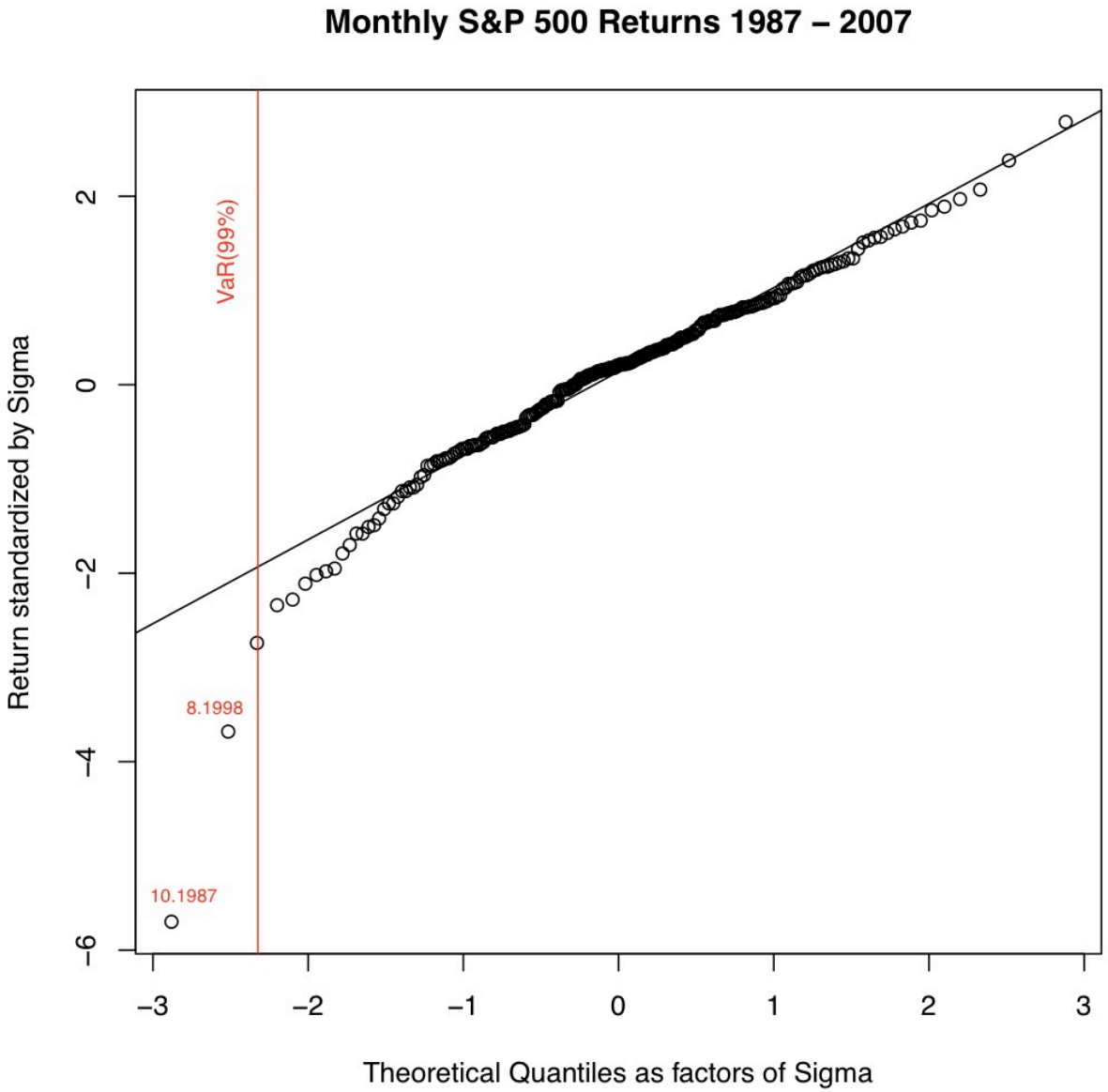
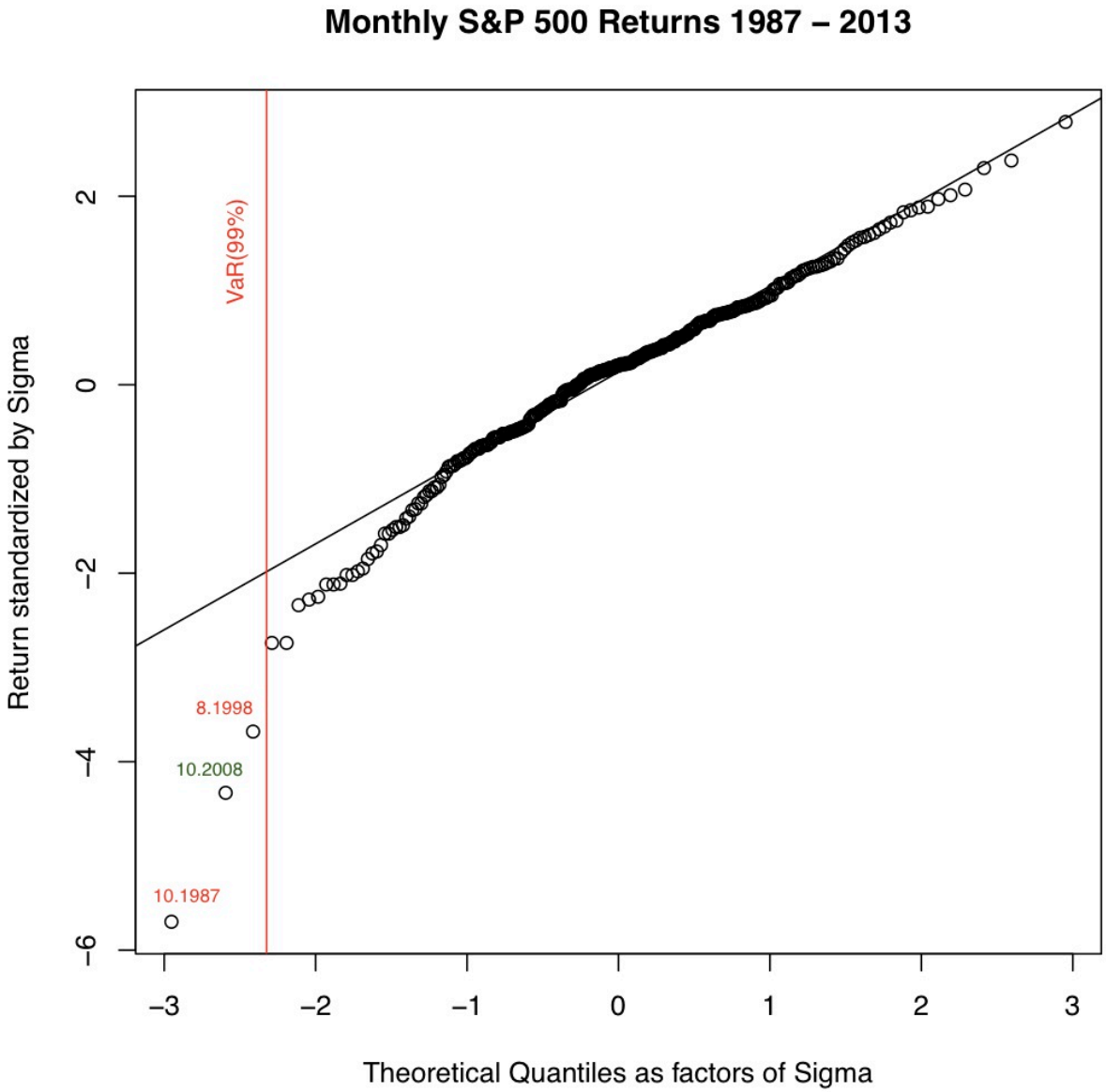
# Extreme Value Theory and its Applications to Insurance and Finance

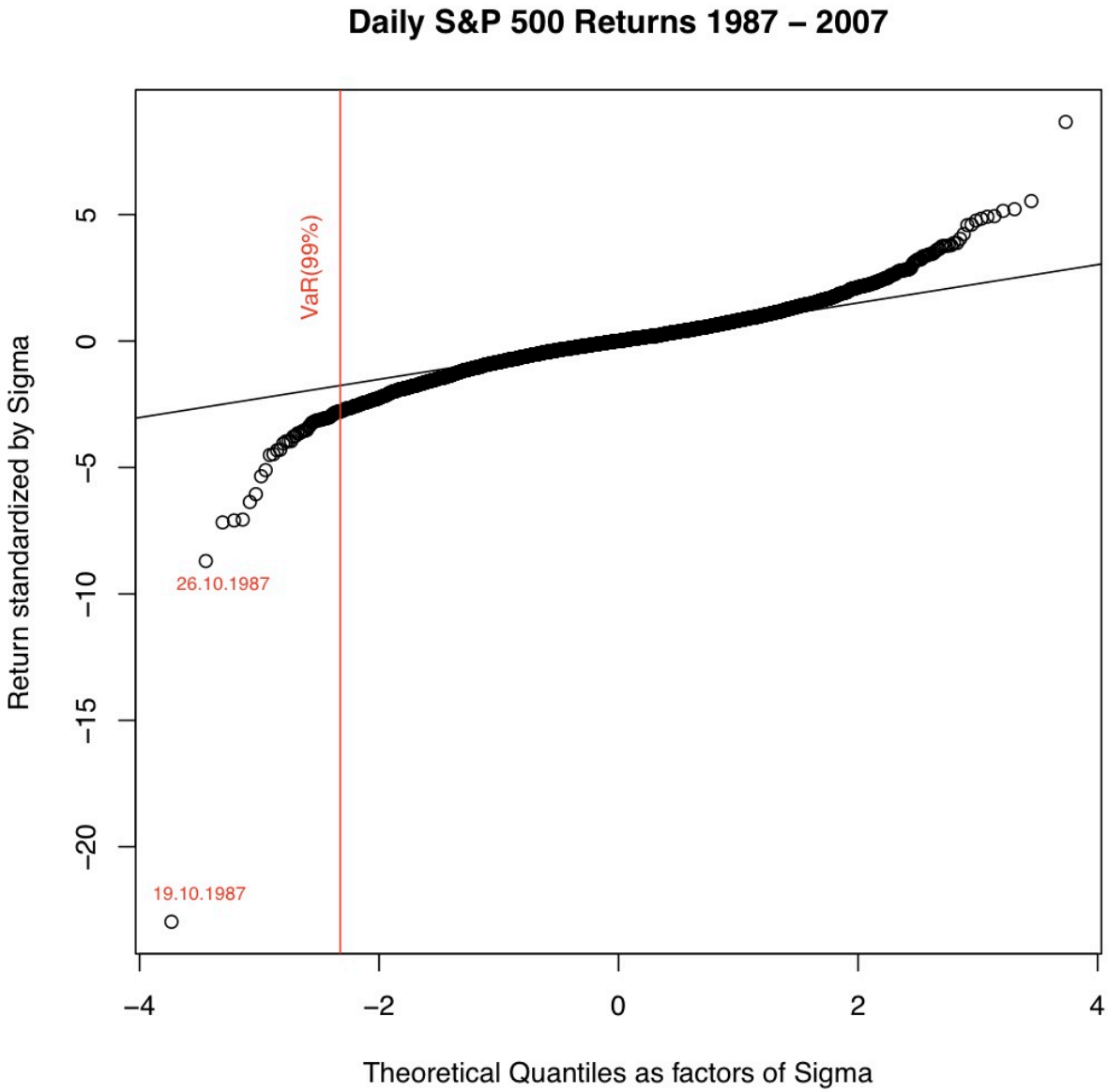
**Marie Kratz**
**ESSEC Business School**

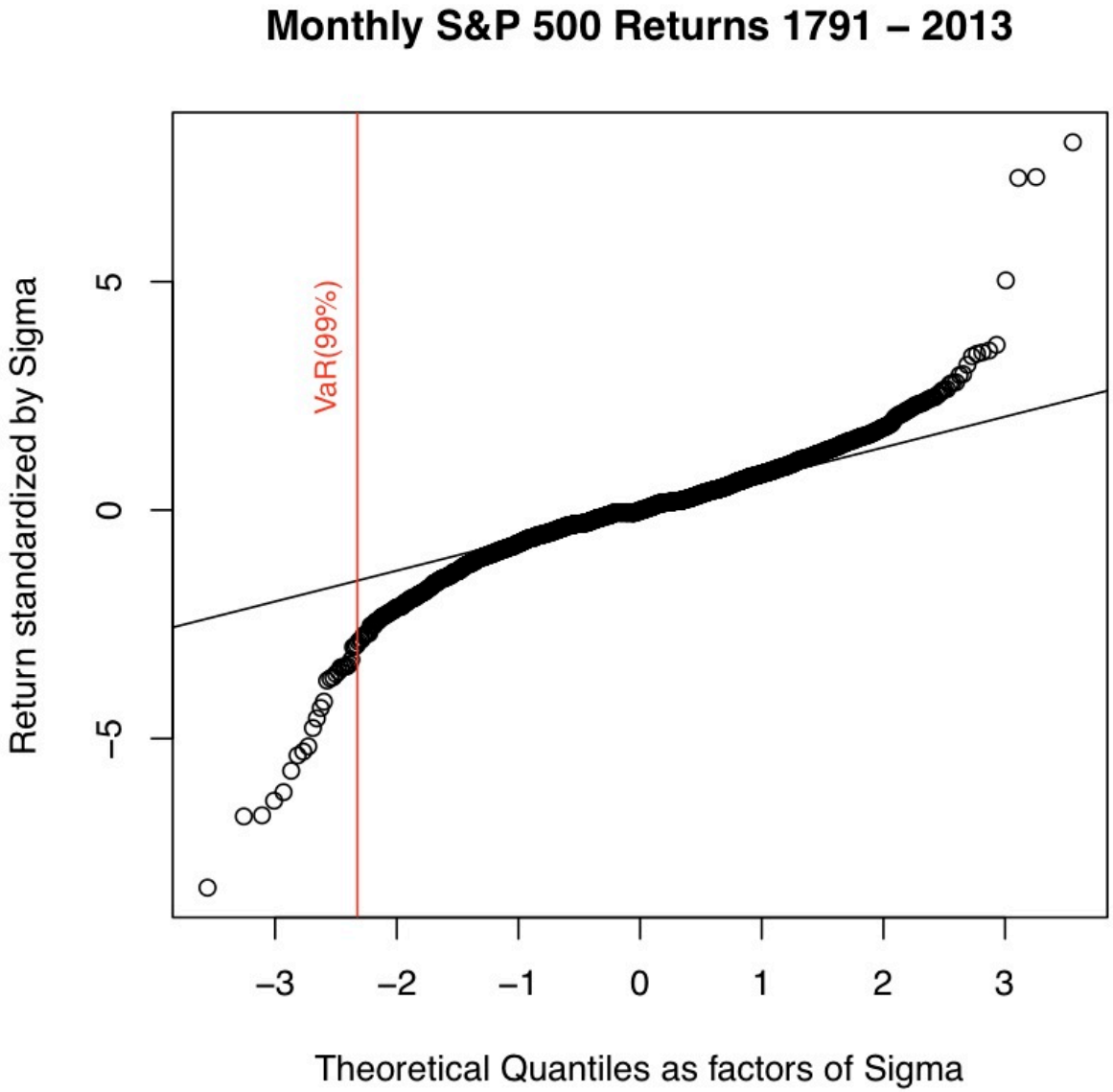

**ETH Risk Center**
**Zurich, March 24, 2017**

# I - Overview of univariate EVT




Monthly S&P 500 Returns 1987 – 2013
Return standardized by Sigma
Theoretical Quantiles as factors of Sigma
VaR(99%)
8.1998
10.2008
10.1987
2
0
−2
−4
−6
−3
−2
−1
0
1
2
3


Daily S&P 500 Returns 1987 – 2007
Return standardized by Sigma
Theoretical Quantiles as factors of Sigma
VaR(99%)
26.10.1987
19.10.1987
5
0
−5
−10
−15
−20
−4
−2
0
2
4




*(data from https://www.globalfinancialdata.com)*

# CLT versus EVT

$(X_i)_{i=1,\cdots,n}$ iid with continuous cdf $F$

↗ **mean behavior** $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ : asymptotically (large $n$)

**Gaussian** if $var(X) < \infty$, when **linearly transformed**/normalized (since $var(\bar{X}_n) = \frac{1}{n} var(X) \underset{n\to\infty}{\to} 0$) (CLT)

$$\frac{\bar{X}_n - \mathbb{E}(\bar{X}_n)}{\sqrt{var(\bar{X}_n)}} = \sqrt{\frac{n}{var(X)}}\,(\bar{X}_n - \mathbb{E}(X)) = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{\sigma^2}} \underset{n\to\infty}{\overset{d}{\to}} \mathcal{N}(0,1)$$

i.e. $\lim_{\mathbf{n}\to\infty} \mathbb{P}[(\bar{\mathbf{X}}_\mathbf{n} - \mathbf{b_n})/\mathbf{a_n} \le \mathbf{x}] = \mathbf{F}_{\mathcal{N}(\mathbf{0},\mathbf{1})}(\mathbf{x})$ with $b_n = \mathbb{E}(X), a_n = \sqrt{\frac{var(X)}{n}}$

↘ **extreme behavior?** consider e.g. the **maximum**

$$\mathbb{P}[\max_{1\le i\le n} X_i \le x] = \prod_{i=1}^{n} \mathbb{P}[X_i \le x] = F^n(x) \underset{n\to\infty}{\to} \begin{cases} 0 \text{ if } F(x) < 1 \\ 1 \text{ if } F(x) = 1 \end{cases}$$

Could we find, as for the CLT, a linear transformation, to avoid such degeneracy, and say that there exist sequences $(a_n)$, $(b_n)$ and a rv $Z$ with cdf H such that

$$\lim_{n\to\infty} \mathbb{P}[(\max X_i - b_n)/a_n \le x] = H(x)\,?$$

Equivalent to look for $(a_n)$ and $(b_n)$, and a non degenerated cdf $H$ s.t.

$$\mathbb{P}\left[\frac{\max X_i - b_n}{a_n} \leq x\right] = \mathbb{P}[\max_{1\leq i\leq n} X_i \leq a_n x + b_n] = F^n(a_n x + b_n) \underset{n\to\infty}{\simeq} H(x)$$

The "three-types theorem". *The rescaled sample extreme (max renormalized) has a limiting distribution $H$ that can only be of three types:*

$$\begin{aligned}
H_{1,a}(x) &:= \exp\{-x^{-a}\}\mathbb{I}_{(x>0)} && (a>0) && : \textbf{Fréchet}\\
H_{2,a}(x) &:= \mathbb{I}_{(x\geq 0)} + \exp\{-(-x)^{a}\}\mathbb{I}_{(x<0)} && (a>0) && : \textbf{Weibull}\\
H_{3,0}(x) &:= \exp\{-e^{-x}\}, && \forall x\in\mathbb{R} && : \textbf{Gumbel}
\end{aligned}$$

(similar result holds for the minimum)

We will classify the distributions according to the three possible limiting distributions of the (rescaled) maximum, introducing the notion of maximum domain of attraction (MDA):

$$F \in \text{MDA}(H) \quad \Leftrightarrow \quad \exists (a_n)>0, (b_n) : \forall x\in\mathbb{R}, \lim_{n\to\infty} F^n(a_n x + b_n) = H(x).$$

The three types of extreme value distribution have been combined into a single three-parameter family (Jenkinson-Von Mises, 1955; Hosking et al., 85) known as **Generalized Extreme Value Distribution** (GEV)

EVT (EV Theorem).
*If $F \in MDA(G)$ then, necessarily, $G$ is of the same type as the GEV cdf $H_\gamma$ (i.e. $G(x) = H_\gamma(ax+b)$, $a>0$) defined by*

$$H_\gamma(x) = \begin{cases} \exp\left[-(1+\gamma x)_+^{-\frac{1}{\gamma}}\right] & \text{if } \gamma \neq 0 \\ \exp(-e^{-x}) & \text{if } \gamma = 0 \end{cases}$$

*where $y_+ = \max(0, y)$.*

The tail index $\gamma$ ($\in \mathbb{R}$) determines the nature of the tail distribution and is called the extreme-value index.
$\gamma > 0$ (Fréchet), $= 0$ (Gumbel) or $< 0$ (Weibull).

$$G(x) = G_{\mu,\sigma,\xi}(x) = \exp\left[-\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-\frac{1}{\xi}}\right], \text{ for } 1+\xi\,\frac{x-\mu}{\sigma} > 0$$

Moments of GEV: the $k$th moment exits if $\gamma < 1/k$

Example:

| cdf | for the maximum |
|---|---|
| Uniform | Weibull |
| Exponential(1) $(F(x) = 1 - e^{-x},\ x > 0)$ | Gumbel |
| Gaussian | Gumbel |
| Log-normal | Gumbel |
| Gamma $(\lambda, r)$ | Gumbel |
| Cauchy $(F(x) = \frac{1}{2} + \frac{1}{\pi} \arctan x)$ | Fréchet |
| Student | Fréchet |
| Pareto $(\beta)$ $(F(x) = 1 - x^{-\beta},\ x \geq 1,\ \beta > 0)$ | Fréchet |

- $F$ belongs to the MDA(Fréchet) ($\xi > 0$) iff its tail $\bar{F} = 1 - F$ is regularly varying of order $-\xi$ ($\xi$ is called the tail index of the distribution):

$$\lim_{x \to \infty} \frac{\bar{F}(ax)}{\bar{F}(x)} = a^{-\xi}, \ a > 0.$$

  Ex: Pareto, Cauchy, loggama, $\alpha$-stable d.f. with $\alpha < 2$.

- For the Weibull domain of attraction: $F$ has support bounded to the right $\Rightarrow$ not used to model extremal events in insurance and finance Ex: uniform d.f. on finite interval, beta d.f.

- For the Gumbel domain of attraction: wide class of d.f. with different tail behaviors

# A limit theorem for Extremes: the Pickands theorem

More information in the tail of a distribution than just that given by the maximum: consider the $k$th $(k \geq 1)$ largest order statistics.

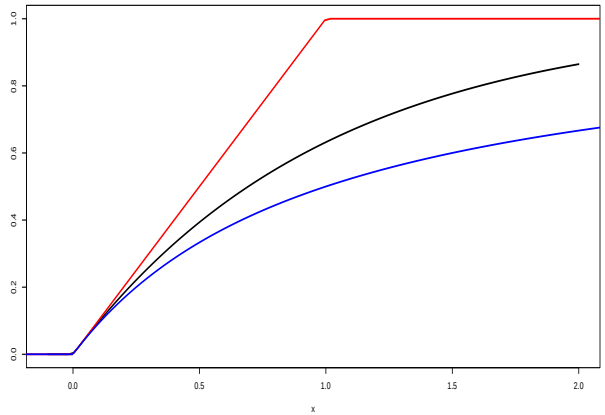
Notion of 'threshold exceedances' where all data are extreme in the sense that they exceed a high threshold. Main alternative approach to classic EVT and based on exceedances over high thresholds.

Pick up a high threshold $u < x_F^+$ (upper-end point of $F$), then study all exceedances above $u$.

Pickands Theorem. *For a sufficiently high threshold* $u$, $\exists\ \beta(u) > 0$ *and* $\xi$ *real number such that the Generalized Pareto Distribution (GPD) is a very good approximation to the excess d.f.* $F_u$*:*

$$F_u(y) := \mathbb{P}\left[X - u \leq y \,|\, X > u\right] \underset{\text{large } u}{\approx} G_{\xi,\beta(u)}(y)\,.$$

Recall - def GPD: $\overline{G}_{\xi,\beta(u)}(y) = \begin{cases} \left(1 + \xi \frac{y}{\beta(u)}\right)^{-\frac{1}{\xi}} & \text{if } \xi \neq 0 \\ e^{-\frac{y}{\beta(u)}} & \text{if } \xi = 0 \end{cases}$



$GPD_{\xi,1}$, *with* $\xi = -1$ *(red),* $\xi = 0$ *(black),* $\xi = 1$ *(blue)*

As for the GEV, 3 different cases for the GPD $G(\xi, \beta)$, depending on the sign of the tail index $\xi$:

- $\xi > 0$: $\overline{G}_{\xi,\beta}(y) \sim cy^{-1/\xi}$, $c > 0$ ("Pareto" tail) : heavy-tail. We have $\mathbb{E}(X^k) = \infty$ for $k \geq 1/\xi$.
- $\xi < 0$: $x_G^+ = \beta/|\xi|$ (upper endpoint of $G$), similar to the Weibull type of GEV(short-tailed, pareto type II distribution)
- $\xi = 0$: $\overline{G}_{\xi,\beta}(y) = e^{-y/\beta}$: light-tail (exponential distribution with mean $\beta$)

The mean of the GPD is defined for $\xi < 1$: $\mathbb{E}(X) = \dfrac{\beta}{1-\xi}$.

**Goal in EVT**: **to model the tail of the distribution**. It means to extract the information from the observations above a high threshold (largest order stat)

↪ **Main issue**: how to **determine this threshold**, to be able to estimate **the tail index**?

## Standard methods in EVT (supervised):

- **MEP (Mean Excess Plot)** method

  If $Y$ has a $\text{GPD}_{\xi,\beta}$, then its Mean Excess function $e$ is linear:

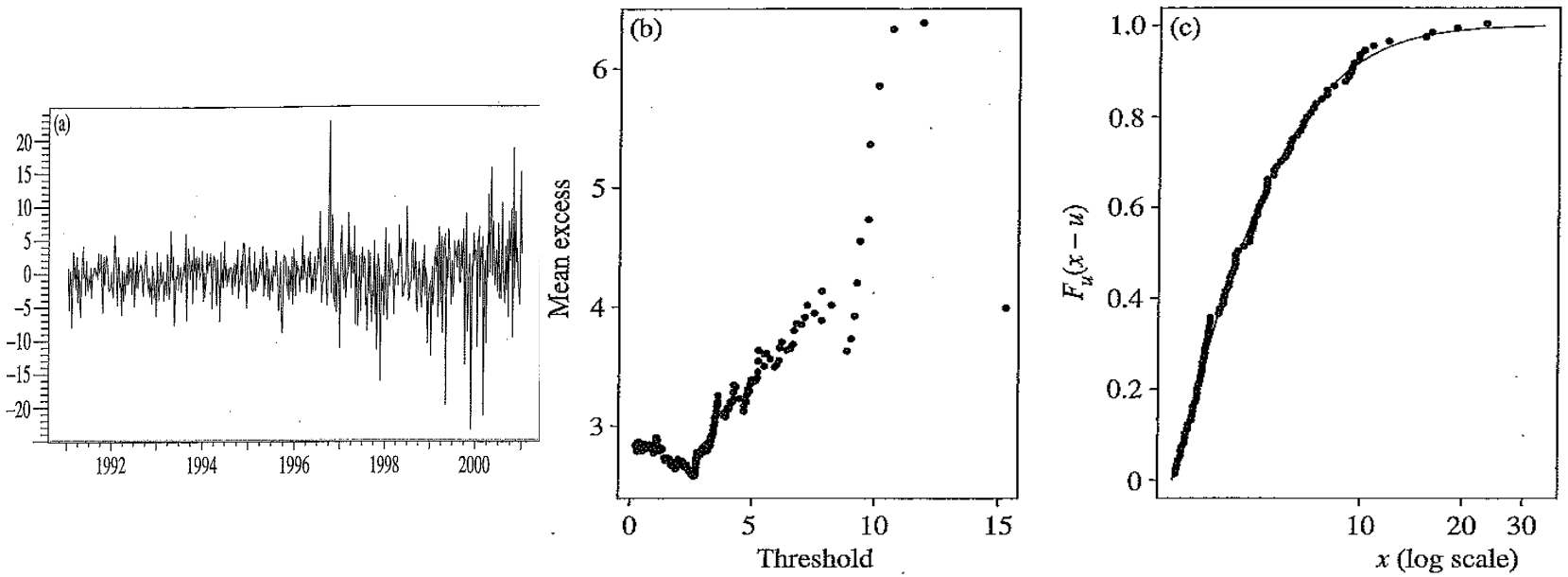
$$e(v) = \mathbb{E}[Y - v \mid Y > v] = \frac{1}{1-\xi}\,(\beta + v\,\xi)\;1_{(\beta + v\;\xi > 0)}.$$

  Hence, **above** $u$ at which the GPD provides a valid approximation to the excess distribution, the **MEP** should stay reasonably close to a **linear** function $\Rightarrow$ way to select $u$

  Then, u being chosen, use ML or Moments estimators for the tail index $\xi$ (and the scaling parameter $\beta$).

**Illustration** (Example from Embrechts et al's book: *Modelling Extremal Events: for Insurance and Finance*)

Data set: time series plot (a) of *AT&T weekly percentage loss data* for the 521 complete weeks in the period 1991-2000




(b) Sample MEP. Selection of the threshold at a loss value of 2.75% (102 exceedances)

(c) Empirical distribution of excesses and fitted GPD, with ML estimators $\hat{\xi} = 0.22$ and $\hat{\beta} = 2.1$ (with SE 0.13 and 0.34, respectively)

## **Tail index estimators** for MDA(Fréchet) distributions

Other graphical methods to determine the tail index than MEP may be used; we present two of them under regular variation framework: the Hill plot (most popular) and the QQ-plot (Kratz & Resnick).

For a sample size $n$, the tail index estimators are built on the $k = k(n)$ upper order statistics, with $k(n) \to \infty$ such that $k(n)/n \to 0$, as $n \to \infty$.

Choosing $k$ is usually the Achilles heel of all these (graphical) supervised procedures, including the MEP one, as already observed.

Nevertheless it is remarkable to notice that for these methods, no extra information is required on the observations before the threshold (the $n-k$th order statistics).

Assume $F \in$ MDA(Fréchet) with tail index $\xi > 0$,
*i.e.* $\bar{F} \sim RV_{-\alpha}$, with $\xi = \alpha^{-1}$
Order statistics: $X_{1,n} \leq X_{2,n} \leq \cdots \leq X_{n-1,n} \leq X_{n,n} = max(X_i)$

Consider the threshold $u = X_{n-k,n}$ with $k = k(n) \to \infty$ and $k(n)/n \to 0$ as $n \to \infty$.

↗ **Hill estimator** $H_{k,n}$, of the tail index $\xi = \alpha^{-1}$ :

$$H_{k,n} := \frac{1}{k} \sum_{i=0}^{k-1} \log \left( \frac{X_{n-i,n}}{X_{n-k,n}} \right) \xrightarrow[n\to\infty]{P} \xi$$

Rk: rate of convergence: $1/\alpha^2$

↘ **QQ-estimator** (Kratz & Resnick), $Q_{k,n}$, of the tail index $\xi$:

Main idea of the QQ-method: suppose $X_{1,n} \leq X_{2,n} \leq \ldots X_{n,n}$ order statistics from an iid $n$-sample with continuous cdf $G$.

Then the plot of $\{(\frac{i}{n+1}\,,\, G(X_{i,n})), 1 \leq i \leq n\}$ should be approximately linear, hence, idem for the plot of $\{(G^{\leftarrow}(\frac{i}{n+1})\,,\, X_{i,n}), 1 \leq i \leq n\}$.

Note that $G^{\leftarrow}(\frac{i}{n+1})$= theoretical quantile, and $X_{i,n}$=corresponding quantile of the empirical distribution function; hence the name QQ-plot.

If $G = G_{\mu,\sigma}(x) = G_{0,1}(\frac{x-\mu}{\sigma})$, since $G^{\leftarrow}_{\mu,\sigma}(y) = \sigma G^{\leftarrow}_{0,1}(y) + \mu$, the plot of $\{(G^{\leftarrow}_{0,1}(\frac{i}{n+1})\,,\, X_{i,n}), 1 \leq i \leq n\}$ should be approximately a line of slope $\sigma$ and intercept $\mu$.

Take the example of $(X_i)$ Pareto-$\alpha$, ie $\bar{F}(x) = x^{-\alpha}$ ; then, for $y > 0$, $G_{0,\alpha}(y) := P[\log X_1 > y] = e^{-\alpha y}$ and the plot of

$$\left\{(G_{0,1}^{\leftarrow}(\frac{i}{n+1}), \log X_{i,n}), 1 \leq i \leq n\right\} = \left\{\left(-\log\left(1 - \frac{i}{n+1}\right), \log X_{i,n}\right), 1 \leq i \leq n\right.$$

should be approximately a line with intercept 0 and slope $\alpha^{-1}$.
Now, just use the least squares estimator for the slope (SL), namely

$$SL(\{(x_i, y_i), 1 \leq i \leq n\}) = \frac{\sum_{i=1}^n x_i y_i - \bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - \bar{x}^2}$$

to conclude that, for the Pareto example, an estimator of $\alpha^{-1}(= \xi)$ is

$$\widehat{\alpha^{-1}} = \frac{\sum_{i=1}^n -\log(\frac{i}{n+1})\{n \log X_{n-i+1,n} - \sum_{j=1}^n \log X_{n-j+1,n}\}}{n\sum_{i=1}^n(-\log(\frac{i}{n+1}))^2 - (\sum_{i=1}^n -\log(\frac{i}{n+1}))^2},$$

which we call the QQ-estimator.

More generally, for $\bar{F} \sim RV_{-\alpha}$, we can define the QQ–estimator $Q_{k,n}$ of the tail index $\xi = \alpha^{-1}$, based on the upper $k$ order statistics, by

$$Q_{k,n} = SL(\{(-\log(1 - \frac{i}{k+1}), \log X_{n-k+i,n}), 1 \leq i \leq k\})$$

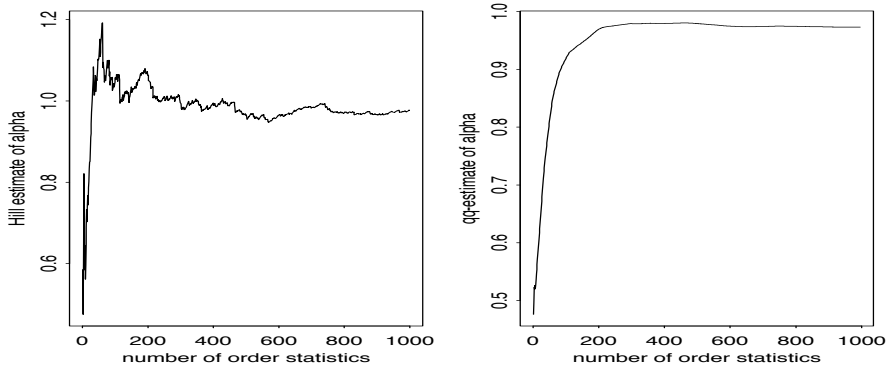
i.e.

$$Q_{k,n} := \frac{\sum_{i=1}^{k} - \log\left(\frac{i}{k+1}\right) \left\{ k \log\left(X_{n-i+1,\,n}\right) - \sum_{j=1}^{k} \log\left(X_{n-j+1,\,n}\right) \right\}}{k \sum_{i=1}^{k} \left(- \log\left(\frac{i}{k+1}\right)\right)^2 - \left(\sum_{i=1}^{k} - \log\left(\frac{i}{k+1}\right)\right)^2} \xrightarrow[n\to\infty]{P} \xi$$

Rk: rate of convergence: $1/(2\alpha^2)$

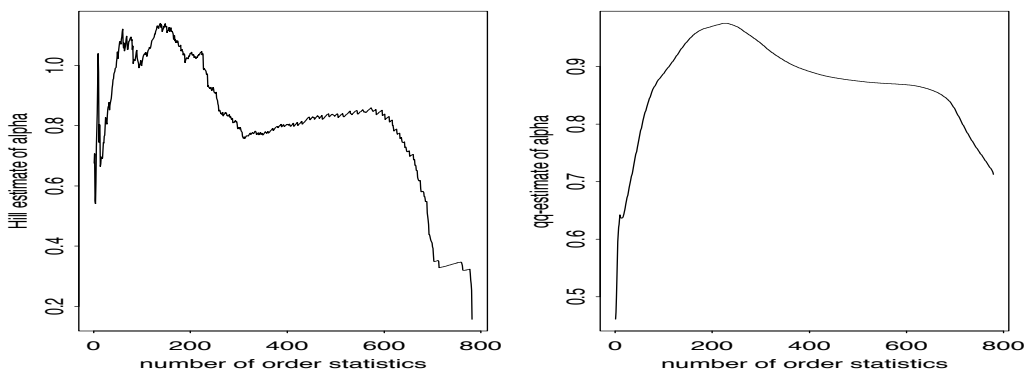
**Illustration**: Comparison of the Hill plot and the QQ–plot of estimates of $\alpha$.

- On Pareto(1) simulated data (sample size $n = 1000$)




The QQ–plot shows $\widehat{\alpha^{-1}} \simeq 0.98$. Seems a bit less volatile than the Hill plot.

- On real data:




The Hill plot is somewhat inconclusive, whereas the QQ–plot indicates a value of about 0.97

▷ the QQ-method in practice:

- Make a QQ–plot of all the data (empirical vs theoretical quantile)
- Choose $k$ based on visual observation of the portion of the graph that looks linear.
- Compute the slope of the line through the chosen upper $k$ order statistics and the corresponding exponential quantiles.

▷ Alternatively, for Hill and QQ methods:

- Plot $\{(k, \widehat{\alpha^{-1}}(k)), 1 \leq k \leq n\}$
- Look for a stable region of the graph as representing the true value of $\alpha^{-1}$.

$\hookrightarrow$ Choosing $k$ with those graphical (supervised) methods is an art as well as a science and the estimate of $\alpha$ is usually rather sensitive to the choice of $k$.

# II - A self-calibrating method for modelling heavy tailed probability distributions (see *N. Debbabi, M. Kratz, M. Mboup*, ESSEC Working Paper 1619 / preprint on arXiv, 2016 )

## Idea of the method

*Frame*: (right) **heavy-tailed continuous** data $\rightarrow$ fit the tail using a **GPD** with a positive tail index (Fréchet domain of attraction)

*Main motivation*: to suggest a **unsupervised method to determine the threshold** above which we fit the GPD, and to have a good fit for the **entire distribution**

*Way*: introduce a **hybrid model** with 3 components (G-E-GPD):

- a Gaussian distribution to model the mean behavior
- a **GPD** for the **tail**
- an exponential distribution to bridge the mean and tail behaviors

*Assumption*: the distribution (which belongs to the Fréchet domain of attraction) has a density that is $C^1$

$$h(x;\theta) = \begin{cases} \gamma_1 \; f(x;\mu,\sigma), & \text{if} \quad x \leq u_1, \\ \gamma_2 \; e(x;\lambda), & \text{if} \quad u_1 \leq x \leq u_2, \\ \gamma_3 \; g(x-u_2;\xi,\beta), & \text{if} \quad x \geq u_2, \end{cases}$$

$f$: Gaussian pdf $(\mu, \sigma^2)$.

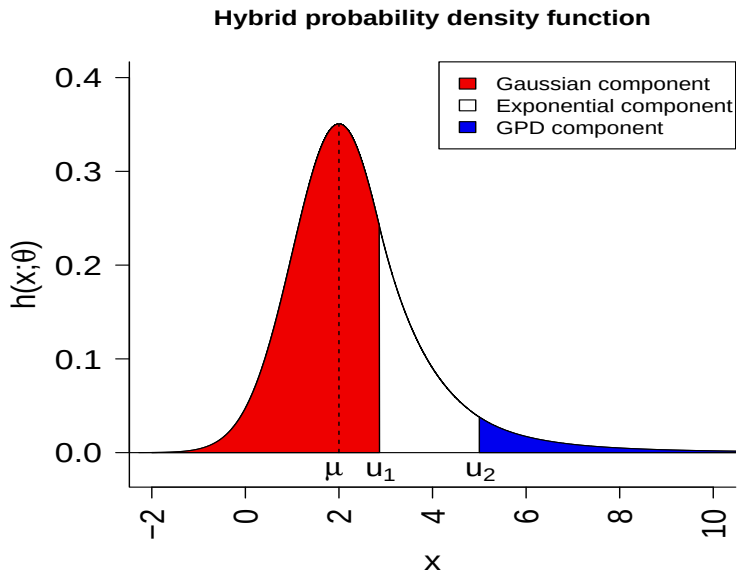
$e$: Exponential pdf with intensity $\lambda$.

$g$: GPD pdf with tail index $\xi$ and scale parameter $\beta$.

$\theta = [\mu, \sigma, u_2, \xi]$: the parameters vector.

$\gamma_1$, $\gamma_2$ and $\gamma_3$: the weights (evaluated from the assumptions (in part. $C^1$) )

$\beta = u_2\xi > 0$; $\lambda = \frac{1+\xi}{\beta}$; $u_1 = \mu + \lambda\sigma^2$



Applications:

- modelling of the **tail only**: determination of the threshold above which the tail distribution is a GPD
- modelling of the **entire distribution**
  - with a G-E-GPD using limit theorems (CLT+Pickands)
  - with a hybrid distribution **including the GPD** and with 1 or 2 other components to be chosen to fit the non-extreme data (if not using the CLT; e.g. lognormal-E-GPD in insurance)
- **general method**: straigthforward to generalize it for data with left and right tails (example: GPD-Gauss-GPD for neural data - proceedings Gretsi 2015)

## Pseudo-code of the algorithm for the G-E-GPD parameters estimation

[1] Initialization of $\widetilde{p}^{(0)} = [\widetilde{\mu}^{(0)}, \widetilde{\sigma}^{(0)}, \widetilde{u}_2^{(0)}]$, $\alpha$, $\varepsilon > 0$, and $k_{max}$, then initialization of $\widetilde{\xi}^{(0)}$ (recall that $\theta = [\mu, \sigma, u_2, \xi]$):

$$\widetilde{\xi}^{(0)} \leftarrow \underset{\xi>0}{argmin} \left\| H(y; \theta \mid \widetilde{p}^{(0)}) - H_n(y) \right\|_2^2,$$

where $H_n$ is the empirical cdf of X (and distance computed on $y = (y_j)_{1 \le j \le m}$).

[2] Iterative process:

- $k \leftarrow 1$

Step 1 - Estimation of $\widetilde{p}^{(k)}$: $\widetilde{p}^{(k)} \leftarrow \underset{\substack{(\mu,\sigma)\in\mathbb{R}\times\mathbb{R}_+^* \\ u_2\in\mathbb{R}_+}}{argmin} \left\| H(y; \theta \mid \widetilde{\xi}^{(k-1)}) - H_n(y) \right\|_2^2$

Step 2 - Estimation of $\widetilde{\xi}^{(k)}$: $\widetilde{\xi}^{(k)} \leftarrow \underset{\xi>0}{argmin} \left\| H(y; \theta \mid \widetilde{p}^{(k)}) - H_n(y) \right\|_2^2,$

- $k \leftarrow k + 1$

until $\Big( d(H(y; \theta^{(k)}), H_n(y)) < \varepsilon$ and $d(H(y_{q_\alpha}; \theta^{(k)}), H_n(y_{q_\alpha})) < \varepsilon \Big)$ or $\big(k = k_{max}\big)$.

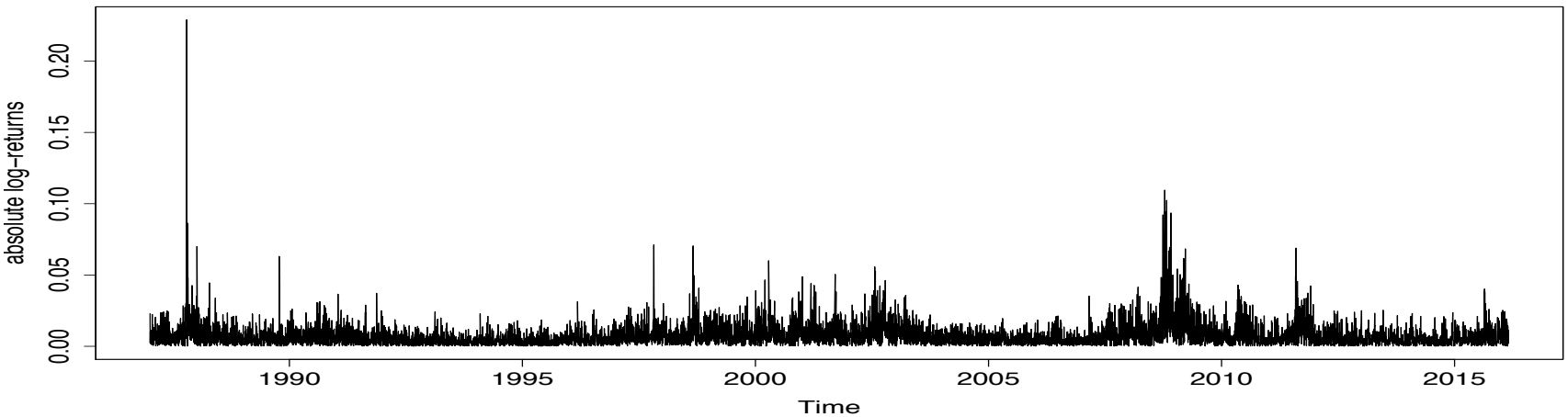
[3] Return $\theta^{(k)} = [\widetilde{\mu}^{(k)}, \widetilde{\sigma}^{(k)}, \widetilde{u}_2^{(k)}, \widetilde{\xi}^{(k)}]$.

# Performance of the method (algorithm) tested via MC simulations. Algorithm also proved to be convergent

1. $\{X^q = (X_p^q)_{1\le p\le n}\}_{1\le q\le N}$: training sets of length $n$
   $\{Y^q = (Y_p^q)_{1\le p\le l}\}_{1\le q\le N}$: test sets of length $l$
   both with G-E-GPD parent distribution, fixed parameters vector $\theta$.
2. On each training set $X^q$, $1 \le q \le N$, evaluate $\widetilde{\theta}^q = [\widetilde{\mu}^q, \widetilde{\sigma}^q, \widetilde{u_2}^q, \widetilde{\xi}^q]$ using our algorithm
3. Compute the empirical mean $\tilde{a}$ and variance $\tilde{S}_a$ of estimates of each parameter $a$ over the $N$ training sets. To evaluate the performance of the estimator, two criteria:
   (i) MSE expressed for any $a$ as: $\text{MSE}_a = \frac{1}{N}\sum_{q=1}^{N}(\widetilde{a}^q - a)^2$.
   A small value of MSE highlights the reliability of parameters estimation using the algorithm
   (ii) Test on the mean (with unknown variance): $\left| \begin{array}{lcl} H0 & : & \widetilde{a} = a \\ H1 & : & \widetilde{a} \neq a \end{array} \right.$
   (use for instance the normal test for a large sample)
4. Compare the hybrid pdf $h$ (with the fixed $\theta$) with the corresponding estimated one $\widetilde{h}$, using $\widetilde{\theta}^q$ on each test set $Y^q$. To do so, compute the average of the log-likelihood function $\mathcal{D}$, over $N$ simulations, between $h(Y^q;\widetilde{\theta}^q)$ and $h(Y^q;\theta)$: $\mathcal{D} = \frac{1}{Nl}\sum_{q=1}^{N}\sum_{p=1}^{l}\log\left(h(Y_p^q;\theta)/\,\widetilde{h}(Y_p^q;\widetilde{\theta}^q)\right)$. The smaller $\mathcal{D}$ is, the better.

# Applications

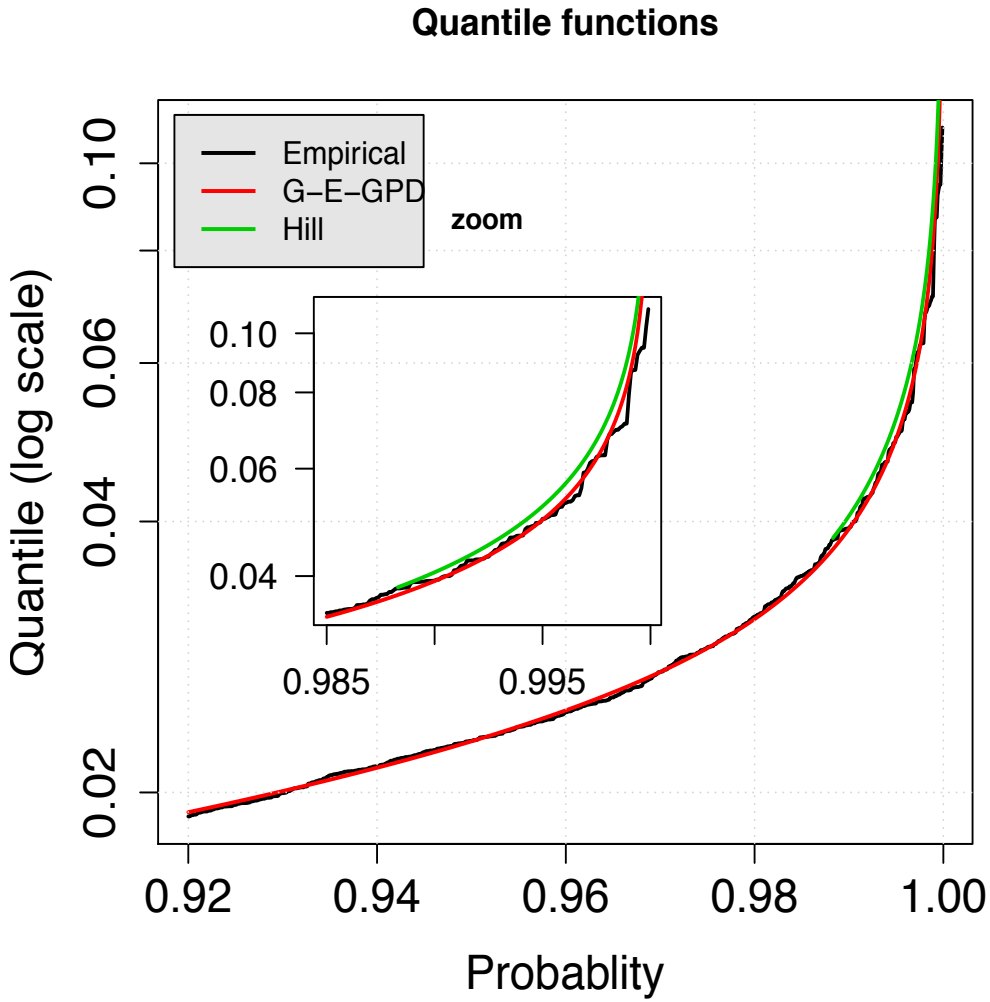
## Example - S&P 500




S&P500 absolute daily log-returns from January 2, 1987 to February 29, 2016. Number of observations: 7349

Comparison between the self-calibrating method and the three graphical methods: MEP, Hill and QQ ones. $N_u$ represents the number of observations above the obtained threshold, and 'distance ' corresponds to the Mean squared Error between the empirical cdf and the estimated one.

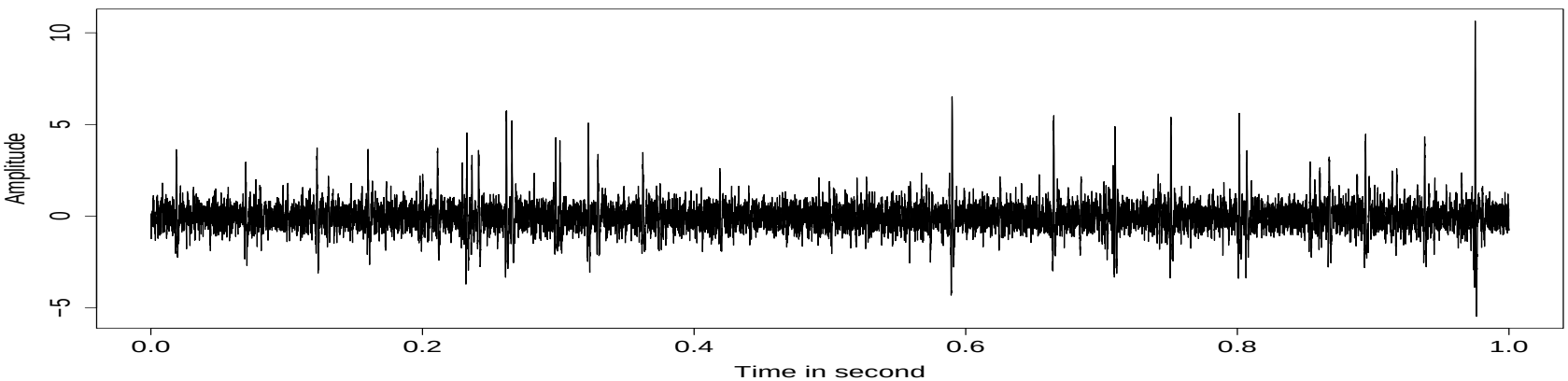
| Model | tail index ($\xi$) | threshold ($u_2$) | $N_u$ | distance (tail distr.) | distance (full distr.) |
|---|---|---|---|---|---|
| GPD | MEP: 0.3025 | $0.0282 = q_{97.21\%}$ | 206 | $1.7811\ 10^{-7}$ | |
| GPD | Hill-estimator: 0.3094 | $0.0382 = q_{98.85\%}$ | 85 | $4.4953\ 10^{-8}$ | |
| GPD | QQ-estimator: 0.3288 | $0.0323 = q_{98.14\%}$ | 137 | $6.0505\ 10^{-8}$ | |
| G-E-GPD | Self-calibrating method: 0.3332 | $0.0289 = q_{97.49\%}$ | 184 | $1.9553\ 10^{-7}$ | $1.0635\ 10^{-5}$ |




Comparison between the estimated quantile functions using our method and the MEP one.

# Example in neuroscience - neural data

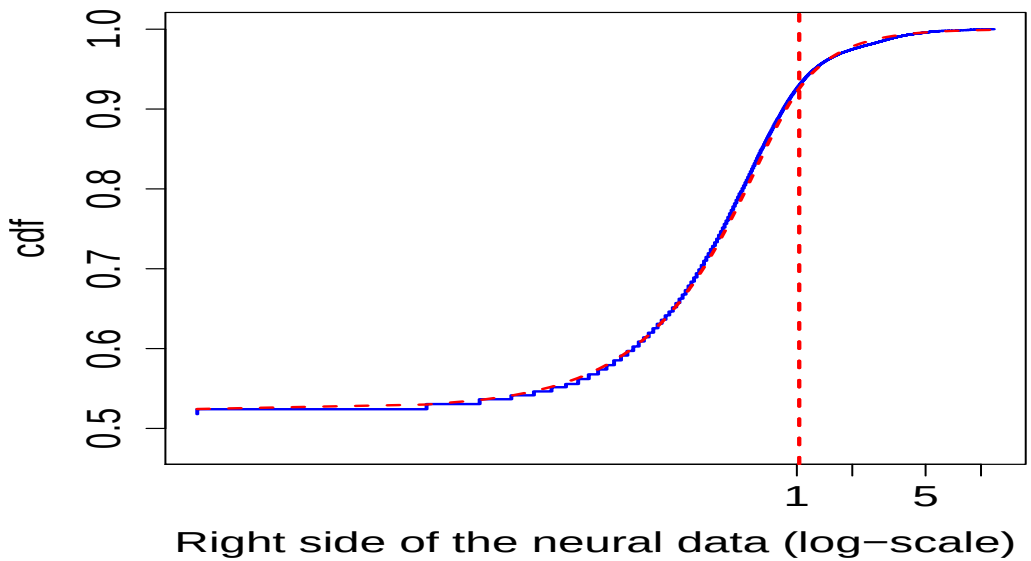
Data corresponding to twenty seconds, equivalent to $n = 3.10^5$ observations, of real extracellular recording of neurons activities. The information to be extracted from these data (spikes or action potentials) lies on the extreme behaviors (left and right) of the data.




One second of neural data, extracellularly recorded.

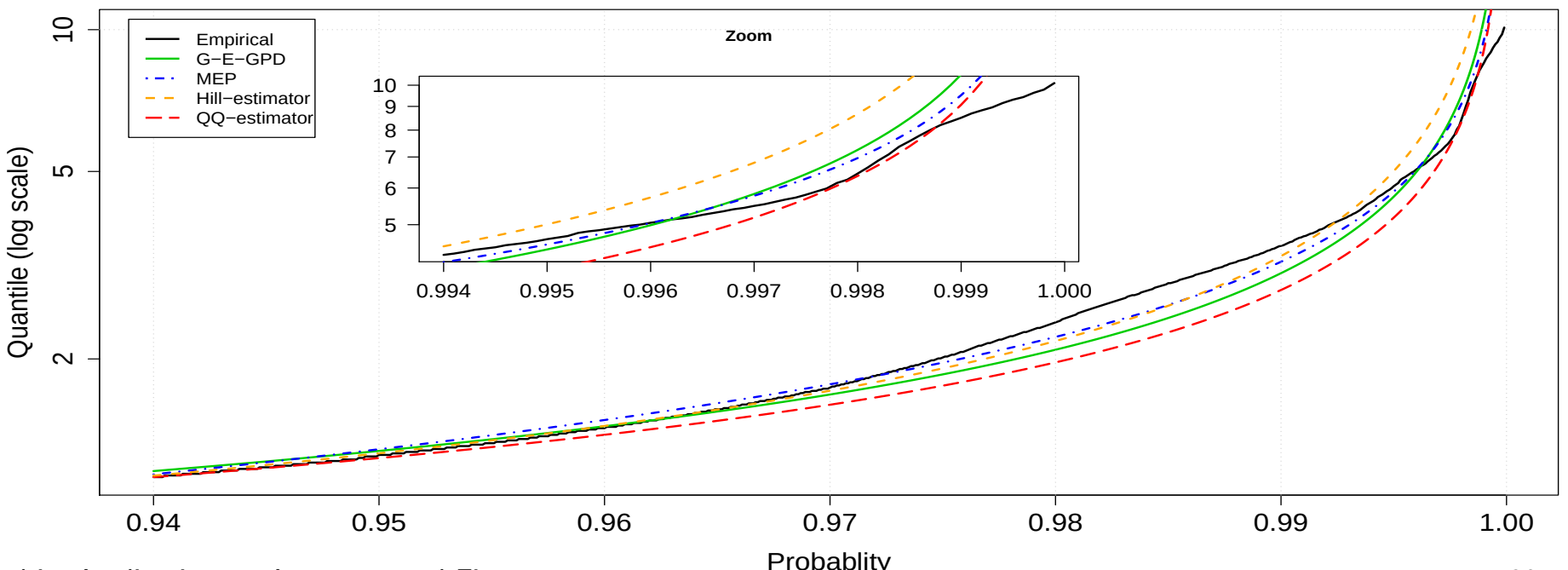
Since the neural data can be considered as symmetric, it is sufficient to evaluate the right side of the distribution with respect to its mode.

Right side of the empirical cdf vs the hybrid cdf obtained with the self-calibrating method



Comparison between the self-calibrating method and the three graphical methods: MEP, Hill and QQ ones. $N_{u_2}$ represents the number of observations above $u_2$. The distance gives the MSE between the empirical (tail or full respectively) distribution and the estimated one from a given model (GPD or hybrid G-E-GPD respectively).

| Model | tail index ($\xi$) | threshold ($u_2$) | $N_{u_2}$ | distance (tail distr.) | distance (full distr.) |
|---|---|---|---|---|---|
| GPD | MEP (PWM): $0.3326$ | $1.0855 = q_{93.64\%}$ | 19260 | $4.0663\ 10^{-6}$ | |
| GPD | Hill-estimator: $0.599$ | $1.0855 = q_{93.64\%}$ | 19260 | $2.0797\ 10^{-6}$ | |
| GPD | QQ-estimator: $0.5104$ | $1.0671 = q_{93.47\%}$ | 19871 | $1.2685\ 10^{-5}$ | |
| G-E-GPD | Self-calibrating method: $0.5398$ | $1.0301 = q_{92.9\%}$ | 21272 | $7.7903\ 10^{-6}$ | $9.3168\ 10^{-5}$ |



# Conclusion
## ▷ Synopsis of the method

- Unsupervised method to model heavy tailed data that may be non-homogeneous and multi-components
- To develop it, we introduced a general hybrid $C^1$ distribution for highly right skewed data modeling, a G-E-GPD that links a Gaussian distribution to a GPD via an exponential distribution that bridges the gap between mean and asymptotic behaviors
- The three distributions are connected to each other at junction points estimated by an iterative algorithm, as are the other parameters of the model. Analytical and numerical study of the convergence of the algorithm
- Performance of the method studied on simulated data. Judicious fit of the asymmetric right heavy tailed data with an accurate determination of the threshold indicating the presence of extremes; good estimation of the parameters of the GPD that fits the extremes over this threshold
- Several applications of the method have been done on real data. Comparison with standard EVT methods.

## ▷ Advantages of the method

- To be unsupervised, avoiding the resort, when fitting the tail, to standard graphical methods (e.g. MEP, Hill, QQ methods) in EVT
- To fit with the same iterative algorithm the full distribution of observed heavy-tailed data, of any type (whenever $C^1$-distribution), providing an accurate estimation of the parameters for the mean and extreme behaviors
- Generality of the method: besides the GPD needed when fitting the heavy tail, the other components might be chosen differently, not using limit behavior (CLT) but distributions chosen specifically for the data that are worked out (as e.g. lognormal for insurance claims). It would not change at all the structure of the algorithm

## ▷ Limits of the method

- Determining in a unsupervised way the threshold over which we have extremes, requires to have information before the threshold. We suggest here an approach that fits the entire distribution.
- Further investigation will follow in order to make this method also available as a pure EVT tool (i.e. to fit the tail only). It means to determine the minimum information required to determine the neighbor distribution of the GPD to have a robust estimation for the tail threshold and the GPD parameters estimation.
- We plan also to tackle the numerical study of the convergence of the algorithm in a analytical way

# III - Popular Risk Measures

- The choice of risk measure has much impact in terms of risk management and model validation.
- Various usages of risk measures
  - The main usage of risk measures is to compute, from the probability distribution of the firm's value, the Risk Adjusted Capital in its different forms:
    - In **insurance**
      1. Solvency Capital Requirements of Solvency II: VaR (99.5% yearly)
      2. Target capital for the Swiss Solvency Test: ES (99% yearly)
    - In **banks**
      1. Basel II: VaR (99% daily)
      2. In the future Basel III: ES (97.5% daily for market risk)
  - Heart of a risk/reward strategy :
    1. to measure the diversification benefit of a risk portfolio
    2. to allow capital allocation among the various risks of the portfolio (very important role of the risk measure to optimize companies value)

- What are the main properties we should expect in practice from a "good" risk measure?

  1. the **subadditivity** i.e. $\rho(L_1 + L_2) \leq \rho(L_1) + \rho(L_2)$ (more generally the *coherence*) and comonotonic additivity (i.e. $\rho(L_1 + L_2) = \rho(L_1) + \rho(L_2)$ with $L_i = f_i(X)$, $f_i \nearrow$), to measure the diversification benefit
  2. good estimates and possibility of **backtesting**

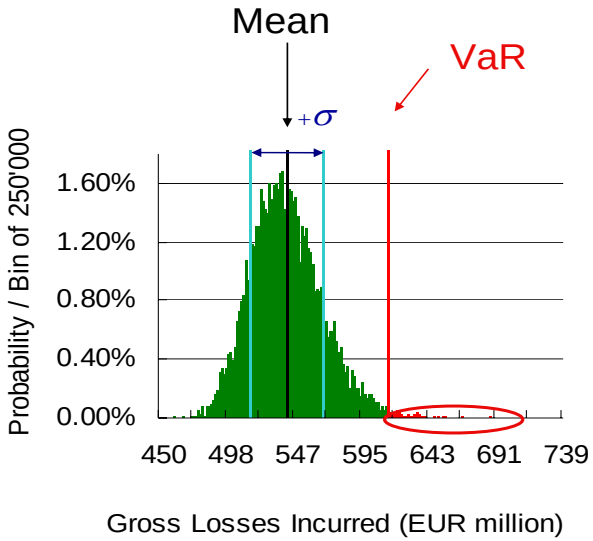
- Popular / regulatory risk measures:

  $\hookrightarrow$ Value-at-Risk ($\text{VaR}_\alpha$) = quantile $q_\alpha(L)$ s.t. $q_\alpha(L) = \inf\{q \in \mathbb{R} : \mathbb{P}(L > q) \leq 1 - \alpha\}$;

  $\hookrightarrow$ Expected Shortfall (ES) =Tail VaR (TVaR) :

$$ES_\alpha(L) = \frac{1}{1-\alpha} \int_\alpha^1 q_\beta(L)\, d\beta \underset{F_L \text{ cont}}{=} \mathrm{E}[L \,|\, L \geq q_\alpha(L)]$$

(may be seen as the average over all losses larger than $\text{VaR}_\alpha$)




*Standard Deviation* measures typical size of fluctuations

*Value-at-Risk (VaR)* measures position of 99th percentile, „happens once in a hundred years“

*Expected Shortfall* (ES) is the weighted average VaR beyond the 1% threshold.

Main advantage of the

- VaR: no constraint on the existence of moments
- ES: coherent risk measure

Main drawback of the VaR:

- The subadditivity fails to hold for VaR **in general**. Contradicts that there should be a diversification benefit associated with merging the portfolios. Cq: a decentralization of risk management using VaR is difficult since we cannot be sure that by aggregating VaR numbers for different portfolios or business units we will obtain a bound for the overall risk of the enterprise.
- VaR gives no information about the severity of losses which occur with a probability less than $1-\alpha$.

# IV - Backtesting Risk Measures

## 1 - VaR

### (a) Optimal point forecast

VaR is *elicited* by the weighted absolute error scoring function

$$s(x,y) = (\mathbf{1}_{\{x \geq y\}} - \alpha)(x-y),\ 0 < \alpha < 1 \text{ fixed}$$

(Thomson (79), Saerens (00), or Gneiting (11) for details)
$\Rightarrow$ VaR : optimal **point** forecast

$\hookrightarrow$ this allows for the *comparison of different forecast methods*.

However, in practice, we have to compare VaR predictions by a single method with observed values, in order to assess the quality of the predictions.

## (b) A popular procedure: a binomial test on the proportion of violations

- Assuming a continuous loss distribution, $\mathbb{P}[L > VaR_\alpha(L)] = 1 - \alpha$
  $\Rightarrow$ the probability of a violation of VaR is $1 - \alpha$

- We define the violation process of VaR as

$$I_t(\alpha) = \mathbf{1}_{\left\{L(t) > VaR_\alpha(L(t))\right\}}.$$

  VaR forecasts are valid iff the violation process $I_t(\alpha)$ satisfies the two conditions (Christoffersen, 03):

  (**i**) $\mathbb{E}[I_t(\alpha)] = 1 - \alpha$ (**ii**) $I_t(\alpha)$ and $I_s(\alpha)$ are independent for $s \neq t$

- Under (i) & (ii), $I_t(\alpha)$'s are iid $\mathcal{B}(1 - \alpha)$ $\Rightarrow$
  $\sum_{t=1}^{n} I_t(\alpha) \overset{d}{\sim} \mathcal{B}(n, 1 - \alpha)$

In practice, it means:

- to estimate the violation process by replacing VaR by its estimates
- check that this process behaves like iid Bernoulli random variables with violation (success) probability $p_0 \simeq 1-\alpha$
- Test on the proportion $p$ of VaR violations,

  estimated by $\frac{1}{n}\sum_{t=1}^{n} I_t(\alpha)$:

  H0: $p = p_0 = 1-\alpha$ against H1: $p > p_0$

If the proportion of VaR violations is not significantly different from $1-\alpha$, then the estimation/prediction method is reasonable.

Note:

- Convenient procedure because it can be performed **straightforwardly** within the algorithms estimating the VaR
- Condition (ii) might be violated in practice $\Rightarrow$ various tests on the independence assumption have been proposed in the literature, as e.g. one developed by Christoffersen and Pelletier (04), based on the *duration of days between the violations of the VaR thresholds*.

## 2 - Expected Shortfall (ES) or TVaR

### (a) Backtesting **distribution** forecasts

Testing the distribution forecasts could be helpful, in particular for tail-based risk measures like ES.

Ex: method for the out-of-sample validation of distribution forecasts, based on the Lévy-Rosenblatt transform, named also Probability Integral Transform (PIT).

See Diebold et al.; based on the Rosenblatt result that $F(X) \stackrel{d}{=} \mathcal{U}(0,1)$; they observed that if a sequence of distribution forecasts coincides with the sequence of unknown conditional laws that have generated the observations, then the sequence of PIT are iid $\mathcal{U}(0,1)$.

Nevertheless, there were still some gaps to fill up before a full implementation and use in practice. Various issues left open studied by Blum (PhD thesis, 04), in part. in situations with overlapping forecast intervals and multiple forecast horizons.

## (b) A component-wise optimal forecast for ES

ES: example of a risk measure whose *conditional elicitability* (see Emmer et al.) provides the possibility to forecast it in two steps.

1. We forecast the quantile ($\text{VaR}_\alpha$) as

$$\hat{q}_\alpha(L) = \arg\min_x E_P[s(x, L)]$$

with $s(x, y) = (\mathbf{1}_{\{x \geq y\}} - \alpha)(x - y)$ strictly consistent scoring function

2. Fixing this value $\hat{q}_\alpha$, $E[L|L \geq \hat{q}_\alpha]$ is just an expected value. Thus we can use strictly consistent scoring function to forecast

$$\text{ES}_\alpha(L) \approx E[L|L \geq \hat{q}_\alpha].$$

If $L$ is $L^2$, the score function can be chosen as the squared error:

$$\widehat{\text{ES}}_\alpha(L) \approx \arg\min_x E_{\tilde{P}}[(x - L)^2] \quad \text{where } \tilde{P}(A) = P(A|L \geq \hat{q}_\alpha).$$

## (c) An implicit backtest for ES: a simple multinomial test

▷ Idea came from the following (Emmer et al.):

$$\begin{aligned} \mathrm{ES}_\alpha(L) &= \frac{1}{1-\alpha}\int_\alpha^1 q_u(L)\,du \\ &\approx \frac{1}{4}\left[q_\alpha(L) + q_{0.75\,\alpha+0.25}(L) + q_{0.5\,\alpha+0.5}(L) + q_{0.25\,\alpha+0.75}(L)\right] \end{aligned}$$

where $q_\alpha(L) = VaR_\alpha(L)$. Hence, if the four $q_{a\alpha+b}(L)$ are successfully backtested, then also the estimate of $\mathrm{ES}_\alpha(L)$ might be considered reliable.

▷ We can then build a backtest based on that intuitive idea of backtesting ES via **simultaneously** backtesting multiple VaR estimates evaluated with the same method as the one used to compute the ES estimate.

Note: the Basel Committee on banking Supervision suggests a variant of this ES-backtesting approach based on testing level violations for two quantiles at 97.5% and 99% level (Jan. 2016).

## V - Building an implicit backtest for ES via a simple multinomial approach (See *M. Kratz, Y. Lok, A. McNeil*, ESSEC working paper 1617 / preprint on arXiv, 2016)

Main questions:

- Does a multinomial test work better than a binomial one for model validation?
- Which particular form of the multinomial test should we use in which situation?
- What is the 'optimal' number of quantiles that should be used for such a test to perform well?

To answer these questions, we build a multi-steps experiment on simulated data:

- ▷ Static view: we test distributional forms (typical for the trading book) to see if the multinomial test distinguishes well between them, in particular between their tails
- ▷ Dynamic view: looking at a time series setup in which the forecaster may misspecify both the conditional distribution of the returns and the form of the dynamics, in different ways.

## A multinomial test
### *Testing set-up*

- We have a series of ex-ante predictive models $\{F_t, t = 1, \ldots, n\}$ and a series of ex-post losses $\{L_t, t = 1, \ldots, n\}$.
- At each time $t$, the model $F_t$ is used to produce estimates (or forecasts) of $VaR_{\alpha,t}$ and $ES_{\alpha,t}$ at various probability levels $\alpha$.
- The VaR estimates are compared with $L_t$ to assess the adequacy of the models in describing the losses, with particular emphasis on the most extreme losses.

We generalize the idea of Emmer et al. by considering VaR probability levels $\alpha_1, \ldots, \alpha_N$ defined, for some starting level $\alpha$, by

$$\alpha_j = \alpha + \frac{j-1}{N}(1-\alpha), \quad j = 1, \ldots, N.$$

We set, for $N > 1$, $\alpha = 0.975$ (level used for ES calculation, and lowest of the two levels used for backtesting under the Basel rules for banks) and for $N = 1$, $\alpha = 0.99$ (usual level for binomial tests of VaR exceptions).

We also set $\alpha_0 = 0$ and $\alpha_{N+1} = 1$.

Testing simultaneously $N$ VaR's (with $N > 1$) leads to a multinomial distribution; we can set the null hypothesis of the multinomial test as

$$(\text{H0}): \ p_j := \mathbb{E}[1_{(L_t > VaR_{j,t})}] (= \mathbb{P}[L_t > VaR_{j,t}]) \ = \ p_{j,0} := 1 - \alpha_j, \quad \forall\, j = 1,$$

Assuming the $n$ observations come from a loss variable $L$ with continuous distribution $F$, introduce the observed cell counts between quantile levels $q_\alpha = F^{\leftarrow}(\alpha)$ as $O_j = \sum_{t=1}^{n} I_{(q_{j-1} < L_t \leq q_j)}$, for $j = 1, ..., N+1$.

Then $(O_1, \ldots, O_{N+1})$ follows a mutinomial distribution:

$$(O_1, \ldots, O_{N+1}) \sim \text{MN}(\beta_1 - \beta_0, \ldots, \beta_{N+1} - \beta_N)$$

for parameters $\beta_1 < \cdots < \beta_N$ with $\beta_0 = 0$ and $\beta_{N+1} = 1$.
Hence the test can be rewriten as

$$\left| \begin{array}{lll} \text{H0}: & \beta_j & = \alpha_j \quad \text{for } j = 1, \ldots, N \\ \text{H1}: & \beta_j & \neq \alpha_j \quad \text{for at least one } j \in \{1, \ldots, N\}. \end{array} \right.$$

To judge the relevance of the test, compute :
its size $\gamma = \mathbb{P}(\text{reject H0}) | \text{H0 true}]$ (type I error)
and its power $1 - \beta = 1 - \mathbb{P}[(\text{accept H0}) | \text{H0 wrong}]$ (1 - type II error).

- Checking the size of the multinomial test: straightforward, by simulating data from a multinomial distribution under the null hypothesis (H0). This can be done by simulating data from any distribution (such as normal) and counting the observations between the true values of the $\alpha_j$-quantiles, or simulating from the multinomial distribution directly.

- To calculate the power: we have to simulate data from multinomial models under the alternative hyp. (H1). Here we chose to simulate from models coming from a distribution $G$, having heavy tails and possibly skewness, with $G \neq F$ (true distr.), where the parameters are given by

$$\beta_j = F\left(G^{\leftarrow}(\alpha_j)\right), \quad \text{with } \beta_j \neq \alpha_j\,.$$

*Example*:
$F=$ true distribution of $L_t$, so that the true quantiles $= F^{\leftarrow}(\alpha_j)$. However a modeller chooses the wrong distribution $G$ and makes estimates $G^{\leftarrow}(\alpha_j)$ of the quantiles. The probabilities associated with these quantile estimates are s.t. $\beta_j = F(G^{\leftarrow}(\alpha_j)) \neq \alpha_j$.

This test is closely linked to the PIT method for the backtest of a probability distribution forecast. Checking that, $\forall j$, $\beta_j = F(G^{\leftarrow}(\alpha_j)) = \alpha_j$, comes back to check that $G(X_j)$ is uniformly distributed, with $X_j \sim F$, with known realizations $x_j$ (for PIT method, we would do it for all the known realizations). Kind of PIT test.

Various test statistics can be used to describe the event (reject of H0) (see e.g. Cai and Krishnamoorthy for five possible tests for testing the multinomial proportions). Here we use, for comparison,

- the Pearson chi-square: $S_N = \sum_{j=0}^{N} \frac{(O_j - n(\alpha_{j+1} - \alpha_j))^2}{n(\alpha_{j+1} - \alpha_j)} \underset{H0}{\overset{d}{\sim}} \chi^2_N$

  and two of its possible modifications:
- the Nass test
- the LR (asymptotic Likelihood Ratio test)

## ▷ Multi-steps experiment: Static view

- Simulate multinomial data where $F$ is normal (benchmark) and $G$ of various types: $t5$, $t3$ and skewed $t3$
- Count the simulated observations lying between the $N$ quantiles of $G$, where $N = 1, 2, 4, 8, 16, 32, 64$
- Choose different lengths $n_1$ for the sample of backtest, namely $n_1 = 250, 500, 1000, 2000$, and estimate the rejection probability for the null hypothesis (H0) using 10 000 replications (changing seeds)
- Why introducing several quantiles (ES) ?

| | $VaR_{0.975}$ | $VaR_{0.99}$ | $\Delta_1$ | $ES_{0.975}$ | $\Delta_2$ |
|---|---|---|---|---|---|
| Normal | 1.96 | 2.33 | 0.00 | 2.34 | 0.00 |
| t5 | 1.99 | 2.61 | 12.04 | 2.73 | 16.68 |
| t3 | 1.84 | 2.62 | 12.69 | 2.91 | 24.46 |
| st3 ($\gamma = 1.2$) | 2.04 | 2.99 | 28.68 | 3.35 | 43.11 |

Values of $VaR_{0.975}$, $VaR_{0.99}$ and $ES_{0.975}$ for four distributions (mean0, var 1) used in simulation study (Normal, Student t5, Student t3, skewed Student t3 with skewness parameter $\gamma = 1.2$). $\Delta_1$ column shows percentage increase in $VaR_{0.99}$ compared with normal distribution; $\Delta_2$ column shows percentage increase in $ES_{0.975}$ compared with normal distribution.

Rejection rate for the null hypothesis (H0) on a sample size of length $n_1$, using a multinomial approach with 3 possible tests ($\chi^2$, Nass, LR) to backtest simultaneously the $N = 2^k$, $1 \leq k \leq 6$, quantiles $\text{VaR}_{\alpha_j}$, $1 \leq j \leq N$, with $\alpha_1 = \alpha = 97.5\%$, on data simulated from various distributions (normal, Student $t3$, $t5$ and skewed $t3$)

| | test | Pearson | | | | | | | Nass | | | | | | | LRT | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $G$ | $n \mid N$ | 1 | 2 | 4 | 8 | 16 | 32 | 64 | 1 | 2 | 4 | 8 | 16 | 32 | 64 | 1 | 2 | 4 | 8 | 16 | 32 | 64 |
| Normal | 250 | 3.9 | 4.7 | 5.6 | 8.5 | 10.5 | 14.1 | 21.5 | 3.9 | 3.5 | 5.0 | 4.7 | 5.1 | 5.0 | 4.8 | 7.5 | 10.0 | 6.5 | 6.5 | 6.5 | 6.2 | 6.1 |
| | 500 | 3.9 | 4.4 | 5.2 | 6.6 | 8.6 | 12.3 | 16.2 | 3.9 | 3.9 | 4.7 | 4.7 | 5.5 | 5.5 | 5.3 | 5.9 | 5.8 | 5.5 | 5.6 | 5.3 | 5.3 | 5.2 |
| | 1000 | 5.0 | 5.2 | 5.0 | 5.6 | 7.2 | 9.0 | 12.0 | 5.0 | 4.8 | 4.7 | 4.9 | 5.1 | 5.3 | 5.1 | 4.1 | 5.5 | 5.5 | 5.8 | 5.6 | 5.6 | 5.7 |
| | 2000 | 5.0 | 4.5 | 4.8 | 5.0 | 6.3 | 7.2 | 8.8 | 5.0 | 4.3 | 4.5 | 4.5 | 5.3 | 5.1 | 4.9 | 4.2 | 4.9 | 4.7 | 5.0 | 5.1 | 5.1 | 5.0 |
| t5 | 250 | 4.1 | 10.2 | 14.1 | 20.8 | 22.4 | 27.0 | 34.2 | 4.1 | 7.7 | 12.8 | 14.1 | 13.4 | 14.4 | 13.0 | 6.9 | 14.4 | 15.8 | 21.6 | 26.6 | 30.7 | 33.7 |
| | 500 | 5.2 | 15.7 | 22.1 | 28.4 | 32.2 | 36.2 | 39.8 | 5.2 | 14.3 | 20.5 | 24.5 | 26.6 | 26.0 | 22.7 | 6.5 | 15.5 | 26.9 | 36.6 | 44.7 | 50.4 | 54.8 |
| | 1000 | 6.9 | 26.7 | 40.2 | 48.2 | 53.0 | 54.8 | 55.8 | 6.9 | 25.5 | 39.5 | 46.2 | 48.6 | 47.7 | 43.8 | 5.2 | 26.1 | 46.4 | 61.8 | 71.4 | 76.7 | 80.5 |
| | 2000 | 7.3 | 47.2 | 70.4 | 79.3 | 82.5 | 82.8 | 82.0 | 7.3 | 47.0 | 69.6 | 78.2 | 80.8 | 80.2 | 77.0 | 5.8 | 48.0 | 77.4 | 89.5 | 94.4 | 96.6 | 97.6 |
| t3 | 250 | 3.6 | 7.3 | 13.7 | 21.1 | 19.4 | 25.8 | 28.1 | 3.6 | 5.6 | 12.1 | 14.8 | 13.4 | 13.2 | 13.6 | 10.3 | 24.4 | 24.4 | 35.4 | 43.2 | 48.0 | 51.9 |
| | 500 | 4.8 | 16.1 | 25.2 | 32.7 | 35.2 | 40.1 | 38.6 | 4.8 | 15.5 | 22.4 | 28.7 | 32.3 | 29.4 | 26.4 | 9.5 | 26.2 | 44.2 | 58.6 | 67.9 | 73.8 | 78.0 |
| | 1000 | 9.9 | 37.4 | 55.6 | 62.9 | 65.2 | 64.8 | 64.2 | 9.9 | 35.2 | 54.1 | 60.3 | 61.4 | 59.9 | 54.7 | 9.7 | 47.2 | 75.4 | 87.7 | 93.2 | 95.5 | 96.8 |
| | 2000 | 16.6 | 73.1 | 91.0 | 94.5 | 94.9 | 93.9 | 92.1 | 16.6 | 72.7 | 90.5 | 94.2 | 94.3 | 92.6 | 89.6 | 16.5 | 79.5 | 96.8 | 99.4 | 99.8 | 99.9 | 100.0 |
| st3 | 250 | 5.4 | 18.9 | 28.8 | 40.0 | 38.7 | 46.3 | 50.5 | 5.4 | 15.3 | 26.3 | 30.5 | 30.2 | 30.5 | 30.7 | 8.0 | 24.6 | 33.5 | 46.5 | 55.1 | 60.8 | 65.4 |
| | 500 | 6.9 | 34.9 | 50.7 | 60.6 | 64.6 | 69.5 | 70.2 | 6.9 | 33.2 | 47.6 | 56.2 | 61.4 | 60.0 | 56.8 | 7.9 | 35.9 | 59.3 | 73.6 | 81.6 | 86.2 | 88.9 |
| | 1000 | 9.5 | 62.3 | 83.0 | 89.1 | 91.3 | 92.1 | 92.0 | 9.5 | 61.4 | 82.3 | 88.1 | 90.0 | 90.0 | 87.9 | 6.9 | 62.3 | 88.1 | 95.3 | 97.9 | 98.9 | 99.2 |
| | 2000 | 12.2 | 90.7 | 98.7 | 99.7 | 99.8 | 99.8 | 99.7 | 12.2 | 90.7 | 98.6 | 99.7 | 99.7 | 99.7 | 99.5 | 9.8 | 91.6 | 99.3 | 99.9 | 100.0 | 100.0 | 100.0 |

*Synopsis for the static view*

- For all non normal distributions, considering only the VaR (1 point) does not reject the normal hypothesis, for all tests. The VaR does not capture enough the heaviness of the tail. Taking 2 quantiles improves slightly but not enough. The mulinomial approach with $N \geq 4$ gives certainly much better results than the traditional binomial backtest

- The heavier the tail of the tested distribution, the more powerful is the multinomial test

- For all the distributions, increasing the number $n_1$ of observations improves the power of all tests

- The Nass test with $N = 4$ or $8$ seems to be a good compromise between an acceptable size and power and to be slightly preferable to the Pearson test with $N = 4$.

- In comparison with Nass, the LRT with $N = 4$ or $N = 8$ is a little oversized but very powerful.

- If obtaining power to reject bad models is the overriding concern, then the LRT with $N > 8$ is extremely effective

Determining an 'optimal' $N$, s.t. $N$ the smallest possible to provide a combination of reasonable size and power of the backtest (to have a backtest comparable with the one of the VaR in terms of simplicity and speed of procedure):

- Select $N$ s.t. the size of the 3 corresponding tests lies below 6%.
- For $n_1 \geq 500$, the size varies between 4.2% and our threshold 6%. For the first two tests (chi-square and Nass), the size increases with $N$, whereas, for the LRT, it is more or less stable (slightly nonincreasing with increasing $N$)
- The power increases with $N$ and the sample size $n_1$, for the 3 tests. It makes sense: the more information we have in the tail, the easier it is to distinguish between light and heavy tails

$\hookrightarrow$ $N = 4$ or $8$ : overall reasonable choice.

| $G$ | $n$ \| test | Bin (0.99) | Pearson (4) | Nass (4) | LRT (4) | LRT (8) |
|---|---|---|---|---|---|---|
| Normal | 250 | 4.0 | 5.6 | 5.0 | 6.5 | 6.5 |
| | 500 | 3.7 | 5.2 | 4.7 | 5.5 | 5.6 |
| | 1000 | 3.8 | 5.0 | 4.7 | 5.5 | 5.8 |
| | 2000 | 5.4 | 4.8 | 4.5 | 4.7 | 5.0 |
| t5 | 250 | 17.7 | 14.1 | 12.8 | 15.8 | 21.6 |
| | 500 | 22.4 | 22.1 | 20.5 | 26.9 | 36.6 |
| | 1000 | 33.0 | 40.2 | 39.5 | 46.4 | 61.8 |
| | 2000 | 59.9 | 70.4 | 69.6 | 77.4 | 89.5 |
| t3 | 250 | 13.5 | 13.7 | 12.1 | 24.4 | 35.4 |
| | 500 | 16.2 | 25.2 | 22.4 | 44.2 | 58.6 |
| | 1000 | 22.3 | 55.6 | 54.1 | 75.4 | 87.7 |
| | 2000 | 41.4 | 91.0 | 90.5 | 96.8 | 99.4 |
| st3 | 250 | 31.2 | 28.8 | 26.3 | 33.5 | 46.5 |
| | 500 | 44.2 | 50.7 | 47.6 | 59.3 | 73.6 |
| | 1000 | 66.2 | 83.0 | 82.3 | 88.1 | 95.3 |
| | 2000 | 92.9 | 98.7 | 98.6 | 99.3 | 99.9 |

*Table 4: Comparison of estimated size and power of one-sided binomial score test with $\alpha = 0.99$ and Pearson, Nass and likelihood-ratio test with $N = 4$ and LRT with $N = 8$. Results are based on 10000 replications*

Results from binomial tests are much more sensitive to the choice of $\alpha$. We have seen before that their performance for $\alpha = 0.975$ is very poor. The multinomial tests using a range of thresholds are much less sensitive to the exact choice of these thresholds, which makes them a more reliable type of test.

## Other experimental design (static view)

The style of backtest is designed to mimic the procedure used in practice where models are continually updated to use the latest market data. We assume that the estimated model is updated every 10 steps.

In each experiment we generate a total dataset of $n + n_2$ values from the true distribution $G$; we use the same four choices as in the previous section. The length $n$ of the backtest is fixed at the value 1000.

The modeller uses a rolling window of $n_2$ values to obtain an estimated cdf $F$, with $n_2$=250, 500 respectively. We consider 4 possibilities for $F$:

**The oracle** who knows the correct distribution and its exact parameter values.

**The good modeller** who estimates the correct type of distribution (normal when $G$ is normal, Student t when $G$ is t5 or t3, skewed Student when $G$ is st3).

**The poor modeller** who always estimates a normal distribution (which is satisfactory only when $G$ is normal).

**The industry modeller** who uses the empirical distribution function by forming standard empirical quantile estimates, a method known as *historical simulation* in industry.

*Results:*

- Again clear that taking values of $N \geq 4$ gives reliable results, superior to those obtained when $N = 1$ or $N = 2$.
- The use of only one or two quantile estimates does not seem sufficient
  - $\hookrightarrow$ to discriminate between light and heavy tails
  - $\hookrightarrow$ a fortiori to construct an implicit backtest of ES based on $N$ VaR levels.

## ▷ Multi-steps experiment: Dynamic view

Backtesting experiment conducted now in a time-series setup. The true data-generating mechanism for the losses is a stationary GARCH(1,1) model with Student innovations (parameters chosen by fitting this model to S&P index log-returns for the period 2000–2012 (3389 values)).

A variety of forecasters use different methods to estimate the conditional distribution of the losses at each time point and deliver VaR estimates.

Length of the backtest: $n = 1000$ (approximately 4 years)

Each forecaster uses a rolling window of $n_2$ values to make their forecasts.

We consider the values $n_2 = 500, 1000$ respectively (longer window lengths than in the static backtest study since more data is generally needed to estimate a GARCH model reliably). All models are re-estimated every 10 time steps. *Experiment repeated* $500$ *times to determine rejection rates for each forecaster.*

**Oracle:** the forecaster knows the correct model and its exact parameter values.

**GARCH.t:** the forecaster estimates the correct type of model (GARCH(1,1)- $t$).

**GARCH.HS:** the forecaster uses a GARCH(1,1) model to estimate the dynamics of the losses, but applies empirical quantile estimation to the residuals to estimate quantiles of the innovation distribution and hence quantiles of the conditional loss distribution (method called 'filtered historical simulation')

**GARCH.EVT:** the forecaster uses a variant on GARCH.HS in which an EVT tail model is used to get slightly more accurate estimates of conditional quantiles in small samples.

**GARCH.norm:** the forecaster estimates a GARCH(1,1) model with normal innovation distribution.

**ARCH.t:** the forecaster misspecifies the dynamics of the losses by choosing an ARCH(1) model but correctly guesses that the innovations are $t$-distributed.

**ARCH.norm:** as in GARCH.norm but the forecaster misspecifies the dynamics to be ARCH(1).

**HS:** the forecaster applies standard empirical quantile estimation to the data. As well as completely neglecting the dynamics of market losses, this method is prone to the drawbacks of empirical quantile estimation in small samples.

Estimated size and power of three different types of multinomial test (Pearson, Nass, likelihood-ratio test (LRT)) based on exceptions of $N$ levels. Results are based on 500 replications of backtests of length 1000

| | Test | Pearson | | | | | | | Nass | | | | | | | LRT | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $n_2$ | $F \mid N$ | 1 | 2 | 4 | 8 | 16 | 32 | 64 | 1 | 2 | 4 | 8 | 16 | 32 | 64 | 1 | 2 | 4 | 8 | 16 | 32 | 64 |
| 500 | Oracle | 6.0 | 4.0 | 3.8 | 5.0 | 5.6 | 9.6 | 9.4 | 6.0 | 3.2 | 3.6 | 4.2 | 2.8 | 4.8 | 4.0 | 3.4 | 4.8 | 5.2 | 5.0 | 5.4 | 5.6 | 5.6 |
| | GARCH.t | 6.8 | 5.6 | 6.2 | 8.0 | 8.2 | 12.8 | 18.4 | 6.8 | 5.0 | 6.0 | 6.2 | 7.0 | 7.6 | 6.4 | 4.6 | 5.0 | 5.4 | 4.8 | 4.6 | 5.2 | 5.2 |
| | GARCH.HS | 1.6 | 1.6 | 4.4 | 11.8 | 25.4 | 92.0 | 98.8 | 1.6 | 1.4 | 4.4 | 10.8 | 20.4 | 85.0 | 97.4 | 0.8 | 1.6 | 3.6 | 2.0 | 2.0 | 5.6 | 13.0 |
| | GARCH.EVT | 2.2 | 3.6 | 3.6 | 7.2 | 7.6 | 12.2 | 16.2 | 2.2 | 3.6 | 3.2 | 6.0 | 5.0 | 6.8 | 7.4 | 0.8 | 3.6 | 2.0 | 0.8 | 1.2 | 1.0 | 1.0 |
| | GARCH.norm | 10.8 | 34.0 | 50.4 | 61.6 | 66.0 | 68.6 | 69.4 | 10.8 | 32.2 | 49.4 | 60.0 | 61.4 | 63.2 | 55.4 | 8.2 | 34.0 | 55.2 | 71.2 | 79.8 | 85.0 | 87.4 |
| | ARCH.t | 34.0 | 32.4 | 32.0 | 29.8 | 29.4 | 33.8 | 39.6 | 34.0 | 31.4 | 31.2 | 28.6 | 26.8 | 25.6 | 28.0 | 30.4 | 31.2 | 31.4 | 31.6 | 31.8 | 31.8 | 31.2 |
| | ARCH.norm | 96.2 | 99.6 | 99.6 | 99.8 | 100.0 | 100.0 | 100.0 | 96.2 | 99.6 | 99.6 | 99.8 | 100.0 | 100.0 | 100.0 | 95.0 | 99.6 | 99.6 | 99.8 | 100.0 | 100.0 | 100.0 |
| | HS | 39.4 | 38.8 | 39.8 | 42.2 | 49.8 | 80.2 | 90.0 | 39.4 | 38.6 | 39.8 | 40.8 | 48.0 | 77.0 | 85.0 | 36.8 | 40.0 | 44.8 | 43.8 | 42.2 | 49.2 | 55.8 |
| 1000 | Oracle | 4.2 | 3.4 | 3.8 | 3.4 | 5.0 | 7.6 | 10.2 | 4.2 | 3.2 | 3.8 | 2.8 | 3.6 | 3.8 | 3.8 | 3.4 | 2.6 | 3.2 | 2.6 | 2.6 | 2.4 | 2.4 |
| | GARCH.t | 5.8 | 4.6 | 6.2 | 5.2 | 6.0 | 11.2 | 12.8 | 5.8 | 3.8 | 5.2 | 3.2 | 4.2 | 6.6 | 6.6 | 4.4 | 2.8 | 3.6 | 3.6 | 3.4 | 4.8 | 4.0 |
| | GARCH.HS | 3.0 | 2.0 | 2.6 | 4.4 | 10.2 | 21.2 | 69.0 | 3.0 | 1.8 | 2.2 | 4.0 | 7.2 | 13.2 | 56.0 | 1.8 | 1.6 | 2.6 | 3.4 | 3.8 | 3.0 | 5.2 |
| | GARCH.EVT | 2.6 | 3.4 | 4.2 | 4.2 | 7.0 | 7.2 | 10.4 | 2.6 | 3.4 | 3.6 | 3.4 | 5.0 | 3.8 | 4.2 | 1.6 | 4.6 | 3.2 | 2.6 | 2.0 | 1.8 | 1.8 |
| | GARCH.norm | 9.4 | 30.6 | 45.6 | 52.2 | 58.4 | 61.4 | 63.6 | 9.4 | 29.8 | 44.6 | 49.6 | 53.0 | 54.6 | 50.6 | 6.4 | 28.4 | 49.8 | 65.2 | 76.6 | 83.0 | 86.6 |
| | ARCH.t | 42.4 | 36.8 | 32.8 | 28.0 | 25.0 | 30.2 | 33.8 | 42.4 | 36.0 | 32.2 | 27.0 | 23.0 | 25.6 | 27.2 | 39.4 | 40.2 | 39.6 | 40.0 | 40.2 | 40.4 | 40.8 |
| | ARCH.norm | 82.8 | 94.6 | 97.6 | 98.2 | 98.6 | 98.2 | 98.6 | 82.8 | 94.4 | 97.6 | 98.0 | 98.2 | 98.0 | 97.8 | 80.8 | 95.2 | 98.8 | 98.8 | 99.2 | 99.2 | 99.4 |
| | HS | 51.4 | 51.0 | 45.0 | 37.2 | 34.6 | 39.8 | 55.6 | 51.4 | 50.6 | 44.4 | 35.0 | 31.8 | 36.2 | 49.8 | 49.2 | 51.8 | 52.6 | 53.8 | 53.8 | 53.0 | 55.4 |

To conclude, this experiment confirms that using $N = 4$ or 8 quantiles gives an effective multinomial test; $N = 4$ is appropriate if using a Pearson or Nass tests and $N = 8$ gives superior power if using the LRT.

## A procedure to implicitely backtest ES

▷ In view of the numerical results we can suggest an 'optimal' multinomial test.

- *Number of quantiles taken at intervals such that $\mathbb{E}(O_j)$ is constant; it turns out that choosing $N = 4$ seems adequate, and $N = 8$ in LRT is the most powerful.*
- *Among the 3 possible tests, the Nass test and the LRT share on average the best results, taking into account both the test size and power, the Nass for the static view and the LRT for the dynamic one. The LRT is in general the most powerful and might be used if we want more sensitivity, in particular w.r.t. the parameters.*

▷ The ES estimated with a model that is not rejected by our multinomial test, is implicitely accepted by our backtest. Hence we can use the same rejection criterion for ES as for the null hypothesis (H0) in the multinomial test.

▷ A traffic light system has been proposed recently by the Basel Committee for Banking Supervision (Jan.2016) based on the backtest of two quantiles, to improve the binomial approach for one quantile. We can use a similar metaphor to illustrate the decision criterion about validating or not the ES estimate, and compare it to the binomial ($N = 1$) or $N = 2$ approaches.

## Conclusion

- We developed several variants of a multinomial test to simultaneously judge the backtesting performance of trading book models at different VaR levels; it gives then an implicit backtest for ES.
- Evaluation of this multinomial approach in a series of Monte Carlo simulation studies of size and power, and further experiments that replicate typical conditions of an industry backtest. It aims as understanding better the test itself and set a benchmark; it has been carried out on real data (one example on S&P500; see the paper)
- The multinomial test distinguishes much better between good and bad models (particularly in longer backtests) than:
  - the standard binomial exception test
  - a multinomial test based on two quantiles

- Backtesting simultaneously 4 or 8 (for LRT) quantile levels seems an optimal choice whatever is the test in terms of balancing simplicity and reasonable size and power properties
- This multinomial backtest could be used for ES as a regular routine, as done usually for the VaR with the binomial backtest, giving even more arguments to move from VaR to ES in the future Basel III.
- Possible to design a traffic-light system for the application of capital multipliers and the imposition of regulatory interventions, completely analogous to the current traffic-light system based on VaR exceptions over a 250 day period at the 99% level.
- We would suggest moving to longer backtesting periods than 250 days to obtain more powerful discrimination between good and bad backtesting results.
- For sharper results (but not for a daily routine), other backtests may complement this one, as the PIT already used for distribution forecasts, or methods based on realized $p$-values, or e.g. joint testing procedures of expected shortfall and VaR proposed by Acerbi and Székely

## *Main references:*

BCBS (2016). *Standards. Minimum capital requirements for market risk*. Basel Committee on Banking Supervision, January 2016.

Y. Cai, K. Krishnamoorthy (2006). Exact size and power properties of five tests for multinomial proportions. *Comm. Statistics - Simulation and Computation* **35(1)**, 149-160.

S.D. Campbell (2006). A review of Backtesting and Backtesting Procedures. *Journal of Risk* **9(2)**, 1-17.

N. Debbabi, M. Kratz, M. Mboup (2016). A self-calibrating method for heavy tailed data modeling. Application in neuroscience and finance. *ESSEC Working Paper 1619 & arXiv1612.03974*

P. Embrechts, C. Klüppelberg, T. Mikosch (1997). Modelling Extremal Events for Insurance and Finance. *Springer-Verlag*

S. Emmer, M. Kratz, D. Tasche (2015). What is the best risk measure in practice? A comparison of standard measures. *Journal of Risk* **18**, 31-60.

M. Kratz, Y. Lok, A. McNeil (2016). A multinomial test to discriminate between models. *ASTIN 2016 proceedings*

M. Kratz, Y.H. Lok, A. McNeil (2016). Multinomial VaR Backtests: A simple implicit approach to backtesting expected shortfall. *ESSEC Working Paper 1617 & arXiv1611.04851*

M. Kratz and S. Resnick (1996). The qq-estimator and heavy tails. *Comm. Statistics - Stochastic Models* **12**(4), 699-724.

A. McNeil, R.Frey, P. Embrechts (2015, 2nd Ed.). Quantitative Risk Management. *Princeton Univ. Press*

S. Resnick (1987; 2nd Ed. 2008). Extreme Values, Regular Variation, and Point Processes. *Springer-Verlag*

S. Resnick (2007). Heavy-tail Phenomena: Probabilistic and Statistical Modeling. *Springer Science & Business Media*