1. Time Series Analysis in Astronomy II

- Lauri Jetsu
  Ph.D. (1994), Docent (1996), Director of the Observatory of the University of Helsinki (2001–2009), University Lecturer (2010 →)
- Jyri Lehtinen, M.Sc. (2010), Assistant of the Course.
- Aim: To learn, program and apply the Continuous Period Search method (CPS) formulated in Lehtinen et al. 2011 (A&A 527, A136)
- Homepage of the Course:
  `http://www.helsinki.fi/~jetsu/time2/time2.html`
- Only English ⇒ Ask about words when necessary.
- We shall use `emacs` to edit our files/programs.
- Our programming language will be IDL (Interactive Data Language)
- You need username/access to `sky.astro.helsinki.fi` $\boxed{\text{work}}$
- Connection `ssh -Y -ljetsu sky.astro.helsinki.fi`
- Evaluation
  - ⊙ No exam(s), only assignments
  - ⊙ Division into groups, if necessary (fill the form).
  - ⊙ Assignments during lectures and also as homework.
  - ⊙ No competion: Groups can/will help each other
  - ⊙ You explain topics to each other, and to me.
  - ⊙ Do not get embarrassed, just say: “I don’t know.”
  - ⊙ Do not memorize, try to understand instead or ask.
  - ⊙ I/we omit details on purpose, e.g. d/dx(sin x)=?
- Lecture: New concept(-s) explained ⇒ assignment(-s).
  ⊙ You will show and explain your solutions of personal assignments to others. Solutions for all assignments always explained to everybody “until it gets boring”.

Sketch

This skecth of our timetable will be constantly updated during this term. Your progress determines the pace.

1. Grid search and Refined Search: One dataset: `Input, Output`
   Exercise 1: Case of K=1. Then extended to K=0 and 2.
2. K criterion: One dataset: `Input, Output`
3. Bootstrap: One dataset: `Input, Output`
   Note that original value of free parameters should not change!
4. Kolmogorov-Smirnov test: One dataset: `Input, Output`
5. Reliability: One dataset: `Input, Output`
6. Format: One dataset: `Input, Output`
7. Segment and dataset division, Outliers, Revised segments and datasets `Input, Output`
8. All datasets: All subroutines `Input, Output`
9. All datasets: All subroutines `Input, Output`
10. All datasets: Time scale of change `Input, Output`
11. Format: Presentation of all parameters `Input, Output`
12. Rayleigh and Kuiper test `Input, Output`
13. Summary of results `Input, Output`

Concepts repeated from Time Series Analysis in Astronomy I (hereafter TSA I)

- Page numbering below is taken from `http://www.helsinki.fi/~jetsu/time1/time1.html/timeseries1.pdf`
  - ⊙ Model for the data (page 29)
  - ⊙ Least squares fit (page 30)
  - ⊙ `CURVEFIT` routine (page 32)
  - ⊙ Error estimates for free parameters (page 37)
  - ⊙ Grid search (hereafter GSch) (pages 40–42)
  - ⊙ Refined search (hereafter RSch) (page 47)
  - ⊙ Combining GSch and RSch for $K = 1$ (page 48)
- AIM: TO UNDERSTAND THE SIMILARITIES AND DIFFERENCES BETWEEN LINEAR AND NONLINEAR MODELS!
- The following data files can be found in the homepage. These will be analysed in the first exercises.
  - ⊙ `STANDARD0.DAT`
  - ⊙ `STANDARD1.DAT`
  - ⊙ `STANDARD2.DAT`
- Let's practice how to copy these with a mouse, as well as some model programs.

Simple first order linear model

- A linear fit example is given in `MODEL6.PRO` in TSA I.
- The first order, $K=1$, **linear** model is
  $g(t;\bar{\beta}) = M + B_1 \cos 2\pi f t + C_1 \sin 2\pi f t = M + B_1 \cos x + C_1 \sin x$
- The free parameters are $\bar{\beta} = [M, B_1, C_1]$ = `A=[A(0),A(1),A(2)]`.
- The tested frequency f =`F(BB)` is **not a free parameter**.
- `FUNCT1` argument $x = 2\pi f(t - t_1)$ =`X=2.D0*!PI*F(BB)*(T-MIN(T))`.
- The first observing time is t_1 =`MIN(T)`.

```
; ----------------------------------------------------------------
PRO FUNCT1,X,A,F,PDER                          ; SUBROUTINE
F=0.+A(0)+A(1)*COS(X)+A(2)*SIN(X)              ; g (model)
PDER=DBLARR(N_ELEMENTS(X),3)                   ; Partial derivatives
PDER(*,0)=1.D0                                 ; dg/dM
PDER(*,1)=COS(X)                               ; dg/dB_1
PDER(*,2)=SIN(X)                               ; dg/dC_1
RETURN & END                                   ; Ends any subroutine.
; ----------------------------------------------------------------
```

- Data error, σ_i= `E`, gives weights $w_i = \sigma_i^{-2}$ = `W=1.D0/(E*E)`

```
W=1.D0/(E*E)
A=DBLARR(3)+0.1D0
X=2.D0*!PI*F(BB)*(T-MIN(T))
YFIT=CURVEFIT(X,Y,W,A,EA,FUNCTION_NAME='FUNCT1')
```

YFIT	$\bar{g}(t, \bar{\beta}_{\text{final}})$	OUTPUT
X	$2\pi\bar{t}$	INPUT
Y	$\bar{y}$	INPUT
W	$\bar{w}$	INPUT
A	$\bar{\beta}_{\text{trial}} \Rightarrow \bar{\beta}_{\text{final}}$	INPUT ⇒ OUTPUT
EA	$\bar{\sigma}_{\beta}$	OUTPUT
PDER(*,0)=1.D0	$\frac{\partial g(t,\bar{\beta})}{\partial M}$	INPUT
PDER(*,1)=SIN(X)	$\frac{\partial g(t,\bar{\beta})}{\partial B_1}$	INPUT
PDER(*,2)=COS(X)	$\frac{\partial g(t,\bar{\beta})}{\partial C_1}$	INPUT

- Let's play with `MODEL6.PRO` from TSA I, "until it gets boring".

General K :th order linear or nonlinear model

- The general K :th order model is

$$\begin{aligned} g(\bar{\beta}) = g(t,\bar{\beta}) &= M + \sum_{\mathrm{j}=1}^{\mathrm{K}} B_{\mathrm{j}} \cos{(j2\pi f t)} + C_{\mathrm{j}} \sin{(j2\pi f t)} \\ &= M + \sum_{\mathrm{j}=1}^{\mathrm{K}} B_{\mathrm{j}} \cos{(jx)} + C_{\mathrm{j}} \sin{(jx)}, \end{aligned}$$

  where argument is $x = 2\pi f(t - t_1)$ and t_1 is first observing time.

- **Linear** model has $Q = 2K + 1$ free parameters: $\bar{\beta} = [M, B_1, C_1, ..., , B_{\mathrm{K}}, C_{\mathrm{K}}] =$`[A(0),A(1),...,A(2K)]`, where $M =$`A(0)`, $B_j =$`A(2j-1)` and $C_j =$`A(2j)`

- **Nonlinear** model has $Q = 2K + 2$ free parameters: $\bar{\beta} = [M, B_1, C_1, ..., B_{\mathrm{K}}, C_{\mathrm{K}}, f] =$ `[A(0),A(1),...,A(2K+1)]`, where $M =$`A(0)`, $B_j =$`A(2j-1)`, $C_j =$`A(2j)` and $f =$`A(2K+1)`

Linear: $x = 2\pi f(t - t_1)$	`FUNCTK`	**Nonlinear:** $x = 2\pi(t - t_1)$	`GUNCTK`
$\frac{\partial g}{\partial M} = 1$	`PDER(*,0)`	$\frac{\partial g}{\partial M} = 1$	`PDER(*,0)`
$\frac{\partial g}{\partial B_j} = \cos{(jx)}$	`PDER(*,2j-1)`	$\frac{\partial g}{\partial B_j} = \cos{(jfx)}$	`PDER(*,2j-1)`
$\frac{\partial g}{\partial C_j} = \sin{(jx)}$	`PDER(*,2j)`	$\frac{\partial g}{\partial C_j} = \sin{(jfx)}$	`PDER(*,2j)`
		$\frac{\partial g}{\partial f} = \sum_{\mathrm{j}=1}^{\mathrm{K}} jx\,[C_k \cos{(jfx)} - B_j \sin{(jfx)}]$	`PDER(*,2K+1)`

- You can use these instructions to edit your own subroutines

`FUNCT1` for GSch K=1	`GUNCT1` for RSch K=1
`FUNCT2` for GSch K=2	`GUNCT2` for RSch K=2

- QUESTION: DO YOU NOW UNDERSTAND THE SIMILARITIES AND DIFFERENCES BETWEEN LINEAR AND NONLINEAR MODELS!

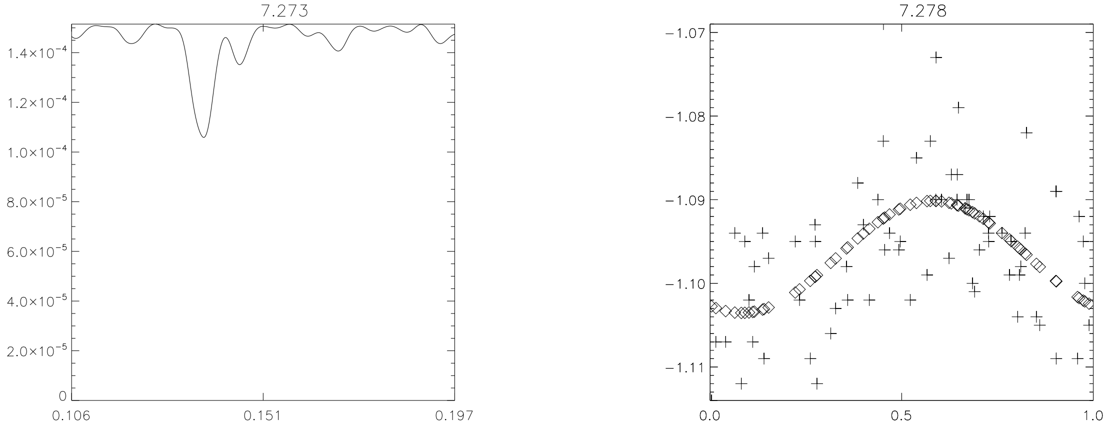

Fig. 1. One example of a solution for **Exercise 1.**

- **Exercise 1.** Write an IDL-program EXERCISE1.PRO that

  (a) Calculates the first order $K = 1$ grid search periodogram $\Theta_{\mathrm{grid}}(f)$ for the data in `STANDARD1.DAT`. The accuracy of these data is $\sigma = 0.008$, i.e. use $w = \sigma^{-2}$ for all y_i. The tested period interval is $P_{\min} = 0.7 * P$ and $P_{\max} = 1.3 * P$, where $P = 7.25$ days and `OFAC=20`.

  (b) Plots the GSch periodogram as a continuous line (`PSYM=0`) and shows the best period above the plot.

  (c) Performs the RSch, which is based on the results of GSch.

  (d) Plots the data, y_i (`Y` denoted with crosses, i.e. `PSYM=1`), and the model of RSch, $g(t_i, \bar{\beta}_{\mathrm{final}})$ (`YFIT` denoted with diamonds, i.e. `PSYM=4`), as a function of phase calculated with this best period. Shows the best period given by RSch above this plot.

∼ • ∼ Exercise ends here. ∼ • ∼

Best modelling order K

- **Problem**: If the model is

  $$g(\bar{\beta}) = g(t, \bar{\beta}) \; = M + \sum_{\mathrm{j}=1}^{\mathrm{K}} B_{\mathrm{j}} \cos{(j 2\pi f t)} + C_{\mathrm{j}} \sin{(j 2 \pi f t)},$$

  what is the best modelling order K for any arbitrary data $\bar{y}$?
- The numbering of equations below is from Lehtinen et al. 2010 which can be found from the homepage.
- Definition of χ^2 in Eq. 5 was

  $\chi^2(\bar{y}; \bar{\beta}) = \sum_{i=1}^{n} w_i \epsilon_i^2$
- Bayesian Information Criterion in Eq. 6 was

  $R_{\mathrm{BIC}} = 2n \ \ln \ \lambda(\bar{y}; \bar{\beta}) + (5K + 1) \ \ln \ n,$

  where $\lambda(\bar{y}; \bar{\beta}) = \chi^2(\bar{y}; \bar{\beta})/[\Sigma_{i=1}^{n} w_i]$.
- Logic: $\chi(\bar{y}; \bar{\beta})$ decreases, if K increases.
  More complicated models decrease the first term $2n \ \ln \ \lambda(\bar{y}; \bar{\beta})$, but at the same time increase the second term $(5K + 1) \ \ln \ n$.
- The best value for K minimizes R_{BIC}.
- Solution: Test the cases $K = 0, 1, 2, 3, ...,$. If R_{BIC} begins to increase with $K = K'$, then the best modelling order is $K = K' - 1$.
- K $= 0$ model (i.e. no periodicity, i.e. $g(t, \bar{\beta})$ is constant) is simply the weighted mean of all data $y_i \pm \sigma_i$, i.e.
  $g(t, \bar{\beta}) = M = [\Sigma_{i=1}^{n} w_i y_i] \, / \, \Sigma_{i=1}^{n} w_i$, where $w_i = \sigma_i^{-2}$.
- **Question**: What would the model, where j begins from zero, i.e.

  $$g(\bar{\beta}) = g(t, \bar{\beta}) \; = M + \sum_{\mathrm{j}=0}^{\mathrm{K}} B_{\mathrm{j}} \cos{(j 2\pi f t)} + C_{\mathrm{j}} \sin{(j 2 \pi f t)},$$

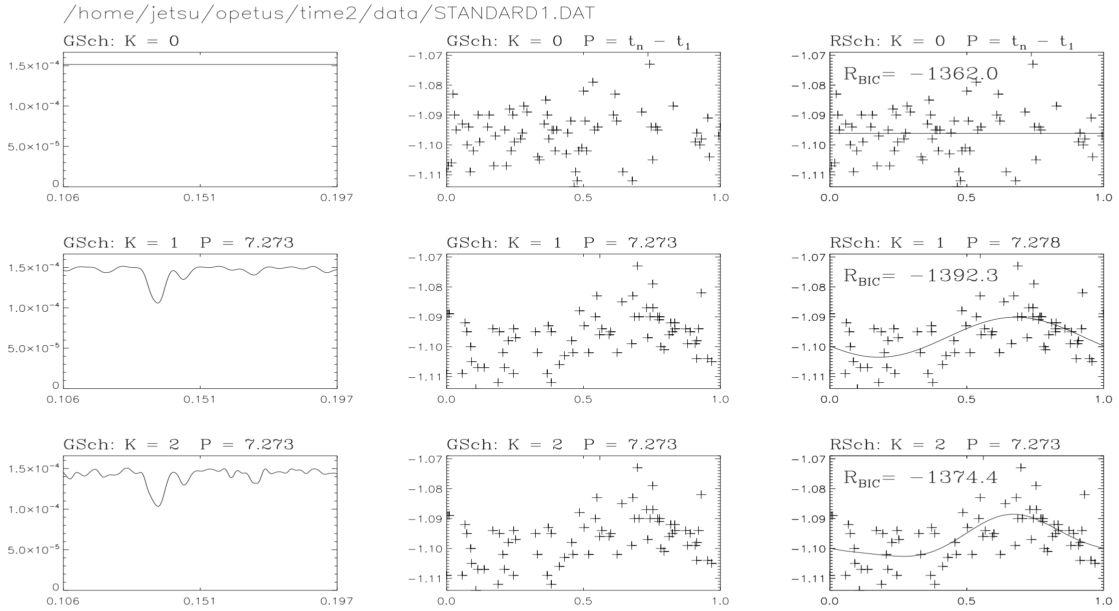

Fig. 2. One example of a solution for **Exercise 2.**

- **Exercise 2.** Test each dataset `STANDARD0.DAT`, `STANDARD1.DAT` and `STANDARD2.DAT` separately. The accuracy is $\sigma_i = 0.008$. Write an IDL-program EXERCISE2.PRO that

  (a) Calculates the $K = 0, 1$ and 2 grid search periodograms $\Theta_{\mathrm{grid}}(f)$. The tested period interval is $P_{\mathrm{min}} = 0.7 * P$ and $P_{\mathrm{max}} = 1.3 * P$, where $P = 7.25$ days and `OFAC=20`. Plots the GSch periodogram as a continuous line (`PSYM=0`). Shows the best period.

  (b) Plots the data as function of phase calculated with the best period P_{best} given by GSch. Use $P_{\mathrm{best}} = t_n - t_1$ for the $K = 0$ model.

  (c) Performs RSch based on the results of GSch. Plots the data, y_i (crosses, i.e. `PSYM=1`), and the model of RSch, $g(t_i, \bar{\beta}_{\mathrm{final}})$ (continuous line, i.e. `PSYM=0`), as a function of phase calculated with this best period. Shows the best period P_{best} given by RSch, as well as the value of R_{BIC}. Use $P_{\mathrm{best}} = t_n - t_1$ for the $K = 0$ model.

  (d) Did this program reveal the best modelling order for each of the three datasets? Write a few lines about this in our e-mail to the assistant.

$\sim \bullet \sim$ **Exercise ends here.** $\sim \bullet \sim$

Best modelling order K for data with a constant accuracy.

- If $w_i = w =$ constant, then
  $\chi^2(\bar{y};\bar{\beta}) = \sum_{i=1}^{n} w_i \epsilon_i^2 = w \sum_{i=1}^{n} \epsilon_i^2$ and $\sum_{i=1}^{n} w_i = wn$
  which gives $\lambda(\bar{y};\bar{\beta}) = n^{-1} \sum_{i=1}^{n} \epsilon_i^2$
- In this particular case, the R_{BIC} of Eq. 6 is

$$\begin{aligned} R_{\mathrm{BIC}} &= 2n \ \ln\,[n^{-1} \sum_{i=1}^{n} \epsilon_i^2] + (5K+1)\ \ln\, n \\ &= 2n\{\ \ln\,[n^{-1}] + \ \ln\,[\sum_{i=1}^{n} \epsilon_i^2]\} + (5K+1)\ \ln\, n \\ &= 2n\{\ \ln\,[\sum_{i=1}^{n} \epsilon_i^2] - \ \ln\, n\} + (5K+1)\ \ln\, n \end{aligned}$$

- The final result is

$$R_{\mathrm{BIC}} = 2n \ \ln\,[\sum_{i=1}^{n} \epsilon_i^2] + (5K - 2n + 1)\ \ln\, n$$

- **Question**: There is something “funny” in this result. Can anybody “see” what it is? Take your time. I spent some time figuring it out.

Bootstrap.

- **Question:** Why is bootstrap used?
- **Answer:** Bootstrap can solve any model parameter and its error numerically!
- Six stages of bootstrap explained in detail in exit TSA I (page 49.)

  ⊙ **1:** Model the original data $\bar{y}$ with $\bar{w}$. It gives $\epsilon_{\mathrm{i}} = y(t_{\mathrm{i}}) - g(t_{\mathrm{i}}, \bar{\beta}_{\mathrm{min}}) = y_{\mathrm{i}} - g_{\mathrm{i}}$.
  Results are one estimate for $\bar{\beta}_{\mathrm{min}}$ and for other model parameters.

  ⊙ **2:** Select a random sample $\bar{\epsilon}^*$ from $\bar{\epsilon}$. Connection of ϵ_i to weights w_i gives $\bar{w}^*$.

  ⊙ **3:** A random data sample $\bar{y}^* = \bar{g} + \bar{\epsilon}^*$ with $\bar{w}^*$ is obtained.

  ⊙ **4:** Model the random data $\bar{y}^*$ with $\bar{w}^*$.
  Results are one estimate for $\bar{\beta'}_{\mathrm{min}}$ and for other model parameters.

  ⊙ **5:** Return 2nd stage, until S estimates of $\bar{\beta'}_{\mathrm{min}}$ and “other” parameters have been obtained.

  ⊙ **6:** The expectation value and variance for any $\bar{\beta}_{\mathrm{min}}$ component are the mean and variance of its S estimates in $\bar{\beta}'_{\mathrm{min}}$. The same applies to the S estimates of “other” model parameters.
- Note:

  ⊙ $\bar{y}$, $\bar{w}$, $\bar{\epsilon}$ and $\bar{g}$ do not change during bootstrap.

  ⊙ $\bar{y}^*$, $\bar{w}^*$ and $\bar{\epsilon}^*$ change during every bootstrap round.

  ⊙ $\bar{\epsilon}^*$ determines $\bar{y}^*$ and $\bar{w}^*$.
- `MODEL10.PRO` in TSA I gave an example of bootstrap, where the epoch of the minimum (`RESULT1`) and total amplitude (`RESULT2`) were solved numerically. Let us copy and test!

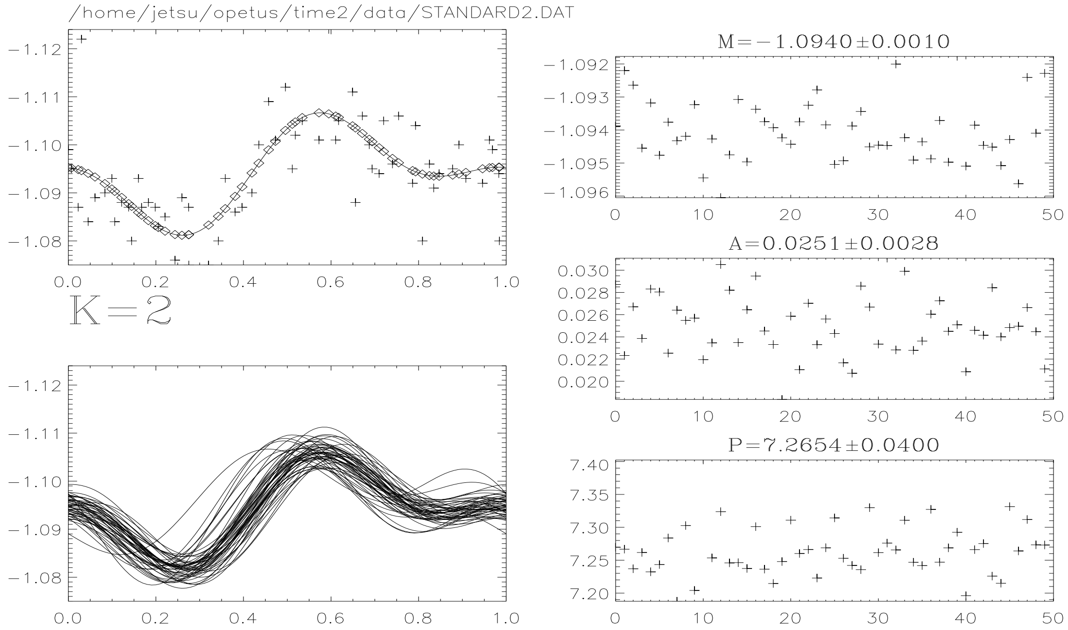

Fig. 3. One example of a solution for **Exercise 3.**

- **Exercise 3.** Edit a program that performs $S = 50$ bootstrap rounds to determine the values and errors of M (mean), A (total amplitude) and P (period) of the $K = 2$ model for `STANDARD2.DAT`. **(a)** Perform $K = 2$ order GSch and Rsch for the original data as in Exercise 2. Plot original data and the original model in a separate panel. **(b)** Perform $K = 2$ order GSch and Rsch for every random sample $\bar{y}^*$ and $\bar{w}^*$. Plot the model for each of these random samples in the same panel. **(c, d, e)** Collect the $S = 50$ estimates for M, A and P. Plot these estimates in separate panels. Give their mean values and standard deviations.

∼ • ∼ Exercise ends here. ∼ • ∼

Verifying Bootstrap

- Kolmogorov–Smirnov test verifies bootstrap validity:

  ⊙ Is the n residuals ϵ_{i} distribution gaussian?

  ⊙ Are the bootstrap distributions of the S estimates for every $\bar{\beta}$ component, and for other model parameters (e.g. A) gaussian?
- A general and very useful test, since any distribution (i.e. $F(u)$ in the next page) can be tested.

Kolmogorov–Smirnov Test for Gaussian Distribution

- 1: H_0: the S values $x_1, x_2, ..., x_{\mathrm{S}}$ represent a random sample drawn from a gaussian distribution.
- 2: Fix the preassigned significance level γ (e.g. 0.1, 0.05, 0.01, etc...) for rejecting H_0. Choose the corresponding $c_0(\gamma, n)$ limit for the Kolmogorov–Smirnov test statistic c, where $P(c \geq c_0) \leq \gamma$ (e.g. Kreyszig 1970: Table 7 on p. 453).
- 3: Arrange $\bar{x}$ into an ascending (i.e. rank) order $x_1 \leq x_2 \leq ... \leq x_{\mathrm{S}}$
- 4: Transform x_{i} to $u_{\mathrm{i}} = (x_{\mathrm{i}} - m_{\mathrm{x}})/s_{\mathrm{x}}$, where m_{x} and s_{x} are the mean and standard deviation of $\bar{x}$.
- 5: Derive the cumulative distribution function

$$F_{\mathrm{S}}(u) = \begin{cases} 0, u < u_1 \\ iS^{-1}, u_{\mathrm{i}} \leq u < u_{\mathrm{i+1}} \\ 1, u > u_{\mathrm{S}} \end{cases}$$

- 6: Reject H_0 if, and only if,

  $c = \max[\ | F_{\mathrm{S}}(u) - F(u) |\ ] > c_0,$

  where $F(u) = (2\pi)^{-1/2} \int_{-\infty}^{u} \mathrm{e}^{-z^2/2} dz$ is the cumulative gaussian distribution function.
- **Note:** The above definition of $F_S(u)$ means that it is step function having two values $F_S = (i-1)/S$ and $F_S = i/S$ at every u_i. This means that you have to calculate $F_S(u_1) = 0, F_S(u_1) = 1/S, F_S(u_2) = 1/S, F_S(u_2) = 2/S, ..., F_S(u_S) = (S-1)/S, F_S(u_s) = S/S = 1$, for example in the next Exercise 4.

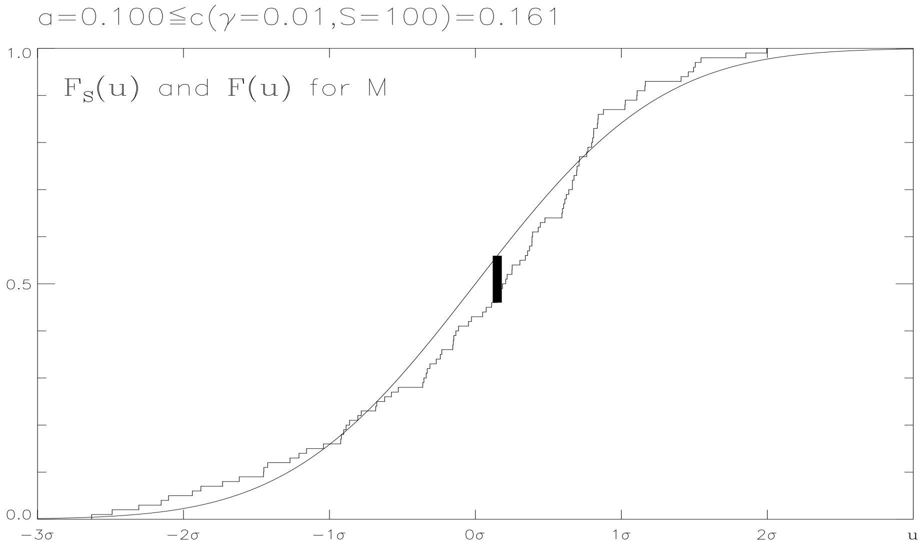

Fig. 4. One example of a solution for **Exercise 4.**

- **Exercise 4.** Write an IDL-program EXERCISE4.PRO that performs Kolmogorov Smirnov test for $S = 100$ bootstrap estimates of the mean M. Test the gaussian distribution hypothesis. Plot $F_{\mathrm{S}}(u)$ and $F(u)$. Give c and c_0. Show the place from which you obtained the value of c. Use the c_0 of $\gamma = 0.01$ on 5th column of `KOLMOROV.DAT` which can be copied from the homepage under text *"The data for the first exercises can copied from HERE."* One example of a solution is given in Fig. 4.

$\sim \bullet \sim$ **Exercise ends here.** $\sim \bullet \sim$

Format used in presenting the results

- All parameters explained in Lehtinen et al. (2011) (see homepage)

[SEG] = integer	$[t_0]$ = HJD	[SET] = integer	
$[t_1]$ = HJD	$[t_n]$ = HJD	$[\tau]$ = HJD	[IND] = integer
$[n]$ = integer	$[m_y]$ = mag	$[s_y]$ = mag	
$[K]$ = integer	$[\sigma_\epsilon]$ = mag	$[T_{\mathrm{C}}]$ = d	
$[M]$ = mag	$[\sigma_M]$ = mag	[R] = integer	
$[P]$ = d	$[\sigma_P]$ = d	[R] = integer	
$[A]$ = mag	$[\sigma_A]$ = mag	[R] = integer	
$[t_{\mathrm{min},1}]$ = HJD	$[\sigma_{t_{\mathrm{min},1}}]$ = d	[R] = integer	
$[t_{\mathrm{min},2}]$ = HJD	$[\sigma_{t_{\mathrm{min},2}}]$ = d	[R] = integer	

- We will use considerable time to go to over these again and again.
- **Exercise 5.** Write an IDL-program EXERCISE5.PRO that calculates similar results for the data in `STANDARD0.DAT`, `STANDARD1.DAT` and `STANDARD2.DAT`. Determine the best modelling order K for each of these three datasets. Model each dataset with this best modelling order K. You do not have to solve $t_{\mathrm{min},1}$, $t_{\mathrm{min},2}$ or T_{C}, just use `-1` for them. Use the same notation also for SEG, t_0, SET and IND. For the $K = 0$ models, also use `-1` for P and A. Give the results in the following format

```
        -1          -1          -1
51200.9434  51225.8738  51214.2168         -1
        22      0.2658      0.0086
         2      0.0055          -1
    0.2654      0.0012           0
    7.1382      0.1267           0
    0.0274      0.0045           0
        -1          -1          -1
        -1          -1          -1
```

Use time units `T=T-2400000.0D0`, i.e. 5 integers and 4 decimals.

∼ • ∼ Exercise ends here. ∼ • ∼

Datasets (SET) and Segments (SEG)

- The data are divided into datasets (SET) and segments (SEG).
- Every dataset (SET) is modelled separately.
- SEG notation is used only to divide/identify the datasets belonging to separate seasons.
- (1) Fix the minimum value of data points, $n_{\rm min}$, in a dataset.
- (2) Fix the maximum length in time, $\Delta T_{\rm max}$, of a dataset.
- (3) Begin from $t' = t_{j=1}$.
- (4) Select n observations $y_i = y(t_i)$ that fulfill $t' \leq t_i < t' + \Delta T_{\rm max}$
- (5) Use the following criterion:
  A: You have found a dataset candidate, if $n \geq n_{\rm min}$.
  B: You have not found a dataset candidate, if $n < n_{\rm min}$.
- (6) Use the next time point, i.e. $t' = t_{j+1}$ and go back to (4).
- (7) Stop when you have gone through all data.
- (8) From all these candidates, choose only those datasets that fulfill

  $$\mathrm{SET}_k \not\subset \mathrm{SET}_{k+1} \text{ and } \mathrm{SET}_{k+1} \not\subset \mathrm{SET}_k.$$

- In short, subsequent chosen datasets must contain at least one observation that does not belong to the other one.
- (9) Identify the segments. A new segment begins, if there is a gap in observations greater than $\Delta T_{\rm max}$.
- (10) Use the temporal order SET=1, 2, 3, ... in the numbering of all datasets of every segment.

- **Exercise 6.** You will receive data via ordinary mail. Divide it into datasets and segments using $n_{\rm min} = 10$ and $\Delta T_{\rm max} = 30$ days. **Note:** This is unpublished real data. Do not distribute it elsewhere!

∼ • ∼ Exercise ends here. ∼ • ∼

- **Exercise 7.** Write an IDL-program EXERCISE7.PRO that calculates the same results as in Exercise 5, but now for all datasets (SET) and segments (SEG) of Exercise 6. Determine the best modelling order K for each of these datasets. Model each dataset with this best modelling order K. You do not have to solve $t_{\mathrm{min},1}$, $t_{\mathrm{min},2}$ or T_{C}, just use `-1` for them. Use the same notation `-1` also for IND. Note that you can now tabulate the values of SEG, t_0 and SET. For the $K = 0$ models, also use `-1` for P and A. Give the results for all datasets and segments in one file using in the following format

```
         1  51200.9434           1
51200.9434  51225.8738  51214.2168        -1
        22      0.2658      0.0086
         2      0.0055          -1
    0.2654      0.0012           0
    7.1382      0.1267           0
    0.0274      0.0045           0
        -1          -1          -1
        -1          -1          -1
         1  51200.9434           2
51202.9563  51227.761  51216.4168        -1
        19      0.2678      0.0076
         1      0.0052          -1
    0.2712      0.0011           0
    7.1452      0.1567           0
    0.0286      0.0029           0
        -1          -1          -1
        -1          -1          -1
 ... continues ...
```

Use time units `T=T-2400000.0D0`, i.e. 5 integers and 4 decimals.

∼ • ∼ **Exercise ends here.** ∼ • ∼

Outliers I

- Data can contain outliers caused, e.g. by measurement errors, clouds, flares, variable comparison stars, ...
- Outliers must be identified and removed before final modelling.
- SEG and SET values may, or may not, change when outliers are removed. Why?
- Problem: How to identify outliers?
- Solution: The residuals of a reasonable model have a gaussian distribution. An outlier deviates from this distribution.
- Use the model that has the residuals $\epsilon_1, \epsilon_2, ... \epsilon_n$. The standard deviation of these residuals is σ_ϵ. Outlier datapoints y_i fulfill

  $$|\epsilon_i / \sigma_\epsilon| > A, \tag{1}$$

  where suitable values are $A \geq 2$.
- **Exercise 8.** Copy your EXERCISE7.PRO to EXERCISE8.PRO. Model all SET and SEG like in the previous exercise. The residuals connected to data points t_i and y_i are $\epsilon_{\mathrm{i}} = y(t_{\mathrm{i}}) - g(t_{\mathrm{i}}, \bar{\beta}_{\mathrm{min}}) = y_{\mathrm{i}} - g_{\mathrm{i}}$. Program a subroutine that collects data points t_i and y_i that have residuals ϵ_i that fulfill Eq. 1 with $A = 3$. Write these outliers into a separate file. Note that the same outlier may, or may not, be identified in several conseqtive models. Remove the outliers and model the remaining data like in the previous exercise. Note that your SEG and SET values may change after removing the outliers.

∼ • ∼ Exercise ends here. ∼ • ∼

Outliers II

- The relation of Eq. 1 does not always identify all outliers. Why?
- There are two solutions.
- **Solution 1:** Apply `EXERCISE8.PRO` to your data again and again, until no new outliers are identified.
- **Solution 2:**

  ⊙ 2A: Select the n_{old} values $y(t_i)$ of your **current dataset**.

  ⊙ 2B: Model a dataset to get the residuals $\epsilon_1, \epsilon_2, ...\epsilon_{n_{\mathrm{old}}}$.

  ⊙ 2C: Calculate the standard deviation σ_ϵ of these residuals.

  ⊙ 2D: Remove the outlier datapoints y_i that fulfill $|\epsilon_i/\sigma_\epsilon| > A$, where suitable values are $A \geq 2$.

  ⊙ 2E: If there are no outliers, select the next dataset as your **current dataset** and begin from 2A.

  ⊙ 2F: If there are outliers, remove those y_i values. The remaining data have $n_{\mathrm{new}} < n_{\mathrm{old}} - 1$ data points.

  ⊙ 2G: If $n_{\mathrm{new}} \geq n_{\mathrm{min}}$, go back to 2A and use these remaining n_{new} values as your new **current dataset**.

  ⊙ 2H: If $n_{\mathrm{new}} < n_{\mathrm{min}}$, select the next dataset as your **current dataset** and begin from 2A.

- **Exercise 9.** Identify all outliers in your data with $A = 3$ in Eq. 1. You can use **Solution 1** or **Solution 2.**

∼ • ∼ Exercise ends here. ∼ • ∼

- The aim of this last exercise: everybody gets the same values for
  – Datasets SET and segments SEG (i.e. Exercise 6 checked)
  – Order (K), Mean (M), Period (P), Amplitude (A) (i.e. Exercise 7 checked)
  – All outliers identified and removed. SEG and SET values are final (i.e. Exercises 8 and 9 checked).