Stick-Breaking Mixture Normalizing Flows with Component-wise Tail Adaptation for Variational Inference

Seungsu Han

November 5, 2025

Seoul National University
Department of Statistics

· Han, S., Hwang, J., & Chang, W. (2025).
Stick-Breaking Mixture Normalizing Flows with Component-Wise Tail Adaptation for Variational Inference. arXiv preprint arXiv:2510.07965. (under review)

Main Contributions and Outline

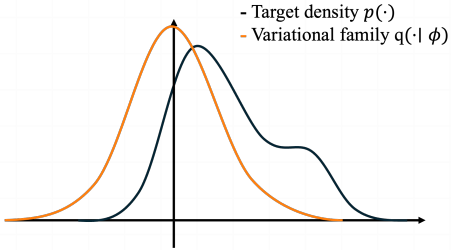
Main Contributions

- We propose **StiCTAF**, a flexible variational inference framework combining stick-breaking mixtures and tail-adaptive normalizing flows.
- It effectively models **multimodal** and **heavy-tailed** target distributions with improved flexibility and sample efficiency.

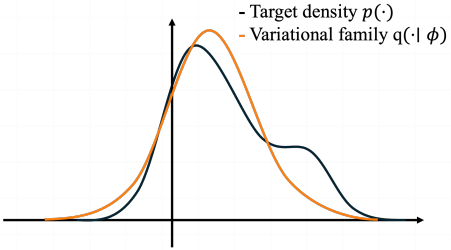
- Introduction
- Stick-breaking mixture model
- Heavy tail transformation
- Numerical experiments
- Summary

Introduction

Variational Inference



(a) Standard Gaussian $q(z) = \mathcal{N}(0, 1)$



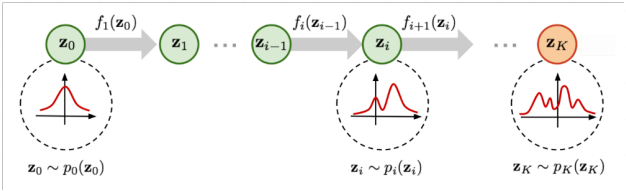
(b) Gaussian optimized by VI

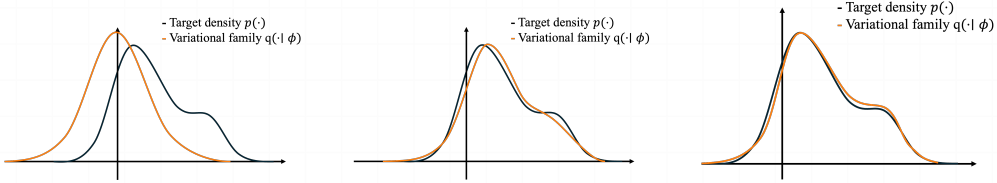
Intractable posterior $p(z \mid D) \propto p(D \mid z)p(z)$ given data D and prior $p(z)$.

- Goal: Approximate the **target** $p(z \mid D)$ with a tractable family $q_\phi(z)$. *Example)* Gaussian family $q_{(\mu,\sigma^2)}(z) = \mathcal{N}(z \mid \mu, \sigma^2)$
- VI enables fast optimization and sampling from the variational posterior, avoiding slow methods such as MCMC.

Normalizing Flows

Normalizing flows extend the variational family via an **invertible**, **differentiable** composition $T_\psi(z_0) = f_n \circ \cdots \circ f_1(z_0)$.





(a) iteration 0 **(b)** iteration 50 **(c)** iteration 100

Optimizing Objective Functions

Evidence of Lower Bound (ELBO): Variational Inference

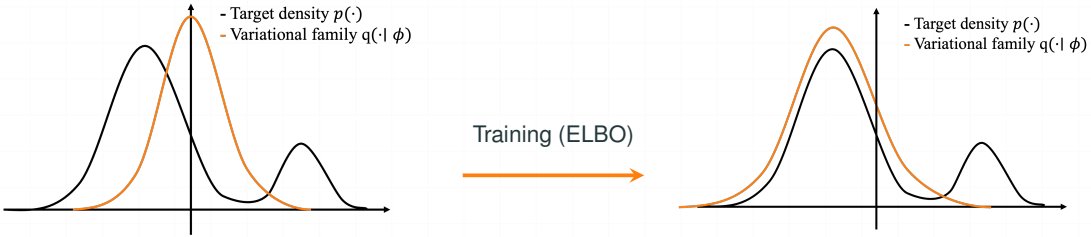
- $\mathrm{KL}(q_\theta(z) \,\|\, p(z \mid D)) = \log p(D) - \underbrace{\mathbb{E}_{z\sim q_\theta}[\log p(D, z) - \log q_\theta(z)]}_{\mathrm{ELBO}(\theta)}$.
- $\mathrm{ELBO}(\theta) \approx \frac{1}{n}\sum_{i=1}^{n}(\log p(D, z^{(i)}) - \log q_\theta(z^{(i)}))$, where $z^{(1)}, \ldots, z^{(n)} \sim q_\theta$.

- Maximize ELBO $\Leftrightarrow$ minimize $\mathrm{KL}(q_\theta \,\|\, p)$.
- We optimize $\mathrm{KL}(q_\theta \| p)$ because expectations under q_θ are tractable, while expectations under $p(z \mid D)$ are not.

Evidence of Lower Bound (ELBO): VI with Normalizing Flows

- For a bijection T_ψ, we minimize $\mathrm{KL}(T_\psi(q_\phi) \,\|\, p(z \mid D))$ using the change-of-variables formula $q_{(\psi,\phi)}(z) = q_\phi(T_\psi^{-1}(z))\left|\det J_{T_\psi^{-1}}(z)\right|$.
- $\mathrm{ELBO}(\psi, \phi) = \mathbb{E}_{z_0\sim q_\phi}\Big[\log p(D, T_\psi(z_0)) - \log q_\phi(z_0) + \log|\det J_{T_\psi}(z_0)|\Big]$

Limitations: Mode-Seeking Behavior

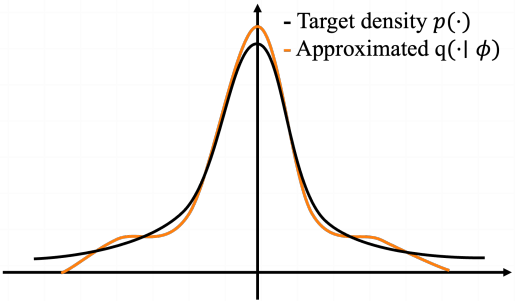


(a) Initial (pre-training)

(b) Converged (post-training)

- Reverse-KL is mode-seeking: it favors high-density regions and under-covers modes.
- We use a **stick-breaking mixture** as the flow's base distribution.

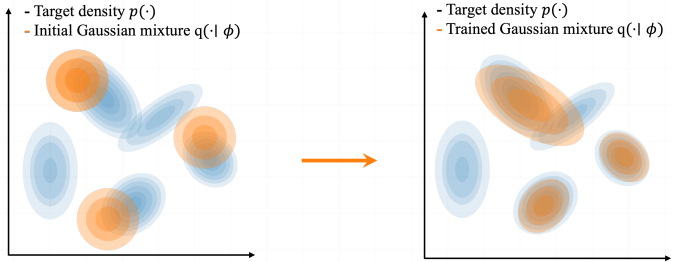
Limitations: Heavy-Tailed Distributions



- State-of-the-art normalizing-flow transformations struggle to produce heavy-tailed targets when starting from a light-tailed Gaussian base.
- Some flows use heavy-tailed bases (e.g., Student's t), but estimating tail thickness (degrees of freedom) is unstable and empirical performance is weak.

Stick-Breaking Mixture Model

Stick-Breaking Mixture: Motivation



(a) Initial

(b) Converged

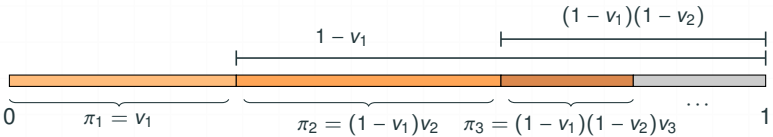
- **Mixture base:** $q_\phi(z) = \sum_{k=1}^{T} \pi_k\, q_k(z)$.
- **Issue with a fixed-T Gaussian mixture:**
  - We must choose T before training; the target's number of modes is rarely known a priori, and changing T mid-training is hard.

Idea: Use a stick-breaking mixture model (SBMM) so the **effective** number of components is learned automatically.

Stick-Breaking Mixture Base Distribution

Stick-breaking weights:

$$\pi_k \;=\; v_k \prod_{j<k}(1-v_j), \qquad v_k \sim \mathrm{Beta}(\alpha_k, \beta_k), \quad k = 1, 2, \ldots$$



Mixture base (e.g., Gaussian components):

$$q_\phi(z) \;=\; \sum_{k=1}^{T_{max}} \pi_k \, \mathcal{N}\big(z \mid \mu_k, \Sigma_k\big).$$

- With a large $T_{\max}$, stick-breaking weights **automatically determine** the effective number of components, while the mixture base **covers multiple modes**.

Heavy Tail Transformation

Previous Works

Definition 1 (Tail Index)

For $p, \alpha > 0$, define

- $\mathcal{E}_\alpha^p := \Big\{X : \ \Pr(|X| \geq x) = e^{-\alpha x^p} L(x), \quad \log L(x) = o(x^p) \ (x \to \infty)\Big\}$,
- $\mathcal{L}_\alpha^p := \Big\{X : \ \Pr(|X| \geq x) = \exp\{-\alpha(\log x)^p\} L(x), \quad \log L(x) = o\Big((\log x)^p\Big) \ (x \to \infty)\Big\}$,

where $L : \mathbb{R}_+ \to \mathbb{R}_+$ is slowly varying at infinity (i.e. $L(cx)/L(x) \to 1$ for every fixed $c > 0$). Specifically, for $X \in \mathcal{L}_\alpha^1$, we call α the **tail index** of X. For multivariate X and a unit vector u, the tail index of $[\langle u, X\rangle]_+$ is the **directional tail index** of X along u.

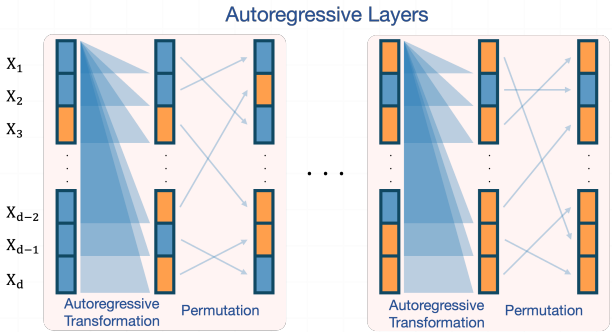
Theorem 2 (Liang at el., 2022)

Let $f : \mathbb{R} \to \mathbb{R}$ be a bi-Lipschitz bijective map (i.e. f and f^{-1} are globally Lipschitz). If $X \in \mathcal{E}_\alpha^p$, then $f(X) \in \mathcal{E}_{\tilde{\alpha}}^p$ for some $\tilde{\alpha} > 0$. Moreover, if $X \in \mathcal{L}_\alpha^p$ then $f(X) \in \mathcal{L}_\alpha^p$.

- Bi-Lipschitz normalizing flows map light-tailed to light-tailed and heavy-tailed to heavy-tailed.

Tail Dominance

Why not use a **heavy-tailed** base distribution?



- Suppose X_3 has the smallest tail index (heaviest tail). Downstream variables become dominated by X_3's tail.
- Permutation layers shuffle tail thickness across coordinates.

Tail Transformation Flow (TTF)

Idea. Start from a Gaussian-base NF and append a non-Lipschitz tail map $T_{\mathrm{TTF}} = (T_{\mathrm{TTF}}^{(1)}, \ldots, T_{\mathrm{TTF}}^{(d)}) : \mathbb{R}^d \to \mathbb{R}^d$ as the **final layer**. For each coordinate $i = 1, \ldots, d$,

$$T_{\mathrm{TTF}}^{(i)}(z_i;\, \xi_{+e_i}, \xi_{-e_i}) = \mu^{(i)} \pm \frac{\sigma^{(i)}}{\xi_{\pm e_i}} \left[\mathrm{erfc}\left(\frac{|z_i - \mu^{(i)}|}{\sigma^{(i)}\sqrt{2}} \right)^{-\xi_{\pm e_i}} - 1 \right],$$

where the sign $\pm$ and the index $\xi_{\pm e_i}$ follow the sign of $z_i - \mu^{(i)}$, and $\mathrm{erfc}(x) = 1 - \mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2}\, dt$.

Notes.

- *Effect.* Final tail map thickens a Gaussian base per coordinate/side via ξ_{+e_i}, ξ_{-e_i}.
- *Placement.* As the last layer, tail effects aren't mixed by permutations.
- *Challenge.* Reverse-KL is mode-seeking; ξ is hard to learn without good initialization.
- *Need.* Estimate directional (side-specific) tail indices to initialize/guide ξ.

Tail Estimation

Directional Tail Estimator

For each direction $\mathbf{u} \in \mathbb{S}^{d-1}$, draw i.i.d. $z_1, \ldots, z_n$ from a known heavy-tailed distribution (e.g., Student-t_ν with small ν), sort $z_{\mathbf{u},i} = z_i \mathbf{u}$ by magnitude $\|z_{\mathbf{u},(1)}\| \geq \cdots \geq \|z_{\mathbf{u},(n)}\|$, and for a top-$j$ subset define

$$\widehat{\xi}_{\mathbf{u}} = -\frac{1}{j} \sum_{i=1}^{j} \frac{\log p\big(z_{\mathbf{u},(i)} \mid D\big) - \log p\big(z_{\mathbf{u},(j+1)} \mid D\big)}{\log \|z_{\mathbf{u},(i)}\| - \log \|z_{\mathbf{u},(j+1)}\|} \; - \; 1.$$

Proposition 3. Consistency

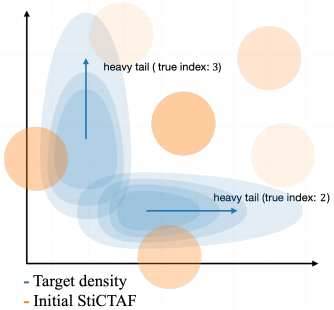
For fixed $u \in \mathbb{S}^{d-1}$, suppose the posterior $p(z \mid D)$ has a directional tail index $\xi_u \in (0, \infty)$ along u, and $p(z \mid D)$ decreases monotonically for all $z \geq z_0$. Then,

$$\widehat{\xi}_u \xrightarrow{P} \xi_u \quad (n \to \infty).$$

For light-tailed classes (e.g., Gaussian) the estimator diverges, while for heavier classes than $\mathcal{L}^1_\alpha$, it converges to 0.

StiCTAF Overview

Step 1 Base mixture training

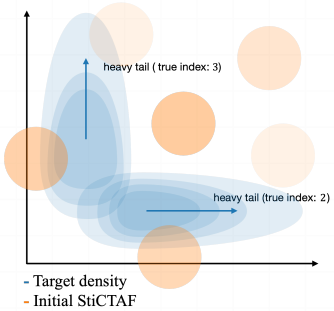


Tail index estimation

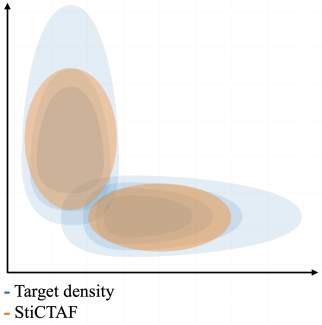
- Applying the same tail transform to all mixture components reduces flexibility and leads to undesirable results.
- Estimate component-wise tail indices and apply **distinct** tail transforms for each component.

StiCTAF Overview

Step 1 Base mixture training



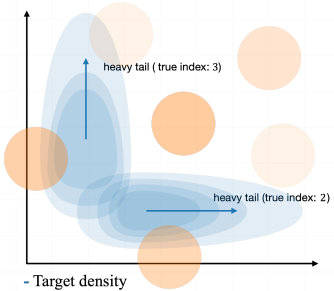
Step 2 Tail index estimation



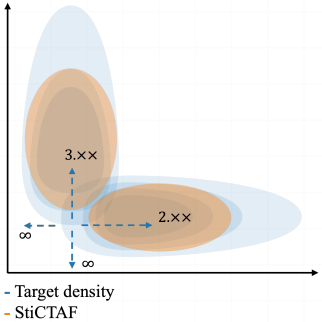
- Applying the same tail transform to all mixture components reduces flexibility and leads to undesirable results.
- Estimate component-wise tail indices and apply **distinct** tail transforms for each component.

StiCTAF Overview

Step 1 Base mixture training



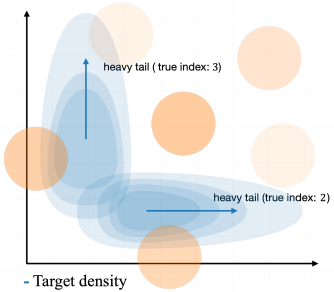
Step 2 Tail index estimation



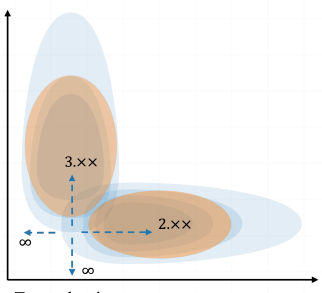
- Applying the same tail transform to all mixture components reduces flexibility and leads to undesirable results.
- Estimate component-wise tail indices and apply **distinct** tail transforms for each component.

StiCTAF Overview

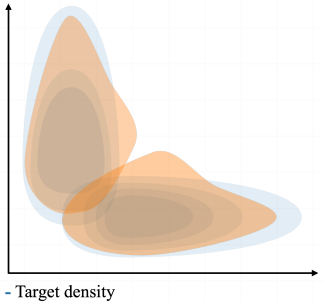
Step 1 Base mixture training



Step 2 Tail index estimation



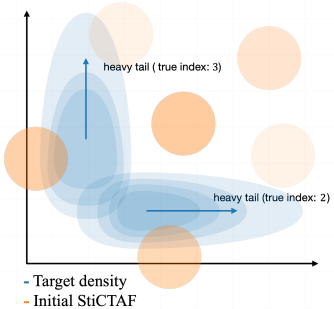
Step 3 Flow training



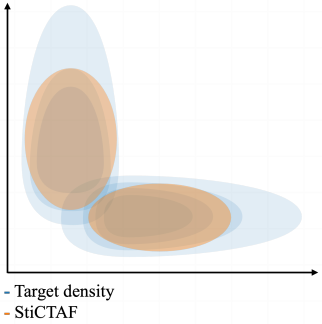
- Applying the same tail transform to all mixture components reduces flexibility and leads to undesirable results.
- Estimate component-wise tail indices and apply **distinct** tail transforms for each component.

StiCTAF Overview

Step 1 Base mixture training



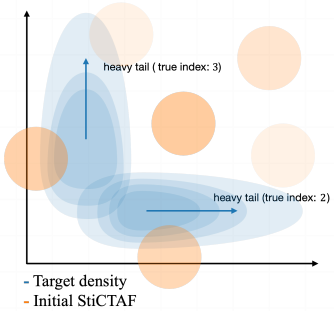
Step 2 Tail index estimation



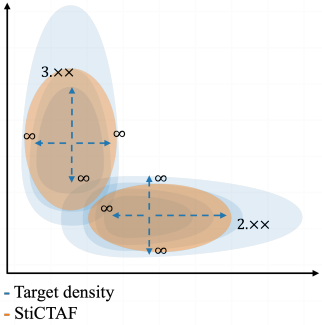
- Applying the same tail transform to all mixture components reduces flexibility and leads to undesirable results.
- Estimate component-wise tail indices and apply **distinct** tail transforms for each component.

StiCTAF Overview

Step 1 Base mixture training



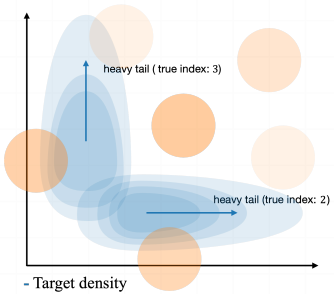
Step 2 Tail index estimation



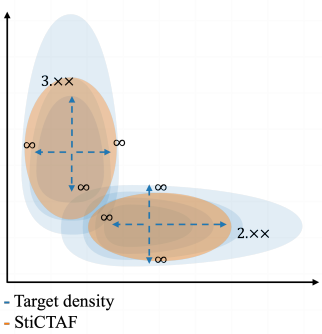
- Applying the same tail transform to all mixture components reduces flexibility and leads to undesirable results.
- Estimate component-wise tail indices and apply **distinct** tail transforms for each component.

StiCTAF Overview

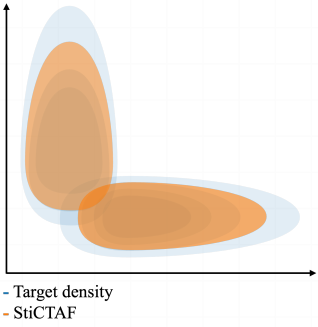
Step 1 Base mixture training



Step 2 Tail index estimation



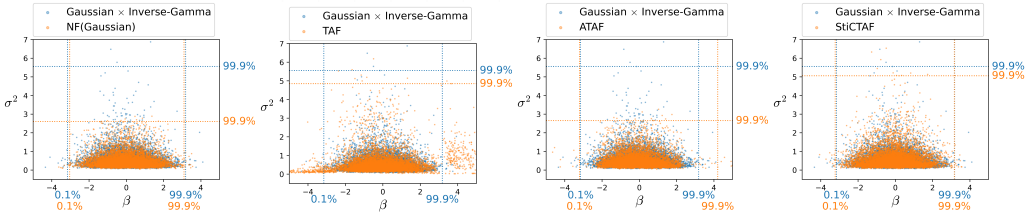
Step 3 Flow training



- Applying the same tail transform to all mixture components reduces flexibility and leads to undesirable results.
- Estimate component-wise tail indices and apply **distinct** tail transforms for each component.

Numerical Experiments

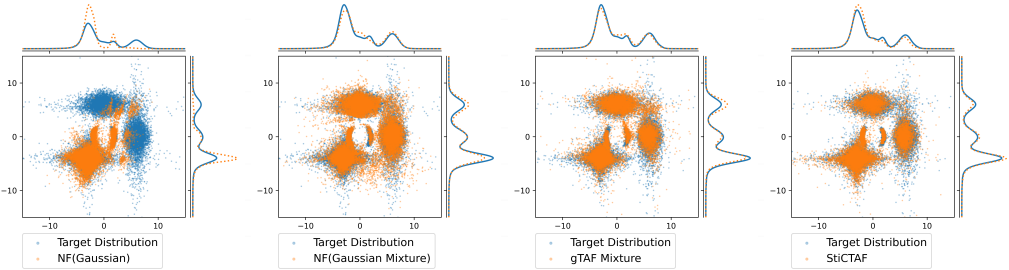
Normal–Inverse-Gamma Distribution



(a) NF(Gaussian) **(b)** TAF **(c)** ATAF **(d)** StiCTAF (Ours)

- Target: $\mathcal{N}(0,1) \times \text{Inv-Gamma}(3,1)$
- True tail index: $(\infty, 3)$, estimated tail index: $(\infty, 3.08)$

Complex Mixture Target Distribution



(a) NF(Gaussian) **(b)** NF (Gaussian mixture) **(c)** gTAF mixture **(d)** StiCTAF (Ours)

- Target: mixture of $t_{\nu=3} \times N(0,1)$, $N(0,1) \times t_{\nu=2}$, Two-moons, and $t_{\nu=2} \times t_{\nu=3}$.
- For samples $z_i \sim q_\theta$ and importance weights $w_i = \frac{p(z_i)}{q(z_i)}$, Effective Sample Size(ESS) is defined as, $\mathrm{ESS} = \frac{(\sum_i w_i)^2}{\sum_i w_i^2}$.

Table 1: (mean ± standard deviation)s over 10 random seeds.

Method	Forward KL	ESS(normalized)
NF (Gaussian)	1.92 ± 1.21	0.31 ± 0.17
NF (Gaussian Mixture)	0.33 ± 0.05	0.65 ± 0.23
gTAF Mixture	0.43 ± 0.12	0.48 ± 0.20
StiCTAF	**0.22** ± 0.09	**0.79** ± 0.19

Real Data Analysis: 2024 Daily Maximum Wind Speeds in Korea

- Use the *bivariate logistic extreme value framework* of **Fawcett & Walshaw (2006)**.
- $X_{(j,s),t}$: the daily maximum wind speed at station $j \in \{1, \ldots, 4\}$, season $s \in \{1, \ldots, 4\}$, and day t.
- Set the joint CDF for a consecutive pair of exceedance as

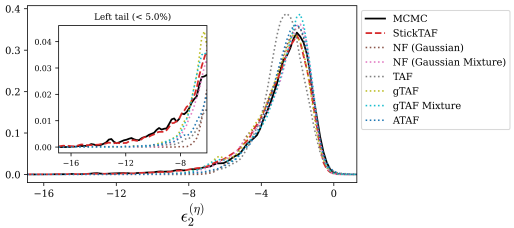
$$F(x_t, x_{t+1} \mid \sigma_{(j,s)}, \eta_{(j,s)}, \alpha_j) = 1 - \left[Z(x_t \mid \sigma_{(j,s)}, \eta_{(j,s)})^{-1/\alpha_j} + Z(x_{t+1} \mid \sigma_{(j,s)}, \eta_{(j,s)})^{-1/\alpha_j} \right]_+^{\alpha_j},$$

  where $Z(x \mid \sigma, \eta) = \Lambda^{-1}\left(1 + \frac{\eta(x-u)_+}{\sigma}\right)^{1/\eta}$ for fixed threshold u, and exceedance rate Λ.
- We decompose station and season effects using the additive models

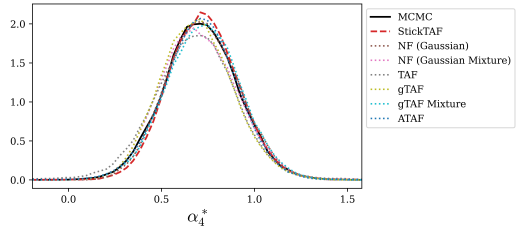
$$\sigma_{j,s} = \text{softplus}(\gamma_j^{(\sigma)}) + \text{softplus}(\varepsilon_s^{(\sigma)}), \quad \eta_{j,s} = \text{softplus}(\gamma_j^{(\eta)}) + \text{softplus}(\varepsilon_s^{(\eta)}), \quad \alpha_j = \text{sigmoid}(\alpha_j^*)$$

- We get 20-dimensional parameter space, $\gamma_{1:4}^{(\sigma)}$, $\varepsilon_{1:4}^{(\sigma)}$, $\gamma_{1:4}^{(\eta)}$, $\varepsilon_{1:4}^{(\eta)}$, and $\alpha_{1:4}^*$.

Real Data Analysis: 2024 Daily Maximum Wind Speeds in Korea (2)



(a) Posterior of $\varepsilon_2^{(\eta)}$



(b) Posterior of α_4^*

Table 2: Estimated mode and the 99% equal-tail credible interval.

Parameter	MCMC	StiCTAF	NF (Gaussian)	TAF
$\varepsilon_1^{(\eta)}$	-1.69 (-11.21, -0.32)	-1.81 (-11.89, -0.27)	-1.72 (-5.89, -0.31)	-1.70 (-4.71, 0.05)
$\varepsilon_2^{(\eta)}$	-2.02 (-11.95, -0.38)	-2.06 (-12.09, -0.51)	-1.95 (-6.29, -0.51)	-2.60 (-7.12, 1.89)
$\varepsilon_3^{(\eta)}$	-1.52 (-9.18, -0.13)	-1.62 (-10.21, -0.26)	-1.50 (-4.71, -0.22)	-1.65 (-6.05, 1.02)
$\varepsilon_4^{(\eta)}$	-2.09 (-11.64, -0.50)	-2.22 (-13.88, -0.71)	-2.14 (-5.98, -0.52)	-2.32 (-5.32, 0.59)
Comp. time (hr)	11.90	0.08	0.03	0.03

Summary

Summary

- **StiCTAF** integrates a stick-breaking mixture base with a tail transformation flow for **component-wise pushforward**, enabling flexible, tail-aware variational inference.
- It accurately models **multimodal** and **heavy-tailed** posteriors while remaining computationally efficient.

→ Flexible VI for complex target distributions.

Reference

- ..

Appendix A.