



HAL
open science

Levels Merging in the Latent Class Model

Christophe Biernacki

► **To cite this version:**

Christophe Biernacki. Levels Merging in the Latent Class Model. Statistical Learning Sustainability and Impact Evaluation, Jun 2023, Ancona (IT), Italy. hal-04370900

HAL Id: hal-04370900

<https://inria.hal.science/hal-04370900v1>

Submitted on 3 Jan 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Levels Merging in the Latent Class Model

C. Biernacki

Statistical Learning Sustainability and Impact Evaluation, June 21-23 2023, Ancona (Italy)



Take home message

potentially numerous

Variable.j = ((Level.1 Level.2) Level.3 (Level.4 Level.5) ... Level.(mj))

Level.{1,2} Level.{4,5}

Level merging could increase clustering accuracy!

Limit of the Latent Class Model (LCM) in case of numerous levels

- Cluster analysis is one of the main data analysis method. It aims at estimating a K groups partition $\mathbf{z} = (\mathbf{z}_1, \dots, \mathbf{z}_n)$ of a data set $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n) \in \mathcal{X}^d$
- Model-based clustering allows to reformulate cluster analysis as a well-posed estimation problem both for the partition \mathbf{z} and for the number K of groups [8]
- Model-based clustering has led practical successes, even in some challenging situations, as the high dimensional context ("large" d value) [3]
- When \mathcal{X} corresponds to d categorical variables (frequent situation), the j th one having m_j response levels, the standard latent class model (LCM, [5]) is used

$$p(\mathbf{x}_i; \theta) = \sum_{k=1}^K \pi_k \prod_{j=1}^d \prod_{h=1}^{m_j} (\alpha_k^{jh})^{x_i^{jh}}, \quad \text{where } \theta = (\pi, \alpha), \pi = (\pi_k), \alpha = (\alpha_k)$$

- When some m_j s are large another kind of dimensionality curse appear. Ex.: US states, French depts, professional social categories, medical nomenclature...

The standard latent class model is affected by such a curse of dimension paradigm since the number of free parameters is equal to $K - 1 + K \sum_{j=1}^d (m_j - 1)$

LCM-LM: a LCM extension for the case of numerous levels

- **Idea:** constrain some levels to share the same parameter value (level partitionning)
- **Notation:** for each variable j , L_j clusters of levels $\mathbf{w}_j = (\mathbf{w}_j^1, \dots, \mathbf{w}_j^{m_j})$
- **New model LCM-LM** (*Latent Class Model by Levels Merging*): from the full level clustering $\mathbf{w} = (\mathbf{w}_1, \dots, \mathbf{w}_d)$, the new LCM is defined by

$$p(\mathbf{x}_i | \mathbf{w}; \boldsymbol{\vartheta}) = \sum_{k=1}^K \pi_k \prod_{j=1}^d \prod_{\ell=1}^{L_j} \left(\frac{\beta_k^{j\ell}}{\sum_{h=1}^{m_j} w_j^{h\ell}} \right)^{\sum_{h=1}^{m_j} w_j^{h\ell} x_i^{jh}}$$

with $\boldsymbol{\beta} = (\beta_1, \dots, \beta_K)$, $\beta_k = (\beta_k^{j\ell}; j = 1, \dots, d; \ell = 1, \dots, L_j)$ and $\boldsymbol{\vartheta} = (\boldsymbol{\pi}, \boldsymbol{\beta})$

- **Remark:** LCM-LM is an extension a simpler model in [6]
- **An extreme case:** if $L_j = 1$ then LCM-LM acts as a variable selection (no effect of variable j on the clustering)

LCM-LM is parsimonious with regards to the number of levels since its number of parameters is equal to $K - 1 + K \sum_{j=1}^d (L_j - 1)$

LCM-LM: parameter estimation and model selection

Hint

Possible to reformulating the log-likelihood of LCM-LM as a classical LCM...

- LCM-LM \Leftrightarrow LCM with the following new data $\mathbf{y}_w = (\mathbf{y}_{w,1}, \dots, \mathbf{y}_{w,n})$
- Parameter estimation: $L(\vartheta | \mathbf{w}; \mathbf{x}) = L(\vartheta; \mathbf{y}_w) + c_w$, thus use a classical LCM EM
- Model selection:
 - ICL criterion [1, 2]: just written as a penalization of the classical LCM ICL

$$\text{ICL}_w^{\text{LCM-LM}} = \ln p(\mathbf{x}, \hat{\mathbf{z}}_w | \mathbf{w}) = \underbrace{\text{ICL}_w^{\text{LCM}}}_{\text{classical LCM ICL}} + \underbrace{c_w}_{\text{penalization}}$$

- MICL criterion [7]: MICL (*Maximum Integrated Complete-data Likelihood*) is a variant of ICL avoiding multiple parameter estimation

$$\text{MICL}_w^{\text{LCM-LM}} = \ln p(\mathbf{x}, \mathbf{z}_w^* | \mathbf{w}) \quad \text{with} \quad \mathbf{z}_w^* = \arg \max_{z \in \mathcal{Z}} \ln p(\mathbf{x}, z | \mathbf{w}).$$

Then, the model \mathbf{w}^* maximizing $\text{MICL}_w^{\text{LCM-LM}}$ is retained: $\mathbf{w}^* = \arg \max_{\mathbf{w} \in \mathcal{W}} \text{MICL}_w$. Finally a specific two step algorithm derived from [7] can be used for performing this optimization (not implemented yet).

Numerical experiments and conclusion

■ Numerical experiments

- **Data set:** 506 patients on the basis of petrial variates alone for the prostate cancer clinical trial data of [4], described by four categorical variables with various numbers of levels: performance rating (so-called PF, with 4 levels), cardiovascular disease history (HX, 2 levels), electrocardiogram code (EKG, 7 levels) and bone metastases (BM, 2 levels). There are also 62 missing values, so about 1% of the whole sample. There exists two stages of the desease (Stage 3 and Stage 4).
- **EM-like package:** RMixtcomp¹ for LCM categorical data and also dealing with missing values)
- **LCM-LM proposal:** since level 4 of variable PF is under-represented (only 2 individuals), we propose to merge it with level 3 of the same variable
- **Results with $K = 2$:**

	LCM	LCM-LM
ICL	-1728.4	-1618.4
error	49.2%	28.8%

■ Conclusion

- Very promizing preliminary numerical results
- Need to still implement MICL

¹<https://cran.r-project.org/web/packages/RMixtComp/index.html>

Main references



C. Biernacki, G. Celeux, and G. Govaert.

Assessing a mixture model for clustering with the integrated completed likelihood.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(7):719–725, 2000.



C. Biernacki, G. Celeux, and G. Govaert.

Exact and Monte Carlo calculations of integrated likelihoods for the latent class model.
Journal of Statistical Planning and Inference, 140(11):2991–3002, 2011.



Christophe Biernacki and Cathy Maugis.

High-dimensional clustering.

In J.-J. Dreesbeke, G. Saporta, and C. Thomas-Agnan, editors, *Choix de modèles et agrégation*. Technip, September 2017.



D.P. Byar and S.B. Green.

The choice of treatment for cancer patients based on covariate information: Application to prostate cancer.
Bulletin du Cancer, 67:477–490, 1980.



L. A. Goodman.

Exploratory latent structure models using both identifiable and unidentifiable models.
Biometrika, 61:215–231, 1974.



R. Lebre, S. Iovleff, F. Langrognet, C. Biernacki, G. Celeux, and G. Govaert.

Rmixmod: The r package of the model-based unsupervised, supervised and semi-supervised classification mixmod library.
Journal of Statistical Software, in press, 2015.



M. Marbac and M. Sedki.

Variable selection for model-based clustering using the integrated complete-data likelihood.
Statistics and Computing, 27:1049–1063, 2017.



G.J. McLachlan and K.E. Basford.

Mixture Models: Inference and Applications to Clustering.
Marcel Dekker, New York, 1988.