

EVALUATION OF NEGENTROPY-BASED CLUSTER VALIDATION TECHNIQUES IN PROBLEMS WITH INCREASING DIMENSIONALITY

L. F. Lago-Fernández, G. Martínez-Muñoz, A. M. González and M. A. Sánchez-Montañés

Escuela Politécnica Superior, Universidad Autónoma de Madrid, Madrid, Spain

{luis.lago, gonzalo.martinez, ana.marcos, manuel.smontanes}@uam.es

Keywords: Clustering, Cluster validation, Model selection.

Abstract: The aim of a crisp cluster validity index is to quantify the quality of a given data partition. It allows to select the best partition out of a set of potential ones, and to determine the number of clusters. Recently, negentropy-based cluster validation has been introduced. This new approach seems to perform better than other state of the art techniques, and its computation is quite simple. However, like many other cluster validation approaches, it presents problems when some partition regions have a small number of points. Different heuristics have been proposed to cope with this problem. In this article we systematically analyze the performance of different negentropy-based validation approaches, including a new heuristic, in clustering problems of increasing dimensionality, and compare them to reference criteria such as AIC and BIC. Our results on synthetic data suggest that the newly proposed negentropy-based validation strategy can outperform AIC and BIC when the ratio of the number of points to the dimension is not high, which is a very common situation in most real applications.

1 INTRODUCTION

Negentropy-based cluster validation has been recently introduced (Lago-Fernández and Corbacho, 2010). It aims at finding well separated and compact clusters, and has a number of advantages such as the simplicity of its calculation, which only requires the computation of the log-determinants of the covariance matrices and the prior probabilities for each cluster. It can deal satisfactorily with clusters with heterogeneous orientations, scales and densities, and has been shown to outperform other classic validation indices on a range of synthetic and real problems.

However, like many other cluster validation approaches (Gordon, 1998; Xu and II, 2005), negentropy based validation presents difficulties when validating clustering partitions with very small clusters. In these cases, the quality of the estimation of the log-determinant of the covariance matrix involved in the computation of the negentropy index can be very poor, with a strong bias towards $-\infty$, as shown in (Lago-Fernández et al., 2011). This can bias the validation index towards solutions with too many clusters if no additional requirements, such as constraints on the minimum number of points per cluster, are imposed. In the mentioned study this problem is for-

mally analyzed, and a correction to the bias is proposed. A heuristic for cluster validation based on the negentropy index is also introduced. This heuristic takes into account the variance in the estimation of the negentropy index, and allows to disregard clustering partitions with a low negentropy index but a high variance.

In this work we propose a more formal heuristic that refines the correction of the negentropy index proposed in (Lago-Fernández et al., 2011) in order to quantify the confidence levels of the index value. Additionally, we improve their analysis studying the performance of different negentropy based validation approaches with respect to the number of dimensions. In order to make a systematical test, we use a benchmark database that spans a broad range of dimensions. This benchmark is based on the *twonorm* classification problem (Breiman, 1996). In order to interpret our results in a proper context, we compare the performance of the different negentropy-based cluster validation approaches with AIC (Akaike, 1974) and BIC (Schwartz, 1978; Fraley and Raftery, 1998).

Our results on synthetic data show that, in general, for low dimensions negentropy-based cluster validation performs well when there is not a high overlap amongst the clusters. On the other hand, when the

clusters are highly overlapped the BIC index can provide better results, as long as the number of points per cluster is high enough. Note however that BIC is intended for fitting distributions rather than for clustering, so it can deal well with the overlap. We also find that, when the ratio of the number of points to the dimension is small, negentropy-based methods can outperform BIC. Given that this is usual in real applications, and given the simplicity in the calculation of negentropy-based indices, we strongly encourage its application for real clustering problems.

2 THE NEGENTROPY INDEX

Let us consider a random variable \mathbf{X} in a d -dimensional space, distributed according to the probability density function $f(\mathbf{x})$. Let $s = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ be a random sample from \mathbf{X} , and let us consider a partition of the space into a set of k non-overlapping regions $\Omega = \{\omega_1, \dots, \omega_k\}$ that cover the full data space. This partition imposes a crisp clustering structure on the data, with k clusters each consisting of the data points falling into each of the k partition regions. The negentropy increment of the clustering partition Ω applied to \mathbf{X} is defined as (Lago-Fernández and Corbacho, 2010):

$$\Delta J(\Omega, \mathbf{X}) = \frac{1}{2} \sum_{i=1}^k p_i \log |\Sigma_i| - \sum_{i=1}^k p_i \log p_i \quad (1)$$

where p_i and Σ_i are the prior probability and covariance matrix respectively for \mathbf{X} restricted to the region ω_i . The negentropy increment is a measure of the average normality that is gained by making a partition on the data. The lower the value of $\Delta J(\Omega)$, the more Gaussian the clusters are on average, therefore the rule for cluster validation is to select the partition that minimizes the negentropy increment index. Of course, in any practical situation we do not have knowledge of the full distribution of \mathbf{X} , and we have to estimate the negentropy increment from the finite sample s . A straightforward estimation can be done using the index:

$$\Delta J_B(\Omega, s) = \frac{1}{2} \sum_{i=1}^k \tilde{p}_i \log |\tilde{\Sigma}_i| - \sum_{i=1}^k \tilde{p}_i \log \tilde{p}_i \quad (2)$$

where \tilde{p}_i and $\tilde{\Sigma}_i$ are the sample estimations of p_i and Σ_i respectively. The subindex B has been introduced to emphasize that this estimation of the negentropy increment is biased due to a wrong estimation of the terms involving the log-determinants (Lago-Fernández et al., 2011). This bias can be corrected

using the expression:

$$\Delta J_U(\Omega, s) = \Delta J_B(\Omega, s) + \frac{1}{2} \sum_{i=1}^k \tilde{p}_i C(n_i, d) \quad (3)$$

where $C(n_i, d)$ is a correction term for the log-determinant which depends only on the number of sample points in region i , n_i , and on the dimension d (Misra et al., 2005):

$$C(n_i, d) = -d \log \frac{2}{n_i - 1} - \sum_{j=1}^d \Psi\left(\frac{n_i - j}{2}\right) \quad (4)$$

Here Ψ is the digamma function (Abramowitz and Stegun, 1965). It can be shown that this new estimator is unbiased, that is:

$$E[\Delta J_U(\Omega, s)]_s = \Delta J(\Omega) \quad (5)$$

And that the variance of $\Delta J_U(\Omega, s)$ can be estimated as:

$$\sigma_s^2(\Delta J_U) \approx \frac{1}{4} \sum_{i=1}^k \tilde{p}_i^2 \sum_{j=1}^d \Psi'\left(\frac{n_i - j}{2}\right) \quad (6)$$

where Ψ' is the first derivative of the digamma function, also known as trigamma function. Different uses of these results lead to the different validation approaches presented in the following section.

3 VALIDATION APPROACHES

3.1 Negentropy-based Approaches

The general rule for cluster validation based on the negentropy increment is that, given a set of clustering partitions $\Pi = \{\Omega_1, \dots, \Omega_M\}$ on a given problem defined by the random variable \mathbf{X} , one should select the partition Ω_i for which $\Delta J(\Omega_i)$ is minimum. That is:

$$\Delta J(\Omega_i) \leq \Delta J(\Omega_j) \quad \forall \Omega_j \in \Pi$$

This means that the clusters resulting from Ω_i are, on average, more Gaussian than those resulting from any other partition in Π . In practical terms, we never know the values $\Delta J(\Omega_i)$, but only estimations obtained from a finite sample s . The different approximations shown in section 2 lead to the following approaches.

Biased Index. The first possibility is to use the estimation $\Delta J_B(\Omega, s)$ in equation 2. Minimization of ΔJ_B over Π will lead to the validated partition.

Unbiased Index VI. A second approach is to consider the estimation $\Delta J_U(\Omega, s)$ in equation 3. As before, minimization of ΔJ_U over Π will lead to the

validated partition.

Unbiased Index V2. The direct minimization of the corrected index $\Delta J_U(\Omega, s)$ does not take into account the variance in the estimation due to the finite sample size. So it could happen that, for two given partitions Ω_1 and Ω_2 , the true values of the negentropy increment satisfy $\Delta J(\Omega_1) < \Delta J(\Omega_2)$, while their sample estimations satisfy $\Delta J_U(\Omega_1, s) > \Delta J_U(\Omega_2, s)$. To minimize this effect we follow here the approach in (Lago-Fernández et al., 2011) and consider the two partitions equivalent if:

$$\frac{\Delta J_U(\Omega_2, s) + \sigma_s(\Delta J_U(\Omega_2))}{\Delta J_U(\Omega_1, s) - \sigma_s(\Delta J_U(\Omega_1))} < 1 \quad (7)$$

In such cases we select the simplest (lower number of regions) partition. We will refer to this approach as ΔJ_{US} .

Unbiased Index V3. If we make the assumption that the real $\Delta J(\Omega)$ is normally distributed around $\Delta J_U(\Omega, s)$ with variance $\sigma_s^2(\Delta J_U)$, we can estimate the probability that $\Delta J(\Omega_1) < \Delta J(\Omega_2)$ by:

$$P(\Delta J(\Omega_1) < \Delta J(\Omega_2)) = \int_{-\infty}^{\infty} dx f_2(x) F_1(x) \quad (8)$$

where $f_i(x)$ and $F_i(x)$ are, respectively, the probability and cumulative density functions of a random Gaussian variable $X \sim N(\Delta J_U(\Omega_i, s), \sigma_s(\Delta J_U))$. Then we can consider the two partitions equivalent if P is lower than a given threshold α . In such a case we must proceed as before and select the simplest partition. We will consider $\alpha = 0.8$, and will refer to this approach as ΔJ_{UG} .

3.2 Reference Approaches

We will consider two additional criteria based on information theoretic approaches: the Akaike Information Criterion, AIC (Akaike, 1974), and the Bayesian Information Criterion, BIC (Schwartz, 1978; Fraley and Raftery, 1998). Both of them are intended to measure the relative goodness of fit of a statistical model by introducing a penalty term to the log-likelihood, and have been extensively used to determine the number of clusters in model-based clustering. It is known that, when fitting a statistical model to a data sample, it is possible to arbitrarily increase the log-likelihood by increasing the complexity of the model, but doing so may result in overfitting. AIC and BIC are defined as follows:

$$\begin{aligned} AIC &= 2p - 2\log(L) \\ BIC &= p\log(n) - 2\log(L) \end{aligned}$$

where p is the number of free parameters in the statistical model, n is the sample size and L is the log-likelihood for the model. Both methods reward the goodness of fit of the model, but also include a penalty term that is an increasing function of the number of free parameters. In both cases the preferred model is the one with the smallest AIC or BIC value.

4 DATA SETS

4.1 Gaussian Clusters in 2D

We use a set of two-dimensional clustering problems generated as in (Lago-Fernández and Corbacho, 2010). Each problem consists of c clusters, with each cluster containing 200 points randomly extracted from a bivariate normal distribution whose means and covariance matrices are also randomly selected. We consider c in the range $[1, 9]$, and generate 100 problems for each c .

4.2 Twonorm Problems

The *twonorm* problem is a synthetic problem initially designed for classification (Breiman, 1996). Given that the classes of the problem are known it also constitutes a good benchmark for testing clustering algorithms. It is a d -dimensional problem where each class is extracted from a d -variate normal distribution with identity covariance matrix and mean located at (a, a, \dots, a) for class/cluster 1 and at $(-a, -a, \dots, -a)$ for class/cluster -1, where $a = 2/\sqrt{20}$. The optimal separation plane is the hyperplane which passes through the origin and whose normal vector is (a, a, \dots, a) . The problem is designed such that the Bayes error is constant (≈ 0.023) and independent of the dimension d . Here we consider even dimensions in the range $[2, 20]$, and generate 1000 points for each cluster.

5 ANALYSIS

5.1 Clustering Algorithm

For a given problem, we use the Expectation-Maximization (EM) algorithm to fit a mixture of k Gaussian components to the data. Different number of components are tried, and a total of 20 different runs of the algorithm are performed for each k . After convergence of the algorithm, a crisp partition of the data is obtained by assigning each data point to a

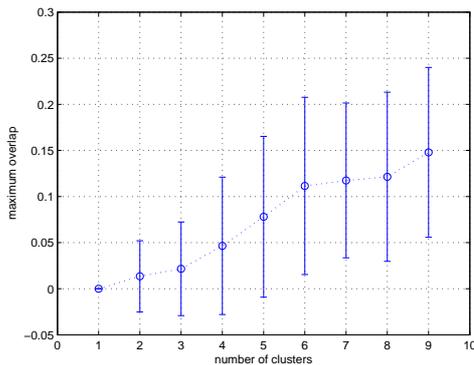


Figure 1: Maximum overlap between pairs of clusters versus number of clusters c for the Gaussian 2D problems. For a given c the average over 100 problems and its standard deviation are shown.

single cluster, represented by the mixture component that most likely explains it. For the 2D Gaussian problems we consider $k \in \{1, \dots, 13\}$. For the twonorm problems we consider $k \in \{1, \dots, 4\}$. This means that we end up with a set of 260 different clustering partitions (only 80 for the twonorm problems) that must be validated in a subsequent stage. This validation is performed using each of the 6 approaches described in section 3. Each approach leads to a single selected partition for each of the problems.

5.2 Evaluation of the Results

To measure the quality of a validated partition, and extensively the quality of a given validation approach, we compare the number of regions in the partition with the real number of clusters in the problem. We consider the number of problems for which a given validation approach provides a partition into the correct number of regions. Additionally, in some cases we also compute the average number of regions to have an idea of whether the validation index tends to under or over-estimate the number of clusters.

Finally, in order to measure the intrinsic difficulty of a given clustering problem, we consider the maximum overlap between any two clusters in the problem. We measure the overlap between two clusters as the Bayes error for a two-class classification problem where each cluster is one class. For the Gaussian 2D problems, this overlap increases with the number of clusters because the total amount of space is fixed (see figure 1). The twonorm problem, on the other hand, is designed such that the Bayes error is constant (≈ 0.023) and independent of the dimension, so all the problems have in this case the same intrinsic difficulty.

Table 1: Gaussian problems in 2D. Number of problems correctly validated by each of the six validation approaches considered. The first column, c , represents the actual number of clusters. The number of problems for a given c is 100.

c	AIC	BIC	ΔJ_B	ΔJ_U	ΔJ_{US}	ΔJ_{UG}
1	48	99	72	88	100	100
2	5	91	11	24	84	95
3	1	91	10	21	75	91
4	1	83	3	20	66	84
5	2	79	9	20	46	61
6	3	83	10	17	37	45
7	3	72	13	13	24	35
8	5	66	12	17	29	32
9	3	42	16	18	16	16

6 RESULTS

6.1 Gaussian Clusters in 2D

In table 1 we show the number of problems for which each validation technique provides the correct number of clusters. Each row shows the results for a given number of real clusters in the problem. There are 100 different problems for each number of clusters, therefore the maximum possible value in the table is 100. We see that the negentropy-based approaches ΔJ_{US} and ΔJ_{UG} outperform the classical BIC index only for problems with a small number of clusters ($c \leq 4$). Of these, the ΔJ_{UG} provides slightly better results. The BIC index is the best for high number of clusters. The ΔJ_B , ΔJ_U and AIC indices perform very poorly for all the problems.

In table 2 we show the average number of clusters for the solutions selected by each of the methods. Note that, in spite of finding a correct solution in more occasions, BIC presents a stronger tendency to over-estimate the number of clusters when it fails. In such a situation, the indices ΔJ_{US} and ΔJ_{UG} tend to underestimate the number of clusters. From a clustering perspective, this kind of error is in more accordance with intuition: it seems more plausible to merge two highly overlapping clusters than to split a single cluster into two components. It was shown in figure 1 that the maximum overlap increases with the number of clusters. This could explain the observed loss of performance of ΔJ_{US} and ΔJ_{UG} with increasing c . Finally, the ΔJ_B , ΔJ_U and AIC indices tend to overestimate the number of clusters even in the low overlap regime.

In figure 2 we show how the overlap is distributed, both for the correctly and the incorrectly validated

Table 2: Gaussian problems in 2D. Average number of clusters in the validated partitions for each of the six validation approaches considered. The column labeled c shows the actual number of clusters.

c	AIC	BIC	ΔJ_B	ΔJ_U	ΔJ_{US}	ΔJ_{UG}
1	2.5 ± 1.7	1.0 ± 0.2	1.8 ± 1.4	1.4 ± 1.1	1.0 ± 0.0	1.0 ± 0.0
2	4.7 ± 1.2	2.1 ± 0.6	4.0 ± 1.7	3.3 ± 1.6	1.8 ± 0.4	1.9 ± 0.2
3	6.0 ± 1.0	3.1 ± 0.4	4.9 ± 1.6	4.3 ± 1.5	2.7 ± 0.5	2.9 ± 0.3
4	6.8 ± 1.1	4.0 ± 0.5	6.2 ± 1.8	5.3 ± 1.7	3.6 ± 0.7	3.8 ± 0.4
5	7.6 ± 1.1	4.9 ± 0.4	6.5 ± 2.0	5.8 ± 1.8	4.3 ± 0.9	4.5 ± 0.6
6	8.8 ± 1.2	5.8 ± 0.4	7.6 ± 2.0	7.2 ± 1.7	5.0 ± 0.9	5.3 ± 0.8
7	9.7 ± 1.1	6.8 ± 0.6	8.6 ± 2.1	7.8 ± 1.8	5.8 ± 1.0	6.1 ± 0.9
8	10.8 ± 1.2	7.8 ± 0.8	9.9 ± 2.0	9.1 ± 1.9	6.9 ± 1.3	6.7 ± 1.2
9	11.7 ± 1.2	8.8 ± 1.0	10.2 ± 2.2	9.6 ± 2.0	7.6 ± 1.5	7.4 ± 1.1

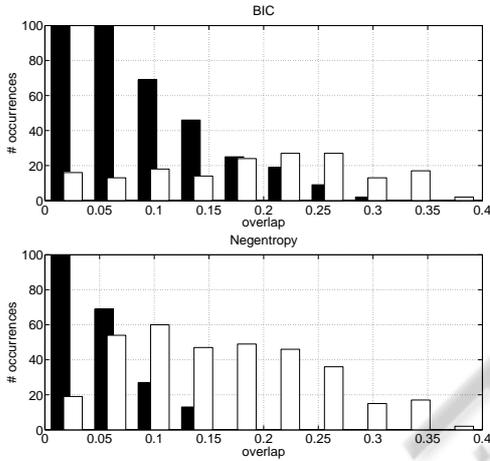


Figure 2: Distribution of overlaps for correctly (black filled) and incorrectly (white filled) validated problems by each of the two methods BIC (top) and ΔJ_{UG} (bottom). The highest bars have been truncated for the sake of clarity.

problems by each of the two methods BIC (top) and ΔJ_{UG} (bottom). The distributions for ΔJ_{US} are similar to those for ΔJ_{UG} (not shown). Observe that ΔJ_{UG} is able to assess the correct number of clusters only for small overlap. The number of failures is also reduced in this small overlap region. This is in clear contradiction with the observation for BIC, which is able to find the correct partition even in high overlap regimes.

If we recompute the values shown in table 1 taking into account only the problems whose maximum overlap is below a given threshold t , we obtain the results shown in table 3. The value of the threshold has been fixed to $t = 0.03$. The column labeled NP shows the number of problems that satisfy this restriction for a given c . Note that now there is almost no difference between the results provided by BIC and ΔJ_{UG} .

Table 3: Gaussian problems in 2D. Number of problems correctly validated by each of the six validation approaches considered. Only problems with overlap lower than $t = 0.03$ are considered. The column labeled NP shows the number of problems that satisfy this constraint for a given c .

c	NP	AIC	BIC	ΔJ_{US}	ΔJ_{UG}
1	100	48	99	100	100
2	86	2	78	75	86
3	79	1	72	65	79
4	63	0	54	50	62
5	42	0	40	32	40
6	28	1	27	19	24
7	12	0	10	7	10
8	16	1	14	12	12
9	7	0	4	4	3

Table 4: Twonorm problems. Number of problems correctly validated by each of the six validation approaches considered. The first column, d , indicates the dimensionality of the problem. The number of problems for a given d is 100.

d	AIC	BIC	ΔJ_B	ΔJ_U	ΔJ_{US}	ΔJ_{UG}
2	31	100	80	89	100	100
4	5	100	38	55	100	100
6	1	100	2	10	100	100
8	0	100	0	2	100	100
10	0	9	0	0	100	100
12	0	0	0	0	100	100
14	0	0	0	0	100	100
16	0	0	0	0	100	97
18	0	0	0	0	93	81
20	7	0	0	0	62	76

6.2 Twonorm Problems

The twonorm problems considered here present an overlap of approximately 0.023. This falls below the threshold $t = 0.03$ used previously to filter high

Table 5: Twonorm problems. Average number of clusters in the validated partitions for each of the six validation approaches considered. The first column shows the dimensionality of the problem.

d	AIC	BIC	ΔJ_B	ΔJ_U	ΔJ_{US}	ΔJ_{UG}
2	3.0±0.8	2.0±0.0	2.3±0.8	2.1±0.4	2.0±0.0	2.0±0.0
4	3.6±0.6	2.0±0.0	2.8±0.7	2.6±0.7	2.0±0.0	2.0±0.0
6	3.8±0.4	2.0±0.0	3.6±0.5	3.4±0.7	2.0±0.0	2.0±0.0
8	3.8±0.4	2.0±0.0	3.9±0.3	3.6±0.5	2.0±0.0	2.0±0.0
10	3.7±0.4	1.1±0.3	4.0±0.1	3.8±0.4	2.0±0.0	2.0±0.0
12	3.8±0.4	1.0±0.0	4.0±0.0	4.0±0.2	2.0±0.0	2.0±0.0
14	3.8±0.4	1.0±0.0	4.0±0.0	4.0±0.1	2.0±0.0	2.0±0.0
16	3.6±0.5	1.0±0.0	4.0±0.0	4.0±0.0	2.0±0.0	2.0±0.2
18	3.4±0.5	1.0±0.0	4.0±0.0	4.0±0.0	2.1±0.3	2.2±0.4
20	3.2±0.5	1.0±0.0	4.0±0.0	4.0±0.0	2.4±0.5	2.2±0.4

overlap problems, which means that the two clusters are quite well separated. The difficulty in this case arises from the high dimensionality. The results for these problems are shown in tables 4 and 5. The first column in both tables is the dimension. Table 4 shows the number of problems for which each of the validation methods provides a solution with 2 clusters. Table 5 shows the average number of clusters in the assessed solutions. All the validation approaches show a loss of performance when the dimension of the problems increases, but the ΔJ_{US} and ΔJ_{UG} indices are more robust than the others. BIC starts to fail at $d = 10$, experimenting a sudden loss of accuracy. For higher dimensions it tends to select one single cluster. On the other hand, ΔJ_{US} and ΔJ_{UG} are very accurate even for $d = 16$, and their loss of accuracy for higher dimensions is more gradual. Finally, the ΔJ_B , ΔJ_U and AIC approaches provide very poor results, and show a strong tendency to overestimate the number of clusters.

7 CONCLUSIONS

The aim of this paper was to systematically study the performance of negentropy-based cluster validation in synthetic problems with increasing dimensionality. Negentropy-based indices are quite simple to compute, as they only need to estimate the probabilities and the log-determinants of the covariance matrices for each cluster. However, the computation of the log-determinants in regions with small number of points introduces a strong bias that must be corrected in order to properly estimate the negentropy index. A heuristic based on a formal analysis of the bias can be obtained to alleviate this effect.

In this paper we refined the correction of the negentropy index proposed in (Lago-Fernández et al., 2011) in order to quantify the confidence levels of

the index value, thus obtaining a new, more formal heuristic for the validation of clustering partitions. Then we studied the performance of this and other negentropy-based validation approaches in problems with increasing dimensionality, and compared the results with two well established techniques such as BIC and AIC. The performance of BIC in problems where the ratio of the number of points to the dimension is high, is quite good. For problems where there are clusters with a high overlap, it clearly outperforms the negentropy-based indices. This was expected since BIC is optimal for Gaussian clusters, which is the case for the synthetic data considered here. The AIC criterion seems to produce very bad results for the set of problems considered, providing a strong overestimation of the number of clusters in all the cases. Negentropy-based indices are designed for crisp clustering, and they seek to detect compact and well separated clusters. When we consider only problems where the clusters are not highly overlapped, the performances of BIC and the negentropy-based index are quite similar.

In order to test the behavior of the indices as a function of the dimensionality, we constructed a clustering benchmark database based on the twonorm classification problem (Breiman, 1996). This database is generated using two Gaussian clusters of increasing dimensionality but constant degree of overlap. The number of points in each cluster is constant independently of the dimension. Therefore, the effect of the dimensionality on the performance of the indices is isolated. As the dimensionality increases, the performance of BIC degrades quickly, but the performance of the negentropy-based index is quite stable, finding the correct solution for all the problems up to $d = 16$, and experimenting a gradual degradation for higher dimension.

In conclusion we showed, using the synthetic database *twonorm*, that our approach to negentropy-

based validation can outperform AIC and BIC in problems where the ratio of the number of points to the dimension is not high, which is a very common situation in most real applications. New experiments with other databases are required in order to check if this property is general. We expect that this finding will be more accentuated in benchmarks with non Gaussian clusters (Biernacki et al., 2000; Lago-Fernández et al., 2009). This will be the subject of future work.

Xu, R. and II, D. W. (2005). Survey of clustering algorithms. *IEEE Trans. Neural Networks*, 16(3):645–678.

ACKNOWLEDGEMENTS

The authors thank the financial support from DGUI-CAM/UAM (Project CCG10-UAM/TIC-5864).

REFERENCES

- Abramowitz, M. and Stegun, A. (1965). *Handbook of mathematical functions*. Dover, New York.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control*, 19:716–723.
- Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Analysis Machine Intelligence*, 22(7):719–725.
- Breiman, L. (1996). Bias, variance, and arcing classifiers. Technical Report 460, Statistics Department, University of California, Berkeley, CA, USA.
- Fraley, C. and Raftery, A. (1998). How many clusters? which clustering method? answers via model-based cluster analysis. Technical Report 329, Department of Statistics, University of Washington, Seattle, WA, USA.
- Gordon, A. (1998). Cluster validation. In Hayashi, C., Ohsumi, N., Yajima, K., Tanaka, Y., Bock, H., and Baba, Y., editors, *Data science, classification and related methods*, pages 22–39. Springer.
- Lago-Fernández, L. F. and Corbacho, F. (2010). Normality-based validation for crisp clustering. *Pattern Recognition*, 43(3):782–795.
- Lago-Fernández, L. F., Sánchez-Montanés, M. A., and Corbacho, F. (2009). Fuzzy cluster validation using the partition negentropy criterion. *Lecture Notes in Computer Science*, 5769:235–244.
- Lago-Fernández, L. F., Sánchez-Montanés, M. A., and Corbacho, F. (2011). The effect of low number of points in clustering validation via the negentropy increment. *Neurocomputing*, 74(16):2657–2664.
- Misra, N., Singh, H., and Demchuk, E. (2005). Estimation of the entropy of a multivariate normal distribution. *Journal of Multivariate Analysis*, 92(2):324–342.
- Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.