Neural network based inference for complex dependence models

KAUST/INRAE/Lancaster workshop

Lídia André

Jenny Wadsworth and Raphaël Huser

1st October 2024

Engineering and Physical Sciences Research Council

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:
  - Weighted copula model (André et al., 2024)
  - 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: $\boldsymbol{Z} := (\boldsymbol{Z}_1', \ldots, \boldsymbol{Z}_n')'$ 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 $\boldsymbol{\theta}$
  - $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(\boldsymbol{z}))$: squared 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: $\hat{\boldsymbol{\theta}}(\boldsymbol{Z} \mid \boldsymbol{\gamma}^*)$

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

- 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 $\boldsymbol{Z}^*$ of the independent replicates in $\boldsymbol{Z}$:

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

- Dense Neural Network (DNN)

$$\hat{\theta}(\boldsymbol{Z};\gamma) = \phi\left(\boldsymbol{T}(\boldsymbol{Z};\gamma_{\psi});\gamma_{\phi}\right)$$
$$\boldsymbol{T}(\boldsymbol{Z};\gamma_{\psi}) = \boldsymbol{a}\left(\{\psi(\boldsymbol{Z}_i;\gamma_{\psi} : i = 1,\ldots,n\}\right)$$

  - $\psi : \mathbb{R}^d \to \mathbb{R}^q$ and $\phi : \mathbb{R}^q \to \mathbb{R}^p$: neural networks
  - $\boldsymbol{a} : (\mathbb{R}^q)^n \to \mathbb{R}^q$: permutation-invariant set function
  - $a_s(\cdot)$: returns the **elementwise average** over its input set for $s = 1,\ldots,q$
  - $\boldsymbol{T}$: summary statistics

Neural Network architecture

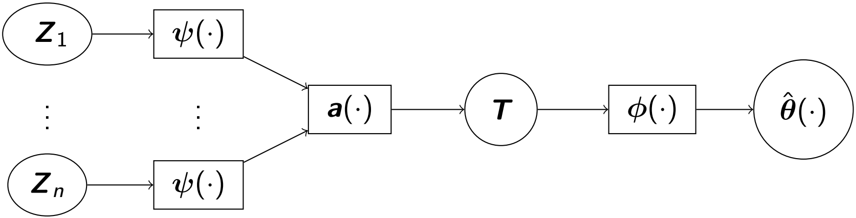
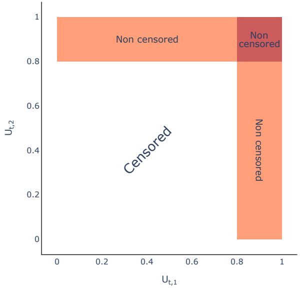


Figure 1: Schematic of the DeepSets architecture.

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 $\boldsymbol{A} = ((\boldsymbol{Z}^*)', \boldsymbol{I}')$
- τ can be treated as **fixed** or **variable**
- If **variable**

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

with $\boldsymbol{T}(\boldsymbol{A}; \gamma_{\boldsymbol{\psi}}, \tau) = \left(\boldsymbol{T}(\boldsymbol{A}; \gamma_{\boldsymbol{\psi}})', \tau\right)'$ and $\boldsymbol{T}(\boldsymbol{A}; \gamma_{\boldsymbol{\psi}})$ is defined as before

Parameter estimation

Model of Wadsworth et al. (2017)

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

- $R \sim \mathrm{GPD}(1, \xi)$ and $V \sim \mathrm{Beta}(\alpha, \alpha)$
- $(V_1, V_2) = (V, 1 - V)/\|(V, 1 - V)\|_\infty$
- $\xi > 0$: asymptotic dependence
- $\xi \leq 0$: asymptotic independence

Priors:

$$\alpha \sim \mathrm{Unif}(0.2, 15), \qquad \xi \sim \mathrm{Unif}(-2, 1),$$
$$\tau \sim \mathrm{Unif}(0.5, 0.99), \qquad n \sim \mathrm{Unif}(100, 1500)$$

(Sample size n and censoring level τ are treated as variable)

Assessment of NBEs

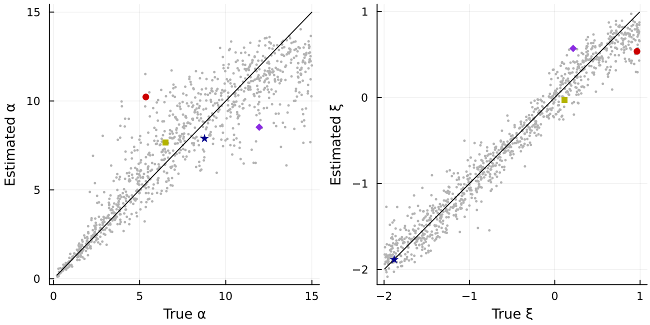
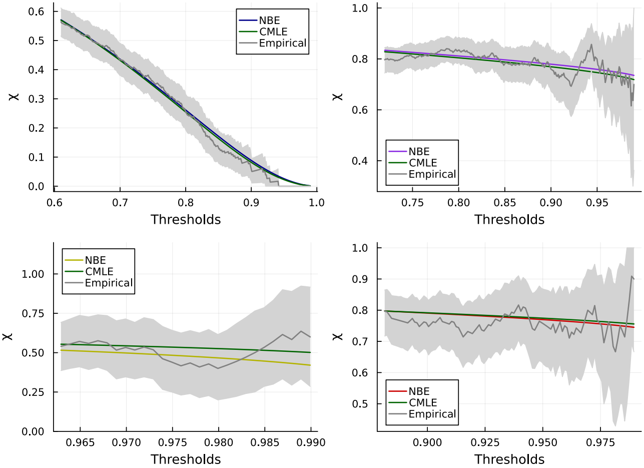


Table 1: Coverage probability and average length of the 95% credible intervals for 1000 parameter estimates obtained through the neural Bayes estimator (rounded to 2dp).

Parameter	Coverage probability	Average length
α	0.72	3.07
ξ	0.76	0.41

Assessment of NBEs



Comparison with censored MLE

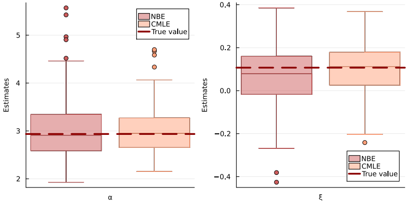


Figure 2: $\boldsymbol{\theta} = (2.94, 0.11)$ and $\tau = 0.79$.

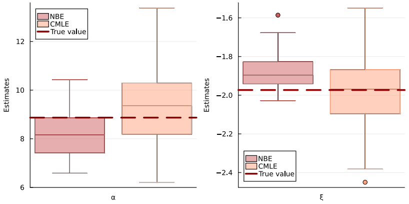


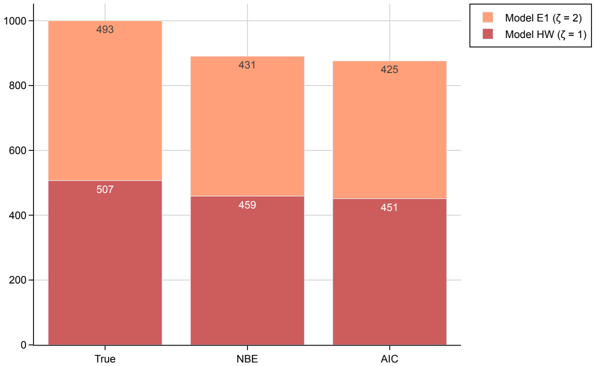
Figure 3: $\boldsymbol{\theta} = (8.87, -1.97)$ and $\tau = 0.60$.

- Once trained, getting an estimate through this NBE takes on average 0.361 seconds.
- An estimate through censored MLE takes on average 92.611 seconds.

Model selection

- Information criteria like AIC/BIC cannot be used
- **Solution:** Neural networks as classifiers for model selection
- Classification problem: M candidate models
- Model index: $\zeta \in \{1, \ldots, M\}$
- Each model has equal probability of being drawn
- DNN that takes $\boldsymbol{Z}_i$ as input, and returns the probabilities of it "belonging" to model with index ζ

$M = 2$ and $\zeta \sim \text{Bernoulli}(1/2)$



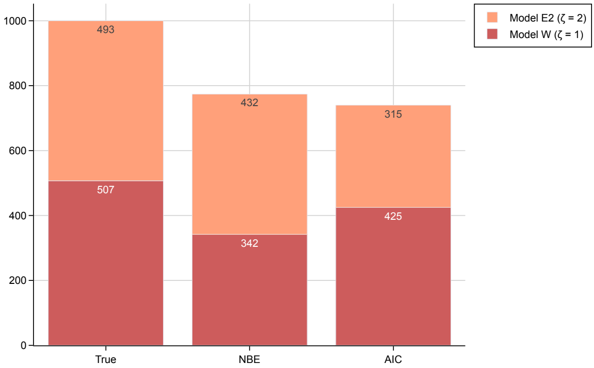
NBE

- Model HW: 92.90% $\in$ (90.59, 95.24)
- Model E1: 87.42% $\in$ (84.45, 90.18)

AIC

- Model HW: 88.95%
- Model E1: 86.21%

$M = 2$ and $\zeta \sim \text{Bernoulli}(1/2)$



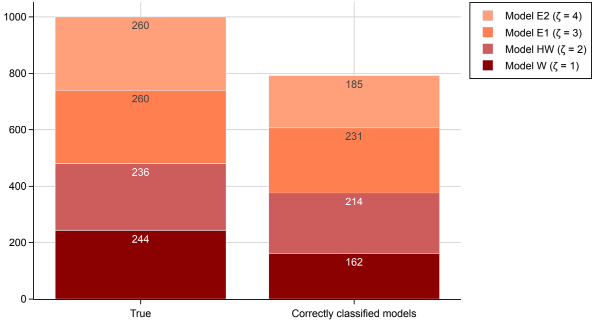
NBE

- Model W: 67.46% $\in$ (63.52, 71.05)
- Model E2: 87.63% $\in$ (84.66, 90.52)

AIC

- Model W: 83.83%
- Model E2: 63.89%

$M = 4$ and $\zeta \sim \text{Multinomial}(1/4, 1/4, 1/4, 1/4)$



- Model W: $66.39\% \in (60.42, 72.69)$
- Model HW: $90.68\% \in (87.60, 93.96)$
- Model E1: $88.85\% \in (85.02, 92.69)$
- Model E2: $71.15\% \in (66.26, 76.34)$

Currently working on

- Quantify the uncertainty of the NBEs for parameter estimation using the quantile loss
- Finish the comparison with classical tools such as AIC for model selection
- Have an ensemble of NBEs to try and get better results in the model selection procedure
  - The current NBEs seem to be sensitive to the model coding

A note on the weighted copula model

$$c(u, v; \boldsymbol{\theta}) = \frac{(1 - \pi(u, v; \gamma))c_b(u, v; \rho) + \pi(u, v; \gamma)c_t(u, v; \alpha)}{K(\boldsymbol{\theta})}$$

- Through likelihood inference it was not feasible to have one of these 4 models as c_t and now we are able to
- For the same model (when likelihood inference is feasible)
  - An MLE estimate takes on average 11536.88 seconds (3 hours and 12 minutes)
  - An NBE estimate takes on average 0.325 seconds
  - This is 35 542 faster!

A note on the weighted copula model

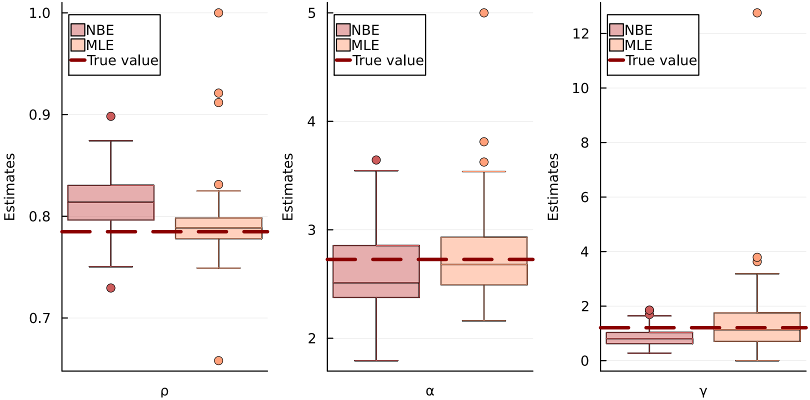


Figure 4: $\boldsymbol{\theta} = (0.78, 2.73, 1.21)$.

Questions?

Thank you all for listening!

References I

André, L., Wadsworth, J., and O'Hagan, A. (2024). Joint modelling of the body and tail of bivariate data. *Computational Statistics and Data Analysis*, 189:107841.

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.

References II

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.

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