COMMUNITY ECOLOGY 8(1): 9-13, 2007
1585-8553/$20.00 © Akadémiai Kiadó, Budapest
DOI: 10.1556/ComEc.8.2007.1.2

Changes in the model of within-cluster distribution of attributes and their effects on cluster analysis of vegetation data

M. B. Dale

Australian School of Environmental Studies, Griffith University, Nathan, Qld. 4111, Australia.

Keywords: Clustering, Gaussian, Minimum message length, t-distribution, Within cluster variation.

Abstract: In previous studies a minimum message length fuzzy clustering method was applied to vegetation data and shown to give sensible estimates for the number of clusters as well as consistent estimates of cluster parameters. The minimum message length method provides a principled method of choosing between models and between classes of models. It comprises 2 components; one coding the model and its associated (meta)parameter values, the other coding the data, given the model. The program uses uncorrelated Gaussian distributions as a model for the distribution of attributes within clusters. This assumption may not be acceptable and in this paper a more general model, the t-distribution, has been examined. The t-distribution provides a class of thick-tailed models, while including the Gaussian as a subclass. This should be appropriate in hierarchical clustering where, even if the final clusters had internal Gaussian distributions, the upper levels would not. In addition, it may provide a better model of within-cluster distribution of the attributes even in the final clusters. Although forcing the use of t-distributions was not profitable, allowing a choice between Gaussian and t-distributions for each attribute in each class resulted in improved results. This was despite only one attribute actually selecting the t-distribution over the Gaussian.

Introduction

In a series of papers (Dale 2000, 2001, 2002, 2005, Dale and Dale 2002, Dale et al. 2001), the application of the minimum message length (MML) method has been used to provide a principled method of model choice. Minimum message length uses 2 components to obtain an overall message length; one is a coding of the model and its associated parameters, the other is a coding of the data, given the selected model. To obtain consistent estimates of the parameters, a fuzzy solution was used and the program also determined the optimal precision for coding parameter values.

These previous studies used the original SNOB program (Wallace and Boulton 1968), which makes some assumptions concerning the nature of the clusters it identifies. In particular, for numeric data, it permits only Poisson and Gaussian distributions within clusters for numeric variables[1]. The restriction to Gaussian within-cluster distributions might seem to be an unnecessary limitation and it is desirable to examine other possible distributions.

Alternative thick-tailed distributions are of interest because of their suitability for use in hierarchical clustering. Recently, Agusta and Dowe (2002) have extended the available models to include the t-distribution, implemented in the program Jsnob. Compared with the Gaussian, the t-distribution, though it remains symmetric, has thicker tails, with the thickness of the tails being determined by the degrees of freedom. Thus, with 1 degree of freedom it becomes the Cauchy distribution, while with large numbers of degrees of freedom ($\rightarrow\infty$) it approximates the Gaussian.

The t-distribution thus provides a more general class of models than the Gaussian, at the cost of extra parameters (the degrees of freedom) to be estimated for each attribute in each cluster. In particular, for a hierarchical clustering, even if the final clusters are well fitted by a within-cluster Gaussian model, clusters at higher levels will not be well fitted since they are a mixture of several

1 A comparison between these choices can be found in Dale (2001). Snob also permits multistate and angular data. Recent developments have also introduced gamma distributions (Agusta and Dowe 2003a) and Dirichlet distributions (Bouguila and Zhiou 2006).

Table 1. General properties of solutions. Message lengths in nits.

Analysis	1-cluster Message Length	# cluster (n)	n-cluster Message Length	1-class – n-class	% capture difference 1-class
Gaussian	5941.9	5	4562.2	1379.7	23.2
t-distribution	7224.7	2	5686.8	1537.9	21.0
optional	5680.0	5	4501.7	1178.3	20.7

clusters. A thick-tailed distribution would therefore be a preferable model.

This paper compares the results obtained from analyses using either Gaussian or t-distributions and one allowing choice between the two distributions. By using the minimum message length (MML) criterion we can both estimate the optimal number of clusters and compare the different models allowing selection of the 'best' model. MML operationalises Ockham's razor (Needham and Dowe 2001) to determine the preferred model.

Data and methods

The data used in these analyses form a spatial sequence of vegetation reported by Gitay and Agnew (1989) and have been previously used in a study of gradients (Dale 2005). This study suggests that the plots do fall along some kind of gradient, although this gradient is inconsistent over species. It is often masked by competitive exclusion leading to bimodality in spatial distributions. The primary data were recorded on a transect of 113 contiguous samples, each 4 cm × 4 cm, from a dune slack in the Ynyslas National Nature Reserve, West Wales. In each sample the combined above- and below- ground biomass for all perennial species was measured, together with the total Calcium, Phosphorus and Organic Matter. In total 12 species were recorded but 3 were very rare and have been ignored here, as have the environmental factors; in any case, these latter were not closely related to the spatial gradient. The nine species used were *Amblystegium serpens, Preissia quadrata, Agrostis stolonifera, Carex arenaria, Carex flacca, Eleocharis uniglumis, Juncus articulatus, Hydrocotyle vulgaris* and *Ranunculus bulbosus*.

Three analyses were performed. These were:

1. A standard analysis using Gaussian, within-cluster distributions.

2. An analysis using t-distributions as the within cluster distribution

3. An analysis which permitted the program to choose between Gaussian and t-distributions for every attribute in each class.

Table 2. Cluster assignments of samples. The first line gives the Gaussian result, the second the t-distribution result and the third the optional result. The plots are contiguous, although arranged in blocks of twenty for display. The cluster labels are arbitrary.

Plots 1-20																				
Gaussian	1	1	2	2	2	2	2	2	2	2	2	2	1	2	2	2	2	2	2	
t-distr.	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	
Optional	4	4	2	2	2	2	2	2	2	2	2	2	4	2	2	2	2	2	2	
Plots 21-40																				
Gaussian	2	2	2	2	2	2	3	3	2	2	2	3	2	2	2	2	2	2	1	1
t-distr.	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Optional	2	2	2	2	2	2	5	5	2	2	2	5	2	2	2	2	2	2	4	4
Plots 41-60																				
Gaussian	1	1	2	2	2	2	2	2	2	2	2	1	2	3	3	3	2	2	2	2
t-distr.	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Optional	4	4	2	2	2	2	2	2	2	2	2	4	2	5	5	5	2	2	2	2
Plots 61-80																				
Gaussian	2	2	2	2	2	3	2	2	2	3	2	2	2	2	2	3	2	2	4	4
t-distr.	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	1
Optional	2	2	2	2	2	5	2	2	2	5	2	2	2	2	2	5	2	2	1	1
Plots 81-100																				
Gaussian	4	4	4	5	4	3	3	4	3	3	4	3	4	5	3	3	3	5	5	5
t-distr.	1	1	1	1	1	2	2	1	2	2	1	2	1	1	2	2	2	1	1	1
Optional	1	3	1	3	3	5	5	3	5	5	3	5	3	3	5	5	5	3	3	3
Plots 101-113																				
Gaussian	5	5	2	5	5	5	5	5	5	5	3	4	5							
t-distr.	1	1	2	1	1	1	1	1	1	1	2	1	1							
Optional	3	3	2	3	3	3	3	1	3	3	1	1	1							

These three analyses will be referred to as *Gaussian, t-distribution* and *optional* analyses. For the t-distribution and optional analyses, the degrees of freedom were regarded as unknown parameters to be estimated by the program. Any estimate of degrees of freedom that exceeded 100 was regarded as indicating a close approximation to a Gaussian distribution. Since the approximation of the Gaussian by the t-distribution is good for degrees of free-

dom of around 30, this results in a bias towards the t-distribution.

None of the analyses was constrained by spatial arrangement; the result is a clustering rather than a segmentation into spatially coherent sections. A comparison with segmentation has been made (Dale et al. 2007).

Results

The results are shown in Tables 1-5. For each analysis the one-class solution and the estimated optimal message lengths are provided, together with their difference and this difference as a proportion of the one-class value (Table 1).

The entries in the attribute tables (Tables 3-5) record the significance of difference between cluster and population means and the direction of that difference. In addition, where t-distributions are involved the degrees of freedom are recorded in parentheses. Table 1 gives general properties of the 3 analyses such as message length and the percentage of variation captured by the estimated number of clusters. All three analyses show significant clustering to be present, with differences between one-class and optimal n-class solutions of over 1000 nits in every case. This indicates odds of less than e^{-1000} in favour of the cluster solutions compared to the single cluster alternative.

The three results are ordered by their optimal n-class message lengths (Table 1) as:

optional (4501.8) < *Gaussian* (4562.2)
< *t-distribution* (5686.8).

Since a smaller message length is to be preferred, this ordering also applies to choice of analysis. The t-distribution has a much longer message length while the Gaussian and optional have more similar values. However, a difference of 60.4 is still significant, with odds in favour of the smaller of $e^{-60.4}$:1, so that the MML selection criterion indicates the optional solution as best. The data are clearly very noisy with structure capture percentages in the low 20's; such low values are not unusual with vegetation data.

The Gaussian and optional solutions both estimated 5 classes, while the t-distribution solution has only 2. However, these 2 classes represent a coarser subdivision of the data along the spatial transect, largely separating it into 2 parts which the other analyses also recognise but further subdivide (see Wallace and Dale 2005). This is detectable in the assignments of plots to clusters given in Table 2. The t-distribution cluster 1 is markedly similar to Gaussian 4 and 5 together, and somewhat less similar to the optional solution clusters 1 and 3.

Examining the discrimination of species (Tables 3, 4 and 5), the t-distribution has fewer attributes with signifi-

Table 3. Gaussian analysis: species discrimination. ++: $p < 0.01$ and more common in cluster, +: $p < 0.2$ and more common in cluster, *: not significant, -: $p < 0.2$ and less common in cluster, —: $p < 0.01$ and less common in cluster.

Species/group	1	2	3	4	5
Amblystegium	--	--	--	*	++
Preissia	--	--	*	++	-
Agrostis	*	*	*	*	*
C. arenaria	*	*	*	*	*
C. flacca	+	*	*	-	--
Eleocharis	--	--	--	*	++
Juncus	*	*	*	*	*
Hydrocotyle	*	*	*	*	*
Ranunculus	++	--	--	--	--

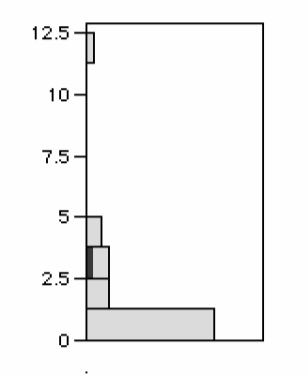
Table 4. t-distribution: species discrimination. ++: $p < 0.01$ and more common in cluster, +: $p < 0.2$ and more common in cluster, *: not significant, -: $p < 0.2$ and less common in cluster, —: $p < 0.01$ and less common in cluster. Entries in parenthesis show the estimated degrees of freedom. ∞ shows degrees of freedom exceeding 100.

Species	Group 1	Group 2
Amblystegium	**++ (7.4)**	-- (∞)
Preissia	**++(0.9)**	*
Agrostis	*	*
C. arenaria	*	*
C. flacca	--(∞)	*
Eleocharis	++ (∞)	*
Juncus	*	*
Hydrocotyle	*	*
Ranunculus	*	*

Table 5. Optional: t-distribution or Gaussian: species discrimination. ++: $p < 0.01$ and more common in cluster, +: $p < 0.2$ and more common in cluster, *: not significant, -: $p < 0.2$ and less common in cluster, —: $p<0.01$ and less common in cluster. Entries in parentheses show the estimated degrees of freedom, with ∞ indicating that a Gaussian distribution was used.

Species/Class	1	2	3	4	5
Amblystegium	++ (∞)	-- (∞)	++ (∞)	-- (∞)	-- (∞)
Preissia	++ (∞)	*	**++(0.8)**	*	++ (∞)
Agrostis	*	*	*	*	*
C. arenaria	*	*	*	*	*
C. flacca	- (∞)	*	-- (∞)	-- (∞)	++ (∞)
Eleocharis	*	*	++ (∞)	*	*
Juncus	+ (∞)	*	+ (∞)	++ (∞)	-- (∞)
Hydrocotyle	*	*	*	*	*
Ranunculus	-- (∞)	-- (∞)	-- (∞)	++ (∞)	-- (∞)

Figure 1. Histogram of the distribution of abundances of *Preissia* in cluster 1 of the t-distribution analysis.



cant differences, and a higher proportion of negatives. The optional analysis has slightly more significant differences than the Gaussian and a higher proportion of positives. The Gaussian, more typically of vegetation data, has the highest proportion of negative discriminators. In general the same species are significant in all analyses, notably *Amblystegium, Preissia, Ranunculus* and *Eleocharis*. Note that the estimates for degrees of freedom are either very large (∞) or quite small, so the bias towards the t-distribution appears to have little or no effect.

Discussion

What is curious is that the optional solution selects almost entirely Gaussian distributions for species, with only 1 species in 1 cluster (*Preissia* in cluster 3) indicating that a t-distribution is preferable. However, this single case is extreme with degrees of freedom estimated at about 1 so the distribution has extremely thick tails. This single exception causes a difference in message length of nearly 61 nits, which is highly significant. Of course it is possible that there exists a still better Gaussian solution but these data have been subjected to many analyses without any sign of such a solution.

Examination of the abundances of *Preissia* within the cluster (Fig. 1) shows that *Preissia* has a bimodal distribution or an extremely long tail. There is a single value of 11.8, but the second largest value is only 4.9 and there are a considerable number of zero values. It seems that separation of the two modes as different clusters does not provide sufficient improvement in fit to compensate for the extra message length needed to code the extra cluster. In any case, Snob has difficulties identifying singleton clusters and the message length contributed by the relevant sample, which could identify outlying status, was not reported by the Jsnob analysis.

While the majority of distributions within classes appear to be well approximated by the Gaussian, the single incidence of a t-distribution has a marked effect. In contrast, forcing all species to have t-distributions is counterproductive and produces the worst result overall in terms of message length, losing considerable cluster structure. Gaussian distributions remain the commonest; only 2 species have t-distributions with low degrees of freedom.

If this is typical, then restriction to Gaussian does not appear to be major limitation of the Snob analysis

There was no examination made here for a possible distribution with thin tails that would indicate a very limited response. This is an unlikely case because of sampling effects. Besides thin tails, two other possibilities remain to be investigated. First, the t-distribution is symmetric, and the distribution may well be asymmetric. The χ^2 distribution provides a possible asymmetric model and has the additional advantage that it is necessarily non-negative. However, asymmetric distributions can be well modelled as mixture distributions, i.e., clusters, so the desirability of using a specific asymmetric distribution remains to be established.

Second, the existence of correlation within clusters could have effects on fit for any of these models. The method for incorporating correlation with Gaussian distributions is known (Agusta and Dowe 2003b) although a suitable program is presently unavailable. The existence of such correlation would lead to the recognition of more clusters than necessary in the present analyses.

In summary, it seems that the assumption of within-cluster Gaussian distributions is unlikely to cause major disturbance to cluster solutions. Assuming thick-tailed distributions like the t-distribution, might, as in the present study, lead to a loss in detail. However, this does leave open the possibility that, in a hierarchical analysis (Wallace and Dale 2005), the upper levels of the hierarchy might be better modelled with a t-distribution or some similar thick-tailed distribution. At high levels, the clusters would be mixtures of Gaussian distributions varying in their degree of overlap, and a thick tailed distribution would be appropriate to the implied polymodality. Further subdivision, obtained by using a Gaussian model, could follow at a lower level.

References

Agusta, Y. and D. L. Dowe. 2002. MML clustering of continuous-valued data using Gaussian and t distributions. In: B. McKay and J. Slaney (eds.), *Lecture Notes on Artificial Intelligence 2557*. Springer, Berlin. pp. 143-154.

Agusta, Y. and D. L. Dowe. 2003a. Unsupervised learning of gamma mixture models using Minimum Message Length, (to appear) In: *Proc. 3rd IASTED International Conference on Artificial In-*

telligence and Applications (AIA 2003), ACTA Press, Calgary. pp. 457-462.

Agusta, Y. and D. L. Dowe. 2003b. Unsupervised learning of correlated multivariate Gaussian mixture models using MML. *Lecture Notes in Artificial Intelligence (LNAI) 2903*, Springer, Berlin. pp. 477-489.

Bouguila, N. and D. Ziou. 2006. Unsupervised selection of a finite dirichlet mixture model: An MML-based approach. *IEEE Transactions on Knowledge and Data Engineering* 18: 993-1009.

Dale, M. B. 2000. Mt Glorious revisited: secondary succession in subtropical rainforest. *Community Ecology* 1: 181-193.

Dale, M. B. 2001. Minimal message length clustering, environmental heterogeneity and the variable Poisson model. *Community Ecology* 2: 171-180.

Dale, M. B. 2002. Models, measures and messages: an essay on the role for induction. *Community Ecology* 3: 191-204.

Dale, M. B. 2005. On gradients and response curves. *Community Ecology* 6: 155-166.

Dale, M. B., L. Allison and P. E. R. Dale. 2007. Segmentation and clustering as complementary sources of information *Acta Oecologica* 30:1-10.

Dale, M. B., L. Salmina and L. Mucina. 2001. Minimum message length clustering: an explication and some applications to vegetation data. *Community Ecology* 2: 231-247.

Dale, P. E. R. and M. B. Dale. 2002. Optimal classification to describe environmental change: pictures from the exposition. *Community Ecology* 3: 19-30.

Gitay, H. and A. D. Q. Agnew. 1989. Plant community structure, connectance, niche limitation and species guilds within a dune slack grassland. *Vegetatio* 83: 241-248.

Needham, S. L. and D. L. Dowe. 2001. Message length as an effective Ockham's razor in decision tree induction. In: Proc. 8th International Workshop on Artificial Intelligence and Statistics (AI+STATS 2001), Key West, Florida, U.S.A. pp. 253-260

Wallace, C. S. and D. M. Boulton. 1968. An information measure for classification, *Computer Journal* 11: 185-194.

Wallace, C. S. and M. B. Dale. 2005. Hierarchical clusters of vegetation types. *Community Ecology* 6: 57-74.

Received November 7, 2005
Revised May 25, 2006
Accepted October 12, 2006