Manual for BIEMS data generator

The BIEMS data generator has been built to generate data files for testing and debugging of the BIEMS program. The format of the data output is consistent with BIEMS input requirements. The data generator itself consists of two files: the executable biems_datagen.exe and the input file input.txt. The specifications of the input file are describe below. Upon running the data generator, it creates a whitespace-separated data file data.txt.

Input File

The properties of the generated data are specified by editing the file *input.txt* with a plain-text editor. The file is divided into input lines each consisting of a header line containing a description and one or more lines containing the specification for that line. All input lines must be separated from each other by a single empty line. Details for the specification of each line are given below, using a running example for illustration.

General settings

The general structure of the data is specified on the first and second input line. In the example the first input line looks as follows:

Input line 1

```
#DV  #groups  #cov  iseed  generate exact data
 2         2    2     23           1
```

The first three values on this input line are the number of dependent variables P, the number of groups J, and the number of explanatory variables K. The first two values must be at least 1. If the number of groups is set to 1, the grouping variable becomes the intercept. The fourth value on the first input line contains the random seed, which can be set to any positive integer or to -1. For a given value of the random seed, if the data specification is kept constant the data generator will output the same data each time it is run, making the process repeatable. If this parameter is set to -1, the program will pick a different random seed every time it is run, ensuring different data is produced each time. The fifth value, *generate exact data*, must be set to either 0 or 1. When set to 1, the maximum likelihood estimate of the mean parameter matrix is equal to the chosen parameter matrix, i.e., $\hat{\mathbf{B}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{B}$ and the maximum likelihood estimate of the error covariance matrix is equal to the chosen error covariance matrix, i.e., $\hat{\boldsymbol{\Sigma}} = N^{-1}\left(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}\right)'\left(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}\right) = \boldsymbol{\Sigma}$.

The second input line specifies the number of data points for each group. A number of values equal to the number of groups must be given, all on the same line. If the number of groups is set to 1, this simply contains the total sample size. If the number of groups is larger than one, the sum of these numbers is the total sample size.

Input line 2

```
number of data points per group
50  40
```

Settings for the error term

The dependent variables are generated as a linear combination of the independent variables plus an error term. This error term is drawn from a multivariate normal distribution with means equal to 0 and the error covariance matrix $\mathbf{\Sigma}$. The standard deviations and correlation matrix of this covariance matrix are specified on input lines 3 and 4 respectively.

The standard deviations for the error terms are all given on a single line, with the number of values equal to the number of dependent variables.

Input line 3
`standard errors for the error terms`
`1  3`

The correlations between the error terms are given as a square matrix. The matrix must be a proper correlation matrix, and as such it should be symmetrical around its diagonal and the diagonal itself should contain all 1's. If only a single dependent variable is used, this line should contain a 1.

Input line 4
`correlation matrix for the error terms`
`1    0.2`
`0.2  1`

Settings for the model parameters

On the fifth input line the parameter matrix $\mathbf{B}$ is specified which contains $J + K$ rows and P columns. Each column corresponds to a dependent variable. The first J rows correspond to a group means and, if $K > 0$, the following K rows are the regression coefficients of the explanatory variables. If there is no grouping in the data, i.e., $J = 1$, the first row contains the intercepts.

Input line 5	
`mean parameter matrix`	
`1  2`	group 1
`4  5`	group 2
`7  8`	expl. var. 1
`9  7`	expl. var. 2

Settings for explanatory variables

The explanatory variables are generated from a multivariate normal distribution, specified separately for each group. The means of these distributions are specified on input line 6, the covariance matrices on input line 7. The means are given in a matrix where column k corresponds to explanatory variable k, for $k = 1, \ldots, K$, and row j corresponds to group j, for $j = 1, \ldots, J$. In the specification below, the first explanatory variable has a mean of 5 in group 1 and a mean of 7 in group 2, and the second explanatory variable has a mean of 10 in group 1 and a mean of 13 in group 2.

Input line 6	
`means for normal covariates per group`	
`5  10`	group 1
`7  13`	group 2

On the seventh input line, covariance matrices are given for each group. Hence, this consists of J blocks of $K \times K$ covariance matrices. In the specification below, the covariance matrix of the explanatory variables in the first group is equal to $\begin{bmatrix} 2 & 3 \\ 3 & 9 \end{bmatrix}$ and equal to $\begin{bmatrix} 11 & 2 \\ 2 & 5 \end{bmatrix}$ in the second group.

Input line 7	
`covariance matrices for normal covariates per group`	
`10  3`	group 1
`3   5`	
`11  2`	group 2
`2   5`	

If no explanatory variables are used, these input lines are not read.

Data file

The generated data is written to the file *data.txt* in the same directory as the other files. Note that if this file already exists, it is overwritten. The data is written in a plain-text, whitespace-separated format with rows corresponding to data points and columns to variables. There is no header with names or descriptions of the variables, but the order of the variables is always the same. From left to right, the file contains: the P dependent variables; the K explanatory variables (if any); and the grouping variable j. If there is no grouping in the data, the grouping variable is equal to 1 for all data points. A part of the data for the example is shown below (rounded for brevity).

data.txt				
107.5	94.6	4.1	8.6	1
126.2	109.6	5.4	9.8	1
⋮	⋮	⋮	⋮	⋮
137.8	121.7	5.9	10.3	2
dv1	*dv2*	*cov1*	*cov2*	*group*