Evaluation of integrative clustering methods for the analysis of multi-omics data

Cécile Chauvel, Alexei Novoloaca, Pierre Veyre, Frédéric Reynier and Jérémie Becker

cecile.chauvel@bioaster.org

StatOmique workshop, November 5th 2019
Integrative clustering methods for multi-omics data

Four different strategies of integration:

1. Analyze each omics separately and combine results at the interpretation step
2. Clustering on each omics separately before applying consensus clustering
3. Concatenation into a single matrix before applying standard clustering approaches
4. Search for common variations across omics by specific models

Several questions to be addressed:

• How are omics data integrated?
• How is clustering performed?
• How are data pre-processed?
• How are the model parameters tuned?
• What are the performances of the methods?
OUTLINE

① Presentation of the methods
② Simulation study
③ Application on the TCGA breast cancer dataset
④ Conclusion
PRESENTATION OF THE METHODS
Presentation of the methods

- The dataset is composed of $K$ matrices $X_1, \ldots, X_K$
- Each matrix $X_k$ is of size $p_k \times n$ ($p_k$ variables/features, $n$ samples)
- All matrices contain measurements on the same $n$ samples
  - The goal is to perform clustering on the samples

- Focus on approaches that
  - can be applied to any omics,
  - do not require any prior biological knowledge (e.g., pathways)
  - and give an insight to omics variables.

### Non-integrative
- (2.) Gaussian mixture models on each omic + consensus clustering
- (3.) Concatenation + Gaussian mixture models

### Matrix factorization
- iCluster
- moCluster
- JIVE
- iNMF

### Bayesian
- BCC
- MDI
iCluster is a Gaussian joint latent variable model:

\[ X_k = W_k Z + \epsilon_k, \]
\[ Z \sim N_q(0, I). \]

- \( W_k (p_k \times q) \) data-specific loading matrix
- \( Z (q \times N) \) shared latent variable matrix
- \( \epsilon_k \sim N(0, \Sigma_k) \), with \( \Sigma_k \) diagonal

**PARAMETERS**

- Number of clusters determined by the Proportion of Deviance or the Rand Index.
- Number of latent variables = Number of clusters - 1

**DATA PRE-PROCESSING**

Centering of the \( X_k \)

**ESTIMATION**

EM algorithm

**CLUSTERING**

K-means on \( E(Z|X_1, ..., X_K) \)
moCluster


Model close to iCluster:

\[ \mathbf{X}_k = \mathbf{W}_k \mathbf{Z} + \epsilon_k, \]

- \( \mathbf{W}_k (p_k \times q) \) data-specific loading matrix
- \( \mathbf{Z} (q \times N) \) shared latent variable matrix
- \( \epsilon_k \sim N(0, \sigma^2 I) \)

Same noise variance across variables and data types ➔ shared and specific variations no longer separable

**PARAMETERS**

- Number of clusters determined by the gap statistic
- Number of latent variables determined by inspection of eigen values (scree plot or permutation test)

**DATA PRE-PROCESSING**

\( \mathbf{X}_k \) standardized and scaled by the inverse of the largest eigen value

**ESTIMATION**

Consensus PCA (NIPALS algorithm)

**CLUSTERING**

HCA on the latent variable matrix \( \mathbf{Z} \)
Addition of a data-specific term:

\[ X_k = W_k Z + W^s_k Z^s_k + \epsilon_k \]

- \( W^s_k (p_k \times q_k) \) data-specific loading matrix
- \( Z^s_k (q_k \times N) \) data-specific latent variable matrix

Constraint of orthogonality for identifiability: \( W_k Z \cdot (W^s_k Z^s_k)^T = 0 \)

**DATA PRE-PROCESSING**

- \( X_k \) centered, and scaled by their Frobenius norm

**ESTIMATION**

Iterative error minimization by fixing one term (shared or specific) at a time + SVD decomposition

**CLUSTERING**

No guidelines

**PARAMETERS**

Number of latent variables estimated by permutation approach on the eigen values
iNMF – integrative Non-negative Matrix Factorization


The model is a particular case of JIVE in which the shared and specific loadings are equal:

\[ X_k = (Z + Z^s_k)W_k + \epsilon_k, \]

with a non-negativity constraint: \( Z, Z^s_k, W_k \geq 0 \)

**ESTIMATION**

Minimization of the penalized loss function:

\[ \min_{Z, Z^s_k, W_k} \sum_{k=1}^{K} ||X_k - (Z + Z^s_k)W_k||^2 + \lambda \sum_{k=1}^{K} ||Z^s_k W_k||^2. \]

\( \lambda \) controls for the homogeneity between shared and specific structure:

High \( \lambda \) \( \Rightarrow \) more emphasis on the shared structure.

**PARAMETERS**

- Number of latent variables maximizing stability (consensus approach)
- \( \lambda \) : ad hoc procedure attributing as much weight as possible to the specific structure

**DATA PRE-PROCESSING**

Variance stabilization (log – transformation), non-negativity transformation and scaling by the Frobenius norm.

**CLUSTERING**

No guidelines
MDI - Multiple Dataset Integration


- Bayesian method: Dirichlet multinomial allocation mixture model
- Cluster assignments are dependant across datasets:

\[ P(c_{i1}, c_{i2}, \ldots, c_{iK} | \phi) \propto \prod_{k=1}^{K} \pi_{c_{ik}} k \prod_{k=1}^{K-1} \prod_{l=k+1}^{K} (1 + \phi_{kl} \mathbb{I}(c_{ik} = c_{il})) \]

Cluster allocation of sample \( i \) in dataset \( k \)

Mixture proportion associated with cluster \( c_{ik} \) in dataset \( k \)

Association strength between datasets \( k \) and \( l \)

DATA PRE-PROCESSING
None

ESTIMATION
Gibbs sampling

GLOBAL CLUSTERING
Maximization of the posterior expected adjusted Rand Index across source-specific clusterings

PARAMETERS
Maximal number of clusters. The authors’ recommendation: \( n/2 \) but instable in our simulations (\( n \) was chosen)
BCC - Bayesian Consensus Clustering


Dirichlet mixture model, aiming at uncovering a single clustering across sources by:

\[
P(L_{kn} = l | C_n) = \begin{cases} \alpha_k & \text{if } C_n = L_{kn} \\ 1 - \alpha_k & \text{if } C_n \neq L_{kn} \end{cases} \]

**DATA PREPROCESSING**
- None

**ESTIMATION**
- Gibbs sampling

**CLUSTERING**
- Estimated by C (shared clustering) or L (source-specific)

**PARAMETERS**
- Maximal number of clusters $q$ maximizing the mean adherence
Main simulation study

- K=3 data matrices with
  - 180, 210 and 240 variables
  - 60 samples
  - 3 shared clusters of 20 samples each
  - 2 levels of Signal to Noise Ratio (SNR)

- 3 simulation strategies
  - iNMF-derived scenario with overlaps between the shared and specific blocks
  - iNMF-derived scenario without overlaps between the shared and specific blocks
  - BCC-derived scenario with 3 to 5 specific clusters

- Evaluation (100 repetitions of each scenario):
  - Estimated number of shared clusters
  - Clustering performance (Adjusted Rand Index)
Number of shared clusters

- 3 clusters chosen on average
- Sharp peak around 3 for high SNR and iNMF overlap scenarios
- Ranking of the methods (by % of times 3 clusters are found):
  1. iNMF
  2. iCluster
  3. JIVE
  4. BCC
  5. moCluster
  6. MDI
Clustering on shared structures

- ARI of integrative methods are higher than those of non-integrative ones
- SNR + simulation design have a great impact on clustering.
- Ranking of the methods:
  1. iCluster, moCluster, iNMF
  2. JIVE, BCC
  3. MDI (extremely sensitive to noise)
Clustering on specific structures

- Not central in our study, classical clustering methods apply such as GMM
- GMM slightly outperforms BCC
- JIVE underperforms ➜ identifiability issues
- MDI sensitive to noise
High-dimension simulation study

• Design of the high dimension study
  On the iNMF-overlap scenario
  – 300, 600 and 3000 variables
  – 60 samples
  – 3 common clusters of 20 samples each
  – 2 levels of signal to noise ratio
  – 100 repetitions
High dimension study – no impact of the data set size

- Shared clustering: same performances and ranking as in the low dimension case (not shown here)

- Specific clustering not impacted by the sample size of the data set:
## Run times

<table>
<thead>
<tr>
<th>Method</th>
<th>Time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>moCluster</td>
<td>0.5</td>
</tr>
<tr>
<td>Consensus clustering</td>
<td>1.3</td>
</tr>
<tr>
<td>GMM</td>
<td>1.3</td>
</tr>
<tr>
<td>Concatenation</td>
<td>1.6</td>
</tr>
<tr>
<td>iCluster</td>
<td>16.0</td>
</tr>
<tr>
<td>JIVE</td>
<td>111.9</td>
</tr>
<tr>
<td>iNMF</td>
<td>102.6</td>
</tr>
<tr>
<td>BCC</td>
<td>1441.4</td>
</tr>
<tr>
<td>MDI</td>
<td>3810.6</td>
</tr>
</tbody>
</table>

### Main study

<table>
<thead>
<tr>
<th>Method</th>
<th>Time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>moCluster</td>
<td>0.1</td>
</tr>
<tr>
<td>Consensus clustering</td>
<td>34.7</td>
</tr>
<tr>
<td>GMM</td>
<td>34.7</td>
</tr>
<tr>
<td>Concatenation</td>
<td>58.9</td>
</tr>
<tr>
<td>iCluster</td>
<td>194.6</td>
</tr>
<tr>
<td>JIVE</td>
<td>20.3</td>
</tr>
<tr>
<td>iNMF</td>
<td>1056.3</td>
</tr>
<tr>
<td>BCC</td>
<td>14300.6</td>
</tr>
<tr>
<td>MDI</td>
<td>63153.4</td>
</tr>
</tbody>
</table>

### High-dimension study
Conclusion on the simulations

- For shared structures, iCluster, moCluster and iNMF have good clustering performances.
- For specific structures, only BCC reaches the performances of non-integrative methods.
- No method can well detect both shared and specific structures at the same time.
- No impact of the number of features in the datasets.
- Ranking supported as well by a sensitivity simulation study.
APPLICATION ON TCGA DATA
TCGA breast cancer data

- Four omics measured on 348 patients:
  - mRNA
  - miRNA
  - DNA methylation
  - proteins

- Practitioners divide patients into 4 subtypes based on:
  - expression of proliferating protein Ki67
  - receptor status for estrogen (ER)
  - receptor status for progesterone (PR)
  - receptor status for human epidermal growth factor 2 (HER2)

<table>
<thead>
<tr>
<th>Subtype</th>
<th>Markers Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basal</td>
<td>ER- PR- HER2-</td>
</tr>
<tr>
<td>HER2</td>
<td>ER- PR- HER2+</td>
</tr>
<tr>
<td>Luminal A</td>
<td>ER+ and/or PR+ HER2-</td>
</tr>
<tr>
<td>Luminal B</td>
<td>ER+ and/or PR+ HER2+ or High Ki67</td>
</tr>
</tbody>
</table>

- Comparison of these classes with clusters from integrative methods
TCGA breast cancer data

<table>
<thead>
<tr>
<th>Method</th>
<th>% ER</th>
<th>% PR</th>
<th>% HER2</th>
<th>ARI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consensus clustering</td>
<td>97</td>
<td>89</td>
<td>11</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>66</td>
<td>45</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>5</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>96</td>
<td>80</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>Concatenation</td>
<td>97</td>
<td>89</td>
<td>7</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>98</td>
<td>77</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>6</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>63</td>
<td>44</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td>iCluster</td>
<td>95</td>
<td>71</td>
<td>26</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>96</td>
<td>85</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>90</td>
<td>81</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>5</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>moCluster</td>
<td>13</td>
<td>6</td>
<td>2</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td>98</td>
<td>83</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>64</td>
<td>40</td>
<td>56</td>
<td></td>
</tr>
<tr>
<td></td>
<td>96</td>
<td>89</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>iNMF</td>
<td>8</td>
<td>56</td>
<td>41</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>97</td>
<td>9</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>6</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>87</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>JIVE</td>
<td>96</td>
<td>84</td>
<td>9</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>4</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>99</td>
<td>84</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>79</td>
<td>63</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>BCC</td>
<td>70</td>
<td>49</td>
<td>43</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>9</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>97</td>
<td>84</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>98</td>
<td>89</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>MDI</td>
<td>94</td>
<td>87</td>
<td>10</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>8</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>94</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td></td>
<td>99</td>
<td>83</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>mRNAs</td>
<td>-</td>
<td>-</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>DNA methylation</td>
<td>-</td>
<td>-</td>
<td>0.41</td>
<td></td>
</tr>
<tr>
<td>miRNAs</td>
<td>-</td>
<td>-</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>Proteins</td>
<td>-</td>
<td>-</td>
<td>0.00</td>
<td></td>
</tr>
</tbody>
</table>

- Performances of the single omics vary: impact of the number of features or biological explanation?
- All integrative methods but iCluster and JIVE overpass single omics
- Ranking of the methods:
  1. moCluster, iNMF (consistent with simulations)
  2. MDI, BCC, Non-integrative
  3. iCluster, JIVE (different from simulations)
- Limit of the comparison:
  - Classification used as gold standard in clinics but no « true » classes
  - Very low prevalence of HER2 subclass ➔ difficult to detect
CONCLUSIONS
Key points

- The integration of multiple omics shows a clear improvement in clustering performance as compared to non-integrative methods.

- Matrix factorization methods are on average better at identifying shared clusters (especially moCluster and iNMF).

- Although iNMF showed a lack of sensitivity, it can finely be tuned to recover either common or specific clusters.

- Despite moderate performances on shared clusters, BCC displayed the best ability to recover both structures.

- MDI highly impacted by noise.

- Bayesian methods easier to parametrize, but longer to run.

- It would be interesting to study variable selection (available in iCluster, moCluster, JIVE and iNMF).
References


Thank you for your attention
Sensitivity study

• Design of the sensitivity study

On iNMF scenarios
- 180, 210 and 240 variables
- 60 samples, in which 3 blocks of \{15, 12, 9, 6, 3, 0\} samples are noise
- 3 common clusters of 20 samples each
- 2 levels of signal to noise ratio
- 20 repetitions
Sensitivity study

- SNR, simulation design (overlaps or not) and cluster sizes impact ARI
- Methods ranking:
  1. iCluster, moCluster, JIVE
  2. iNMF, BCC
  3. No results for MDI (too sensitive to noise)
Grid search on parameters for iNMF