Data Units as (Co-)Clustering Model Enlargement

Christophe Biernacki, Alexandre Lourme

Data Units as (Co-)Clustering Model Enlargement

Christophe Biernacki[(1)], Alexandre Lourme[(2)]

(1) INRIA Lille - Nord Europe (France), University Lille I, Modal Team.
(2) University of Bordeaux (France), Department of Economics.

Clustering searches a hidden structure among data. Co-Clustering searches a structure among data and variables at the same time. Mixture Models are commonly used for Clustering [9] and Co-Clustering [7]. But, most of the Mixture Models used in (Co-)Clustering are scale sensitive : changing one descriptor unit/coding may change the estimated structure(s). Instead of being a drawback such a mathematical unsustainability is an opportunity to enlarge the existing set of (Co-)Clustering models ; indeed, combining one standard model with several data units provides a new collection of (Co-)Clustering models needing few creative efforts.

Model-Based Clustering and Co-Clustering

Targets

$\mathbf{x} = (x_{i,j})$: a $n \times d$ data matrix of n individuals (rows) described by d features (columns).

Clustering Target : finding a K-class partition $\mathbf{z} = (z_{i,k}) \in \{0,1\}^{n \times K}$ of the rows : $z_{i,k} = 1$ iif $\mathbf{x}_{i,\bullet} = (x_{i,1}, \ldots, x_{i,d}) \in$ Individual Class k.

Co-Clustering Target : finding $\mathbf{z}$ and a L-block partition $\mathbf{w} = (w_{j,l}) \in \{0,1\}^{d \times L}$ of the columns : $w_{j,l} = 1$ iif $\mathbf{x}_{\bullet,j} = (x_{1,j}, \ldots, x_{n,j})' \in$ Feature Block l.

Models

Clustering Model : the pdf of $\mathbf{x}$ is the likelihood of a K-component mixture model :

$$p(\mathbf{x}; \boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f(\mathbf{x}_{i,\bullet}; \boldsymbol{\alpha}_k) \tag{1}$$

where π_k and $f(\circ; \boldsymbol{\alpha}_k)$ denote respectively the weight and the pdf of Individual Class k.

Co-Clustering Model (Latent Block Model) : the pdf of $\mathbf{x}$ is :

$$p(\mathbf{x}; \boldsymbol{\theta}) = \sum_{\mathbf{z},\mathbf{w}} \prod_{i,k} \pi_k^{z_{i,k}} \prod_{j,l} \rho_l^{w_{j,l}} \prod_{i,j,k,l} \{f(x_{i,j}; \boldsymbol{\alpha}_{k,l})\}^{z_{i,k} w_{j,l}} \tag{2}$$

where π_k is the weight of individual Class k, ρ_l the weight of Feature Block l and $f(\circ; \boldsymbol{\alpha}_{k,l})$ the pdf of one feature of Feature Block l, in Individual Class k.

Inference

Iterative procedures deriving from the EM algorithm [5] can be used to maximize (1) or (2) with respect to $\boldsymbol{\theta}$, providing the Maximum Likelihood estimate $\hat{\boldsymbol{\theta}}$.

Clustering Inference. A SEM algorithm [4] implemented into the MixtComp software[a] estimates $\boldsymbol{\theta}$ even when (i) $\mathbf{x}$ includes missing data (ii) $\mathbf{x}$ columns are mixed type (nominal, count, continuous, etc.). A SE final step (SEM without M step) provides the estimated partition $\hat{\mathbf{z}}$.

Co-Clustering Inference. A SEM-Gibbs algorithm [7, p.74] implemented into the BlockCluster software[a] estimates $\boldsymbol{\theta}$ when features are all continuous/binary/categorical/contigency. A SE final step (SEM without M step) provides the estimated partitions $\hat{\mathbf{z}}, \hat{\mathbf{w}}$.

Model Selection

Clustering Model Selection. Noting Θ the space of θ, a clustering model is a 3-tuple $\boldsymbol{m} = (K, f, \boldsymbol{\Theta})$ and BIC criterion [11] defined by : $BIC(\boldsymbol{m}) = -\log p(\mathbf{x}; \hat{\boldsymbol{\theta}}) + (dof/2)\log(n)$ enables to : (i) assess a tradeoff between $\boldsymbol{m}$ goodness of fit and parsimony (ii) compare competing models inferred on $\mathbf{x}$ (iii) select the class pdf f as the class number K.

Co-Clustering Model Selection. ICL criterion [2] computed on a co-clustering model $\boldsymbol{m} = (K, L, f, \boldsymbol{\Theta})$ is the logarithm of the integrated likelihood of the complete data $(\mathbf{x}, \mathbf{z}, \mathbf{w})$: $ICL(\boldsymbol{m}) = \log \int_{\boldsymbol{\Theta}} p(\mathbf{x}, \mathbf{z}, \mathbf{w}; \boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}$. So, ICL favours well separated Individual Classes and well separated Feature Blocks. Moreover, when all features are categorical ICL is tractable without approximation [7, p.97].

a. freely available on the MASSICCC web platform https ://modal-research-dev.lille.inria.fr

Unit Change and Model Enlargement

Data Types

Each column of $\mathbf{x}$ is a series of numbers since any M-level nominal variable can be coded as a M-dimensional vector of dummy variables.

The type of Feature j depends on the set $\mathcal{X}_j$ where $x_{1,j}, \ldots, x_{n,j}$ leave. According to $\mathcal{X}_j = \mathbb{R}$, $\mathcal{X}_j = \mathbb{N}$ or $\mathcal{X}_j = \{0,1\}$, $x_{1,j}, \ldots, x_{n,j}$ are continuous/count/binary/etc. data.

Unit Changes

Feature j can be rescaled/recoded through ϕ_j, a bijective map matching $\mathcal{X}_j$ with a space of rescaled data : $\phi_j(\mathcal{X}_j)$.

Remark. The global scaling map $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_d)$ is :
- feature wise : the rescaled series $\mathbf{x}^{\phi}_{\bullet,j}$ only depends on $\mathbf{x}_{\bullet,j}$
- homogeneous across classes : $\phi_j(x_{i,j})$ does not depend on $\mathbf{z}_{i,\bullet}$
- non parametric

Enlarging the (Co-)Clustering Model

Assuming (1) as a pdf for the rescaled data $\mathbf{x}^{\boldsymbol{\phi}} = (\phi_j(x_{i,j}))$ leads to set as a pdf for $\mathbf{x}$:

$$p^{\boldsymbol{\phi}}(\mathbf{x}, \boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{k=1}^{K} [\pi_k f(\mathbf{x}^{\boldsymbol{\phi}}_{i,\bullet}, \boldsymbol{\alpha}_k) \prod_{j \in J} |\phi'_j(x_{i,j})|] \tag{3}$$

where $\mathbf{x}^{\boldsymbol{\phi}}_{i,\bullet} = (\phi_1(x_{i,1}), \ldots, \phi_d(x_{i,d}))$ and $J \subset \{1, \ldots, d\}$ contains the indices of the continuous features.

Assuming (2) as a pdf for the rescaled data $\mathbf{x}^{\boldsymbol{\phi}}$ leads to set as a pdf for $\mathbf{x}$:

$$p^{\boldsymbol{\phi}}(\mathbf{x}, \boldsymbol{\theta}) = \sum_{\mathbf{z},\mathbf{w}} \prod_{i,k} \pi_k^{z_{i,k}} \prod_{j,l} \rho_l^{w_{j,l}} \prod_{i,j,k,l} [f(\phi_j(x_{i,j}), \boldsymbol{\alpha}_{k,l}) \gamma_{i,j}]^{z_{i,k} w_{j,l}} \tag{4}$$

where $\gamma_{i,j} = |\phi'_j(x_{i,j})|$ if Feature j is continuous and $\gamma_{i,j} = 1$ otherwise.

Any (co-)clustering model $p(\circ; \boldsymbol{\theta})$ on $\mathbf{x}^{\boldsymbol{\phi}}$ data produces a new model $p^{\boldsymbol{\phi}}(\circ; \boldsymbol{\theta})$ on $\mathbf{x}$ data. Both models $p(\circ; \boldsymbol{\theta})$ and $p^{\boldsymbol{\phi}}(\circ; \boldsymbol{\theta})$ can be compared on $\mathbf{x}$ data through BIC or ICL.

Consequences

In a clustering context.

- Cluster analysis of d-dimensional continuous data with the R package mixmod [8].
  ↳ 28 Gaussian Mixture Models among which 12 are scale dependent
- Only one alternative measurement unit is considered for each continuous feature
  ↳ $12 \times (2^d - 1)$ additional models are immediately available
- Similar enlargement can be applied to other packages involving Gaussian mixtures : bgmm [1], mclust [6], pgmm [10], etc.

In a co-clustering context.

- Co-Cluster analysis of d-dimensional binary data (0/1) into K classes and L blocks.
  ↳ a convenient model : the Latent Block Model.
- Each one of the d series is possibly recoded (permutation of 0 and 1).
  ↳ $2^d - 2$ additional models immediately available.

First Experiments and Future Prospects

A Cluster Analysis Example (using MixtComp[a] software)

· R dataset `rwm1984{COUNT}` consists on $n = 3,874$ patients of German hospitals decribed by 11 variables : 4 count, 1 categorical, 5 binary, 1 continuous.
· MixtComp model : count $\sim$ Poisson, categorical $\sim$ Multinomial, binary $\sim$ Bernoulli, continuous $\sim$ Gaussian + local independence
· Data Units. 4 maps $\boldsymbol{\phi}$ are considered rescaling some of the counts one by one : (a) none of the variables are rescaled (raw units) (b) time spent into hospital is counted in half days instead of days (c) ages are shifted (youngest age taken as origin) (d) duration of education is shifted (shortest duration taken as origin).

BIC values obtained by combining several units and class numbers on `rwm1984{COUNT}` data

Data units	best BIC	$\hat{K}$
(a) raw counts (original units)	51647	21
(b) half days into hospital	52327	20
(c) shifted ages	51833	21
(d) shifted years of education	**50044**	23

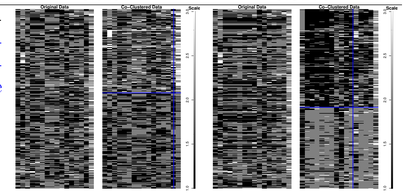
Combining the shifted duration of education with a Poisson model improves the best BIC clustering model

A Co-Cluster Analysis Example (using BlockCluster[a] software)

· The Congressional Voting Records Data Set[b] provides the response (y/n/ ?) of $n = 435$ U.S. Congressmen on 16 votes.
· Standard coding : $(1,0,0)$ for 'y', $(0,1,0)$ for 'n', $(0,0,1)$ for ' ?' on each vote
For each vote, an alternative coding : $(0,1,0)$ for 'y' and $(1,0,0)$ for 'n'
· $K = 2$ since the party (Democrat/Republican) of each congressman is known and $L = 2$

	(a) *standard coding*	(b) *best ICL coding*
ICL	5,916.13	**5,458.15**
error rate	0.4229	**0.1403**
ARI	0.0234	**0.5175**

original data (left), co-clustered data (right), Individual Class rule (horizontal) and Feature Block rule (vertical)



Recoding five votes (i) provides the best ICL model (ii) enables to retrieve more accurately the party of each congressman (iii) gives more coherence to the vote blocks

Other clustering and co-clustering examples. see [3]

Chalenging Issues

· Each scaling map ϕ_j could (i) become parametric (ii) depend on all features (iii) depend on the class. Parametric classwise and featurewise maps are considered in [12].
· All combinations : unit $\times$ model cannot be browsed exhaustively in a reasonable computational time when d is large. User friendly processes are needed to preselect a subset of features to be rescaled.

a. MASSICCC web platform https ://modal-research-dev.lille.inria.fr
b. http ://archive.ics.uci.edu/ml/machine-learning-databases/voting-records/house-votes-84.data

Références

[1] Przemyslaw Biecek, Ewa Szczurek, Martin Vingron, and Jerzy Tiuryn. The r package bgmm : Mixture modeling with uncertain knowledge. *Journal of Statistical Software*, 47(i03), 2012.

[2] Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(7) :719–725, 2000.

[3] Christophe Biernacki and Alexandre Lourme. Unifying data units and models in (co-) clustering. *Advances in Data Analysis and Classification*, pages 1–25, 2017.

[4] Gilles Celeux. The sem algorithm : a probabilistic teacher algorithm derived from the em algorithm for the mixture problem. *Computational statistics quarterly*, 2 :73–82, 1985.

[5] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[6] C Fraley et al. mclust version 4 for r : Normal mixture modeling for model-based clustering, classification, and density estimation, 2012. *University of Washington : Seattle*.

[7] Grard Govaert and Mohamed Nadif. *Co-clustering*. Wiley-IEEE Press, 2013.

[8] Rémi Lebret, Serge Iovleff, Florent Langrognet, Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Rmixmod : the r package of the model-based unsupervised, supervised and semi-supervised classification mixmod library. *Journal of Statistical Software*, 67(6) :241–270, 2015.

[9] Geoffrey McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2004.

[10] PD McNicholas, KR Jampani, AF McDaid, TB Murphy, and L Banks. pgmm : Parsimonious gaussian mixture models. *R package version*, 1(1), 2011.

[11] Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2) :461–464, 1978.

[12] Xuwen Zhu and Volodymyr Melnykov. Manly transformation in finite mixture modeling. *Computational Statistics & Data Analysis*, 2016.