Research Techniques Made Simple: Latent Class Analysis

Luigi Naldi[1,2] and Simone Cazzaniga[2,3]

Latent class analysis (LCA) is a statistical technique that allows for identification, in a population characterized by a set of predefined features, of hidden clusters or classes, that is, subgroups that have a given probability of occurrence and are characterized by a specific and predictable combination of the analyzed features. Compared with other methods of so called data segmentation, such as hierarchical clustering, LCA derives clusters using a formal probabilistic approach and can be used in conjunction with multivariate methods to estimate parameters. The optimal number of classes is the one that minimizes the degree of relationship among cases belonging to different classes, and it is decided by relying on methods such as the Bayesian Information Criterion that capitalize on the value of the negative log-likelihood function, a well-established measure of the goodness of fit of a statistical model. LCA has not been extensively used in dermatology. The areas of application are manifold, from the phenotype classification to the analysis of behavior in relation with risk factors to the performance of diagnostic tests.

Journal of Investigative Dermatology (2020) **140,** 1676–1680; doi:10.1016/j.jid.2020.05.079

INTRODUCTION

Latent class analysis (LCA) is a statistical way to uncover hidden clusters in data by grouping subjects with a number of prespecified multifactorial features or manifest variables into latent classes (LCs), that is, subgroups with similar characteristics based on unobservable membership (Banfield and Raftery, 1993). The assumption is that, theoretically, any combination of a set of features could happen, but in reality, only a few of them do happen, forming a limited set of clusters where the individual features have a specific probability of occurrence, exactly the LCs. For example, one may question if expert clinicians confronted with a series of patients with different clinical features are similarly consistent when posing a diagnosis of psoriatic arthritis. In a study using LCA, it was documented that expert clinicians group together into two clusters labeled as high and low diagnosers of psoriatic arthritis. No intermediate category was found (Symmons et al., 2006). Less experienced clinicians or residents might have been clustered in a larger number of categories, expressing diagnostic uncertainty.

Terminology matters and we refer to the Glossary for definitions (see Supplementary Materials). The term latent implies that the analysis is based on an error-free underlying variable that is not directly measurable or observable but that can cause effects, for example, the diagnostic attitude of clinicians confronted with a given clinical scenario.

HOW TO USE LCA

Clustering and the concept of finite mixture modeling

Technically speaking, LCA is a special kind of finite mixture model (FMM) (Bouveyron et al., 2019), which assumes that an observed set of data derives from several underlying subpopulations and makes statistical inferences about the properties of these subpopulations having information on the pooled population only. In the previous example, the only distribution available is the distribution of the diagnostic decisions by experienced clinicians. How these clinicians group together, considering their propensity to make a diagnosis of psoriatic arthritis when confronted with specific clinical features, is not known. Unlike other clustering techniques such as hierarchical clustering that try to find clusters with some arbitrary chosen distance measure, FMM derives clusters using a probabilistic approach.

There are two sets of parameters in an LCA. The first is the set of inclusion probabilities (or class membership probabilities), that is, the probability that any random case in a population will be included in any LC. In the previous example, there are two classes, high and low diagnosers, and each experienced clinician has a given probability of belonging to one or the other class. The second parameter is the conditional probability that, given a specific class, a variable takes a certain value, for example, the probability that a patient with a specific feature, being assessed by a clinician in the class of low diagnoser, is classified as a patient with psoriatic arthritis. These probabilities are usually presented in a tabular format, as in Table 3, showing data from a study of hidradenitis suppurativa (HS). In the table, for example, the inclusion probabilities were 0.48 and 0.26 for LC1 and LC2, respectively. Comedones had a conditional probability of 0.25 in LC1 and 0.74 in LC2.

Parameters of the subdistributions are usually determined by maximum likelihood estimation with the expectation-

[1]Department of Dermatology, AULSS8 Ospedale San Bortolo, Vicenza, Italy; [2]Centro Studi GISED, Bergamo, Italy; and [3]Department of Dermatology, Inselspital University Hospital, Bern, Switzerland

Correspondence: Luigi Naldi, Centro Studi GISED, Via Clara Maffei 4, 24121 Bergamo, Italy. E-mail: luigi.naldi@gised.it

Abbreviations: AIC, Akaike Information Criterion; BIC, Bayesian Information Criterion; FMM, finite mixture model; HS, hidradenitis suppurativa; LC, latent class; LCA, latent class analysis; RF, rheumatoid factor; RMLCA, repeated measures latent class analysis

SUMMARY POINTS

What is Latent Class Analysis?

Latent class analysis (LCA) is a statistical way to uncover hidden clusters in data. This technique divides a set of observations (cases) characterized by several variables into mutually exclusive groups or classes, such that the observed variables are unrelated to each other within each class (local independence) and observations are similar in each class but different from those in other classes. Technically speaking, LCA is a special kind of finite mixture model (FMM), also known as unsupervised learning models, which model a statistical distribution by a mixture (or weighted sum) of other distributions and group similar data together based on selected parameters (i.e., data segmentation).

The optimal number of clusters in the set of observations is decided based on explicit probabilistic rules such as the Akaike Information Criterion and the Bayesian Information Criterion.

One advantage of FMMs compared with other methods of data segmentation, such as cluster analysis, is that they can be used in conjunction with multivariate methods.

Variables in LCA should be qualitative and nominal. When continuous variables are used, alone or in combination with categorical variables, the term latent profile analysis is usually preferred.

What are the major applications of LCA in clinical research?

LCA has a potentially extensive field of applications in clinical research, where qualitative and nominal variables are frequently used, ranging from phenotype classification to the analysis of behavior in relation to risk factors to the performance of diagnostic tests. The actual applications in dermatology have been, however, rather limited.

maximization algorithm, a well-established measure of the goodness of fit of a statistical model. The typical equation is:

$$p(x_i) = \sum_{k=1}^{K} p_k \prod_{n=1}^{N} p_n(x_{in}|k),$$

where $p(x_i)$ is the probability of observing a particular combination of responses in a group of N variables, p_k is the probability of membership in LC k, and $p_n(x_{in}|k)$ is the probability of response to variable n, conditional on membership in LC k.

As already mentioned, LCA divides a set of observations (cases) into mutually exclusive groups, or classes, such that manifest variables are unrelated to each other within each class (local independence) and observations are similar in each class but different from those in other classes. Additional covariates can also be used to predict class membership (Figure 1). Going back to the initial example, each clinician is confronted with a set of variables characterizing the patient as a case or noncase of psoriatic arthritis. These manifest variables, such as family history of psoriasis, rheumatoid factor (RF) titer, nail dystrophy, and toenails dactylitis, distribute differently in the two classes of high versus low diagnosers of psoriatic arthritis. In the analysis, to adjust for covariates that may affect classification, such as sex or age of the clinician, multivariate methods can be employed.

Format of variables

In LCA, measurable variables and covariates should be qualitative and nominal, with one or more categories per variable (e.g., sex [males and females] and age categories). Depending on the software, variables can be directly handled as nominal entities or must be entered as dichotomous (dummy) variables. For example, RF titers were not taken as a continuous variable in the analysis we mentioned previously, but two dummy variables were employed, RF titer higher than 40 and RF higher than 80.

When manifest variables are taken as continuous, alone or in combination with categorical variables, the term latent profile analysis is the preferred one (Oberski, 2016).

Longitudinal LCA, also known as repeated measures LCA (RMLCA), is an extension of LCA (Collins and Lanza, 2010). By examining repeats of the same categorical indicator, RMLCA allows to see how many common patterns of change over time emerge and what the probability of a target outcome is for each repeat in each class.

Different free and commercial software packages are available for LCA. Table 1 presents a list of the main available ones.

Choosing the number of classes

LCA usually provides several options for data grouping, and a crucial problem is to choose the optimal number of classes (k). The decision should be based on statistical ground. The methods more frequently adopted, such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), involve finding the k value that minimizes the negative log-likelihood function, increased by some penalty function that reflects the complexity of the model (Bouveyron et al., 2019). BIC is usually the preferred choice.

Interpretability is also an important issue but may give rise to some arbitrary decisions. In a study, comorbidity patterns were assessed in a series of more than 110,000 incident patients with psoriasis (Wu et al., 2018). The value of BIC was 10,320 for a four-class model and 7,814 for a five-class model. In spite of a higher BIC value, the four-class model was chosen because it was more easily interpretable based on the following classes: multi-comorbid class (patients with a variety of conditions), metabolic syndrome class, hypertension and chronic obstructive pulmonary disease class, and relatively healthy class. The decision of forgetting BIC values appears questionable and further validation is needed.

LCA IN DERMATOLOGY

LCA has not been extensively used in dermatology. We searched Medline up to 28 February 28 2020 and retrieved a total 6,159 papers using an LCA methodology. Out of these, only 37 papers dealt with dermatological conditions (see Supplementary Material).

Table 1. Details of the Main Available Software Packages for LCA

Software name	License	Package/plugin	Covariates	Polytomous manifest variables	Continuous manifest variables	Longitudinal LCA	Other features
R (R Foundation for Statistical Computing, Vienna, Austria)	Open source	poLCA	Yes	Yes	No	No	Results visualization, dataset simulation
		e1071 (lca)	No	No	No	No	—
		BayesLCA	No	No	No	No	Bayesian setting LCA
		RandomLCA	No	No	No	Yes	Random effects LCA
		LCAvarsel	Yes	Yes	No	No	Variable selection framework
SAS (SAS Institute Inc, Cary, NC)	Commercial	proc LCA	Yes	Yes	No	Yes (with proc LTA)	Accounting for sampling weights and clusters
STATA (StataCorp LLC, College Station, TX)	Commercial	LCA plugin	Yes	Yes	No	No	Accounting for sampling weights and clusters
MPLUS (Muthén & Muthén Computer Software, Los Angeles, CA)	Commercial	—	Yes	Yes	Yes	Yes	Ordinal, censored, and count manifest variables; FMMs and mixture regression; Random effects LCA
Latent GOLD (Statistical Innovations Inc Belmont, MA)	Commercial	—	Yes	Yes	Yes	Yes	Ordinal and count manifest variables; Multilevel models; Random effects LCA

Abbreviations: FMM, finite mixture model; LCA, latent class analysis; LTA, latent transition analysis.

The areas of application were in rank order of frequency: the phenotype classification of allergic diseases and eczema (n = 12), the analysis of behavior in relation to several different risk factors (n = 10; out of these, six studies were dealing with sexually transmitted disease), the phenotype classification of skin diseases other than eczema (n = 8, including psoriasis, dermatomyositis, vitiligo, HS, chronic skin ulcers, and psychodermatological conditions), the performance of diagnostic tests (n = 4), and the pattern of response to drugs and adverse reactions (n = 3). Overall, the absence of studies in the area of cutaneous oncology and the limited number of studies dealing with chronic inflammatory diseases other than eczema is remarkable.

The use of LCA is well established when assessing patterns of behavior: a total of 2,301 (39%) studies in our search were dealing with behavioral issues. This high prevalence is partly because of the fact that a number of leading researchers have published papers encouraging the use of LCA in behavioral studies (Lanza and Cooper, 2016), because multiple aspects of individual functioning in mental health can be studied holistically.

Table 2. Prevalence and Conditional Probabilities of LC Indicators for Sleep Disturbance in Allergic Disease

		Class Conditional Probability[1]				
Disease	Prevalence (%)	LC1	LC2	LC3	LC4	LC5
Eczema	7.2%	0.27	0.16	0.10	0.35	0.02
Asthma	8.0%	0.08	0.24	0.08	0.48	0.02
Hay fever	7.5%	0.08	0.34	0.03	0.33	0.03
Resp allergy	11.2%	0.00	0.80	0.07	0.73	0.01
Dig allergy	3.9%	0.09	0.10	0.03	0.31	0.00
Fatigue	15.8%	0.14	0.11	0.72	0.83	0.02
Sleepiness	12.8%	0.04	0.03	0.63	0.71	0.02
Insomnia	19.3%	0.30	0.16	0.62	0.75	0.07

Abbreviations: Dig, digestive; LC, latent class; Resp, respiratory.
[1]Probabilities were estimated from the available graph and may show slight variations when compared with the actual data in the study.
Adapted from Silverberg et al., 2015.

Fatigue, sleep disturbance, and allergic disorders

Although poor sleep quality has been well documented in childhood eczema, few studies have examined the quality of sleep in the adult eczema population. Despite the high variability of presentation, is it possible to define consistent patterns of association of fatigue, sleep disturbance, and allergic disease in the adult population? The question was addressed

Table 3. Prevalence and Conditional Probabilities of LC Indicators for Clinical Patterns of HS

		Class Conditional Probability		
Clinical Pattern	Prevalence (%)	LC1	LC2	LC3
Probability of class membership		*0.48*	*0.26*	*0.26*
Armpit/breast	71.8%	0.74	0.96	0.45
Gluteal area	30.3%	0.12	0.37	0.54
Ears/chest/other	22.3%	0.06	0.55	0.18
Hypertrophic scars	33.3%	0.41	0.54	0.01
Comedones	44.3%	0.25	0.74	0.49
Epidermal cysts/macrocysts	8.4%	0.04	0.23	0.01
Papules and folliculitis	45.3%	0.23	0.71	0.71
Pilonidal sinus	30.4%	0.27	0.48	0.18
Family history of HS	35.4%	0.29	0.44	0.37
Acne/history of severe acne	26.7%	0.21	0.47	0.16

Abbreviations: HS, hidradenitis suppurativa; LC, latent class.
Adapted from Canoui-Poitrine et al., 2013.

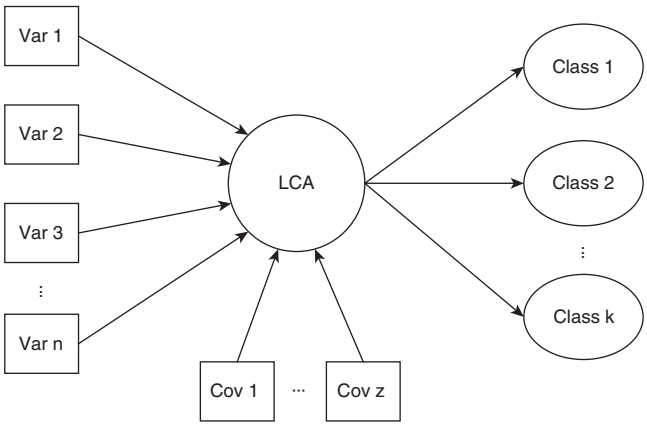

Figure 1. Representation of an LCA model. Var n represent the manifest variables; Cov z, the additional covariates; and Class k, the latent classes predicted by LCA. LCA, latent class analysis.

in a study analyzing data obtained in the context of the 2012 National Health Interview Survey (Silverberg et al., 2015). The data analyzed pertained, in particular, to history of eczema, sleep problems, and overall health. BIC and AIC were used to select the best fitting model. The model had five classes, LC1–5 (Figure 2). Two classes presented high probabilities of sleep disturbance: LC4, characterized by high probabilities of eczema, asthma, hay fever, and food allergy, and LC3 with low probabilities of these disorders. LC1 had an intermediate probability of insomnia but not fatigue or sleepiness and an intermediate probability of eczema (Table 2). The study presented data from a cross-sectional study; patterns of changes over time may represent an interesting issue to explore in future studies by using RMLCA.

HS phenotype

Considerable variability occurs in the clinical presentation and disease severity of HS. In a cross-sectional study, LCA was applied to a series of 648 consecutive patients with HS with the aim of building an empirical classification scheme without any a priori hypotheses (Canoui-Poitrine et al., 2013).

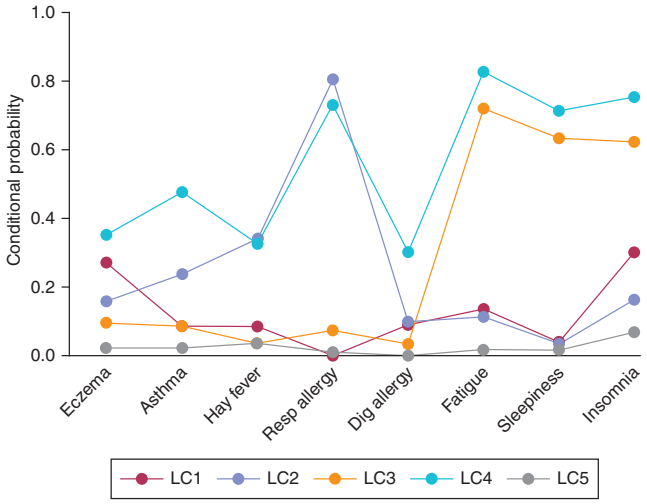

Figure 2. Identification of five classes of sleep disturbance in allergic disease and the conditional probabilities of the items studied (redrawn from Silverberg et al., 2015). LC, latent class.

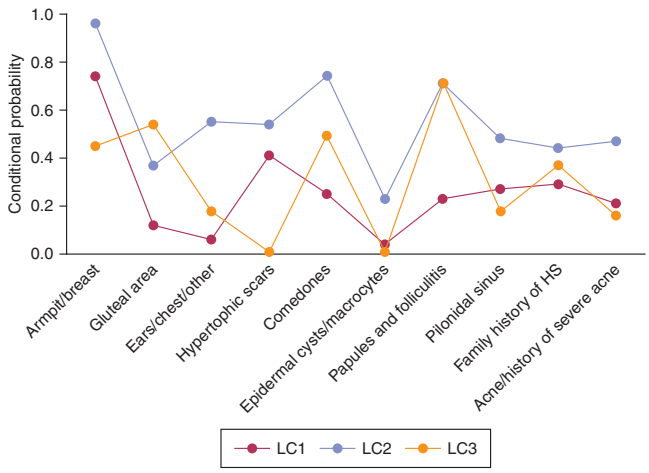

Figure 3. Identification of three main clinical patterns of HS and conditional probabilities of the individual features in each class (redrawn from Canoui-Poitrine et al., 2013). HS, hidradenitis suppurativa; LC, latent class.

Ten indicators pertaining to clinical features, namely, sites involved, lesion type (nodules, hypertrophic scars, comedones, papules and folliculitis, epidermal cysts, macrocysts, and pilonidal sinuses), severity assessment (by Sartorius score and Hurley stage), family history, and previous history of severe acne, were chosen to inform the clinical classification.

A classification into three LCs (LC1–3) provided the best fit of data as estimated by using BIC (Figure 3). LC1 patients (n = 299, 48%) had high probabilities for breast and armpit involvement and for hypertrophic scars; LC2 patients (n = 161, 26%) had high probabilities for involvement of the ears, chest, back, or legs and also for follicular lesions and a history of severe acne; and LC3 patients (n = 158, 26%) were characterized by gluteal involvement, follicular papules, and folliculitis (Table 3). Significant differences were found among the three LCs for sex, body mass index, smoking status, severity scores, age at disease onset, and HS duration. The identification of subgroups may allow for further investigation of matters such as biological markers, class changes over time, and prognosis.

THE FUTURE OF CLASSIFICATION METHODS

With the progress of information systems in medicine, a huge amount of data can be routinely collected, that is, big data. LCA can be used to analyze these data to find clusters, especially when rare LCs (with <5% data) are present. However, when a lot of information is available, a possible drawback is that redundant or noninformative variables present in the dataset may potentially introduce biases or reduce the efficiency of clustering algorithms. For this reason, standard stepwise selection or more complex search procedures via genetic algorithms may be used to create a reduced dataset for the subsequent analysis (Dean and Raftery, 2010). Alternatively, more complex models, such as artificial neural networks, able to find data clusters without assumptions about data distribution or parameters can be considered.

There are several areas where LCA can be efficiently employed in dermatological research. For example, little has been done for profiling endotypes of complex disorders such as atopic dermatitis or psoriasis; LCA can be used in a way similar to what has been proposed for asthma where clinical data, functional data, comorbidities,

MULTIPLE CHOICE QUESTIONS

1. What is meant by the term latent classes?
   A. A type of qualitative data
   B. Subgroups of data with similar characteristics based on unobservable membership
   C. A statistical model to find hidden clusters in data
   D. Variables that are not directly measurable or observable but that can cause effects
   E. A synthetic representation of multidimensional variables
2. What is the Bayesian Information Criterion (BIC)?
   A. A function of the positive log-likelihood
   B. A likelihood ratio test
   C. A statistical test to reject the null hypothesis
   D. A measure of correlation among variables
   E. A measure of corrected model fit based on the value of the negative log-likelihood augmented by a penalty function
3. What is the simplest way to validate results in latent class analysis (LCA)?
   A. k-fold cross-validation
   B. Bootstrap method
   C. Splitting procedure
   D. A combination of bootstrap and k-fold cross-validation
   E. No validation is required for LCA results
4. In a study, the BIC values for different number of classes (k) are reported in the table. Which classification would you prefer?

Number of classes (k)	BIC
1	16,739.0
2	16,636.1
3	16,604.2
4	16,707.4
5	16,813.3

   A. A single class classification
   B. The classification into two groups
   C. The classification into three groups
   D. The classification into four groups
   E. The classification into five groups
5. How does cluster analysis differ from LCA?
   A. There is no difference. The two terms are synonyms.
   B. At variance with cluster analysis, LCA can be used in conjunction with multivariate methods, avoiding a two-step approach in estimating parameters.
   C. Cluster analysis works well with any kind of data, whereas LCA works with continuous data only.
   D. Cluster analysis requires a preliminary assumption of the number of classes present in the data at hand.
   E. Cluster analysis is a nonparametric technique.

See online version of this article for a detailed explanation of correct answers.

and inflammatory parameters were considered together (Howard et al., 2015). Similarly, an area of potential development is the better characterization of symptoms such as itching or pain where clinical, psychological, and behavioral factors may interact. Finally, large opportunities for the use of LCA exist in oncology to analyze patterns of presentation and/or progression of cancer and to assess variables affecting the impact of preventive measures or treatment.

CONFLICT OF INTEREST

The authors state no conflict of interest.

AUTHOR CONTRIBUTIONS

Data Curation: SG; Formal Analysis: SG; Investigation: LN, SG; Methodology: LN, SG; Project Administration: LN; Resources: LN, SG; Software: SG; Supervision: LN; Validation: LN, SG; Visualization: SG; Writing - Original Draft Preparation: SG, LN; Writing - Review and Editing: LN, SG

SUPPLEMENTARY MATERIAL

Supplementary material is linked to this paper. Teaching slides are available as supplementary material.

REFERENCES

Banfield JD, Raftery E. Model-based Gaussian and non-Gaussian clustering. Biometrics 1993;49:803–21.

Bouveyron C, Celeux G, Murphy TB, Raftery AE. Model-based clustering and classification for data science. Cambridge: Cambridge University Press; 2019.

Canoui-Poitrine F, Le Thuaut A, Revuz JE, Viallette C, Gabison G, Poli F, et al. Identification of three hidradenitis suppurativa phenotypes: latent class analysis of a cross-sectional study. J Invest Dermatol 2013;133:1506–11.

Collins LM, Lanza ST. RMLCA and LTA. In: Latent class and latent transition analysis with applications in the social, behavioral, and health sciences. Hoboken, NJ: John Wiley & Sons, Inc; 2010. p. 181–224.

Dean N, Raftery AE. Latent class analysis variable selection. Ann Inst Stat Math 2010;62:11–35.

Howard R, Rattray M, Prosperi M, Custovic A. Distinguishing asthma phenotypes using machine learning approaches. Curr Allergy Asthma Rep 2015;15:38.

Lanza ST, Cooper BR. Latent class analysis for developmental research. Child Dev Perspect 2016;10:59–64.

Oberski D. Mixture models: latent profile and latent class analysis. In: Robertson J, Kaptein M, editors. Modern statistical methods for HCI (Human–computer interaction series). Cham, Switzerland: Springer International Publishing Switzerland; 2016. p. 275–90.

Silverberg JI, Garg NK, Paller AS, Fishbein AB, Zee PC. Sleep disturbances in adults with eczema are associated with impaired overall health: a US population-based study. J Invest Dermatol 2015;135:56–66.

Symmons DP, Lunt M, Watkins G, Helliwell P, Jones S, McHugh N, et al. Developing classification criteria for peripheral joint psoriatic arthritis. Step I. Establishing whether the rheumatologist's opinion on the diagnosis can be used as the "gold standard. J Rheumatol 2006;33:552–7.

Wu CY, Hu HY, Li CP, Chou YJ, Chang YT. Comorbidity profiles of psoriasis in Taiwan: a latent class analysis. PLoS One 2018;13:e0192537.

DETAILED ANSWERS

1. **What is meant by the term latent classes?**

Answer: B. Subgroups of data with similar characteristics based on unobservable membership

Latent classes are subgroups of data or subjects, that is, clusters, with similar characteristics based on unobservable (latent) membership, where the individual features have a specific probability of occurrence.

2. **What is the Bayesian Information Criterion (BIC)?**

Answer: E. A measure of corrected model fit based on the value of the negative log-likelihood augmented by a penalty function

BIC is a measure of corrected model fit. It reflects the negative log-likelihood augmented by a penalty function depending both on the complexity of the model (represented by the number of free estimated parameters) and the logarithm of the sample size. It is also the commonly used decision rule to select the optimal number of classes in latent class analysis (LCA).

3. **What is the simplest way to validate results in latent class analysis (LCA)?**

Answer: C. Splitting procedure

As for any classification procedure, validation on a separate sample is required. The splitting procedure is the simplest method to validate the results of LCA. It involves a random splitting of the dataset into two equal groups: one is used to build the LCA model and the other to test its stability in terms of number of classes, classification accuracy, and predicted posterior probabilities.

4. **In a study, the BIC values for different number of classes (k) are reported in the table. Which classification would you prefer?**

Number of classes (k)	BIC
1	16,739.0
2	16,636.1
3	16,604.2
4	16,707.4
5	16,813.3

Answer: C. The classification into three groups

The k value that minimizes the BIC score is the preferred one. The BIC score reflects the negative log-likelihood function, increased by a penalty function influenced by the number of independent parameters estimated and by the sample size.

5. **How does cluster analysis differ from LCA?**

Answer: B. At variance with cluster analysis, LCA can be used in conjunction with multivariate methods avoiding a two-step approach in estimating parameters

One advantage of LCA compared with other methods of data segmentation, such as cluster analysis, is that it can be used in conjunction with multivariate methods, avoiding a two-step approach in estimating parameters.