DEPARTMENT OF STATISTICS

University of Wisconsin
1210 West Dayton Street
Madison, WI 53706-1685

TECHNICAL REPORT NO. 978

MAY 1997

Probability Models and Limit Theorems for Random Interval Graphs with Applications to Cluster Analysis

by

Bernard Harris and Erhard Godehardt
PROBABILITY MODELS AND LIMIT THEOREMS FOR RANDOM INTERVAL GRAPHS WITH APPLICATIONS TO CLUSTER ANALYSIS

Bernard Harris¹, Erhard Godehardt²

¹University of Wisconsin, Madison, Wisconsin, U.S.A.
²Heinrich Heine Universität, Düsseldorf, Germany

Abstract

Assume that n k-dimensional data points have been obtained and subjected to a cluster analysis algorithm. A potential concern is whether the resulting clusters have a "causal" interpretation or whether they are merely consequences of "random" fluctuation. In this report, the asymptotic properties of a number of potentially useful combinatorial tests based on the theory of random interval graphs are described. Some preliminary numerical results illustrating their possible application as a method of resolving the above question are provided. Regrettably, much of the mathematical development is omitted here, due to space limitations and will be published elsewhere.

1. Introduction and Summary

Let $F_X(x)$ be a cumulative distribution function on $E_k$, k-dimensional Euclidean space. We assume that $F_X(x)$ is absolutely continuous with respect to k-dimensional Lebesgue measure and denote the corresponding probability density function by $f_X(x)$. Assume that a random sample of size $n$ has been obtained from $F_X(x)$ and denote the realizations by $x_1, x_2, ..., x_n$. In cluster analysis, similar objects are to be placed in the same cluster. We will interpret similarity as being close with respect to some distance on $E_k$. The relationship between graph theory and cluster analysis has been described in the books by Bock (1974) and Godehardt (1990). Mathematical results related to those used here are given in Eberl and Hafner (1971), Hafner (1972), Godehardt and Harris (1995), Jammalamadaka and Janson (1986), Jammalamadaka and Zhou (1990) and Machara (1990).

In order to proceed, we need to introduce some notions from graph theory.

2. Graph Theoretic Concepts.

A graph $G = (V, E)$ is defined as follows. $V$ is a set with $|V| = n$ and $E$ is a set of (unordered) pairs of elements of $V$. The elements of $V$ are called the vertices of the graph $G$ and the pairs in $E$ are referred to as the edges of the graph $G$. With no loss of generality, we can assume $V = \{1, 2, ..., n\}$. For the purposes at hand, we choose a distance $\rho$ on $E_k$ and a threshold $d > 0$. Then for $i \neq j$, place $(i, j) \in E$ if $\rho(x_i, x_j) < d$. Since $x_1, x_2, ..., x_n$ are
realizations of random variables, the set \( E \) is a random set and the graph \( G \) is a random graph. In particular, these graphs are generalizations of interval graphs. Specifically, if \( I_1, I_2, \ldots, I_n \) are intervals on the real line, then the interval graph \( G(I_n) \) is defined by \( V = \{1, 2, \ldots, n\} \) and \((i,j) \in E \) if \( I_i \cap I_j \neq \emptyset \), \( 1 \leq i < j \leq n \). Thus, for the model under consideration, if \( k = 1 \), the intervals \( I_i \), \( i = 1, 2, \ldots, n \), are the intervals \([x_i - d/2, x_i + d/2]\). Let \( V_m \subset V \) with \(| V_m | = m < n \). \( K_{m,d} \) is a complete subgraph of order \( m \), if all \( \binom{m}{2} \) pairs of elements of \( V_m \) are in \( E \). If \( m = 1 \), then \( K_{1,d} \) is a vertex, if \( m = 2 \), then \( K_{2,d} \) is an edge and if \( m = 3 \), then \( K_{3,d} \) is called a triangle. A complete subgraph of order \( m \) is said to be a maximal complete subgraph, denoted by \( K^*_{m,d} \), if there is no vertex in \( V \neg V_m \) such that adjoining that vertex to \( V_m \) results in a complete subgraph of order \( m + 1 \). A vertex has degree \( \nu \), \( \nu = 0, 1, 2, \ldots, n - 1 \), if there are exactly \( \nu \) edges incident with that vertex. If \( \nu = 0 \), then that vertex is said to be an isolated vertex.


We now describe the probability that a specified set of \( m \) vertices form a \( K_{m,d} \) or a \( K^*_{m,d} \). With no loss of generality, we can assume that these vertices are \( \{1, 2, \ldots, m\} \). Then,

\[
P\{\max_{1 \leq i \leq m} X_i - \min_{1 \leq i \leq m} X_i \leq d\}
\]

\[
= m \int_{-\infty}^{\infty} [F(x + d) - F(x)]^{m-1} f(x) \, dx.
\]

Similarly,

\[
P\{(1, 2, \ldots, m) \text{ is a } K^*_{m,d}\} = m(m-1) \int_{-\infty}^{\infty} \int_{x_1}^{x_1+d} \int_{x_2}^{x_2+d} [F(x_2) - F(x_1)]^{m-2} \cdot [1 - F(x_1 + d) + F(x_2 - d)]^{m-3} f(x_1) f(x_2) \, dx_2 \, dx_1.
\]

The probability that a specified vertex forms a \( K^*_{1,d} \) (i.e. is isolated) is

\[
P\{1 \text{ is isolated}\} = \int_{-\infty}^{\infty} \left[ 1 - [F(x + d) - F(x - d)] \right]^{m-1} f(x) \, dx.
\]

The probability that a specified vertex has degree \( \nu \), \( \nu = 0, 1, \ldots, n - 1 \) is

\[
P\{\text{vertex } 1 \text{ has degree } \nu\}
\]

\[
= \int_{-\infty}^{\infty} \binom{n-1}{\nu} [F(x + d) - F(x - d)]^{\nu} [1 - F(x + d) + F(x - d)]^{n-\nu-1} f(x) \, dx.
\]

To obtain asymptotic approximations to the above distributions, some assumptions concerning the behavior of the probability density function \( f_X(x) \) are needed. Hence we will assume that the probability density function is uniformly continuous on every compact subset of the carrier set for \( X \) and let \( f'X(x) \) exist and be uniformly bounded on the carrier set of \( X \).

4. Asymptotic Behavior of Probability Distributions of Properties of Random Interval Graphs
In this section, we examine the asymptotic behavior of the probability distributions introduced in the preceding section, under the conditions $n \to \infty$ and $d \to 0$ and also assuming the regularity conditions for $f(x)$ given above. The asymptotic probability ($d \to 0$) that the vertices $\{1, 2, \ldots, m\}$ form a $K_{m, d}$ is

$$m \, d^{m-1} \int_{-\infty}^{\infty} f^m(x) \, dx. \quad (5)$$

The asymptotic probability ($d \to 0$, $n \to \infty$, so that $nd \to 0$) that the vertices $\{1, 2, \ldots, m\}$ form a $K_{m, d}^*$ is

$$d^{m-1} \int_{-\infty}^{\infty} f^m(x) \{ m + (n - m) f(x) [(m-1)d - 2m] \} \, dx. \quad (6)$$

The asymptotic probability ($d \to 0$, $n \to \infty$, so that $nd \to 0$) that vertex 1 is isolated is

$$1 - 2nd \int_{-\infty}^{\infty} f^2(x) \, dx. \quad (7)$$

The asymptotic probability ($d \to 0$, $n \to \infty$, so that $nd \to 0$) that vertex 1 has degree $\nu$ is

$$\left( \frac{n-1}{\nu} \right) (2d)^\nu \int_{-\infty}^{\infty} (f(x))^{\nu+1} \{ 1 - 2nd f(x) \} \, dx. \quad (8)$$

5. Asymptotic Poisson Distributions and Asymptotic Normal Distributions

In this section, we give specific limiting distributions for the graph theoretic characteristics previously described.

If $n^m d^{m-1} \to \tau > 0$ as $n \to \infty$ and $d \to 0$, then the number of complete subgraphs of order $m$ is asymptotically Poisson distributed with mean $\lambda = \left( \begin{array}{c} n \\ m \end{array} \right) m d^{m-1} \int_{-\infty}^{\infty} f^m(x) \, dx$. If $n \to \infty$ and $d \to 0$ so that $n^m d^{m-1} \to \tau > 0$, then the number of maximal complete subgraphs of order $m$ has an asymptotic Poisson distribution with expected value $\left( \begin{array}{c} n \\ m \end{array} \right) m d^{m-1} \int_{-\infty}^{\infty} f^m(x) \, dx$, $m = 2, 3, \ldots$. If $n \to \infty$ and $d \to 0$ so that $dn/(\ln n) \to \tau_1 > 0$, then the number of isolated vertices has an asymptotic Poisson distribution with expected value $n \int_{-\infty}^{\infty} e^{-2ndf(x)} f(x) \, dx$. If $n \to \infty$ and $d \to 0$ so that $n^{\nu+1} d^{\nu} \to \tau_2 > 0$, then the number of vertices of degree $\nu$, $\nu = 2, 3, \ldots$ is asymptotically Poisson distributed with mean $n^{\nu+1} d^{\nu} 2^\nu \int_{-\infty}^{\infty} [f(x)]^\nu \, dx$. 
For each of these characteristics, limiting normal distributions have also been obtained. Clearly, if the asymptotic Poisson means tend to infinity, then the random variables described above, when suitably normalized, will be asymptotically normally distributed with mean zero and variance unity. Alternatively, the theory of U-statistics can also be successfully exploited to establish normal limits. Due to space limitations, the specific details will be provided in a more extensive manuscript.

6. Multidimensional Extensions

In this section, we assume that $X_1, X_2, ..., X_n$ are independent, identically distributed random variables, taking values in $E_k$, Euclidean $k$-space, $k > 1$. We assume that these random variables are distributed by the cumulative distribution function $F_X(x)$, where $F_X(x)$ is absolutely continuous with respect to $k$-dimensional Lebesgue measure and has probability density function $f_X(x)$. We will also assume that an $L_p$ norm is specified on $E_k$, $1 \leq p \leq \infty$. For the mathematical development, the primary difficulty in making the transition to more than one dimension, is that the realizations of the random variables can no longer be ordered.

As before, the vertices $\{1, 2, ..., m\}$ form a complete subgraph of order $m$, denoted by $K_{m,d}$, whenever $\rho(x_i, x_j) < k$, $1 \leq i, j \leq m$. Let $A(K_{m,d})$ be the event that the vertices $\{1, 2, ..., m\}$ form a $K_{m,d}$. Let $S(x, r)$ be the ball of radius $r$ with center at $x$. Let $B(m, d) = \{x_1, x_2, ..., x_m \subset \rho(x_j, x_1) < k, j = 2, ..., m\}$. Let $C(m, d)$ be the event that $x_j \in S(x_1, \frac{d}{2})$, $j = 2, ..., m$. It is easy to see that $C(m, d) \subset A(K_{m,d}) \subset B(m, d)$. Therefore, we can write the following:

$$P(C(m, d) \subset P(A(K_{m,d}) \leq B(m, d))$$

and hence

$$\int_{E_k} \left[ \int_{S(x_1, \frac{d}{2})} \prod_{i=2}^{m} f(x_i) dx_i \right] f_X(x_1) dx_1 \leq P(A(K_{m,d}) \leq \int_{B(m,d)} \left[ \int_{S(x_1, \frac{d}{2})} \prod_{i=2}^{m} f_X(x_i) dx_i \right] f_X(x_1) dx_1.$$

Similarly, we can obtain upper and lower bounds on the probability that a given set of $m$ vertices form a maximal complete subgraph of order $m$.

It can be shown that these inequalities are adequate for establishing the limiting behavior of the number of complete subgraphs of order $m$ or the number of maximal complete subgraphs of order $m$. This is accomplished by using indicator functions and the method of moments for establishing Poisson limits and the theory of U-statistics for establishing normal limits.

Further, the probability that vertex $1$ is of degree $\nu > 0$ is given by
\[ P\{ \text{vertex 1 is of degree } \nu \} = \binom{n-1}{\nu} \int_{E_k} \left( \int_{s(d,x)} f_X(y) dy \right)^\nu. \]

\[ \left\{ 1 - \int_{s(d,x)} f_X(w) dw \right\}^{n-\nu} f_X(x) dx. \]

Similarly, the probability that vertex 1 is isolated is given by

\[ P\{ \text{vertex 1 is isolated} \} = \int_{E_k} \left( 1 - \int_{s(d,x)} f_X(y) dy \right)^{n-1} f_X(x) dx. \]

As in the one-dimensional case, in order to obtain uniform approximations to the above integrals, various smoothness conditions on \( f_X(x) \) are needed; these include the requirement that \( f_X(x) \) is uniformly continuous on every compact set and that the partial derivatives are uniformly bounded. Then, for example, it is possible to establish that the number of complete subgraphs of order \( m > 1 \), will have an asymptotic Poisson distribution whenever \( d^{n(m-1)/m} \rightarrow \tau > 0 \), as \( n \rightarrow \infty \) and \( d \rightarrow 0 \).


A widely used application of cluster analysis is the detection of mixtures of distributions. Specifically, assume \( f_X(x) = \alpha f_1(x) + (1-\alpha)f_2(x) \), \( 0 \leq \alpha \leq 1 \), where \( f_1(x) \) and \( f_2(x) \) are distinct probability density functions. Then, under various conditions on \( \alpha \), \( f_1(x) \) and \( f_2(x) \), a cluster analysis would be expected to detect two clusters; one cluster consisting largely of data from \( f_1(x) \) and the other consisting primarily of data from \( f_2(x) \). This leads in a natural way to the following question in statistical inference. Is the data that has been obtained from a homogeneous population (i.e. \( \alpha = 0 \) or \( 1 \), or \( f_1(x) = f_2(x) \)), or from a nontrivial mixture of two populations?

To illustrate some of the techniques proposed in this report as potential test criteria, the following simple case is treated. Let \( f_1(x) \) and \( f_2(x) \) be univariate normal distributions. The mean of \( f_1(x) \) will be zero and we will vary the mean \( \mu \) and variance \( \sigma^2 \) of \( f_2(x) \).

For this case, we will examine the asymptotic distribution of the number of complete subgraphs of order \( m \) and provide specific numerical values for \( m = 2 \). These values will be "normalized", so that they can be interpreted for all "large" values of \( n \).

Then, the probability that \( m \) specified vertices form a \( K(m,d) \) is

\[ P\{ K(m,d) \} = m \int_{-\infty}^{\infty} \left[ F(x+d)-F(x) \right]^{m-1} f(x) dx, \]
where \( f(x) = \alpha f_1(x) + (1-\alpha)f_2(x) \) and \( F(x) \) is the corresponding cumulative distribution function. Using the approximation given above, we have

\[
P\{K(m,d)\} \approx 2md^{m-1} \int_{-\infty}^{\infty} f^m(x) \, dx.
\]

For the specific case at hand,

\[
f(x) = \alpha f_1(x) + (1-\alpha)f_2(x)
\]

and thus,

\[
P\{K(m,d)\} \approx 2md^{m-1} \sum_{k=0}^{m} \binom{m}{k} \frac{\alpha^k(1-\alpha)^l}{(2\pi)^{\frac{m}{2}l} \sigma^l} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2} + \frac{\alpha^2}{2\alpha(\sigma^2+d)}} \, dx.
\]

where \( l = m - k \). Evaluating the integral, we obtain,

\[
P\{K(m,d)\} \approx 2md^{m-1} \sum_{k=0}^{m} \binom{m}{k} \frac{\alpha^k(1-\alpha)^l}{(2\pi)^{\frac{m}{2}l} \sigma^l} \frac{\frac{\alpha^2}{2\alpha(\sigma^2+d)}}{\sqrt{\pi^2 \sigma^4 k + l}}.
\]

Hence, the expected number of complete subgraphs of order \( m \) is approximately given by

\[
S(m) = 2md^{m-1} \left( \sum_{k=0}^{m} \binom{m}{k} \frac{\alpha^k(1-\alpha)^l}{(2\pi)^{\frac{m}{2}l} \sigma^l} \frac{\frac{\alpha^2}{2\alpha(\sigma^2+d)}}{\sqrt{\pi^2 \sigma^4 k + l}} \right).
\]

In a similar manner, we can approximate the variance of \( K(m,d) \) and specifically for \( K(2,d) \), we get

\[
\frac{4m^3d^2}{2\pi\sigma^{l+1}((\sigma^2+d)^{l+1})} \sum_{k=0}^{3} \binom{3}{k} \alpha^k(1-\alpha)^l e^{\frac{\alpha^2}{2\alpha(\sigma^2+d)}} - \frac{1}{2} \frac{\sigma^2}{d^2}.
\]

Below is a brief table of values for the asymptotic means and variances of the number edges, when \( n^2d \to \infty \), and hence the asymptotic normal approximations are valid. To provide specific values, I have set \( \alpha = \frac{1}{2} \) and \( \sigma^2 = 1 \). So that the results can be interpreted for "large \( n \" general, the means have been normalized by the factor \( n^2d \) and the variances by the factor \( 4n^3d^2 \). In particular note that when \( n^2d \to \infty \), rescaling the number of edges by this factor, forces the asymptotic variances to zero. Thus, asymptotic consistency of tests for mixtures has been established under the given conditions.
## TABLE OF ASYMPOTIC MEANS AND VARIANCES

<table>
<thead>
<tr>
<th>$\mu$</th>
<th>Means (Normalized)</th>
<th>Variances (Normalized)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.282</td>
<td>.092</td>
</tr>
<tr>
<td>1</td>
<td>.251</td>
<td>.072</td>
</tr>
<tr>
<td>2</td>
<td>.193</td>
<td>.041</td>
</tr>
<tr>
<td>3</td>
<td>.156</td>
<td>.026</td>
</tr>
<tr>
<td>$\infty$</td>
<td>.141</td>
<td>.023</td>
</tr>
</tbody>
</table>

**References**


