# Efficient Evaluation of Frequency Test for Overlapping Vectors Statistic

Krzysztof Mańk
*Military University of Technology, Warsaw, Poland*

March 8, 2023

**Abstract**

Randomness testing is one of the essential and easiest tools for evaluating cryptographic primitives. The faster we can test, the greater volume of data that can be tested. Thus a more detailed analysis is possible. This paper presents a range of observations made for a well-known frequency test for overlapping vectors in binary sequence testing. We have obtained precise chi-square statistic computed in $O\left(dt2^{dt}\right)$ instead of $O\left(2^{2dt}\right)$ time, without precomputed tables.

*keywords: chi-square test, overlapping vectors testing, randomness testing*

## 1 Introduction

Randomness testing is widely used in the evaluation of cryptographic primitives by reduction to examine appropriately crafted binary sequences. It is obvious that the quality of the analysis increases with the volume of tested data, but the time and cost increase as well.

In this paper we will take a closer look at well-known frequency test for overlapping vectors. One of the first results comes from Good [1]. He proposed using two statistics computed for $t$ and $t-1$ element vectors. Supose we have sequence of $n$ $d$-bit nonoverlapping blocks: $B_1, B_2, ..., B_n$, which we extend adding $t-1$ of the initial elements at the end. Now we create two sequences, the first – of $t \geq 2$ element vectors:

$$\begin{gathered}(B_1, B_2, ..., B_t), (B_2, B_3, ..., B_{t+1}), (B_3, B_4, ..., B_{t+2}), \ldots, \\ (B_{n-t+1}, B_{n-t+2}, ..., B_n), (B_{n-t+2}, B_{n-t+3}, ..., B_n, B_1), \\ (B_{n-t+3}, B_{n-t+4}, ..., B_n, B_1, B_2), \ldots, (B_n, B_1, B_2, ..., B_{t-1}),\end{gathered}$$

and the second of $t-1$ element vectors:

$$\begin{gathered}(B_1, B_2, ..., B_{t-1}), (B_2, B_3, ..., B_t), (B_3, B_4, ..., B_{t+1}), \ldots, \\ (B_{n-t+2}, B_{n-t+3}, ..., B_n), (B_{n-t+3}, B_{n-t+4}, ..., B_n, B_1), \\ (B_{n-t+4}, B_{n-t+5}, ..., B_n, B_1, B_2), \ldots, (B_n, B_1, B_2, ..., B_{t-2}).\end{gathered}$$

Both sequences consist of $n$ vectors.

Let $v_{i_t}$ for $i = 0..2^{dt}-1$ be numbers of observed occurrences of all possible $t$ elements vectors ($dt$ bit blocks), and $v_{j_{t-1}}$ for $j = 0..2^{d(t-1)}-1$ – numbers of observed occurrences of all possible $t-1$ elements vectors ($d \cdot (t-1)$ bit blocks), where each $dt$ bit block is identified by an integer value, which binary representation this block constitutes.

The test statistic is a simple difference of the two usual Pearson statistics:

$$\psi_t^2 = \frac{2^{dt}}{n} \sum_{i=0}^{2^{dt}-1} v_{i_t}^2 - \frac{2^{d(t-1)}}{n} \sum_{i=0}^{2^{d(t-1)}-1} v_{i_{t-1}}^2,$$

has chi-square distrubion with $2^{dt} - 2^{d(t-1)}$ degrees of freedom asymptotically.

This approach was adopted by authors of Statistical Test Suite [2] in Serial Test. Two papers published in 2004: [3] and [4] gave means for evaluating exact test statistic. From the first one, we can derive the formula for the covariance matrix of the vector of counts – $v_{i_t}$, but to get efficient implementations, one needs to store the weak inverse of the covariance matrix. In the second paper, Alhakim proposed a workaround using the matrix's eigenvectors for eigenvalue 1. Alhakim's method is intended for a more general case in which the vector's elements come from any range of natural numbers. This causes the construction of eigenvectors to be unnecessarily complicated, and to obtain equivalent results to using a covariance matrix, it requires repetition of procedure for all vector lengths from 1 up to desired $t$. That means a lot of computations.

The observations below lead to the exact test statistic with the number of arithmetic calculations similar to Good's approach.

# 2 Construction of eigenvectors for all eigenvalues

In [3] we can find a formula for test statistic identical to the one received from a quadratic form with the weak inverse of the covariance matrix:

$$S_{d,t} = \frac{1}{n} \sum_{w=1}^{t} \frac{1}{w} \sum_{i=1}^{L(w,d,t)} \left(\Psi_i^w \circ v\right)^2,$$

where $w$ are eigenvalues and $\Psi_i^w$ are corresponding eigenvectors, and $L(w,d,t) = \left(2^d-1\right)^{\min\{2,t-w+1\}} \cdot \left(2^d\right)^{\max\{0,t-w-1\}}$ is the number of eigenvectors for eigenvalue $w$ [3]. Alhakim's formula from [4] is a truncation to $w = 1$ only, of presented above.

Let $H^m = H^1 \otimes H^{m-1}$, where $H^1 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ and $\otimes$ denotes Kronecker product, denote Walsh-Hadamard matrix of degree $2^m$. By $H_i^m$ for $i = 0..2^m - 1$ we will denote rows of the matrix $H^m$.

For a given pair $d$ and $t$, we will derive all eigenvectors from the matrix $H^{dt}$.

There are $L(1,d,t) = \left(2^d-1\right)^t \cdot 2^{d(t-2)}$ eigenvectors for $w = 1$, they are:

$$H^{dt}_{i+j\cdot 2^d+2^{d(t-1)}},\ i = 1..2^d-1,\ j = 0..\left(2^d-1\right)2^{d(t-2)}-1.$$

There are $L(2,d,t) = \left(2^d-1\right)^{t-1} \cdot 2^{d(t-3)}$ eigenvectors for $w = 2$ and they are:

$$\frac{1}{\sqrt{2}}\left(H^{dt}_{i+j\cdot 2^d+2^{d(t-2)}} + H^{dt}_{\left(i+j\cdot 2^d+2^{d(t-2)}\right)2^d}\right),$$

$$i = 1..2^d-1,\ j = 0..\left(2^d-1\right)2^{d(t-3)}-1.$$

For any $w < t$ its eigenvectors are given by:

$$\frac{1}{\sqrt{w}} \sum_{k=1}^{w} H^{dt}_{\left(i+j\cdot 2^d+2^{d(t-2)}\right)2^{d(k-1)}},$$

$$i = 1..2^d - 1,\ j = 0..\left(2^d - 1\right)2^{d(t-w-1)} - 1,$$

and for $w = t$ we have:

$$\frac{1}{\sqrt{t}}\sum_{k=1}^{t} H^{dt}_{i2^{d(k-1)}},\ i = 1..2^d - 1.$$

Due to the linearity of the dot product, we can rewrite the formula for the test statistic:

$$S_{d,t} = \frac{1}{n}\sum_{w=1}^{t-1}\frac{1}{w^2}\sum_{i=1}^{2^d-1}\sum_{j=0}^{\left(2^d-1\right)2^{d(t-w-1)}-1}\left(\sum_{k=0}^{w-1}\left(H^{dt}_{\left(i+j\cdot 2^d+2^{d(t-2)}\right)2^{dk}}\circ v\right)\right)^2 +$$

$$+\frac{1}{n}\frac{1}{t^2}\sum_{i=1}^{2^d-1}\left(\sum_{k=0}^{t-1}\left(H^{dt}_{i2^{dk}}\circ v\right)\right)^2.$$

Because all dot products above constitute elements of the Walsh-Hadamard transform of vector $v$, which we will denote as $V$, we finally get:

$$S_{d,t} = \frac{1}{n}\sum_{w=1}^{t-1}\frac{1}{w^2}\sum_{i=1}^{2^d-1}\sum_{j=0}^{\left(2^d-1\right)2^{d(t-w-1)}-1}\left(\sum_{k=0}^{w-1}V_{\left(i+j\cdot 2^d+2^{d(t-2)}\right)2^{dk}}\right)^2 + \frac{1}{n}\frac{1}{t^2}\sum_{i=1}^{2^d-1}\left(\sum_{k=0}^{t-1}V_{i2^{dk}}\right)^2.$$

Vector $V$ should be computed by means of the fast Walsh-Hadamard transform, which time complexity is $O\left(dt2^{dt}\right)$.

# 3 Structure of the count vector and its utilization

Since consecutive observed vectors $(B_i, B_{i+1}, ..., B_{i+t-1})$ overlap on $t-1$ blocks and we have extended examined sequence by as many its initial blocks, then for every possible $d\,(t-1)$ bit block value we can write an equation:

$$\sum_{k=0}^{2^d-1} v_{k\cdot 2^{d(t-1)}+i} = \sum_{k=0}^{2^d-1} v_{k+i\cdot 2^{d(t-1)}},\ i = 0..2^{d(t-1)} - 1.$$

The additional equation is obvious:

$$\sum_{k=0}^{2^{dt}-1} v_k = n.$$

This system of equations has order $2^{d(t-1)}$ thus allows to determine $2^{d(t-1)}$ of the elements of the vector $v$ as a linear combination of $n$ and the rest of them, that is $2^{dt} - 2^{d(t-1)}$, which is consistent with the stated number of degrees of freedom of the test statistic $S_{d,t}$.

Walsh-Hadamard transform of modified this way vector $v$ has an interesting property – elements $V_{\left(i+j\cdot 2^d+2^{d(t-2)}\right)2^{dk}}$ for a given trio $(w, i, j)$ and every $k = 0..w-1$ are equal. The same applies to $V_{i2^{dk}}$ of course. This leads to further simplification of the test statistic:

$$S_{d,t} = \frac{1}{n}\sum_{w=1}^{t-1}\sum_{i=1}^{2^d-1}\sum_{j=0}^{\left(2^d-1\right)2^{d(t-w-1)}-1}\left(V_{\left(i+j\cdot 2^d+2^{d(t-2)}\right)2^{d(w-1)}}\right)^2 + \frac{1}{n}\sum_{i=1}^{2^d-1}\left(V_{i2^{d(t-1)}}\right)^2.$$

In a such setup, initial $2^{d(t-1)}$ elements of $V$ are obsolete.

# 4 Conclusion

As shown by us, the method of determining eigenvectors for all eigenvalues of the covariance matrix seems important for two reasons. First, it allows for avoiding repetitions of the sequence evaluation for consecutive vector lengths and a significant acceleration of calculations. An additional bonus, described in our earlier work [5] is the possibility of determining the test statistic values for all vector dimensions, from 1 to the assumed t, after one run of the sequence.

Finally, we would like to draw attention to the groundbreaking nature of the theoretical work of Alhakim, Kawczak, and Molchanov, while ours is the result of observations made while implementing those results.

Further work should lead to strict algebraic proof of the correctness of the eigenvectors' construction.

# References


[1] Good I. The serial test for sampling numbers and other tests for randomness. *Mathematical Proceedings of the Cambridge Philosophical Society*, 1953. **49**:276–284.

[2] Rukhin A, Soto J, Nechvatal J, Smid M, Barker E, Leigh S, Levenson M, Vangel M, Banks D, Heckert A, Dray J, Vo S. A statistical test suite for random and pseudorandom number generators for cryptographic applications. *NIST Special Publication SP 800-22 Revision 1a*, 2010. doi:10.6028/NIST.SP.800-22r1a.

[3] Alhakim A, Kawczak J, Molchanov S. On the Class of Nilpotent Markov Chains, I. The Spectrum of Covariance Operator. *Markov Processes and Related Fields*, 2004. **4**.

[4] Alhakim A. On the eigenvalues and eigenvectors of an overlapping Markov chain. *Probability Theory and Related Fields*, 2004. **128**:589–605. doi:10.1007/s00440-003-0321-z.

[5] Mank K. Test czestosci dla nakladajacych sie wektorow (in Polish: Frequency test for overlapping vectors). *Cyberprzestepczosc i ochrona informacji Bezpieczenstwo w internecie tom II, Wydawnictwo Wyzszej Szkoly Menedzerskiej w Warszawie*, 2013.

[6] L'Ecuyer P, Simard R. TestU01: A C library for empirical testing of random number generators. *ACM Trans. Math. Softw.*, 2007. **33**.