An improvement of the NEC criterion for assessing the number of clusters in a mixture model

Christophe Biernacki, Gilles Celeux

INRIA Rhône-Alpes

Gérard Govaert

UMR CNRS 6599 - UTC Compigne

Abstract: The entropy criterion NEC showed good performances for choosing the number of clusters arising from a mixture model. But it was not valid to decide between one and more than one clusters. This note presents a natural extension of this criterion to deal with this situation. Illustrative experiments exhibit good behavior of this modified entropy criterion.

Keywords: Cluster analysis; Mixture model; Normalized entropy criterion

1 The improved NEC criterion

Celeux and Soromenho (1996) proposed an entropy criterion called NEC (Normalized Entropy Criterion) to estimate the number of clusters arising from a mixture model. This criterion, derived from a relation linking the likelihood and the classification likelihood of a mixture works well, but, contrary to other criteria as AIC, BIC or ICOMP (see for instance Cutler and Windham 1993), suffers of limitation to decide between one and more than one clusters. Celeux and Soromenho (1996) proposed a somewhat cumbersome rule of thumb to deal with this problem, but their procedure was restricted to Gaussian mixtures and has shown a disappointing behavior (see Biernacki and Govaert 1997). In this note, we propose a simpler and general procedure to deal with this problem. This procedure is well justified and numerical experiments reported here show good performances.

In the mixture model, the data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are assumed to be a sample from a probability distribution with density

$$\phi(\mathbf{x}) = \sum_{k=1}^{K} p_k f(\mathbf{x}, \mathbf{a}_k) \tag{1.1}$$

where the p_k's are the mixing proportions ($0 < p_k < 1$ for all $k = 1, \ldots, K$ and $\sum_k p_k = 1$) and the $f(\mathbf{x}, \mathbf{a}_k)$ are densities from the same parametric family. The maximized log-likelihhod of the sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is

$$L(K) = \sum_{i=1}^{n} \ln \left[\sum_{k=1}^{K} \hat{p}_k f(\mathbf{x}_i, \hat{\mathbf{a}}_k) \right], \tag{1.2}$$

$\hat{p}_k$ and $\hat{\mathbf{a}}_k$ denoting the maximum likelihood estimates of the corresponding parameters. The NEC criterion is derived from a relation between the log-likelihhod $L(K)$ and a classification-type log-likelihhod $C(K)$. Denoting

$$t_{ik} = \frac{\hat{p}_k f(\mathbf{x}_i, \hat{\mathbf{a}}_k)}{\sum_{j=1}^{K} \hat{p}_j f(\mathbf{x}_i, \hat{\mathbf{a}}_j)}$$

as the conditional probability that x_i arises from the kth mixture component ($1 \leq i \leq n$ and $1 \leq k \leq K$), direct calculations show that

$$L(K) = C(K) + E(K), \tag{1.3}$$

with

$$C(K) = \sum_{k=1}^{K} \sum_{i=1}^{n} t_{ik} \ln \left[\hat{p}_k f(\mathbf{x}_i, \hat{\mathbf{a}}_k)\right],$$

and

$$E(K) = -\sum_{k=1}^{K} \sum_{i=1}^{n} t_{ik} \ln t_{ik} \geq 0.$$

Relation (1.3) provides a decomposition of the log-likelihhod $L(K)$ in a classification log-likelihhod term $C(K)$ and the entropy $E(K)$ of the fuzzy classification matrix $\mathbf{t} = [t_{ik}]$ which measures the overlap of the mixture components. The entropy $E(K)$ cannot be used directly as a criterion to assess the number of clusters in a mixture since $L(K)$ is an increasing function of K and has to be normalized. We have

$$1 = \frac{C(K) - C(1)}{L(K) - L(1)} + \frac{E(K) - E(1)}{L(K) - L(1)}, \quad K > 1, \tag{1.4}$$

and the normalized entropy criterion to be minimized for assessing the number of clusters arising from a mixture is, since $E(1) = 0$,

$$\mathrm{NEC}(K) = \frac{E(K)}{L(K) - L(1)}.$$

But, NEC(1) is not defined. Thus, we are unable to compare the situations $K = 1$ versus $K > 1$ directly when using the criterion NEC(K). And there is a need to extend NEC to this particular problem.

An ad hoc procedure for the one cluster case for Gaussian mixtures. Celeux and Soromenho (1996) proposed a procedure to circumvent this difficulty in the Gaussian mixture setting, which works as follows:

- Let K^* be the number of clusters minimizing $\mathrm{NEC}(K), K > 1$. To decide $K = K^*$ or $K = 1$, we estimate the parameters of a K^* component Gaussian mixture with equal means $\mu_1 = \ldots = \mu_{K^*} = \overline{\mathbf{x}}$, where $\overline{\mathbf{x}}$ represents the sample mean of the data

$(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, equal proportions $p_1 = \ldots = p_{K^*} = 1/K^*$ and free variance matrices. Denoting $\tilde{L}(1)$ and $\tilde{E}(1)$ the resulting log-likelihood and entropy, the ratio

$$\mathrm{NEC}(1) = \frac{\tilde{E}(1)}{\tilde{L}(1) - L(1)}$$

is computed and K^* clusters is chosen if $\mathrm{NEC}(K^*) \leq \mathrm{NEC}(1)$, otherwise no clustering structure in the data is declared.

As mentioned in the introduction, this procedure can perform poorly in some situations. Possible reasons for this deceptive behavior are the followings:

- Constraining the proportions to be equal to avoid degenerate solutions is not really justified.
- When a Gaussian mixture with equal variance matrices is considered as clustering model, the mixture models considered in the situation $K = 1$ (unequal variance matrices) and the situation $K > 1$ are different.

A general procedure for the one cluster case. As stressed in Celeux and Soromenho (1996) and shown through numerical experiments in Biernacki and Govaert (1997) and Celeux and Soromenho (1996), the NEC criterion is designed to choose the mixture model providing the greatest evidence for partitioning data. To decide between $K > 1$ and $K = 1$, we take the same point of view.

First, note that $C(K)$ measures the fit of partitioning the data into K clusters and $C(1) = L(1)$ measures the fit of the one cluster solution. When comparing two number of clusters, K and K', the more complex model, with the larger number of clusters, can be expected to provide a better fit. Hence, in a parsimony perspective, it is reasonable to do not favor K to K' if $K > K'$ and $C(K) < C(K')$. As a consequence, to choose $K > 1$ rather than $K = 1$ it is natural requiring that the classification likelihood $C(K) > C(1) = L(1)$. It is worthwhile to remark that if $C(K) > L(1)$ then all the terms in (1.4) are non negative and, thus, $0 \leq \mathrm{NEC}(K) \leq 1$: the only numbers $K > 1$ of clusters of interest must verify this condition. If there is no $K > 1$ such that $\mathrm{NEC}(K) \leq 1$, there is no reason to select more than one cluster. Finally, to decide between $K > 1$ and $K = 1$, we propose to simply proceed as follows:

Let K^* be the value minimizing $NEC(K), (2 \leq K \leq K_{sup})$, K_{sup} being an upper bound for the number of mixture components. We choose K^* clusters if $NEC(K^*) \leq 1$, otherwise we declare no clustering structure in the data.

Remark that it simply consists in choosing K minimizing $\mathrm{NEC}(K)$ with the convention $\mathrm{NEC}(1) = 1$. It is worth noting that this procedure is consistent in the following sense: It is easily seen that if $K = 1$ is preferred to $K = K^*$ then $K = 1$ is preferred to any other $K > 1$.

2 Numerical experiments

We examine through numerical experiments if this improved version of the NEC criterion does the job and avoids the tendency of the old NEC criterion to prefer $K > 1$ clusters when the true number of mixture components is $K = 1$. We consider experiments for Gaussian mixtures and for Bernoulli mixtures.

2.1 Gaussian mixtures

At first, 30 samples of 200 points are simulated from a Gaussian distribution in R^2 with mean $\mu = (0, 0)'$ and variance matrix unity $\Sigma = I$. Two different mixture models are considered: both models assumed a common variance matrix of the form $\Sigma = \lambda I$, but the first model assumes equal proportions while the second one considers no constraint on the proportions. In all our experiments, we used the EM algorithm to estimate the parameters of the mixture at hand. The numbers of clusters in competition were $K = 1, \ldots, 5$. We initiate the EM algorithm with the true parameter values when K was the true number of components. Otherwise, we ran 20 times the EM algorithm from 20 random initial positions and selected the solution providing the largest likelihood. Results are displayed in Table 1. In this table, oNEC denotes the old NEC criterion and NEC the NEC criterion incorporating our new procedure to deal with the 'one cluster' case. It appears that NEC outperforms the old procedure oNEC in this situation. Note that the mean value of NEC(K^*), $K^* \neq 1$,

	equal prop.		free prop.	
K	oNEC	NEC	oNEC	NEC
1	37	**97**	0	**90**
2-5	**63**	3	**100**	10

Table 1: Percentage of choosing K clusters from a single Gaussian distribution with the old NEC criterion (oNEC) and the new one (NEC).

is 75.8 with standard error 66.2 in the equal proportion situation and 5.1 with standard error 2.8 in the free proportion situation.

Now, it is important to examine if the new NEC procedure has a tendency to favor the answer $K = 1$ when there is more than one cluster present in a data set. In this purpose, we simulated 30 samples of 200 points from a two component Gaussian mixture with the following parameter values: $p_1 = p_2 = 0.5$, $\mu_1 = (0, 0)'$, $\mu_2 = (d, 0)'$ and $\Sigma_1 = \Sigma_2 = I$. We assume a mixture model with equal proportions and a common spherical variance matrix

$\Sigma = \lambda I$. We only compare $K = 1$ and $K = 2$. Varying the value of d from 0 to 4 by step of 0.1, we compute the oNEC and NEC values. In Figure 1, we display the percentage of choosing $K = 1$ for both criteria in function of d. We determine for both procedures oNEC and NEC the smallest value of d leading to decide $K = 2$ in more than 80% of the 30 experiments. For oNEC this value is $d = 1.5$, and for NEC this value is $d' = 3.0$. No surprisingly, $d' > d$. In Figure 2 we display two samples obtained with $d = 1.5$ on the left hand and $d = 3.0$ on the right hand. From this figure, it is apparent that the evidence of a two cluster structure is weak for $d = 1.5$ and clear for $d = 3$. Thus, the choice of NEC is in accordance with the aim of selecting a model providing the greatest evidence to partitioning data.

Figure 1 about here

Figure 2 about here

2.2 Bernoulli mixtures

We consider now binary data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \{0, 1\}^d$ is a sample from a Bernoulli mixture (see for instance Goodman 1994). The kth mixture component has density

$$f(\mathbf{x}, \mathbf{a}_k) = \prod_{j=1}^{d} (a_k^j)^{x^j} (1 - a_k^j)^{1-x^j},$$

with $\mathbf{a}_k = (a_k^1, \ldots, a_k^d)$, where a_k^j gives the probability that $x^j = 1$ in the class k for $j = 1, \ldots, d$ and $k = 1, \ldots, K$.

At first, 30 samples of size $n = 500$ were simulated from a Bernoulli mixture with one cluster in dimension $d = 6$ with the parameter value $\mathbf{a} = (0.9\ 0.9\ 0.1\ 0.1\ 0.1\ 0.1)$. For $K = 1, \ldots, 6$, we ran 20 times the EM algorithm from 20 random initial positions and selected the solution providing the largest likelihood. Pourcentage of choosing K clusters and mean value of NEC into parentheses are displayed in Table 2: the NEC criterion prefers dramatically the one cluster solution.

Then, 30 samples in $\{0, 1\}^6$ of size $n = 500$ were simulated from three well-separated component Bernoulli mixture with the parameter values $\mathbf{a}_1 = (0.9\ 0.9\ 0.1\ 0.1\ 0.1\ 0.1)$, $\mathbf{a}_2 = (0.1\ 0.1\ 0.9\ 0.9\ 0.1\ 0.1)$, $\mathbf{a}_3 = (0.1\ 0.1\ 0.1\ 0.1\ 0.9\ 0.9)$ and equal mixing proportions. We used the same process as before and results are displayed in the same way in Table 2: the NEC criterion always selects the three cluster solution. It is also worth noting that mixtures with more than one component are highly preferred to the one component mixture.

K	none cluster case		three cluster case	
1	**100**	(1)	0	(1)
2	0	(42.32)	0	(0.15)
3	0	(39.10)	**100**	(0.09)
4-6	0	(45.70)	0	(0.30)

Table 2: Percentage of choosing K clusters and, into parentheses, mean of NEC values in the one and the three cluster cases for Bernoulli mixtures.

3 Discussion

With this improvement to decide between $K = 1$ and $K > 1$, the NEC criterion appears to be a powerful criterion to estimate the number of clusters arising from a mixture model. Actually, this criterion is less sensitive than other criteria as AIC, BIC and Classification Likelihood criteria to violations of the assumptions of the considered mixture model for partitioning data (cf. Biernacki and Govaert, 1997 and Biernacki, 1997). Moreover, it is interpretable in its own terms. From a practical point of view, a NEC value near zero indicates that the associated partition is of interest. Thus, examining the value of NEC for different number of clusters can suggest several possible pertinent partitions. Let us illustrate this fact with an example.

We simulate a sample from a Gaussian mixture with three well-separated components: $p_1 = p_2 = p_3 = 1/3$, $\mu_1 = (0,0)'$, $\mu_2 = (3,0)'$, $\mu_3 = (7,7)'$, $\Sigma_1 = \Sigma_2 = \Sigma_3 = I$. This sample is depicted in Figure 3. In Table 3, we display values of NEC derived from this sample when using the mixture model assuming equal proportions and a common identity variance matrix.

Figure 3 about here

NEC prefers the two cluster solution because when compared to the three cluster solution, NEC favors the more evident partitioning structure. But the small NEC value for $K = 3$ (NEC=0.0728) shows clearly that the three cluster solution is also quite sensible.

K	2	3	4
NEC	0.0003	0.0728	0.2264

Table 3: NEC values for the data depicted Figure 3.

References

Biernacki C. (1997). Choix de modèles en classification. PhD thesis, Compiègne university of technology.

Biernacki C. and Govaert G. (1997). Using the classification likelihood to choose the number of clusters. *Computing Science and Statistics*, **29**, 2, 451-457.

Celeux G. and Soromenho G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, **13**, 195-212.

Cutler A. and Windham M. P. (1993). Information-based validity functionals for mixture analysis. *Proceedings of the first US-Japan Conference on the Frontiers of Statistical Modeling.* Ed. H. Bozdogan, Amsterdam: Kluwer, pp.149-170.

Goodman L. A. (1974) Exploratory latent structure models using both identifiable and unidentifiable models. *Biometrika*, **61**, 215-231.

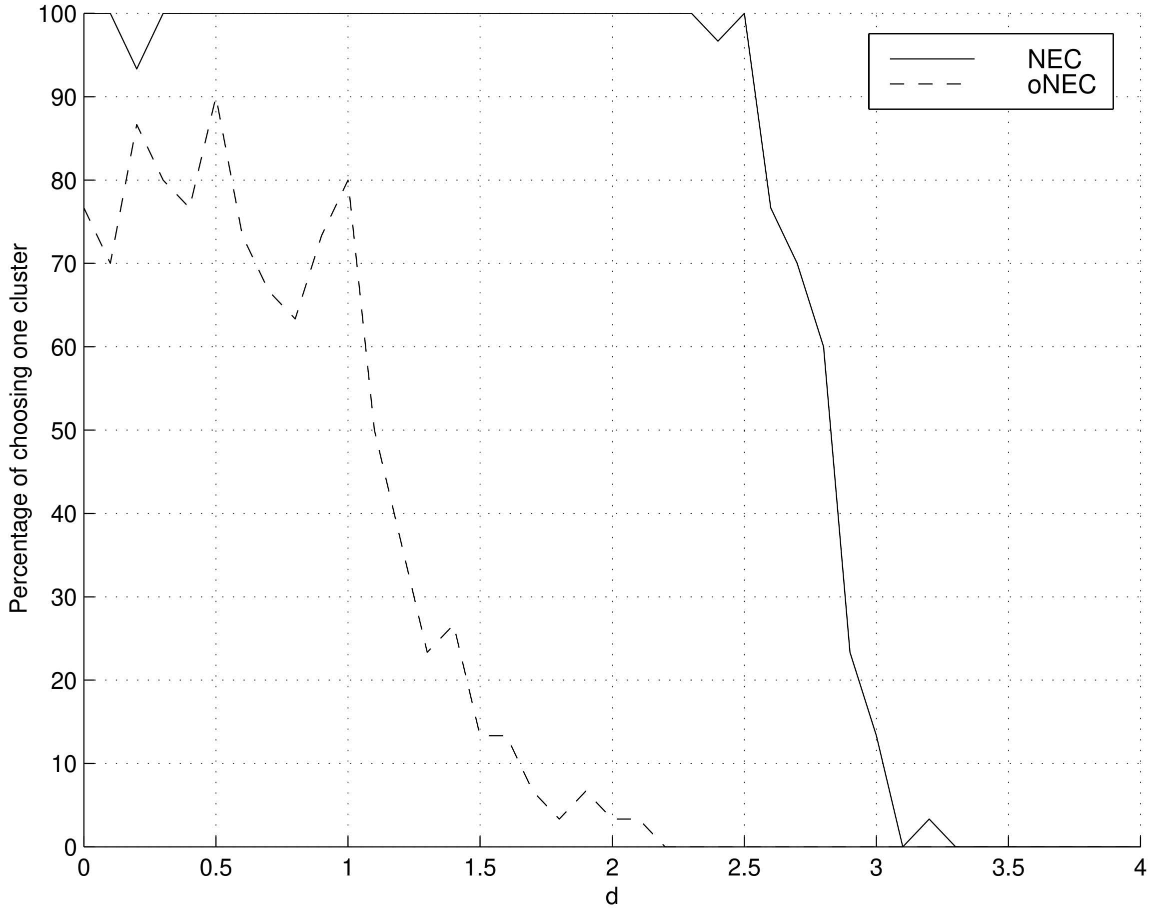

Figure 1: Percentage of choosing one cluster for oNEC and NEC criteria in function of the separation d.

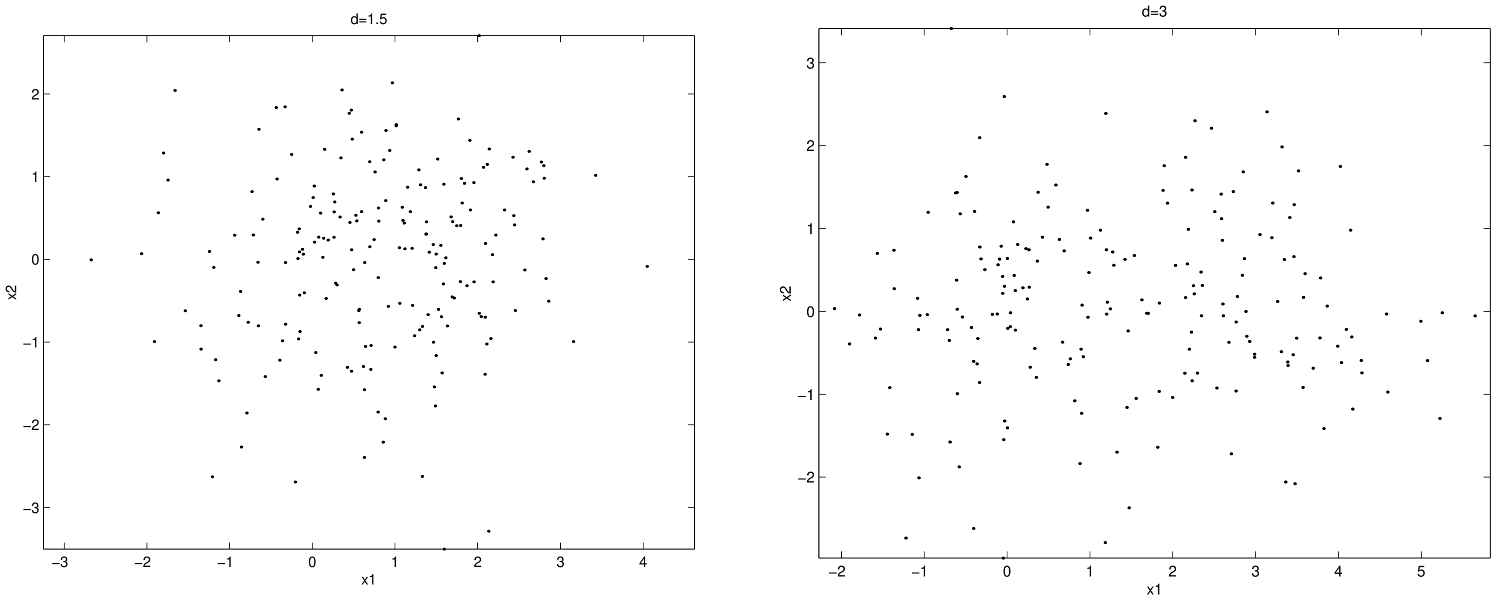

Figure 2: Two samples respectively with the separation $d = 1.5$ (left) and $d = 3.0$ (right).

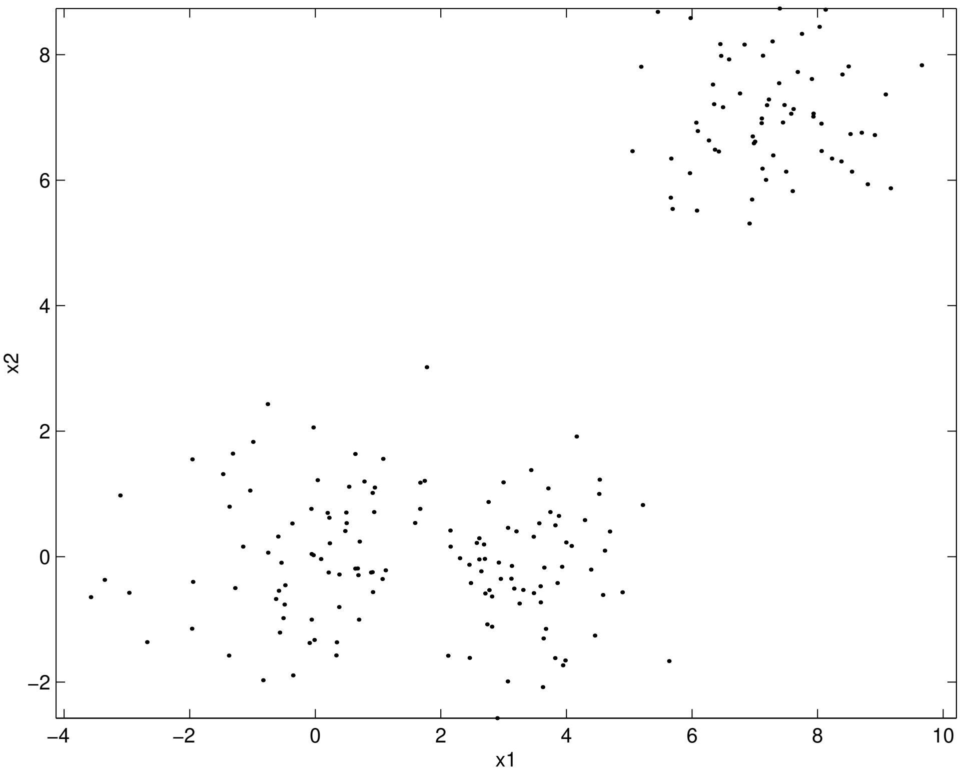

Figure 3: A sample of three well-separated clusters.