Latent Class Modeling as a Probabilistic Extension of K-Means Clustering

Latent Class Cluster Models

According to Kaufman and Rousseeuw (1990), cluster analysis is "the classification of similar objects into groups, where the number of groups, as well as their forms are unknown". This same definition could be used for exploratory Latent Class (LC) analysis where a K-class latent variable is used to explain the associations among a set of observed variables. Each latent class, like each cluster, groups together similar cases.

Contrary to traditional ad hoc clustering approaches, the LC approach to clustering is model-based. The fundamental assumption underlying LC models is that of *local independence* which states that objects (persons, cases) in the same latent class share a common joint probability distribution among the observed variables. Since persons in the same latent class (cluster) cannot be distinguished from each other based on their observed responses, they are similar to each other (homogeneous) with respect to these observed variables. Persons are classified into that class having the highest posterior membership probability of belonging given the set of responses for that case.

LC is most similar to the K-Means approach to cluster analysis in which cases that are "close" to one of K centers are grouped together. In fact, LC clustering can be viewed as a probabilistic variant of K-Means clustering where probabilities are used to define "closeness" to each center (McLachlan and Basford, 1988). As such, LC clustering provides a way not only to formalize the K-Means approach in terms of a statistical model, but also to extend the K-Means approach in several directions.

LC Extensions of the K-Means Approach

1. **Probability-based classification**. While K-Means uses an ad-hoc distance measure for classification, the LC approach allows cases to be classified into clusters using model based posterior membership probabilities estimated by maximum likelihood (ML) methods. This approach also yields ML estimates for misclassification rates.

2. **Determination of number of clusters**. K-Means provides no assistance in determining the number of clusters. In contrast, LC clustering provides various diagnostics such as the BIC statistic, which can be useful in determining the number of clusters.

3. **Inclusion of variables of mixed scale types**. K-Means clustering is limited to interval scale quantitative variables, for which Euclidean distance measures can be calculated. In contrast, LC clustering can be performed on variables of mixed metrics. Variables may be continuous, categorical (nominal or ordinal), or counts or any combination of these.

   No need to standardize variables. Prior to performing K-Means clustering, variables must be standardized to have equal variance prior to avoid obtaining clusters that are dominated by variables having the largest amounts of variation. In contrast, the LC clustering solution is invariant of linear transformations on the variables; thus, standardization of variables is not necessary.

4. **Inclusion of demographics and other exogenous variables**. A common practice following a K-Means clustering is to use discriminant analysis to describe differences among the clusters on one or more exogenous variables. In contrast, the LC cluster model can be easily extended to include exogenous variables (covariates). This allows both classification and cluster description to be performed simultaneously using a single uniform ML estimation algorithm.

The General LC Cluster Model

The basic LC cluster model can be expressed as:

$$f(\mathbf{y}_i) = \sum_k p(x=k)\, f(\mathbf{y}_i|x=k)$$

while the LC cluster model with covariates is:

$$f(\mathbf{y}_i|\mathbf{z}_i) = \sum_k p(x=k|\mathbf{z}_i)\, f(\mathbf{y}_i|x=k)$$

or

$f(\mathbf{y}_i|\mathbf{z}_i) = \sum_k p(x=k|\mathbf{z}_i)\, f(\mathbf{y}_i|x=k,\mathbf{z}_i)$

where:

- $\mathbf{y}_i$: vector of dependent/endogenous/indicators for case i
- $\mathbf{z}_i$: vector of independent/exogenous/covariates for case i
- x: nominal latent variable (k denotes a class, k=1,2,…,K)

and $f(\mathbf{y}_i|x=k)$ denotes the joint distribution specified for the $\mathbf{y}_i$ given latent class x=k.

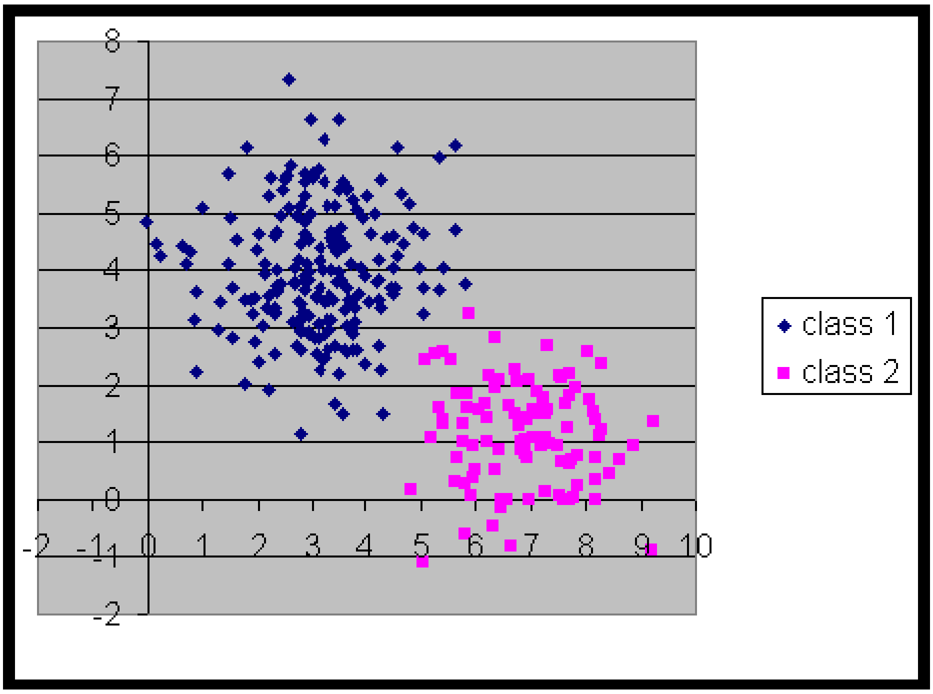
> For $\mathbf{y}_i$ continuous, the multivariate normal distribution is used with class-specific means. In addition, the within-class covariance matrices can be assumed to be equal or unequal across classes (ie., class independent or class dependent), and the local independence assumption can be relaxed by applying various structures to the within-class covariance matrices:
>
> - diagonal (local independence)
> - free or partially free -- allow non-zero correlations (direct effects) between selected variables

For variables of other/mixed scale types, local independence among the variables imposes restrictions on second-order as well as to higher-order moments. Within a latent class, the likelihood function under the assumption of independence is specified using the product of the following distributions:

- continuous: normal
- nominal: multinomial
- ordinal: restricted multinomial
- count: Poisson / binomial

LC Cluster vs. K-Means – Comparisons with Simulated Data

To examine the kinds of differences that might be expected in practice between LC cluster and K-Means clustering, we generated data of the type most commonly assumed when using K-Means clustering. Specifically, we generated several data sets containing two normally distributed variables within each of K=2 clusters. For data sets 1-3, the first cluster consists of 200 cases with mean (3,4), the second 100 cases with mean (7,1).



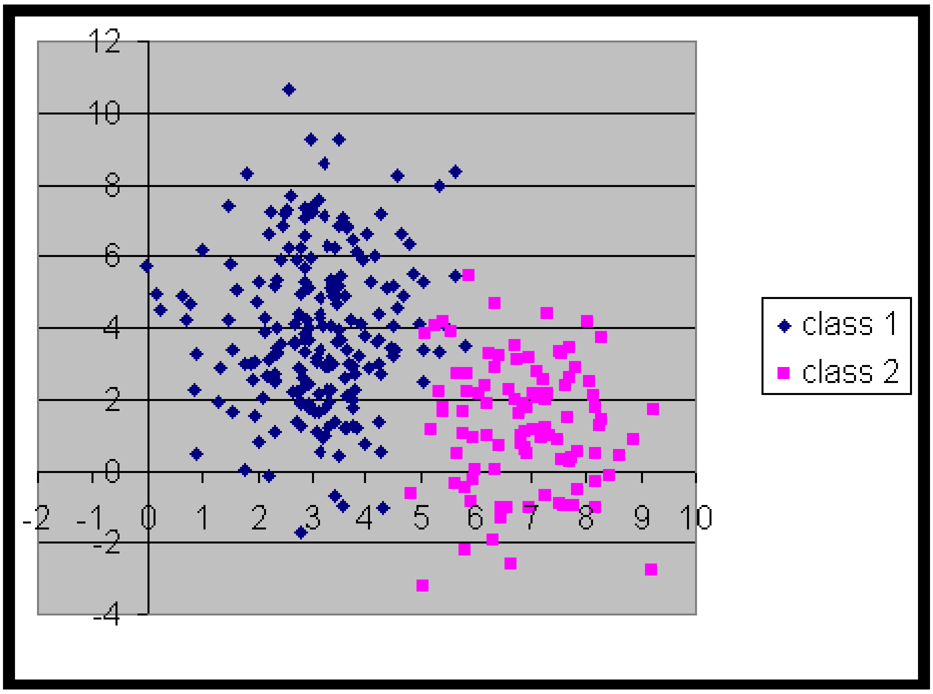
Data Set 1: Within each Class, Variables are Independent with Std. Dev. σ = 1

In data set 1, within each cluster the variables were generated to be independent with

standard deviation equal to 1. Data set 1 was generated to make discrimination easy and not exploit the inability of the K-Means approach to properly handle variables having different variances.

The LC models correctly identify this data set as arising from 2 clusters, having equal within-cluster covariance matrices (i.e., the "2-cluster, equal" model has the lowest BIC = 2154). The ML estimate for the expected misclassification rate is 1.1%. Classification based on the modal posterior membership probability resulted in all 200 cluster 1 cases being classified correctly and only 1 of the 100 cluster 2 cases, (y_1,y_2) = (5.08,2.43), being misclassified into class 1. For data set 1, use of K-means clustering with 2 clusters produced a comparable result – all 100 cluster 2 cases were classified correctly and only 1 of the 200 cluster 1 cases were misclassified, (y_1,y_2) = (4.32,1.49).

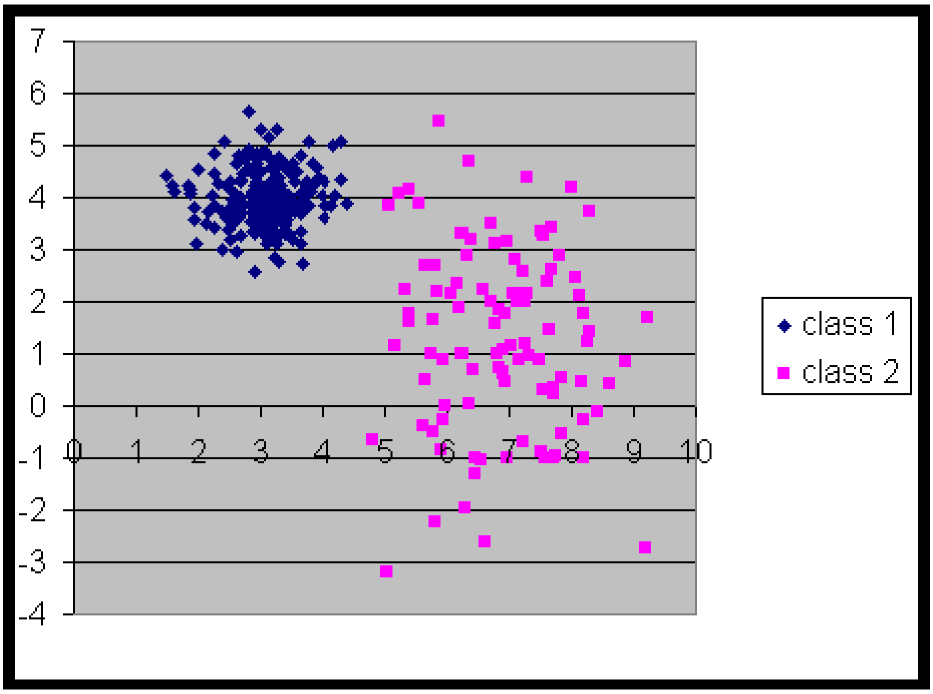
Data set 2 was identical to data set 1 except that the standard deviation for the second variable was doubled so the standard deviation for Y_2 was twice that of Y_1.



Data Set 2: Within each Class, Std. Dev. for y2 = 2σ

The LC models again correctly identify this data set as arising from 2 clusters, having equal within-cluster covariance matrices (i.e., the "2-cluster, equal" model has the lowest BIC = 2552). The ML estimate for the expected misclassification rate is 0.9%. Classification based on the modal posterior membership probability resulted in 3 of the cluster 1 cases and 1 of the cluster 2 cases being misclassified. For these data, K-Means performed much worse than LC clustering. Overall, 24 (8%) of the cases were misclassified (18 cluster 1 cases and 6 cluster 2 cases). When the variables were standardized to have equal variances prior to the K-Means analysis, the number of misclassifications dropped to 15 (5%), 10 of the cluster 1 and 5 of the cluster 2 cases, still markedly worse than the LC clustering.

Data set 3 threw in a new wrinkle of unequal standard deviations across clusters. To accomplish this, for cluster 1 the standard deviations were reduced to 0.5 for both variables. For cluster 2, the data remained the same as used in data set 2.

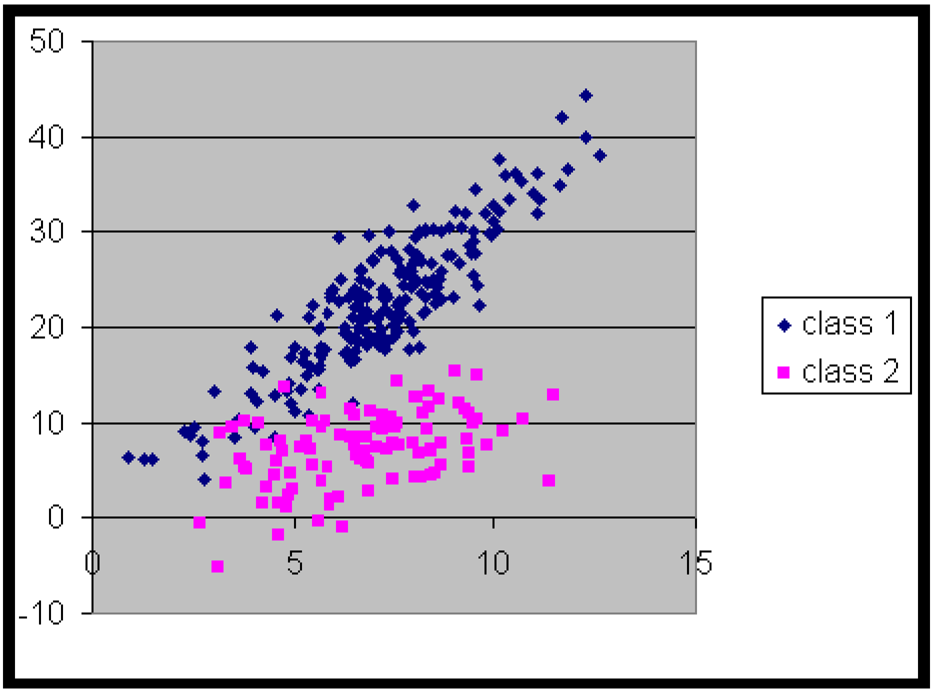


Data Set 3: Within Class 1, Std. Dev. for y1 and y2 = 0.5σ

The LC models correctly identify this data set as arising from 2 clusters, having unequal within-cluster covariance matrices (i.e., the "2-cluster, unequal" model has the lowest BIC = 1750). The ML estimate for the expected misclassification rate was 0.1%, and use of the modal posterior membership probabilities resulted in perfect classification. K-Means correctly classified all cluster 1 cases for these data but misclassified 6 cluster 2 cases. When the variables were standardized to have equal variances prior to a K-Means analysis, the results were identical, markedly worse than the LC clustering.

For data set 4 we added some within-class correlation to the variables so that the local independence assumption no longer held true. For class 1 the correlation added was moderate, while for class 2 only a slight amount of correlation was added.

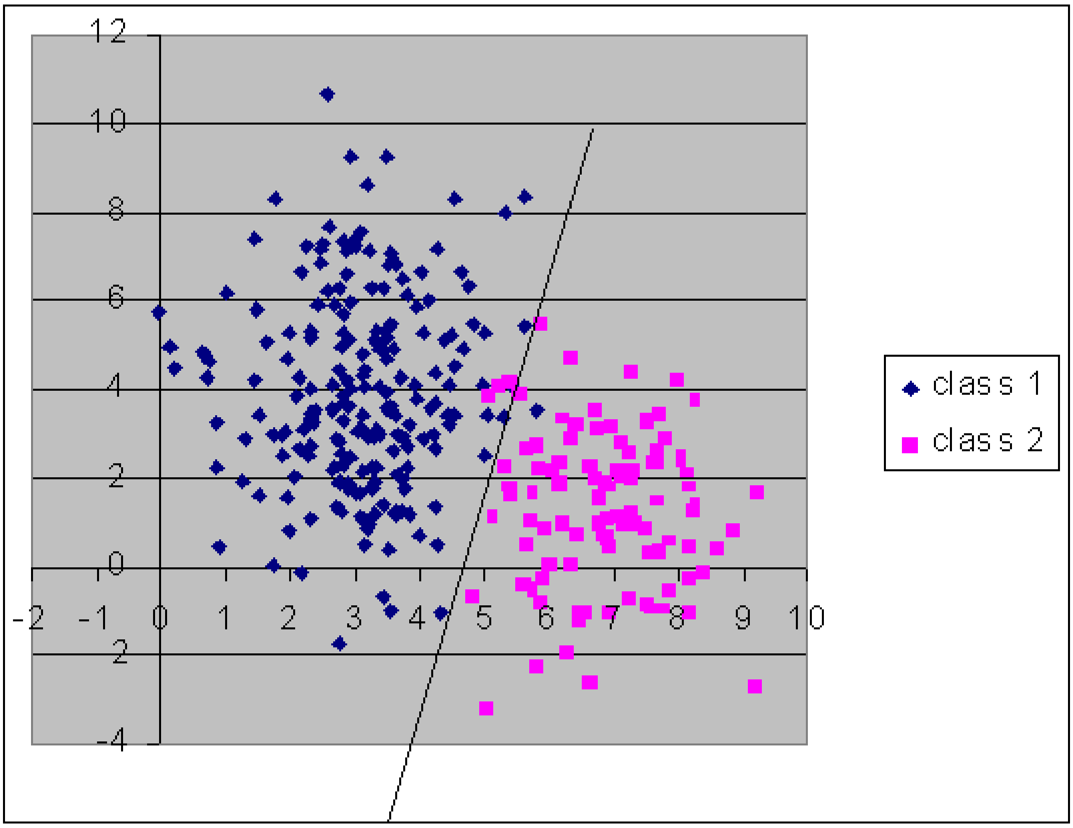
The LC models correctly identify this data set as arising from 2 clusters, having a "free" covariance structure – i.e., unequal within-cluster covariance matrices that included nonzero correlations within each class (i.e., the "2-cluster, free" model has the lowest BIC = 3263). The ML estimate for the expected misclassification rate was 3.3%, and use of the modal posterior membership probabilities resulted in 10 misclassifications among the 300 cases. K-Means performed very poorly for these data. While all 100 cluster 2 cases were classified correctly, 44 cluster 1 cases were misclassified, for an overall misclassification rate of almost 15%. If the recommended standardization procedure was followed prior to a K-Means analysis, the results would have been even worse -- 14 of the cluster 1 and 66 of the cluster 2 cases being misclassified, an error rate of over 26%!



Data Set 4: Moderate Correlation within Class 1, Slight Correlation within Class 2

Comparison with Discriminant Analysis

Since data set 2 satisfies the assumptions made in discriminant analysis, if we now pretend that the true class membership is *known* for all cases, the linear discriminant function can be calculated and used as the gold standard. We computed the linear discriminant function and appended it to the data set in Figure 5. Remarkably, it can be seen that the results are identical to that of latent class analysis – the same 4 cases are misclassified! These results show that it is not possible to obtain better classification results for these data than that given by the LC model.



Data Set 5: Data Set 2 with linear discriminant added

Summary and Conclusion

Recent developments in LC modeling offer an alternative approach to cluster analysis which can be viewed as a probabilistic extension of the K-Means approach to clustering. Using 4 data sets which simulate the occurrance of data from 2 homogeneous populations we compared LC with K-Means clustering. For all situations considered the LC approach does exceptionally well in classification. In contrast, the K-Means approach only does well when the variables have equal variance and the assumption of local independence holds true. Further research is recommended to explore other simulated settings.

Data set 1: diagonal / class-independent

Model	LL	BIC	Npar
1-Cluster equal	-1226	2475	4
2-Cluster equal	-1057	2154 *	7
3-Cluster equal	-1051	2159	10
1-Cluster unequal	-1132	2293	5
2-Cluster unequal	-1057	2160	8
3-Cluster unequal	-1051	2164	11

Data set 2: diagonal / class-independent

Model	LL	BIC	Npar

1-Cluster equal	-1333	2689	4
2-Cluster equal	-1256	2552 *	7
3-Cluster equal	-1251	2558	10
1-Cluster unequal	-1333	2689	5
2-Cluster unequal	-1252	2557	8
3-Cluster unequal	-1249	2561	11

Data set 3: diagonal / class-dependent

Model	LL	BIC	Npar
1-Cluster equal	-1209	2440	4
2-Cluster equal	-962	1964	7
3-Cluster equal	-906	1869	10
1-Cluster unequal	-1209	2440	4
2-Cluster unequal	-850	1750 *	9
3-Cluster unequal	-846	1772	14

Data set 4: free / class-dependent

Model	LL	BIC	Npar
1-Cluster diagonal	-1750	3522	4
2-Cluster diagonal	-1700	3450	9
3-Cluster diagonal	-1645	3370	14
1-Cluster free	-1686	3400	5
2-Cluster free	-1600	3263 *	11
3-Cluster free	-1596	3289	17