Asymptotic behavior of some robust estimators under long-range dependence.

Hélène Boistard, Université Toulouse 1, France
Joint work with
C. Lévy-Leduc, E. Moulines, M. Taqqu, V. Reisen

CLAPEM 2012, Viña del Mar, Chile

helene@boistard.fr - www.boistard.fr

Available for download at : www.boistard.fr/boistard-clapem2012.pdf

A robust estimator of the autocovariance

A robust estimator of the scale

$X_1, \ldots, X_n$ r.v.'s having a common c.d.f. F and p.d.f. f.
Robust scale estimator introduced in [Rousseeuw and Croux, 1993]:

$$Q_n^{\mathrm{RC}}(\{X_1, \ldots, X_n\}) = c\{|X_i - X_j|;\ i < j\}_{(k)} ,$$

where c is a fixed constant which depends on the shape of the distribution F and $k \approx \binom{n}{2}/4$.
Good robustness properties of $Q_n^{\mathrm{RC}}(\{X_1, \ldots, X_n\})$ (see [Rousseeuw and Croux, 1993]):

- the highest possible breakdown point (50%)
- bounded influence function
- simple and explicit formula, easily and efficiently implementable

Tools: *U*-statistics and Delta-method

In terms of a U-statistic:

$$Q_n^{\mathrm{RC}}(\{X_1, \dots, X_n\}) = c U_n^{-1}(1/4) ,$$

where U_n^{-1} is the generalized inverse of

$$\begin{aligned} r \mapsto U_n(r) &= \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} \mathbf{1}_{\{|X_i - X_j| \le r\}} \\ &= \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} \mathbf{1}_{\{G(X_i, X_j) \le r\}}, \end{aligned}$$

where $G(x, y) = |x - y|$.
Strategy: limit theorems for U_n, then functional Delta-method.

Polarization identity: for any r.v.'s X and Y and nonzero a and b,

$$\mathrm{Cov}(X, Y) = \frac{1}{4ab} \left\{ \mathrm{Var}(aX + bY) - \mathrm{Var}(aX - bY) \right\} \ .$$

Robust autocovariance estimator (see [Ma and Genton, 2000])

$$\begin{aligned} 4\hat{\gamma}_Q(h) &= \left(Q^{\mathrm{RC}}_{n-h}(\{X_1 + X_{1+h}, \ldots, X_{n-h} + X_n\}) \right)^2 \\ &\quad - \left(Q^{\mathrm{RC}}_{n-h}(\{X_1 - X_{1+h}, \ldots, X_{n-h} - X_n\}) \right)^2 \end{aligned}$$

in order to estimate

$$4\gamma(h) = 4\mathrm{Cov}(X_1, X_{1+h}).$$

Other examples of estimators

Hodges-Lehman location estimator : $Y_i = \theta + X_i$, $i = 1, \ldots, n$, where X_i is a centered stationary process and θ is a location parameter.

$$\begin{aligned} \hat{\theta}_{HL} &= \text{median}\left\{\frac{Y_i + Y_j}{2}, 1 \leq i < j \leq n\right\} \\ &= \theta + \text{median}\left\{\frac{X_i + X_j}{2}, 1 \leq i < j \leq n\right\} = \theta + U_n^{-1}\left(\frac{1}{2}\right) \end{aligned}$$

with $G(x, y) = \frac{x+y}{2}$.

Shamos-Bickel scale estimator : $Y_i = \sigma X_i$.

$$\hat{\sigma}_{SB} = b \,\text{median}\{|Y_i - Y_j|\} = b\sigma \,\text{median}\{|X_i - X_j|\}.$$

Similar expression with $G(x, y) = |x - y|$.

Limit distribution under long-range dependence

$$U_n(r) = \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} \mathbf{1}_{\{|X_i - X_j| \le r\}}$$

for $r \in I$ where $(X_i)_{i \ge 1}$ is a stationary mean-zero Gaussian process with covariances $\rho(k) = \mathbb{E}(X_1 X_{k+1})$ satisfying:

$$\rho(0) = 1 \text{ and } \rho(k) = k^{-D} L(k),\ 0 < D < 1\ ,$$

where L is slowly varying at infinity and is positive for large k.

Limit distribution under long-range dependence with $D > 1/2$

Hoeffding decomposition of the U-statistic $U_n(r)$ for $r \in I$: define

$$h(x, y, r) = \mathbf{1}_{\{|x-y| \leq r\}}, \; h_1(x, r) = \int h(x, y, r)\varphi(y)dy$$

$$U(r) = \int \int h(x, y, r)\varphi(x)\varphi(y)dxdy,$$

then

$$U_n(r) = U(r) + W_n(r) + R_n(r)$$

where

- $W_n(r) = \frac{2}{n} \sum_{i=1}^{n} \{h_1(X_i, r) - U(r)\}$ is the leading term (dealt with using [Arcones, 1994]),
- $R_n(r) = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \{h(X_i, X_j, r) - h_1(X_i, r) - h_1(X_j, r) + U(r)\}$ is a negligible term satisfying: $\sup_{r \in I} \sqrt{n} R_n(r) = o_P(1)$.

Theorem
(i) Scale estimator. Denote by $\sigma(F)$ the standard deviation of F.

$$\sqrt{n}(Q_n^{RC} - \sigma(F)) \xrightarrow{d} \mathcal{N}(0, \tilde{\sigma}^2) \ ,$$

where $\tilde{\sigma}^2 = \mathbb{E}[\mathrm{IF}(X_1, F)^2] + 2\sum_{k\geq 1} \mathbb{E}[\mathrm{IF}(X_1, F)\mathrm{IF}(X_{k+1}, F)]$ and

$$\mathrm{IF} : (x, F) \mapsto c\left(\frac{1/4 - F(x + \sigma(F)/c) + F(x - \sigma(F)/c)}{\int f(y)f(y + \sigma(F)/c)dy}\right) \ .$$

(ii) Covariance estimator. Let h be a non negative integer and denote by F_+ the common c.d.f of $(X_i + X_{i+h})_{i\geq 1}$, by F_- the common c.d.f of $(X_i - X_{i+h})_{i\geq 1}$ and by $\sigma(F_+)$, $\sigma(F_-)$ the respective standard deviations. Let $\gamma(h) = \mathbb{E}[X_1 X_{1+h}]$,

$$\psi : (x, y) \mapsto \frac{1}{2}\left\{\sigma(F_+)\mathrm{IF}(x + y, F_+) - \sigma(F_-)\mathrm{IF}(x - y, F_-)\right\}$$

and
$\check{\sigma}_h^2 = \mathbb{E}[\psi(X_1, X_{1+h})^2] + 2\sum_{k\geq 1}\mathbb{E}[\psi(X_1, X_{1+h})\psi(X_{k+1}, X_{k+1+h})]$, then,

$$\sqrt{n}\,(\hat{\gamma}_Q(h) - \gamma(h)) \xrightarrow{d} \mathcal{N}(0, \check{\sigma}_h^2)\ . \qquad (1)$$

Limit distribution under long-range dependence with $D < 1/2$

Decomposition based on the expansion of h on the basis of Hermite polynomials (here, the Hermite rank of h is $m = 2$):

$$h(x, y, r) - U(r) = \sum_{\substack{p,q \geq 0 \\ p+q \geq m}} \frac{\alpha_{p,q}(r)}{p!q!} H_p(x) H_q(y).$$

Decomposition of U_n:

$$n(n-1)\left(U_n(r) - U(r)\right) = \tilde{W}_n(r) + \tilde{R}_n(r), \text{ where}$$

- $\tilde{W}_n(r) = \sum_{1 \leq i \neq j \leq n} \sum_{\substack{p,q \geq 0 \\ p+q \leq m}} \frac{\alpha_{p,q}(r)}{p!q!} H_p(X_i) H_q(X_j)$ is the leading term (argument based on the identification of cumulants),
- $\tilde{R}_n(r) = \sum_{1 \leq i \neq j \leq n} \sum_{\substack{p,q \geq 0 \\ p+q > m}} \frac{\alpha_{p,q}(r)}{p!q!} H_p(X_i) H_q(X_j)$ is a negligible term satisfying $\sup_{r \in I} n^{mD/2-2} L(n)^{-m/2} \tilde{R}_n(r) = o_P(1)$, $n \to \infty$.

Theorem
(i) Scale estimator.

$$\beta(D)\frac{n^D}{L(n)}(Q_n^{RC} - \sigma(F)) \xrightarrow{d} \frac{\sigma(F)}{2}\left(Z_2(1) - Z_1(1)^2\right) ,$$

where $\beta(D) = \mathbf{B}((1-D)/2, D+2)$, $\mathbf{B}$ denoting the Beta function and the processes $Z_1(.)$ and $Z_2(.)$ being respectively a fractional Brownian motion and a Rosenblatt process.
(ii) Covariance estimator. Under some regularity conditions on L,

$$\beta(D)\frac{n^D}{\tilde{L}(n)}\left(\hat{\gamma}_Q(h) - \gamma(h)\right) \xrightarrow{d} \frac{\sigma(F_+)^2}{4}\left(Z_2(1) - Z_1(1)^2\right) ,$$

where $\tilde{L}(n) = 2L(n) + L(n+h)(1+h/n)^{-D} + L(n-h)(1-h/n)^{-D}$.

Efficiency of the scale estimator

- For $D > 1/2$, the efficiency of Q_n^{RC} with respect to the classical scale estimator (standard deviation) is greater than 86.31%.
- For $D < 1/2$, the efficiency is 1: there is no loss of efficiency.

Some numerical results

Monte Carlo studies

- Generate a process $Y_i, i = 1 \ldots n$ with distribution ARFIMA $(0, d, 0)$ (with $D = 1 - 2d$)
- Generate i.i.d. $W_i, i = 1 \ldots n$ with $P(W_i = -1) = P(W_i = 1) = p/2$, $P(W_i = 0) = 1 - p$ for
  some *contamination proportion* p
- Define $X_i = Y_i + \omega W_i$ for some *contamination magnitude* ω
- Compute Q_n^{RC} and the classical estimator denoted by *sd* for original/contaminated data
- Here, $d = 0.2$, $n = 500$, $p = 10\%$ and $\omega = 10$.

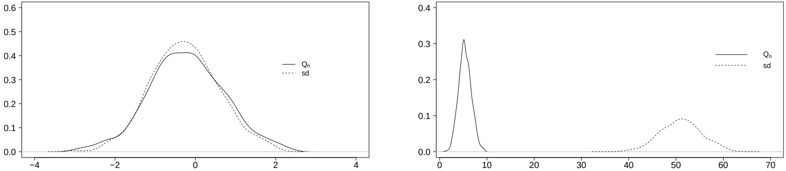

Figure: Empirical densities of $\sqrt{n}(Q_n^{\mathrm{RC}} - \sigma(F))$ and $\sqrt{n}(sd - \sigma(F))$, without outliers (left) and with outliers (right)

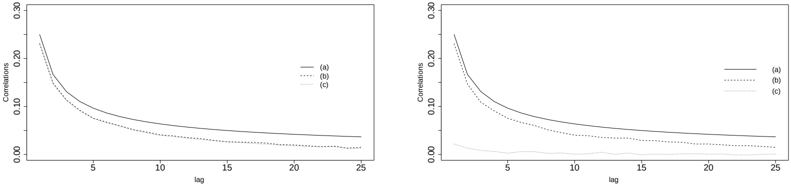

Figure: Sample correlations without outliers (left) and with outliers (right). (a) is the population correlation, (b) the robust sample correlation, (c) the classical sample correlation.

Real data: the Nile data plot

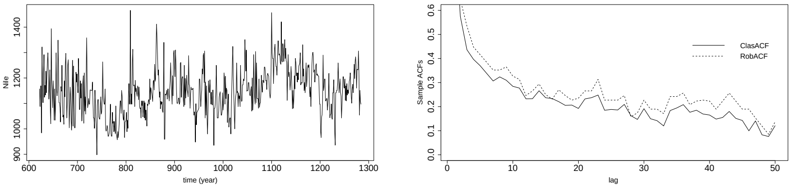

Figure: Left: the Nile Data plot. Right: sample autocorrelation functions of the Nile River data.

Remark: possible loss of memory due to the presence of outliers.

Some references

- Rousseeuw, P., Croux, C.: Alternatives to the median absolute deviation. J. Amer. Statist. Assoc., 88, no. 424, 1273–1283, (1993).
- Ma, Y., Genton, M.: Highly robust estimation of the autocovariance function. J. Time Ser. Anal., 21, no. 6, 663–684, (2000).
- Lévy-Leduc, C., Boistard, H., Moulines, E., Taqqu, M., Reisen, V. : Robust estimation of the scale and of the autocovariance function of Gaussian short and long-range dependent processes. J. Time Ser. Anal., 32, no. 2, 135–156 (2011).
- LL, B, M, T, V : Large sample behavior of some well-known robust estimators under long-range dependence. Statistics, 45, no. 1, 59–71 (2011).
- LL, B, M, T, V : Asymptotic properties of U-processes under long-range dependence. Ann. Stat., 39, no. 3, 1399–1426 (2011).