ANOVEX: ANalysis Of Variability for heavy-tailed EXtremes

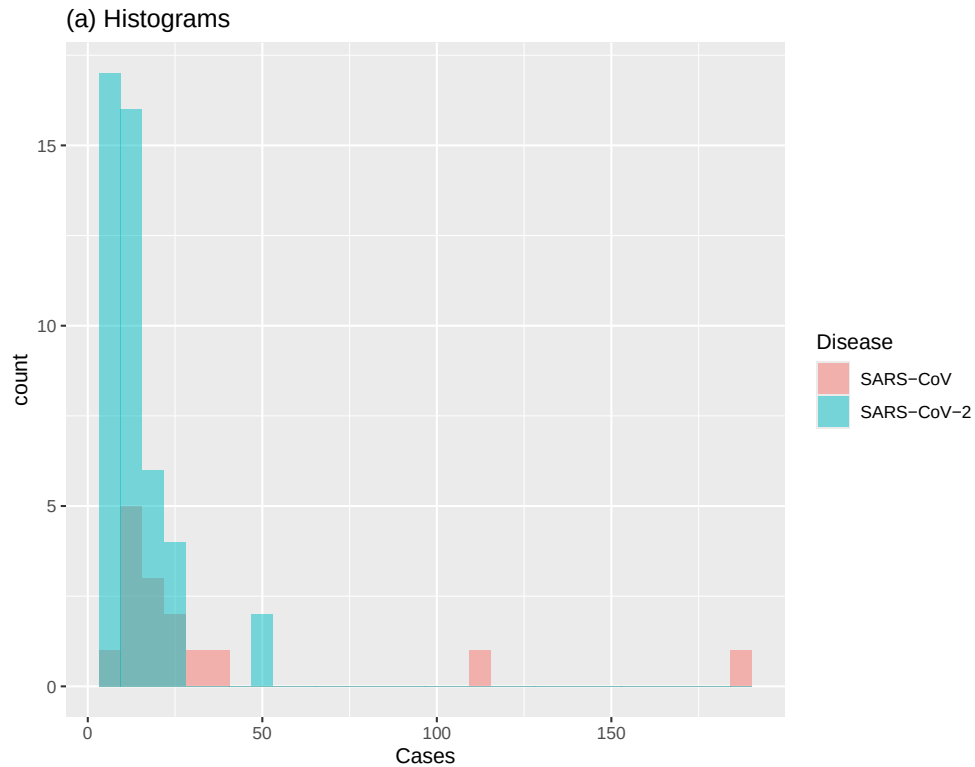
Antoine Usseglio-Carleve[1]

[1]Avignon Université, Laboratoire de Mathématiques d'Avignon – LMA UPR 2151
antoine.usseglio-carleve@univ-avignon.fr

Joint work with Stéphane Girard and Thomas Opitz

Motivation

In practice, it is frequent to ask whether two or more groups of individuals come from the same population (for data pooling issues or the study of the impact of a categorical variable for instance).



A widespread tool is the classical ANOVA.

From ANOVA...

Let us consider $J > 1$ samples $E_j = \{X_i^{(j)},\ i = 1, \ldots, n_j\}$, $j = 1, \ldots, J$, with independence between samples and possibly different sample sizes $n_j > 1$ (such that $n_j = O(n)$, for all j). We assume that the random variables in each E_j are identically distributed according to a cumulative distribution function F_j, with mean $\mu_j = \int_{\mathbb{R}} (1 - F_j(t))\, dt$.

The classical ANalysis Of VAriance (ANOVA) aims at testing the equality of the means $\mu_1, \ldots, \mu_J$, based on the decomposition:

$$\underbrace{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(X_i^{(j)} - \overline{X}\right)^2}_{\text{total var.}} = \underbrace{\sum_{j=1}^{J} n_j \left(\overline{X^{(j)}} - \overline{X}\right)^2}_{\text{var. between samples}} + \underbrace{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(X_i^{(j)} - \overline{X^{(j)}}\right)^2}_{\text{var. within samples}}.$$

From ANOVA...

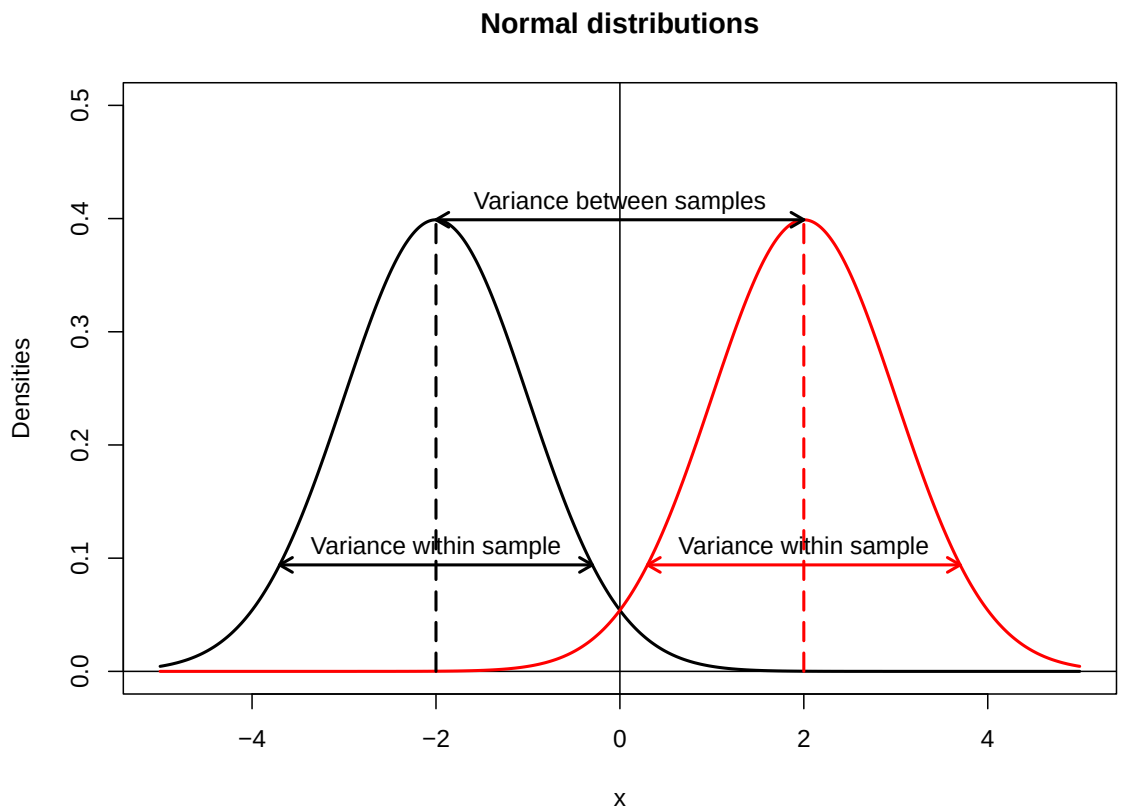


Figure 1: ANalysis Of VAriance (ANOVA) for two normal distributions.

... to ANOVEX

Instead of testing the equality of the means (i.e. the body) of the samples, we propose to test the equality of the tails:

(H_0) For all $(j, j') \in \{1, \dots, J\}^2$ with $j \neq j'$, we have $q_{j'}(\alpha)/q_j(\alpha) \to 1$ as $\alpha \to 1$.

For that purpose, we consider $L \in \mathbb{N}^\star$ extreme quantile levels $\alpha_{1,n}, \dots, \alpha_{L,n}$ (i.e. such that $n(1-\alpha_{\ell,n}) = O(1)$ for all $\ell \in \{1, \dots, L\}$). The key will be the decomposition of the term:

$$\Delta_n = \underbrace{\frac{1}{JL}\sum_{j=1}^{J}\sum_{\ell=1}^{L}\left(\log \widetilde{q}_j(\alpha_{\ell,n}) - \mu_{\alpha,n}\right)^2}_{\text{total var.}}, \text{ with } \mu_{\alpha,n} = \frac{1}{JL}\sum_{j=1}^{J}\sum_{\ell=1}^{L}\log \widetilde{q}_j(\alpha_{\ell,n})$$

and $\widetilde{q}_j(\alpha_{\ell,n})$ is an estimator of $q_j(\alpha_{\ell,n})$.

Note that the existence of μ_j is not necessary !

... to ANOVEX

Proposition 1

The decomposition $\Delta_n = \Delta_{1,n} + \Delta_{2,n}$ *holds where*

$$\Delta_{1,n} = \underbrace{\frac{1}{JL}\sum_{\ell=1}^{L}\sum_{j=1}^{J}\left(\log \widetilde{q}_j(\alpha_{\ell,n}) - \frac{1}{J}\sum_{j=1}^{J}\log \widetilde{q}_j(\alpha_{\ell,n})\right)^2}_{\text{var. due to the different samples}},$$

$$\Delta_{2,n} = \underbrace{\frac{1}{L}\sum_{\ell=1}^{L}\left(\frac{1}{J}\sum_{j=1}^{J}\log \widetilde{q}_j(\alpha_{\ell,n}) - \mu_{\alpha,n}\right)^2}_{\text{var. due to the different quantile levels}}.$$

... to ANOVEX

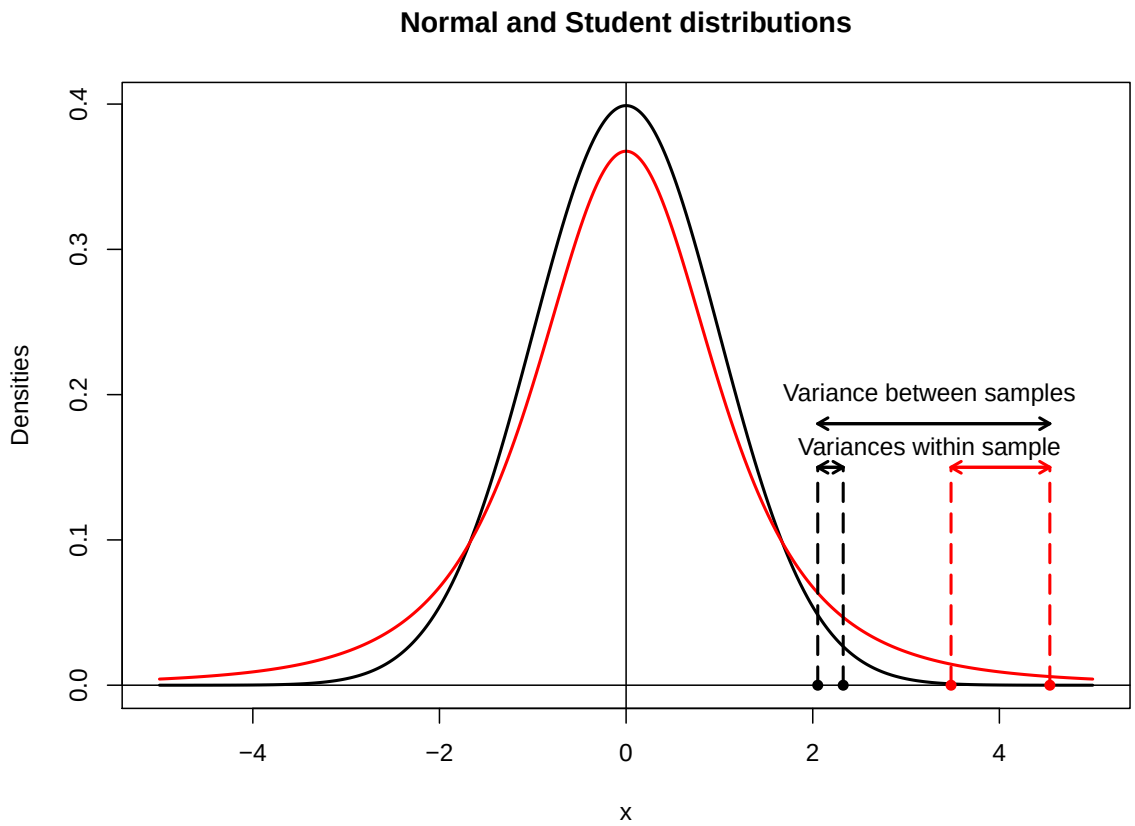


Figure 2: Decomposition of extreme log-quantile variance for a standard normal (black curve) and a Student's t distribution 3 degrees of freedom (red curve), for two extreme quantiles at probability levels $\{0.98, 0.99\}$.

Assumptions

Unfortunately, estimating an extreme quantile is not an obvious task, and some conditions are required.

- $\mathcal{C}_2(\xi, \rho, A)$: A cumulative distribution function F belongs to the class $\mathcal{C}_2(\xi, \rho, A)$ with tail index $\xi > 0$ and second-order parameter $\rho < 0$, if there exists a measurable auxiliary function A with constant sign, satisfying $A(t) \to 0$ as $t \to \infty$, such that

$$\lim_{t\to\infty} \frac{1}{A(1/\overline{F}(t))} \left(\frac{\overline{F}(ty)}{\overline{F}(t)} - y^{-1/\xi} \right) = y^{-1/\xi} \, \frac{y^{\rho/\xi} - 1}{\xi\rho}, \quad \text{for all } y > 0.$$

Extreme quantile estimator

In order to estimate the extreme quantiles in $\Delta_{1,n}$ and $\Delta_{2,n}$, we use the approach proposed by Weissman 1978:

- We carefully choose intermediate quantile levels $\beta_{j,n} \to 1$ such that $n_j(1-\beta_{j,n}) = n(1-\beta_n)(1+o(1)) \to \infty$ as $n \to \infty$.
- We estimate the tail indices $\xi_1, \ldots, \xi_J$ using the Hill estimator:

$$\widehat{\xi}_j(\beta_{j,n}) = \frac{1}{\lfloor (1-\beta_{j,n})n_j \rfloor} \sum_{i=0}^{\lfloor (1-\beta_{j,n})n_j \rfloor - 1} \log X^{(j)}_{n_j-i,n_j} - \log X^{(j)}_{n_j-\lfloor (1-\beta_{j,n})n_j \rfloor, n_j},$$

  where $X^{(j)}_{1,n_j} \leq X^{(j)}_{2,n_j} \leq \ldots \leq X^{(j)}_{n_j,n_j}$ (see Hill 1975).
- We deduce the Weissman extreme quantile estimator

$$\widehat{q}^{\,\mathrm{W}}_j(\alpha_{\ell,n} \,|\, \beta_{j,n}) = \widehat{q}_j(\beta_{j,n}) \left(\frac{1-\beta_{j,n}}{1-\alpha_{\ell,n}} \right)^{\widehat{\xi}_j(\beta_{j,n})}.$$

ANOVEX test statistic

For convenience, let us consider in the sequel

$$\alpha_{\ell,n} = 1 - \tau_\ell / n, \tau_\ell > 0 \text{ for } \ell = 1, \ldots, L.$$

We thus define the ANOVEX test statistic

$$T_n = \frac{J \operatorname{varlog}(\tau_{1:L})\, n(1-\beta_n)}{S_n(\beta_n, \tau_{1:L})} \frac{\Delta_{1,n}}{\Delta_{2,n}},$$

where

$$\operatorname{varlog}(\tau_{1:L}) = \frac{1}{L}\sum_{\ell=1}^{L} (\log(\tau_\ell))^2 - \left(\frac{1}{L}\sum_{\ell=1}^{L} \log(\tau_\ell)\right)^2 \text{ and}$$

$$S_n(\beta_n, \tau_{1:L}) = \frac{1}{L}\sum_{\ell=1}^{L} \left(\log\left(\frac{n(1-\beta_n)}{\tau_\ell}\right)\right)^2.$$

ANOVEX test statistic

Theorem 1

Suppose that $F_j \in \mathcal{C}_2(\xi_j, \rho_j, A_j)$ *for* $j = 1, \ldots, J$. *Moreover, we assume*

$$\sqrt{n(1-\beta_n)}\, A_j\left((1-\beta_n)^{-1}\right) \to 0, \quad n \to \infty, \quad \text{for all } j = 1, \ldots, J.$$

Then, under (H_0),

$$T_n \xrightarrow{d} \chi^2_{J-1}, \quad n \to \infty.$$

The ANOVEX test rejects (H_0) with asymptotic level $\gamma \in (0,1)$ if

$$T_n = \frac{J\,\mathrm{varlog}(\tau_{1:L})\, n(1-\beta_n)}{S_n(\beta_n, \tau_{1:L})} \frac{\Delta_{1,n}}{\Delta_{2,n}} > \chi^2_{J-1,1-\gamma},$$

where $\chi^2_{J-1,1-\gamma}$ denotes the quantile of level $1-\gamma$ of the chi-square distribution with $J-1$ degrees of freedom.

Type-I and type-II error approximations

Let us denote for convenience

$$s_n(\beta_n, \tau_{1:L}) = \frac{1}{L}\sum_{\ell=1}^{L}\sqrt{1+\left(\log\left(\frac{n(1-\beta_n)}{\tau_\ell}\right)\right)^2},$$

$$\text{and } \mathfrak{s}_n(\beta_n, \tau_{1:L}) = \frac{1}{L}\sum_{\ell=1}^{L}\log\left(\frac{n}{\tau_\ell}\right)\sqrt{1+\left(\log\left(\frac{n(1-\beta_n)}{\tau_\ell}\right)\right)^2}.$$

We will deal with three situations (all in the case $J = 2$):

1. Two identically distributed Pareto samples,
2. Two Pareto samples with different scale parameters,
3. Two Pareto samples with different shape parameters.

Identically distributed Pareto samples

Proposition 2

Consider two independent samples $E_j = \{X_1^{(j)}, \ldots, X_n^{(j)}\}$, $j = 1, 2$, *of i.i.d. variables following the same Pareto distribution* $\mathcal{P}(1/\xi)$, $\xi > 0$. *Assume that* (β_n) *is an intermediate probability level such that* $(1 - \beta_n)\log(n) \to 0$ *as* $n \to 0$. *Then, as* $n \to \infty$,

$$T_n \overset{d}{=} \Gamma^2 \left(1 + \frac{1}{S_n(\beta_n, \tau_{1:L})}\right) \times \left(1 + O_{\mathbb{P}}\left(\frac{1}{\sqrt{n(1-\beta_n)}}\right) + O_{\mathbb{P}}\left(\frac{1-\beta_n}{\log(n(1-\beta_n))}\right)\right)$$

where Γ *is a standard normal random variable.*

Identically distributed Pareto samples

The probability $\mathbb{P}_{H_0}\left(T_n > \chi^2_{1,1-\gamma}\right)$ to wrongly reject (H_0) with asymptotic level $\gamma \in (0,1)$ is for large n approximately equal to

$$p_n(\gamma) = 2\bar{\Phi}\left(\bar{\Phi}^{-1}(\gamma/2)\left(1+\frac{1}{S_n(\beta_n,\tau_{1:L})}\right)^{-1/2}\right),$$

where $\overline{\Phi}(\cdot)$ is the standard Gaussian survival function.

Identically distributed Pareto samples

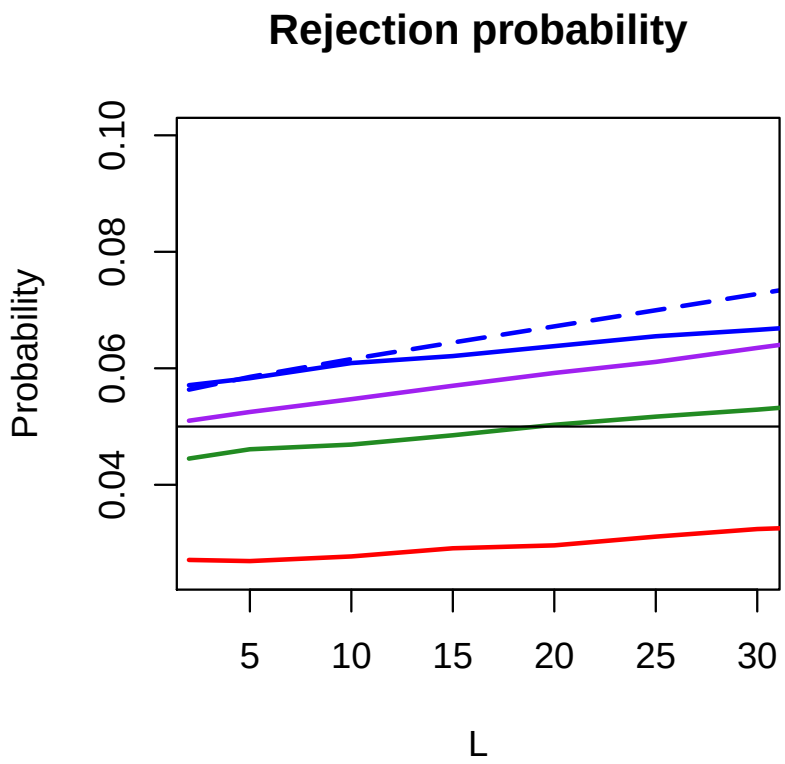


Figure 3: Empirical (solid curves) and approximated (dashed curve) type I errors obtained for $10,000$ replications with $n = 1,000$, $\beta_n = 0.9$ and $\xi = 0.25$, as functions of L. The underlying distribution is a Pareto (blue), Fréchet (purple), Burr with $\rho = -0.75$ (green) and GPD (red) distribution.

Pareto samples with different scale parameters

Consider two independent samples denoted by $E_1 = \{X_1^{(1)}, \ldots, X_n^{(1)}\}$ and $E_2 = \{X_1^{(2)}, \ldots, X_n^{(2)}\}$ where

$(H_{1,n})$ $X_i^{(1)} \sim \mathcal{P}(1/\xi)$ and $X_i^{(2)} \stackrel{d}{=} \lambda_n X_i^{(1)}$, $i = 1, \ldots, n$ and $\lambda_n \to 1$ as $n \to \infty$.

The condition $\lambda_n \to 1$ (and in a sense $(H_{1,n}) \to (H_0)$) is a concept known in the literature as the contiguity.

Pareto samples with different scale parameters

Proposition 3

Assume that

$$\frac{\log(n(1-\beta_n))}{(n(1-\beta_n))^{3/4}} \vee \sqrt{\frac{\log(n(1-\beta_n))}{n}} << \log(\lambda_n) << \frac{1}{\sqrt{n(1-\beta_n)}}.$$

Then, as $n \to \infty$,

$$T_n \overset{d}{=} \left(\frac{(\log(\lambda_n))^2 \, n(1-\beta_n)}{2\xi^2 S_n(\beta_n, \tau_{1:L})} - \frac{\sqrt{2n(1-\beta_n)}\log(\lambda_n) s_n(\beta_n, \tau_{1:L})}{\xi S_n(\beta_n, \tau_{1:L})}\Gamma \right.$$
$$\left. + \frac{1 + S_n(\beta_n, \tau_{1:L})}{S_n(\beta_n, \tau_{1:L})}\Gamma^2 \right) \times \left(1 + O_{\mathbb{P}}\left(\frac{1}{\sqrt{n(1-\beta_n)}} \right) + O_{\mathbb{P}}\left(\frac{1-\beta_n}{\log(n(1-\beta_n))} \right) \right)$$

where Γ *is a standard normal random variable.*

Pareto samples with different scale parameters

By assuming the slightly stronger condition

$$\frac{\left(\log(n(1-\beta_n))\right)^2}{(n(1-\beta_n))^{3/4}} \vee \sqrt{\frac{\left(\log(n(1-\beta_n))\right)^3}{n}} = o(\log(\lambda_n)),$$

then the probability $\mathbb{P}_{H_{1,n}}\left(T_n \leq \chi^2_{1,1-\gamma}\right)$ to (wrongly) not reject (H_0) with asymptotic level $\gamma \in (0,1)$ is for large n approximately equal to

$$\bar{\Phi}\left(\Omega_{1,n} - \sqrt{\Omega_{2,n}}\right) - \bar{\Phi}\left(\Omega_{1,n} + \sqrt{\Omega_{2,n}}\right), \text{ where}$$

$$\Omega_{1,n} = \frac{\log(\lambda_n)\sqrt{n(1-\beta_n)}s_n(\beta_n, \tau_{1:L})}{\sqrt{2}\xi\left(1 + S_n(\beta_n, \tau_{1:L})\right)},$$

$$\Omega_{2,n} = \frac{\left(\log(\lambda_n)\right)^2 n(1-\beta_n)}{2\xi^2} \frac{s_n(\beta_n, \tau_{1:L})^2 - 1 - S_n(\beta_n, \tau_{1:L})}{\left(1 + S_n(\beta_n, \tau_{1:L})\right)^2}$$

$$+ \frac{S_n(\beta_n, \tau_{1:L})}{1 + S_n(\beta_n, \tau_{1:L})}\chi^2_{1,1-\gamma} > 0.$$

Pareto samples with different scale parameters

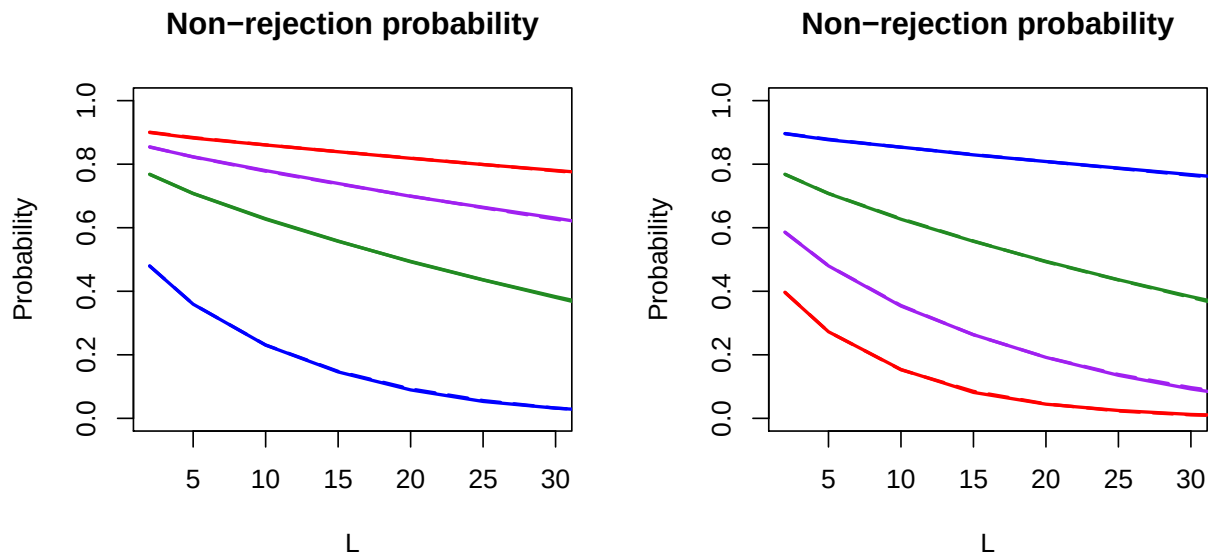


Figure 4: Empirical (solid curves) and approximated (dashed curves) type II errors obtained for $10,000$ replications with $n = 1,000$ and $\beta_n = 0.9$, shown as functions of L. Left: $\lambda_n = 1 + 2n^{-1/3} = 1.2$ and $\xi = 0.15$ (blue curves), $\xi = 0.25$ (green curves), $\xi = 0.35$ (purple curves) and $\xi = 0.5$ (red curves). Right: $\xi = 0.25$ and $\lambda_n = 1.1$ (blue curves), $\lambda_n = 1.2$ (green curves), $\lambda_n = 1.3$ (purple curves) and $\lambda_n = 1.4$ (red curves).

Pareto samples with different shape parameters

Consider two independent samples $E_1 = \{X_1^{(1)}, \ldots, X_n^{(1)}\}$ and $E_2 = \{X_1^{(2)}, \ldots, X_n^{(2)}\}$, where

$(H'_{1,n})$ $X_i^{(1)} \sim \mathcal{P}(1/\xi)$ and $X_i^{(2)} \stackrel{d}{=} (X_i^{(1)})^{\theta_n}$, $i = 1, \ldots, n$ and $\theta_n \to 1$ as $n \to \infty$.

Let us assume that $(\theta_n)_n$ satisfies the same conditions as $(\lambda_n)_n$.

Pareto samples with different shape parameters

Proposition 4

$$T_n \overset{d}{=} 2\left(\frac{n(1-\beta_n)(1-\theta_n)^2}{(1+\theta_n)^2}\frac{\mathrm{smlog}(n/\tau_{1:L})}{S_n(\beta_n,\tau_{1:L})} + \frac{(1+\theta_n^2)}{(1+\theta_n)^2}\frac{1+S_n(\beta_n,\tau_{1:L})}{S_n(\beta_n,\tau_{1:L})}\Gamma^2 \right.$$
$$\left. +2\frac{\sqrt{n(1-\beta_n)}\sqrt{1+\theta_n^2}(1-\theta_n)}{(1+\theta_n)^2}\frac{\mathfrak{s}_n(\beta_n,\tau_{1:L})}{S_n(\beta_n,\tau_{1:L})}\Gamma\right)$$
$$\times\left(1+O_{\mathbb{P}}\left(\frac{1-\beta_n}{\log(n(1-\beta_n))}\right)+O_{\mathbb{P}}\left(\frac{1}{\sqrt{n(1-\beta_n)}}\right)\right), \text{ as } n\to\infty,$$

where Γ is a standard normal random variable and

$$\mathrm{smlog}(n/\tau_{1:L}) = \frac{1}{L}\sum_{\ell=1}^{L}\left(\log\left(n/\tau_\ell\right)\right)^2$$

Pareto samples with different shape parameters

Under a slightly stronger assumption on θ_n, then the probability $\mathbb{P}_{H'_{1,n}}\left(T_n \leq \chi^2_{1,1-\gamma}\right)$ to (wrongly) not reject (H_0) with asymptotic level $\gamma \in (0,1)$ may be approximated by for n large enough by

$$\bar{\Phi}\left(\Psi_{1,n} - \sqrt{\Psi_{2,n}}\right) - \bar{\Phi}\left(\Psi_{1,n} + \sqrt{\Psi_{2,n}}\right)$$

where

$$\Psi_{1,n} = \frac{\sqrt{n(1-\beta_n)}(\theta_n - 1)\mathfrak{s}_n(\beta_n, \tau_{1:L})}{\sqrt{1+\theta_n^2}\,(1 + S_n(\beta_n, \tau_{1:L}))},$$

$$\Psi_{2,n} = \frac{(\theta_n - 1)^2 n(1-\beta_n)}{(1+\theta_n^2)} \frac{\mathfrak{s}_n(\beta_n, \tau_{1:L})^2 - (1 + S_n(\beta_n, \tau_{1:L}))\,\mathrm{smlog}(n/\tau_{1:L})}{(1 + S_n(\beta_n, \tau_{1:L}))^2}$$

$$+ \frac{(1+\theta_n)^2}{(1+\theta_n^2)} \frac{S_n(\beta_n, \tau_{1:L})}{1 + S_n(\beta_n, \tau_{1:L})} \frac{\chi^2_{1,1-\gamma}}{2} > 0.$$

Pareto samples with different shape parameters

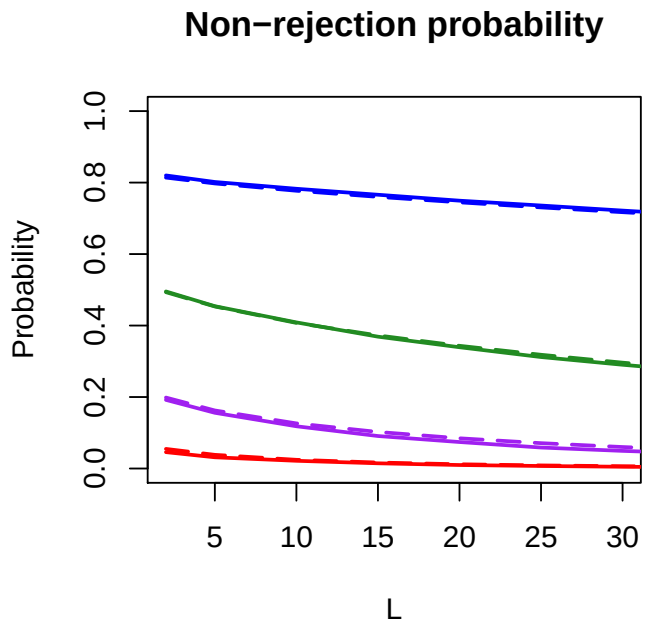


Figure 5: Empirical (solid curves) and approximated (dashed curves) type II errors obtained for $10,000$ replications with $n = 1,000$ and $\beta_n = 0.9$, shown as functions of L. $\xi = 0.25$ and $\theta_n = 1 + n^{-1/3} = 1.1$ (green curves), $\theta_n = 1.2$ (brown curves), $\theta_n = 1.3$ (blue curves) and $\theta_n = 1.4$ (red curves).

A note on the proofs

We denote $k_n = \lfloor n(1-\beta_n) \rfloor$. If E_1 and E_2 are Pareto distributed, the Rényi's representation[1] provides

$$\widehat{\xi}_1(\beta_n) - \widehat{\xi}_2(\beta_n) \stackrel{d}{=} \frac{(\mathcal{E}_1^{(1)} - \mathcal{E}_1^{(2)}) + \ldots + (\mathcal{E}_{k_n}^{(1)} - \mathcal{E}_{k_n}^{(2)})}{k_n} \stackrel{d}{=} \frac{\mathcal{L}_1 + \ldots + \mathcal{L}_{k_n}}{k_n},$$

where $\{\mathcal{E}_1^{(j)}, \ldots, \mathcal{E}_{k_n}^{(j)}\}$ and $\{\mathcal{L}_1, \ldots, \mathcal{L}_{k_n}\}$ are i.i.d. realizations of an exponential distribution with mean ξ and a centered Laplace distribution with variance $2\xi^2$, respectively.

Since the Laplace distribution is log-concave, centered and symmetric, Klartag 2009[2] proved that the Berry-Esseen bound is refined with k_n instead of $\sqrt{k_n}$.

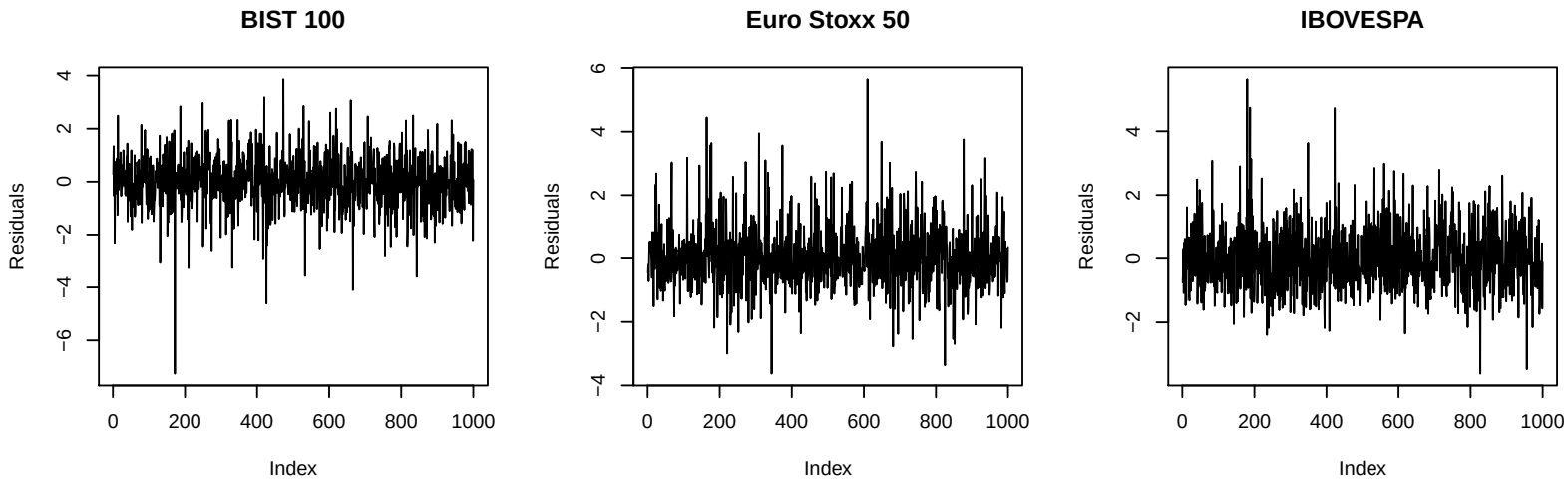
---

[1] Rényi, A. (1953). On the theory of order statistics. *Acta Mathematica Academiae Scientiarum Hungarica*, **4(2)**:191–231.

[2] Klartag, B. (2009). A Berry-Esseen type inequality for convex bodies with an unconditional basis. *Probability Theory and Related Fields*, **145(1)**:1–33.

Real data example

We use the last $n = 1,000$ negative daily log-returns for $J = 12$ stock market indices (before June 16, 2023 included), BIST 100 (Turkey), IBOVESPA (Brazil), IPC Mexico, KOSPI Composite (South Korea), MOEX Russia, PSEi (Philippines), S&P BSE 500 (India), S&P MERVAL (Argentina), SSE Composite (China), TA-125 (Israel) and Tadawul All Shares (Saudi Arabia). We also added the European Euro Stoxx 50 in the study.



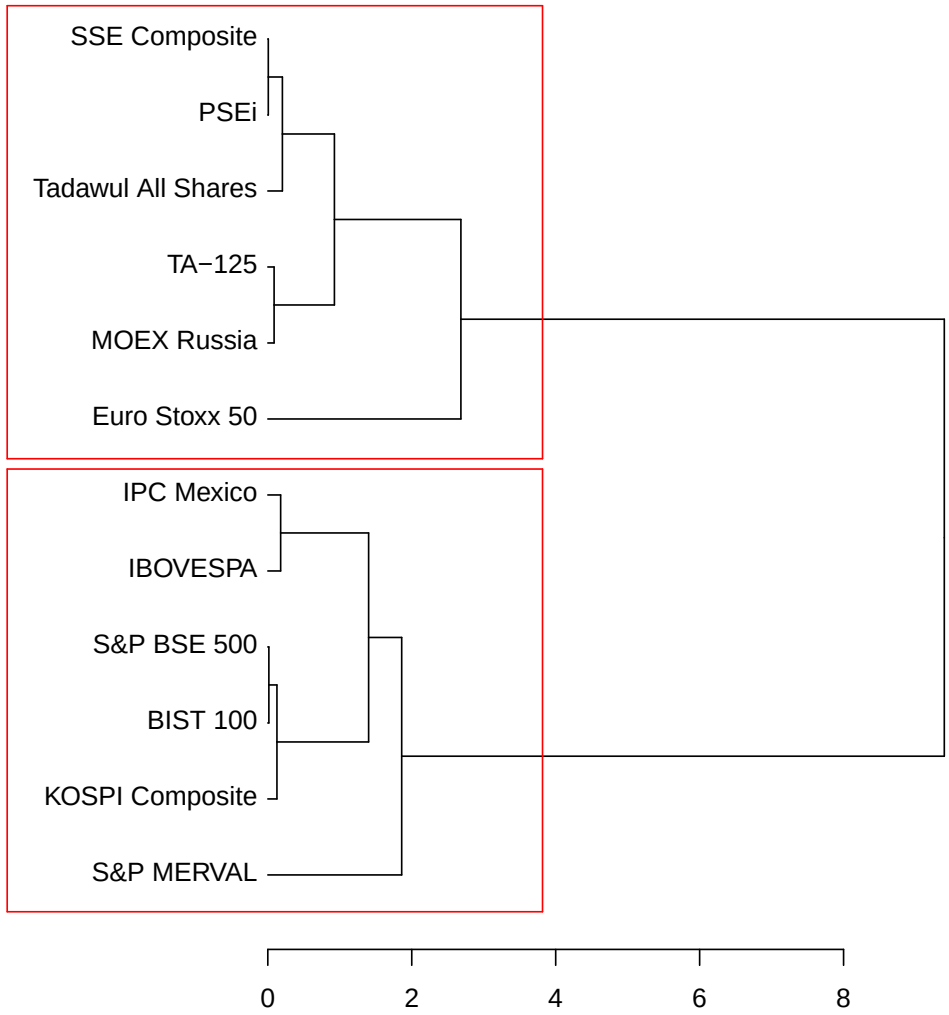
Real data example

As in Jondeau and Rockinger 2003[3], we propose to test the equality of the tails, and identify clusters of stock indices having the same extreme quantiles.

For that purpose, we use the ANOVEX statistic as a dissimilarity measure to construct a dendrogram, and select the optimal number of groups by hierarchically applying the ANOVEX test.

[3] Jondeau, E. and Rockinger, M. (2003). Testing for differences in the tails of stock-market returns. *Journal of Empirical Finance* **10** 559–581.

Real data example



Real data example

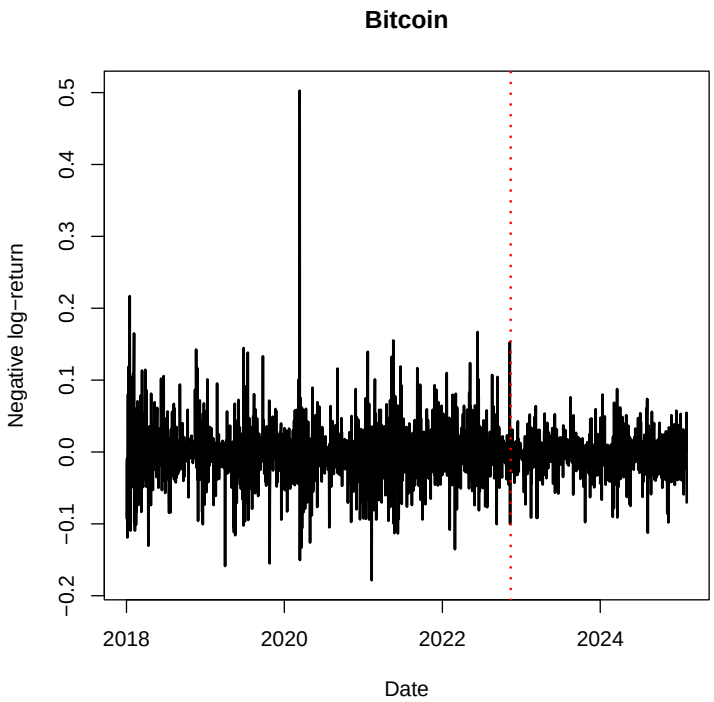
It is interesting to notice this procedure tends to split the samples mainly into two groups:

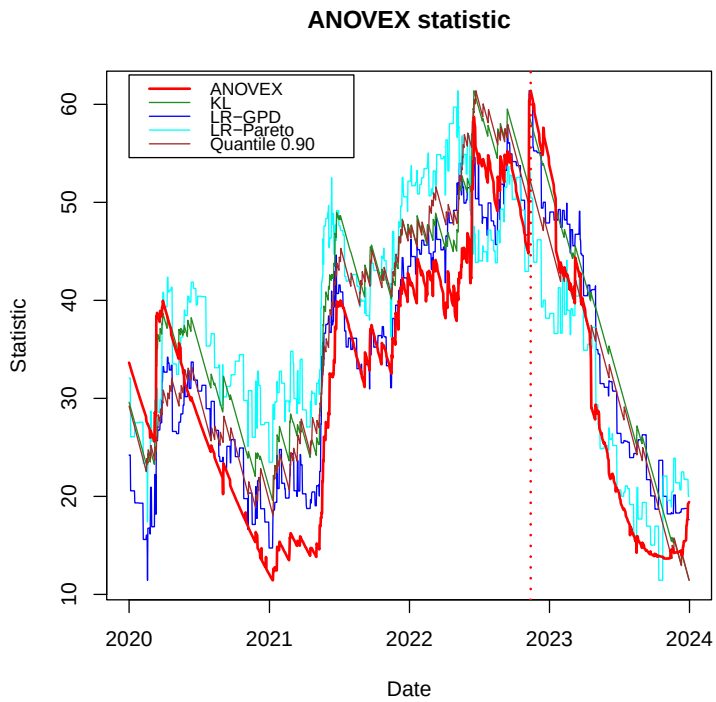
- a first one containing the indices from Central and South America (and the indian, korean and turkish indices as well),
- a second one containing the other Eurasian indices (including the Euro Stoxx 50).

Conclusion

- Girard, S., Opitz, T. and Usseglio-Carleve, A. (2024). ANOVEX: ANalysis Of Variability for heavy-tailed EXtremes, *Electronic Journal of Statistics*, **18(2)**, 5258-5303.
- `R` package `ANOVEX` `https://github.com/AntoineUC/ANOVEX`.

Application in change-point detection coming soon (with C. Yan)...





References I

[Hil75] Bruce M. Hill. "A Simple General Approach to Inference About the Tail of a Distribution". In: *The Annals of Statistics* 3.5 (Sept. 1975), pp. 1163–1174. DOI: 10.2307/2958370. URL: http://gen.lib.rus.ec/scimag/index.php?s=10.2307/2958370.

[JR03] Eric Jondeau and Michael Rockinger. "Testing for differences in the tails of stock-market returns". In: *Journal of Empirical Finance* 10.5 (2003), pp. 559–581.

[Kla09] B. Klartag. "A Berry-Esseen type inequality for convex bodies with an unconditional basis". In: *Probability Theory and Related Fields* 145.1 (2009), pp. 1–33.

[Wei78] Ishay Weissman. "Estimation of parameters and large quantiles based on the k largest observations". In: *Journal of the American Statistical Association* 73.364 (1978), pp. 812–815.