The MODECLUS Procedure: Nonparametric Significance Tests for Clusters

A. H. Kuo, SAS Institute Inc., Cary, NC

ABSTRACT
The MODECLUS procedure, available in release 6.08 of SAS/STAT® software, clusters the observations in a SAS® data set using any of several algorithms based on nonparametric density estimates. The data can be numeric coordinates or distances. PROC MODECLUS can perform approximate significance tests for the number of clusters and can hierarchically join nonsignificant clusters.

This paper gives a brief overview of the density estimation, clustering methods, and significance tests in PROC MODECLUS. How PROC MODECLUS differs from other clustering methods is explained, and simulations results for sample sizes ranging from 20 to 2000 are shown.

PROBLEM
The purpose of the significance tests is as follows: given a random sample of objects from a population, you would like to obtain an estimate of the number of the clusters in the population such that the probability in repeated sampling that the estimate exceeds the true number of clusters is not much greater than α, 1% ≤ α ≤ 10%. In other words, a sequence of null hypotheses such as

\[ H_0^{(i)}: \text{The number of population clusters is } i \text{ or less} \]

where \( i = 1, 2, \ldots, n \), is tested against the alternatives such as

\[ H_a^{(i)}: \text{The number of population clusters exceeds } i \]

with a maximum experimentwise error rate of approximately α. The tests protect you from overestimating the number of population clusters. It is impossible to protect against underestimating the number of population clusters without introducing much stronger assumptions than are used here, since the number of population clusters could conceivably exceed the sample size.

DESIRABLE FEATURES OF A SOLUTION
The method used in the MODECLUS procedure has the following useful features:

- No distributional assumptions are required.
- The data can be coordinates or distances.
- Time and space requirements for the significance tests are no worse than those for obtaining the clusters.
- The power is high enough to be useful for practical purposes.

A SOLUTION HAVING THE PRECEDING FEATURES
The method for conducting significance tests is as follows:

1. Estimate densities using fixed-radius uniform kernels.
2. Obtain preliminary clusters by a "valley-seeking" method. Other clustering methods could be used but would yield less power.
3. Compute an approximate p-value for each cluster by comparing the estimated maximum density in the cluster with the estimated maximum density on the cluster boundary.
4. Repeatedly join the least significant cluster with a neighboring cluster until all remaining clusters are significant.
5. Estimate the number of population clusters as the number of significant sample clusters.
6. The preceding steps can be repeated for any number of different radii, and the estimate of the number of population clusters can be taken to be the maximum number of significant sample clusters for any radius.

COMPARISON WITH OTHER CLUSTERING PROCEDURES
Unlike the methods used in the CLUSTER procedure, the methods used in the MODECLUS procedure are not inherently hierarchical. However, the MODECLUS procedure can do approximate nonparametric significance tests for the number of clusters by obtaining an approximate p-value for each cluster and can hierarchically join nonsignificant clusters.

Another important difference between the MODECLUS procedure and many other clustering methods is that you do not tell the MODECLUS procedure how many clusters you want. Instead, you give the MODECLUS procedure a "smoothing parameter" and, optionally, a significance level, and the MODECLUS procedure determines how many clusters there are. You can specify a list of smoothing parameters, and the MODECLUS procedure will perform a separate cluster analysis for each value in the list.
WHAT IS A CLUSTER?

To make inferences regarding population clusters, it is first necessary to define what is meant by a cluster. For clustering methods using nonparametric density estimation, a cluster is usually loosely defined as a region surrounding a local maximum of the probability density function or a maximal connected set of local maxima.

Figure 1. Population Density -- 2 or 14 Clusters?

This definition may not be satisfactory for very rough densities with many local maxima. For instance, it is not clear whether the number of clusters is 14 or 2 in Figure 1, since there are 14 local maxima.

This definition is also not applicable at all to discrete distributions for which the density does not exist. As another example in which this definition is not intuitively reasonable, consider a uniform distribution in two dimensions with support in the shape of a "figure eight" (including the interior). This density might be considered to contain two clusters even though it does not have two distinct modes.

These difficulties can be avoided by defining clusters in terms of the local maxima of a smoothed probability density or mass function. For example, define the neighborhood distribution function (NDF) with radius \( r \) at a point \( x \) as the probability that a randomly selected point will lie within a radius \( r \) of \( x \), that is, the probability integral over a hypersphere of radius \( r \) centered at \( x \):

\[
s(x) = P(d(x, X) \leq r)
\]

where \( X \) is the random variable being sampled, \( r \) is a user-specified radius, and \( d(x, y) \) is the distance between points \( x \) and \( y \).

After you apply this smoothing technique with \( r = 0.063 \) to Figure 1, the number of clusters is quite obvious, as shown in Figure 2.

Figure 2. Smoothed Population Density

The NDF exists for all probability distributions. The radius can be selected according to the degree of resolution required. The minimum-variance unbiased estimate of the NDF at a point \( x \) is proportional to the uniform-kernel density estimate with corresponding support.

You can define a modal region as a maximal connected set of local maxima of the NDF. A cluster is a connected set containing exactly one modal region. This definition seems to give intuitively reasonable results in most cases.

UNIFORM KERNEL DENSITY ESTIMATES

The MODECLUS procedure uses (hyper)spherical uniform kernels of fixed or variable radius. It is convenient to refer to the sphere of support of the kernel at observation \( x_i \) as the neighborhood of \( x_i \). The observations within the neighborhood of \( x_i \) are the neighbors of \( x_i \). The estimated density at \( x_i \) is then

\[
f_i = \frac{n_i}{n v_i},
\]

that is, the number of neighbors of \( x_i \) divided by the product of the sample size (which is \( n \)) and the volume of the neighborhood at \( x_i \) (which is \( v_i \)). The size of the sphere is determined by the smoothing parameters that you are required to specify.

Uniform kernels do not provide quite as good density estimates as some other types of kernels, but they are quite satisfactory for clustering. The significance tests for the number of clusters require the use of fixed-size uniform kernels.

To illustrate density estimation in the MODECLUS procedure, a set of 30 observations with two variables was randomly generated in Figure 3. Three groups of observations were constructed. In each group, the points were generated according to a bivariate normal distribution with given mean vector \( (\mu_x, \mu_y) \) and standard deviation vector \( (\sigma_x, \sigma_y) \) for both \( x \) and \( y \):
group1: 10 observations
\[ \mu_x = 20 \quad \mu_y = 20 \quad \sigma_x = 6 \quad \sigma_y = 6 \]
group2: 10 observations
\[ \mu_x = 70 \quad \mu_y = 25 \quad \sigma_x = 7 \quad \sigma_y = 9 \]
group3: 10 observations
\[ \mu_x = 50 \quad \mu_y = 60 \quad \sigma_x = 25 \quad \sigma_y = 3 \]

The first group has a more compact structure while the second group has a less compact structure. The third group contains a more elongated elliptical structure. Note that the observations of the third group have been rotated 45 degrees counterclockwise. Also observation numbers are marked for the purpose of tracing clustering process.

Figure 3. Sample Data from 3 Bivariate Normal Distributions

With a fixed smoothing parameter (radius=15), the volume of the neighborhood at each observation will be the same for all observations. Therefore you can use the count, the number of observations within a neighborhood not including the observation itself, to represent the density estimates for each observation as shown in Figure 4. In this figure, all observations of the first group have greater counts since the compact structure (thus shorter distances between observations) allows more neighbors to be included in each neighborhood.

Figure 4. Count Within Neighborhood When Radius=15

CLUSTERING METHODS

The clustering methods used by the MODECLUS procedure use spherical clustering neighborhoods of fixed or variable radius that are similar to the spherical kernels used for density estimation. Several clustering methods are currently implemented in PROC MODECLUS. Here are some of those methods:

Method 1: Join each object to the neighbor with maximum estimated density.
This method is similar to the first stage of an algorithm proposed by Mizoguchi and Shimura (1980).

Method 2: Join each object to the neighbor with maximum estimated gradient.
This method was invented by Koontz, Narendra, and Fukunaga (1976).

Method 3: Join each object to the nearest neighbor with greater estimated density.
This method is similar to the classification or assignment stage of algorithms described by Gitman (1973) and Huizinga (1978).

Detail:
Objects with neighbors with equal density estimates but not greater density estimates require special treatment.

Other methods: Please refer to references for Koontz & Fukunaga (1972) and Wong & Lane (1983).

If you apply Method 3 described above, the clustering process is illustrated in the following three graphs, which use the sample data from Figure 3. All the observations are marked with counts, and an edge between two observations represents how two adjacent observations are joined.
In Figure 5, three clusters are formed when a given radius equals 15. All observations are correctly assigned to their corresponding groups.

![Figure 5. Clustering with Radius = 15](image)

In Figure 6, with a smaller smoothing parameter (radius=10), seven clusters are formed. Observations 21, 23, 28, and 30 are separated from the third group, and observations 2, 18, and 20 are separated from the second group.

![Figure 6. Clustering with Radius = 10](image)

In Figure 7, with an increased smoothing parameter (radius=30), all observations belong to one cluster.

![Figure 7. Clustering with Radius = 30](image)

ESTIMATING MAXIMUM MODAL AND BOUNDARY DENSITIES

If a population has two clusters, it must obviously have two modal regions. If there are two modal regions, there must be a "valley" between them. It seems intuitively desirable that the boundary between the two clusters should follow the bottom of this valley.

Regardless of the precise location of the cluster boundary, it is clear that the maximum of the NDF along the boundary between two clusters must be strictly less than the value of the NDF in either modal region; otherwise there would be only a single modal region. According to Hartigan and Hartigan (1985), there must be a "dip" between the two modes. The MODECLUS procedure assesses the significance of a sample cluster by comparing the NDF in the modal region with the maximum of the NDF along the cluster boundary. If the NDF has second-order derivatives in the region of interest and if the boundary between the two clusters is indeed at the bottom of the valley, then the maximum value of the NDF along the boundary occurs at a saddle point. Hence this test is called a saddle test.

The obvious estimate of the maximum NDF in a sample cluster is the maximum estimated NDF at an observation in the cluster. Let \( m(k) \) be the index of the observation for which the maximum is attained in cluster \( k \).

Estimating the maximum NDF on the cluster boundary is more complicated. One approach is to take the maximum NDF estimate at an observation in the cluster that has a neighbor belonging to another cluster. This method yields excessively large estimates when the neighborhood is large. Another approach is to try to choose an object closer to the boundary by taking the observation with the maximum sum of estimated densities of neighbors belonging to a different cluster.
After some experimentation, it was found that a combination of these two methods worked well. Let \( B_k \) be the set of indices of observations in cluster \( k \) that have neighbors belonging to a different cluster, and compute
\[
\max_{i \in B_k} \left( 0.2j_i n_i + \sum_{j \in N_i \cup C_k} j_j \right)
\]
where
\[
j_j = \text{the estimated density at } x_j
\]
\( n_i = \text{the number of observations within the neighborhood of } x_i, \text{ including } x_i \text{ itself} \)
\( N_i = \text{the set of indices of neighbors of } x_i, \text{ including } i \)
\( C_k = \text{the set of indices of observations assigned to cluster } k \)

Let \( s(k) \) be the index of the observation for which the maximum is attained.

**SIGNIFICANT TESTS**

Using the notation \#(S) for the cardinality of set \( S \), let
\[
n_{i_j} = \#(N_i^r \cap N_j^r)
\]
\( cm(k) = n_{m(k)} - n_{m(k) \cup \{k\}} \)
\( cr(k) = n_{r(k)} - n_{m(k) \cup \{k\}} \text{ if } B_k \neq \emptyset, \)
\( = 0 \text{ otherwise} \)
\( px = 1/2 \text{ if } B_k \neq \emptyset, \)
\( = 2/3 \text{ otherwise} \)
\( z_k = \frac{cm(k) - px (cm(k) + cr(k)) - 1/2}{\sqrt{px(1-px)(cm(k) + cr(k))}} \)
\( u = \left[ \frac{(2 + .05 \sqrt{n}) \sum_{c_{ni} > 1} \frac{1}{n_i + 1}}{n_{i}} \right] \)

where
\[
n_{i}^r = \text{the number of observations within the neighborhood of } x_i, \text{ not including } x_i \text{ itself} \)
\( N_i^r = \text{the set of indices of neighbors of } x_i, \text{ not including } i \)

**Note:** \( cr(k) = 0 \) and \( px = 1/2 \) when \( B_k = \emptyset \).

Cluster \( k \) is "isolated" since no member of the cluster has a neighbor belonging to a different cluster.

Let \( R(u) \) be a random variable distributed as the range of a random sample of \( u \) observations from a standard normal distribution. Then the approximate \( p \)-value for cluster \( k \), \( px \), is
\[
px = Pr(z_k > R(u)/\sqrt{2})
\]

If points \( m(k) \) and \( s(k) \) were fixed a priori, \( x_k \) would be the usual approximately normal test statistic for comparing two binomial random variables. In fact, \( m(k) \) and \( s(k) \) are selected in such a way that \( m(k) \) tends to be large and \( s(k) \) tends to be small. For this reason, and because there may be a large number of clusters, each with its own \( z_k \) to be tested, each \( z_k \) is referred to the distribution of \( R(u) \) instead of a standard normal distribution. If the tests were conducted for only one radius and if \( u \) were chosen equal to \( n \), then the \( p \)-values would be very conservative because (1) you are not making all possible pair-wise comparisons of observations in the sample and (2) \( n_{i}^r \) and \( n_{j}^r \) are positively correlated if the neighborhoods overlap. In the formula for \( u \), the summation overcorrects somewhat for the conservativeness due to correlated \( n_{i}^r \)'s. The factor .2+.05\( \sqrt{n} \) was empirically estimated from simulation results to adjust for the use of more than one radius.

**JOIN NONSIGNIFICANT CLUSTERS**

After computing the approximate \( p \)-value for each cluster, the MODECLUS procedure repeatedly joins nonsignificant clusters by using the following steps until all remaining clusters are significant at the chosen level. The number of population clusters is then estimated by the number of significant sample clusters.

1. Find the least-significant cluster.
2. If it is isolated, make all the members unassigned.
3. Otherwise, find the sum of density estimates of neighbors of the estimated saddle point belonging to other clusters. Join the nonsignificant cluster to the cluster with the greatest sum.

**SIMULATION RESULTS**

The saddle test implemented in the MODECLUS procedure has been evaluated by simulation from known distributions. Some results are given in the three tables on the next two pages. In Table 1, samples of 20 to 2000 observations were generated from a one-dimensional uniform distribution. For sample sizes of 1000 or less, 2000 samples were generated and analyzed by the MODECLUS procedure. For a sample size of 2000, only 1000 samples were generated. The analysis was done with at least twenty different values of the \( R \) option spread across the range of radii most likely to yield significant results. The six central columns of the table give the observed error rates at the nominal error rates (\( \alpha \)) at the head of each column. The standard errors of the observed error rates are given at the bottom of the table. The observed error rates are conservative for \( \alpha \leq 5\% \), but they increase with \( \alpha \) and become slightly liberal for sample sizes in the middle of the range tested.
Table 1. Observed Error Rates (%) for Uniform Distribution

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Type 1 Error Rate</th>
<th>No. of Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>Nominal</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>100</td>
<td>0.35</td>
<td>0.70</td>
</tr>
<tr>
<td>200</td>
<td>0.35</td>
<td>0.70</td>
</tr>
<tr>
<td>500</td>
<td>0.45</td>
<td>0.90</td>
</tr>
<tr>
<td>1000</td>
<td>0.50</td>
<td>1.30</td>
</tr>
<tr>
<td>Std. err.</td>
<td>0.22</td>
<td>0.31</td>
</tr>
</tbody>
</table>

Table 2 shows results from a unimodal mixture of two normal distributions with equal variances and equal sampling probabilities and with means separated by 2 standard deviations. Any greater separation would produce a bimodal distribution. The observed error rates are quite conservative.

Table 2. Observed Error Rates (%) for Normal Mixture with 2σ Separation

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Type 1 Error Rate</th>
<th>No. of Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>200</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>500</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Std. err.</td>
<td>0.22</td>
<td>0.31</td>
</tr>
</tbody>
</table>

Table 3 gives results in terms of power for identifying two clusters in samples from a bimodal mixture of two normal distributions with equal variances and equal sampling probabilities separated by 4 standard deviations. In this simulation, the MODECLUS procedure never indicated more than two significant clusters.

Table 3. Power (%) for Normal Mixture with 4σ Separation

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Type 1 Error Rate</th>
<th>No. of Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>350</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>750</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>1500</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>2000</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

The method for computing the p-values is based on a series of plausible approximations. There are as yet no rigorous proofs that the method is infallible. Neither are there any asymptotic results. However, simulations for sample sizes ranging from 20 to 2000 indicate that the p-values are almost always conservative. The only case discovered so far in which the p-values are liberal is a uniform distribution in one dimension for which the simulated error rates exceed the nominal significance level only slightly for a limited range of sample sizes.

References


SAS and SAS/STAT are registered trademarks or trade-marks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.