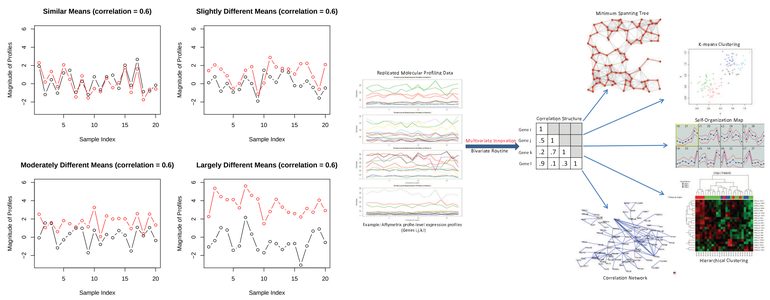
A Generalized Multivariate Approach for Correlation-based Pattern Discovery from Replicated Molecular Profiling Data

Dongxiao Zhu, Guorong Xu and Lipi R. Acharya

Department of Computer Science
University of New Orleans, New Orleans, LA 70148

Problem Statement

Correlation-based pattern discovery from replicated molecular profiling data enables essential data mining tasks. Unfortunately, the existing approaches are not tailored to analyze replicated measurements, which is further confused by various replication mechanisms. With few exception, existing approaches average or summarize over replicates of diverse magnitude, which might wipe out important patterns of low magnitude and/or cancel out patterns of similar magnitude.



(a) Left panel: correlation is scale-free. (b) Right panel: (left) Four replicated molecular profiles with 11 replicates in each. The magnitude of each sibling molecular profile (one color curve) differs significantly from the others. (middle) A scale-free correlation matrix of four replicated molecular profiles. (right) Five popular correlation-based pattern discovery algorithms. Our *multivariate innovation* is to estimate multivariate scale-free correlation structure from replicated molecular profiles directly.

Goals:

- Develop a multivariate parsimonious correlation model for replicated molecular profiling data with blind replication mechanisms.
- Develop a constrained (less parsimonious) correlation model explicitly considering the informed replication mechanisms.
- Develop a correlation-based pattern discovery software with Graphical User Interface (GUI) for analyzing replicated molecular profiling data.

Strategies and existing approaches

Instead of averaging over replicates, our strategies are:

- Treat each replicate individually.
- Parsimonious Multivariate Gaussian Model (PMGM) for general-sense replicates with unknown replication mechanisms (Zhu et al 2007).
- Infinite Bayesian Mixture Model (IBMM) (Medvedovic et al 2004).
- Constrained (less parsimonious) multivariate normal model for narrow-sense replicates with known replication mechanism (this work).
- Maximum Likelihood (ML) parameter estimation (this work).

Statistical models

Parsimonious model for blind replication mechanisms: Without *a priori* known replication mechanisms, $(m_1 + m_2)$ by $(m_1 + m_2)$ parsimonious correlation matrix Σ_p given by

$$\Sigma_p = \begin{pmatrix} 1 & \dots & \rho_x & \rho & \dots & \rho \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \rho_x & \dots & 1 & \rho & \dots & \rho \\ \rho & \dots & \rho & 1 & \dots & \rho_y \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \rho & \dots & \rho & \rho_y & \dots & 1 \end{pmatrix} = \begin{bmatrix} \Sigma_x^b & \Sigma_{xy}^b \\ \Sigma_{xy}^{b\ T} & \Sigma_y^b \end{bmatrix}. \quad (1)$$

The Maximum Likelihood Estimates (MLE's) of the parameters Σ_p are given by

$$\hat{\Sigma_p} = \frac{1}{n}\sum_{j=1}^{n}(Z_j - \hat{\mu})(Z_j - \hat{\mu})^T. \quad (2)$$

Constrained model for informed replication mechanisms: With *a priori* known replication mechanisms, e.g., 3 biological replicates with 2 technical replicates nested within each. Σ is 12×12 constrained (less-parsimonious) correlation matrix (Σ_c) with 13 parameters given by

$$\Sigma_c = \begin{pmatrix} 1 & \rho^{tt} & \rho_x^{12} & \rho_x^{12} & \rho_x^{13} & \rho_x^{13} & \rho_{xy}^{11} & \rho_{xy}^{11} & \rho_{xy}^{12} & \rho_{xy}^{12} & \rho_{xy}^{13} & \rho_{xy}^{13} \\ \rho^{tt} & 1 & \rho_x^{12} & \rho_x^{12} & \rho_x^{13} & \rho_x^{13} & \rho_{xy}^{11} & \rho_{xy}^{11} & \rho_{xy}^{12} & \rho_{xy}^{12} & \rho_{xy}^{13} & \rho_{xy}^{13} \\ \rho_x^{21} & \rho_x^{21} & 1 & \rho^{tt} & \rho_x^{23} & \rho_x^{23} & \rho_{xy}^{21} & \rho_{xy}^{21} & \rho_{xy}^{22} & \rho_{xy}^{22} & \rho_{xy}^{23} & \rho_{xy}^{23} \\ \rho_x^{21} & \rho_x^{21} & \rho^{tt} & 1 & \rho_x^{23} & \rho_x^{23} & \rho_{xy}^{21} & \rho_{xy}^{21} & \rho_{xy}^{22} & \rho_{xy}^{22} & \rho_{xy}^{23} & \rho_{xy}^{23} \\ \rho_x^{31} & \rho_x^{31} & \rho_x^{32} & \rho_x^{32} & 1 & \rho^{tt} & \rho_{xy}^{31} & \rho_{xy}^{31} & \rho_{xy}^{32} & \rho_{xy}^{32} & \rho_{xy}^{33} & \rho_{xy}^{33} \\ \rho_x^{31} & \rho_x^{31} & \rho_x^{32} & \rho_x^{32} & \rho^{tt} & 1 & \rho_{xy}^{31} & \rho_{xy}^{31} & \rho_{xy}^{32} & \rho_{xy}^{32} & \rho_{xy}^{33} & \rho_{xy}^{33} \\ \rho_{xy}^{11} & \rho_{xy}^{11} & \rho_{xy}^{21} & \rho_{xy}^{21} & \rho_{xy}^{31} & \rho_{xy}^{31} & 1 & \rho^{tt} & \rho_y^{12} & \rho_y^{12} & \rho_y^{13} & \rho_y^{12} \\ \rho_{xy}^{11} & \rho_{xy}^{11} & \rho_{xy}^{21} & \rho_{xy}^{21} & \rho_{xy}^{31} & \rho_{xy}^{31} & \rho^{tt} & 1 & \rho_y^{12} & \rho_y^{12} & \rho_y^{13} & \rho_y^{12} \\ \rho_{xy}^{12} & \rho_{xy}^{12} & \rho_{xy}^{22} & \rho_{xy}^{22} & \rho_{xy}^{32} & \rho_{xy}^{32} & \rho_y^{21} & \rho_y^{21} & 1 & \rho^{tt} & \rho_y^{23} & \rho_y^{23} \\ \rho_{xy}^{12} & \rho_{xy}^{12} & \rho_{xy}^{22} & \rho_{xy}^{22} & \rho_{xy}^{32} & \rho_{xy}^{32} & \rho_y^{21} & \rho_y^{21} & \rho^{tt} & 1 & \rho_y^{23} & \rho_y^{23} \\ \rho_{xy}^{13} & \rho_{xy}^{13} & \rho_{xy}^{23} & \rho_{xy}^{23} & \rho_{xy}^{33} & \rho_{xy}^{33} & \rho_y^{31} & \rho_y^{31} & \rho_y^{32} & \rho_y^{32} & 1 & \rho^{tt} \\ \rho_{xy}^{13} & \rho_{xy}^{13} & \rho_{xy}^{23} & \rho_{xy}^{23} & \rho_{xy}^{33} & \rho_{xy}^{33} & \rho_y^{31} & \rho_y^{31} & \rho_y^{32} & \rho_y^{32} & \rho^{tt} & 1 \end{pmatrix} = \begin{bmatrix} \Sigma_x^{(3,2)} & \Sigma_{xy}^{(3,2)} \\ \Sigma_{xy}^{(3,2)T} & \Sigma_y^{(3,2)T} \end{bmatrix}. \quad (3)$$

The superscript $(3, 2)$ specifies the underlying replication mechanisms, i.e., number of biological replicates followed by number of technical replicates. The MLE of Σ_c is given by eq. 4,

$$\hat{\Sigma_c} = \frac{1}{n}\sum_{j=1}^{n}\begin{bmatrix} (Z_j^{m_1} - \hat{\mu}^{m_1})(Z_j^{m_1} - \hat{\mu}^{m_1})^T & (Z_j^{m_1} - \hat{\mu}^{m_1})(Z_j^{m_2} - \hat{\mu}^{m_2})^T \\ (Z_j^{m_2} - \hat{\mu}^{m_2})(Z_j^{m_1} - \hat{\mu}^{m_1})^T & (Z_j^{m_2} - \hat{\mu}^{m_2})(Z_j^{m_2} - \hat{\mu}^{m_2})^T \end{bmatrix}, \quad (4)$$

where $Z_j^{m_1}$ and $Z_j^{m_2}$ represent the vectors formed by the first m_1 and the last m_2 entries in Z_j. The notations $\hat{\mu}^{m_1}$ and $\hat{\mu}^{m_2}$ have similar explanations corresponding to $\hat{\mu}$. The experimental design information was specified by J_{m_1}, J_{m_2} (numbers of biological replicates) and I_{m_1}, I_{m_2} (numbers of technical replicates nested within each biological replicate).

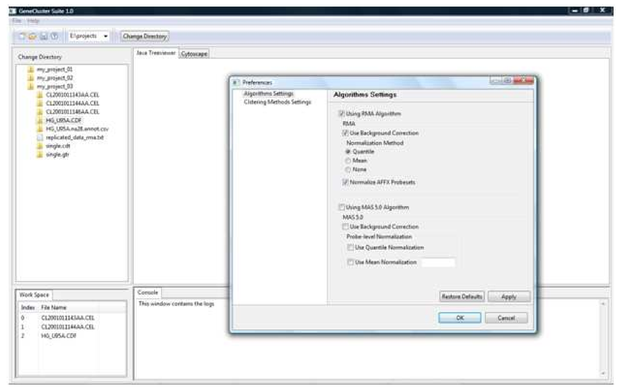
Summarization of a correlation structure: The Likelihood Ratio (LR) static for summarizing a correlation structure Σ against a null correlation structure Σ_0 is given by:

$$\wedge = \frac{|\hat{\Sigma}_0|^{-n/2}\exp(\frac{-1}{2}\sum_{j=1}^{n}(Z_j - \hat{\mu})^T\hat{\Sigma}_0^{-1}(Z_j - \hat{\mu}))}{|\hat{\Sigma}|^{-n/2}\exp(\frac{-1}{2}\sum_{j=1}^{n}(Z_j - \hat{\mu})^T\hat{\Sigma}^{-1}(Z_j - \hat{\mu}))}. \quad (5)$$

Graphical User Interface (GUI) software

Three layers of packages:

- Presentation Layer (PL): User Interface (UI) package and Calculator package.
- Business Layer (BL): clustering algorithms taking inputs from the PL.
- Third-party Layer (TL): a programmer friendly interface.



A screenshot of the GUI implemented using Java Eclipse technology.

Data analysis results

- Number of simulations $N = 1000$.
- Small, medium and large sample size, we fix $n = 10,\ 20,\ 30$ and 50.
- Number of biological replicates($m_1 = m_2 = m$) is set as 3 with 2 technical replicates (t) nested within each, $m = 4$ with $t = 2$ nested within each and $m = 5$ with $t = 3$ nested within each.
- Intramolecular and intermolecular correlation values are set at 3 different levels low(L)(0.2-0.3), medium(M)(0.3-0.5) and clean(H)(0.5-0.6).
- Triplet $(\rho^{ij}, \rho_x^{ij}, \rho_y^{ij})$ represent true correlation values used to generate the data set.