Generating unobserved multivariate extremes: a variational auto-encoder approach

Nicolas Lafon, Philippe Naveau and Ronan Fablet

LSCE
Laboratoire des sciences du climat & de l'environnement

Problem at hand

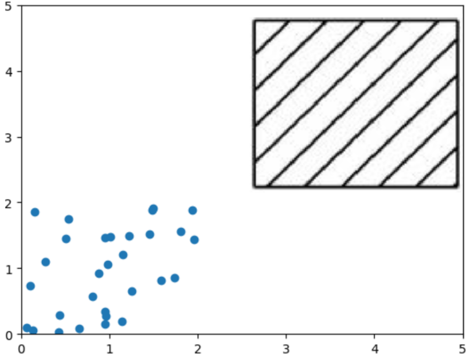


Figure – How to sample from observations (blue dots) in extreme regions (black square) to estimate probability of rare events ?

Who cares ?

Risk managers

- How to determine a flood plain area ?
- How to choose the height of a dam or a levy ?
- How to protect population from unprecedented climate events ?

Machine learners

- How to learn outside of the training sample ?

A hydrological example

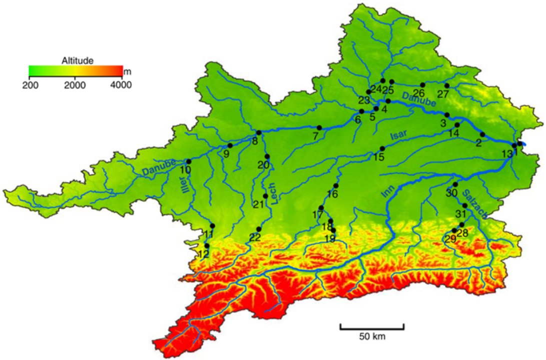


Figure – Gauging stations on the Danube river basin (*Asadi et al., 2015*).

A hydrological example

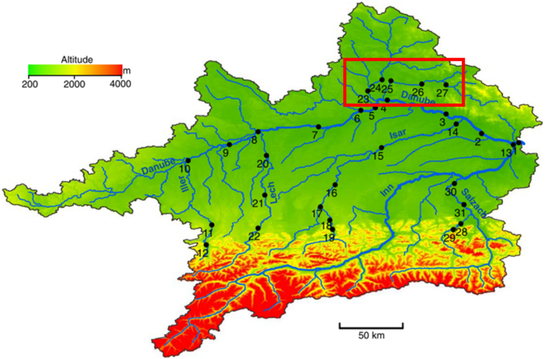


Figure – Gauging stations on the Danube river basin (*Asadi et al., 2015*).

A hydrological example

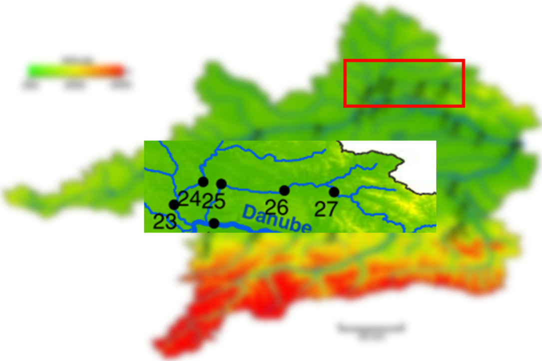


Figure – Gauging stations on the Danube river basin (*Asadi et al., 2015*).

Probability of joint extremes

$$A_{23:j}^{(q)} = \bigcap_{i=23}^{j} \left\{ X_i > u_i^{(q)} \right\}$$

with $u_i^{(q)}$ discharge at q quantile for station i in test set

Probability of joint extremes

$$A_{23:j}^{(q)} = \bigcap_{i=23}^{j} \left\{ X_i > u_i^{(q)} \right\}$$

with $u_i^{(q)}$ discharge at q quantile for station i in test set

Table – Proportion (in %) .

	quantile level = 90%				
	Training set	Test set			
$A_{23:25}^{(.9)}$	5.9	6.6			
$A_{23:26}^{(.9)}$	4.9	6.0			
$A_{23:27}^{(.9)}$	3.8	5.1			

Probability of joint extremes

$$A_{23:j}^{(q)} = \bigcap_{i=23,\ldots,j} \left\{ X_i > u_i^{(q)} \right\}$$

with $u_i^{(q)}$ discharge at q quantile for station i in test set

Table – Proportion (in %) .

	quantile level = 99%				
	Training set	Test set			
$A_{23:25}^{(.99)}$	0.0	0.48			
$A_{23:26}^{(.99)}$	0.0	0.4			
$A_{23:27}^{(.99)}$	0.0	0.25			

How to model multivariate extreme distributions

Gaussexit (31/12/2001)

Gaussexit (31/12/2001)

Why was it hard to leave the Deutschmark and Gauss ?

- The assumption of normality is very prevalent in the theoretical and applied statistical research, e.g. the cost function are often L2 norm
- Asymptotic justification : Central Limit Theorem
- Nice properties of Gaussian vectors
- Completely characterized by its first two moments
- Stability under linearity
- Stability under summation
- Stability under conditioning

Why was it hard to leave the Deutschmark and Gauss ?

- The assumption of normality is very prevalent in the theoretical and applied statistical research, e.g. the cost function are often L2 norm
- Asymptotic justification : Central Limit Theorem
- Nice properties of Gaussian vectors
- Completely characterized by its first two moments
- Stability under linearity
- Stability under summation
- Stability under conditioning

ML techniques and Gauss

- Diffusion example : $p(x_t|x_{t-1}) \sim N\left(\sqrt{1-\beta_t}x_{t-1}, \beta_t I\right)$ and $p(x_{t-1}|x_t) \sim N(\mu_t, \Sigma_t)$
- Scoring rules for additive models $Y = X + \sigma^2 N(0,1)$

$$E(X|Y=y) = y + \sigma^2 \frac{f_Y'(y)}{f_Y(y)}$$

Why was it hard to leave the Deutschmark and Gauss ?

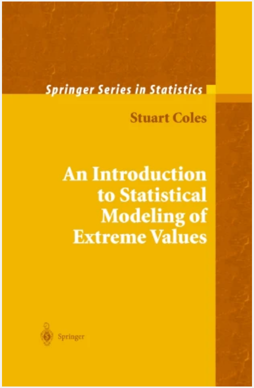
- The assumption of normality is very prevalent in the theoretical and applied statistical research, e.g. the cost function are often L2 norm
- Asymptotic justification : Central Limit Theorem
- Nice properties of Gaussian vectors
- Completely characterized by its first two moments
- Stability under linearity
- Stability under summation
- Stability under conditioning

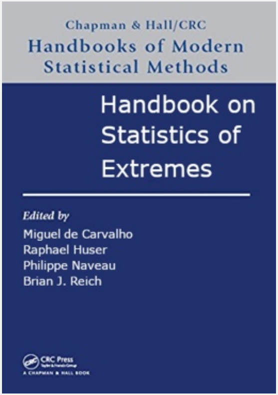
A main drawback for extremes

A light tail distribution may not model the blue points of interest

Extreme Value Theory goal is little bit like the Brexit

Applied MEVT resources





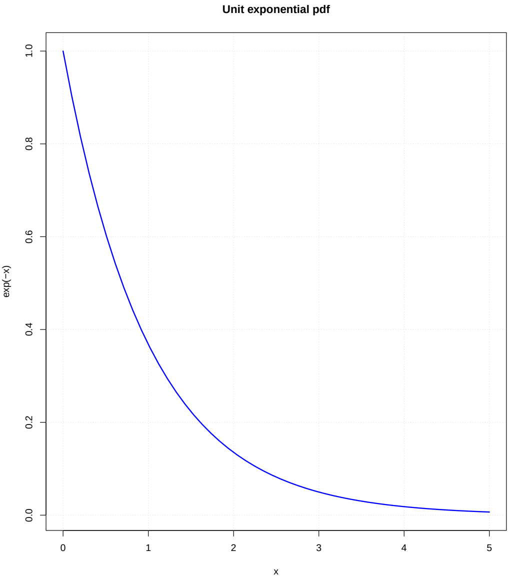
https://extremestats.github.io/Handbook/

- A statistical journal called Extremes
- A more applied journal called Weather and Climate Extremes

How to model exceedances above a high threshold

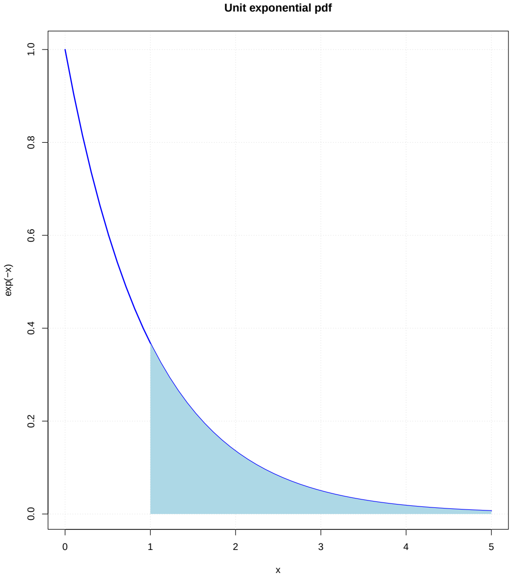
Threshold stability in 1D

Unit exponential $P(E > x) = \exp(-x)$



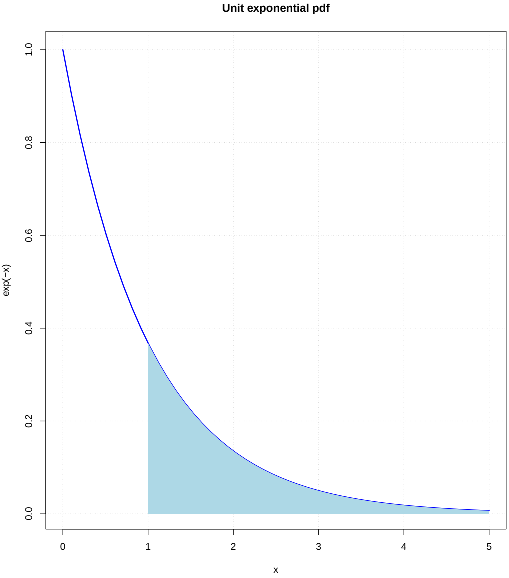
Unit exponential $P(E > x) = \exp(-x)$

$$P(E > x + 1) = \exp(-(x + 1)) = P(E > 1)P(E > x)$$



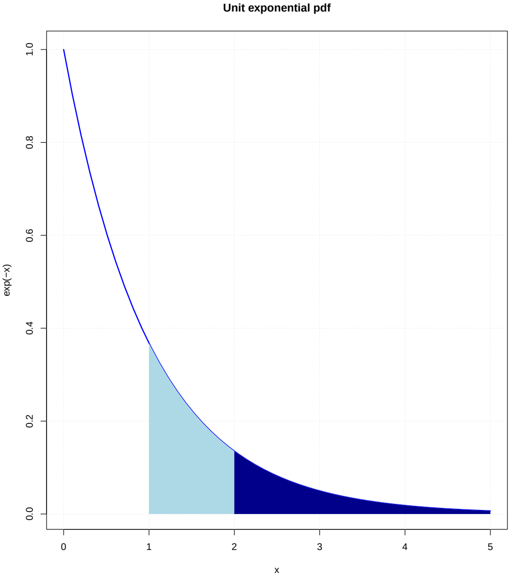
Unit exponential $P(E > x) = \exp(-x)$

$$P(E > x + 1 | E > 1) = P(E > x)$$



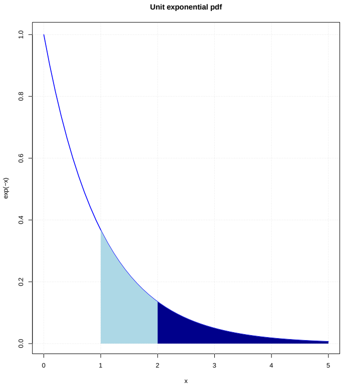
Unit exponential $P(E > x) = \exp(-x)$

$$P(E > x + u | E > u) = P(E > x)$$



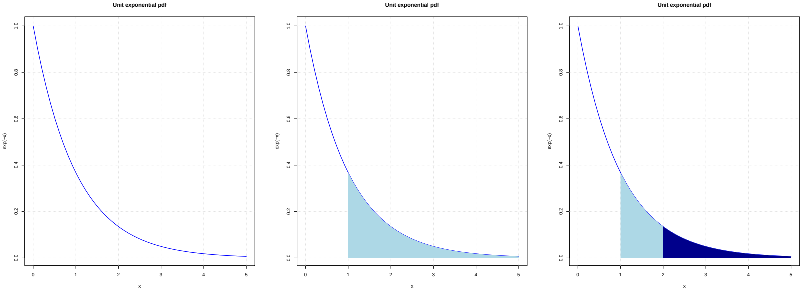
Unit exponential $P(E > x) = \exp(-x)$

$$P(E > x + M | E > M) = P(E > x)$$



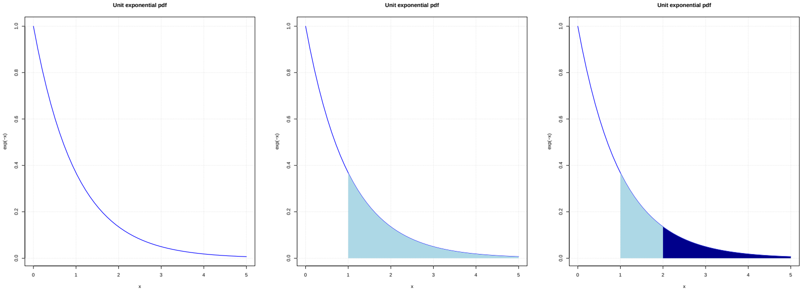
for any positive random variable M

Measure point of view



$$P(E \in B + u) = \int_{B+u} e^{-x}\, \mathrm{d}x = e^{-u} \int_{B} e^{-x}\, \mathrm{d}x = e^{-u} P(E \in B)$$

Measure point of view

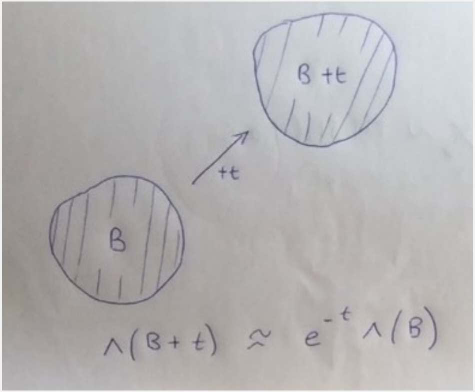


$$P(E \in B + u) = \int_{B+u} e^{-x}\, \mathrm{d}x = e^{-u} \int_B e^{-x}\, \mathrm{d}x = e^{-u} P(E \in B)$$

$$\Lambda(B + u) = e^{-u} \Lambda(B)$$

where Λ is a measure with density $\lambda(x) = e^{-x}$

Is it possible to have the same property in 2D ?



Is the exponential the only threshold invariant pdf ?

Thresholding : the Generalized Pareto Distribution (GPD)

$$\mathrm{pr}\{\mathbf{X}-u > x | \mathbf{X} > u\} = \overline{H}_\xi(x/\sigma) = \left(1 + \frac{\xi\, x}{\sigma}\right)_+^{-1/\xi}$$

Vilfredo Pareto : 1848-1923

Born in France and trained as an engineer in Italy, he turned to the social sciences and ended his career in Switzerland. He formulated the power-law distribution (or "Pareto's Law"), as a model for how income or wealth is distributed across society.

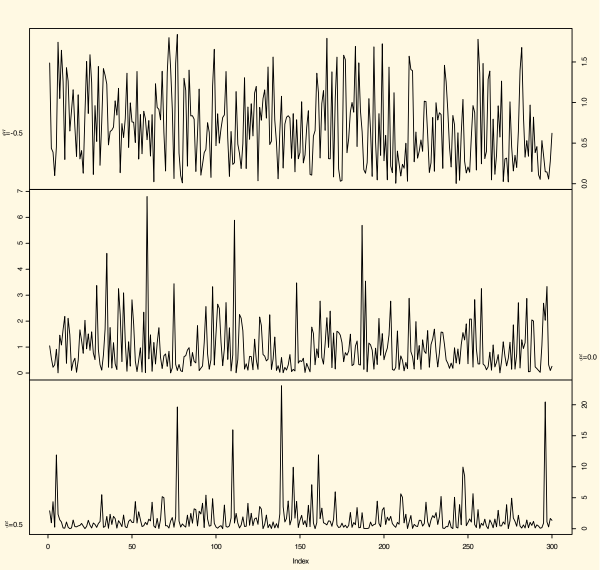
Stability by thresholding

Pickands-Balkema-de Haan theorem (1974,1975)

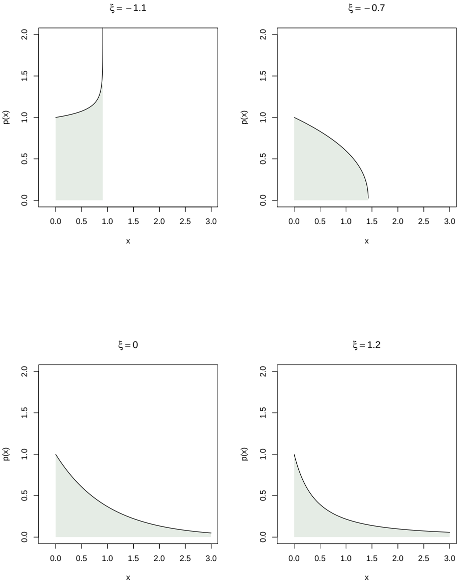
For a large class of random variables, we have

$$\lim_{u \to y_F} \sup_y |P(Y > y | Y > u) - \overline{H}(y, \sigma(u), \xi)| = 0$$

From Bounded to Heavy tails



Probability densities of a Generalized Pareto law

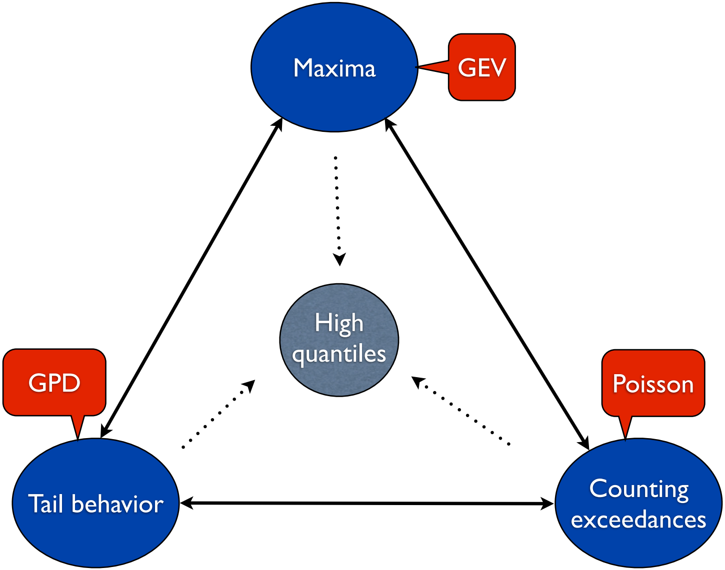


A fun fact about the Pareto for heavy tails ($\xi > 0$)

Pareto as a ratio of two random variables behavior

A Pareto can always be simulated by multiplying an unit exponential random variable by a independent inverse Gamma random variable

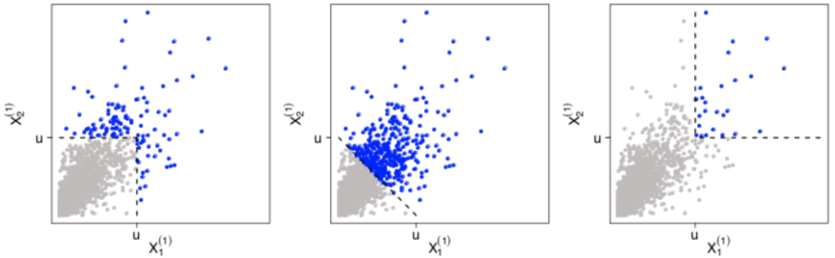
A summary in 1D case



Multivariate extreme value theory

Risk function [1]

$r(x) = \max(x_1, x_2)$ $\quad$ $r(x) = x_1 + x_2$ $\quad$ $r(x) = \min(x_1, x_2)$



1. Dombry and Ribatet (2015, Statistics and its Interface)

Transforming the positive vector $\mathbf{X} = (X_1, \ldots, X_d)^T$

Radial component

$$R = X_1 + \cdots + X_d$$

Transforming the positive vector $\mathbf{X} = (X_1, \ldots, X_d)^T$

Radial component

$$R = X_1 + \cdots + X_d$$

"Angular" coordinates

$\boldsymbol{\Theta} = (\theta_1, \ldots, \theta_d)^T$ with

$$\theta_i = \frac{X_i}{X_1 + \cdots + X_d}$$

Multivariate regular variations

Radial component

$$R = X_1 + \cdots + X_d$$

with

$$P(R > tr | R > t) \sim cst \times r^{-\alpha}$$

(threshold invariant like a GDP with $\alpha = 1/\xi$)

Multivariate regular variations

Radial component

$$R = X_1 + \cdots + X_d$$

with

$$P(R > tr | R > t) \sim cst \times r^{-\alpha}$$

(threshold invariant like a GDP with $\alpha = 1/\xi$)

“Angular" coordinates

$\mathbf{\Theta} = (\theta_1, \ldots, \theta_d)^T$ with

$$\theta_i = \frac{X_i}{X_1 + \cdots + X_d}$$

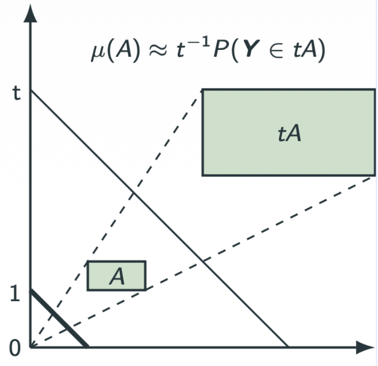
with $\mathbf{\Theta}$ independent of R for large R, i.e.

$$\lim_{t \to \infty} P(\mathbf{\Theta} \in A | R > t) = \nu(A),$$

a angular measure on the simplex

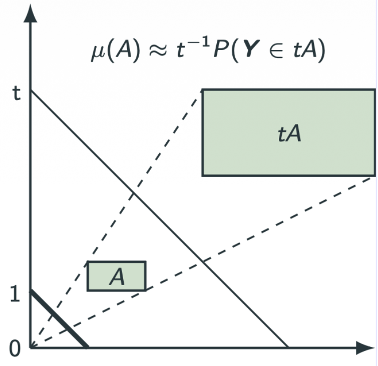
Multivariate regular variations (see, e.g. Bucher and Boulin (2025)

Exponent measure



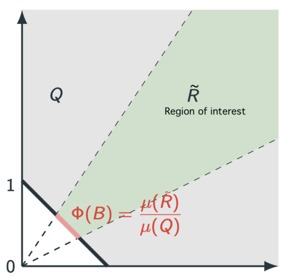
Multivariate regular variations (see, e.g. Bucher and Boulin (2025)

Exponent measure

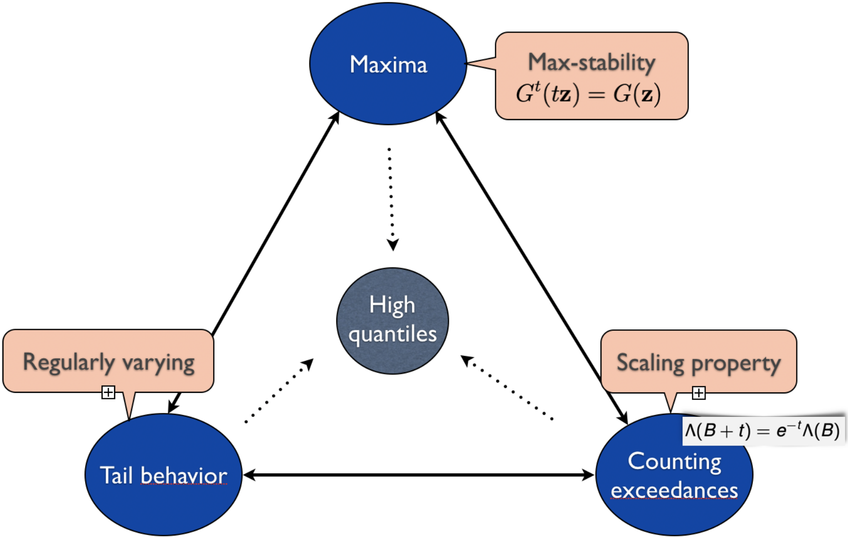


Angular (spectral) measure

$$\phi(B) = \frac{\mu\left(\{\frac{\mathbf{X}}{R} \in B\} \cap \{R > 1\}\right)}{\mu\left(\{R > 1\}\right)}$$



A multivariate summary



Variational auto-encoder (VAE) : how does it work

Sampling strategy of the VAE

Training data set : N samples $\mathbf{y}^{(i)}$ of $\mathbf{y}$.

VAE learnable sampling strategy :

- Sample $\mathbf{x}$ from $p(\mathbf{x})$;
- Sample $\mathbf{y}'$ from $p_\theta(\mathbf{y}'|\mathbf{x})$ parameterized by θ.

Purpose : Find θ such that $\boldsymbol{y}' \overset{d.}{=} \boldsymbol{y}$.

Learning criterion : Evidence Lower BOund (ELBO) maximization

VAE framework introduces a target distribution $q_\phi(\mathbf{x}|\mathbf{y})$ parameterized by ϕ to approximate the true posterior distribution.
Whatever q_ϕ, the following holds :

$$\log(p(\mathbf{x}^{(i)})) \geq L(\mathbf{x}^{(i)}, \theta, \phi),$$

with L the ELBO cost given by

$$\begin{aligned} L(\mathbf{y}^{(i)}, \theta, \phi) &= E_{q_\phi(\mathbf{x}|\mathbf{y}^{(i)})}\left[\log p_\theta(\mathbf{y}^{(i)}|\mathbf{x})\right] - D_{KL}\left(q_\phi(\mathbf{x}|\mathbf{y}^{(i)})||p(\mathbf{x})\right) \\ &\approx \frac{1}{L}\sum_{l=1}^{L} p_\theta(\mathbf{y}^{(i)}|\mathbf{x}^{(i,l)}) - D_{KL}\left(q_\phi(\mathbf{x}|\mathbf{y}^{(i)})||p(\mathbf{x})\right) \end{aligned}$$

VAE learning strategy aims to maximize $\sum_{i=1}^{n} L(\mathbf{y}^{(i)}, \theta, \phi)$ with respect to ϕ and θ.

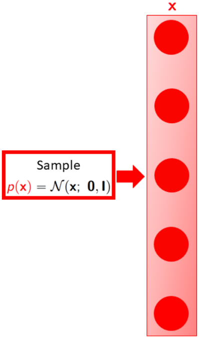
Standard VAE parameterization

Classically,

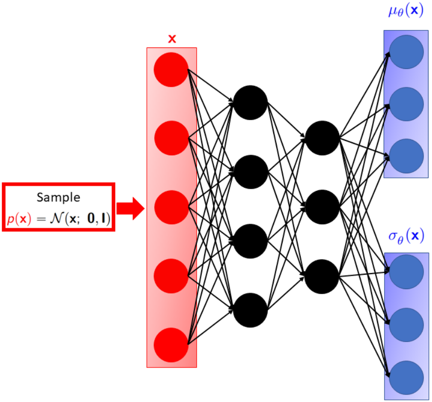
$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x};\ \mathbf{0}, \mathbf{I}),$$
$$q_\phi(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x};\ \mu_\phi(\mathbf{y}), \sigma^2_\phi(\mathbf{y})\mathbf{I}),$$
$$p_\theta(\mathbf{y}'|\mathbf{x}) = \mathcal{N}(\mathbf{y}\prime;\ \mu_\theta(\mathbf{x}), \sigma^2_\theta(\mathbf{x})\mathbf{I}),$$

with μ_θ, σ_θ, μ_ϕ and σ_ϕ neural network functions.

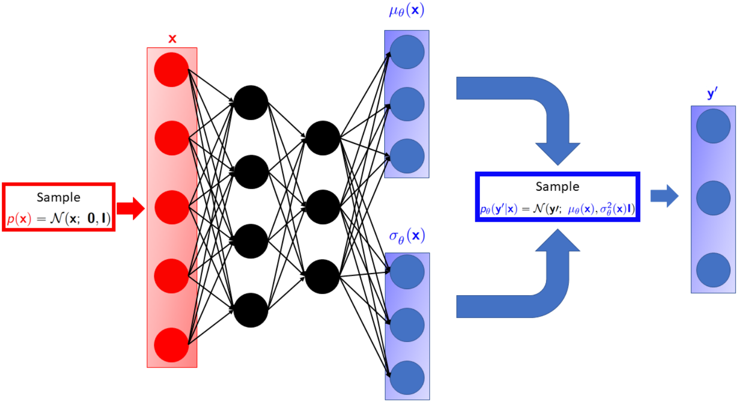
Generator architecture



Generator architecture



Generator architecture



Application to hydrological data

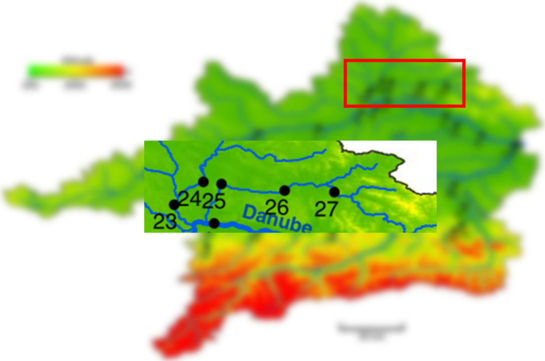


Figure – Gauging stations on the Danube river basin (*Asadi et al., 2015*).

Application to hydrological data

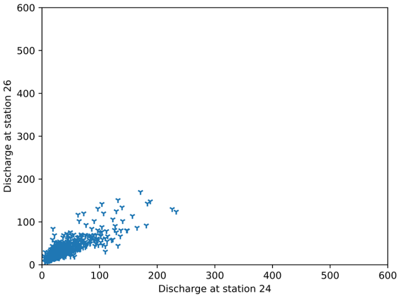


Figure – Training set of discharges at stations 24 and 26.

Application to hydrological data

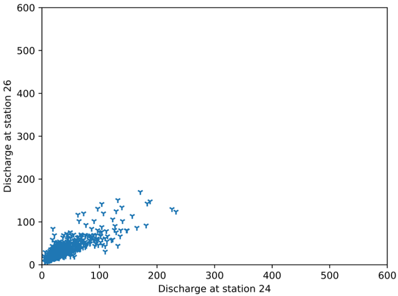


Figure – Training set of discharges at stations 24 and 26.

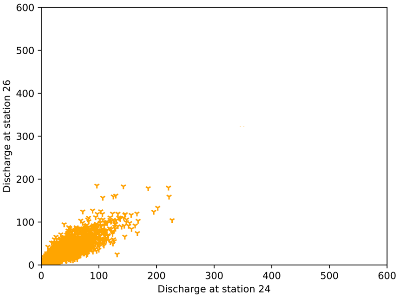


Figure – Emulation of 50 years of data with the standard VAE at stations 24 and 26.

Application to hydrological data

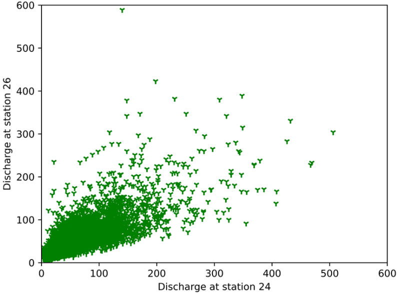


Figure – Test set of discharges at stations 24 and 26.

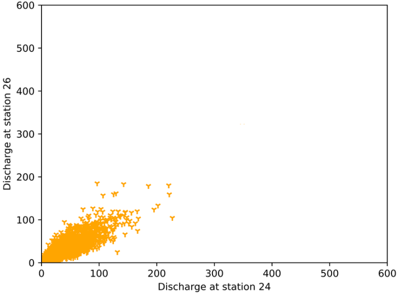


Figure – Emulation of 50 years of data with the standard VAE at stations 24 and 26.

Application to hydrological data

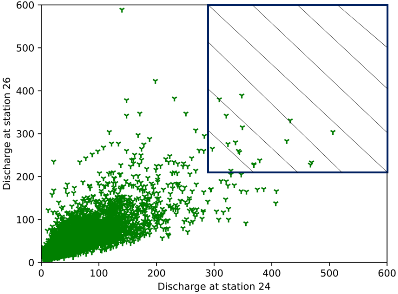


Figure – Test set of discharges at stations 24 and 26.

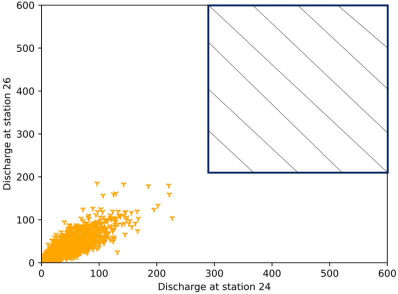


Figure – Emulation of 50 years of data with the standard VAE at stations 24 and 26.

Tail of standard VAE output distribution

Proposition

Using the Gaussian parameterization, the VAE cannot produce heavy-tail distributions ($p(y) \sim y^{-(\alpha+1)}$, α tail index).

Coupling VAE and extreme value theory

Generative models for extremes

GAN	VAE
Boulaguiem et al. (2022) *Allouche et al. (2022)* *Girard et al. (2025)* *Lhaut et al. (2026)*	*Zhang et al. (2026)*
Diffusion models	**Normalizing flows**
Yoon et al. (2023) *Pandey et al. (2024)* *Shariatian et al. (2025)*	*Wiese et al. (2019)* *Jaini et al. (2020)* *McDonald et al. (2022)* *Laszkiewicz et al. (2022)* *Hickling and Prangle (2024)*

Generative models for extremes : goal

Tail only
vs
Tail + Bulk

Boulaguiem et al. (2022)
Allouche et al. (2022)
Girard et al. (2025)
Lhaut et al. (2026)

Zhang et al. (2026)

Yoon et al. (2023)
Pandey et al. (2024)
Shariatian et al. (2025)

Wiese et al. (2019)
Jaini et al. (2020)
McDonald et al. (2022)
Laszkiewicz et al. (2022)
Hickling and Prangle (2024)

Generative models for extremes : strategy

Marginal preprocessing
vs
Intrinsic changes

Boulaguiem et al. (2022)
Allouche et al. (2022)
Girard et al. (2025)
Lhaut et al. (2026)

Zhang et al. (2026)

Yoon et al. (2023)
Pandey et al. (2024)
Shariatian et al. (2025)

Wiese et al. (2019)
Jaini et al. (2020)
McDonald et al. (2022)
Laszkiewicz et al. (2022)
Hickling and Prangle (2024)

Fun facts about the Pareto for heavy tails ($\xi > 0$)

Multivariate behaviors

By dividing the dataset by a “radius", say R, the radius becomes independent of the “angle" $\mathbf{\Theta} = X_i/R$ when the radius becomes large

$r(x) = \max(x_1, x_2)$ $r(x) = x_1 + x_2$ $r(x) = \min(x_1, x_2)$

Architecture of VAE for sampling multivariate regularly varying vector

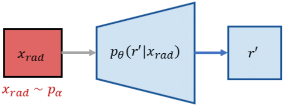


Figure – Scheme of our generative strategy

Architecture of VAE for sampling multivariate regularly varying vector

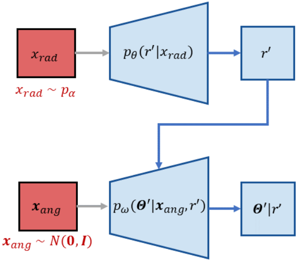


Figure – Scheme of our generative strategy

Architecture of VAE for sampling multivariate regularly varying vector

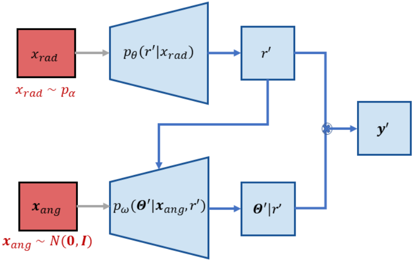


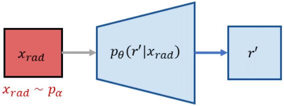
Figure – Scheme of our generative strategy

Fun fact about the Pareto for heavy tails ($\xi > 0$)

Marginal behavior

A Pareto can always be simulated by multiplying an unit exponential random variable by a independent inverse Gamma random variable

VAE for sampling heavy-tailed radius : setting



New parameterization :

$$p_\alpha(x_{rad}) = Inv\Gamma(x_{rad}; \alpha, 1),$$
$$q_\phi(x_{rad}|r) = Inv\Gamma(x_{rad};\ \alpha_\phi(r), \beta_\phi(r)),$$
$$p_\theta(r|x_{rad}) = \Gamma(r;\ \alpha_\theta(x_{rad}), \beta_\theta(x_{rad})).$$

VAE for sampling heavy-tailed radius : learning criterion

α, θ, ϕ are learned to optimise $\sum_{i=1}^{n} L(r^{(i)}, \alpha, \theta, \phi)$, with

$$L(r^{(i)}, \alpha, \theta, \phi) \approx -D_{KL}\left(q_\phi(x_{rad}|r^{(i)})||p_\alpha(x_{rad})\right) + \frac{1}{L}\sum_{l=1}^{L} p_\theta(r^{(i)}|x_{rad}^{(i,l)}),$$

$$\approx (\alpha_\phi(r^{(i)}) - \alpha)\psi(\alpha) - \log\frac{\Gamma(\alpha_\phi(r^{(i)}))}{\Gamma(\alpha)}$$

$$+ \alpha\log\beta_\phi(r^{(i)}) + \alpha_\phi(r^{(i)})\frac{1-\beta_\phi(r^{(i)})}{\beta_\phi(r^{(i)})}$$

$$+ \frac{1}{L}\sum_{l=1}^{L} p_\theta(r^{(i)}|x_{rad}^{(i,l)}),$$

with $x_{rad}^{(i,l)}$ sample from $q_\phi(x_{rad}|r^{(i)})$.

VAE for sampling heavy-tailed radius : ELBO maximization

- With some additional constraints on β_θ, the generated radius is heavy-tailed with tail index α,
- Tail index α could be learned directly during the training phase by maximizing L, without threshold selection.

Conditional VAE for sampling on the simplex : setting

$$
\begin{aligned}
p(\mathbf{x_{ang}}) &= \mathcal{N}(\mathbf{z_{ang}}\ ;\ \mathbf{0}, \mathbf{I}), \\
p_\omega(\mathbf{\Theta}' \mid \mathbf{x}_{ang}, r') &= \mathrm{Dir}(\mathbf{\Theta}'\ ;\ \mathbf{a}_\omega(\mathbf{x}_{ang}, r')). \\
q_\nu(\mathbf{x}_{ang}|\mathbf{s}, r) &\sim \mathcal{N}\left(\mathbf{x}_{ang}\ ;\ \mu_\nu(\mathbf{\Theta}, r), \Sigma_\nu(\mathbf{\Theta}, r)\right),
\end{aligned}
$$

μ_ν, Σ_ν, $\mathbf{a}_\omega$ neural network functions with parameters ν and ω.

A condition for sampling regularly varying random vector

The independence between the radius distribution and the angular distribution when $r \to +\infty$ is enforced :

$$\lim_{r \to +\infty} \mathbf{a}_{\omega}(\mathbf{x}_{ang}, r) = \mathbf{a}_{\infty}(\mathbf{z}_{ang})$$

Datasets

Synthesized dataset

$$R \sim 2\mathbf{U} \times \mathbf{Inv`}(\alpha = 1.5 \; ; \; \beta = 0.6)$$
$$\mathbf{\Theta}|r \sim \mathbf{Dir}\left(\alpha_1(r), \alpha_1(r), \alpha_2(r), \alpha_2(r), \alpha_2(r)\right),$$

Training set : 250 samples
Test set : 10000

Danube dataset

5 stations
Training set : 750 samples
Test set : 18000 samples

Visualization of limit angular measures for synthesized dataset

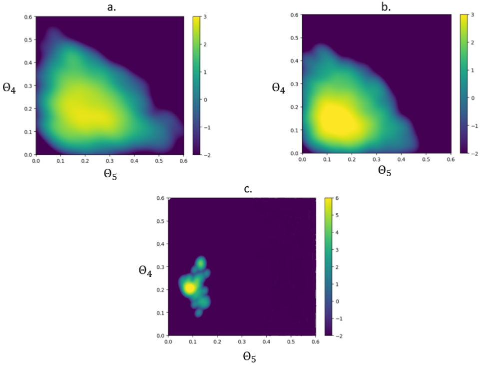


Figure – Log probability of the estimated limit angular measure obtained with a. ExtVAE, b. true distribution, c. ParetoGAN, projected on two axes (named θ_4 and θ_5).

Danube discharges : probability of joint extremes

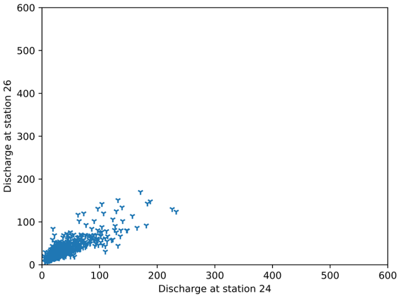


Figure – Training set of discharges at stations 24 and 26.

Danube discharges : probability of joint extremes

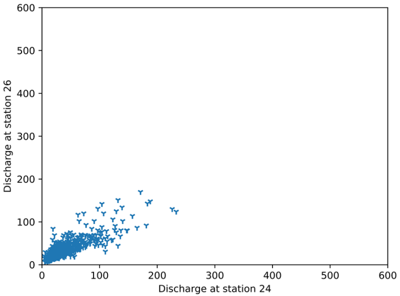


Figure – Training set of discharges at stations 24 and 26.

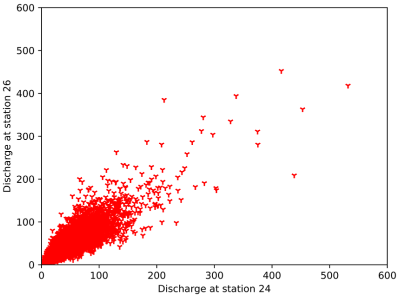


Figure – Emulation of 50 years of data with our VAE at stations 24 and 26.

Danube discharges : probability of joint extremes

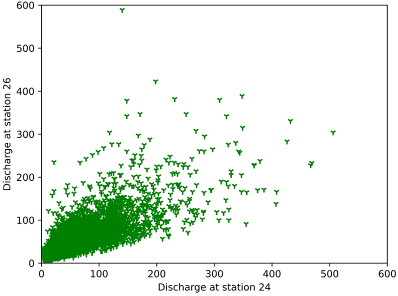


Figure – Test set of discharges at stations 24 and 26.

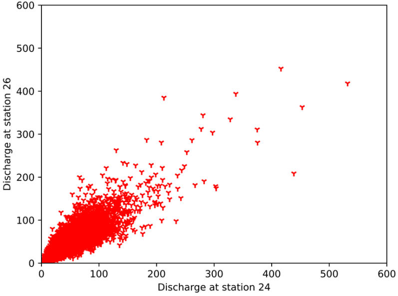


Figure – Emulation of 50 years of data with our VAE at stations 24 and 26.

Danube discharges : probability of joint extremes

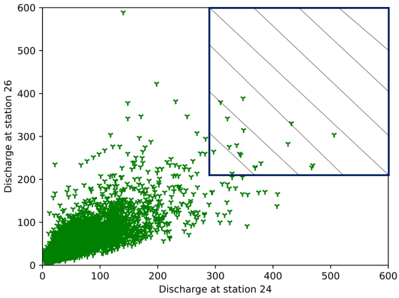


Figure – Test set of discharges at stations 24 and 26.

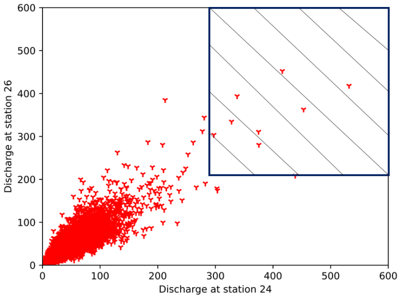


Figure – Emulation of 50 years of data with our VAE at stations 24 and 26.

Danube discharges : probability of joint extremes

Table – Proportion (in %) of elements satisfying $A_j^{(q)}$ in the training and test datasets as well as datasets sampled from StdVAE ExtVAE and ParetoGAN, where $A_j^{(q)} = \bigcap_{i=23}^{j} X_i > u_i^{(q)}$ with $u_i^{(q)}$ the value of the flow at q quantile for station i in test set.

	quantile = 0.9				
	Train	Test	UExtVAE	Std VAE	Pareto GAN
$A_{25}^{(0.9)}$	5.9	6.6	5.0	3.8	5.5
$A_{26}^{(0.9)}$	4.9	6.0	4.6	3.3	5.5
$A_{27}^{(0.9)}$	3.8	5.1	4.1	2.5	4.4

Danube discharges : probability of joint extremes

Table – Proportion (in %) of elements satisfying $A_j^{(q)}$ in the training and test datasets as well as datasets sampled from StdVAE ExtVAE and ParetoGAN, where $A_j^{(q)} = \bigcap_{i=23}^{j} X_i > u_i^{(q)}$ with $u_i^{(q)}$ the value of the flow at q quantile for station i in test set.

	q = 0.99				
	Train	Test	UExtVAE	Std VAE	Pareto GAN
$A_{25}^{(0.99)}$	0.0	0.48	0.22	0.01	0.13
$A_{26}^{(0.99)}$	0.0	0.4	0.2	0.0	0.13
$A_{27}^{(0.99)}$	0.0	0.25	0.18	0.0	0.09

Take home messages from this study

General setup

- Risk managers have an interest to sample outside of the range of the training set, especially for unprecedented extremes
- Coupling machine learning generative approaches with extreme value theory is a new and active research field

Take home messages from this study

General setup

- Risk managers have an interest to sample outside of the range of the training set, especially for unprecedented extremes
- Coupling machine learning generative approaches with extreme value theory is a new and active research field

Our specific contribution

- Variational auto-encoders based on extreme value theory outperform classical methods
- Better accuracy for predicting joint extremes than ParetoGAN for tested datasets
- Precise inference of tail index based on an ELBO cost
- No need for threshold selection, either to estimate tail index, or to study limit angular measure

Open questions concerning extremes and ML techniques

General setup

- How to frame the question (extremes in the covariate and/or outputs)
- How to deal with multi-model approximation
- How to insert physical in extreme representations
- Other types of approaches (Bouchet et al.)
- Suggestions ?

References

Allouche, M., Girard, S., and Gobet, E.
EV-GAN : Simulation of extreme events with ReLU neural networks.
Journal of Machine Learning Research, 23(150), 1-39. (2022).

Bhatia, S., Jain, A., and Hooi, B.
Exgan : Adversarial generation of extreme samples.
In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 8, pp. 6750-6758). (2021, May).

Huster, T., Cohen, J., Lin, Z., Chan, K., Kamhoua, C., Leslie, N. O., ... and Sekar, V.
Pareto gan : Extending the representational power of gans to heavy-tailed distributions.
International Conference on Machine Learning (pp. 4523-4532). PMLR, (2021, July).

References

Kingma, D. P., and Welling, M.
Stochastic gradient VB and the variational auto-encoder.
In Second international conference on learning representations, ICLR (Vol. 19, p. 121), (2014, April).

Lafon, N., Naveau, P., and Fablet, R.
A VAE approach to sample multivariate extremes.
(2023).

Resnick, S. I.
Heavy-tail phenomena : probabilistic and statistical modeling.
Springer Science and Business Media, (2007).

Rezende, D. J., Mohamed, S., and Wierstra, D.
Stochastic backpropagation and approximate inference in deep generative models.
In International conference on machine learning (pp. 1278-1286). PMLR., (2014, June).

Wiese, M., Knobloch, R., Korn, R., and Kretschmer, P.
Quant GANs : deep generation of financial time series.
Quantitative Finance, 20(9), 1419-1440. (2020).