Extension of Rank Test for Sequences over GF(q)

Ali Kakhbod
(ali_kakhbod@ec.iut.ac.ir, Department .of Electrical Engineering Isfahan University of Technology)
Mohammad Dakhil Alian
(mdalian@cc.iut.ac.ir, Department .of Electrical Engineering Isfahan University of Technology)
Soheil Mohajer
(soheil.mohajer@epfl.ch, School of Computer and Communication Sciences École Polytechnique Fèdèrale De Lausanne)

Abstract

Pseudo-random and random sequences are widely used in many applications, such as stream cipher systems. Statistical tests are usually used to evaluate randomness of sequences. ***Binary Matrix Rank Test*** *is one method to evaluate randomness of sequences [2]. This test is based on constructing matrices whose rows are successive sub-strings of the sequence, and check for linear dependency among the rows or columns of the constructed matrices. In this paper, we represent new method to testing for randomness based on linear dependency among fixed-length substrings of sequences over GF(q).*

1. Introduction

In many applications such as cryptography systems and especially in stream cipher systems, randomness or in better words, how much the output of algorithms look random- has a special significance? Also in many computer simulations, generating random or pseudo-random sequences is highly significant and brings the results of simulations closer to reality. The important question is that how can the amount of randomness of a sequence generated by a source be determined, or in another concept, how can randomness of a source be determined by the sequences it has generated. In this paper we assume an ideal random sequence will be a sequence with independent and uniform distribution elements. There are several methods for comparing sequences generated by different sources with an ideal random sequence. But unfortunately, recognizing an ideal random sequence from another sequence (a pseudo-random sequence) is not a simple task. Therefore different methods are used for this purpose and the most general one is using statistical tests. By subtle mathematical examination of the ideal random sequences, some standards have been stated, for sequences having desired random behavior (pseudo-random sequences). Subsequently and according to the probabilistic modeling of some specifications of random sequences, numerous tests have been introduced for studying the randomness of sequences [2]. The main point of a statistical test is to compare the randomness of the sequence undergoing the test (and subsequently the randomness of the source that has generated it), to a mathematical probabilistic model. Statistical tests are applied to the output of random (or pseudo-random) generators or sources with specific lengths and through that test; the sequence either fails, or passes. To evaluate a source, often a considerably large number of sequences are tested to determine the amount of compatibility between those sequences and an ideal random sequence. It is necessary to mention that if a sequence passes the test, it does not mean that it is random; because in these tests, the general behavior of a sequence is studied in the form of a limited probabilistic model, thus the sequences that match this general behavior, it will be passed. Therefore it is difficult to compare different statistical tests, because each test uses a different concept to model the random behavior of an ideal random sequence. Therefore when a sequence passing one test, it does not mean it's an ideal random sequence, except for universal *tests* that are considered from a theoretical point of view and no practical methods have been mentioned yet.

The ***B****inary* ***M****atrix* ***R****ank (BMR)* test is one of the statistical tests which are based on the probability function of a matrix with uniformly distributed independent binary elements, being full-rank. In the present paper we explain the appropriate approximation of the probability function of a random matrix being full-rank, by stating the probability function of rank of random matrices with elements belonging to GF(q). Afterwards, on that basis, we present a method to upgrade the *BMR* in the general case for matrices with elements belonging to GF(q).

II. The Binary Matrix Rank test (BMR)

The BMR test is based on this question, if we build a matrix with consecutive bits of a binary sequence, with desired numbers of rows and columns, how can we study the dependency between rows and columns? [1]. So we assume that we have binary matrix $m \times q$ with independent elements uniformly distributed (or an ideal random matrix). It can be proved that the rank of this matrix has following probability function,

0-7803-9521-2/06/$20.00 ©2006 IEEE.

$$P(R = r) = P_r = 2^{r(q+m\ \ r)\ \ mq}\prod_{i=0}^{r\ \ 1}\frac{(1 - 2^{i\ \ q})(1 - 2^{i\ \ m})}{1 - 2^{i\ \ r}}$$

$$r = 0,\ \ 1, \ldots, m$$
$$m = Min(m, q)$$

Where R is the rank of the random binary matrix $m\times q$, and P_r is the probability if rank of the matrix is r.

If $m=q$ then $P_m \approx 0.2888$, $P_{m-1} \approx 0.5776$, $P_{m-2} \approx 0.1284$ and if $m \geq 10$ other values of P_r will be very small. To perform the statistical test, if the length of the binary sequence is assumed to be $n=m^2N$ bits, this sequence will be formed into N matrices, $m \times m$ which have ranks $R_1, R_2, \ldots, R_N$. In fact this sequence consists of discrete random variables with probability functions (1). Now if we form N matrices $m \times m$ in the sequence undergoing this test, which has a length of m^2N, and let the number of matrices with a rank of M be F_M and the number of matrices with a rank of $M-1$ be F_{M-1}, then,

$$F_M = \{R_l = M\} \quad l = 1,2,\cdots, N$$
$$F_{M-1} = \{R_l = M-1\}$$

Statistic of the *BMR* test will be,

$$\chi^2_{BMR} = \frac{(F_M - 0.2888N)^2}{0.2888N} + \frac{(F_{M-1} - 0.5776N)^2}{0.5776N} + \frac{(N - F_M - F_{M-1} - 0.13336N)^2}{0.13336N}$$

In limit when $N \to \infty$, χ^2_{BR1} will be a Chi squared random variable with 2 degree of freedom. The calculated χ^2_{BMR} for a sequence provides a basis to determine whether that sequence fails or passes the test.

III. The Probability Function for the Rank of Matrices on GF(q)

If elements of matrix $A_{m\times n}$ belong to GF(q), and these elements are randomly chosen from the elements of GF(q), then it is proved that the probability that the matrix has a rank of r is equal to [4,1],

$$P_r = P_r[rank(A_{m\times n}) = r] = \frac{1}{q^{(m-r)(n-r)}}\prod_{i-0}^{r-1}\frac{(1-q^{i-n})(1-q^{i-m})}{1-q^{i-r}} \quad r = 0,1,\ldots,m \qquad (1)$$

Above expression shows that if r and m ($m < n$) are great, calculating the probability function is not easily possible. Therefore, comparatively good approximation can be used for equation (4), especially for $q > 2$. First, through a lemma we state an approximation for $P_r[rank(A_{m\times n}) = r]$, and then for the special case $m = n$, we calculate a lower bound for this probability and prove that this lower bound tends towards the desired probability. Calculating this bound is very much easier than (1).

Lemma:

In this Lemma we calculate the number of full-rank $\mathrm{m}\times\mathrm{m}$ matrices over GF(q). There are M matrices over GF(q). Where

$$\mathrm{M} = \mathrm{q}^{\mathrm{m}^2} \qquad (2)$$

$\prod_{i-1}^{m-1}(q^m - q^i)$ of these are full rank, Therefore,

$$Pr[\mathrm{A}_{m\times m}\ \text{full rank}] = \prod_{i-1}^{m}(1 - \frac{1}{q^i}). \qquad (3)$$

This probability is a bounded and decreasing function of m. Thus, it will be greater than

$$\lim_{m\to\infty}\prod_{i-1}^{m}(1 - \frac{1}{q^i}) \qquad (4)$$

We denote this lower bound by L. According to convexity of the logarithm function, it is obvious that,

$$\log(1 - x) \geq xq\log(1 - \frac{1}{q}) \qquad (5)$$

for $0 \leq x \leq \frac{1}{q}$

Thus for any integer i, by choosing x=1/q^i, we have

$$\log(1 - \frac{1}{q^i}) \geq \log(1 - \frac{1}{\mathrm{q}})\mathrm{q}^{\mathrm{i}-1} \qquad (6)$$

By using the above inequality, we have

$$\log L = \sum_{i-1}^{m}\log(1 - \frac{1}{q^i}) \geq \sum_{i-1}^{m}\log(1 - \frac{1}{q})q^{i-1}$$
$$\geq \log(1 - \frac{1}{q})\sum_{i-1}^{m}q^{i-1}$$
$$= \frac{q}{q-1}\log(1 - \frac{1}{q}) \qquad (7)$$

Therefore,

$$L \geq (\frac{q-1}{q})^{\frac{q}{q-1}} \tag{8}$$

Considering above lemma, it can be seen that for q > 2 and great values of m, P_m can be calculated with simpler expressions than (4). Also considering (4) and assuming to have the value of P_m, we can also calculate P_{m-1}, P_{m-2}, using P_m. For example,

$$P_m = \prod_{i=1}^{m} \left(1 - \frac{1}{q^i}\right) \tag{9}$$

$$P_{m-1} = \frac{1}{q} \prod_{i=0}^{m-2} \frac{(1-q^{i-m})^2}{(1-q^{i-(m-1)})} = \frac{q}{(q-1)^2} P_m \tag{10}$$

$$P_{m-2} = \frac{1}{q^4} \prod_{i=0}^{m-3} \frac{(1-q^{i-m})^2}{(1-q^{i-(m-2)})} = \left(\frac{q}{(q^2-1)^2(q-1)}\right)^2 P_m \tag{11}$$

If in a particular application, only P_m, P_{m-1} and P_{m-2} are required and the approximation

$$P_m \approx (\frac{q-1}{q})^{\frac{q}{q-1}}$$

is also used, calculations are very easier than (4). In the following paragraph, we present a statistical test, considering that it only uses P_m, P_{m-1} and P_{m-2}.In the following sections first we explain the *Chi Squared* test, and then we try to upgrade the *BMR* test.

IV. The χ^2 Test

Samples of a random variable with a particular probability function have a behavior compatible with that probability function. Now, having a number of samples taken from a random variable, how can it be determined whether or not the mentioned samples are compatible with the probability function assumed for that random variable? Statistical test method is a subject that tries to answer this question. In this test a probability function is assumed for the subject random variable, and then the test determines whether or not this probability function is acceptable for the samples. Generally speaking, if the results of the samples seem to be compatible with the assumed probability function, we tend to accept the assumption and otherwise, we tend to reject it. In practical matters, the case that the probability function is known for a random variable whose samples are available, is tested against the case that the probability function of the samples is not of the assumed type. A usual method to test these cases is the Chi Squared test.

If $(x_1, x_2, ..., x_n)$ is a random vector with parameters n, $p_1, p_2, \ldots, p_m$, in which x_i's each get values of a_j (j = 1,…, m) with the following probability,

$$P(x_i = a_j) = P_j \qquad i = 1, \ldots, n \quad j = 1, \ldots, m$$

Then for enough big n, the following expression will tend towards the Chi Squared distribution with m-1 degree of freedom.

$$T_n(obs) = \sum_{i-1}^{m} \frac{(N_i - np_i)^2}{np_i} \tag{12}$$

Where N_i is the number of times that a_i has appeared in the sample random variable (or observed samples). For acceptable approximation in (12), the following condition must be met,

$$nP_i \geq 5 \qquad i = 1, 2, ..., m$$

The Chi Squared random variable distribution function is as follows,

$$P(y \leq k) = \frac{\int_0^k 2^{m/2} y^{\frac{(m\ \ 2)}{2}} e^{y/2}}{\Gamma(m/2)} dy \tag{13}$$

Where $\Gamma(\cdot)$ is the Gamma function which is[2],

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx \tag{14}$$

T_n(obs) is called "observed value" in the test. By getting T_n(obs) and considering the theory of the test, the probability value, which is the area under the curve of the Chi Squared distribution density function, can be derived from the value of T_n(obs) to infinity, as follows.

$$PV = \Gamma(m-1, T_n(obs))$$

To make the decision, it is enough to calculate PV according to the significance level (α) which is usually chosen smaller than 0.01 [2]. If the PV is greater than α, then the sequence is assumed to be random, and if PV is smaller than the chosen value of α, then the sequence fails.

V. Test of Ranks of Matrices with Elements Belonging to GF(q)

0-7803-9521-2/06/$20.00 ©2006 IEEE.

Suppose that S^{n^*} is a random sequence with elements belonging to GF(q), where $n^* = n^*m^2$, and $S^{n^*} = S_1, S_2, \ldots, S_{n^*}$, $S_i \in$ GF(q). Now if we look at each m^2 consecutive elements of this sequence as an $m \times m$ matrix ($\mathrm{M_i}$), a sequence of $m \times m$ matrices Mat^n can be formed,

$$Mat^n = M_1, M_2, \ldots, M_n$$

If we let the rank of each matrix M_i (i = 1, 2, …, n) in the Mat^n sequence be called R_i, we have a sequence Ran^n matching Mat^n like this,

$$Ran^n = R_1, R_2, \ldots, R_n, \quad R_i \in \{0, 1, 2, \ldots, m\}$$

If we assume that elements of S^{n^*} are chosen from elements of GF(q) to be completely random, independent and uniformly distributed, then sequence Ran^n is a sequence of independent random variables, which according to equation (4) has the probability function below,

$$P[\mathrm{R_i} = r_i] = P_r, \quad r_i = 0, 1, \ldots, m$$

In section III we introduced a simple method to approximate P_m and as a result, P_{m-1}, P_{m-2},…. .As it is evident, if elements of the S^{n^*} sequence are independent and uniformly distributed, then the sequence Ran^n will be equivalent to a random vector with parameters n, P_1, P_2, …, P_m, and according to theorem 1 it can be said that $\mathrm{T_n}$(obs) will tend to a Chi Squared random variable with m degree of freedom..

$$T_n(obs) = \sum_{i-0}^{m} \frac{(F_i - np_i)^2}{np_i} \tag{15}$$

Considering that using $\mathrm{T_n}$(obs) as the statistic of the Chi Squared test, is bound to the condition $nP_i > 5$, it is possible that for some cases this condition is not met. Therefore, a number of events with high probability can be picked, and other cases can be combined such as *BMR* test. Therefore if we assume that except for P_m, P_{m-1} and P_{m-2}, probabilities are not significant, then we can count all matrices having a rank smaller than or equal to $m-2$ as one, and call it F^*. It is evident that the probability that matrices with a rank of $m-2$ or less occur is equal to,

$$P^* = 1 - (P_m + P_{m-1})$$

Therefore statistic of the Rank of Matrix on GF(q) test can be looked at this way,

$$T_n(obs) = \frac{(F_m - nP_m)^2}{nP_m} + \frac{(F_{m\,1} - nP_{m\,1})^2}{nP_{m\,1}} + \frac{(F^* - nP^*)^2}{nP^*} \tag{16}$$

It is evident that $\mathrm{T_n}$(obs) will tend towards a Chi Squared random variable with 2 degree of freedom. The condition for appropriate approximation to perform the test will be, nP_m, nP_{m-1}, $nP^* > 5$.

If a fewer number of P_i's are desired to be combined, then $\mathrm{T_n}$(obs) will be,

$$T_n(obs) = \sum_{i=0}^{t} \frac{(F_{m\ i} - nP_{m\ i})^2}{nP_{m\ i}} + \frac{(F^* - nP^*)^2}{nP^*}$$

$$P^* = 1 - \sum_{i=0}^{t} P_{m\ i}$$

In this case $\mathrm{T_n}$(obs) will tend towards a Chi Squared random variable with $t+1$ degree of freedom. Therefore the probability value for $\mathrm{T_n}$(obs) in the two mentioned cases will be,

$$\mathrm{PV} = gamma(2, \mathrm{T_n(obs)})$$

$$\mathrm{PV} = gamma(t+1, \mathrm{T_n(obs)})$$

Therefore, to perform the test, by choosing α and comparing it to the probability value (PV), we can make a decision about the sequence undergoing the test. If the PV for a sequence is greater than or equal to α, then from these tests, the sequence is random, and if the PV is smaller than α, then the sequence will not be taken as random. It is necessary to mention that the value of α is suggested to be in the span 0.001<α<0.01 [2,3].

VI. Conclusion

In this paper we consider the presented *BMR* test that converted the binary sequence to $m \times m$ matrices, and then the statistical test was performed, according to the probability function of resulting binary matrices' ranks. Then we went through the process of upgrading this test we suppose that the sequences have elements belonging to GF(q). In this regard, considering that the ranks of matrices with elements belonging to GF(q), do not have a simple expression, first we tried to simplify an approximate this function, then considering the achieved expressions and the simplified probability function, we presented the method to perform the test on the basis of the ranks of matrices with elements belonging to GF(q).

VII. References:

[1] J. Kahn, J. Komlos, and E. Szemeredi, "On the probability that a random matrix is singular", J. Am. Math. Soc., *(1), 233-240 (1995)

[2] "A Statistical Test Suite For Random And Pseudorandom Number Generators Cryptographic

Applications" NIST Special Publication 800-22,May2001
[3]M. Dakhil-Alian, M. R. Aref, B. Sadeghian, M. Modarres, "A New Statistical Test Based on Linear Complexity Profile" (LCP), Amirkabir, Vol.11,No.42
[4] N. Linial, D. Weits, "Random vectors of bounded weight and their linear dependencies" preprint