MASSICCC: A SaaS Platform for Clustering and Co-Clustering of Mixed Data

https://massiccc.lille.inria.fr/

C. Biernacki
with B. Auder, G. Celeux, J. Demont, F. Langrognet, V. Kubicki, C. Poli, J. Renault, S. Iovleff

21-22 June 2018, Workshop MixStatSeq, Paris
“Mixture models: Theory and applications"
| 1 | Introduction |
| 2 | Model-based clustering |
| 3 | Mixmod in MASSICCC |
| 4 | MixtComp in MASSICCC |
| 5 | BlockCluster in MASSICCC |
| 6 | Conclusion |
MASSICCC?

massiccc.lille.inria.fr

Massive Clustering with Cloud Computing
Clustering of heterogeneous data with missing values. Hosted in the cloud. No installation or configuration required. Upload your data, and get results straight away.

Developed by Inria

SaaS: Software as a Service
A high quality and easy to use web platform where are transferred mature research clustering (and more) software towards (non academic) professionals
Here is the computer you need!
Clustering?

Detect hidden structures in data sets

- low income
- average income
- high income

1st MCA axis
2nd MCA axis

Introduction
Model-based clustering
Mixmod in MASSICCC
MixtComp in MASSICCC
BlockCluster
Conclusion
Clustering everywhere

Decision trees, regression, and cluster analysis continue to form a triad of core algorithms for most data miners. This has been very consistent over time. However, a wide variety of algorithms are being used.

Consultants are more likely to use Ensemble Models:

<table>
<thead>
<tr>
<th></th>
<th>Corporate</th>
<th>Consultants</th>
<th>Academic</th>
<th>NGO / Gov't</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Trees</td>
<td>69%</td>
<td>32%</td>
<td>20%</td>
<td>10%</td>
</tr>
<tr>
<td>Regression</td>
<td>68%</td>
<td>31%</td>
<td>19%</td>
<td>7%</td>
</tr>
<tr>
<td>Cluster Analysis</td>
<td>32%</td>
<td>27%</td>
<td>17%</td>
<td>6%</td>
</tr>
<tr>
<td>Time Series</td>
<td>26%</td>
<td>26%</td>
<td>15%</td>
<td>5%</td>
</tr>
<tr>
<td>Neural Nets</td>
<td>33%</td>
<td>25%</td>
<td>21%</td>
<td>8%</td>
</tr>
<tr>
<td>Factor Analysis</td>
<td>27%</td>
<td>21%</td>
<td>17%</td>
<td>6%</td>
</tr>
<tr>
<td>Text Mining</td>
<td>22%</td>
<td>21%</td>
<td>16%</td>
<td>5%</td>
</tr>
<tr>
<td>Association Rules</td>
<td>22%</td>
<td>21%</td>
<td>15%</td>
<td>5%</td>
</tr>
<tr>
<td>Ensemble Models</td>
<td>21%</td>
<td>18%</td>
<td>14%</td>
<td>4%</td>
</tr>
<tr>
<td>Support Vector</td>
<td>19%</td>
<td>14%</td>
<td>12%</td>
<td>3%</td>
</tr>
<tr>
<td>Bayesian</td>
<td>21%</td>
<td>14%</td>
<td>12%</td>
<td>5%</td>
</tr>
<tr>
<td>Anomaly Detection</td>
<td>16%</td>
<td>13%</td>
<td>10%</td>
<td>3%</td>
</tr>
<tr>
<td>Survival Analysis</td>
<td>14%</td>
<td>12%</td>
<td>9%</td>
<td>3%</td>
</tr>
<tr>
<td>Rule Induction</td>
<td>13%</td>
<td>11%</td>
<td>8%</td>
<td>3%</td>
</tr>
<tr>
<td>Social Network Analysis</td>
<td>12%</td>
<td>9%</td>
<td>6%</td>
<td>2%</td>
</tr>
<tr>
<td>Genetic Algorithms</td>
<td>11%</td>
<td>8%</td>
<td>6%</td>
<td>2%</td>
</tr>
<tr>
<td>Link Analysis</td>
<td>9%</td>
<td>7%</td>
<td>5%</td>
<td>2%</td>
</tr>
<tr>
<td>Uplift Modeling</td>
<td>9%</td>
<td>8%</td>
<td>6%</td>
<td>2%</td>
</tr>
<tr>
<td>MARS</td>
<td>8%</td>
<td>7%</td>
<td>5%</td>
<td>2%</td>
</tr>
</tbody>
</table>

Consultants and corporate data miners are more likely to use Uplift Modeling:

<table>
<thead>
<tr>
<th></th>
<th>Corporate</th>
<th>Consultants</th>
<th>Academic</th>
<th>NGO / Gov't</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Trees</td>
<td>10%</td>
<td>12%</td>
<td>4%</td>
<td>5%</td>
</tr>
<tr>
<td>Regression</td>
<td>21%</td>
<td>27%</td>
<td>20%</td>
<td>18%</td>
</tr>
<tr>
<td>Cluster Analysis</td>
<td>21%</td>
<td>20%</td>
<td>17%</td>
<td>15%</td>
</tr>
<tr>
<td>Time Series</td>
<td>21%</td>
<td>19%</td>
<td>17%</td>
<td>15%</td>
</tr>
<tr>
<td>Neural Nets</td>
<td>21%</td>
<td>19%</td>
<td>17%</td>
<td>15%</td>
</tr>
<tr>
<td>Factor Analysis</td>
<td>20%</td>
<td>18%</td>
<td>16%</td>
<td>14%</td>
</tr>
<tr>
<td>Text Mining</td>
<td>20%</td>
<td>18%</td>
<td>16%</td>
<td>14%</td>
</tr>
<tr>
<td>Association Rules</td>
<td>19%</td>
<td>18%</td>
<td>16%</td>
<td>14%</td>
</tr>
<tr>
<td>Ensemble Models</td>
<td>19%</td>
<td>18%</td>
<td>16%</td>
<td>14%</td>
</tr>
<tr>
<td>Support Vector</td>
<td>18%</td>
<td>17%</td>
<td>15%</td>
<td>13%</td>
</tr>
<tr>
<td>Bayesian</td>
<td>18%</td>
<td>17%</td>
<td>15%</td>
<td>13%</td>
</tr>
<tr>
<td>Anomaly Detection</td>
<td>17%</td>
<td>16%</td>
<td>14%</td>
<td>12%</td>
</tr>
<tr>
<td>Survival Analysis</td>
<td>16%</td>
<td>15%</td>
<td>13%</td>
<td>11%</td>
</tr>
<tr>
<td>Rule Induction</td>
<td>15%</td>
<td>14%</td>
<td>12%</td>
<td>10%</td>
</tr>
<tr>
<td>Social Network Analysis</td>
<td>14%</td>
<td>13%</td>
<td>11%</td>
<td>9%</td>
</tr>
<tr>
<td>Genetic Algorithms</td>
<td>13%</td>
<td>12%</td>
<td>10%</td>
<td>8%</td>
</tr>
<tr>
<td>Link Analysis</td>
<td>12%</td>
<td>11%</td>
<td>9%</td>
<td>7%</td>
</tr>
<tr>
<td>Uplift Modeling</td>
<td>11%</td>
<td>10%</td>
<td>8%</td>
<td>6%</td>
</tr>
<tr>
<td>MARS</td>
<td>10%</td>
<td>9%</td>
<td>7%</td>
<td>5%</td>
</tr>
</tbody>
</table>

Question: What algorithms/analytic methods do you TYPICALLY use? (Select all that apply)

1Rexer Analytics’s Annual Data Miner Survey is the largest survey of data mining, data science, and analytics professionals in the industry (survey of 2011)
Data sets structure
Large data sets

An opportunity for detecting weak signal

Low frequent classes appear when $n$ grows
Todays features: full mixed/missing

- **Categorical**
  - Marital status: married
  - Drink preference: beer > soda > water

- **Integer**
  - Children: 3

- **Missing**
  - Size (m): ?

- **Ordinal**
  - Intelligence: low

- **Continuous**
  - Weight (kg): 119.5

- **Functional**
  - Drink consumption

- **Graph**
  - Family

And so on...
Notations

- **Data:** $n$ individuals: $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} = \{\mathbf{x}^O, \mathbf{x}^M\}$ in a space $\mathcal{X}$ of dimension $d$
  - Observed individuals $\mathbf{x}^O$
  - Missing individuals $\mathbf{x}^M$

- **Aim:** estimation of the partition $\mathbf{z}$ and the number of clusters $K$
  Partition in $K$ clusters $G_1, \ldots, G_K$: $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$, $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})'$

$$\mathbf{x}_i \in G_k \iff z_{ih} = \mathbb{I}\{h=k\}$$

### Mixed, missing, uncertain

<table>
<thead>
<tr>
<th>Individuals $\mathbf{x}$</th>
<th>Partition $\mathbf{z}$</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>0.5 red</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>0.3 green</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>0.3 red, green</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>0.9 red</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>[0.25 0.45] red</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>↓</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>continuous</td>
<td>continuous</td>
<td>categorical</td>
</tr>
</tbody>
</table>

- [0.25 0.45]
Outline

1. Introduction

2. Model-based clustering

3. Mixmod in MASSICCC

4. MixtComp in MASSICCC

5. BlockCluster in MASSICCC

6. Conclusion
Parametric mixture model

- **Parametric assumption:**
  \[ p_k(x_1) = p(x_1; \alpha_k) \]
  
  thus
  \[
  p(x_1) = p(x_1; \theta) = \sum_{k=1}^{K} \pi_k p(x_1; \alpha_k)
  \]

- **Mixture parameter:**
  \[ \theta = (\pi, \alpha) \text{ with } \alpha = (\alpha_1, \ldots, \alpha_K) \]

- **Model:** it includes both the family \( p(\cdot; \alpha_k) \) and the number of groups \( K \)
  \[
  \mathbf{m} = \{ p(x_1; \theta) : \theta \in \Theta \}
  \]

  The number of free continuous parameters is given by
  \[ \nu = \dim(\Theta) \]

Clustering becomes a well-posed problem...
The clustering process in mixtures

1. Estimation of $\theta$ by $\hat{\theta}$
2. Estimation of the conditional probability that $x_i \in G_k$

$$t_{ik}(\hat{\theta}) = p(Z_{ik} = 1|X_i = x_i; \hat{\theta}) = \frac{\hat{\pi}_k p(x_i; \hat{\alpha}_k)}{p(x_i; \hat{\theta})}$$

3. Estimation of $z_i$ by maximum a posteriori (MAP)

$$\hat{Z}_{ik} = \mathbb{I}\{k = \arg\max_{h=1,…,K} t_{ih}(\hat{\theta})\}$$

4. Model selection: BIC, ICL, …
Only continuous features: 14 models on $\Sigma_k$

$$\Sigma_k = \lambda_k \cdot D_k \cdot A_k \cdot D'_k$$

- volume
- orientation
- shape
Only categorical variables: latent class model

- **Categorical variables**: \( d \) variables with \( m_j \) modalities each, \( x_i^j \in \{0, 1\}^{m_j} \) and

\[
x_i^{jh} = 1 \iff \text{variable } j \text{ of } x_i \text{ takes level } h
\]

- **Conditional independence**:

\[
p(x_i; \alpha_k) = \prod_{j=1}^{d} \prod_{h=1}^{m_j} (\alpha_{jh}^k)^{x_i^{jh}}
\]

and

\[
\alpha_{jh}^k = p(x_i^{jh} = 1 \mid z_{ik} = 1)
\]

with \( \alpha_k = (\alpha_{jh}^k; j = 1, \ldots, d; h = 1, \ldots, m_j) \)
Mixing continuous and categorical data: full local independence

Combine continuous and categorical data

\[ x_1 = (x_1^{cont}, x_1^{cat}) \]

The proposed solution is to mixed both types by inter-type conditional independence

\[ p(x_1; \alpha_k) = p(x_1^{cont}; \alpha_k^{cont}) \times p(x_1^{cat}; \alpha_k^{cat}) \]

In addition, for symmetry between types, intra-type conditional independence

Only need to define the univariate pdf for each variable type!

- **Continuous**: Gaussian
- **Categorical**: multinomial
Estimation of $\theta$ by complete-likelihood

Maximize the complete-likelihood over $(\theta, z)$

$$\ell_c(\theta; x, z) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \ln \{ \pi_k p(x_i; \alpha_k) \}$$

- **Equivalent** to traditional methods
  
<table>
<thead>
<tr>
<th>Metric</th>
<th>$M = I$</th>
<th>$M$ free</th>
<th>$M_k$ free</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian model</td>
<td>$[\pi \lambda I]$</td>
<td>$[\pi \lambda C]$</td>
<td>$[\pi \lambda_k C_k]$</td>
</tr>
</tbody>
</table>

- **Bias** of $\hat{\theta}$: heavy if poor separated clusters
- Associated optimization algorithm: **CEM** (see later)
- **CEM** with $[\pi \lambda I]$ is strictly equivalent to $K$-means
- **CEM** is simple et fast (convergence with few iterations)
Estimation of $\theta$ by observe-likelihood

Maximize the observe-likelihood on $\theta$

$$\ell(\theta; x) = \sum_{i=1}^{n} \ln p(x_i; \theta)$$

- Convergence of $\hat{\theta}$, asymptotic efficiency, asymptotically unbiased
- General algorithm for missing data: EM
- EM is simple but slower than CEM
- Interpretation: it is a kind of fuzzy clustering
**Principle of EM and CEM**

- **Initialization:** $\theta^0$
- **Iteration $n^q$:**
  - **Step E:** estimate probabilities $t^q = \{t_{ik}(\theta^q)\}$
  - **Step C:** classify by setting $t^q = \text{MAP}(\{t_{ik}(\theta^q)\})$
  - **Step M:** maximize $\theta^{q+1} = \text{arg max}_\theta \ell_c(\theta; x, t^q)$
- **Stopping rule:** iteration number or criterion stability

**Properties**
- $\oplus$: simplicity, monotony, low memory requirement
- $\ominus$: local maxima (depends on $\theta^0$), linear convergence (EM)
Prostate cancer data (without mixing data)

- **Individuals:** $n = 475$ patients with prostatic cancer grouped on clinical criteria into two Stages 3 and 4 of the disease.
- **Variables:** $d = 12$ pre-trial variates were measured on each patient, composed by eight continuous variables (age, weight, systolic blood pressure, diastolic blood pressure, serum haemoglobin, size of primary tumour, index of tumour stage and histologic grade, serum prostatic acid phosphatase) and four categorical variables with various numbers of levels (performance rating, cardiovascular disease history, electrocardiogram code, bone metastases).
- **Model:** cond. indep. $p(x_1; \alpha_k) = p(x_1; \alpha_k^{cont}) \cdot p(x_1; \alpha_k^{cat})$
Prostate cancer data (without missing data)

<table>
<thead>
<tr>
<th>Variables</th>
<th>Continuous</th>
<th>Categorical</th>
<th>Mixed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error (%)</td>
<td>9.46</td>
<td>47.16</td>
<td>8.63</td>
</tr>
<tr>
<td>True \ estimated group</td>
<td>1 2</td>
<td>1 2</td>
<td>1 2</td>
</tr>
<tr>
<td>Stage 3</td>
<td>247 26</td>
<td>142 131</td>
<td>252 21</td>
</tr>
<tr>
<td>Stage 4</td>
<td>19 183</td>
<td>120 82</td>
<td>20 182</td>
</tr>
</tbody>
</table>

### Continuous data

![PCA plot for continuous data](image1.png)

### Categorical data

![MCA plot for categorical data](image2.png)
Continuous data
Continuous data
Mixed data

<table>
<thead>
<tr>
<th>Model</th>
<th>Criterion</th>
<th>Nb Clusters</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heterogeneous_pk_Ekjh_Lk_Bk</td>
<td>ICL(23198.3)</td>
<td>2</td>
<td>No error</td>
</tr>
<tr>
<td>Heterogeneous_pk_Ekjh_Lk_Bk</td>
<td>ICL(23327.2)</td>
<td>3</td>
<td>No error</td>
</tr>
<tr>
<td>Heterogeneous_pk_Ekjh_Lk_Bk</td>
<td>ICL(23402.6)</td>
<td>4</td>
<td>No error</td>
</tr>
<tr>
<td>Heterogeneous_pk_Ekjh_Lk_Bk</td>
<td>ICL(23464.3)</td>
<td>5</td>
<td>No error</td>
</tr>
<tr>
<td>Heterogeneous_pk_Ekjh_Lk_Bk</td>
<td>ICL(23762.2)</td>
<td>1</td>
<td>No error</td>
</tr>
</tbody>
</table>

**Variables**

- Model Criterion
  - This chart represents the ICL value for each model that was built. The lower the value (close to 0), the better the model.
  - Criterion: ICL

**Variable Importance**

This chart represents the discriminating level of each variable. A high value (close to one) means that the variable is highly discriminating. A low value (close to zero) means that the variable is poorly discriminating.

**Variable Parameters**

- AP (Gaussian)
  - Class 1: mean: 3.9268, sigma: 2.8781
  - Class 2: mean: 1.6476, sigma: 0.2509
Outline

1. Introduction
2. Model-based clustering
3. Mixmod in MASSICCC
4. MixtComp in MASSICCC
5. BlockCluster in MASSICCC
6. Conclusion
Full mixed data: conditional independence everywhere

The aim is to combine continuous, categorical, integer data, ordinal, ranking and functional data

\[ \mathbf{x}_1 = (x_1^{cont}, x_1^{cat}, x_1^{int}, \ldots) \]

The proposed solution is to mixed all types by inter-type conditional independence

\[ p(\mathbf{x}_1; \alpha_k) = p(x_1^{cont}; \alpha_k^{cont}) \times p(x_1^{cat}; \alpha_k^{cat}) \times p(x_1^{int}; \alpha_k^{int}) \times \ldots \]

In addition, for symmetry between types, intra-type conditional independence

Only need to define the univariate pdf for each variable type!

- **Continuous**: Gaussian
- **Categorical**: multinomial
- **Integer**: Poisson
- ...
Missing data: MAR assumption and estimation

Assumption on the missingness mechanism

**Missing At Random (MAR):** the probability that a variable is missing does not depend on its own value given the observed variables.

Observed log-likelihood...

$$
\ell(\theta; \mathbf{x}^O) = \sum_{i=1}^{n} \log \left( \sum_{k=1}^{K} \pi_k p(x_i^O; \alpha_k) \right) = \ln \left[ \sum_{k=1}^{K} \pi_k \int_{x_i^M} p(x_i^O, x_i^M; \alpha_k) dx_i^M \right]
$$

MAR assumption
A SEM algorithm to estimate $\theta$ by maximizing the observed-data log-likelihood

- **Initialisation:** $\theta^{(0)}$
- **Iteration nb $q$:**
  - **E-step:** compute conditional probabilities $p(x^M, z|x^0; \theta^{(q)})$
  - **S-step:** draw $(x^{M(q)}, z^{(q)})$ from $p(x^M, z|x^0; \theta^{(q)})$
  - **M-step:** maximize $\theta^{(q+1)} = \arg\max_\theta \ln p(x^O, x^{M(q)}, z^{(q)}; \theta)$
- **Stopping rule:** iteration number

### Properties: simpler than EM and interesting properties!

- Avoid possibly difficult E-step in an EM
- Classical M steps
- Avoids local maxima
- The mean of the sequence $(\theta^{(q)})$ approximates $\hat{\theta}$
- The variance of the sequence $(\theta^{(q)})$ gives confidence intervals

---

4 MixtComp software on the MASSICCC platform: https://massiccc.lille.inria.fr/
Prostate cancer data (with missing data)\(^5\)

- **Individuals**: 506 patients with prostatic cancer grouped on clinical criteria into two Stages 3 and 4 of the disease
- **Variables**: \(d = 12\) pre-trial variates were measured on each patient, composed by eight continuous variables (age, weight, systolic blood pressure, diastolic blood pressure, serum haemoglobin, size of primary tumour, index of tumour stage and histologic grade, serum prostatic acid phosphatase) and four categorical variables with various numbers of levels (performance rating, cardiovascular disease history, electrocardiogram code, bone metastases)
- **Some missing data**: 62 missing values (\(\approx 1\%\))

We forget the classes (Stages of the disease) for performing clustering

**Questions**

- How many clusters?
- Which partition?

### Data upload without preprocessing

#### Preview

<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th>Wt</th>
<th>PF</th>
<th>HX</th>
<th>SBP</th>
<th>DBP</th>
<th>EKG</th>
<th>HG</th>
<th>SZ</th>
<th>SG</th>
<th>AP</th>
<th>BM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>75</td>
<td>76</td>
<td>1</td>
<td>1</td>
<td>15</td>
<td>9</td>
<td>5</td>
<td>138</td>
<td>1.4142</td>
<td>8</td>
<td>1.0986</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>54</td>
<td>116</td>
<td>1</td>
<td>1</td>
<td>13</td>
<td>7</td>
<td>4</td>
<td>146</td>
<td>6.4807</td>
<td>?</td>
<td>1.9459</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>69</td>
<td>102</td>
<td>1</td>
<td>2</td>
<td>14</td>
<td>8</td>
<td>5</td>
<td>134</td>
<td>1.7321</td>
<td>9</td>
<td>1.0986</td>
<td>1</td>
</tr>
</tbody>
</table>
Run clustering analysis

Parameters

- Title: Run demo on cancer data set
- Data File: MixtComp-Example.csv
- Package: MixMod, MixComp, BlockCluster
- Function: Cluster
- Labels Column: 
- Cluster Groups: 1-7

Create
It is running on the (Inria) cloud...
Several quick result overviews... without post-processing
Variable significance on global partition

- **Variable Importance**

  This chart represents the **discriminating** level of each variable. A high value (close to one) means that the variable is highly discriminating. A low value (close to zero) means that the variable is poorly discriminating. Click on one of the bars to display the distribution of this variable and, to also display the similarities between this variable and all the others. The color of the bars reflects the similarities between all the variables and the selected variable.

  Sort Variables: ↓↑

  + similarity between variables
Variable “Age” difference between clusters

This chart summarizes the distribution of the selected variable.

<table>
<thead>
<tr>
<th>Class</th>
<th>Mean (Age)</th>
<th>Sigma (Age)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1</td>
<td>71.534</td>
<td>6.760</td>
</tr>
<tr>
<td>Class 2</td>
<td>71.313</td>
<td>7.463</td>
</tr>
</tbody>
</table>
Variable “SG” difference between clusters

Variable Parameters

This chart summarizes the distribution of the selected variable.

SG

Boxplot of the distribution per class for SG

SG (Gaussian)

Hide model parameters

<table>
<thead>
<tr>
<th>Class</th>
<th>1</th>
<th>Class</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>8.940</td>
<td>mean</td>
<td>12.087</td>
</tr>
<tr>
<td>sigma</td>
<td>1.154</td>
<td>sigma</td>
<td>1.405</td>
</tr>
</tbody>
</table>
Variable “BM” difference between clusters

This chart summarizes the distribution of the selected variable.

BM (Multinomial)

Hide model parameters

<table>
<thead>
<tr>
<th></th>
<th>Class 1</th>
<th>Class 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>scatter</td>
<td>[0.993, 0.007]</td>
<td>[0.633, 0.367]</td>
</tr>
</tbody>
</table>
Individual cluster separation (with the cluster weight)
Two strategies in competition

- **Strategy “mice$^6$ + MixtComp”:** MixtComp on the dataset completed by mice

  ```r
  > data.imp=mice(data)
  > data.comp.mice=complete(data.imp)
  ```

- **Strategy “full MixtComp”:** MixtComp on the observed (no completed) dataset

<table>
<thead>
<tr>
<th>Strategy</th>
<th>mice + MixtComp</th>
<th>full MixtComp</th>
</tr>
</thead>
<tbody>
<tr>
<td>% misclassified</td>
<td>12.8</td>
<td>8.1</td>
</tr>
</tbody>
</table>

---

$^6$http://cran.r-project.org/web/packages/mice/mice.pdf
Choosing $K$ with the ICL criterion

mice + MixtComp
$\hat{K} = 7$

full MixtComp
$\hat{K} = 2$

...may lose some cluster information when imputation before clustering
Scoring cancer data following the clustering task
The Kneading dataset comes from Danone Vitapole Paris Research Center and concerns the quality of cookies and the relationship with the flour kneading process\textsuperscript{7}. There are 115 different flours for which the dough resistance is measured during the kneading process for 480 seconds. One obtains 115 kneading curves observed at 241 equispaced instants of time in the interval $[0; 480]$. The 115 flours produce cookies of different quality: 50 of them have produced cookies of good quality, 25 produced medium quality and 40 low quality.

\textsuperscript{7}Léveder et al, 04
Upload curves data
Run a clustering task with three clusters

Parameters

- Title: Clustering of cookies into three clusters
- Data File: Functional-Example.csv
- Package: MixMod
- Function: Cluster
- Cluster Groups: 3
- Variable Params: Yes

Create
Overview of the three clusters of cookies
Outline

1. Introduction
2. Model-based clustering
3. Mixmod in MASSICCC
4. MixtComp in MASSICCC
5. BlockCluster in MASSICCC
6. Conclusion
High-dimensional (HD) data

Bias/variance in HD: reduce variance, accept bias

A two-component $d$-variate Gaussian mixture with intra-dependency:

$$\pi_1 = \pi_2 = \frac{1}{2}, \quad \mathbf{X}_1 | z_{11} = 1 \sim \mathcal{N}_d(\mathbf{0}, \Sigma), \quad \mathbf{X}_1 | z_{12} = 1 \sim \mathcal{N}_d(\mathbf{1}, \Sigma)$$

- Each variable provides equal and own separation information
- Theoretical error decreases when $d$ grows: $\text{err}_{\text{theo}} = \Phi(-\|\mu_2 - \mu_1\|\Sigma^{-1}/2)$
- Empirical error rate with the (true) intra-correlated model worse with $d$
- Empirical error rate with the (false) intra-independent model better with $d$!
Some alternatives for reducing variance

- Dimension reduction in non-canonical space (PCA-like typically)
- Dimension reduction in the canonical space (variable selection)
- Model parsimony in the initial HD space (constraints on model parameters)

But which kind of parsimony?
- Remember that clustering is a way for dealing with large $n$
- Why not reusing this idea for large $d$?

Co-clustering
It performs parsimony of row clustering through variable clustering
From clustering to co-clustering

[Govaert, 2011]
Notations

- $z_i$: the cluster of the row $i$
- $w_j$: the cluster of the column $j$
- $(z_i, w_j)$: the block of the element $x_{ij}$ (row $i$, column $j$)

- $z = (z_1, \ldots, z_n)$: partition of individuals in $K$ clusters of rows
- $w = (w_1, \ldots, w_d)$: partition of variables in $L$ clusters of columns
- $(z, w)$: bi-partition of the whole data set $\mathbf{x}$
- Both space partitions are respectively denoted by $Z$ and $W$

Restriction

All variables are of the same kind (research in progress for overcoming that... )
The latent block model (LBM)

- Generalization of some existing non-probabilistic methods
- Extend the latent class principle of local (or conditional) independence
- Thus $x_{ij}$ is assumed to be independent once $z_i$ and $w_j$ are fixed ($\alpha = (\alpha_{kl})$):
  \[ p(x|z, w; \alpha) = \prod_{i,j} p(x_{ij}; \alpha_{z_iw_j}) \]

- $\pi = (\pi_k)$: vectors of proba. $\pi_k$ that a row belongs to the $k$th row cluster
- $\rho = (\rho_k)$: vectors of proba. $\rho_k$ that a row belongs to the $l$th column cluster
- Independence between all $z_i$ and $w_j$
- Extension of the traditional mixture model-based clustering ($\alpha = (\alpha_{kl})$):
  \[ p(x; \theta) = \sum_{(z, w) \in Z \times W} \prod_{i,j} \pi_{z_i} \rho_{w_j} p(x_{ij}; \alpha_{z_iw_j}) \]
Distribution for different kinds of data

[Govaert and Nadif, 2014] The pdf $p(\cdot; \alpha_z,w_j)$ depends on the kind of data $x_{ij}$:

- **Binary** data: $x_{ij} \in \{0, 1\}$, $p(\cdot; \alpha_{kl}) = B(\alpha_{kl})$
- **Categorical** data with $m$ levels:
  $x_{ij} = \{x_{ijh}\} \in \{0, 1\}^m$ with $\sum_{h=1}^m x_{ijh} = 1$ and $p(\cdot; \alpha_{kl}) = M(\alpha_{kl})$ with $\alpha_{kl} = \{\alpha_{kjh}\}$
- **Count** data: $x_{ij}^l \in \mathbb{N}$, $p(\cdot; \alpha_{kl}) = P(\mu_k \nu_l \gamma_{kl})$\(^9\)
- **Continuous** data: $x_{ij}^l \in \mathbb{R}$, $p(\cdot; \alpha_{kl}) = N(\mu_{kl}, \sigma_{kl}^2)$

\(^9\)The Poisson parameter is here split into $\mu_k$ and $\nu_l$ the effects of the row $k$ and the column $l$ respectively and $\gamma_{kl}$ the effect of the block $kl$. Unfortunately, this parameterization is not identifiable. It is therefore not possible to estimate simultaneously $\mu_k$, $\nu_l$ and $\gamma_{kl}$ without imposing further constraints. Constraints $\sum_k \pi_k \gamma_{kl} = \sum_l \rho_l \gamma_{kl} = 1$ and $\sum_k \mu_k = 1$, $\sum_l \nu_l = 1$ are a possibility.
Extreme parsimony ability

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binary</td>
<td>(\dim(\pi) + \dim(\rho) + KL)</td>
</tr>
<tr>
<td>Categorical</td>
<td>(\dim(\pi) + \dim(\rho) + KL(m - 1))</td>
</tr>
<tr>
<td>Contingency</td>
<td>(\dim(\pi) + \dim(\rho) + KL)</td>
</tr>
<tr>
<td>Continuous</td>
<td>(\dim(\pi) + \dim(\rho) + 2KL)</td>
</tr>
</tbody>
</table>

Very parsimonious so well suitable for the (ultra) HD setting

\[\text{nb. param}_{\text{HD}} = \text{nb. param}_{\text{classic}} \times \frac{L}{d}\]

Other advantage: stay in the canonical space thus meaningful for the end-user
Binary illustration: easy interpretation

[Govaert, 2011]
Binary illustration: user-friendly visualization

\[ n = 500, \ d = 10, \ K = 6, \ L = 4 \]

[Govaert, 2011]
MLE estimation: log-likelihood(s)

- Remember Lesson 3: first estimate \( \theta \), then deduce estimate of \((z, w)\)
- Observed log-likelihood: \( \ell(\theta; x) = \ln p(x; \theta) \)
- MLE:
  \[ \hat{\theta} = \arg \max_{\theta} \ell(\theta; x) \]

- Complete log-likelihood:
  \[
  \ell_c(\theta; x, z, w) = \ln p(x, z, w; \theta) = \sum_{i,k} z_{ik} \log \pi_k + \sum_{k,l} w_{jl} \log \rho_l + \sum_{i,j,k,l} z_{ik} w_{jl} \log p(x_i^j; \alpha_{kl})
  \]

Be careful with asymptotics...

If \( \ln(d)/n \to 0, \ln(n)/d \to 0 \) when \( n \to \infty \) and \( d \to \infty \), then the MLE is consistent [Brault et al., 2017]
MLE estimation: EM algorithm

- **E-step** of EM (iteration $q$):

$$Q(\theta, \theta^{(q)}) = E[\ell_c(\theta; x, z, w)|x; \theta^{(q)}]$$

$$= \sum_{i,k} p(z_i = k|x; \theta^{(q)}) \ln \pi_k + \sum_{j,l} p(w_i = l|x; \theta^{(q)}) \ln \rho_l$$

$$+ \sum_{i,j,k,l} p(z_i = k, w_j = l|x; \theta^{(q)}) \ln p(x_{ij}; \alpha_{kl})$$

- **M-step** of EM (iteration $q$): classical. For instance, for the Bernoulli case, it gives

$$\pi_k^{(q+1)} = \frac{\sum_i t_{ik}^{(q)}}{n}, \quad \rho_l^{(q+1)} = \frac{\sum_j s_{jl}^{(q)}}{d}, \quad \alpha_{kl}^{(q+1)} = \frac{\sum_{i,j} e_{ijkl}^{(q)} x_{ij}}{\sum_{i,j} e_{ijkl}^{(q)}}$$
MLE: intractable E step

\[ e^{(q)}_{ijkl} \text{ is usually intractable...} \]

- Consequence of dependency between \( x_{ij} \)s (link between rows and columns)
- Involve \( K^n L^d \) calculus (number of possible blocks)
- Example: if \( n = d = 20 \) and \( K = L = 2 \) then \( 10^{12} \) blocks
- Example (cont’d): 33 years with a computer calculating 100,000 blocks/second

Alternatives to EM

- Variational EM (numerical approx.): conditional independence assumption
  \[ p(z, w|x; \theta) \approx p(z|x; \theta)p(w|x; \theta) \]

- SEM-Gibbs (stochastic approx.): replace E-step by a S-step approx. by Gibbs
  \[ z|x, w; \theta \text{ and } w|x, z; \theta \]
MLE: variational EM (1/2)

- Use a general variational result from [Hathaway, 1985]
- Maximizing $\ell(\theta; x)$ on $\theta$ is equivalent to maximize $\tilde{\ell}_c(\theta; x, e)$ on $(\theta, e)$

$$
\tilde{\ell}_c(\theta; x, e) = \sum_{i,k} t_{ik} \ln \pi_k + \sum_{j,l} s_{jl} \ln \rho_l + \sum_{i,j,k,l} e_{ijkl} \ln p(x_{ij}; \alpha_{kl})
$$

where $e = (e_{ijkl})$, $e_{ijkl} \in \{0, 1\}$, $\sum_{k,l} e_{ijkl} = 1$, $t_{ik} = \sum_{j,l} e_{ijkl}$, $s_{jl} = \sum_{i,k} e_{ijkl}$

- Of course maximizing $\ell(\theta; x)$ or $\tilde{\ell}_c(\theta; x, e)$ are both intractable
- Idea: restriction on $e$ to obtain tractability $e_{ijkl} = t_{ik}s_{jl}$
- New variables are thus now $t = (t_{ik})$ and $s = (s_{jl})$
- As a consequence, it is a maximization of a lower bound of the max. likelihood

$$
\max_\theta \ell(\theta; x) \geq \max_{\theta, t, s} \tilde{\ell}_c(\theta; x, e)
$$
MLE: variational EM (2/2)

Approximated E-step

\[ Q(\theta, \theta^{(q)}) \approx \sum_{i,k} t_{ik}^{(q)} \ln \pi_k + \sum_{j,l} s_{jl}^{(q)} \ln \rho_l + \sum_{i,j,k,l} t_{ik}^{(q)} s_{jl}^{(q)} \ln p(x_{ij}; \alpha_{kl}) \]

- We called it now VEM
- Also known as mean field approximation
- Consistency of the variational estimate [Brault et al., 2017]
MLE: local maxima

- More local maxima than in classical mixture models
- It is a consequence of many more latent variables (blocks)
- Thus: either many VEM runs, or use the SEM-Gibbs algorithm
MLE: SEM-Gibbs

- We have already seen the SEM algorithm in Lesson 3 (thus we do not detail more)
- It limits dependency to starting point, so it limits local maxima
- The S-step: a draw \((z^{(q)}, w^{(q)}) \sim p(z, w|x; \theta^{(q)})\) instead an expectation
- But it is still intractable, thus use a Gibbs algorithm to approx. this draw

**Approximated S-step**

Two easy draws

\[ z^{(q)} \sim p(z|w^{(q-1)}, x; \theta^{(q)}) \]

and

\[ w^{(q)} \sim p(w|z^{(q)}, x; \theta^{(q)}) \]

- Rigorously speaking, many draws within the S-step should be performed
- Indeed, Gibbs has to reach a stochastic convergence
- In practice it works well while saving computation time
MLE: degeneracy

- More degenerate situations than in classical mixture models
- It is again a consequence of many more latent variables (blocks)
- The Bayesian regularization (instead MLE) can be an answer
Illustration of a degenerate situation
Bayesian estimation: pitch

- Everything passes by the **posterior distribution of** $\theta$

\[ p(\theta|x) \propto p(x|\theta) \cdot p(\theta) \]

- log-likelihood  
  prior

- Then, take (for instance) the **MAP** as a $\theta$ estimate (use a VEM like algo. . . )

\[ \hat{\theta} = \arg \max_{\theta} p(\theta|x) \]
Bayesian estimation: limiting degeneracy

- Interest for avoiding degeneracy is the prior: it acts as a penalization term
- Typical choices are Dirichlet for $\pi$ and $\rho$ (with independence between $\pi$, $\rho$, $\alpha$)

$$p(\theta) = \underbrace{p(\pi)}_{D_K(a,\ldots,a)} \times \underbrace{p(\rho)}_{D_L(a,\ldots,a)} \times \underbrace{p(\alpha)}_{\text{model dependent}}$$

- The Dirichlet distribution is conjugate, thus easy calculus
- Control degeneracy frequency with the $a$ value:
  - $a = 1$: uniform prior, so $\hat{\theta}$ is strictly the MLE (no regularisation)
  - $a = 1/2$: Jeffreys prior, classical (no informative prior) but may favor degeneracy
  - $a = 4$: a rule of thumb working well for limiting degeneracy frequency
Bayesian estimation: prior overview

\[ \text{Be} \left( \frac{1}{2}, \frac{1}{2} \right) \]

\[ \text{Be}(1,1) \]

\[ \text{Be}(4,4) \]

Graphs showing the probability density functions for different prior distributions.
Document clustering (1/2)

- Mixture of 1033 medical summaries and 1398 aeronautics summaries
- **Lines**: 2431 documents
- **Columns**: present words (except stop), thus 9275 unique words
- **Data matrix**: cross counting document $\times$ words
- **Poisson model**
### Document clustering (2/2)

\[
p(\hat{z} \neq z) \leq 2n \exp \left\{ -\frac{1}{8} d \left[ \min_{k \neq k'} |\tau_k - \tau_{k'}| \right] \right\} + K(1 - \min_k \pi_k)^n
\]

**Results with 2×2 blocs**

<table>
<thead>
<tr>
<th></th>
<th>Medline</th>
<th>Cranfield</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medline</td>
<td>1033</td>
<td>0</td>
</tr>
<tr>
<td>Cranfield</td>
<td>0</td>
<td>1398</td>
</tr>
</tbody>
</table>
Running BlockCluster

Configuration

If you change the configuration of your job and save it, it will start a new process with the updated parameters. This will erase previous results.

Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Trial BlockCluster</td>
</tr>
<tr>
<td>Data File</td>
<td>Blockcluster-Example.csv</td>
</tr>
<tr>
<td>Data Type</td>
<td>Categorical</td>
</tr>
<tr>
<td>Rows Cluster Groups</td>
<td>1:5</td>
</tr>
<tr>
<td>Column Cluster Groups</td>
<td>1:5</td>
</tr>
</tbody>
</table>
Running BlockCluster

Select a job execution from the list below:

- **69** Trial BlockCluster
  - Blockcluster-Example.csv
  - 23 May 20:47

- **68** Genes K1-12
  - lcpm.txt
  - 23 May 08:12

- **67** Genes
  - lcpm.txt
  - 22 May 15:38

- **65** Genes K1-10
  - lcpm.txt
  - 22 May 15:27
Running BlockCluster

<table>
<thead>
<tr>
<th>Model</th>
<th>Criterion</th>
<th>NbClusters</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>pük_rhel_multi</td>
<td>ICL (-45557.1)</td>
<td>[2, 3]</td>
<td>No error</td>
</tr>
<tr>
<td>pük_rhel_multi</td>
<td>ICL (-45563.3)</td>
<td>[3, 3]</td>
<td>No error</td>
</tr>
<tr>
<td>pük_rhel_multi</td>
<td>ICL (-45866.6)</td>
<td>[2, 4]</td>
<td>No error</td>
</tr>
<tr>
<td>pük_rhel_multi</td>
<td>ICL (-45873.9)</td>
<td>[4, 3]</td>
<td>No error</td>
</tr>
<tr>
<td>pük_rhel_multi</td>
<td>ICL (-45874.6)</td>
<td>[5, 5]</td>
<td>No error</td>
</tr>
<tr>
<td>pük_rhel_multi</td>
<td>ICL (-45877.7)</td>
<td>[3, 4]</td>
<td>No error</td>
</tr>
<tr>
<td>pük_rhel_multi</td>
<td>ICL (-45878.8)</td>
<td>[2, 5]</td>
<td>No error</td>
</tr>
</tbody>
</table>

Model Criterion

This chart represents the criterion value for each model that was built. The higher the value (closer to 0) the better the model.
Running BlockCluster

Cluster Plot

This image presents the original data matrix, and the matrix obtained after performing co-clustering.
Outline

1. Introduction
2. Model-based clustering
3. Mixmod in MASSICCC
4. MixtComp in MASSICCC
5. BlockCluster in MASSICCC
6. Conclusion
- Use probabilistic modelling as a mathematical guideline
- Use the MASSICCC platform for user-friendly implementation

https://massiccc.lille.inria.fr/