Guide to the STACMR R package

21 August 2017

STACMR is a set of R functions that implement the Conjoint Monotonic Regression (CMR) approach to State-Trace Analysis (STA). This is a quick guide to its use.

Contents

1 Starting up

1. Unzip the file `STACMR-R.zip`.
2. Make the folder `../STACMR-R` the working directory.
3. Run `staCMRsetup`. This loads the relevant functions and locates and links the most recent java runtime library.

You should now be good to go.

2 Partial order

The various STACMR functions often make use of an optional partial order. A partial order on a vector, x, is a set of pairs, (i, j), such that $x_i \leq x_j$. This is represented in STACMR in two (equivalent) ways:

1. As a list containing the set of (i, j) pairs, such as, `list(c(1,2), c(2,3), c(1,4))`. There is a shorthand for a linear order such as, `list(c(1,2), c(2,3))`, which can be written as, `list(c(1,2,3)` or, even more simply, as `list(c(1:3))`.
2. As an adjacency matrix in which entry $(i, j) = 1$ if (i, j) is an element of the partial order, otherwise $(i, j) = 0$.

The R function, `list2adj`, converts a partial order in list form into its corresponding adjacency matrix form. For example,

```
> a=list2adj(c(1:4),list(c(1:3)))
> a
     [,1] [,2] [,3] [,4]
[1,]    0    1    0    0
[2,]    0    0    1    0
[3,]    0    0    0    0
[4,]    0    0    0    0
```

In the above call to `list2adj`, the vector, `c(1:4)`, specifies the set that the partial order applies to. It is almost invariably the sequence of numbers, 1 to k, where k is the total number of conditions. In the above example, $k = 4$. The function `adj2list` converts an adjacency matrix into its corresponding list.

3 Continuous data

Continuous data has the form that each observation is a number drawn from a continuous distribution.

3.1 Input data structures

At present, STACMR accepts two kinds of data structure for continuous data:

1. **List format**. In this format, the data are contained in a $b \times n$ list where b is the number of between-participant conditions (groups) and n is the number of dependent variables. Each component of the list is itself an $N \times w$ matrix of observations where N is the number of subjects (which may vary across groups and dependent variables) and w is the number of within-participant conditions (fixed across groups and dependent variables). The dependent variable may be either within-participant or between-participant it doesnt matter because the correlation between dependent variables is assumed to be zero (although this might change in future implementations).

2. **General format**. This is a fixed column format organised as a matrix in which each row corresponds to an observation and each column is defined as follows:

column 1: Participant number (for identification only, not used directly)

column 2: Between-participant condition or group (if none, then set this value to 1)

column 3: Dependent variable (numbered 1, 2, and so on)

columns 4 to end: Values for each within-participant condition

While STACMR accepts data in general format it always converts it to listformat using the function,

```
> y = gen2list (data=NULL, varnames=list())
```

Here, `varnames` is an optional vector of the names of each within-participant condition.

3.2 Principal functions

The principle functions are located in the folder, `.../STACMR-R`. Their operation is illustrated with respect to the dataset, `delay`, located in folder, `.../STACMR-R/Data files`.

The `delay` data set comes from the study by Dunn, J. C., Newell, B. R., & Kalish, M. L. (2012). The effect of feedback delay and feedback type on perceptual category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition, 38*(4), 840-859.

Participants in this study completed one of two category-learning tasks over 4 blocks of training (a within-participant factor) under one of 2 different conditions (a between-participant factor). The dependent variables are the proportions correct for each of two tasks defined according to the category structure that participants learned. For the rule-based (RB) group, the category structure was defined by a simple rule. For the information-integration (II) group, the category structure was more complex and could not be defined by a simple rule.

The data are in general format and is contained in the file, `delay.dat`. That is,

```
> delay = read.table('Data files/delay.dat')
> vnames = c('B1','B2','B3','B4')
> y = gen2list(data=delay, varnames=vnames)
```

3.2.1 staSTATS.R

This function computes summary statistics of a data structure in cell array format. Example call:

```
> output = staSTATS (data=NULL, varnames=list(),
  shrink=-1)
```

Input:

- `data` is the name of a data set in either general or list format.
- `varnames` is an optional list of names of the within-participant variables.
- `shrink` is an optional parameter denoting how much shrinkage to apply to the estimated covariance matrix. Generally, the covariance matrix needs to be shrunk during the bootstrap cycle to avoid ill-conditioning. If `shrink = 0` then no shrinkage is applied. If `shrink = 1` then maximum shrinkage is applied. This means that the covariance matrix is diagonalized with all off-diagonal entries set to zero. If `shrink < 0` (the default) then an optimal shrinkage value is estimated for each within-participant block and applied according to an algorithm developed by: Ledoit, O. & Wolf, M. (2004). Honey, I shrunk the sample covariance matrix, *The Journal of Portfolio Management, 30*(4), 110-119.

Output:

`staSTATS` returns in the variable `delaystats` a list of length equal to the number of dependent variables in `y`. Each component of `delaystats` is a named list. For the dependent variable, `ivar`:

- `output[[ivar]]$means` = vector of means across all conditions
- `output[[ivar]]$cov` = the covariance matrix (for information only)
- `output[[ivar]]$regcov` = the adjusted covariance matrix following application of shrinkage
- `output[[ivar]]$n` = matrix of number of observations (subjects) in each within-participant block

- `output[[ivar]]$lm` = matrix of Loftus-Masson within-participant standard errors (used by staPLOT)
- `output[[ivar]]$weights` = matrix of weights defined by: `output[[ivar]]$n.*solve(output[[ivar]]$regcov)`
- `output[[ivar]]$shrinkage` = a vector of length b (where b is the number of levels of the between-participant independent variable) containing the specified or estimated shrinkage values

Not yet implemented

Try the following:

```
> delaystats = staSTATS(delay, vnames)
```

The output `delaystats` has two components corresponding to the two dependent variables in `delay`. If you look at the means, they should look like this:

```
> delaystats[[1]]$means
      b1        b2        b3        b4        b1
  b2        b3        b4
0.3676471 0.4676471 0.5757353 0.6117647 0.3444853
  0.4433824 0.5080882 0.5169118

> delaystats[[2]]$means
      b1        b2        b3        b4        b1
  b2        b3        b4
0.3308333 0.4550000 0.5345833 0.5491667 0.2835937
  0.3031250 0.3179688 0.3097656
```

3.2.2 staMR.R

This function conducts monotonic regression on a data structure according to a given partial order. We say it fits the partial order model to the data (i.e., the set of dependent variables).

Example call:

```
> output = staMR (data=NULL, partial=list(), shrink=-1)
```

Input:

Here, data is either a data structure (in list or general format) or structured output from `staSTATS`; `partial` is a partial order (required) in

either list or adjacency matrix format; `shrink` is an optional shrinkage parameter (defined in Section **??**). If data is a list of structured output from `staSTATS`, then the shrinkage specified by this output is used whether the argument `shrink` is specified or not.

Output:

- `output$x` is a n-element list that contains the best-fitting values for each dependent variable
- `output$fval` is the value of the least squares fit
- `output$shrinkage` is a $b \times n$ matrix of shrinkage values (where b is the number of levels of the between-participant independent variable)

Try `staMR` with `delay`. To do so, we have to specify a partial order. Use the following:

```
> E = list(c(1:4),c(5:8),c(5,1),c(6,2),c(7,3),c(8,4))
> out2 = staMR (delay, E)
> cbind(out2$x[[1]],out2$x[[2]]) # simplify presentation
          [,1]      [,2]
[1,] 0.3676471 0.3308333
[2,] 0.4676471 0.4550000
[3,] 0.5757353 0.5345833
[4,] 0.6117647 0.5491667
[5,] 0.3444853 0.2830441
[6,] 0.4433824 0.3033609
[7,] 0.5080882 0.3149231
[8,] 0.5169118 0.3149231
> out2$fval
[1] 0.1721286
> out2$shrinkage
           [,1]       [,2]
[1,] 0.05197173 0.04909396
[2,] 0.03135966 0.26502326
```

3.2.3 staCMR.R

This is the main function that conducts the CMR (state-trace) analysis. It takes a data structure or a list of structured output from `staSTATS` and an optional partial order and returns the best fitting values (to the data means)

and the least squares fit. It fits the monotonic model to the data.

Example call:

```
> output = staCMR (data, partial=list(), shrink=-1)
```

On the input side, `data` is a data structure, `partial` is an optional partial order, and `shrink` is the optional shrinkage parameter (defined in Section **??**).

On the output side, `output$x` is a list of the best-fitting values, `output$fval` is the value of the least squares fit, and `output$shrinkage`, is a $b \times n$ matrix of shrinkage values.

Now try `staCMR` with `delay`:

```
> out1 = staCMR (delay, E)
> cbind(out1$x[[1]],out1$x[[2]])
          [,1]      [,2]
[1,] 0.3758963 0.3149845
[2,] 0.4850354 0.4358350
[3,] 0.5898416 0.5166958
[4,] 0.6265142 0.5317816
[5,] 0.3352864 0.2855973
[6,] 0.4185966 0.3149845
[7,] 0.4805813 0.3226823
[8,] 0.4850354 0.3226823
> out1$fval
[1] 1.749307
> out1$shrinkage
           [,1]       [,2]
[1,] 0.05197173 0.04909396
[2,] 0.03135966 0.26502326
```

3.2.4 staMRFIT.R

This function tests the fit of the partial order model.
Example call:

```
> output = staMRFIT(data, partial=list(), nsample=1,
  shrink=-1)
```

Input:

The `staMRFIT` function takes the following arguments:

- `data` is the name of a data structure (either in general format, list format), or list output from `staSTATS`. If a data structure is specified then the bootstrap re-sampling is non-parametric. If only the summary statistics are provided (e.g., `delaystats`) then the bootstrap is parametric and assumes that observations are distributed normally for each dependent variable in each condition.

- `partial` is a required partial order. This may be in either cell array or adjacency matrix form.

- `nsample` is the number of Monte-Carlo samples to be drawn in computing the empirical sampling distribution of the fit value.

- `shrink` is the optional shrink parameter, defined in Section **??**.

Output:

- `output$p` is the estimated p-value for the hypothesis that the fit of the model is zero. It is the proportion of bootstrap fit values that are greater than or equal to the observed fit value. Note that it will be be different from run to run, as it is a Monte Carlo estimate.

- `output$datafit` is the observed fit value. It is the same as `output$fval` returned by `staMR` above.

- `fits` is a vector of length `nsample` of computed bootstrap fit values. Thus, `output$p` is the proportion of values of `output$fits` that are greater than or equal to `output$datafit`.

Applying `staMRFIT` as follows produced the following outputs:

```
> out = staMRFIT(delay, partial=E, nsample=10000)
> out$p
[1] 0.7470
> out$datafit
[1] 0.1721
```

3.2.5 staCMRFIT.R

This function estimates the empirical distribution (and hence p-value) of the difference in the fit of the conjoint monotonic and the fit of the partial order model. The function call is analogous to `staMRFIT`:

```
> output = staCMRFIT(data, partial=list(), nsample=1,
  shrink=-1)
```

Input:

- `data` is the name of the data structure (in either general or list format) or the name of structured output from `staSTATS`). If specified as a data structure then the bootstrap re-sampling is non-parametric, otherwise the bootstrap is parametric and assumes that observations are distributed normally for each dependent variable in each condition.
- `partial` is an optional partial order in either list or adjacency matrix form.
- `nsample` is the number of Monte-Carlo samples to be drawn in computing the empirical sampling distribution of the fit value. We recommend using about 10,000 for estimating p to the nearest 100th.
- `shrink` is the optional shrink parameter, defined in Section **??**.

Output:

- `output$p` is the estimated p-value for the hypothesis that the difference in fit between the monotonic model and the partial order model is zero. It is the proportion of differences of bootstrap fits values that are greater than or equal to the observed difference in fit value.
- `output$datafit` is the observed difference in fit value between the monotonic model and the partial order model.
- `output$fits` is a vector of length `nsample` of computed differences in bootstrap fit values. Thus, `output$p` is the proportion of components of `output$fits` that are greater than or equal to `output$datafit`.

Applying staCMRFIT to the delay data produces the following output:

```
> out = staCMRFIT(delay, partial=E, nsample=10000)
> out$p
[1] 0.1753
> out$datafit
[1] 1.5772
```

4 Binary data

Binary data has the form that each observation is either a 'success' or a 'failure'. STACMR assumes that these observations have been aggregated into counts of the total number of successes and failures for each participant and each condition.

4.1 Input data structures

At present, STACMR accepts two kinds of data structure for binary data:

1. **List format**. In this format, the data are contained in a $N \times n$ list where N is the number of observation units (typically, participants) and n is the number of dependent variables. Each component of this list is itself an $k \times 2$ matrix of counts of successes and failures (here called hits and misses), respectively, where k is the number of conditions.

2. **General format**. This is a fixed column format organised as a matrix in which each row corresponds to a particular observation unit (participant) and condition. Each column is defined as follows:

   column 1: Participant number

   column 2: Condition number (if none, then set this value to 1)

   column 3: Dependent variable (numbered 1, 2, and so on)

   column 4: Count of number of successes

   column 5: Count of number of failures

Data in general format are converted to list format using the function,

```
> y = gen2listBN (data);
```

4.2 Principle functions

The principle functions are located in the folder, `.../STACMR-R`. Their operation is illustrated with respect to the dataset, `dfie`, located in folder, `.../STACMR-R/Data files`.

We illustrate the binary STACMR functions using The `dfie` data set reported by Prince, M., Hawkins, G., Love, J., & Heathcote, A. (2012). An R package for state-trace analysis. *Behavior Research Methods, 44*(3), 644-655.

In this study, the dependent variables were accuracy of memory for pictures of faces and accuracy of memory for pictures of houses. There were $k = 6$ conditions defined by the combination of two factors; stimulus orientation (upright, inverted), and study duration (short, medium, and long). There were $N = 18$ participants each of whom were tested under all six conditions on both dependent variables. The data from each participant therefore can be analyzed individually.

The data (counts of hits and misses) for each participant are contained in the text file, `dfie.dat`:

```
> dfie = read.table('Data files/dfie.dat')
> y = gen2listBN(dfie)
```

4.2.1 staSTATSBN.R

This function returns summary statistics for binary data.

Example call:

```
> dfiestats = staSTATSBN (dfie);
```

The output, `dfiestats`, consists of a 18×2 list in which each component is a participant (indexed below by `isub`) and a dependent variable (indexed below by `dvar`). That is,

- `dfiestats[[isub]][[dvar]]$count` = matrix of counts of successes (hits) and failures (misses) for each condition.
- `dfiestats[[isub]][[dvar]]$means` = vector of 'means' corresponding to the proportion of successes for each condition.
- `dfiestats[[isub]][[dvar]]$weights` = vector of weights corresponding to the number of trials (i.e., number of successes + number of failures) for each condition.
- `dfiestats[[isub]][[dvar]]$n` = vector of counts corresponding to the number of trials for each condition. Identical to weights.

4.2.2 staMRBN.R

The function `staMRBN` fits a partial order model to a binary data structure.

Example call:

```
> output = staMRBN (data, partial=list());
```

Input

- `data` is the name of the binary data structure.
- `partial` is the name of a partial order (required) in either list or adjacency matrix format.

Output:

- `output$x` is an $N \times n$-element list, where N is the number of observation units (participants) and n is the number of dependent variables. Each component of `output$x` is a k-element vector containing the best-fitting values for the corresponding participant and dependent variable.
- `output$fval` is a N-vector containing the values of the least squares fit for each participant.
- `output$g.squared` is a N-vector containing the G^2 fit for each participant.

Execute the following commands:

```
> E = {1:3,4:6};
> out2 = staMRBN (dfiestats, E);
```

For participant 1, the output is:

```
> out2$x[[1]]
          [,1]      [,2]
[1,] 0.7051282 0.6730769
[2,] 0.7051282 0.6730769
[3,] 0.8717949 0.7692308
[4,] 0.4487179 0.6153846
[5,] 0.6410256 0.6794872
[6,] 0.7179487 0.7564103
> out2$fval[1]
[1] 0.03205128
> out2$g.squared[1]
[1] 0.1524826
```

4.2.3 staCMRBN.R

This function fits the monotonic model to a binary data structure.
Example call:

```
> output = staCMRBN (data, E);
```

Input:

- `data` is the name of a binary data structure.
- `E` is an optional partial order in either list or adjacency matrix format.

Output:

- `output$x` is an $N \times n$-element list, where N is the number of observation units (participants) and n is the number of dependent variables. Each component of `output$x` is a k-element vector containing the best-fitting values for the corresponding participant and dependent variable.
- `output$fval` is a N-vector containing the values of the least squares fit for each participant.
- `output$g.squared` is a N-vector containing the G^2 fit for each participant.

Execute the following commands:

```
> out1 = staCMRBN (dfie, E)
```

For participant 1, the output is:

```
> out1$x[[1]]
          [,1]      [,2]
[1,] 0.7051282 0.6752137
[2,] 0.7051282 0.6752137
[3,] 0.8717949 0.7692308
[4,] 0.4487179 0.6153846
[5,] 0.6410256 0.6752137
[6,] 0.7179487 0.7564103
> out1$fval[[1]]
[1] 0.03418803
> out1$g.squared[[1]]
[1] 0.1622374
```

4.2.4 staMRFITBN.R

This function tests the fit of the partial order model to binary data.
Example call:

```
> output = staMRFITBN (data, partial=list(), nsample=1,
  shrink=-1)
```

Input:

- `data` is the name of the binary data structure.

- `partial` is the name of a partial order in list or adjacency matrix form.
- `nsample` is the number of Monte-Carlo samples.
- `shrink` is the shrinkage parameter defined in Section **??**.

Output:

- `output$p` is an N-vector containing the estimated p-values for each observation unit (participant).
- `output$datafit` is an N-vector containing the observed G^2 value for each each observation unit (participant).
- `output$fits` is a matrix of computed bootstrap G^2 values. Each row corresponds to each of N observation units (participants) and each row consists of a vector of length `nsamples`.

Example output (note that `p` and `fits` will be differ from run-to-run as they are Monte Carlo estimates):

```
> out = staMRFITBN (dfie, partial=E, nsample=10000)
> out$.p
[1] 0.0.7311
> out$datafit
[1] 0.1525
```

To be completed.

4.2.5 staCMRFITBN.R

This function tests the difference between the fit of the partial order model and the fit of the monotonic model to binary data.
Example call:

```
> output = staCMRFITBN (data, partial=list(), nsample=1,
  shrink=-1)
```

Input:

- `data` is the name of the binary data structure.
- `partial` is an optional partial order in list or adjacency matrix form.

- `nsample` is the number of Monte-Carlo samples.
- `shrink` is the shrinkage parameter defined in Section **??**.

Output:

- `output$p` is an N-vector containing the estimated p-values for each observation unit (participant).
- `output$datafit` is an N-vector containing the observed G^2 value for each each observation unit (participant).
- `output$fits` is a matrix of computed bootstrap G^2 values. Each row corresponds to each of N observation units (participants) and each row consists of a vector of length `nsamples`.

Example output (note that `output$p` and `output$fits` will be differ from run-to-run as they are Monte Carlo estimates):

```
> out = staCMRFITBN (dfie, partial=E, nsample=10000)
> out$.p
[1] 0.1847
> out$datafit
[1] 0.0098
```