numerical results obtained by scanning the region $\mathcal{R}_3$ of size $60 \times 60 \times 60$, with two windows of the same volume but different sizes, first a cubic window of size $4 \times 4 \times 4$ and second a rectangular region of size $8 \times 4 \times 2$. We observe that the results are closely related, but significantly different. In Table 3.11, we have included numer-

n	$\hat{\mathbb{P}}(S \leq n)$	AppH Eq.(3.99)	$E_{sapp}(3)$ Eq.(3.101)	$E_{sf}(3)$ Eq.(3.100)	Total Error
		$m_1 = m_2 = m_3 = 4$			
5	0.961691	0.963506	0.000038	0.003622	0.003660
6	0.999006	0.999023	0.000000	0.000071	0.000071
7	0.999980	0.999980	0.000000	0.000001	0.000001
8	0.999999	0.999999	0.000000	2×10^{-9}	2×10^{-9}
		$m_1 = 8, m_2 = 4, m_3 = 2$			
5	0.969189	0.969110	0.000007	0.003387	0.003395
6	0.999297	0.999228	0.000000	0.000071	0.000071
7	0.999984	0.999984	0.000000	0.000001	0.000001
8	0.999999	0.999999	0.000000	2×10^{-9}	2×10^{-9}

Table 3.10: Approximation for $\mathbb{P}(S_{\mathbf{m}}(\mathbf{T}) \leq n)$ over the region $\mathcal{R}_3$ with windows of the same volume by different sizes: $T_1 = T_2 = T_3 = 60$, $p = 0.0025$, $ITER_{app} = 10^5$, $ITER_{sim} = 10^3$

ical values emphasizing the situation described by Remark 3.2.1. We consider the Bernoulli model of parameter $p = 0.0001$ over the region $\mathcal{R}_3$ of size $185 \times 185 \times 185$ and scan it with a cubic window of length 10. The second and forth columns gives the values corresponding to the bounds described in Eq.(3.23), while in the third column we presented the simulated values for $\mathbb{P}\left(S_{10,10,10}(185, 185, 185) \leq n\right)$.

n	$\mathbb{P}\left(S(L_1+1,L_2+1,L_3+1)\leq n\right)$	$\hat{\mathbb{P}}(S\leq n)$	$\mathbb{P}\left(S(L_1,L_2,L_3)\leq n\right)$
4	0.97524633	0.97465263	0.97491935
	(±0.00754004)	(±0.00618987)	(±0.00643099)
5	0.99931055	0.99935163	0.99938629
	(±0.00015833)	(±0.00014759)	(±0.00013490)
6	0.99998641	0.99998632	0.99998784
	(±0.00000272)	(±0.00000326)	(±0.00000230)

Table 3.11: Approximation for $\mathbb{P}(S_{\mathbf{m}}(\mathbf{T})\leq n)$ based on Remark 3.2.1: $m_1=m_2=m_3=10$, $T_1=T_2=T_3=185$, $L_1=L_2=L_3=20$, $ITER_{app}=10^5$, $ITER_{sim}=10^3$

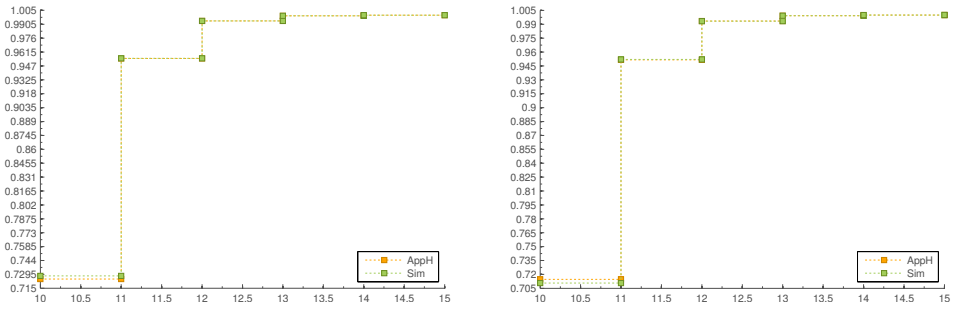
In order to compare the binomial and Poisson models, in Table 3.12 we have evaluated the distribution of the scan statistics over a region of size $84\times84\times84$, scanned with a $4\times4\times4$ cubic window, in the two situations. In the first case, we have a binomial random field with parameters m and p, that is $X_{s_1,s_2,s_3}\sim B(m,p)$, while in the second we considered that $X_{s_1,s_2,s_3}\sim P(\lambda)$, with $\lambda=mp$. We observe that the

n	$\hat{\mathbb{P}}(S\leq n)$	AppH Eq.(3.99)	$E_{sapp}(3)$ Eq.(3.101)	$E_{sf}(3)$ Eq.(3.100)	Total Error
			$Bin(10,0.0025)$		
10	0.726386	0.723224	0.007763	0.032197	0.039960
11	0.954605	0.955417	0.000123	0.003079	0.003202
12	0.993938	0.993906	0.000001	0.000331	0.000333
13	0.999289	0.999284	0.000000	0.000033	0.000033
14	0.999923	0.999921	0.000000	0.000003	0.000003
15	0.999992	0.999992	0.000000	3×10^{-7}	3×10^{-7}
			$Poiss(0.025)$		
10	0.713184	0.708481	0.009211	0.035294	0.044506
11	0.950947	0.950197	0.000143	0.003345	0.003488
12	0.993624	0.993452	0.000002	0.000365	0.000367
13	0.999218	0.999210	0.000000	0.000038	0.000038
14	0.999912	0.999911	0.000000	0.000003	0.000003
15	0.999990	0.999990	0.000000	3×10^{-7}	3×10^{-7}

Table 3.12: Approximation for $\mathbb{P}(S_{\mathbf{m}}(\mathbf{T})\leq n)$ in the binomial and Poisson models: $m_1=m_2=m_3=4$, $T_1=T_2=T_3=84$, $ITER_{app}=10^5$, $ITER_{sim}=10^3$

cumulative function, in the two models, are close to each other (see also Figure 3.7).

Lastly, we have consider a numerical example for the Gaussian model. In Table 3.13, we present numerical values for the distribution of the three dimensional discrete scan statistics over a region of size $256\times256\times256$, scanned with a cubic window of size $10\times10\times10$. The underlying random field is taken here to be formed of i.i.d. random variables distributed according to a $\mathcal{N}(0,1)$ law. We compare our results with the simulated value ($ITER_Sim=10^2$) and with the product type



(a) Binomial model (b) Poisson model

Figure 3.7: The empirical cumulative distribution function for the binomial and Poisson models in Table 3.12

approximation introduced in [Guerriero et al., 2010a] for the Bernoulli case (see also Chapter 1, Eq.(1.87)).

n	$\hat{\mathbb{P}}(S \leq n)$	AppPT. Eq.(1.87)	AppH Eq.(3.99)	$E_{sapp}(3)$ Eq.(3.101)	$E_{sf}(3)$ Eq.(3.100)	Total Error
175	0.894045	0.886580	0.893375	0.000775	0.017527	0.018302
176	0.901135	0.903979	0.909782	0.000528	0.014624	0.015152
177	0.922387	0.917918	0.923584	0.000368	0.012175	0.012543
178	0.930162	0.930604	0.935843	0.000257	0.010126	0.010383
179	0.950977	0.941291	0.945966	0.000183	0.008422	0.008605
180	0.947742	0.950626	0.954560	0.000122	0.006977	0.007099
181	0.958603	0.958162	0.962098	0.000086	0.005797	0.005883
182	0.968953	0.965074	0.968498	0.000059	0.004815	0.004874
183	0.968504	0.970581	0.973523	0.000041	0.003974	0.004015
184	0.973949	0.975465	0.977604	0.000029	0.003282	0.003311
185	0.981032	0.979428	0.981513	0.000019	0.002716	0.002736
186	0.984730	0.982610	0.984792	0.000013	0.002244	0.002258
187	0.986089	0.985640	0.987422	0.000009	0.001846	0.001855
188	0.987624	0.987935	0.989483	0.000006	0.001522	0.001529
189	0.989902	0.989956	0.991301	0.000004	0.001251	0.001255
190	0.992026	0.991744	0.992910	0.000003	0.001031	0.001034
191	0.993535	0.993126	0.994085	0.000002	0.000848	0.000850
192	0.994097	0.994416	0.995122	0.000001	0.000696	0.000697
193	0.995560	0.995346	0.995966	0.000001	0.000569	0.000570
194	0.996642	0.996142	0.996707	0.000001	0.000467	0.000468
195	0.997113	0.996809	0.997288	0.000000	0.000382	0.000382

Table 3.13: Approximations for $\mathbb{P}(S_{\mathbf{m}}(\mathbf{T}) \leq n)$ in the Gaussian $\mathcal{N}(0,1)$ model: $m_1 = m_2 = m_3 = 10$, $T_1 = T_2 = T_3 = 256$, $ITER_{app} = 10^5$, $ITER_{sim} = 10^3$

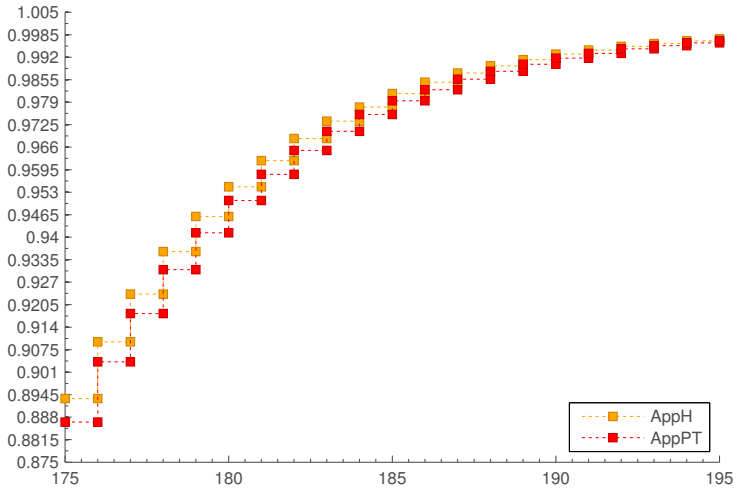


Figure 3.8: The empirical cumulative distribution functions (AppH = Our Approximation, AppPT = Product Type Approximation) for the Gaussian model in Table 3.13

Notice that the contribution of the approximation error (E_{sapp}) to the total error is almost negligible in most of the cases with respect to the simulation error (E_{sf}). Thus, the precision of the method will depend mostly on the number of iterations ($ITER$) used to estimate Q_{t_1,t_2,t_3}.

Chapter 4

Scan statistics over some block-factor type dependent models

In this chapter, we present an estimate for the distribution of the multidimensional discrete scan statistics over a random field generated by a block-factor type model. This dependent model generalizes the i.i.d. model studied in Chapter 3 and is introduced in Section 4.1. The approximation process, as well as the associated error bounds, are described in Section 4.2. Section 4.3 includes examples and numerical applications for particular block-factor models in one and two dimensions. Some of the results presented in this chapter appeared in [Amărioarei and Preda, 2014], for the special case of two dimensions (see also [Amărioarei and Preda, 2013b]).

Contents

4.1 Block-factor type model

Most of the research devoted to the one, two or three dimensional discrete scan statistic considers the independent and identically distributed (i.i.d.) model for the random variables that generate the random field which is to be scanned. In this section, we define a dependence structure for the underlying random field based on a block-factor type model. Throughout this chapter, we adopt the definitions and the notations introduced in Section 3.1 for the d dimensional setting.

Recall from Definition 2.1.3 (see also [Burton et al., 1993]), that a sequence $(W_l)_{l\geq 1}$ of random variables with state space S_W is said to be a k block-factor of the sequence $(\tilde{W}_l)_{l\geq 1}$ with state space $S_{\tilde{W}}$, if there is a measurable function $f : S_{\tilde{W}}^k \to S_W$ such that

$$W_l = f\left(\tilde{W}_l, \tilde{W}_{l+1}, \ldots, \tilde{W}_{l+k-1}\right)$$

for all l.

Our block-factor type model is defined in the following way. Let $\tilde{T}_1$, $\tilde{T}_2$, $\ldots$, $\tilde{T}_d$ be positive integers, $d \geq 1$ and $\left\{\tilde{X}_{s_1,\ldots,s_d} \mid 1 \leq s_j \leq \tilde{T}_j, 1 \leq j \leq d\right\}$ be a family of independent and identically distributed real valued random variables over the d dimensional rectangular region $\tilde{\mathcal{R}}_d = [0, \tilde{T}_1] \times \cdots \times [0, \tilde{T}_d]$ in such a way that to each rectangular elementary subregion $\tilde{r}_d(s_1, s_2, \ldots, s_d) = [s_1 - 1, s_1] \times \cdots \times [s_d - 1, s_d]$ it corresponds a variable $\tilde{X}_{s_1,\ldots,s_d}$. As we saw in Chapter 3, it is customary to interpret these random variables as the number of some events (of interest) that occurred in the subregion $\tilde{r}_d(s_1, s_2, \ldots, s_d)$.

For $j \in \{1, 2, \ldots, d\}$, let $x_1^{(j)}$, $x_2^{(j)}$ be nonnegative integers such that $x_1^{(j)} + x_2^{(j)} + 1 \leq \tilde{T}_j$. Define $c_j = x_1^{(j)} + x_2^{(j)} + 1$ and take $T_j = \tilde{T}_j - c_j + 1$. To each d-tuple $(s_1, \ldots, s_d)$, with $s_j \in \left\{x_1^{(j)} + 1, \ldots, \tilde{T}_j - x_2^{(j)}\right\}$, $j \in \{1, \ldots, d\}$, we associate a d-way (or of order d) tensor (see [Kolda and Bader, 2009]) $\mathfrak{X}_{s_1,\ldots,s_d} \in \mathbb{R}^{c_1 \times \cdots \times c_d}$ whose elements are given by

$$\mathfrak{X}_{s_1,\ldots,s_d}(j_1, \ldots, j_d) = \tilde{X}_{s_1 - x_1^{(1)} - 1 + j_1, \ldots, s_d - x_1^{(d)} - 1 + j_d}, \tag{4.1}$$

where $(j_1, \ldots, j_d) \in \{1, \ldots, c_1\} \times \cdots \times \{1, \ldots, c_d\}$. We observe that for the particular cases of one and two dimensions, the d-way tensor $\mathfrak{X}_{s_1,\ldots,s_d}$ reduces to a vector (a tensor of order one) and a matrix (a tensor of order two), respectively.

If $\Pi : \mathbb{R}^{c_1 \times \cdots \times c_d} \to \mathbb{R}$ is a measurable real valued function defined on the set of the tensors $\mathfrak{X}_{s_1,\ldots,s_d}$ (measurable with respect to the usual Borel σ-fields of $\mathbb{R}^{c_1 \times \cdots \times c_d}$ and $\mathbb{R}$), then we define the *block-factor type* model by

$$X_{s_1,\ldots,s_d} = \Pi\left(\mathfrak{X}_{s_1 + x_1^{(1)}, \ldots, s_d + x_1^{(d)}}\right), \tag{4.2}$$

for all $1 \leq s_j \leq T_j$, $1 \leq j \leq d$.

To illustrate the intuition behind the above definition, we consider the special case of two dimensional block-factor model ($d = 2$). In this setting, the underlying random field is defined as $X_{s_1,s_2} = \Pi\left(\mathfrak{X}_{s_1 + x_1^{(1)}, s_2 + x_1^{(2)}}\right)$, where the tensor $\mathfrak{X}_{s_1,s_2}$, which encapsulates the dependence structure, takes the matrix form

$$\mathfrak{X}_{s_1,s_2} = \begin{pmatrix} \tilde{X}_{s_1 - x_1^{(1)}, s_2 - x_1^{(2)}} & \cdots & \tilde{X}_{s_1 + x_2^{(1)}, s_2 - x_1^{(2)}} \\ \vdots & \ddots & \vdots \\ \tilde{X}_{s_1 - x_1^{(1)}, s_2 + x_2^{(2)}} & \cdots & \tilde{X}_{s_1 + x_2^{(1)}, s_2 + x_2^{(2)}} \end{pmatrix}. \tag{4.3}$$

Figure 4.1 illustrates the construction of the block-factor model: on the left (see Figure 4.1(a)) is presented the configuration matrix defined by Eq.(4.3) and the resulted random variable after applying the transformation Π; on the right (see

Figure 4.1(b)) is exemplified how the i.i.d. model $\tilde{X}_{s_1,s_2}$ is transformed into the block-factor model X_{s_1,s_2}.

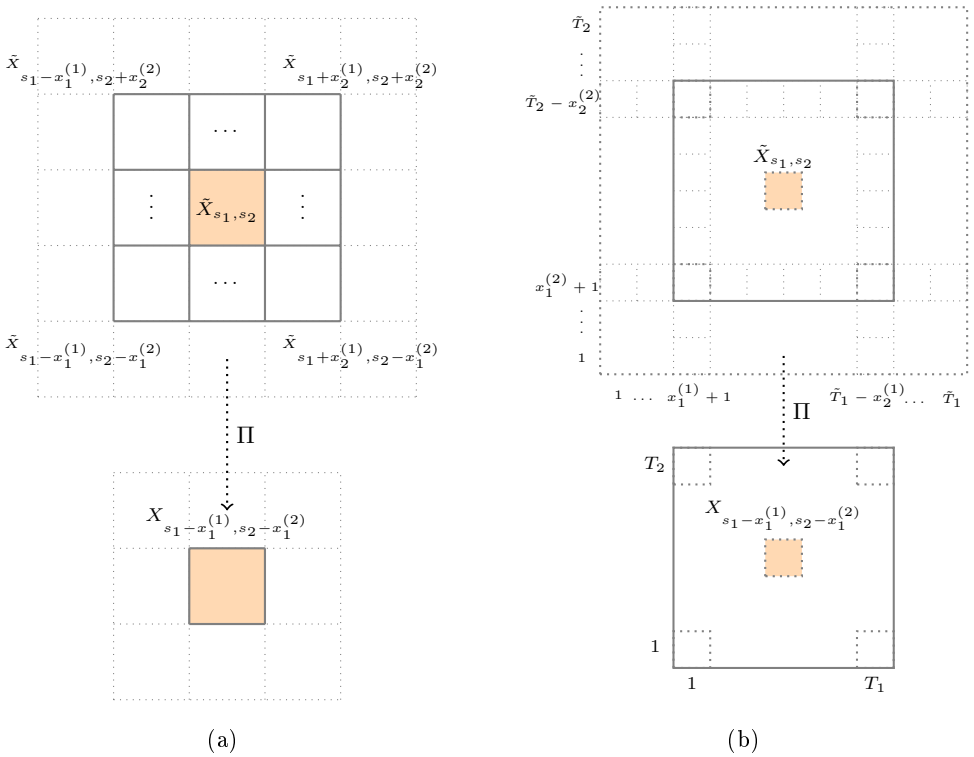


Figure 4.1: Illustration of the block-factor type model in two dimensions ($d = 2$)

It is clear, due to the overlapping structure, that $\{X_{s_1,s_2} \mid 1 \leq s_1 \leq T_1, 1 \leq s_2 \leq T_2\}$ forms a dependent family of random variables (see Figure 4.2) and the same is true for the general case.
Recall from Definition 2.1.1 that a sequence $(W_l)_{l\geq 1}$ is m-dependent with $m \geq 1$ (see [Burton et al., 1993]), if for any $h \geq 1$ the σ-fields generated by $\{W_1, \ldots, W_h\}$ and $\{W_{h+m+1}, \ldots\}$ are independent. Thus, from the definition of the block-factor model in Eq.(4.2), we observe that the sequence $\left(X_{s_1,\ldots,s_{h-1},s_h,s_{h+1},\ldots,s_d}\right)_{1\leq s_h \leq T_h}$ is $(c_h - 1)$-dependent, for each $1 \leq h \leq d$. In Figure 4.2 we illustrate, for the two dimensional case, the dependence structure of the field X_{s_1,s_2} emphasizing its (c_1-1) and $(c_2 - 1)$-dependent character.

Remark 4.1.1. *Notice that if $c_1 = \cdots = c_d = 1$ (that is $x_1^{(j)} = x_2^{(j)} = 0$ for $j \in \{\mathbf{1}, \ldots, d\}$) and Π is the identity function, then $X_{s_1,\ldots,s_d} = \tilde{X}_{s_1,\ldots,s_d}$ and we are in the i.i.d. situation. In this case, an approximation method for the distribution of the d-dimensional discrete scan statistics was studied in Chapter 3.*

Remark 4.1.2. *If we take $d = 1$, we find ourselves in the particular situation of one dimensional discrete scan statistics over a $(c_1 - 1)$-dependent sequence. The distribution of one dimensional scan statistics over this type of dependence was studied by*

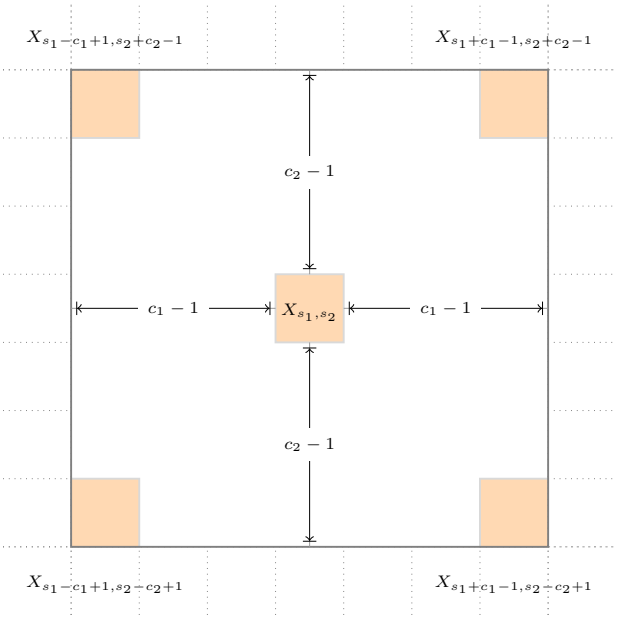


Figure 4.2: The dependence structure of X_{s_1,s_2} in two dimensions

[Haiman and Preda, 2013] in the particular case of Gaussian stationary 1*-dependent sequences* $W_i \sim \mathcal{N}(0,1)$ *of random variables generated by a two block-factor of the form*

$$W_i = aU_i + bU_{i+1}, \quad i \geq 1,$$

where $a^2 + b^2 = 1$ *and* $(U_i)_{i\geq 1}$ *is an i.i.d. sequence of* $\mathcal{N}(0,1)$ *random variables. Clearly, their model can be obtained from the one described in this section by putting* $x_1^{(1)} = 0$, $x_2^{(1)} = 1$, *which implies* $c_1 = 2$, *and choosing* $\Pi(\alpha_1, \alpha_2) = a\alpha_1 + b\alpha_2$. *Applications of the one dimensional discrete scan statistics over a sequence of moving average of order* q $(c_1 = q + 1)$ *was recently given by [Wang and Glaz, 2013] (see also [Wang, 2013, Chapter 4]). We include in Section 4.3, a numerical example in which we compare their results with the ones obtained by our method.*

Based on the dependent model introduced in this section, in Section 4.2 we give an approximation for the distribution of the d-dimensional discrete scan statistic over the random field generated by the random variables $X_{s_1,\ldots,s_d}$ and the corresponding error bounds.

4.2 Approximation and error bounds

In this section, we give an estimate for the distribution function, $Q_{\mathbf{m}}(\mathbf{T})$, of the d-dimensional discrete scan statistics evaluated over a random field generated by the block-factor model introduced in the foregoing section. The methodology used for the derivation of the approximation formula follows closely the approach adopted in Chapter 3 for the i.i.d. model.

4.2.1 The approximation process

Let us consider that we are in the framework of the block-factor model introduced in Section 4.1 and we scan the generated random field with a window of size $m_1 \times \cdots \times m_d$, $2 \leq m_j \leq T_j$, $j \in \{1, \ldots, d\}$. As in the case of the i.i.d. model (see Section 3.2), the main idea behind the estimation process is to express the scan statistics random variable, $S_{\mathbf{m}}(\mathbf{T})$, $\mathbf{m} = (m_1, \ldots, m_d)$, $\mathbf{T} = (T_1, \ldots, T_d)$, as the maximum of a 1-dependent stationary sequence of properly selected random variables.
Assume that for each $j \in \{1, \ldots, d\}$, $\tilde{T}_j = L_j(m_j + c_j - 2)$, where L_1, $\ldots$, L_d are positive integers and observe that the generated region $\mathcal{R}_d$, after applying the transformation Π, has the sides of length

$$T_j = (L_j - 1)(m_j + c_j - 2) + m_j - 1. \tag{4.4}$$

We define for each $k_1 \in \{1, 2, \ldots, L_1 - 1\}$ the random variables

$$Z_{k_1} = \max_{\substack{(k_1-1)(m_1+c_1-2)+1 \leq i_1 \leq k_1(m_1+c_1-2) \\ 1 \leq i_j \leq (L_j-1)(m_j+c_j-2) \\ j \in \{2,\ldots,d\}}} Y_{i_1, i_2, \ldots, i_d}, \tag{4.5}$$

and we claim that $\{Z_1, \ldots, Z_{L_1-1}\}$ forms a 1-dependent and stationary sequence of random variables in the view of Definition 2.1.1 and Definition 2.1.2. To see this, we first remark that Z_{k_1}'s represent the d-dimensional scan statistics on the overlapping rectangular strips

$$[(k_1 - 1)(m_1 + c_1 - 2), k_1(m_1 + c_1 - 2) + m_1 - 1] \times [0, T_2] \times \cdots \times [0, T_d], \tag{4.6}$$

of size $(2m_1 + c_1 - 3) \times T_2 \times \cdots \times T_d$. In Figure 4.3 we illustrate, in the particular situation of two dimensions ($d = 2$), the overlapping structure of Z_1, Z_2 and Z_3 emphasizing the 1-dependent character of the sequence.
The one dependence structure of $\{Z_1, \ldots, Z_{L_1-1}\}$ follows immediately from

$$\begin{aligned} Z_{k_1-1} &\in \sigma\left(\{X_{s_1,\ldots,s_d} | (k_1 - 2)(m_1 + c_1 - 2) + 1 \leq s_1 \leq (k_1 - 1)(m_1 + c_1 - 2) + m_1 - 1\}\right) \\ &\in \sigma\left(\left\{\tilde{X}_{s_1,\ldots,s_d} \middle| (k_1 - 2)(m_1 + c_1 - 2) + 1 \leq s_1 \leq k_1(m_1 + c_1 - 2)\right\}\right) \end{aligned}$$

and similarly

$$\begin{aligned} Z_{k_1} &\in \sigma\left(\left\{\tilde{X}_{s_1,\ldots,s_d} \middle| (k_1 - 1)(m_1 + c_1 - 2) + 1 \leq s_1 \leq (k_1 + 1)(m_1 + c_1 - 2)\right\}\right), \\ Z_{k_1+1} &\in \sigma\left(\left\{\tilde{X}_{s_1,\ldots,s_d} \middle| k_1(m_1 + c_1 - 2) + 1 \leq s_1 \leq (k_1 + 2)(m_1 + c_1 - 2)\right\}\right) \end{aligned}$$

and the independence of the family $\left\{\tilde{X}_{s_1,\ldots,s_d} \mid 1 \leq s_j \leq \tilde{T}_j, 1 \leq j \leq d\right\}$ of random variables. Moreover, since $\tilde{X}_{s_1,\ldots,s_d}$ are identically distributed, the stationarity of the random variables Z_{k_1} is verified.
We observe that from the definition of the sequence $(Z_{k_1})_{1 \leq k_1 \leq L_1 - 1}$, the scan statistics random variable can be written as

$$S_{\mathbf{m}}(\mathbf{T}) = \max_{\substack{1 \leq i_j \leq T_j - m_j + 1 \\ j \in \{1,2,\ldots,d\}}} Y_{i_1, i_2, \ldots, i_d} = \max_{1 \leq k_1 \leq L_1 - 1} Z_{k_1}. \tag{4.7}$$

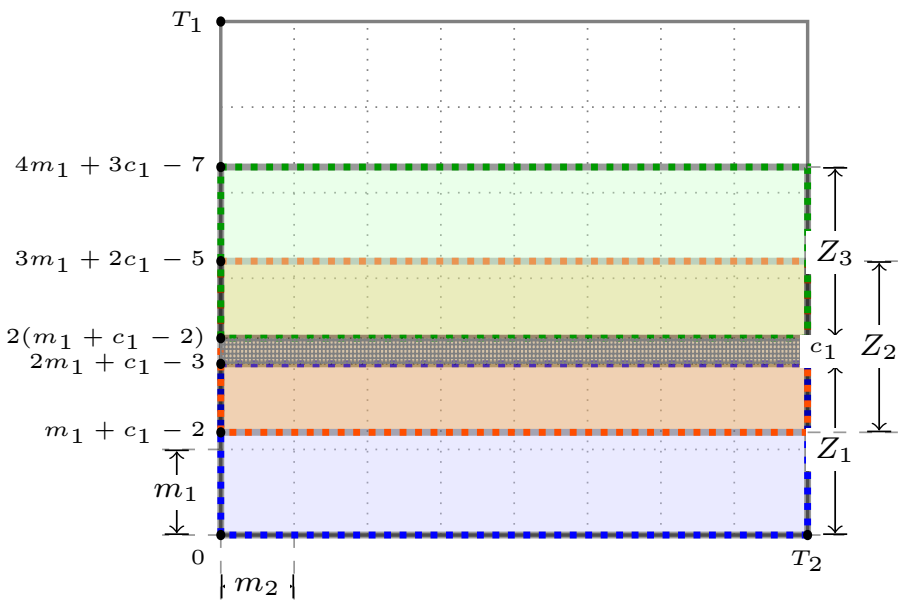


Figure 4.3: Illustration of Z_{k_1} emphasizing the 1-dependence

The foregoing relation is the key identity in the approximation process since it shows that $S_{\mathbf{m}}(\mathbf{T})$ is the maximum of a 1-dependent stationary sequence. Tools for the estimation of the distribution of this type of random variables were developed in Chapter 2. Similar with the methodology used in the previous chapter for the i.i.d. model, for the approximation process we employ the estimate from Theorem 2.2.9, repeatedly.
If, for $t_1 \in \{2, 3\}$, we take

$$Q_{t_1} = Q_{t_1}(n) = \mathbb{P}\left(\bigcap_{k_1=1}^{t_1-1} \{Z_{k_1} \leq n\}\right) = \mathbb{P}\left(\max_{\substack{1\leq i_1\leq (t_1-1)(m_1+c_1-2)\\ 1\leq i_j\leq (L_j-1)(m_j+c_j-1)\\ j\in\{2,\ldots,d\}}} Y_{i_1,i_2,\ldots,i_d} \leq n\right) \tag{4.8}$$

and n is such that $Q_2(n) \geq 0.9$, then, based on Theorem 2.2.9, we get the first step estimate

$$|Q_{\mathbf{m}}(\mathbf{T}) - H(Q_2, Q_3, L_1)| \leq (L_1 - 1)F(Q_2, L_1 - 1)(1 - Q_2)^2, \tag{4.9}$$

where $H(x, y, m)$ and $F(x, m)$ are evaluated via Eqs.(3.13) and (3.10), respectively. Observe that, even if we employed the same notations as in Section 3.2, Q_2, Q_3 and L_1, L_2 in the above equations differs from the corresponding ones in the i.i.d. case. For example, here Q_{t_1}'s represents the distribution of the d-dimensional discrete scan statistics over the strip

$$[0, (t_1 - 1)(m_1 + c_1 - 2) + m_1 - 1] \times [0, T_2] \cdots \times [0, T_d],$$

as opposed to the i.i.d. model where we scanned the smaller region

$$[0, t_1(m_1 - 1)] \times [0, T_2] \cdots \times [0, T_d].$$

In order to derive an approximation expression for $Q_{\mathbf{m}}(\mathbf{T})$, that involves simpler quantities, we need to iterate the above procedure at most d steps. In the subsequent, we present the general s-step, $1 \leq s \leq d$, routine. In this phase, the goal is to find an estimate for the distribution of the d-dimensional discrete scan statistics evaluated over the multidimensional rectangular regions

$$[0,(t_1-1)(m_1+c_1-2)+m_1-1]\times\dots[0,(t_{s-1}-1)(m_{s-1}+c_{s-1}-2)+m_{s-1}-1]\times[0,T_s]\cdots\times[0,T_d].$$

These distribution functions are denoted, for $(t_1, \dots, t_{s-1}) \in \{2,3\}^{s-1}$, by

$$Q_{t_1,t_2,\dots,t_{s-1}} = Q_{t_1,t_2,\dots,t_{s-1}}(n) = \mathbb{P}\left(\max_{\substack{1\leq i_l\leq (t_l-1)(m_l+c_l-2)\\ l\in\{1,\dots,s-1\}\\ 1\leq i_j\leq (L_j-1)(m_j+c_j-2)\\ j\in\{s,\dots,d\}}} Y_{i_1,i_2,\dots,i_d} \leq n \right). \quad (4.10)$$

To reduce the s dimension, we take for $t_l \in \{2,3\}$, $l \in \{1,\dots,s-1\}$ and $k_s \in \{1,2,\dots,L_s-1\}$ the random variables

$$Z_{k_s}^{(t_1,t_2,\dots,t_{s-1})} = \max_{\substack{1\leq i_l\leq (t_l-1)(m_l+c_l-2)\\ l\in\{1,2,\dots,s-1\}\\ (k_s-1)(m_s+c_s-2)+1\leq i_s\leq k_s(m_s+c_s-2)\\ 1\leq i_j\leq (L_j-1)(m_j+c_j-2)\\ j\in\{s+1,\dots,d\}}} Y_{i_1,i_2,\dots,i_d}, \quad (4.11)$$

which form, based on the same arguments as in step one, a 1-dependent stationary sequence. Moreover, these random variables satisfy the relation

$$Q_{t_1,t_2,\dots,t_{s-1}} = \mathbb{P}\left(\max_{1\leq k_s\leq L_s-1} Z_{k_s}^{(t_1,t_2,\dots,t_{s-1})} \leq n \right). \quad (4.12)$$

Notice that, from Eq.(4.10) and Eq.(4.11), we get the distribution function

$$Q_{t_1,t_2,\dots,t_s} = Q_{t_1,t_2,\dots,t_s}(n) = \mathbb{P}\left(\bigcap_{k_s=1}^{t_s-1} \{Z_{k_s}^{(t_1,t_2,\dots,t_{s-1})} \leq n\} \right). \quad (4.13)$$

If we take n such that $Q_{t_1,t_2,\dots,t_{s-1},2}(n) \geq 0.9$ then, applying the estimate from Theorem 2.2.9, we deduce the s step approximation

$$\left|Q_{t_1,\dots,t_{s-1}} - H\left(Q_{t_1,\dots,t_{s-1},2}, Q_{t_1,\dots,t_{s-1},3}, L_s\right)\right| \leq (L_s-1)F(Q_{t_1,\dots,t_{s-1},2}, L_s-1) \times (1-Q_{t_1,\dots,t_{s-1},2})^2. \quad (4.14)$$

Depending on the problem dimension, by substituting repeatedly, for each $s \in \{2,\dots,d\}$, the estimate from Eq.(4.14) in Eq.(4.9), we obtain an approximation

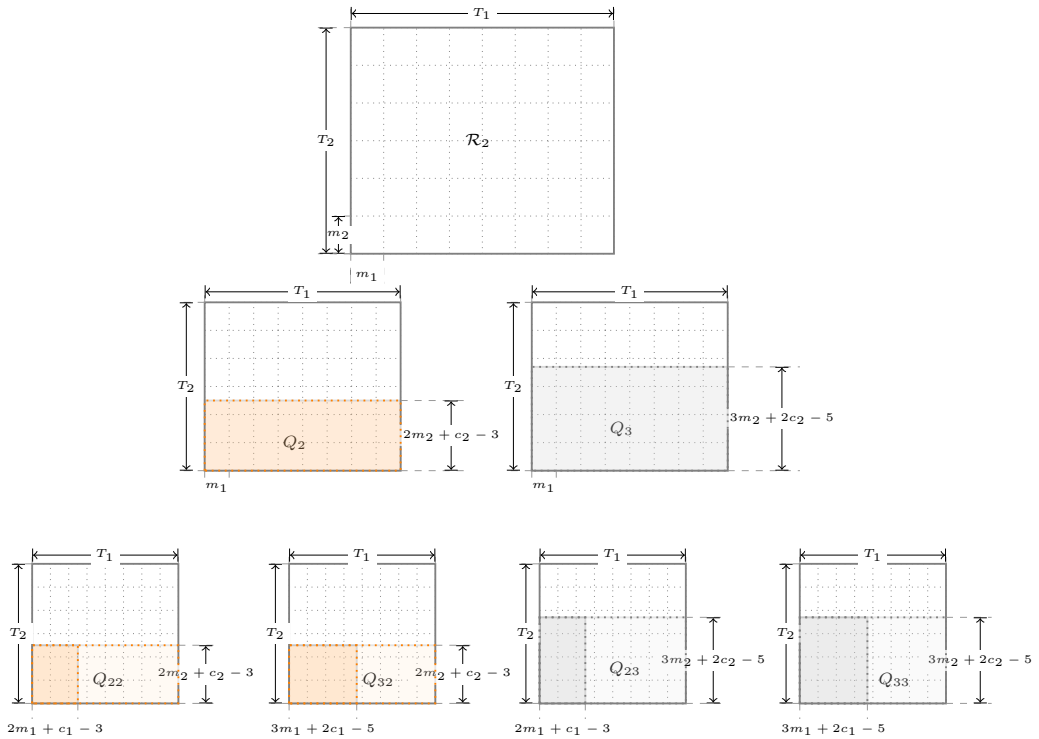


Figure 4.4: Illustration of the approximation process for $d = 2$

formula for the distribution of the d-dimensional scan statistics depending on the 2^d quantities $Q_{t_1,\dots,t_d}$, which will be evaluated by Monte Carlo simulation. To get a better understanding of the approximation process described above, we consider the particular case of two dimensional scan statistics. Figure 4.4 illustrates the diagram that summarizes the steps involved in this process.

Remark 4.2.1. *We should point out that if $X_{s_1,\dots,s_2} = \tilde{X}_{s_1,\dots,s_d}$, that is $c_1 = \cdots = c_d = 1$ and $\Pi(x) = x$, then all of the above formulas reduces to the ones given in Section 3.2.*

Remark 4.2.2. *Notice that if there are indices $j \in \{1, 2, \dots, d\}$ such that $\tilde{T}_j$ are not multiples of $m_j + c_j - 2$ then, by taking $L_j = \left\lfloor \frac{\tilde{T}_j}{m_j + c_j - 2} \right\rfloor$, we can approximate the distribution of the scan statistics $Q_{\mathbf{m}}(\mathbf{T})$ by the multi-linear interpolation procedure described in Remark 3.2.1.*

4.2.2 The associated error bounds

Since in the estimation process described in Section 4.2.1, by adopting the notations introduced in Section 3.1, we have obtained exactly the same type of relations as in the i.i.d. case, the resulted error bounds will also take the same expressions as those from Section 3.3. Such being the case, in this section we include only their formulas and not their derivation. As mentioned in Section 3.3, there are three

errors involved: the theoretical approximation error ($E_{app}(d)$), the simulation error that corresponds to the approximation error ($E_{sapp}(d)$) and the simulation error associated to the approximation formula ($E_{sf}(d)$). Their expressions are given in the subsequent.

a) *The theoretical approximation error*

Following the methodology described in Section 3.3.1 for the i.i.d. model, we have

$$E_{app}(d) = \sum_{s=1}^{d} (L_1 - 1)\cdots(L_s - 1) \times \sum_{t_1,\ldots,t_{s-1}\in\{2,3\}} F_{t_1,\ldots,t_{s-1}} \left(1 - \gamma_{t_1,\ldots,t_{s-1},2} + B_{t_1,\ldots,t_{s-1},2}\right)^2, \tag{4.15}$$

where

$$\gamma_{t_1,\ldots,t_s} = \begin{cases} H\left(\gamma_{t_1,\ldots,t_s,2}, \gamma_{t_1,\ldots,t_s,3}, L_{s+1}\right), \text{ for } 1 \le s \le d-1 \\ Q_{t_1,\ldots,t_d}, \text{ for } s = d, \end{cases} \tag{4.16}$$

$$\hat{Q}_{t_1,\ldots,t_s} = \begin{cases} H\left(\hat{Q}_{t_1,\ldots,t_s,2}, \hat{Q}_{t_1,\ldots,t_s,3}, L_{s+1}\right), \text{ for } 1 \le s \le d-1 \\ \hat{Q}_{t_1,\ldots,t_d}, \text{ for } s = d \end{cases} \tag{4.17}$$

and where for $2 \le s \le d$

$$F_{t_1,\ldots,t_{s-1}} = F\left(\hat{Q}_{t_1,\ldots,t_{s-1},2}, L_s - 1\right), \tag{4.18}$$

$$B_{t_1\ldots,t_{s-1}} = (L_s - 1)\left[F_{t_1,\ldots,t_{s-1}}\left(1 - \gamma_{t_1,\ldots,t_{s-1},2} + B_{t_1\ldots,t_{s-1},2}\right)^2 + \sum_{t_s\in\{2,3\}} B_{t_1\ldots,t_s}\right] \tag{4.19}$$

with $F = F\left(\hat{Q}_2, L_1 - 1\right)$ and $B_{t_1\ldots,t_d} = 0$.
Observe that in Eq.(4.15) we adopted the following convention for $s = 1$: $\sum_{t_1,t_0\in\{2,3\}} x = x$, $F_{t_1,t_0} = F$, $\gamma_{t_1,t_0,2} = \gamma_2$ and $B_{t_1,t_0,2} = B_2$.

b) *The simulation error that corresponds to the approximation error*

If the error between the true value of $Q_{t_1,\ldots,t_d}$ and the estimated value $\hat{Q}_{t_1,\ldots,t_d}$, resulted from a simulation methodology, is denoted by $\beta_{t_1,\ldots,t_d}$ then, the simulation error that appears from the approximation error given by Eq.(4.15) has the expression

$$E_{sapp}(d) = \sum_{s=1}^{d} (L_1 - 1)\cdots(L_s - 1) \sum_{t_1,\ldots,t_{s-1}\in\{2,3\}} F_{t_1,\ldots,t_{s-1}} \Big(1 - \hat{Q}_{t_1,\ldots,t_{s-1},2} + A_{t_1,\ldots,t_{s-1},2} + C_{t_1,\ldots,t_{s-1},2}\Big)^2, \tag{4.20}$$

where for $s = 1$: $\sum_{t_1,t_0\in\{2,3\}} x = x$, $F_{t_1,t_0} = F$, $\hat{Q}_{t_1,t_0,2} = \hat{Q}_2$, $A_{t_1,t_0,2} = A_2$ and $C_{t_1,t_0,2} = C_2$. The quantities $A_{t_1,\dots,t_s}$ and $C_{t_1,\dots,t_s}$, that appear in the above formula, are computed via

$$A_{t_1,\dots,t_{s-1}} = (L_s - 1)\dots(L_d - 1) \sum_{t_s,\dots,t_d\in\{2,3\}} \beta_{t_1,\dots,t_d} \tag{4.21}$$

and

$$C_{t_1\dots,t_{s-1}} = (L_s - 1)\left[F_{t_1,\dots,t_{s-1}}\left(1 - \hat{Q}_{t_1,\dots,t_{s-1},2} + A_{t_1\dots,t_{s-1},2} + C_{t_1\dots,t_{s-1},2}\right)^2 + \sum_{t_s\in\{2,3\}} C_{t_1\dots,t_s}\right], \tag{4.22}$$

for $2 \le s \le d$ while for $s = d$ we have $A_{t_1,\dots,t_d} = \beta_{t_1,\dots,t_d}$ and $C_{t_1\dots,t_d} = 0$.

c) *The simulation error associated to the approximation formula*

This simulation error arise from the difference

$$\left|H(\gamma_2,\gamma_3,L_1) - H\left(\hat{Q}_2,\hat{Q}_3,L_1\right)\right|$$

and is given by the relation

$$E_{sf}(d) = (L_1 - 1)\dots(L_d - 1) \sum_{t_1,\dots,t_d\in\{2,3\}} \beta_{t_1,\dots,t_d}. \tag{4.23}$$

The total error is obtained by adding the two simulation error terms from Eq.(4.20) and Eq.(4.23)

$$E_{total}(d) = E_{sf}(d) + E_{sapp}(d). \tag{4.24}$$

4.3 Examples and numerical results

In order to illustrate the efficiency of the approximation and the error bounds obtained in Section 4.2, in this section we present several examples for the particular cases of one and two dimensional discrete scan statistics.

4.3.1 Example 1: A one dimensional Bernoulli model

We start this section with an example of a one dimensional block-factor model, for which we can compute exactly the distribution functions that appear in our proposed approximation: Q_2 and Q_3. This model is based on the parametrization introduced by [Haiman, 2012] (see also [Haiman and Preda, 2013]).
In the block-factor framework introduced in Section 4.1, we take $d = 1$, $x_1^{(1)} = 0$, $x_2^{(1)} = 1$ and we have $c_1 = 2$ and $T_1 = \tilde{T}_1 - 1$.

Let, for $1 \leq t \leq \tilde{T}_1$, $W_t, W'_t \sim \mathcal{B}(p)$, with $0 \leq p \leq \frac{1}{2}$ and $W''_t \sim \mathcal{B}\left(\frac{1}{2}\right)$ be three sequences of $0-1$ Bernoulli random variables such that $(W_t, W'_t)_{1\leq t\leq \tilde{T}_1}$ are independent random variables satisfying $\mathbb{P}(W_t = i, W'_t = j) = q(i,j)$ and W''_t's are independent and independent of $(W_t, W'_t)_{1\leq t\leq \tilde{T}_1}$. The joint probabilities $q(i,j)$ are evaluated via the relation

$$q(i,j) = \begin{cases} 1-2p, \text{ if } (i,j) = (0,0) \\ p, \text{ if } (i,j) = (0,1) \text{ or } (i,j) = (1,0) \\ 0, \text{ if } (i,j) = (1,1). \end{cases} \tag{4.25}$$

Let $\tilde{X}_{s_1} = \left(W_{s_1}, W'_{s_1}, W''_{s_1}\right)$ for $1 \leq s_1 \leq \tilde{T}_1$, which forms an i.i.d. sequence of random variables and observe that the 1-way tensor that encapsulates the block-factor model is given by

$$\mathfrak{X}_{s_1} = \left(\tilde{X}_{s_1}, \tilde{X}_{s_1+1}\right). \tag{4.26}$$

We define, for $1 \leq s_1 \leq T_1$, our dependent model by

$$X_{s_1} = \Pi\left(\tilde{X}_{s_1}, \tilde{X}_{s_1+1}\right) = \begin{cases} W_{s_1+1}, \text{ if } W''_{s_1+1} = 0 \\ W'_{s_1}, \text{ if } W''_{s_1+1} = 1 \end{cases} \tag{4.27}$$

and we notice that X_{s_1}'s, being expressed as a 2 block-factor, form a 1-dependent, stationary sequence of $0-1$ Bernoulli of parameter p random variables. This type of 1-dependent sequence was studied in [Haiman, 2012]. The author showed (see [Haiman, 2012, Lemma 3]) that the joint distribution of (X_1, X_2) satisfies the equation

$$\mathbb{P}(X_1 = i, X_2 = j) = p(i,j), \tag{4.28}$$

where

$$p(i,j) = \begin{cases} 1-2p+\frac{3}{4}p^2, \text{ if } (i,j) = (0,0) \\ p-\frac{3}{4}p^2, \text{ if } (i,j) = (0,1) \text{ or } (i,j) = (1,0) \\ \frac{3}{4}p^2, \text{ if } (i,j) = (1,1). \end{cases} \tag{4.29}$$

Moreover, from [Haiman, 2012, Theorem 1], we have that the joint distribution of the sequence $(X_{s_1})_{1\leq s_1\leq T_1}$ verifies the recurrence formula

$$\begin{aligned} \mathbb{P}(X_1 = a_1, \ldots, X_{k+1} = a_{k+1}) &= \mathbb{P}(X_1 = a_1, \ldots, X_k = a_k)\,\mathbb{P}(X_{k+1} = a_{k+1}) \\ &+ \mathbb{P}(X_1 = a_1, \ldots, X_{k-1} = a_{k-1})\left[\mathbb{P}(X_k = a_k, X_{k+1} = a_{k+1})\right. \\ &\left.-\mathbb{P}(X_k = a_k)\,\mathbb{P}(X_k = a_k)\right], \end{aligned} \tag{4.30}$$

for $2 \leq k \leq T_1 - 1$ and $(a_1, \ldots, a_{k+1}) \in \{0,1\}^{k+1}$, which gives the proper tools for finding the exact values of Q_2 and Q_3. Following the approach from [Haiman and Preda, 2013, Section 2], let $K \geq 2$ be a positive integer and denote with

$$V_K(b) = \left\{ \mathbf{a} = (a_1, \ldots, a_K) \in \{0,1\}^K \middle| \max_{t=1,\ldots,K-m_1+1} \sum_{i=t}^{t+m_1-1} a_i = b \right\}, \tag{4.31}$$

the set of all binary sequences of length K for which the scan statistics with window of size m_1 takes the value b. Therefore, the value of the distribution function $Q_{m_1}(K)$ is computed by

$$Q_{m_1}(K) = \mathbb{P}\left(S_{m_1}(K) \leq n\right) = \sum_{b=0}^{n} \sum_{\mathbf{a} \in V_K(b)} \mathbb{P}\left(X_1 = a_1, \ldots, X_K = a_K\right), \qquad (4.32)$$

where the last quantity is evaluated via the recurrence formula from Eq.(4.30). Particularizing for $K = 2m_1 - 1$ and $K = 3m_1 - 1$ in Eq.(4.32), we obtain the exact values of Q_2 and Q_3, respectively.
To show the accuracy of the proposed approximation for the one dimensional scan statistics distribution $\mathbb{P}(S_{m_1}(T_1) \leq n)$ and the sharpness of the error bounds, we present in Table 4.1, a numerical study for $m_1 = 8$, $T_1 = 1000$ and selected values of the parameter p. The second and third column gives the exact values of Q_2 and Q_3, while the next two columns show the approximation and the corresponding error bound[1].

n	Q_2 Eq.(4.32)	Q_3 Eq.(4.32)	AppH	$E_{app}(1)$ Eq.(4.15)	LowB Eq.(4.35)	UppB Eq.(4.36)
			$p = 0.1$			
3	0.985914	0.974354	0.231796	0.031264	0.225598	0.236072
4	0.998911	0.997931	0.885221	0.000153	0.885081	0.885340
5	0.999958	0.999917	0.995014	0.000000	0.995014	0.995014
6	0.999999	0.999999	0.999914	0.000000	0.999914	0.999914
7	1.000000	1.000000	1.000000	0.000000	1.000000	1.000000
			$p = 0.2$			
5	0.997750	0.995697	0.774337	0.000667	0.773801	0.774784
6	0.999915	0.999834	0.989987	0.000001	0.989986	0.989988
7	0.999999	0.999998	0.999890	0.000000	0.999890	0.999890
8	1.000000	1.000000	1.000000	0.000000	1.000000	1.000000
			$p = 0.3$			
5	0.980187	0.963136	0.113912	0.067341	0.107535	0.118347
6	0.998801	0.997676	0.869460	0.000186	0.869288	0.869611
7	0.999981	0.999963	0.997748	0.000000	0.997748	0.997748
8	1.000000	1.000000	1.000000	0.000000	1.000000	1.000000

Table 4.1: One dimensional Bernoulli block-factor model: $m_1 = 8$, $T_1 = 1000$

Notice that, for comparison reasons, we have also included a lower ($LowB$) and an upper bound ($UppB$). These margins are derived from the following result, which extends the inequalities developed by [Glaz and Naus, 1991] (see also [Wang et al., 2012]), presented in Section 1.1.3.

[1] Here we only have one error bound, namely the theoretical one, since the values of Q_2 and Q_3 are computed exactly

Proposition 4.3.1. *Let $\bar{Y}_1, \bar{Y}_2, \ldots$ be a sequence of associated*[2] *(see [Esary et al., 1967]) and l-dependent random variables generated by a $(l+1)$ block-factor. If we denote the distribution of the maximum of the sequence with $Q(M) = \mathbb{P}\left(\max_{1\le t\le M} \bar{Y}_t \le n\right)$, then for $M \ge l+2$ we have*

$$Q(M) \ge \frac{Q(l+2)}{\left[1+\frac{Q(l+1)-Q(l+2)}{Q(l+1)Q(l+2)}\right]^{M-l-2}}, \tag{4.33}$$

$$Q(M) \le Q(l+2)\left[1-Q(l+1)+Q(l+2)\right]^{M-l-2}. \tag{4.34}$$

In the particular case of one dimensional discrete scan statistics over the block-factor model from Section 4.1, we have, in accordance with Remark 4.1.2, that X_{s_1} is (c_1-1) dependent so the sequence of moving sums $Y_1, \ldots, Y_{T_1-m_1+1}$ is $l = m_1+c_1-2$ dependent. Thus, the distribution function $Q_{m_1}(T_1)$, for $T_1 \ge 2m_1+c_1-1$, is bounded by

$$Q_{m_1}(T_1) \ge \frac{Q_{m_1}(2m_1+c_1-1)}{\left[1+\frac{Q_{m_1}(2m_1+c_1-2)-Q_{m_1}(2m_1+c_1-1)}{Q_{m_1}(2m_1+c_1-2)Q_{m_1}(2m_1+c_1-1)}\right]^{T_1-(2m_1+c_1-1)}}, \tag{4.35}$$

$$\begin{aligned} Q_{m_1}(T_1) \le Q_{m_1}(2m_1+c_1-1)\left[1-Q_{m_1}(2m_1+c_1-2)\right. \\ \left.+Q_{m_1}(2m_1+c_1-1)\right]^{T_1-(2m_1+c_1-1)}. \end{aligned} \tag{4.36}$$

We observe that, if in the foregoing relations we take $c_1 = 1$, that is we consider the i.i.d. model, then we get exactly the bounds presented in Section 1.1.3. From the numerical values in Table 4.1, we see that these bounds are very tight and also that our estimate is very accurate.

Remark 4.3.2. *Notice that in the text of Proposition 4.3.1 we assumed that the random variables $\bar{Y}_1, \bar{Y}_2, \ldots$ are associated and generated by a $(l+1)$ block-factor. To construct such sequences, it is enough to take the function that defines the block-factor to be increasing (or decreasing) in each argument (see for example [Oliveira, 2012, Chapter 1] or [Esary et al., 1967, Theorem 2.1 and property (P_4)]).*

Remark 4.3.3. *The proof of Proposition 4.3.1, due to the additional hypothesis of associated random variables, follows closely the proof steps of the corresponding result in [Glaz and Naus, 1991] and will be omitted here.*

[2] We say that the random variables $\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_n$ are associated if, given two coordinatewise nondecreasing functions $f, g : \mathbb{R}^n \to \mathbb{R}$ then $Cov\left[f\left(\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_n\right), g\left(\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_n\right)\right] \ge 0$, whenever the covariance exists.

4.3.2 Example 2: Length of the longest increasing run

We continue our list of examples with one that belongs to the class of runs statistics. In the subsequent we employ the notations introduced in Section 4.1.
Let $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_{\tilde{T}_1}$ be a sequence of independent and identically distributed random variables with the common distribution F. We say that the subsequence $\left(\tilde{X}_k, \ldots, \tilde{X}_{k+l-1}\right)$ forms an *increasing run* (or ascending run) of length $l \geq 1$, starting at position $k \geq 1$, if it verifies the following relation

$$\tilde{X}_{k-1} > \tilde{X}_k < \tilde{X}_{k+1} < \cdots < \tilde{X}_{k+l-1} > \tilde{X}_{k+l}. \tag{4.37}$$

We denote the length of the longest increasing run among the first $\tilde{T}_1$ random variables by $M_{\tilde{T}_1}$. This run statistics plays an important role in many applications in fields such computer science, reliability theory or quality control. The asymptotic behaviour of $M_{\tilde{T}_1}$ has been investigated by several authors depending on the common distribution, F. In the case of a continuous distribution, [Pittel, 1981] (see also [Frolov and Martikainen, 1999]) has shown that this behaviour does not depend on the common law. For the particular setting of uniform $\mathcal{U}([0,1])$ random variables, this problem was addressed by [Révész, 1983], [Grill, 1987] and [Novak, 1992]. Under the assumption that the distribution F is discrete, the limit behaviour of $M_{\tilde{T}_1}$ depends strongly on the common law F, as [Csaki and Foldes, 1996] (see also [Grabner et al., 2003] and [Eryilmaz, 2006]) proved for the case of geometric and Poisson distribution. In [Louchard, 2005], the case of discrete uniform distribution is investigated, while in [Mitton et al., 2010], the authors study the asymptotic distribution of $M_{\tilde{T}_1}$ when the variables are uniformly distributed but not independent.
In this example, we evaluate the distribution of the length of the longest increasing run using the methodology developed in this chapter. The idea is to express the distribution of the random variable $M_{\tilde{T}_1}$ in terms of the distribution of the scan statistics random variable. In the block-factor setting described in Section 4.1, we take $d = 1$, $x_1^{(1)} = 0$, $x_2^{(1)} = 1$ and $T_1 = \tilde{T}_1 - 1$. It follows that, for $1 \leq s_1 \leq T_1$, the 1-way tensor $\mathfrak{X}_{s_1}$ is $\left(\tilde{X}_{s_1}, \tilde{X}_{s_1+1}\right)$ and if we define the block-factor transformation $\Pi : \mathbb{R}^2 \to \mathbb{R}$ by

$$\Pi(x,y) = \begin{cases} 1, \text{ if } x < y \\ 0, \text{ otherwise} \end{cases} \tag{4.38}$$

then, our block-factor model becomes

$$X_{s_1} = \mathbf{1}_{\tilde{X}_{s_1} < \tilde{X}_{s_1+1}}. \tag{4.39}$$

Clearly, the foregoing equation shows that $X_1, \ldots, X_{T_1}$ form a 1-dependent and stationary sequence of random variables.
Notice that the distribution of $M_{\tilde{T}_1}$ and the distribution of the length of the longest run of ones[3], L_{T_1}, among the first T_1 binary random variables X_{s_1}, are related and

[3] This statistic is also known as the length of the longest success run or head run and was extensively studied in the literature. One can consult the monographs of [Balakrishnan and Koutras, 2002] and [Fu and Lou, 2003] for applications and further results concerning this statistic.

satisfy the following identity

$$\mathbb{P}\left(M_{\tilde{T}_1} \leq m_1\right) = \mathbb{P}\left(L_{T_1} < m_1\right), \text{ for } m_1 \geq 1. \tag{4.40}$$

Moreover, the random variable L_{T_1} can be interpreted as a particular case of the scan statistics random variable and between the two we have the relation

$$\mathbb{P}\left(L_{T_1} \geq m_1\right) = \mathbb{P}\left(S_{m_1}(T_1) \geq m_1\right) = \mathbb{P}\left(S_{m_1}(T_1) = m_1\right). \tag{4.41}$$

Hence, combining Eq.(4.40) and Eq.(4.41), we can express the distribution of the length of the longest increasing run as

$$\mathbb{P}\left(M_{\tilde{T}_1} \leq m_1\right) = \mathbb{P}\left(S_{m_1}(T_1) < m_1\right). \tag{4.42}$$

Thus, we can estimate the distribution of $M_{\tilde{T}_1}$ using the foregoing identity and the approximations developed in this chapter for the discrete scan statistics random variable.

We should also note that [Novak, 1992] studied the asymptotic behaviour of L_{T_1} over a sequence of m-dependent binary random variables. The author showed that, given a stationary m-dependent sequence of random variables with values 0 and 1, $(W_k)_{k\geq 1}$, if there exist positive constants t, C such that

$$\mathbb{P}\left(W_{k+1} = 1 | W_1 = \cdots = W_k = 1\right) \geq \frac{1}{Ck^t}, \text{ for all } k \geq C, \tag{4.43}$$

then, as $N \to \infty$

$$\max_{1\leq k\leq N} \left|\mathbb{P}\left(L_N < k\right) - e^{-Nr(k)}\right| = \mathcal{O}\left(\frac{(\ln N)^h}{N}\right), \tag{4.44}$$

where $r(k) = \mathbb{P}\left(W_1 = \cdots = W_k = 1\right) - \mathbb{P}\left(W_1 = \cdots = W_{k+1} = 1\right)$ and $h = mt \vee 1$.

In order to illustrate the accuracy of the approximation of $M_{\tilde{T}_1}$ based on scan statistics, using the methodology developed in Section 4.2.1, we consider that the random variables $\tilde{X}_{s_1}$'s have a common uniform $\mathcal{U}\left([0,1]\right)$ distribution. Simple calculations show that $\mathbb{P}\left(X_1 = \cdots = X_k = 1\right) = \frac{1}{(k+1)!}$ and

$$\mathbb{P}\left(X_{k+1} = 1 | X_1 = \cdots = X_k = 1\right) = \frac{1}{k+2} \geq \frac{1}{2k}, \tag{4.45}$$

thus $C = 2$, $t = 1$ and $h = 1$. In the context of our particular situation, the result of [Novak, 1992] in Eq.(4.44) becomes:

$$\max_{1\leq m_1\leq T_1} \left|\mathbb{P}\left(L_{T_1} < m_1\right) - e^{-T_1 r(m_1)}\right| = \mathcal{O}\left(\frac{\ln T_1}{T_1}\right), \tag{4.46}$$

where $r(m_1) = \mathbb{P}\left(X_1 = \cdots = X_{m_1} = 1\right) - \mathbb{P}\left(X_1 = \cdots = X_{m_1+1} = 1\right) = \frac{m_1+1}{(m_1+2)!}$.

In Table 4.2, we consider a numerical compared study between the simulated value (column *Sim*), the approximation based on scan statistics (column *AppH*) and the limit distribution (column *LimApp*) of the distribution of the length of the longest increasing run, $\mathbb{P}\left(M_{\tilde{T}_1} \leq m_1\right)$, in a sequence of $\tilde{T}_1 = 10001$ random variables distributed uniformly over $[0,1]$. The results show that both our method and the asymptotic approximation in Eq.(4.43) are very accurate.

m_1	Sim	AppH	$E_{total}(1)$ Eq(4.24)	LimApp Eq(4.43)
5	0.00000700	0.00000733	0.14860299	0.00000676
6	0.17567262	0.17937645	0.01089628	0.17620431
7	0.80257424	0.80362353	0.00110990	0.80215088
8	0.97548510	0.97566460	0.00011579	0.97550345
9	0.99749821	0.99751049	0.00001114	0.99749792
10	0.99977074	0.99977183	0.00000098	0.99977038
11	0.99998075	0.99998083	0.00000008	0.99998073
12	0.99999851	0.99999851	0.00000001	0.99999851
13	0.99999989	0.99999989	0.00000000	0.99999989
14	0.99999999	0.99999999	0.00000000	0.99999999
15	1.00000000	1.00000000	0.00000000	1.00000000

Table 4.2: The distribution of the length of the longest increasing run: $\tilde{T}_1 = 10001$, $ITER_{sim} = 10^4$, $ITER_{app} = 10^5$

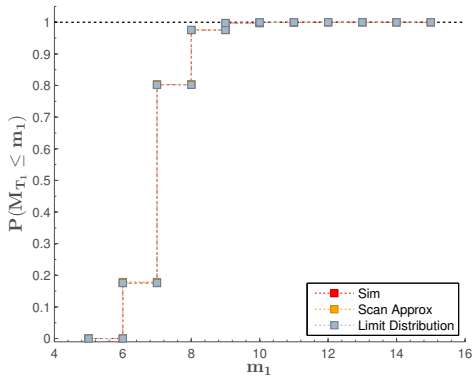


Figure 4.5: Cumulative distribution function for approximation, simulation and limit law

Remark 4.3.4. *For the simulated values and the estimation of the distribution functions Q_2 and Q_3 that appear in the scan approximation formula (see Eq.(3.91)), we employed the importance sampling algorithm (Algorithm 1) presented in the context of i.i.d. random variables in Section 3.4.2, with $ITER_{sim} = 10^4$ and $ITER_{app} = 10^5$ iterations of the algorithm, respectively. We see that in our setting, the Bonferroni bound can be easily computed via $B(1) = \frac{T_1 - m_1 + 1}{(m_1+1)!}$. The second step in the algorithm is similar to the one described in Example 3.4.1 and is implemented using a simple sorting procedure of $m_1 + 1$ independent $\mathcal{U}([0,1])$ random variables.*

4.3.3 Example 3: Moving average of order q model

In this third example, we consider the particular situation of the one dimensional discrete scan statistics over a $MA(q)$ model. In the block-factor model introduced in Section 4.1 we take $d = 1$, $x_1^{(1)} = 0$, $x_2^{(1)} = q$ for $q \geq 1$ a positive integer and $\tilde{X}_1$, $\tilde{X}_2$, ..., $\tilde{X}_{\tilde{T}_1}$ a sequence of independent and identically distributed Gaussian random variables with known mean μ and variance σ^2. We observe that, based on the notations used in Section 4.1, $c_1 = q + 1$, $T_1 = \tilde{T}_1 - q$ and for $s_1 \in \{1, \ldots, T_1\}$, the 1-way tensor $\mathfrak{X}_{s_1}$ becomes

$$\mathfrak{X}_{s_1} = \left(\tilde{X}_{s_1}, \tilde{X}_{s_1+1}, \ldots, \tilde{X}_{s_1+q}\right). \tag{4.47}$$

Let $\mathbf{a} = (a_1, \ldots, a_{q+1}) \in \mathbb{R}^{q+1}$ be a fixed non null vector and take $\Pi : \mathbb{R}^{q+1} \to \mathbb{R}$, the (measurable) transformation that defines the block-factor model, to be equal with

$$\Pi(x_1, \ldots, x_{q+1}) = a_1 x_1 + a_2 x_2 + \cdots + a_{q+1} x_{q+1}. \tag{4.48}$$

Following Eq.(4.2), our dependent model is defined by the relation

$$X_{s_1} = \Pi\left(\mathfrak{X}_{s_1}\right) = a_1 \tilde{X}_{s_1} + a_2 \tilde{X}_{s_1+1} + \cdots + a_{q+1} \tilde{X}_{s_1+q}, \ 1 \leq s_1 \leq T_1. \tag{4.49}$$

Clearly, from Eq.(4.49), the random variables $X_1, \ldots, X_{T_1}$ form a moving average of order q. Notice that the moving sums Y_{i_1}, $1 \leq i_1 \leq T_1 - m_1 + 1$ can be expressed as

$$Y_{i_1} = \sum_{s_1=i_1}^{i_1+m_1-1} X_{s_1} = b_1 \tilde{X}_{i_1} + b_2 \tilde{X}_{i_1+1} + \cdots + b_{m_1+q} \tilde{X}_{i_1+m_1-1+q}, \tag{4.50}$$

where the coefficients $b_1, \ldots, b_{m_1+q}$ are evaluated by

a) For $m_1 \geq q$,

$$b_k = \begin{cases} \sum_{j=1}^{k} a_j & , k \in \{1, \ldots, q\} \\ \sum_{j=1}^{q+1} a_j & , k \in \{q+1, \ldots, m_1\} \\ \sum_{j=k-m_1+1}^{q+1} a_j & , k \in \{m_1+1, \ldots, m_1+q\} \end{cases} \tag{4.51}$$

b) For $m_1 < q$,

$$b_k = \begin{cases} \sum_{j=1}^{k} a_j & , k \in \{1, \ldots, m_1\} \\ \sum_{j=k-m_1+1}^{k} a_j & , k \in \{m_1+1, \ldots, q\} \\ \sum_{j=k-m_1+1}^{q+1} a_j & , k \in \{q+1, \ldots, m_1+q\}. \end{cases} \tag{4.52}$$

Therefore, for each $i_1 \in \{1, \ldots, T_1 - m_1 + 1\}$, the random variable Y_{i_1} follows a normal distribution with mean $\mathbb{E}[Y_{i_1}] = (b_1 + \cdots + b_{m+q})\mu$ and variance $Var[Y_{i_1}] = \left(b_1^2 + \cdots + b_{m+q}^2\right)\sigma^2$. Moreover, it is not hard to see (one can use the same type of arguments as in Lemma 3.4.3) that the covariance matrix $\Sigma = \{Cov[Y_t, Y_s]\}$ has the entries

$$Cov[Y_t, Y_s] = \begin{cases} \left(\sum_{j=1}^{m_1+q-|t-s|} b_j b_{|t-s|+j} \right) \sigma^2 & , |t-s| \leq m_1 + q - 1 \\ 0 & , \text{otherwise.} \end{cases} \tag{4.53}$$

Given the mean and the covariance matrix of the vector $(Y_1, \ldots, Y_{T_1-m_1+1})$, one can use the importance sampling algorithm of [Naiman and Priebe, 2001] (see also [Malley et al., 2002]) detailed in Section 3.4.2 (Example 3.4.2) or the one of [Shi et al., 2007] presented in Section 3.4.4 to estimate the distribution of the one dimensional discrete scan statistics $S_{m_1}(T_1)$. Another way is to use the quasi-Monte

Carlo algorithm developed by [Genz and Bretz, 2009] to approximate the multivariate normal distribution. Following the comparison study presented in Section 3.4.4, in this example we adopt the first importance sampling procedure.
In order to evaluate the accuracy of the approximation developed in Section 4.2.1, we consider $q = 2$, $T_1 = 1000$, $m_1 = 20$, $\tilde{X}_i \sim \mathcal{N}(0, 1)$ and the coefficients of the moving average model to be $(a_1, a_2, a_3) = (0.3, 0.1, 0.5)$. We compare our approximation with the one (column $AppPT$) given in [Wang and Glaz, 2013] (see also [Wang, 2013, Chapter 4]). In Table 4.3, we present numerical results for the setting described above. In our algorithms we used $ITER_{app} = 10^6$ iterations for the approximation and $ITER_{sim} = 10^5$ replicas for the simulation.

n	Sim	AppPT Eq.(1.46)	AppH	$E_{sapp}(1)$ Eq.(4.20)	$E_{sf}(1)$ Eq.(4.23)	Total Error
11	0.582252	0.589479	0.584355	0.011503	0.003653	0.015156
12	0.770971	0.773700	0.771446	0.002319	0.001691	0.004010
13	0.889986	0.890009	0.889431	0.000434	0.000733	0.001167
14	0.951529	0.954536	0.951723	0.000073	0.000297	0.000370
15	0.980653	0.982433	0.980675	0.000011	0.000113	0.000124
16	0.992827	0.993690	0.992791	0.000001	0.000040	0.000042
17	0.997486	0.995471	0.997499	0.000000	0.000013	0.000014
18	0.999186	0.999411	0.999188	0.000000	0.000004	0.000004
19	0.999754	0.999717	0.999754	0.000000	0.000001	0.000001
20	0.999930	1	0.999930	0.000000	0.000000	0.000000

Table 4.3: MA(2) model: $m_1 = 20$, $T_1 = 1000$, $X_i = 0.3\tilde{X}_i + 0.1\tilde{X}_{i+1} + 0.5\tilde{X}_{i+2}$, $ITER_{app} = 10^6$, $ITER_{sim} = 10^5$

In Figure 4.6, we illustrate the cumulative distribution functions obtained by approximation and simulation. For the approximation we present also the corresponding lower and upper bounds (computed from the total error of the approximation process, the last column in Table 4.3).

4.3.4 Example 4: A game of minesweeper

This example draws its motivation (and name) from the well known computer game of *Minesweeper*, whose objective is to detect all the mines from a grided minefield in such a way that no bomb is detonated. Our two dimensional model can be described as follows.
Let $d = 2$, $\tilde{T}_1$, $\tilde{T}_2$ be positive integers and $\left\{\tilde{X}_{s_1,s_2} \mid 1 \leq s_1 \leq \tilde{T}_1, 1 \leq s_2 \leq \tilde{T}_2\right\}$ be a family of i.i.d. Bernoulli random variables of parameter p. We interpret the random variable $\tilde{X}_{s_1,s_2}$ as representing the presence ($\tilde{X}_{s_1,s_2} = 1$) or the absence ($\tilde{X}_{s_1,s_2} = 0$) of a mine in the elementary square subregion $\tilde{r}_2(s_1, s_2) = [s_1 - 1, s_1] \times [s_2 - 1, s_2]$, within the region $\tilde{\mathcal{R}}_2 = [0, \tilde{T}_1] \times [0, \tilde{T}_2]$.
In this example, we consider $x_1^{(1)} = x_2^{(1)} = 1$ and $x_1^{(2)} = x_2^{(2)} = 1$. Based on the notations introduced in Section 4.1, we observe that $c_1 = c_2 = 3$, $T_1 = \tilde{T}_1 - 2$ and

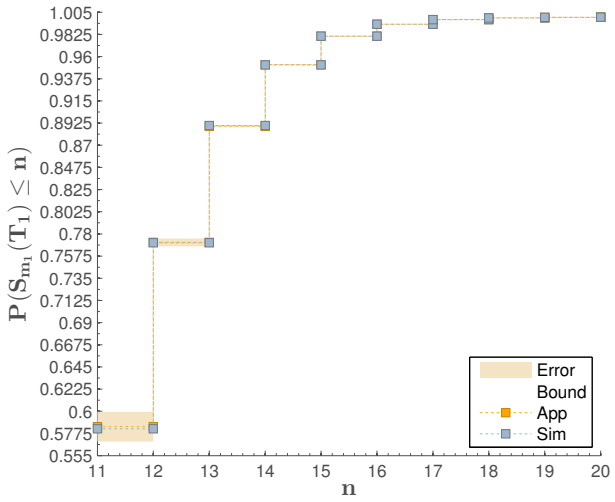


Figure 4.6: Cumulative distribution function for approximation and simulation along with the corresponding error under MA model

$T_2 = \tilde{T}_2 - 2$. For each pair $(s_1, s_2) \in \{2, \ldots, \tilde{T}_1 - 1\} \times \{2, \ldots, \tilde{T}_2 - 1\}$, the 2-way tensor $\mathfrak{X}_{s_1,s_2}$ is given by

$$\mathfrak{X}_{s_1,s_2}(j_1, j_2) = \tilde{X}_{s_1+j_1-2,s_2+j_2-2}, \text{ where } 1 \leq j_i \leq c_i,\ i \in \{1, 2\}. \tag{4.54}$$

Since the elements of the tensor $\mathfrak{X}_{s_1,s_2}$ are arranged in a 3×3 matrix, we define the block-factor real valued transformation Π on the set of matrices $\mathcal{M}_{3,3}(\mathbb{R})$, such that it verifies the relation

$$\Pi \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \sum_{1 \leq s,t \leq 3} a_{st} - a_{22}. \tag{4.55}$$

From Eq.(4.2) and Eq.(4.55), our dependent model is defined as

$$X_{s_1,s_2} = \Pi\left(\mathfrak{X}_{s_1+1,s_2+1}\right) = \sum_{\substack{(i,j) \in \{0,1,2\}^2 \\ (i,j) \neq (1,1)}} \tilde{X}_{s_1+i,s_2+j}. \tag{4.56}$$

From the foregoing relation, we can interpret the random variable X_{s_1,s_2} as the number of neighboring mines associated with the location (s_1, s_2). In Figure 4.7, we present a realization of the introduced model. On the left, we have the realization of the initial set of random variables where a gray square represents the presence of a mine, while the white square signifies the absence of a mine. On the right side, we have the realization of the X_{s_1,s_2} according to the transformation Π from Eq.(4.55), that is the corresponding number of neighboring mines associated to each site.
We present numerical results (Table 4.4-Table 4.11) for the described block-factor model with $\tilde{T}_1 = \tilde{T}_2 = 44$ (that is $T_1 = T_2 = 42$), $m_1 = m_2 = 3$ and the underlying

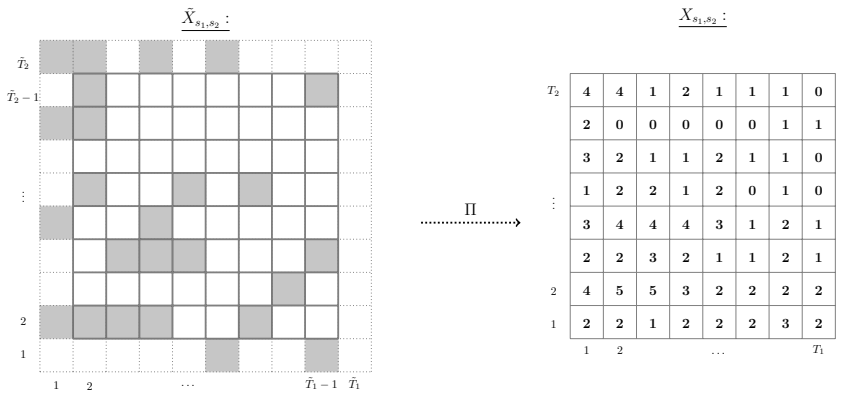


Figure 4.7: A realization of the minesweeper related model

random field generated by i.i.d. Bernoulli random variables of parameter p ($\tilde{X}_{s_1,s_2} \sim \mathcal{B}(p)$) in the range $\{0.1, 0.3, 0.5, 0.7\}$. We also include numerical values for the corresponding i.i.d. model: $T_1 = T_2 = 42$, $m_1 = m_2 = 3$ and $X_{s_1,s_2} \sim \mathcal{B}(8, p)$.

n	Sim	AppH	$E_{sapp}(2)$ Eq.(4.20)	$E_{sf}(2)$ Eq.(4.23)	Total Error
29	0.828763	0.813457	0.018678	0.024528	0.043205
30	0.886702	0.875875	0.006135	0.010670	0.016805
31	0.930094	0.922997	0.001912	0.005374	0.007286
32	0.957297	0.953079	0.000628	0.003290	0.003918
33	0.974541	0.971980	0.000204	0.002239	0.002443
34	0.985523	0.984022	0.000063	0.001588	0.001651
35	0.991524	0.990718	0.000020	0.001171	0.001191
36	0.995301	0.994885	0.000006	0.000854	0.000860
37	0.997492	0.997253	0.000002	0.000617	0.000619
38	0.998668	0.998547	0.000000	0.000447	0.000447
39	0.999313	0.999272	0.000000	0.000319	0.000319
40	0.999653	0.999629	0.000000	0.000231	0.000231
41	0.999826	0.999808	0.000000	0.000164	0.000164
42	0.999916	0.999911	0.000000	0.000116	0.000116
43	0.999963	0.999959	0.000000	0.000079	0.000079
44	0.999981	0.999979	0.000000	0.000054	0.000054
45	0.999991	0.999993	0.000000	0.000037	0.000037
46	0.999995	0.999997	0.000000	0.000022	0.000022
47	0.999999	0.999999	0.000000	0.000017	0.000017
48	1.000000	0.999999	0.000000	0.000009	0.000009

Table 4.4: Block-factor: $m_1 = m_2 = 3$, $\tilde{T}_1 = \tilde{T}_2 = 44$, $T_1 = T_2 = 42$, $\mathbf{p} = \mathbf{0.1}$, $ITER = 10^8$

For all of our results presented in the tables we used Monte Carlo simulations with 10^8 iterations for the block-factor model and with 10^5 replicas for the i.i.d. model. Notice that the contribution of the simulation error that corresponds to the approximation error (E_{sapp}) to the total error is almost negligible in most of the cases with respect to the other simulation error (E_{sf}). Thus, the precision of the method will depend mostly on the number of iterations ($ITER$) used to estimate Q_{t_1,t_2}. The cumulative distribution function and the probability mass function for the block-factor and i.i.d. models are presented in Figure 4.8 and Figure 4.9.

n	Sim	AppH	$E_{sapp}(2)$ Eq.(3.97)	$E_{sf}(2)$ Eq.(3.96)	Total Error
17	0.789376	0.788934	0.005813	0.011393	0.017206
18	0.925456	0.925186	0.000529	0.002095	0.002625
19	0.976889	0.976763	0.000045	0.000455	0.000500
20	0.993444	0.993447	0.000003	0.000105	0.000108
21	0.998288	0.998287	0.000000	0.000023	0.000024
22	0.999584	0.999583	0.000000	0.000005	0.000005
23	0.999905	0.999905	0.000000	0.000001	0.000001
24	0.999980	0.999980	0.000000	0.000000	0.000000

Table 4.5: Independent: $m_1 = m_2 = 3$, $T_1 = T_2 = 42$, $\mathcal{B}(\mathbf{r} = \mathbf{8}, \mathbf{p} = \mathbf{0.1})$, $ITER = 10^5$

n	Sim	AppH	$E_{sapp}(2)$ Eq.(4.20)	$E_{sf}(2)$ Eq.(4.23)	Total Error
48	0.768889	0.749275	0.046577	0.053831	0.100408
49	0.844717	0.829918	0.014207	0.019702	0.033908
50	0.899398	0.889501	0.004574	0.008810	0.013384
51	0.936771	0.930795	0.001499	0.004769	0.006269
52	0.961836	0.958113	0.000485	0.002988	0.003472
53	0.977672	0.975326	0.000152	0.002045	0.002197
54	0.987307	0.985922	0.000047	0.001463	0.001510
55	0.993022	0.992251	0.000014	0.001056	0.001070
56	0.996333	0.995917	0.000004	0.000761	0.000765
57	0.998151	0.997954	0.000001	0.000539	0.000540
58	0.999091	0.998992	0.000000	0.000381	0.000381
59	0.999576	0.999522	0.000000	0.000265	0.000265
60	0.999794	0.999802	0.000000	0.000178	0.000178
61	0.999908	0.999920	0.000000	0.000115	0.000115
62	0.999965	0.999973	0.000000	0.000077	0.000077
63	0.999993	0.999991	0.000000	0.000044	0.000044
64	0.999999	0.999998	0.000000	0.000028	0.000028
65	1.000000	0.999999	0.000000	0.000017	0.000017

Table 4.6: Block-factor: $m_1 = m_2 = 3$, $\tilde{T}_1 = \tilde{T}_2 = 44$, $T_1 = T_2 = 42$, $\mathbf{p} = \mathbf{0.3}$, $ITER = 10^8$

n	Sim	AppH	$E_{sapp}(2)$ Eq.(3.97)	$E_{sf}(2)$ Eq.(3.96)	Total Error
35	0.716804	0.716395	0.012836	0.021243	0.034079
36	0.867167	0.866643	0.001951	0.005093	0.007044
37	0.943946	0.944024	0.000285	0.001409	0.001694
38	0.978505	0.978400	0.000039	0.000419	0.000457
39	0.992274	0.992262	0.000005	0.000126	0.000131
40	0.997395	0.997399	0.000001	0.000037	0.000037
41	0.999176	0.999178	0.000000	0.000010	0.000010
42	0.999753	0.999754	0.000000	0.000003	0.000003
43	0.999931	0.999931	0.000000	0.000001	0.000001
44	0.999982	0.999982	0.000000	0.000000	0.000000
45	0.999995	0.999995	0.000000	0.000000	0.000000

Table 4.7: Independent: $m_1 = m_2 = 3$, $T_1 = T_2 = 42$, $\mathcal{B}(\mathbf{r} = \mathbf{8}, \mathbf{p} = \mathbf{0.3})$, $ITER = 10^5$

n	Sim	AppH	$E_{sapp}(2)$ Eq.(4.20)	$E_{sf}(2)$ Eq.(4.23)	Total Error
61	0.725109	0.701781	0.085110	0.093544	0.178654
62	0.828019	0.888902	0.004453	0.008665	0.013118
63	0.899560	0.888902	0.004453	0.008665	0.013118
64	0.945304	0.939436	0.001049	0.004054	0.005103
65	0.972203	0.969026	0.000235	0.002334	0.002569
66	0.986999	0.985439	0.000047	0.001460	0.001507
67	0.994506	0.993814	0.000008	0.000927	0.000935
68	0.997851	0.997605	0.000001	0.000572	0.000573
69	0.999326	0.999230	0.000000	0.000320	0.000320
70	0.999826	0.999786	0.000000	0.000171	0.000171
71	0.999968	0.999952	0.000000	0.000083	0.000083
72	1.000000	1.000000	0.000000	0.000000	0.000000

Table 4.8: Block-factor: $m_1 = m_2 = 3$, $\tilde{T}_1 = \tilde{T}_2 = 44$, $T_1 = T_2 = 42$, $\mathbf{p} = \mathbf{0.5}$, $ITER = 10^8$

n	Sim	AppH	$E_{sapp}(2)$ Eq.(3.97)	$E_{sf}(2)$ Eq.(3.96)	Total Error
50	0.741089	0.735210	0.010514	0.018002	0.028516
51	0.882209	0.880827	0.001499	0.004196	0.005695
52	0.952545	0.952389	0.000200	0.001098	0.001299
53	0.982842	0.982891	0.000024	0.000307	0.000331
54	0.994328	0.994337	0.000002	0.000084	0.000087
55	0.998282	0.998278	0.000000	0.000022	0.000022
56	0.999517	0.999518	0.000000	0.000005	0.000005
57	0.999876	0.999876	0.000000	0.000001	0.000001
58	0.999971	0.999971	0.000000	0.000000	0.000000
59	0.999994	0.999994	0.000000	0.000000	0.000000
60	0.999999	0.999999	0.000000	0.000000	0.000000

Table 4.9: Independent: $m_1 = m_2 = 3$, $T_1 = T_2 = 42$, $\mathcal{B}(\mathbf{r} = \mathbf{8}, \mathbf{p} = \mathbf{0.5})$, $ITER = 10^5$

n	Sim	AppH	$E_{sapp}(2)$ Eq.(4.20)	$E_{sf}(2)$ Eq.(4.23)	Total Error
70	0.729239	0.705944	0.074290	0.082392	0.156682
71	0.876484	0.864370	0.006976	0.011623	0.018600
72	1.000000	1.000000	0.000000	0.000000	0.000000

Table 4.10: Block-factor: $m_1 = m_2 = 3$, $\tilde{T}_1 = \tilde{T}_2 = 44$, $T_1 = T_2 = 42$, $\mathbf{p} = \mathbf{0.7}$, $ITER = 10^8$

n	Sim	AppH	$E_{sapp}(2)$ Eq.(3.97)	$E_{sf}(2)$ Eq.(3.96)	Total Error
62	0.620295	0.611819	0.030328	0.042319	0.072646
63	0.847421	0.846730	0.002591	0.005851	0.008442
64	0.952524	0.952588	0.000194	0.000978	0.001172
65	0.987854	0.987887	0.000011	0.000168	0.000179
66	0.997472	0.997460	0.000000	0.000026	0.000027
67	0.999568	0.999568	0.000000	0.000003	0.000003
68	0.999943	0.999943	0.000000	0.000000	0.000000
69	0.999994	0.999994	0.000000	0.000000	0.000000

Table 4.11: Independent: $m_1 = m_2 = 3$, $T_1 = T_2 = 42$, $\mathcal{B}(\mathbf{r} = \mathbf{8}, \mathbf{p} = \mathbf{0.7})$, $ITER = 10^5$

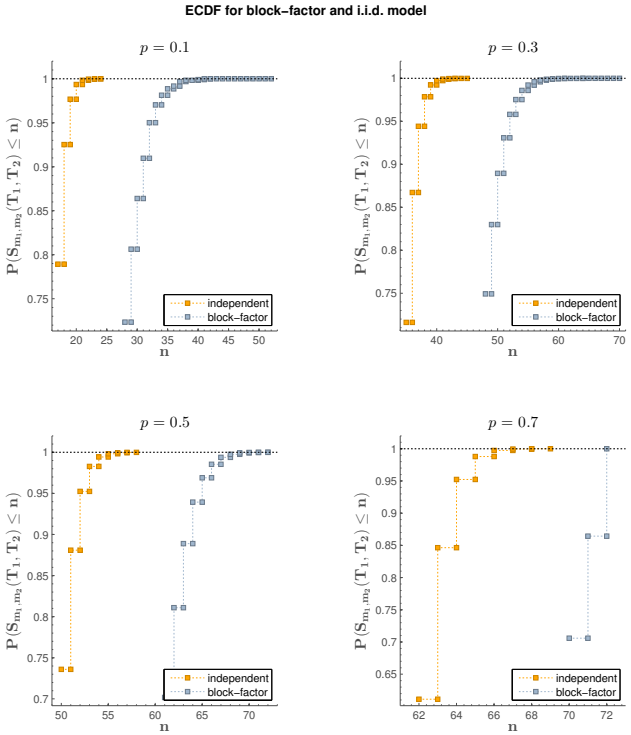


Figure 4.8: Cumulative distribution function for block–factor and i.i.d. models

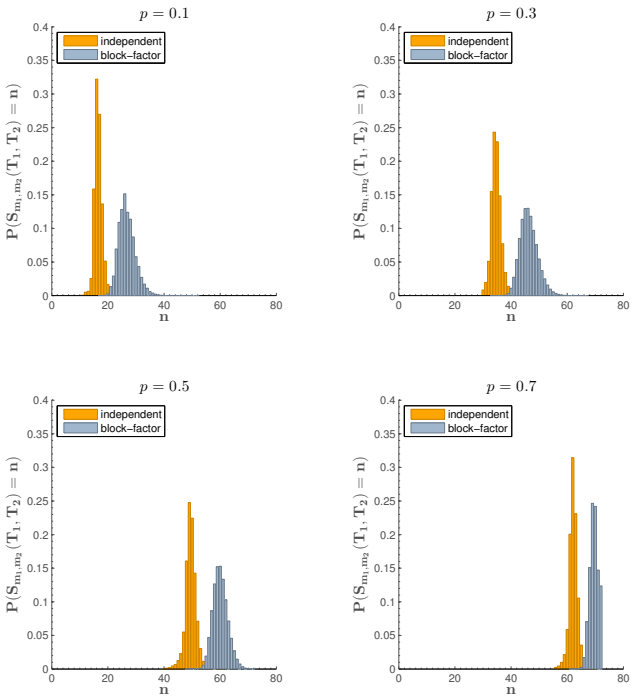


Figure 4.9: Probability mass function for blockâĹŠ-factor and i.i.d. models

Conclusions and perspectives

In this thesis, we provide a unified method for estimating the distribution of the multidimensional discrete scan statistics based on a result concerning the extremes of 1-dependent sequences of random variables. This approach has two main advantages over the existing ones presented in the literature: first, it includes, beside an approximation formula, expressions for the corresponding error bounds and second, the approximation can be applied no matter what the distribution of the observations under the null hypothesis is. We consider two models for the underlying distribution of the random field over which the scan process is performed: the widely studied i.i.d. model and a new dependent model based on a block-factor construction. For each of these models we give detailed expressions for the approximation, as well as for the associated error bounds formulas. Since the simulation process plays an important part in the estimation of the multidimensional discrete scan statistics distribution, we present, for the particular case of i.i.d. observations, a general importance sampling algorithm that increases the efficiency of the proposed approximation. From the numerical applications conducted in both, the i.i.d. and the block-factor models, we conclude that our results are accurate especially for the high order quantiles.

Currently, we are working on adapting the methodology developed for the multidimensional discrete scan statistics to the continuous scan statistics framework. One of the main challenge in this problem is to develop fast and efficient algorithms for the simulation of the continuous scan statistics, in more than three dimensions. Deriving accurate approximations for the distribution of the scan statistics over a random field generated by independent but not identically distributed random variables, constitutes a problem of future interest. Another future direction of research, is to consider, in the multidimensional discrete scan statistics problem, that the scanning window has a general convex shape and to study the influence of this shape in the scanning process. We should mention that for the multidimensional continuous scan statistics for Poisson processes, this problem was already addressed by [Alm, 1998].

Appendix A

Supplementary material for Chapter 1

A.1 Supplement for Section 1.1

In this section, we outline the algorithm of [Karwe and Naus, 1997], used for finding the distribution function of the one dimensional scan statistics over a sequence of i.i.d. discrete random variables of length at most $3m_1$. Based on the notations introduced in Section 1.1, we define

$$\begin{aligned} b^n_{2(m_1)}(y) &= \mathbb{P}\left(S_{m_1}(2m_1) \leq n, Y_{m_1+1} = y\right), \\ f(y) &= \mathbb{P}(X_1 = y), \\ Q^n_{2m_1} &= \mathbb{P}\left(S_{m_1}(2m_1) \leq n\right). \end{aligned}$$

We have the following recurrence relations for $Q_{m_1}(2m_1 - 1)$ and $Q_{m_1}(2m_1)$:

$$\begin{aligned} b^n_{2(1)}(y) &= \left(\sum_{j=0}^{n} f(j)\right) f(y), \\ b^n_{2(m_1)}(y) &= \sum_{\eta=0}^{y} \sum_{\nu=0}^{n-y+\eta} b^{n-\nu}_{2(m_1-1)}(y-\eta) f(\nu) f(\eta), \\ Q^n_{2m_1} &= \sum_{y=0}^{n} b^n_{2(m_1)}(y), \\ Q^n_{2m_1-1} &= \sum_{x=0}^{n} f(x) Q^{n-x}_{2(m_1-1)}. \end{aligned}$$

For the computation of $Q_{m_1}(3m_1 - 1)$ and $Q_{m_1}(3m_1)$, we take $b^{k_1,k_2}_{3(m_1)}(x, y) = 0$ if $x > k_1 \wedge k_2$ or $y > k_2$ and for the other cases as

$$b^{k_1,k_2}_{3(m_1)}(x,y) = \mathbb{P}\left(\bigcap_{i=1}^{m_1} \{Y_i \leq k_1\} \cap \{Y_{m_1+1} = x\} \bigcap_{j=m_1+2}^{2m_1} \{Y_j \leq k_2\} \cap \{Y_{2m_1+1} = y\}\right).$$

Considering that the random variables X_i take values in the set $\{0, \ldots, c\}$, we have the following recurrences (see [Karwe and Naus, 1997, Appendix] for closed form

formulas):

$$b^{k_1,k_2}_{3(m_1)}(x,y) = \sum_{\alpha=0}^{c}\sum_{\beta=0}^{c}\sum_{\gamma=0}^{c} b^{k_1-\alpha,k_2-\beta}_{3(m_1-1)}(x-\beta,y-\gamma)f(\alpha)f(\beta)f(\gamma),$$

$$Q^n_{3m_1} = \sum_{x=0}^{n}\sum_{y=0}^{n} b^{n,n}_{3(m_1)}(x,y),$$

$$Q^n_{3m_1-1} = \sum_{x_1=0}^{n}\sum_{x_2=0}^{n-x_1}\sum_{x_3=0}^{n-x_1}\left[\sum_{x_4=0}^{n-x_3} b^{n-x_2,n-x_3}_{3(m_1-1)}(x_1,x_4)\right] f(x_2)f(x_3).$$

A.2 Supplement for Section 1.2

Some supplementary materials for the results presented in Section 1.2 are given in this section.

A.2.1 Supplement for Section 1.2.1

In this section, we include the formulas for the unknown quantities that appear in the computation of the product-type approximation given in Eq.(1.64) in the case of binomial and Poisson observations. These formulas were presented in [Glaz et al., 2009, Chapter 5], for the case of $T_1 = T_2$ and $m_1 = m_2$.

The following notation, $W^{j_1,j_2}_{i_1,i_2} = \sum_{i=i_1}^{j_1}\sum_{j=i_2}^{j_2} X_{i,j}$, will be used throughout this section.

a) $X_{i,j}$ are i.i.d. binomial random variables with parameters r and p

We have

$$Q(m_1+1,m_2) = \sum_{y=0}^{n\wedge(m_1-1)m_2r} \mathbb{P}^2\left(W^{1,m_2}_{1,1} \le n-y\right)\mathbb{P}\left(W^{m_1,m_2}_{2,1} = y\right)$$

and

$$\begin{aligned} Q(m_1+1,m_2+1) = &\sum_{y_1=0}^{n\wedge(m_1-1)(m_2-1)r}\ \sum_{y_2=0}^{(n-y_1)\wedge(m_2-1)r}\ \sum_{y_3=0}^{(n-y_1)\wedge(m_2-1)r} \\ &\sum_{y_4=0}^{[n-y_1-y_2\vee y_3]\wedge(m_1-1)r}\ \sum_{y_5=0}^{[n-y_1-y_2\vee y_3]\wedge(m_1-1)r} \mathbb{P}\left(W^{1,1}_{1,1} \le a_1\right) \\ &\times \mathbb{P}\left(W^{1,m_2+1}_{1,m_2+1} \le a_2\right)\mathbb{P}\left(W^{m_1+1,1}_{m_1+1,1} \le a_3\right)\mathbb{P}\left(W^{m_1+1,m_2+1}_{m_1+1,m_2+1} \le a_4\right) \\ &\times \mathbb{P}\left(W^{m_1,m_2}_{2,2} = y_1\right)\mathbb{P}\left(W^{1,m_2}_{1,2} = y_2\right)\mathbb{P}\left(W^{m_1+1,m_2}_{m_1+1,2} = y_3\right) \\ &\times \mathbb{P}\left(W^{m_1,1}_{2,1} = y_4\right)\mathbb{P}\left(W^{m_1,m_2+1}_{2,m_2+1} = y_5\right), \end{aligned}$$

where

$$a_1 = n - y_1 - y_2 - y_4, \quad a_2 = n - y_1 - y_2 - y_5,$$
$$a_3 = n - y_1 - y_3 - y_4, \quad a_4 = n - y_1 - y_3 - y_5.$$

The random variables $W_{1,1}^{1,1}$, W_{1,m_2+1}^{1,m_2+1}, $W_{m_1+1,m_2+1}^{m_1+1,m_2+1}$ are binomial distributed with parameters r and p, $W_{2,2}^{m_1,m_2}$ is a binomial random variable with parameters $(m_1-1)(m_2-1)r$ and p, $W_{1,2}^{1,m_2}$ and $W_{m_1+1,2}^{m_1+1,m_2}$ are binomial random variable with parameters $(m_2-1)r$ and p and $W_{2,1}^{m_1,1}$ and $W_{2,m_2+1}^{m_1,m_2+1}$ are binomial random variable with parameters $(m_1-1)r$ and p.

b) $X_{i,j}$ are i.i.d. Poisson random variables of parameter λ

We have

$$Q(m_1+1, m_2) = \sum_{y=0}^{n} \mathbb{P}^2\left(W_{1,1}^{1,m_2} \leq n-y\right) \mathbb{P}\left(W_{2,1}^{m_1,m_2} = y\right)$$

and

$$\begin{aligned} Q(m_1+1, m_2+1) = &\sum_{y_1=0}^{n} \sum_{y_2=0}^{n-y_1} \sum_{y_3=0}^{n-y_1} \sum_{y_4=0}^{n-y_1-y_2\vee y_3} \sum_{y_5=0}^{n-y_1-y_2\vee y_3} \mathbb{P}\left(W_{1,1}^{1,1} \leq a_1\right) \\ &\times \mathbb{P}\left(W_{1,m_2+1}^{1,m_2+1} \leq a_2\right) \mathbb{P}\left(W_{m_1+1,1}^{m_1+1,1} \leq a_3\right) \mathbb{P}\left(W_{m_1+1,m_2+1}^{m_1+1,m_2+1} \leq a_4\right) \\ &\times \mathbb{P}\left(W_{2,2}^{m_1,m_2} = y_1\right) \mathbb{P}\left(W_{1,2}^{1,m_2} = y_2\right) \mathbb{P}\left(W_{m_1+1,2}^{m_1+1,m_2} = y_3\right) \\ &\times \mathbb{P}\left(W_{2,1}^{m_1,1} = y_4\right) \mathbb{P}\left(W_{2,m_2+1}^{m_1,m_2+1} = y_5\right), \end{aligned}$$

where

$$a_1 = n - y_1 - y_2 - y_4, \quad a_2 = n - y_1 - y_2 - y_5,$$
$$a_3 = n - y_1 - y_3 - y_4, \quad a_4 = n - y_1 - y_3 - y_5.$$

The random variables $W_{1,1}^{1,1}$, W_{1,m_2+1}^{1,m_2+1}, $W_{m_1+1,m_2+1}^{m_1+1,m_2+1}$ are Poisson distributed of parameter λ, $W_{2,2}^{m_1,m_2}$ is a Poisson random variable of mean $(m_1-1)(m_2-1)\lambda$, $W_{1,2}^{1,m_2}$ and $W_{m_1+1,2}^{m_1+1,m_2}$ are Poisson of mean $(m_2-1)\lambda$ and $W_{2,1}^{m_1,1}$ and $W_{2,m_2+1}^{m_1,m_2+1}$ are Poisson random variable of mean $(m_1-1)\lambda$.

A.2.2 Supplement for Section 1.2.2

Following the methodology presented in Section 1.1.3 for finding the upper bound of the one dimensional scan statistics, in this section we extend the results to the two dimensional case. Assume, for simplicity, that $T_1 = L_1 m_1$ and $T_2 = L_2 m_2$.

Applying [Kuai et al., 2000] inequality, we can write

$$\mathbb{P}\left(S_{m_1,m_2}\ (T_1,T_2)\leq n\right)=1-\mathbb{P}\left(\bigcup_{i_1=1}^{L_1-1}\bigcup_{i_2=1}^{L_2-1}E^c_{i_1,i_2}\right)$$

$$\leq 1-\sum_{i_1=1}^{L_1-1}\sum_{i_2=1}^{L_2-1}\left[\frac{\theta_{i_1,i_1}\mathbb{P}(E^c_{i_1,i_2})^2}{\sum_{j_1=1}^{L_1-1}\sum_{j_2=1}^{L_2-1}\mathbb{P}\left(E^c_{i_1,i_2}\cap E^c_{j_1,j_2}\right)+(1-\theta_{i_1,i_2})\mathbb{P}(E^c_{i_1,i_2})}\right.$$

$$\left.+\frac{(1-\theta_{i_1,i_2})\mathbb{P}(E^c_{i_1,i_2})^2}{\sum_{j_1=1}^{L_1-1}\sum_{j_2=1}^{L_2-1}\mathbb{P}\left(E^c_{i_1,i_2}\cap E^c_{j_1,j_2}\right)-\theta_{i_1,i_2}\mathbb{P}(E^c_{i_1,i_2})}\right]=:UB,$$

where for $1\leq i_1\leq L_1-1$ and $1\leq i_2\leq L_2-1$, the events E_{i_1,i_2} are defined by

$$E_{i_1,i_2}=\left\{\max_{\substack{(i_1-1)m_1+1\leq s_1\leq i_1m_1+1\\(i_2-1)m_2+1\leq s_2\leq i_2m_2+1}}Y_{s_1,s_2}\leq n\right\}$$

and where

$$\theta_{i_1,i_2}=\frac{\sum_{j_1=1}^{L_1-1}\sum_{j_2=1}^{L_2-1}\mathbb{P}\left(E^c_{i_1,i_2}\cap E^c_{j_1,j_2}\right)}{\mathbb{P}(E^c_{i_1,i_2})}-\left\lfloor\frac{\sum_{j_1=1}^{L_1-1}\sum_{j_2=1}^{L_2-1}\mathbb{P}\left(E^c_{i_1,i_2}\cap E^c_{j_1,j_2}\right)}{\mathbb{P}(E^c_{i_1,i_2})}\right\rfloor.$$

To simplify the results, we adopt the notations $M=(L_1-1)(L_2-1)$ and

$$\Sigma_{i_1,i_2}=\sum_{j_1=1}^{L_1-1}\sum_{j_2=1}^{L_2-1}\mathbb{P}\left(E^c_{i_1,i_2}\cap E^c_{j_1,j_2}\right).$$

We observe, from the definition of E_{i_1,i_2}, that if $|i_1-j_1|\geq 2$ or $|i_2-j_2|\geq 2$, then E_{i_1,i_2} and E_{j_1,j_2} are independent and, since the events are also stationary, we have

$$\mathbb{P}(E^c_{i_1,i_2})=1-\mathbb{P}(E_{i_1,i_2})=1-Q(2m_1,2m_2),$$

and

$$\begin{aligned}\mathbb{P}(E^c_{i_1,i_2}\cap E^c_{j_1,j_2})&=1-\mathbb{P}(E_{i_1,i_2})-\mathbb{P}(E_{i_1,i_2})+\mathbb{P}(E_{i_1,i_2}\cap E_{j_1,j_2})\\&=1-2Q(2m_1,2m_2)+\mathbb{P}(E_{i_1,i_2}\cap E_{j_1,j_2}).\end{aligned}$$

Simple calculations lead to

$$\Sigma_{i_1,i_2} = \begin{cases} 4 - 7Q(2m_1,2m_2) + Q(2m_1,3m_2) + Q(3m_1,2m_2) + Q(3m_1,3m_2) \\ +(M-4)\left[1-Q(2m_1,2m_2)\right]^2, \text{ if } i_1 \in \{1, L_1-1\},\ i_2 \in \{1, L_2-1\} \\ \\ 6 - 11Q(2m_1,2m_2) + Q(2m_1,3m_2) + 2Q(3m_1,2m_2) + 2Q(3m_1,3m_2) \\ +(M-6)\left[1-Q(2m_1,2m_2)\right]^2, \text{ if } 2 \le i_1 \le L_1-2,\ i_2 \in \{1, L_2-1\} \\ \\ 6 - 11Q(2m_1,2m_2) + 2Q(2m_1,3m_2) + Q(3m_1,2m_2) + 2Q(3m_1,3m_2) \\ +(M-6)\left[1-Q(2m_1,2m_2)\right]^2, \text{ if } i_1 \in \{1, L_1-1\},\ 2 \le i_2 \le L_2-2 \\ \\ 9 - 15Q(2m_1,2m_2) + 2\left[Q(2m_1,3m_2) + Q(3m_1,2m_2) + Q(3m_1,3m_2)\right] \\ +(M-9)\left[1-Q(2m_1,2m_2)\right]^2, \text{ if } 2 \le i_1 \le L_1-2,\ 2 \le i_2 \le L_2-2. \end{cases}$$

Hence, the upper bound becomes

$$UB = 1 - 4B_1 - 2(L_1-3)B_2 - 2(L_2-3)B_3 - (L_1-3)(L_2-3)B_4,$$

where for $i \in \{1,2,3,4\}$,

$$B_i = \frac{\theta_i\left[1-Q(2m_1,2m_2)\right]^2}{\Sigma_i + (1-\theta_i)\left[1-Q(2m_1,2m_2)\right]} + \frac{(1-\theta_i)\left[1-Q(2m_1,2m_2)\right]^2}{\Sigma_i - \theta_i\left[1-Q(2m_1,2m_2)\right]},$$
$$\theta_i = \frac{\Sigma_i}{1-Q(2m_1,2m_2)} - \left\lfloor \frac{\Sigma_i}{1-Q(2m_1,2m_2)} \right\rfloor,$$

and Σ_1, Σ_2, Σ_3 and Σ_4 correspond to the first, second, third and fourth branch, respectively of Σ_{i_1,i_2} given above. The unknown quantities $Q(2m_1,2m_2)$, $Q(2m_1,3m_2)$, $Q(3m_1,2m_2)$ and $Q(3m_1,3m_2)$ are evaluated by simulation.

A.3 Supplement for Section 1.3

Using the same approach as in [Guerriero et al., 2010a] for the case of $T_1 = T_2 = T_3$, $m_1 = m_2 = m_3$ and where $X_{i,j,k}$ were Bernoulli random variables, we present computational expressions for the unknown quantities involved in the approximation formula given in Eq.(1.87). We consider detailed formulas only for the situation of binomial observations, the Poisson case being similar (see Remark A.3.1).
Throughout this section, we will use the following shorthand notation:

$$W_{i_1,i_2,i_3}^{j_1,j_2,j_3} = \sum_{i=i_1}^{j_1}\sum_{j=i_2}^{j_2}\sum_{k=i_3}^{j_3} X_{i,j,k}.$$

Assume that $X_{i,j,k}$ are i.i.d. binomial random variables with parameters r and p.

- $\underline{Q(m_1,m_2,m_3)}$

It is clear that

$$Q(m_1,m_2,m_3) = \mathbb{P}\left(W_{1,1,1}^{m_1,m_2,m_3} \leq n\right) = \sum_{i=0}^{n\wedge m_1m_2m_3r} \binom{m_1m_2m_3r}{i} p^i(1-p)^{m_1m_2m_3r-i}.$$

- $\underline{Q(m_1+1,m_2,m_3),\, Q(m_1,m_2+1,m_3),\, Q(m_1,m_2,m_3+1)}$

We have

$$Q(m_1+1,m_2,m_3) = \sum_{x_1=0}^{n\wedge(m_1-1)m_2m_3r} \mathbb{P}^2\left(W_{1,1,1}^{1,m_2,m_3} \leq n-x_1\right)\mathbb{P}\left(W_{2,1,1}^{m_1,m_2,m_3} = x_1\right),$$

$$Q(m_1,m_2+1,m_3) = \sum_{x_1=0}^{n\wedge m_1(m_2-1)m_3r} \mathbb{P}^2\left(W_{1,1,1}^{m_1,1,m_3} \leq n-x_1\right)\mathbb{P}\left(W_{1,2,1}^{m_1,m_2,m_3} = x_1\right),$$

$$Q(m_1,m_2,m_3+1) = \sum_{x_1=0}^{n\wedge m_1m_2(m_3-1)r} \mathbb{P}^2\left(W_{1,1,1}^{1,m_2,m_3} \leq n-x_1\right)\mathbb{P}\left(W_{1,1,2}^{m_1,m_2,m_3} = x_1\right).$$

- $\underline{Q(m_1,m_2+1,m_3+1)}$

We can write

$$Q(m_1,m_2+1,m_3+1) \approx \sum_{x_1=0}^{y_1}\sum_{x_2=0}^{y_2}\sum_{x_3=0}^{y_3}\sum_{x_4=0}^{y_4}\sum_{x_5=0}^{y_5}\left(\prod_{i=1}^{4}a_i\right)\left(\prod_{j=1}^{5}d_j(x_j)\right),$$

where

$$\begin{aligned}
y_1 &= n\wedge m_1(m_2-1)(m_3-1)r,\\
y_2 &= (n-x_1)\wedge m_1(m_2-1)r,\\
y_3 &= (n-x_1)\wedge m_1(m_2-1)r,\\
y_4 &= [n-x_1-(x_2\vee x_3)]\wedge m_1(m_3-1)r,\\
y_5 &= [n-x_1-(x_2\vee x_3)]\wedge m_1(m_3-1)r
\end{aligned}$$

and

$$\begin{aligned}
a_1 &= \mathbb{P}\left(W_{1,1,1}^{m_1,1,1} \leq n-x_1-x_2-x_4\right),\\
a_2 &= \mathbb{P}\left(W_{1,m_2+1,1}^{m_1,m_2+1,1} \leq n-x_1-x_3-x_4\right),\\
a_3 &= \mathbb{P}\left(W_{1,1,m_3+1}^{m_1,1,m_3+1} \leq n-x_1-x_2-x_5\right),\\
a_4 &= \mathbb{P}\left(W_{1,m_2+1,m_3+1}^{m_1,m_2+1,m_3+1} \leq n-x_1-x_3-x_5\right)
\end{aligned}$$

and

$$\begin{aligned}
d_1(x_1) &= \mathbb{P}\left(W_{1,2,2}^{m_1,m_2,m_3} = x_1\right),\\
d_2(x_2) &= \mathbb{P}\left(W_{1,2,1}^{m_1,m_2,1} = x_2\right),\\
d_3(x_3) &= \mathbb{P}\left(W_{1,2,m_3+1}^{m_1,m_2,m_3+1} = x_3\right),\\
d_4(x_4) &= \mathbb{P}\left(W_{1,1,2}^{m_1,1,m_3} = x_4\right),\\
d_5(x_5) &= \mathbb{P}\left(W_{1,m_2+1,2}^{m_1,m_2+1,m_3} = x_5\right).
\end{aligned}$$

- $\underline{Q(m_1+1, m_2, m_3+1)}$

Similarly, we can write

$$Q(m_1+1, m_2, m_3+1) \approx \sum_{x_1=0}^{y_1}\sum_{x_2=0}^{y_2}\sum_{x_3=0}^{y_3}\sum_{x_4=0}^{y_4}\sum_{x_5=0}^{y_5}\left(\prod_{i=1}^{4} a_i\right)\left(\prod_{j=1}^{5} d_j(x_j)\right),$$

where

$$\begin{aligned}
y_1 &= n \wedge (m_1-1)m_2(m_3-1)r,\\
y_2 &= (n-x_1) \wedge (m_1-1)m_2 r,\\
y_3 &= (n-x_1) \wedge m_1(m_2-1)r,\\
y_4 &= [n-x_1-(x_2 \vee x_3)] \wedge m_2(m_3-1)r,\\
y_5 &= [n-x_1-(x_2 \vee x_3)] \wedge m_2(m_3-1)r
\end{aligned}$$

and

$$\begin{aligned}
a_1 &= \mathbb{P}\left(W_{1,1,1}^{1,m_2,1} \le n-x_1-x_2-x_4\right),\\
a_2 &= \mathbb{P}\left(W_{1,1,m_3+1}^{1,m_2,m_3+1} \le n-x_1-x_3-x_4\right),\\
a_3 &= \mathbb{P}\left(W_{m_1+1,1,1}^{m_1+1,m_2,1} \le n-x_1-x_2-x_5\right),\\
a_4 &= \mathbb{P}\left(W_{m_1+1,1,m_3+1}^{m_1+1,m_2,m_3+1} \le n-x_1-x_3-x_5\right)
\end{aligned}$$

and

$$\begin{aligned}
d_1(x_1) &= \mathbb{P}\left(W_{2,1,2}^{m_1,m_2,m_3} = x_1\right),\\
d_2(x_2) &= \mathbb{P}\left(W_{2,1,1}^{m_1,m_2,1} = x_2\right),\\
d_3(x_3) &= \mathbb{P}\left(W_{2,1,m_3+1}^{m_1,m_2,m_3+1} = x_3\right),\\
d_4(x_4) &= \mathbb{P}\left(W_{1,1,2}^{1,m_2,m_3} = x_4\right),\\
d_5(x_5) &= \mathbb{P}\left(W_{m_1+1,1,2}^{m_1+1,m_2,m_3} = x_5\right).
\end{aligned}$$

- $\underline{Q(m_1+1, m_2+1, m_3)}$

As before,

$$Q(m_1+1, m_2+1, m_3) \approx \sum_{x_1=0}^{y_1}\sum_{x_2=0}^{y_2}\sum_{x_3=0}^{y_3}\sum_{x_4=0}^{y_4}\sum_{x_5=0}^{y_5}\left(\prod_{i=1}^{4} a_i\right)\left(\prod_{j=1}^{5} d_j(x_j)\right),$$

where

$$\begin{aligned}
y_1 &= n \wedge (m_1-1)(m_2-1)m_3 r,\\
y_2 &= (n-x_1) \wedge (m_1-1)m_3 r,\\
y_3 &= (n-x_1) \wedge (m_1-1)m_3 r,\\
y_4 &= [n-x_1-(x_2 \vee x_3)] \wedge (m_2-1)m_3 r,\\
y_5 &= [n-x_1-(x_2 \vee x_3)] \wedge (m_2-1)m_3 r
\end{aligned}$$

and

$$\begin{aligned}
a_1 &= \mathbb{P}\left(W_{1,1,1}^{1,1,m_3} \leq n-x_1-x_2-x_4\right),\\
a_2 &= \mathbb{P}\left(W_{1,m_2+1,1}^{1,m_2+1,m_3} \leq n-x_1-x_3-x_4\right),\\
a_3 &= \mathbb{P}\left(W_{m_1+1,1,1}^{m_1+1,1,m_3} \leq n-x_1-x_2-x_5\right),\\
a_4 &= \mathbb{P}\left(W_{m_1+1,m_2+1,1}^{m_1+1,m_2+1,m_3} \leq n-x_1-x_3-x_5\right)
\end{aligned}$$

and

$$\begin{aligned}
d_1(x_1) &= \mathbb{P}\left(W_{2,2,1}^{m_1,m_2,m_3} = x_1\right),\\
d_2(x_2) &= \mathbb{P}\left(W_{2,1,1}^{m_1,1,m_3} = x_2\right),\\
d_3(x_3) &= \mathbb{P}\left(W_{2,m_2+1,1}^{m_1,m_2+1,m_3} = x_3\right),\\
d_4(x_4) &= \mathbb{P}\left(W_{1,2,1}^{1,m_2,m_3} = x_4\right),\\
d_5(x_5) &= \mathbb{P}\left(W_{m_1+1,2,1}^{m_1+1,m_2,m_3} = x_5\right).
\end{aligned}$$

- $\underline{Q(m_1+1, m_2+1, m_3+1)}$

We have

$$\begin{aligned}
Q(m_1+1, m_2+1, m_3+1) \approx &\sum_{x_1=0}^{y_1}\sum_{x_2=0}^{y_2}\sum_{x_3=0}^{y_3}\sum_{x_4=0}^{y_4}\sum_{x_5=0}^{y_5}\sum_{x_6=0}^{y_6}\sum_{x_7=0}^{y_7}\sum_{x_8=0}^{y_8}\sum_{x_9=0}^{y_9} A_1 A_2\\
&\times\left(\prod_{j=1}^{9} d_j(x_j)\right),
\end{aligned}$$

with

$$\begin{aligned}
y_1 &= n \wedge (m_1-1)(m_2-1)(m_3-1)r,\\
y_2 &= (n-x_1) \wedge (m_1-1)(m_2-1)r,\\
y_3 &= (n-x_1) \wedge (m_1-1)(m_2-1)r,\\
y_4 &= [n-x_1-(x_2 \vee x_3)] \wedge (m_1-1)(m_3-1)r,\\
y_5 &= [n-x_1-(x_2 \vee x_3)] \wedge (m_1-1)(m_3-1)r,\\
y_6 &= (n-x_1-x_2-x_4) \wedge (m_1-1)r,\\
y_7 &= (n-x_1-x_2-x_5) \wedge (m_1-1)r,\\
y_8 &= (n-x_1-x_3-x_4) \wedge (m_1-1)r,\\
y_9 &= (n-x_1-x_3-x_5) \wedge (m_1-1)r,
\end{aligned}$$

and

$$\begin{aligned}
d_1(x_1) &= \mathbb{P}\left(W_{2,2,2}^{m_1,m_2,m_3} = x_1\right),\\
d_2(x_2) &= \mathbb{P}\left(W_{2,2,1}^{m_1,m_2,1} = x_2\right),\\
d_3(x_3) &= \mathbb{P}\left(W_{2,2,m_3+1}^{m_1,m_2,m_3+1} = x_3\right),\\
d_4(x_4) &= \mathbb{P}\left(W_{2,1,2}^{m_1,1,m_3} = x_4\right),\\
d_5(x_5) &= \mathbb{P}\left(W_{2,m_2+1,2}^{m_1,m_2+1,m_3} = x_5\right),\\
d_6(x_6) &= \mathbb{P}\left(W_{2,1,1}^{m_1,1,1} = x_6\right),\\
d_7(x_7) &= \mathbb{P}\left(W_{2,m_2+1,1}^{m_1,m_2+1,1} = x_7\right),\\
d_8(x_8) &= \mathbb{P}\left(W_{2,1,m_3+1}^{m_1,1,m_3+1} = x_8\right),\\
d_9(x_9) &= \mathbb{P}\left(W_{2,m_2+1,m_3+1}^{m_1,m_2+1,m_3+1} = x_9\right).
\end{aligned}$$

The unknown A_1 is computed by

$$A_1 = \sum_{x_{10}=0}^{y_{10}} \sum_{x_{11}=0}^{y_{11}} \sum_{x_{12}=0}^{y_{12}} \sum_{x_{13}=0}^{y_{13}} \sum_{x_{14}=0}^{y_{14}} \left(\prod_{i=1}^{4} a_i\right) \left(\prod_{j=10}^{14} d_j(x_j)\right),$$

with

$$\begin{aligned}
y_{10} &= (n-x_1-x_2-x_4-x_6) \wedge (m_2-1)(m_3-1)r,\\
y_{11} &= [n-x_1-x_4-x_{10}-(x_2+x_6) \vee (x_3+x_8)] \wedge (m_3-1)r,\\
y_{12} &= [n-x_1-x_5-x_{10}-(x_2+x_7) \vee (x_3+x_9)] \wedge (m_3-1)r,\\
y_{13} &= [n-x_1-x_2-x_{10}-(x_4+x_6) \vee (x_5+x_7)] \wedge (m_2-1)r,\\
y_{14} &= [n-x_1-x_3-x_{10}-(x_4+x_8) \vee (x_5+x_9)] \wedge (m_2-1)r,
\end{aligned}$$

and where

$$
\begin{aligned}
a_1 &= \mathbb{P}\left(W_{1,1,1}^{1,1,1} \leq n - x_1 - x_2 - x_4 - x_6 - x_{10} - x_{11} - x_{13}\right), \\
a_2 &= \mathbb{P}\left(W_{1,1,m_3+1}^{1,1,m_3+1} \leq n - x_1 - x_3 - x_4 - x_8 - x_{10} - x_{11} - x_{14}\right), \\
a_3 &= \mathbb{P}\left(W_{1,m_2+1,1}^{1,m_2+1,1} \leq n - x_1 - x_2 - x_5 - x_7 - x_{10} - x_{12} - x_{13}\right), \\
a_4 &= \mathbb{P}\left(W_{1,m_2+1,m_3+1}^{1,m_2+1,m_3+1} \leq n - x_1 - x_3 - x_5 - x_9 - x_{10} - x_{12} - x_{14}\right)
\end{aligned}
$$

and

$$
\begin{aligned}
d_{10}(x_{10}) &= \mathbb{P}\left(W_{1,2,2}^{1,m_2,m_3} = x_{10}\right), \\
d_{11}(x_{11}) &= \mathbb{P}\left(W_{1,1,2}^{1,1,m_3} = x_{11}\right), \\
d_{12}(x_{12}) &= \mathbb{P}\left(W_{1,m_2+1,2}^{1,m_2+1,m_3} = x_{12}\right), \\
d_{13}(x_{13}) &= \mathbb{P}\left(W_{1,2,1}^{1,m_2,1} = x_{13}\right), \\
d_{14}(x_{14}) &= \mathbb{P}\left(W_{1,2,m_3+1}^{1,m_2,m_3+1} = x_{14}\right).
\end{aligned}
$$

Similarly, to compute A_2, we use

$$
A_2 = \sum_{x_{15}=0}^{y_{15}} \sum_{x_{16}=0}^{y_{16}} \sum_{x_{17}=0}^{y_{17}} \sum_{x_{18}=0}^{y_{18}} \sum_{x_{19}=0}^{y_{19}} \left(\prod_{i=5}^{8} a_i\right) \left(\prod_{j=15}^{19} d_j(x_j)\right),
$$

with

$$
\begin{aligned}
y_{15} &= (n - x_1 - x_2 - x_4 - x_6) \wedge (m_2 - 1)(m_3 - 1)r, \\
y_{16} &= [n - x_1 - x_4 - x_{15} - (x_2 + x_6) \vee (x_3 + x_8)] \wedge (m_3 - 1)r, \\
y_{17} &= [n - x_1 - x_5 - x_{15} - (x_2 + x_7) \vee (x_3 + x_9)] \wedge (m_3 - 1)r, \\
y_{18} &= [n - x_1 - x_2 - x_{15} - (x_4 + x_6) \vee (x_5 + x_7)] \wedge (m_2 - 1)r, \\
y_{19} &= [n - x_1 - x_3 - x_{15} - (x_4 + x_8) \vee (x_5 + x_9)] \wedge (m_2 - 1)r
\end{aligned}
$$

and where

$$
\begin{aligned}
a_5 &= \mathbb{P}\left(W_{m_1+1,1,1}^{m_1+1,1,1} \leq n - x_1 - x_2 - x_4 - x_6 - x_{15} - x_{16} - x_{18}\right), \\
a_6 &= \mathbb{P}\left(W_{m_1+1,1,m_3+1}^{m_1+1,1,m_3+1} \leq n - x_1 - x_3 - x_4 - x_8 - x_{15} - x_{16} - x_{19}\right), \\
a_7 &= \mathbb{P}\left(W_{m_1+1,m_2+1,1}^{m_1+1,m_2+1,1} \leq n - x_1 - x_2 - x_5 - x_7 - x_{15} - x_{17} - x_{18}\right), \\
a_8 &= \mathbb{P}\left(W_{m_1+1,m_2+1,m_3+1}^{m_1+1,m_2+1,m_3+1} \leq n - x_1 - x_3 - x_5 - x_9 - x_{15} - x_{17} - x_{19}\right),
\end{aligned}
$$

and

$$
\begin{aligned}
d_{15}(x_{15}) &= \mathbb{P}\left(W_{m_1+1,2,2}^{m_1+1,m_2,m_3} = x_{15}\right), \\
d_{16}(x_{16}) &= \mathbb{P}\left(W_{m_1+1,1,2}^{m_1+1,1,m_3} = x_{16}\right), \\
d_{17}(x_{17}) &= \mathbb{P}\left(W_{m_1+1,m_2+1,2}^{m_1+1,m_2+1,m_3} = x_{17}\right), \\
d_{18}(x_{18}) &= \mathbb{P}\left(W_{m_1+1,2,1}^{m_1+1,m_2,1} = x_{18}\right), \\
d_{19}(x_{19}) &= \mathbb{P}\left(W_{m_1+1,2,m_3+1}^{m_1+1,m_2,m_3+1} = x_{19}\right).
\end{aligned}
$$

Remark A.3.1. *The above expressions remain valid in the i.i.d. Poisson model, the only difference is that the upper bounds that appear in the sums (the y_j's) do not contain the minimum part. For example, consider the case of $Q(m_1, m_2+1, m_3+1)$. The upper bounds y_1 to y_5 that appear in the formula, now become*

$$
\begin{aligned}
y_1 &= n, \\
y_2 &= n - x_1, \\
y_3 &= n - x_1, \\
y_4 &= n - x_1 - (x_2 \vee x_3), \\
y_5 &= n - x_1 - (x_2 \vee x_3).
\end{aligned}
$$

Appendix B

Supplementary material for Chapter 3

B.1 Proof of Lemma 3.3.1

From the mean value theorem in two dimensions we have

$$H(x_1,y_1)-H(x_2,y_2)=\frac{\partial H(x^*,y^*)}{\partial x}(x_1-x_2)+\frac{\partial H(x^*,y^*)}{\partial y}(y_1-y_2), \tag{B.1}$$

where (x^*,y^*) is a point on the segment $(x_1,y_1)-(x_2,y_2)$. Notice that

$$\frac{\partial H(x,y)}{\partial x}=\frac{2\left[1+x-y+2(x-y)^2\right]-(m-1)(2x-y)\left[1+4(x-y)\right]}{\left[1+x-y+2(x-y)^2\right]^m}, \tag{B.2}$$

$$\frac{\partial H(x,y)}{\partial y}=\frac{-\left[1+x-y+2(x-y)^2\right]+(m-1)(2x-y)\left[1+4(x-y)\right]}{\left[1+x-y+2(x-y)^2\right]^m} \tag{B.3}$$

and that when $y_i\leq x_i$ for $i\in\{1,2\}$, we have $y^*\leq x^*$ (since both points, (x_1,y_1) and (x_2,y_2), are lying between ($0x$ and the first bisector the same is true for any point on the segment determined by them). Applying Bernoulli inequality to the denominator in Eqs.(B.2) and (B.3)

$$\left[1+x-y+2(x-y)^2\right]^m\geq 1+(m-1)\left[x-y+2(x-y)^2\right], \tag{B.4}$$

we see, by elementary calculations (study of the sign of the equation of degree 2), that

$$\begin{aligned}\left|\frac{\partial H(x,y)}{\partial x}\right| &\leq \frac{4(x-y)^2(m-2)+(x-y)\left[4x(m-1)+m-3\right]+(mx-x-2)}{1+m\left[x-y+2(x-y)^2\right]}\\ &\leq \begin{cases} m-1, \text{ for } 3\leq m\leq 5\\ m-2, \text{ for } m\geq 6.\end{cases}\end{aligned} \tag{B.5}$$

In the same way,

$$\begin{aligned}\left|\frac{\partial H(x,y)}{\partial y}\right| &\leq \frac{2(x-y)^2(2m-3)+(x-y)\left[4x(m-1)+m-2\right]+(mx-x-1)}{1+m\left[x-y+2(x-y)^2\right]}\\ &\leq \begin{cases} m-1, \text{ for } 3\leq m\leq 5\\ m-2, \text{ for } m\geq 6.\end{cases}\end{aligned} \tag{B.6}$$

and in combination with Eq.(B.5), we obtain the requested result. ■

B.2 Proof of Lemma 3.4.3

Obviously, the random variables $Y_{i_1,\dots,i_d}$ follow a multivariate normal distribution since are equal, by definition, with a sum of i.i.d normals ($X_{s_1,\dots,s_d} \sim \mathcal{N}(\mu,\sigma^2)$, $1 \leq s_j \leq T_j$, $j \in \{1,\dots,d\}$). Thus, the mean is, clearly, equal with $\bar{\mu} = m_1 \cdots m_d \mu$. We show that the covariance matrix is given by the formula

$$Cov\left[Y_{i_1,\dots,i_d}, Y_{j_1,\dots,j_d}\right] = \begin{cases} (m_1 - |i_1 - j_1|)\cdots(m_d - |i_d - j_d|)\,\sigma^2 & , |i_s - j_s| < m_s \\ & s \in \{1,\dots,d\}, \\ 0 & , \text{otherwise.} \end{cases} \tag{B.7}$$

Take $u_q = |i_q - j_q|$, for each $q \in \{1,\dots,d\}$, and observe that if there is an index $r \in \{1,\dots,d\}$ such that $u_r \geq m_r$, then the random variables $Y_{i_1,\dots,i_d}$ and $Y_{j_1,\dots,j_d}$ are independent (they do not share any random variable $X_{s_1,\dots,s_d}$), so the covariance is zero.

Assume now that $u_q \leq m_q - 1$, for all $q \in \{1,\dots,d\}$. We observe that the two random variables, $Y_{i_1,\dots,i_d}$ and $Y_{j_1,\dots,j_d}$, can be rewritten as

$$Y_{i_1,\dots,i_d} = Z^{j_1,\dots,j_d}_{i_1,\dots,i_d} + \overline{X}, \tag{B.8}$$

$$Y_{j_1,\dots,j_d} = Z^{j_1,\dots,j_d}_{i_1,\dots,i_d} + \overline{Y}, \tag{B.9}$$

where

$$Z^{j_1,\dots,j_d}_{i_1,\dots,i_d} = \sum_{s_1 = i_1 \wedge j_1 + u_1}^{i_1 \wedge j_1 + m_1 - 1} \cdots \sum_{s_d = i_d \wedge j_d + u_d}^{i_d \wedge j_d + m_d - 1} X_{s_1,\dots,s_d} \tag{B.10}$$

and $\overline{X} = Y_{i_1,\dots,i_d} - Z^{j_1,\dots,j_d}_{i_1,\dots,i_d}$ and $\overline{Y} = Y_{j_1,\dots,j_d} - Z^{j_1,\dots,j_d}_{i_1,\dots,i_d}$, respectively.

Since $X_{s_1,\dots,s_d} \sim \mathcal{N}(\mu,\sigma^2)$ are independent, $1 \leq s_j \leq T_j$, $j \in \{1,\dots,d\}$, the random variables $Z^{j_1,\dots,j_d}_{i_1,\dots,i_d}$, $\overline{X}$ and $\overline{Y}$ are pairwise independent. The result now follows from the following simple property:

Fact B.2.1. *Assume that W_1 and W_2 are two random variables such that $W_1 = X+Y$, $W_2 = X+Z$, where X, Y and Z are pairwise independent random variables. Then, the covariance between W_1 and W_2 satisfies*

$$Cov\left[W_1, W_2\right] = Var\left[X\right]. \tag{B.11}$$

From the foregoing relation and since the variance of $Z^{j_1,\dots,j_d}_{i_1,\dots,i_d}$ is given by

$$Var\left[Z^{j_1,\dots,j_d}_{i_1,\dots,i_d}\right] = (m_1 - u_1)\cdots(m_d - u_d)\sigma^2, \tag{B.12}$$

we conclude that if $u_q \leq m_q - 1$ for all $q \in \{1,\dots,d\}$, then

$$Cov\left[Y_{i_1,\dots,i_d}, Y_{j_1,\dots,j_d}\right] = (m_1 - |i_1 - j_1|)\cdots(m_d - |i_d - j_d|)\,\sigma^2. \ \blacksquare \tag{B.13}$$

B.3 Validity of the algorithm in Example 3.4.2

To verify the validity of the algorithm proposed by [Hoffman and Ribak, 1991], reduces to show that $\overline{\mathbf{W}}_i \sim \mathcal{N}(\mu_{w_i|t}, \Sigma_{w_i|t})$, where

$$\mu_{w_i|t} = \mathbb{E}[\mathbf{W}_i] + \frac{1}{Var[Z_l]} Cov[\mathbf{W}_i, Z_l](t - \mathbb{E}[Z_l]), \tag{B.14}$$

$$\Sigma_{w_i|t} = Cov(\mathbf{W}_i) - \frac{1}{Var[Z_l]} Cov[\mathbf{W}_i, Z_l] Cov^T[\mathbf{W}_i, Z_l]. \tag{B.15}$$

Clearly, from the definition, the random vector $\overline{\mathbf{W}}_i$ follow a multivariate normal distribution and satisfies the relation

$$\overline{\mathbf{W}}_i = \mathbf{W}_i + \mu_{w_i|t} - \mathbb{E}\left[\mathbf{W}_i | Z_l\right]. \tag{B.16}$$

By conditioning with Z_l in Eq.(B.16), we have

$$\mathbb{E}\left[\overline{\mathbf{W}}_i | Z_l\right] = \mathbb{E}\left[\mathbf{W}_i | Z_l\right] + \mu_{w_i|t} - \mathbb{E}\left[\mathbf{W}_i | Z_l\right] = \mu_{w_i|t} \tag{B.17}$$

hence, by taking the mean,

$$\mathbb{E}\left[\overline{\mathbf{W}}_i\right] = \mathbb{E}\left[\mathbb{E}\left[\overline{\mathbf{W}}_i | Z_l\right]\right] = \mu_{w_i|t}. \tag{B.18}$$

For the covariance matrix, we notice that

$$\begin{aligned} Cov\left[\overline{\mathbf{W}}_i | Z_l\right] &= Cov\left[\mathbf{W}_i + \mu_{w_i|t} - \mathbb{E}\left[\mathbf{W}_i | Z_l\right] | Z_l\right] \\ &= Cov\left[\mathbf{W}_i - \mathbb{E}\left[\mathbf{W}_i | Z_l\right] | Z_l\right] \end{aligned} \tag{B.19}$$

and considering the (r, q) element, we have

$$\begin{aligned} Cov\left[\overline{\mathbf{W}}_i | Z_l\right](r,q) &= \mathbb{E}\left[\left(Z_r - \mathbb{E}\left[Z_r | Z_l\right]\right)\left(Z_q - \mathbb{E}\left[Z_q | Z_l\right]\right) | Z_l\right] \\ &\quad - \mathbb{E}\left[Z_r - \mathbb{E}\left[Z_r | Z_l\right] | Z_l\right] \mathbb{E}\left[Z_q - \mathbb{E}\left[Z_q | Z_l\right] | Z_l\right] \\ &= \mathbb{E}[Z_r Z_q | Z_l] - \mathbb{E}[Z_r | Z_l]\mathbb{E}[Z_q | Z_l] - \mathbb{E}[Z_q | Z_l]\mathbb{E}[Z_r | Z_l] \\ &\quad + \mathbb{E}[Z_r | Z_l]\mathbb{E}[Z_q | Z_l] \\ &= \mathbb{E}[Z_r Z_q | Z_l] - \mathbb{E}[Z_r | Z_l]\mathbb{E}[Z_q | Z_l] = \Sigma_{w_i|t}(r,q), \end{aligned} \tag{B.20}$$

since the posterior covariance is independent of the specific observations from Eq.(B.15).
By taking the average in Eq.(B.19) and using the last relation, we get

$$Cov\left[\overline{\mathbf{W}}_i\right] = \Sigma_{w_i|t}. \tag{B.21}$$

B.4 Proof of Lemma 3.4.4

The result in Lemma 3.4.4 is a direct consequence of the following proposition (by an extension to the d-dimensional setting):

Proposition B.4.1. *Let $2 \leq m \leq N-1$ be positive integers and $X_1, \ldots, X_N$ be independent and identically distributed normal random variables with mean μ and variance σ^2. If $i \in \{1, \ldots, N-m+1\}$, then conditionally given $\sum_{j=i}^{i+m-1} X_j = t$, the random variables X_s, $s \neq i$, are jointly distributed as the random variables*

$$\tilde{X}_s = \left[\frac{t-\mu\sqrt{m}}{m} - \frac{1}{m-1}\left(1 - \frac{1}{\sqrt{m}}\right)\left(\sum_{j=i+1}^{i+m-1} X_j\right)\right] \mathbf{1}_{\{i+1,\ldots,i+m-1\}}(s) + X_s, \tag{B.22}$$

where $\mathbf{1}_A(x)$ is the indicator of the set A, i.e. $\mathbf{1}_A(x) = 1$ if $x \in A$ and zero otherwise.

For simplicity, we will prove the result for standard normal ($X_j \sim \mathcal{N}(0,1)$) random variables, since the general case will be straightforward after the usual change of variable $X_j \to \mu + \sigma X_j$.
Thus, we consider $X_1, \ldots, X_N$ be i.i.d. standard normal random variables and we want to show that

$$\tilde{X}_s = \left[\frac{t}{m} - \frac{1}{m-1}\left(1 - \frac{1}{\sqrt{m}}\right)\left(\sum_{j=i+1}^{i+m-1} X_j\right)\right] \mathbf{1}_{\{i+1,\ldots,i+m-1\}}(s) + X_s, \tag{B.23}$$

are jointly distributed as $\left\{ (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_N) \,|\, \sum_{j=i}^{i+m-1} X_j = t \right\}$.
To simplify the notations, let $f_{\tilde{\mathbf{X}}}$ and $g_{\mathbf{X}}$ be the joint density functions corresponding to the random vectors $\tilde{\mathbf{X}} = \left(\tilde{X}_1, \ldots, \tilde{X}_{i-1}, \tilde{X}_{i+1}, \ldots, \tilde{X}_N\right)$ and $\mathbf{X} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_N)$, respectively. Also, denote by $\Phi(x)$ the density function of a standard normal random variable.
By the change of variable in $N-1$ dimensions formula, we have

$$f_{\tilde{\mathbf{X}}}(u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_N) = g_{\mathbf{X}}\left(h^{-1}(u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_N)\right) \left|det(J_{h^{-1}})\right|, \tag{B.24}$$

where the function h is given by

$$h(v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_N) = \left(v_1, \ldots, v_{i-1}, \alpha - \beta\left(\sum_{j=i+1}^{i+m-1} v_j\right) + v_{i+1}, \ldots, \right.$$
$$\left. \alpha - \beta\left(\sum_{j=i+1}^{i+m-1} v_j\right) + v_{i+m-1}, v_{i+m}, \ldots, v_N \right), \tag{B.25}$$

with $\alpha = \frac{t}{m}$, $\beta = \frac{1}{m-1}\left(1 - \frac{1}{\sqrt{m}}\right)$ and where the Jacobian matrix of h^{-1} is equal with

$$J_{h^{-1}}(x) = \frac{\partial h^{-1}}{\partial \mathbf{u}}(x). \tag{B.26}$$

From Eq.(B.25), we obtain that h^{-1} has the form

$$h^{-1}\left(u_1\ ,\ldots,u_{i-1},u_{i+1},\ldots,u_N\right)=\left(u_1,\ldots,u_{i-1},-\frac{t}{\sqrt{m}}+\frac{1}{\sqrt{m}+1}\left(\sum_{j=i+1}^{i+m-1}u_j\right)\right.$$
$$\left.+u_{i+1},\ldots,-\frac{t}{\sqrt{m}}+\frac{1}{\sqrt{m}+1}\left(\sum_{j=i+1}^{i+m-1}u_j\right)+u_{i+m-1},u_{i+m},\ldots,u_N\right), \tag{B.27}$$

hence, by simple calculations, the Jacobian is equal with

$$|det(J_{h^{-1}})|=\sqrt{m}. \tag{B.28}$$

Since the random variables $X_1,\ldots,X_N$ are independent, the joint density function $g_{\mathbf{X}}$ is the product of the standard normal densities, that is

$$g_{\mathbf{X}}(v_1,\ldots,v_{i-1},v_{i+1},\ldots,v_N)=\Phi(v_1)\ldots\Phi(v_{i-1})\Phi(v_{i+1})\ldots\Phi(v_N), \tag{B.29}$$

thus, from Eqs.(B.24), (B.27) and (B.28), we obtain

$$f_{\tilde{\mathbf{X}}}(u_1,\ldots,u_{i-1},u_{i+1},\ldots,u_N)=\sqrt{m}\Phi(u_1)\ldots\Phi(u_{i-1})\Phi(u_{i+m})\ldots\Phi(u_N)$$
$$\times\prod_{s=i+1}^{i+m-1}\Phi\left(-\frac{t}{\sqrt{m}}+\frac{1}{\sqrt{m}+1}\left(\sum_{j=i+1}^{i+m-1}u_j\right)+u_s\right). \tag{B.30}$$

Taking the exponent of the joint density function $f_{\tilde{\mathbf{X}}}$ (denoted with $Exp\left(f_{\tilde{\mathbf{X}}}\right)$), we can write

$$Exp\left(f_{\tilde{\mathbf{X}}}\right)=-\frac{1}{2}\left[\sum_{\substack{s=1\\ s\notin\{i,\ldots,i+m-1\}}}^{N}u_s^2+\sum_{s=i+1}^{i+m-1}\left(u_s-\frac{t}{\sqrt{m}}+\frac{1}{\sqrt{m}+1}\left(\sum_{j=i+1}^{i+m-1}u_j\right)\right)^2\right]. \tag{B.31}$$

We have, by denoting with $S=\sum_{s=i+1}^{i+m-1}\left(u_s-\frac{t}{\sqrt{m}}+\frac{1}{\sqrt{m}+1}\left(\sum_{j=i+1}^{i+m-1}u_j\right)\right)^2$ and with $A=\sum_{j=i+1}^{i+m-1}u_j$,

$$S=\sum_{s=i+1}^{i+m-1}\left[u_s-\frac{t}{\sqrt{m}(\sqrt{m}+1)}+\frac{1}{\sqrt{m}+1}(A-t)\right]^2$$
$$=\sum_{s=i+1}^{i+m-1}\left[\left(u_s-\frac{t}{\sqrt{m}(\sqrt{m}+1)}\right)^2+\frac{2(A-t)}{\sqrt{m}+1}\left(u_s-\frac{t}{\sqrt{m}(\sqrt{m}+1)}\right)+\frac{(A-t)^2}{(\sqrt{m}+1)^2}\right]. \tag{B.32}$$

Since

$$\sum_{s=i+1}^{i+m-1} \frac{(A-t)^2}{(\sqrt{m}+1)} = \frac{\sqrt{m}-1}{\sqrt{m}+1}(A-t)^2, \tag{B.33}$$

$$\sum_{s=i+1}^{i+m-1} \frac{2(A-t)}{\sqrt{m}+1}\left(u_s - \frac{t}{\sqrt{m}(\sqrt{m}+1)}\right) = \frac{2(A-t)^2}{\sqrt{m}+1} + \frac{2t(A-t)}{\sqrt{m}(\sqrt{m}+1)} \tag{B.34}$$

and

$$\sum_{s=i+1}^{i+m-1}\left(u_s - \frac{t}{\sqrt{m}(\sqrt{m}+1)}\right)^2 = \sum_{s=i+1}^{i+m-1} u_s^2 - \frac{2t(A-t)}{\sqrt{m}(\sqrt{m}+1)} + \frac{t^2(\sqrt{m}-1)}{\sqrt{m}(\sqrt{m}+1)}, \tag{B.35}$$

we deduce, by substituting Eqs.(B.33), (B.34) and (B.35) in Eq.(B.32) , that

$$S = \sum_{s=i+1}^{i+m-1} u_s^2 + \left(\sum_{j=i+1}^{i+m-1} u_j - t\right)^2 - \frac{t^2}{m}, \tag{B.36}$$

so exponent $Exp\left(f_{\tilde{\mathbf{X}}}\right)$ becomes

$$Exp\left(f_{\tilde{\mathbf{X}}}\right) = -\frac{1}{2}\left[\sum_{s=1}^{N} u_s^2 + \left(\sum_{j=i+1}^{i+m-1} u_j - t\right)^2 - \frac{t^2}{m}\right]. \tag{B.37}$$

Combining Eqs.(B.30) and (B.37), we get that the density function $f_{\tilde{\mathbf{X}}}$ is given by

$$f_{\tilde{\mathbf{X}}}(u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_N) = \sqrt{m}\left(\frac{1}{\sqrt{2\pi}}\right)^{N-1} e^{Exp(f_{\tilde{\mathbf{X}}})}. \tag{B.38}$$

We note that the joint density function $\bar{f}$ of the random variables X_s, $s \neq i$, conditionally given $\sum_{j=i}^{i+m-1} X_j = t$, is by definition

$$\bar{f}(y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_N) = \frac{f_1(y_1, \ldots, y_{i-1}, t, y_{i+1}, \ldots, y_N)}{f_2(t)}, \tag{B.39}$$

where f_1 is the density function of $\left(X_1, \ldots, X_{i-1}, \sum_{j=i}^{i+m-1} X_j, X_{i+1}, \ldots, X_N\right)$ and f_2 is the density of $\sum_{j=i}^{i+m-1} X_j$. Since $\sum_{j=i}^{i+m-1} X_j \sim \mathcal{N}(0, m)$, we have

$$f_2(t) = \frac{1}{\sqrt{2\pi m}} e^{-\frac{t^2}{2m}}. \tag{B.40}$$

The density f_1 is given by, after applying the change of variable,

$$f_1(y_1, \ldots, y_{i-1}, t, y_{i+1}, \ldots, y_N) = \bar{g}(\bar{h}^{-1}(y_1, \ldots, y_N))\left|det(J_{\bar{h}^{-1}})\right|, \tag{B.41}$$

where $\bar{g}$ is the density of the vector $(X_1, \ldots, X_N)$ and

$$\bar{h}^{-1}(y_1, \ldots, y_N) = \left(y_1, \ldots, y_{i-1}, y_i - \sum_{j=i+1}^{i+m-1} y_j, y_{i+1}, \ldots, y_N \right). \tag{B.42}$$

We can easily see that $|det(J_{\bar{h}^{-1}})| = 1$ and, from the independence of X_s,

$$\bar{g}(v_1, \ldots, v_N) = \Phi(v_1) \ldots \Phi(v_N). \tag{B.43}$$

Substituting Eqs.(B.40)-(B.43) in (B.39), we obtain

$$\begin{aligned} \bar{f} &= \frac{\Phi(y_1) \ldots \Phi(y_{i-1}) \Phi\left(t - \sum_{j=i+1}^{i+m-1} y_j \right) \Phi(y_{i+1}) \ldots \Phi(y_N)}{\frac{1}{\sqrt{2\pi m}} e^{-\frac{t^2}{2m}}} \\ &= \sqrt{m} \left(\frac{1}{\sqrt{2\pi}} \right)^{N-1} e^{-\frac{1}{2}\left[\sum_{s=1}^{N} y_s^2 + \left(\sum_{j=i+1}^{i+m-1} y_j - t \right)^2 - \frac{t^2}{m} \right]}, \end{aligned} \tag{B.44}$$

which is exactly the formula in Eq.(B.38), what we needed to show. ■

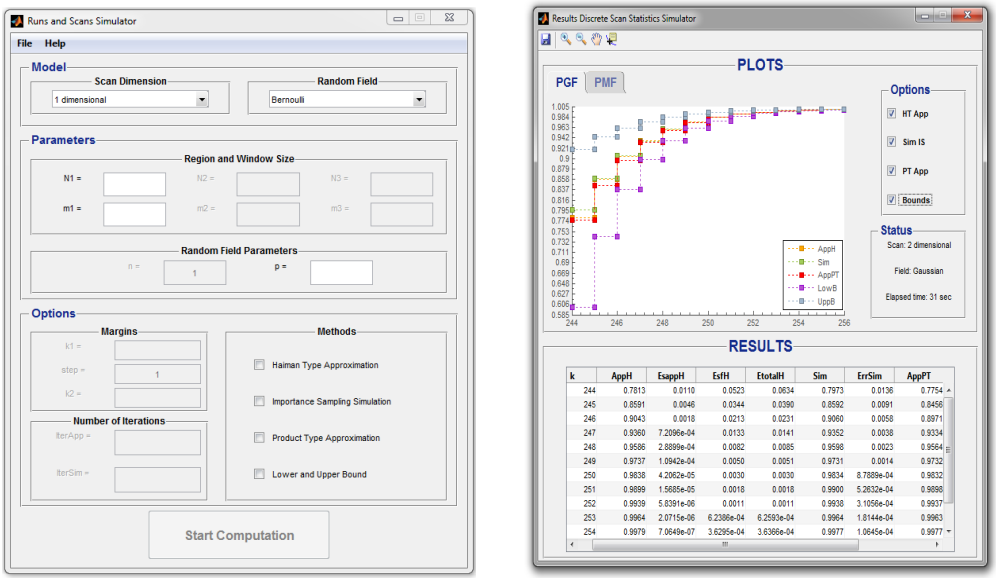
Remark B.4.2. *It is interesting to notice that the result in Proposition B.4.1 can be generalized by conditioning with a linear combination instead of just the sum. The proof follows the lines of the above proof and we will omit it.*

Appendix C

A Matlab graphical user interface for discrete scan statistics

We present here a graphical user interface (GUI), developed in Matlab®, that permits to estimate the distribution of the discrete scan statistics for different scenarios. The purpose of this GUI application is to illustrate the accuracy of the existing methods used for the approximation process of the scan statistic variable. We consider the cases of one, two and three dimensional scan statistics over a random field distributed according to a Bernoulli, binomial, Poisson or Gaussian law. In the particular situation of one dimensional scan statistics, we have also included a moving average of order q model.
We should emphasize that almost all of the numerical results included in this thesis were obtained with the help of this GUI program.

C.1 How to use the interface



(a) The *Input Window* (b) The *Output Window*

Figure C.1: The *Scan Statistics Simulator* GUI

Once the program is executed, the user interface window from Figure C.1(a) appears. This window, denoted in the subsequent as *Input Window*, is divided into three main panels: *Model*, *Parameters* and *Options*, which are interconnected. The *Options* panel becomes active only after the first two panels are correctly filled in, the same being true for the *Computation Button*.
From the *Model* section, the user can choose the dimension of the problem and the distribution of the model that needs to be evaluated. The following table illustrates the available options for each dimension.

	1 dimensional	2 dimensional	3 dimensional
Bernoulli	√	√	√
Binomial	√	√	√
Poisson	√	√	√
Gaussian	√	√	√
Moving Average	√		

Table C.1: *Scan Dimension* versus *Random Field*

Based on the values selected for the *Scan Dimension*, the first section of the *Parameters* panel activates accordingly. The first line gives the size of the region to be scanned (at most $N_1 \times N_2 \times N_3$), while the second line shows the size of the scanning window. The *Random Field Parameters* section change in accordance with the selection made in the *Random Field* panel. For example, if we select the *Bernoulli* distribution, then we get the possibility to insert only the success probability value p.
If all the values in the *Parameters* panel are correctly inserted, the *Options* panel is activated. This panel is composed of three sections: *Margins*, *Number of Iterations* and *Methods*. In the *Margins* section, we insert the values in-between we need to evaluate the distribution of the discrete scan statistics. For example, if $k_1 = 1$ and $k_2 = 4$, the distribution $\mathbb{P}(S_{\mathbf{m}}(\mathbf{T}) \leq k)$ is computed for $k \in \{1, 2, 3, 4\}$. In the case of a Gaussian model, we also can choose the step. As the name mention, the *Number of Iterations* section permits the user to select the number of replications needed to run the algorithms for both simulation (*IterSim*) or approximation (*IterApp*).
The *Methods* section is probably the most important. This section lets the user to choose between different methods used in the estimation of scan statistics distribution. There are two approximation methods, the *Haiman* type approximation developed in Chapter 3 and the product type approximation presented in Chapter 1, a method used for the simulation of the distribution and lower and upper bounds (when there are available). For the simulated value and the *Haiman* type approximation we used the importance sampling algorithms described in Chapter 3. Table C.2 summarizes the approaches used for evaluating the distribution of the discrete scan statistics implemented in the program.
As soon as the computation button is pushed and the calculations are done, a second window pops-up, namely the *Output Window*. This window, as we can see

		AppH	E_{sapp}	E_{sf}	*AppPT*	*LowB*	*UppB*
Bernoulli	1*d*	Eq.(3.91)	Eq.(3.92)	–	Eq.(1.46)	Eq.(1.47)	Eq.(1.49)
	2*d*	Eq.(3.95)	Eq.(3.96)	Eq.(3.97)	Eq.(1.67)	Eq.(1.76)	Eq.(1.77)
	3*d*	Eq.(3.99)	Eq.(3.101)	Eq.(3.100)	Eq.(1.87)	–	–
Binomial	1*d*	Eq.(3.91)	Eq.(3.93)	Eq.(3.94)	Eq.(1.46)	Eq.(1.47)	Eq.(1.49)
	2*d*	Eq.(3.95)	Eq.(3.96)	Eq.(3.97)	Eq.(1.67)	Eq.(1.76)	Eq.(1.77)
	3*d*	Eq.(3.99)	Eq.(3.101)	Eq.(3.100)	Eq.(1.87)	–	–
Poisson	1*d*	Eq.(3.91)	Eq.(3.93)	Eq.(3.94)	Eq.(1.46)	Eq.(1.47)	Eq.(1.49)
	2*d*	Eq.(3.95)	Eq.(3.96)	Eq.(3.97)	Eq.(1.67)	Eq.(1.76)	Eq.(1.77)
	3*d*	Eq.(3.99)	Eq.(3.101)	Eq.(3.100)	Eq.(1.87)	–	–
Gaussian	1*d*	Eq.(3.91)	Eq.(3.93)	Eq.(3.94)	Eq.(1.46)	Eq.(1.47)	Eq.(1.49)
	2*d*	Eq.(3.95)	Eq.(3.96)	Eq.(3.97)	Eq.(1.67)	Eq.(1.76)	Eq.(1.77)
	3*d*	Eq.(3.99)	Eq.(3.101)	Eq.(3.100)	Eq.(1.87)	–	–
Moving Average	1*d*	Eq.(3.91)	Eq.(4.20)	Eq.(4.23)	Eq.(1.46)	Eq.(4.35)	Eq.(4.36)

Table C.2: Relations used for estimating the distribution of the scan statistics

from Figure C.1(b), is partitioned into two parts: *Plots* and *Results*.
The *Plots* section is divided in three subsections: *Axes*, *Options* and *Status*. The main part (the *Axes*) shows the graphical representation of the distribution functions selected from the *Option* panel. The figure is plotted in accordance to the given computation parameters (bounds, methods, etc.) from the *Input Window*.
The *Options* panel is in one to one correspondence with the *Methods* panel from the *Input Window*. Notice that there are active only the options that are conform to the methods previously selected by the user. By checking an active option, on the Axes panel it is drawn the distribution function associated with the method. We should mention that the obtained image can be saved, by right clicking on the figure, in several formats: .bmp, .jpeg, .png or .pdf.
On the bottom of the *Output Window*, the user can find the *Results* panel. Here it is displayed the numerical information obtained from the algorithms. The table contains, for each method selected by the user, a different number of columns as follows: for the Product Type Approximation method one column with "AppT" caption, for the Lower and Upper Bounds method two columns entitled "LowB" and "UppB", for the Importance Sampling Simulation method two columns "Sim" and "ErrSim" (the simulation error) and for the Haiman Type Approximation (IS) four columns "AppH", "EsappH", "EsfH" and "EtotalH" (these captions corresponds to Approximation, Approximation Error, Simulation Error, Total Error; for more details see Table C.2). The resulted values can be saved, by right clicking on the table, in .txt, .xlsx and even .tex formats.

C.2 Future developments

This graphical user interface is just a part of a larger project that will include also the estimation of the distribution of the multidimensional continuous scan statistics.

Bibliography

[Aaronson et al., 1989] Aaronson, J., Gilat, D., Keane, M., and de Valk V. (1989). An algebraic construction of a class of one-dependent processes. *Ann. Probab.*, 17:475–480. (Cited on page 28.)

[Aaronson and Gilat D., 1992] Aaronson, J. and Gilat D., K. M. (1992). On the structure of one-dependent Markov chains. *J. Theoret. Probab.*, 5:545–561. (Cited on page 28.)

[Aki, 1992] Aki, S. (1992). Waiting time problems for a sequence of discrete random variables. *Ann. Inst. Statist. Math.*, 44:363–378. (Cited on page 12.)

[Akiba and Yamamoto, 2004] Akiba, T. and Yamamoto, H. (2004). Upper and lower bounds for 3-dimensional r-within n consecutive -(r_1, r_2, r_3)-out-of -(n_1, n_2, n_3): F-system. In *Advanced Reliability Modelling*. (Cited on page 26.)

[Akiba and Yamamoto, 2005] Akiba, T. and Yamamoto, H. (2005). Evaluation for the reliability of a large 2-dimensional rectangular k-within-consecutive -(r, s)-out-of-(m, n): F system. *Journal of Japan Industrial Management Association*, 53:208–219. (Cited on page 23.)

[Alm, 1998] Alm, S. (1998). Approximation and simulation of the distributions of scan statistics for Poisson processes in higher dimensions. *Extremes*, 1:111–126. (Cited on page 119.)

[Amărioarei, 2012] Amărioarei, A. (2012). Approximation for the distribution of extremes of one dependent stationary sequences of random variables. *arXiv:1211.5456v1, submitted.* (Cited on page 27.)

[Amărioarei and Preda, 2013a] Amărioarei, A. and Preda, C. (2013a). Approximation for the distribution of three-dimensional discrete scan statistic. *Methodol Comput Appl Probab*. (Cited on pages 55, 56, and 57.)

[Amărioarei and Preda, 2013b] Amărioarei, A. and Preda, C. (2013b). Approximations for two-dimensional discrete scan statistics in some dependent models. In *Proceedings, 15th Applied Stochastic Models and Data Analysis (ASMDA2013)*. (Cited on page 95.)

[Amărioarei and Preda, 2014] Amărioarei, A. and Preda, C. (2014). Approximations for two-dimensional discrete scan statistics in some block-factor type dependent models. *Journal of Statistical Planning and Inference*, 151-152:107–120. (Cited on page 95.)

[Balakrishnan and Koutras, 2002] Balakrishnan, N. and Koutras, M. V. (2002). *Runs and scans with applications.* Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], New York. (Cited on pages 9 and 108.)

[Balakrishnan et al., 1997] Balakrishnan, N., Mohanty, S. G., and Aki, S. (1997). Start-up demonstration tests under Markov dependence model with corrective actions. *Ann. Inst. Statist. Math.*, 49:155–169. (Cited on page 12.)

[Barton and Mallows, 1965] Barton, D. and Mallows, C. (1965). Some aspects of the random sequence. *Annals of Mathematical Statistics*, 36:236–260. (Cited on page 7.)

[Boutsikas and Koutras, 2000] Boutsikas, M. V. and Koutras, M. V. (2000). Reliability approximation for Markov chain imbeddable systems. *Methodol. Comput. Appl. Probab.*, 2:393–411. (Cited on pages 6, 20, 21, and 25.)

[Boutsikas and Koutras, 2003] Boutsikas, M. V. and Koutras, M. V. (2003). Bounds for the distribution of two-dimensional binary scan statistics. *Probab. Eng. Inform. Sci.*, 17:509–525. (Cited on page 22.)

[Broman, 2005] Broman, E. (2005). One dependent trigonometric determinantal processes are two-block factors. *Ann. Probab.*, 33:601–609. (Cited on page 29.)

[Burton et al., 1993] Burton, R. M., Goulet, M., and Meester, R. (1993). On one-dependent processes and k-block factors. *Ann. Probab.*, 21:2157–2168. (Cited on pages 28, 29, 96, and 97.)

[Chaderjian et al., 2012] Chaderjian, B., Ebneshahrashoob, M., and Gao, T. (2012). Exact distributions of waiting time problems of mixed frequencies and runs in Markov dependent trials. *Applied Mathematics*, 3:1689–1696. (Cited on page 12.)

[Chang, 2002] Chang, Y. (2002). *Waiting times distributions of runs and patterns.* PhD thesis, University of Manitoba. (Cited on page 10.)

[Chen and Glaz, 1996] Chen, J. and Glaz, J. (1996). Two-dimensional discrete scan statistics. *Statist. Probab. Lett.*, 31:59–68. (Cited on pages 19, 22, 23, 56, 57, and 87.)

[Chen and Glaz, 1997] Chen, J. and Glaz, J. (1997). Approximations and inequalities for the distribution of a scan statistic for 0-1 Bernoulli trials. *Advances in the Theory and Practice of Statistics*, 1:285–298. (Cited on page 15.)

[Chen and Glaz, 2009] Chen, J. and Glaz, J. (2009). *Scan statistics*, chapter 5, Approximations for two-dimensional variable window scan statistics., pages 109–128. Birkhäuser Boston, Inc., Boston. (Cited on page 21.)

[Cressie, 1993] Cressie, N. (1993). *Statistics for spatial data.* Wiley-Interscience [John Wiley & Sons], New York. (Cited on page 57.)

[Csaki and Foldes, 1996] Csaki, E. and Foldes, A. (1996). On the length of theh longest monnotone block. *Studio Scientiarum Mathematicarum Hungarica*, 31:35–46. (Cited on page 108.)

[Darling and Waterman, 1986] Darling, R. W. R. and Waterman, M. S. (1986). Extreme value distribution for the largest cube in a random lattice. *SIAM J. Appl. Math.*, 46:118–132. (Cited on pages 6 and 25.)

[Dawson and Sankoff, 1967] Dawson, A. and Sankoff, D. (1967). An inequality for probabilities. *Proceedings of the American Mathematical Society*, 18:504–507. (Cited on pages 18 and 19.)

[de Valk, 1988] de Valk, V. (1988). *One-dependent processes.* PhD thesis, Technische Universiteit Delft (The Netherlands). (Cited on page 29.)

[de Valk and Ruschendorf, 1993] de Valk, V. and Ruschendorf, L. (1993). On regression representations of stochastic processes. *Stochastic Process. Appl.*, 46:183–198. (Cited on page 29.)

[Devroye, 1986] Devroye, L. (1986). *Non uniform random variate generation.* Springer-Verlag, New York. (Cited on page 76.)

[Doucet, 2010] Doucet, A. (2010). A note on efficient conditional simulation of Gaussian distributions. (Cited on page 77.)

[Ebneshahrashoob et al., 2004] Ebneshahrashoob, M., Gao, T., and Sobel, M. (2004). Double window acceptance sampling. *Naval Research Logistics*, 51:297–306. (Cited on page 12.)

[Ebneshahrashoob et al., 2005] Ebneshahrashoob, M., Gao, T., and Wu, M. (2005). An efficient algorithm for exact distribution of discrete scan statistics. *Methodol. Comput. Appl. Probab.*, 7:1423–1436. (Cited on pages 12, 13, and 14.)

[Ebneshahrashoob and Sobel, 1990] Ebneshahrashoob, M. and Sobel, M. (1990). Sooner and later waiting time problems for Bernoulli trials: frequency and run quotas. *Statist. Probab. Lett.*, 9:5–11. (Cited on page 12.)

[Eryilmaz, 2006] Eryilmaz, S. (2006). A note on runs of geometrically distributed random variables. *Discrete Mathematics*, 306:1765–1770. (Cited on page 108.)

[Esary et al., 1967] Esary, J., Proschan, F., and Walkup, D. (1967). Association of random variables, with applications. *The Annals of Mathematical Statistics*, 38:1466–1474. (Cited on page 107.)

[Feller, 1968] Feller, W. (1968). *An Introduction to Probability Theory and Its Applications.*, volume 1. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], New York. (Cited on pages 12 and 13.)

[Fishman, 1996] Fishman, G. (1996). *Monte Carlo: Concepts, Algorithms and Applications.* Springer Series in Operations Research. Springer-Verlag, New York. (Cited on pages 66, 69, and 71.)

[Flak and Schmid, 1995] Flak, T. and Schmid, W. (1995). Extreme sums of strictly stationary sequences of m-dependent variables. *Indian J. Stat.*, 57:186–201. (Cited on page 30.)

[Frigessi and Vercellis, 1984] Frigessi, A. and Vercellis, C. (1984). An analysis of Monte Carlo algorithms for counting problems. *Department of Mathematics, University of Milan.* (Cited on page 71.)

[Frolov and Martikainen, 1999] Frolov, A. and Martikainen, A. (1999). On the length of the longest increasing run in R^d. *Stat. and Prob. Letters*, 41:153–161. (Cited on page 108.)

[Fu, 1986] Fu, J. (1986). Reliability of consecutive k out of n: f system with $(k-1)$ step Markov dependence. *IEEE Trans. Reliab.*, 35:602–606. (Cited on page 9.)

[Fu, 1996] Fu, J. (1996). Distribution theory of runs and patterns associated with a sequence of multistate trials. *Stat. Sininca*, 6:957–974. (Cited on page 9.)

[Fu, 2001] Fu, J. (2001). Distribution of the scan statistic for a sequence of bistate trials. *J. Appl. Probab.*, 38:908–916. (Cited on pages 9, 10, 11, 13, and 15.)

[Fu and Chang, 2002] Fu, J. and Chang, Y. (2002). On probability generating functions for waiting time distributions of compound patterns in a sequence of multistate trials. *J. Appl. Prob.*, 39:70–80. (Cited on pages 9 and 10.)

[Fu and Koutras, 1994] Fu, J. C. and Koutras, M. V. (1994). Distribution theory of runs: a Markov chain approach. *J. Amer. Statist. Assoc.*, 89:1050–1058. (Cited on page 9.)

[Fu and Lou, 2003] Fu, J. C. and Lou, W. (2003). *Distribution theory of runs and patterns and its applications. A finite Markov chain imbedding approach.* World Scientific Publishing Co., Inc., River Edge, NJ. (Cited on pages 9, 11, and 108.)

[Galambos, 1972] Galambos, J. (1972). On the distribution of the maximum of random variables. *Ann. Math. Stat.*, 43:516–521. (Cited on page 30.)

[Galambos and Simonelli, 1996] Galambos, J. and Simonelli, I. (1996). *Bonferroni-type Inequalities with Applications.* Springer-Verlag, New York. (Cited on page 18.)

[Gao and Wu, 2006] Gao, T. and Wu (2006). A combinatorial problem from sooner waiting time problems with run and frequency quotas. *J. Math. Anal. Appl.*, 321:949–960. (Cited on page 14.)

[Genz and Bretz, 2009] Genz, A. and Bretz, F. (2009). *Computation of Multivariate Normal and T Probabilities.* Springer-Verlag, New York. (Cited on pages v, 82, 83, and 112.)

[Glaz, 1990] Glaz, J. (1990). A comparison of product-type and Bonferroni-type inequalities in presence of dependence. In *Symposium on Dependence in Probability and Statistics.*, volume 16 of *IMS Lecture Notes-Monograph Series*, pages 223–235. IMS Lecture Notes. (Cited on page 17.)

[Glaz and Balakrishnan, 1999] Glaz, J. and Balakrishnan, N. (1999). *Scan Statistics and Applications.* Springer Sciences+Business Media. (Cited on pages 6 and 15.)

[Glaz and Naus, 1991] Glaz, J. and Naus, J. (1991). Tight bounds and approximations for scan statistic probabilities for discrete data. *Annals of Applied Probability*, 1:306–318. (Cited on pages 6, 15, 17, 71, 106, and 107.)

[Glaz et al., 2001] Glaz, J., Naus, J., and Wallenstein, S. (2001). *Scan statistics.* Springer Series in Statistics. Springer-Verlag, New York. (Cited on pages 6, 7, 13, 15, 17, 20, 21, 22, 25, 56, 57, and 71.)

[Glaz et al., 2009] Glaz, J., Pozdnyakov, V., and Wallenstein, S. (2009). *Scan statistics. Methods and applications.* Birkhäuser Boston, Inc., Boston, MA. (Cited on pages 20, 25, and 122.)

[Goulet, 1992] Goulet, M. (1992). *One-dependence and k-block factors.* PhD thesis, Oregon State Univ. (Cited on page 29.)

[Grabner et al., 2003] Grabner, P., Knopfmacher, A., and Prodinger, H. (2003). Combinatorics of geometrically distributed random variables: run statistics. *Theoret. Comput. Sci.*, 297:261–270. (Cited on page 108.)

[Grill, 1987] Grill, K. (1987). Erdos-Révész type bounds for the length of the longest run from a stationary mixing sequence. *Probab. Theory Relat. Fields*, 75:169–179. (Cited on page 108.)

[Guerriero et al., 2010a] Guerriero, M., Glaz, J., and Sen, R. (2010a). Approximations for a three dimensional scan statistic. *Methodol. Comput. Appl. Probab.*, 12:731–747. (Cited on pages 24, 26, 56, 57, 89, 90, 92, and 125.)

[Guerriero et al., 2010b] Guerriero, M., Pozdnyakov, V., Glaz, J., and Willett, P. (2010b). A repeated significance test with applications to sequential detection in sensor networks. *IEEE Trans. Signal Process.*, 58:3426–3435. (Cited on page 6.)

[Guerriero et al., 2009] Guerriero, M., Willett, P., and Glaz, J. (2009). Distributed target detection in sensor networks using scan statistics. *IEEE Trans. Signal Process.*, 57:2629–2639. (Cited on pages 6, 22, and 57.)

[Haiman, 1999] Haiman, G. (1999). First passage time for some stationary processes. *Stochastic Process. Appl.*, 80:231–248. (Cited on pages 27, 30, 32, 36, and 45.)

[Haiman, 2000] Haiman, G. (2000). Estimating the distributions of scan statistics with high precision. *Extremes*, 3:349–361. (Cited on pages 6, 32, and 57.)

[Haiman, 2007] Haiman, G. (2007). Estimating the distribution of one-dimensional discrete scan statistics viewed as extremes of 1-dependent stationary sequences. *J. Statist. Plann. Inference*, 137:821–828. (Cited on pages 17, 32, and 57.)

[Haiman, 2012] Haiman, G. (2012). 1-dependent stationary sequences for some given joint distributions of two consecutive random variables. *Methodol. Comput. Appl. Probab.*, 14:445–458. (Cited on pages 104 and 105.)

[Haiman et al., 1995] Haiman, G., Kiki, M., and Puri, M. (1995). Extremes of Markov sequences. *J. Stat. Plan. Infer.*, 45:185–201. (Cited on page 32.)

[Haiman and Preda, 2002] Haiman, G. and Preda, C. (2002). A new method for estimating the distribution of scan statistics for a two-dimensional Poisson process. *Methodol. Comput. Appl. Probab.*, 4:393–407. (Cited on pages 32 and 57.)

[Haiman and Preda, 2006] Haiman, G. and Preda, C. (2006). Estimation for the distribution of two-dimensional discrete scan statistics. *Methodol. Comput. Appl. Probab.*, 8:373–381. (Cited on pages 22, 32, 57, and 68.)

[Haiman and Preda, 2013] Haiman, G. and Preda, C. (2013). One dimensional scan statistics generated by some dependent stationary sequences. *Statist. Probab. Lett.*, 83:1457–1463. (Cited on pages 32, 98, 104, and 105.)

[Han and Hirano, 2003] Han, Q. and Hirano, K. (2003). Waiting time problem for an almost perfect match. *Stat. and Prob. Letters*, 65:39–49. (Cited on page 12.)

[Harrelson and Houdre, 2003] Harrelson, D. and Houdre, C. (2003). A characterization of m-dependent stationary infinitely divisible sequences with applications to weak convergence. *Ann. Probab.*, 31:849–881. (Cited on page 29.)

[Hoffman and Ribak, 1991] Hoffman, Y. and Ribak, E. (1991). Constrained realizations of Gaussian Fields - a simple algorithm. *The Astrophysical Journal*, 380:L5–L8. (Cited on pages 77 and 135.)

[Hoh and Ott, 2000] Hoh, J. and Ott, J. (2000). Scan statistics to scan markers for susceptibility genes. *Proc Natl Acad Sci*, 97:9615–9617. (Cited on page 6.)

[Hoover, 1990] Hoover, D. (1990). Subset complement addition upper bounds - an improved inclusion-exclusion method. *J. Stat. Plan. Infer.*, 24:195–202. (Cited on page 23.)

[Hunter, 1976] Hunter, D. (1976). An upper bound for the probability of a union. *Journal of Applied Probability*, 13:597–603. (Cited on page 18.)

[Ibragimov and Linnik, 1971] Ibragimov, I. A. and Linnik, Y. V. (1971). *Independent and stationary sequences of random variables.* Wolters-Noordhoff Publishing. (Cited on page 28.)

[Karlin and McGregor, 1959] Karlin, S. and McGregor, G. (1959). Coincidence probability. *Pacific J. Math.*, 9:1141–1164. (Cited on page 7.)

[Karwe and Naus, 1997] Karwe, V. and Naus, J. (1997). New recursive methods for scan statistic probabilities. *Computational Statistics & Data Analysis*, 17:389–402. (Cited on pages 16, 17, 19, 21, 23, and 121.)

[Kolda and Bader, 2009] Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review*, 51:455–500. (Cited on page 96.)

[Koutras and Alexandrou, 1995] Koutras, M. V. and Alexandrou, V. A. (1995). Runs, scans and urn model distributions: a unified Markov chain approach. *Ann. Inst. Statist. Math.*, 47:743–766. (Cited on page 9.)

[Kuai et al., 2000] Kuai, H., Alajaji, F., and Takahara, G. (2000). A lower bound on the probability of a finite union of events. *Discrete Mathematics*, 215:147–158. (Cited on pages 19, 23, and 124.)

[Kwerel, 1975] Kwerel, M. (1975). Most stringent bounds on aggregated probabilities of partially specified dependent probability systems. *Journal of the American Statistical Association*, 70:472–479. (Cited on page 18.)

[Louchard, 2005] Louchard, G. (2005). Monotone runs of uniformly distributed integer random variables: a probabilistic analysis. *Theoret. Comput. Sci.*, 346:358–387. (Cited on page 108.)

[Malley et al., 2002] Malley, J., Naiman, D. Q., and Bailey-Wilson, J. (2002). A compresive method for genome scans. *Human Heredity*, 54:174–185. (Cited on pages 79 and 111.)

[Marcos and Marcos, 2008] Marcos, R. and Marcos, C. (2008). From star complexes to the field: open cluster families. *Astrophysical Journal*, 672:342–351. (Cited on page 6.)

[Marsaglia, 1963] Marsaglia, G. (1963). Generating a variable from the tail of the normal distribution. Technical report, Boeing Scientific Research Laboratories. (Cited on page 76.)

[Matus, 1998] Matus, F. (1998). Combining m-dependence with Markovness. *Ann. Inst. Henri Poincare*, 34:407–423. (Cited on page 29.)

[Mitton et al., 2010] Mitton, N., Paroux, K., Sericola, B., and Tixeuil, S. (2010). Ascending runs in dependent uniformly distributed random variables: application to wireless networks. *Methodol Comput Appl Probab*, 12:51–62. (Cited on page 108.)

[Naiman and Priebe, 2001] 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. (Cited on pages 6, 25, 71, and 111.)

[Naiman and Wynn, 1997] 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. (Cited on pages 68 and 73.)

[Naus, 1974] Naus, J. (1974). Probabilities for a generalized birthday problem. *Journal of American Statistical Association*, 69:810–815. (Cited on page 7.)

[Naus, 1982] Naus, J. (1982). Approximations for distributions of scan statistics. *Journal of American Statistical Association*, 77:177–183. (Cited on pages 6, 8, 15, and 16.)

[Neil, 2006] Neil, D. (2006). *Detection of spatial and spatio-temporal clusters*. PhD thesis, School of Computer Science, Carnegie Mellon University. (Cited on page 79.)

[Neil, 2012] Neil, D. (2012). Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society*, 74(2):337–360. (Cited on page 79.)

[Neill et al., 2005] Neill, D., Moore, A., Pereira, F., and Mitchell, T. (2005). Detecting significant multidimensional spatial clusters. *Advances in Neural Information Processing Systems*, 17:969–976. (Cited on page 79.)

[Newell, 1964] Newell, G. (1964). Asymptotic extremes for m-dependent random variables. *Ann. Math. Stat.*, 35:1322–1325. (Cited on page 30.)

[Novak, 1992] Novak, S. (1992). Longest runs in a sequence of m-dependent random variables. *Probab. Theory Relat. Fields*, 91:269–281. (Cited on pages 108 and 109.)

[Nuel, 2008a] Nuel, G. (2008a). Pattern Markov chains: optimal Markov chain embedding through deterministic finite automata . *Journal of Applied Probability*, 45:226–243. (Cited on page 12.)

[Nuel, 2008b] Nuel, G. (2008b). Waiting time distribution for pattern occurrence in a constrained sequence: an embedding Markov chain approach. *Discrete Mathematics and Theoretical Computer Science*, 10:149–160. (Cited on page 12.)

[Oliveira, 2012] Oliveira, P. E. (2012). *Asymptotics for associated random variables*. Springer Sciences+Business Media. (Cited on page 107.)

[Pittel, 1981] Pittel, B. (1981). Limiting behavior of a process of runs. *Ann. Probab.*, 9:119–129. (Cited on page 108.)

[Priebe et al., 2001] Priebe, C. E., Naiman, D. Q., and Cope, L. (2001). Importance sampling for spatial scan analysis: computiing scan statistic p-values for marked processes. *Computational Statistics & Data Analysis*, 35:475–485. (Cited on page 82.)

[Révész, 1983] Révész, P. (1983). Three problems on the llength of increasing runs. *Stochastic Process. Appl.*, 5:169–179. (Cited on page 108.)

[Riesz and Nagy, 1990] Riesz, F. and Nagy, B. (1990). *Functional Analysis (Reprint of the 1955 original)*. Dover Publications, Inc., New York. (Cited on page 52.)

[Ross, 2012] Ross, S. (2012). *Simulation (Fifth Edition)*. Elsevier Academic Press Publications. (Cited on pages 69 and 70.)

[Rubino and Tuffin, 2009] Rubino, G. and Tuffin, B. (2009). *Rare event simlation using Monte Carlo methods*. Wiley-Interscience [John Wiley & Sons], New York. (Cited on pages 69 and 70.)

[Rubinstein and Kroese, 2008] Rubinstein, R. and Kroese, D. (2008). *Simulation and the Monte Carlo method*. Wiley-Interscience [John Wiley & Sons], New York. (Cited on page 69.)

[Sheng and Naus, 1994] Sheng, K.-N. and Naus, J. (1994). Pattern matching between two non aligned random sequences. *Bulletin of Mathematical Biology*, 56:1143–1162. (Cited on page 6.)

[Shi et al., 2007] 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. (Cited on pages 80, 82, and 111.)

[Shinde and Kotwal, 2008] Shinde, R. and Kotwal, K. (2008). Distributions of scan statistics in a sequence of Markov Bernoulli trials. *International Journal of Statisitcs*, LXVI:135–155. (Cited on page 15.)

[Tong, 1990] Tong, Y. (1990). *The multivariate normal distribution*. Springer Series in Statistics. Springer-Verlag, New York. (Cited on page 77.)

[Uchida, 1998] Uchida, M. (1998). On generating functions of waiting time problems for sequence patterns of discrete random variables. *Ann. Inst. Statist. Math.*, 50:650–671. (Cited on page 12.)

[Uchida and Aki, 1995] Uchida, M. and Aki, S. (1995). Sooner and later waiting time problems in a two-state Markov chain. *Ann. Inst. Statist. Math.*, 47:415–433. (Cited on page 12.)

[Wang, 2013] Wang, X. (2013). *Scan statistics for normal data*. PhD thesis, University of Connecticut. (Cited on pages 98 and 112.)

[Wang and Glaz, 2013] 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. (Cited on pages 98 and 112.)

[Wang et al., 2012] Wang, X., Glaz, J., and Naus, J. (2012). Approximations and inequalities for moving sums. *Methodol. Comput. Appl. Probab.*, 14:597–616. (Cited on pages 6, 15, 17, and 106.)

[Watson, 1954] Watson, G. (1954). Extreme values in samples from m-dependent stationary stochastic processes. *Ann. Math. Stat.*, 25:798–800. (Cited on page 30.)

[Worsley, 1982] Worsley, K. (1982). An improved Bonferroni inequality and applications. *Biometrika*, 69:297–302. (Cited on page 18.)

[Wu, 2013] Wu, T.-L. (2013). On finite Markov chain imbedding technique. *Methodol Comput Appl Probab*, 15:453–465. (Cited on pages 9, 11, and 13.)

[Wu and Naiman, 2005] Wu, X. and Naiman, D. Q. (2005). p-Value Simulation for Affected Sib Pair Multiple Testing. *Human Heredity*, 59:190–200. (Cited on page 82.)

[Yaglom and Yaglom, 1987] Yaglom, A. and Yaglom, I. (1987). *Challenging mathematical problems with elementary solutions.*, volume 1. Dover Publications, Inc., New York. (Cited on page 49.)