Nonparametric Tests for Randomness

Ying Wang

Abstract

To decide whether a given sequence is "truely" random, or independent and identically distributed, we need to resort to nonparametric tests for randomness. Six tests: the ordinary run test, the sign test, the runs up and down test, the Mann-Kendall test, the Bartels' rank test and the test based on entropy estimators are introduced in this report and their weaknesses are analyzed. Combining the decisions made by each test, we can further improve the confidence on the randomness of a given sequence. As an example, the tests are applied to test the randomness of DCT coefficient channels of images. Surprisingly, the results show that almost half of DCT AC coefficient channels are decided "i.i.d" for the image *Lena*, while only three are decided "i.i.d" for the image *Baboon*.

I. Introduction

In statistical literature, a "truly random" process refers to a process that can produce independent and identically distributed (i.i.d) samples. If an observed value in the sequence is influenced by its position in the sequence, or by the observations which proceed it, the process is not truly random. Here, the "randomness" really equals to the property of i.i.d. This property is essential for the theoretical bases of many classical statistical tests [1], signal detection and estimation methods [3], capacity calculation formula [4] and so on. Even when the observations are not truly random, which is true in many practical applications, we can still perform those simple and neat theoretical results with a certain degree of confidence if we can tell how close to random the data can be. Also, it is of great interests in cryptographic security, where it is necessary to examine the real randomness of various "random" number generators, and Monte Carlo simulations, where the randomness of casted number greatly affect the accuracy of integrals.

Investigations of randomness of a given sequence often require statistical tools for distribution comparison. Among them, goodness-of-fit tests and entropy estimates are two well-understood concepts [5]. However, when the distribution of the observed data is unknown, the hypotheses simply are

$$\begin{cases} H_0 : \text{Sequence is i.i.d (random)} \\ H_1 : \text{Sequence is not i.i.d (random).} \end{cases} \tag{1}$$

Then we have to resort to nonparametric tests, using some distribution-invariant properties of random processes. For example, if the observations can be transformed to some symbols

that can reflect some properties of their relative positions or magnitudes, then the pattern of the resulting symbol sequence can serve as a measure of the randomness of the original process.

The pattern of the symbols can be analyzed using runs, or entropy estimators if we can know the distribution of the symbols. In this report, we first introduce the tests of randomness based on runs or trends, such as the ordinary run test, the runs up and down test, the sign test, the Mann-Kendall test and the Bartels' rank test. Then the test based on entropy estimators is described in section III. Since these tests are only based on partial information of the sequence, some of them even fail to detect the randomness of a deterministic sequence as discussed in section IV. In section V, we applied these tests to examine the usual i.i.d assumption for the block DCT AC coefficient channels and to find out whether watermarking can affect the randomness of certain sequences taken from the image processes. The last section concludes this report.

II. Tests based on Runs

In any ordered sequence with two types of symbols, a run is defined as a succession of one or more identical symbols, which are followed and proceeded by a different symbol or no symbol at all [1]. For example, the males and females in a line can have patterns such as $M\ F\ M\ F\ M\ F\ M\ F$ and $M\ M\ M\ M\ F\ F\ F\ F$, which have 8 and 2 runs, respectively. Both the number of runs and their lengths can be used as a measure of the randomness of the ordered symbol sequence. Too few runs, too many runs, a run of excessive length, etc., are very rare in truly random sequences, therefore they can serve as statistical criteria for the rejection of H_0. Also, these criteria are related with each other. Too few runs means that some runs are too long; too many runs results in short runs. So we can be only concerned with the total number of runs.

In the above two examples, symbols arise naturally. For quantitative observations, we need to impose some dichotomizing criterion to symbolize the sequence. Every number is compared to a focal point, commonly the median or mean of the samples, and is denoted as "+" or "-", according to whether the number is larger or smaller than the focal point. Also, the relative magnitudes or ranks of adjacent numbers can provide information on the trend or autocorrelation of the sequence.

A. The ordinary run test based on the median

Using the median of all observations as a focal point, the dichotomy for the ordinary run tests compares each observation with the median and assigns "+" to those samples larger than the median and "-" to the samples less than or equal to the median. Under the null hypothesis of randomness, every arrangement of "+" and "-" signs is supposedly equiprobable. Assume that the ordered sequence has n samples, n_1 of "+", n_2 of "-" and $n = n_1 + n_2$. Also, we denote the number of runs of "+" as R_1 and the number of runs of "-" as R_2, then the total number of runs is $R = R_1 + R_2$.

By simple knowledge of permutations and combinations, we can get the joint probability distribution of R_1 and R_2, the respective marginal probability distributions of R_1 and R_2, and the probability distribution of R [2]. The last one is

$$f_R(r) = \begin{cases} \binom{n_1-1}{r/2-1}\binom{n_2-1}{r/2-1} \Big/ \binom{n_1+n_2}{n_1} & \text{, if r is even;} \\ \left[\binom{n_1-1}{(r-1)/2}\binom{n_2-1}{(r-3)/2} + \binom{n_1-1}{(r-3)/2}\binom{n_2-1}{(r-1)/2}\right] \Big/ \binom{n_1+n_2}{n_1} & \text{, if r is odd.} \end{cases} \tag{2}$$

When both n_1 and n_2 are large, we can get the normal approximation for the null distribution of R, for which the mean is

$$\mu_1 = 2n_1n_2/n + 1, \tag{3}$$

and the variance is

$$\sigma_1^2 = 2n_1n_2(2n_1n_2 - n)/n^2(n-1). \tag{4}$$

Set the probability of false alarm P_{FA} as α, then the critical region is

$$|R - \mu_1| \geq \mathcal{Q}(\alpha/2)\sigma_1, \tag{5}$$

where $\mathcal{Q}(\alpha) = \frac{1}{\sqrt{2\pi}} \int_\alpha^\infty e^{-t^2/2} \mathrm{d}t$.

B. The runs up and down test

For numerical observations, we can also look at the difference between two consecutive samples: $X_2 - X_1$, $X_3 - X_2$, ..., $X_n - X_{n-1}$. Ignoring the zero differences, we record the sequence difference using plus signs for $X_i - X_{i-1} > 0$, and minus signs otherwise. Naturally, for a random process, we expect that there are roughly equal numbers of both signs. By the central limit theorem, the number of positive signs P converges weakly to

$\mathcal{N}(m/2, m/4)$ if there are m non-zero values of $X_i - X_{i-1}$. This simple fact can also serve as a test for randomness, which is called the sign test.

However, the sign test certainly can not reject the case such as half positive signs followed by half negative signs. The cluster of positive signs means that an up trend happens in the sequence; The cluster of negative signs corresponds to a down trend. Lasting up or down trends certainly should not appear in a random sequence. So, again we look at the number of runs of consecutive positive or negative differences: R. A run up starts with a plus sign and a run down starts with a minus sign. The exact distribution of R under the null hypothesis of randomness can be obtained by calculating all possible permutations and combinations [2]. When the number of observations is large, say $n > 25$, the asymptotic distribution of R is $R \sim \mathcal{N}(\mu_2, \sigma_2^2)$, where $\mu_2 = (2m-1)/3$ and $\sigma_2^2 = (16m-29)/90$. Similarly, the critical region is

$$|R - \mu_2| \geq \mathcal{Q}(\alpha/2)\sigma_2, \tag{6}$$

if the false alarm probability P_{FA} is α.

C. Other test against trends

As mentioned in last subsection, we should reject the null hypothesis if there is any trend in the sequence. The following are two tests against trend by looking at the signs of sample differences and the rank variation of successive samples.

C.1 Mann-Kendall test

Originally, Kendall's tau statistic is used as a measure of association in a bivariate population (X, Y) [2]. If we treat the time, $\{1,\ 2, ...,\ n\}$, of an observed sequence as X and the set of time-ordered observations, $\{Y_1,\ Y_2,\ ...,\ Y_n\}$, as Y, then the association between X and Y can be considered as an indication of a trend. Unlike the runs up and down test, the signs of relative magnitude of each observation relative to every preceding observation are considered in the Kendall's sample tau coefficient. The test statistic is

$$T = \sum_{i=2}^{n} \sum_{j=1}^{i-1} \text{sign}(Y_i - Y_j), \tag{7}$$

which converges to a normal random variable under the null hypothesis of randomness: $T \sim \mathcal{N}(0, \sigma_3^2)$, where $\sigma_3^2 = n(n-1)(2n+5)/18$. The critical region is

$$|T| \geq \mathcal{Q}(\alpha/2)\sigma_3, \tag{8}$$

if the false alarm probability P_{FA} is α.

C.2 Bartels' rank test

Instead of comparing the magnitude of each observation with its preceding samples, Bartels' rank test ranks all the samples from the smallest to the largest. The rank is the corresponding sequential number of X_i: $R(X_i)$. Under the null hypothesis of randomness, any rank arrangement from all $n!$ possibilities should be equiprobable. Once again, through calculating permutations and combinations, the probability for the test statistic $NM = \sum_{i=1}^{n-1}[R(X_i) - R(X_{i+1})]^2$ can be obtained. The large-sample approximation for $NM/(n(n^2-1)/12)$ is a normal random variable with mean 2 and variance $\sigma_4^2 = 4(n-2)(5n^2-2n-9)/5n(n+1)(n-1)^2$. Correspondingly, the critical region is

$$|NM - 2| \geq \mathcal{Q}(\alpha/2)\sigma_4, \tag{9}$$

if the false alarm probability P_{FA} is α.

III. Tests Based on Entropy Estimators

As discussed earlier in section II, the positive and negative signs for the ordinary run test and the runs up and down test should be equiprobable, i.e. $P(+) = 1/2$ and $P(-) = 1/2$. Also every element in the sequence of signs should be i.i.d. This mean we know exactly how the sign sequence should behave. Under the null hypothesis of randomness, the entropy rate of the sequence should be 1. If there is any dependence between the samples, the entropy rate should be strictly less than 1. That is, the new hypotheses are

$$\begin{cases} H_0 : H(P) = 1 \\ H_1 : H(P) < 1. \end{cases} \tag{10}$$

When we estimate $H(P)$ from a given sequence $\{X_i\}_{i=1}^n$, the entropy rate $H(P)$ can be calculated by many means if the sequence is i.i.d. The direct one is to get the approximate

probability distribution $\hat{P}(a) = \#\{0 < i \leq n, X_i = a\}/n$, then $H(\hat{P})$. Both $\hat{P}$ and $H(\hat{P})$ converge to their true value almost surely by the strong law of large numbers. Moreover, we can obtain $\hat{H}$ by looking at the approximate probability distribution $\hat{P}^{(r)}(n)$ of overlapping r-tuples of successive sign samples: $\tilde{X}_l^{(r)} = (\tilde{X}_l, \tilde{X}_{l+1}, ..., X_{l+r-1})$. For an i.i.d sequence,

$$\hat{H}_f^{(r)} = H\left(\hat{P}^{(r)}(n)\right) - H\left(\hat{P}^{(r-1)}(n)\right) \stackrel{a.s.}{\rightarrow} H(P). \tag{11}$$

Moreover, it is proved that even if the sequence is an ergodic chain of order k, (11) holds true for all $r > k$ [5]. Therefore, (11) can actually have an extra merit of providing us a measure to decide the correlation between samples since $\hat{H}_f$ will fall to the true value after r passes k.

To form a test with certain false alarm probability P_{FA}, it is necessary to find a statistic with a distribution function. It is proved that for an ergodic and independent sequence,

$$2n(log2(m) - \hat{H}_f^{(r)}) \stackrel{d.}{\rightarrow} \chi^2_{m^r - m^{r-1}}, \tag{12}$$

where m is the cardinality of the random variable space. For our case, $m = 2$. The critical region is

$$\hat{H}_f^{(r)} < 1 - \left(\chi^2_{2^r - 2^{r-1}}\right)^{-1}(\alpha),\ \forall\ r > 0. \tag{13}$$

IV. Comparison of tests for randomness

By far, we have introduced six tests for randomness: run test based on sample median, sign test, runs up and down test, Mann-Kendall test, Bartels' rank test and test based on entropy estimators. Apparently, the sign test is the weakest one since it accepts H_0 as long as there are approximately equal numbers of positive and negative signs from $sign(X_i - X_{i-1})$. Other tests are also likely to flounder for some specific processes according to their own weaknesses. We can compare the power of these tests by using some typical processes. In the following comparison, we will use $P_T = 2\mathcal{Q}(|T - \mu_T|/\sigma_T)$ as a measure of detection power since all the statistics for the above tests, except the test based on entropy estimators, can be approximated as normal random variables in large sample scenario. Therefore, all t's such that $|t - \mu_T| > |T - \mu_T|$ fall out of the confidence level of P_T for accepting H_0 and conversely H_0 is rejected with error probability of P_T. Certainly, $T = \mu_T$ gives full confidence on accepting H_0 and a large deviation of T from μ_T gives

high confidence on rejecting H_0. For the test based on entropy estimators, the statistic $T = 2n(log2(m) - \hat{H}_f^{(r)})$ is one-sided, so we define $P_T = P(T > t)$. $T = 0$ gives full confidence on accepting H_0.

First, we perform all the tests on a 1000-sample sequence with alternating $+1$ and -1's. We observe the following phenomena:

• The sign test accepts H_0 with confidence level of 0.9563 since the positive and negative signs for $\{X_i - X_{i-1}\}$ are almost equiprobable.

• The Mann-Kendall test also accepts H_0 with confidence level of 0.9622 since the alternating structure of $+1$ and -1 makes $\sum_{j=1}^{i-1} sign(X_i - X_j)$ and $\sum_{j=1}^{i} sign(X_{i+1} - X_j)$ almost cancel out with each other, hence $T \approx \mu_T = 0$.

• The entropy test accepts H_0 with confidence level of 0.9994 if just $\hat{H}_f^{(1)} = H(\hat{P}^{(1)})$ is used. But once the entropy is estimated using $\hat{H}_f^{(r)} = H(\hat{P}^{(r)}) - H(\hat{P}^{(r-1)})$, $r > 1$, the estimated entropy drops to 0 abruptly from $r = 2$, which indicate that the signs of difference between samples and their median form a totally deterministic sequence, therefore the original sequence must not be a random process.

Then, we gradually increase the repetition periods of $+1$'s and -1's. The sign test and Mann-Kendall test will more and more likely to reject H_0 with the increasing number of consecutive $+1$'s or -1's in one repetition. The test based on $\hat{H}_f^{(1)}$ will always accept H_0. However, it is interesting to notice that the ordinary run test and the test based on $\hat{H}_f^{(2)}$ accepts H_0 only when there are $\{+1, +1, -1, -1\}$ in one repetition, otherwise they can always correctly reject H_0. The reason for the ordinary run test's failure is that the mean for the number of runs is $\mu_1 \approx n/2$ and coincidentally $T \approx \mu_1$ in this case. For the calculation of $H(\hat{P}^{(2)})$, there happens to be $(1, 1)$, $(1, -1)$, $(-1, -1)$ and $(-1, 1)$ in one repetition. Both the rank test and the runs up and down test can correctly reject H_0 irrespective to the change. P_T, the confidence of accepting H_0, is shown in Fig.1 versus the number of consecutive $+1$'s or -1's in one reptition period for all the random tests.

It seems that the rank test and the runs up and down test are the most powerful in the above setup. However, we can always construct some deterministic sequence such that the sequence can make their test statistics T near to μ_T. For example, a sequence $1, 2, 3, 2, 3, 2, 1, 2, 3, 2, 3, 2, 1, 2, ...$ induces $+, +, -, +, -, -, +, +, -, +, -, -, +...$, in which

the number of runs is about the mean $\mu_2 \approx 2n/3$. Similarly, we may also find a deterministic sequence that flounders the rank test. Certainly, it will be very difficult to make it periodic since the rank of one sample is relative to the whole sequence. It is verified by Bartels [6] that the rank test is superior to the runs up and down test in many cases. "Its asymptotic relative efficiency is 0.91 with respect to the ordinary serial correlation coefficient against the alternative of first-order autocorrelation under normality." [2]

Although the asymptotic relative efficiency of test based on entropy estimators compared with the rank test is unknown, the nice thing about test based on entropy estimators is that the order of the sequence k can be estimated according to the property of $\hat{H}_f^{(r)}$, which will quickly converge to the true value of $H(P)$ once r is larger than k. If the final convergence is zero, the sequence must be deterministic.

From the above analysis, most of the tests are vulnerable to a certain set of sequences, which are deterministic but accepted as random processes. Since different tests have different weaknesses, we can combine the decisions from all the tests to minimize the error probability of missing H_1. Moreover, we can only to be sure a sequence is not random if any of the tests rejects H_0. However, even all above tests decide a sequence to be random, it is still very possible that the sequence is actually non-random since only partial information about the sequence is used in any of the above tests: signs of magnitude difference of samples relative to the median, signs of relative magnitude of samples and so on.

V. Testing the Randomness of 64 8×8 DCT Coefficient Channels

The 8×8 DCT transformation generates 64 equal-size data. Usually, it is assumed that the data in each AC channel is i.i.d, then goodness-of-fit tests are used to estimate the probability distribution, which is widely accepted as generalized Gaussian or Laplacian [7]. Using the above tests for randomness, we decide a channel to be i.i.d only if it passes all the tests with the false alarm probability of 0.05 for each test. The number in each entry of the following table shows the decision of randomness on 8×8 DCT coefficients from image *Lena*, where 1 means that the channel is i.i.d. Almost half of the channels can be accepted as i.i.d with some confidence. Most i.i.d channels are from higher AC frequencies. Usually the coefficients in the DC channel are correlated. Fig.2 shows the estimated entropy of the signs for the relative magnitude between samples and their median: $\hat{H}_f^{(r)}$ vs. the tuples of

overlapping samples used in the estimation r. For the DC channel, the entropy of signs drops abruptly at $r = 2$, which implies that the sign sequence may be close to an order-1 Markov chain. The trends for AC coefficient channels at $(1, 2)$ and $(1, 3)$ are almost the same. The curves drop deeply to about 0.1 from $r = 10$ to $r = 15$, which may imply that the sign sequence is more like a high-order Morkov chain with low entropy. But we should be cautious on making any more claim on the original DCT coefficient sequence since we are only observing the behavior of $\{sign(X_i - X_{median})\}$.

0	0	0	0	0	1	0	0
0	0	1	1	1	0	0	1
0	0	0	1	0	0	1	0
0	0	0	1	1	1	1	1
0	0	0	0	0	1	1	0
0	1	0	1	1	1	1	1
1	0	1	1	1	0	1	1
0	1	0	1	0	1	1	0

It is interesting to notice that for image *Baboon*, only AC channels of $(1, 3)$, $(1, 4)$ and $(2, 6)$ are decided to be i.i.d by the combination of all the tests. Especially, the rank test rejects the randomness of most AC channels. Taking the AC coefficient channel at (4,4) as an example, Fig.3 compares the values of the 500th to 700th samples in that channel for images *Baboon* and *Lena*, while Fig.4 compares the ranks corresponding to these samples. Looking at the values alone, we can only conclude that the dynamic range of the DCT coefficients is large in *Baboon* than in *Lena* since the former has large areas with noise-like textures. However, the ranks for *Baboon* are more clustered than for *Lena*, which hence means that the sample values for *Baboon* is more clustered that they appear because of the large dynamic range. The test statistic $NM = \sum_{i=1}^{n-1}[R(X_i) - R(X_{i+1})]^2$ for *Baboon* is therefore much smaller than that of an i.i.d sequence and H_0 is rejected.

VI. Conclusions

We combines decisions of six tests for randomness on deciding a sequence's randomness. Since all the tests only use partial information of the sequence, be it the signs of value differences or the ranks of the samples, we actually only have valid confidence when we

reject the null hypothesis of randomness. When we accept a null hypothesis, the acceptance is conditional on what aspects of the sequence we have examined.

The above tests of randomness can also be applied to some sequences from an original image and its watermarked version, which can be the DCT coefficients in each AC channels or the pixel value difference sequence from the image[8], to see whether and how the watermarking process affects the randomness of those sequences. However, some preliminary results show that the watermark detectable by [8] does not significantly change the decision on the randomness of selected sequences. So tests of randomness may not be a good steganalysis tool. The final conclusion still needs to be investigated.

References

[1] J. D. Gibbons, *Nonparametric methods for quantitative analysis*. New York: Holt, Rinehart and Winston, 1976.

[2] J. D. Gibbons and S. Chakraborti, *Nonparametric statistical inference*. New York: Marcel Dekker, 1992.

[3] H. V. Poor, *An introduction to signal detection and estimation*. New York: Springer-Verlag, 1988.

[4] T. M. Cover and J. A. Thomas, *Elements of information theory*. New York: John Wiley and Sons, Inc.

[5] S. Wegenkittl, " Entropy estimators and serial tests for ergotic chains", *IEEE Trans. Inform. Theory*, vol 47, no. 6, pp. 2480-2489, Sept. 2001.

[6] R. Bartels, "The rank version of von Neumann's ratio test for randomness", *Journal of the American Statistical Association*, vol.77, pp.40-46.

[7] F. Muller, "Distribution shape of two-dimensional DCT coefficient of natural images", *Electronic Letters*, vol. 29, no. 22, pp.1935-36, Oct. 23, 1993.

[8] Y. Wang and P. Moulin, "Steganalysis of block-DCT based stego-images", submitted to *Statistical Signal Processing Workshop, 2003*.