Bayesian Generalized Biclustering Analysis via Adaptive Structured Shrinkage

Supplementary materials

Ziyi Li[1,†], Changgee Change[2,†], Suprateek Kundu[1] and Qi Long[2,*]

[1]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, 30322, USA.
[2]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, 423 Guardian Drive, 19104 Philadelphia, PA, USA.

[†] Ziyi Li and Changgee Chang contributed equally to this paper.
[*] Correspondance: qlong@pennmedicine.upenn.edu.

November 22, 2018

1 Details of EM algorithm

The optimization problem in the M step at the t-th iteration is defined as follows,

$$(\boldsymbol{W}^{(t)}, \boldsymbol{Z}^{(t)}, \boldsymbol{m}^{(t)}, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\xi}^{(t)}) = \underset{W,Z,m,\alpha,\xi}{\operatorname{argmax}}\ \tilde{\mathbb{E}}_t \log \pi(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{m}, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\rho}, \boldsymbol{\Omega}, \boldsymbol{X}),$$

where the expectation $\tilde{\mathbb{E}}_t$ is taken with respect to $\tilde{\pi}_t(\boldsymbol{\rho}, \boldsymbol{\Omega}) = \pi(\boldsymbol{\rho}, \boldsymbol{\Omega}|\boldsymbol{W}^{(t-1)}, \boldsymbol{Z}^{(t-1)}, \boldsymbol{m}^{(t-1)}, \boldsymbol{\alpha}^{(t-1)}, \boldsymbol{\xi}^{(t-1)}, \boldsymbol{X})$. The objective function to be optimized at the t-th EM iteration step is given by

$$\begin{aligned}\mathbf{Q}_t(\mathbf{Z}, \mathbf{W}, \boldsymbol{m}, \boldsymbol{\alpha}, \boldsymbol{\xi}) = &-\frac{1}{2}\sum_{i,j} \rho_{ji}^{(t)} (\mu_{ji} - \psi_{ji})^2 + \sum_{i,j} \kappa_{ji}\mu_{ji} + \sum_{j,l} \alpha_{jl} - \sum_{j,l} \lambda_{jl}|w_{jl}| \\ &+ \nu_3 \sum_{l,i} \log \xi_{i,l} - \sum_{i,l} \xi_{l,i}(|z_{li}| + \frac{1}{\nu_4}) - \frac{1}{2\nu_2}\sum_l (\boldsymbol{\alpha_l} - \nu_1 \mathbf{1})^T \mathbf{\Omega}^{(t)} (\boldsymbol{\alpha_l} - \nu_1\mathbf{1})\end{aligned}$$

where $\boldsymbol{\mu} = \boldsymbol{m}^{(t-1)} + \mathbf{W}^{(t-1)}\mathbf{Z}^{(t-1)}$, $\rho_{ij}^{(t)} = \mathbb{E}(\rho_{ij}|\mathbf{X}, \mathbf{W}^{(t-1)}, \mathbf{Z}^{(t-1)}, \boldsymbol{m}^{(t-1)}, \boldsymbol{\alpha}^{(t-1)}, \boldsymbol{\xi}^{(t-1)})$, and $\mathbf{\Omega}^{(t)} = \mathbb{E}(\omega_{ij}|\mathbf{X}, \mathbf{W}^{(t-1)}, \mathbf{Z}^{(t-1)}, \boldsymbol{m}^{(t-1)}, \boldsymbol{\alpha}^{(t-1)}, \boldsymbol{\xi}^{(t-1)})$. The detailed steps of the EM algorithm are explained as follows.

E-step for $\boldsymbol{\rho}$: If the data type is Gaussian,

$$\widetilde{\mathbb{E}}(\rho_j) = \frac{\zeta_j + n}{\zeta_j + \sum_i (x_{ji} - \mu_{ji})^2}.$$

Otherwise,

$$\widetilde{\mathbb{E}}(\rho_{ji}) = \frac{b_{ji}(e^{\mu_{ji}} - e^{\psi_{ji}})}{2(\mu_{ji} - \psi_{ji})(e^{\mu_{ji}} + e^{\psi_{ji}})}.$$

E-step for $\boldsymbol{\omega_{jk}}$.

$$\widetilde{\mathbb{E}}(\omega_{jk}) = \frac{2\nu_2 \alpha_\omega}{2\nu_2 b_\omega + \Sigma_l (\alpha_{jl} - \alpha_{kl})^2}, \quad j < k.$$

M-step for $\boldsymbol{Z}$: Solve the lasso problem

$$\mathbf{z}_{\cdot i} = \underset{\mathbf{z}}{\operatorname{argmin}} \left(\frac{1}{2} \mathbf{z}'\mathbf{W}'\widetilde{\mathbb{E}}(\mathbf{D}_i)\mathbf{W}\mathbf{z} - \mathbf{z}'\mathbf{W}'\widetilde{\mathbb{E}}(\mathbf{c}_i) + \sum_l \xi_{li}|z_l| \right),$$

where $i = 1, \dots, n$, $\mathbf{D}_i = \operatorname{diag}(\rho_{1i}, \dots, \rho_{pi})$ and $\mathbf{c}_i$ is the i-th column of the $p \times n$ matrix $\boldsymbol{\kappa} + \boldsymbol{\rho} \circ (\boldsymbol{\psi} - \mathbf{m}\mathbf{1}^\top)$. Here, $\circ$ denotes the Hadamard product.

M-step for $\boldsymbol{W}$: Solve the lasso problem

$$\mathbf{w}'_{j\cdot} = \underset{\mathbf{w}}{\operatorname{argmin}} \left(\frac{1}{2} \mathbf{w}'\mathbf{Z}\widetilde{\mathbb{E}}(\mathbf{G}_j)\mathbf{Z}'\mathbf{w} - \mathbf{w}'\mathbf{Z}\widetilde{\mathbb{E}}(\mathbf{f}_j) + \sum_l \lambda_{jl}|w_l| \right),$$

where $j = 1, \dots, p$, $\mathbf{G}_j = \operatorname{diag}(\rho_{j1}, \dots, \rho_{jn})$, and $\mathbf{f}'_j$ is the j-th row of $\boldsymbol{\kappa} + \boldsymbol{\rho} \circ (\boldsymbol{\psi} - \mathbf{m}\mathbf{1}^\top)$.

M-step for m:

$$m_j = \frac{-\mathbf{w}_j^{'}\mathbf{Z}\boldsymbol{\rho}_j + \boldsymbol{\psi}_j^{'}\boldsymbol{\rho}_j + \boldsymbol{\kappa}_j^{'}\mathbf{1}}{\boldsymbol{\rho}_j^{'}\mathbf{1}}, \quad j = 1, \dots, p.$$

M-step for $\boldsymbol{\alpha_l}$:

$$\boldsymbol{\alpha}_l^{(new)} = \boldsymbol{\alpha}_l^{(old)} - \mathbf{H}_l^{-1}\mathbf{g}_l.$$

Here $\mathbf{H}_l = -\operatorname{diag}(e^{\boldsymbol{\alpha}_l|w_l|}) - \frac{\boldsymbol{\Omega}}{\nu_2}$ and $\mathbf{g}_l = \mathbf{1}_{p\times 1} - e^{\boldsymbol{\alpha}_l|w_l|} - \frac{\boldsymbol{\Omega}(\alpha_l - v_1\mathbf{1})}{\nu_2}$, where $l = 1, \dots, L$,

M-step for $\boldsymbol{\xi}$.

$$\xi_{li} = \frac{\nu_3}{|z_{li}| + \frac{1}{\nu_4}}.$$

$\mathcal{C} \leftarrow \{j : x_j \text{ is Gaussian}\}$ and $\mathcal{D} \leftarrow \{j : x_j \text{ is discrete}\}$;
$t \leftarrow 0$;
repeat
  for $j \leftarrow 1$ **to** p **do**
    if $j \in \mathcal{D}$ **then** $\rho_{ji} \leftarrow \frac{b_j(e^{\mu_{ji}} - e^{\psi_{ji}})}{2(\mu_{ji}-\psi_{ji})(e^{\mu_{ji}}+e^{\psi_{ji}})}$ for $1 \leq i \leq n$;
    else $\rho_j \leftarrow \frac{\zeta_j + n}{\zeta_j + \sum_i (x_{ji} - \mu_{ji})^2}$;
  end
  for $i \leftarrow 1$ **to** n **do**
  $\mathbf{z}_{\cdot i} \leftarrow \text{argmin}_{\mathbf{z}} \left(\frac{1}{2}\mathbf{z}'(\mathbf{W})' D_i \mathbf{W}\mathbf{z} - \mathbf{z}'(\mathbf{W})'\mathbf{c}_i + \sum_l \xi_{li}|z_l|\right)$;
  for $j \leftarrow 1$ **to** p **do**
    $(\mathbf{w}_{j\cdot})' \leftarrow \text{argmin}_{\mathbf{w}} \left(\frac{1}{2}\mathbf{w}'\mathbf{Z}G_j(\mathbf{Z})'\mathbf{w} - \mathbf{w}'\mathbf{Z_j} + \sum_{\mathbf{l}} \lambda_{\mathbf{jl}}|\mathbf{w_l}|\right)$;
    $m_j \leftarrow \frac{-\mathbf{w}_j'\mathbf{Z}\boldsymbol{\rho}_j + \boldsymbol{\psi}_j'\boldsymbol{\rho}_j + \boldsymbol{\kappa}_j'\mathbf{1}}{\boldsymbol{\rho}_j'\mathbf{1}}$.
  end
  $\phi_{jl} \leftarrow \begin{cases} \frac{N(e^{\alpha_{jl}} - |w_{jl}|/N)}{2(\alpha_{jl} + \log(|w_{jl}|/N))(e^{\alpha_{jl}} + |w_{jl}|/N)}, & w_{jl} \neq 0, \\ 0, & w_{jl} = 0; \end{cases}$
  $\boldsymbol{\alpha}_l \leftarrow (\boldsymbol{\Omega}/\nu_2 + \Phi_l)^{-1}(\boldsymbol{\Omega}\underline{1}\nu_1/\nu_2 + \Phi_l\boldsymbol{\chi}_l + (1 - N/2)\underline{1})$;
  $\omega_{jk} \leftarrow \frac{2\nu_2 a_\omega}{2\nu_2 b_\omega + \sum_l (\alpha_{jl} - \alpha_{kl})^2}, \qquad 1 \leq j < k \leq p, (j,k) \in E$;
  $t \leftarrow t + 1$;
until convergence;

Algorithm 1: EM algorithm for GBC with Adaptive Structured Shrinkage prior

2 Computation performance

Our algorithm provides great computational performance and is scalable to high dimensional settings for genomic data analysis since EM algorithm is used. We benchmark the computational performances of our algorithm on a computer with 4GB RAM and Intel Core i5 CPU. For a moderate dataset with 1000 genes and 300 samples, one run from GBC takes 54.8 seconds and from sGBC takes 62.6 seconds (biological information includes 4764 edges).

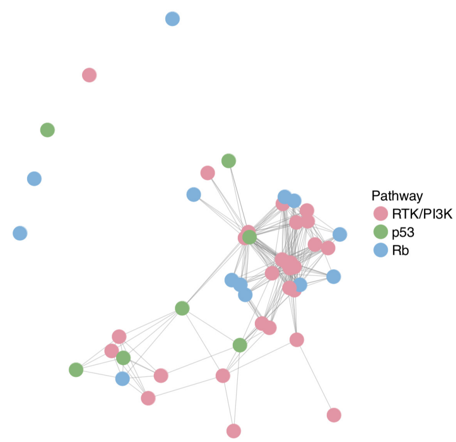

Figure S1: Interactions of 48 genes that overlap with the three critical signaling pathways - RTK/PI3K, p53, and Rb, which closely relate with migration, survival and apoptosis progression of cell cycles. This gene network information is extracted from the KEGG pathway and is utilized in the integrative analysis by the proposed method (section 4.4).

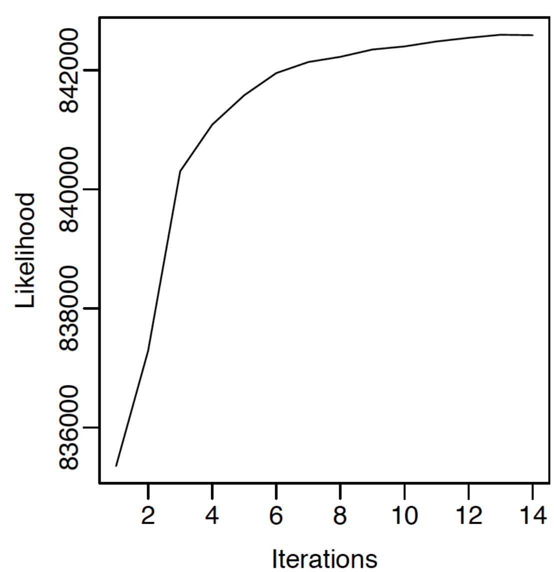

Figure S2: Graph of likelihood by number of iterations to show that GBC converges very quickly. A synthetic data set with 300 samples and 1000 variables are used to plot. Overlap is set to 0 and no biological information is used.

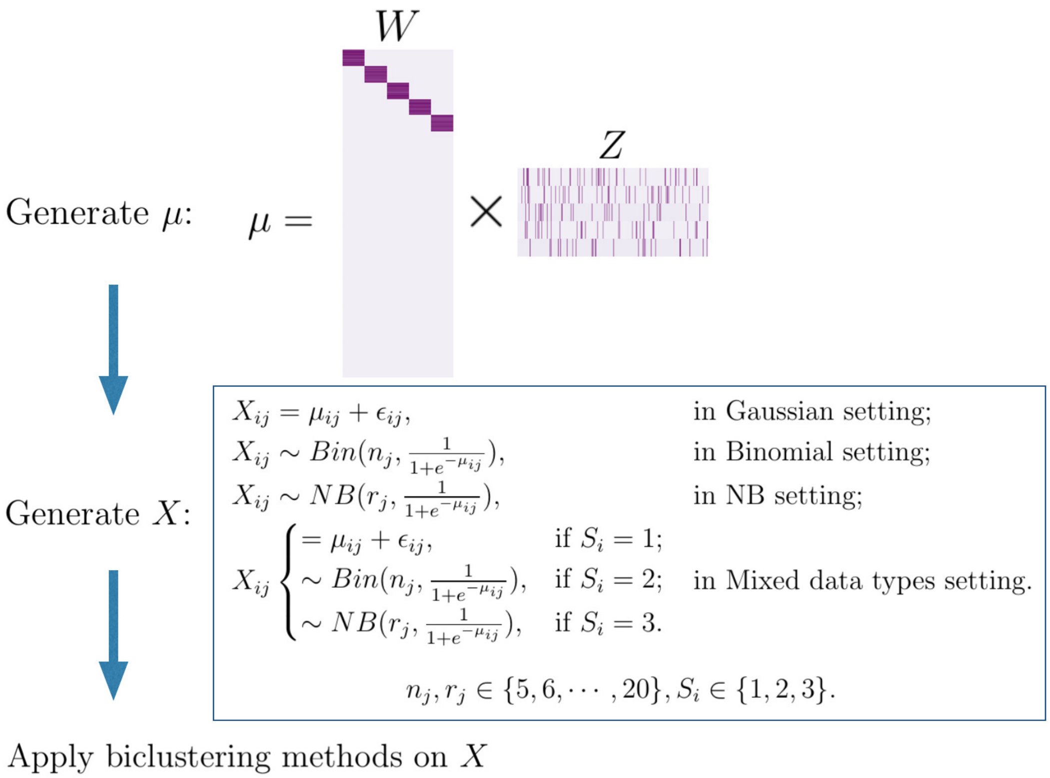

Figure S3: Work flow of the simualtion study.

		Negative Binomial				
overlap	Method	CE	CS	SEN	SPE	MCC
0	Plaid	0.012(1e-03)	0.15(2e-02)	0.35(3e-02)	1(0e+00)	0.04(6e-03)
	CC	0.00043(3e-04)	0.00039(3e-04)	0.00047(4e-04)	1(3e-05)	-7.7e-05(1e-03)
	FABIA	0.21(3e-02)	0.21(3e-02)	0.21(3e-02)	1(1e-04)	0.42(6e-02)
	XMotifs	0.0028(1e-03)	0.0028(1e-03)	0.0031(1e-03)	1(4e-05)	0.0075(4e-03)
	ISA	0(0e+00)	0(0e+00)	0(0e+00)	0(0e+00)	0(0e+00)
	GBC	0.49(2e-01)	0.53(1e-01)	1(9e-04)	0.98(1e-02)	0.72(8e-02)
	sGBC	0.48(2e-01)	0.52(1e-01)	1(0e+00)	0.97(1e-02)	0.71(8e-02)
15	Plaid	0.036(3e-02)	0.089(3e-02)	0.26(5e-02)	1(1e-02)	0.09(6e-02)
	CC	0.00023(2e-04)	0.00022(2e-04)	0.00025(3e-04)	1(3e-05)	-0.00083(9e-04)
	FABIA	0.17(3e-02)	0.17(3e-02)	0.17(3e-02)	1(1e-04)	0.38(5e-02)
	XMotifs	0.0024(1e-03)	0.0024(1e-03)	0.0027(1e-03)	1(3e-05)	0.007(5e-03)
	ISA	0.0039(3e-03)	0.0035(3e-03)	0.004(3e-03)	1(6e-05)	0.024(2e-02)
	GBC	0.42(2e-01)	0.47(1e-01)	1(8e-04)	0.97(2e-02)	0.69(9e-02)
	sGBC	0.49(1e-01)	0.53(1e-01)	1(4e-04)	0.97(1e-02)	0.73(6e-02)

Table S1: Simulation results for Negative Binomial settings. Results are generated based on 100 simulated datasets: mean(sd).

Mixed						
overlap	Method	CE	CS	SEN	SPE	MCC
0	Plaid	0.011(1e-03)	0.073(1e-02)	0.23(3e-02)	1(1e-02)	0.027(6e-03)
	CC	6.3e-05(9e-05)	6e-05(9e-05)	6.8e-05(1e-04)	1(2e-05)	-0.0011(4e-04)
	FABIA	0.1(2e-02)	0.1(2e-02)	0.11(2e-02)	1(5e-04)	0.3(5e-02)
	XMotifs	1.2e-06(1e-05)	1.1e-06(1e-05)	1.2e-06(1e-05)	1(4e-05)	-0.00012(3e-04)
	ISA	0(0e+00)	0(0e+00)	0(0e+00)	0(0e+00)	0(0e+00)
	GBC	0.48(2e-01)	0.51(1e-01)	0.85(7e-02)	0.98(1e-02)	0.69(9e-02)
	sGBC	0.7(1e-01)	0.71(1e-01)	0.99(1e-02)	0.99(6e-03)	0.84(6e-02)
15	Plaid	0.019(1e-02)	0.043(1e-02)	0.16(3e-02)	1(1e-02)	0.042(3e-02)
	CC	4.1e-05(7e-05)	4e-05(7e-05)	4.4e-05(7e-05)	1(3e-05)	-0.0013(3e-04)
	FABIA	0.1(2e-02)	0.1(2e-02)	0.1(2e-02)	1(7e-04)	0.29(6e-02)
	XMotifs	5.1e-06(3e-05)	4.7e-06(3e-05)	5.2e-06(3e-05)	1(5e-05)	-0.00014(4e-04)
	ISA	0(0e+00)	0(0e+00)	0(0e+00)	0(0e+00)	0(0e+00)
	GBC	0.51(1e-01)	0.53(1e-01)	0.89(6e-02)	0.98(1e-02)	0.72(7e-02)
	sGBC	0.64(1e-01)	0.66(8e-02)	0.97(3e-02)	0.99(6e-03)	0.81(5e-02)

Table S2: Simulation results for mixed data types. Results are generated based on 100 simulated datasets: mean(sd).