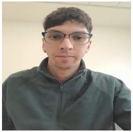
Bayesian Model Selection for Heavy-Tailed Data via Transformations: Exploring Possibilities for Amortized Inference

Luis Gutiérrez

PUC, Chile, llgutier@uc.cl

(a) Daniel Saavedra

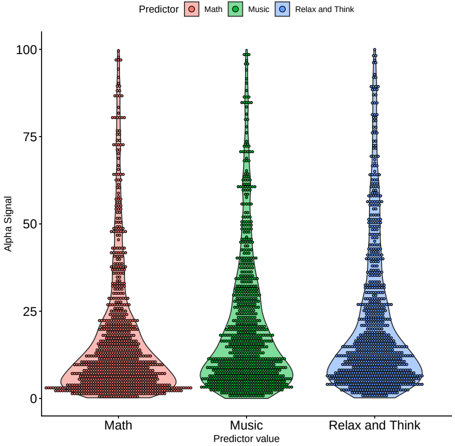


(b) Cristian Capetillo

(c) Clemente Ferrer

Outline

- Let $Y \in \mathbb{R}^+$ be a heavy-tailed response variable associated with a known vector of predictors $x \in \mathbb{R}^p$
- We assume that $Y \mid x \sim \mathrm{F}_{Y|x}$, and that $\mathrm{F}_{Y|x}$ belongs to the class of regularly varying distributions
- Let $\{x_j : j \in 2^p\}$ be the set of all subsets of x
- We aim to find an optimal model index k, say x_k, such that, $\mathrm{F}_{Y|x} = \mathrm{F}_{Y|x_k}$ and estimate the density $f_{Y|x_k}$ for high but finite quantiles



1. The responses are brain signals.
2. w denotes different activities, Math, Music, Relax and Think
3. Design vector $x = (1, x_1, x_2)^\top$

$$x_1 = \begin{cases} 1 & w = \text{Music} \\ 0 & w = \text{In other case} \end{cases}$$

$$x_2 = \begin{cases} 1 & w = \text{Relax and Think} \\ 0 & w = \text{In other case} \end{cases}$$

4. Compute the posterior for $f_{y|\{1\}}, f_{y|\{1,x_1\}}, f_{y|\{1,x_2\}}, f_{y|\{1,x_1,x_2\}}$, select the most probable and estimate the density

- The analysis of heavy-tailed data has been extensively studied using Extreme Value Theory (de Haan and Ferreira, 2006)
  - Block maxima (BM) method (Gumbel, 1958)
  - Peak-over-threshold (PoT) (Pickands, 1975)

- Recent contributions have proposed modeling heavy-tailed data without explicitly selecting a threshold or block
  - Bladt and Rojas-Nandayapa (2018) propose to use Phase-Type distributions to model the bulk and the tail of heavy-tailed distributions. Extremes 21(2), 285–313 (2018)
  - Palacios et. al. (2025) characterize the tails of the normalized generalized gamma process and develop classes of heavy-tailed mixture models. Bayesian Analysis 20(4),1315–1343 (2025)

- Let $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ be a sample from Y, and $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ a matrix with predictors

- To search over the model space induced by predictors, we define a binary vector $\gamma \in \{0, 1\}^p$

- If the $k-$th element of γ is equal to 1, then the $k-$th element of x is relevant

- For instance, if $p = 2$, then the model space is given by

  $$\mathcal{M} = \{\{1\}, \{1, x_1\}, \{1, x_2\}, \{1, x_1, x_2\}\} \iff \mathcal{A} = \{(0,0), (1,0), (0,1), (1,1)\}$$

We propose the following general model

$$\begin{aligned} y_i \mid x_i, \gamma_j, F_{Y|x\odot\gamma_j} &\sim F_{Y|x\odot\gamma_j}, \\ F_{Y|x\odot\gamma_j} &\sim \Pi_1(F_{Y|x\odot\gamma_j}), \\ \gamma_j &\sim \Pi_2(\gamma_j). \end{aligned} \tag{1}$$

- $F_{Y|x\odot\gamma_j}$ is not fixed in a particular parametric family, just its prior Π_1 is fixed
- Prior Π_2 enables the procedure to search for models in $\mathcal{M}$
- $F_{Y|x\odot\gamma_j}$ can change because of the model or because it is random itself

- The selection of prior Π_1 poses a challenge
  - Existing models are not sufficiently flexible or do not meet the conditions necessary for an exploration of the model space
    - Under the spike-and-slab prior is necessary to integrate out the model parameters (regression coefficients) to maintain the ergodic properties of the Markov chain
  - The Dirichlet process is a well-known flexible representation; however
  - Doss and Sellke (1982) showed that it does not exhibit heavy tails

Given these limitations, our proposal first applies a transformation that converts the heavy-tailed response into a light-tailed one, enabling us to fit a flexible, analytically tractable model in the transformed space

Theorem

Let $Y \in \mathbb{R}^+$ be a random variable following a regularly varying distribution F_Y with tail-index $\alpha > 0$. Let $Z = \ln(Y)$ be a transformation. Then, the random variable Z follows a light-tailed distribution F_Z.

- The proof resorts to Karamata's characterization and Karamata's representation theorem (Karamata, 1933)

Proposition

Let $Y \in \mathbb{R}^+$ be a random variable following a regularly varying distribution F_Y with tail-index $\alpha > 0$ and scale parameter σ. Then the random variable $Z = \ln(Y)$ has a location parameter $\ln(\sigma)$ and α behaves like a scale parameter in the tail of the distribution (in the limit at infinity).

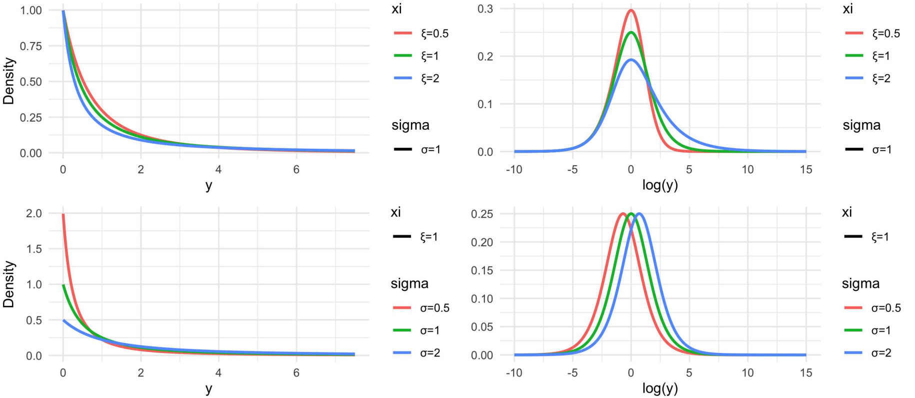


Figure: Generalized Pareto and its logarithm transformation for different values of the parameters $\xi = (0.5, 1, 2)$ and $\sigma = (0.5, 1, 2)$.

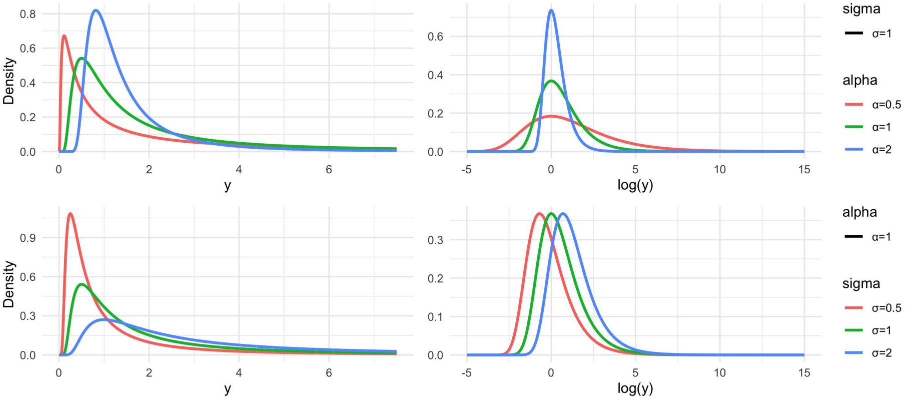
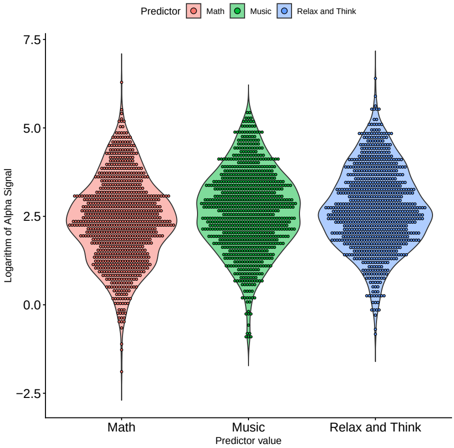


Figure: Fréchet density and its logarithm transformation for different values of the parameters $\alpha = (0.5, 1, 2)$ and $\sigma = (0.5, 1, 2)$.



1.

$$
\begin{aligned}
z_i = \ln(y_i) \mid x_i, \gamma_j, F_{Z|x\odot\gamma_j} &\sim F_{Z|x_i\odot\gamma_j}, \\
F_{Z|x_i\odot\gamma_j} &\sim \Pi_1^*(F_{Z|x\odot\gamma_j}).
\end{aligned}
$$

2.

$$
F_{Y|x} = \begin{cases} F_{Z|x}(\ln y), & \text{if } y > 0, \\ 0, & \text{if } y \leq 0. \end{cases}
$$

3. $F_{Z|x}$ has to be flexible!

We propose the following model

$$
\begin{aligned}
z_i \mid x_i, \gamma_j, s_i, &\overset{\text{ind}}{\sim} \mathsf{N}(x_i^\top \beta_{s_i}, \sigma^2_{s_i}), \\
(s_1, \ldots, s_n) &\sim \mathsf{CRP}(M), \\
M &\sim \mathsf{Gamma}(a_M, b_M), \\
(\beta_{s_i}, \sigma^2_{s_i}) &\overset{\text{iid}}{\sim} \mathsf{G}_{0|\gamma_j} \\
\gamma_j &\sim \Pi_2(\gamma_j),
\end{aligned}
\tag{2}
$$

where the density of $\mathsf{G}_{0|\gamma_j}$ is given by

$$
\mathsf{g}_{0|\gamma_j} = \prod_{l=1}^{p} \left[\gamma_{j,l}\mathsf{N}(\beta_l \mid 0, \tau\sigma^2) + (1-\gamma_{j,l})\delta_0(\beta_l)\right] \times \mathsf{IG}(\sigma^2 \mid a_\sigma, b_\sigma)
$$

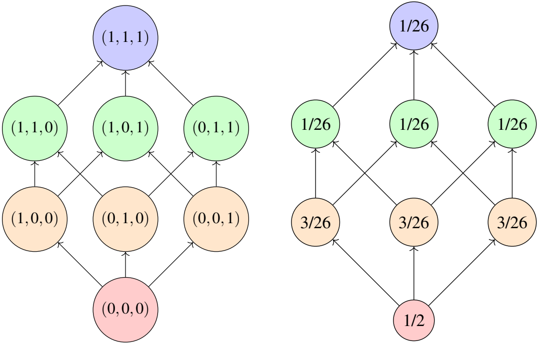


Figure: Space of models for $p = 3$ (left panel). Womack prior probability for $\rho = 1$ (right panel)

$$\pi_2(\gamma) = \rho \sum_{\gamma' : \gamma \prec \gamma'} \pi_2(\gamma'), \qquad (3)$$

where $\gamma \prec \gamma'$ if and only if $\gamma' \neq \gamma$ and $\gamma_i \leq \gamma'_j$ for all $1 \leq i, j \leq p$, $\rho \in (0, \infty)$ is a hyperparameter.

$$\Pi(\gamma = \gamma^* \mid \boldsymbol{z}, \boldsymbol{x}) = \frac{\mathcal{L}(\boldsymbol{z} \mid \boldsymbol{x}, \gamma^*)\Pi_2(\gamma^*)}{\sum_{\gamma' \in \mathcal{M}} \mathcal{L}(\boldsymbol{z} \mid \boldsymbol{x}, \gamma')\Pi_2(\gamma')} = \left[\sum_{\gamma' \in \mathcal{M}} B_{\gamma', \gamma^*} \frac{\Pi_2(\gamma')}{\Pi_2(\gamma^*)} \right]^{-1},$$

Theorem

Let Z be a random variable whose true density is $f^0_{Z|x^\star}$, satisfying conditions (v)-(viii)) of Theorem 5 and Corollary 1 in Barrientos, et. al., (2012). Let $x_1 = x \odot \gamma_1$ be the set of all relevant covariates, and let $x_2 = x \odot \gamma_2$ be any other subset of observed covariates. Let the space of density that can be represented as Dirichlet process mixtures of Gaussian kernels, i.e., $\mathcal{F}_{x,\gamma_j} = \{f : f(z) = \int \mathrm{N}(z \mid x^\top \beta, \sigma^2)\, G_{\gamma_j}(d\beta, d\sigma^2), G_{\gamma_j} \sim \mathrm{DP}(\alpha, G_{0|\gamma_j})\}$, where $g_{0|\gamma_j}$ is the defined base measure. Then $B_{\gamma_2,\gamma_1} \to 0$ when $n \to \infty$.

- With the proposed model is possible:
- Perform conditional density estimation, because density estimation is a local problem
    - The logarithm transformation facilitates the analytical work with the model
    - The model is useful for identifying relevant predictors and estimating the density for high but finite quantiles, that is, $q_p < \infty$.
- With the proposed model is not possible:
- Estimate tail index, because tail estimation is an asymptotic problem
    - Our model is not adequate to estimate the tail index, that is q_p, with $p \to 1$.

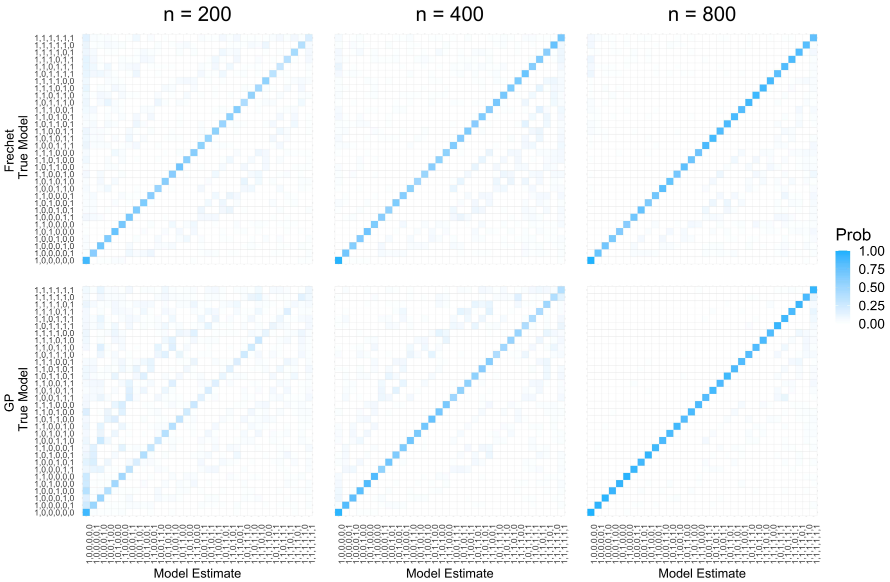


Figure: Monte Carlo average of the posterior probability over each model. Fréchet distribution (top panel), Generalized Pareto Distribution (bottom panel). For both distributions, predictors were indexed in the tail index parameter.

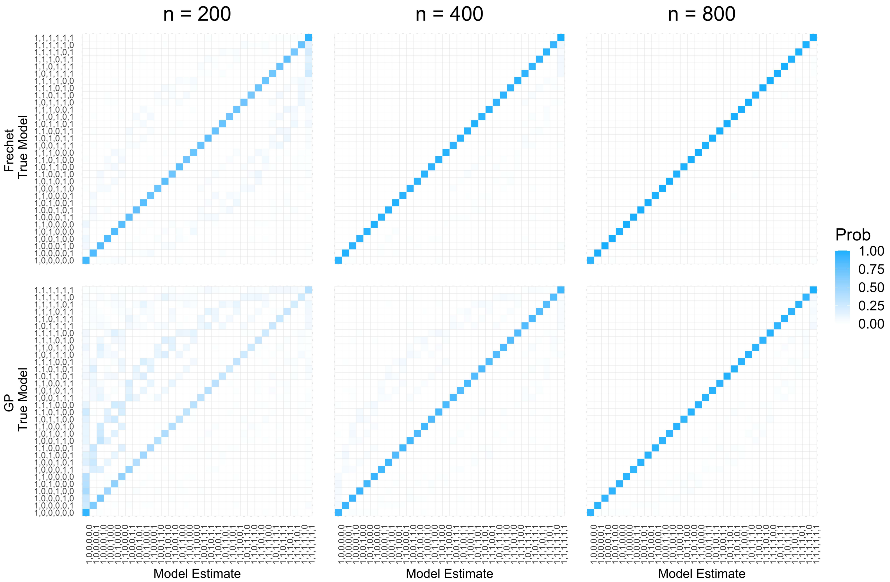


Figure: Monte Carlo average of the posterior probability over each model. Fréchet distribution (top panel), Generalized Pareto Distribution (bottom panel). For both distributions, predictors were indexed in the scale parameter.

$$f(y) = w_1 \cdot \text{Frechet}(y \mid \mu, \sigma, \alpha = e^{x^\top \beta_1/2}) + w_2 \cdot \text{Frechet}(y \mid \mu, \sigma, \alpha = e^{x^\top \beta_2/2})$$

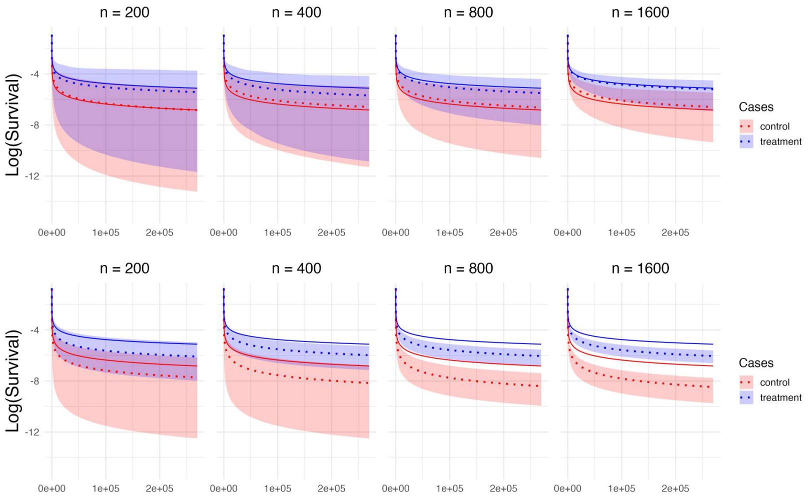


Figure: Tail behavior (quantiles 0.9–0.99): proposed model (top) vs. `VGAM` BIC Fréchet (bottom). Solid line: ground truth; dotted line: MC average. Shaded areas: 95% credibility/confidence bands.

- Electroencephalography (EEG) signals are markers of human brain activity
- EEG signals enable the identification of neural activation patterns linked to
  - Cognitive stimuli
  - Behavioral responses
  - Baselines states
- We focus on two key frequency bands:
  - Alpha band (8–13 Hz), associated with relaxed wakefulness and inhibitory cortical processing
  - Beta band (13–20 Hz), linked to active cognitive processing and sensorimotor activity

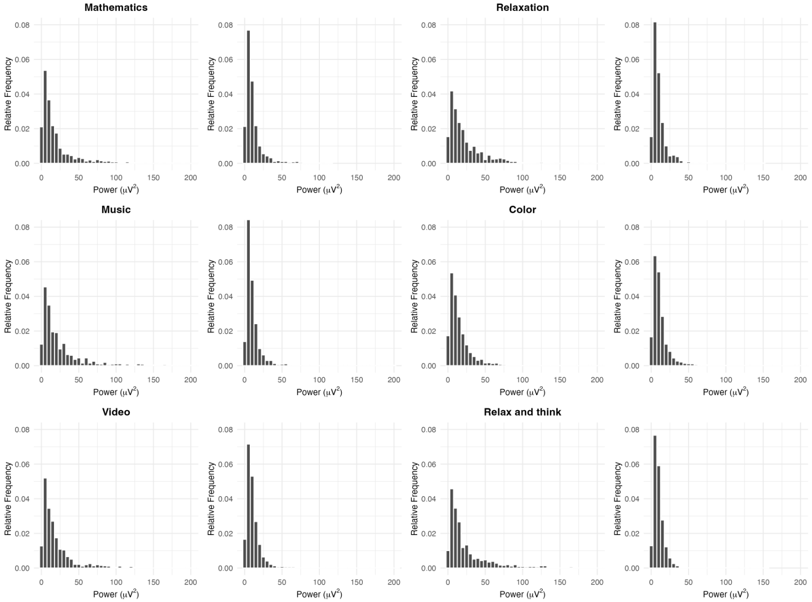


Figure: Histogram of potential alpha and beta wave for each stimulus.

- The most probable model in the posterior was:
  - $\gamma = (1, 1, 1, 1, 1, 1)$ with probability 0.560 for the Alpha band
  - $\gamma = (1, 1, 0, 0, 1, 0, 0)$ with probability 0.993 for the Beta band

Stimulus	Marginal Posterior	
	Alpha	Beta
Relaxation	0.999	0.000
Music	0.994	0.000
Color	1.000	1.000
Video	0.559	0.007
Relax and think	0.999	0.000

Table: Marginal posterior probabilities for each γ associated with the corresponding stimuli, relative to the control stimulus *Mathematics* estimated by the proposed model.

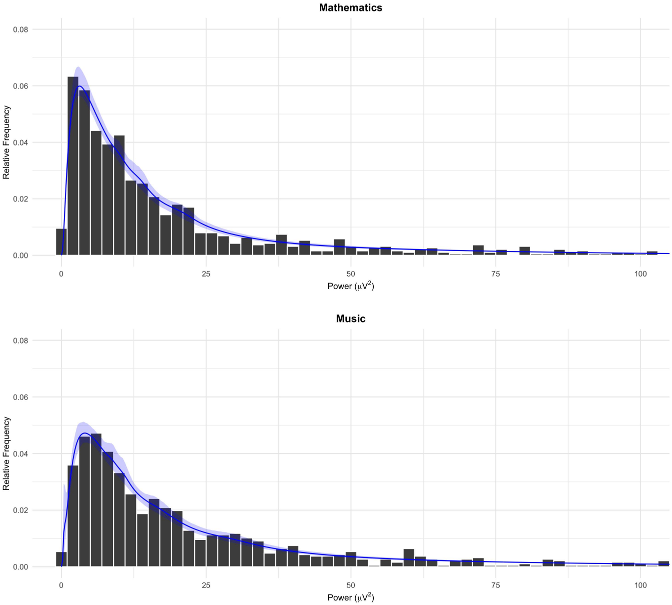


Figure: Density of spectral power for the alpha band

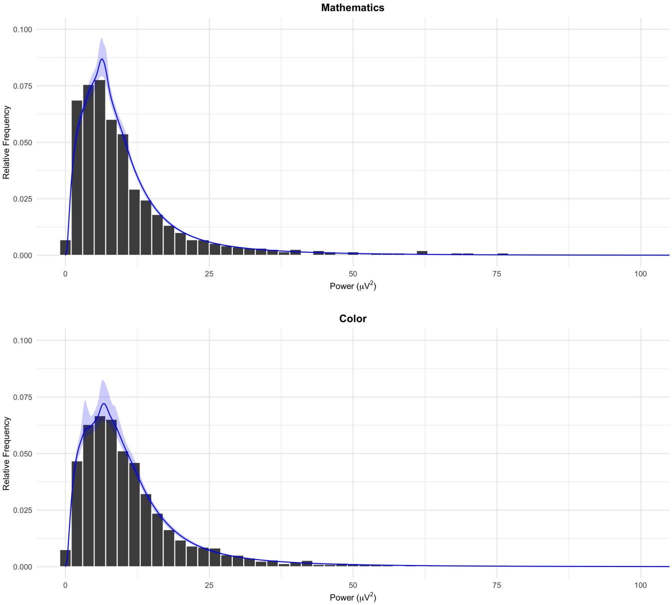


Figure: Density of spectral power for the beta band

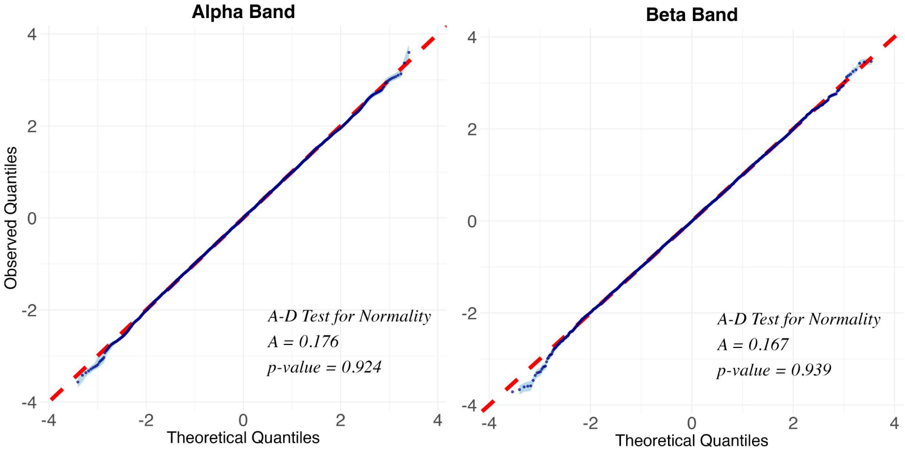


Figure: Q-Q plots of quantile residuals for the alpha (left) and beta (right) bands. In the Q-Q plots, the red line denotes theoretical normal quantiles, and the light blue shaded area represents the 95% credibility band.

Generative model

$$
\begin{aligned}
p_1 &\sim \text{Geom}(\phi_1) \\
p_2 &\sim \text{Geom}(\phi_2) \\
x_{i,j}^{(1)} &\sim \text{Bern}(\pi_j), \quad j = 1, \ldots, p_1 \\
x_{i,j}^{(2)} &\sim \text{N}(0,1), \quad j = 1, \ldots, p_2 \\
M &\sim \text{Gamma}(a_M, b_M), \\
s_1, \ldots, s_n &\sim \text{CRP}(M), \\
\beta_{s_i}, \sigma_{s_i}^2 &\sim \text{N}(\beta \mid 0, \tau\sigma^2 \boldsymbol{I})\text{IG}(\sigma^2 \mid a_\sigma, b_\sigma) \\
z_i &\overset{\text{ind}}{\sim} \text{N}(x_i^\top \beta_{s_i}, \sigma_{s_i}^2),
\end{aligned}
$$

Assuming that the predictors $x_i = (x_i^{(1)}, x_i^{(2)})$ are exogenous, and as the base measure is conjugate with the kernel, the posterior inference of the linear DDP reduces to:

$$p(\boldsymbol{s}, M \mid \boldsymbol{z}, \boldsymbol{x}) = p(\boldsymbol{s} \mid M, \boldsymbol{z}, \boldsymbol{x})p(M \mid \boldsymbol{z}, \boldsymbol{x})$$

- Marginal posterior

$$p(\boldsymbol{s}, M \mid \boldsymbol{z}, \boldsymbol{x}) = p(\boldsymbol{s} \mid M, \boldsymbol{z}, \boldsymbol{x})p(M \mid \boldsymbol{z}, \boldsymbol{x})$$

- Amortized posterior

$$q(\boldsymbol{s}, M \mid \boldsymbol{z}, \boldsymbol{x}) = q_D(\boldsymbol{s} \mid M, h(\boldsymbol{w}), S(\boldsymbol{w}))q_F(M \mid S(\boldsymbol{w}))$$

- Where
  - q_F denotes a normalized flow
  - q_D denotes an autoregressive decoder
  - $\boldsymbol{w} = \{x_i, y_i\}_{i=1}^n$
  - $h(\boldsymbol{w})$ and $S(\boldsymbol{w})$ are permutation-invariante summaries
- The empirical training loss is

$$\mathcal{L}(\psi) = \frac{1}{B}\sum_{b=1}^{B}\left[\log q_D(\boldsymbol{s}^{(b)} \mid M^{(b)}, h(\boldsymbol{w}^{(b)}), S(\boldsymbol{w}^{(b)})) + \log q_F(M^{(b)} \mid S(\boldsymbol{w}^{(b)}))\right]$$

  - Where $(\boldsymbol{w}^{(b)}, \boldsymbol{s}^{(b)}, M^{(b)})$ are fresh draws from the generative model at each gradient step

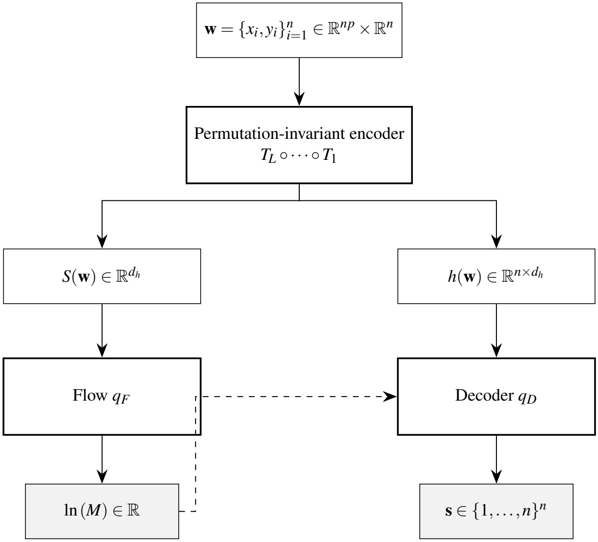


Figure: Overview of the amortized posterior architecture.

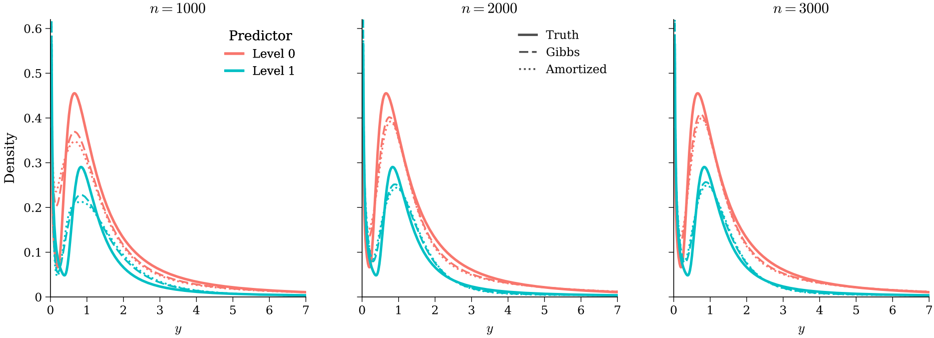


Figure: Amortized versus Gibbs estimation of a mixture of Frechet distributions.

References

- Haan, L., Ferreira, A.: Extreme Value Theory. Springer, New York, USA (2006)
- Gumbel, E.J.: Statistics of Extremes. Columbia University Press, New York, USA (1958)
- Pickands, J.: Statistical inference using extreme order statistics. The Annals of Statistics 3(1), 119–131 (1975)
- Bladt, M., Rojas-Nandayapa, L.: Fitting phase–type scale mixtures to heavy–tailed data and distributions. Extremes 21(2), 285–313 (2018)
- Palacios, V., Carvalho, M., Gutiérrez, L.: Heavy-Tailed NGG-Mixture Models. Bayesian Analysis 20(4),1315–1343 (2025)

- Doss, H., Sellke, T.: The tails of probabilities chosen from a Dirichlet prior. The Annals of Statistics 10(4)(1982)
- Karamata, J.: Sur un mode de croissance régulière. Théorèmes fondamentaux. Bulletin de La Société Mathématique de France 61, 55–62 (1933)
- Barrientos, A.F., Jara, A., Quintana, F.A.: On the support of Maceachern's dependent Dirichlet processes and extensions. Bayesian Analysis 7(2), 277–310 (2012)