Supplementary Materials for paper “AICM: A Genuine Framework for Correcting Inconsistency Between Large Pharmacogenomics Datasets"

Zhiyue Tom Hu
Department of Biostatistics
University of California, Berkeley
zyhu95@berkeley.edu

Yuting Ye
Department of Biostatistics
University of California, Berkeley
yeyt@berkeley.edu

Patrick A. Newbury
Department of Pediatrics and Human Development
Michigan State University
newburyp@mse.edu

Haiyan Huang[*]
Department of Statistics
University of California, Berkeley
hyh0110@berkeley.edu

Bin Chen[*]
Department of Pediatrics and Human Development
Michigan State University
Bin.Chen@hc.msu.edu
** corresponding author*

Synthetic Data Generation Process

We denote the true drug sensitivity matrix by $M_0 \in \mathbb{R}^{n\times p}$. With the assumption that this ground truth can be decomposed into two independent sources, the drug part and the cell-line part, we can simplify M_0 as $M_0 = \mathbf{a} \cdot \mathbf{b}^T$, where $\mathbf{a} \in \mathbb{R}^n$ contains the information about the n drugs, while $\mathbf{b} \in \mathbb{R}^p$ summarizes the structure of the cell lines. Then, what we finally observed can be given by a model

$$M = \alpha \mathbf{1} \cdot \mathbf{1}^T + \mathbf{a} \cdot \mathbf{b}^T + W, \tag{1}$$

where α is the baseline, and $W \in \mathbb{R}^{n\times p}$ is a random matrix from a matrix normal distribution, which reflects the composite of the noise.

Specifically, we use the following set of parameters to generate the synthetic datasets in Section 3.1 of the original paper. Set $n = 50$, $p = 40$, $\alpha = 0$. The vectors **a** and **b** are generated by independently sampling two instances from standard multivariate Gaussian distributions; then taking the absolute of the two instances. Here, to simulate ineffective drugs, e.g. placebo, **a** is further refined by setting its first 10 entries to 0s. Let $W \sim \mathcal{MN}_{n,p}(\mathbf{0}_{n\times p}, \Sigma_1, \Sigma_2)$, where Σ_1, Σ_2 are strictly positive definite and represent the noise structures of drugs and cells respectively. We simply take

$$\Sigma_1 = I_{n\times n} \quad \Sigma_2 = \begin{bmatrix} 1 & r & r^2 & \dots & r^{p-1} \\ 1 & 1 & 1 & \dots & r^{p-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ r^{p-1} & r^{p-2} & r^{p-3} & \dots & 1 \end{bmatrix},$$

where $r = 0.5$.

Preprint.