Journal of Classification 9:63-70 (1992)

The RUNT Test for Multimodality

J. A. Hartigan
Yale University

Surya Mohanty
Yale University

Abstract: Single linkage clusters on a set of points are the maximal connected sets in a graph constructed by connecting all points closer than a given threshold distance. The complete set of single linkage clusters is obtained from all the graphs constructed using different threshold distances. The set of clusters forms a hierarchical tree, in which each non-singleton cluster divides into two or more subclusters; the runt size for each single linkage cluster is the number of points in its smallest subcluster. The maximum runt size over all single linkage clusters is our proposed test statistic for assessing multimodality. We give significance levels of the test for two null hypotheses, and consider its power against some bimodal alternatives.

Keywords: Single linkage clusters; Minimal spanning tree; Tests for modes; The RUNT test.

1. Introduction

Several statistical tests have been developed for detecting multimodality in a distribution. The dip statistic (Hartigan and Hartigan 1985), is based on the distance between the empirical distribution F_n and the unimodal distribution F closest to it:

Research partially supported by NSF Grant No. DMS-8617919.
Authors' Addresses: J. A. Hartigan and Surya Mohanty, Department of Statistics, Yale University, New Haven, CT 06520-2179, USA.

$$\text{dip} = \min_{F\text{unimodal}} \max_{x} |F(x) - F_n(x)| .$$

Multimodality is suggested if the dip is large. A generalization of the dip statistic to higher numbers of dimensions is proposed in the SPAN statistic of Hartigan (1987), which relies on an analogue of the empirical distribution function defined on the minimum spanning tree. The SPAN statistic is conceptually and computationally complex.

In the one dimensional case, a test using a dip intensity statistic based on intervals between successive order statistics is suggested by J. B. Kruskal in Giacomelli, Wiener, Kruskal, Pomeranz and Loud (1971), but this test requires that the modes be specified in advance.

Silverman (1981) considers a quite different approach, in which the test statistic is the smallest window size in a kernel estimator that will produce a unimodal density estimate; Silverman suggests computing the significance level of the test differently for each sample, so that considerable computation is necessary on each sample.

We consider in this paper a simple test, the RUNT test, based on single linkage clusters. The RUNT test is defined as follows. Consider all the single linkage clusters. Since the clusters form a hierarchical tree, each non-singleton cluster divides into a number of subclusters. Associate with each cluster C the number of points $n(C)$ in the smallest subcluster (or "runt") of the cluster. The RUNT is $\max_C n(C)$.

The justification of the test statistic is based on the asymptotics of single linkage clusters described in Hartigan (1981). (The same asymptotic behavior does not hold for other hierarchical clustering techniques such as complete linkage which do not identify modes, so this test is appropriate only for single linkage clusters.) If there are at least two modes in the population density, then asymptotically, just one of the single linkage clusters will split into two clusters of points about the different modes. The smaller of these two clusters will be the runt. A large number of points in the smaller cluster indicates bimodality. On the other hand, if there is a single mode, we expect each cluster to divide into two clusters, the smaller of which contains very few points. Thus a large value of the RUNT should indicate multimodality.

2. Percentage Points of the RUNT Test

We need to consider the distribution of the RUNT for unimodal population distributions. Since there are many unimodal distributions, a conservative approach would use that unimodal distribution that tends to give the largest values to the RUNT. We consider two unimodal distributions for the null hypothesis, the spherical normal and the spherical uniform. The uniform is

the more conservative null hypothesis, but in many problems it is reasonable to suppose that the distribution deviates from normality by the addition of a few modes, and we would like to discover evidence of such deviation. It is therefore advisable to use the normal density as a null unimodal density to cover cases of this type. RUNTs that are significant by either null suggest some clustering is present.

In using the tables, we suggest selecting the dimension with some attention to how the distances are computed in the single linkage clustering. Ideally, the distribution of points should be spherically symmetric to justify using these null hypotheses. Suppose instead that the points have covariance matrix V that is not proportional to the identity, and suppose that its eigenvalues are $\lambda_1, \lambda_2, \ldots, \lambda_p$. Then $(\Sigma_i \lambda_i)^2 / (\Sigma_i \lambda_i^2)$ is the suggested dimensionality for use with the table. This covers the case when only the first three of ten variables, say, make a substantial contribution to the distances; we will use dimensionality close to three.

Tables 1 and 2 were constructed using the spherical normal distribution. For a particular sample size and dimension, the RUNT was computed for 999 different samples. Consider for example, the 95% point 22 for sample size 200 and dimension 20. This is the value such that no greater than 5% of the 999 computed runt values exceeds it. If an observed runt is 22 or greater, then the corresponding tail probability is less than 5%. Since the RUNT cannot exceed half the sample size, note that it is not possible to obtain 1% tail probabilities for sample size 10 unless the dimension exceeds 10.

Tables 3 and 4 were constructed using a uniform distribution over a sphere. The RUNT values are always higher than the corresponding normal runt values, so that it is more difficult to reject the null hypothesis of unimodality, if the uniform distribution expresses that null hypothesis.

The exact distribution of the RUNT, for tail probabilities less than 0.5, is available in the one dimensional uniform case: let $X_1, X_2, \ldots, X_n$ be n uniform order statistics, and let G_i be the interval length $X_{i+1} - X_i$. Then $G_1, \ldots, G_{n-1}$ have an exchangeable distribution.

For a fixed integer k, let E_i denote the event

$$G_{i-k}, G_{i-k+1}, \ldots, G_{i-1} < G_i > G_{i+1}, \ldots, G_{i+k} .$$

If the event E_i occurs, the cluster containing G_i and no greater interval divides into two clusters: the first cluster contains at least the $k + 1$ points $X_{i-k}, \ldots, X_i$ and the second cluster contains at least the $k + 1$ points $X_{i+1}, \ldots, X_{i+k+1}$. (Note that there are $2(k + 1)$ X's that are endpoints of the $2k + 1$ G's.) Thus if the event E_i occurs, the RUNT must exceed k. Also if the RUNT exceeds k, some event E_i must occur, for $k + 1 \leq i \leq n - k - 1$.

TABLE 1
95% POINT OF THE RUNT DISTRIBUTION, NORMAL NULL
999 iterations

	dimension						
sample size	1	2	3	4	5	10	20
10	6	5	5	5	5	5	5
20	10	9	9	8	8	8	7
50	22	18	16	15	14	12	11
100	42	31	26	24	21	17	17
200	75	50	41	36	33	27	22

TABLE 2
99% POINT OF THE RUNT DISTRIBUTION, NORMAL NULL
999 iterations

	dimension						
sample size	1	2	3	4	5	10	20
10	6	6	6	6	6	6	5
20	11	10	10	9	9	9	8
50	24	21	19	18	18	16	13
100	45	38	29	28	28	24	21
200	83	62	49	45	41	33	28

TABLE 3
95% POINT OF THE RUNT DISTRIBUTION, UNIFORM NULL
999 iterations

	dimension						
sample size	1	2	3	4	5	10	20
10	6	6	6	6	6	6	5
20	11	10	10	10	10	9	9
50	25	25	23	21	21	19	19
100	49	47	42	40	37	32	30
200	96	92	81	73	69	54	41

TABLE 4
99% POINT OF THE RUNT DISTRIBUTION, UNIFORM NULL
999 iterations

	dimension						
sample size	1	2	3	4	5	10	20
10	6	6	6	6	6	6	6
20	11	11	11	11	11	10	10
50	26	26	25	24	24	22	21
100	51	50	48	45	43	38	37
200	100	97	91	82	79	63	49

And finally, if $k + 1 > n / 3$, it is not possible for different events E_i, E_j to occur for $i \neq j$. Thus the probability that the RUNT exceeds k is the sum of the probabilities of the $n - 2k - 1$ E_i, and each of these probabilities is, by the exchangeability of the distribution of the G_i, just $\frac{1}{2k + 1}$. Thus

$$P\{\text{RUNT} > k\} = \frac{n - 2k - 1}{2k + 1} \text{ whenever } k + 1 > n / 3 .$$

$$P\{\text{RUNT} \geq k\} \leq .05 \text{ if } k \geq \frac{n}{2.1} + 0.5 .$$

For example, if $n = 200$, the 95% point is 96 in agreement with the simulations.

3. Power of the RUNT Test

We evaluate the power of the RUNT test, under both null hypotheses, for departures that are mixtures of normals, mixtures of uniforms, and for a log normal distribution. The last distribution is unimodal, and we would like to establish that the RUNT test will not be sensitive to it, so that we can conclude that large values of the RUNT are due to clusters rather than long tails.

In each case we use 999 iterations to estimate the probability that the RUNT will exceed the .05 percentage point under the null hypotheses, when samples are taken from the various alternatives. The test is powerful when this probability is high. The standard error for each entry is obtained by taking the square root of the entry ignoring the decimal point. For example, the first entry .061 has standard error approximately .008.

TABLE 5
POWER OF THE RUNT TEST, NORMAL MIXTURE AGAINST NORMAL NULL

Power for a mixture of $1/3 \times N(0, I) + 2/3 \times N(3, I)$

	dimension						
sample size	1	2	3	4	5	10	20
10	0.061	0.093	0.072	0.081	0.073	0.064	0.063
20	0.072	0.112	0.097	0.103	0.084	0.078	0.072
50	0.082	0.120	0.148	0.153	0.098	0.111	0.081
100	0.084	0.144	0.151	0.155	0.159	0.104	0.085

TABLE 6
POWER OF THE RUNT TEST, NORMAL MIXTURE AGAINST NORMAL NULL

Power for a mixture of $1/3 \times N(0, I) + 2/3 \times N(5, I)$

	dimension						
sample size	1	2	3	4	5	10	20
10	0.082	0.128	0.117	0.151	0.148	0.134	0.143
20	0.126	0.202	0.247	0.260	0.298	0.329	0.304
50	0.122	0.347	0.511	0.592	0.656	0.588	0.411
100	0.102	0.684	0.836	0.873	0.841	0.691	0.525

TABLE 7
POWER OF THE RUNT TEST, LOG-NORMAL AGAINST NORMAL NULL

Power for a Log-Normal alternative

	dimension						
sample size	1	2	3	4	5	10	20
10	0.069	0.043	0.032	0.024	0.017	0.007	0.005
20	0.019	0.015	0.011	0.009	0.006	0.002	0.002
50	0.013	0.009	0.004	0.004	0.001	0.002	0.001
100	0.003	0.004	0.003	0.002	0.003	0.001	0.001

TABLE 8
POWER OF THE RUNT TEST, UNIFORM MIXTURE AGAINST UNIFORM NULL

Power for a Uniform Mixture $1/3 \times U(0,1) + 2/3 \times U(3,4)$

	dimension						
sample size	1	2	3	4	5	10	20
10	0.050	0.061	0.067	0.098	0.089	0.091	0.096
20	0.044	0.044	0.107	0.147	0.119	0.191	0.233
50	0.018	0.030	0.077	0.107	0.119	0.305	0.434
100	0.002	0.011	0.020	0.102	0.162	0.625	0.801

An adjustment in the power calculations is made to allow for the discreteness of the RUNT distribution. The 0.95 point given in the tables is that value such that no more than 0.05 of the distribution exceeds it. For the power calculations we use a randomized test that sometimes rejects the null hypothesis when the RUNT is one less than the stated value, to bring the rejection probability to .05 exactly. For example, in the normal null with sample size 10, $P\{\text{RUNT} \geq 6\} = .042$ and $P\{\text{RUNT} = 5\} = .014$, so the randomized test rejects when RUNT ≥ 6, but also when RUNT = 5 with probability $\frac{.05 - .042}{.014}$.

Considering first the case of the normal mixtures, Tables 5 and 6, note that $N(3,I)$ refers to a spherical normal distribution with mean that is distance 3 from zero. We see that when the mixture is of $N(0,I)$ and $N(3,I)$, maximum power is achieved for dimension about 3 or 4, but the test is not powerful even then. The picture is more satisfactory for mixing $N(0,I)$ and $N(5,I)$ which we can detect with reasonable probability that increases with sample size, but again tends to be largest for dimensions 3 to 5. The RUNTs for the log-normal alternative tends to be lower than for the normal null hypothesis, which is what we want, so that we won't be falsely discovering modes when we really only have long tails.

For the uniform alternative, Tables 7 and 8, note that $U(0,1)$ denotes a spherical uniform of diameter 1 about zero, and that $U(3,1)$ denotes a spherical uniform of diameter 1 about a point distant 3 from zero. In general the RUNT statistic becomes more powerful as the dimensionality increases; for dimensions less than 4, the power decreases with sample size; for dimensions greater than 4, the power increases with sample size. The reason is that for low dimensions the critical value of the RUNT is about half the sample size, but the alternative gives RUNT values about one third the sample size; for

high dimensions, the critical value of the RUNT declines and the test becomes more powerful as the sample size increases.

4. Conclusions and Discussion

The RUNT is a simply computed test statistic that may be used for detecting the presence of multimodality in populations with the aid of single linkage clustering. It shows some power in detecting well-separated normal mixtures when the normal distribution is taken as the null distribution. It works best for dimensions 3, 4, and 5.

References

GIACOMELLI, F., WIENER, J., KRUSKAL, J. B., POMERANZ, J. W., and LOUD, A. V. (1971), "Subpopulations of Blood Lymphocytes as Demonstrated by Quantitative Cytochemistry," *Journal of Histochemistry and Cytochemistry, 19*, 426-433.

HARTIGAN, J. A., and HARTIGAN, P. M. (1985), "The Dip Test for Unimodality," *The Annals of Statistics, 13*, 70-84.

HARTIGAN, J. A. (1981), "Consistency of Single Linkage for High Density Clusters," *Journal of the American Statistical Association, 76*, 388-394.

HARTIGAN, J. A. (1988), "The Span Test for Unimodality," in *Classification and Related Methods of Data Analysis*, Ed. H. H. Bock, Amsterdam: North Holland, 229-236.

SILVERMAN, B. W. (1981), "Using kernel Density Estimates to Investigate Multimodality," *Journal of the Royal Statistical Society, Series B, 43*, 97-99.