CCBER - Software Description Document

Brian Vegetabile

June 2, 2018

# Overview of Software

is an R package for the estimation of behavioral entropy rate, developed for the Conte Center @ UCI. The package reads in an individual, or set of, excel files and processes the file to estimate the entropy rate from the supplied data. This document describes usage and input formats that are expected with this software.

See reference: Davis, E.P., Stout, S.A., Molet, J., Vegetabile, B., Glynn, L.M., Sandman, C.A., Heins, K., Stern, H., Baram, T.Z. (2017). **Exposure to unpredictable maternal sensory signals influences cognitive development across-species**. *Proceedings of the National Academy of Sciences*. September 26, 2017. 114 (39) 10390-10395

An overview of behavioral entropy rate estimation in the context of human behavior is found in the supporting information (SI) for the reference above.

The source files for ccber are found at [github.com/bvegetabile/ccber](https://www.github.com/bvegetabile/ccber)

## Installation of ccber

The package devtools is required to install this R package from this Github repository. Install this package first if it is not already installed.

From within an R console, enter the following:

install.packages('devtools', dependencies = TRUE)

Once that package has been installed, use the following to install ccber

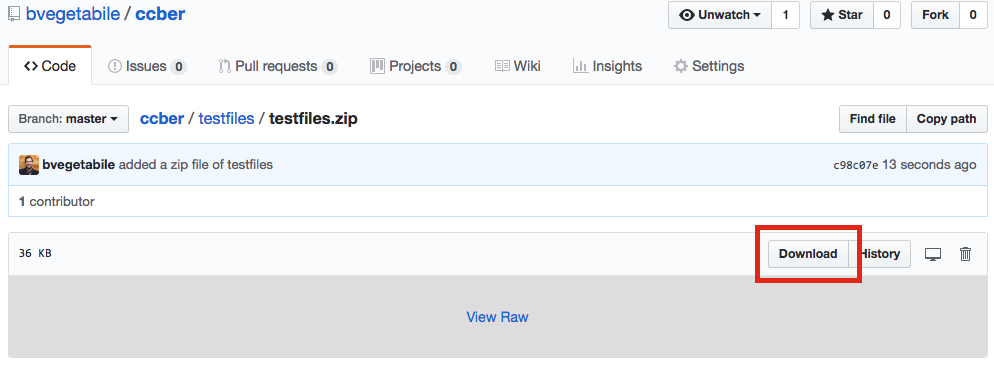
devtools::install\_github('bvegetabile/ccber')

Load the package to begin analysis!

library('ccber')

## Quick Start

Download files from [github.com/bvegetabile/ccber/tree/master/testfiles/testfiles.zip](https://github.com/bvegetabile/ccber/blob/master/testfiles/testfiles.zip).



Screenshot of the download location for testfiles.zip“

Navigate to the directory where the files are located using the following R command. The setwd command sets the working directory for R. ( *Note in the below, the path should be changed to the location of where the files have been uncompressed* )

setwd('~/git/ccber/testfiles/')

Then run the following,

test\_output <- ccber::ber\_analyze\_dir('.')

By setting the working directory in the first step, any output files will be put in the directory specified.

If successful, you fill will see the following:

> ccber::ber\_analyze\_dir('.')  
Completed without issue : Entropy\_6m - 88888HE - Event Logs.xlsx  
Completed without issue : Entropy\_6m - 99999LE - Event Logs.xlsx  
Script total run time: 0.013 minutes  
-------------------- Check the log for files below --------------------

The object test\_output contains the entropy rates and some additional measures. The output will look as follows:

> test\_output  
 SubjectID CanEstimateEntropy EntropyRate  
1 88888HE TRUE 1.2755499  
2 99999LE TRUE 0.6442886  
 TotalNumberOfTransitions CombinedVideoDuration PercentMissing  
1 119 600.027 0  
2 69 600.027 0  
 AuditoryCounts AuditoryTotalTime AuditoryAverageTime  
1 23 25.001 1.087  
2 15 15.000 1.000  
 VisualCounts VisualTotalTime VisualAverageTime TactileCounts  
1 16 309.9636 19.37273 21  
2 11 185.0270 16.82064 10  
 TactileTotalTime TactileAverageTime  
1 362.9628 17.28394  
2 285.0273 28.50273

To save the output data to a .csv file to be read into excel later, use the following:

write.csv(test\_output,   
 file = file = paste(Sys.Date(),   
 '-ber-estimates.csv',   
 sep=''),  
 row.names = F)

The command Sys.Date() prepends the date to document when the data was created

## Function Overviews

Ordinary use of the package would not require any of the information below this point in the description.

These function overviews provide additional information that may be helpful if there is a need to modify the settings (e.g., an error has occurred, there is a new data type or you want to try an alternative padding in event definitions).

### Function : ber\_analyze\_file

The following function from ccber is one of the primary functions for the estimation of entropy rate:

ber\_analyze\_file(f\_loc,  
 plot\_all=F,  
 plots\_to\_file=F,  
 tactile\_padding = 1.0,  
 auditory\_padding = 1.0,  
 behavior\_types=list(  
 "mom\_auditory\_types" = c('Vocal'),  
 "mom\_tactile\_types" = c('TouchBaby',  
 'HoldingBaby'),  
 "mom\_visual\_types" = c('ManipulatingObject'),  
 "baby\_visual\_types" = c('LookAtMomActivity'),  
 "missing\_types" = c('CantTellHolding',  
 'ActivityNotVisible',  
 'CantTellLooking')),  
 missing\_threshold = 0.1)

Below is a more detailed description of each input:

|  |  |  |
| --- | --- | --- |
| Input | Input Type | Description |
| f\_loc | String | String indicating the location of the file of interest |
| plot\_all | Logical | Indicator of whether to provide visualization to the user. Values : True or False |
| plots\_to\_file | Logical | Indicator of whether or not to save visuals. Currently not implemented. |
| tactile\_padding | Numeric | Value (in seconds) to right pad each tactile point event |
| auditory\_padding | Numeric | Value (in seconds) to right pad each auditory point event |
| behavior\_types | List | List outlining the behavioral states expected. Required : mom\_auditory\_types, mom\_tactile\_types, mom\_visual\_types, baby\_visual\_types, missing\_types. These categories define the sensory domains as in the original paper. See the example code above for the default values expected in each category. |
| missing\_threshold | Numeric | Value (a proportion) that indicators how much missingness is acceptable. This threshold is defined to be the percentage of the behavioral sequence that is represented by missing types defined in missing\_types. |

#### Example Usage

ber\_analyze\_file('./testfiles/Entropy\_6m - 88888HE - Event Logs.xlsx')

#### Output

ber\_analyze\_file returns an R list whose first element is a data.frame named estimates and a second element that is a list called file\_checks. The element estimates is the primary object of interest and has the following column headings

|  |  |
| --- | --- |
| Column Header | Description |
| SubjectID | Subject ID found during analysis of file |
| CanEstimateEntropy | Indicator of success or failure in estimating entropy rate |
| EntropyRate | Estimate of entropy rate between 0 and where is the total number of states |
| TotalNumberOfTransitions | The total number of events considered in the final event sequence |
| CombinedVideoDuration | Total duration of the video (endtime + final duration) |
| PercentMissing | Proportion of time represented by missing\_types category |
| AuditoryCounts | Total number of events in auditory category |
| AuditoryTotalTime | Total time representing events in auditory category |
| AuditoryAverageTime | Average duration of events in auditory category |
| VisualCounts | Total number of events in visual category |
| VisualTotalTime | Total time representing events in visual category |
| VisualAverageTime | Average duration of events in visual category |
| TactileCounts | Total number of events in tactile category |
| TactileTotalTime | Total time representing events in tactile category |
| TactileTotalTime | Average duration of events in tactile category |

During the course of running, the script performs a series of quality assurance tests of the file that has been input. Specifically it performs the following tests:

|  |  |
| --- | --- |
| Test Name | Description |
| header\_pass | Checks for the following column headers : c('Observation', 'Behavior', 'Time\_Relative\_sf', 'Duration\_sf', 'Event\_Type') |
| subjid\_pass | Checks for an entry in the first cell of the column Observation. Does not check format of subject IDs |
| misdat\_pass | Checks for empty cells in the columns list above using sum(is.na(behavior\_data$Behavior)) > 0 |
| blabel\_pass | Finds the unique entries in the column Behavior, then compares that with the labels in behavior\_types. Reports unused labels, i.e., labels found in Behavior that were not including in behavior\_types. |
| elabel\_pass | Finds the unique entries in the column Event\_Type, then compares with the following list : c('State start', 'State point', 'Point', 'State stop', 'State Stop'). These are currently supported events types. |
| misnes\_pass | Checks the proportion of time representing missing data based off of the inputs to missing\_types within behavior\_types. |

#### Example output from script:

When running ber\_analyze\_file by itself, the following is an example of what is printed to the console at runtime to help diagnose issues with files:

-----------------------------------------------------------------------  
Filename: Entropy\_6m - 88888HE - Event Logs.xlsx   
Time of Analysis: 2018-06-04 10:33:12   
-----------------------------------------------------------------------  
\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Performing File Check \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
- Checking for required Columns:  
 Observation : First cell used to set "SubjectID"  
 Behavior : Set of used behavior labels  
 Time\_Relative\_sf : Sets the start point for each action  
 Duration\_sf : Time\_Relative\_sf + Duration\_sf sets end points  
 Event\_Type : Defines point events and states  
--- PASSED : Found all Required Column Headers  
- Checking "Observation" Column For Subject ID  
--- PASSED: Using Subject ID from Column J, Cell 1: 88888HE   
- Checking for Missing Data in Columns  
--- "Behavior" : PASSED  
--- "Time\_Relative\_sf" : PASSED  
--- "Duration\_sf" : PASSED  
--- "Event\_Type" : PASSED  
- Checking "Behavior" Column For Unused Labels:  
--- WARNING : Unused Labels in "Behavior" Column, See Below:  
 Expected Label : "NotHoldingBaby", not used in analysis  
 Expected Label : "NoObjectInHand", not used in analysis  
--- NOTE: Investigate this if these do not look familar  
- Checking "Event\_Type" Column For Labels:  
--- PASSED : No Unused Labels in "Event\_Type" Column  
- Checking Missingness based on "missing\_types"  
--- Percent Missingness: 0   
--- PASSED : Percent missing less than threshold  
-----------------------------------------------------------------------  
\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* File Completed Successfully \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
-----------------------------------------------------------------------

### Function : ber\_analyze\_dir

The following function is another primary functions for the estimation of entropy rate. Given a directory, the function makes multiple calls to ber\_analyze\_file to analyze the excel files within that directory. Note that there is no plotting capability currently included when operating on directories.

ber\_analyze\_dir(dir\_loc,  
 tactile\_padding = 1.0,  
 auditory\_padding = 1.0,  
 behavior\_types=list(  
 "mom\_auditory\_types" = c('Vocal'),  
 "mom\_tactile\_types" = c('TouchBaby',  
 'HoldingBaby'),  
 "mom\_visual\_types" = c('ManipulatingObject'),  
 "baby\_visual\_types" = c('LookAtMomActivity'),  
 "missing\_types" = c('CantTellHolding',  
 'ActivityNotVisible',  
 'CantTellLooking')),  
 missing\_threshold = 0.1,  
 log\_file = paste(Sys.Date(), '-ber-logfile.txt', sep=''))

|  |  |  |
| --- | --- | --- |
| Input | Input Type | Description |
| dir\_loc | String | String indicating the location of the directory of interest |
| tactile\_padding | Numeric | Value (in seconds) to right pad each tactile point event |
| auditory\_padding | Numeric | Value (in seconds) to right pad each auditory point event |
| behavior\_types | List | List outlining the behavioral states expected. Required : mom\_auditory\_types, mom\_tactile\_types, mom\_visual\_types, baby\_visual\_types, missing\_types. These categories define the sensory domains as in the original paper. See the example code above for the default values expected in each category. |
| missing\_threshold | Numeric | Value (a proportion) that indicators how much missingness is acceptable. This threshold is defined to be the percentage of the behavioral sequence that is represented by missing types defined in missing\_types |
| log\_file | String | String indicating the location for where to save the diagnostic information from each run of ber\_analyze\_file |

#### Example Usage

From the directory where the files of interest are located type the following:

ber\_analyze\_dir('.')

#### Output

ber\_analyze\_dir returns a R data.frame with column headings as in ber\_analyze\_file. Each row of the data.frame represents the results from calling ber\_analyze\_file on a file.

## Input file formats

This section describes the expected input file format required for the software. It is also identifies important columns utilized in the estimation of entropy rate.

#### File Format Type

At the moment, the required format is an Excel .xlsx file.

Note: In the future this may be modifed to include .csv files, by altering specific lines of code in ber.R. Specifically, line 37 within ber\_analyze\_dir should be altered to accept different string patterns:

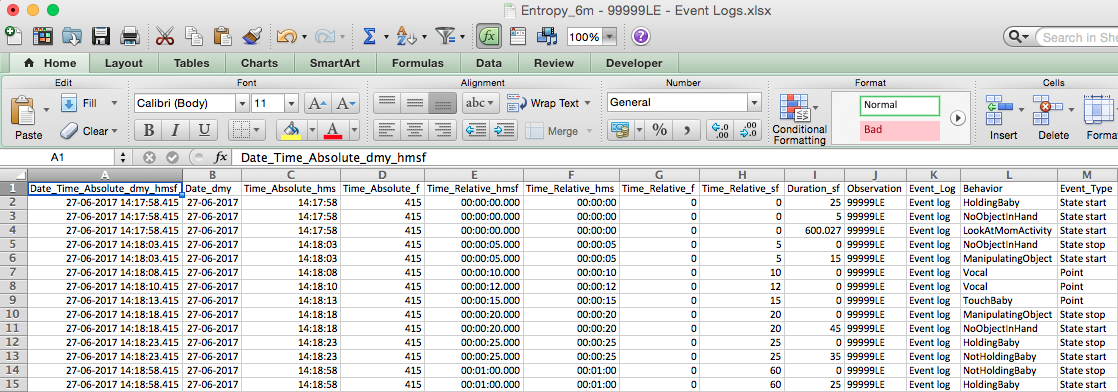
all\_files = list.files(dir\_loc, pattern="\*.xlsx")

and line 116 within ber\_analyze\_file should be changed read .csv files:

behavior\_data <- data.frame(readxl::read\_xlsx(f\_loc))

#### File Organization

Below is a visualization of a current input file that is located within the testfiles subdirectory of the source files and was used earlier in the software demonstration:



Screenshot of the test file “Entropy\_6m - 99999LE - Event Logs.xlsx”

The expected column headings are:

|  |  |  |
| --- | --- | --- |
| Column Heading | Used? | Description |
| Date\_Time\_Absolute\_dmy\_hmsf | No | Absolute Time - Day-Month-Year Hour:Min:Sec.Millisec (Military Time) |
| Date\_dmy | No | Absolute Time - Day-Month-Year |
| Time\_Absolute\_hms | No | Absolute Time - Hour:Min:Sec |
| Time\_Absolute\_f | No | ??? Appears to be millisecond part ??? |
| Time\_Relative\_hmsf | No | Relative Time - Hour:Min:Sec.Millisec |
| Time\_Relative\_hms | No | Relative Time - Hour:Min:Sec.Millisec |
| Time\_Relative\_f | No | ??? Appears to be millisecond part ??? |
| Time\_Relative\_sf | Yes | Relative Time - Seconds. Used as starting point for each instance in the video |
| Duration\_sf | Yes | Relative Time - Seconds. Duration of the the instance. The time of an event is Time\_Relative\_sf + Duration\_sf |
| Observation | Yes | Typically Subject ID. Cell J1 is used as the Subject ID that is reported from ber\_analyze\_dir and ber\_analyze\_file |
| Event\_Log | No | Unknown |
| Behavior | Yes | Used to find and match against behavior\_types specified in ber\_analyze\_dir and ber\_analyze\_file |
| Event\_Type | Yes | State start indicates cells which will be used as *events*. Point instances that are included will be right padded to become events. State stop rows are ignored. |

The order of the rows should **not** matter, but it is best to be safe and keep the organization that is above for ease of processing.

### Input File Tests

As stated above the following tests are run using ber\_analyze\_file and are shown again below:

|  |  |
| --- | --- |
| Test Name | Description |
| header\_pass | Checks for the following column headers : c('Observation', 'Behavior', 'Time\_Relative\_sf', 'Duration\_sf', 'Event\_Type') |
| subjid\_pass | Checks for an entry in the first cell of the column Observation. Does not check format of subject IDs |
| misdat\_pass | Checks for empty cells in the columns list above using sum(is.na(behavior\_data$Behavior)) > 0 |
| blabel\_pass | Finds the unique entries in the column Behavior, then compares that with the labels in behavior\_types. Reports unused labels, i.e., labels found in Behavior that were not including in behavior\_types. |
| elabel\_pass | Finds the unique entries in the column Event\_Type, then compares with the following list : c('State start', 'State point', 'Point', 'State stop', 'State Stop'). These are currently supported events types. |
| misnes\_pass | Checks the proportion of time representing missing data based off of the inputs to missing\_types within behavior\_types. |

Be sure to check the log files and output using ber\_analyze\_dir to ensure that the files are processed correctly.