IMPERIAL

Scale-Preserving Nonlinear Dimension Reduction for Heavy-Tailed Data

Homogeneous Autoencoders

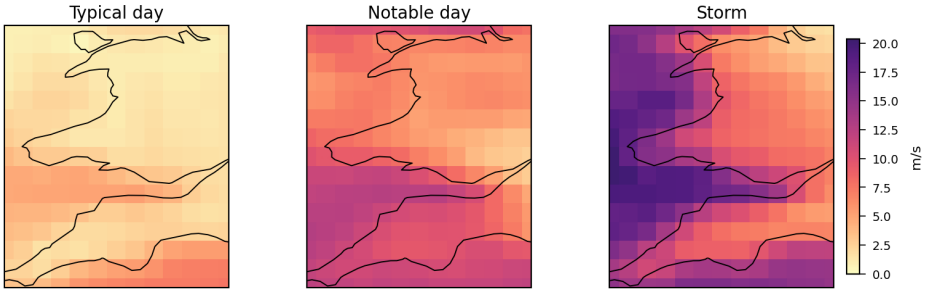
Ashley Turner
Supervised by Nicola Gnecco and Almut Veraart
Department of Mathematics, Imperial College London

GAME Bologna, June 2026

High-dimensional, heavy-tailed data

High-dimensional, heavy-tailed data are common

Wind speeds over SW England & Wales



Heavy-tailed and high-dimensional settings are commonly found in many areas.
E.g., climate, finance, insurance and astronomy

ERA5: [Hersbach et al., 2020]

Modelling directly in $\mathbb{R}^D$ brings significant challenges

A single observation typically lives in hundreds to thousands of dimensions (or more!).

- **Computation.** Sampling, fitting, evaluating likelihoods all scale with D.
- **Curse of dimensionality.** Statistical estimators degrade exponentially in D.
- **Measure collapse.** High-dimensional ambient measures concentrate and may become singular: density modelling becomes degenerate.

This creates the need for a lower-dimensional representation.

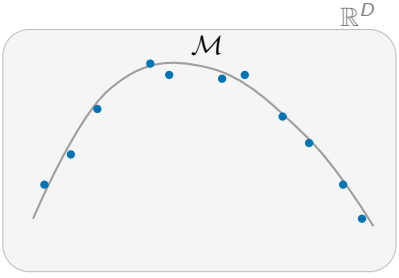
Real data have low-dimensional structure

Manifold hypothesis:
high-dimensional observations often concentrate on a smooth low-dimensional manifold.

$$X \in \mathcal{M} \subset \mathbb{R}^D, \qquad \dim \mathcal{M} = m \ll D.$$

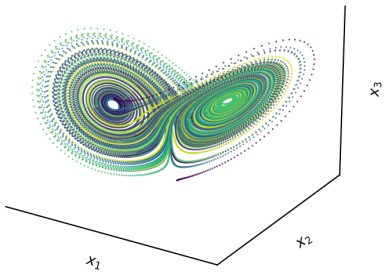
We would ideally like to work *intrinsically*, not in the ambient coordinates.

[Cayton, 2008; Whiteley et al., 2025]



Data concentrate *near* or on the manifold (some noise allowed).

Examples of low-dimensional structure



Lorenz attractor

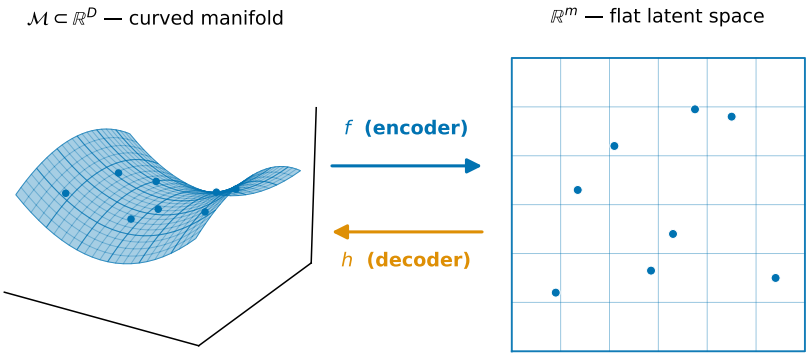
Cat manifold[1]

[1] not a technical term.

One approach: autoencoders

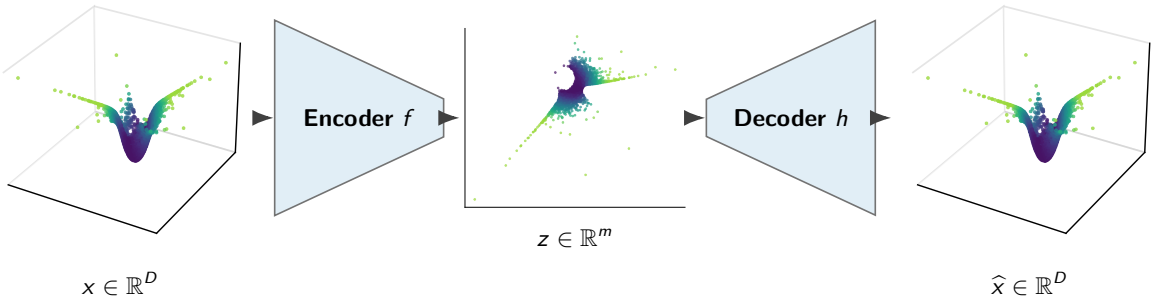
(. . . and related models)

Latent representation of the manifold



A good latent representation represents on-manifold points via *intrinsic* coordinates, so the manifold geometry parameterises as a flat $\mathbb{R}^m$ latent space.

Autoencoders: encoder, decoder, reconstruction loss



Encoder $f : \mathbb{R}^D \to \mathbb{R}^m$, decoder $h : \mathbb{R}^m \to \mathbb{R}^D$.

Parameterised as a neural network and trained to minimise reconstruction error $\|X - h(f(X))\|^2$.

Example: heavy-tailed curved surface, D=3, m=2 with Student-t radii (ν=2.8).

[Kingma & Welling, 2014]

But the values we care about lie in the tail

Climate risk, market crashes, insurance losses...

Rare events drive practical interest,
not the bulk of observations.

In order to remain useful in this context,
a good latent representation should preserve a notion of
extremeness relative to the ambient space.

Background on multivariate extremes

Univariate regular variation (heavy tails)

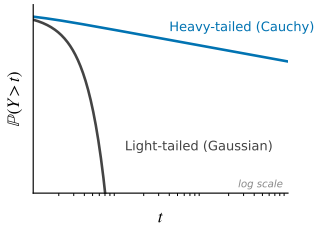
Definition (Univariate regular variation — heavy tails)

A random variable $Y \in \mathbb{R}$ has a heavy tail if extremes are rare but not vanishingly so:

$$\mathbb{P}(Y > t) \;\sim\; L(t)\, t^{-\alpha} \quad \text{as } t \to \infty,$$

with L slowly varying[1] and $\alpha > 0$ the tail index.

- Smaller $\alpha \iff$ heavier tail.
- $\alpha \leq 2$: divergent variance (e.g. Student-t with $\nu = 2$).
- $\alpha \leq 1$: divergent mean (e.g. Cauchy).
- Examples: financial returns, wind speeds.



[Resnick, 2007; de Haan & Ferreira, 2006]

[1] $L(tx)/L(t) \to 1$ as $t \to \infty$ for every $x > 0$; e.g. $L(t) = \log t$.

Multivariate regular variation

Definition (Multivariate regular variation)

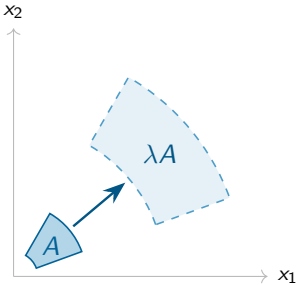
$X \in \mathbb{R}^D$ is regularly varying if there exist a scaling $b(t) \to \infty$ and a non-zero tail measure μ on $\mathbb{R}^D \setminus \{0\}$ with

$$t\,\mathbb{P}\Big(\frac{X}{b(t)} \in \cdot\Big) \xrightarrow[t\to\infty]{} \mu(\cdot).$$

Two facts about μ:

- α-homogeneous: $\mu(\lambda A) = \lambda^{-\alpha}\,\mu(A)$.
- Examining the radius $\mathbb{P}(\|X\| > t) \sim L(t)\,t^{-\alpha}$ recovers the univariate case.

α represents the heaviness of the (joint) tail.
μ controls which directions are extreme.



$\mu(\lambda A) = \lambda^{-\alpha}\mu(A)$

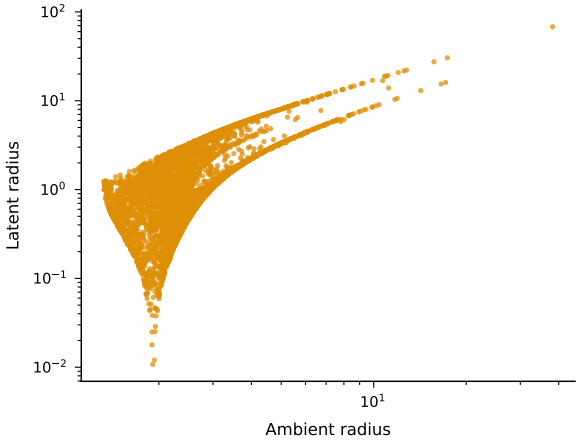
[Resnick, 2007; Lindskog et al., 2014]

The problem

Standard autoencoding methods do not preserve tails

Recall that large $\|X\|$ are the rare events that drive practical interest. When we train a standard unconstrained MLP AE on heavy-tailed data and examine the latent space...

- large $\|X\|$ aren't reliably large $\|Z\|$
- the latent tail index α_{lat} drifts from α_{amb}
- yet reconstruction error still looks fine



Latent vs. ambient radius of a standard AE applied to synthetic heavy-tailed data.

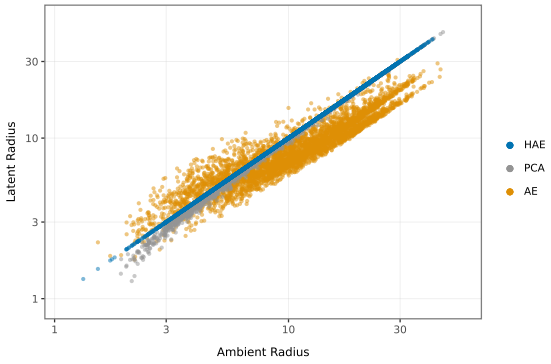
[Jaini et al., 2020]

Homogeneous Autoencoder

The core principle

Decompose each ambient point in polar form $x = r(x)\,\theta$.

- Build an encoder that transforms only the angle θ, leaving the radius intact.
- Radial tail properties (magnitude, scaling, regular variation…) then transport into latent space by construction.
- This is the core principle behind the Homogeneous Autoencoder (HAE).



HAE demonstrated on ERA5 10-m wind.
(D=195 grid cells, m=8)

Encoding

Define an encoder $f : \mathbb{R}^D \setminus \{0\} \to \mathbb{R}^m \setminus \{0\}$.

Construction (p-homogeneous encoder)

$$f(x) = \underbrace{r(x)^p}_{\text{radial}} \underbrace{e(\theta)}_{\text{angular}}$$

where $e : \mathbb{S}^{D-1} \to \mathbb{S}^{m-1}$ is a unit-sphere map parameterised by a neural network.

Exactly p-homogeneous by construction: $f(\lambda x) = \lambda^p f(x)$.

Ambient radial properties are transported analytically into latent space.

Decoding

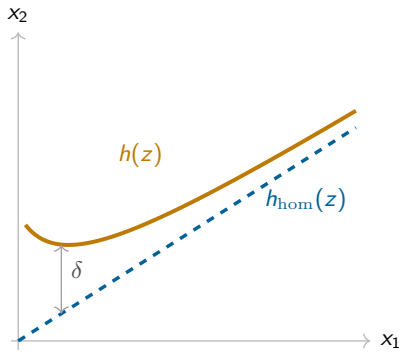
Latent polar form $z = r(z)\,\eta$ on $\mathbb{R}^m \setminus \{0\}$.

Construction (asym. $1/p$-homogeneous decoder)

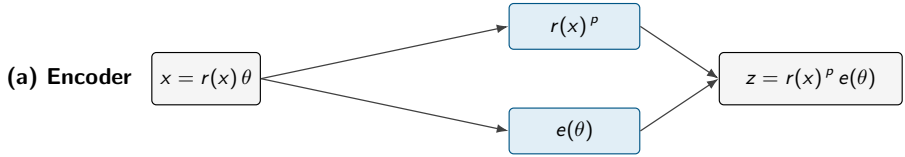
$$h(z) \;=\; \underbrace{r(z)^{1/p}\,\tilde{e}(\eta)}_{h_{\mathrm{hom}}(z)} \;+\; \underbrace{\delta(\eta, r(z))}_{\text{finite-scale correction}}$$

where:

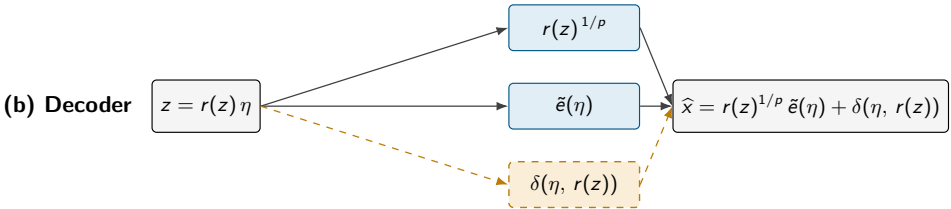
- $\tilde{e} : \mathbb{S}^{m-1} \to \mathbb{S}^{D-1}$.
- $\delta : \mathbb{S}^{m-1} \times (0,\infty) \to \mathbb{R}^D$.
- and importantly, $\delta \to 0$ as $r \to \infty$.



Architecture summary



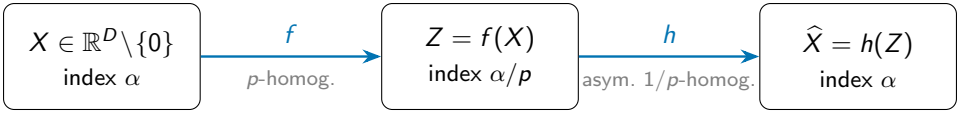
$\theta{=}x/r(x) \;\Rightarrow\; f(\lambda x) = \lambda^{p} f(x)$ *exactly*



asymptotically homogeneous: $\sup_{\eta} \|\delta(\eta, r(z))\| \to 0$ *as* $r(z) \to \infty$

Theory

Theoretical result

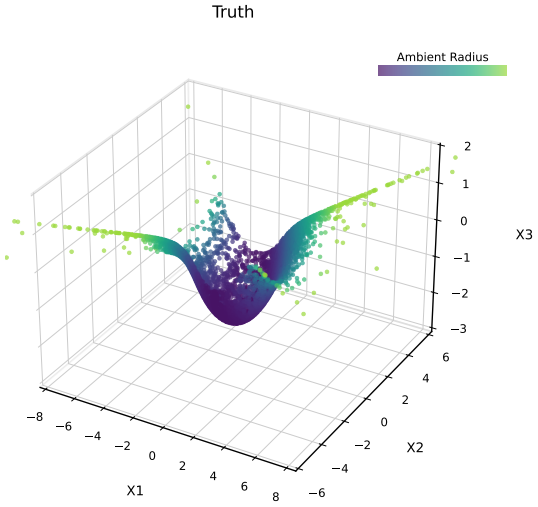


The HAE round-trip preserves the ambient tail index.

Full statements in appendix.

In practice: synthetic data

Synthetic heavy-tailed data on a curved manifold

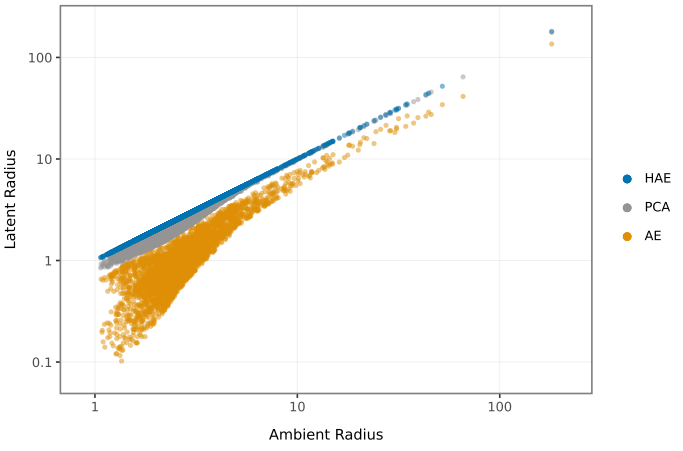


$D = 3$, $m = 2$ curved manifold with Student-t radii ($\nu = 2.8$), so ambient tail index $\alpha = \nu = 2.8$.

A heavy-tailed sample on a smooth low-dimensional manifold.

Experiments use a similar construction extended to $D = 10$, $m = 3$ (details in appendix).

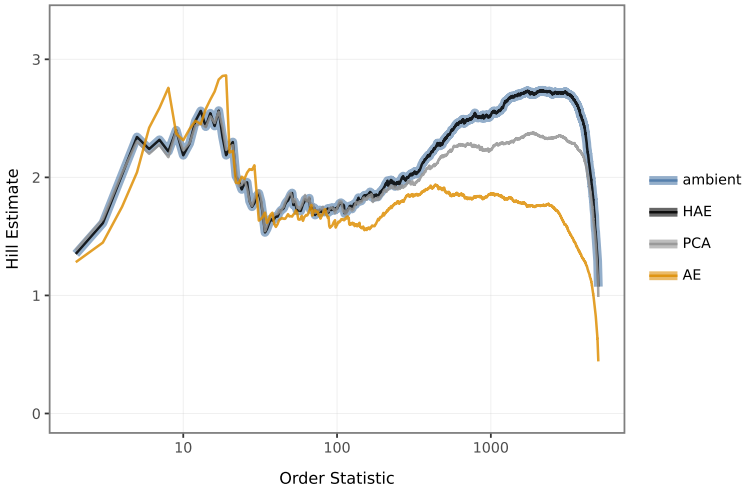
HAE preserves the ambient radius in the latent space



$D = 10$, $m = 3$.
Student-t radii with $\alpha = 1.8$.

PCA included as a linear reference. The Homogeneous Autoencoder (HAE) and PCA both trace clean radial scaling. The AE drifts.

HAE preserves the latent tail index



The latent distribution encoded via the HAE again inherits the same tail behaviour as the ambient distribution. The standard AE drifts.

Reported $\widehat{\alpha}$ uses the top 10% of order statistics, $k = \lfloor 0.1\, n \rfloor$; $n = 5000$, $k = 500$ here.

Reconstructive performance is approximately matched

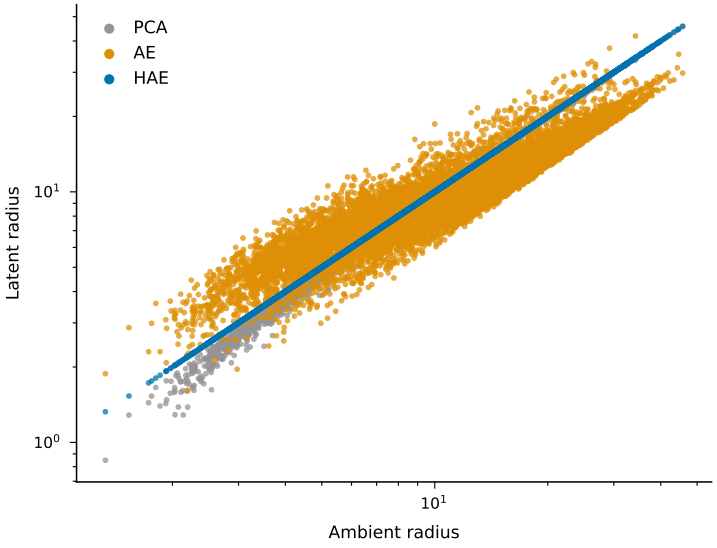
	Average per-sample MSE
HAE	3.9×10^{-2}
AE	1.2×10^{-2}
PCA	2.7×10^{-1}

HAE matches AE on the flexible synthetic dataset. PCA much worse.

D=10, m=3, α=1.8.

In practice: real atmospheric data (ERA5)

ERA5 atmospheric fields (u_{10} wind)



ERA5 10-m wind, $D = 195$ grid cells over SW England and Wales, $m = 8$.

The Homogeneous Autoencoder (HAE) traces the diagonal.

The unconstrained AE's latent radii drift into a cloud.

[Hersbach et al., 2020]

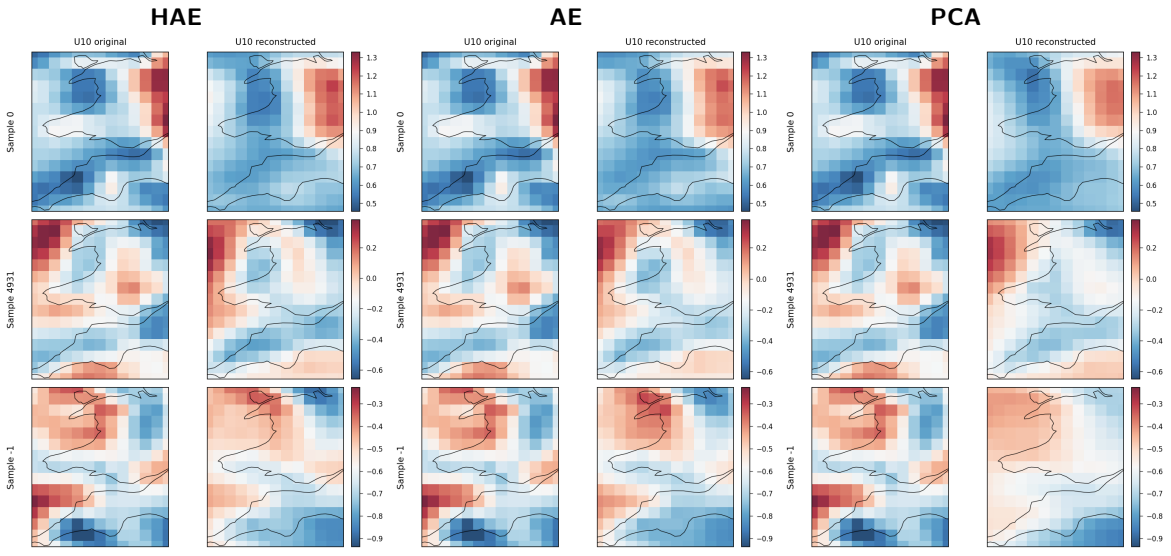
Reconstruction quality is matched on ERA5

	Average per-sample MSE
HAE	9.5×10^{-3}
AE	9.1×10^{-3}
PCA	1.4×10^{-2}

HAE and AE are practically tied. PCA is an order of magnitude worse.

ERA5 u_{10} wind, D=195 grid cells, m=8.

ERA5 reconstructions, original — reconstructed



Closing

Related work

- **Generative modelling of extremes.** VAEs, GANs, heavy-tailed normalising flows. Sampling generally done in ambient space.

  [Lafon et al., 2023; Allouche et al., 2026; Monte et al., 2025; Jaini et al., 2020; Łaszkiewicz et al., 2021, 2022; Wiese et al., 2019]

- **Dimension reduction / structure for extremes.** Latent linear factor models, graphical / sparse models, clustering.

  [Boulin & Bücher, 2026; Boulin et al., 2025; Engelke & Hitz, 2020; Engelke & Ivanovs, 2021; Huser & Wadsworth, 2024]

- **Nonlinear DR / generative modelling.** VAEs, flows, diffusion, t-SNE, UMAP, Isomap, manifold-aware flows (no tail focus).

  [Kingma & Welling, 2014; Rezende & Mohamed, 2015; Ho et al., 2020; van der Maaten & Hinton, 2008; McInnes et al., 2018; Tenenbaum et al., 2000; Brehmer & Cranmer, 2020; Caterini et al., 2021; Loaiza-Ganem et al., 2024; Kalatzis et al., 2022; Horvat & Pfister, 2022]

- **Multivariate EVT.** Regular variation, angular measures, mapping theorems.

  [Resnick, 2007; de Haan & Ferreira, 2006; Lindskog et al., 2014; Basrak & Segers, 2009; Basrak et al., 2025]

- **Homogeneous / equivariant networks.** Spherical / rotation-equivariant CNNs.

  [Cohen et al., 2018; Worrall et al., 2017; Weiler & Cesa, 2019; Lyu & Li, 2020]

Conclusion

Scale-preserving nonlinear dimension reduction for heavy-tailed data.

- Tail preservation is enforced architecturally.
- Reconstruction performance practically matches a parameter-matched unconstrained AE.

Thank you

- **Paper:** *Scale-Preserving Autoencoders for High-Dimensional Heavy-Tailed Data*
  arXiv preprint and code coming soon!

Thank you. Questions?

Appendix: future directions

- **Generative modelling in latent space.** Parametric EVT / diffusion / flow-based / conditional models to Z, then decode back with tail-index consistency ensured by Cor. 1.
- **Extrapolative performance.** Pre-transform the data to better align with a conic support so the homogeneous decoder has the right asymptotic target geometry beyond the training range.
- **Learned homogeneity degree p.** Currently fixed beforehand, could learn p jointly via a term in the loss for more flexible modelling.
- **Full tail-measure recovery.** Going beyond tail-index preservation to the round-trip angular-measure pushforward, $(h_{\mathrm{hom}} \circ f)_{\#}\mu_X = \mu_X$, e.g., via pseudo-invertible architectures.
- **Tail-aware reconstruction loss.** Weight the reconstruction term by an increasing function of the ambient radius for better sampling in the tail.
- **Any other ideas are welcome! Can you think of more?**

Appendix: why not just penalise the tail in the loss?

An obvious attempt at the same problem: add a term such as $\|\widehat{\alpha}_{\mathsf{lat}} - \widehat{\alpha}_{\mathsf{amb}}\|$ to the training objective.

- Hill estimators (and other tail-informed loss functions) have significant finite-sample bias.
- The penalty is batch-dependent: the tail isn't visible in most mini-batches!
- Training-time guarantees do not transfer to deployment where out-of-sample extremes can still misbehave.
- Tradeoffs are difficult to balance as compound loss functions get more complex: bulk reconstruction vs tail consistency.
- More broadly, modern ML methods typically need *lots* of data to perform well. This juxtaposes the fact that, by definition, we have few samples in the tail of our data distribution!

We want the guarantee baked into the architecture, not just incentivised by the loss.

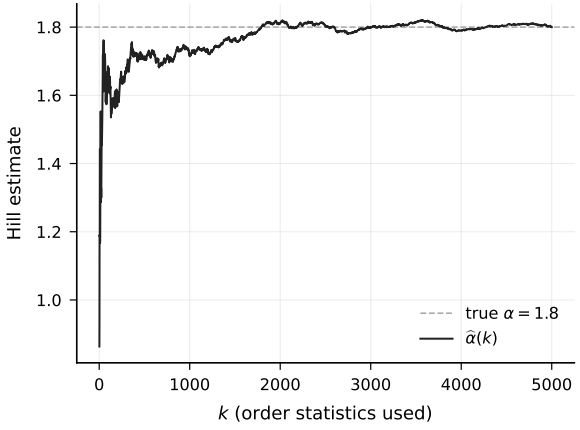
Appendix: Hill estimator

Definition (Hill estimator)

Given the top k order statistics $Y_{(1)} \geq \cdots \geq Y_{(k)}$ of a sample of size n,

$$\widehat{\alpha}(k) = \frac{1}{\widehat{H}_{k,n}}, \quad \widehat{H}_{k,n} = \frac{1}{k}\sum_{i=1}^{k} \log \frac{Y_{(i)}}{Y_{(k+1)}}.$$

Choosing k isn't always straightforward but is not the focus of this work.
Broadly: look for a *plateau* as k varies



$\widehat{\alpha}$ stabilises around the true value;
Pareto($\alpha = 1.8$) sample, $n = 10\,000$.

[Hill, 1975]

Appendix: ReLU networks

ReLU networks *are* asymptotically homogeneous: $f(x) = \mathrm{ReLU}(Wx + b)$ has asymptotic homogeneous limit $f_\infty(\theta) = (W\theta)_+$. But this isn't enough for tail-index preservation.

$f(X)$ remains regularly varying only when f_∞ is non-vanishing on $\mathrm{supp}(\mu)$. Trained weights are not constrained to keep this so. If f_∞ vanishes μ-a.e. on $\mathrm{supp}(\mu)$, the heavy tail of X is mapped asymptotically to 0 and $f(X)$ ceases to be regularly varying altogether: the latent tail index is destroyed.

Example. Let $X \sim \mathrm{Pareto}(\alpha)$ on $(0, \infty)$, so $\mathrm{supp}(\mu) = \{+1\}$. The negative-weight ReLU encoder $f(x) = \mathrm{ReLU}(-wx + b)$ with $w > 0$, $b \geq 0$ has $f_\infty(\theta) = (-w\theta)_+$ vanishing at $\theta = +1$. For $x > b/w$, $-wx + b < 0$, so $f(x) = 0$ and the entire heavy tail of X is mapped to 0. The latent $f(X)$ is bounded above by $\max(0, b)$, hence not regularly varying.

Appendix: encoder tail transport

Proposition (encoder tail transport)

Let X be regularly varying on $\mathbb{R}^D \setminus \{0\}$ with scaling b, tail index α, tail measure μ and let $f : \mathbb{R}^D \setminus \{0\} \to \mathbb{R}^m \setminus \{0\}$ be positively p-homogeneous, continuous μ-a.e., and with $f^{-1}(A)$ bounded away from 0 whenever A is.

Then $f(X)$ is regularly varying on $\mathbb{R}^m \setminus \{0\}$ with scaling b^p, tail index α/p and tail measure $\mu_f = f_{\#}\mu$.

Tail (angular) measure transports as $f_{\#}\mu$.

Essentially a special case of the mapping theorem: [Lindskog et al., 2014, Thm. 2.3]

Appendix: decoder tail transport

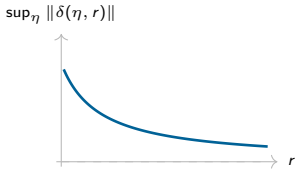
Theorem (decoder tail transport)

Let Z be regularly varying on $\mathbb{R}^m \setminus \{0\}$ with scaling b, tail index β, tail measure μ_Z and let $h : \mathbb{R}^m \setminus \{0\} \to \mathbb{R}^D \setminus \{0\}$ decompose as

$$h(z) = h_{\mathrm{hom}}(z) + \delta(\eta, r(z)),$$

where h_{hom} is positively $1/p$-homogeneous (constructed analogously to f) and $\sup_\eta \|\delta(\eta, r)\| \to 0$ as $r \to \infty$.

Then $h(Z)$ is regularly varying on $\mathbb{R}^D \setminus \{0\}$ with scaling $b^{1/p}$, tail index $p\beta$ and tail measure $\mu_h = (h_{\mathrm{hom}})_{\#}\mu_Z$.



The correction δ does not affect regular variation, vanishing in the tail by construction.

Appendix: round-trip tail-index preservation

Corollary (Round-trip tail-index preservation)

Let X be regularly varying on $\mathbb{R}^D \setminus \{0\}$ with scaling b, tail index α, tail measure μ_X. Let f be a p-homogeneous encoder and h an asymptotically homogeneous decoder of exponent $1/p$. Then $h(f(X))$ is regularly varying on $\mathbb{R}^D \setminus \{0\}$ with scaling b, tail index α and tail measure $(h_{\mathrm{hom}} \circ f)_{\#}\mu_X$.

Appendix: architecture (radial-angular factorisation)

Any positively p-homogeneous $f : \mathbb{R}^D \setminus \{0\} \to \mathbb{R}^m$ factorises as

$$f(x) \;=\; r(x)^p \, g\big(x / r(x)\big), \qquad g : \mathbb{S}_r^{D-1} \to \mathbb{R}^m,$$

and we parameterise g as a unit-sphere map $e : \mathbb{S}_r^{D-1} \to \mathbb{S}_r^{m-1}$.

Encoder:

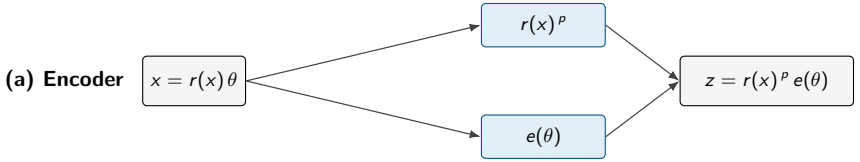
- radial: $r(x)^p$, $r := \|\cdot\|$
- angular direction:
  $e(\theta) = \mathrm{normalise}(E_\phi(\theta))$
  $$z = r(x)^p \, e(\theta)$$

Decoder:

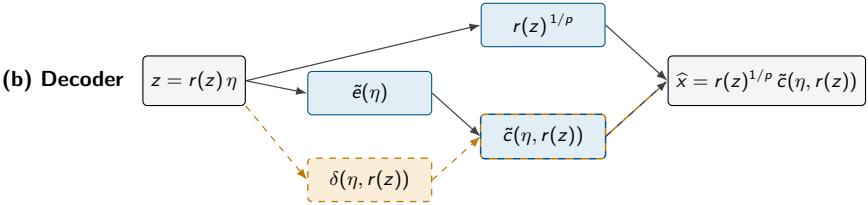
- radial: $r(z)^{1/p}$
- angular direction: $\tilde{c}(\eta, r(z)) = \mathrm{normalise}\big(\tilde{e}(\eta) + \delta(\eta, r(z))\big)$
- correction:
  $\delta(\eta, r) = \Delta(\eta, s) - \Delta(\eta, 0),\ s = \frac{1}{1+r}$
  $$\widehat{x} = r(z)^{1/p} \, \tilde{c}(\eta, r(z))$$

All angular components $e, \tilde{e}, \Delta$ are MLPs. The round trip $r(\widehat{x}) = r(x)$ is algebraically exact.

Appendix: full architecture



$\theta{=}x/r(x) \;\Rightarrow\; f(\lambda x) = \lambda^p f(x)$ *exactly,* $r(f(x)) = r(x)^p$



asymp. homogeneous: $\sup_\eta \|\delta(\eta, r(z))\| \to 0$ *as* $r(z) \to \infty$*;* $r(\widehat{x}) = r(z)^{1/p}$ *exactly*

Appendix: synthetic data construction

Curved surface (D=3, m=2)

1. Latent pair (t, v) from a correlated bivariate Student-t with d.f. ν.
2. Angular chart $(z_1, z_2) = (a_1 \tanh t,\ a_2 \tanh v)$, lifted to $\mathbb{S}^2$ via inverse stereographic projection $e(z)$.
3. Radial profile $r(t) = r_0 + \sqrt{t^2 + \varepsilon^2}$.
4. Ambient sample

$$X = r(t)\, e(z) \ \in \ \mathbb{R}^3.$$

Tail index $\alpha = \nu$ inherited from the Student-t.

Flexible toy-model (D=10, m=3 presented)

1. Radial $R \sim \mathrm{Pareto}(\alpha)$ on $[1, \infty)$, $\mathbb{P}(R > t) = t^{-\alpha}$.
2. Direction $U \sim \mathrm{Unif}(\mathbb{S}^{m-1})$.
3. Latent $Y = R\, U \in \mathbb{R}^m$.
4. Embedding $\varphi : \mathbb{R}^m \to \mathbb{R}^D$,

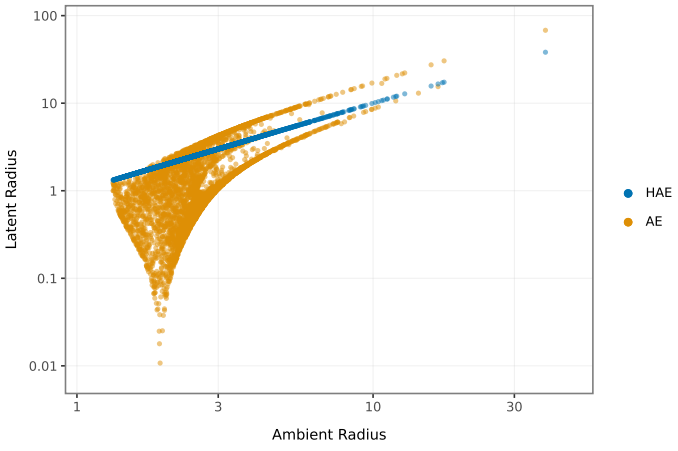
$$\varphi(y) \ = \ A\, y \ + \ \kappa\, B \, \tanh(C\, y),$$

A a random orthonormal frame, B, C fixed, κ tunes curvature.

5. Ambient sample $X = \varphi(Y)$.

Asymptotically linear (tanh saturates) $\Rightarrow$ X inherits tail index α.

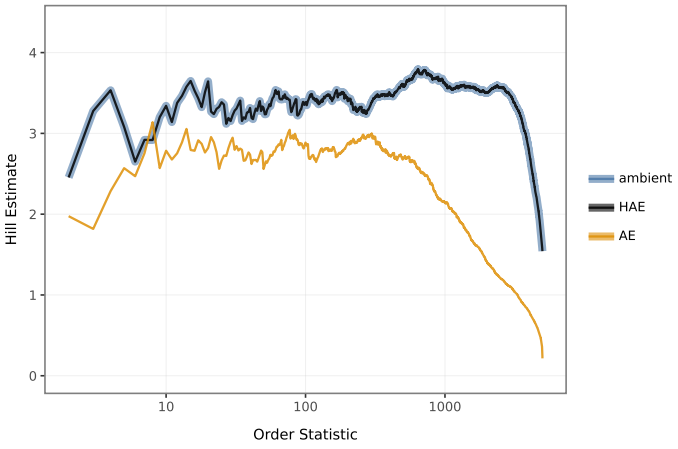
Appendix: ambient radius preservation (curved manifold)



Latent vs. ambient radius for HAE and a standard AE, log-log. $D = 3$, $m = 2$.

Clear correspondence between latent and ambient radii under the HAE; the standard AE drifts into a broad cloud.

Appendix: latent tail index (curved manifold)



Hill estimates of the tail index at different order statistics, ambient vs latent. $D = 3$, $m = 2$.

HAE's latent Hill curve sits on top of the ambient curve; the unconstrained AE drifts away.

Reported $\widehat{\alpha}$ uses the top 10% of order statistics, $k = \lfloor 0.1\, n \rfloor$; $n = 4000$, $k = 400$ here.

Appendix: reconstruction (curved manifold)

	Average per-sample MSE
HAE	1.8×10^{-5}
AE	7.6×10^{-6}

Both reconstruct the smooth surface accurately. Absolute error is tiny for either model.

D=3, m=2, α=2.8.