Evaluating information criteria in latent class analysis: application to identify classes of breast cancer dataset

Abdallah Abarda*

Laboratoire de Modélisation Mathématiques
et de Calculs Economiques, FSJES,
Université Hassan 1er,
Settat, Morocco
Email: abardabdallah@gmail.com
*Corresponding author

Mohamed Dakkon

Département de Statistique et Informatique de Gestion, FSJES,
Université Abdelmalek Essaadi,
Tetouan, Morocco
Email: m.dakkoun@gmail.com

Khawla Asmi

LRIT, Associated Unit to CNRST (URAC No 29),
Rabat IT Center – Faculty of Sciences,
Mohammed V University in Rabat, Morocco
Email: asmi.smi.khawla@gmail.com

Youssef Bentaleb

EECOMAS-Lab,
Ibn Tofail University,
Kenitra, Morocco
Email: ybentaleb@ymail.com

Abstract: In recent studies, latent class analysis (LCA) modelling has been proposed as a convenient alternative to standard classification methods. It has become a popular tool for clustering respondents into homogeneous subgroups based on their responses on a set of categorical variables. The absence of a common accepted statistical indicator for deciding the number of classes in the study of population represents one of the major unresolved issues in the application of the LCA. Determining the number of classes constituting the profiles of a given population is often done by using the likelihood ratio test, however the use of such methodology is not correct theoretically. To overcome this problem, we propose an alternative for the classical latent class models selection methods based on the information criteria. This article aims to investigate the performance of information criteria for selecting the latent class analysis models. Nine information criteria are compared under various sample sizes and model dimensionality. We propose also an application of ICs to select the best model of breast cancer dataset.

Keywords: latent class analysis; model selection; information criteria; classification methods.

Reference to this paper should be made as follows: Abarda, A., Dakkon, M., Asmi, K. and Bentaleb, Y. (2021) 'Evaluating information criteria in latent class analysis: application to identify classes of breast cancer dataset', *Int. J. Data Analysis Techniques and Strategies*, Vol. 13, Nos. 1/2, pp.72–87.

Biographical notes: Abdallah Abarda is a PhD in the Department of Mathematics at the Faculty of Sciences, Kenitra, Ibn Tofail University, Morocco. He received his Engineering degree in Statistics at the National Institute of Statistics and Applied Economic of Rabat in 2013. Currently, he is a Research Professor at the Hassan 1st University, Settat, Morocco. He has a number of publications in the fields of statistics, data analysis, classification and big data.

Mohamed Dakkon is a PhD in the Department of Mathematics at the Faculty of Sciences, Kenitra, Ibn Tofail University, Morocco. Currently, he is a Professor at the Abdelmalek Essaadi University, Settat, Morocco. He has a number of publications in the fields of statistics and probability.

Khawla Asmi received his Master's degree in Computer Science and Telecommunications from the Faculty of Science, Mohammed V University in Rabat, Morocco in 2015. Currently, she is pursuing his PhD at Mohammed V-Agdal University, LRIT, Associated Unit to CNRST (URAC 29). Her research interests include social network analysis, community detection, statistics, classification and big data.

Youssef Bentaleb is a Research Professor. He holds a PhD in Mathematics and Computer Science. He is the President of the Moroccan Center for Polytechnic Research and Innovation (CMRPI). He is also the Director of the *International Journal of Scientific Research and Innovation IJRSI-CMRPI*, the Editor-in-Chief of JCCE journal, a member of the EECOMAS Research Laboratory and a guest speaker.

This paper is a revised and expanded version of a paper entitled 'Solving the problem of latent class selection' presented at LOPAL'2018, Rabat, Morocco, 2–5 May 2018.

1 Introduction

The selection of the best model of latent class analysis represents a critical problem because it can affect substantive interpretations of the studied phenomenon (Yang, 2006). Indeed, an incorrect selection of the latent classes can lead to an incorrect analysis. Therefore, the choice of models that fit the data represents very interesting topics in the classification research. In our tests of several models of latent class analysis, two problems arise. The first one concerns the choice of the number of classes, where the second one concerns the form of the model that provides the number of classes (Youness, 2004). The definition of the number of classes in a given population is commonly achieved by using a likelihood ratio test, which is often used to compare two models (nested models deriving from each other by adding or deleting terms) under the assumption that these two models correctly fit the data (McCullagh and Nelder, 1989). In the comparison of many models, the risk of rejecting the null hypothesis when it is true increases substantially (Lancelot and Lesnoff, 2005). Clogg (1995), Aitkin and Rubin (1985) and Aitkin et al. (1981) demonstrated that the chi-square test of the likelihood ratio is not theoretically correct to select LC models. The problem relies on the fact that under the null hypothesis of an LC model with $C - 1$ latent classes against the alternative hypothesis of C latent classes, the regularity conditions of the likelihood ratio test are often not satisfied asymptotically. In addition, Everitt (1981, 1988) and Yang (1998) performed simulation studies to prove the inadequacy of the likelihood ratio test. To overcome this problem, several alternative methods have been proposed such as information criteria, parametric re-sampling, etc., note that the information criteria represents the most practical one with less computational effort (Dziak et al., 2012).

Recent works are focused on model selection by using information criteria (Abarda et al., 2018; Zhang et al., 2018); Seo and Thorne (2018) used information criteria for phylogenetic partitioning decisions. For the purpose of selecting an optimal partitioning scheme, they used the AIC with an idea of splitting errors. As well as they introduced a similar adjustment to the Bayesian information criteria (BIC) via simulation and empirical data analysis. Mangan et al. (2017) employ sparse regression and information criteria for dynamical systems selection. Where they introduce that AIC criteria scores place each candidate model in the strong support, weak support or no support category and that the method correctly recovers several canonical dynamical systems. Kitagawa (2018) investigates information criteria for statistical modelling in data-rich in which the criteria such as Akaike information criteria (AIC), generalised information criterion (GIC), bootstrap information criterion (EIC), and so on are employed for evaluating prediction accuracy by statistical models. Pels et al. (2018) focus on the caution in the information criteria to select the working correlation structure in generalised estimating equations.

In this paper, we use the information criteria to improve the selection of the best model of latent class analysis. We introduce also new information's criteria for the LCA method. The rest of the paper is organised as follows. In Sections 2 and 3, we adapt the new information criteria to the LCA model and we compare the different performance of this CIs with a numerical simulation. Section 4 describes an application of those ICs to select the best model of breast cancer dataset. Section 5 concludes this paper and outlines future research directions.

2 Comparison of information criteria for the selection of latent class models

2.1 *Criteria for selection of LCA models*

The information criteria for the selection of models were originally introduced by Akaike (1992), who used the Kullback-Leibler (KL) measure to discriminate among competing models. Schwarz (1978) proposed another important class of information criteria based on the Bayesian statistics. Since then, many modified information criteria have been proposed and developed.

Various ICs for model selection can be discussed in a unified way (Dziak et al., 2012) such as the Akaike information criteria (AIC) (Akaike, 1992), the Bayesian information criteria (BIC) (Schwarz, 1978), The coherent AIC (CAIC) of Bozdogan (1987), the adjusted BIC (ABIC) and others (Sclove, 1987). The ICs include an adjustment term (a log-likelihood function) and a control penalty.

These ICs involve the choice of a model with the least penalised log-likelihood function, that is, the highest value of the term $l - A_N p$ where $l = \log(L(\theta_j))$ is the log-likelihood function, A_N is a constant or a function that may depend on the sample size N, and p is the number of parameters in the model. In the literature, instead of maximising this term precisely, we minimise the function $l - A_N p$. Thus, the general form of information criteria is written as follows:

$$IC = -2l + A_N p \tag{1}$$

The term (1) is called the generalised information criteria (Atkinson, 1980). In this expression, the log-likelihood term is occasionally replaced in practice by G^2, which is called deviance and is equal to $-2l$ plus a function of the saturated model (it is a variation of the log-likelihood), where $A_N p$ is the penalty term.

The term (1) cannot be used directly in practice without the choice of A_N. Note that a specific choice of A_N makes the difference between AIC, BIC, adjusted BIC, CAIC and other criteria. Indeed, this choice is motivated by the theory and objectives of the criteria as well as the different contexts of use and relative degrees of emphasis of parsimony (Claeskens and Hjort, 2008; Lin and Dayton, 1997). Therefore, what is the best use of each of the information criteria in practice? This information is especially important, since the use of ICs can lead to different results for the same data, and this influences the choice of the right number of classes.

2.2 Comparison of information criteria

2.2.1 Akaike information criteria

The most known among information criteria is the AIC. AIC represents a compromise between bias, decreasing with the number of free parameters, and parsimony, minimising number of parameters in the model. The AIC (Akaike, 1992) defines $A_N = 2$ in (1). It estimates the KL relative distance (a non-parametric distance measure) of the likelihood function specified by a suitable candidate model from the unknown likelihood that produced the data. More precisely, KL can be written as $KL = E_t(l_t(y)) - E_t(l(y))$, where E_t is the expectation of the true unknown distribution, l_t is the log-likelihood of the data (such that the true distribution is unknown), and $E_t(l_t(y))$ will be the same for all models considered. Thus, the minimisation of KL is equivalent to the maximisation of $E_t(l(y))$. Akaike (1992) showed that the term

$$l - tr(\widehat{J}^{-1}\widehat{K}) + c$$

is an approximate estimate of $E_t(l(y))$, where J and K are $(p \times p)$ matrices, $tr()$ is the trace of the matrix, and c is a constant. $\widehat{J}$ is an estimator of the parameter covariance matrix (based on the matrix of the second derivatives of l), and $\widehat{K}$ is an estimator based on the crossed products of the first derivatives (Claeskens and Hjort, 2008). Akaike (1992) showed that $\widehat{J}$ and $\widehat{K}$ are asymptotically equal for the true model. The trace therefore becomes approximately equal to p, where p is the number of parameters in the model. Thus, for models that are not well-adjusted to the true model, the approximation cannot be as good. However, it is likely that they have values less than 1, so the precise size of the penalty is unimportant (Burnham and Anderson, 2002). In the term $l - p$, it is suggested to use $A_N = 2$ in (1) for the AIC, assuming that the models adjusted with low values of (1) will likely provide a likelihood function closer to the truth.

The AIC for LCA represents a good compromise between the parsimony (the need to describe the data with as few parameters as possible) and the bias (which decreases with the number of parameters). In the AIC, the deviance of the model $-2\ln(l)$ is penalised by 2 times the number of parameters. According to Tibshirani and Knight (1999) and Hastie et al. (2001), the approximation $tr(\widehat{J}^{-1}\widehat{K}) \approx P$ is too 'optimistic', and the resulting penalty for the complexity of the model is too low, which represents a disadvantage of the AIC. This has led other researchers to think about the penalty issue for CIs by introducing other forms of the term A_N. AIC is defined as follows:

$$AIC = -2\ln(L) + 2p \tag{2}$$

2.2.2 Bayesian information criteria

The Bayesian information criteria represent a widely used information criteria. In the model selection with the Bayesian criteria, a priori probability is established for each model M_i, and a priori distributions is established for the non-zero coefficients in each model. If we suppose that a single model with its associated priors is true, we can consider the Bayes' theorem.

Let $Pr(M_i)$ be the a priori probability established by the researcher and $Pr(y/M_i)$ be the probability density of the data under model M_i, calculated as the value of the expectation of the likelihood function of y taking into account the model and the parameters under the a priori distribution of the parameters. According to the Bayes' theorem, the posterior probability $Pr(M_i/y)$ of a model is proportional to $Pr(M_i)Pr(y/M_i)$.

The degree to which the data promote model M_i compared to another model M_j is given by the ratio of the posterior ratings to the previous ratings: $Pr(M_i/y)/Pr(M_j/y)/Pr(M_i)/Pr(M_j)$.

If we assume that the priori probabilities are equal for each model, this is simplified in the Bayes factor:

$$B_{ij} = Pr(M_i/y)/Pr(M_j/y) = Pr(y/M_i)/Pr(y/M_j) \tag{3}$$

Thus, the model that has the greatest Bayes factor also has the greatest posterior probability. Schwarz (1978) and Kass and Raftery (1995) have shown that in many types of models, B_{ij} can be approximated by

$$\exp\left(-\frac{1}{2}BIC_i + \frac{1}{2}BIC_j\right)$$

with

$$BIC = -2\ln(L) + \ln(N)p \tag{4}$$

The model with the highest posterior probability is probably the one with the lowest BIC. The BIC is sometimes preferable to the AIC because the BIC is consistent. Assuming a fixed number of models is available and one of them is the real model, a consistent CI is one that will select the real model with a probability close to 100% as $N \longrightarrow \infty$. In this context, the true model is the smallest adequate model (i.e., the only model that minimises the distance KL, or the smallest among the others (Claeskens and Hjort, 2008)). The AIC is considered not consistent as it fails to identify the right model when the value of n becomes high.

2.2.3 Corrected Akaike information criteria

Many researchers suggested corrected Akaike information criteria (AICc). This measure represents an improved version of AIC (Sugiura, 1978; Hurvich and Tsai, 1989; Burnham and Anderson, 2004). AICc is used in the context of time series and regression. AICc criteria applies a heavy penalty that depends on p and N and gives results very close to those of the AIC in the case where N is large relative to p. For small N, Hurvich and Tsai (1989) have shown that AICc is better than AIC when $n/k < 40$. For LCA models, AICc is defined as follows:

$$AICc = -2\ln(L) + 2p + \frac{2p(p+1)}{N-p-1} \tag{5}$$

2.2.4 AIC3 criteria

Some researchers have suggested using $A_N = 3$ (Andrews and Currim, 2003) in (1) instead of $A_N = 2$. This criterion is sometimes called AIC3. There is little theoretical basis for AIC3, despite its relatively good simulation performance. AIC3 is therefore defined by:

$$AIC3 = -2\ln(L) + 3p \tag{6}$$

2.2.5 Adjusted Bayesian information criteria

Sclove (1987) proposed an adjusted BIC (ABIC or BIC* or BICA) based on the work of Rissanen (1978) and Boekee and Buss (1981). He takes $A_N = \ln(\frac{N+2}{24})$ in (1) instead of $A_N = ln(N)$. Thus, the ABIC is defined as follows:

$$ABIC = -2\ln(L) + p\ln\left(\frac{N+2}{24}\right) \tag{7}$$

This penalty is much lighter than the penalty used in BIC and it can be slightly lighter than penalty used in AIC criteria.

2.2.6 Coherent Akaike information criteria

The coherent Akaike information criteria (CAIC) is similar to the BIC. It was proposed by Bozdogan (1987) to make the AIC consistent (a corrected version of the AIC). The author uses $A_n = ln(n) + 1$ (not to be confused with the AICc that will be studied later). The criteria is thus written as follows:

$$CAIC = -2\ln(L) + p(\ln(N) + 1) \tag{8}$$

This penalty seems more appropriate for improving the parsimony of the model and providing more adjustment compared to the BIC with the penalty $A_n = \ln(n)$ (Dziak et al., 2012). Thus, the term A_n for the CAIC was chosen arbitrarily so that it was consistent. On the other hand, any criteria proportional to $ln(n)$ provides consistency in model selection. Thus, the CAIC has no clear advantage comparing to the BIC, which is very 'popular'.

Other information criteria have been proposed, such as Hurvich and Tsai criteria (HT), CAIU, Hannan and Quinn criteria (HQ), approximate weight of evidence criteria (AWE) and Draper BIC criteria (DBIC). For the choice of the number of classes, we rely on a comparison of these criteria. The criteria that will be used for the determination of the number of classes vary simultaneously and is consistent. To eliminate plausible models, the best model is the one that minimises these information criteria. It should be emphasised that the choice of the final model depends on some considerations such as the results of previous research, the parsimony of the model and the consistency with the theory. Thus, we define criteria HT, CAIU, HQ, AWE and DBIC as follows:

2.2.7 Hurvich and Tsai criteria

Hurvich and Tsai (1989) have proposed another information criteria that is a variant of the AIC and is used at the beginning in the regression and the time series. The first two terms for the AIC and HT are the same, except that in the latter, a third term is added to reduce the bay in the case of small samples. The HT criteria is given as follows:

$$HT = -2\ln(L) + 2p + \frac{2(p+1)(p+2)}{N-p-2} \tag{9}$$

2.2.8 CAIU criteria

The CAIU or AICU is equal to the sum of the CAIC and a term that depends on N and the number of parameters p (McQuarrie et al., 1997). It is given by the following formula:

$$CAIU = AICc + N\ln\left(\frac{N}{N-p-1}\right) \tag{10}$$

2.2.9 Hannan and Quinn criteria

Hannan and Quinn (1979) suggested that the term 'penalty' should be $(\ln(\ln N))$. This gives birth to another criteria:

$$HQ = -2\ln(L) + 2p\ln(\ln(N)) \tag{11}$$

2.2.10 Approximate weight of evidence criteria

Banfield and Raftery (1993) suggested a Bayesian solution to the choice of the number of classes based on an approximation of the probability of classification and an approximate weight of evidence (AWE). This penalises highly complex models such as the BIC, and it will select more parsimonious models than the BIC criteria, with the exception of well-separated clusters; it selects a model with clusters that minimises the expression (1) where:

$$AWE = -2\ln(L) + p(\ln(N) + 1.5) \qquad (12)$$

2.2.11 Draper criteria

Draper (1995) suggested that the BIC's penalty could be $A_N = ln(N) - ln(2\pi)$, which gives birth to a new criteria, DBIC:

$$DBIC = -2\ln(L) + p(\ln(N) - \ln(2\pi)) \qquad (13)$$

Thus, the addition of the term $-p\ln(2\pi)$ is asymptotically unimportant when $N \longrightarrow \infty$. On the other hand, the DBIC has an effect on the log-likelihood of small and medium sample sizes. Draper (1995) concluded that this criterion is important, especially for the selection of classes in real problems.

Table 1 Comparison of information criteria

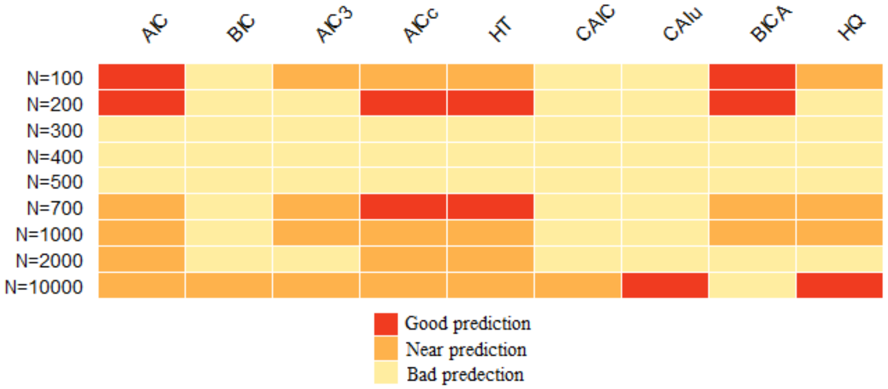
Criteria	*Expression*
AIC	$\text{AIC} = -2\ln(L) + 2p$
AICc	$\text{AICc} = -2\ln(L) + 2p + \frac{2p(p+1)}{N-p-1}$
AIC3	$\text{AIC3} = -2\ln(L) + 3p$
BIC	$\text{BIC} = -2\ln(L) + \ln(N)p$
ABIC	$\text{ABIC} = -2\ln(L) + p\ln(\frac{N+2}{24})$
CAIC	$\text{CAIC} = -2\ln(L) + p(\ln(N) + 1)$
HT	$\text{HT} = -2\ln(L) + 2p + \frac{2(p+1)(p+2)}{N-p-2}$
CAIU	$\text{CAIU} = \text{AICc} + N\ln(\frac{N}{N-p-1})$
HQ	$\text{HQ} = -2\ln(L) + 2p\ln(\ln(N))$
AWE	$\text{AWE} = -2\ln(L) + p(\ln(N) + 1.5)$

Table 1 summarises and compares the general expressions of the information criteria.

3 Performance of information criteria in selecting latent class analysis models

First, we simulate a LCA model with four dichotomous variables and four latent classes for different sample sizes: N = 100, N = 200, N = 300, N = 400, N = 500, N = 700, N = 1,000, N = 2,000, N = 10,000. In order to test the relevance of the different criteria to predict the exact number of classes for each N, we construct models with 2, 3, 4, 5 and six classes, then, we calculate the following criteria: AIC, BIC, AIC3, AICc, HT, CAIC, CAIU and HQ. In each case, to determine the number of classes, we choose the class number corresponding to the minimum value of the criteria.

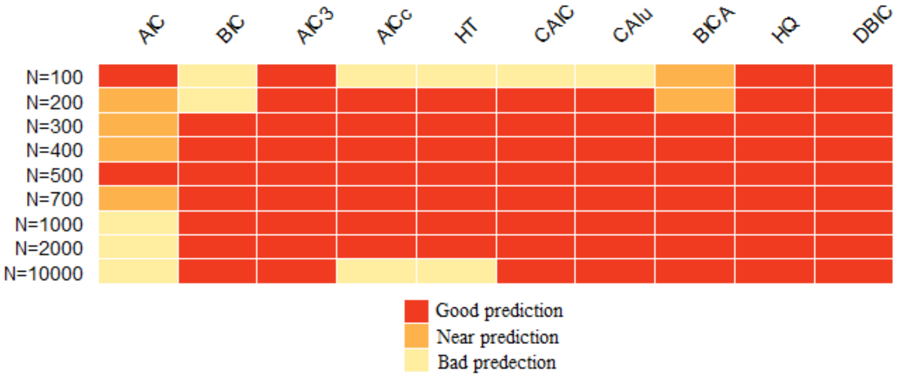
Figure 1 Prediction of the number of classes by CIs for a model with four variables (see online version for colours)



Thus, we can see in Figure 1 that for four variables:

- The AIC and ABIC predicted the number of classes for the N = 100 and N = 200 samples. This means that these two criteria predict models well when the samples are small.
- The BIC, AIC3, AICc, HT, CAIU and HQ tend to predict fairly the number of classes for large samples.

Figure 2 Prediction of the number of classes by CIs for models with 12 variables (see online version for colours)



From Figure 2, we conclude that for 12 variables:

- The AIC tends to predict models for samples N = 100, N = 200, N = 300, N = 400, N = 500 and N = 700. On the other hand, this criteria could not estimate the exact number of classes for samples greater than N = 1,000.

- The BIC is not the best choice when estimating models for small samples, however, it provides more reasonable number of classes when the samples are over than 300.
- The AIC3, HQ, and DBIC are more efficient than the other criteria in estimating the exact number of classes. Moreover, they perfectly predict the number of classes, whatever the size of the sample.

By comparing the obtained results from Figure 2, we note that all information criteria are very sensitive to the number of employed variables. Indeed, the general prediction of CIs with four-variable model is worse than the prediction for a 12-variable model.

To measure the effect of sample size, we calculate a prediction score for samples below and above 500. Thus, we construct a scale to measure the prediction quality of the models:

- a good prediction = 1
- a near prediction = 0.5
- a bad prediction = 0.

The values obtained in Tables 2 and 3 are the percentages of scores for the different sample sizes. This represents a sort of probability of selection and then a consistent criteria is the one that has a high probability of selecting the right model.

Table 2 Exact classification rate by sample size (V = 4)

V = 4	*AIC*	*BIC*	*AIC3*	*AICc*	*HT*	*CAIC*	*CAIU*	*ABIC*	*HQ*
N < 500	50%	0%	25%	38%	38%	0%	0%	67%	20%
N > 500	50%	100%	75%	63%	63%	100%	100%	33%	80%

Table 3 Exact classification rate by sample size (V = 12)

V = 12	*AIC*	*BIC*	*AIC3*	*AICc*	*HT*	*CAIC*	*CAIU*	*ABIC*	*HQ*	*DBIC*
N < 500	83%	33%	50%	50%	50%	43%	43%	43%	50%	50%
N > 500	17%	67%	50%	50%	50%	57%	57%	57%	50%	50%

4 Application: identifying classes of breast cancer dataset

We apply the latent class analysis to identify hidden classes from breast cancer dataset. The first goal of this application is to measure the LCA's ability to detect the classes forming the data. Another important aim of this application is to assess the performance of the proposed information criteria in choosing the right number of classes. The proposed data is mainly composed of two classes. We use the PoLCA package of the R software, and then we test LCA models with two to six classes, we calculate for each model the criteria (AIC, CAIC, CAIU, BIC, AIC3, AICc, HT, ABIC and HQ) and we evaluate among the proposed criteria which gives the right model (the correct number of classes). The classic employed criteria are AIC and BIC. In this experiment, we add seven other criteria.

4.1 Data 1: Breast cancer dataset

For ours first experiment, we used breast cancer dataset taken from the UCI machine learning repository in our tests (this breast cancer domain was obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. Data provided by Zwitter and Soklic). This dataset is commonly used among researchers in the machine learning fields and it is devoted for the breast cancer classification. Which enables us to compare the performance of the proposed information criteria with that of other. The dataset contains 286 samples (including 201 instances of one class and 85 instances of another class). It consists of nine categorical attributes (+ the class attribute), we represent each of category as an integer between 1 and the max of category. The attributes are as follows:

1 *class:* no-recurrence-events, recurrence-events

2 *age:* 10–19, 20–29, 30–39, 40–49, 50–59, 60–69, 70–79, 80–89, 90–99

3 *menopause:* lt40, ge40, premeno

4 *tumor.size:* 0–4, 5–9, 10–14, 15–19, 20–24, 25–29, 30–34, 35–39, 40–44, 45–49, 50–54, 55–59

5 *inv.nodes:* 0–2, 3–5, 6–8, 9–11, 12–14, 15–17, 18–20, 21–23, 24–26, 27–29, 30–32, 33–35, 36–39

6 *node.caps:* yes, no

7 *deg.malig:* 1, 2, 3

8 *breast:* left, right

9 *breast.quad:* left-up, left-low, right-up, right-low, central

10 *irradiat:* yes, no.

Table 4 Information criteria for choosing the number of classes (data 1)

Classes	*Gsq*	*Llik*	*AIC*	*AIC3*	*AICc*	*HT*	*CAIC*	*CAIU*	*BIC*	*ABIC*	*HQ*
2	2,019	–2,599	5,348	5,423	5,401	5,404	5,697	5,490	5,622	5,384	5,458
3	1,833	–2,506	5,238	5,351	5,386	5,391	5,764	5,532	5,651	5,292	5,403
4	1,746	–2,463	5,227	5,378	5,567	5,577	5,930	5,784	5,779	5,300	5,448
5	1,687	–2,433	5,244	5,433	5,988	6,008	6,124	6,300	5,935	5,335	5,521
6	1,653	–2,416	5,286	5,513	7,063	7,118	6,343	7,519	6,116	5,396	5,619

Table 5 Results generated for the breast cancer dataset (data 1)

Information criteria	*Decision*
AIC	4 classes
CAIC, CAIU, BIC	2 classes
AIC3, AICc, HT, aBIC, HQ	3 classes

The results show that the AIC criteria gives four classes, which is not suitable for the data structure (two classes). The AIC3, AICc, HT, ABIC and HQ criteria are close in choosing the right model. The CAIC, CAIU, BIC adjust the data well and represent the suitable choice for two class model, which is adapted to the data structure. Given the two-class model, The estimated class proportion are 0.71 for the first class and 0.29 for the second.

We can conclude that:

- The AIC as well as the statistical indicator linked to the PoLCA packages does not give the exact number of classes.
- Among the proposed criteria, the criteria CAIC, CAIU and BIC identify the best model, where the criteria AIC3, AICC, HT, HQ and ABIC are close to good choice. However, the AIC criterion is far from identifying the exact number of classes.
- It is not enough to be satisfied with a single criterion, but it is necessary to take into account several criteria at the same time to decide the number of classes.

4.2 Data 2: Wisconsin Breast Cancer Database

For the second experiment, we used Wisconsin Breast Cancer Database provided by the UCI machine learning repository (Mangasarian and Wolberg, 1990; Wolberg and Mangasarian, 1990). This dataset is commonly also used by researchers in the machine learning fields and it is devoted for the breast cancer classification. The dataset contains 699 samples (including 458 instances of one class and 241 instances of another class). It consists of nine categorical attributes (+ the class attribute), the attributes are as follows:

1. *class:* benign, malignant.
2. *clump thickness:* 1–10
3. *uniformity of cell size:* 1–10
4. *uniformity of cell shape:* 1–10
5. *marginal adhesion:* 1–10
6. *single epithelial cell size:* 1–10
7. *bare nuclei:* 1–10
8. *bland chromatin:* 1–10
9. *normal nucleoli:* 1–10
10. *mitoses:* 1–10.

Table 6 Results generated for the Wisconsin Breast Cancer dataset (data 2)

Information criteria	*Decision*
AIC	5 classes
CAIC, CAIU, BIC	2 classes
AIC3, AICc, HT, aBIC, HQ	3 classes

Table 7 Information Criteria for choosing the number of classes for Wisconsin Breast Cancer Database

Classes	*Gsq*	*Llik*	*AIC*	*AIC3*	*AICc*	*HT*	*CAIC*	*CAIU*	*BIC*	*aBIC*	*HQ*
2	7,764	–7,846	16,018	16,181	16,117	16,119	16,923	16,304	16,760	16,242	1,6305
3	7,370	–7,649	15,788	16,033	16,053	16,056	17,147	16,356	16,902	16,124	1,6219
4	7,169	–7,549	15,751	16,078	16,327	16,334	17,566	16,770	17,239	16,200	1,6326
5	6,954	–7,441	15,700	16,109	16,857	16,870	17,970	17,475	17,561	16,262	1,6419
6	6,840	–7,384	15,750	16,241	18,079	18,105	18,475	18,930	17,984	16,425	1,6614

The same conclusion is projected on the second application. The criteria BIC, CAIC and CAIU are efficient and predict the number of classes. Note also that the AIC criterion gives five classes which is far from reality. This may be explained by the fact that the criterion is sensitive to the size of the sample which confirms the results obtained in Figure 2. We concluded that for samples larger than 500, some information criteria tend to predict the exact number of classes, but in some conditions:

1 number of variables greater than 12

2 all variables are two modalities

Thus, we can enhance this comparison in future work by adding another hypothesis which suppose that also the number of modalities for the chosen variables have an effect on the behaviour of these criteria of information.

5 Conclusions

The selection of the best model of latent class analysis is a critical issue because it can affect substantive interpretations of the studied phenomenon. This paper showed that the information criteria do not have the same decision of choosing the best LCA model, because these criteria are affected by the sample size and the number of variables.

Table 8 Comparison of the roles of the information criteria

ICs	*Remarks*
AIC	Good prediction but not consistent
AICc	Correction of AIC bias
AIC3	Predicts a parsimonious model
BIC	Consists of and predicts a parsimonious model
ABIC	Adjusts the BIC criteria
CAIC	Making AIC asymptotically coherent
HT	Reduces bias in the case of small samples
CAIU	Achieves a good performance
HQ	Alternative to AIC but little used
AWE	Adds a third dimension to the information criteria

The application of Latent class analysis to the breast cancer dataset, in order to detect the number of Hidden classes, showed that it is necessary to take into account all proposed criteria and not only classical criteria AIC and BIC.

The choice of suitable criteria to decide the right model is essential, this choice must consider the size of the population and the number of variables. Therefore, we provide some important remarks (Table 8) which can guide the choice of the relevant criteria.

References

Abarda, A., Bentaleb, Y., El Moudden, M., Dakkon, M., Azhari, M., Zerouaoui, J. and Ettaki, B. (2018) 'Solving the problem of latent class selection', *Proceedings of the International Conference on Learning and Optimization Algorithms: Theory and Applications*, pp.15, ACM.

Aitkin, M. and Rubin, D. (1985) 'Estimation and hypothesis testing in finite mixture models', *Journal of the Royal Statistical Society. Series B*, Vol. 47, No. 1, pp.67–75.

Aitkin, M., Anderson, D. and Hinde, J. (1981) 'Statistical modeling of data on teaching styles', *Journal of the Royal Statistical Society. Series A*, Vol. 144, No. 4, pp.419–461.

Akaike, H. (1992) 'Information theory and an extension of the maximum likelihood principle', in Kotz, S. and Johnson, N.L. (Eds.): *Breakthroughs in Statistics. Springer Series in Statistics (Perspectives in Statistics)*, pp.610–624, Springer, New York.

Andrews, R.L. and Currim, I.S. (2003) 'A comparison of segment retention criteria for finite mixture logit models', *Journal of Marketing Research*, Vol. 40, No. 2, pp.235–243.

Atkinson, A.C. (1980) 'A note on the generalized information criterion for choice of a model', *Biometrika*, Vol. 67, No. 2, pp.413–418.

Banfield, J.D. and Raftery, A.E. (1993) 'Model-based Gaussian and non-Gaussian clustering', *Biometrics*, Vol. 49, No. 3, pp.803–821.

Boekee, D.E. and Buss, H.H. (1981) 'Order estimation of autoregressive models', *Proceedings of the 4th Aachen Colloquium: Theory and Application of Signal Processing*, pp.126–130.

Bozdogan, H. (1987) 'Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions', *Psychometrika*, Vol. 52, No. 3, pp.345–370.

Burnham, K.P. and Anderson, D.R. (2002) *Model Selection and Multimodel Inference: A Practical Information Theoretic Approach*, Springer Science and Business Media, New York.

Burnham, K.P. and Anderson, D.R. (2004) 'Multimodel inference: understanding AIC and BIC in model selection', *Sociological Methods and Research*, Vol. 33, No. 2, pp.261–304.

Claeskens, G. and Hjort, N.L. (2008) *Model Selection and Model Averaging*, Cambridge Books, New York, NY; Cambridge.

Clogg, C.C. (1995) 'Latent class models', *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, pp.311–359, Springer, Boston, USA.

Draper, D. (1995) 'Assessment and propagation of model uncertainty', *Journal of the Royal Statistical Society. Series B*, Vol. 57, No. 1, pp.45–97.

Dziak, J., Coffman, D.L., Lanza, S.T. and Li, R. (2012) 'Sensitivity and specificity of information criteria', *The Methodology Center and Department of Statistics, Penn State, The Pennsylvania State University*, Vol. 16, No. 30, p.140.

Everitt, B.S. (1981) 'A Monte Carlo investigation of the likelihood ratio test for number of components in a mixture of normal distributions', *Multivariate Behavioral Research*, Vol. 16, No. 2, pp.171–180.

Everitt, B.S. (1988) 'A Monte Carlo investigation of the likelihood ratio test for number of classes in latent class analysis', *Multivariate Behavioral Research*, Vol. 23, No. 4, pp.531–538.

Hannan, E.J. and Quinn, B.G. (1979) 'The determination of the order of an autoregression', *Journal of the Royal Statistical Society. Series B*, Vol. 41, No. 2, pp.190–195.

Hastie, T., Tibshirani, R. and Friedman, J. (2001) *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer-Verlag, New York.

Hurvich, C.M. and Tsai, C. (1989) 'Regression and time series model selection in small samples', *Biometrika*, Vol. 76, No. 2, pp.297–307.

Kass, R.E. and Raftery, A.E. (1995) 'Bayes factors', *Journal of the American Statistical Association*, Vol. 90, No. 430, pp.773–795.

Kitagawa, G. (2018) 'Information criteria for statistical modeling in data-rich era', in Anh, L., Dong, L., Kreinovich, V. and Thach, N. (Eds.): *Econometrics for Financial Applications. ECONVN 2018. Studies in Computational Intelligence*, Vol. 760, pp.20–43, Springer, Cham.

Lancelot, R. and Lesnoff, M. (2005) *Sélection de modèles avec l'AIC et critères d'information dérivés*, CIRAD, Montpellier.

Lin, T.H. and Dayton, C.M. (1997) 'Model selection information criteria for non-nested latent class models', *Journal of Educational and Behavioral Statistics*, Vol. 22, No. 3, pp.249–264.

Linzer, D.A. and Lewis, J.B. (2011) 'poLCA: an RPackage for polytomous variable latent class analysis', *Journal of Statistical Software*, Vol. 42, No. 10, pp.1–29.

Mangan N.M., Kutz J.N., Brunton S.L. and Proctor J.L. (2017) 'Model selection for dynamical systems via sparse regression and information criteria', *Proc. R. Soc. A*, Vol. 473, No. 2204, DOI: 10.1098/rspa.2017.0009.

Mangasarian, O.L. and Wolberg W.H., (1990) 'Cancer diagnosis via linear programming', *SIAM News*, Vol. 23, No. 5, pp.1–18.

McCullagh, P. and Nelder, J.A. (1989) *Generalised Linear Models II*, Vol. 37, CRC Press, , New York.

McQuarrie, A., Shumway, R. and Tsai, C.L. (1997) 'The model selection criterion AICu', *Statistics and Probability Letters*, Vol. 34, No. 3, pp.285–292.

Pels, W.A., Alam, S., Carpp, L.N. and Moodie, E.E. (2018) 'A call for caution in using information criteria to select the working correlation structure in generalized estimating equations', *Epidemiology*, Vol. 29, No. 6, pp.e51–e52.

Rissanen, J. (1978) 'Modeling by shortest data description', *Automatica*, Vol. 14, No. 5, pp.465–471.

Schwarz, G. (1978) 'Estimating the dimension of a model', *The Annals of statistics*, Vol. 6, No. 2, pp.461–464.

Sclove, S.L. (1987) 'Application of model-selection criteria to some problems in multivariate analysis', *Psychometrika*, Vol. 52, No. 3, pp.333–343.

Seo, T.K. and Thorne J.L. (2018) 'Information criteria for comparing partition schemes', *Systematic Biology*, Vol. 67, No. 4, pp.616–632.

Sugiura, N. (1978) 'Further analysis of the data by Akaike's information criterion and the finite corrections', *Communications in Statistics, Theory, and Methods*, Vol. 7, No. 1, pp.13–26.

Tibshirani, R. and Knight, K. (1999) 'The covariance in inflation criterion for adaptive model selection', *Journal of the Royal Statistical Society B*, Vol. 61, No. 3, pp.529–546.

Wolberg, W.H. and Mangasarian, O.L. (1990) 'Multisurface method of pattern separation for medical diagnosis applied to breast cytology', *Proceedings of the National Academy of Sciences*, Vol. 87, pp. 9193–9196.

Yang, C.C. (1998) *Finite Mixture Model Selection with Psychometrics Applications*, PhD dissertation, University of California, Los Angeles.

Yang, C.C. (2006) 'Evaluating latent class analysis models in qualitative phenotype identification', *Computational Statistics and Data Analysis*, Vol. 50, No. 4, pp.1090–1104.

Youness, G. (2004) *Contributions à une méthodologie de comparaison de partitions*, Thèse de doctorat, Paris 6.

Zhang, Z., Abarda, A., Contractor, A.A., Wang, J. and Dayton, C.M. (2018) 'Exploring heterogeneity in clinical trials with latent class analysis', *Annals of Translational Medicine*, Vol. 6, No. 7, doi: 10.21037/atm.2018.01.24.

Zwitter, M. and Soklic, M. (1988) *Breast Cancer Data*, Institute of Oncology University Medical Center, Ljubljana, Yugoslavia [online] http://archive.ics.uci.edu/ml/datasets/Breast+Cancer (accessed 15 February 2018).