

On Statistical Testing of Random Numbers Generators

El Haje, F.¹, Golubev, Yu.², Liardet, P.-Y.³, Teglia, Ya.⁴

^{1,2} Université de Provence (Aix-Marseille 1)

CMI, 39 rue F. Joliot-Curie, 13453 Marseille, France

^{3,4} ST Microelectronics, ZI Rousset BP2 13106 Rousset, France

Maurer's test is nowadays a basic statistical tool for testing physical random number generators in cryptographic applications. Based on a statistical analysis of this test we propose simple and effective methods for its improvement. These methods are related to the m - spacing technique common in goodness-of-fit problems and the L - leave out method used for a noise reduction in the final Maurer test statistic. We also show that the spacing distribution test represents a serious competitor for Maurer's test in the case when the random number generator is governed by a Markov chain with a long memory. Finally we discuss some approaches to the multiple testing problem which seems essential for constructing new powerful statistical tests.

Key words: maximum likelihood test, multiple testing, entropy, Maurer's test, spacings.

1 Introduction

1.1 Cryptographic applications of statistical tests

Generating random numbers is not only a key issue in cryptographic applications, but also in counter measures against side-channel attacks on secure tokens like Smartcards (see [8] and [16] for some instances). The cornerstone character of these problems have brought institutional organizations, like the NIST in the USA and the BSI (Bundersamt für Sicherheit der Informationstechnik) in Germany, to develop standards to define Random Number Generators (RNGs), to classify them regarding their intended use and to analyze the confidence that one can have in claimed properties of RNGs.

The first approach developed in [12] and [13] was to qualify RNG using statistical tests. Canonical statistical tests include for instance the frequency test aimed to check uniformity of the outputs of the RNG and the long run test that verifies whether the RNG is not stuck at a given value during a defined period.

In the meantime, under the concern of completing security evaluation criteria of ITSEC or Common Criteria concerning random supply, RNG have been classified in two categories by the BSI:

- Pseudo Random Number Generators (PRNG) that apply iterative numerical algorithms to an initial seed,

- True Random Number Generators (TRNG) that apply a numerical processing merged to a noise that come from the "real world", like thermal noise.

Typically, in testing of PRNGs, classical statistical tests are combined with thorough theoretical analysis of the cryptographic properties of the underlying algorithm [14]. On the other hand, TRNGs are more tricky to evaluate. So, the standard AIS 31 [15] has defined a way to qualify the expected quality of TRNG.

It classifies TRNG in two classes P1 and P2 regarding their intended use:

- P1 is the class of TRNG that will be used as nonce generators for authentication protocols,
- P2 embodies the class of "strong" TRNG that might be used as key generators

Notice that in contrast to testing of TRNG within the class P1, where statistical tests may be applied directly to the output of the generator, for strong TRNG, AIS 31 requires extra tests of the noise source. Testing the noise source aims at evaluating its intrinsic entropy, namely the degree of uncertainty that relies in the underlying physical phenomenon. The entropy of the noise source is a very delicate notion and nowadays there are still vast discussions regarding the best way to quantify and qualify it (see [19] for detail). In [11] Ueli Maurer has proposed the famous test based on a statistic asymptotically related to the source entropy. Roughly speaking, this test consists in counting the distances between patterns in the output data stream. In [5], J-S Coron and D. Naccache have proposed to modify this test in order to fit more precisely the source entropy (see also [6] for the latest version of this test which is now part of [15]). For industrial applications related to building new noise sources, failing the Maurer's test means failing AIS 31 certification. This means that applications and markets requiring AIS 31 certification are no longer accessible for suppliers whose devices did not succeed in this certification scheme.

In this paper, we will see how Maurer's test and its counterparts behave in the presence of specific statistical defects in the random source, how Maurer's test can be defeated and last but not least how it can be improved. We would like to stress in this context that essential applications of statistical tests are related to testing of physical random number generators.

1.2 Statistical backgrounds of RNG testing

From the mathematical viewpoint, the problem of testing of a random bit generator can be easily stated. Let \mathbb{B}^n be the set of all n -bit vectors $\mathbf{b} = (b_1, \dots, b_n)$. The distribution of a random vector $\mathbf{b} \in \mathbb{B}^n$ is described by a discrete distribution $p(\cdot)$ on \mathbb{B}^n

$$p(\mathbf{x}) = \mathbf{P}(b_1 = x_1, \dots, b_n = x_n). \quad (1)$$

Recall that the random vector \mathbf{b} is said uniformly distributed if $p(\cdot)$ equals the uniform distribution

$$p(\mathbf{x}) = \mu(\mathbf{x}) \stackrel{\text{def}}{=} \left(\frac{1}{2}\right)^n, \quad \mathbf{x} \in \mathbb{B}^n. \quad (2)$$

With these notations, the problem of testing of a random bit generator can be formulated as follows. Suppose we are given a random bit vector \mathbf{b} distributed according an unknown law p . Then on the basis of \mathbf{b} we want to test the null hypothesis

$$H_0 : \quad p(\mathbf{x}) = \mu(\mathbf{x}) \quad \text{for all } \mathbf{x} \in \mathbb{B}^n$$

against the composite alternative

$$H_1 : p(\mathbf{x}) \neq \mu(\mathbf{x}) \quad \text{for some } \mathbf{x} \in \mathbb{B}^n.$$

In other words, we want to decide whether \mathbf{b} is uniformly distributed on \mathbb{B}^n or not. Our decision can be viewed as a measurable function $\varphi(\mathbf{b})$ (called *critical function*) taking two values $\{0, 1\}$. If $\varphi(\mathbf{b}) = 0$, then H_0 is accepted, otherwise H_1 is accepted. Usually the quality of testing is measured by two types of error probabilities: the probability of the first kind error

$$\alpha(\varphi) \stackrel{\text{def}}{=} \mathbf{P}_0(\varphi(\mathbf{b}) = 1),$$

where $\mathbf{P}_0(\cdot)$ is the probability measure corresponding to the uniform measure μ , and the probability of the second kind error

$$\beta(\varphi, \mathbf{P}) \stackrel{\text{def}}{=} \mathbf{P}(\varphi(\mathbf{b}) = 0).$$

Here $\mathbf{P}(\cdot)$ is any probability measure different from the uniform distribution. The value $1 - \beta(\varphi, \mathbf{P})$ is called *power of test*. Statistical sense of $\alpha(\varphi)$ is very transparent, since this is the probability to reject a good RNG. In contrast to classical statistical testing, where $\alpha(\varphi)$ varies typically from 0.01 to 0.05, in cryptographic applications, we deal with smaller probabilities of the first kind error residing in the range $(10^{-7}, 10^{-3})$. Usually this error probability is fixed regarding the losses which we shall have rejecting H_0 . For instance, in nature, there exist chaotic processes such as the thermal noise in a transistor, and it is a difficult engineering task to design an electronic circuit that exploits this randomness. So, rejection of a good generator might be very expensive. On the other hand, with very small $\alpha(\varphi)$ we can accept bad generators. Therefore a reasonable choice of the probability of the first kind error is a delicate issue (compare [12] and [13]).

From mathematics viewpoint, fixing α , we define the set of statistical tests

$$\Phi_\alpha = \{\varphi : \alpha(\varphi) \leq \alpha\}.$$

and the main goal of statistical testing is to find the most powerful test φ^* within the class Φ_α . In other words, we are looking for the test φ^* such that

$$\beta(\varphi^*, \mathbf{P}) \leq \beta(\varphi, \mathbf{P}) \quad \text{for all } \varphi \in \Phi_\alpha \text{ and for all } \mathbf{P} \neq \mathbf{P}_0.$$

It is easy to see that when the alternative contains all probability distributions, the most powerful test doesn't exist and any attempt to use directly maximum likelihood or Bayesian tests will immediately fail. This happens because we cannot recover the underlying probability distribution on the basis of the data at hand when the set of alternative is too rich. Therefore the basic idea to overcome this difficulty is to consider a smaller alternative family \mathcal{P} satisfying the following properties

- the probability distributions within \mathcal{P} can be recovered with a sufficiently high accuracy for large n
- the maximum likelihood test

$$\varphi_{ML}(\mathbf{b}, \mathcal{P}) = \mathbf{1} \left\{ \max_{P \in \mathcal{P}} \frac{p(\mathbf{b})}{\mu(\mathbf{b})} > t_\alpha \right\} \quad (3)$$

is feasible from numerical complexity viewpoint.

Recall that the critical value t_α is defined by

$$t_\alpha = \inf \left\{ t > 0 : \mathbf{P}_0 \left(\max_{P \in \mathcal{P}} \frac{p(\mathbf{b})}{\mu(\mathbf{b})} > t \right) \leq \alpha \right\}.$$

In order to shed some light on typical problems related to this approach, let us look at the classical frequency test. To construct this test, assume that a RNG generates independent identically distributed blocks $B_i = (b_{1i}, \dots, b_{di})$ containing d bits. Our goal is to check whether B_i are uniformly distributed in \mathbb{B}^d or not. If we associate with the block B_i the integer

$$x_i = \sum_{k=1}^d 2^{k-1} b_{ki},$$

our problem is reduced to the simplest goodness of fit testing: based on the sample $\mathbf{x} = (x_1, \dots, x_N)$ of i.i.d. random variables to test the null hypothesis

$$H_0 : \quad \mathbf{P}(x_i = l) = 2^{-d} \quad \text{for all } l \in \{0, \dots, 2^d - 1\}$$

against the alternative

$$H_1 : \quad \mathbf{P}(x_i = l) \neq 2^{-d} \quad \text{for some } l \in \{0, \dots, 2^d - 1\}.$$

It is easy to check with a simple algebra, that the maximum likelihood test has the following form

$$\varphi_{ML}^d(\mathbf{x}) = \mathbf{1} \left\{ N \sum_{s=0}^{2^d-1} \hat{p}_s(\mathbf{x}) \log \frac{\hat{p}_s(\mathbf{x})}{2^{-d}} > t_\alpha \right\}, \quad (4)$$

where

$$\hat{p}_s = \frac{1}{N} \sum_{i=1}^N \mathbf{1}(x_i = s)$$

is the empirical distribution.

At the first glance everything goes smoothly with this test, but our approach has a serious drawback related to the fact that a priori it was assumed that the RNG generates independent blocks of length d . In fact, there is no reasonable argument justifying this hypothesis. For instance a RNG may work according to a Markov model. In this case, simple simulations reveal that the power of the test depends strongly on d and on the underlying Markov model, and fitting d , we can improve significantly the performance of the test. Since the statistical model of the RNG is hardly known in practice, we should choose the block length based on the data at hand. From statistical viewpoint it means that (4) doesn't really define the statistical test but the family of statistical tests. In order to construct a good test, we have to pick it up within this family on the basis of the observations. Some approaches to this very delicate statistical problem called *multiple testing problem* are discussed in Section 5. Notice that if the statistical model of the RNG were known, then we could easily find the best test within the given family, but the goal of multiple testing is to make this choice without a priori information about the RNG.

In this paper, we are interested in the question "how could statistical tests be improved with the proper choice of the generating alternative family \mathcal{P} " (see (3)). In particular, we will discuss simple methods for improving Maurer's test and finally we will compare numerically this test with a test based on distribution of 1 - spacings in the data flow.

2 Maurer's test

2.1 Uniformity tests

Standard motivations of Maurer's test are related to the notion of entropy of ergodic bit flow (see [11]). In this paper, we present a slightly different viewpoint based on classical uniformity tests. This new interpretation will help us to understand why Maurer's test could be improved. Testing of uniformity means the following. Let $\mu(x) = \mathbf{1}$, $x \in [0, 1]$ be the uniform probability density on the interval $[0, 1]$. Suppose we observe n i.i.d. random variables $\mathbf{X}^n = (X_1, \dots, X_n)$ with an unknown probability density $p(x)$, $x \in [0, 1]$. The goal of the uniformity testing is to test on the basis of \mathbf{X}^n the null hypothesis

$$H_0 : p(x) = \mu(x) \quad \text{for all } x \in [0, 1]$$

against the composite alternative

$$H_1 : p(x) \neq \mu(x) \quad \text{for some } x \in [0, 1].$$

In statistics, the most powerful tests are usually constructed with the help of the maximum likelihood principle which can be motivated by the famous Neyman-Pearson lemma. In order to explain how this principle works in our setting, let us assume for a moment that H_1 is a simple alternative, say $H_1 : p(x) = p_1(x)$, where $p_1(x)$ is a known smooth probability density on $[0, 1]$. In this case, in view of the Neyman-Pearson lemma the maximum likelihood test defined by

$$\varphi(\mathbf{X}^n) = \mathbf{1}\{L(\mathbf{X}^n) \geq h_\alpha\},$$

where

$$L(\mathbf{X}^n) = \sum_{i=1}^n \log p_1(X_i), \tag{5}$$

is the most powerful test. Recall also that the critical value of the test h_α is computed as a root of the equation $\mathbf{P}_0(\varphi(\mathbf{X}^n) = 1) = \alpha$.

Let's now return back to the composite alternative when the density p is unknown. A simple heuristic idea to overcome this difficulty is to construct a non parametric density estimator $\hat{p}(x, \mathbf{X}^n)$ and then to plug-in it in (5). Thus we arrive at the following test statistics

$$S(\mathbf{X}^n) = \sum_{i=1}^n \log \hat{p}(X_i, \mathbf{X}^n)$$

and the principal issue is to find a reasonable density estimator. Standard methods of nonparametric density estimation are motivated by the definition of probability density

$$p(x) = \lim_{h \rightarrow 0} \frac{\mathbf{P}\{X_1 \in [x, x + h]\}}{h}.$$

Roughly speaking, the above formula says that for all sufficiently small h

$$p(x) \approx \frac{\mathbf{P}\{X_1 \in [x, x + h]\}}{h}.$$

Estimating $\mathbf{P}\{X_1 \in [x, x+h]\}$ in the above display by the empirical probability $n^{-1} \sum_{i=1}^n \mathbf{1}\{X_i \in [x, x+h]\}$, we get the classical kernel density estimator

$$\hat{p}_h(X_j, \mathbf{X}^n) = \frac{1}{nh} \sum_{i=1}^n \mathbf{1}\{X_i \in [X_j, X_j+h]\} = \frac{\#\{X_i \in [X_j, X_j+h]\}}{nh}.$$

Our final step is based on the fact that the bandwidth h in this formula might be data-dependent $h = h(X_j, \mathbf{X}^n)$. For instance, one can take

$$h = h(X_{(j)}, \mathbf{X}^n) = X_{(j+m)} - X_{(j)},$$

where $X_{(k)}$ stay for the order statistics $X_{(1)} \leq \dots \leq X_{(n)}$. The increments $X_{(j+m)} - X_{(j)}$, $j = 1, \dots, n-m$ are called m -spacings. Thus we get the following m -spacing density estimator

$$\hat{p}_m(X_{(j)}, \mathbf{X}^n) = \frac{m}{n[X_{(j+m)} - X_{(j)}]}$$

or equivalently, the test statistic

$$S_m(\mathbf{X}^n) = \sum_{i=1}^n \log \frac{m}{n[X_{(j+m)} - X_{(j)}]}.$$

Certainly, the idea to use this statistics is well known and widely used in goodness of fit testing (see e. g. [18], [7], [20]). In order to shed some light on statistical properties of $S_m(\mathbf{X}^n)$, it is very instructive to look at its limit distribution under the alternative. The simplest way to do this is to apply the famous Pyke's theorem [17] about the distribution of order statistics.

Theorem 1 *Let U_1, \dots, U_n be i.i.d. uniformly distributed on $[0, 1]$ and e_1, \dots, e_n be i.i.d. standard exponentially distributed random variables. Then*

$$\left\{ U_{(k+1)} - U_{(k)}, 1 \leq k \leq n-1 \right\} \stackrel{D}{=} \left\{ e_k / \sum_{s=1}^n e_s, 1 \leq k \leq n-1 \right\}. \quad (6)$$

With this theorem, one can find the limit distribution of $S_m(\mathbf{X}^n)$. Unfortunately, the rigorous argument involve a lot of technical details, therefore we provide here only a simple heuristic motivation. Since $F(X_k) = U_k$, where F is the distribution function of X_1 , we have by the Taylor formula

$$U_{(j+m)} - U_{(j)} = F[X_{(j+m)}] - F[X_{(j)}] \approx p(X_{(j)})[X_{(j+m)} - X_{(j)}].$$

This yields the following asymptotic ($n \rightarrow \infty$) formula for the test statistics

$$\begin{aligned} S_m(\mathbf{X}^n) &\approx \sum_{i=1}^n \log p(X_i) - \sum_{i=1}^n \log \frac{n[U_{(j+m)} - U_{(j)}]}{m} \\ &= -nH(p) + \sqrt{n} \frac{1}{\sqrt{n}} \sum_{i=1}^n [\log p(X_i) - \mathbf{E}_p \log p(X_i)] \\ &\quad + nC(m) + \sqrt{n} \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[\log \left(\frac{1}{m} \sum_{l=0}^{m-1} e_{i+l} \right) - C(m) \right] \\ &\quad + n \log \left[1 + \frac{1}{n} \sum_{i=1}^n (e_i - 1) \right], \end{aligned} \quad (7)$$

where $H(p)$ is the entropy

$$H(p) = - \int_0^1 p(x) \log p(x) dx \quad \text{and} \quad C(m) = \mathbf{E} \log \left(\frac{1}{m} \sum_{i=1}^m e_i \right).$$

Notice also that the last term at the right-hand side of (7) can be simplified by the Taylor formula

$$n \log \left[1 + \frac{1}{n} \sum_{i=1}^n (e_i - 1) \right] \approx \sqrt{n} \frac{1}{\sqrt{n}} \sum_{i=1}^n (e_i - 1).$$

Even a quick look at (7) shows that

$$\lim_{n \rightarrow \infty} \frac{S_m(\mathbf{X}^n)}{n} \stackrel{a.s.}{=} -H(p) + C(m).$$

Moreover, it is also well known (see e.g. [10]) that if $p(x)$ is strictly bounded for below on $[0, 1]$, then $S_m(\mathbf{X}^n)$ is asymptotically Gaussian

$$\lim_{n \rightarrow \infty} \frac{S_m(\mathbf{X}^n) - nH(p) - nC(m)}{\sqrt{n}} \stackrel{D}{=} \mathcal{N}\left(0, \sigma^2(p) + \sigma_m^2\right),$$

where the asymptotic variance of this Gaussian law is defined by (see also [4])

$$\sigma_m^2(p) = \int_0^1 [\log p(x) + H(p)]^2 p(x) dx, \quad \sigma_m^2 = (2m^2 - 2m + 1)\psi'(m) - 2m + 1,$$

with

$$\psi'(m) = \frac{\pi^2}{6} - \sum_{j=1}^{m-1} \frac{1}{j^2}.$$

From the plot of σ_m^2 shown on Figure 1, we see that this function vanishes very rapidly. It means that with moderate m we could improve the performance of testing. In order to explain this phenomenon, notice that under the hypothesis, for large n

$$S_m(\mathbf{X}^n) \sim \mathcal{N}(nC(m), n\sigma_m^2).$$

Therefore, for a sufficiently small α , the critical value can be computed by

$$h_\alpha \approx nC(m) + \sqrt{2n\sigma_m^2 \log(1/\alpha)}$$

and if the entropy is large enough

$$H(p) \gg \sqrt{\frac{2\sigma_1^2 \log(1/\alpha)}{n}},$$

then the probability of the second kind error is given by

$$\log \mathbf{P}\left(S_m(\mathbf{X}^n) \leq h_\alpha\right) \approx - \frac{\left[\sqrt{n}H(p) + \sqrt{2\sigma_m^2 \log(1/\alpha)}\right]^2}{2[\sigma^2(p) + \sigma_m^2]}. \quad (8)$$

It is easy to check that the right-hand side in (8) is monotone in σ_m^2 . Therefore the probability of the second kind error reaches its minimum when $m = \infty$. However, since σ_m^2 vanishes rapidly, we can get almost the minimal error probability with a relatively small m . In fact, the optimal choice of m should be data-driven and we discuss some approaches to this problem in Section 5.

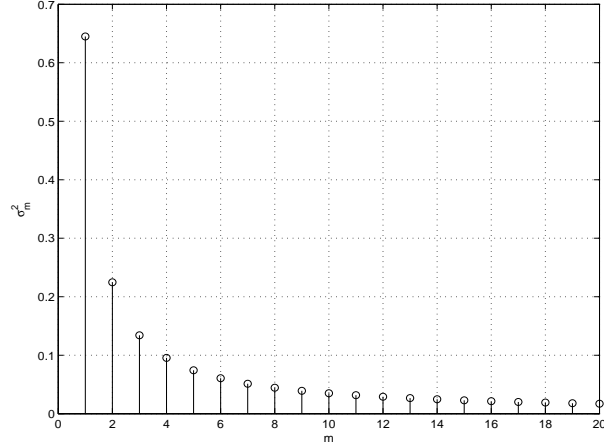


Figure 1: The variance σ_m^2

2.2 Maurer's test

This test has the standard form $\varphi^{ma}(\mathbf{b}) = \mathbf{1}(T(\mathbf{b}) \geq t_\alpha)$, where the test's statistic $T(\mathbf{b})$ is computed by the following principal steps:

1. Transform the input bit sequence $\mathbf{b} = (b_1, \dots, b_n)$ into the sequence of integers $\mathbf{x} = (x_1, \dots, x_s)$, $s = \lfloor n/d \rfloor$ taking values in $\mathbb{A}^d = \{0, \dots, 2^d - 1\}$

$$x_k = \sum_{i=1}^d 2^{i-1} b_{(k-1)d+i}$$

2. For each motif $q \in \mathbb{A}^d$ compute its positions in \mathbf{x} :

$$N^q = \{k : x_k = q\}.$$

3. For each motif $q \in \mathbb{A}^d$ compute the intermediate statistics

$$S^q(\mathbf{b}) = - \sum_i \log(N_{i+1}^q - N_i^q).$$

4. Compute the final test's statistic

$$T(\mathbf{b}) = \sum_{q \in \mathbb{A}^d} S^q(\mathbf{b}).$$

In order to describe completely the test, remember that the critical value t_α is defined as a root of the equation

$$\mathbf{P}_0(\varphi^{ma}(\mathbf{b}) = 1) = \alpha. \quad (9)$$

There are two standard ways to compute t_α

- compute the empirical distribution function of $T(\mathbf{b})$ by the Monte-Carlo method and solve the empirical counterpart of (9)

- use the fact that asymptotically ($n \rightarrow \infty$) the distribution of $T(\mathbf{b})$ is Gaussian.

We intentionally decomposed Maurer’s test into 4 steps in order to stress its relations with the uniformity testing. From the viewpoint of the uniformity testing, the underlying ideas of Maurer’s test are related to steps 3 and 4 that clearly show what does the test do: it checks whether the positions of patterns are uniformly distributed in the bit stream. Mathematically, this principal idea is based on the assumption that all x_k are independent (see [11], [6]). In other words, this means that d should be large. This hypothesis immediately entails that

- $N_{i+1}^q - N_i^q$ are almost independent and follow an exponential law under the null hypothesis and under the alternative (step 3)
- under the null hypothesis and under the alternative, the covariance matrix

$$r_{pq} = \mathbf{cov}(S^q(\mathbf{b}), S^p(\mathbf{b})), \quad 0 \leq p, q \leq 2^d - 1$$

has always the form

$$r_{pq} \approx \begin{cases} 1, & p = q \\ c, & p \neq q, \end{cases}$$

where c is a constant (step 4).

Under the null hypothesis all these assumptions hold true, but unfortunately, they may fail for alternatives related to stationary ergodic processes. The reason is that we cannot take d very large, since there is a natural upper bound $d \leq \log_2(n/10)$ (see [11]). Therefore in practical applications, the cornerstone hypothesis that d is really large, is not well justified. It is surprising that this fact opens some perspectives for significant improvements of Maurer’s test.

In order to illustrate numerically statistical phenomena in this paper, we shall use two statistical models for random bit generators. The first one called *Markov chain model* (see also [19]) works as follows. Let ξ_i be i.i.d. bits such that $\mathbf{P}(\xi = 0) = p$, then the random bits are generated as follows

$$b_i = \begin{cases} b_{i-m} & \text{if } \xi_i = 0 \\ 1 - b_{i-m} & \text{otherwise,} \end{cases}$$

where $m \geq 1$ is called memory of the chain and $p \in [0, 1]$ is called transition probability. In all our numerical experiments, the length of the bit vector is $n = 20000$ and the probability of the first kind error is 0.001. We use these basic simulation parameters from now on.

Another statistical mechanism for random bit generation is called *season drift model*. In this case the bits b_i are independent but not identically distributed. Namely, it assumes that

$$\mathbf{P}(b_i = 0) = (0.5 + A) \cos\left(\frac{2\pi i}{\tau}\right),$$

where $A \in [0.5, 1]$ is called the amplitude and τ is called *season period*.

First of all let us look at the covariance matrix of the intermediate test’s statistic. Figure 2 illustrates the fact that this covariance matrix may substantially differ under the null hypothesis and the alternative. The left panel of this figure shows the covariance matrix under the null hypothesis whereas the right panel shows this matrix under an alternative. As an alternative we used the Markov chain with memory 5 and transition probability 0.7.

We start to analyze Maurer’s test with the question whether the log-function in $\log(N_{i+1}^q - N_i^q)$ is good. At the first glance the answer is negative since log results from the hypothesis

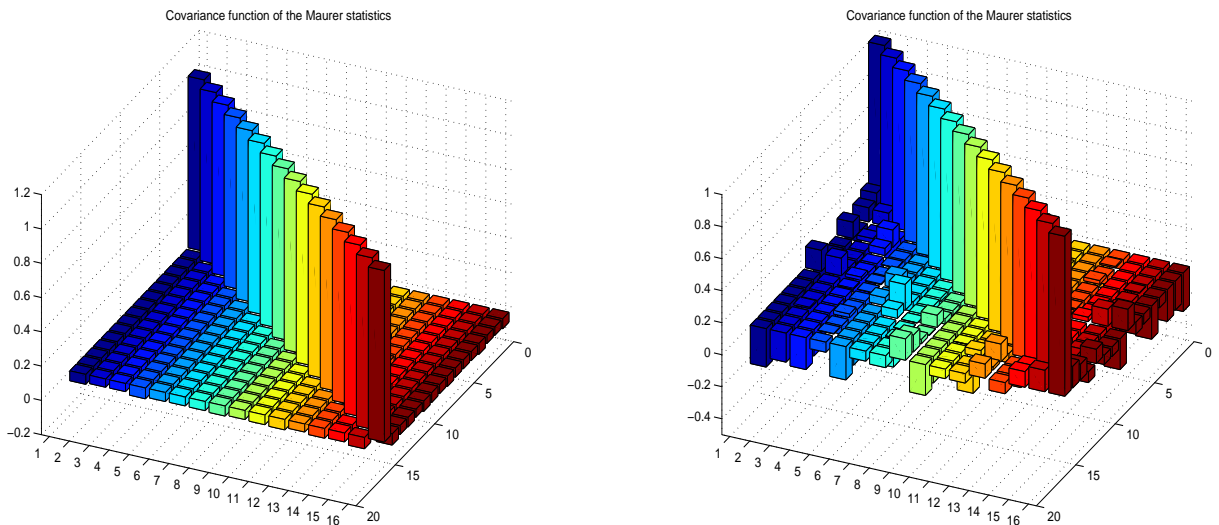


Figure 2: The covariance matrix under the hypothesis and the alternative.

that $N_{i+1}^q - N_i^q$ are exponentially distributed. In fact, under the null hypothesis, these random variables follow a geometric law. It means that if the only one intermediate test's statistic, say S^0 , was used, then $\log(x)$ should be replaced by $l(x) = x \log(x) - (x-1) \log(x-1)$. This fact doesn't mean that $l(x)$ is optimal in the case when we are dealing with the sum of S^q . This curious phenomenon can be explained by the correlation between statistics S^q , $q \in \mathcal{A}^d$. Moreover, $\log(x)$ is not optimal too and we illustrate this fact in the following statistical experiment. On Figure 3 we plotted the probability of the second kind error as function of the transition probability for standard Maurer's test (dotted line) and for Maurer's test with $\log(N_{i+1}^q - N_i^q + 10)$ (solid line) for 6-tuples ($d = 6$) partition. The bit vectors were generated by the Markov chain model with memory 1. As we see that there is a slight improvement of Maurer's test. In other numerical

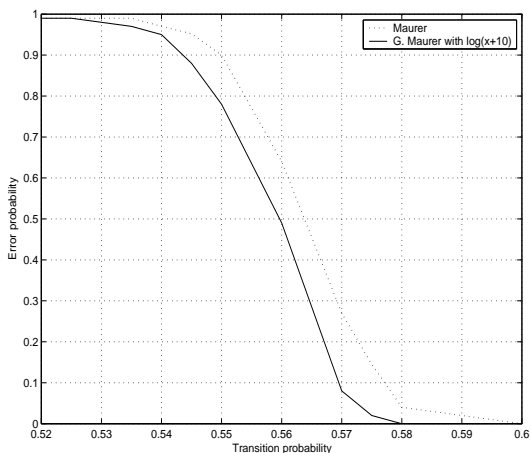


Figure 3: Maurer's test with $\log(x + 10)$

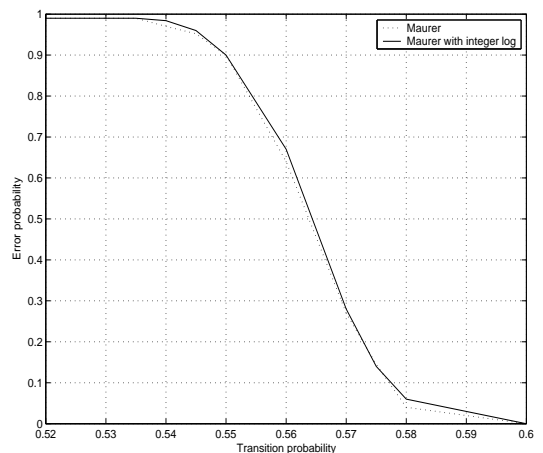


Figure 4: Coron's modification

simulation the authors looked at, the function $\log(x + 10)$ always improves the power of the test but it seems to us that improvements are not very significant and therefore in practical

applications $\log(x)$ may be considered as a reasonable choice. Another idea to use $\sum_{i=1}^{x-1} i^{-1}$ instead of $\log(x)$, was proposed in [6] and [5]. Unfortunately in our simulation study, this method doesn't result in visible improvements of the test's power. Typical behavior of Coron's test and Maurer's test are shown on Figure 4. Here we used the statistical model for the random bit source from the previous example.

Let us discuss another problem related to the step 3. Namely, the optimality of the first order spacings $N_{i+1}^q - N_i^q$. For the uniformity testing problem, we have seen that m -spacings may improve the test power. Similar effect takes place for Maurer's test, it turns out that using m -spacings $N_{i+m}^q - N_i^q$ with $m > 1$ it is possible to improve the power of this test.

The next natural question related to the final step 4 is: whether the sum of $S^q(\mathbf{b})$ is a good idea for testing or not? This statistics would be optimal if the covariance matrix of the vector $(S^1(\mathbf{b}), \dots, S^{2^d}(\mathbf{b}))$ doesn't change its form under the alternative. Unfortunately, we have seen see that it isn't true. This phenomenon opens another way to improve Maurer's test.

3 Motif Uniformity test

The term *Motif Uniformity* (MU) test is referred to Maurer's test with the following modifications:

- in place of $S^q(\mathbf{b})$ computed at the step 3, we use m -spacing statistics

$$S^{q,m}(\mathbf{b}) = - \sum_i \log(N_{i+m}^q - N_i^q)$$

- the final test statistics $\sum_{q \in \mathbb{A}^d} S^q(\mathbf{b})$ computed at the step 4 is replaced by a special non-linear transform based on p -leave out method.

3.1 m -spacing method

In this section, we present two examples showing that m -spacing technique improves the power of Maurer's test. Figure 5 shows the power of Maurer's test (dotted line) and the power of its modification based on 4-spacings (solid line) as function of the transition probability. We see that the improvement is clear. The next example demonstrates a more significant improvement of Maurer's test. In this example, we deal with the season drift model with $T = 3000$ and plot the power of the tests as function of amplitude A . For this random bit model Maurer's test with $d = 1$ is, in some sense, optimal. Since p_k is a very smooth function of k , m -spacings with large m may improve substantially the test power. Figure 6 distinctly illustrates this fact.

3.2 L -leave out method

We have seen that the covariance matrix of Maurer's test statistics $S^q(\mathbf{b})$, $q \in \mathcal{A}^d$ may change substantially under the alternative and now we use this phenomenon to improve the performance of this test. The underlying idea is very simple. We order the test statistics $(S^1(\mathbf{b}), \dots, S^{2^d}(\mathbf{b}))$ such that

$$S^{(1)}(\mathbf{b}) \geq S^{(2)}(\mathbf{b}) \geq \dots \geq S^{(2^d)}(\mathbf{b})$$

and compute the final test statistics

$$T^L(\mathbf{b}) = \sum_{i=1}^{2^d-L} S^{(i)}(\mathbf{b}).$$

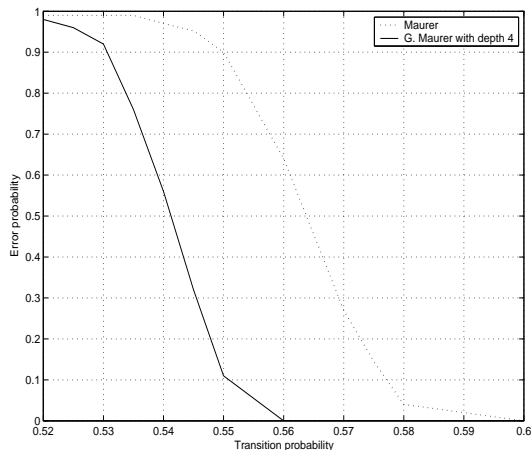


Figure 5: The MU test with 4-spacings

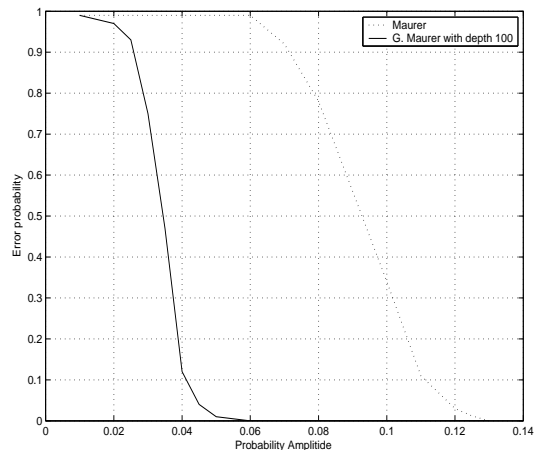


Figure 6: The MU test with 100-spacings

Figure 7 illustrates improvements in the test power based on 4 - leave out technique. Here we plotted the probability of the second kind error as function of transition probability for 1-memory Markov chain.

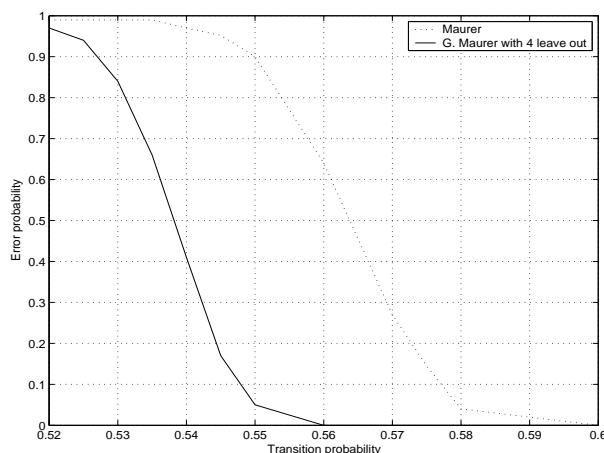


Figure 7: Performance of Maurer's test with 4 leave out statistics

4 Spacings distribution test

In contrast to Maurer's test which checks whether the positions of motifs in the input vector are uniformly distributed or not, the goal in the Spacing Distribution (SD) test is to compare the empirical distribution of 1-spacings with a geometric distribution. For typical alternatives, the distribution of $N_{i+1}^q - N_i^q$ may be very far from geometric thus providing an additional and significant statistical information about RNG. From the statistical viewpoint, we can retrieve this information testing the hypothesis that the law of $N_{i+1}^q - N_i^q$ is geometric. In this section, we propose to use the maximum likelihood method to test this hypothesis.

Remember that under the null hypothesis 1-spacing follows a geometric law

$$\mathbf{P}_0\left(N_{i+1}^q - N_i^q = k\right) = p_0(k) = \frac{1}{2^d - 1} \left(1 - 2^{-d}\right)^k.$$

We define the SD test as the maximum likelihood test assuming that for a given q the spacings $N_{i+1}^q - N_i^q$ are i.i.d. This test consists in the following steps

- For all motifs q compute the empirical distribution of $N_{i+1}^q - N_i^q$

$$\widehat{p}^q(k) = \frac{1}{\#N^q} \sum_{i=1}^{\#N^q} \mathbf{1}\left(N_{i+1}^q - N_i^q = k\right)$$

and compute the intermediate test statistics

$$S^q(\mathbf{b}) = \sum_{i=1}^{\#N^q} \log \frac{\widehat{p}^q(N_{i+1}^q - N_i^q)}{p_0(N_{i+1}^q - N_i^q)}.$$

- Compute the critical function

$$\varphi_{SD}(\mathbf{b}) = \mathbf{1}\left(\sum_{q=0}^{2^d-1} S^q(\mathbf{b}) \geq h_\alpha\right).$$

In some sense, the SD test can be viewed as a very good complementary of Maurer's test since this test is very stable and powerful for the Markov chains alternatives. Figure 8 illustrates this fact. In this numerical experiment we try to find out how the powers of Maurer's test and the SD test depend on the parameters of the Markov chain alternative. On left panel we plotted the power of Maurer's test with $d = 7$ as function of the memory of the chain varying from 1 to 20 and the transition probability belonging to $[0.5, 0.8]$. The left panel represents the power of SD test for the same alternatives.

This figure distinctly shows the principle differences Maurer's and SD test. First of all, Maurer's test detects very badly alternatives with memories greater than d . This is the principle drawback of the test since the block length d cannot be large. We have already mentioned that $d \leq \log_2(n/10)$, where n is the length of the bit flow at hand, otherwise the test statistics $S^q(\mathbf{b})$ may have no sense. On the other hand, Maurer's test with large d may detect badly the Markov alternatives with short memories. Therefore, it seems to us that Maurer's test with a priori fixed large d is not good for practical implementations. The only way to overcome this difficulty of Maurer's test is to use a multiple testing approach which will be discussed in the next section.

Fortunately, these drawbacks are not inherent to SD test. Even with small d this test can detect the Markov chains with large memories. However, we would like to stress that the SD test should not be used as an universal test. It not surprising for instance that its power may be low for season drift models. In this case, the 1-spacing follows a geometric law and from the viewpoint of the SD test, there is no big difference between the hypothesis and the alternative.

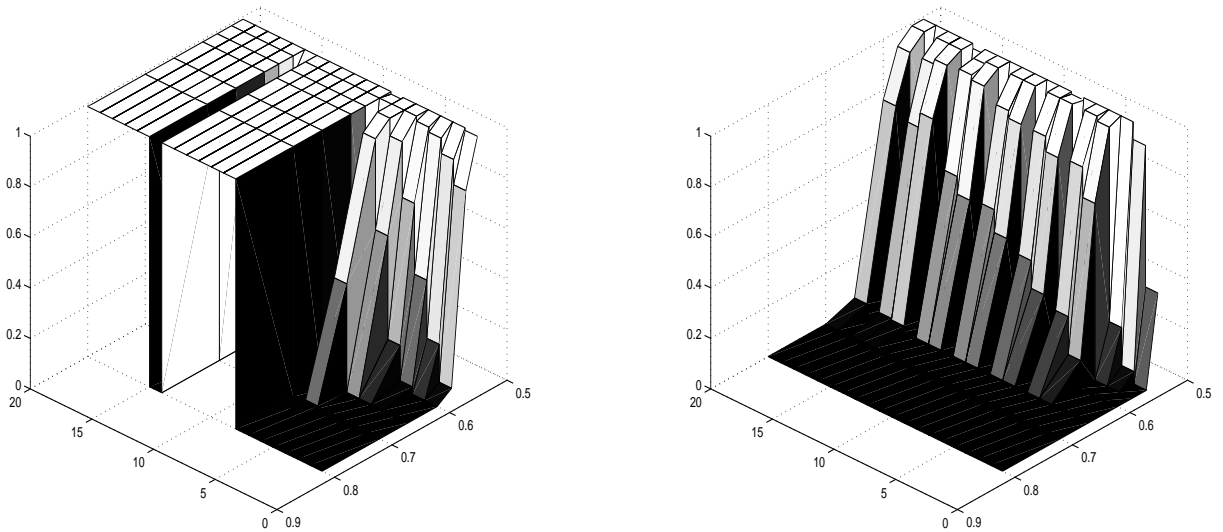


Figure 8: Statistical performance of Maurer's and SD tests.

5 Multiple testing

The literature on multiple testing is so vast that it would be impractical to cite it here, we refer the interested reader to [1] and [9] for the modern state of art of this old issue and further references. In this section, we present only a heuristic approach to this problem. Suppose we are given a family of N statistical tests

$$\mathbb{T} = \left\{ \mathbf{1}(T_1(\mathbf{b}) - t_1 > 0), \dots, \mathbf{1}(T_N(\mathbf{b}) - t_N > 0) \right\}.$$

The test statistics here might be thought as empirical versions of pseudo-distances between a probability measure $p(\cdot)$ and the uniform distribution $\mu(\cdot)$ on \mathbb{B}^n . For instance, one could have in mind the family of Maurer's tests with different $d = 1, \dots, N$. If we decide to use the test's statistic $T_k(\mathbf{b})$, then we should choose the critical value t_k as a root of the following equation (if it exists)

$$\mathbf{P}_0(T_k(\mathbf{b}) > t_k) = \alpha.$$

Then under the alternative, the power of this test is computed by

$$\beta_p(T_k) = \mathbf{P}(T_k(\mathbf{b}) > t_k).$$

Remember that p is the probability distribution of the data under the alternative (see (1)). Suppose for a moment that there is an oracle which provides us with p . Then we can easily find the best statistical test within \mathbb{T} . Obviously, we have to choose the test $\mathbf{1}(T_{k^*}(\mathbf{b}) - t_{k^*} > 0)$ with

$$k^* = k_p = \arg \max_{k=1, \dots, N} \beta_p(T_k).$$

In other words, with the help of the oracle we can easily construct the test with the maximal power. The goal of the multiple testing is to find a data-driven method of choosing a test within \mathbb{T} which mimics the oracle method.

A standard solution of this problem is based on idea of *tests intersection*. This method works as follows. Let $\bar{t}_k(\alpha)$ be a sequence of reals such that

$$\mathbf{P}_0\left(\max_{k=1,\dots,N}[T_k(\mathbf{b}) - \bar{t}_k(\alpha)] > 0\right) \leq \alpha \quad (10)$$

We shall call $t_k(\alpha)$ α -envelope of the family \mathbb{T} . This definition of $\bar{t}_k(\alpha)$ evidently entails that for any data-driven test choice $k(\mathbf{b})$

$$\mathbf{P}_0\left([T_{k(\mathbf{b})}(\mathbf{b}) - \bar{t}_{k(\mathbf{b})}(\alpha)] > 0\right) \leq \alpha.$$

If so, we can define the new test by

$$\hat{k}(\mathbf{b}) = \arg \max_{k=1,\dots,N}[T_k(\mathbf{b}) - \bar{t}_k(\alpha)].$$

In other words, we accept H_1 when $\max_{k=1,\dots,N}[T_k(\mathbf{b}) - \bar{t}_k(\alpha)] > 0$.

Intuitively, in order to construct the most powerful test, we should find the smallest α -envelope. So we arrive at a very delicate optimization problem of computing the smallest $\bar{t}_k(\alpha)$ for which (10) holds true.

Probably, the first idea of construction of an α -envelope was proposed by Bonferroni [2], [3]. Notice that

$$\mathbf{P}_0\left(\max_{k=1,\dots,N}[T_k(\mathbf{b}) - \bar{t}_k(\alpha)] > 0\right) \leq \sum_{i=1}^N \mathbf{P}_0\left(T_i(\mathbf{b}) - \bar{t}_i(\alpha) > 0\right).$$

Therefore if we take $\bar{t}_i^{bon}(\alpha)$ as a root of equation

$$\mathbf{P}_0\left(T_i(\mathbf{b}) - \bar{t}_i^{bon}(\alpha) > 0\right) = \frac{\alpha}{N},$$

then we get an α -envelope of \mathbb{T} .

The Bonferroni method is good only when N is small enough and the test's statistics $T_i(\mathbf{b})$, $i = 1, \dots, N$ are almost independent. For instance, this method may be used for multiple testing with Maurer's tests family with different d . On the other hand, we cannot hope that this idea results in a good test for multiple testing with MU tests having different spacing orders, since in this case the statistics are strongly dependent.

6 Concluding remarks

There are two main objectives in the present paper. First of all, we propose a new interpretation of Maurer's test which is based on nonparametric maximum likelihood uniformity tests. This approach explains why and how Maurer's test can be improved, and we provide three methods to improve it

- using m -spacing technique
- L -leave out correction of the test statistics
- spacing distribution test

In numerical examples, we demonstrate that all these methods improve significantly the test's power.

The second objective of the paper is related to the multiple testing approach. The main goal of this method is to find the "best" method within the family of given statistical tests. Roughly speaking, multiple testing provides us with a test which adapts automatically to the random bit source. In our opinion, this idea seems to be very fruitful and might be considered as a cornerstone for future developments of powerful statistical tests.

7 Acknowledgments

We would like to thank Stefan Weggenkitl and Schindler Werner for interesting discussions and remarks.

References

- [1] Benjamini, Y. and Yekutieli, D. (2001). Controlling the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29**, 1165–1188.
- [2] Bonferroni, C. E. (1935). Il calcolo delle assicurazioni su gruppi di teste. In *Studi in Onore del Professore Salvatore Ortu Carboni* Rome: Italy, 13-60.
- [3] Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilit. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze* **8**, 3-62.
- [4] Cressie, N. (1976) On the logarithms of high-order spacings. *Biometrika*, **63**, no. 2, 343-355.
- [5] Coron, J.-S. and Naccache, D. (1998). An Accurate Evaluation of Maurer's Universal Test, in *Proceedings of Selected Areas in Cryptography 98*. LNCS 1556 P57-71. Springer-Verlag.
- [6] Coron, J.S. (1999). On the security of Random sources, in *Public Key Cryptography: Second International Workshop on Practice and Theory in Public Key Cryptography, PKC'99, Kamakura, Japan, March 1999. Proceedings Editors: H. Imai, Y. Zheng (Eds.):* vol. 1560. Lecture Notes in Computer Science, <http://www.gemplus.com/smart/rd/publications>
- [7] Del Pino, G. (1979). On the asymptotic distribution of k -spacings with applications to goodness-of-fit tests. *Ann. Statist.* **8**, 1058–1065.
- [8] Goubin, L. and Patarin, J. (1999) DES and Differential Power Analysis the "Duplication" Method LNCS *Proceedings of CHES'99*
- [9] Finner, H. Rotters, M. (2002). Multiple hypothesis testing and expected number of type 1 errors. *Ann. Statist.* **30**, no 1, 220–238.
- [10] Hall, P. limit theorems for sums of general functions of m -spacings. *Math. Proc. Cambridge Phil. Society*, **96**, 517-532.
- [11] Maurer, U. (1992). A universal statistical test for random bit generators. *J. Cryptology*, **5**, no. 2, 89–105.

- [12] National Institute of Standards and Technology, Security Requirements for Cryptographic Modules, FIPS 140-1, Jan. 1994
<http://www.nist.gov/itl/div897/pubs/fip140-1.htm>.
- [13] National Institute of Standards and Technology, Security Requirements for Cryptographic Modules, FIPS 140-2, May. 2001
<http://csrc.nist.gov/publications/fips/fips140-2/fips1402.pdf>
- [14] Functionality Classes and Evaluation Methodology for Deterministic Random Number Generators, Version 2.0 December 1999.
www.bsi.de/zertifiz/zert/interpr/ais20e.pdf
- [15] Functionality Classes and Evaluation Methodology for True (physical) Random Number Generators, Version 3.1 September 2001.
<http://www.bsi.bund.de/zertifiz/zert/interpr/trngk31e.pdf>
- [16] Oswald, E. and Preneel, B. A Survey on Passive Side Channel Attacks and their Counter Measures for the Nessie Public-Key Cryptosystems.
<https://www.cosic.esat.kuleuven.be/nessie/reports/phase2/kulwp5-027-1.pdf>
- [17] Pyke, R. (1965) Spacings. *J.R. Statist. Soc. B* **27**, 395–436.
- [18] Shilling, M. (1983) Goodness of fit in \mathbb{R}^n based on weighted empirical distributions of certain neighbor statistics *Ann. Statist.* **11**, 1–12.
- [19] Weggenkitl, S. (2001). Entropy estimators and Serial Tests for Ergodic Chains. *IEEE Trans. Inform. Theory.* **47**, no. 6, 2480-2489.
- [20] Weiss, L. (1957) Asymptotic of certain tests of fit based on sample spacings. *Ann. Math. Statist.*, **28**, 783-786.