BicMix: Biclustering with mixture of sparse and dense components

Chuan Gao

Sep 2014

1 Compilation

BicMix is written in C++. Both the source code and a compiled binary executable are provided. BicMix used the scientific library GSL, Eigen and Boost, all three libraries need to be installed before compilation. To compile, edit the Makefile file by changing the -L/ and -I/ line to your linker and header file location, then in the terminal,

```
$ make
```

this generates the binary executable file: BicMix. This binary only support Linux for the time being.

2 Run BicMix

BicMix takes a tab or space delimited file of all numbers (strictly numbers, no headers) as its input, and writes its estimates in a specified directory. It uses the default value of $a = b = c = d = e = f = 0.5$ to recapitulate the horseshoe, and a starting value of $\alpha = \beta = 1$ for the mixing proportion of the sparse and dense components. BicMix infer the factor numbers non-parametrically by shrinking a big starting value down. It evaluates convergence by checking the number of nonzero values for the loading matrix, and assume convergence if the number fixes for 200 iterations. It also writes parameter values to a specified directory for every 200 iterations, so that when the algorithms takes unbearably long, some intermediate values are available.

To run BicMix, issue the following command in terminal:

```
$ BicMix --nf number_of_components --y your_gene_expression_file --out
dir_result --sep separation_character
```

where each argument is specified by a flag, more details about the flags are:

- `--nf` specifies the starting factor number, users should make it reasonably big but not too big to burden the program.

- `--y` specifies the input file.
- `--out` specifies the output directory.
- `--sep` specifies the delimiter of the file, takes two values, "space" or "tab".

3 Output

The model is set up as $\mathbf{Y} = \mathbf{\Lambda X} + \epsilon$, the output exactly reflect that. So for a $\mathbf{Y}$ matrix of $p \times n$, where n is the number of samples and p is the number of genes. The files that are written into the specified directory are:

- command.txt: a file that records the command that is given.
- LAM: the MAP estimates of the loadings.
- EX: the expected values of the factors.
- EXX: the value of $\langle X^T X \rangle = \langle X \rangle^T \langle X \rangle + p * \Sigma_X$
- Z: a $2 \times K$ matrix indicating whether a loading is dense or sparse. For a loading, a top row value of 1 and bottom row value of 0 indicate that it is sparse, vice versa. Because the expected value of this hidden variable is used, a value in the range of [0, 1] is sometimes observed.
- O: a $2 \times K$ matrix indicating whether a factor is dense or sparse. Similar to the loading, a top row value of 1 and bottom row value of 0 indicate that it is sparse, vice versa.
- PSI: the diagonal values for the variance matrix $\mathbf{\Psi}$

4 Simulations

A toy data has been provided, where a gene expression file of $n = 300$ samples and $p = 500$ genes is simulated. There are 15 components in total, 10 of them are sparse and 5 of them are dense. Files in the data directory are:

- Y: The gene expression file
- LAM.txt: the true loading matrix
- X.txt: the true factor matrix
- Z.txt: the true indicator matrix for the loading.
- O.txt: the true indicator matrix for the factor.

To run BicMix on the data

```
mkdir result
./BicMix --nf 50 --y ./Y.txt --out result --sep space
```