Single Cell Response Curve (SCRC) - R package

User Manual

K. Nienałtowski*, R.E. Rigby, J.Walczak, K.Zakrzewska, J.Rehwinkel, M. Komorowski

03 maja 2020

Contents

*author and maintainer, please contact via karol.nienaltowski a t gmail.com

Preliminaries

Requirements - Hardware

- A 32 or 64 bit processor (recommended: 64bit)
- 1GHz processor (recommended: multicore for a comprehensive analysis)
- 2GB MB RAM (recommended: 4GB+, depends on the size of experimental data)

Requirements - Software

The main software requirement is the installation of the R environment (version: >= 3.2), which can be downloaded from R project website and is distributed for all common operating systems. We tested the package in R environment installed on Windows 7, 10; Mac OS X 10.11 - 10.13 and Ubuntu 18.04 with no significant differences in the performance. The use of a dedicated Integrated development environment (IDE), e.g. RStudio is recommended.

Apart from a base installation of R, SLEMI requires the following R packages:

1. for installation

- devtools

2. for estimation

- nnet
- doParallel (if parallel computation are needed)

3. for visualisation

- ggplot2
- ggthemes
- grDevices
- viridis

4. for data handling

- data.table
- reshape2
- dplyr
- foreach

Each of the above packages can be installed by executing

```
install.packages("name_of_a_package")
```

in the R console.

Importantly, during installation availability of the above packages will be verified and missing packages will be automatically installed.

Installation

The package can be directly installed from GitHub. For installation, open RStudio (or base R) and run following commands in the R console

```
install.packages("devtools") # run if 'devtools' is not installed
library(devtools)
install_github("sysbiosig/SCRC")
```

Are required packages not found, they will be installed automatically.

Citing and support

All problems, issues and bugs can be reported here:

https://github.com/sysbiosig/SCRC/issues

or directly via e-mail: karol.nienaltowski a t gmail.com.

Package functionalities

- Define SCRC and confusion matrix

Package structure

The SCRC package provides their functionalities with three main functions:

1. `SCRC()`- estimates the single cell dose-response curve and cell-to-cell heterogeneity structure.
2. `plotConfusionMatrix()` - visualises the cell-to-cell heterogeneity structure as fractions of cells stimulated with one dose (rows) with responses typical to either of the doses (columns).
3. `plotSCRCWaves()` visualises single cell dose-response curve, line correspons do the fractions of cells that exhibit different response levels as the dose increases, whereas the bands representing the cell-to-cell heterogeneity structure.

Morevoer, package contains examplary datasets, that were used in the publication:

1. `data.scrc.cytof`
2. `data.scrc.ps1`
3. `data.scrc.ps3`
4. `data.scrc.nfkb`

Input data

The function `SCRC()` takes `data` in the form of the object `data.frame` with a specifc structure of rows and columns. Responses y_j^i are assumed to be measured for a finite set of stimuli levels $x_1, x_2, \ldots, x_m$. The responses y_j^i can be multidimensional. Usually, experimental dataset is represented as a table with rows and columns organized as shown in Figure 1.

Data example

An example of the input `data.frame`, which contains the multivariate dose-responses to IFN-a2a in **monocytes CD14+ CD16-** presented in the **MP** is available within the package under the variable `SCRC::data.scrc.cytof`. It has the following format

Stim	pSTAT1	pSTAT3	pSTAT4	pSTAT5	pSTAT6
25	9.423005	0.0000000	0.000000	0.2226261	1.0420985
25	14.077361	0.0000000	0.000000	4.4516195	0.0000000
25	6.634479	2.3057582	0.000000	0.0000000	0.8014894
25	5.739593	0.0000000	3.807480	0.3100144	0.6536241
25	24.686699	0.0000000	0.000000	0.1629745	1.2835708
25	21.104772	0.3036974	1.324072	3.2628903	0.6757941

where each row represents measurements of a single-cell, the column named `Stim` specifies the dose level of IFN-a2a, while `pSTAT1`, `pSTAT3`, `pSTAT4`,`pSTAT5`,`pSTAT6` are the normalized levels of phosporylated STATs in an individual cell. The above table can be shown in R by calling

input		output 1	output 2	output 3	...
n_1	x_1	$y^1_{1,1}$	$y^1_{1,2}$	$y^1_{1,m}$	
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
	x_1	$y^1_{n_1,1}$	$y^1_{n_1,2}$	$y^1_{n_1,m}$	
n_2	x_2	$y^2_{1,1}$	$y^2_{1,2}$	$y^2_{1,m}$	
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
	x_2	$y^2_{n_2,1}$	$y^2_{n_2,2}$	$y^2_{n_2,m}$	
					...
$\vdots$		$\vdots$	$\vdots$	$\vdots$	
n_m	x_m	$y^m_{1,1}$	$y^m_{1,2}$	$y^m_{1,m}$	
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
	x_m	$y^m_{n_m,1}$	$y^m_{n_m,2}$	$y^m_{n_m,m}$	

Figure 1: Standard output graph presenting probabilities of correct discrimination between each pair of input values.

```
head(SCRC::data.scrc.cytof)
```

Esitmation of the SCRC

Then, main function is called as:

```
model <-  SCRC::SCRC(
  data = data,
  signal = "input",
  response = c("output_1", "output_2", "output_3", ...),
  bootstrap.number = bootstrap.number,
  ...
)
```

Variables `signal` and `response` describes respectively dose level and single-cell responses. These columns should be of type `numeric`; order and number of outputs should be the same for all cells. Number of observations in `data` shoulb large, possibly >100, per input value is required.

The variable `bootstrap.number` represents number of bootstrap samples required for estimation of cell-to-cell heterogeneity. It is crucial to choose this value carefully, as it induce estimator accuracy.

The result of the function is an object of class `SCRCModel`, that contains results of the estimator. To see the results call :

```
print(model)
```

To get the SCRC (cumulative frequency) and cell-to-cell heterogeneity (confusion matrix) please call `model$scrc` and `model$confusion.matrix`, respectively.

The result can be visualised using one of our plots, as it was presented in the publication.

SCRC can be plotted using function:

```
plotSCRCWaves(model)
```

The cell-to-cell heterogeneity can be plotted using function:

```
plotConfusingMatrix(model)
```

Example of usage

Below, we present an application of `SCRC` package to the case of the multivariate dose-responses to IFN-a2a in **monocytes CD14+ CD16-** described above.

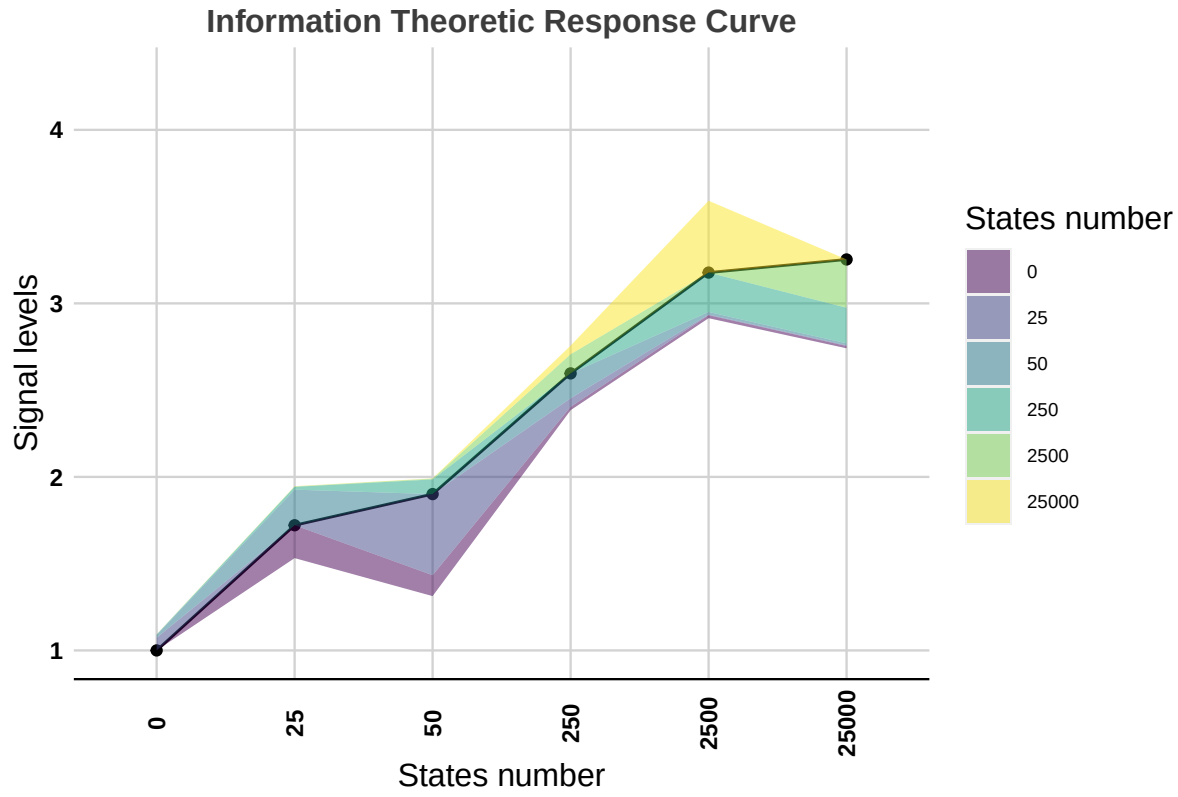
```
library(SCRC)
model <-
  SCRC(
    data = SCRC::data.itrc.cytof,
    signal = "Stim",
    response = c("pSTAT1", "pSTAT3", "pSTAT4", "pSTAT5", "pSTAT6"),
    bootstrap.number = 64)
```

The result is called by:

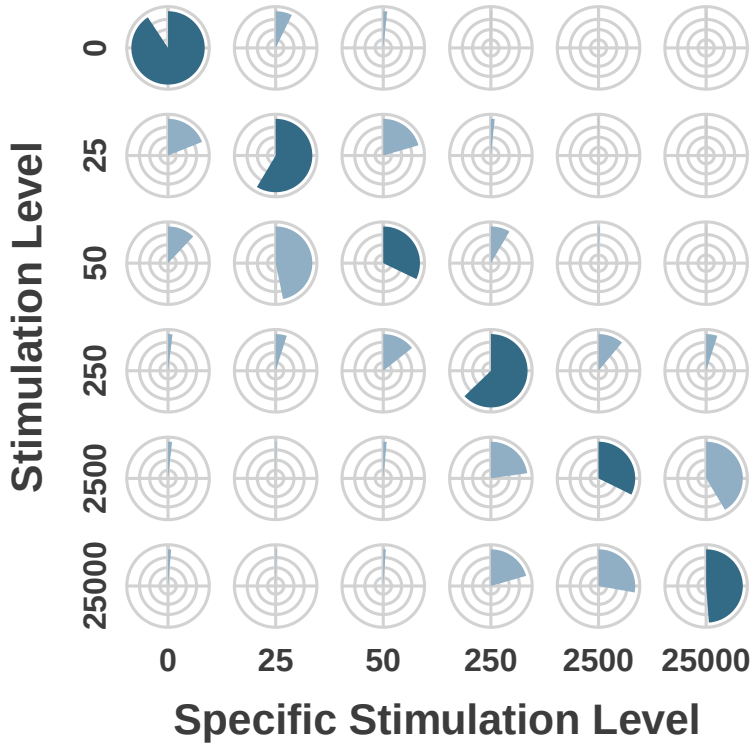
```
print(model)
```

```
## SCRCModel
## formula : Stim ~ pSTAT1+pSTAT3+pSTAT4+pSTAT5+pSTAT6
## SCRC :
##     0    25    50   250  2500 25000
##  1.00  1.72  1.90  2.60  3.18  3.25
## confusion matrix :
##          0   25   50  250 2500 25000
## 0     0.91 0.07 0.02 0.00 0.00  0.00
## 25    0.19 0.59 0.20 0.02 0.00  0.00
## 50    0.12 0.47 0.32 0.08 0.00  0.00
## 250   0.02 0.05 0.14 0.63 0.11  0.05
## 2500  0.02 0.00 0.02 0.23 0.32  0.41
## 25000 0.01 0.00 0.01 0.21 0.28  0.49
```

```
SCRC::plotSCRCWaves(model = model)
```




```
SCRC::plotCofusionMatrix(model = model)
```



Details of SCRC packgae functions

SCRC

Calculation of the SCRC and cell-to-cell heterogeneity with default settings is perfomed by the command

```
model <-   SCRC(
  data,
  signal = "signal",
  response = "response",
  sample = "sample",
  bootstrap.number = 0,
  bootstrap.sample_size = 1000,
  parallel_cores = 1,
  lr_maxit = 1000,
  MaxNWts = 5000,
  ...
)
```

The required arguments are:

- `data` - a data.frame or data.table object in a wide format that describe response (might be multi-dimmensional) of the samples to the signal (now only one dimmensional); data.frame data consists columns of names defined by sample, signal (optional), and response; each row represents a response of one sample to the input signal; column signal define the input signal; columns response define the multidimmensional (optional) response to the input signal; column sample specify identifaction of sample; if sample is not defined then sample is identified by row number;
- `signal` - character, specify name of the column that represents the input signal;

- `response` vector of characters, that specify names of the columns that represents the output response;
- `sample` - character (optional), specify name of the column that consists identifiaction of sample;
- `parallel_cores` - specify number of cores used for computations, `default = 1`
- `bootstrap.number` (`default = 1`) - numeric, `bootstrap.number >= 1`, specify nymber of bootstrap samples used for estimation SCRC and cell-to-cell heterogeneity. It is crucial to choose this value carefully, as it induce estimator accuracy. The proper value depends on data dimmensions and density distribution. The practice indicates that the higher number of bootstrap samples are required to obtain satisfying level of the accuracy of the cell-to-cell heterogeneity estimator. The `bootstrap.number = 1` denotes that one bootstrap sampling is performed to guarantee equipotence between number of cells for each dose, that is assumed in method;
- `bootstrap.sample_size` - numeric, size of the bootstrap sample;
- `lr_maxit` (`default = 1000`) - a maximum number of iterations of fitting step of logistic regression algorithm in `nnet` function. If a warning regarding lack of convergence of logistic model occurs, should be set to a larger value (possible if data is more complex or of a very high dimension);
- `MaxNWts` (`default = 5000`) - a maximum number of parameters in logistic regression model. A limit is set to prevent accidental over-loading the memory. It should be set to a larger value in case of exceptionally high dimension of the output data or very high number of input values. In principle, logistic model requires fitting $(m-1)\cdot(d+1)$ parameters, where m is the number of unique input values and d is the dimension of the output.

The function returns a list with the following elements

plotSCRCWaves

```
plotSCRCWaves(
  model,
  title_ = "Information Theoretic Response Curve",
  xlab_ = "States number",
  ylab_ = "Signal levels",
  fill.guide_ = "legend",
  ylimits_ = TRUE,
  alpha_ = 0.5,
  getScaleY = getScaleY.SCRC,
  theme.signal = NULL,
  ...
)
```

The required arguments are:

- model - SCRCModel object return by SCRC function
- title_ - character, specify title of plot, default "Information Theoretic Response Curve"
- xlab_ - character, label of x axes, default "States number"
- ylab_ - character, label of y axes and legend title, default "Signal levels"
- fill.guide_ - logical, specify if legend should be displayed
- ylimits_ - logical (TRUE or FALSE) or vector of minimum and maximum of y axes,
- theme.signal - optional, object returned by GetRescaledSignalTheme

... Arguments passed on to rescaleSignalsValues

rescale.fun parameter, that defines a function used for rescaling signals in plots. There are three built-in functions, that can be chosen: (1) 'factor' - signals treated as factors (default) with levels defined in list rescale.fun.args, (2) 'numeric', (3) logarithmic - with base defined in rescale.fun.args - default: e = exp(1). Function must be defined as a lambda construct function(x, ...){...}.

rescale.fun.args list of the arguments to defaults rescale.fun

plotConfusionMatrix