On the Estimation of conditional quantiles.

Véronique Maume-Deschamps
contains joint works with Kevin Elie-Dit-Cosaque, Didier Rullière and Antoine Usseglio-Carleve.

Mascotnum 2021 meeting

Lyon 1
Math Institut Camille Jordan

Plan

Plan

1. Introduction
   - Why conditional quantiles?
   - Quantile Oriented Sensitivity indices
   - Computing / estimating conditional quantiles
2. Elliptic distributions
3. Random forest estimation
4. Conclusions

Quantiles

Quantile is widely used as a risk measure (VaR). Recall: for X a random variable (a risk) with distribution function F_X,

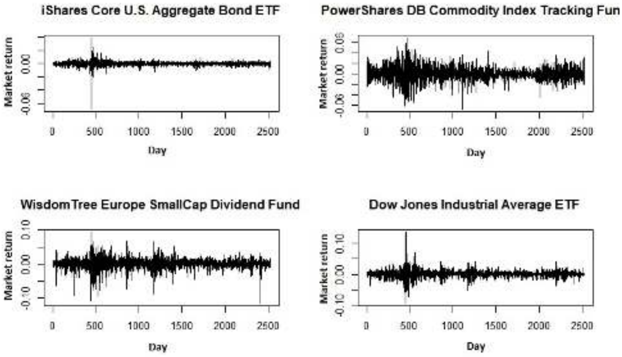
- $q_\alpha(X) = \mathrm{VaR}_\alpha(X) = \inf\{t \ / \ F_X(t) \geq \alpha\} = F_X^{-1}(\alpha)$,
- RiskMetrics popularized the use of VaR as a risk measure (1994).
- Basel Committee : Internal approach to capital management using VaR (1996),

Many natural examples where conditional quantiles are relevant: some variables are better known than others, you may estimate quantiles of the later knowing the first ones.

Quantile Oriented Sensitivity Analysis (QOSA).

A financial example

Consider four assets: iShares Core U.S. Aggregate Bond ETF, PowerShares DB Commodity Index Tracking Fund, WisdomTree Europe SmallCap Dividend Fund and SPDR Dow Jones Industrial Average ETF.

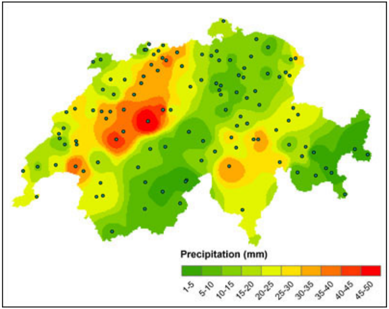


A financial example

Consider four assets: iShares Core U.S. Aggregate Bond ETF, PowerShares DB Commodity Index Tracking Fund, WisdomTree Europe SmallCap Dividend Fund and SPDR Dow Jones Industrial Average ETF.

How to use the knowledge of the 4 variables in order to estimate risk measures for WisdomTree Japan Hedged Equity Fund.

A spatial example



Source: Geographic Information Technology Training Alliance.

How to estimate risk measures related to $X(s)$ knowing $\mathbf{X}_{s_1,\dots,s_p}$?

Uncertainty

Model

$$f : \left| \begin{array}{ccc} \mathbb{R}^d & \to & \mathbb{R} \\ \mathbf{x} = (x_1, \ldots, x_d) & \mapsto & y = f(\mathbf{x}) \end{array} \right.$$

with

- f: mathematical or numerical model,
- $\mathbf{x}$: uncertain input parameters,
- y: model's output.

E.g. f is the Profit & Loss amount at time $t = 1$, the x_i's are different lines of insurance portfolio (automobile claims, home insurance, asset management, ...).

Uncertainty

Model

$$f : \left| \begin{array}{ccc} \mathbb{R}^d & \to & \mathbb{R} \\ \mathbf{x} = (x_1, \ldots, x_d) & \mapsto & y = f(\mathbf{x}) \end{array} \right.$$

The uncertainty on the input parameters is modelled by a probability distribution $\mathbb{P}$ on $\mathbb{R}^d$ and we get

$$Y = f(X_1, \ldots, X_d)$$

with the vector $\mathbf{X} = (X_1, \ldots, X_d)$ distributed as $\mathbb{P}$.

Sensitivity Analysis (SA)

The study of how uncertainty in the output of a model (numerical or otherwise) can be apportioned to different sources of uncertainty in the model's inputs (Saltelli et al. *(2004) e.g.).*

Sobol indices

Independent X_i's. Defined by Sobol (1993)[1].

$$S_i = \frac{\text{var}\left(\mathbb{E}[Y|X_i]\right)}{\text{var}(Y)}$$

$$S_i = \frac{\text{var}(Y) - \mathbb{E}\left(\text{var}[Y|X_i]\right)}{\text{var}(Y)}$$

$$S_i = \frac{\mathbb{E}\left[(Y - \mathbb{E}[Y])^2\right] - \mathbb{E}\left(\mathbb{E}\left[(Y - \mathbb{E}[Y|X_i])^2 |X_i\right]\right)}{\mathbb{E}\left[(Y - \mathbb{E}[Y])^2\right]}$$

$$S_i = \frac{\min_\theta \mathbb{E}\left[(Y - \theta)^2\right] - \mathbb{E}\left(\min_\theta \mathbb{E}\left[(Y - \theta)^2 |X_i\right]\right)}{\min_\theta \mathbb{E}\left[(Y - \theta)^2\right]}$$

[1] Ilya M Sobol (1993). In: *Mathematical Modelling and Computational Experiments*

Quantile oriented sensitivity analysis

QOSA: Quantile Oriented Sensitivity Analysis index: (Fort *et al.* 2016)

$$S_i^\alpha = \frac{\min_{\theta \in \mathbb{R}} \mathbb{E}\left[\psi_\alpha(Y,\theta)\right] - \mathbb{E}\left[\min_{\theta \in \mathbb{R}} \mathbb{E}\left[\psi_\alpha\left(Y,\theta\right)|X_i\right]\right]}{\min_{\theta \in \mathbb{R}} \mathbb{E}\left[\psi_\alpha(Y,\theta)\right]}$$

$$S_i^\alpha = \frac{\mathbb{E}\left[\psi_\alpha\left(Y, q_\alpha(Y)\right)\right] - \mathbb{E}\left[\psi_\alpha\left(Y, q_\alpha(Y|X_i)\right)\right]}{\mathbb{E}\left[\psi_\alpha(Y, q_\alpha(Y))\right]}$$

with the contrast function $\psi_\alpha : (y,\theta) \mapsto (y-\theta)(\alpha - \mathbf{1}_{y \leq \theta})$, $\alpha \in [0,1]$.

Remark ψ is related to quantiles:

$$q_\alpha(Y) = \arg\min_{\theta \in \mathbb{R}} \mathbb{E}(\psi_\alpha(Y,\theta)).$$

Quantile oriented sensitivity analysis

QOSA: Quantile Oriented Sensitivity Analysis index: (Fort *et al.* 2016)

$$S_i^\alpha = \frac{\mathbb{E}\left[\psi_\alpha\left(Y, q_\alpha(Y)\right)\right] - \mathbb{E}\left[\psi_\alpha\left(Y, q_\alpha(Y|X_i)\right)\right]}{\mathbb{E}\left[\psi_\alpha(Y, q_\alpha(Y))\right]}$$

Properties:

- $0 \leq S_i^\alpha \leq 1$
- $S_i^\alpha = 0 \iff Y$ and X_i are independent
- $S_i^\alpha = 1 \iff Y$ is X_i measurable

Application example: Y is the observed ozone concentration, $\mathbf{X}$ contains several variables such as: day type, deterministic prevision of ozone concentration, temperature, humidity ... Which of these variables have influence on the quantiles of Y?

Estimating QOSA

Estimating Sobol' index may avoid the estimation of the conditional distribution by using $\mathrm{var}\left(\mathbb{E}[Y|X_i]\right) = \mathrm{Cov}(Y, Y')$ with

$$Y' = f(\mathbf{X}'),\ \mathbf{X}' = (X_1', \ldots, X_{i-1}', X_i, X_{i+1}', \ldots, X_n')$$

X_j' independent copy of X_j.
The estimation of QOSA' index requires to estimate the conditional distribution $Y|X_i$.

- Kernel methods[2] optimal window width difficult to calibrate, requires a large number of calls to the costly function f.
- Random Forest method Less calls to f, time consuming nevertheless.

[2] Véronique Maume-Deschamps and Ibrahima Niang (2018). In: *Statistics & Probability Letters*
Thomas Browne et al. (2017). In: *hal.archives-ouvertes.fr*

Gaussian case

Assume $(Y, \mathbf{X})$ is Gaussian with expectation $\mu = (\mu_Y, \mu_{\mathbf{X}})$ and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_Y^2 & \Sigma_{Y\mathbf{X}}^T \\ \Sigma_{Y\mathbf{X}} & \Sigma_{\mathbf{X}} \end{pmatrix}$$

If $\Sigma_{\mathbf{X}}$ is invertible, then $Y|\mathbf{X}$ follows a normal law with expectation $\mu_{Y|\mathbf{X}} = \mu_Y + \Sigma_{Y\mathbf{X}}^T \Sigma_X^{-1}(\mathbf{X} - \mu_{\mathbf{X}})$ and variance $\sigma_{Y|\mathbf{X}}^2 = \sigma_Y^2 - \Sigma Y \mathbf{X}^T \Sigma_{\mathbf{X}}^{-1} \Sigma_{Y\mathbf{X}}$.
Then, the conditional quantiles are easily computable:

$$q_{Y|\mathbf{X}}^{\alpha} = \mu_{Y|\mathbf{X}} + \phi^{-1}(\alpha)\sigma_{Y|\mathbf{X}}.$$

Quantile regression

Approximate the conditional quantile by[3]:

$$\hat{q}_\alpha(X_2|\mathbf{X}_1) = {\beta^*}^T \mathbf{X}_1 + \beta_0^*$$

where β^* and β_0^* are the solutions of the following minimization problem:

$$(\beta^*, \beta_0^*) = \underset{\beta \in \mathbb{R}^N, \beta_0 \in \mathbb{R}}{\arg\min}\ \mathbb{E}[\psi_\alpha(X_2, \beta^T \mathbf{X}_1 + \beta_0)]$$

Recall:

$$\psi_\alpha(x, \theta) = (x - \theta)(\alpha - \mathbf{1}_{x \leq \theta})$$

and

$$q_\alpha(X_2|\mathbf{X}_1) = \underset{\theta \in \mathbb{R}}{\arg\min}\, \mathbb{E}(\psi_\alpha(X_2, \theta)|\mathbf{X}_1).$$

[3] R. Koenker and G. Jr. Bassett (1978). In: *Econometrica*

Other methods

- Random Forest,
- Neural networks,
- Nearest neighbors.

A survey with various methods is proposed by Torosian et al.[4]

[4] Léonard Torossian et al. (2020). In: *Reliability Engineering & System Safety*

Plan

Consistent Elliptic distributions

Definition

A $\mathbb{R}^d$ random vector $\mathbf{X}$ *has a consistent elliptic distribution if it writes*[5]*:*

$$\mathbf{X} \stackrel{d}{=} \mu + \epsilon \mathcal{N}(0, \Sigma)$$

with ϵ a positive random variable, independent of the underlying normal vector. This means that, a consistent elliptical distribution is a normal distribution with random variance $\epsilon^2\Sigma$.

[5] Y. Kano (1994). In: *Journal of Multivariate Analysis*

An equivalent definition

$\mathbf{X} \stackrel{d}{=} \mu + \epsilon\mathcal{N}(0,\Sigma)$ rewrites as[6]

$$\mathbf{X} \stackrel{d}{=} \mu + R\Lambda U^{(d)}$$

where $\Lambda\Lambda^T = \Sigma$, $U^{(d)}$ is a $d-$dimensional random vector uniformly distributed on $\mathcal{S}^{d-1}$, $R \stackrel{d}{=} \chi_d\epsilon$, R and $U^{(d)}$ are independent.
R is called the radius of $\mathbf{X}$, χ_d^2 is a χ-squared distribution, independent of ϵ and of the underlying Gaussian process.
$\mathbf{X}$ is said to be a consistent $(R,d)-$elliptical random vector with parameters μ and Σ.

[6] S. Cambanis, S. Huang, and G. Simons (1981). In: *Journal of Multivariate Analysis*

Properties of elliptic distributions

- Sub-vectors of elliptical vectors are elliptical, more precisely, Let $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ be a consistent $(R, d)-$elliptical random vector with parameters μ and Σ. $\mathbf{X}_1$ and $\mathbf{X}_2$ are d_1 and d_2-dimensional subvectors of $\mathbf{X}$. Let us write Σ :

  $$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

  Then $\mathbf{X}_1$ and $\mathbf{X}_2$ are respectively $(R, d_1)-$ and $(R, d_2)-$elliptical with parameters μ_1, Σ_{11} and μ_2, Σ_{22}, respectively.
- Conditional distributions of elliptical vectors are also elliptical.
- Linear combinations of coordinates / sub-vectors of elliptic distributions are also elliptic.

Properties of elliptic distributions

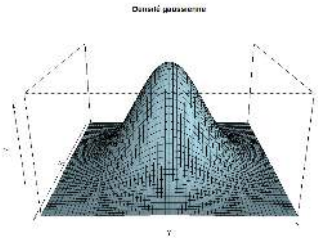
- Sub-vectors of elliptical vectors are elliptical,
- Conditional distributions of elliptical vectors are also elliptical. More precisely,
  $\mathbf{X}_2 | (\mathbf{X}_1 = x_1)$ is still elliptical, with radius R^* given by:

$$R^* \stackrel{d}{=} R\sqrt{1-\beta} | \left(R\sqrt{\beta} U^{(d)} = C_{11}^{-1}(x_1 - \mu_1) \right),$$

  where C_{11} is the root of Σ_{11}, and $\beta \sim Beta(\frac{d_1}{2}, \frac{d_2}{2})$.
- Linear combinations of coordinates / sub-vectors of elliptic distributions are also elliptic.

Examples

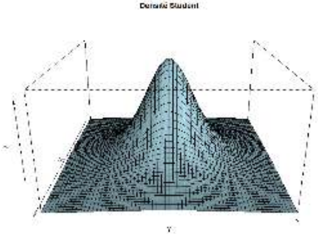
- Normal distributions: $\epsilon = 1$.



- Student distributions: with ν degrees of freedom: $\epsilon \stackrel{d}{=} \sqrt{\frac{\nu}{\chi_d^2}}$.
- Slash distributions: $\epsilon \stackrel{d}{=} \mathcal{P}(1, a)$.
- Laplace distibution: $\epsilon \stackrel{d}{=} \sqrt{\mathcal{E}(\lambda)}$.
- Many other.

Examples

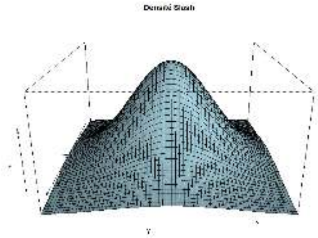
- Normal distributions: $\epsilon = 1$.
- Student distributions: with ν degrees of freedom: $\epsilon \stackrel{d}{=} \sqrt{\frac{\nu}{\chi_d^2}}$.



- Slash distributions: $\epsilon \stackrel{d}{=} \mathcal{P}(1, a)$.
- Laplace distibution: $\epsilon \stackrel{d}{=} \sqrt{\mathcal{E}(\lambda)}$.
- Many other.

Examples

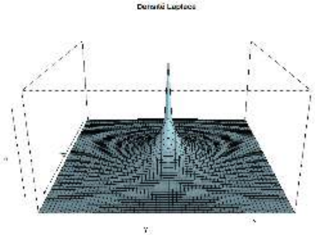
- Normal distributions: $\epsilon = 1$.
- Student distributions: with ν degrees of freedom: $\epsilon \overset{d}{=} \sqrt{\frac{\nu}{\chi^2_d}}$.
- Slash distributions: $\epsilon \overset{d}{=} \mathcal{P}(1, a)$.



- Laplace distibution: $\epsilon \overset{d}{=} \sqrt{\mathcal{E}(\lambda)}$.
- Many other.

Examples

- Normal distributions: $\epsilon = 1$.
- Student distributions: with ν degrees of freedom: $\epsilon \overset{d}{=} \sqrt{\frac{\nu}{\chi^2_d}}$.
- Slash distributions: $\epsilon \overset{d}{=} \mathcal{P}(1, a)$.
- Laplace distibution: $\epsilon \overset{d}{=} \sqrt{\mathcal{E}(\lambda)}$.



- Many other.

Conditional quantiles

We are interested in conditional quantiles for elliptical distributions. We have seen that conditional elliptical distribution are still elliptical. Assume a X is a $(R, 1)$ elliptical random vector with parameters μ and $\sigma^2 \in \mathbb{R}^+$, then

$$X = \mu + \sigma R U^{(1)}$$

where $U^{(1)} = -1$ or 1 with probability $\frac{1}{2}$. Thus, for $\alpha > \frac{1}{2}$,

$$q_\alpha(X) = \mu + \sigma \Phi_R^{-1}(2\alpha - 1)$$

where Φ_R is the distribution function of R.

Conditional quantiles

Proposition

Let $X = (\mathbf{X}_1, X_2)$ *a* $(R, N+1)-$*elliptical random vector with parameters* μ *and* Σ. *Write*

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix}.$$

Then for $\alpha \geq \frac{1}{2}$,

$$q_\alpha\left(X_2 | \mathbf{X}_1 = \mathbf{x}_1\right) = \mu_{2|1} + \sqrt{\Sigma_{2|1}}\Phi_{R^*}^{-1}(2\alpha - 1)$$

with $\begin{cases} \mu_{2|1} = & \mu_2 + \Sigma_{12}^T \Sigma_{11}^{-1}(x_1 - \mu_1) \\ \Sigma_{2|1} = & \Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12} \end{cases}$

Problem: the distribution of R^* is hardly accessible.

Quantile Regression for elliptic distributions

Write α' for $2\alpha - 1$.

Theorem

Let $X = (\mathbf{X}_1, X_2)$ be an elliptical distribution, the optimal quantile regression β^* is given by :

$$\beta^* = \Sigma_{11}^{-1}\Sigma_{12}$$

The Quantile Regression Predictor with level $\alpha \in [\frac{1}{2}, 1]$ is given by:

$$\hat{q}_\alpha(X_2|\mathbf{X}_1 = \mathbf{x}_1) = \mu_{2|1} + \sqrt{\Sigma_{2|1}}\Phi_R^{-1}(\alpha')$$

It satisfies

$$\hat{q}_\alpha(X_2|X_1) \sim \mathcal{E}_1\left(\mu_2 + \Sigma_{2|1}\Phi_R^{-1}(\alpha'), \Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12}, R\right)$$

How good is the quantile regression?

Gaussian case

$$\begin{cases} q_\alpha(X_2|\mathbf{X}_1 = \mathbf{x}_1) = & \mu_{2|1} + \sigma_{2|1}\Phi^{-1}(\alpha') \\ \hat{q}_\alpha(X_2|\mathbf{X}_1 = \mathbf{x}_1) = & \mu_{2|1} + \sigma_{2|1}\Phi^{-1}(\alpha') \end{cases}$$

The Quantile Regression Predictor is exactly the conditional quantile.

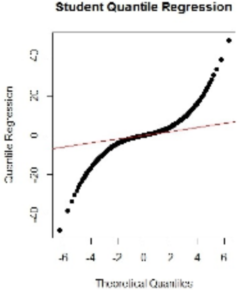
How good is the quantile regression?

Student case

$$\left\{\begin{array}{ll} q_\alpha(X_2|\mathbf{X_1}=\mathbf{x}_1) = & \mu_{2|1}+\sigma_{2|1}\sqrt{\frac{\nu}{\nu+N}}\sqrt{1+\frac{1}{\nu}d_1}\Phi^{-1}_{\nu+N}(\alpha') \\ \hat{q}_\alpha(X_2|\mathbf{X_1}=\mathbf{x}_1) = & \mu_{2|1}+\sigma_{2|1}\Phi^{-1}_{\nu}(\alpha') \end{array}\right. ;$$

where Φ_ν is the distribution function of a Student law with ν degrees of freedom.

The error may be huge, especially if the Mahalanobis distance $d_1 = (\mathbf{x}_1-\mu_1)^T\Sigma_{11}^{-1}(\mathbf{x}_1-\mu_1)$ is high. The picture is for $N=5$.



Extreme approximations

In case $\alpha \sim 1$, alternative methods have to be proposed. More precisely, we found an equivalent of $\Phi_{R^*}^{-1}(\alpha')$.

Some asymptotic relationships

Theorem

Under some technical assumptions, their exist $0 < \ell < +\infty$ *and* $\eta \in \mathbb{R}$ *such that :*

$$\left[\Phi_R^{-1}\left(1 - \frac{1}{\frac{\ell}{1-\alpha}+2(1-\ell)}\right)\right]^{\frac{1}{\eta}} \underset{\alpha\to 1}{\sim} \Phi_{R^*}^{-1}(\alpha)$$

This allows to approximate the conditional quantiles.

Examples

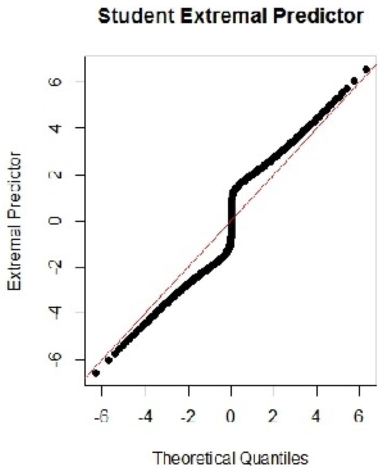
Property

The Gaussian, Student and Slash distributions satisfy the previous assumptions, with coefficients η and ℓ given in the table below.

Distribution	η	ℓ
Gaussian	1	1
Student, $\nu > 0$	$\frac{N}{\nu} + 1$	$\frac{\Gamma\left(\frac{\nu+N+1}{2}\right)\Gamma\left(\frac{\nu}{2}\right)}{\Gamma\left(\frac{\nu+N}{2}\right)\Gamma\left(\frac{\nu+1}{2}\right)} \left(1 + \frac{q_1}{\nu}\right)^{\frac{N+\nu}{2}} \frac{\nu^{\frac{N}{2}+1}}{\nu+N}$
Slash, $a > 0$	$\frac{N}{a} + 1$	$\frac{\Gamma\left(\frac{N+1+a}{2}\right)q_1^{\frac{N+a}{2}}}{\Gamma\left(\frac{N+a}{2}\right)(N+a)\chi^2_{N+a}(q_1)2^{\frac{a}{2}-1}\Gamma\left(\frac{1+a}{2}\right)}$

Examples

Extremal correction in the Student case

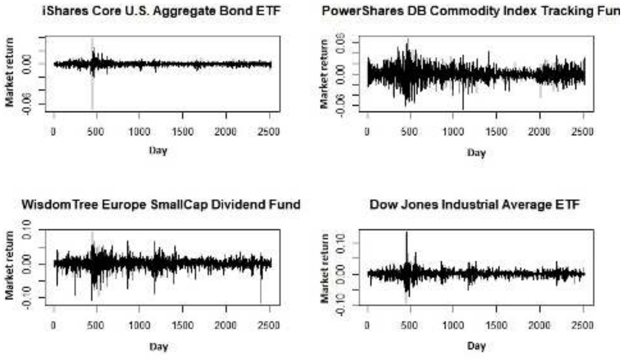


Estimations

Under additional assumptions (heavy tail + order two condition), estimations of the parameters ℓ, η, γ + asymptotic normality of the estimators are given[7].

[7] Antoine Usseglio-Carleve (2018). In: *Electronic Journal of Statistics*

Financial example



These four values are the first available every day $\Rightarrow$ anticipate the behaviour of the return of WisdomTree Japan Hedged Equity Fund X_2.

Financial example

The sample size is 2520. The first 2519 days (from January 3, 2007 to December 5, 2016) = learning sample, and we focus on the 2520th day: $\mathbf{x}_1 = (-0.0185\%, -0.4464\%, 0.9614\%, 0.1405\%)$.
Estimate quantiles of $X_2|\mathbf{X}_1 = \mathbf{x}_1$.

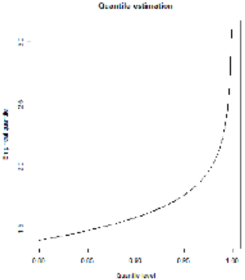
Financial example

The sample size is 2520. The first 2519 days (from January 3, 2007 to December 5, 2016) = learning sample, and we focus on the 2520th day: $\mathbf{x}_1 = (-0.0185\%, -0.4464\%, 0.9614\%, 0.1405\%)$.
Estimate quantiles of $X_2|\mathbf{X}_1 = \mathbf{x}_1$.
Data exploration:

- the daily returns can be considered as independent.
- the marginals seem symmetrical.
- the measured tail index is approximately the same for the marginals.

Could be assumed to be elliptical.

Financial example



E.g., for $\alpha = 0.999$, the estimated VaR is 3.1%.

Conclusion / perspectives for part I.

- Regression methods are not satisfactory for non gaussian distributions.
- Framework adapted to a large class of risk measures (TVaR, L^p quantile, Haezendonck-Goovaerts risk measures).
- New technics needed in the high dimension case (N large).
- More details in references below[8].
- Mixed approaches for non central but non extreme risk levels?
- Non symetric distributions?

[8] V. Maume-Deschamps, D. Rullière, and A. Usseglio-Carleve (2017a). In: *Journal of Multivariate Analysis*
V. Maume-Deschamps, D. Rullière, and A. Usseglio-Carleve (2017b). In: *Methodology and Computing in Applied Probability*
Antoine Usseglio-Carleve (2018). In: *Electronic Journal of Statistics*

Plan

Methods for conditional quantiles estimation

- Quantile regression is bad if you are far from gaussian,
- Kernel methods to estimate the conditional distribution function $F_{Y|\mathbf{X}}(t) = \mathbb{P}(Y \leq t|\mathbf{X})$, difficulty to adapt the window.
- Random forest methods,
- Neural networks methods.

In this part, we focus on random forest methods, having in mind that we aim at estimating QOSA indices:

Methods for conditional quantiles estimation

In this part, we focus on random forest methods, having in mind that we aim at estimating QOSA indices:

$$S_i^\alpha = \frac{\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha(Y,\theta)\right] - \mathbb{E}\left[\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha\left(Y,\theta\right)|X_i\right]\right]}{\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha(Y,\theta)\right]}$$

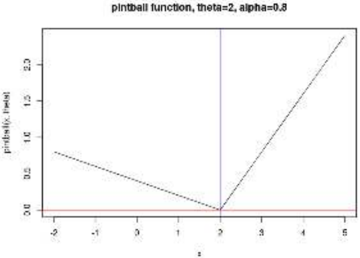
$$S_i^\alpha = \frac{\mathbb{E}\left[\psi_\alpha\left(Y, q_\alpha(Y)\right)\right] - \mathbb{E}\left[\psi_\alpha\left(Y, q_\alpha(Y|X_i)\right)\right]}{\mathbb{E}\left[\psi_\alpha(Y, q_\alpha(Y))\right]},$$

with $\psi_\alpha(x,\theta) = (x-\theta)(\alpha - \mathbf{1}_{x\leq\theta})$.

A remark on the definition of QOSA

$\psi_\alpha(x,\theta)$: a non symetric distance.

$\mathbb{E}\left[\psi_\alpha\left(Y,\theta\right)\right]$ is a mean dispersion measure of Y which is minimized for $\theta = q^\alpha(Y)$. So that QOSA indices compare the dispersion of Y around its quantile with its conditional counterpart.

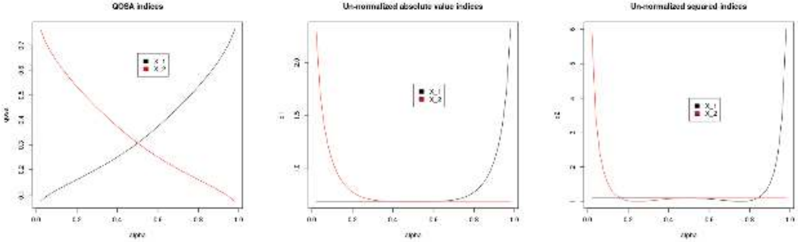


Other indices have been proposed by Kucherenko *et al.* in order to assess the impact of Y over quantiles, but their interpretation is questionnable:

$$\bar{k}^\alpha_{i,1} = \mathbb{E}\left[\left|q^\alpha\left(Y\right) - q^\alpha\left(\left.Y\right|X_i\right)\right|\right] \quad \bar{k}^\alpha_{i,2} = \mathbb{E}\left[\left(q^\alpha\left(Y\right) - q^\alpha\left(\left.Y\right|X_i\right)\right)^2\right].$$

A remark on the definition of QOSA

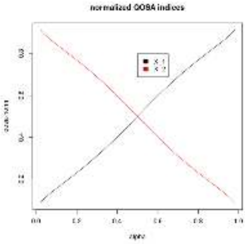
Comparison on the toy model: $Y = X_1 - X_2$ with $X_i \rightsquigarrow \mathcal{E}(1)$.

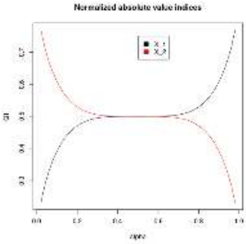


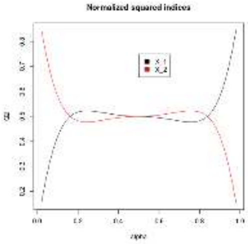
A remark on the definition of QOSA

Normalized versions

$$K_{i,1}^{\alpha} = \frac{\overline{k}_{i,1}^{\alpha}}{\sum_{j=1}^{d} \overline{k}_{j,1}^{\alpha}} \quad \text{and} \quad K_{i,2}^{\alpha} = \frac{\overline{k}_{i,2}^{\alpha}}{\sum_{j=1}^{d} \overline{k}_{j,2}^{\alpha}} .$$





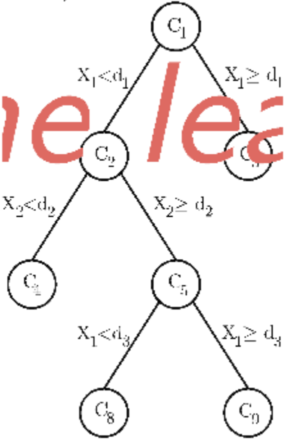


Recall CART

Classification And Regression Tree[9].

Input variables: $\mathbf{X} = (X_1, \ldots, X_d)$, Output variable: Y.

- Tree: constant piecewise predictor, obtained by binary recursive partitioning.
- Separate the data from the current node, by looking for the split reducing the most the heterogeneity of Y at the two child nodes.



Random Forests

Agregate several CART's to reduce the estimation variance

- Training sample: $\mathcal{D}_n = \left(\mathbf{X}^i, Y^i\right), i = 1, \ldots, n$
- $\Theta_\ell, \ell = 1, \ldots, k$ are independent random variables which determine how a tree is constructed (bootstrap on $\mathcal{D}_n$ and which variables are considered for the splits of each node), Θ_ℓ is assumed to be independent of $\mathcal{D}_n$.
- $A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)$: the leaf that is obtained when dropping $\mathbf{x}$ down the tree.
- $N_n(\mathbf{x}, \Theta_\ell, \mathcal{D}_n)$: the number of points which are in $A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)$.
- $N_n^b(\mathbf{x}, \Theta_\ell, \mathcal{D}_n)$: the number of points of the bootstrapped sample, which are in $A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)$.

Random forest conditional distributions functions estimation

Methods with random forest are often using the bootstrap sample, consider the random variable $B_j\left(\Theta_\ell, \mathcal{D}_n\right)$ as the number of times that the observation $\left(\mathbf{X}^j, Y^j\right)$ has been drawn from the original dataset for the ℓ-th tree construction. Consider the weights:

$$\omega_{n,i}(\mathbf{x},\Theta) = \frac{1}{k}\sum_{j=1}^{k}\frac{\mathbf{1}_{\mathbf{X}^i\in A_n(\mathbf{x},\Theta_j,\mathcal{D}_n)}}{N_n(\mathbf{x},\Theta_j,\mathcal{D}_n)},$$

$$\omega_{n,i}^{b}\left(\mathbf{x},\Theta\right) = \frac{1}{k}\sum_{\ell=1}^{k}\frac{B_i\left(\Theta_\ell,\mathcal{D}_n\right)\mathbf{1}_{\mathbf{X}^i\in A_n\left(\mathbf{x};\Theta_\ell,\mathcal{D}_n\right)}}{N_n^b\left(\mathbf{x};\Theta_\ell,\mathcal{D}_n\right)},$$

and the corresponding estimations of $F(y|\mathbf{X}=\mathbf{x})$:

$$\hat{F}_n^{b}\left(y|\mathbf{X}=\mathbf{x}\right) = \sum_{i=1}^{n}\omega_{n,i}^{b}\left(\mathbf{x}\right)\mathbf{1}_{\{Y^i\leqslant y\}}.$$

Random forest conditional quantiles estimation

Once the conditional distribution function is estimated, the conditional quantiles are estimated straightforwardly:

$$\hat{q}_\alpha(Y|\mathbf{X}) = \inf\{t \in \mathbb{R}, \hat{F}_n(t|\mathbf{X}) \geq \alpha\}.$$

With standard arguments, the consistency of $\hat{F}_n(t|\mathbf{X})$ leads to the consistency of $\hat{q}_\alpha(Y|\mathbf{X})$, provided that for all $\mathbf{x}$, the conditional $y \mapsto F(y|\mathbf{X} = \mathbf{x})$ is continuous and increasing.

Consistency of random forests

Results by Scornet, Biau, Vert (2015) in a linear model context:

$$Y = m(X) + \varepsilon \text{ with } \varepsilon \rightsquigarrow \mathcal{N}(0, \sigma^2) \text{ and } m(X) = \sum_{i=1}^{d} m_i(X_i).$$

$$m_n(\mathbf{x}, \Theta) = \sum_{i=1}^{n} \omega_{n,i}(\mathbf{x}, \Theta)\, Y^i,$$

$\omega_{n,i}$ as before. Under various assumptions including tree size wrt n and a forest correlation control, for $\mathbf{X} \rightsquigarrow \mathcal{U}[0,1]^d$,

$$\mathbb{E}[(m_n(\mathbf{X}) - m(\mathbf{X}))^2] \longrightarrow 0, \ \textit{with } m_n = \mathbb{E}_\Theta(m_{n,k}).$$

- No results for $m(\mathbf{x})$
- Results for fully grown trees and for limited grown trees.

Consistency of conditional distribution

Assume $Y = f(\mathbf{X}) + \varepsilon$, with ε a centred random variable, independent on $\mathbf{X}$.
In Meinshausen (2006)[10], convergence results for $\widehat{F}(y|\mathbf{X} = \mathbf{x})$ for a simplified random forest model. The $\omega_{n,i}(\mathbf{x})$'s are considered as constant (while they are random variables - depending on Θ, $\mathbf{X}^i$, Y^i, $i = 1, \ldots, n$)
$+$ various assumptions including tree growth and some regularity on $F(y|\mathbf{X} = \mathbf{x})$.

[10] Nicolai Meinshausen (2006). In: *Journal of Machine Learning Research*

Consistency: assumptions

Conditions

Relations between k (number of trees) and $N_n^b(\mathbf{x};\Theta,\mathcal{D}_n)$ (number of bootstrap observations in a leaf node):

1. $k = \mathcal{O}(n^{\alpha})$, with $\alpha > 0$.
2. $\forall \mathbf{x}$, $N_n^b(\mathbf{x};\Theta,\mathcal{D}_n) = \Omega\left(\sqrt{n}\,(\ln(n))^{\beta}\right)$, with $\beta > 1$, *a.s.*[a]
   or

[a] $f(n) = \Omega(g(n)) \iff \exists k > 0, \exists n_0 > 0 \mid \forall n \geqslant n_0 \quad |f(n)| \geqslant k \cdot |g(n)|$

Consistency: assumptions

Conditions

Relations between k (number of trees) and $N_n^b(\mathbf{x};\Theta,\mathcal{D}_n)$ (number of bootstrap observations in a leaf node):

1. $k=\mathcal{O}\left(n^{\alpha}\right)$, with $\alpha>0$.
2. $\forall \mathbf{x}, \quad \mathbb{E}\left[N_n^b\left(\mathbf{x};\Theta,\mathcal{D}_n\right)\right]=\Omega\left(\sqrt{n}\left(\ln\left(n\right)\right)^{\beta}\right)$, with $\beta>1$, and

   $\forall \mathbf{x}, \quad CV\left(N_n^b(\mathbf{x};\Theta,\mathcal{D}_n)\right)=\mathcal{O}\left(\frac{1}{n^{(1+\alpha)/2}\left(\ln\left(n\right)\right)^{\gamma/2}}\right)$, with $\gamma>1$.[a]

---

[a] $\mathrm{CV}\left(X\right)=\sigma_X/\mathbb{E}(X)$

Consistency: assumptions

Conditions

Relations between k (number of trees) and $N_n^b(\mathbf{x};\Theta,\mathcal{D}_n)$ (number of bootstrap observations in a leaf node):

1. $k = \mathcal{O}\left(n^{\alpha}\right)$, with $\alpha > 0$.
2. $\forall \mathbf{x}, \quad \mathbb{E}\left[N_n^b\left(\mathbf{x};\Theta,\mathcal{D}_n\right)\right] = \Omega\left(\sqrt{n}\left(\ln\left(n\right)\right)^{\beta}\right)$, with $\beta > 1$, and

   $\forall \mathbf{x}, \quad CV\left(N_n^b(\mathbf{x};\Theta,\mathcal{D}_n)\right) = \mathcal{O}\left(\frac{1}{n^{(1+\alpha)/2}\left(\ln\left(n\right)\right)^{\gamma/2}}\right)$, with $\gamma > 1$.[a]

The variations of function $F(y|\mathbf{X}=\cdot)$ is small on the trees' leaves:
$\forall \mathbf{x}, \forall y$,

$$\sup_{\mathbf{z},\mathbf{z}'\in A_n(\mathbf{x},\Theta_j)} \left|F\left(y|\,\mathbf{z}\right) - F\left(y|\,\mathbf{z}'\right)\right| \xrightarrow[n\to\infty]{a.s.} 0.$$

[a] $\mathrm{CV}\left(X\right) = \sigma_X/\mathbb{E}(X)$

Consistency: result[11]

Theorem

Assume the 3 conditions above are verified and $F(\cdot|\mathbf{X}=\mathbf{x})$ is continuous and increasing, $\forall \mathbf{x} \in \mathbb{R}^d$. Let F_n be either $\hat{F}_n^b$ or $\hat{F}_n$,

$$\sup_{y\in\mathbb{R}} \left|F_n\left(y\right|\mathbf{X}=\mathbf{x}) - F\left(y\right|\mathbf{X}=\mathbf{x})\right| \xrightarrow[n\to\infty]{a.s.} 0$$

Idea of the proof: The main idea is to use an auxiliary sample: let $(\mathbf{X}^{i\diamond}, Y^{i\diamond}, i=1,\dots n)$ be a second sample, independent from $(\mathbf{X}^i, Y^i, i=1,\dots,n)$ and consider the weights and the corresponding estimation of $F(y|\mathbf{X}=\mathbf{x})$:

$$\omega_{n,i}^{\diamond}(\mathbf{x},\Theta) = \frac{1}{k}\sum_{j=1}^{k}\frac{\mathbf{1}_{\mathbf{X}^{i\diamond}\in A_n(\mathbf{x},\Theta_j,\mathcal{D}_n)}}{N_n(\mathbf{x},\Theta_j,\mathcal{D}_n)},\quad F_n^{\diamond}\left(y|\mathbf{X}=\mathbf{x}\right) = \sum_{i=1}^{n}\omega_i^{\diamond}\left(\mathbf{x}\right)\mathbf{1}_{\{Y^{i\diamond}\leqslant y\}}.$$

[11] Kevin Elie-Dit-Cosaque and Véronique Maume-Deschamps (2020). In: *hal.archives-ouvertes.fr*

The two samples method.

We prove:

1. $|F_n(y|\mathbf{X}=\mathbf{x}) - F_n^\diamond(y|\mathbf{X}=\mathbf{x})| \xrightarrow[n\to\infty]{a.s.} 0$, uses a Hoeffding like inequality + Vapnik-Chervonenkis classes[12] (proximity of $N^\diamond$ and N^b),
2. $|F_n^\diamond(y|\mathbf{X}=\mathbf{x}) - F(y|\mathbf{X}=\mathbf{x})| \xrightarrow[n\to\infty]{a.s.} 0$, uses Vapnik-Chervonenkis classes again.
3. use a Dini argument to conclude with the $\sup_{y\in\mathbb{R}}$.

[12] V. N. Vapnik and A. Ya. Chervonenkis (1971). In: *Theory of Probability and its Applications*

Estimation strategies for the QOSA indices

Recall:

$$S_i^\alpha = 1 - \frac{\mathbb{E}\left[\psi_\alpha\left(Y, q^\alpha(Y|X_i)\right)\right]}{\mathbb{E}\left[\psi_\alpha(Y, q^\alpha(Y))\right]} = 1 - \frac{\mathbb{E}\left[\min\limits_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha\left(Y, \theta\right) \middle| X_i\right]\right]}{\mathbb{E}\left[\psi_\alpha(Y, q^\alpha(Y))\right]}.$$

Training sample: $\mathcal{D}_n = \left(\mathbf{X}^j, Y^j\right)_{j=1,\dots,n}$, the denominator is easily estimated with $\widehat{P}_1 = \frac{1}{n}\sum_{j=1}^{n} \psi_\alpha\left(Y^j, \widehat{q}^\alpha(Y)\right)$.

Two strategies to estimate the numerator:

- Quantile based estimators $\mathbb{E}\left[\psi_\alpha\left(Y, q^\alpha(Y|X_i)\right)\right]$,
- Minimum based estimators $\mathbb{E}\left[\min\limits_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha\left(Y, \theta\right) \middle| X_i\right]\right]$.

Quantile based estimators

Methods based on **two** training samples:

$\mathcal{D}_n^\star = \left(\mathbf{X}^{\star j}, Y^{\star j}\right)_{j=1,\dots,n}$ for computing the index,

$\mathcal{D}_n = \left(\mathbf{X}^{j}, Y^{j}\right)_{j=1,\dots,n}$ for estimating the conditional quantile.

$$\widehat{R}_i = \frac{1}{n}\sum_{j=1}^{n} \psi_\alpha\left(Y^{\star j}, \widehat{q}^{\alpha}\left(Y \middle| X_i = X_i^{\star j}\right)\right)$$

Construct the forest with $\mathcal{D}_n^i = \left(X_i^j, Y^j\right)_{j=1,\dots,n}$ from $\mathcal{D}_n$.

1. Quantile estimation with a weighted approach: $\widehat{R}_i^{1,b}$ or $\widehat{R}_i^{1,o}$

$$F_{k,n}^{b}\left(y \middle| X_i = x_i\right) = \sum_{j=1}^{n} w_{n,j}^{b}\left(x_i\right) \mathbf{1}_{\{Y^j \leqslant y\}}$$
$$\widehat{q}^{\alpha}\left(Y \middle| X_i = x_i\right) = \inf\left\{Y^p,\ p = 1, \dots, n : F_{k,n}^{b}\left(Y^p \middle| X_i = x_i\right) \geqslant \alpha\right\}$$

2. Quantile estimation within a leaf: $\widehat{R}_i^{2,b}$ or $\widehat{R}_i^{2,o}$.

Quantile based estimators

Methods based on **two** training samples:
$\mathcal{D}_n^\star = \left(\mathbf{X}^{\star j}, Y^{\star j}\right)_{j=1,\dots,n}$ for computing the index,
$\mathcal{D}_n = \left(\mathbf{X}^{j}, Y^{j}\right)_{j=1,\dots,n}$ for estimating the conditional quantile.

$$\widehat{R}_i = \frac{1}{n} \sum_{j=1}^{n} \psi_\alpha \left(Y^{\star j}, \widehat{q}^{\alpha} \left(Y | X_i = X_i^{\star j} \right) \right)$$

Construct the forest with $\mathcal{D}_n^i = \left(X_i^j, Y^j\right)_{j=1,\dots,n}$ from $\mathcal{D}_n$.

1. Quantile estimation with a weighted approach: $\widehat{R}_i^{1,b}$ or $\widehat{R}_i^{1,o}$
2. Quantile estimation within a leaf: $\widehat{R}_i^{2,b}$ or $\widehat{R}_i^{2,o}$.
   For one tree, $\widehat{q}_\ell^{b,\alpha}\left(Y | X_i = x_i \right)$ on the leaf containing x_i. On the forest:

$$\widehat{q}^{\alpha}\left(Y | X_i = x_i \right) = \frac{1}{k} \sum_{\ell=1}^{k} \widehat{q}_\ell^{b,\alpha}\left(Y | X_i = x_i \right).$$

Minimum based estimators

Minimum estimation with a weighted approach:
$\mathcal{D}_n = \left(\mathbf{X}^j, Y^j\right)_{j=1,\dots,n}$ and $\left(\mathbf{X}^{\star j}\right)_{j=1,\dots,n}$, i.e. requires 1.5 training samples.

Estimate $\mathbb{E}\left[\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha\left(Y,\theta\right)\middle| X_i\right]\right]$ with

$$\frac{1}{n}\sum_{m=1}^{n} \min_{p=1,\dots,n} \sum_{j=1}^{n} w_{n,j}^{b}\left(X_i^{\star m}\right)\psi_\alpha\left(Y^j, Y^p\right)$$

$\Longrightarrow \widehat{Q}_i^{1,b}$ or $\widehat{Q}_i^{1,o}$.

Minimum based estimators

Minimum estimation within a leaf: $\mathcal{D}_n = \left(\mathbf{X}^j, Y^j\right)_{j=1,\dots,n}$. Estimate $\mathbb{E}\left[\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha\left(Y,\theta\right)\middle| X_i\right]\right]$ with

$$\frac{1}{k}\sum_{\ell=1}^{k}\left[\frac{1}{N_{leaves}^{\ell}}\sum_{m=1}^{N_{leaves}^{\ell}}\left(\min_{p\in\mathcal{L}_{\ell,m}^{b}}\sum_{j\in\mathcal{L}_{\ell,m}^{b}}\frac{\psi_\alpha\left(Y^j,Y^p\right)}{\left|\mathcal{L}_{\ell,m}^{b}\right|}\right)\right]$$

$\Longrightarrow$ $\widehat{Q}_i^{2,b}$ or $\widehat{Q}_i^{2,o}$.

Principles of the Cross-Validation

Preliminary studies have showned that size's leaves is crutial in the estimation $\Longrightarrow$ cross-validation strategy in order to choose the number of elements in the leaves.

1. Shuffle the dataset randomly and split the dataset in **k** folds
2. For each unique group: Take the group as a test dataset; Take the remaining groups as a training dataset; Fit a model on the training set and evaluate it on the test set; Retain the evaluation score and discard the model
3. Summarize the skill of the model using the sample of model evaluation scores

Leaf size issue

We use $\widehat{R}_i^1 = \frac{1}{n} \sum_{j=1}^{n} \psi_\alpha \left(Y^j, \widehat{q}^\alpha \left(Y | X_i = X_i^j \right) \right)$ as score.

- In the cross validation process, among a grid of possible sizes, construct a forest with leaf size realizing the minimal score.
- Using the Out of Bag (OoB) sample.
  1. For a given observation (X_i^j, Y^j) from $\mathcal{D}_n^i$, consider the set of trees built with the bootstrap samples not containing this observation (it is *Out of Bag*).
  2. Aggregate the estimations from these trees to make the OoB estimation: $\widehat{q}_{oob}^{b,\alpha} \left(Y | X_i = X_i^j \right)$ of $q^\alpha \left(Y | X_i = X_i^j \right)$.
  3. Calculate the OoB score:

$$\widehat{OOB}_i^b = \frac{1}{n} \sum_{j=1}^{n} \psi_\alpha \left(Y^j, \widehat{q}_{oob}^{b,\alpha} \left(Y | X_i = X_i^j \right) \right) .$$

Among a grid of possible sizes, construct a forest with leaf size realizing the minimal OoB score.

Leaf size issue

We use $\widehat{R}_i^1 = \frac{1}{n}\sum_{j=1}^{n} \psi_\alpha\left(Y^j, \widehat{q}^\alpha\left(Y | X_i = X_i^j\right)\right)$ as score.

- In the cross validation process, among a grid of possible sizes, construct a forest with leaf size realizing the minimal score.
- Using the Out of Bag (OoB) sample. Among a grid of possible sizes, construct a forest with leaf size realizing the minimal OoB score.

Using the OoB sample is much less time consuming since, it does not require cutting out the training sample and it takes place during the forest construction process.

Sum of exponential laws

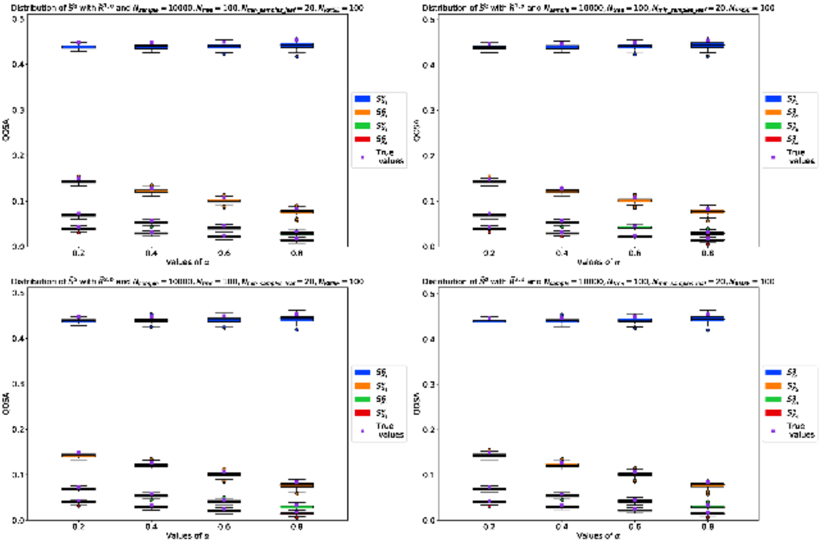
case $X_i \rightsquigarrow \mathcal{E}(\lambda_i)$, $\lambda_i \in \mathbb{R}^+$ distinct;
$Y = \sum_{i=1}^{n} X_i$ a semi-closed form formula may be obtained by using calculations from Marceau (2014).

Simulation study for $\lambda_1 = 0.5$, $\lambda_2 = 1$, $\lambda_3 = 1.5$, $\lambda_4 = 2$.

`sample size` $= 10^4$,
`nb trees` $= 100$,
`boxplots on 100 repetitions.`

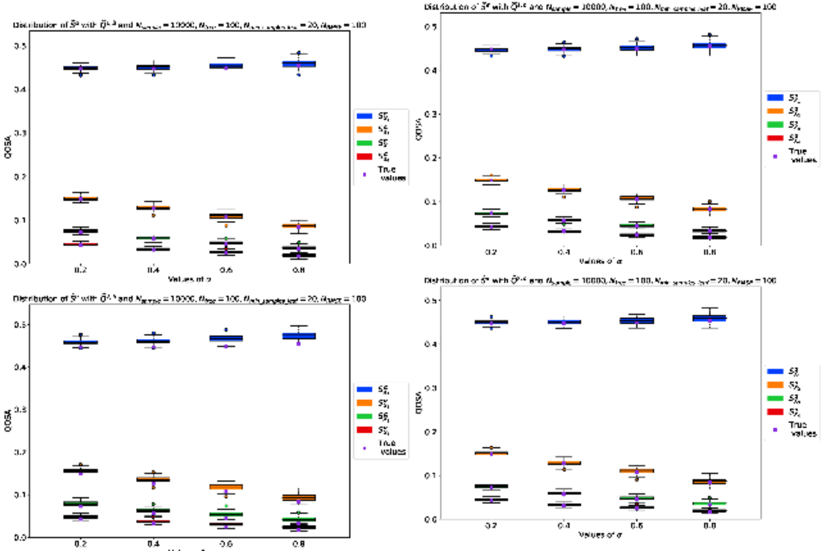
Sum of exponential laws

Quantile based methods



Sum of exponential laws

Minimum based methods



Comparison with kernel methods

Consider a toy model: $Y = X_1 - X_2$ with $X_i \rightsquigarrow \mathcal{E}(1)$ independent.
RMSE and run time of the random forest based estimators: with $\widehat{Q}_i^{1,o}$ and $\widehat{Q}_i^{2,o}$ as well as those based on kernel: $\widetilde{S}_i^{\alpha}$[13] and $\check{S}_i^{\alpha}$[14], sample size is 10^4.

	RF with $\widehat{Q}_i^{1,o}$	RF with $\widehat{Q}_i^{2,o}$	$\widetilde{S}_i^{\alpha}$	$\check{S}_i^{\alpha}$
$\alpha = 0.1$	0.007	0.009	0.061	0.020
$\alpha = 0.25$	0.008	0.009	0.042	0.013
$\alpha = 0.5$	0.008	0.008	0.027	0.019
$\alpha = 0.75$	0.008	0.008	0.014	0.035
$\alpha = 0.99$	0.006	0.006	0.013	0.084
run time	1 hr	18 min 24 sec	1 min 51 sec	1 hr 55 min

[13] Véronique Maume-Deschamps and Ibrahima Niang (2018). In: *Statistics & Probability Letters*
[14] Thomas Browne et al. (2017). In: *hal.archives-ouvertes.fr*

A real dataset

Bias between the predictions from **MOCAGE** (**M**odèle de **C**himie **A**tmosphérique à **G**rande **E**chelle) and the observed ozone concentration.

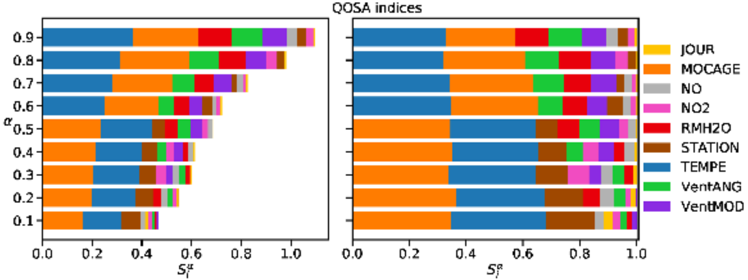
This dataset[15] contains 10 variables with 1041 observations.

O3obs: observed ozone concentration will be explained by the 9 other variables.

JOUR: type of day (holiday vs no holiday) MOCAGE: ozone concentration predicted by a fluid mechanics model RMH2O: humidity ratio VentMOD: wind force NO: nitric oxide concentration	STATION: site of observations (5 different sites) TEMPE: officially predicted temperatures NO2: nitrogen dioxide concentration VentANG: wind direction

[15] Philippe Besse et al. (2007). In: *Pollution atmosphérique*

Application on a real dataset: results



Evolution of the ranking of the QOSA indices (brut indices on the left, in %ages on the right)in function of the levels α.
Considering the central effects leads[16] to consider MOCAGE and TEMPE as the most influencial variables, then STATION and NO2. We see that for quantile levels ≥ 0.6 wind is also important.

[16] Philippe Besse et al. (2007). In: *Pollution atmosphérique*
Baptiste Broto, Francois Bachoc, and Marine Depecker (2020). In: *SIAM/ASA Journal on Uncertainty Quantification*

Conclusion / perspectives for part II.

- Random forest methods usefull for conditional quantile and QOSA estimations, but costly.
- Methods implemented in Python[17] (QOSA) and Julia[18] (conditional distributions).
- Asymptotic distributions to get confidence intervals?
- To be compared with Generalized Random Forest[19].

[17] Kévin Elie-Dit-Cosaque (2020).
[18] Benoit Fabrège and Véronique Maume-Deschamps (2020).
[19] Susan Athey, Julie Tibshirani, Stefan Wager, et al. (2019). In: *The Annals of Statistics*

Plan

Conclusions

- Importance of conditional quantile estimations
  - Various methods exists, we have seen only few.
  - Specific methods available for some classes of distributions such as elliptical distributions.
  - Specific attention for high level quantiles (uses extreme value theory).
- Interest of QOSA indicies
  - Give different informations than Sobol indices, pertinent if you are interested in different quantile levels.
  - Interpretation not so easy, especially if inputs are dependent $\longrightarrow$ go the qosa-Shapley (mixture of Shapley effect[20] and QOSA indices (work in progress).

[20] Art B Owen and Clémentine Prieur (2016). In: *arXiv preprint arXiv:1610.02080*

References I

Athey, Susan, Julie Tibshirani, Stefan Wager, et al. (2019). "Generalized random forests". In: *The Annals of Statistics* 47.2, pp. 1148–1178.

Besse, Philippe et al. (2007). "Comparaison de techniques de «Data Mining» pour l'adaptation statistique des prévisions d'ozone du modèle de chimie-transport MOCAGE". In: *Pollution atmosphérique* 49.195, pp. 285–292.

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32.

Broto, Baptiste, Francois Bachoc, and Marine Depecker (2020). "Variance reduction for estimation of Shapley effects and adaptation to unknown input distribution". In: *SIAM/ASA Journal on Uncertainty Quantification* 8.2, pp. 693–716.

Browne, Thomas et al. (2017). "Estimate of quantile-oriented sensitivity indices". In: *hal.archives-ouvertes.fr.*

Cambanis, S., S. Huang, and G. Simons (1981). "On the theory of elliptically contoured distributions". In: *Journal of Multivariate Analysis* 11, pp. 368–385.

Elie-Dit-Cosaque, Kévin (2020). *qosa-indices, a python package available at:* `https://gitlab.com/qosa_index/qosa`.

Elie-Dit-Cosaque, Kevin and Véronique Maume-Deschamps (2020). "Random forest estimation of conditional distribution functions and conditional quantiles". In: *hal.archives-ouvertes.fr.*

Fabrège, Benoit and Véronique Maume-Deschamps (2020). *Conditional Distribution Forest: a Julia package available at* `https://github.com/bfabreges/ConditionalDistributionForest.jl`.

Fort, Jean-Claude, Thierry Klein, and Nabil Rachdi (2016). "New sensitivity analysis subordinated to a contrast". In: *Communications in Statistics-Theory and Methods* 45.15, pp. 4349–4364.

References II

Kano, Y. (1994). “Consistency Property of Elliptical Probability Density Functions”. In: *Journal of Multivariate Analysis* 51, pp. 139–147.

Koenker, R. and G. Jr. Bassett (1978). “Regression Quantiles”. In: *Econometrica* 46.1, pp. 33–50.

Kucherenko, Sergei, Shufang Song, and Lu Wang (2019). “Quantile based global sensitivity measures”. In: *Reliability Engineering & System Safety* 185, pp. 35–48.

Marceau, Etienne (2013). *Modélisation et évaluation quantitative des risques en actuariat: Modèles sur une période*. Springer.

Maume-Deschamps, V., D. Rullière, and A. Usseglio-Carleve (2017a). “Quantile predictions for elliptical random fields”. In: *Journal of Multivariate Analysis* 159, pp. 1 –17.

— (2017b). “Spatial expectile predictions for elliptical random fields”. In: *Methodology and Computing in Applied Probability* 20.2, pp. 643–671.

Maume-Deschamps, Véronique and Ibrahima Niang (2018). “Estimation of quantile oriented sensitivity indices”. In: *Statistics & Probability Letters* 134, pp. 122–127.

Meinshausen, Nicolai (2006). “Quantile regression forests”. In: *Journal of Machine Learning Research* 7.Jun, pp. 983–999.

Owen, Art B and Clémentine Prieur (2016). “On Shapley value for measuring importance of dependent inputs”. In: *arXiv preprint arXiv:1610.02080*.

Saltelli, Andrea et al. (2004). *Sensitivity analysis in practice: a guide to assessing scientific models*. John Wiley & Sons.

References III

Scornet, Erwan, Gérard Biau, and Jean-Philippe Vert (2015). "Consistency of random forests". In: *The Annals of Statistics*.

Sobol, Ilya M (1993). "Sensitivity estimates for nonlinear mathematical models". In: *Mathematical Modelling and Computational Experiments* 1.4, pp. 407–414.

Torossian, Léonard et al. (2020). "A review on quantile regression for stochastic computer experiments". In: *Reliability Engineering & System Safety*.

Usseglio-Carleve, Antoine (2018). "Estimation of conditional extreme risk measures from heavy-tailed elliptical random vectors". In: *Electronic Journal of Statistics*.

Vapnik, V. N. and A. Ya. Chervonenkis (1971). "On the Uniform Convergence of Relative Frequencies of Events to their Probabilities". In: *Theory of Probability and its Applications* 16.2, pp. 264–280.

Thank you

Thanks for your attention.

Merci pour votre attention.

AMIES

N'oubliez pas qu'AMIES peut vous aider dans vos collaborations avec les entreprises

LES MATHS
VECTEUR D'INNOVATION