Introduction to Biclustering Methods

Jingxuan Bao

Advisor: Dr. Li Shen

School of Arts and Sciences
Applied Math and Computational Science
University of Pennsylvania

November 12, 2020

Overview

1. Biclustering
2. Selected Biclustering Methods
3. Other Biclustering Methods
4. Relevant Reference

What is biclustering?

Definition

Biclustering, block clustering, co-clustering, or two-mode clustering is a data mining technique which allows simultaneous clustering of the rows and columns of a matrix.

Given a set of m samples represented by an n-dimensional feature vector, the entire dataset can be represented as n rows in m columns (i.e., an $n \times m$ matrix). The biclustering algorithm generates biclusters – a subset of rows which exhibit similar behavior across a subset of columns, or vice versa.

Type of Biclusters

Bicluster with constant values

a) Bicluster with constant values

2.0	2.0	2.0	2.0	2.0
2.0	2.0	2.0	2.0	2.0
2.0	2.0	2.0	2.0	2.0
2.0	2.0	2.0	2.0	2.0
2.0	2.0	2.0	2.0	2.0

Figure: Biclustering with constant values

Type of Biclusters

Biclusters with constant values on rows or columns

b) Bicluster with constant values on rows

1.0	1.0	1.0	1.0	1.0
2.0	2.0	2.0	2.0	2.0
3.0	3.0	3.0	3.0	3.0
4.0	4.0	4.0	4.0	4.0
5.0	5.0	5.0	5.0	5.0

c) Bicluster with constant values on columns

1.0	2.0	3.0	4.0	5.0
1.0	2.0	3.0	4.0	5.0
1.0	2.0	3.0	4.0	5.0
1.0	2.0	3.0	4.0	5.0
1.0	2.0	3.0	4.0	5.0

Type of Biclusters

Biclusters with constant values on coherent values

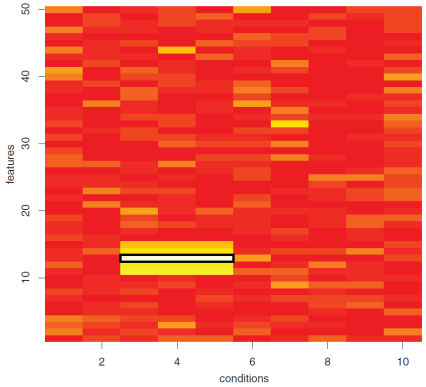
d) Bicluster with coherent values (additive)

1.0	4.0	5.0	0.0	1.5
4.0	7.0	8.0	3.0	4.5
3.0	6.0	7.0	2.0	3.5
5.0	8.0	9.0	4.0	5.5
2.0	5.0	6.0	1.0	2.5

e) Bicluster with coherent values (multiplicative)

1.0	0.5	2.0	0.2	0.8
2.0	1.0	4.0	0.4	1.6
3.0	1.5	6.0	0.6	2.4
4.0	2.0	8.0	0.8	3.2
5.0	2.5	10.0	1.0	4.0

xMotif Algorithm



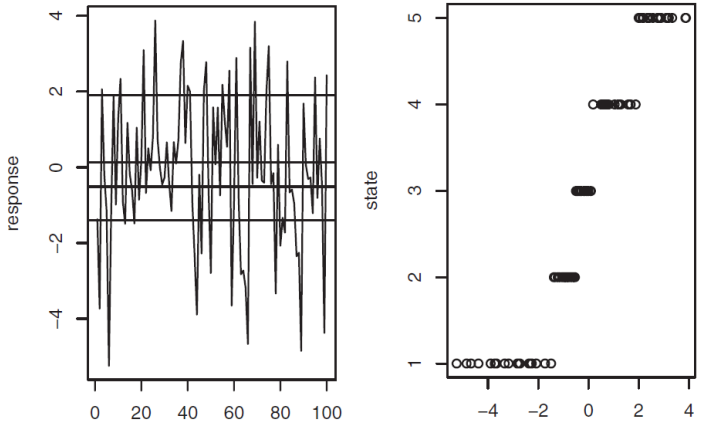
The basic idea is to find conserved gene expression motifs (xMotifs). An xMotif is defined as a subset of rows (genes) that is simultaneously conserved across a subset of the columns (conditions). We say a gene is conserved across a subset of conditions if the gene is in the same state in all conditions that belong to this subset.

xMotif Algorithm

What is "state"?

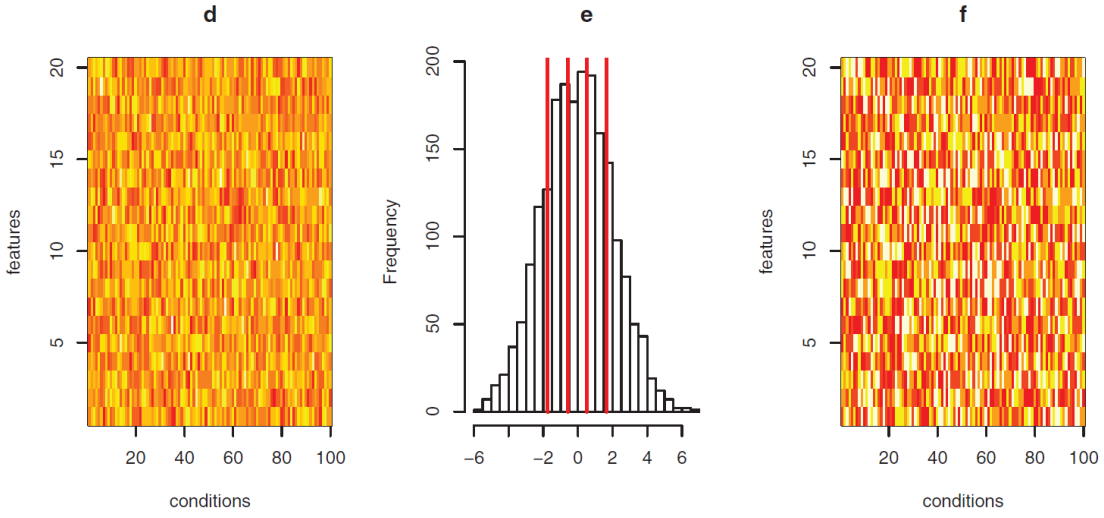
A gene state is a range of expression values and it is further assumed that there are a fixed number of states.

Example



xMotif Algorithm

Example



xMotif Algorithm

What are we looking for?

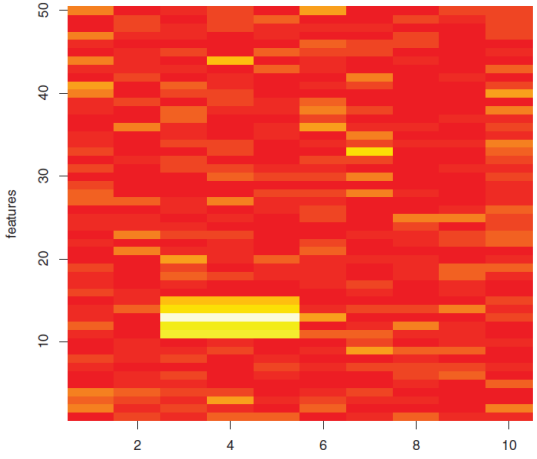
Bicluster (xMotif): Genes that belong to the bicluster are in the same state for a given set of conditions. We are aiming to find the largest xMotif.

- The number of columns must be at least a fraction of all the columns in the data matrix;
- For every row not belonging to the xMotif, the row must be conserved only in a fraction of the columns in the biclustering module;
- The expression levels of the genes in the bicluster can increase or decrease from one condition to another.

xMotif Algorithm

What are we looking for?

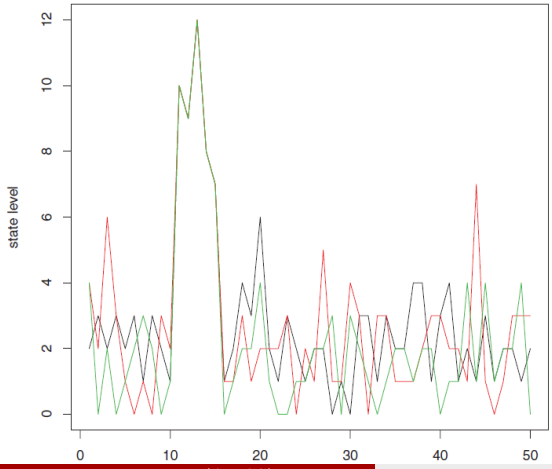
Bicluster (xMotif): Genes that belong to the bicluster are in the same state for a given set of conditions. We are aiming to find the largest xMotif.

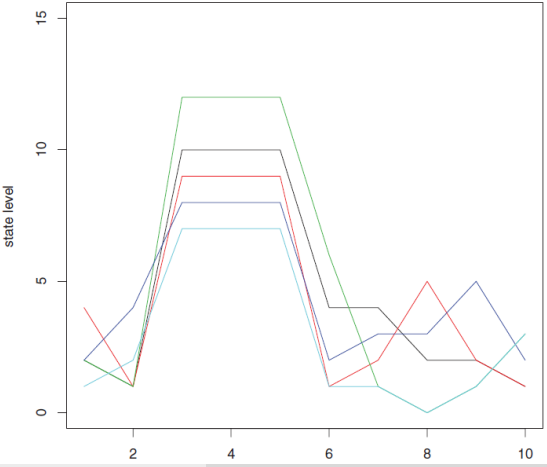


xMotif Algorithm

What are we looking for?

Bicluster (xMotif): Genes that belong to the bicluster are in the same state for a given set of conditions. We are aiming to find the largest xMotif.





xMotif Algorithm

Algprithm 1

- For $i = 1, \ldots, n_s$ do
  - Choose a sample c uniformly at random.
  - For $j = 1, \ldots, n_d$ do
    - Choose a subset D of the samples of size s_d uniformly at random.
    - For each gene g, if g is in the same state s in c and all the samples in D, include the pair (g, s) in the set G_{ij}.
    - C_{ij} = set of samples that agree with c in all the gene-states in G_{ij}.
    - Discard (C_{ij}, G_{ij}) if C_{ij} contains less than $\alpha \times n$ samples.
- Return the xMotif (C^*, G^*) that maximizes $|G_{ij}|$, where $1 \leq i \leq n_s$, $1 \leq j \leq n_d$.

Plaid Model

Plaid Model Description

The plaid model is an additive biclustering method that defines the expression level of the i-th gene under the j-th condition as a sum of biclusters (layers) in the expression matrix.

Mathematical Description

Let Y be $N \times M$ data matrix for which the rows represent genes and the columns conditions. For K biclusters, the gene expression level is expressed as a linear model of the form

$$Y_{ij} = \mu_0 + \sum_{k=1}^{K} \theta_{ijk}\rho_{ik}\kappa_{jk} + \epsilon_{ij}, \quad i = 1, \ldots, N; \quad j = 1, \ldots, M.$$

Plaid Model

$$Y_{ij} = \mu_0 + \sum_{k=1}^{K} \theta_{ijk}\rho_{ik}\kappa_{jk} + \epsilon_{ij}, \quad i = 1, \ldots, N; \quad j = 1, \ldots, M.$$

where

- μ_0 is an background effect;
- ϵ_{ij} is a Gaussian error with mean zero and variance σ^2;
- The parameters ρ_{jk} and κ_{jk} are binary parameters that represent the membership of the gene/condition in bicluster k in the following way:

$$\rho_{ik} = \begin{cases} 1 & \textit{gene } i \textit{ belongs to bicluster } k \\ 0 & \textit{otherwise} \end{cases}$$

$$\kappa_{jk} = \begin{cases} 1 & \textit{condition } j \textit{ belongs to bicluster } k \\ 0 & \textit{otherwise.} \end{cases}$$

Plaid Model

$$Y_{ij} = \mu_0 + \sum_{k=1}^{K} \theta_{ijk}\rho_{ik}\kappa_{jk} + \epsilon_{ij}, \quad i = 1, \ldots, N; \quad j = 1, \ldots, M.$$

where

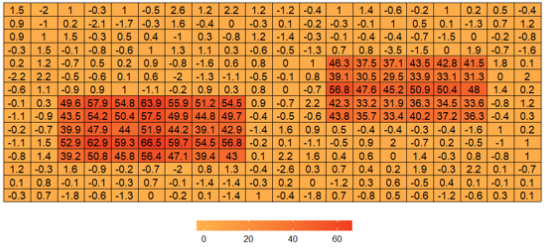
- θ_{ijk} is the mean gene expression for the k-th bicluster which takes one of four possible forms:

$$\theta_{ijk} = \begin{cases} \mu_k, & \textit{constant bicluster}; \\ \mu_k + \alpha_{ik}, & \textit{biclusters with constant rows}; \\ \mu_k + \beta_{jk}, & \textit{biclusters with constant columns}; \\ \mu_k + \alpha_{ik} + \beta_{jk}, & \textit{bicluster with coherent values}. \end{cases}$$

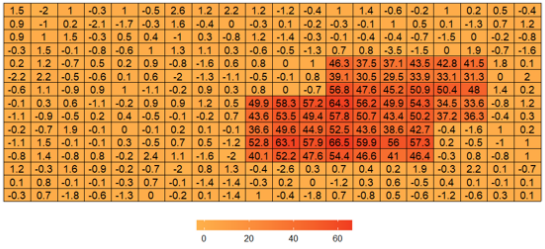
Plaid Model

$$Y_{ij} = \mu_0 + \sum_{k=1}^{K} \theta_{ijk}\rho_{ik}\kappa_{jk} + \epsilon_{ij}, \quad i = 1, \ldots, N; \quad j = 1, \ldots, M.$$

The binary parameters ρ_{ik} and κ_{jk} define the membership status of each gene and condition in the expression matrix and determine the underlying bicluster structure of the expression matrix. Hence,



Two Non-overlapping Biclusters



Two Overlapping Biclusters

Plaid Model

$$Y_{ij} = \mu_0 + \sum_{k=1}^{K} \theta_{ijk}\rho_{ik}\kappa_{jk} + \epsilon_{ij}, \quad i = 1, \ldots, N; \quad j = 1, \ldots, M.$$

The binary parameters ρ_{ik} and κ_{jk} define the membership status of each gene and condition in the expression matrix and determine the underlying bicluster structure of the expression matrix. Hence,

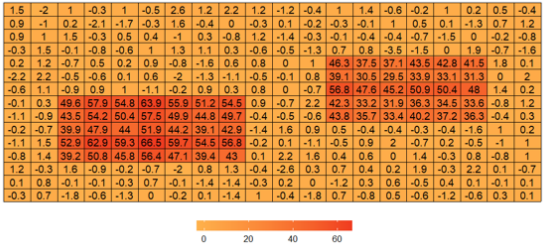
$$\sum_k \rho_{ik} = \begin{cases} 1 & \textit{feature } i \textit{ belongs to one bicluster}, \\ \geq 2 & \textit{feature } i \textit{ belongs to more than one cluster}, \\ 0 & \textit{feature } i \textit{ does not belong to any bicluster}. \end{cases}$$

$$\sum_k \kappa_{jk} = \begin{cases} 1 & \textit{observation } j \textit{ belongs to one bicluster}, \\ \geq 2 & \textit{observation } j \textit{ belongs to more than one cluster}, \\ 0 & \textit{observation } j \textit{ does not belong to any bicluster}. \end{cases}$$

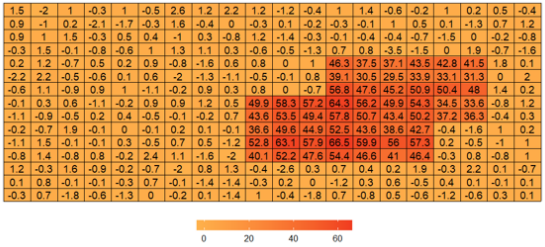
Plaid Model

$$Y_{ij} = \mu_0 + \sum_{k=1}^{K} \theta_{ijk}\rho_{ik}\kappa_{jk} + \epsilon_{ij}, \quad i = 1, \ldots, N; \quad j = 1, \ldots, M.$$

The binary parameters ρ_{ik} and κ_{jk} define the membership status of each gene and condition in the expression matrix and determine the underlying bicluster structure of the expression matrix. Hence,



Two Non-overlapping Biclusters



Two Overlapping Biclusters

Plaid Model - Estimation

For a given number of biclusters K the residual sum of squares is given by

$$Q = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(Y_{ij} - \theta_{ij0} - \sum_{k=1}^{K}\theta_{ijk}\rho_{ik}\kappa_{jk}\right)^2$$

where $\theta_{ijk} = \mu_k + \alpha_{ik} + \beta_{jk}$. Given the estimations for the model parameters and the membership parameters for $l-1$ biclusters, we define the residuals by

$$\hat{Z}_{ij} = Y_{ij} - \hat{\theta}_{ij0} - \sum_{k=1}^{l-1}\hat{\theta}_{ijk}\hat{\rho}_{ik}\hat{\kappa}_{jk} = (\mu_l + \alpha_{il} + \beta_{jl})\rho_{il}\kappa_{jl} + \epsilon_{ij}.$$

The current residual matrix $\boldsymbol{Z}$ is the input matrix for searching the l-th bicluster, whose parameters can be estimated by minimizing the residual sum of squares for the l-th bicluster

$$Q = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(\hat{Z}_{ij} - (\mu_l + \alpha_{il} + \beta_{jl})\rho_{il}\kappa_{jl}\right)^2.$$

The estimation of the unknown parameters is done in an iterative procedure in which one set of parameters is the estimated condition of the other set.

FABIA Model

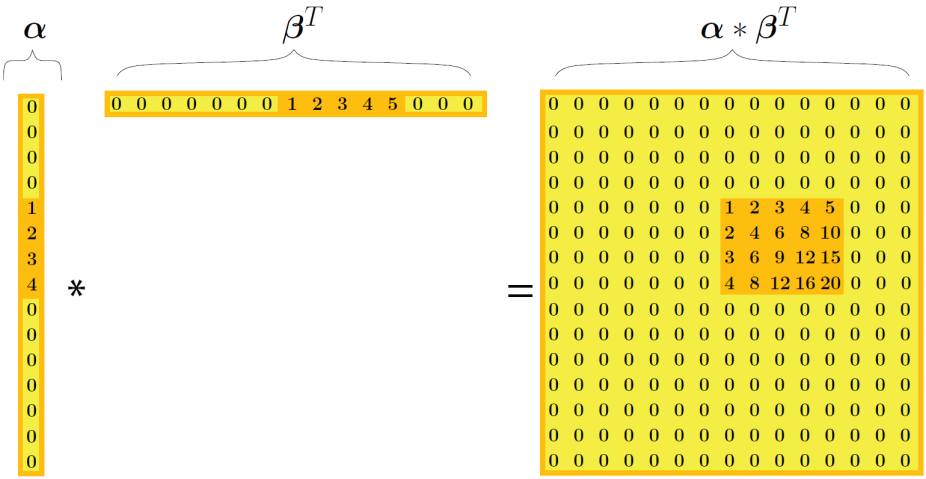


Figure: **Idea of FABIA Model:** Sparse matrix factorization model

FABIA Model - Formulation

Assume that the data matrix $\boldsymbol{Y} \in \mathbb{R}^{N \times M}$ is given and that it is row-centered, and may be normalized row-wise. The FABIA model for K biclusters and additive noise is

$$\boldsymbol{Y} = \sum_{k=1}^{K} \boldsymbol{\alpha}_k \boldsymbol{\beta}_k^\top + \boldsymbol{Z} = \boldsymbol{\Lambda}\boldsymbol{\Gamma} + \boldsymbol{Z}$$

where

- $\boldsymbol{Z} \in \mathcal{R}^{N \times M}$ is additive noise;
- $\boldsymbol{\alpha}_k \in \mathcal{R}^N$ is the sparse row (feature) membership vector of the k-th bicluster;
- $\boldsymbol{\beta}_k \in \mathcal{R}^M$ is the sparse column (sample) membership vector of the k-th bicluster;
- $\boldsymbol{\Lambda} = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_K] \in \mathbb{R}^{N \times K}$ is called loading matrix;
- $\boldsymbol{\Gamma} = \begin{bmatrix} \boldsymbol{\beta}_1^\top \\ \boldsymbol{\beta}_2^\top \\ \vdots \\ \boldsymbol{\beta}_K^\top \end{bmatrix} \in \mathbb{R}^{K \times M}$ is called factor matrix.

FABIA Model - Formulation

The FABIA model for K biclusters and additive noise is

$$\boldsymbol{Y} = \sum_{k=1}^{K} \boldsymbol{\alpha}_k \boldsymbol{\beta}_k^\top + \boldsymbol{Z} = \boldsymbol{\Lambda}\boldsymbol{\Gamma} + \boldsymbol{Z}$$

where the j-th sample $\boldsymbol{y}_j$ (the j-th column of $\boldsymbol{Y}$) is

$$\boldsymbol{y}_j = \sum_{k=1}^{K} \boldsymbol{\alpha}_k \beta_{kj} + \epsilon_j = \boldsymbol{\Lambda}\tilde{\boldsymbol{\beta}}_j + \boldsymbol{\epsilon}_j$$

where

- $\boldsymbol{\epsilon}_j \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\psi}) \in \mathbb{R}^N$ is the j-th column of the error matrix $\boldsymbol{Z}$ where $\boldsymbol{\psi} \in \mathbb{R}^{N \times N}$ is assumed to be diagonal to account for independent Gaussian noise;
- $\tilde{\boldsymbol{\beta}}_j = \begin{bmatrix} \beta_{1j} \\ \beta_{2j} \\ \vdots \\ \beta_{Kj} \end{bmatrix}$ is the j-th column of the factor matrix $\boldsymbol{\Gamma}$;
- $\tilde{\boldsymbol{\beta}}_{kj}$ represents the k-th value of the j-th factor.

FABIA Model - Estimation

$$\boldsymbol{y}_j = \sum_{k=1}^{K} \boldsymbol{\alpha}_k \beta_{kj} + \epsilon_j = \boldsymbol{\Lambda}\tilde{\boldsymbol{\beta}}_j + \boldsymbol{\epsilon}_j$$

Component-wise Independent Laplace Prior for $\tilde{\boldsymbol{\beta}}_j$

$$p(\tilde{\boldsymbol{\beta}}_j) = \left(\frac{1}{\sqrt{2}}\right)^K \prod_{k=1}^{K} e^{-\sqrt{2}|\tilde{\beta}_{kj}|}$$

Component-wise Independent Laplace Prior for $\boldsymbol{\alpha}_k$

$$p(\boldsymbol{\alpha}_k) = \left(\frac{1}{\sqrt{2}}\right)^N \prod_{i=1}^{N} e^{-\sqrt{2}|\tilde{\beta}_{ik}|}$$

FABIA Model - Estimation

$$\boldsymbol{y}_j = \sum_{k=1}^{K} \boldsymbol{\alpha}_k \beta_{kj} + \epsilon_j = \boldsymbol{\Lambda}\tilde{\boldsymbol{\beta}}_j + \boldsymbol{\epsilon}_j$$

Variational Expectation Maximization (EM) Approach

We want to estimate the parameters by maximizing the posterior distribution. When dealing with latent variables, we often need to first marginalize the latent variables out by integration or by expectation. However, the Laplace prior of the factor $\tilde{\boldsymbol{\beta}}_j$ leads to an analytically intractable[a] posterior distribution:

$$p(\tilde{\boldsymbol{\beta}}_j, \boldsymbol{\Lambda}, \boldsymbol{\psi} | \boldsymbol{y}_j) = p(\boldsymbol{\Lambda}, \boldsymbol{\psi} | \tilde{\boldsymbol{\beta}}_j, \boldsymbol{y}_j) p(\tilde{\boldsymbol{\beta}}_j).$$

Therefore, we apply variational EM algorithm to estimate the model parameters.

[a] Intractable: It is analytically impossible to marginalize $\tilde{\boldsymbol{\beta}}_j$ by integration or by calculating expectation.

FABIA Model - Variational EM Algorithm

Idea - Variational EM Algorithm

- Consider a variational measure for $\tilde{\boldsymbol{\beta}}_j$

$$\hat{p}(\tilde{\boldsymbol{\beta}}_j|\boldsymbol{\eta}_j, \boldsymbol{\Lambda}, \boldsymbol{\psi}, \boldsymbol{y}_j),$$

  where $\boldsymbol{\eta}_j$ is called variational measure.
- Minimize the Kullback–Leibler divergence between $\hat{p}(\tilde{\boldsymbol{\beta}}_j|\boldsymbol{\eta}_j, \boldsymbol{\Lambda}, \boldsymbol{\psi}, \boldsymbol{y}_j)$ and $p(\tilde{\boldsymbol{\beta}}_j|\boldsymbol{\Lambda}, \boldsymbol{\psi}, \boldsymbol{y}_j)$ with respect to $\boldsymbol{\eta}_j$ to get the update rule for variational parameter $\boldsymbol{\eta}_j$.
- E-step: take expectation to the log-posterior distribution $\log p(\tilde{\boldsymbol{\beta}}_j, \boldsymbol{\Lambda}, \boldsymbol{\psi}|\boldsymbol{y}_j)$ with respect to $\tilde{\boldsymbol{\beta}}_j$ using variational measure $\hat{p}(\tilde{\boldsymbol{\beta}}_j|\boldsymbol{\eta}_j, \boldsymbol{\Lambda}, \boldsymbol{\psi}, \boldsymbol{y}_j)$.
- M-step: maximize the expected log-posterior distribution with respect to the model parameters $\boldsymbol{\Lambda}, \boldsymbol{\psi}$ to obtain updated rule.

Biclustering on Expression Data: A Review

Iterative Greedy Search

1. Direct Clusteriang (DC)
2. Biclustering of expression data by Chengu and Church (CC)
3. SMSR-based biclustering (SMSR-CC)
4. HARP algorithm
5. Maximum Similarity Bicluster algorithm (MSB)
6. Weighted Fuzzy-Based Maximum Similarity Bicluster algorithm (WF-MSB)
7. Biclustering by Iteratively Sorting with Weighted Coefficients (BISWC)
8. Biclustering by Correlated and Large Number of Individual Clustered seeds (BICLIC)
9. Intensive Correlation Search (ICS)

Biclustering on Expression Data: A Review

Stochastic Iterative Greedy Search

1. Flexible Overlapped biClustering (FLOC)
2. Random Walk Biclustering (RWB)
3. Reactive GRASP Biclustering (RGRASP-B)
4. Pattern-Driven Neighborhood Search (PDNS)

Nature-Inspired Meta-Heuristics

1. Simulated Annealing Biclustering (SA-B)
2. Crowding distance based Multi-Objective Particle Swarm Optimization Biclustering (CMOPSOB)
3. Multi-Objective Multi-population artificial immune Network (MOM-aiNet)
4. Evolutionary algorithms for biclustering
5. Multi-objective evolutionary algorithms

Biclustering on Expression Data: A Review

Clustering-Based Approaches

1. SVD and clustering-based approaches
2. Biclustering based on Related Genes and Conditions Extraction (RGCE-B)

One-way clustering-based approaches

1. Coupled Two-way Clustering (CTWC)
2. Interrelated Two-way Clustering (ITWC)

Probabilistic models

1. Plaid Models (PM)
2. Rich Probabilistic Models (RPM)
3. Gibbs Sampling (GS)
4. Bayesian Biclustering model (BBC)
5. Conserved gene expression Motifs (xMOTIFs)
6. cMonkey
7. Penalized Plaid Model (PPM)

Biclustering on Expression Data: A Review

Linear algebra

1. Spectral Biclustering (SB)
2. Iterative Signature Algorithm (ISA)
3. Non-smooth Non-negative Matrix Factorization (nsNMF)
4. Pattern-based Biclustering (BicPAM)

Optimal reordering of rows and columns

1. OPSM
2. OREO

Relevant Reference

1. Kasim, A., Shkedy, Z., Kaiser, S., Hochreiter, S., Talloen, W. (2017). Applied Biclustering Methods for Big and High-Dimensional Data Using R. New York: Chapman and Hall/CRC, https://doi.org/10.1201/9781315373966.

2. L. Lazzeroni and A. Owen. Plaid Models for Gene Expression Data. *Statistica Sinica*,12, 05 2000.

3. T. M. Murali and Simon Kasif. Extracting conserved gene expression motifs from gene expression data. *Proceedings of the Pacific Symposium on Biocomputing*, volume 8, pages 77–88, 2003.

4. S. Hochreiter et al. FABIA: factor analysis for bicluster acquisition. *Bioinformatics*, 26(12): 1520–1527, 04 2010.

5. Beatriz Pontes, Raúl Giráldez, Jesús S. Aguilar-Ruiz, Biclustering on expression data: A review, *Journal of Biomedical Informatics*, Volume 57, 2015.