RobPer: An R Package to Calculate Periodograms for Light Curves Based on Robust Regression

Anita M. Thieler
TU Dortmund University

Roland Fried
TU Dortmund University

Jonathan Rathjens
TU Dortmund University

Abstract

An important task in astroparticle physics is the detection of periodicities in irregularly sampled time series, called light curves. The classic Fourier periodogram cannot deal with irregular sampling and with the measurement accuracies that are typically given for each observation of a light curve. Hence, methods to fit periodic functions using weighted regression were developed in the past to calculate periodograms.

We present the R package **RobPer** which allows to combine different periodic functions and regression techniques to calculate periodograms. Possible regression techniques are least squares, least absolute deviations, least trimmed squares, M-, S- and τ-regression. Measurement accuracies can be taken into account including weights. Our periodogram function covers most of the approaches that have been tried earlier and provides new model-regression-combinations that have not been used before.

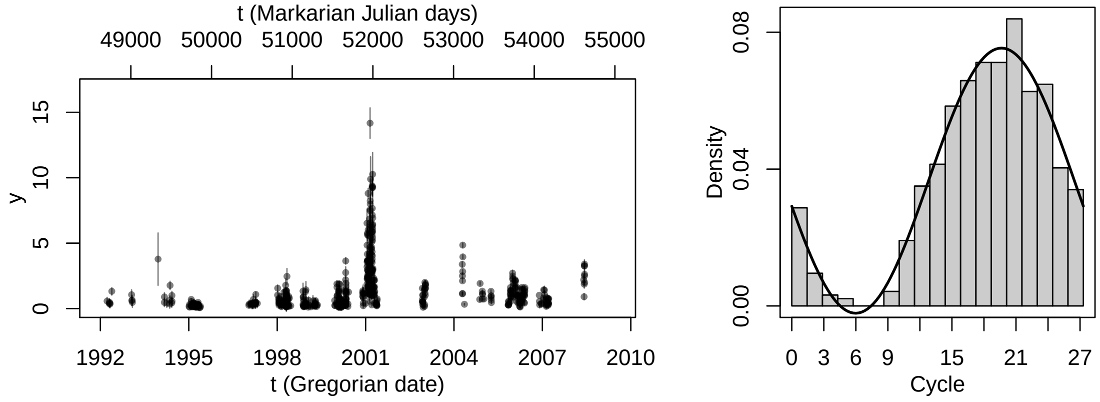
To detect valid periods, **RobPer** applies an outlier search on the periodogram instead of using fixed critical values that are theoretically only justified in case of least squares regression, independent periodogram bars and a null hypothesis allowing only normal white noise. Finally, the package also includes a generator to generate artificial light curves.

Keywords: periodogram, light curves, period detection, irregular sampling, robust regression, outlier detection, Cramér-von-Mises distance minimization, time series analysis, beta distribution, measurement accuracies, astroparticle physics, weighted regression, regression model.

1. Introduction

We introduce the R (R Core Team 2015) package **RobPer** (Thieler, Rathjens, and Fried 2015), which can be used to calculate periodograms and detect periodicities in irregularly sampled time series. Our special objective are light curves, which occur in astroparticle physics and are irregularly sampled times series $(t_i, y_i, s_i)_{i=1,\dots,n}$ consisting of unequally spaced observation times $t_1, \dots, t_n$, observed values $y_1, \dots, y_n$ and measurement accuracies $s_1, \dots, s_n$. The measurement accuracies s_i give information about how precise the y_i were measured. They can be interpreted as estimates for the standard deviations of the observed values. The observed values possibly contain a periodic fluctuation y_f with fluctuation period p_f and the irregularly spaced observation times t_i are realizations of random variables with a periodically shaped density.

Such periodicity in the pattern of the observation times is a typical phenomenon, as the sampling of astroparticle physics' time series is influenced among others by astronomical constellations. For example, plotting a histogram of the observation times for the gamma



(a)

(b)

Figure 1: Light curve with gamma particle emissions for the very high energy gamma particle source Mrk 421 (see Tluczykont *et al.* 2010, and references therein). Panel 1a shows the light curve, vertical lines at each point show the reported measurement accuracies. Panel 1b depicts a histogram of the observation times t_i modulo the period $p_s = 27.31$. A sine represents the shape rather well.

particle source Mrk 421 modulo the period $p_s = 27.31$ shows an unequal distribution over a cycle of this length (see Figure 1). This is due to the fact that observations cannot be sampled during full moon and the moon period is similar to p_s.

So we assume the following model for the observations indexed by $i = 1, \dots, n$:

$$\begin{aligned}
&T_i = T^\star_{(i)}, && T^\star_1, \dots, T^\star_n \sim \mathcal{D}(p_s) \text{ i.i.d.}, && (1)\\
&Y_i = Y_{f;i} + Y_{w;i}, && && (2)\\
&Y_{f;i} = f\left(\frac{T_i}{p_f}\right), && f(\xi) = f(\xi+1)\ \forall \xi \in \mathbb{R} && (3)\\
&Y_{w;i} \sim \mathcal{N}(0, \sigma_i^2), && && (4)\\
&s_i : \text{given estimate for } \sigma_i \text{ independent from } Y_1, \dots, Y_n, && &&
\end{aligned}$$

where $T^\star_{(i)}$ denotes the ith ordered observation time in $T^\star_1, \dots, T^\star_n$ and $\mathcal{D}(p_s)$ is a periodic sampling density with period p_s. The observation times $t_1, \dots, t_n$ and the observed values $y_1, \dots, y_n$ are realizations of $T_1, \dots, T_n$ and $Y_1, \dots, Y_n$, respectively. We assume the observation times to be measured without error. $Y_{f;i}$ is the systematic periodic component in the observations, corresponding to an unknown periodic function f and the period p_f we are searching for. $Y_{w;i}$ is additive noise.

To detect a periodic fluctuation with period p_f in the observed values y_i, it is not possible to use the standard periodogram of Fourier analysis. This method can only be applied to time series with equidistant observation times, while light curves are typically irregularly sampled. A setting-adapted procedure, the Deeming periodogram (Deeming 1975), is not recommendable either in this case, because it is known to react to a periodicity p_s in the sampling (see Hall and Li 2006).

In order to determine periodicity in light curves, other methods than the classical Fourier periodogram or the Deeming periodogram should be used. Popular periodogram methods in astroparticle physics are for example the Lomb-Scargle periodogram (Scargle 1982) or the phase dispersion minimization periodogram (Stellingwerf 1978). These and many other approaches can be generalized to fitting periodic functions to the light curve using least squares regression and calculating periodogram bars based on SE and SY, where SE is the remaining variance in the residuals of the fit and SY is the overall variance in the observed values y_i. An even broader class of periodogram methods additionally allows application of robust regression instead of least squares regression and weighted regression to take the measurement accuracies s_i into account.

The function `RobPer` in our homonymous R package calculates a periodogram of a light curve based on fitting periodic functions to $(t_i, y_i)_{i=1,\dots,n}$ using least squares or a robust regression technique, optionally taking measurement accuracies s_i into account using weighted regression. The coefficient of determination corresponding to the objective function of the regression technique is used as periodogram bar. This proceeding incorporates analogues to most of the existing periodograms and introduces several new techniques. Preliminary implementations of most of these periodogram methods have been compared by Thieler, Backes, Fried, and Rhode (2013). Here, we explain the usage of the R package **RobPer**, which makes improved and extended methods for period detection publicly available.

This article is organized as follows: In Section 2, the usage and the structure of the function `RobPer` are explained. Especially, the different periodic functions and regression techniques are discussed and related to the existing periodogram methods. Diagrams which show how this R function is implemented in detail are displayed in Appendix A. Section 3 is devoted to the question how to find valid periods using a periodogram. Thieler *et al.* (2013) propose robust fitting of a beta distribution combined with outlier detection. The function `betaCvMfit` in the package **RobPer** performs this. In Section 4, the function `tsgen` is presented which allows to generate artificial light curves. Some examples for how to use the package are given in Section 5. Section 6 concludes with a summary.

The **RobPer** software package is available from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=RobPer`. Other R packages implementing periodograms of irregularly sampled time series are the packages **lomb** (Ruf 1999, function `lsp`), **cts** (Wang 2013, function `spec.ls`) and **nlts** (Bjornstad 2013, function `spec.lomb`). They calculate the Lomb-Scargle periodogram, which is based on the least squares fit of a sine function. Furthermore, the package **GeneCycle** (Ahdesmäki, Fokianos, and Strimmer 2012, function `robust.spectrum`) fits sine functions using robust M-regression to calculate a periodogram based on the square of the estimated amplitude. None of these functions permits taking measurement accuracies using weighted regression into account and most of them (apart from the function `spec.lomb`) have restrictions concerning the trial periods fitted.

2. Calculate periodograms with `RobPer`

The R function `RobPer` calculates a periodogram of a given light curve $(t_i, y_i, s_i)_{i=1,\dots,n}$. This is done by fitting a periodic function g to the data $(t_i, y_i)_{i=1,\dots,n}$. The function g has m parameters entering g linearly. It has a period of 1 and is transformed by $g\left(\frac{t}{p_j}\right)$ for each given trial period $(p_j)_{j=1,\dots,q}$. A simple example is $g(t) = \sin(2\pi t)\beta_1 + \cos(2\pi t)\beta_2$. The periodogram bars for the different trial periods are defined as the coefficients of determination of the respective fits. Using weighted regression with weights $1/s_i$ makes it possible to take

Argument	Comment
`ts` $\in \mathbb{R}^{n\times 3}$ or $\mathbb{R}^{n\times 2}$	Light curve (t_i, y_i, s_i) or (t_i, y_i), $i = 1, \ldots, n$; If `weighting = FALSE` the measurement accuracies s_i column may be omitted.
`weighting` $\in$ {`T`, `F`}	If `TRUE`, weighted regression is performed to take into account the s_i.
`periods` $\in \mathbb{R}^q_{>0}$	Trial periods $p_1, \ldots, p_q$.
`regression`	Regression technique (see Section 2.2), possible choices: `"L2"`, `"L1"`, `"LTS"`, `"S"`, `"huber"`, `"bisquare"`, `"tau"`.
`model`	Periodic fluctuation to be fitted (see Section 2.1), possible choices: `"step"`, `"2step"`, `"sine"`, `"fourier(2)"`, `"fourier(3)"`, `"splines"`.
`steps` $\in \mathbb{N}$	Number of steps per cycle for periodic step functions. Default: 10
`var1` $\in$ {`T`, `F`}	`TRUE` sets variance estimate to one for weighted M-regression. Default: `weighting`
`tol` $\in \mathbb{R}_{>0}$	Precision for convergence criteria. Used in case of M-regression and in case of LTS regression if `LTSopt = TRUE`. Default: 10^{-3}

`genoudcontrol` $\in \mathbb{N}^3$	Settings for `genoud` (see paragraph about LTS regression in Section 2.2):
	`max.generations`, `wait.generations`, `pop.size`
	Used if `regression = "bisquare"` or `LTSopt = TRUE & regression = "LTS"`.
	Default: {50, 5, 50}

`LTSopt` $\in$ {`T`, `F`}	Determines whether the regression result of `ltsReg` should be optimized. Default: `TRUE` if `regression = "LTS"`
`taucontrol` $\in \mathbb{N}^4\times$ {`T`, `F`}	Settings for τ-regression: `N`, `kk`, `tt`, `rr`, `approximate`. Used if `regression = "tau"`, `rr` only necessary for `approximate = TRUE`. Default: {100, 2, 5, 2, `FALSE`}
`Scontrol` $\in \mathbb{N}^3 \times \mathbb{R}^2_{>0} \times \mathbb{N}$	Settings for S-regression: `N`, `kk`, `tt`, `b`, `cc`, `seed`. Used in case of `regression = "S"`. `seed` can be fixed in order to get reproducible results or can be left empty. Default: {`N`, 2, 5, 0.5, 1.547, `NULL`} with `N` $= 50$ if `weighting = FALSE` and `N = 200` if `weighting = TRUE`.

Return value	
`periodogram` $\in \mathbb{R}^q$	Vector of periodogram bars belonging to the trial periods.
Possibly warnings	

Table 1: Arguments and return values of the function `RobPer`. {`T`, `F`} means {`TRUE`, `FALSE`}.

the measurement accuracies into account. As the shape of the true fluctuation f in Equation 3 is usually unknown, we will typically have $g \neq f$.

Table 1 gives an overview over all arguments of `RobPer`. The possible shapes of the function

g that may be fitted by `RobPer` are presented in Section 2.1. Fitting them using least squares regression is in many cases equivalent to already existing periodogram methods (see Table 2 or Thieler *et al.* 2013 for a more detailed discussion).

In addition to least squares regression, `RobPer` offers a selection of robust regression techniques to fit $g\left(\frac{t}{p_j}\right)$, see Section 2.2. All regression techniques implemented in `RobPer` are based on minimizing an objective value

$$\mathrm{SE} = \zeta\left(y - X\beta\right) \tag{5}$$

with respect to the unknown parameter value $\beta \in \mathbb{R}^m$, where $X \in \mathbb{R}^{n\times m}$ is the design matrix containing the known components of $g\left(\frac{t}{p}\right)$ at the measurement times $t_1, \ldots, t_n$ with p being a trial period and y the vector of observations $y_1, \ldots, y_n$. In the simple example mentioned above, the ith row of X has the elements $\sin(2\pi t_i/p)$ and $\cos(2\pi t_i/p)$. The function $\zeta : \mathbb{R}^n \to [0, \infty[$ is chosen according to the regression method, e.g., $\zeta(r) = \sum_{i=1}^n r_i^2$ for least squares regression. Using the same regression technique, the location μ of the observations $y_1, \ldots, y_n$ can be estimated minimizing

$$\mathrm{SY} = \zeta\left(y - \mathfrak{i}\mu\right) \tag{6}$$

with $\mathfrak{i} = \mathbb{1}_n$ being an n-variate vector of ones in case of unweighted regression. The periodogram bar can then be calculated as $R^2 = 1 - \frac{\mathrm{SE}}{\mathrm{SY}}$. This definition for the coefficient of determination does not only apply for least squares regression, but also for least absolute deviation- (L_1) and M-regression in general (see Maronna, Martin, and Yohai 2006, p. 171) as well as for S-, least trimmed squares- (LTS) and τ-regression (see Croux and Dehon 2003).

If it is intended to take given measurement accuracies $s_1, \ldots, s_n$ into account, weighted regression can be performed. In this case, the terms y, X and $\mathfrak{i}$ in the two fitted models

$$y = X\beta + \epsilon \qquad \text{(full model)}, \tag{7}$$

$$y = \mathfrak{i}\mu + \epsilon \qquad \text{(location model)}, \tag{8}$$

$$\text{with } \epsilon \in \mathbb{R}^n, \epsilon_i \underset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2), \tag{9}$$

are replaced by $\widetilde{y}_i = y_i/s_i$, $\widetilde{X}_{ij} = X_{ij}/s_i$ and $\widetilde{\mathfrak{i}}_i = \mathfrak{i}_i/s_i = \mathbb{1}_n/s_i$, respectively. In the following, we will focus on the case of unweighted regression and only point out the handling of weighted regression, when both procedures differ.

Table 2 displays periodogram methods following the principle of fitting periodic functions. Up to now, weighted regression or robust regression in affiliation with periodic step functions has only been performed by Thieler *et al.* (2013), though the unweighted least squares versions belong to the most popular periodogram methods in this area of research. S- or τ-regression, which are also available in `RobPer`, have not been investigated up to now in this context.

2.1. Periodic function fitted: Argument `model`

For each trial period p_i, $i \in \{1, \ldots, q\}$ (given by the argument `periods`, see Table 1), a periodic function (defined by `model`) is fitted to the light curve (using regression technique `regression`). Implemented periodic functions include step functions, sine functions, Fourier series and spline functions.

Model	Regression technique	Publication (Name of the method)
`step`	`L2`	Leahy *et al.* (1983) (epoch folding)
	`L2`	Schwarzenberg-Czerny (1989) (analysis of variance)
	`L2`, `L1`, `huber`, `bisquare`	Thieler *et al.* (2013)
`2step`	`L2`	Stellingwerf (1978) (phase dispersion minimization)
	`L2`, `L1`, `huber`, `bisquare`	Thieler *et al.* (2013)
`sine`	`L2`	Scargle (1982) (Lomb-Scargle)
	`L2`	Zechmeister and Kürster (2009) (generalized Lomb-Scargle)
	`L2`	Cumming *et al.* (1999) (floating mean)
	`L2`	Ferraz-Mello (1981) (date compensated Fourier transform*)
	`L2`	Reegen (2007) (SigSpec*)
	`L1`	Li (2009)*, Li (2010)*
	`LTS`	Ahdesmäki *et al.* (2007)*
	`bisquare`	Ahdesmäki *et al.* (2007)*
	`huber`	Zhang and Chan (2005)*
	`L2`, `L1`, `huber`, `bisquare`	Thieler *et al.* (2013)
`fourier(2)`, `fourier(3)`	`L2`	Hall *et al.* (2000)
	`L2`	Palmer (2009) (Fast-χ^2)
	`L2`, `L1`, `huber`, `bisquare`	Thieler *et al.* (2013)
`splines`	`L2`	Akerlof *et al.* (1994)
	`L2`	Hall *et al.* (2000)
	`L2`	Oh *et al.* (2004) (generalized cross validation)
	`huber`	Oh *et al.* (2004) (robust cross validation)
	`L2`, `L1`, `huber`, `bisquare`	Thieler *et al.* (2013)

Table 2: Published periodogram methods that rely on fitting a periodic model g to a light curve using a regression technique. Models (see Section 2.1): periodic step functions and pairwise overlapping step functions (`step` and `2step`), the sine function (`sine`), Fourier series of second and third degree and periodic spline functions (`fourier(2)`, `fourier(3)` and `splines`). Regression techniques: See Table 3 for labels. The underlined methods can take into account measurement accuracies using weighted regression. The periodogram bars of methods marked by * do not base on SE or SY, but on the parameter vector of the function fitted (e.g., squared amplitude).

Step functions

Many periodogram methods from astroparticle physics such as the epoch folding periodogram (Leahy *et al.* 1983) or the analysis of variance periodogram (Schwarzenberg-Czerny 1989) can be interpreted as fitting a step function to a light curve (see Schwarzenberg-Czerny 1998 or Thieler *et al.* 2013). They use periodogram bars related to R^2, n and the numbers of steps per cycle.

Another typical periodogram method in astroparticle physics is the phase dispersion minimization periodogram (PDM, Stellingwerf 1978). Depending on the particular setting the periodogram bar in many cases equals the mean of the coefficients of determination of two fits with different step functions with staggered jumps (see Thieler *et al.* 2013 or Thieler 2013 for more details).

`RobPer` provides two options to fit periodic step functions. The number of steps per cycle is controlled by the argument `steps`. Using `model = "step"`, a single periodic step function with steps of equal width is fitted for each trial period. Performing `regression = "L2"`, `model = "step"` is equivalent to calculating an epoch folding- or analysis of variance periodogram. Using `model = "2step"`, two different step functions with opposed jump times and steps of equal width are fitted separately and the periodogram bar is the mean of both coefficients of determination. This is the only option where two periodic functions are fitted for one trial period. It is included to provide the PDM periodogram with overlapping bins.

Sine functions

Sine functions are periodic and quite popular for investigating periodicity. The classic periodogram of Fourier analysis for equally sampled time series represents the explained variance SE of a least squares fit of a sine model to the zero-centered time series. The Lomb-Scargle periodogram (Scargle 1982) works equivalently for unequally sampled time series.

As the mean of an irregularly sampled time series is not identical to the least squares fit of an intercept in a sine model, more recent methods use the uncentered data and fit a model with intercept, e.g., the floating mean periodogram by Cumming *et al.* (1999) and the generalized Lomb-Scargle periodogram by Zechmeister and Kürster (2009). Performing `regression = "L2"`, `model = "sine"` is equivalent to calculating those periodograms and in case of equidistant observation times also equivalent to the Fourier periodogram.

Some other methods as the Date Compensated Fourier Transform by Ferraz-Mello (1981), the SigSpec periodogram by Reegen (2007) or robust approaches by Ahdesmäki *et al.* (2007) and Zhang and Chan (2005) apply the same regression step as the floating mean- and the generalized Lomb-Scargle periodogram, but use the squared amplitude of the fitted sinusoid as the periodogram bar. In case of regular sampling, this is another representation of the classical periodogram of Fourier analysis. As the amplitude is a concept closely related to trigonometric functions, `RobPer` uses the coefficient of determination only, to obtain a general method independent of the periodic function chosen.

Further periodic functions

Recently, fitting more complex periodic functions has been proposed for periodograms. Fourier series (see Hall *et al.* 2000 and Palmer 2009) and periodic splines (see Akerlof *et al.* 1994, Hall *et al.* 2000 and Oh *et al.* 2004) may provide better adaptivity compared to sine functions, but still present a continuous function, unlike the step function. `RobPer` offers the possibility to fit Fourier series of second (`model = "fourier(2)"`) or third (`model = "fourier(3)"`)

Regression technique	`regression`	R function (**package**)
Least squares	`"L2"`	`lm` (**stats**, R Core Team 2014)
Least absolute deviations	`"L1"`	`rq` (**quantreg**, Koenker 2015)
Least trimmed squares	`"LTS"`	`ltsReg` (**robustbase**, Rousseeuw *et al.* 2015)
M-regression		
. . . with Huber function	`"huber"`	Own implementation.
. . . with Bisquare function	`"bisquare"`	`lmrob..M..fit` (**robustbase**, Rousseeuw *et al.* 2015)
S-regression	`"S"`	Slightly modified code from Salibian-Barrera and Yohai (2006).
τ-regression	`"tau"`	Slightly modified code from Salibian-Barrera *et al.* (2008).

Table 3: Regression techniques implemented in `RobPer` and R functions used to perform the regression technique. For more details see Section 2.2.

degree or a periodic spline function with four knots per cycle (`model = "splines"`). For the latter option, B-splines are generated using the function `spline.des` from the package **splines** (Bates and Venables 2016).

2.2. Regression techniques: Argument `regression`

Instead of fitting the models mentioned above by the popular least squares regression (see Table 2), `RobPer` also allows application of six robust regression techniques, see Table 3. Robust regression techniques like least absolute deviations, least trimmed squares (Rousseeuw and Yohai 1984) and M-regression (Huber and Ronchetti 1981) have already been used to fit sines (evaluating the squared amplitude) by Zhang and Chan (2005), Ahdesmäki *et al.* (2007), Li (2009) and Li (2010). M-regression with the Huber function was applied to fit periodic splines by Oh *et al.* (2004). Thieler *et al.* (2013) use least absolute deviations and M-regression and all models described in this article to calculate periodograms based on the coefficient of determination.

To the best of our knowledge, S- (Rousseeuw and Yohai 1984) and τ-regression (Yohai and Zamar 1988) have not been used before in periodogram calculation. For the latter, `RobPer` uses the algorithms Fast-S from Salibian-Barrera and Yohai (2006) and Fast-τ from Salibian-Barrera, Willems, and Zamar (2008) and slightly modified versions of the code distributed with the respective publication (see the respective paragraphs entitled in Section 2.2). The following paragraphs outline the algorithms used by `RobPer` for calculating the different regression estimators. For the basic definitions of these regression techniques we refer to the literature mentioned above and the book by Maronna *et al.* (2006).

LTS regression

The R function `ltsReg` from package **robustbase** (Rousseeuw *et al.* 2015) is used to perform LTS regression in `RobPer`. In preliminary studies we observed that the function can have problems finding a good solution for some of the candidate periods. This results in coefficients of determination which are too small or sometimes even negative. By setting `LTSopt = TRUE`, it is possible to let `RobPer` further optimize the solution of `ltsReg` by using the R function `genoud` from package **rgenoud** (Mebane, Jr. and Sekhon 2011). This function uses an evolutionary approach to improve the given solution, locally optimizing the temporarily best solutions in a gradient descent algorithm. Further arguments `pop.size` (size of one generation), `max.generations` (maximum of generations before stopping the algorithm) and `wait.generations` (maximum number of generations to wait for an improvement of the optimization criterion) control the behavior of the algorithm and can be set in `RobPer` by the argument `genoudcontrol` (see Table 1). The argument `tol` controls the precision for convergence criteria.

A further problem we observed is that `ltsReg` sometimes aborts the fit. However, it is typically able to perform the fit if it is run again. In case of a crash, `RobPer` calls `ltsReg` up to three times. After the third failed attempt, the respective periodogram bar is set to `NA`, or a least absolute deviation regression is performed. The latter is done, if the `ltsReg` regression result should be further processed, using the `genoud` algorithm or using the LTS result as initial estimate for an M-regression fit (see next paragraph).

M-regression

In case of M-regression, a periodogram bar, i.e., the coefficient of determination $R^2 = 1 - \frac{\text{SE}}{\text{SY}}$ is calculated from the values

$$\text{SE} = \min_{\beta} \sum_{i=1}^{n} \rho \left(\frac{y_i - x_i^\top \beta}{\widehat{\sigma}} \right) \tag{10}$$

and

$$\text{SY} = \min_{\mu} \sum_{i=1}^{n} \rho \left(\frac{y_i - \mathfrak{i}_i \mu}{\widehat{\sigma}} \right), \tag{11}$$

where $\widehat{\sigma}$ is an estimate of the error scale σ in the regression model. As explained above, Equations 10 and 11 represent the minimization criteria of the fits of the chosen periodic fluctuation (SE in Equation 5) and of a location estimate (SY in Equation 6), respectively. The function ρ is a distance measure. The vector $\mathfrak{i}$ consists of ones in case of unweighted regression. As mentioned before, in case of weighted regression, y_i, $\mathfrak{i}_i$ and the rows x_i of the design matrix are standardized by the measurement accuracy s_i (see Figure 12 in Appendix A).

The value $\widehat{\sigma}$ is obtained in an initial estimation of the periodic fluctuation, calculating a scale estimate of the fitted residuals. In principle, one could use a different estimate of σ calculated from fitting only an intercept in Equation 11, but Maronna *et al.* (2006, p. 171) recommend using the scale estimate from the (larger) regression model. In our context this means that SY depends on the trial period and cannot be calculated globally. On the other hand this ensures that the regression model $Y = X\beta + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is a generalization of the intercept model $Y = \mathfrak{i}\mu + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and thus SE $\leq$ SY and $R^2 \geq 0$.

So for this regression technique, an implementation is needed where the scale estimate can be fixed in advance. For M-regression using the biweight function, the function `lmrob..M..fit` from package **robustbase** by Rousseeuw *et al.* (2015) is used. This R function includes Huber M-regression only as a limiting case of Hampel M-regression with all but one of its tuning constants set to very large values. In other R functions known to us for M-regression (`rlm` from package **MASS** byVenables and Ripley 2002, `iwlsm` from package **RSiena** by Ripley, Boitmanis, and Snijders 2013 and `robustregBS` and `robustRegH` from package **robustreg** by Johnson 2015), the scale estimate cannot be fixed in advance. Hence M-regression using the Huber function is newly implemented for **RobPer**. Like the functions specified before, this implementation is based on an iteratively reweighted least squares (IRWLS) approach (see Maronna *et al.* 2006, pp. 104–105), and meets our special requirements. For M-regression using the biweight function, the implementation makes also use of the function `genoud` from package **rgenoud** (see previous paragraph) to overcome possible problems with local optima.

As noted above, weighted regression scales observed values and design matrices by the measurement accuracies. The variance of the error is expected to be about one then. Hence it can be reasonable to set $\widehat{\sigma}$ to one. This can be done in `RobPer` setting the argument `var1` to `TRUE`, as is recommendable in our experience in case of weighted M-regression.

To calculate a periodogram bar using M-regression with IRWLS, three initial estimates are needed: A scale estimate $\widehat{\sigma}$ (if not set to one) and initial location estimates $\widehat{\beta}^{(0)}$ and $\widehat{\mu}^{(0)}$ for β and μ. The initial estimates should be obtained using robust techniques. As proposed by Maronna *et al.* (2006, p. 105) we use the median (weighted if the s_i shall be taken into account) to initially estimate μ. For β, LTS regression (see previous paragraph) is used. It has a high breakdown point and is appropriate in situations with many observations not agreeing with the best fit. This situation will often occur in periodogram calculation, as many trial periods and thus many wrong models are fitted to the light curve. The scale estimate $\widehat{\sigma}$ is calculated as the (weighted) median of the residuals of the LTS fit.

S-regression

In case of `regression = "S"`, `RobPer` uses the Fast-S algorithm by Salibian-Barrera and Yohai (2006) to perform S-regression for fitting the periodic function efficiently. The algorithm starts with a set of `N` parameter candidates, locally optimizes them using `kk` iterations, then optimizes the `tt` best of these candidates until convergence and finally chooses the best parameter candidate.

The R function `FastS` used in **RobPer** is a slightly modified version of the function `fast.s` published by Salibian-Barrera and Yohai (2006). It was changed in order to work more efficiently in the context given here, especially when fitting step functions, and to specify one parameter candidate in advance. This candidate is set to

$$\widehat{\beta}_\mu = \begin{cases} (\hat{\mu}, \ldots, \hat{\mu})^\top \in \mathbb{R}^m & \texttt{model} \in \{\texttt{"step"}, \texttt{"2step"}, \texttt{"splines"}\} \\ (\hat{\mu}, 0, \ldots, 0)^\top \in \mathbb{R}^m & \texttt{model} \in \{\texttt{"sine"}, \texttt{"fourier(2)"}, \texttt{"fourier(3)"}\} \end{cases} \tag{12}$$

where m denotes the dimension of the linear model of the periodic function and $\hat{\mu}$ denotes the location estimate. $\widehat{\beta}_\mu$ arises from plugging in the fit obtained from the location model into the parametrization of the full model. This ensures that fitting the full periodic function will not give a worse fit than fitting only a location parameter. Otherwise it could happen that SY $<$ SE and the coefficient of determination (which has to be in $[0, 1]$) would be negative.

Further changes in `FastS` are:

1. The arguments `k` and `best.r` are renamed to `kk` and `tt` to unify notation as in `FastTau`. The arguments `int`, `N`, `kk`, `tt`, `b`, `cc` and `seed` are merged to a list `Scontrol`, which is also an argument of `RobPer` (except for `int`, which is fixed in `RobPer`).

2. If an intercept column is added to the design matrix (using `Scontrol$int = TRUE`), this is done before the dimension of the design matrix is determined (instead of doing this first and redoing it in case of `Scontrol$int = TRUE`).

3. To find a subsample in general position, regressors $x_{i\star}^{\top}$ are sampled from the set of rows of the design matrix X ignoring the frequency of occurrence in X. For each regressor $x_{i\star}^{\top}$, one value y_i is then sampled from the entries of y belonging to this regressor. In case of a step function to be fitted, one observation per step is drawn to get a subsample.

4. If no subsample can be found in 100 trials, `FastS` returns `NA`. `RobPer` then releases a warning, but can calculate further periodogram bars for other trial periods.

5. The internal functions `loss.S`, `re.s`, `f.w`, `scale1`, `our.solve` and `rho` are now defined outside `FastS`. Otherwise R would have to redefine them for each periodogram bar.

6. The subfunction `norm` is replaced by the function `norm(..., "2")` from the package **base** (R Core Team 2016).

7. The labels of the return values are changed for better interpretation.

τ-regression

In case of `regression = "tau"`, τ-regression is used to fit the periodic function. `RobPer` uses the Fast-τ algorithm of Salibian-Barrera *et al.* (2008) which works according to the same optimizing principle as `FastS` for S-regression (see previous paragraph), i.e., optimizing `N` candidates in `kk` iterations and further optimizing the `tt` best of these until convergence. Since computation of the objective value is expensive, it is possible to approximate it with `rr` iteration steps when choosing `approximate = TRUE`. For more details see Salibian-Barrera *et al.* (2008).

The R function `FastTau` used in **RobPer** is a slightly modified version of the R code published in Salibian-Barrera *et al.* (2008) with similar changes as in `FastS` compared to `fast.s` (see previous paragraph). The changes are:

1. A candidate for β_μ, see Equation 12, is allowed.

2. Arguments `N`, `kk`, `tt`, `rr` and `approximate` are combined to a list `taucontrol`, which is also an argument for `RobPer`.

3. Subsamples in general position are found as in `FastS` (change 3 in the previous paragraph).

4. If no subsample can be found, `FastTau` returns `NA` instead of a break using the `stop` function. This allows `RobPer` to release a warning, while calculating further periodogram bars for other trial periods.

5. A block of code used several times to check new regression parameter candidates for providing the best optimization value so far has been modularized into the subfunction `checkbest`.

6. Due to rounding errors, it may happen in the IRWLS algorithm that negative values close to zero occur, although they have to be non-negative by theory. This is avoided by setting such values to zero.

7. The subfunction `randomset` is replaced by the R function `sample` from the **base** package as both functions fulfill the same task and `sample` is faster.

8. The labels of the return values are changed for better interpretation.

3. Fit beta distributions with `betaCvMfit`

In this section we present the function `betaCvMfit`, which robustly fits a beta distribution to a sample using Cramér-von-Mises (CvM) distance minimization. The function is adapted from R code by Brenton R. Clarke for fitting a gamma distribution (see Clarke, McKinnon, and Riley 2012) using CvM distance minimization. Section 3.1 motivates the application of this function, while its usage is explained in more detail in Section 3.2.

3.1. Motivation

After a periodogram is calculated, one might be interested in the automatic detection of significant periods. A period shall be called significant, if the respective periodogram bar is atypical from the distribution of the applied criterion under the null hypothesis of no periodic fluctuation. To determine significance, this distribution needs to be known or estimated. Let Q_α be the α-quantile of this distribution. Assuming independent identically distributed periodogram bars $\mathrm{Per}(p_1), \ldots, \mathrm{Per}(p_q)$ we get

$$P\Big(\max\left(\mathrm{Per}(p_1), \ldots, \mathrm{Per}(p_q)\right) \geq Q_{\sqrt[q]{1-\alpha}}\Big) = \alpha. \tag{13}$$

A single periodogram bar calculated as described in Section 2 using unweighted least squares regression is $\mathcal{B}(\frac{m-1}{2}, \frac{n-m}{2})$-distributed, where $\mathcal{B}$ denotes the beta distribution and m is the dimension of the model. This result can be found in Schwarzenberg-Czerny (1998) or easily be deduced from Seber and Lee (2003, p. 110) and Gupta and Nadarajah (2004, p. 51). Already small violations of the assumptions made about the method or the light curve disturb this proceeding. In this work, we consider weighted and robust regression in addition to ordinary least squares. Besides, we have to take into account small deviations from our model assumptions like bad estimates s_i. An example is shown in Figure 2. Panel 2a shows the weighted least squares periodogram (using a sine model) of a light curve only consisting of white noise. The observed values were generated as

$$y_i = y_{w;i} + c \cdot y_{r;i}, \qquad i = 1, \ldots, n \tag{14}$$

with $y_{w;i}$ and $y_{r;i}$ being realizations from

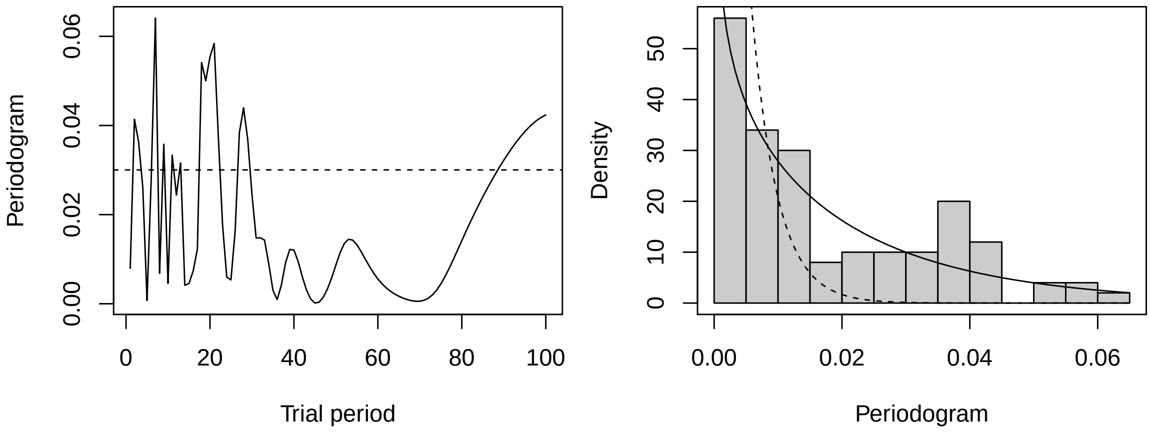
$$Y_{w;i} \sim \mathcal{N}(0, s_i^2), \tag{15}$$

$$Y_{r;i} \sim \mathcal{N}(0, 1). \tag{16}$$

The value of s_i is given for all i, and c is chosen to fulfill

$$\frac{\mathrm{var}(c \cdot y_r)}{\mathrm{var}(y_w) + \mathrm{var}(c \cdot y_r)} = 0.2, \tag{17}$$

where var() denotes the empirical variance. This means, there is roughly an extra 20 percent noise which is not explained by the measurement accuracies. Evidently, no periodogram bar



(a)

(b)

Figure 2: Example illustrating that a predefined $\mathcal{B}\left(\frac{m-1}{2}, \frac{n-m}{2}\right)$ distribution is sometimes not flexible enough if the model restrictions are slightly violated (see text for details). Panel 2a shows the periodogram of a light curve not completely following the assumed data model with the $\sqrt[q]{0.95}$ quantile of a $\mathcal{B}(\frac{m-1}{2}, \frac{n-m}{2})$ distribution (dashed line). Panel 2b shows a histogram of the periodogram bars, with the density of the $\mathcal{B}(\frac{m-1}{2}, \frac{n-m}{2})$ (dashed) and the CvM-fitted beta distribution with parameters $0.8 < 1 = \frac{m-1}{2}$ and $40.18 < 248.5 = \frac{n-m}{2}$ (solid).

is outstanding, but using the $\sqrt[q]{0.95}$ quantile of a $\mathcal{B}(\frac{m-1}{2}, \frac{n-m}{2})$ distribution (dashed line), several periods are found automatically.

To circumvent these problems, Thieler *et al.* (2013) propose to relax the assumption of a predefined $\mathcal{B}\left(\frac{m-1}{2}, \frac{n-m}{2}\right)$-distribution and only assume that the periodogram values can be approximated by any beta distribution. As peculiar periods are expected to show up as outliers, robustly fitting a $\mathcal{B}(\theta_1, \theta_2)$-distribution to $\mathrm{Per}(p_1), \ldots, \mathrm{Per}(p_q)$ is proposed. The authors use CvM distance minimization for this, which has been recommended by Clarke *et al.* (2012) for fitting gamma distributions in the presence of outliers. The CvM is defined as

$$\int_0^\infty \left(F_n(u) - F_\theta(u)\right)^2 dF_\theta(u) = \frac{1}{n}\sum_{i=1}^{n}\left(F_\theta(u_{(i)}) - \frac{i-0.5}{n}\right)^2 + \frac{1}{12n^2}, \tag{18}$$

where $u_{(1)}, \ldots, u_{(n)}$ is the ordered sample, F_n is the empirical distribution function and F_θ is the distribution function of $\mathcal{B}(\theta_1, \theta_2)$.

Panel 2b shows the predefined (solid) and the CvM-fitted (dashed) beta density for a periodogram calculated from the only-noise-data described above. While the $\sqrt[q]{0.95}$ quantile of the predefined distribution is about 0.03, the related quantile of the fitted distribution is 0.16 and no period is detected automatically.

The above approach falls within the framework of outlier detection described by Davies and Gather (1993) and is successfully used by Thieler *et al.* (2013) in the context discussed here. However, it assumes independent periodogram bars. This may cause problems when the periodogram peaks are broad (because the assumption of independency of the periodogram bars is violated): Then it can be hard for the automatism to find any outlying periodogram value, as there are many high values. One might try to ease this problem choosing a selection of trial periods with large distances or considering only the periods referring to local maxima in the periodogram as (roughly) independent trial periods (modifying and expanding an approach of Zechmeister and Kürster 2009) and fit the beta distribution to them using a CvM fit.

Simulations indicate that the beta distribution describes the distribution under the null hypothesis rather well for the different periodograms. Nevertheless, in the following we will call detected periods "valid" and not "significant" to stress that our approach to detect periods lacks a theoretical justification.

3.2. The R function `betaCvMfit`

The function `betaCvMfit` fits a $\mathcal{B}(\theta_1, \theta_2)$-distribution with mean $\theta_1/(\theta_1 + \theta_2)$ to a sample vector `data` using CvM distance minimization and has been applied in Thieler *et al.* (2013) for fitting beta distributions to periodograms to detect valid periods.

As it may happen that the periodogram bars become negative due to fitting problems, the function sets all negative entries of `data` to zero. If the logical argument `CvM` is set to `TRUE`, a CvM fit is calculated. As initial values for the optimization, the moment estimates of the beta distribution

$$\widehat{\theta}_1 = -\frac{\bar{x} \cdot (-\bar{x} + \bar{x}^2 + \hat{s}^2)}{\hat{s}^2}, \qquad \widehat{\theta}_2 = \frac{\widehat{\theta}_1 - \widehat{\theta}_1 \cdot \bar{x}}{\bar{x}} \tag{19}$$

are used. If the argument `rob` is set to `TRUE`, the median and the median absolute deviation from the median (MAD) are used instead of the arithmetic mean for $\bar{x}$ and the standard deviation for $\hat{s}$, respectively. In case of a very small estimate $\hat{s}$ (which happens particularly if $\hat{s}$ is the MAD), the function stops as it is not possible to calculate the estimates $\widehat{\theta}_1$ and $\widehat{\theta}_2$ shown above. The parameters of a beta distribution are strictly positive. Since it can happen that $\widehat{\theta}_1$ or $\widehat{\theta}_2$ are negative, the initial estimates are clipped to be at least 0.00001. If `CvM` is set to `FALSE`, the CvM distance is not optimized, and the initial estimates $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are returned.

Figure 3 shows the different fits varying the arguments `CvM` and `rob` for 50 $\mathcal{B}(4, 15)$-distributed observations containing 10 percent outliers between 0.8 and 1.

4. Generate light curves with `tsgen`

To investigate our periodogram methods in simulations, we implemented the R function `tsgen` to generate artificial light curves. A preliminary version of this function is used in Thieler

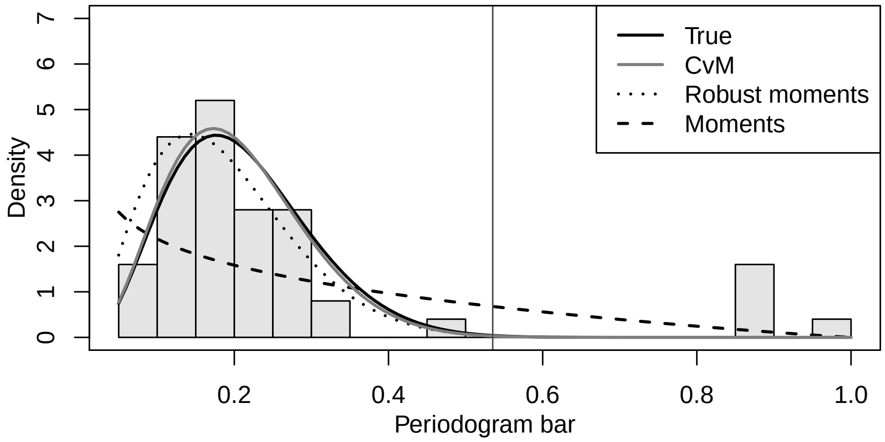


Figure 3: Gray-scale-version of the example for `betaCvMfit` given in the **RobPer** manual: Histogram of 45 $\mathcal{B}(4,15)$-distributed observations and 5 outliers uniformly distributed between 0.8 and 1. The black solid line shows the $\mathcal{B}(4,15)$-distribution, the other curves show different fits using `betaCvMfit` (in case of `CvM = TRUE`, the different settings for `rob` lead to the same result).

et al. (2013). The light curves $(t_i, y_i, s_i)_{i=1,\dots,n}$ are generated as realizations of the model

$$T_i = T^\star_{(i)}, \quad T^\star_1, \dots, T^\star_n \sim \mathcal{D}(p_s), \tag{20}$$

$$Y_i = \begin{cases} Y_{f;i} + Y_{w;i} + Y_{r;i}, & Y_i \text{ “behaves regularly”} \\ Y^\star_i, & Y_i \text{ is an outlier} \end{cases}, \tag{21}$$

$$Y_{f;i} = f\left(\frac{T_i}{p_f}\right), \quad f(\xi) = f(\xi+1)\ \forall \xi \in \mathbb{R} \tag{22}$$

$$Y_{w;i} \sim \mathcal{N}(0, \sigma_i^2), \tag{23}$$

$$s_i = \begin{cases} \text{given estimate for } \sigma_i \text{ independent from } Y_1, \dots, Y_n, & s_i \text{ “behaves regularly”} \\ s^\star_i, & s_i \text{ is an outlier} \end{cases}, \tag{24}$$

where $T^\star_{(i)}$ denotes the ith ordered observation time in $T^\star_1, \dots, T^\star_n$ and $\mathcal{D}(p_s)$ is a periodic sampling density with period p_s. The noise component Y_r is a power law noise (see Timmer and König 1995) with power exponent α and is white noise in case of $\alpha = 0$. Inserting another noise component and two types of outliers, this extended model allows to generate data violating the model introduced in Section 1.

The function calls several autonomous subfunctions one by one which perform individual simulation steps. These are:

1. Generate a sampling $t_1, \dots, t_n$ (using `sampler`, see Section 4.1).

2. Generate a periodic signal $y_{f;1}, \dots, y_{f;n}$ (using `signalgen`, see Section 4.2).

3. Add noise $y_{w;1}, \dots, y_{w;n}$ with related measurement accuracies $s_1, \dots, s_n$ and a noise component $y_{r;1}, \dots, y_{r;n}$ unrelated to the s_i (using `lc_noise`, see Section 4.3).

4. Disturb the light curve replacing measurement accuracies s_i by outliers, or replacing observations $y_i = y_{f;i}+y_{w;i}+y_{r_i}$ by aperiodic features (using `disturber`, see Section 4.4).

Table 4 lists all arguments for the subfunctions. The gray-shaded arguments are also arguments for `tsgen`, which passes them to the respective subfunction.

4.1. Generate sampling using `sampler`

The R function `sampler` is used to sample observation times $t_1, \ldots, t_n$ in the interval $[0, n_s \cdot p_s]$ with a possibly periodic sampling of period p_s. The sampling pattern depends on the argument `ttype` (see Table 4). If a periodic pattern is chosen, the observed time interval covers n_s cycles of it.

In case of `ttype = "equi"`, the observation times are equidistantly sampled with $t_i = i\frac{p_s \cdot n_s}{n}$. For `ttype = "unif"`, the observation times are drawn independently from a uniform distribution on $[0, n_s \cdot p_s]$. Both these sampling schemes are aperiodic, the sampling period p_s only influences the duration $t_n - t_1$ of the sampling.

For `ttype = "sine"` and `ttype = "trian"`, the observation times are sampled from a periodic density with sampling period p_s. First, observation cycles $z_i^\star$ are drawn from a discrete uniform distribution on $\{1, \ldots, n_s\}$ to determine the cycle the ith observation is part of. Second, observation phases $\varphi_i^\star$ are sampled with density

$$d_{sine}(x) = \sin(2\pi x) + 1 \qquad \text{(for } \texttt{ttype = "sine"}\text{)} \tag{25}$$

$$\text{or } d_{trian}(x) = \begin{cases} 3x, & 0 \le x \le \frac{2}{3}, \\ 6-6x, & \frac{2}{3} < x \le 1 \end{cases} \qquad \text{(for } \texttt{ttype = "trian"}\text{).} \tag{26}$$

To sample from d_{sine}, the function `BBsolve` from package **BB** (Varadhan and Gilbert 2009), is used. The unsorted observation times $t_i^\star$ are then generated using

$$t_i^\star = \varphi_i^\star + (z_i^\star - 1)p_s. \tag{27}$$

The sine-shaped density is motivated by sampling patterns observed in real data, see Panel 1b. The triangular shaped density offers an alternative periodic sampling design. Separately sampling observation cycle and phase was proposed by Hall and Yin (2003).

As the result, `sampler` returns the ordered observation times $t_1, \ldots, t_n$.

4.2. Generate periodic signal using `signalgen`

To generate the periodic component in the observed values, the R function `signalgen` is used. The values $y_{f;1}, \ldots, y_{f;n}$ with fluctuation period p_f at observation times $t_1, \ldots, t_n$ are generated using

$$y_{f;i} = f\left(\frac{t_i}{p_f}\right), \quad i = 1, \ldots, n. \tag{28}$$

The observation times, the fluctuation period and the shape of f are arguments of `signalgen` (see Table 4). In case of `ytype = "const"`, f is defined as

$$f(t) = 0, \tag{29}$$

Argument	Subfunction	Comment
`ps` $\in \mathbb{R}_{>0}$	`sampler`, `disturber`	Sampling period p_s. Default 1.
`ncycles` $\in \mathbb{N}$	`sampler`	Number n_s of sampling cycles.
`npoints` $\in \mathbb{N}$	`sampler`	Sample size n.
`ttype`	`sampler`	Distribution $\mathcal{D}(p_s)$ of the unsorted observation times. Options are: `"equi"` (equidistant sampling), `"unif"` (uniform sampling), `"sine"` (sine-shaped density, see Section 4.1) and `"trian"` (triangular density, see Section 4.1).
`tt` $\in \mathbb{R}^n$	`signalgen`, `lc_noise`, `disturber`	Observation times $t_1, \ldots, t_n$, e.g., return value from `sampler`.
`pf` $\in \mathbb{R}_{>0}$	`signalgen`	Fluctuation period p_f. Default 1.
`ytype`	`signalgen`	Type of periodic fluctuation f. Options: `"const"` (constant), `"sine"` (sine), `"trian"` (triangular function) and `"peak"` (peak function).
`sig` $\in \mathbb{R}^n$	`lc_noise`	Values $y_{f;1}, \ldots, y_{f;n}$ of the periodic fluctuation, e.g., return value from `signalgen`.
`SNR` $\in \mathbb{R}_{>0}$	`lc_noise`	Relation $\text{var}(y_f)/\,\text{var}(y_w + y_r)$.
`redpart` $\in [0,1]$	`lc_noise`	Fraction $\text{var}(y_r)/(\text{var}(y_w) + \text{var}(y_r))$ of noise not related to measurement accuracies.
`alpha`	`lc_noise`	Power law coefficient of the noise component y_r. Set to zero for $y_{r;1}, \ldots, y_{r;n} \underset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$.
`y` $\in \mathbb{R}^n$	`disturber`	Observed values $y_1, \ldots, y_n$, e.g., return value from `lc_noise`.
`s` $\in \mathbb{R}^n_{>0}$	`disturber`	Measurement accuracies $s_1, \ldots, s_n$, e.g., return value from `lc_noise`.
`s.outlier.fraction` $\in [0,1]$	`disturber`	Fraction of measurement accuracies to be replaced by outliers.
`interval` $\in \{$`TRUE`, `FALSE`$\}$	`disturber`	If `TRUE`, the y_i belonging to a random time interval are disturbed.

Table 4: Arguments for the subfunctions of `tsgen`. See the respective section for more details. Gray-shaded values are also arguments for `tsgen`, which passes the values to the respective subfunction. “var” denotes the empirical variance.

so there is no (periodic) fluctuation. This setting can be used to investigate the false alarm probability of a period detection method. In case of `ytype = "sine"`, f is defined as

$$f(t) = \sin\left(\frac{2\pi t}{p_f}\right). \tag{30}$$

This is a typical assumption in the literature. For `ytype = "trian"`,

$$f(t) = \begin{cases} 3\varphi_1(t), & 0 \leq \varphi_1(t) \leq \frac{2}{3} \\ 6 - 6\varphi_1(t), & \frac{2}{3} < \varphi_1(t) \leq 1 \end{cases} \tag{31}$$

with $\varphi_1(t) = t \mod 1 = (t - \lfloor t \rfloor)$ is used. This triangular shaped function was originally implemented in order to be able to choose between different periodic shapes. The light curve observed for CoRoT ID 0105288363 (Chadid *et al.* 2011) shows that functions with a similar shape are quite realistic. When choosing `ytype = "peak"`, y_f is generated using

$$f(t) = \begin{cases} 9\exp\left(-3p_f^2\left(\varphi_1(t) - \frac{2}{3}\right)^2\right), & 0 \leq \varphi_1(t) \leq \frac{2}{3} \\ 9\exp\left(-12p_f^2\left(\varphi_1(t) - \frac{2}{3}\right)^2\right), & \frac{2}{3} < \varphi_1(t) \leq 1 \end{cases}. \tag{32}$$

This function mostly shows values close to zero and large values for only one time unit per cycle. This "peak" occurring in each cycle has an asymmetric shape.

As the result, `signalgen` returns the periodic component $y_{f;1}, \ldots, y_{f;n}$ of the observed values.

4.3. Add noise and measurement accuracies using `lc_noise`

The R function `lc_noise` is used to generate measurement accuracies $s_1, \ldots, s_n$ and add noise to a periodic fluctuation (see Table 4). The measurement accuracies are sampled from a gamma$(3, 10)$ distribution. This choice is motivated by real data from Tluczykont *et al.* (2010). As shown in Equation 4, the noise component $y_w = (y_{w;1}, \ldots, y_{w;n})^\top$ is a realization of Y_w with $Y_{w;i} \sim \mathcal{N}(0, s_i^2)$.

A second noise component y_r does not depend on the s_i. It is generated as red noise, i.e., following a power law with power law index α. For $\alpha = 0$ we get white noise. Flicker noise (pink noise) is generated using $\alpha = 1$ and brown noise using $\alpha = 2$. The power law noise is generated using subfunctions `TK95_uneq` and `TK95`. The latter generates an equidistant time series of power law noise according to Timmer and König (1995). For irregular observation times, a noise series resulting from `TK95` is used and an unequally sampled noise series is generated following Uttley, McHardy, and Papadakis (2002).

The noise components are scaled so that the variance of the $y_{r;i}$ has approximately the proportion `redpart` in the overall noise variance and that `SNR` is the ratio $\mathrm{var}(y_f)/\,\mathrm{var}(y_w + y_r)$, where $\mathrm{var}(x)$ is the empirical variance of vector x. Note that the white noise components' variances are exactly s_i^2, so that the s_i are not estimates but true values. In this sense, the measurement accuracies of a generated light curve are more informative for our artificial data than for real light curves, where the measurement accuracies are estimates. Allowing for a second noise component makes it possible to lower the information of the measurement accuracies with respect to the overall noise in the observed values.

The function `lc_noise` returns the observed values $y_i = y_{f;i} + y_{w;i} + y_{r;i}$, $i = 1, \ldots, n$.

4.4. Disturb light curve using `disturber`

The last subfunction applied in `tsgen` is `disturber`, which can be used to disturb a given light curve (see Table 4). It replaces a given fraction of measurement accuracies by the smaller value

$s_i^\star = \frac{1}{2}\min(s_1, \ldots, s_n)$, i in a subset of $\{1, \ldots, n\}$. As small measurement accuracies stand for precise observations, the influence of observations with disturbed measurement accuracies $s_i^\star$ rises in case of a weighted fit. For unweighted regression, this type of disturbance does not affect the result of the fit.

Optionally, `disturber` also replaces observed values y_i by atypical values. For this, a time interval $[t_{\text{start}}, t_{\text{start}} + 3p_s]$ within the interval $[t_1, t_n]$ is randomly chosen and all observed values belonging to this time interval are replaced by a peak function:

$$y_i^\star = 6\ \tilde{y}_{0.9}\ \phi\left(\frac{t_i - t_{\text{start}} - 1.5p_s}{p_s}\right)/\phi(0) \quad \forall\ i\ :\ t_i \in [t_{\text{start}}, t_{\text{start}} + 3p_s], \tag{33}$$

where ϕ denotes the density of the standard normal distribution. If the y_i are intended to be disturbed and the light curve is shorter than $3p_s$, the function will stop with an error message.

The function returns the modified vectors $y = (y_1, \ldots, y_n)^\mathsf{T}$ and $s = (s_1, \ldots, s_n)^\mathsf{T}$. If the option to change y values is not used (see Table 4) and the fraction of outlying measurement accuracies is set to zero, y and s are returned unchanged.

5. Application

In this section, we give examples how to use the **RobPer** package for light curve analysis. We start with an artificial example, also given in the manual, and then analyze some real data.

5.1. Artificial example

To generate an artificial light curve, `tsgen` can be used:

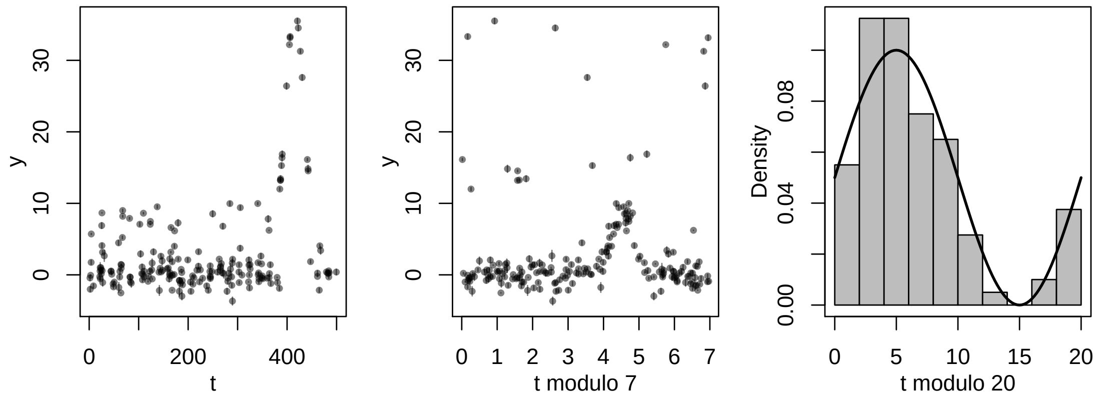
```
R> library("RobPer")
R> set.seed(22)
R> lightcurve <- tsgen(ttype = "sine", ytype = "peak", pf = 7,
+       redpart = 0.1, s.outlier.fraction = 0.1, interval = TRUE,
+       npoints = 200, ncycles = 25, ps = 20, SNR = 3, alpha = 0)
```

This light curve has a sine-shaped sampling (`ttype`) with sampling period 20 (`ps`) and covers a time interval of about 25 sampling cycles (`ncycles`), so 500 time units. It consists of 200 observations (`npoints`) and the observed values contain a peak-shaped periodic fluctuation (`ytype`) with fluctuation period 7 (`pf`). The measurement accuracies are related to about 90 percent of the noise component (`1-redpart`), the rest of the noise is white as well (`alpha`). The empirical variance of the periodic fluctuating component in the observed values is three times larger than the empirical variance in the noise component (`SNR`). The light curve contains 10 per cent outliers in the measurement accuracies (`s.outlier.fraction`) and atypical observed values (`interval`).

Alternatively, the functions `sampler`, `signalgen`, `lc_noise` and `disturber` can be used to generate the same light curve, see Section 4.

Sampling observation times:

```
R> set.seed(22)
R> tt <- sampler(ttype = "sine", npoints = 200, ncycles = 25, ps = 20)
```



(a) (b) (c)

Figure 4: Artificial light curve in Panel 4a with vertical bars marking the s_i. Plotting time axis modulo 7 in Panel 4b reveals the periodic fluctuation of $p_f = 7$. Histogram and sampling density of the observation times modulo 20 in Panel 4c shows the sampling periodicity of $p_s = 20$.

Generate periodic fluctuation:

```
R> yf <- signalgen(tt, ytype = "peak", pf = 7)
```

Add noise and scale signal to the right `SNR`:

```
R> temp <- lc_noise(tt, sig = yf, SNR = 3, redpart = 0.1, alpha = 0)
R> y <- temp$y
R> s <- temp$s
```

Replace measurement accuracies by tiny outliers and include a peak:

```
R> temp <- disturber(tt, y, s, ps = 20, s.outlier.fraction = 0.1,
+      interval = TRUE)
```

The result is the same:

```
R> all(cbind(tt, temp$y, temp$s) == lightcurve)
```

Figure 4 shows plots of the generated light curve.

In the next step, we calculate a periodogram of the light curve. The periodogram is calculated fitting a step model using unweighted M-regression with the Huber function. The light curve spans a time interval of approximately `ncycles` $\cdot$ `ps` $= 500$ time units, so it is sensible to investigate periods up to 50 (one tenth, see Halpern, Leighly, and Marshall 2003).

```
R> PP <- RobPer(lightcurve, model = "splines", regression = "huber",
+      weighting = FALSE, var1 = FALSE, periods = 1:50)
```

Outstanding periodogram bars are sought fitting a beta distribution to the periodogram values using Cramér-von-Mises distance minimization (CvM) and determining the $\sqrt[q]{0.95}$-quantile with $q = 50$ as the number of periodogram bars.

```
R> betavalues <- betaCvMfit(PP)
R> crit.val <- qbeta((0.95)^(1 / 50), shape1 = betavalues[1],
+      shape2 = betavalues[2])
```

Panel 5a depicts the histogram of the periodogram bars, the beta distribution fitted (solid line) and its $\sqrt[50]{0.95}$-quantile (solid vertical line). Further fits of a beta distribution (method of moments, dashed, and robust method of moments, dotted) and their respective $\sqrt[50]{0.95}$-quantiles are shown as well.

```
R> hist(PP, breaks = 20, freq = FALSE, xlim = c(0, 0.08), col = "grey",
+      main = "", xlab="Periodogram")
R> betafun <- function(x) dbeta(x, shape1 = betavalues[1],
+      shape2 = betavalues[2])
R> curve(betafun, add = TRUE, lwd = 2)
R> abline(v = crit.val, lwd = 2)
```

Application of method of moments:

```
R> par.mom <- betaCvMfit(PP, rob = FALSE, CvM = FALSE)
R> myf.mom <- function(x) dbeta(x, shape1 = par.mom[1], shape2 = par.mom[2])
R> curve(myf.mom, add = TRUE, lwd = 2, lty = 2)
R> crit.mom <- qbeta((0.95)^(1 / 50), shape1 = par.mom[1],
+      shape2 = par.mom[2])
R> abline(v = crit.mom, lwd = 2, lty = 2)
```

Application of robust method of moments:

```
R> par.rob <- betaCvMfit(PP, rob = TRUE, CvM = FALSE)
R> myf.rob <- function(x) dbeta(x, shape1 = par.rob[1], shape2 = par.rob[2])
R> curve(myf.rob, add = TRUE, lwd = 2, lty = 3)
R> crit.rob <- qbeta((0.95)^(1 / 50), shape1 = par.rob[1],
+      shape2 = par.rob[2])
R> abline(v = crit.rob, lwd = 2, lty = 3)
R> legend("topright", lty = 1:3, legend = c("CvM", "Moments",
+      "Robust moments"), bg = "white", lwd = 2)
R> box()
```

Using the $\sqrt[50]{0.95}$ quantile of the CvM fit (solid line), a period of 7 time units seems to be valid, see Panel 5b. Twice this period, which is 14, might be valid, too. So the real periodic fluctuation of $p_f = 7$ is well recognized within the disturbed signal, as intended. Of course, a periodic function with period p is also periodic with period $k \cdot p$, $k \in \mathbb{N}$.

```
R> plot(1:50, PP, xlab = "Trial period", ylab = "Periodogram", main = "",
+      type = "l")
R> abline(h = crit.val, lwd = 2)
R> text(7, PP[7]-0.002,7, pos=4)
R> text(14, PP[14]+0.002,14, pos=4)
```

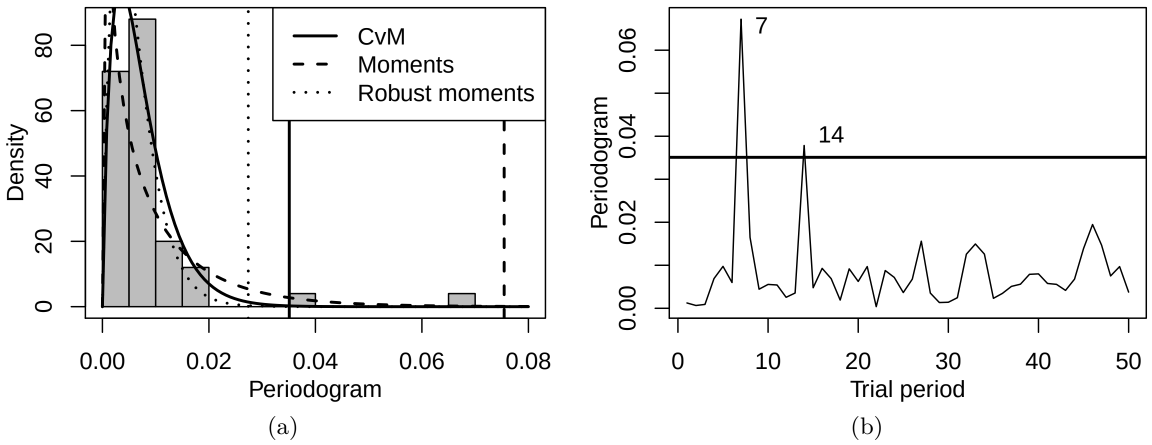


Figure 5: Periodogram bars calculated fitting a spline model using unweighted M-regression with the Huber function to the artificial example from Figure 4: Robustly fitting a beta distribution to the periodogram bars in Panel 5a leads to two outstanding trial periods in Panel 5b.

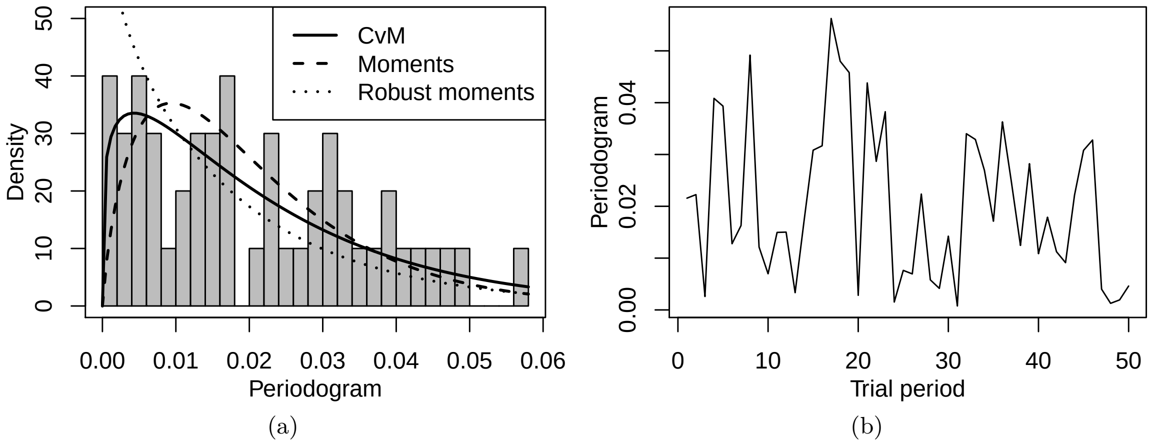


Figure 6: Analysis of the artificial example as in Figure 5, now using least squares regression.

While the robust M-regression recognizes the real periodic fluctuation, fitting the same model by least squares regression does not, as shown in Figure 6. Only the periodogram is calculated in another way.

```
R> PP <- RobPer(lightcurve, model = "splines", regression = "L2",
+      weighting = FALSE, var1 = FALSE, periods = 1:50)
```

The analysis proceeds as before.

5.2. Disturbed data from GROJ0422+32

The first real data set we analyze is a light curve for gamma ray emission of the source

GROJ0422+32, obtained by the BATSE Earth Occultation Monitoring project of the NASA. These experiments are described in Harmon, Fishman, Wilson, Paciesas, Zhang, Finger, Koshut, McCollough, Robinson, and Rubin (2002) and Harmon, Wilson, Fishman, Connaughton, Henze, Paciesas, Finger, McCollough, Sahi, Peterson, Shrader, Grindlay, and Barret (2004). The data have been kindly provided by the NASA, are available from `http://gammaray.nsstc.nasa.gov/batse/occultation`, and are shown in Panel 7a.

A large peak is visible starting at about 48900 Markarian Julian days (which corresponds to December 10 1991 in the Gregorian calendar), a so called gamma ray burst. It occasionally occurs in gamma ray observations and can be considered as outlier. The light curve covers a time interval of about 3312 days, so following Halpern *et al.* (2003) we consider periods up to 330 days (about one tenth of the overall duration of the light curve). Figure 7b shows the periodogram obtained fitting a sine function using least squares regression, which is the classical approach in astroparticle physics. It is calculated using

```
R> data(star_groj0422.32)
R> PP <- RobPer(star_groj0422.32, periods = 1:330, model = "sine",
+      regression = "L2", weighting = FALSE)
```

Periodograms for τ-regression and M-regression using the Huber function are obtained replacing `"L2"` by `"tau"` or `"huber"` in the code above. The respective periodograms are shown in Panels 7c and 7d. All three periodograms do not show any outstanding peak. Apart from this, the periodograms using robust regression have a completely different shape than the least squares periodogram, which seems to have problems with the gamma ray burst. It might be questionable if the least squares periodogram can find a periodic structure in the observations in the presence of the gamma ray burst. We add a sine with period 30 and amplitude 0.005 to the observed values and repeat the analysis. The results can be seen in Figure 8. In Panel 8a it is visible that we did not introduce a strong periodic behavior. Nevertheless, the robust periodograms, Panels 8c and 8d, easily detect it, while there is only a small local peak in the least squares periodogram in Panel 8b. The horizontal lines in Panels 8c and 8d show the respective $\sqrt[330]{0.95}$-quantiles of the CvM-fitted beta distribution and are calculated from a periodogram `PP` using

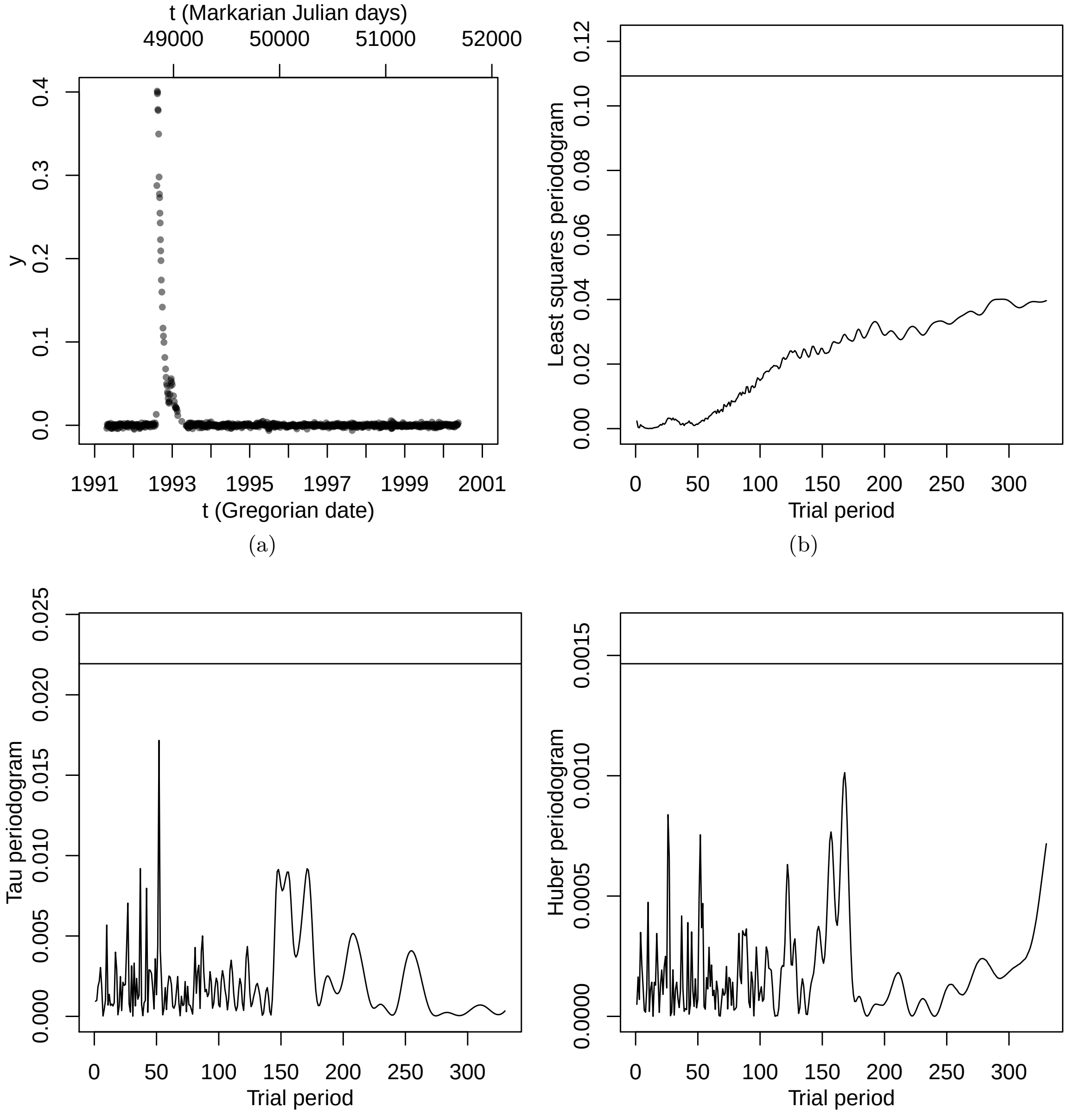
```
R> shapes <- betaCvMfit(PP)
R> Crit <- qbeta(0.95^(1 / 330), shape1 = shapes[1], shape2 = shapes[2])
```

So, as opposed to least squares regression, robust techniques are able to detect an (added) periodic fluctuation although the data are disturbed seriously by the gamma ray burst.

5.3. Data from Markarian 421 and 501

A further real data example are gamma ray light curves from Markarian 421 (Mrk 421) and Markarian 501 (Mrk 501), kindly provided by the Gamma Astronomy group of the Deutsches Elektronen-Synchrotron. The data have been collected from various original sources, combined, and published by Tluczykont *et al.* (2010), and are available from `http://astro.desy.de/gamma_astronomy/magic/projects/light_curve_archive/index_eng.html`. See the **RobPer** manual for details about the original sources and references.

The light curve obtained for Mrk 421 is shown in Panel 1a on page 2. Periodograms obtained fitting a sine are shown in Figure 9. Using the least squares periodogram in Panel 9a, no valid



(a) (b) (c) (d)

Figure 7: Analysis of GROJ0422+32: Panel 7a shows the light curve, while the other panels show the periodograms fitting a sine using least squares in Panel 7b, τ- in Panel 7c, Huber M-regression in Panel 7d. No periodogram bar exceeds the respective $\sqrt[330]{0.95}$-quantile of the CvM-fitted beta distribution (horizontal line).

period is detected, but considering the shape of the periodogram, one might wonder if there is a periodicity of 31 hidden in the same way as when adding a small periodic fluctuation to the GROJ0422+32 data, see Panel 8b. However, the periodograms for τ-regression in Panel 9b and Huber M-regression in Panel 9c show a different behavior from Figure 8, so this does not seem to be the case. Especially, the least squares and the Huber M periodogram show a

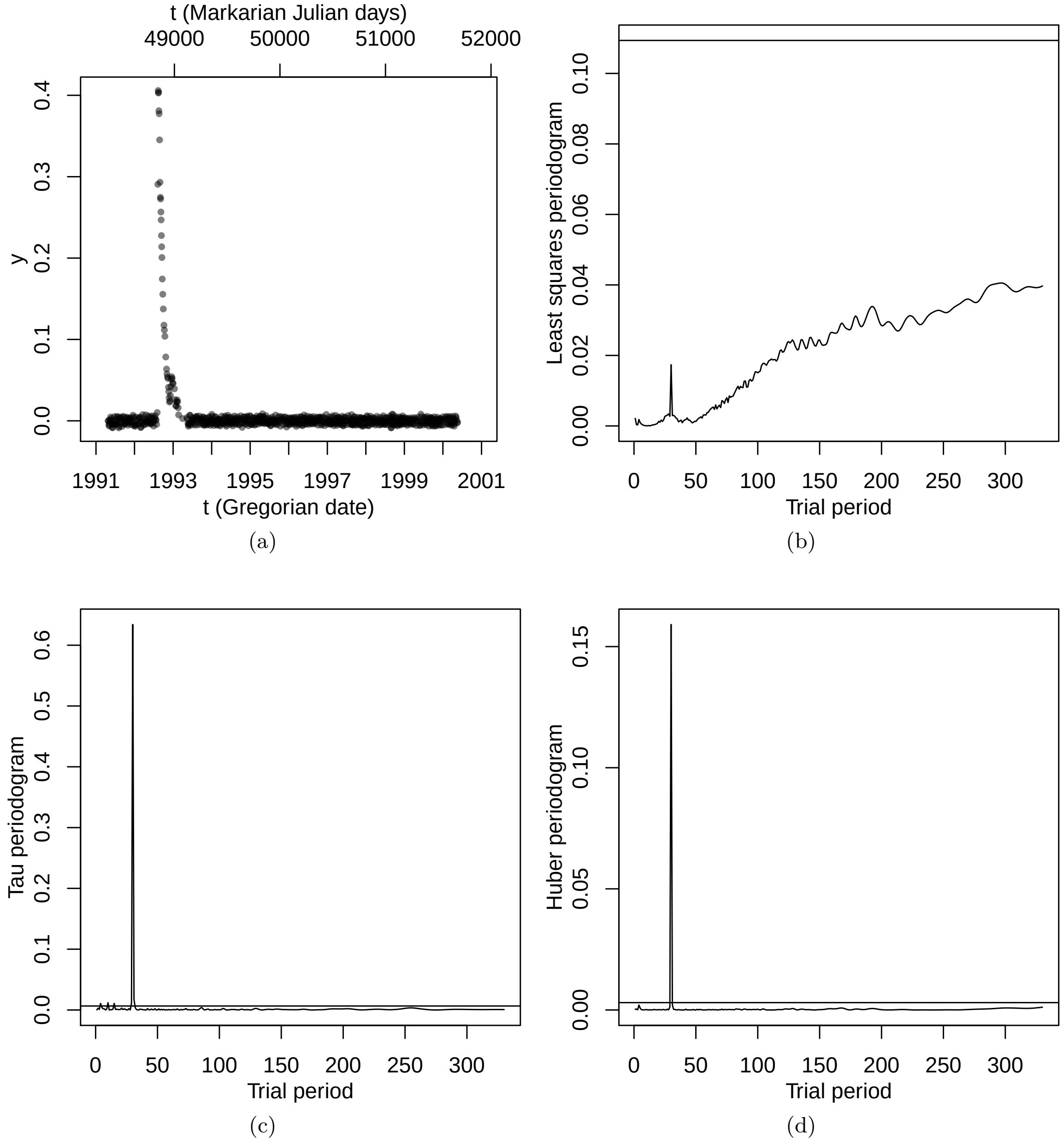
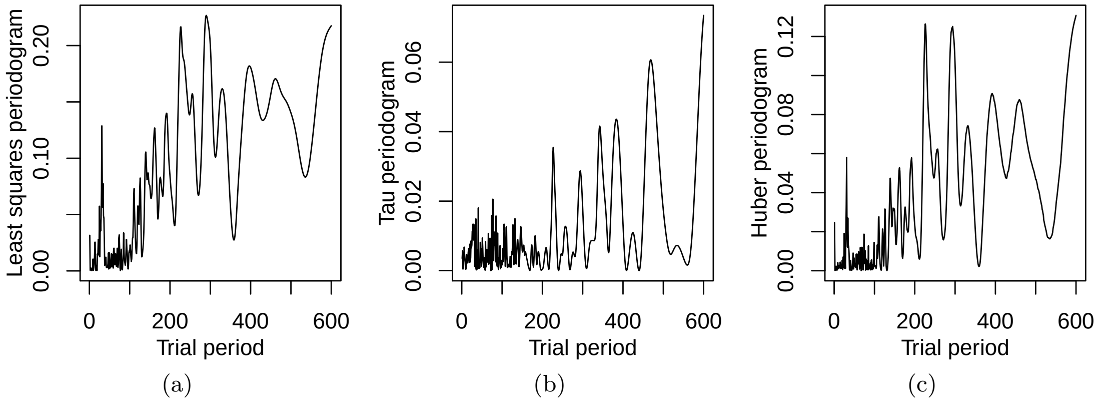


Figure 8: Adding a sine with amplitude 0.005 to the light curve of GROJ0422+32. Panel 8a shows the modified light curve, while the other panels show the periodograms fitting a sine using least squares in Panel 8b, τ- in Panel 8c, Huber M-regression in Panel 8d. The horizontal lines in those three panels show the respective $\sqrt[330]{0.95}$-quantile of the CvM-fitted beta distribution.

quite similar behavior regarding the local maxima. This could mean that there are not many observations weighted down in Huber M-regression.

Another light curve, obtained for Mrk 501, and periodograms using least squares regression, τ-regression and Huber M-regression are shown in Figure 10. Here we apply step regression,



(a) (b) (c)

Figure 9: Periodograms for Mrk 421, see Panel 1a, obtained fitting a sine with least squares regression in Panel 9a, τ-regression in Panel 9b, Huber M-regression in Panel 9c.

which is equivalent to epoch folding or phase dispersion minimization when using least squares regression (see Section 2). The periodogram is calculated applying

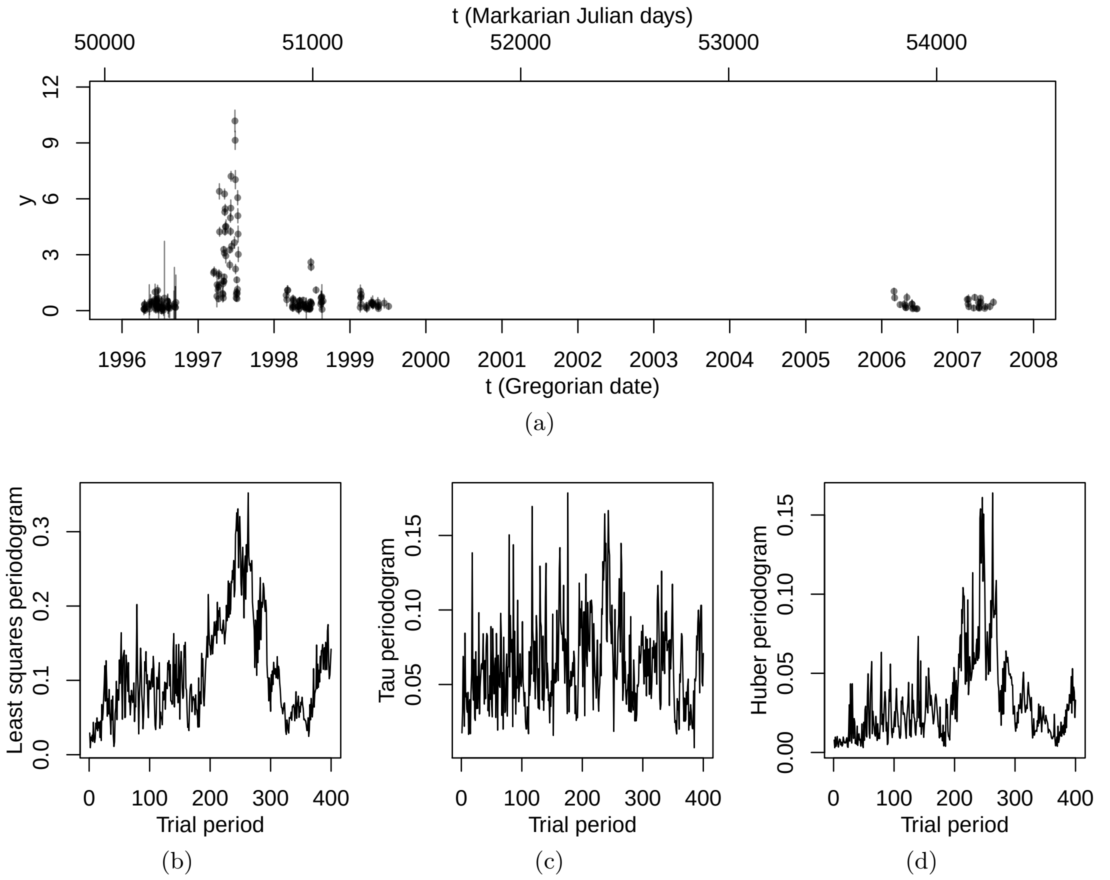
```
R> data(Mrk501)
R> RobPer(Mrk501, periods = 1:400, model = "step", regression = "L2",
+      weighting = FALSE)
```

in case of least squares regression and with `regression = "tau"` or `regression = "huber"` in case of τ- or Huber M-regression, respectively. For least squares regression in Panel 10b and Huber M-regression in Panel 10d we see a broad peak between the trial periods 200 and 300, much too broad to be considered as valid period (see Halpern *et al.* 2003). For τ-regression in Panel 10c, this behavior is not observed.

In the examples from the previous section, robust techniques recognize some periodicity in a light curve, while the least squares periodogram only provides a slightly atypical behavior for the trial period in question. Here it is the other way round: the least squares periodogram does not indicate a valid period, but exhibits some interesting feature similar to the previous data set, where a periodicity was hidden in noisy data. This initial suspicion cannot be confirmed by using robust regression instead of least squares regression. In summary, using our methods, we do not find a periodicity in the light curves for Mrk 421 and Mrk 501, neither using least squares nor robust regression.

6. Conclusions

The R package **RobPer** presented in this work allows searching for periodicity in irregularly sampled time series, possibly taking into account additional information on the precision of the measurement, if available. These are the typical characteristics of light curves, that is time series occurring in astroparticle physics. The periodogram is calculated fitting periodic functions to the light curve. The user can choose between six different periodic functions and seven different regression techniques, meaning that 42 possible combinations are offered, not taking into account further options like choosing the number of steps for the step model



(a)

(b)

(c)

(d)

Figure 10: Light curve in Panel 10a and periodograms for Mrk 501 obtained fitting a periodic step function with least squares regression in Panel 10b, τ-regression in Panel 10c, Huber M-regression in Panel 10d.

or using weighted regression. The function `betaCvMfit` allows to search for prominent periodogram bars as outliers in a beta distribution robustly fitted to the periodogram. The function `tsgen` allows generation of artificial light curves for investigative use.

Acknowledgments

The authors are grateful to the reviewers for their constructive comments which led to several improvements of the paper and the programs. The financial support of the Deutsche Forschungsgemeinschaft (DFG) (within the Collaborative Research Center SFB 876 “Providing Information by Resource-Constrained Data Analysis”, project C3, and Graduate School GK 1032 “Statistische Modellbildung”) is gratefully acknowledged. We thank the ITMC at TU Dortmund University for providing computer resources on LiDO. We thank the NASA and the Deutsches Elektronen-Synchrotron for providing data.

References

Ahdesmäki M, Fokianos K, Strimmer K (2012). ***GeneCycle**: Identification of Periodically Expressed Genes*. R package version 1.1.2, URL https://CRAN.R-project.org/package=GeneCycle.

Ahdesmäki M, Lähdesmäki H, Gracey A, Shmulevich I, Yli-Harja O (2007). "Robust Regression for Periodicity Detection in Non-Uniformly Sampled Time-Course Gene Expression Data." *BMC Bioinformatics*, **8**(1), 233–248. doi:10.1186/1471-2105-8-233.

Akerlof C, Alcock C, Allsman R, Axelrod T, Bennett D, Cook K, Freeman K, Griest K, Marshall S, Park H (1994). "Application of Cubic Splines to the Spectral Analysis of Unequally Spaced Data." *The Astrophysical Journal*, **436**, 787–794. doi:10.1086/174954.

Bates DM, Venables WN (2016). ***splines**: Regression Spline Functions and Classes*. Base R package version 3.2.4, URL https://www.R-project.org/.

Bjornstad ON (2013). ***nlts**: (Non)Linear Time Series Analysis*. R package version 0.2-0, URL https://CRAN.R-project.org/package=nlts.

Chadid M, Perini C, Bono G, Auvergne M, Baglin A, Weiss W, Deboscher J (2011). "CoRoT Light Curves of Blazhko RR Lyrae Stars." *Astronomy & Astrophysics*, **527**, A146. doi:10.1051/0004-6361/201016048.

Clarke B, McKinnon P, Riley G (2012). "A Fast Robust Method for Fitting Gamma Distributions." *Statistical Papers*, **53**(4), 1001–1014. doi:10.1007/s00362-011-0404-3.

Croux C, Dehon C (2003). "Estimators of the Multiple Correlation Coefficient: Local Robustness and Confidence Intervals." *Statistical Papers*, **44**(3), 315–334. doi:10.1007/s00362-003-0158-7.

Cumming A, Marcy G, Butler R (1999). "The Lick Planet Search: Detectability and Mass Thresholds." *The Astrophysical Journal*, **526**(2), 890–915. doi:10.1086/308020.

Davies L, Gather U (1993). "The Identification of Multiple Outliers." *Journal of the American Statistical Association*, **88**(423), 782–792. doi:10.1080/01621459.1993.10476339.

Deeming T (1975). "Fourier Analysis with Unequally-Spaced Data." *Astrophysics and Space Science*, **36**(1), 137–158. doi:10.1007/bf00681947.

Ferraz-Mello S (1981). "Estimation of Periods from Unequally Spaced Observations." *The Astronomical Journal*, **86**(4), 619–624. doi:10.1086/112924.

Gupta A, Nadarajah S (2004). *Handbook of Beta Distribution and Its Applications*. Dekker, New York, Basel.

Hall P, Li M (2006). "Using the Periodogram to Estimate Period in Nonparametric Regression." *Biometrika*, **93**(2), 411–424. doi:10.1093/biomet/93.2.411.

Hall P, Reimann J, Rice J (2000). "Nonparametric Estimation of a Periodic Function." *Biometrika*, **87**(3), 545–557. doi:10.1093/biomet/87.3.545.

Hall P, Yin J (2003). "Nonparametric Methods for Deconvolving Multiperiodic Functions." *Journal of the Royal Statistical Society B*, **65**(4), 869–886. doi:10.1046/j.1369-7412.2003.00420.x.

Halpern J, Leighly K, Marshall H (2003). "An Extreme Ultraviolet Explorer Atlas of Seyfert Galaxy Light Curves: Search for Periodicity." *The Astrophysical Journal*, **585**, 665–676. doi:10.1086/346106.

Harmon B, Fishman G, Wilson C, Paciesas W, Zhang S, Finger M, Koshut T, McCollough M, Robinson C, Rubin B (2002). "The Burst and Transient Source Experiment Earth Occultation Technique." *The Astrophysical Journal Supplement Series*, **138**(1), 149–183. doi:10.1086/324018.

Harmon B, Wilson C, Fishman G, Connaughton V, Henze W, Paciesas W, Finger M, McCollough M, Sahi M, Peterson B, Shrader C, Grindlay J, Barret D (2004). "The Burst and Transient Source Experiment (BATSE) Earth Occultation Catalog of Low-Energy Gamma-Ray Sources." *The Astrophysical Journal Supplement Series*, **154**(2), 585–622. doi:10.1086/421940.

Huber P, Ronchetti E (1981). *Robust Statistics*, volume 1. John Wiley & Sons.

Johnson IM (2015). ***robustreg**: Robust Regression Functions.* R package version 0.1-9, URL https://CRAN.R-project.org/package=robustreg.

Koenker R (2015). ***quantreg**: Quantile Regression.* R package version 5.21, URL https://CRAN.R-project.org/package=quantreg.

Leahy D, Darbro W, Elsner R, Weisskopf M, Kahn S, Sutherland P, Grindlay J (1983). "On Searches for Pulsed Emission with Application to Four Globular Cluster X-Ray Sources: NGC 1851, 6441, 6624, and 6712." *The Astrophysical Journal*, **266**(1), 160–170. doi:10.1086/160766.

Li T (2009). "A Robust Spectral Analyzer for One-Dimensional and Multi-Dimensional Data Analysis." 2009/0112954 A1. US Patent Application.

Li T (2010). "A Nonlinear Method for Robust Spectral Analysis." *IEEE Transactions on Signal Processing*, **58**(5), 2466–2474. doi:10.1109/tsp.2010.2042479.

Maronna R, Martin R, Yohai V (2006). *Robust Statistics: Theory and Methods.* John Wiley & Sons, Chichester. doi:10.1002/0470010940.

Mebane, Jr WR, Sekhon JS (2011). "Genetic Optimization Using Derivatives: The **rgenoud** Package for R." *Journal of Statistical Software*, **42**(11), 1–26. doi:10.18637/jss.v042.i11.

Norm DIN 66261 (1985). "Sinnbilder für Struktogramme nach Nassi-Shneiderman."

Oh H, Nychka D, Brown T, Charbonneau P (2004). "Period Analysis of Variable Stars by Robust Smoothing." *Journal of the Royal Statistical Society C*, **53**(1), 15–30. doi:10.1111/j.1467-9876.2004.00423.x.

Palmer D (2009). "A Fast Chi-Squared Technique for Period Search of Irregularly Sampled Data." *The Astrophysical Journal*, **695**, 496–502. doi:10.1088/0004-637x/695/1/496.

R Core Team (2014). ***stats**: R Statistical Functions.* R package version 3.0.3, part of R 3.0.3.

R Core Team (2016). ***base**: Base R Functions.* Base R package version 3.2.4, URL https://www.R-project.org/.

R Core Team (2015). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Reegen P (2007). "SigSpec – I. Frequency- And Phase-Resolved Significance in Fourier Space." *Astronomy & Astrophysics*, **467**(3), 1353–1371. doi:10.1051/0004-6361:20066597.

Ripley R, Boitmanis K, Snijders TAB (2013). ***RSiena**: Siena – Simulation Investigation for Empirical Network Analysis.* R package version 1.1-232, URL https://CRAN.R-project.org/package=RSiena.

Rousseeuw P, Croux C, Todorov V, Ruckstuhl A, Salibian-Barrera M, Verbeke T, Koller M, Maechler M (2015). ***robustbase**: Basic Robust Statistics.* R package version 0.92-5, URL https://CRAN.R-project.org/package=robustbase.

Rousseeuw P, Yohai V (1984). "Robust Regression by Means of S-Estimators." In J Franke, W Härdle, D Martin (eds.), *Robust and Nonlinear Time Series Analysis*, Lecture Notes in Statistics No. 26, pp. 256–272. Springer-Verlag, Berlin, New York.

Ruf T (1999). "The Lomb-Scargle Periodogram in Biological Rhythm Research: Analysis of Incomplete and Unequally Spaced Time-Series." *Biological Rhythm Research*, **30**, 178–201. doi:10.1076/brhm.30.2.178.1422.

Salibian-Barrera M, Willems G, Zamar R (2008). "The Fast-τ Estimator for Regression." *Journal of Computational and Graphical Statistics*, **17**(3), 659–682. doi:10.1198/106186008x343785.

Salibian-Barrera M, Yohai V (2006). "A Fast Algorithm for S-Regression Estimates." *Journal of Computational and Graphical Statistics*, **15**(2), 414–427. doi:10.1198/106186006x113629.

Scargle J (1982). "Studies in Astronomical Time Series Analysis. II. Statistical Aspects of Spectral Analysis of Unevenly Spaced Data." *The Astrophysical Journal*, **263**, 835–853. doi:10.1086/160554.

Schwarzenberg-Czerny A (1989). "On the Advantage of Using Analysis of Variance for Period Search." *Monthly Notices of the Royal Astronomical Society*, **241**, 153–165. doi:10.1093/mnras/241.2.153.

Schwarzenberg-Czerny A (1998). "The Distribution of Empirical Periodograms: Lomb-Scargle and PDM Spectra." *Monthly Notices of the Royal Astronomical Society*, **301**(3), 831–840. doi:10.1111/j.1365-8711.1998.02086.x.

Seber G, Lee A (2003). *Linear Regression Analysis.* 2nd edition. John Wiley & Sons, Hoboken, New Jersey. doi:10.1002/9780471722199.

Stellingwerf R (1978). "Period Determination Using Phase Dispersion Minimization." *The Astrophysical Journal*, **224**, 953–960. doi:10.1086/156444.

Thieler A (2013). *Robuste Verfahren zur Periodendetektion in ungleichmäßig beobachteten Lichtkurven*. Doctoral thesis, TU Dortmund University.

Thieler A, Backes M, Fried R, Rhode W (2013). "Periodicity Detection in Irregularly Sampled Light Curves by Robust Regression and Outlier Detection." *Statistical Analysis and Data Mining*, **6**(1), 73–89. doi:10.1002/sam.11178.

Thieler AM, Rathjens J, Fried R (2015). ***RobPer**: Robust Periodogram and Periodicity Detection Methods*. R package version 1.2.1, with contributions from Clarke BR, Ligges U, Salibian-Barrera M, Willems G, Yohai V, URL https://CRAN.R-project.org/package=RobPer.

Timmer J, König M (1995). "On Generating Power Law Noise." *Astronomy and Astrophysics*, **300**, 707–710.

Tluczykont M, Bernardini E, Satalecka K, Clavero R, Shayduk M, Kalekin O (2010). "Long-Term Lightcurves from Combined Unified Very High Energy Gamma-Ray Data." *Astronomy and Astrophysics*, **524**, A48. doi:10.1051/0004-6361/201015193.

Uttley P, McHardy I, Papadakis I (2002). "Measuring the Broad-Band Power Spectra of Active Galactic Nuclei with RXTE." *Monthly Notices of the Royal Astronomical Society*, **332**(1), 231–250. doi:10.1046/j.1365-8711.2002.05298.x.

Varadhan R, Gilbert P (2009). "**BB**: An R Package for Solving a Large System of Nonlinear Equations and for Optimizing a High-Dimensional Nonlinear Objective Function." *Journal of Statistical Software*, **32**(4), 1–26. doi:10.18637/jss.v032.i04.

Venables W, Ripley B (2002). *Modern Applied Statistics with S*. 4th edition. Springer-Verlag, New York. doi:10.1007/978-0-387-21706-2.

Wang Z (2013). "**cts**: An R Package for Continuous Time Autoregressive Models via Kalman Filter." *Journal of Statistical Software*, **53**(5), 1–19. doi:10.18637/jss.v053.i05.

Yohai V, Zamar R (1988). "High Breakdown-Point Estimates of Regression by Means of the Minimization of an Efficient Scale." *Journal of the American Statistical Association*, **83**(402), 406–413. doi:10.2307/2288856.

Zechmeister M, Kürster M (2009). "The Generalised Lomb-Scargle Periodogram. A New Formalism for the Floating-Mean and Keplerian Periodograms." *Astronomy and Astrophysics*, **496**(2), 577–584. doi:10.1051/0004-6361:200811296.

Zhang Z, Chan S (2005). "Robust Adaptive Lomb Periodogram for Time-Frequency Analysis of Signals with Sinusoidal and Transient Components." In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*, volume 4, pp. 493–496.

A. Implementation diagrams for RobPer

In this appendix, the structure of the `RobPer` function is displayed as Nassi-Shneiderman diagram (structogram after Norm DIN 66261). Figure 11 contains a reading guidance for the blocks used in the structogram. The structogram for `RobPer` is displayed in Figure 12, for the algorithm `singleFUN` in Figure 13 and for the function `IRWLS` in Figure 14. The arguments and return values of the latter are shown in Table 5. The following definitions are used:

$$\zeta_{L_2}(r) = \sum_{i=1}^{n} r_i^2 \tag{34}$$

$$\zeta_{LTS}(r) = \sum_{i=1}^{h(m)} r_{(i)}, \qquad h(m) = \left\lfloor \frac{n}{2} \right\rfloor + \left\lfloor \frac{m+1}{2} \right\rfloor, \tag{35}$$

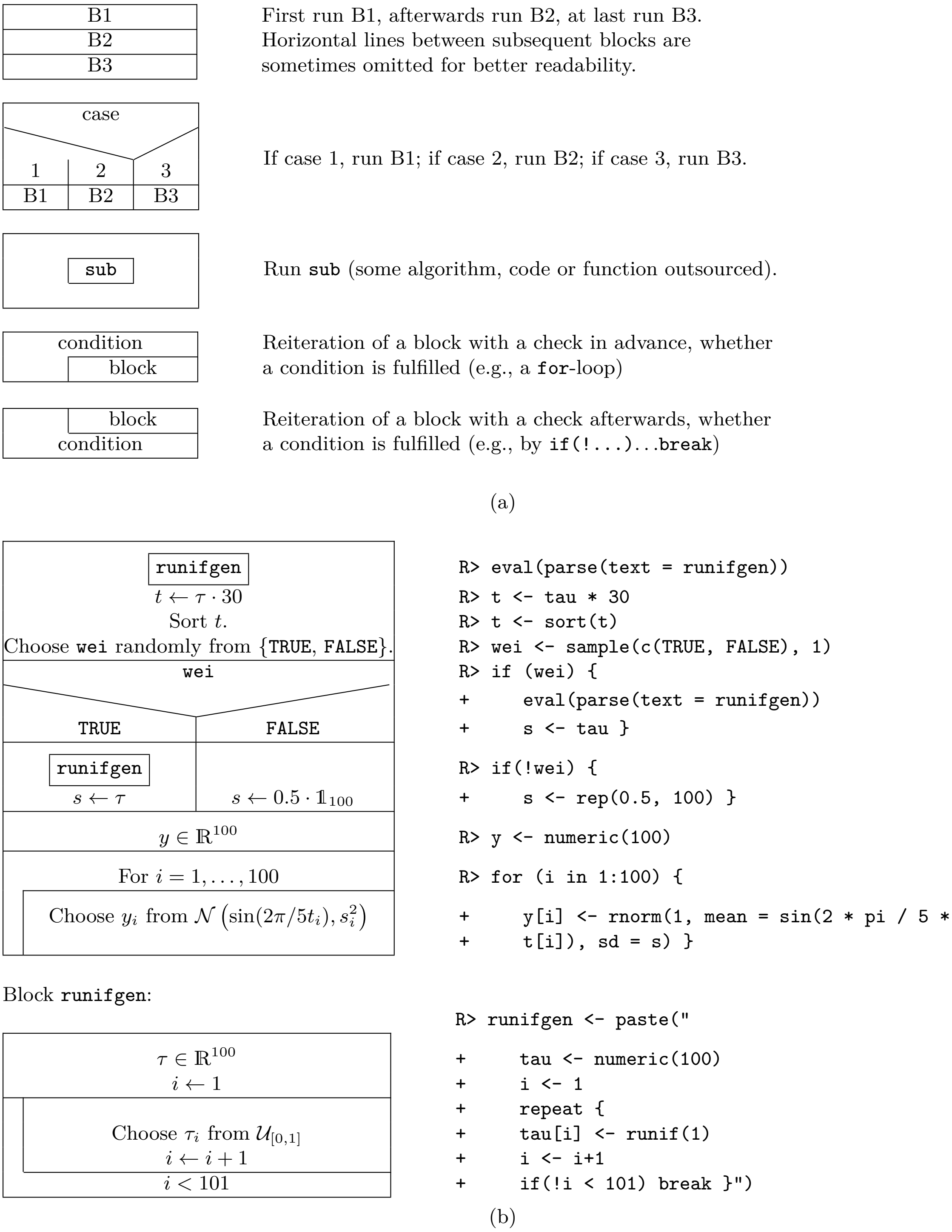
$$\zeta_{L_1}(r) = \sum_{i=1}^{n} |r_i|, \tag{36}$$

$$\rho_{MH}(\nu) = \begin{cases} \nu^2 & |\nu| \leq k \\ 2k|\nu| - k^2 & |\nu| > k \end{cases}, \qquad \rho_{MB}(\nu) = \begin{cases} 1 - \left(1 - \left(\frac{\nu}{k}\right)^2\right)^3 & |\nu| \leq k \\ 1 & |r| > k \end{cases}, \tag{37}$$

$$\zeta_{MH}(r) = \sum_{i=1}^{n} \rho_{MH}\left(\frac{r_i}{\widehat{\sigma}}\right), \qquad \zeta_{MB}(r) = \sum_{i=1}^{n} \rho_{MB}\left(\frac{r_i}{\widehat{\sigma}}\right), \tag{38}$$

$$W_{MH}(\nu) = \begin{cases} c_{MH} & |\nu| \leq k \\ c_{MH} \cdot \frac{k}{|\nu|} & |\nu| > k \end{cases}, \qquad W_{MB}(\nu) = \begin{cases} c_{MB} \cdot \left(1 - \left(\frac{\nu}{k}\right)^2\right)^2 & |\nu| \leq k \\ 0 & |\nu| > k \end{cases}. \tag{39}$$

The normalization constant can be set to $c_{MH} = c_{MB} = 1$ due to the scale invariance of the least squares estimation used in the iteratively reweighted least squares (IRWLS) step.


B1
B2
B3
case
1
2
3
B1
B2
B3
sub
condition
block
block
condition
(a)
runifgen
$t \leftarrow \tau \cdot 30$
Sort t.
Choose wei randomly from {TRUE, FALSE}.
wei
TRUE
FALSE
runifgen
$s \leftarrow \tau$
$s \leftarrow 0.5 \cdot \mathbb{1}_{100}$
$y \in \mathbb{R}^{100}$
For $i = 1, \dots, 100$
Choose y_i from $\mathcal{N}\left(\sin(2\pi/5t_i), s_i^2\right)$
Block runifgen:
$\tau \in \mathbb{R}^{100}$
$i \leftarrow 1$
Choose τ_i from $\mathcal{U}_{[0,1]}$
$i \leftarrow i + 1$
$i < 101$
(b)

First run B1, afterwards run B2, at last run B3. Horizontal lines between subsequent blocks are sometimes omitted for better readability.

If case 1, run B1; if case 2, run B2; if case 3, run B3.

Run `sub` (some algorithm, code or function outsourced).

Reiteration of a block with a check in advance, whether a condition is fulfilled (e.g., a `for`-loop)

Reiteration of a block with a check afterwards, whether a condition is fulfilled (e.g., by `if(!...)...break`)

```
R> eval(parse(text = runifgen))
R> t <- tau * 30
R> t <- sort(t)
R> wei <- sample(c(TRUE, FALSE), 1)
R> if (wei) {
+     eval(parse(text = runifgen))
+     s <- tau }
R> if(!wei) {
+     s <- rep(0.5, 100) }
R> y <- numeric(100)
R> for (i in 1:100) {
+     y[i] <- rnorm(1, mean = sin(2 * pi / 5 *
+     t[i]), sd = s) }
```

```
R> runifgen <- paste("
+     tau <- numeric(100)
+     i <- 1
+     repeat {
+     tau[i] <- runif(1)
+     i <- i+1
+     if(!i < 101) break }")
```

Figure 11: Reading guidance for the structograms: In Panel 11a, the blocks used for the representation of an algorithm. In Panel 11b, a structogram (left) for a simple R code (right), which generates the observations $(t_i, y_i, s_i)_{i=1,\dots,100}$ of a simple light curve with fluctuation period 5. This R code is for demonstration only and not programmed efficiently.

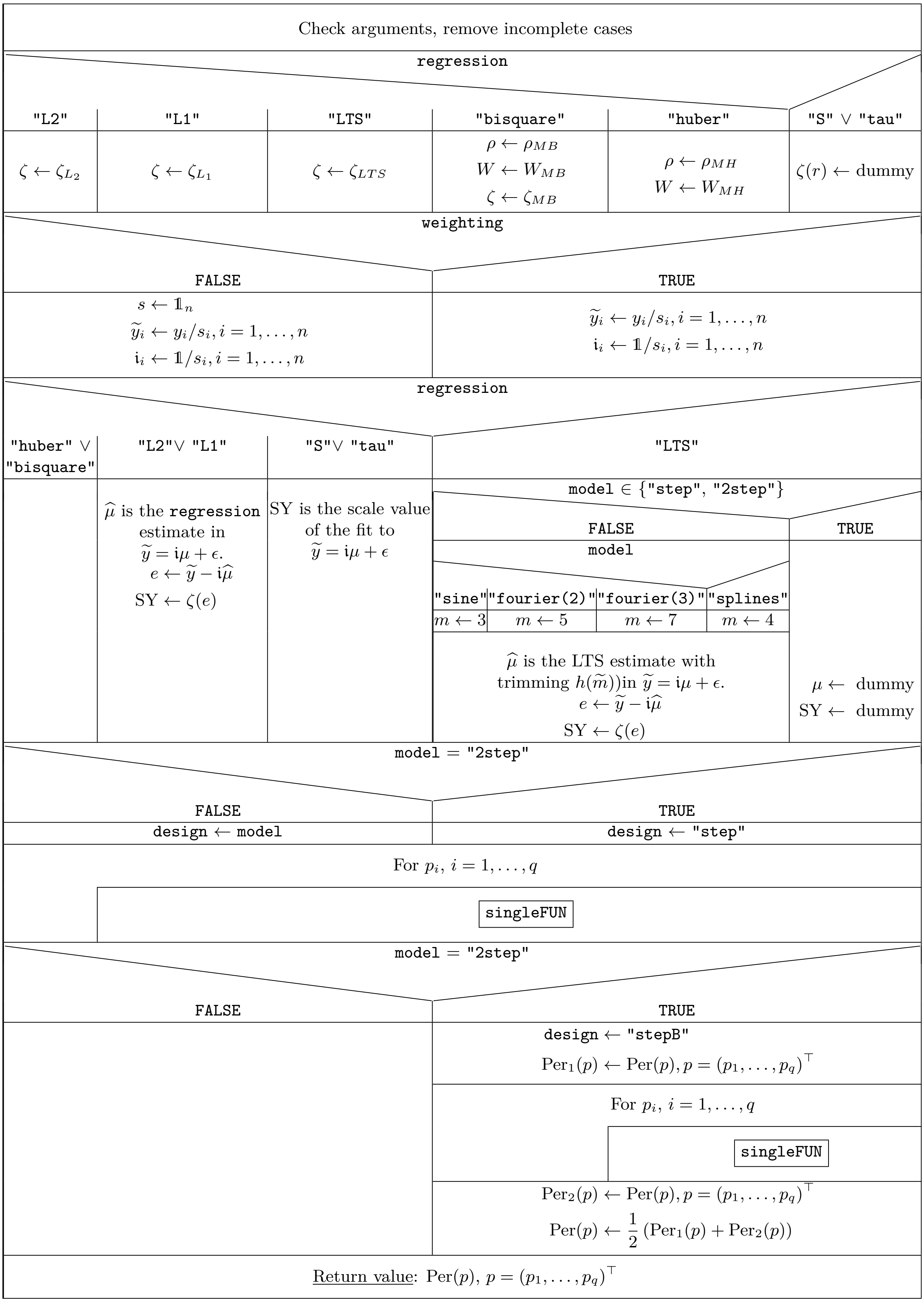

Check arguments, remove incomplete cases
regression
"L2"
"L1"
"LTS"
"bisquare"
"huber"
"S" $\vee$ "tau"
$\zeta \leftarrow \zeta_{L_2}$
$\zeta \leftarrow \zeta_{L_1}$
$\zeta \leftarrow \zeta_{LTS}$
$\rho \leftarrow \rho_{MB}$
$W \leftarrow W_{MB}$
$\zeta \leftarrow \zeta_{MB}$
$\rho \leftarrow \rho_{MH}$
$W \leftarrow W_{MH}$
$\zeta(r) \leftarrow$ dummy
weighting
FALSE
TRUE
$s \leftarrow \mathbb{1}_n$
$\widetilde{y}_i \leftarrow y_i/s_i, i = 1, \ldots, n$
$\mathbb{i}_i \leftarrow \mathbb{1}/s_i, i = 1, \ldots, n$
$\widetilde{y}_i \leftarrow y_i/s_i, i = 1, \ldots, n$
$\mathbb{i}_i \leftarrow \mathbb{1}/s_i, i = 1, \ldots, n$
regression
"huber" $\vee$ "bisquare"
"L2"$\vee$ "L1"
"S"$\vee$ "tau"
"LTS"
$\widehat{\mu}$ is the regression estimate in $\widetilde{y} = \mathbb{i}\mu + \epsilon$.
$e \leftarrow \widetilde{y} - \mathbb{i}\widehat{\mu}$
SY $\leftarrow \zeta(e)$
SY is the scale value of the fit to $\widetilde{y} = \mathbb{i}\mu + \epsilon$
model $\in$ {"step", "2step"}
FALSE
TRUE
model
"sine"
"fourier(2)"
"fourier(3)"
"splines"
$m \leftarrow 3$
$m \leftarrow 5$
$m \leftarrow 7$
$m \leftarrow 4$
$\widehat{\mu}$ is the LTS estimate with trimming $h(\widetilde{m})$)in $\widetilde{y} = \mathbb{i}\mu + \epsilon$.
$e \leftarrow \widetilde{y} - \mathbb{i}\widehat{\mu}$
SY $\leftarrow \zeta(e)$
$\mu \leftarrow$ dummy
SY $\leftarrow$ dummy
model = "2step"
FALSE
TRUE
design $\leftarrow$ model
design $\leftarrow$ "step"
For p_i, $i = 1, \ldots, q$
singleFUN
model = "2step"
FALSE
TRUE
design $\leftarrow$ "stepB"
$\mathrm{Per}_1(p) \leftarrow \mathrm{Per}(p), p = (p_1, \ldots, p_q)^\top$
For p_i, $i = 1, \ldots, q$
singleFUN
$\mathrm{Per}_2(p) \leftarrow \mathrm{Per}(p), p = (p_1, \ldots, p_q)^\top$
$\mathrm{Per}(p) \leftarrow \frac{1}{2}(\mathrm{Per}_1(p) + \mathrm{Per}_2(p))$
Return value: $\mathrm{Per}(p)$, $p = (p_1, \ldots, p_q)^\top$

Figure 12: Structogram of RobPer. The block singleFUN is displayed in detail in Figure 13.

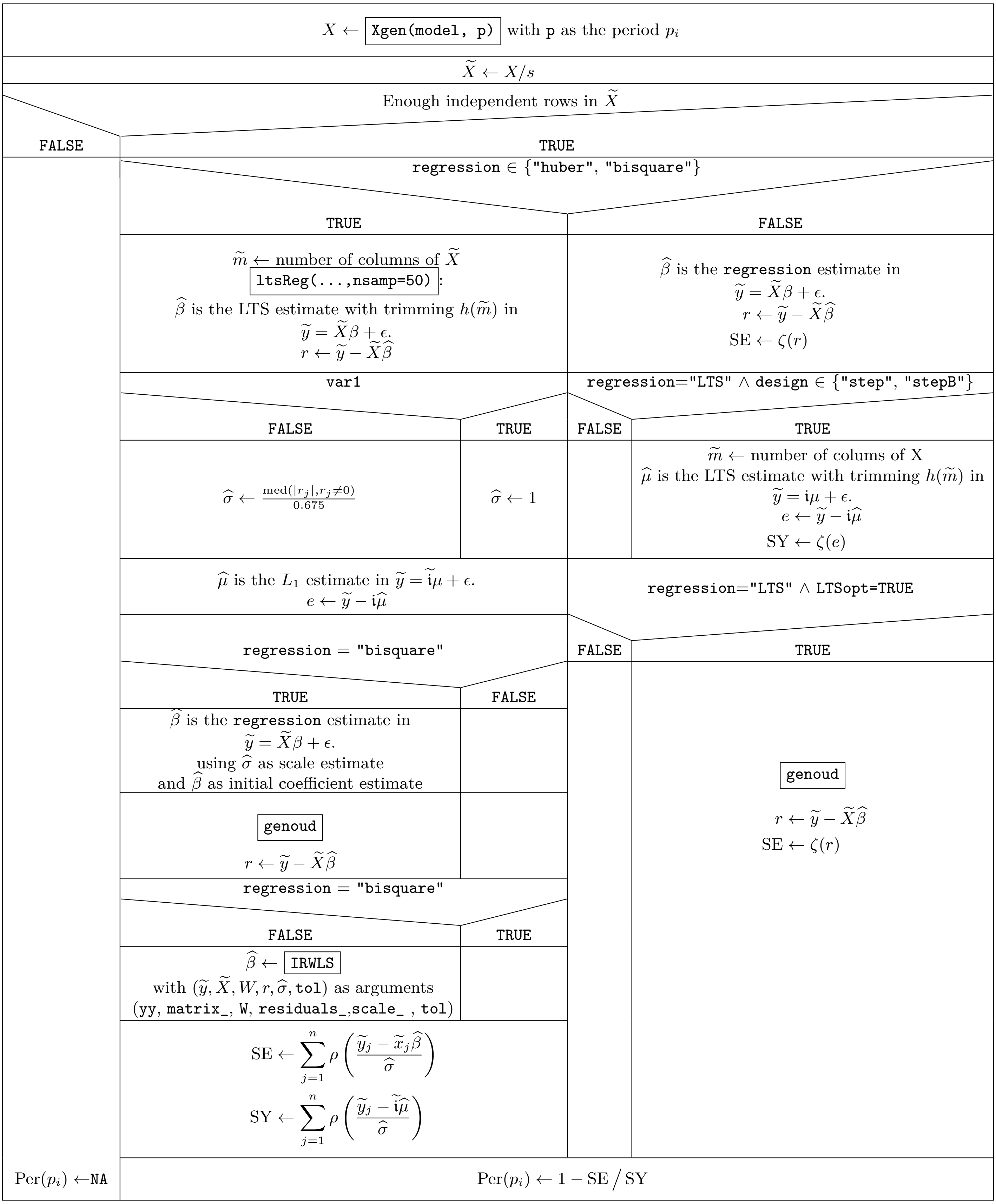
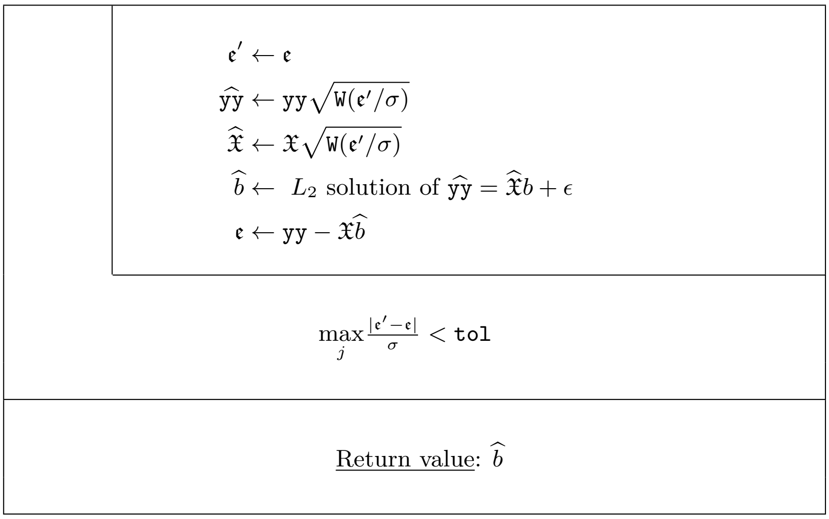

$X \leftarrow$ Xgen(model, p) with p as the period p_i
$\widetilde{X} \leftarrow X/s$
Enough independent rows in $\widetilde{X}$
FALSE
TRUE
regression $\in$ {"huber", "bisquare"}
TRUE
FALSE
$\widetilde{m} \leftarrow$ number of columns of $\widetilde{X}$
ltsReg(...,nsamp=50) :
$\widehat{\beta}$ is the LTS estimate with trimming $h(\widetilde{m})$ in
$\widetilde{y} = \widetilde{X}\beta + \epsilon.$
$r \leftarrow \widetilde{y} - \widetilde{X}\widehat{\beta}$
$\widehat{\beta}$ is the regression estimate in
$\widetilde{y} = \widetilde{X}\beta + \epsilon.$
$r \leftarrow \widetilde{y} - \widetilde{X}\widehat{\beta}$
SE $\leftarrow \zeta(r)$
var1
FALSE
TRUE
$\widehat{\sigma} \leftarrow \frac{\mathrm{med}(|r_j|, r_j \neq 0)}{0.675}$
$\widehat{\sigma} \leftarrow 1$
regression="LTS" $\wedge$ design $\in$ {"step", "stepB"}
FALSE
TRUE
$\widetilde{m} \leftarrow$ number of colums of X
$\widehat{\mu}$ is the LTS estimate with trimming $h(\widetilde{m})$ in
$\widetilde{y} = \mathrm{i}\mu + \epsilon.$
$e \leftarrow \widetilde{y} - \mathrm{i}\widehat{\mu}$
SY $\leftarrow \zeta(e)$
$\widehat{\mu}$ is the L_1 estimate in $\widetilde{y} = \widetilde{\mathrm{i}}\mu + \epsilon.$
$e \leftarrow \widetilde{y} - \mathrm{i}\widehat{\mu}$
regression="LTS" $\wedge$ LTSopt=TRUE
regression = "bisquare"
FALSE
TRUE
TRUE
FALSE
$\widehat{\beta}$ is the regression estimate in
$\widetilde{y} = \widetilde{X}\beta + \epsilon.$
using $\widehat{\sigma}$ as scale estimate
and $\widehat{\beta}$ as initial coefficient estimate
genoud
$r \leftarrow \widetilde{y} - \widetilde{X}\widehat{\beta}$
genoud
$r \leftarrow \widetilde{y} - \widetilde{X}\widehat{\beta}$
SE $\leftarrow \zeta(r)$
regression = "bisquare"
FALSE
TRUE
$\widehat{\beta} \leftarrow$ IRWLS
with $(\widetilde{y}, \widetilde{X}, W, r, \widehat{\sigma}, \mathtt{tol})$ as arguments
(yy, matrix_, W, residuals_,scale_ , tol)
$\mathrm{SE} \leftarrow \sum_{j=1}^{n} \rho\left(\frac{\widetilde{y}_j - \widetilde{x}_j\widehat{\beta}}{\widehat{\sigma}}\right)$
$\mathrm{SY} \leftarrow \sum_{j=1}^{n} \rho\left(\frac{\widetilde{y}_j - \widetilde{\mathrm{i}}\widehat{\mu}}{\widehat{\sigma}}\right)$
Per(p_i) $\leftarrow$NA
Per(p_i) $\leftarrow 1 - \mathrm{SE} \big/ \mathrm{SY}$

Figure 13: Structogram of `singleFUN`. `NA` indicates a missing value. The block `IRWLS` is displayed in detail in Figure 14.

Argument	Symbol	Explanation
`yy` $\in \mathbb{R}^n$	`yy`	Observed values
`matrix_` $\in \mathbb{R}^{n\times m}$	$\mathfrak{X}$	Design matrix
`W`: $\mathbb{R} \to \mathbb{R}_{\geq 0}$	`W`	Weight function
`residuals_` $\in \mathbb{R}^n$	$\mathfrak{e}$	Vector of residuals
`scale_` $\in \mathbb{R}_{>0}$	σ	(Estimate of) Standard deviation
`tol` $\in \mathbb{R}_{>0}$	`tol`	Precision for convergence
Return value		
`tempIRWLS$coeff`	$\widehat{b}$	Fitted vector of parameters

Table 5: Arguments and return value of the function `IRWLS`.



```
    $\mathfrak{e}' \leftarrow \mathfrak{e}$
    $\widehat{\texttt{yy}} \leftarrow \texttt{yy}\sqrt{\texttt{W}(\mathfrak{e}'/\sigma)}$
    $\widehat{\mathfrak{X}} \leftarrow \mathfrak{X}\sqrt{\texttt{W}(\mathfrak{e}'/\sigma)}$
    $\widehat{b} \leftarrow$ $L_2$ solution of $\widehat{\texttt{yy}} = \widehat{\mathfrak{X}}b + \epsilon$
    $\mathfrak{e} \leftarrow \texttt{yy} - \mathfrak{X}\widehat{b}$
$\max_j \frac{|\mathfrak{e}' - \mathfrak{e}|}{\sigma} < \texttt{tol}$
Return value: $\widehat{b}$
```

Figure 14: Function `IRWLS` in `RobPer`

Affiliation:

Anita M. Thieler
Faculty of Statistics
TU Dortmund University
44221 Dortmund, Germany
E-mail: anita.thieler@tu-dortmund.de
URL: http://www.statistik.tu-dortmund.de/thieler-en.html