GENERALIZED BLOMOVIST CORRELATION: A NEW U STATISTIC AND SOME E--ETC(U)

NOV 81 S M DUNN

AFOSR-77-3271

AFOSR-TR-82-0268

END

5 82
GENERALIZED BLOMQVIST CORRELATION: 
A NEW U STATISTIC AND SOME EXAMPLES 

Stanley M. Dunn 
Computer Vision Laboratory 
Computer Science Center 
University of Maryland 
College Park, MD 20742 

ABSTRACT 

Generalized Blomqvist Correlation, a generalization of 
the double median test, is first formulated as a new U sta-
tistic with a lower variance. Several open questions are 
answered, and some examples are given.

The support of the U.S. Air Force Office of Scientific Research 
under Grant AFOSR-77-3271 is gratefully acknowledged, as is the 
help of Janet Salzman in preparing this paper.
Generalized Blomqvist Correlation: A New U Statistic and Some Examples

Stanley M. Dunn

Computer Vision Laboratory
Computer Science Center
University of Maryland
College Park, MD 20742

Mathematical & Information Sciences Directorate
Air Force Office of Scientific Research
Bolling AFB DC 20332

Approved for public release; distribution unlimited

Random variables
Dependence
U statistics
Distribution free statistics
Medial axis test

Generalized Blomqvist Correlation, a generalization of the double median test, is first formulated as a new U statistic with a lower variance. Several open questions are answered, and some examples are given.
1. Introduction

Generalized Blomqvist Correlation (GBC), originally presented by the author [1], is a nonparametric test for independence of two random variables, say X and Y. The xy plane is divided into $n^2$ regions by n-1 order statistics of each of the X and Y samples. This is a generalization of the double median test since we allow the number of partitions to increase by using the additional information of order statistics other than the median.

We will first present a brief review of the double median test and the generalization for the population. Secondly, we present the new $U$ statistic to calculate the sample correlation coefficient and present its mean and variance, along with a discussion of some asymptotic properties. Finally, several examples are presented comparing GBC with Kendall's $\tau$ and Spearman's $\rho_s$.

The problem which led to the development of GBC was that of determining the correlation between a sample of points in a digital image and a template in an attempt to locate edges in the original image. We desire to know whether or not the additional information provided by added order statistics will give us a more statistically significant estimate of the correlation.
2. **Medial axis correlation and the generalization**

We define as a measure of correlation the difference in probabilities

\[ \phi = \pi_s - \pi_d \]  

where

\[ \pi_s = \text{Prob}\{(x>x_0) \text{ and } (y>y_0) \text{ or } (x<x_0) \text{ and } (y<y_0)\} \]

\[ \pi_d = \text{Prob}\{(x>x_0) \text{ and } (y<y_0) \text{ or } (x<x_0) \text{ and } (y>y_0)\} \]  

The probability \( \pi_s \) is the probability that the deviations of \( x \) from the chosen \( x_0 \) and \( y \) from \( y_0 \) have the same sign. The probability \( \pi_d \) is the probability that the deviations have different signs. If we let \( x_0=x_m \) and \( y_0=y_m \) where \( x_m \) is the median of \( x \) and \( y_m \) is the median of \( y \), we have medial axis correlation.

The sample analog is constructed by dividing the xy plane into four regions by the lines \( x=x_m \) and \( y=y_m \). The sample correlation coefficient \( q' \) (after Blomqvist [2]) is given by

\[ q' = \frac{n_1-n_2}{n_1+n_2} \]  

where

\[ n_1 = \text{the number of samples } (x_i,y_i) \text{ such that } x_i<x_m \text{ and } y_i<y_m \]  

\[ n_2 = \text{the number of samples } (x_i,y_i) \text{ such that } x_i>x_m \text{ and } y_i>y_m \]  

Clearly, \( n_1+n_2=N \), the number of samples. Procedures for dealing with odd sample sizes are discussed in [1] and [2].
The above statistic is generalized by further subdividing the xy plane by additional equally spaced order statistics. The xy plane is divided into\( n^2 \) regions by \( n-1 \) \( x \) order statistics and \( n-1 \) \( y \) order statistics. \( \xi_i^{(n)} \) denotes the \( i \)th \( x \) order statistic from GBC of order \( n \) and \( \eta_i^{(n)} \) denotes the \( i \)th \( y \) order statistic from GBC of order \( n \). Thus

\[
\begin{align*}
\pi_s^{(n)} &= \text{Prob}\{ \bigwedge_{i=1}^{n} (\xi_{i-1}^{(n)} < x < \xi_i^{(n)}) \text{ and } (\eta_{i-1}^{(n)} < y < \eta_i^{(n)}) \} \\
\pi_d^{(n)} &= \text{Prob}\{ \bigwedge_{i=1}^{n} (\xi_{i-1}^{(n)} < x < \xi_i^{(n)}) \text{ and } (\eta_{n-i+1}^{(n)} < y < \eta_n^{(n)}) \}
\end{align*}
\]

where \( \xi_0^{(n)} \) and \( \eta_0^{(n)} \) are taken to be \( -\infty \) and \( \xi_n^{(n)} \) and \( \eta_n^{(n)} \) are taken to be \( +\infty \).

Let \( r_{ij}^{(n)} \) denote the region of the xy plane where

\[ \xi_{i-1}^{(n)} < x < \xi_i^{(n)} \text{ and } \eta_{j-1}^{(n)} < y < \eta_j^{(n)} . \]

Let \( |r_{ij}^{(n)}| \) denote the number of samples in \( r_{ij}^{(n)} \). The sample statistic \( q'_n \) is computed by

\[
q'_n = \frac{1}{N} \left( \sum_{i=1}^{n} |r_{ij}^{(n)}| - \sum_{i=1}^{n} |r_{i,n-i+1}^{(n)}| \right) \quad (5)
\]

where \( N \) is the number of samples and \( n \) is the number of subdivisions along the \( x \) or \( y \) axis, which results in \( n^2 \) regions in the xy plane.

An alternative method for computing the sample analog is presented in the following section. This will allow us to investigate the asymptotic properties of the statistics including the asymptotic relative efficiency.
3. **The U statistic for computing \( q'_n \)**

In equation (5) we computed the sample correlation coefficient by counting the number of samples that lie in the regions on the diagonals of the xy plane; see Figure 1. Those points in regions on diagonal 1 are counted in the first sum of equation (5), whereas those points in regions along diagonal 2 are counted in the second sum. Points along diagonal 1 are seen as contributing to positive correlation, and points along diagonal 2 are seen as contributing to negative correlation.

We can reformulate the computation of \( q'_n \) by viewing this statistic as a sum of functions of the sample points, \( \phi(x_i, y_i) \). We let \( \phi(\cdot) \) be 1 if the point \( (x_i, y_i) \) is in one of the regions along diagonal 1, \( \phi(\cdot) \) is -1 if \( (x_i, y_i) \) is in a region along diagonal 2, and otherwise \( \phi(\cdot) \) is zero. This reformulated statistic will be referred to as \( U \) which is given by

\[
U = \frac{1}{N} \sum_{i=1}^{N} \phi(x_i, y_i) \tag{6}
\]

where \( N \) is the number of sample points, and \( \phi(\cdot) \) is described above. We now show how to compute \( \phi(x_i, y_i) \).

Recall that we have subdivided each axis (x or y) into \( n \) intervals. A point \( (x_i, y_i) \) is said to be in x(y) interval \( k \) if \( \xi_{k-1}^{(n)} < x_i < \xi_k^{(n)} \) \( (\eta_{k-1}^{(n)} < y_i < \eta_k^{(n)}) \). Clearly \( k \) varies from 1 to \( n \). Thus a point is in one of the regions along diagonal 1 if its x interval number equals its y interval number. Also a point is in one of the regions along diagonal 2 if its x interval number equals \( n+1-(y \text{ interval number}) \).
Clearly, \( N/n \) is the number of points in each interval along either axis. Thus the interval number of \( x_i \) is 1 if the rank of \( x_i \) (denoted \( R_i \)) is between 1 and \( N/n \). We compute the \( x \) interval number by

\[
\text{x interval number} = \left\lfloor \frac{R_i}{(N/n)} \right\rfloor . \tag{7}
\]

Equivalently the \( y \) interval number is computed by replacing \( R_i \) with \( Q_i \), the rank of \( y_i \). We can now express the condition that the \( x \) interval number equals the \( y \) interval number by

\[
\left\lfloor \frac{R_i}{(N/n)} \right\rfloor = \left\lfloor \frac{Q_i}{(N/n)} \right\rfloor \tag{8}
\]

where \( \lfloor x \rfloor \) is the ceiling function, the least integer \( \geq x \).

\( \phi(x_i, y_i) \) is defined as

\[
\phi(x_i, y_i) = \begin{cases} 
1, & \text{if } \left\lfloor \frac{R_i}{(N/n)} \right\rfloor = \left\lfloor \frac{Q_i}{(N/n)} \right\rfloor \\
0, & \text{otherwise} \\
-1, & \text{if } \left\lfloor \frac{R_i}{(N/n)} \right\rfloor = n+1- \left\lfloor \frac{Q_i}{(N/n)} \right\rfloor
\end{cases} \tag{9}
\]

We are careful to note that \( \phi \) has one sample as an argument. We denote by \( Z_i \) the single sample point \((x_i, y_i)\) and we write \( q'_n ) as

\[
q'_n = \frac{1}{N} \sum_{i=1}^{N} \phi(Z_i) \tag{10}
\]

for \( \phi(\cdot) \) defined above. By showing that \( q'_n ) is estimable of degree 1, the statistic as given in equation (9) is a U statistic. Recall that this statistic is being used to test the hypotheses.
\[ H_0: \ F(x,y) = F(x)F(y) \]
\[ H_1: \ F(x,y) \neq F(x)F(y) \]
or some subclass of \( H_1 \). We now show

**Theorem:** \( q'_n \) is estimable of degree 1.

**Proof:** To show that \( q'_n \) is estimable of degree 1, we show that there exists a function \( \phi \) of one argument such that

\[ E(\phi(Z_1)) = q'_n \]

We let \( \phi(Z_1) \) be the function defined in equation (9). Under the null hypothesis of independence, all regions of the x-y plane are equally likely; thus

\[ E(\phi(Z_1)) = 0 \]

which is precisely \( q'_n \) under the null hypothesis. Hence \( q'_n \) is estimable of degree 1. \[ \]

From Hoeffding [3], we know that \( q'_n \) is asymptotically normally distributed with mean given by

\[ E(q'_n) = 0 \]

(11)

and the variance is

\[ \text{Var}(q'_n) = \frac{1}{N}\xi_1 \]

(12)

where \( \xi_1 \) is

\[ \xi_1 = E(\phi^2(Z_1)) - (q'_n)^2 \]

(13)

To determine the variance of the statistic \( q'_n \), we first compute

\[ \xi_1 = E(\phi^2(Z_1)) - (q'_n)^2 \]

\[ = \frac{2n}{n^2} - 0 \]

\[ = \frac{2}{n} \]

(14)
since $\phi^2(z_1)$ is 0 or 1, and as before $q'_n(n)$ is 0. The variance is

$$\text{var}(q'_n(n)) = \frac{1}{N} \xi_1$$

$$= \frac{1}{N} \left( \frac{2}{n} \right)$$

$$= \frac{2}{Nn}$$  \hspace{1cm} (15)

Appendix I contains tables of critical values of $q'_n(n)$ for $n=2$ to 8 and for sample sizes from $N=2$ to 30. The significance levels indicated are one-tailed; thus for two-tailed tests they should be doubled. These tables will be used in the examples to follow. First, we present some asymptotic results concerning $q'_n(n)$.
4. **Asymptotic properties of \( q'_n \)**

The asymptotic relative efficiency (ARE) of two statistical tests is a ratio of the sample sizes required to achieve the same level of statistical significance. The sample size in using the first test need only be \((100 \cdot \text{ARE})\%\) of the sample size of the second test to achieve the same statistical significance. We investigate the ARE of \( q'_n \) (Generalized Blomqvist Correlation) relative to \( q' \) (medial axis correlation). The reader is referred to Gibbons [4] for a detailed explanation with examples. The ARE is defined as

\[
\text{ARE}(q'_n, q') = \lim_{N \to \infty} \frac{e(q'_n)}{e(q')} \quad (16)
\]

where the efficacy \( e(\cdot) \) is

\[
e(T) = \left[ \frac{dE(T)/d\theta}{\sigma^2(T)|_{\theta=\theta_0}} \right]^2 \quad (17)
\]

for a test statistic \( T \). For both \( q'_n \) and \( q' \) \( dE(T)/d\theta \) is 1, so that

\[
\text{ARE}(q'_n, q') = \lim_{N \to \infty} \frac{\sigma^2(q'_n)}{\sigma^2(q'_n)} \quad (18)
\]

The variance of \( q'_n \) is given in equation (15), and from Blomqvist [2] the variance of \( q' \) is

\[
\frac{4a_0(1-2a_0)}{k}
\]

where \( a_0 \) is \( \text{Prob}\{x<x_m \text{ and } y<y_m\} \) in the neighborhood of \((x_m, y_m)\). The ARE is
\[
\text{ARE}(q'_n, q') = \lim_{N \to \infty} \frac{\text{lim}_{H_1 \to H_0} \left[ 4a_0(1-2a_0) \right]/k}{2/Nn}
\]

\[
= \lim_{N \to \infty} \frac{4a_0(1-2a_0)}{k} \cdot \frac{Nn}{2}
\]

(19)

Since \(N=2k\) (see Blomqvist [2])

\[
\text{ARE}(q'_n, q') = \lim_{N \to \infty} \frac{(4a_0(1-2a_0) \cdot n}{H_1 \to H_0}
\]

(20)

Now taking the limit, as \(H_1 \to H_0\), \(a_0 = \frac{1}{4}\) and

\[
\text{ARE}(q'_n, q') = \frac{n}{2}
\]

(21)

The sample size for \(q'_n\) need only be \(\frac{2}{n}\) times the sample size of \(q'\) to achieve the same level of significance. We can check this by recalling that for \(n=2\), GSC is exactly medial axis correlation and we would expect the ARE to be 1, which it is. In the limit, there would be one sample in each interval when \(n=N\); thus

\[
\text{ARE}(q'_N, q') = N/2
\]

(22)
5. Examples using Generalized Blomqvist Correlation

In this section, we will present several examples of using GBC in practice. We will compare GBC of order 4 with Spearman's rho, Kendall's tau, and the medial axis correlation. Sample sizes of 20 and 30 will be used from bivariate normal and exponential distributions with known correlation.

The first example is Hajek's data [5]. This sample is of size 20 with known correlation of $\rho = \frac{1}{2}$. Hajek showed that with both Spearman's test and Kendall's test we could reject the hypothesis of independence at $\alpha = .10$, but we were forced to accept the null hypothesis at $\alpha = .05$. Hajek's results for the quadrant test which is precisely GBC of order 2 indicate that the null hypothesis cannot be rejected at $\alpha = .10$. The same is true of GEC of order 4. The results of this and the following examples are presented in Figure 2. In all cases, we are testing the null hypothesis of independence versus the alternate of positive dependence.

In Figure 2, we present results for two more bivariate normal distributions and three bivariate exponential distributions whose form is from Mardia [6]. The real correlation coefficient is indicated, along with those computed using Spearman's, Kendall's, the medial axis test, and GBC of order 4. GBC(2) is the medial axis test which is identically GBC of order 2, and GBC(4) is GBC of order 4. The associated
significance levels are indicated below the correlation coefficient, where ns means not significant at $\alpha=.10$. The first entry of the three normal distribution examples is Hajek's data. Since sample sizes were varied, the sample size is indicated along with the actual correlation.
6. Conclusion

In this paper a generalization of the medial axis correlation test was presented. By Hoeffding's results on U-statistics, we were able to determine the asymptotic distribution of GBC. The asymptotic relative efficiency of GBC to medial axis correlation was given and shown to be the ratio of the number of intervals used in GBC to the number of intervals used in the medial axis correlation, which is two. Thus one can see that the ARE of $q_{(n_1)}^1$ to $q_{(n_2)}^1$ is simply $n_1/n_2$.

The critical levels of $q_{(n)}^1$ for $n = 2$ to 8 were tabulated and they are used in several examples presented in this paper. The null hypothesis is that of independence versus an alternative of positive dependence. The first three examples are from a bivariate normal distribution with known correlation. The results indicate that any order of GBC, involving the medial axis test, is inferior to Spearman's rho and Kendall's tau. This is expected since it is well known that the medial axis is less efficient than either Spearman's or Kendall's test relative to the normal alternatives.

The final three examples were taken from bivariate exponential distributions with known correlation. In these three cases, both GBC(2) and GBC(4) performed better than either Spearman's or Kendall's test. In all three examples, both GBC(2) and GBC(4) achieved a higher level of significance. The second example with $p=0.5$ and $N=30$ shows the real power
of GBC. We were able to achieve a higher level of significance by increasing the number of order statistics used to partition the xy plane. This behavior is intuitively possible as one can see upon closer examination of a single quadrant of the xy plane in the medial axis test.

If we partition the xy plane for GBC(4), the above chosen quadrant is itself divided into four quadrants. Two of these are "on" diagonal (upper left, lower right) and the other two are "off" diagonal (lower left, upper right). It is natural to expect that if the two random variables are positively correlated, the points that lie in the chosen quadrant will actually lie in the "on" diagonal regions when the quadrant is further partitioned. This is true for the upper left and lower right quadrants. For the lower left and upper right quadrants, if the two random variables are negatively correlated, we expect the points to lie in the "off" diagonal regions when partitioned. Figure 3 illustrates this concept.

The purpose of developing Generalized Blomqvist Correlation was to investigate whether or not the additional information provided by additional order statistics will give a test with a higher level of statistical significance. The results presented in this paper illustrate that this is indeed possible.
References

1. Dunn, S. M., "Generalized Blomqvist Correlation," TR-1043, Computer Vision Laboratory, Computer Science Center, University of Maryland, College Park, MD, April 1981.


Figure 1. Relevant regions of the xy plane.
<table>
<thead>
<tr>
<th>Actual correlation</th>
<th>$\rho_s$</th>
<th>$t$</th>
<th>GBC(2)</th>
<th>GBC(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>.5, N=20</td>
<td>.3278</td>
<td>.2526</td>
<td>.2</td>
<td>.1</td>
</tr>
<tr>
<td></td>
<td>.1</td>
<td></td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>.5, N=30</td>
<td>.5537</td>
<td>.4023</td>
<td>.2000</td>
<td>.1333</td>
</tr>
<tr>
<td></td>
<td>.001</td>
<td></td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>.25, N=30</td>
<td>.3593</td>
<td>.2644</td>
<td>.0667</td>
<td>.0000</td>
</tr>
<tr>
<td></td>
<td>.001</td>
<td></td>
<td>ns</td>
<td>ns</td>
</tr>
</tbody>
</table>

**Bivariate Normal Samples**

<table>
<thead>
<tr>
<th>Actual correlation</th>
<th>$\rho_s$</th>
<th>$t$</th>
<th>GBC(2)</th>
<th>GBC(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>.5, N=20</td>
<td>.0000</td>
<td>.0105</td>
<td>.4000</td>
<td>.3000</td>
</tr>
<tr>
<td></td>
<td>ns</td>
<td></td>
<td>.05</td>
<td>.05</td>
</tr>
<tr>
<td>.5, N=30</td>
<td>.2800</td>
<td>.1860</td>
<td>.3333</td>
<td>.2667</td>
</tr>
<tr>
<td></td>
<td>.10</td>
<td></td>
<td>.05</td>
<td>.025</td>
</tr>
<tr>
<td>.25, N=30</td>
<td>.2133</td>
<td>.2230</td>
<td>.4667</td>
<td>.2667</td>
</tr>
<tr>
<td></td>
<td>ns</td>
<td>.05</td>
<td>.01</td>
<td>.025</td>
</tr>
</tbody>
</table>

**Bivariate Exponential Samples**

Figure 2. Experimental results of correlation tests.
a. Partitioning under the medial axis test.

b. Partitioning under GBC(4).

Figure 3. xy plane partitioning.
Appendix 1: Tables of critical levels of the statistic $q(n)$ for $n=2$ to $8$