The Korean Communications
in Statistics Vol. 12 No. 2, 2005
pp. 305-312

Estimating the Number of Clusters using Hotelling's T^2[1]

Kyungmee Choi[2]

Abstract

In the cluster analysis, Hotelling's T^2 can be used to estimate the unknown number of clusters based on the idea of multiple comparison procedure. Especially, its threshold is obtained according to the probability of committing the type one error. Examples are used to compare Hotelling's T^2 with other classical location test statistics such as Sum-of-Squared Error and Wilks' Λ The hierarchical clustering is used to reveal the underlying structure of the data. Also related criteria are reviewed in view of both the between variance and the within variance.

Keywords : Multiple Comparison Procedure, Type One Error, Bonferroni-Type Significance Level

1. Introduction

Recently the cluster analysis has been widely used. However estimating the number of optimum clusters has lead to a variety of different clustering methods. Early works related to this include works by Ward (1963) and Mojena (1975). Latest works include Rousseeuw and Driessen (1999), Duda et al. (2001), Hastie et al. (2001), and Gallegos (2002). Duda, Hart, and Stork (2001) pointed out that in general the number of optimum clusters is not known.

Most previous methods have used, as criteria, functions of variances within each clusters (within-variance). Smaller within-variances tend to provide well separated clusters. Bigger variances between clusters (between-variance) also imply that clusters are well separated. If the between-variance is big compared to the within-variance, it is very clear that the clusters are well separated. However if the between-variance is small compared to the within-variance, clusters are often unlikely to be well separated even with the small within-variance. Therefore it is more reasonable to use the criteria which consider both the within-variance and the between-variance.

[1] This work was supported by the 2003 Hongik University Academic Research Support Fund.

[2] Associate Professor, Department of Mathematics, College of Science and Technology, Hongik University at Jochiwon, Sinan 300, Jochiwon, Yungi Chung-Nam, 339-701, Korea.
E-mail : kmchoi@hongik.ac.kr

The classical Hotelling's T^2 is used as a criterion to estimated the number of clusters, and other related location test statistics such as Sum-of-Squared Error (SSE) and Wilks' Λ are explored for comparison. See Mardia et al.(1979), Duda et al. (2001), Hastie et al. (2001), and Rencher (2002). While SSE does not consider the between-variance, the rest two consider a ratio of the between-variance to the within-variance. However SSE has been one of the most widely used criteria based on the within-variance. Compared to SSE, Wilks' Λ has not been used as a criterion. Also Hotelling's T^2 has been only once used by Kim and Chung (2003) even though they used a much bigger threshold than the one that will be proposed in this paper. Since both Wilks' Λ and Hotelling's T^2 are powerful location tests, they are expected to make potentially powerful criteria to estimate the number of clusters.

When the data follow a normal distribution, SSE is approximately related to a χ^2 distribution, Wilks' Λ follows a Wilks' Λ distribution and also approximately follows a χ^2 distribution. Hotelling's T^2 follows an F distribution. Thus their distributions provide us clear thresholds for the given significance levels (significant error rates) along with nice statistical interpretation.

The hierarchical clustering is used because it reproduces the hierarchical structure of the data or the underlying structure of the data (Mojena, 1975). However all the criteria mentioned in this paper can be also used for the k-means clustering. These criteria will be calculated at each hierarchical level and presented in graphs to depict the estimates of the number of clusters. However the partitions are not necessarily optimal (Ward,1963).

In Section 2, three criteria are reviewed and their thresholds are sought for the given significance levels (significant error rates). In Section 3, as examples, seven equally spaced clusters of size 15 are simulated from bivariate normal distributions with different locations. For each set of data, three criteria are calculated and depicted.

2. The notations and the Criterion Functions

Suppose that for $x_j \in R^p, j=1,\cdots,n$, let the data be a set of $D = \{x_1, x_2, \cdots, x_n\}$ and cluster them into the c disjoint clusters, $D_1, D_2, \cdots, D_c$. Let n_i be the size of D_i. For each cluster D_i, let us define the mean and variance, $m_i = \sum_{x \in D_i} x/n_i$ and $S_i = \sum_{x \in D_i} (x-m_i)(x-m_i)^T$. The grand mean is $m = \sum_{x \in D} x/n$. Then $S_T = S_W + S_B$, where S_W is the within-cluster scatter matrix (within-variance), and S_B is between-cluster scatter matrix (between-variance) defined in the following way:

$$S_W = \sum_{i=1}^{c} S_i \text{ and } S_B = \sum_{i=1}^{c} n_i (m_i - m)(m_i - m)^T.$$

At the hierarchy of clusters, the level c corresponds to c clusters. Let the given significance level at each clustering level be α, which is controlled by the thresholds.

2.1 The Sum-of-Squared-Error

Let us define the Sum-of-Squared-Error as

$$SSE=\sum_{i=1}^{c}\sum_{x\in D_i}\|x-m_i\|^2$$

Note that $SSE=tr(S_W)=\sum_{i=1}^{c}tr(S_i)$. Since $tr[S_T]=tr[S_B]+tr[S_W]$ and $tr[S_T]$ is fixed, minimizing S_W implies maximizing S_B. Duda, Hart, and Stork (2001) suggested to find the number of clusters by minimizing SSE and pointed out that SSE worked best when the clusters are compact and well-separated. They also mentioned that when there was no minimum, the natural number of clusters was determined at the big gap. However often SSE decreases monotonically in c and tends to converge, so that there is not always the minimum. Also there could be multiple big gaps.

Ward (1963) tried to estimate the number of clusters by minimizing increase of SSE, which lead to the use of both the within-variance and the between-variance. Mojena (1975) evaluated Ward's Incremental Sum of Squares as the best among seven criteria studied at that time. On the other hand Rousseeuw and Driessen (1999) used $\prod_{i=1}^{c}det(S_i)^{|S_i|}$, where $|S_i|$ is the cardinality of ith cluster. Gallegos (2002) used $\prod_{i=1}^{c}det(S_i)$ as a criterion, and showed that m_i and S_i were Maximum Likelihood Estimators of means and variances of each clusters when data were generated from normal distributions. Using the trace considers only diagonals of the variance matrices, while using the determinant considers correlations, too.

2.2 Wilks' Λ

Wilks' Λ is one of the traditional statistics which test whether the locations of more than two groups are equal. This measure can be expressed as a function of the ratio of the between-variance to the within-variance, which is defined by

$$\Lambda=\frac{det(S_W)}{det(S_B+S_W)}=\frac{1}{det(S_W^{-1}S_B+I)}\,.$$

See Mardia (1979). The number of clusters is sought where Λ is minimized. However like SSE, Λ decreases monotonically in c. When the data follow a multivariate normal distribution, this statistic follows a Wilks' Λ distribution $\Lambda(p,n-c,c-1)$. When the sample size is large enough, its log transformation approximately follows a χ^2 distribution.

To obtain the statistically meaningful threshold which controls the significant level (significant error rate), let us define the p-value of a given value Λ_o at the cth clustering level as follows : $p=P(\Lambda \le \Lambda_o)$. A small p-value provides a strong evidence of two separate clusters. If there is not a significant decrease in p-value from the cth clustering level to the $(c-1)$th clustering level, then c is closer to the optimal number of clusters, where the criteria reaches the minimum.

The related statistics have been introduced by Pillai, Lawley-Hotelling, and Roy. $tr(S_W^{-1}S_B)$ and $tr(S_T^{-1}S_B)$ are also closely related to Λ See Hastie (2001), Duda, Hart, and Stork (2001), and Rencher (2002). Rencher (2002) introduced an analog of the univariate analysis of variance, $[tr(S_B)/(c-1)]/[tr(S_W)/(n-c)]$, which has a local maximum.

2.3 Hotelling's T^2

The classical Hotelling's T^2 tests whether the locations of two clusters are equal or not. For D_i and D_j clusters with $i \neq j$, it is defined by

$$T^2_{ij} = \frac{n_i n_j (n-2)}{(n_i+n_j)^2}(m_i - m_j)^T S^{-1}_{pij}(m_i - m_j),$$

where $S_{pij} = (S_i + S_j)/(n_i + n_j - 2)$. When the data follow a multivariate normal distribution, $(n_i+n_j-p-1)/p(n_i+n_j-2)T^2$ follows an $F(p,\ n_i+n_j-p-1)$. This statistic can be interpreted as the Mahalanobis distance between the centers of two clusters. See Mardia (1979).

To start finding the number of clusters, let us consider two clustering levels with $(c-1)$ and c clusters. It is necessary to decide which level is more optimal than the other. If no more merging occurs, then the final level is called the optimal in view of this criterion. To be more precise let us consider $\binom{c}{2}$ of Hotelling's T^2s at the cth clustering level, and they are used to decide which pair of clusters to be merged. Note that this leads to a classical multiple comparison (multiple-inference) procedure. If no significant merging occurs, then c is more optimal than $(c-1)$. Otherwise $(c-1)$ is more optimal.

Let us assume the value T_o be Hotelling's T^2_{ij} for the pair of D_i and D_j clusters at the cth clustering level. Then the corresponding p-value, p_{ij}, is defined by

$$p_{ij} = P((n_i+n_j-p-1)/p(n_i+n_j-2)\,T^2_{ij} \ge T_o) = P(F(p,\ n_i+n_j-p-1) \ge T_o),$$

where $F(p,\ n_i+n_j-p-1)$ is the F distribution with p and n_i+n_j-p-1 degrees of freedom. A small p_{ij} is a good evidence of two separate clusters. Especially if $\max_{1 \le i \neq j \le c} p_{ij}$

(MPH) is less than the given threshold, all c clusters are separated and so c is closer to the optimal than $(c-1)$. Thus in order to obtain a meaningful threshold, the bound of MPH should be studied. See Proposition 1 below.

Traditional multiple comparison procedures control the significance level (significant error rate) α by controlling the probability of committing any falsely declared significant inference under simultaneous consideration of $\binom{c}{2}$ multiple inferences. Yet, Kim and Chung (2003) have used Hotelling's T^2 as an individual inference to decide whether each pair of clusters were to be merged, so that each individual inference used α as a threshold. Ignoring the multiplicity of the inference, however, leads to a greatly increased false significant error rate. In this paper we control the multiplicity effect using the Bonferroni-Type Significance Level procedure ($BSLP$)(Rencher, 2002).

Let R be the number of pairs of clusters which are declared to be separated. Let V be the number of pairs of clusters which are falsely declared to be separated. $BSLP$ tests individually each pair of clusters at level $\alpha_s=\alpha/\binom{c}{2}$, which guarantees the probability of at least one falsely declared significant to be less than α. That is, $P(V\geq 1)\leq\alpha$. Since α_s gets usually very small as c grows, $BSLP$ is known to be very conservative and relatively lose its power. So the further studies can develop the thresholds based on the different multiple comparison procedures.

As an example, let us assume that α=0.05 as the total significance level (significant error rate) at the 4th clustering level. There are $\binom{4}{2}$ pairs of clusters. In the $BSLP$, $\alpha_s=(0.0083,\cdots,0.0083)$ for all $\binom{4}{2}$ pairs. So $\max_{1\leq i\neq j\leq c} p_{ij}$ is compared to 0.0083.

Proposition 1 Let α be the given significance level (significant error rate) at each clustering level and p_{ij} for $i\neq j$ be p-value of the individual test T^2_{ij}. Then,

$$\text{in the } BSLP: \quad P\left(\max_{1\leq i\neq j\leq c} p_{ij}\leq\alpha/\binom{c}{2} \text{ occurs falsely}\right)\leq\alpha.$$

Proof follows directly from the definitions of the BSL. Therefore using MPH guarantees α as the significance level (significant error rate). So the algorithm follows right away with the threshold α_s based on the $BSLP$.

3. Examples and Discussion

In order to compare three criteria, two sets of simulated data were generated. In R^2 seven equally spaced clusters of size 15 were generated from bivariate distributions. The data were clustered using hierarchical clustering. Then at each clustering level three criterion functions were calculated and plotted. Splus was used for the program. Clusters of outliers were carefully removed.

In order to generate seven equally spaced clusters, an equilateral hexagon is adopted. An equilateral hexagon is decomposed into six equal equilateral triangles, and the centers of six clusters are located at the vertexes of the hexagon. The seventh cluster is centered at the center of the hexagon. Thus for $k=1,\cdots,6$, the centers of six clusters are expressed as

$$\mu_k = (\cos(2k\pi/6), \sin(2k\pi/6)) \times d,$$

where d is the Euclidean distance of the centers from the origin. The center of the seventh cluster is defined by $\mu_0=(0,0)$. The random seven clusters can be generated by randomizing k and d if randomization is needed.

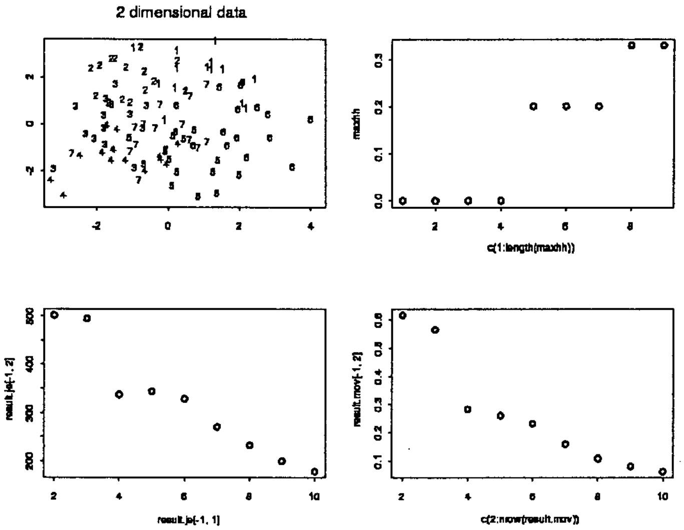
In <Figure 1> and <Figure 2>, the horizontal axis represents the number of clusters, and the vertical axis represents the criteria. In <Figure 1>, data were generated for d=2. *SSE* and Λ drop sharply at c=4 and then decrease more smoothly afterwards, which recommends that the number of clusters can be estimated as $\hat{c}=4$ For Hotelling's T^2, *MPH* jumps at c=4 and c=7. Based on α_s is 0.0083 at c=4, and 0.0024 at c=7 the *BSLP* returns $\hat{c}=4$ as the estimate of the number of clusters.

In <Figure 2>, d=4. *SSE* hits the bottom at c=7, while Λ converges smoothly to c=7. For Hotelling's T^2 *MPH* shows a clear jump at c=7. The *BSLP* provides α_s as 0.0024 at c=7, so it found 7 as the estimate. So this time, *SSE*, Λ and the *BSLP* coincide.

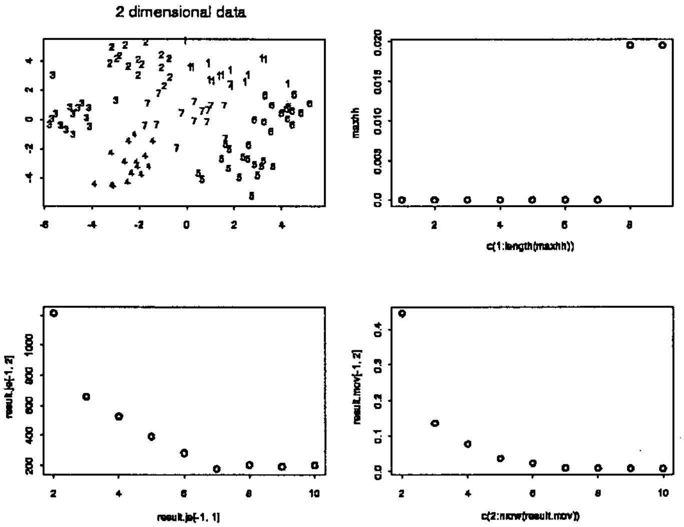
In most cases *SSE* and Λ behave similarly, which means that Λ is strongly related to *SSE*. However Compared to *SSE*, Λ is a good robust criterion because it considers both the within-variance and the between-variance. Also its p-value can be found and used as a clear threshold based on its exact distribution theory. For the similar reasons, Hotelling's T^2 can be used to make a good merging criterion with a clear threshold and it provides a clear jump when a significant split of clusters occurs, so that the number of clusters are easily estimated at the threshold.

Since there have been many other multiple comparison procedures developed, they can be further adopted to be compared with Bonferroni-Type Significance Level.

When the dimension gets bigger than 2, other types of equilateral geometric shape should be considered and the performance could change. Yet, the performance would not collapse down as long as the sample size is big enough and outliers are carefully controlled. Therefore in the future, more simulation study should be done for this matter.



<Figure 1> (Top Left) Plot of data with inter-cluster distance 2 (Top Right) *MPH* (Bottom Left) *MSE* (Bottom Right) Wilks′ Λ



<Figure 2> (Top Left) Plot of data with inter-cluster distance 4 (Top Right) *MPH* (Bottom Left) *MSE* (Bottom Right) Wilks′ Λ

References

[1] Duda, R.D., Hart, P. E., Stork, D.G. (2001). *Pattern Classification.* John Wiley Sons, Inc. New York.

[2] Gallegos, M. T. (2002). Maximum likelihood clustering with outliers, *Classification, Clustering, and Data Analysis*(Jajuga et al Ed.), Springer.

[3] Hastie, T., Tibshirani,R., Friedman, J. (2001). *The Elements of Statistical Learning, Data Mining, Inference, and Prediction.* Springer.

[4] Jajuga,K., Sokolowski A., Bock H.-H. (Eds.) (2002). *Classification, Clustering, and Data Analysis.* Springer.

[5] Kim,D. H. and Chung, C. W. (2003). Qcluster Relevance Feedback Using Adaptive Clustering for Content-Based Image Retrieval. *Proceedings of the ACM SIGMOD Conference.*

[6] Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979). *Multivariate Analysis.* Academic Press.

[7] Mojena, R. (1975). Hierarchical grouping methods and stopping rules: An evaluation. *The Computer Journal,* vol. 20, no. 4.

[8] Rencher, A.C. (2002). *Methods of Multivariate Analysis.* John Wiley and Sons.

[9] Rousseeuw, P. J. and Van Driessen, K. (1999). A first algorithm for the minimum covariance determinant estimator, *Technometrics*, vol. 41, 212-223.

[10] Ward, J. H. (1963). Hierarchical Grouping to optimize an objective function. *Journal of American Statistical Association,* vol. 58, 236-244 .

[Received August 2004, Accepted April 2005]