Analysis methods of heavy-tailed data

Natalia Markovich

Institute of Control Sciences
Russian Academy of Sciences, Moscow, Russia

Exercises

Practical exercises to Moduls 1-4

Rough tests and estimation of heavy-tailed features: generators.

1. Generate 100 Fréchet distributed r.v.s with the distribution function

$$F(x) = \exp\left(-(\gamma x)^{-1/\gamma} 1\{x > 0\}\right)$$

and $\gamma = 1.5$.
To do it

- generate 100 uniformly distributed r.v.s U_i on $[0, 1]$;
- calculate 100 Fréchet distributed r.v.s X_i by formula

$$X_i = \frac{1}{\gamma} \left(-\ln U_i\right)^{-\gamma}.$$

Rough tests and estimation of heavy-tailed features: ratio of the maximum to the sum.

2. Calculate the following statistic

$$R_n(p) = \frac{M_n(p)}{S_n(p)}, \qquad n \geq 1, \qquad p > 0,$$

where

$$M_n(p) = \max\left(|X_1|^p, ..., |X_n|^p\right),$$

$$S_n(p) = |X_1|^p + ... + |X_n|^p$$

by the sample $X^n = X_1, ..., X_n$ for $n = 1, 2, ...$ (a sample X^n may be generated using some random generator or X^n is real data). Draw the plot of dependence $R_n(p)$ against n for different p.

Investigate this plot for the large n and make conclusions regarding the amount of finite moments $\mathbb{E}|X|^p$ of the distribution.

Rough tests and estimation of heavy-tailed features: QQ-plot or "quantiles against quantiles"-plot.

3. To construct a QQ-plot draw the dependence

$$\left\{\left(X_{(k)}, F^{\leftarrow}\left(\frac{n-k+1}{n+1}\right)\right) : k = 1, ..., n\right\},$$

where $X_{(1)} \geq ... \geq X_{(n)}$ are the order statistics of the sample $X^n = \{X_1, ..., X_n\}$ [1], and $F^{\leftarrow}$ is an inverse function of the distribution function F.

Check different alternatives of $F(x)$, e.g. normal, lognormal, exponential, the generalized Pareto distribution

$$\Psi_{\sigma,\gamma}(x) = \begin{cases} 1-(1+\gamma x/\sigma)^{-1/\gamma}, & \gamma \neq 0, \\ 1-\exp\left(-x/\sigma\right), & \gamma = 0, \end{cases} \quad (1)$$

where $\sigma > 0$ and $x \geq 0$, as $\gamma \geq 0$; $0 \leq x \leq -\sigma/\gamma$, as $\gamma < 0$.

If the QQ-plot is linear for some $F(x)$ then the underlying sample is distributed according to this $F(x)$.

[1] X^n is the real data or generated by a random generator.

QQ-plot: exclusion of outliers

3. Continuation. Exclude 10 largest observations (outliers) from the sample X^n and construct QQ-plot by the rest of points.
   Observe the correspondence of the obtained QQ-plot to the linear line.
   Repeat the exclusion of the next 10 largest observations (outliers) from the rest sample and construct a QQ-plot by the rest of points. Make conclusions regarding the influence of the outliers at the QQ-plot.

Rough tests and estimation of heavy-tailed features: plot of the mean excess function.

4. Having the empirical or generated data $X^n = \{X_1, ..., X_n\}$ calculate the empirical mean excess function by formula

$$e_n(u) = \sum_{i=1}^{n}(X_i - u)\mathbf{1}\{X_i > u\} / \sum_{i=1}^{n}\mathbf{1}\{X_i > u\}$$

   Investigate the behavior of $e_n(u)$ for the large u.
   For heavy-tailed distributions the function $e(u)$ tends to infinity. A linear plot $u \to e(u)$ corresponds to a Pareto distribution, the constant $1/\lambda$ corresponds to an exponential distribution and $e(u)$ tends to 0 for light-tailed distributions.

Rough tests and estimation of heavy-tailed features: estimation of the tail index.

5. Having the empirical or generated data $X^n = \{X_1, ..., X_n\}$ reorder the data as $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$.
   Calculate and compare the following estimates of the tail index of your data. Investigate the sign of an estimate and make conclusion regarding the heavy tails.
   - Hill's estimator

     $$\hat{\gamma}_{n,k} = \frac{1}{k} \sum_{i=1}^{k} \ln X_{(n-i+1)} - \ln X_{(n-k)} \tag{2}$$

     for some $k = 1, ..., n-1$.
   - Ratio estimator

     $$a_n = a_n(x_n) = \sum_{i=1}^{n} \ln(X_i/x_n)\mathbf{1}\{X_i > x_n\} / \sum_{i=1}^{n} \mathbf{1}\{X_i > x_n\}$$

     for some $X_{(1)} < x_n < X_{(n)}$.

Rough tests and estimation of heavy-tailed features: estimation of the tail index.

5. Continuation.
   - Moment estimator

$$\hat{\gamma}^{M}_{n,k} = \hat{\gamma}^{H}(n,k) + 1 - 0.5\left(1 - (\hat{\gamma}^{H}(n,k))^2/S_{n,k}\right)^{-1},$$

     where $S_{n,k} = (1/k)\sum_{i=1}^{k}\left(\log X_{(n-i+1)} - \log X_{(n-k)}\right)^2$.
   - *UH* estimator

$$\hat{\gamma}^{UH}_{n,k} = (1/k)\sum_{i=1}^{k}\log UH_i - \log UH_{k+1}, \qquad (3)$$

     where $UH_i = X_{(n-i)}\hat{\gamma}^{H}(n,i)$
   - Pickands's estimator

$$\hat{\gamma}^{P}_{k,n} = \frac{1}{\log 2}\log\frac{X_{(n-k+1)} - X_{(n-2k+1)}}{X_{(n-2k+1)} - X_{(n-4k+1)}}$$

     for some $k \leq n/4$.

Rough tests and estimation of heavy-tailed features: the choice of parameter k of the Hill's estimator by a Hill-plot.

6. Having a sample $X^n = \{X_1, ..., X_n\}$ calculate the Hill's estimate (**??**).
   Draw the dependence $\{(k, \hat{\gamma}_{n,k}), 1 \leq k \leq n-1\}$) and then choose the estimate of $\hat{\gamma}_{n,k}$ from an interval in which these functions demonstrate stability.
   Make conclusions regarding the amount of finite moments[2] of the underlying distribution and the existence of heavy tails.[3]

[2] For light-tailed distributions all moments $\mathbb{E}[(X^+)^k]$ exist and finite as far as for regularly varying distributions (i.e., such that $1 - F(x) = \mathbb{P}\{X > x\} = x^{-1/\gamma}\ell(x)$, $\forall x > 0$, where ℓ is called slowly varying function) the moments $\mathbb{E}X^\beta$ are finite only, as $\beta < 1/\gamma$.

[3] The positive estimate $\hat{\gamma}_{n,k}$ may indicate on a heavy tail existence.

Hill-plot and its bootstrap confidence interval

6. Continuation. Having the Hill's estimates $\gamma_1^*, ..., \gamma_B^*$ of γ obtained by B bootstrap re-samples, construct the tolerant confidence interval of the Hill's estimate by formula

$$(u_1, u_2) = (Mean\gamma - \rho \cdot StDev\gamma; Mean\gamma + \rho \cdot StDev\gamma),$$

where the mean $Mean\gamma$ and standard deviation $StDev\gamma$ are calculated by $\gamma_1^*, ..., \gamma_B^*$.
The interval is constructed in such a way that the $(1-p)$th part of the distribution falls into this interval with the probability P:

$$\rho = \rho_\infty \left(1 + \frac{t_p}{\sqrt{2B}} + \frac{5t_p^2 + 10}{12B}\right).$$

Hill-plot and its bootstrap confidence interval

6. Continuation. ρ_∞ is defined by the equation

$$\frac{1}{\sqrt{2\pi}}\int_{-\rho_\infty}^{\rho_\infty} e^{-t^2/2}dt = 2\Phi_0(\rho_\infty) = 1-p,$$

where $\Phi_0(z) = \frac{1}{\sqrt{2\pi}}\int_0^z e^{-t^2/2}dt$.
Normal distribution function $N(z;0,1) = 0.5 + \Phi_0(z)$ for $z > 0$.
t_p is calculated by the equation

$$\frac{1}{\sqrt{2\pi}}\int_{t_p}^{\infty} e^{-t^2/2}dt = 0.5 - \Phi_0(t_p) = 1 - P.0.$$

Select $P \in \{0.75, 0.95, 0.99\}$ and $p \in \{0.025, 0.05, 0.1\}$. Draw the Hill-plot with 75%, 95%, 99% confidence intervals. Select $B \in \{100, 200, 500\}$.

Rough tests and estimation of heavy-tailed features: the investigation of the Hill's estimator.

7. Generate several samples distributed by the regularly varying distributions

$$1 - F(x) = \mathbb{P}\{X > x\} = x^{-1/\gamma}\ell(x)$$

were $\ell(x) = 1$, $\ell(x) = 2$ and $\gamma = 0.5$; and Weibull distribution

$$1 - F(x) = \exp\left(-cx^{1/\gamma}\right), c = 1, \gamma = 2; c = 2, \gamma = 3$$

Calculate the Hill's estimate (**??**) and investigate the influence of a slowly varying function $\ell(x)$ on the estimate. Compare the true values of the EVI γ with results of estimation for different distributions.[4]

[4] The estimation in the case of the Weibull distribution should be worse.

Bias-reduced Jackknife estimator

8. Having a sample $X^n = \{X_1, ..., X_n\}$ calculate the Jackknife estimator

$$\widehat{\gamma}_k^{GJ} = 2\widehat{\gamma}_k^V - \widehat{\gamma}_{n,k},$$

where $\widehat{\gamma}_{n,k}$ is Hill's estimator,

$$\widehat{\gamma}_k^V = \frac{M_{n,k}}{2\widehat{\gamma}_{n,k}}, \qquad M_{n,k} = \frac{1}{k}\sum_{i=1}^{k} Y_{(i,k)}^2,$$

$$Y_{(i,k)} = \log(\frac{X_{(n-i+1)}}{X_{(n-k)}}).$$

Draw the plot $(k, \widehat{\gamma}_k^{GJ})$ and observe its stability in comparison with Hill's plot.

8. Continuation. Calculate k for $\widehat{\gamma}_k^{GJ}$ by formulae[5]

$$\widehat{k}_{SAMSEE} = \arg \min_{1<k<K^*} SAMSEE(k),$$

where

$$SAMSEE(k) = \frac{(\widehat{\gamma}_{K^*}^{GJ})^2}{k} + 4\widehat{b}_{k,K^*}^2, \qquad \widehat{b}_{k,K} = \overline{\gamma}_{k,K} - \overline{\gamma}_K$$

$$\overline{\gamma}_{k,K} = \frac{1}{K-k+1}\sum_{i=k}^{K}\widehat{\gamma}_{n,i}, \qquad \overline{\gamma}_k = \overline{\gamma}_{1,k} = \frac{1}{k}\sum_{i=1}^{k}\widehat{\gamma}_{n,i}$$

Take $K^* = 400$ or select it as follows...

[5] Schneider, Krajina, Krivobokova 2021

8. Continuation. Calculate

$$AD(K) = \frac{1}{K}\sum_{k=1}^{K}\left(\widehat{\gamma}_k^V + \widehat{b}_{k,K} - \widehat{\gamma}_{n,k}\right)^2.$$

Find K such that provides the stabilized numerical approximation of the derivative of AD:

$$K^* = \arg\min_K \left\{ \sum_{i=-2, i\neq 0}^{2} \left|\frac{AD(K) - AD(K+i)}{i}\right| \right\}.$$

Calculate Hill's estimate $\widehat{\gamma}_{n,K^*}$ and draw plots $(SAMSEE(k), k)$ and $(\widehat{\gamma}_{n,k}, k)$ for $1 \le k \le K^*$.

9. Having a sample $X^n = \{X_1, ..., X_n\}$ divide it into l groups $V_1, ..., V_l$, each group containing m r.v.s, i.e. $n = l \cdot m$. Calculate the Group estimate

$$z_l = (1/l)\sum_{i=1}^{l} k_{li} = \frac{\hat{\alpha}}{\hat{\alpha}+1} = \frac{1}{1+\hat{\gamma}_l} \qquad \Rightarrow \hat{\gamma}_l = 1/z_l - 1,$$

where

$$k_{li} = M_{li}^{(2)}/M_{li}^{(1)}, \qquad M_{li}^{(1)} = \max\{X_j : X_j \in V_i\}$$

and $M_{li}^{(2)}$ is the second largest element in the same group V_i.

Draw the plot $\{(m, 1/z_m - 1)\}$, where $m = 10, 11, ...,$ together with the confidence interval. Take $n \in \{150, 500, 1000\}$.

9. Continuation. We have

$$l\left(l^{-1}\sum_{i=1}^{l}k_{li}-(1+\gamma)^{-1}\right)\left(\sum_{i=1}^{l}k_{li}^2-l^{-1}\left(\sum_{i=1}^{l}k_{li}\right)^2\right)^{-1/2}\to^d N(0,1)$$

$\mathbb{P}\{-z\le Z\le z\}=1-\alpha=0.95$,
Gaussian DF $\Phi(z)=\mathbb{P}\{Z\le z\}=1-\alpha/2=0.975$,
$z=\Phi^{-1}(0.975)=1.96$
Calculate the 95%-confidence interval of the Group estimate for each m by formula $\widehat{\gamma}\in(\gamma_1,\gamma_2)$,

$$\gamma_{1,2}=\left(\overline{k}-\frac{\pm 1.96\sqrt{A_l}}{l}\right)^{-1}-1,$$

where $\overline{k}=(1/l)\sum_{i=1}^{l}k_{li}$,
$A_l=\sum_{i=1}^{l}k_{li}^2-(1/l)\left(\sum_{i=1}^{l}k_{li}\right)^2.$

Estimation of the heavy-tailed density function: kernel estimates.

10. Generate X^n according to some heavy-tailed distribution or take a heavy-tailed real data. Calculate the Hill's estimate (**??**) of the EVI γ.

For heavy-tailed data transform the sample X^n to a new one Y^n by the transformations $T(x) = \ln x$, $T(x) = (2/\pi)\arctan x$ and $T(x) = 1 - (1 + \hat{\gamma}x)^{-1/(2\hat{\gamma})}$ ($Y_i = T(X_i)$, i=1,...,n).
Calculate the kernel estimate

$$\hat{g}_h(x) \quad = \quad \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - Y_i}{h}\right). \tag{4}$$

Take $h = \sigma n^{-1/5}$, where σ^2 is an empirical variance calculated by a sample Y^n and $K(x) = (3/4)(1 - x^2)1\{|x| \leq 1\}$.

10. Continuation.
    Calculate the density of the initial r.v. X_1 by formula

$$\hat{f}_h(x) = \hat{g}_h(T(x))T'(x). \tag{5}$$

For generated data, compare the estimates for different transformations and the true density.

Estimation of the heavy-tailed density function: kernel estimates, comparison of smoothing methods.

11. Generate X^n according to some heavy-tailed distribution or take a heavy-tailed real data.
    Transform the sample X^n to Y^n by the adapted transformation

$$T_{\hat{\gamma}}(x) \quad = \quad 1 - (1 + \hat{\gamma} x)^{-1/(2\hat{\gamma})}. \qquad (6)$$

    Using a sample Y^n calculate a kernel estimate $\hat{g}_h(x)$ by (**??**) and then $\hat{f}_h(x)$ by (**??**).

Find h in (**??**) as a solution of discrepancy equations

$$\sum_{i=1}^{n} \left(\widehat{F}_h(Y_{(i)}) - \frac{i - 0.5}{n} \right)^2 + \frac{1}{12n} = 0.05, \qquad \omega^2 - \text{method},$$

where $\widehat{F}_h(x) = \int_0^x \hat{f}_h(t)dt$,

11. Continuation.

$$\sqrt{n}\hat{D}_n = \sqrt{n}\max(\hat{D}_n^+, \hat{D}_n^-) = 0.5, \qquad D - \text{method},$$

where

$$\sqrt{n}\hat{D}_n^+ = \sqrt{n}\max_{1\le i\le n}\left(\frac{i}{n} - \widehat{F}_h(Y_{(i)})\right),$$

$$\sqrt{n}\hat{D}_n^- = \sqrt{n}\max_{1\le i\le n}\left(\widehat{F}_h(Y_{(i)}) - \frac{i-1}{n}\right),$$

$Y_{(1)} \le Y_{(2)} \le \ldots Y_{(n)}$ are order statistics.

For generated data, compare D-, ω^2- methods and $h = \sigma n^{-1/5}$.

21. Generate the process $MA(q)$:

$$X_t = \sum_{j=0}^{q} \psi_j Z_{t-j}, \qquad t \in \{0, 1, ..., n\}$$

$\{Z_t\}$ are i.i.d. Fréchet distributed r.v.s with $\gamma \in \{0.3, 1, 1.5, 2.5\}$
Take $n = 1000$, $q = 10$, $\{\psi_j = 1/2^j\}$ and $\{\psi_j \equiv 1\}$
Construct the standard sample ACF at lag $h \in Z$ by formula

$$\rho_{n,X}(h) = \frac{\sum_{t=1}^{n-h}(X_t - \overline{X}_n)(X_{t+h} - \overline{X}_n)}{\sum_{t=1}^{n}(X_t - \overline{X}_n)^2},$$

where $\overline{X}_n = \frac{1}{n}\sum_{t=1}^{n} X_t$ represents the sample mean.
Draw the plot $\rho_{n,X}(h)$ versus h and Bartlett's confidence interval $\pm 1.96/\sqrt{n}$.

22. For generated data of the process $MA(q)$ estimate the Hurst parameter by
Kettani& Gübner's method:

$$\hat{H}_n = 0.5\left(1 + \log_2(1 + \rho_{n,X}(1))\right)$$

Make conclusion regarding the long-range dependence.

23. by Aggregated Variance Method:
Let $\{X_i, i = 1, 2, ..., n\}$ be the original time series. Calculate averages within each block of $\{X_i\}$ with number $k = 1, 2, ..., [n/m]$ of size m

$$X^{(m)}(k) = \frac{1}{m} \sum_{i=(k-1)m+1}^{km} X_i,$$

and, the sample variance of $X^{(m)}(k)$

$$\widehat{Var} X^{(m)} = \frac{m}{n} \sum_{k=1}^{n/m} \left(X^{(m)}(k)\right)^2 - \left(\frac{m}{n} \sum_{k=1}^{n/m} X^{(m)}(k)\right)^2.$$

Plot $\log \widehat{Var} X^{(m)}$ versus $\log m$. The line approximating the points has the slope $\beta = 2H - 2$, $-1 \leq \beta < 0$

24. by Ljung-Box test
For generated data of MA(q) process with Fréchet distributed r.v.s $\{Z_t\}$ and $\gamma = 0.3$ calculate statistic

$$Q_h = n(n+2)\sum_{j=1}^{h}\frac{\rho_{n,x}^2(j)}{n-j},$$

where $\rho_{n,x}(j)$ is sample ACF at lag j
Check the inequality $Q_h > \chi^2_\eta(h)$ for $h \in \{10, 20, 30\}$
If the inequality is valid than independence should be rejected.
$\chi^2_\eta(h)$ is η-quantile of χ^2 distribution with h degrees of freedom, i.e. $\mathbb{P}\{\chi^2 > \chi^2_\eta(h)\} = \eta$, $\eta = 0.05$
(see quantiles in tables of χ^2 distribution)

24. Continuation

Table: Ljung-Box test: critical points

Lags, h	$\chi^2_{0.05}(h)$
10	18.3
20	31.4
30	43.8

25. by Runde's test
For generated data of MA(q) process with Fréchet distributed r.v.s $\{Z_t\}$ and $\gamma \in \{1, 1.5, 2.5\}$ calculate statistic

$$Q_R = \left(\frac{n}{\ln n}\right)^{2\gamma} \sum_{j=1}^{h} \rho_{n,X}^2(j),$$

where $\rho_{n,X}(j)$ is sample ACF at lag j
Check the inequality

$$Q_R > Q_h(0.05)$$

for $h \in \{2, 3, 4, 5\}$
If the inequality is valid than independence should be rejected.

25. Continuation

Table: Runde's test: critical points

Lags	$Q_h(0.05)$
2	13.53
3	16.32
4	18.28
5	19.17

26. Generate Fréchet distributed rvs $\{X_i\}$ and lognormal distributed rvs $\{Y_i\}$ (or rvs $\{Y_i = 2 \cdot X_i\}$).
Partition $X_1, X_2, ..., X_n$ and $Y_1, Y_2, ..., Y_n$, $n = 10000$ into r blocks of equal size $m \in \{20, 50, 100\}$, $r = [n/m]$

Calculate block-maxima

$\{X_1^*, ..., X_r^*\}, \qquad \{Y_1^*, ..., Y_r^*\}.$

Estimate Pickands A-function by

Hall and Tajvidi (2000):

$$\widehat{A}_r^{HT}(t) = \left((1/r)\sum_{i=1}^{r}\min\left(\frac{\hat{\xi}_i/\overline{\xi}_r}{1-t}, \frac{\hat{\eta}_i/\overline{\eta}_r}{t}\right)\right)^{-1},$$

26. Continuation.

Estimate Pickands A-function by Capéraà et al. (1997):

$$\begin{aligned}\log \widehat{A}_r^C(t) &= \frac{1}{r}\sum_{i=1}^{r}\log\max\left(t\hat{\xi}_i,(1-t)\hat{\eta}_i\right)\\ &- t\frac{1}{r}\sum_{i=1}^{r}\log\hat{\xi}_i-(1-t)\frac{1}{r}\sum_{i=1}^{r}\log\hat{\eta}_i.\end{aligned}$$

Here

$\hat{\xi}_i = -\log\widehat{G}_1(X_i^*)$ and $\hat{\eta}_i = -\log\widehat{G}_2(Y_i^*)$, $i = 1, ..., r$,
$\overline{\xi}_r = r^{-1}\sum_{i=1}^{r}\hat{\xi}_i$, $\qquad \overline{\eta}_r = r^{-1}\sum_{i=1}^{r}\hat{\eta}_i$.

26. Continuation.

Distribution functions (dfs) estimate by empirical dfs

$$\widehat{G}_1(x) = 1/r \sum_{i=1}^{r} \theta(x - X_i^*), \qquad \widehat{G}_2(y) = 1/r \sum_{i=1}^{r} \theta(y - Y_i^*),$$

where $\theta(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0. \end{cases}$

Draw a plot of A-function against t within a triangle determined by points $(0,1)$, $(1,1)$ and $(0.5, 0.5)$ Conclude regarding the dependence of rvs X_1 and Y_1.

- $A(t) \equiv 1$ corresponds to a total independence
- $A(t) = (1-t) \vee t$ corresponds to a total dependence
- $A(t)$ located inside the triangle corresponds to some kind of dependence

Density estimation

27. Generate 100 Normal, lognormal and Fréchet distributed r.v.s $X^n = \{X_1, X_2, ..., X_n\}$

Estimate both densities by

- the kernel estimator $f_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)$
  with Epanechnikov's kernel $K(x) = \frac{3}{4}(1-x^2)\mathbf{1}\{|x| \leq 1\}$
  Take $h \in \{0.05, 0.1, 0.5, 1\}$
- by polygram (histogram with variable bin width)

$$f_{L,n}(t) = \frac{L}{(n+1)\lambda(\Delta_{rL})}, \qquad t \in \Delta_{rL}$$

We set $\Delta_{1L} = [x_{(1)}, x_{(L)}]$, $\Delta_{2L} = (x_{(L)}, x_{(2L)}]$, $\Delta_{3L} = (x_{(2L)}, x_{(3L)}], \ldots$
$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ are order statistics of the sample X^n
$\lambda(\Delta)$ is the length of Δ. Take $L \in \{2, 5, 10\}$. Compare results.

Generating log-normally distributed r.v.s:
If $X \sim N(\mu, \sigma^2)$ then $\exp(X) \sim Log - N(\mu, \sigma^2)$

30.

- Generate B re-samples with replacement from the original data set $X^n = \{X_1, ..., X_n\}$. This can be done by uniform random consecutive selection of any X_i and returning it back to X^n.
- The size of re-samples $\{X_1^*, ..., X_{n_1}^*\}$ is smaller than n
  $n_1 = n^\beta, \qquad 0 < \beta < 1,$
- The corresponding smaller k_1 and an optimal k are related by:
  $k = k_1(n/n_1)^\alpha, \qquad 0 < \alpha < 1,$

  where $\beta = 1/2$ and $\alpha = 2/3$.

  Such k provides the minimum of $MSE(\hat{\gamma})$.

30. (Continuation)
Empirical bootstrap estimate of the $MSE(\hat{\gamma})$ is

$$MSE^*(n_1, k_1) = \left(\hat{b}^*(n_1, k_1)\right)^2 + \widehat{var}^*(n_1, k_1) \to \min_{k_1},$$

where

$$\hat{b}^*(n_1, k_1) = \frac{1}{B}\sum_{b=1}^{B} \hat{\gamma}_b^*(n_1, k_1) - \hat{\gamma}(n, k),$$

$$\widehat{var}^*(n_1, k_1) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\gamma}_b^*(n_1, k_1) - \frac{1}{B}\sum_{b=1}^{B}\hat{\gamma}_b^*(n_1, k_1)\right)^2$$

are the **empirical bootstrap estimates of the bias and the variance,**

$\hat{\gamma}_b^*$ is the Hill's estimate constructed by some re-sample of the size n_1 with the parameter k_1.