Neural Bayes Estimators for Fast and Efficient Inference with Spatial Peaks-Over-Threshold Models

Jordan Richards[1] Matthew Sainsbury-Dale[1,2]
Andrew Zammit-Mangion[2] Raphaël Huser[1]

[1]King Abdullah University of Science and Technology (KAUST)

[2]Centre for Environmental Informatics, National Institute for Applied Statistics Research Australia, University of Wollongong

Motivation

Likelihood-based inference for spatial extremal processes is computationally problematic in moderate-to-high dimension (sites) D.

- Intractable, or **computationally-expensive**, likelihood functions and/or they require (left) censoring to mitigate bias.
- We construct a likelihood-free inference technique to **emulate censored likelihood-based inference** for these models.

Motivating example: max-stable processes

Max-stable processes (MSPs), which arise as the only possible non-degenerate limit of pointwise maxima of i.i.d random fields, are popular models for spatial extremal dependence.

- Number of terms in the likelihood grows faster-than-exponentially;
- **Computational tractability** of the likelihood is limited (generally) to $D \leq 12$ (Castruccio et al., 2016);
- A lot of time has been spent on researching efficient likelihood-based inference techniques for MSPs, e.g., via **pairwise likelihoods**;
- Computational issues are compounded by censoring.

Castruccio, S., Huser, R., and Genton, M. G. (2016). High-order composite likelihood inference for max-stable distributions and processes. *JCGS* 25.4: 1212-1229.

Motivation: censoring

- Likelihood estimators for spatial extremal dependence models are typically highly biased if spatial extreme events include marginally non-extreme values (Huser et al., 2016);
- Models can also be **misspecified**, e.g., we may fit a MSP (**defined for pointwise maxima**) to all observations;
- Can be mitigated in a peaks-over-threshold framework:
  - treat non-extreme observations as **censored**, i.e., **not fully-observed if below some high threshold** c,
  - where c is typically taken to be the τ-quantile, for $\tau < 1$ close to one;
  - decreases the contribution of low observations to the likelihood;

Huser, R., Davison, A. C., and Genton, M. G. (2016). Likelihood estimators for multivariate extremes. *Extremes*, 19:79–103.

Censoring (cont.)

- The contribution of an observation to the **censored-likelihood** is a C-variate integral, where $C \leq D$ is the number of censored values;
- Likely to be **intractable** for any $C > 0$ and **expensive** for large C;
- Solution: use likelihood-free methods, e.g., neural estimators.
- We want to build a neural estimator that **imitates** censoring, i.e., takes censored data as input and **learns to utilise this censoring in a meaningful way**.

Neural estimators

- A **neural estimator** $\hat{\theta}(\mathbf{Z})$ is a neural network that takes in data $\mathbf{Z}$ as input and provides a parameter point estimate $\boldsymbol{\theta}$ as an output.
- Their construction is simple:
  - Sample (many) parameter vectors $\boldsymbol{\theta}$ from a prior π.
  - Simulate $\boldsymbol{Z}$ from the model, conditional on these parameters.
  - Train a neural network that maps the simulated data $\boldsymbol{Z} \mapsto \boldsymbol{\theta}$ to the true parameters by minimising some loss function $L(\boldsymbol{\theta}, \hat{\theta}(\mathbf{Z}))$.
- Strengths:
  - **Likelihood free**.
  - **Very fast** (once trained) with **predictable run-time**.
  - **Accurate**.
  - An example of **amortised inference**.
- We adapt neural Bayes estimators to allow for censored data as input.

Neural Bayes estimators

- Neural Bayes estimators (NBEs) are **neural estimators** designed to minimise the **Bayes risk** (Sainsbury-Dale et al., 2022);

$$r_\pi(\hat{\boldsymbol{\theta}}(\cdot)) \equiv \int_\Theta L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(\mathbf{Z}))p(\mathbf{Z} \mid \boldsymbol{\theta})\mathrm{d}\mathbf{Z}\mathrm{d}\pi(\boldsymbol{\theta}),$$

  associated with $L(\cdot,\cdot)$ and π.

- They inherit the attractive properties of Bayes estimators (e.g., consistency, asymptotic efficiency, asymptotic normality);
- We minimise the Bayes risk with L as the absolute error loss, which targets the posterior median;
- NBEs have been shown to work well for **spatial models** and **fully-observed data**, but cannot handle censored Z.

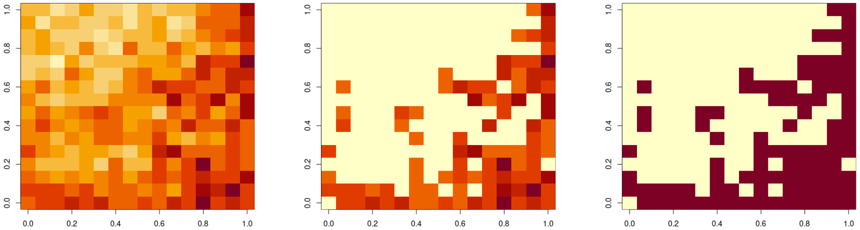
Sainsbury-Dale, M., Zammit-Mangion, A., and Huser, R. (2022) .Neural Point Estimation for Fast Optimal Likelihood-Free Inference. *arXiv:2208.12942*

Handling censored inputs

- NBEs are usually trained on uncensored data $\mathbf{Z}$;
- To emulate censoring, we **communicate to the neural network**:
  - i) which values should be treated as censored;
  - ii) these values should be treated differently to non-censored values.
- NBE input specification:
  - Transform input data $\mathbf{Z} \mapsto \mathbf{Z}^*$ onto **standard margins with a finite lower-endpoint** (this **does not alter** the dependence structure in $\mathbf{Z}$),
  - Set "censored values" to constant c^* outside distribution support,
  - Train NBE on new input data $(\mathbf{Z}^*, \mathcal{I})$ (a two-channel image), where $\mathcal{I}$ is a one-hot encoded map of sites without censoring.
- i) Implicitly encoded in $\mathcal{I}$ is info. about the dependence model and τ;
- ii) Censored values outside of "normal" range, so treated differently.

New input

Left: Realisation $\mathbf{Z}$ from a max-stable process.
Centre: $\mathbf{Z}^*$ with $\tau = 0.9$ censoring and $c^* = 0$.
Right: one-hot encoding $\mathcal{I}$.



Models

We consider inference with 3 **popular models**:

- Max-stable process (MSP) and inverted MSP (1/MSP),
- HW process (Huser and Wadsworth, 2019),

$$\{Z(\mathbf{s})\} = R^{\delta}\{W(\mathbf{s})^{1-\delta}\},$$

  where W is a standard Matérn Gaussian process with the same margins as the heavy-tailed r.v. R and $\delta \in [0, 1]$;
- If $\delta \geq 1/2$, then $Z(\cdot)$ is **asymptotically dependent.**

We illustrate gains in both **comp. and stat. efficiency**, relative to a censored likelihood-based approach, using a NBE.

Huser, R. and Wadsworth, J. L. (2019). Modeling spatial processes with unknown extremal dependence class. *JASA. 114(525):434–444*

Simulation study 1: outline

- Consider MSP and IMSP (1/MSP) with $\tau = 0.9$;
- Both have **range** $\lambda > 0$ and **smoothness** $\kappa \in (0, 2]$, with unif. priors;
- Simulate 200 replicates on a 16×16 grid;
- Compare to the **competing likelihood-based approach**, i.e., censored pairwise-likelihood (cPL);
- cPL(∞): all pairs; cPL(3), only those within 3 units.

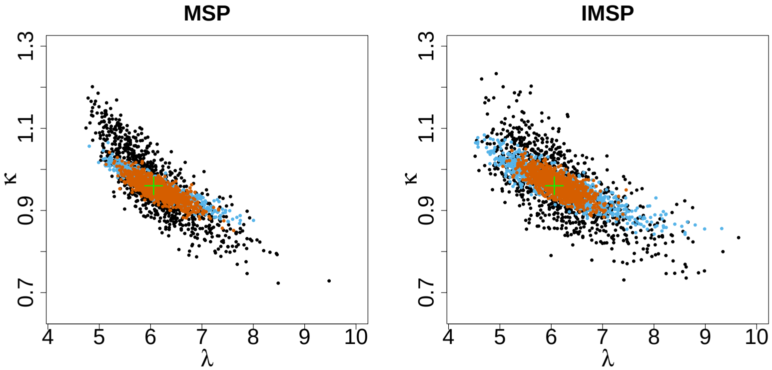
Simulation study 1: results

Marginal test risk (s.d.) evaluated on 1000 **test parameter sets**.

	MSP		IMSP	
	λ	κ	λ	κ
NBE	**2.4 (0.1)**	**1.8 (0.1)**	**2.6 (0.1)**	**2.2 (0.1)**
cPL (3)	3.5 (0.1)	2.2 (0.1)	4.6 (0.2)	3.2 (0.1)
cPL (∞)	4.3 (0.1)	6.4 (0.2)	5.4 (0.2)	6.8 (0.2)

Simulation study 1: joint distribution

- Empirical joint dist. of estimates with single true vector $\boldsymbol{\theta}$;
- Black: cPL(∞). Blue: cPL(3). Brown: NBE.
- **NBE captures well the joint distribution, but with lower variance than the competing likelihood approach.**



Simulation study 1: conclusion

- Takeaways:
  - NBE gives **large improvements in statistical efficiency**;
  - **Improvements in computational efficiency**!
    Amortised NBE takes exactly 0.0016 seconds to evaluate; cPL takes $\approx$ 2 to 10 minutes.
- We also showcase similar gains for r-**Pareto, Gaussian and HW processes.**
- These NBEs are now ready-to-ship! Anyone with data observed on a similar grid can immediately get parameter estimates (for these two models) in milliseconds...but only if they use $\tau = 0.9$.
- We can train an estimator for a general τ if we **supply** τ **to the estimator** as an input.
- The NBE **learns relationship between** τ, $\mathbf{Z}$ and $\mathcal{I}$.

Simulation study 2: outline

- Simulate $m = 200$ replicates of a HW process on a 16×16 grid in $[0, 16] \times [0, 16]$;
- Model has three parameters with priors $\lambda \sim \text{Unif}(0.2, 10)$, $\kappa \sim \text{Unif}(0.5, 2)$ and $\delta \sim \text{Unif}(0, 1)$;
- **For a test censoring level** $\tau^* = 0.919$, we compare two NBEs; one trained with τ fixed at $\tau = \tau^*$ and one with τ randomly drawn from a $\text{Unif}(0.85, 0.95)$ for each set of replicates used for training;

Simulation study 2: results

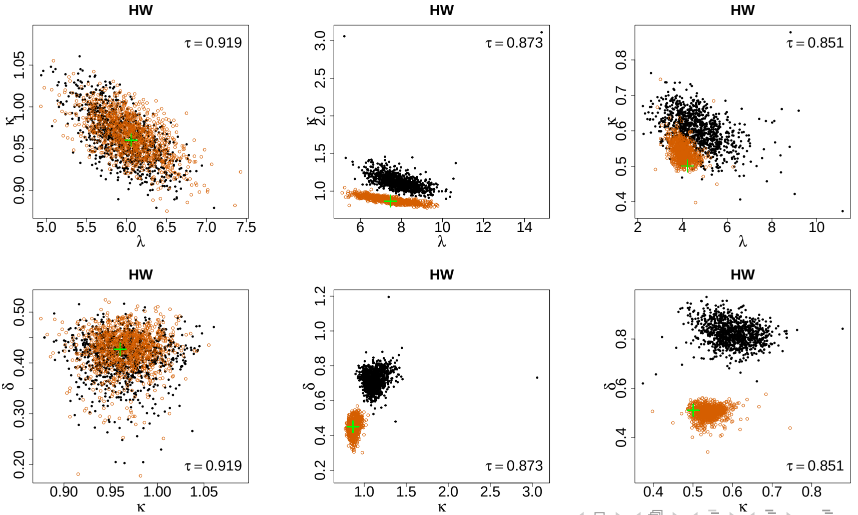
Marginal test risk (s.d.) evaluated on 1000 **test parameter sets with censoring level** τ^*.

τ	λ	κ	δ
random	**2.62 (0.07)**	**2.13 (0.05)**	**2.98 (0.09)**
fixed	2.75 (0.06)	2.41 (0.06)	3.25 (0.10)

- We can train an estimator for a general τ.
- Randomising τ during training improves the estimator performance.
- **Implication**: a new user will not need to retrain an estimator if they want to use a different censoring level.

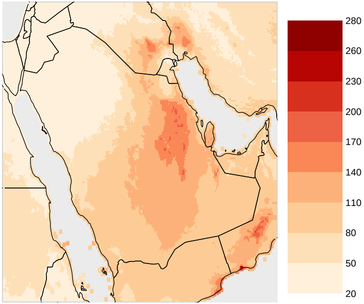
Simulation study 2: joint distribution

Different τ: (left) 0.919, (centre) 0.873, (right) 0.851.



Application

Application to monthly Saudi Arabian $PM_{2.5}$ (Van Donkelaar et al., 2021) concentrations shows the computational gains of our amortised estimator.



Observation of surface average $PM_{2.5}$ conc. ($\mu g/m^3$) for Jul. 2012.

Van Donkelaar, A., et al. (2021). Monthly global estimates of fine particulate matter and their uncertainty. *Environmental Science & Technology*, 55(22):15287–15300.

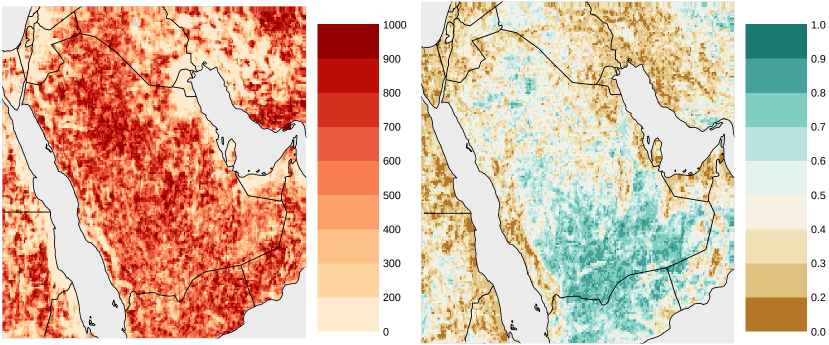
Application (cont.)

- Data are arranged on a 242×182 regular grid; monthly, 1998–2020.
- Fit local anisotropic HW processes with $\tau = 0.9$ (five params.);
- To all possible subsets of data on $G \times G$ grids for smoothing level $G \in \{4, 8, 16, 24, 32\}$. This is over 130,000 fits!
- Once an estimator is trained (roughly 24 to 72 hours), a single model fit takes between 1 and 4 milliseconds to estimate.
- Speed-up/dimension comparison:
  - Full censored likelihood-based inference is limited to $D \approx 6^2 = 36$ and takes roughly 12 hours per estimate;
  - NBE with $D = 32^2 = 1024$ and ≈ 10 million times faster.

Application (cont.)

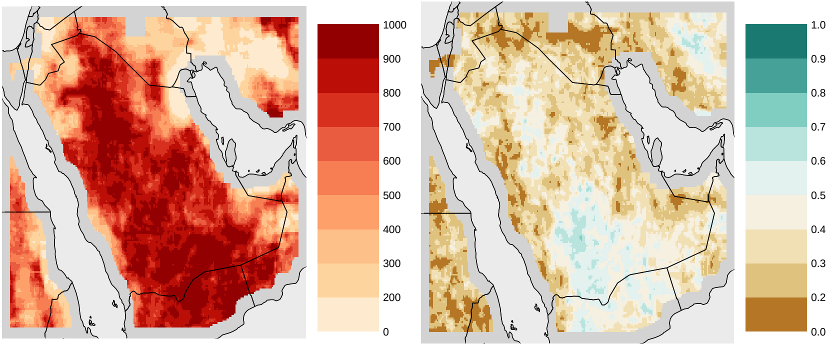
- Data are arranged on a 242×182 regular grid; monthly, 1998–2020.
- Fit local anisotropic HW processes with $\tau = 0.9$ (five params.);
- To all possible subsets of data on $G \times G$ grids for smoothing level $G \in \{4, 8, 16, 24, 32\}$. This is over 130,000 fits!
- Once an estimator is trained (roughly 24 to 72 hours), a single model fit takes between 1 and 4 milliseconds to estimate.
- Speed-up/dimension comparison:
  - Full censored likelihood-based inference is limited to $D \approx 6^2 = 36$ and takes roughly 12 hours per estimate;
  - NBE with $D = 32^2 = 1024$ and $\approx$ **10 million times faster.**

Results



λ (left) and δ (right) estimates for $G = 4$.

Results (cont.)



λ (left) and δ (right) estimates for $G = 16$.

Application (cont.)

- We can also perform parameter uncertainty assessment for free, with 1000 bootstrap estimates obtained within seconds;

- In total, our analysis uses 130 million model fits...

- **...which is far more than any comparable application[1]!**

[1] as far as we know.

Application (cont.)

- We can also perform parameter uncertainty assessment for free, with 1000 bootstrap estimates obtained within seconds;
- In total, our analysis uses 130 million model fits...
- **...which is far more than any comparable application[1]!**

[1] as far as we know.

Conclusion and further work

- We adapt NBEs to allow for **censored inputs** and construct **general estimators that are readily-applicable to new user data and censoring levels**;
- We showcase massive gains in computational and statistical efficiency when using our approach to inference;
- Perform a study of Arabian $PM_{2.5}$ concentration extremes (of unprecedented scale!).
- Further work includes:
  - Irregularly-sampled spatial data (watch this space!);
  - Extension to **high-dim.** priors;
  - **Full posterior estimation**;
- "*Likelihood-free neural Bayes estimators for censored inference with peaks-over-threshold models*" has just gone up on arXiv.
- `R` and `Julia` packages are in development.

References

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.

Sainsbury-Dale, M., Zammit-Mangion, A., and Huser, R. (2022). Neural point estimation for fast optimal likelihood-free inference. *arXiv preprint arXiv:2208.12942*.

Thanks for your attention!

SCAN ME

Scan for full details of my research.