ON STATISTICAL HYPOTHESES TESTING OF EMBEDDING

YU.S. KHARIN, E.V. VECHERKO
Belarusian State University
Minsk, Belarus
e-mail: kharin@bsu.by

Abstract

The goal of this paper is to find out when we are able reliably to distinguish a random sequence that does not have an embedded random sequence from a sequence that has it.

1 Introduction

Nowadays models of embedding are used in solving problems of intellectual property rights [1, 3]. Statistical issues in this application are underdeveloped, so the actual problem is a construction and analysis of the statistical tests for embedding.

2 Probabilistic models of embedding

The mathematical model of a random sequence that does not have an embedded random sequence can be represented as a sequence of N-dimensional random binary column vectors $x_1, x_2, \ldots, x_n$ [3]:

$$x_t = (x_{t1}, x_{t2}, \ldots, x_{tN})' \in V_N,\ t = 1, \ldots, n,$$

where n is the size of the sequence; $V_N = \{J = (j_k) :\ j_k \in V = \{0, 1\},\ k = 1, \ldots, N\}$ is the set of 2^N binary N-vectors. A binary vector x_t will be represented by the number $< x_t >= x_{t1} + 2^1 x_{t2} + \ldots + 2^{N-1} x_{tN},\ \ < x_t > \in A = \{0, 1, \ldots, 2^N - 1\}$.

Define: $N = N_1 + N_2,\ 1 \le N_1 \le N,\ m_\tau = (m_{\tau 1}, \ldots, m_{\tau N_1}) \in V_{N_1},\ J = (J'_{(1)}, J'_{(2)})' \in V_N,\ J_{(1)} \in V_{N_1},\ J_{(2)} \in V_{N-N_1}$. If a hidden sequence $\{m_t\}$ is embedded into $\{x_t\}$ we observe the following sequence:

$$\tilde{x}_t = (\tilde{x}'_{t(1)}, \tilde{x}'_{t(2)})',\ \tilde{x}_{t(1)} \in V_{N_1}, \tilde{x}_{t(2)} \in V_{N-N_1},$$

$$\tilde{x}_{t(2)} = x_{t(2)},\ \tilde{x}_{t(1)} = \xi_t m_{\tau_t} + (1 - \xi_t) x_{t(1)} = \begin{cases} x_{t(1)}, & \xi_t = 0, \\ m_{\tau_t}, & \xi_t = 1 \end{cases}, \tag{1}$$

$$\tau_t = \tau_t(\xi_1, \ldots, \xi_t) = \sum_{i=1}^{t} \xi_i,\ P\{\xi_t = 1\} = 1 - P\{\xi_t = 0\} = \beta,\ t = 1, \ldots, n, \tag{2}$$

where ξ_t is a sequence of independent Bernoulli random variables that determine the embedding process; m_{τ_t} is a sequence of independent and identically distributed random binary N_1-vectors, $P\{m_\tau = J_{(1)}\} = p_{<J_{(1)}>},\ J_{(1)} \in V_{N_1}$, that represents an embedded random sequence. The random sequences $\{x_t\},\ \{\xi_t\},\ \{m_\tau\}$ are independent.

Theorem 1. *Assume that $x_t \in V_N$ is a stationary Markov chain with the state space V_N, a stationary probability distribution $\pi = (\pi_i)$ and any one-step transition probability matrix $P = (p_{ij})$, $m_\tau \in V_{N_1}$ is a random sequence of N_1-vectors that represents an embedded random sequence, $P\{m_\tau = J_{(1)}\} = p_{<J_{(1)}>}$, $J_{(1)} \in V_{N_1}$, and the observed sequence $\tilde{x}_t$ is obtained according to (1). Then the two-dimensional probability distribution $\tilde{\pi}_{<J>,<K>} = P\{\tilde{x}_{t-1} = J, \tilde{x}_t = K\}$, $J, K \in V_N$ of the sequence $\tilde{x}_t$ is:*

$$\tilde{\pi}_{<J>,<K>} = (1-\beta)^2 \pi_{<J>} p_{<J>,<K>} + \beta(1-\beta)\times$$

$$\times\left(p_{<K_{(1)}>} \sum_{v=0}^{2^{N_1}-1} \pi_{<J>} p_{<J>,2^{N_1}<K_{(2)}>+v} + p_{<J_{(1)}>} \sum_{v=0}^{2^{N_1}-1} \pi_{2^{N_1}<J_{(2)}>+v} p_{2^{N_1}<J_{(2)}>+v,<K>}\right)+$$

$$+\beta^2 p_{<J_{(1)}>} p_{<K_{(1)}>} \sum_{v,h=0}^{2^{N_1}-1} \pi_{2^{N_1}<J_{(2)}>+v} p_{2^{N_1}<J_{(2)}>+v,2^{N_1}<K_{(2)}>+h}.$$

3 Statistical testing of embedding

Assume that the the sequence $\{x_t\}$, that does not contain an embedded random sequence, is a sequence of independent and identically distributed (i.i.d.) random binary N-vectors which have a fixed discrete probability distribution: $\pi^0 = (\pi_i^0)$, $\pi^0_{<J>} = P\{x_t = J\}$, $J \in V_N$. A marginal probability distribution of the subvector $(x_{t1}, \ldots, x_{tp})' \in V_p$, $p \le N$ is:

$$\kappa^0_{<J>} = P\{x_{t1} = j_1, \ldots, x_{tp} = j_p\} = \sum_{v=0}^{2^{N-p}-1} \pi^0_{2^p v+<J>}, \; J \in V_p, \; i = 0, \ldots, 2^p - 1. \quad (3)$$

We observe a sequence $\tilde{X} = (\tilde{x}_1', \tilde{x}_2', \ldots, \tilde{x}_n')' \in V_{nN}$ of the finite size n. Let us construct the Neyman-Pearson statistical test for two hypotheses:

$$H_0: \; \{\tilde{X} \text{ does not have an embedded sequence}\},$$

$$H_1 = \bar{H}_0: \; \{\tilde{X} \text{ has an embedded sequence}\},$$

where β is a parameter defined in (2). The error probability of the first type is $\epsilon \in (0, 1)$. Translating our problem into precise terms, we come to the hypotheses:

$$H_0 : \tilde{\kappa} = \kappa^0, \; H_1 : \tilde{\kappa} \ne \kappa^0, \quad (4)$$

where $\tilde{\kappa} = (\tilde{\kappa}_i)$ is the marginal probability distribution of the subvector $(\tilde{x}_{t1}, \ldots, \tilde{x}_{tp})' \in V_p$, $N_1 \le p \le N$: $\tilde{\kappa}_{<J>} = P\{\tilde{x}_{t1} = j_1, \ldots, \tilde{x}_{tp} = j_p\}$, $J = (j_1, \ldots, j_p)' \in V_p$. Then the statistical test is:

$$d = d(\tilde{X}) = \begin{cases} 0, & T_p(\tilde{X}) < \delta, \\ 1, & T_p(\tilde{X}) \ge \delta \end{cases}, \; T_p(\tilde{X}) = \sum_{i=0}^{2^p-1} \frac{(\nu_i - n\kappa_i^0)^2}{n\kappa_i^0}, \quad (5)$$

where $\nu_i = \sum_{t=1}^n \delta_{<(\tilde{x}_{t1}, \ldots, \tilde{x}_{tp})'>, i}$, $i = 0, \ldots, 2^p - 1$; δ is a critical value.

When the null hypothesis is true, the asymptotic ($n \to \infty$) probability distribution of the test statistic $T_p(\tilde{X})$ is a χ^2-distribution with $2^p - 1$ degrees of freedom. If ϵ is a significance level of the test then the critical value of the test (5) is $\delta = (\chi^2_{2^p-1})^{-1}(1-\epsilon)$, where $(\chi^2_{2^p-1})^{-1}(1-\epsilon)$ is the $(1-\epsilon)$-quantile of the χ^2-distribution with $2^p - 1$ degrees of freedom, $0 < \epsilon < 1$.

To find the power of the statistical test (5) we assume that the alternative hypothesis H_1 is true and the probability distribution of the observed sequence that has an embedded sequence is:

$$\tilde{\kappa} = \kappa^1 = \kappa^0 + \Delta,\ \kappa^0 = (\kappa^0_0, \kappa^0_1, \ldots, \kappa^0_{2^p-1}),\ \Delta = (\Delta_0, \Delta_1, \ldots, \Delta_{2^p-1}),$$

where $\sum_{i=0}^{2^p-1} \Delta_i = 0,\ -\kappa^0_i < \Delta_i < 1 - \kappa^0_i,\ i = 0, 1, \ldots 2^p - 1.$

Theorem 2. *When the alternative hypothesis is true: $\tilde{\kappa} = \kappa^1 = \kappa^0 + \Delta$, the asymptotic ($n \to \infty$) probability distribution of the test statistic $T_p(\tilde{X})$, defined by (5), is the noncentral χ^2-distribution with $2^p - 1$ degrees of freedom and the non-centrality parameter*

$$\lambda^2_n = ng(\kappa^0, \Delta),\ g(\kappa^0, \Delta) = \sum_{i=0}^{2^p-1} \frac{\Delta_i^2}{\kappa^0_i}. \tag{6}$$

Note that the results of Theorem 2 are in accordance with the results in [1].

Now we find out what is the minimal size $n^* = n^*(w^*)$ of the observed sequence $\{\tilde{x}_t\}$ such that we are able reliably to identify the existence of an embedded sequence:

$$n^* = n^*(w^*) = \min_n \{\chi^2_{2^p-1,\lambda^2_n}(\delta) \le 1 - w^*\},$$

where w^* is a fixed power of the test (5); $\chi^2_{2^p-1,\lambda^2_n}(\cdot)$ is a function of the noncentral χ^2-distribution with $2^p - 1$ degrees of freedom and the non-centrality parameter λ^2_n.

Theorem 3. *Let $m_\tau \in V_{N_1}$ be an embedded random sequence, $P\{m_\tau = J_{(1)}\} = 1/2^{N_1}$, $J_{(1)} \in V_{N_1}$, and the observed sequence $\tilde{x}_t$ is obtained according to (1). Now if $\pi^0 = (\pi^0_i)$ is a probability distribution of x_t: $\pi^0_i = P\{< x_t >= i\}$, $i \in A$, then, when $H_1 : \{\beta > 0\}$ is true, the non-centrality parameter of the asymptotic noncentral χ^2-distribution of the test statistic $T_N(\tilde{X})$, defined by (5), is:*

$$\lambda^2_n = n\beta^2 \sum_{J \in V_N} \frac{s_1^2}{\pi^0_{<J>}},\ s_1 = \frac{1}{2^{N_1}} \sum_{l=0}^{2^{N_1}-1} \pi^0_{2^{N_1} \cdot <J_{(2)}> + l} - \pi^0_{<J>}. \tag{7}$$

This theorem allows to analyze a dependence of the power of the test on the proportion β of an embedded sequence. In the following example we show how the proportion β of an embedded sequence affects the power of the statistical test.

Example. We calculate a power of the test using Theorem 3 under the following parameter values: $n = 512,\ N = 3,\ N_1 = 1,\ \pi^0 = \frac{1}{40} \cdot (5, 3, 7, 5, 5, 4, 8, 3)'$. In fig. 1 there is a plot of dependence of the power of the test on the proportion β of embedded sequence. For example, if $\beta > 0.76$, then the probability of the correct rejection of the

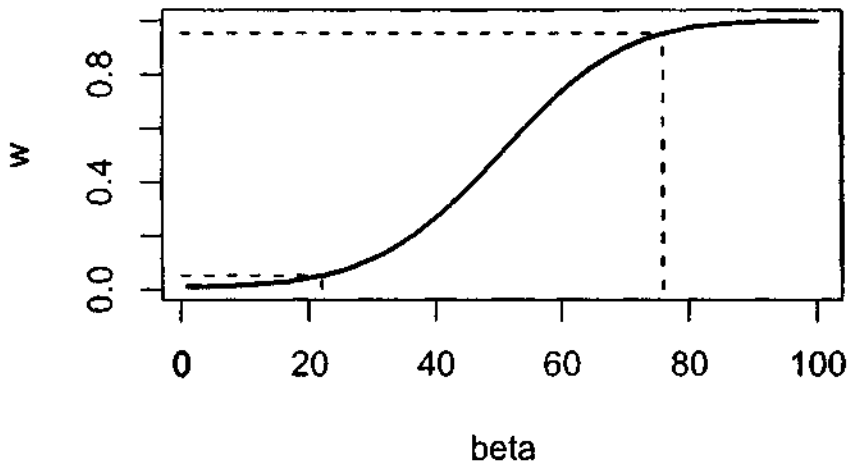

Figure 1: Dependence of power w of the test on proportion β of embedded sequence

null hypothesis is more 0.95. If $\beta < 0.22$, then the probability of rejection of the null hypothesis, when the alternative hypothesis is true, is less than 0.05.

Similarly, we construct a statistical test when a sequence, that does not contain an embedded sequence, is a stationary Markov chain and the one-step transition probability matrix is:

$$P^0 = (p^0_{ij}),\ p^0_{ij} = P\{< x_t >= j | < x_{t-1} >= i\},\ i, j \in A.$$

Let $\tilde{X} = (\tilde{x}'_1, \tilde{x}'_2, \ldots, \tilde{x}'_n)' \in V_{nN}$ be an observed sequence of finite size n. The null and the alternative hypotheses are $H_0 : \tilde{P} = P^0$ and $H_1 : \tilde{P} \neq P^0$ respectively. The statistical test [2] is:

$$d = d(\tilde{X}) = \begin{cases} 0, & T(\tilde{X}) < \delta, \\ 1, & T(\tilde{X}) \geq \delta \end{cases},\ T(\tilde{X}) = \sum_{i,j=0}^{2^{N_1}-1} \frac{(f_{ij} - f_i p^0_{ij})^2}{f_i p^0_{ij}}, \tag{8}$$

where $f_{ij} = \sum_{t=2}^{n} \delta_{<\tilde{x}_{t-1}>,i} \delta_{<\tilde{x}_t>,j}$, $f_i = \sum_{t=1}^{n} \delta_{<\tilde{x}_t>,i}$; δ is a critical value.

When the null hypothesis is true, the asymptotic ($n \to \infty$) probability distribution of the test statistic (8) is the χ^2-distribution with $2^N(2^N - 1)$ degrees of freedom [2].

References

[1] Ponomarev. K.I. (2009). A parametric model of embedding and its statistical analysis. *Discrete Mathematics and Applications.* Vol. 19.

[2] Billingsley. P. (1961) Statistical methods in Markov chains. *Ann. Math. Statist.* Vol. 32.

[3] Waterman M.S. (1989). *Mathematical methods for DNA sequences.* CRC Press, Boca Ration.