A Study of Biclustering Coherence Measures for Gene Expression Data – Supplementary Material

Victor A. Padilha and André C. P. L. F. de Carvalho
Institute of Mathematics and Computer Sciences
University of São Paulo, São Carlos, SP, Brazil
Email: victorpadilha@usp.br, andre@icmc.usp.br

I. Algorithms

The 9 algorithms selected for this study are:

- *Cheng and Church's Algorithm (CCA)* [1], which starts with the whole data matrix as a bicluster and iteratively optimizes the MSR measure by removing or adding rows or columns to it.
- *Statistical-Algorithmic Method for Bicluster Analysis (SAMBA)* [2], which constructs a bipartite graph for the input dataset and searches for biclusters represented as statistically significant complete bipartite subgraphs.
- *Order Preserving Sub-Matrix (OPSM)* [3], which consists of a greedy procedure that finds order preserving biclusters by expanding their columns.
- *Plaid* [4], that fits a linear layer model of the dataset, where each layer corresponds to a bicluster, through a binary least squares procedure.
- *Binary Inclusion Maximal Biclustering Algorithm (Bimax)* [5], which discretizes the input dataset into a binary matrix and searches for biclusters whose values are all equal to one.
- *Bayesian Biclustering (BBC)* [6], that fits the plaid model to the input dataset using a Gibbs sampling procedure.
- *Large Average Submatrices (LAS)* [7], which assumes a Gaussian random matrix as a null model for the data and searches for biclusters with average values that significantly deviate from such a model.
- *Qualitative Biclustering (QUBIC)* [8], that represents the data as a graph, with genes as vertices, edge weights equal to the number of samples for which two genes are similar and a bicluster is equivalent to a heavy subgraph.
- *Factor Analysis for Bicluster Acquisition (FABIA)* [9], which fits a sum of multiplicative models of the dataset, where each model is a bicluster, using a maximum-likelihood approach.

II. Software packages

All the implementations of the algorithms and the measures used in this paper are publicly available. Python implementations and wrappers for publicly available software were run with biclustlib[1] [10]. The GO statistical evaluation was performed with a modified version of the GOATOOLS package [11] that allows over-representation tests[2]. At last, all the coherence bicluster measures were implemented in Python and are freely available in https://padilha.github.io/bracis-2018-suppl/.

III. Notation

Let $A = (R, C)$ be a gene expression matrix, where R is a set of N rows (genes) and C is a set of M columns (samples). A bicluster consists of a submatrix $B = (I, J)$, $I \subseteq R$, $J \subseteq C$.

The ith row, the jth column and each element of B are denoted as a_{i*}, a_{*j} and a_{ij}, respectively.

Coherence measures may include row means, column means or the bicluster mean in their definitions. In this paper, a_{iJ}, a_{Ij} and a_{IJ} correspond to the mean of the ith row, the jth column and all elements of B, respectively.

Other coherence measures are based on the Pearson or Spearman correlations. In this material, the former is denoted as $r(\cdot,\cdot)$ while the latter is represented as $\rho(\cdot,\cdot)$.

IV. Variance-based measures

A. Variance (VAR) [12]

$$\text{VAR}(B) = \frac{1}{|I||J|}\sum_{i=1}^{|I|}\sum_{j=1}^{|J|}(a_{ij} - a_{IJ})^2 \tag{1}$$

Time complexity analysis. The calculation of a_{IJ} costs $O(|I||J|)$. The sum of the squared differences between the bicluster elements and a_{IJ} costs $O(|I||J|)$. Thus, the time complexity of VAR is $O(|I||J|)$.

B. Mean Squared Residue (MSR) [1]

$$\text{MSR}(B) = \frac{1}{|I||J|}\sum_{i=1}^{|I|}\sum_{j=1}^{|J|}(a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2 \tag{2}$$

Time complexity analysis. The calculations of $a_{iJ}\,\forall i \in I$, $a_{Ij}\,\forall j \in J$ and a_{IJ} cost $O(|I||J|)$ each. The sum of the squared differences between the bicluster elements and their expected values predicted by the row, column and bicluster means costs $O(|I||J|)$. So, the time complexity of MSR is $O(|I||J|)$.

[1] https://github.com/padilha/biclustlib

[2] https://github.com/padilha/goatools

C. Mean Absolute Residue (MAR) [13]

$$\text{MAR}(B) = \frac{1}{|I||J|}\sum_{i=1}^{|I|}\sum_{j=1}^{|J|}|a_{ij} - a_{iJ} - a_{Ij} + a_{IJ}| \tag{3}$$

Time complexity analysis. The only distinction between the MAR and MSR measures is the use of the absolute value instead of the square of the differences. Thus, the time of complexity of MAR is $O(|I||J|)$.

D. Relevance Index (RI) [14]

$$\text{RI}(B) = \sum_{j=1}^{|J|} R_j, \tag{4}$$

$$R_j = 1 - \frac{\sigma_{Ij}^2}{\sigma_{Rj}^2}, \tag{5}$$

where σ_{Ij}^2 corresponds to the variance of the jth column of B and σ_{Rj}^2 indicates the variance of the jth column of the full dataset.

Time complexity analysis. The calculation of each σ_{Ij}^2 costs $O(|I|)$. The calculation of each σ_j^2 costs $O(N)$. So, any R_j requires $O(|I|) + O(N) = O(N)$ steps and the RI measure runs in $O(N|J|)$.

E. Overall Constancy (OC) [15]

$$\text{OC}(B) = \frac{|I|C_R(B) + |J|C_C(B)}{|I| + |J|} \tag{6}$$

where

$$C_R(B) = \frac{1}{|I|}\sum_{i=1}^{|I|-1}\sum_{k=i+1}^{|I|} d(a_{i*}, a_{k*}), \tag{7}$$

$$C_C(B) = \frac{1}{|J|}\sum_{j=1}^{|J|-1}\sum_{l=j+1}^{|J|} d(a_{*j}, a_{*l}), \tag{8}$$

and $d(\cdot,\cdot)$ is the Euclidean distance.

Time complexity analysis. Each $d(a_{i*}, a_{k*})$ and each $d(a_{*j}, a_{*l})$ cost $O(|J|)$ and $O(|I|)$, respectively. $C_R(B)$ and $C_C(B)$ cost $O(|I|^2\,|J|)$ and $O(|I|\,|J|^2)$, respectively. In all, the time complexity of OC is $O(|I|^2\,|J| + |I|\,|J|^2)$

F. Scaling Mean Squared Residue (SMSR) [16]

$$\text{SMSR}(B) = \frac{1}{|I||J|}\sum_{i=1}^{|I|}\sum_{j=1}^{|J|}\frac{(a_{iJ}\,a_{Ij} - a_{ij}\,a_{IJ})^2}{a_{iJ}^2\,a_{Ij}^2}. \tag{9}$$

Time complexity analysis. SMSR requires the same quantities as MSR and MAR ($a_{iJ}\,\forall i \in I$, $a_{Ij}\,\forall j \in J$ and a_{IJ}) to determine the differences among the values of the bicluster elements and their expected values. Therefore, the complexity of SMSR is $O(|I||J|)$.

G. Minimal Mean Squared Error (MMSE) [17]

$$\text{MMSE}(B) = \sum_{i=1}^{|I|}\sum_{j=1}^{|J|}(a_{ij} - a_{iJ})^2 - \lambda_{\max}(BB^T), \tag{10}$$

where $\lambda_{\max}(BB^T)$ is the eigenvalue of BB^T with maximum absolute value.

The time complexity of MMSE is $O(\min(|I|, |J|)\,|I||J|)$. The complete analysis is provided in the original paper [17].

V. Correlation-based measures

A. Average Correlation (AC) [18]

$$\text{AC}(B) = \frac{2}{|I|(|I|-1)}\sum_{i=1}^{|I|-1}\sum_{k=i+1}^{|I|} r(a_{i*}, a_{k*}). \tag{11}$$

Time complexity analysis. The calculation of each $r(a_{i*}, a_{k*})$ costs $O(|J|)$. There are $|I|\,(|I|-1)\,/\,2 = O(|I|^2)$ different pairs of rows in B. So, AC requires $O(|I|^2\,|J|)$ steps.

B. Sub-matrix Correlation Score (SCS) [19]

$$\text{SCS}(B) = \min\{S_{\text{row}}(B), S_{\text{col}}(B)\} \tag{12}$$

where

$$S_{\text{row}} = \min_{i=1,\cdots,|I|}\left\{1 - \frac{1}{|I|-1}\sum_{\substack{k=1\\k\neq i}}^{|I|}|r(a_{i*}, a_{k*})|\right\} \tag{13}$$

$$S_{\text{col}} = \min_{j=1,\cdots,|J|}\left\{1 - \frac{1}{|J|-1}\sum_{\substack{l=1\\l\neq j}}^{|J|}|r(a_{*j}, a_{*l})|\right\} \tag{14}$$

Time complexity analysis. The calculation of each $r(a_{i*}, a_{k*})$ and each $r(a_{*j}, a_{*l})$ costs $O(|J|)$ and $O(|I|)$, respectively. The calculations of all S_{row} values and all S_{col} values require $O(|I|^2\,|J|)$ and $O(|I|\,|J|^2)$ steps, respectively. In all, the time complexity of SCS is $O(|I|^2\,|J| + |I|\,|J|^2)$.

C. Average Correlation Value (ACV) [20]

$$\text{ACV}(B) = \max\left\{\frac{2}{|I|(|I|-1)}\sum_{i=1}^{|I|-1}\sum_{k=i+1}^{|I|}|r(a_{i*}, a_{k*})|,\right.$$
$$\left.\frac{2}{|J|(|J|-1)}\sum_{j=1}^{|J|-1}\sum_{l=j+1}^{|J|}|r(a_{*j}, a_{*l})|,\right\}. \tag{15}$$

Time complexity analysis. The average absolute correlation among the rows of B requires $O(|I|^2\,|J|)$ steps. The average absolute correlation between the columns of B costs $O(|I|\,|J|^2)$. So, ACV runs in $O(|I|^2\,|J| + |I|\,|J|^2)$.

D. Average Spearman's Rho (ASR) [21]

$$\text{ASR}(B) = \max\left\{\frac{2}{|I|(|I|-1)}\sum_{i=1}^{|I|-1}\sum_{k=i+1}^{|I|}\rho(a_{i*}, a_{k*}), \right. \\ \left. \frac{2}{|J|(|J|-1)}\sum_{j=1}^{|J|-1}\sum_{l=j+1}^{|J|}\rho(a_{*j}, a_{*l})\right\}. \tag{16}$$

Time complexity analysis. Each $\rho(a_{i*}, a_{k*})$ and each $\rho(a_{*j}, a_{*l})$ cost $O(|J|\log|J|)$ and $O(|I|\log|I|)$, respectively[3]. The first and the second term of $\max$ are performed in $O(|I|^2\,|J|\log|J|)$ and $O(|J|^2\,|I|\log|I|)$ steps, respectively. Thus, ASR runs in $O(|I|^2\,|J|\log|J| + |J|^2\,|I|\log|I|)$.

E. Spearman's Biclustering Measure (SBM) [22]

$$\text{SBM}(B) = \alpha(B)\;\beta(B)\;\bar{\rho}_I(B)\;\bar{\rho}_J(B) \tag{17}$$

where

$$\bar{\rho}_I(B) = \frac{2}{|I|(|I|-1)}\sum_{i=1}^{|I|-1}\sum_{k=i+1}^{|I|}|\rho(a_{i*}, a_{k*})|, \tag{18}$$

$$\bar{\rho}_J(B) = \frac{2}{|J|(|J|-1)}\sum_{j=1}^{|J|-1}\sum_{l=j+1}^{|J|}|\rho(a_{*j}, a_{*l})|, \tag{19}$$

$\alpha(B)$ and $\beta(B)$ are parameters that refer to the importance of the rows and the columns of a bicluster.

Time complexity analysis. SBM is calculated in constant time after $\bar{\rho}_I(B)$ and $\bar{\rho}_J(B)$ are obtained. Therefore, SBM has the same time complexity of ASR: $O(|I|^2\,|J|\log|J| + |J|^2\,|I|\log|I|)$.

VI. Standartization-based measures

The measures included in this category are calculated on the standartized bicluster B', whose elements are defined as:

$$a'_{ij} = \frac{a_{ij} - \mu_i}{\sigma_i}, \tag{20}$$

where μ_i and σ_i are the mean and the standard deviation of the ith row (gene), respectively.

A. Maximal Standard Area (MSA) [23]

$$\text{MSA}(B) = \sum_{j=1}^{|J|-1}\left|\frac{M_j^{B'} - m_j^{B'} + M_{j+1}^{B'} - m_{j+1}^{B'}}{2}\right| \tag{21}$$

where $M_j^{B'}$ and $m_j^{B'}$ correspond to the maximum and minimum values of the jth column of the standartized bicluster B', respectively.

Time complexity analysis. $M_j^{B'}$ and $m_j^{B'}$ require $O(|I|)$ steps. Since we have $|J|$ columns in the bicluster, MSA runs in $O(|I|\,|J|)$.

B. Virtual Error (VE) [24]

$$\text{VE}(B) = \frac{1}{|I||J|}\sum_{i=1}^{|I|}\sum_{j=1}^{|J|}|a'_{ij} - p'_j| \tag{22}$$

where

$$p = \frac{1}{|I|}\sum_{i=1}^{|I|}a_{i*}, \tag{23}$$

and $\hat{p}$ is calculated with Equation 20.

Time complexity analysis. p requires $O(|I||J|)$ steps to be calculated. The standartization of B takes $O(|I||J|)$ steps. The standartization of p costs $O(|J|)$. The absolute differences between the elements of B' and the elements of p' require $O(|I||J|)$. So, VE runs in $O(|I||J|)$.

C. Transposed Virtual Error (VEt) [25]

$$\text{VEt}(B) = \frac{1}{|I||J|}\sum_{i=1}^{|I|}\sum_{j=1}^{|J|}|\hat{a}_{ij} - \hat{p}_i| \tag{24}$$

where

$$p = \frac{1}{|J|}\sum_{j=1}^{|J|}a_{*j}, \tag{25}$$

Time complexity analysis. VEt is the transposed version of VE. Thus, it requires the same number of steps: $O(|I||J|)$.

References

[1] Y. Cheng and G. M. Church, "Biclustering of expression data." in *ISMB*, vol. 8. AAAI Press, 2000, pp. 93–103.

[2] A. Tanay, R. Sharan, and R. Shamir, "Discovering statistically significant biclusters in gene expression data," *Bioinformatics*, vol. 18, no. suppl_1, pp. S136–S144, 2002.

[3] A. Ben-Dor, B. Chor, R. Karp, and Z. Yakhini, "Discovering local structure in gene expression data: the order-preserving submatrix problem," *Journal of computational biology*, vol. 10, no. 3-4, pp. 373–384, 2003.

[4] H. Turner, T. Bailey, and W. Krzanowski, "Improved biclustering of microarray data demonstrated through systematic performance tests," *Computational statistics & data analysis*, vol. 48, no. 2, pp. 235–254, 2005.

[5] A. Prelić, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler, "A systematic comparison and evaluation of biclustering methods for gene expression data," *Bioinformatics*, vol. 22, no. 9, pp. 1122–1129, 2006.

[6] J. Gu and J. S. Liu, "Bayesian biclustering of gene expression data," *BMC genomics*, vol. 9, no. 1, p. S4, 2008.

[7] A. A. Shabalin, V. J. Weigman, C. M. Perou, A. B. Nobel *et al.*, "Finding large average submatrices in high dimensional data," *The Annals of Applied Statistics*, vol. 3, no. 3, pp. 985–1012, 2009.

[8] G. Li, Q. Ma, H. Tang, A. H. Paterson, and Y. Xu, "Qubic: a qualitative biclustering algorithm for analyses of gene expression data," *Nucleic acids research*, vol. 37, no. 15, pp. e101–e101, 2009.

[9] S. Hochreiter, U. Bodenhofer, M. Heusel, A. Mayr, A. Mitterecker, A. Kasim, T. Khamiakova, S. Van Sanden, D. Lin, W. Talloen *et al.*, "Fabia: factor analysis for bicluster acquisition," *Bioinformatics*, vol. 26, no. 12, pp. 1520–1527, 2010.

[10] V. A. Padilha and R. J. G. B. Campello, "A systematic comparative evaluation of biclustering techniques," *BMC bioinformatics*, vol. 18, no. 1, p. 55, 2017.

[11] H. Tang, D. Klopfenstein, B. Pedersen, P. Flick, K. Sato, F. Ramirez, J. Yunes, and C. Mungall, "Goatools: tools for gene ontology," *Zenodo. doi*, vol. 10, 2015.

[12] J. A. Hartigan, "Direct clustering of a data matrix," *Journal of the american statistical association*, vol. 67, no. 337, pp. 123–129, 1972.

[3] Note that the Speaman coefficient measures the correlation between the ranks of the elements of two vectors. For such, it requires a sorting step, which can be performed in $O(n\log n)$ for n elements.

[13] J. Yang, W. Wang, H. Wang, and P. Yu, "δ-clusters: capturing subspace correlation in a large data set," in *IEEE ICDE*. IEEE, 2002, pp. 517–528.

[14] K. Y. Yip, D. W. Cheung, and M. K. Ng, "Harp: A practical projected clustering algorithm," *IEEE TKDE*, vol. 16, no. 11, pp. 1387–1397, 2004.

[15] R. Santamaría, L. Quintales, and R. Therón, "Methods to bicluster validation and comparison in microarray data," in *International Conference on Intelligent Data Engineering and Automated Learning*. Springer, 2007, pp. 780–789.

[16] A. Mukhopadhyay, U. Maulik, and S. Bandyopadhyay, "A novel coherence measure for discovering scaling biclusters from gene expression data," *Journal of bioinformatics and computational biology*, vol. 7, no. 05, pp. 853–868, 2009.

[17] S. Chen, J. Liu, and T. Zeng, "Measuring the quality of linear patterns in biclusters," *Methods*, vol. 83, pp. 18–27, 2015.

[18] J. A. Nepomuceno, A. Troncoso, and J. S. Aguilar-Ruiz, "Biclustering of gene expression data by correlation-based scatter search," *BioData mining*, vol. 4, no. 1, p. 3, 2011.

[19] W.-H. Yang, D.-Q. Dai, and H. Yan, "Finding correlated biclusters from gene expression data," *IEEE TKDE*, vol. 23, no. 4, pp. 568–584, 2011.

[20] L. Teng and L. Chan, "Discovering biclusters by iteratively sorting with weighted correlation coefficient in gene expression data," *Journal of Signal Processing Systems*, vol. 50, no. 3, pp. 267–280, 2008.

[21] W. Ayadi, M. Elloumi, and J.-K. Hao, "A biclustering algorithm based on a bicluster enumeration tree: application to dna microarray data," *BioData mining*, vol. 2, no. 1, p. 9, 2009.

[22] J. L. Flores, I. Inza, P. Larrañaga, and B. Calvo, "A new measure for gene expression biclustering based on non-parametric correlation," *Computer methods and programs in biomedicine*, vol. 112, no. 3, pp. 367–397, 2013.

[23] R. Giraldez, F. Divina, B. Pontes, and J. S. Aguilar-Ruiz, "Evolutionary search of biclusters by minimal intrafluctuation," in *FUZZ-IEEE*. IEEE, 2007, pp. 1–6.

[24] B. Pontes, F. Divina, R. Giráldez, and J. S. Aguilar-Ruiz, "Virtual error: a new measure for evolutionary biclustering," in *5th EvoBIO*. Springer, 2007, pp. 217–226.

[25] B. Pontes, R. Giráldez, and J. S. Aguilar-Ruiz, "Measuring the quality of shifting and scaling patterns in biclusters," in *IAPR PRIB*. Springer, 2010, pp. 242–252.