How many bee species? A case study in determining the number of clusters

Christian Hennig[1]

Department of Statistical Science, University College London, Gower St, London WC1E 6BT, United Kingdom c.hennig@ucl.ac.uk

Abstract. It is argued that the determination of the best number of clusters k is crucially dependent on the aim of clustering. Existing supposedly "objective" methods of estimating k ignore this. k can be determined by listing a number of requirements for a good clustering in the given application and finding a k that fulfils them all. The approach is illustrated by application to the problem of finding the number of species in a data set of Australasian tetragonula bees. Requirements here include two new statistics formalising the largest within-cluster gap and cluster separation. Due to the typical nature of expert knowledge, it is difficult to make requirements precise, and a number of subjective decisions is involved.

1 Introduction

Determining the number of clusters is a notoriously hard key problem in cluster analysis. There is a large body of literature about it (for some references beyond those given below see Jain (2010)).

One of the reasons why the problem is so hard is that most of the literature is based on the implicit assumption that there is a uniquely best or "true" clustering for given data or a given underlying statistical model assumed to be true without defining unambiguously what is meant by this. This obscures the fact that there are various ways of defining a "best clustering" which may lead to different solutions for the same data set or model. Which of these definitions is appropriate depends on the meaning of the data and the aim of analysis. Therefore there is no way to find a uniquely best clustering considering the data (or a model assumption) alone.

For example, the "true clusters" to be counted could correspond to, among others,

- Gaussian mixture components,
- density modes,
- connected data subsets that are strongly separated from the rest of the data set,

- intuitively clearly distinguishable patterns,
- the smallest number of data subsets with a given maximum within-cluster distance.

It is clear that these definitions can lead to different "true" numbers of clusters for the same data. For example it is well known that a mixture of several Gaussians can be unimodal or have more than two modes, two density modes are not necessarily separated by deep gaps, connected and well separated data subsets may include very large within-cluster distances etc. Note further that finding the "true" number of density modes or Gaussian mixture components is an ill posed problem, because it is impossible to distinguish models with k density modes or Gaussian mixture components from models with arbitrarily more of them based on finite data for which a model with k modes/mixture components fits well. A well known implication of this (Hennig (2010)) is that the BIC, a consistent method for estimating the number of Gaussian mixture components, will estimate a k tending to infinity for $n \to \infty$ (n being the number of observations) because of the fact that the Gaussian mixture model does not hold precisely for real data, and therefore more and more mixture components will fit real data better and better if there are only enough observations to fit a large number of parameters.

Different concepts to define the number of clusters are required for different applications and different research aims. For example, in social stratification, the poorest people with the lowest job status should not be in the same cluster (social stratum) as the richest people with the highest job status, regardless of whether there is a gap in the data separating them, or whether these groups correspond to different modes, i.e., large within-cluster dissimilarities should not occur. On the other hand, in pattern recognition on images one often wants to only separate subsets with clear gaps between them regardless of whether there may be large distances or even multiple weak modes within the clusters.

In the present paper I suggest a strategy to determine the number of clusters depending on the research aim and the researcher's cluster concept, which requires input based on an expert's subject matter knowledge. Subject matter knowledge has already been used occasionally in the literature to determine the number of clusters, see e.g., Chaturvedi et al. (2001), Morlini and Zani (2012), but mostly informally.

Section 2 introduces a number of methods to estimate the number of clusters. The new approach is illustrated in Section 3 by applying it to the problem of determining the number of species in a data set of tetragonula bees. Some limitations are discussed in Section 4.

2 Some methods to determine the number of clusters

Here are some standard approaches from the literature to determine the number of clusters. Assume that the data are $x_1, \ldots, x_n$ in some space $\mathcal{S}$, are to

be partitioned into exhaustive and non-overlapping sets $C_1, \ldots C_k$, and that there is a dissimilarity measure d defined on $\mathcal{S}^2$.

Calinski and Harabasz (1974) index. k_{CH} maximises $\frac{B(k)(n-k)}{W(k)(k-1)}$, where

$$W(k) = \sum_{h=1}^{k} \frac{1}{|C_h|} \sum_{x_i, x_j \in C_h} d(x_i, x_j)^2, \text{ and}$$
$$B(k) = \frac{1}{n} \sum_{i,j=1}^{n} d(x_i, x_j)^2 - W(k).$$

Note that k_{CH} was originally defined for Euclidean distances and use with k-means, but the given form applies to general distances.

Average silhouette width (Kaufman and Rousseeuw (1990)). k_{ASW} maximises $\frac{1}{n} \sum_{i=1}^{n} s(i,k)$, where

$$s(i,k) = \frac{b(i,k) - a(i,k)}{\max(a(i,k), b(i,k))},$$
$$a(i,k) = \frac{1}{|C_j| - 1} \sum_{x \in C_j} d(x_i, x), \; b(i,k) = \min_{x_i \notin C_l} \frac{1}{|C_l|} \sum_{x \in C_l} d(x_i, x),$$

C_j being the cluster to which x_i belongs.

Pearson-version of Hubert's Γ (Halkidi et al. (2001)). k_{PG} maximises the Pearson correlation between a vector of all dissimilarities and the corresponding binary vector with 0 for a pair of observations in the same cluster and 1 for a pair of observations in different clusters.

Bootstrap stability selection (Fang and Wang (2012)). This is one of a number of stability selection methods in the literature. For each number of clusters k of interest, B pairs of standard nonparametric bootstrap subsamples are drawn from the data. For each pair, both subsamples are clustered, and observations not occurring in any subsample are classified to a cluster in both clusterings in a way adapted to the used clustering method. For example, in Section 3, average linkage clustering is used and unclustered points are classified to the cluster to which they have the smallest average dissimilarity. For each pair of clusterings the relative frequency of point pairs in the same cluster in one of the clusterings but not in the other is computed, these are averaged over the B bootstrap samples, and k_{BS} is the k that minimised the resulting instability measure.

As many methods in the literature, the former three methods all try to find a compromise between within-cluster homogeneity (which generally improves with increasing k) and between-cluster separation (which usually is better for smaller k). The terms "within-cluster homogeneity" and "between-cluster separation" are meant here in a general intuitive sense and admit various ways of measuring them, which are employed by the various different criteria. The k optimising these indexes may differ. For example, experiments indicate that k_{ASW} may lump together relatively weakly separated data subsets if their union is strongly separated from what is left, whereas k_{CH} may leave them separated if putting them together makes the resulting cluster too heterogeneous. k_{PG} tends less than the two former methods to integrate single outliers in clusters.

A general remark on stability selection is that although *good* stability is a reasonable requirement in many applications, *optimal* stability is more difficult to motivate, because there is no reason why "bad" clusterings cannot be stable.

3 Analysis of the tetragonula bees data set

Franck et al. (2004) published a data set giving genetic information about 236 Australasian tetragonula bees, in which it is of interest to determine the number of species. The data set is incorporated in the package "fpc" of the software system R (`www.r-project.org`). Bowcock et al. (1994) defined the "shared allele dissimilarity" formalising genetic dissimilarity appropriately for species delimitation, which is used for the present data set. It yields values in $[0, 1]$.

In order to apply the approach taken here and in fact also in order to choose an appropriate clustering method, it is important to specify formal requirements of species delimitation. The following list was compiled with help of the species expert Bernhard Hausdorf, museum of zoology, University of Hamburg.

- Large within-cluster gaps should be avoided, because genetic gaps are essential for the species concept. Some caution is needed, though, because gaps could be caused by incomplete sampling and by regional separation within a species.
- Species need to be well separated for the same reason. Experts would normally speak of different species even in case of rather moderate separation among regionally close individuals, so to what extent separation is required depends on the location of the individuals to some extent.
- In order to count as species, a group of individuals needs to have a good overall homogeneity, which can be measured by the average within-species dissimilarity.
- Cluster stability is needed in order to have confidence that the clustering is not a random structure, although there is no specific reason why the best clustering needs to have maximum stability.

The third criterion motivates the average linkage hierarchical clustering, which is applied here, see Figure 1 (it is beyond the scope of the paper to give a more conclusive justification). Determining the number of species amounts to finding the best height at which the dendrogram is cut. Values of k between 2 and 15 were examined.

The criteria introduced in Section 2 do not yield a consistent decision about the number of clusters, with $k_{CH} = k_{ASW} = 10$, $k_{PG} = 9$, $k_{BS} = 5$. Note that for $k > 3$ all instability values are smaller than 0.08, so all clusterings are rather stable and fulfil the fourth requirement. k_{BS} may generally be rather low, because splitting up somewhat ambiguous data subsets may harm

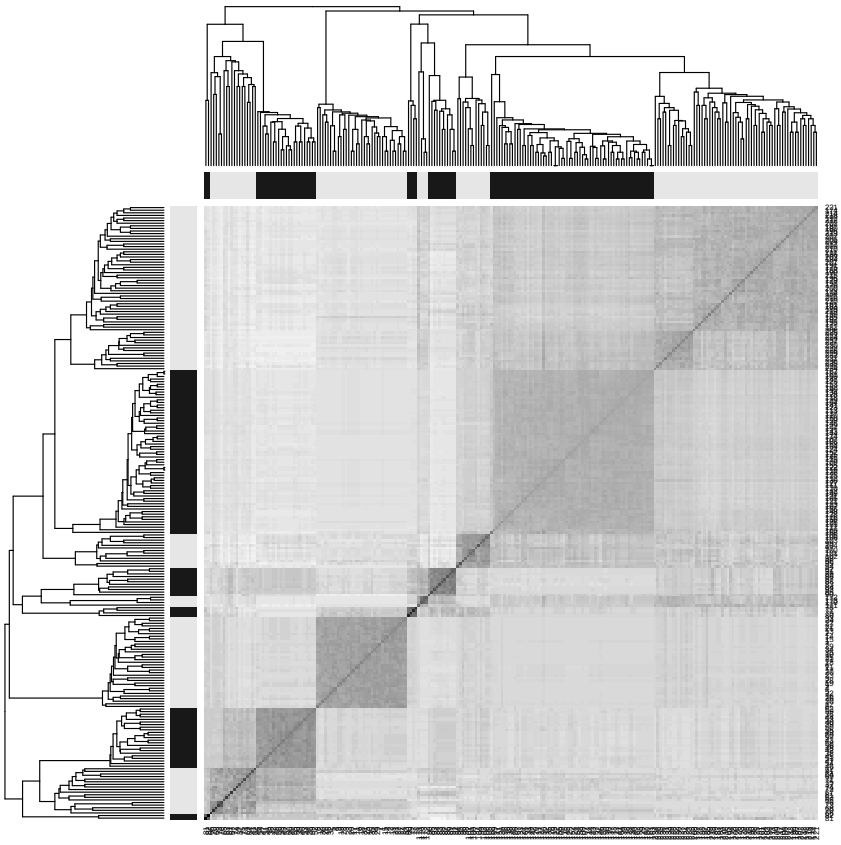

Fig. 1. Heatplot and average linkage clustering for tetragonula bee data. Colour bars at the left side and on top indicate the clustering with $k = 10$.

stability. Just taking the general behaviour of the criteria into account, k_{ASW} with its strong emphasis on separation looks closest to the listed requirements.

An approach driven stronger by the aim of clustering is to find a number of clusters that fulfils all listed requirements separately instead of using a criterion that aggregates them without caring about specific details.

To this end, the *largest within-cluster gap wg* of a clustering can be defined as the maximum over all clusters of the dissimilarity belonging to the the last connecting edge of the minimum spanning tree within each cluster.

Cluster separation se of a clustering can be measured by computing, for all observations, the distance to the closest cluster to which the observation does not belong. Then the average of the minimum 10% of these distances is taken in order to consider only points close to the cluster borders (one should take

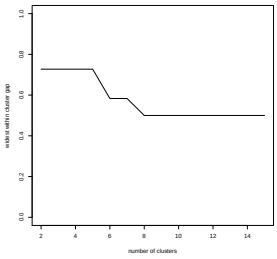

Fig. 2. Largest within-cluster gap for tetragonula bee data.

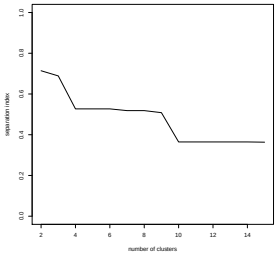

Fig. 3. Cluster separation for tetragonula bee data.

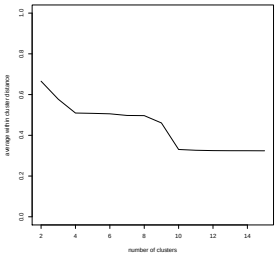

Fig. 4. Average within-cluster dissimilarity for tetragonula bee data.

a certain percentage of points into account in order to not make the index too dependent on a single observation).

Neither the *average within-cluster dissimilarity ad* nor the two statistics just introduced can be optimised over k, because increasing k will normally decrease all three of them.

Examining the statistics (see Figures 2-4), it turns out that *wg* does not become smaller than 0.5 for $k \leq 15$ and 0.5 is reached for $k \geq 8$. *se* falls from about 0.46 at $k = 9$ (which is fairly good) to below 0.4 for larger k. *ad* is 0.33 for $k = 10$, does not improve much for larger k, and is much higher for $k \leq 9$; 0.46 for $k = 9$. Overall this means that $k = 9$ and $k = 10$ can be justified, with

$k = 9$ much better regarding separation and $k = 10$ much better regarding *ad*, i.e., homogeneity.

Automatic aggregation of these aspects by formal criteria such as k_{ASW} or k_{PG} obscures the fact that in this situation the decision which of the requirements is more important really must come from subject matter expertise and cannot be determined from the data alone.

For the given problem, the importance of separation depends on how closely together the sampled individuals actually were taken geographically. From existing information on the sampling of individuals it can be seen that the two clusters merged going from $k = 10$ to 9 consist of individuals that are rather close together, in which case according to B. Hausdorf one would accept a weaker separation and demand more homogeneity. This favours the solution with $k = 10$. This solution is illustrated in Figure 1 by the colour bars on the left side and above (for $k = 9$, the two clusters in the upper right are merged).

Note that for this data set an expert decision about the existing species exists (cf. Franck et al. (2004); we did not use this in order to define criteria make decisions here), using information beyond the analysed data. This could be taken as a "ground truth" but one needs to keep in mind that there is no precise formal definition of a "species". Therefore experts will not always agree regarding species delimitation. According to the expert assessment there are 9 species in the data, but in fact the solution with $k = 10$ is the best possible one (in the sense of matching the species decided by Franck et al.) in the average linkage tree, because it matches the expert delimitation precisely except that one "expert species" is split up, which is in fact split up in all average linkage clusterings with $k \geq 6$ including $k = 9$, which instead merges two species that should be separated according to Franck et al. (2004).

4 Conclusion

The number of clusters for the tetragonula bees data set has been determined by listing a number of formal requirements for clustering in species delimitation and examining them all. This is of course strongly dependent on subjective judgements by the experts. Note though that subjective judgement is always needed if in fact the number of clusters depends on features such as separation and homogeneity, of which it is necessary to decide how to balance them. Supposedly objective criteria such as the ones discussed in Section 2 balance features automatically, but then the user still needs to choose a criterion, and this is a more difficult decision, because the meaning of the criteria in terms of the aim of the cluster analysis is more difficult to understand than statistics that formalise the requirements directly.

Optimally the statistician would like the expert to specify precise cutoff values for all criteria, which would mean that the best k could be found by a formal rule (e.g., the minimum k that fulfils all requirements). Unfortunately,

required cluster concepts such as the idea of a "species" are rarely precise enough to allow such exact formalisation.

The biggest obstacle for the presented approach is in fact that the requirements of clustering are in most cases ambiguous and formalisations are difficult to obtain. The fact that the subject matter experts often do not have enough training in mathematical thinking does not improve matters. However, using supposedly "objective" criteria in a more traditional fashion does not solve these problems but rather hides them.

The tetragonula bees data set has also been analysed by Hausdorf and Hennig (2010), using a method that allows for leaving some outliers out of all clusters. Indexes formalising separation and homogeneity would need an adaptation for such methods.

The new indexes for the largest within-cluster gap and cluster separation introduced above will soon be available in the R-package "fpc".

References

BOWCOCK, A.M., RUIZ-LINARES, A., TOMFOHRDE, J., MINCH, E., KIDD, J.R. and CAVALLI-SFORZA, L.L. (1994) High resolution of human evolutionary trees with polymorphic microsatellites. *Nature 368, 455–457.*

CALINSKI, R.B. and HARABASZ, J. (1974) A Dendrite Method for Cluster Analysis. *Communications in Statistics, 3, 1–27.*

CHATURVEDI, A.D., GREEN, P.E., CARROL, J.D. (2001) K-modes clustering. *Journal of Classification, 18, 35–55.*

FANG, Y. and WANG, J. (2012) Selection of the number of clusters via the bootstrap method. *Computational Statistics and Data Analysis, 56, 468–477.*

FRANCK, P., CAMERON, E., GOOD, G., RASPLUS, J.-Y. and OLDROYD, B.P. (2004) Nest architecture and genetic differentiation in a species complex of Australian stingless bees. *Molecular Ecology, 13, 2317–2331.*

HALKIDI, M., BATISTAKIS, Y. and VAZIRGIANNIS, M. (2001) On Clustering Validation Techniques. *Journal of Intelligent Information Systems 17, 107–145.*

HAUSDORF, B. and HENNIG, C. (2010) Species Delimitation Using Dominant and Codominant Multilocus Markers. *Systematic Biology, 59 491–503.*

HENNIG, C. (2010) Methods for merging Gaussian mixture components. *Advances in Data Analysis and Classification, 4, 3–34.*

JAIN, A.K. (2010) Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31, 651–666.*

KAUFMAN, L. and ROUSSEEUW, P.J. (1990) *Finding Groups in Data*. Wiley, New York.

MORLINI, I. and ZANI, S. (2012) A new class of weighted similarity indices using polytomous variables. *Journal of Classification, 29, 199–226.*