Appendix

A Detailed procedure for generating simulated datasets

This section outlines our method of generating datasets for this purpose. The datasets for the first and the second rows require several functions which are reported in Table A.1. Rest of the data is generated using a single function.

Table A.1: Functions used to generate datasets of type Non-Linear 1 and Non-Linear 2. Both datasets are similar but range of some features is different.

h_i	Non-Linear 1	Non-Linear 2
$h_1(x)$	$I(x) = x$	$I(x) = x$
$h_2(x)$	$\sin(x)$	$\sin(x)$
$h_3(x)$	x^2	x^2
$h_4(x)$	x^{10}	x^{10}
$h_5(x)$	$\sin(\pi x)$	$0.5\sin(\pi x)$
$h_6(x)$	$\sin(2\pi x)$	$0.5\sin(2\pi x) + 0.5$
$h_7(x)$	x^3	x^3
$h_8(x)$	$4x^2$	x^2
$h_9(x)$	$\sin(4\pi x)$	$0.5\sin(4\pi x) + 0.5$
$h_{10}(x)$	$4x^3$	x^3

In Table 1, the first row shows the results on datasets generated in the following manner. A random matrix of size 1000×20 is generated. Now we will replace 500×10 submatrix of data in a way so that the i^{th} column of this submatrix is given by $h_i(x)$, where x is the value in the first column and the definitions of h_i for $i = 1, 2, \cdots, 10$ are given in Table A.1. The row and column numbers that form the bicluster submatrix are chosen randomly. Thus we embed a bicluster of size 500×10 in the data. Note that for some columns, the range in which the bicluster elements lie is different from the range of remaining data. Thus the background for the biclusters is highly variable.

Datasets used for results shown in the second row of Table 1 are generated using a set of functions, which are modifications of the functions used in the first row. The modifications are made so that data in the bicluster lies in the same range as the rest of the data. The modified functions have been reported in the second column of Table A.1. It may be noted that the function x^2 occurs both as h_3 and h_8 in the second column and x^3 occurs both as h_7 and h_{10}, amounting to the repetition of a column in the bicluster. However, we have retained the repeated columns to retain the similar structure of bicluster as in datasets for the first row.

The third row presents the results on datasets of size 1000×20 drawn from uniform distribution. It has a bicluster of size 500×10 generated using functions $h_1(x) = I(x) = x$, where I is the identity function and $h_i(x) = a_i * x$ with $i = 2, 3 \cdots, 10$. a_i with $i = 2, 3, \cdots, 10$ are different random values lying in interval $(0, 1)$. This is our base data for Uniform distribution datasets. Different transformations have been applied to this data for analysing properties of biclustering algorithms and corresponding results have been reported in the following rows.

The fourth row contains results for datasets obtained by scaling each column of data generated for the third row with a random number lying in the interval $(0, 1)$. The fifth row contains results for datasets obtained by adding a random number lying in $(0, 1)$ to each column of the data generated for the third row. These are used to analyze the scaling and translation properties of the algorithms. The sixth row contains results for datasets obtained by linear transform to each column of data generated for the third row. In a way, this is a combination of scaling and translation. Two random numbers r_1 and r_2 are generated for each column and the transform $r_1 x + r_2$ is applied. Scaling, translation and linear transforms are special cases of distance preserving transforms.

The seventh row and the eighth row contain results for datasets obtained by applying square transform $f(x) = x^2$ and exponential transform $f(x) = \exp(x)$, to each observation and each feature of data generated for the third row. Since data used here lies in the range $(0, 1)$ both of these are monotone transforms, but not necessarily distance preserving.

The ninth row contains results for datasets obtained by duplicating each observation of datasets used for reporting the results in the third row. Thus

these datasets contain 2000 observations. The tenth row contains results for datasets obtained by duplicating each observation of bicluster in datasets generated for the third row. Thus these datasets contain 1500 observations. The eleventh row contains results for datasets obtained by adding a random number in the range (0,0.1) to each element of data matrices. The random noise is drawn from a uniform distribution. The robustness of the algorithms regarding noise is checked through this. The twelfth row contains results for datasets obtained by randomly shuffling rows and columns of the data generated for the third row.

The thirteenth row presents the results on datasets of size 1000×20 drawn from Gaussian distribution. The bicluster submatrix of size 500×10 is generated in the same way as the third row. The fourteenth row contains results for datasets obtained by adding noise to each element of the data matrix generated in the thirteenth row. Noise is generated using random numbers drawn from Gaussian distribution with a standard deviation of 0.1. Thus we test the robustness of the algorithms for Gaussian data.

A.1 Generating large simulated dataset

Initially, data is generated using uniform distribution on $[0, 1] \times [0, 1]$. Then, bicluster of size 10000×30 using functions $h_1(x) = I(x) = x$, where I is the identity function and $h_i(x) = a_i * x$ for $i = 2, 3, \cdots, 30$. The term a_i for $i = 2, 3, \cdots, 30$ are different random values lying in interval $(0, 1)$.

B List of features used from WDI dataset

Air transport, passengers carried
Cause of death, by communicable diseases and maternal, prenatal and nutrition conditions (% of total)
Cause of death, by non-communicable diseases (% of total)
Current health expenditure per capita, PPP (current international $)
Death rate, crude (per 1,000 people)
GDP per capita, PPP (current international $)
Incidence of tuberculosis (per 100,000 people)
International migrant stock, total
International tourism, number of arrivals
Labor force participation rate, total (% of total population ages 15+) (modeled ILO estimate)
Life expectancy at birth, total (years)
Mortality from CVD, cancer, diabetes or CRD between exact ages 30 and 70 (%)
Mortality rate, adult, female (per 1,000 female adults)
Mortality rate, adult, male (per 1,000 male adults)
Out-of-pocket expenditure (% of current health expenditure)
People using at least basic sanitation services (% of population)
PM2.5 air pollution, population exposed to levels exceeding WHO guideline value (% of total)
Population ages 15-64 (% of total)
Population ages 65 and above (% of total)
Population density (people per sq. km of land area)
Population, total
Survival to age 65, female (% of cohort)
Survival to age 65, male (% of cohort)
Trade (% of GDP)
Tuberculosis case detection rate (%, all forms)
Tuberculosis treatment success rate (% of new cases)
Urban population (% of total)