

README: completely synthetic data

The completely synthetic data sets used in our empirical study are available. Each set is available as a compressed archive of *tab-delimited text* files.

To uncompress archive “Xrep_low_noise.tar.gz”:

- On linux: tar -xzf Xrep_low_noise.tar.gz
- On windows: winzip would extract the archive

Each archive contains 5 synthetic data sets for a given number of repeated measurements and noise level. Each data set contains 400 genes, 20 experiments and 6 classes. Classes 1 – 4 are periodic sine functions, while classes 5-6 are linear. The third column contains the class numbers, from 1 to 6. The patterns are illustrated in Figure 1 of our manuscript.

How did we generate these synthetic data sets?

Let $\phi(i,j)$ be the artificial pattern of gene i and experiment j before error is added, and suppose gene i belongs to class k , where $i = 1, 2, 3, \dots, 400, j = 1, 2, 3, \dots, 20, k=1,2,\dots, 6$. When $k = 1, 2, 3, 4$, we set $\phi(i,j) = \sin(2\pi j/10 - \pi k/4)$. The size of each of these periodic classes is 67. When $k = 5$, $\phi(i,j) = j/20$. When $k = 6$, $\phi(i,j) = -j/20$. The size of class 5 and class 6 is 66 each. Let $X(i,j,r)$ be the error-added value for gene i , experiment j and repeated measurement r . Let the randomly sampled error from the yeast galactose data be σ_{ij} for gene i and experiment j . Let λ be the multiplicative factor that controls the noise level. When $\lambda=1$, we call it the “low noise level. When $\lambda = 6$, we call it the “high noise level”. $X(i,j,r)$ is generated from a random normal distribution with mean equal to $\phi(i,j)$, and standard deviation equal to $\lambda\sigma_{ij}$.