Repeated Hypothesis Testing on a Growing Data Set

G. V. TRUNK AND J. O. COLEMAN

Abstract—In many problems, especially when data are collected over a long period of time, hypothesis testing is done repeatedly as new data arrive. It is shown both for one particular problem and for a more general class of problems that if testing is performed each time on the total amount of data accumulated, the true simple null hypothesis will be rejected at least once as the number of tests approaches infinity. Furthermore, it is conjectured that the conclusion holds for most problems of interest in which the null hypothesis is simple.

Index Terms—Decision theory, detection theory, growing data set, hypothesis testing.

INTRODUCTION

In many problems, especially when data are collected over a long period of time, hypothesis testing is done repeatedly as new data arrive. If each test is performed only on data that have arrived since the previous test, all of the tests are independent. Consequently, if binary hypothesis tests with a constant false alarm rate were always performed, the probability of getting at least one false alarm would approach unity as the number of tests approached infinity. People who perform such testing are aware of this behavior and take it into account when setting the false alarm rate and interpreting the results.

There is, however, a related type of testing in which the certainty of an eventual false alarm is not obvious at all. In some systems in which data are accumulated slowly and a need for obtaining maximum performance is perceived, testing is performed each time on the total amount of data accumulated, with the total amount increasing with each test. Rather than successive tests being independent as in the first type of testing described, the latter tests grow more and more dependent as data are acquired (i.e., the probability of making the same decision when using n and n + 1 samples approaches one as n approaches infinity). The purposes of this correspondence are 1) to prove both for one particular problem of this type and for a more general class of problems of this type that the probability of an eventual false alarm is still unity and 2) to suggest that a similar proof could probably be applied to almost any problem of this type as long as the null hypothesis is simple.

PROBLEM

Given a sequence of independent samples \( \{x_i\} \) where the \( x_i \) are Gaussian distributed with mean \( \mu \) and unit variance, consider the following binary hypothesis test:

\[
H_0: \mu = 0 \\
H_1: \mu \neq 0.
\]

While the general proof is applicable to this simple problem, a different proof, which provides insight on why the probability of an eventual false alarm is unity, will be used. Since hypothesis \( H_1 \) is two-sided, no uniformly most powerful test exists. However, the accepted procedure is to perform the following test:

- accept \( H_0 \) if \( |S| < T \)
- accept \( H_1 \) if \( |S| > T \)

where the test statistic \( S \) is

\[
S = \sum_{i=1}^{n} x_i
\]

and the threshold \( T \) is set by

\[
\int_{T}^{\infty} e^{-z^2/2n} \sqrt{2\pi n} dz = \alpha/2.
\]

Solving for the threshold \( T \), it is of the form

\[
T = K\sqrt{n}
\]

where \( K \) is a constant determined by the false alarm rate \( \alpha \).

It will be assumed in the discussion which follows that \( H_0 \) is the true hypothesis. Let the probability of rejecting \( H_0 \) on or before the \( m \)th test be \( P_m \). Then

\[
P_m = P_{m-1} + (1 - P_{m-1}) C_m
\]

(1)

where \( C_m \) is the conditional probability that the null hypothesis is rejected on the \( m \)th test given that it has not been rejected by the preceding \( m - 1 \) tests. We will now show that \( P_m \to 1 \) as \( m \to \infty \). Since \( P_m \) is monotonic increasing and bounded by 1, \( P_m \) converges. However, if one performs a test every time one accumulates \( N_0 \) new samples, and if \( C_m \to 0^+ \) as \( m \to \infty \), then the rate of convergence of \( C_m \) determines whether \( P_m \to 1 \). Unfortunately, it is difficult to calculate \( C_m \) to determine its convergence rate. Consequently, a different approach will be taken. Rather than perform a test every time \( N_0 \) new samples are accumulated, an increasing number of samples will be accumulated between tests.

Specifically, let \( m \) be related to the sample size \( n \) by

\[
n = N_0^m
\]

(2)

where \( N_0 \) is any integer greater than 1. That is, if \( n \) samples are used for test \( m \), \( n^2 \) samples are used for test \( m + 1 \). By this

\[1\text{If } C_m \to e > 0 \text{ as } m \to \infty, \text{ then } P_m \to 1.\]
judicious choice of $n$ for each $m$, we will be able to show that $C_m$ is bounded away from zero and that $P_m \to 1$ as $m \to \infty$ since

$$P_\infty = P_\infty + (1 - P_\infty) C_\infty.$$  

The intuitive idea is that by squaring the number of samples in the test statistic, the outcome of the new test is only slightly dependent on the outcome of the previous test. It is therefore reasonable to expect that as $m \to \infty$, $C_m$ will approach $\alpha$, the unconditional probability of acceptance of $H_0$ on a given test. It should also be clear that $P_\infty$, determined with the present choice of $m$, represents a lower bound on the $P_\infty$ that would be obtained with tests performed at uniform intervals because of the more frequent opportunities for prior rejection of $H_0$ the latter situation would imply at any given sample size.

We will now demonstrate that $C_m \to \alpha$ as $m \to \infty$. First of all, $C_m = Pr\{(S_m > T_m) | S_m = -T_m\}$ where $S_i$ and $T_i$ are, respectively, the statistic and the threshold associated with the $i$th test, $T_i$ being calculated using (1) and (2). Since $S_{m+1}$ can be written as

$$S_{m+1} = S_m + \sum_{i=1}^{n} x_i$$

(3)

$C_{m+1}$ can be bounded by

$$C_{m+1} > B_{m+1} \equiv Pr\{S_{m+1} > T_{m+1} | S_m = -T_m\}$$

= $2Pr\{S_{m+1} > T_{m+1} | S_m = -T_m\}$

$$= 2Pr\{S_{m+1} > T_{m+1} | S_m = -T_m\}.$$  

(4)

This bound is obtained by noting that every sequence of samples $\{x_{n+1}, \ldots, x_{n+2}\}$ which causes $S_{m+1} > T_{m+1}$ given $S_m = -T_m$ will also cause $S_{m+1} > T_{m+1}$ given $|S_i| \leqslant T_i, i = 1, \cdots, m - 1$ since the last inequality implies $S_m \geqslant -T_m$. Similarly, every sequence of samples which causes $S_{m+1} < -T_{m+1}$ given $S_m = T_m$ will also cause $S_{m+1} < -T_{m+1}$ given $|S_i| \leqslant T_i, i = 1, \cdots, m$.

From (3), the conditional density of $S_{m+1}$ given $S_m = -T_m$ is Gaussian with mean $-T_m$ and variance $\sigma^2 = n^2 - n$. Now $B_{m+1}$ can be written in integral form as

$$B_{m+1} = 2 \int_{K_{n+1}}^{\infty} \frac{1}{\sqrt{2\pi} \sigma} e^{-((y + T_m)^2/2\sigma^2)} dy.$$  

Letting $z = (y + T_m)/\sigma$,

$$B_{m+1} = 2 \int_{K_1}^{\infty} \frac{1}{\sqrt{2\pi} \sigma} e^{-z^2/2} dz$$

where

$$T' = \frac{K_n + K\sqrt{n}}{\sqrt{n^2 - n}} = K \left[ 1 + \frac{1/n}{\sqrt{1 - 1/n}} \right].$$

Since $T' = (1 + \sqrt{2}) K$ when $n = 2$ and approaches $K$ as $n \to \infty$, $B_{m+1}$ and, consequently, $C_{m+1}$ are bounded away from zero. Thus, since $C_\infty \geqslant B_\infty = \alpha > 0$, $P_\infty$ must equal 1, i.e., the probability of rejecting the true hypothesis at least once approaches one as the number of samples approaches infinity.

The key element of the previous proof is that by squaring the number of samples involved in the statistic, the outcome of the new test is only slightly dependent on the outcome of the previous test. In fact, as the number of samples approaches infinity, the results of consecutive tests (involving the square of the previous number of samples) approach independence (i.e., as $n \to \infty$, $T' \to K$, and $C_{m+1} \to \alpha$). Since for large $m$ the probability of rejecting $H_0$ is essentially $\alpha$ on each test and consecutive tests are essentially independent, the probability of rejecting $H_0$ at least once will approach 1.

**Generalization**

Let $x_1, x_2, \ldots$ be an infinite sequence of independent and identically distributed samples with finite mean $\mu$ and unit variance. Consider the hypothesis test

$$H_0: \mu = 0$$

$$H_1: \mu > 0.$$  

While again it can be proved that $P_\infty = 1$ by squaring the number of samples between consecutive tests, the simplest proof involves the law of iterated logarithm [1] which states that with probability one

$$\lim_{n \to \infty} \sup \left\{ \frac{\sum_{i=1}^{n} x_i}{n} - n\mu_0 \right\} \sqrt{n} = 1.$$  

This implies that given an infinite sequence, the standard test statistic

$$\left( \frac{\sum_{i=1}^{n} x_i}{n} - n\mu_0 \right) \sqrt{n}$$

will be greater than any finite threshold $K$ an infinite number of times as $n \to \infty$. Thus, the true null hypothesis will be rejected an infinite number of times.

**Conjecture**

Unfortunately, we have not been able to obtain the general conditions under which the probability of an eventual false rejection is unity when the hypothesis testing problem does not involve a shift of means. We believe, however, that if

$$\lim_{m \to \infty} Pr\{U_m > T_m\} > 0$$

then $P_m \to 1$ as $m \to \infty$ where $U_m$ is the test statistic and $T_m$ is the corresponding threshold. (Note: $U_m$ need not be the sum of samples; for instance, $U_m$ may be the likelihood ratio.) On the other hand, if

$$\lim_{m \to \infty} Pr\{U_m > T_m\} = 0$$

we believe $P_m \to P < 1$ as $m \to \infty$. Since for a simple null hypothesis the threshold $T_m$ is set by

$$Pr\{U_m > T_m\} = \alpha$$

we believe that $P_m \to 1$ as $m \to \infty$ if $H_0$ is simple. However, if $H_0$ is a composite hypothesis, for example, $H_0: \mu \leqslant 0$, $H_1: \mu > 0$,

the false alarm rate is set at $\alpha$ for the point $\mu = 0$, (5) is true for $\mu = 0$, and (6) is true for all $\mu < 0$. Thus, if our conjecture is true, (6) implies $P_m \to P < 1$ as $m \to \infty$ for all $\mu < 0$.

**Summary**

It was shown for a particular problem that the true hypothesis will be rejected at least once as the number of tests approaches infinity if testing is performed each time on the total amount of data accumulated. The proof was then generalized to testing the mean using the law of iterative
logarithm. Furthermore, it was conjectured that any simple null hypothesis will eventually be rejected, and that the probability of eventual rejection of a composite null hypothesis will be less than unity for all points where the unconditioned false alarm rate is less than \( \alpha \).

### References


### Similarity Measures Between Strings

**Extended to Sets of Strings**

KAREN A. LEMONE

*Abstract—*Similarity measures between strings (finite-length sequences of symbols) are extended to apply to sets of strings in an intuitive way which also preserves some of the desired properties of the initial similarity measure. Two quite different measures are used in the examples: the first for applications where numerical computation of pointwise similarity is needed; the second for applications which depend more heavily on substring similarity. It is presumed throughout that alphabets may be nondenumerable; in particular, the unit interval \([0, 1]\) is used as the alphabet in the examples.

*Index Terms—*Nearest neighbor rule, nondenumerable alphabet, similarity measures between strings, substring.

### I. INTRODUCTION

This correspondence extends similarity measures between strings to their counterparts for sets of strings. The motivations for this research are threefold: 1) to continue the study of properties of similarity measures between strings, 2) to study the properties of similarity measures between sets of strings, and 3) to study the property-preserving characteristics of suitable methods of extending similarity measures between strings to similarity measures between sets of strings.

A string \( w \) is defined to be a finite-length sequence of symbols from some (possibly infinite) alphabet. The alphabet utilized in the examples is the unit interval \([0, 1]\). Five properties of similarity will be considered. The first four of these—\( c_0-c_3 \)—are suggested in [6] as "universal criteria which any reasonable similarity measure must satisfy." The last, \( c_4 \), is essentially a "triangle law" for similarity measures, useful for "pointwise" similarity measures such as Similarity Measure 1 defined below. In the following:

\[
w = w_1w_2 \cdots w_n, \quad v = v_1v_2 \cdots v_m, \quad u = u_1u_2 \cdots u_n,
\]

\[0 \leq w_i, v_i, u_i \leq 1\]

\[c_0: \quad 0 \leq s(w, v) \leq 1\]

\[c_1: \quad s(w, v) = \begin{cases} 0 & \text{if } w \text{ and } v \text{ are "totally dissimilar"} \\ 1 & \text{if } w \text{ and } v \text{ are "the same"} \end{cases} \]

\[c_2: \quad s(w, v) = s(v, w).\]

Manuscript received August 23, 1979; revised November 20, 1981.

The author is with the Department of Computer Science, Worcester Polytechnic Institute, Worcester, MA 01609.

\( c_3: \quad s(w, v) = s(w^R, v^R) \)

where \( w^R, v^R \) denote \( w \) and \( v \) reversed.

\[c_4: \quad s(w, v) + s(v, u) \leq 1 + s(w, u).\]

Two similarity measures will be represented here. The first is based on a normalized city-block metric, the second on the number of occurrences of identical substrings.

*Similarity Measure 1:*

\[s_1(w, v) = 1 - 1/n \sum_{i=1}^n |w_i - v_i|\]

where \( w = w_1w_2 \cdots w_n, \)

\[v = v_1v_2 \cdots v_m, \]

\[0 \leq w_i, v_i \leq 1.\]

Note that this measure is defined between strings of equal length and would not be meaningful in applications such as nonencoded text where \( |w_i - v_i| \) cannot be evaluated.

*Similarity Measure 2:*

\[
s_2(w, v) = \frac{\sum_{i=1}^n p(i)}{\sum_{i=1}^n i}
\]

where

\[w = w_1w_2 \cdots w_k \]

\[v = v_1v_2 \cdots v_m \]

\[n = \max(k, m)\]

\[p(i) \text{ is the number of substrings of length } i \text{ which } w \text{ and } v \text{ have in common.}\]

Here, the two strings to be compared need not be of equal length.

It can be shown that \( s_1 \) satisfies \( c_0-c_4 \) [11] and that \( s_2 \) satisfies \( c_0-c_3 \).

**Example 1:**

\[
s_1(1, 1) = 1 \quad s_2(1, 1) = 1
\]

\[
s_1(0, 1) = 0 \quad s_2(0, 1) = 0
\]

\[
s_1(\frac{1}{2}, \frac{3}{4}) = \frac{3}{4} \quad s_2(\frac{1}{2}, \frac{3}{4}) = 0
\]

\[
s_1(00, 11) = 0 \quad s_2(00, 11) = 0
\]

\[
s_1(00, \frac{1}{2} \frac{1}{2}) = \frac{1}{2} \quad s_2(00, \frac{1}{2} \frac{1}{2}) = 0
\]

\[
s_1(001, 010) = \frac{1}{3} \quad s_2(001, 010) = \frac{3}{3}
\]

\[
s_1(0010, 0110) = \frac{3}{4} \quad s_2(0010, 0110) = \frac{1}{2}
\]

\[
s_1(\frac{1}{2} \frac{1}{2} \frac{1}{2} \frac{1}{2}, 0110) = \frac{1}{2} \quad s_2(\frac{1}{2} \frac{1}{2} \frac{1}{2} \frac{1}{2}, 0110) = 0
\]

\[
s_2(00, 00) = \frac{1}{3} \quad s_2(001, 00100) = \frac{3}{5}
\]

### II. EXTENSION TO SETS OF STRINGS

In this section the similarity between a single string \( w \) and a set of strings \( S \) is defined according to the usual "nearest neighbor" rule from classification theory. Then this definition is in turn used to define the similarity between two sets of strings.

\[s(w, S) = \sup_{v \in S} s(w, v).\]