Neural Bayes inference for extremal dependence models

EDT-day

Lídia André

Jenny Wadsworth and Raphaël Huser

28th March 2025

Motivation

- Maximum likelihood inference requires the knowledge of a likelihood function
- Likelihood evaluation might be **computationally costly** when there is
  - inversion of functions;
  - numerical integration;
  - both
- Examples:
  - Models that are available to **interpolate** between two classes of extremal dependence (Wadsworth et al., 2017; Huser and Wadsworth, 2019; Engelke et al., 2019)

Point estimation

- General setting:
  - Replicate data: $\boldsymbol{Z} := (\boldsymbol{Z}_1', \ldots, \boldsymbol{Z}_n')' \in \mathcal{S}^n$ where $\boldsymbol{Z}_i \sim f(\boldsymbol{z}_i \mid \boldsymbol{\theta})$
  - Sampling space: $\mathcal{S} = \mathbb{R}^d$
  - Parameter space: $\Theta = \mathbb{R}^p$
- Point estimators: $\hat{\boldsymbol{\theta}} : \mathcal{S}^n \to \Theta$
- Bayes estimators: minimise a weighted average of the risk at $\boldsymbol{\theta}$ (Bayes risk)

$$r_\Omega(\hat{\boldsymbol{\theta}}(\cdot)) = \int_\Theta \int_{\mathcal{S}^n} L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(\boldsymbol{z})) f(\boldsymbol{z} \mid \boldsymbol{\theta}) \mathrm{d}\boldsymbol{z} \mathrm{d}\Omega(\boldsymbol{\theta})$$

  - $\Omega(\cdot)$: prior measure for $\boldsymbol{\theta}$
  - $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(\boldsymbol{z}))$: absolute error loss

Neural Bayes estimators (Sainsbury-Dale et al., 2024)

- Bayes estimator that is approximated using a **neural network** as function approximator
- Neural point estimator: $\hat{\boldsymbol{\theta}}(\boldsymbol{Z} \mid \boldsymbol{\gamma})$
  - $\boldsymbol{\gamma}$: parameters of the neural network
- Neural Bayes estimator (NBE): $\hat{\boldsymbol{\theta}}(\boldsymbol{Z} \mid \boldsymbol{\gamma}^*)$

$$\boldsymbol{\gamma}^* = \arg\min_{\boldsymbol{\gamma}} r_{\Omega}(\hat{\boldsymbol{\theta}}(\cdot; \boldsymbol{\gamma}))$$

- NBEs just need to be trained **once**!
  - subsequent estimates are obtained in (milli)seconds

Neural Network architecture

- DeepSets framework (Zaheer et al., 2017)
  - For any permutation $\tilde{\boldsymbol{Z}}$ of the independent replicates in $\boldsymbol{Z}$:

$$\hat{\boldsymbol{\theta}}(\boldsymbol{Z};\gamma) = \hat{\boldsymbol{\theta}}(\tilde{\boldsymbol{Z}};\gamma)$$

- Multilayer perceptron (MLP)

$$\hat{\boldsymbol{\theta}}(\boldsymbol{Z};\gamma) = \boldsymbol{\phi}\left(\boldsymbol{S}(\boldsymbol{Z};\gamma_{\boldsymbol{\psi}});\gamma_{\boldsymbol{\phi}}\right)$$

$$\boldsymbol{S}(\boldsymbol{Z};\gamma_{\boldsymbol{\psi}}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}\left(\boldsymbol{Z}_i;\gamma_{\boldsymbol{\psi}}\right)$$

  - $\boldsymbol{\psi} : \mathbb{R}^d \to \mathbb{R}^q$ and $\boldsymbol{\phi} : \mathbb{R}^q \to \mathbb{R}^p$: neural networks
  - $\boldsymbol{S}$: summary statistics

Neural Network architecture

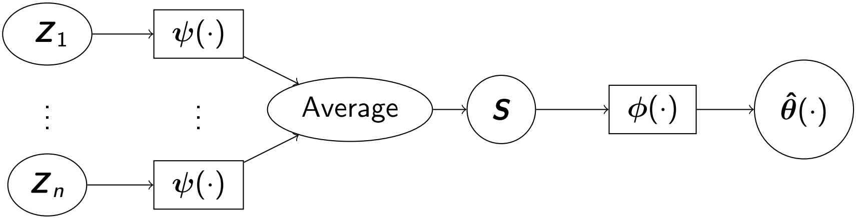
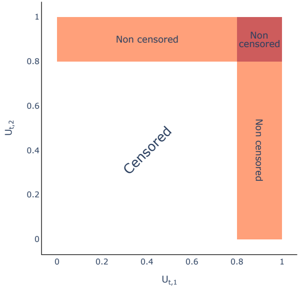


Figure 1: Schematic of the DeepSets architecture.

NBEs for censored data (Richards et al., 2023)

Censor non-extreme values to prevent them affecting the extremal dependence estimation

- $\boldsymbol{Z}^* = ((\boldsymbol{Z}_1^*)', \ldots, (\boldsymbol{Z}_n^*)')'$
- Censored values set to $c \in \mathbb{R}$ outside the support
- $\boldsymbol{I}_i$: indicator vectors
  - if 1 then the observations are censored



NBEs for censored data (Richards et al., 2023)

- NBEs are trained using an augmented data set $\boldsymbol{A} = ((\boldsymbol{Z}^*)', \boldsymbol{I}')$
- Censoring level τ can be treated as **fixed** or **variable**
- If **variable**

$$\hat{\boldsymbol{\theta}}(\boldsymbol{A}; \tau, \gamma) = \boldsymbol{\phi}\left(\boldsymbol{S}(\boldsymbol{A}; \gamma_{\boldsymbol{\psi}}, \tau); \gamma_{\boldsymbol{\phi}}\right)$$

with $\boldsymbol{S}(\boldsymbol{A}; \gamma_{\boldsymbol{\psi}}, \tau) = \left(\boldsymbol{S}(\boldsymbol{A}; \gamma_{\boldsymbol{\psi}})', \tau\right)'$ and $\boldsymbol{S}(\boldsymbol{A}; \gamma_{\boldsymbol{\psi}})$ defined as before

Parameter estimation: Model of Wadsworth et al. (2017)

$$(Z_1, Z_2) = R(V_1, V_2), \quad R \perp\!\!\!\perp (V_1, V_2)$$

$$R \sim \mathrm{GPD}(1, \xi) \text{ and } V \sim \mathrm{Beta}(\alpha, \alpha)$$

$$(V_1, V_2)' = \frac{(V, 1 - V)'}{\max(V, 1 - V)} \in \Sigma$$

with $\Sigma = \{\boldsymbol{v} = (v_1, v_2)' \in \mathbb{R}_+^2 : \max(v_1, v_2) = 1\}$.

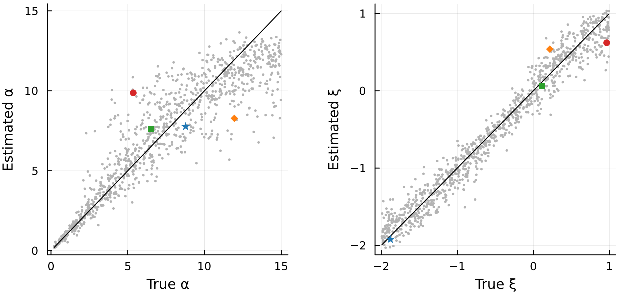
- $\xi > 0$: AD (asymptotic dependence, i.e., large values occur together)
- $\xi \leq 0$: AI (asymptotic independence)

Parameter estimation: Priors

- $\alpha \sim \mathrm{Unif}(0.2, 15)$
- $\xi \sim \mathrm{Unif}(-2, 1)$
- $T \sim \mathrm{Unif}(0.5, 0.99)$
- $N \sim \mathrm{Unif}(\{100, 101, \ldots, 1500\})$

Sample size and censoring level are treated as random variables, N and T respectively.

Assessment of NBEs



Assessment of NBEs: Uncertainty quantification

1. Non-parametric bootstrap procedure:
   - $B = 400$ bootstrap samples
   - $\boldsymbol{\theta}$ is re-estimated
   - 95% confidence intervals are obtained
2. Neural interval estimator:
   - trained under the quantile loss function: $L_q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \sum_{k=1}^{p} (\hat{\theta}_k - \theta_k)(I_{(\hat{\theta}_k > \theta_k)} - q)$, for probability quantiles $q = \{0.025, 0.975\}$
   - marginal 95% central credible intervals are approximated

Assessment of NBEs: Uncertainty quantification

Table 1: Coverage probability and average length of the 95% uncertainty intervals obtained via a non-parametric bootstrap procedure and via the neural interval estimator averaged over 1000 models fitted using a NBE (rounded to 2 decimal places).

Parameter	Bootstrap procedure		Interval estimator	
	Coverage	Length	Coverage	Length
α	0.76	3.59	0.96	6.68
ξ	0.84	0.49	0.98	0.81

Assessment of NBEs: Extremal structure

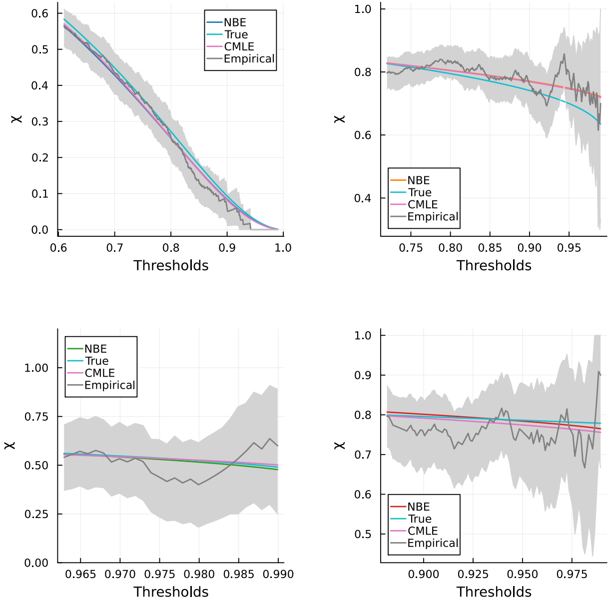
$$\chi = \lim_{u \to 1} \chi(u) = \lim_{u \to 1} \Pr(F_Y(Y) > u \mid \Pr(F_X(X) > u)$$

We have AD if $\chi > 0$ and AI if $\chi = 0$.

Table 2: Coverage probability and average length of the 95% confidence intervals for $\chi(u)$ at levels $u = \{0.80, 0.95, 0.99\}$ obtained via a non-parametric bootstrap procedure averaged over 1000 models fitted using a NBE (rounded to 2 decimal places).

$\chi(u)$	Coverage	Length
$\chi(0.80)$	0.91	0.06
$\chi(0.95)$	0.89	0.09
$\chi(0.99)$	0.88	0.09

Assessment of NBEs: Extremal structure



Comparison with censored MLE

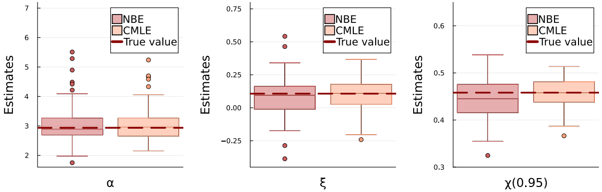


Figure 2: $\boldsymbol{\theta} = (2.94,\ 0.11)'$ and $\tau = 0.79$.

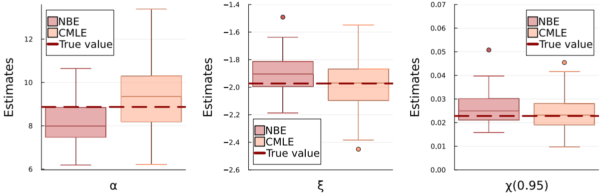


Figure 3: $\boldsymbol{\theta} = (8.87,\ -1.97)'$ and $\tau = 0.60$.

Comparison with censored MLE

- Once trained, getting an estimate through this NBE takes on average 0.676 seconds.
- An estimate through censored MLE takes on average 92.611 seconds.
- This is a 137 fold speed-up

Model selection: neural Bayes classifier (NBC)

- Information criteria like AIC/BIC cannot be used
- **Solution:** Treat model type as a random variable M
- $K \geq 2$ candidate models and M takes values in $\{1, \ldots, K\}$
- M is inferred jointly with $\boldsymbol{\theta}$ (based on $\boldsymbol{Z}$): $(\boldsymbol{\theta}', M)' \mid \boldsymbol{Z}$
- Can be decomposed as the product of $\boldsymbol{\theta} \mid (\boldsymbol{Z}', M)'$ and $M \mid \boldsymbol{Z}$
- $\boldsymbol{\theta} \mid (\boldsymbol{Z}', M)'$ is split into m problems: $\boldsymbol{\theta}_m \mid (\boldsymbol{Z}', M = m)'$– trained with NBEs

Model selection: neural Bayes classifier (NBC)

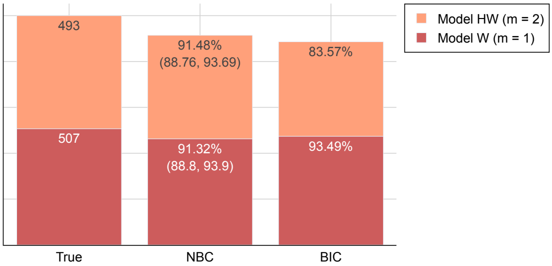
- Construct a neural network that approximates $M \mid \boldsymbol{Z} = \boldsymbol{z}$ for any data input $\boldsymbol{Z} = z$
- Neural Bayes classifier (NBC): $\hat{\boldsymbol{p}}(\boldsymbol{Z}; \boldsymbol{\gamma})$

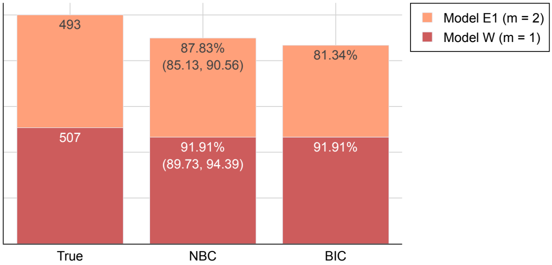
$$\boldsymbol{\gamma}^* = \underset{\boldsymbol{\gamma}}{\arg\min} \; - \sum_{m=1}^{K} p_m \int_{\Omega_m} \int_{\mathcal{S}_m^n} \log\left(\hat{p}_m(\boldsymbol{z}; \boldsymbol{\gamma})\right) f_m\left(\boldsymbol{z} \mid \boldsymbol{\theta}_m\right) \mathrm{d}\boldsymbol{z} \mathrm{d}\Omega_m(\boldsymbol{\theta}_m)$$

  - $p_m = \Pr(M = m) = 1/K$, and $\sum_{m=1}^{K} p_m = 1$
  - $\hat{p}_m(\boldsymbol{z}; \boldsymbol{\gamma})$: approximate posterior probability of model m

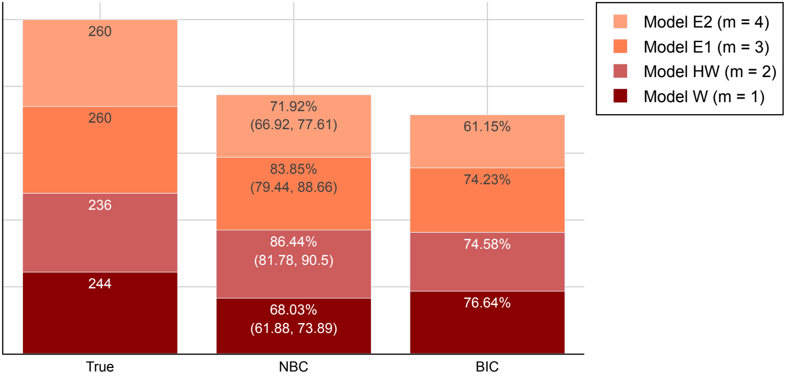
- Identical to a classification problem
- Loss function: categorical cross-entropy
- MLP similar to that of the parameter estimation procedure

$K = 2$ candidate models





$K = 4$ candidate models



Misspecified scenarios

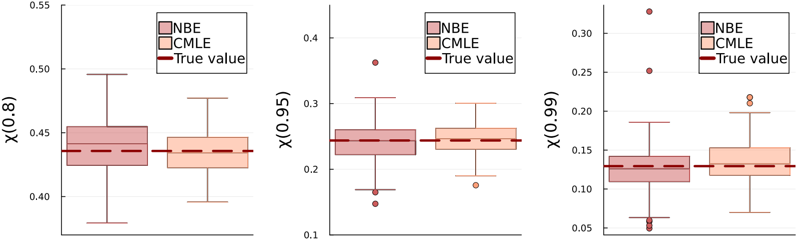
- Data from a Gaussian copula with $\rho = 0.5$ (AI) and $\tau = 0.65$
- 100 samples each with $n = 1000$

Table 3: Proportion of times each model was selected through the NBC and through BIC (left), and proportion of AD and AI samples identified by the NBE and CMLE (right). All the values are rounded up to 2 decimal places.

Model	NBC	BIC
Model W	0.02	0.30
Model HW	0.88	0.69
Model E1	0.02	0.00
Model E2	0.08	0.01

Method	AD	AI
NBE	0.02	0.98
CMLE	0.03	0.97

Misspecified scenarios



Case study: changes in geomagnetic field fluctuations

- Space weather events cause large fluctuations in the geomagnetic field - geomagnetically induced currents (GICs)
- GICs can cause: disruptions on power grids, railway systems, etc
- **Interest:** assess whether a large magnitude of GICs occurring in one location has an effect on another location
- Pairwise χ of the rate of change of the horizontal component of geomagnetic field $\mathrm{d}B_H/\mathrm{t}$ as a measure of magnitude of GICs

Case study

- $n = 1500$ and $\tau \in \{0.60, 0.65, \ldots, 0.95\}-$ results for $\tau = 0.85$
- Pairs: (SCO, STF), (SCO, STJ) and (STF, STJ)

Table 4: International Association of Geomagnetism and Aeronomy (IAGA) code, and location of the observatory for the three locations considered.

IAGA code	Country	Latitude	Longitude
SCO	Greenland	70.48	−21.97
STF	Greenland	67.02	−50.72
STJ	Canada	47.60	−52.68

Case study

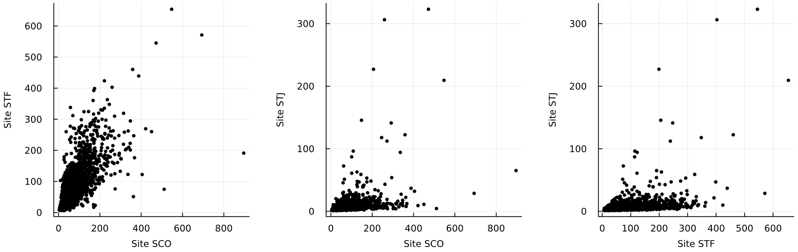


Figure 4: Daily maxima absolute one-minute changes in $\mathrm{d}B_H/\mathrm{d}t$ measurements between three pairs of locations: (SCO, STF) on the left, (SCO, STJ) in the middle, and (STF, STJ) on the right.

Case study

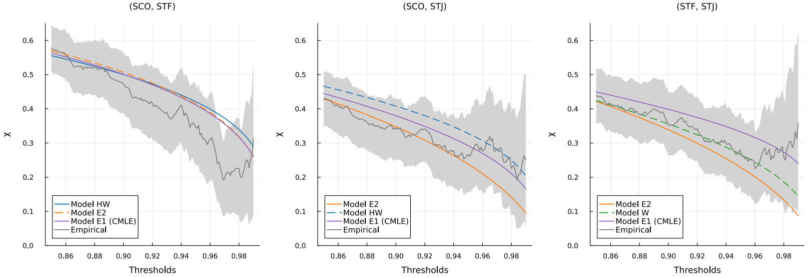


Figure 5: Empirical (in grey) and model $\chi(u)$ estimated via the NBE for $u \in [0.85, 0.99]$ for the models with the two highest posterior probabilities. Estimated model $\chi(u)$ for the selected model through BIC is given by the purple line. The 95% confidence bands were obtained by block boostrapping.

Conclusion: Advantages

- Robust and amortised statistical toolbox
- Fast inference method
- Well calibrated extremal dependence properties
- Sensitivity analysis for censoring level

Conclusion: Limitations

- Biased results
- Poor coverage of bootstrap-based uncertainty results
- Subjectivity in the neural network architecture
- Need to choose prior distributions

Questions?

Thank you all for listening!

References I

Engelke, S., Opitz, T., and Wadsworth, J. (2019). Extremal dependence of random scale constructions. *Extremes*, 22:623–666.

Huser, R. and Wadsworth, J. L. (2019). Modeling spatial processes with unknown extremal dependence class. *Journal of the American Statistical Association*, 114(525):434–444.

Richards, J., Sainsbury-Dale, M., Zammit-Mangion, A., and Huser, R. (2023). Neural Bayes estimators for censored inference with peaks-over-threshold models.

Sainsbury-Dale, M., Zammit-Mangion, A., and Huser, R. (2024). Likelihood-Free Parameter Estimation with Neural Bayes Estimators. *The American Statistician*, 78(1):1–14.

Wadsworth, J. L., Tawn, J. A., Davison, A. C., and Elton, D. M. (2017). Modelling across extremal dependence classes. *J. R. Statist. Soc. B*, 79:149–175.

References II

Zaheer, M., Kottur, S., Ravanbhakhsh, S., Póczos, B., Salakhutdinov, R., and Smola, A. J. (2017). Deep Sets. *Advances in Neural Information Processing Systems*, 30.