# Data Units as (Co-)Clustering Model Enlargement <br> Université <br> de Lille <br> Christophe Biernacki ${ }^{(1)}$, Alexandre Lourme ${ }^{(2)}$ <br> 1) INRIA Lille - Nord Europe (France), University Lille I, Modal Tean <br> Université <br> (2) University of Bordeaux (France), Department of Economics <br> de BORDEAUX 

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-)Clusterin 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 on are scale sensitive : changing one descriptor unit/ coding may change the estimated structure(s). Instead of being a drous (Co-) Clustering models needing few creative efforts.
standard model with several data units provides a new collection of

Model-Based Clustering and Co-Clustering

Targets
$\mathbf{x}=\left(x_{i, j}\right):$ a $n \times d$ data matrix of $n$ individuals (rows) described by $d$ features (columns). $\frac{\text { Clustering Target }}{z_{i, k}=1 \text { iif } \mathbf{x}_{i, 0}}:$ finding a $K$-class partition $\mathbf{z}=\left(z_{i, k}\right) \in\{0,1\}^{n \times K}$ of the rows $\overline{z_{i, k}}=1$ if $\mathbf{x}_{i, \bullet}=\left(x_{i, 1}, \ldots, x_{i, d}\right) \in$ Individual Class $k$
Co-Clustering Target : finding $\mathbf{z}$ and a $L$-block partition $\mathbf{w}=\left(w_{j, l}\right) \in\{0,1\}^{d \times L}$ of the columns : $w_{j, l}=1$ iif $\mathbf{x}_{\bullet}, j=\left(x_{1, j}, \ldots, x_{n, j}\right)^{\prime} \in$ Feature Block $l$.
Models
Clustering Model : the pdf of $\mathbf{x}$ is the likelihood of a $K$-component mixture mode :

$$
p(\mathbf{x} ; \boldsymbol{\theta})=\prod_{i=1}^{n} \sum_{k=1}^{K} \pi_{k} f\left(\mathbf{x}_{i, \bullet} ; \alpha_{k}\right)
$$

where $\pi_{k}$ and $f\left(\circ ; \boldsymbol{\alpha}_{k}\right)$ denote respectively the weight and the pdf of Individual Class $k$
Co-Clustering Model (Latent Block Model) : the pdf of $\mathbf{x}$ is

$$
p(\mathbf{x} ; \boldsymbol{\theta})=\sum_{\mathbf{z}, \mathbf{w}} \prod_{i, k} \pi_{k}^{z_{i, k}} \prod_{j, l} \rho_{l}^{w_{j, l}} \prod_{i, j, k, l}\left\{f\left(x_{i, j} ; \boldsymbol{\alpha}_{k, l}\right)\right\}^{z_{i, k} w_{j, l}}
$$

where $\pi_{k}$ is the weight of individual Class $k, \rho_{l}$ the weight of Feature Block $l$ and $f\left(\circ ; \boldsymbol{\alpha}_{k, l}\right)$ 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 ${ }^{\text {a }}$ estimates $\boldsymbol{\theta}$ even when (i) $\mathbf{x}$ includes missing data (ii) $\mathbf{x}$ columns are mixed type (nominal count, contin
 into the BlockCluster softwarea ${ }^{\text {a }}$ estimates $\boldsymbol{\theta}$ when features are all conti-
nuous/binary/categorical/contigency. A SE final step (SEM without M step) provide nuous/binary/categorical/contigency. A SE final step (SEM without M step) provides he estimated partitions $\hat{\mathbf{z}}_{\mathbf{W}} \hat{\mathbf{W}}$.

Model Selection
Clustering Model Selection. Noting $\Theta$ the space of $\theta$ a a clustering model is a 3 -tuple $\boldsymbol{m}=$ $(K, f, \boldsymbol{\Theta})$ and $B I C$ criterion $[11]$ defined by : $B I C(\boldsymbol{m})=-\log p(\mathbf{x} ; \hat{\boldsymbol{\theta}})+(d o f / 2) \log (n)$ competing models inferred on $\mathbf{x}$ (iii) select the class pdff $f$ as the class number $K$.
$\frac{\text { Co-Clustering Model Selection. } I C L \text { criterion [2] computed on a co-clustering mode }}{m=(K, I f \Theta)}$ $(\mathbf{x}, \mathbf{z}, \mathbf{w}): I C L(\boldsymbol{m})=\log \int_{\boldsymbol{\Theta}} p(\mathbf{x}, \mathbf{z}, \mathbf{w} ; \boldsymbol{\theta}) \mathrm{d} \boldsymbol{\theta}$. So, ICL favours well separated Individual Classes and well separated Feature Blocks. Moreover, when all features are categorical $I C L$ is tractable without approximation $[7$, p. 97 .

## Références

[1] Przemyslaw Biecek, Ewa Szczurek, Martin Vingron, and Jerzy Tiuryn. The r packag 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 ctustering with the integrated completed likel hood Pattern Analysis and Machine for clustering with the integrated completed likelihood. Pattern Analysis and Machine
Intelligence, IEEE Transactions on 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 Classifcation clue,


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

eature $j$ can be rescaled/recoded through $\phi_{j}$, a bijective map matching $\mathcal{X}_{j}$ with a spac of rescaled data : $\phi_{j}\left(\mathcal{X}_{j}\right)$
Remark. The global scaling map $\boldsymbol{\phi}=\left(\phi_{1}, \ldots, \phi_{d}\right)$ is :
feature wise : the rescaled series $\mathbf{x}_{\bullet}^{\dagger}$, only depends on $\mathbf{x}_{\bullet}$.
non parametric
Enlarging the (Co-)Clustering Model
Assuming (1) as a pdf for the rescaled data $\mathbf{x}^{\boldsymbol{\phi}}=\left(\phi_{j}\left(x_{i, j}\right)\right)$ leads to set as a pdf for $\mathbf{x}$ :

$$
p^{\phi}(\mathbf{x}, \boldsymbol{\theta})=\prod_{i=1}^{n} \sum_{k=1}^{K}\left[\pi_{k} f\left(\mathbf{x}_{i, \boldsymbol{,}}^{\phi}, \boldsymbol{\alpha}_{k}\right) \prod_{j \in J}\left|\phi_{j}^{\prime}\left(x_{i, j}\right)\right|\right]
$$

here $\mathbf{x}_{i, \mathbf{\bullet}}^{\phi}=\left(\phi_{1}\left(x_{i, 1}, \ldots, \phi_{d}\left(x_{i, d}\right)\right)\right.$ and $J \subset\{1, \ldots, d\}$ countains the indices of the ontimuois features.
sssuming (2) as a pdf for the rescaled data $\mathbf{x}^{\boldsymbol{\phi}}$ leads to set as a pdf for $\mathbf{x}$

$$
p^{\phi}(\mathbf{x}, \boldsymbol{\theta})=\sum_{\mathbf{z}, \mathbf{w}} \prod_{i, k} \pi_{k}^{z_{i, k}} \prod_{j, l} \rho_{l}^{w_{j, l}} \prod_{i, j, k, l}\left[f\left(\phi_{j}\left(x_{i, j}\right), \boldsymbol{\alpha}_{k, l}\right) \gamma_{i, j} \bar{z}^{z_{i, k} w_{j, l}}\right.
$$

here $\gamma_{i, j}=\left|\phi_{j}^{\prime}\left(x_{i, j}\right)\right|$ 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. 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 $p(0 ; \boldsymbol{\theta})$ and $p^{\phi}(0 ; \boldsymbol{\theta})$ can be compared on $\mathbf{x}$ data through $B I C$ or $I C L$. Consequences
In a clustering context.
Cluster analysis of $d$-dimensional continuous data with the $R$ package mixmod $[8$.
Cluster analysis of $d$-dimensional continuous data with the $R$ package mixmod
$\square 28$ Gaussian Mixture Models among which 12 are scale dependent Only one alternative measurement unit is considered for each continuous feature enlargement simm [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. corverient model: the Latent Block Mode.
Each one of the $d$ series is possibly recoded (permutation of 0 and 1 .

- 2 additional models immediately available


## em al 1985.

5) Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood fron incomplete data via the en 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 Ioveff, Florent Langrognet, Christophe Biernacki, Gilles Celeux and Gerrard Govaert. Rmixmod : the r packase of the model-based unsupervised, su-


A Cluster Analysis Example (using MixtComp ${ }^{\text {a }}$ software) - R dataset rwm $1984\{$ 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 $\phi$ 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 rum1984\{COUNT\} date Data units best BIC $\hat{K}$ $\begin{array}{lllll}\text { (a) raw counts (original units) } & 51647 & 21 & \text { Combining the shifted } \\ \text { (b) half days into hospital } & 52327 & 20 & \text { durbiton } \\ \text { (c) } & 53\end{array}$ $\begin{array}{lllll}\text { (b) half days into hospital } & 52327 & 20 & \text { duration of education with a } \\ \text { (c) shifted ages } & 51833 & 21 & \text { Poisson model improves the }\end{array}$ (d) shifted years of education $\begin{array}{lllll}50044 & 23 & \left.\begin{array}{l}\text { Poisson model improves } \\ \text { best BIC clustering model }\end{array}\right]\end{array}$

A Co-Cluster Analysis Example (using BlockCluster ${ }^{\text {a }}$ software)
-The Congressional Voting Records Data Set ${ }^{\mathrm{b}}$ provides the response ( $\mathrm{y} / \mathrm{n} /$ ?) 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
$\begin{array}{lll} & \text { (a) standard coding } & \text { (b) best ICL coding }\end{array}$


Recoding five votes (i) provides the best $I C L$ 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 $\phi_{j}$ could (i) become parametric (ii) depend on all features (iii) depend on the class. Parametric classwise and featurewise maps are considered in [12]. putational time when $d$ is large. User friendly processes are needed to preselect a subset

pervised 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.
f10] 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.

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

