Levels Merging in the Latent Class Model

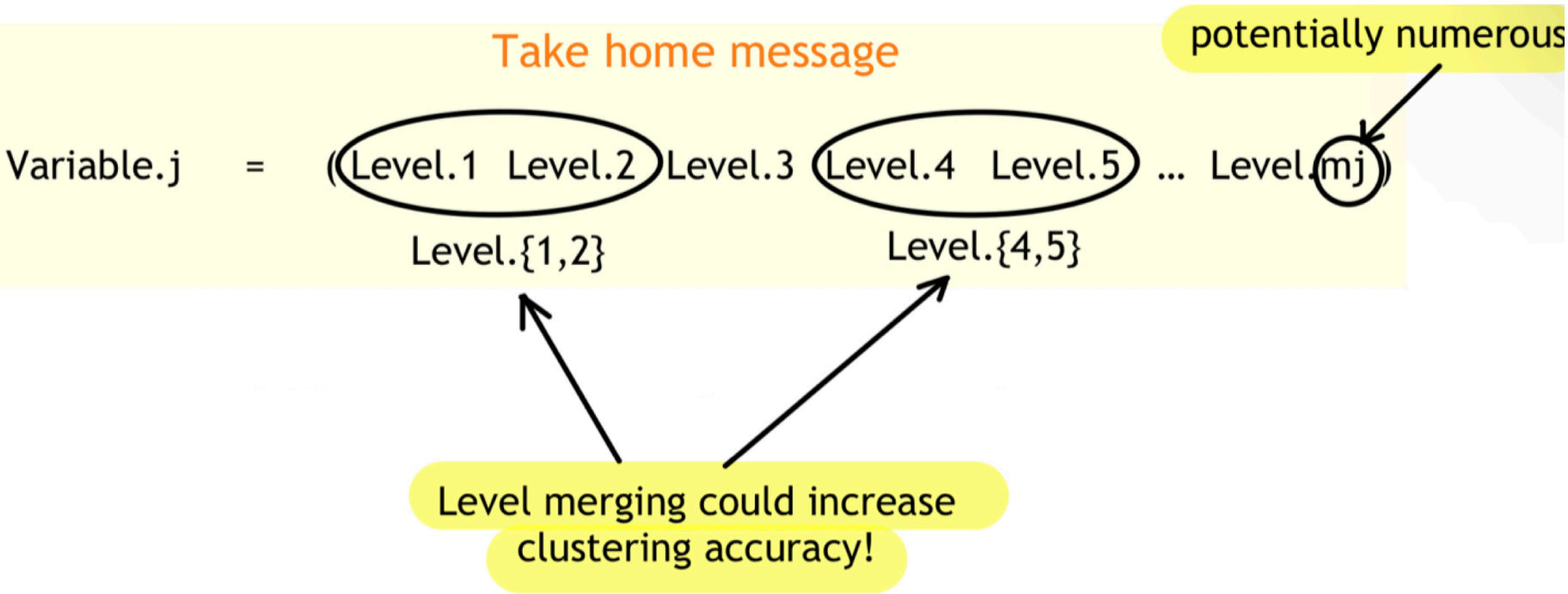
C Biernacki

Levels Merging in the Latent Class Model

C. Biernacki

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

Take home message



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 $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$of a data set $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathcal{X}^d$
- 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 jth one having m_j response levels,the standard latent class model (LCM, [5]) is used

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

- When some m_js 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 if 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 $\mathbf{w}_j = (\mathbf{w}_j^1, \ldots, \mathbf{w}_j^{m_j})$
- New model LCM-LM (*Latent Class Model by Levels Merging)*: from the full level clustering $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_d)$, 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 $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$, $\boldsymbol{\beta}_k = (\beta_k^{j\ell}; j = 1, \ldots, d; \ell = 1, \ldots, L_j)$ 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 $\mathbf{y}_{\mathbf{w}} = (\mathbf{y}_{\mathbf{w},1}, \ldots, \mathbf{y}_{\mathbf{w},n})$
- Parameter estimation: $L(\vartheta|\mathbf{w};\mathbf{x}) = L(\vartheta;\mathbf{y}_{\mathbf{w}}) + c_{\mathbf{w}}$, thus use a classical LCM EM
- Model selection:
  - ICL criterion [1, 2]: just written as a penalization of the classical LCM ICL

$$\mathrm{ICL}_{\mathbf{w}}^{\text{LCM-LM}} = \ln \mathrm{p}(\mathbf{x}, \hat{\mathbf{z}}_{\mathbf{w}}|\mathbf{w}) = \mathrm{ICL}_{\mathbf{w}}^{\text{LCM}} + \underbrace{c_{\mathbf{w}}}_{!}$$

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

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

    Then, the model $\mathbf{w}^*$ maximizing $\mathrm{MICL}_{\mathbf{w}}^{\text{LCM-LM}}$ is retained: $\mathbf{w}^* = \arg\max_{\mathbf{w}\in\mathcal{W}} \mathrm{MICL}_{\mathbf{w}}$. 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: RMixtcomp[1] 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. Droesbeke, 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. Lebret, 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.
Statitics 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.