Usage of package rain

Paul F. Thaben

September 1, 2015

Contents

1 Introduction

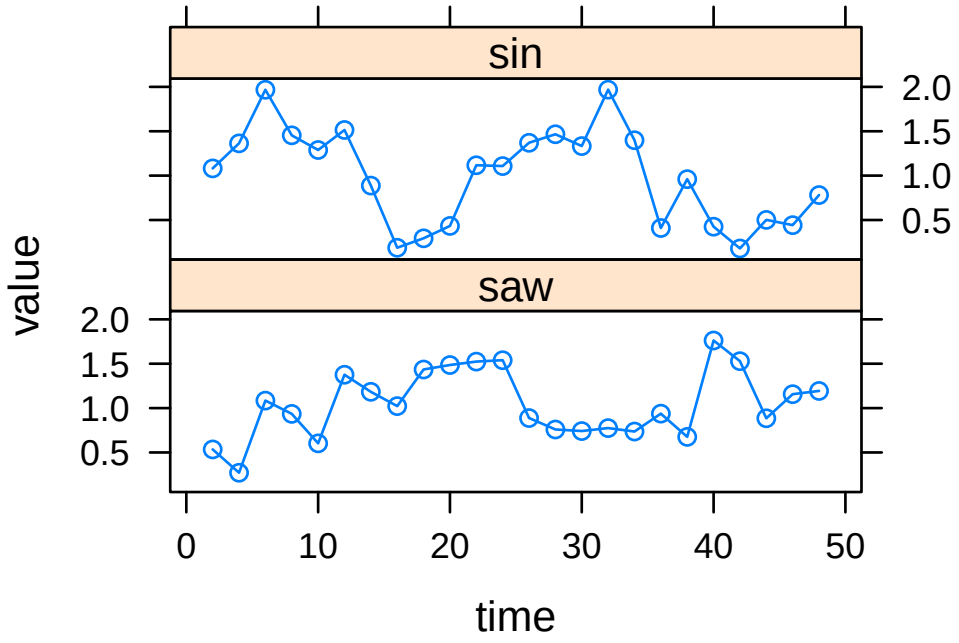
Rain is an algorithm that detects rhythms in time series. It tests against umbrella alternatives (Mack & Wolfe (1981)) to detect rhythmic rising and falling patterns.

2 Usage

2.1 General Usage

To demonstrate rains capability we generate 2 example time series. This set of time series must be given as a matrix, with time points as rows and series as columns.

```
> set.seed(123)
> times <- c(1: 24) * 2
> sin <- 1 + 0.5 * sin(times / 24 * 2 * pi) + rnorm(24, 0, 0.3)
> saw <- rep(13:24 / 18 , 2) + rnorm(24, 0, 0.3)
> measure <- cbind(sin, saw)
> require('lattice')
> xyplot(t(measure)~rep(times, each=2) | c('sin', 'saw'),
+     layout = c(1, 2), type = 'o', xlab = 'time', ylab = 'value', cex.lab = 0.6)
```



This matrix is entered into rain with some additional parameters, which are described below.

```
> require(rain)
> rainresult <- rain(measure, period=24,
+                    deltat=2, peak.border=c(0.1,0.9),
+                    verbose=FALSE
+ )
> rainresult
```

```
            pVal phase peak.shape period
sin 4.062931e-07     8         10     24
saw 9.617898e-03    24          2     24
```

To evaluate a time series in rain, the specification of the searched period length and the sample interval are essential:

Key arguments for calling rain

`x`	The set of time series as a matrix, one column per series, one row per time point
`period`	period to test for
`period.delta`	if a range of periods should be searched this interval is specified according to [`period` − `period.delta`; `period` + `period.delta`]
`deltat`	time difference between two data points. `deltat` uses the same scale as `period` and `period.delta`

Additional parameters facilitate the testing of more complex time series, and take care of special properties of the time series, such as missing values or damping effects.

Other arguments

`na.rm`	if the measurements contain NA values, these are treated as never measured and null distributions for all series are calculated individually (takes longer)
`method`	`'independent'` or `'longitudinal'`: different variants of data interpretation (see subsection)
`nr.series` and `measure.sequence`	different possibilities to specify multiple experiments and irregular time series (see subsection)
`peak.border`	range of skewness to look for (see subsection)
`verbose`	logical value: show progress status while running

2.2 Application on a real dataset

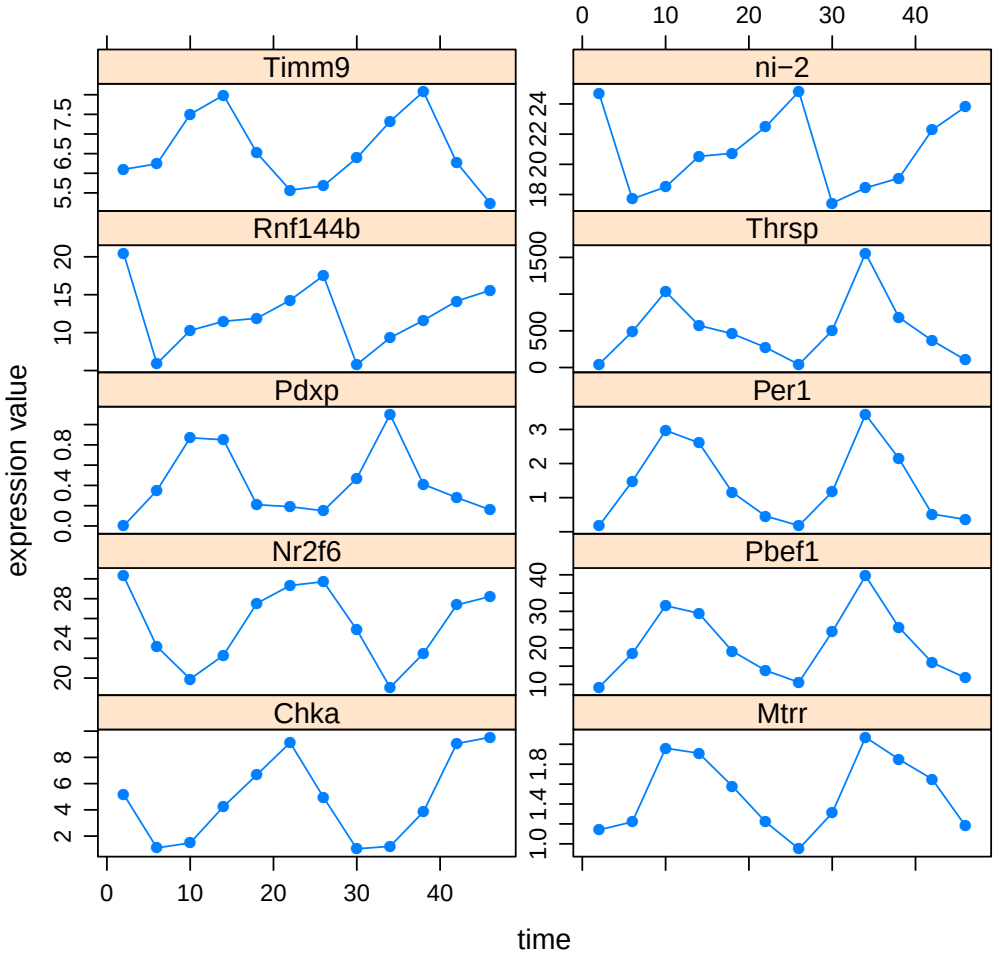
The usage on a realistic dataset is similar. As an example we use the high throughput sequencing profiles of gene expression in mouse liver from Menet et. al. 2012:

```
> data(menetRNASeqMouseLiver)
> colnames(menetRNASeqMouseLiver)
```

```
 [1] "ZT_2_1"  "ZT_2_2"  "ZT_6_1"  "ZT_6_2"  "ZT_10_1" "ZT_10_2" "ZT_14_1" "ZT_14_2"
 [9] "ZT_18_1" "ZT_18_2" "ZT_22_1" "ZT_22_2"
```

We have one series per gene, and for each gene 12 measurement at 6 different time points, so each time point has two independent repeats. To treat these repeats correctly, we have to set `nr.series=2` or `measure.sequence=c(2,2,2,2,2,2)`. Furthermore, data have to be transposed as they contain one column per time point.

```
> results <- rain(t(menetRNASeqMouseLiver), deltat=4, period=24, nr.series=2,
+                 peak.border=c(0.3, 0.7), verbose=FALSE)
> best <- order(results$pVal)[1:10]
> xyplot(as.matrix(menetRNASeqMouseLiver
+             [best, (0:5 * 2 + rep(c(1, 2), each = 6))])
+         ~rep(0:11 * 4 + 2, each = 10) |rownames(menetRNASeqMouseLiver)[best],
+         scales = list(y = list(relation = 'free')),
+         layout = c(2, 5), type = 'b', pch = 16, xlab = 'time',
+         ylab = 'expression value', cex.lab = 1)
```



The Top 10 Results showing, besides Per1, a gene of the central molecular mechanism that generates daily oscillations in vertebrate cells, also some genes with highly asymmetric oscillations (ni-2, Rnf144b, Mtrr) which might be overseen when only sine waves are assumed.

2.3 Dataset format and specification

The algorithm allows exact calculation also for time series with multiple measurements of the same time points or missing values. These properties can be triggered by the argument `nr.series` and `measure.sequence`. Also, `x` has constraints which are necessary to interpret all in a correct manner. This section explains what is possible and how the settings are provided to the function.

Regular time series, no repeats

In the simplest case where all time points are equally spaced and there is one measurement per time point, the default settings are valid. Then the ordering of the the rows in the matrix `x` is equal to the temporal order of the measurements.

Regular time series, regular repeats

A regular time series is measured with multiple independent repeats with the the same number of repeats per time point. The number of repeats for each time series is provided by the argument `nr.series`. In the matrix `x`, the repeats are ordered as followed

time	repeat
1	1
1	2
2	1
2	2
3	1
3	2

Regular time series, irregular repeats

If the number of repeats differ between the time points, the number of repeats is stated for each time point individually using the argument `measure.sequence`. The numbers of repeats for the time points are stated in temporal order. The matrix `x` is created according to the same logic as for regular repeats.

Example: A time series with the `measure.sequence` {1, 3, 1, 2} have a corresponding matrix `x`

time	repeat
1	1
2	1
2	2
2	3
3	1
4	1
4	2

Irregular time series

If the time series is not equally spaced it can be regularized by introducing time points with zero repeats. This may be combined with any of the above repeat settings.

Example: A series measured at times {1, 3, 4, 6} is usable with the `measure.sequence`{1, 0, 1, 1, 0, 1} and a matrix `x`

time	repeat
1	1
3	1
4	2
6	3

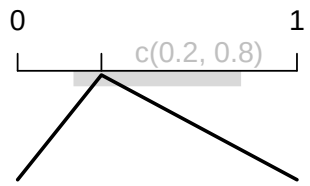
2.4 Peak shape definition and output

A key advantage of rain is the possibility to detect arbitrarily asymmetric oscillations.

Argument

The limits of the asymmetry are controlled via the `peak.border` argument. The argument consists of a vector with 2 numbers between 0 and 1. It assigns an interval of possible positions of the peak between two troughs.

Example for a peak at the point 0.3 and the interval c(0.2, 0.8)



The possible phases are mapped to the time points in the measurement by rounding down/up to the next measurement time point for the lower/upper border.

Output

The best matching peak shape is given in the `peak.shape` column in the result array. It is the time between a peak an the next trough calculated by the time points and the `deltat` argument.

As an example lets have a second look on the first running example with artificial time series. There is an sine wave with symmetric shape and a sawtooth shaped time series, with a long rising and a short falling part. The result table reflects this:

```
> rainresult
```

```
            pVal phase peak.shape period
sin 4.062931e-07     8         10     24
saw 9.617898e-03    24          2     24
```

In the sine wave the falling part of the best matching model contains 10 of 24 so approximately the half of the time points from one period. The sawtooth shaped series has a `peak.shape = 2` so only 2 of 24 time points are in the falling part of the oscillation.

2.5 "longitudinal" versus "independent"

There are different ways to apply the umbrella statistic on time series. The optimal method depends on the experimental setting.

longitudinal

Used for time series sampled from the same individual or cell culture. Resistant to artifacts such as trends or dampening of oscillations.

In this variant there is no reordering of the samples. If the time series contains more than one period, multiple peaks are treated independently, resulting in insensitivity with respect to different amplitudes (damping) and underlying trends. Detection of strong asymmetries could lead to false positives that in fact are pure trends.

independent

Used for time series with time points sampled independently from different biological specimen. An example would be an experiment in which different animals are scarified and assayed at each time point.

In this variant, the statistic is the same for each combination of `period` and `peak.shape`. Time points at the same phase are treated as independent repeats.

3 References

Mack, G. A., & Wolfe, D. A. (1981). K-Sample Rank Tests for Umbrella Alternatives. *Journal of the American Statistical Association*, **76(373)**, 175–181.

Menet, J. S., Rodriguez, J., Abruzzi, K. C., and Rosbash, M. (2012). Nascent-Seq reveals novel features of mouse circadian transcriptional regulation. *eLife*, **1(0)**, e00011. doi:10.7554/eLife.00011