What Clusters Are Generated by Normal Mixtures?

Christian Hennig[1]

Institut für Mathematische Stochastik,
Bundesstr. 55, D-20146 Hamburg
(e-mail: hennig@math.uni-hamburg.de)

Abstract. Model based cluster analysis is often carried out by estimation of the parameters of a normal mixture. But mixture components do not necessarily reflect the idea of a "cluster". I discuss how to formalize the concept of "clusters" w.r.t. probability distributions on the real line by means of fixed point clusters, i.e., sets that do not contain any outlier and with respect to which the rest of the real line consists of outliers. The concept is applied to some normal mixtures.

1 Theoretical clusters of distributions

There are lots of different goals of cluster analysis (CA) and therefore there are lots of different meanings and formal definitions of the term "cluster". Most of them have in common that a cluster should be a group of entities which is "homogeneous" in some manner, i.e., the entities are similar to each other, and which is in some manner "separated" from the rest.

In CA based on probability models, it is assumed that the entities (usually from $\mathbb{R}^p$; here I consider only $\mathbb{R}$) are generated by some probability distribution. The clusters of the data are determined by using estimators of the features of the distribution (i.e., parameters or regions of high density). That is, the found clusters of data can be viewed as estimators of certain properties of the distribution as well, which I call "theoretical clusters". In order to explain the very meaning of the data clusters, it is necessary to think about the theoretical clusters generated by the considered model.

Often a CA is based on a normal mixture model of the form

$$P = \sum_{i=1}^{s} \epsilon_i \mathcal{N}(b_i, \sigma_i^2), \quad \sum_{i=1}^{s} \epsilon_i = 1,\ \epsilon_i > 0,\ i = 1, \ldots, s, \tag{1}$$

where $\mathcal{N}(b, \sigma^2)$ denotes the normal distribution with mean b and variance σ^2, the (b_i, σ_i^2), $i = 1, \ldots, s$, being pairwise distinct. Most of the following considerations apply also to normal fixed partition models. See Bock (1996) for literature concerning the use of these kinds of models in CA.

How to define the theoretical clusters of such normal mixtures? To my knowledge there are three approaches up to now used in the literature more or less implicitly (see Bock (1996) for references):

1. **Mixture components.** When data points are assigned to the clusters on the basis of the a posteriori probabilities of their generation by the mixture components, the component memberships define the theoretical clusters.
2. **Mode clusters.** Normal mixtures with many components may be unimodal. This illustrates that these components do not necessarily form reasonable clusters. Theoretical clusters may be defined alternatively by assigning each point to a "nearest" mode (to be defined) of the density.
3. **Areas of high density.** Since not every mode may be meaningful, all connected subsets of $\mathbb{R}$ where the density exceeds some threshold t can be defined as theoretical clusters.

It is reasonable to demand of a formal definition of a theoretical cluster, that it should not lead to a very different clustering if the distribution is only weakly modified in the sense that the modification can only be observed in case of a very large number of data points. This may be formalized by the Kolmogorov metric. The definitions above have certain drawbacks:

- In arbitrary small Kolmogorov-neighborhoods of a normal mixture there are mixtures with arbitrarily many components. The description of clusters as mixture components looses any sense if the underlying distribution is not precisely a normal mixture.
- In the neighborhoods of a distribution there are distributions with arbitrarily many modes. Moreover, the concept of a probability density is discontinuous. In arbitrary small neighborhoods of a distribution there are other ones with a completely different, or even without, Lebesgue density.
- Areas of density "above t" are not scale equivariant.

In the next section, fixed point clusters (FPCs) are defined. They provide an alternative definition of a "cluster" of a probability distribution. As mentioned above, there may be lots of ideas about a "cluster" and I do not claim that the approach presented here is more "correct" than others. But it has some reasonable properties that are not shared by the concepts above:

- It does not assume the existence of a Lebesgue density.
- It is a local concept in the sense that it is not based on a model for the whole dataset but only on a model for a single cluster and the relation to its neighborhood. This means in particular that an FPC of a distribution does not loose this property if the distribution is changed in areas far from its location.
- "Homogeneity" and "separation" are formalized directly in a way that varying ideas of a cluster may be implemented by choice of constants defining the degree of similarity to a homogeneous reference distributions and the degree of separation required for a theoretical cluster.

The practical applicability of the approach is discussed briefly in the conclusion.

2 Definition of fixed point clusters

The idea of FPCs is that a set of points is homogeneous if it does not contain any outlier, and separated from the rest if all other points are outliers with respect to it. Davies and Gather (1993) define outliers as follows: First they choose a reference model for the "good" part of the data, say, the family of $\mathcal{N}(b, \sigma^2)$-distributions, $b \in \mathbb{R}$, $\sigma^2 > 0$. Then they specify an "outlier region" with probability smaller than α, reflecting the idea of "outlyingness":

$$A(\mathcal{N}(b, \sigma^2), \alpha) = \{(x - b)^2 > c(\alpha)\sigma^2\}, \tag{2}$$

where $c(\alpha)$ is the $(1-\alpha)$-quantile of the χ^2_1-distribution. All points of $A(\mathcal{N}(b, \sigma^2), \alpha)$ are called "α-**outliers**" with respect to $\mathcal{N}(b, \sigma^2)$. This definition may be extended to other distributions P on the real line:

$$A(P, \alpha) = \begin{cases} \emptyset & \text{if E or Var do not exist,} \\ \{(x - \mathrm{E}P)^2 > c(\alpha)\mathrm{Var}P\} & \text{else.} \end{cases} \tag{3}$$

That is, P is treated as if it were a normal distribution, the reference for a single cluster, and expectation and variance should be meaningful to P. Replacement of EP and $\mathrm{Var}P$ by the more robust functionals Median and 1.483 MAD leads to similar results in section 3.

Definition 1. Given some $0 < \alpha < 1$, a subset B of the real line is called a (theoretical) **fixed point cluster** w.r.t. the distribution P, if

$$B = A(P_B, \alpha)^c \text{ and } P(B) > 0, \text{ where } P_B(C) = P(C|B)\ \forall C.$$

That is, B has to be exactly the set of non-outliers w.r.t. P_B, which is the restriction of P to B itself. The given definition may be adapted to other definitions of outlier regions A, i.e., to other classes of reference distributions (Hennig (1998)) defining other concepts of "homogeneity". The constant α defines the degree of required separatedness: The smaller α, the more points are non-outliers, and the more separation is needed for a set to be an FPC.

The equation (2) forces all FPCs to be closed intervals. Here is an equivalent representation: The interval $I_{a,s^2} = \{x : (x-a)^2 \le c(\alpha)s^2\}$ with $s^2 > 0$ is an FPC w.r.t. P iff (a, s^2) is a fixed point of f_P, where

$$f_P(a, s^2) = (\mathrm{E}P_{I_{a,s^2}}, \mathrm{Var}P_{I_{a,s^2}}). \tag{4}$$

E and Var always exist for distributions of the form $P_{I_{a,s^2}}$, $|a|, s^2 < \infty$.

The following result is a trivial consequence of the definition. It shows that the FPC property of some set B does not depend on untypical areas of $\mathbb{R}$, "untypical" meaning that they consist only of outliers w.r.t. P_B:

Corollary 1. Let P and Q be distributions on $\mathbb{R}$ and B a set with $P(B) > 0$, $Q(B) > 0$, $P_B = Q_B$. Then B is an FPC w.r.t. P iff it is an FPC w.r.t. Q.

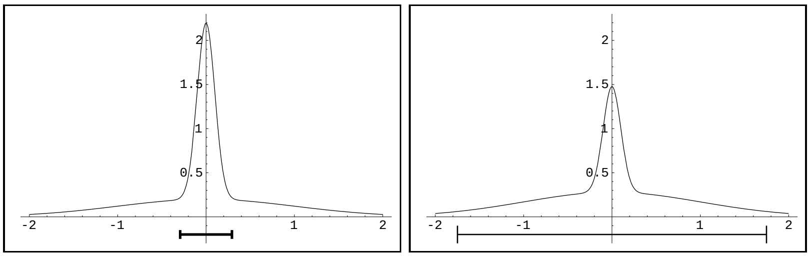

Fig. 1. Densities of $P_{2,0.5}$ and $P_{2,0.7}$ with FPCs.

3 Fixed point clusters of normal mixtures

If P is a normal mixture, f_P may be calculated explicitly. In the following the FPCs of such distributions are discussed. This requires the choice of α which will be discussed below. By default, $\alpha = 0.01$.

Example 1. For $P_1 = \mathcal{N}(0,1)$ there exists a unique FPC at $(a_1, s_1^2) = (0, 0.9001)$. If $\alpha \leq 0.1$, there always exists a unique FPC with $a = 0$ and $s^2(\alpha) \nearrow 1$ with $\alpha \searrow 0$. This is reasonable: A homogeneous $\mathcal{N}(b, \sigma^2)$-distribution leads to a single cluster, the interval $[b - 2.44\sigma, b + 2.44\sigma]$ for $\alpha = 0.01$ (FPCs are scale and location equivariant). Corollary 1 says that this interval remains FPC for all distributions of the form $(1-\epsilon)\mathcal{N}(b, \sigma^2) + \epsilon P^*$, where P^* puts mass 0 on $[b - 2.44\sigma, b + 2.44\sigma]$. If α gets smaller, this interval gets larger.

Example 2. $P_{2,p} = p\mathcal{N}(0,1) + (1-p)\mathcal{N}(0, 0.01)$, see Figure 1. This is a mixture with two components, but unimodal. For $p = 0.5$ there exists a unique FPC with $(a_{21}, s_{21}^2) = (0, 0.013)$. The FPC corresponds obviously to the mixture component with variance 0.01. This component is separated well from the other one in the sense that there is only a very small mass from $\mathcal{N}(0,1)$ in the area of its non-outliers. On the other hand, the area of non-outliers of $\mathcal{N}(0,1)$ contains almost the whole mass of $\mathcal{N}(0, 0.01)$ so that it does not form a cluster in the sense of fixed point clustering. The situation changes when increasing p: For $p = 0.7$ (Figure 1, right side), the first component puts so much mass in the area of the second one that the second component is no longer clearly separated from the first. A kind of compromise, namely $(a_{22}, s_{22}^2) = (0, 0.461)$, results as the unique FPC. If $p \nearrow 1$, the FPC converges to that of $\mathcal{N}(0,1)$.

Example 3. $P_{3,d} = 0.5\mathcal{N}(0,1) + 0.5\mathcal{N}(d, 0.25)$, see Figure 2. For $d = 5$ there are five FPCs defined by $(a_{31}, s_{31}^2) = (0, 0.9)$, $(a_{32}, s_{32}^2) = (5, 0.23)$, $(a_{33}, s_{33}^2) = (0.4, 2.39)$, $(a_{34}, s_{34}^2) = (4.45, 1.89)$ and $(a_{35}, s_{35}^2) = (2.5, 6.87)$. The first two FPCs (marked fat in Figure 2) correspond clearly to the well separated mixture components. One may not expect further theoretical clus-

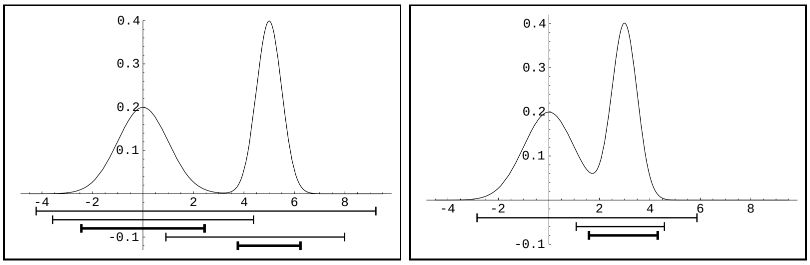

Fig. 2. Densities of $P_{3,5}$ and $P_{3,3}$ with FPCs.

(a, s^2, P)	$D(a, s^2, P)$	(a, s^2, P)	$D(a, s^2, P)$
$(a_{21}, s_{21}^2, P_{2,0.5})$	**0.010**	$(a_{31}, s_{31}^2, P_{3,5})$	$\mathbf{6 * 10^{-5}}$
$(a_{22}, s_{22}^2, P_{2,0.7})$	0.100	$(a_{32}, s_{32}^2, P_{3,5})$	**0.003**
$(a_{36}, s_{36}^2, P_{3,3})$	**0.015**	$(a_{33}, s_{33}^2, P_{3,5})$	0.114
$(a_{37}, s_{37}^2, P_{3,3})$	0.068	$(a_{34}, s_{34}^2, P_{3,5})$	0.236
$(a_{38}, s_{38}^2, P_{3,3})$	0.141	$(a_{35}, s_{35}^2, P_{3,5})$	0.207

Table 1. Kolmogorov distances of mixture FPCs to an FPC from a homogeneous normal distribution. Fat numbers correspond to fat marked FPCs in the figures.

ters for this density, but there are three of them. The reason is that the classification of outliers defined in (3) is based on the normal distribution and works well if P is at least similar to a normal. But f_P may have fixed points at intervals I where P_I deviates strongly from a normal distribution.

In section 2 it was discussed that the meaning of "homogeneity" is defined by the class of reference distributions, i.e. the normal family. The homogeneous FPCs in this sense can be distinguished from the non-homogeneous by the requirement that they should be similar to FPCs from a single normal distribution by means of the Kolmogorov metric.

Definition 2. A closed interval $I = [a - \sqrt{c(\alpha)s^2}, a + \sqrt{c(\alpha)s^2}]$ is called a **normal fixed point cluster** w.r.t. P, if it is a fixed point cluster and

$$D(a, s^2, P) = \sup_{x \in \mathbb{R}} \left| F_I(x) - \Phi_{a, \frac{s^2}{s^2(\alpha)}, I}(x) \right| < q,$$

where F_I denotes the c.d.f. of P_I, $\Phi_{b,\sigma^2,I}$ denotes the c.d.f. of $\mathcal{N}(b, \sigma^2)$ restricted to I, q is a constant that specifies how close to a normal distribution the theoretical clusters should lie, say, $q = 0.05$ or $q = 0.1$. $s^2(\alpha)$ is defined as in Example 1 and denotes the variance of the FPC of an $\mathcal{N}(0, 1)$-distribution. Table 1 shows that the Kolmogorov distance serves well to distinguish optically reasonable FPCs from others. The FPCs defined by (a_{22}, s_{22}^2) and

(a_{37}, s^2_{37}) (see below) illustrate the degree of normal similarity required by $q = 0.05,\ q = 0.1$ respectively.

Example 3. Continued. For $d = 3$ there are three FPCs at $(a_{36}, s^2_{36}) = (2.95, 0.28)$, $(a_{37}, s^2_{37}) = (2.83, 0.46)$, $(a_{38}, s^2_{38}) = (1.51, 2.86)$, of which the first one (fat in Figure 2) is the only normal FPC for $q = 0.05$. If d decreases, the mixture components stop to be well enough separated to produce FPCs. Here, the second mixture component with small variance is still sufficiently "cluster-shaped" as opposed to the first one, which disappeared at about $d = 3.7$. At about $d = 2.8$, the FPC with the small variance vanishes as well. The amount of separation necessary for an FPC decreases with increasing α: For $\alpha = 0.025$, the FPC of the second component vanishes at about $d = 2.4$, and for $\alpha = 0.05$, even the first component keeps its FPC down to $d = 2.6$.

4 Conclusion

A proposal is made to define the theoretical clusters for distributions by direct formalization of "homogeneity" and "separation". While the definition of "homogeneity" is based on the normal reference distributions here, the FPCs may be calculated for arbitrary distributions. It is clear that the meaning of a cluster has to depend on individual aims and judgements. The approach of fixed point clustering shows where such judgements may enter: The required separation can be varied by choice of α, the homogeneity by the choice of the definition of the outlier region A and by the cutoff q for the Kolmogorov distance. There are two ways to apply the FPC concept to concrete data:

- The computation of theoretical FPCs of an estimated distribution can help to decide about the occurence of well separated, homogeneous clusters, if a density or the parameters of a normal mixture have been estimated.
- The definition of FPCs can be transferred directly to datasets. The resulting CA method is sketched for various definitions of outlier regions in Hennig (1998) and will be analyzed elsewhere in greater detail. Software for this is available from
`http://www.math.uni-hamburg.de/home/hennig/`

References

BOCK, H.H.(1996): Probability Models and Hypotheses Testing in Partitioning Cluster Analysis. In: P. Arabie, L. J. Hubert, G. De Soete (Eds.): *Clustering and Classification.* World Scientific Publishers, New Jersey, 377–453.

DAVIES, P.L. and GATHER, U.(1993): The identification of multiple outliers. *Journal of the American Statistical Association*, *88*, 782–801.

HENNIG, C.(1998): Clustering and Outlier Identification: Fixed Point Cluster Analysis. In: A. Rizzi, M. Vichi, H.-H. Bock (Eds.): *Advances in Data Science and Classification.* Springer, Berlin, 37–42.