Bayesian Decomposition Files
Bioinformatics, Fox Chase Cancer Center
7 June 2002

General Information

The primary goal of Bayesian Decomposition (BD) is to simultaneously determine two matrices which when multiplied together reproduce a known data matrix, i.e.

$$[D] = [A][P]$$

where [D] represents a series of measurements (rows), [A] represents a measure of the amplitude of each underlying set of measurements (row of [P]) in each set of data (rows of D). For example, the rows of [D] can be expression at different time points, the rows of [P] changes in regulation due to some underlying biological activity (e.g. activation of a signaling pathway), and the rows of [A] measures of how strongly a gene is activated by such biological activity. At present BD requires a measure of the dimensionality of the solution (number of rows in [P]).

Input Files

For gene expression analysis, genpat (the executable name for this implementation of Bayesian Decomposition) requires two input files with names NAME.ctrl and NAME.data, where NAME is some user-defined specifier. These files allow comments on most lines to help keep clear what the input variables are. Sample files are given below.

FILE Test.ctrl

```
Test        // This is a label to identify output files
This is control file for the genetic simulation
1           // Number of members in ENSEMBLE (must be 1)
1000        // Number of iterations
1000        // Number of iterations between snapshots
2331821     // Random Number Seed
200.0       // A Atoms Parameter, 0 for automatic (0 not implemented)
200.0       // F Atoms Parameter, 0 for automatic (0 not implemented)
0.5         //  This parameter presently ignored
```

In this file the first line specifies the label which will be used to name the output files which are detailed below. The second line is a comment line that is discarded by the program and can be used to describe the data if desired. The third line gives the number of members in the ensemble. The fourth line is the number of iterations for each step during equilibration. The fifth line determines how often the present sample is written to an output file. The sixth line is the random seed (unsigned). The seventh line is the expected number of atoms needed to describe the [A] matrix, while the eighth line is the expected number of atoms needed to describe the [P] matrix. The final line allows an additional parameter into the program, but is not used in this implementation.

FILE Test.data

```
Fake2
Simulation test for installation
3 15 30    // number of patterns, number of genes, number of conditions
1          // 1 indicates universal uncertainty measure, 0 individual
0.2        // if 1 above, this is uncertainty
1.0        // parameter presently not used
0.51028
0.16945
0.074458
-0.06353
0.038197
    …
```

In this file the actual data from an experiment is kept. The first line is a label which is not used by the program but which can be used to keep track of the specific data file used for output files if this is desired. The second line is discarded by the program and can be used to describe the source of the data. The third line contains the number of solutions (rows in [P]), the number of genes (rows in [D]), and the number of points per gene (columns in [D]). The fourth line is a flag indicating whether there is a single noise value for all data points (1 = single value). The fifth line will be used as the noise value if there is a single noise value, otherwise it will be ignored. The sixth line is a parameter describing the data, but is unused in this implementation. After this the data is entered in any form which can be read by the C function fscanf without comments. The data is entered row by row, so in this case the first 51 points are the first data spectrum. If the uncertainty level is not universal (i.e. parameter is not 1 in line 4), then each input data line needs two values. The first is the expression level while the second is the uncertainty for that expression level.

In all input files the '//' is used as a delimiter to note the start of comments. If comments are not included the end of the line is taken as the linebreak (reading is done by the C function fgets with the function strtok breaking up the inputs using '/', '\n', and ' ' as tokens).

Output Files

The programs generate a number of output files containing means and uncertainties for the elements in the matrices [A] and [P], some samples taken during the sampling, and the Mock data created from the model. These files are described below.

FILE TestOut.out

```
TestOut
4.366924e+02  2.366300e+01   4.233000e+01  -7.546976e+01
1  1000  1000  2331821
3  15  30
5.000000e-01   1.000000e+00
5.000000e-01   1.000000e+00   30
 5.102800e-01  1.810720e-02   2.000000e-01
 1.694500e-01  1.488370e-02   2.000000e-01
 7.445800e-02  1.099662e-02   2.000000e-01
    …
```

The first line contains the experimental tag used for file names. The second line has four float values, the χ2 value, the average number of atoms for A then P, and the evidence. The third line has four integer values, the number of members in the ensemble, the number of iterations during sampling, the number of iterations between movie samples, and the seed value for the random number generator. The fourth line has three integer values, the number of patterns, of genes (number of rows of [D]), and of points for each gene (number of columns of [D]). The fifth line has two float values, which relate to the unused input parameters. The sixth line will contain a copy of the fifth line with an additional parameter, but is unused in this implementation. These are followed by lines containing three float values each with the actual data input value, the mock data value for each point, and the noise at each point. The values are listed row by row through the [D] matrix.

FILE TestOut.inf

```
TestOut.inf
3  15  30
Thu Jun  6 12:44:14 2002
1
1.000000e+00
  1.742543e-02    1.003423e-01
  2.230431e-01    4.565692e-01
  4.081820e-03    3.094442e-02
   …
```

The inf files contain the inference variable outputs. The first line contains the name of the file. The second line has three integer values, the number of patterns, of genes (number of rows of [D]), and of points for each gene (number of columns of [D]). The third line has the date as a char* string generated by the C function ctime. The fourth line has a value 1 for this implementation. The following line also has the value 1 as a float for this implementation. The next lines (number of elements of [A]) contain the mean and standard deviation of the mean for [A] written column by column (i.e. amplitudes for solution 1, then solution 2, etc.). The following lines (number of elements in [P]) contain the mean and standard deviation of the mean for [P] written row by row (i.e. first pattern, second pattern, etc.).

FILE TestOut.mov

The final output file is a collection of samples containing the elements of [A] followed by the elements of [P] in the same order as in the inf file. There are (# of iterations / # of iterations between samples) copies in the file.