33rd International Conference on Coastal Engineering
ICCE2012
Santander, Spain 1-6 July 2012

POT METHODS: A NEW INSIGHT INTO THE ESTIMATION OF EXTREME VALUE DISTRIBUTIONS – APPLICATION TO EXTREME WAVE HEIGHTS

Franck Mazas, **Luc Hamm**, ARTELIA Maritime
Philippe Garat, Laboratoire Jean-Kuntzmann, University of Grenoble

CONTEXT AND PURPOSE OF THE STUDY

CURRENT SITUATION The widely spread methodology for determining extreme values of environmental variables has converged towards the so-called **GPD-Poisson model**:

- Extraction of extreme i.i.d. data from a time series using the **Peaks-Over-Threshold (POT)** approach (for wave heights: storm peaks above a physical threshold) ;
- Determination of a statistically meaningful threshold u_s;
- Computation of the excesses above u_s then **Maximum Likelihood** estimation of a **2-parameter Generalized Pareto Distribution (GPD-2)** : $F_{Y;k,\sigma}(y) = 1 - \left(1 + k\frac{y}{\sigma}\right)^{-1/k}$, with $Y = X - u_s | X > u_s$;
- Derivation of extreme values for desired return periods (quantiles) and confidence intervals.

WHAT'S BOTHERING US? Unexpected sensitivity of ML estimates to the lowest and largest data values.

HOW DID WE ADDRESS THIS?

- Classical IAHR **Haltenbanken** extreme wave heights dataset: comparison of 2-parameter MLE (MLE/GPD-2) and 3-parameter L-Moments (LMOM/GPD-3) estimations of the GPD;
- **GPD simulated data**: examination of the behavior of the model likelihood and parameter estimates;
- **Re-sampling of Haltenbanken data**: sensitivity of the estimates to the largest value of the dataset.

The GPD-3 is defined by: $F_{Y;k,\sigma,\mu}(y) = 1 - \left(1 + k\frac{y-\mu}{\sigma}\right)^{-1/k}$, with$Y = X - u_s | X > u_s$, $k \in \mathbb{R}$, $\sigma > 0$, $\mu < y$

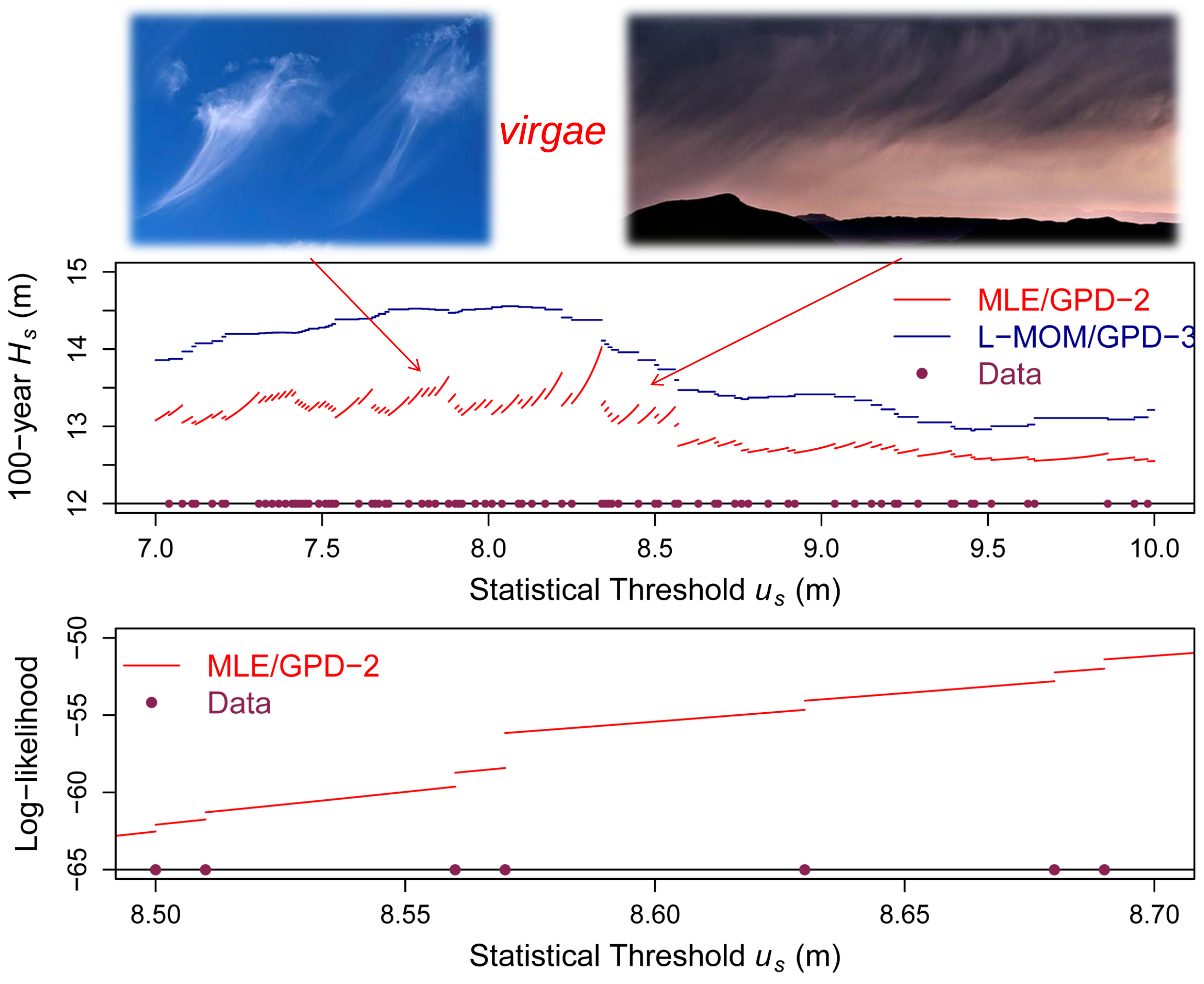

***Figure 1 -** Haltenbanken - Evolution of 100-year Hs (top) and maximized log-likelihood (bottom, zoom) with the threshold u_s*

SENSITIVITY TO THE LOWEST DATA VALUE

WHAT'S WRONG DOC? Let's carry out an accurate sensitivity study with respect to the threshold (***Fig. 1 - top***) using the classical IAHR Haltenbanken dataset: the quantile (and parameter) ML estimates are **unstable** between 2 consecutive data values *("**virgae**").* A slight translation of the sample may lead to a significant change in the estimated quantiles! The maximum likelihood of the GPD-2 model (***Fig. 1 - bottom***) is continuously increasing with the threshold between two consecutive data values and shows a positive jump when the threshold reaches a new sample value.

HOW DO WE EXPLAIN THAT? Considering that:

- you still work with the **same physical events** (storms) when u_s varies between 2 data so the final result shouldn't change;
- letting u_s vary between two data sets the **origin of the distribution** while this role is usually devoted to a location parameter;
- the role of the threshold should be limited to the **selection of extreme data** to be fitted;

we claim the **need for a location parameter μ**, distinct from the threshold, to improve the stability of the results.

COMPARISON WITH A GPD-3 For comparative purpose, we fit a 3-parameter GPD (GPD-3) with the **L-Moments estimator** (Hosking, 1990). Quantile (***Fig. 1 - top***) as well as shape k and scale σ parameter estimates are **stable**! But the 100-yr H_s is 1 m higher…

EXAMINATION OF THE GPD-3 LIKELIHOOD Test: random generation of an ordered sample of simulated GPD data $Y_{i:N}$ $(N = 25)$, then computation of the likelihood for each triple (k, σ, μ) within$\mathbb{R} \times \mathbb{R}^* \times]-\infty; Y_{1:N}[$, with the condition $\sigma > -k(Y_{N:N} - \mu)$. For each constant value of μ, there is a maximum likelihood in the (k, σ) plane (***Fig. 2***). But the global maximum for the full (3D) parameter space is reached at the **open upper bound** of the interval of validity of μ (i.e. $Y_{1:N}$), with **non-null derivatives** (***Fig. 3***), while $(\hat{k}, \hat{\sigma})$ doesn't converge (***Fig. 4***). Still, the asymptotic properties of the MLE require that the maximum be reached on an **interior point of an open set** (Lehmann, 1983). These properties are not proven!

→ The use of MLE is quite **dubious**.

→ An accurate and robust estimation of μ is **necessary**!

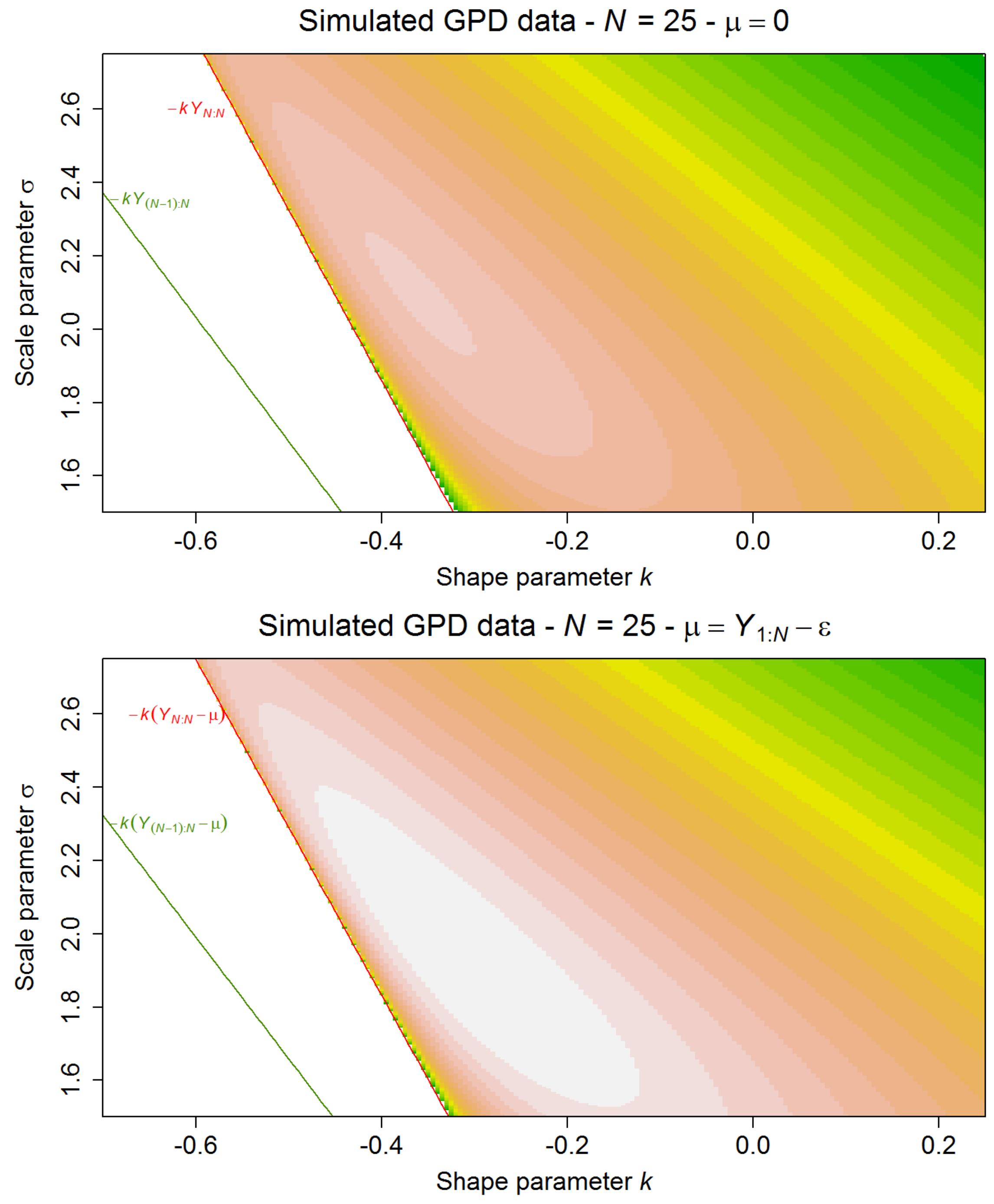

***Figure 2 -** GPD simulated data (N = 25) – 3-parameter log-likelihood function for two hyperplanes of the parameter space μ = constant (max for grey shades)*

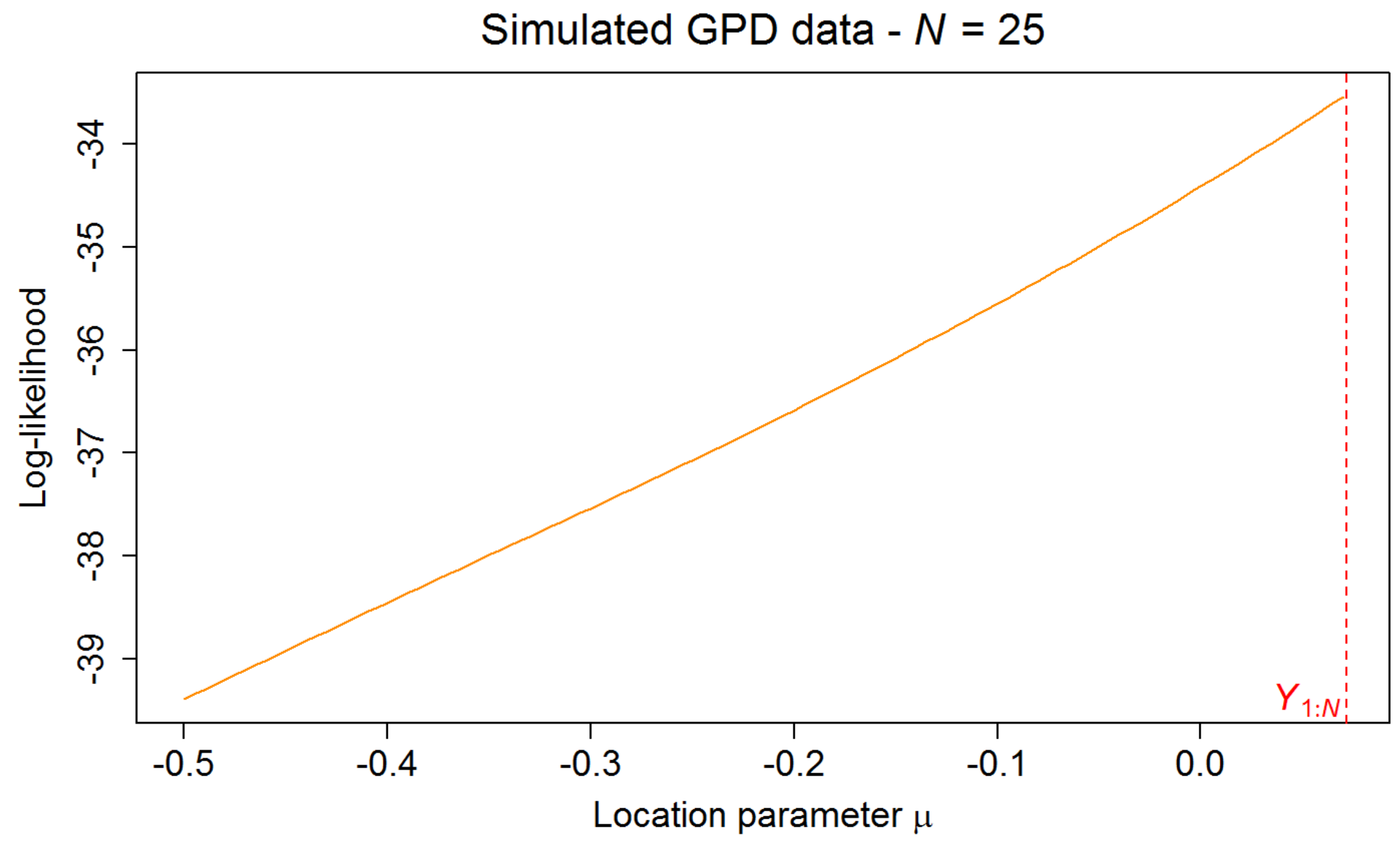

***Figure 3 -** GPD simulated data (N = 25) – Evolution of the 2-parameter maximized log-likelihood profile function with respect to μ*

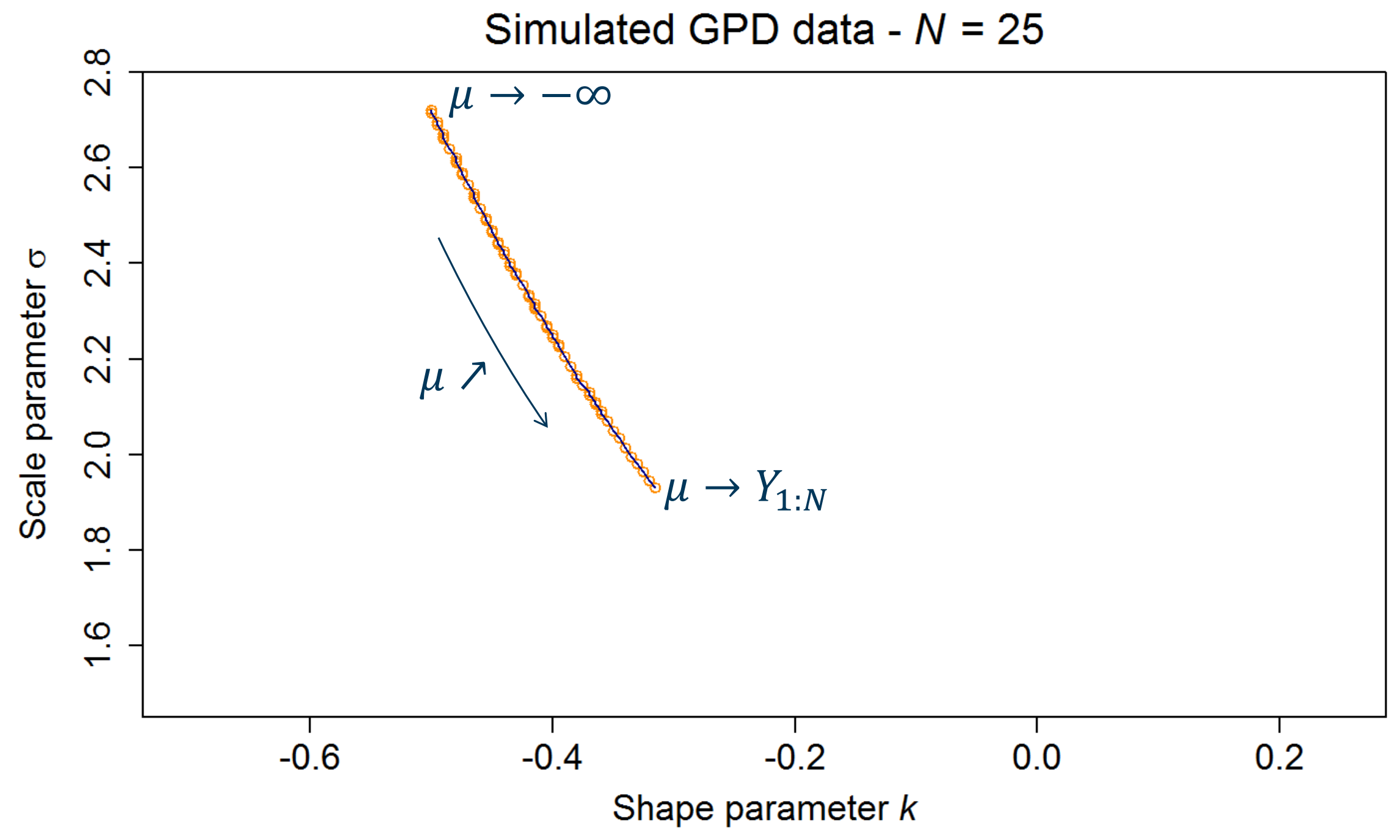

***Figure 4 -** GPD simulated data (N = 25) – Evolution of the ML estimates of $(\hat{k}; \hat{\sigma})$ when μ increases*

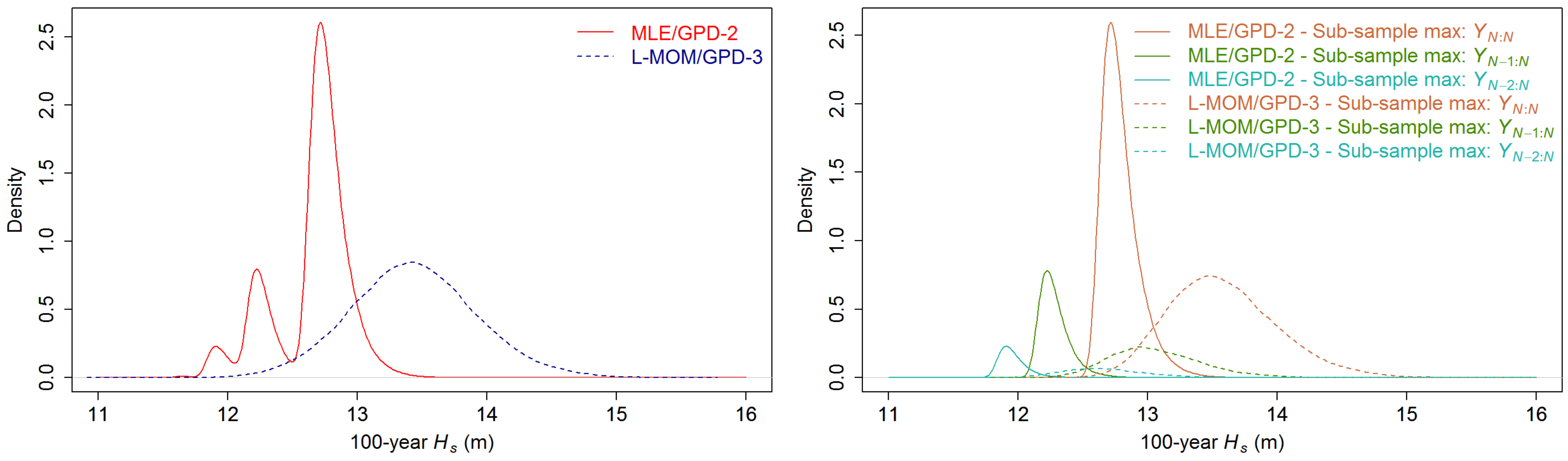

***Figure 5 -** Re-sampling of Haltenbanken dataset (sample rate = 75 %) – Density of 100-year H_s*

SENSITIVITY TO THE LARGEST DATA VALUE

LINK WITH REAL WORLD, PLEASE! Imagine wave height measurements from a buoy that could randomly fail 25% of the time… or never! What will be the influence of statistics on the results?

STATISTICAL RECIPE Ingredients: 1 appropriate dataset, 1 computer, 1 statistical software, 1 brain, coffee.

- Take the Haltenbanken dataset ($N = 46$ H_s storm peak measurements).
- Create a **random sub-sample** with 75% of the data… then reiterate to get 100,000 sub-samples out of the 13,340,783,196 possibilities!
- Fit both models (MLE/GPD-2 and LMOM/GPD-3) to the 100,000 subsamples and compute 100-yr H_s.
- Compare. Get some coffee. Search an explanation. Get some more coffee.

WHAT WE SEE The density (***Fig. 5 - left***) of 100-yr H_s for ML-estimated GPD-2 exhibits **several narrow and sharp peaks**, while the L-Moments-estimated GPD-3 shows a **single broad and flat peak**.

WHY WE SEE THAT Each peak corresponds to the family of the sub-samples whose maximum is the largest ($Y_{N:N}$), second largest ($Y_{N-1:N}$), third largest ($Y_{N-2:N}$)… of the original sample (***Fig. 5 - right***).

There is a strong sensitivity of both models to the value of the sample maximum. This sensitivity is less visible for the LMOM/GPD-3 because of a large variance, but still exists: bad luck, not a perfect model!

→ In practice, if the analysis is based on incomplete data, this may yield incorrect estimates of extreme waves or other environmental data.

→ Significant difference between estimates from both models. Large dispersion for LMOM/GPD-3 while estimates of MLE/GPD-2 are focused on a few peaks linked with the sub-sample maximum.

CONCLUSIONS AND PERSPECTIVES

CONCLUSIONS

- The statistical threshold u_s selects the data to be fitted and defines the "extreme domain", while the location parameter μ sets the origin of the distribution. These two roles are distinct and u_s and μ should NOT be confused.
- A comparison MLE/GPD-2 vs LMOM/GPD-3 shows that introducing μ yields stable results between two consecutive storm peaks, consistent with the physics.
- The global maximum of the 3-parameter likelihood function is reached at an open bound of the parameter space with non-null derivatives: the asymptotic properties of the MLE are not proven.
- Estimation may be quite sensitive to the value of the sample maximum.
- Results are similar with other distributions such as Weibull or Gamma (Mazas & Hamm, 2011).

PERSPECTIVES

→ Which estimator for GPD-3? Can hybrid estimators perform better than L-Moments?

→ Application of local influence approach (Cook, 1986) in progress.

References :

Lehmann, E.L. (1983). *Theory of point estimation.* Wiley & Sons, Australia.
Hosking, J. R. M. (1990). L-moments: Analysis and Estimation of Distributions using Linear Combinations of Order Statistics. *J. R. Statist. Soc.*, **52**(1), 105-124.
Mazas, F., Hamm, L. (2011). A multi-distribution approach to POT methods for determining extreme wave heights. *Coastal Engineering*, **58**, 385-394.
Cook, R. D. (1986). Assessment of Local Influence. *J. R. Statist. Soc.*, series B, **48**(2), 133-169

Contact:

Franck Mazas - ARTELIA Maritime
6, rue de Lorraine
38130 Echirolles – FRANCE
franck.mazas@arteliagroup.com