Pareto Variational Autoencoder

Mincheol Cho[1*] Yedarm Seong[2*] Joong-ho Won[1,2†]

[1]Department of Statistics, Seoul National University [2]Interdisciplinary programs in Artificial Intelligence, Seoul National University

Highlights

- We introduce the symmetric Pareto (**symPareto**) distribution, a heavy-tailed and robust alternative to Gaussian/Laplace.
- We propose the **ParetoVAE**, derived from a joint minimization view of variational inference via the γ-power divergence.
- Across sparse heavy-tailed data and image denoising, ParetoVAE consistently improves robustness and tail reconstruction.

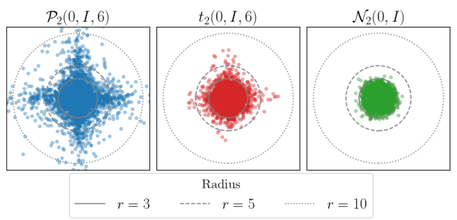
Symmetric Pareto Distribution

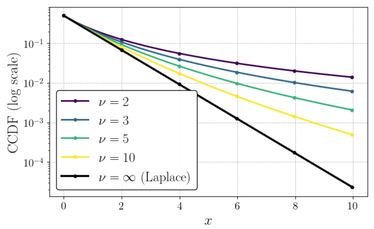
- Its density is defined as

$$\mathcal{P}_n(x \mid \mu,\sigma,\nu) = \frac{C_{n,\nu,\nu}}{\bar{\sigma}}\left(1+\frac{1}{\nu}\left\|\frac{x-\mu}{\sigma}\right\|_1\right)^{-(\nu+n)}, \quad C_{n,\nu_1,\nu_2} = \frac{\Gamma(\nu_1+n)}{(2\nu_2)^n\Gamma(\nu_1)}.$$

- Two Interpretations:

1. ℓ_1-norm–based counterpart of the multivariate t distribution.
2. Heavy-tailed analogue of the product of i.i.d. univariate Laplace distributions.





(a) 2D scatter plots of SymPareto, Student's t, and Gaussian. (b) Log-scale CCDF of SymPareto with varying ν.

γ-power Divergence and Its Geometry

- γ-power divergence [1] serves a tractable alternative for optimization of heavy-tailed families with similar information geometric structure.

$$\mathcal{D}_\gamma(q\,\|\,p) := \frac{\mathcal{C}_\gamma(q,p)-\mathcal{H}_\gamma(q)}{\gamma}, \text{ where } \mathcal{H}_\gamma(p) = -\|p\|_{1+\gamma}, \mathcal{C}_\gamma(q,p) = -\mathbb{E}_q\left[\left(\frac{p}{\|p\|_{1+\gamma}}\right)^\gamma\right].$$

- This divergence parallels the e-geodesic for the KL divergence by introducing the γ-power geodesic, which induces γ-flat submanifolds as

$$\mathcal{S}_\gamma = \left\{p_\theta(x) \propto (1+\gamma\theta^\top s(x))^{\frac{1}{\gamma}} : \theta\in\Theta\right\}, \tag{1}$$

where $s(x)$ is the sufficient statistic of the given distribution.

- When $\mu=0$, $s(x)=|x|$ is valid and thus the symPareto is γ-flat with $\gamma=-\frac{1}{\nu+n}$.
- When $\mu\neq 0$, location shifts can be removed by translation to a zero-centered coordinate system, and this transformation preserves the γ-flat structure.

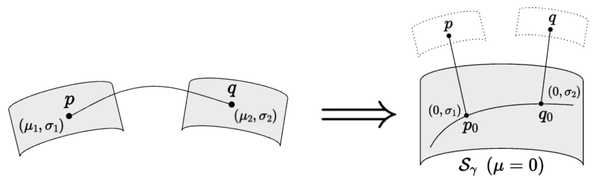


Figure: Illustration of two symPareto manifolds with different μ and their translation to the γ-flat manifold S_γ.

[1] : Juno Kim, Jaehyuk Kwon, Mincheol Cho, Hyunjong Lee, and Joong-Ho Won. t3-variational autoencoder: Learning heavy-tailed data with student's t and power divergence. (ICLR 2024)

VAE as Joint Minimization

- Data: $x\in\mathbb{R}^n$, latent: $z\in\mathbb{R}^m$. Evidence lower bound (ELBO):

$$\text{ELBO}(x;\theta,\phi) = \mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)] - \mathcal{D}_{\text{KL}}(q_\phi(z|x)\,\|\,p_Z(z))$$

- Model distribution manifold $\mathcal{P} := \{p_\theta(x,z) = p_\theta(x|z)p_Z(z) : \theta\in\Theta\}$
- Data distribution manifold $\mathcal{Q} := \{q_\phi(x,z) = p_{\text{data}}(x)q_\phi(z|x) : \phi\in\Phi\}$
- ELBO can be recast as KL divergence between $\mathcal{P},\mathcal{Q}$:

$$\text{argmin}_{\theta,\phi}\,\mathcal{D}_{\text{KL}}(q_\phi(x,z)\,\|\,p_\theta(x,z)) = -\mathbb{E}_{x\sim p_{\text{data}}}[\text{ELBO}(x;\theta,\phi)] - H(p_{\text{data}}).$$

- Key Point: Rather than deriving a new ELBO, we directly solve a joint divergence minimization problem over $\mathcal{P}$ and $\mathcal{Q}$.
- This motivates replacing the KL divergence with alternative divergences better suited to heavy-tailed models.

ParetoVAE Model

- The construction of ParetoVAE starts from the heavy-tailed joint decoder model:

$$p_\theta(x,z) \propto \left[1+\frac{1}{\nu}\left(\|z\|_1 + \frac{\|x-\mu_\theta(z)\|_2^2}{\sigma^2}\right)\right]^{-\frac{2(\nu+m)+n}{2}}. \tag{2}$$

- (2) yields the symPareto prior $p(z)$ and t-decoder $p_\theta(x\mid z)$:

$$p(z) = \mathcal{P}_m(z\mid 0,\mathbf{1}_m,\nu), \quad p_\theta(x\mid z) = t_n\left(x\mid \mu_\theta(z), \frac{\nu+\|z\|_1}{2(\nu+m)}\sigma^2 I_n,\ \nu+m\right)$$

- Encoder: use a symPareto with degrees of freedom increased by $n/2$ to reflect the contribution of the data dimension.

$$q_\phi(z|x) = \mathcal{P}_m(z\mid \mu_\phi(x),\sigma_\phi(x),\nu+n/2).$$

- If $Z\sim\mathcal{L}_n(0,I_n)$ and $W\sim\Gamma(\nu,1)$ are independent, then Then

$$T := (\nu/W)Z \sim \mathcal{P}_n(0,\mathbf{1}_n,\nu),$$

yielding a simple Laplace-Gamma representation for sampling.

- From $\mathcal{D}_\gamma(q(z|x)p_{\text{data}}(x)\,\|\,p_\theta(x|z)p(z))$, we derive γ-loss with $q_{\phi,0} = \mathcal{P}_m(0,\sigma_\phi(x),\nu+n)$, $p_{\text{alt}} = \mathcal{P}_m(0,k\mathbf{1}_m,\nu+n)$:

$$L_\gamma(\theta,\phi) = \mathbb{E}_{x\sim p_{\text{data}}}\left[\underbrace{\frac{1}{2\sigma^2}\mathbb{E}_{z\sim q_\phi(\,\cdot\,|x)}\|x-\mu_\theta(z)\|_2^2}_{\text{reconstruction}} + \underbrace{\alpha\mathcal{D}_\gamma(q_{\phi,0}\,\|\,p_{\text{alt}}) + \alpha\beta\left\|\mu_\phi(x)\right\|_1}_{\text{regularizer}}\right].$$

Decoder Selection: SymPareto ℓ_1-Decoder

- Modifying the joint decoder distribution to symPareto replaces the ℓ_2^2 reconstruction error term $\|x-\mu_\theta(z)\|_2^2$ in (2) with an ℓ_1-norm-based variant:

$$p_\theta(x,z) \propto \left[1+\frac{1}{\nu}\left(\|z\|_1 + \frac{\|x-\mu_\theta(z)\|_1}{\sigma}\right)\right]^{-(\nu+m+n)}. \tag{3}$$

- This leads to the γ-loss function in which the MSE is replaced by the mean absolute error (MAE) when $\gamma=-\frac{1}{\nu+m+n}$, thereby improving robustness to extreme values:

$$\mathcal{L}_\gamma(\theta,\phi) = \mathbb{E}_{x\sim p_{\text{data}}}\left[\frac{1}{\sigma}\mathbb{E}_{z\sim q_\phi(\,\cdot\,|x)}\|x-\mu_\theta(z)\|_1 + 2\alpha D_\gamma(q_{\phi,0}\,\|\,p_{\text{alt}}) + 2\alpha\beta\left\|\mu_\phi(x)\right\|_1\right].$$

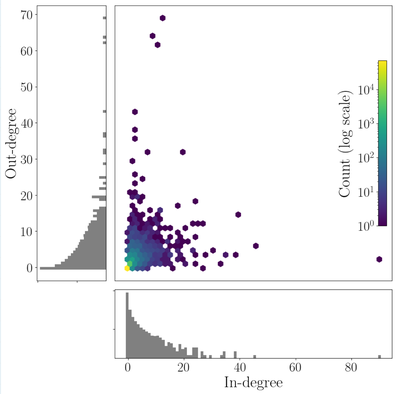
Experiments

- We evaluate ParetoVAE on low- and high-dimensional data under different decoder choices.
- Baselines: VAE (Gaussian), t^3VAE (Student's t), LVAE (Laplace), and AE.

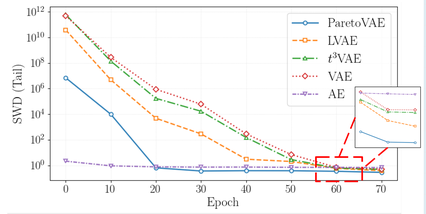
Sparse Heavy-tailed Data Analysis with t-decoder

Graph Degree Reconstruction on SNAP Epinions social network

- We measure SWD and conduct a two-sample MMD test on the joint in-/out-degree count distribution.
- ParetoVAE achieves the best tail reconstruction while preserving overall fit.



(a) 2D hexbin of joint in-/out-degree counts.



(b) SWD convergence curve on the tail region.

Model	SWD (Overall) ↓	SWD (Tail) ↓	p-value (Tail)
ParetoVAE	$\mathbf{0.044\pm0.005}$	$\mathbf{0.170\pm0.029}$	0.221 ± 0.144 ✓
LVAE	0.055 ± 0.009	0.301 ± 0.084	0.119 ± 0.090 ✓
t^3VAE	0.055 ± 0.005	0.389 ± 0.040	0.181 ± 0.110 ✓
VAE	0.061 ± 0.018	0.402 ± 0.025	0.042 ± 0.030 ✗
AE	0.074 ± 0.030	0.621 ± 0.304	0.028 ± 0.027 ✗

(c) Overall SWD, tail SWD, and tail MMD p-values.

Word Frequency Reconstruction on Wikitext-2

- ParetoVAE best preserves rare-word structure while maintaining head-word fidelity.

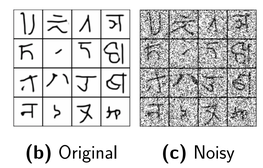
Model	Head			Tail		
	Overlap ↑	Jaccard ↑	p-value	Overlap ↑	Jaccard ↑	p-value
ParetoVAE	$\mathbf{0.981\pm0.001}$	$\mathbf{0.964\pm0.001}$	0.417 ± 0.237 ✓	$\mathbf{0.717\pm0.035}$	$\mathbf{0.560\pm0.043}$	0.178 ± 0.161 ✓
LVAE	0.772 ± 0.008	0.629 ± 0.010	0.233 ± 0.148 ✓	0.230 ± 0.003	0.130 ± 0.002	0.001 ± 0.000 ✗
t^3VAE	0.739 ± 0.002	0.586 ± 0.002	0.665 ± 0.116 ✓	0.226 ± 0.001	0.127 ± 0.001	0.001 ± 0.000 ✗
VAE	0.775 ± 0.017	0.633 ± 0.022	0.229 ± 0.200 ✓	0.224 ± 0.009	0.126 ± 0.006	0.001 ± 0.001 ✗
AE	0.642 ± 0.007	0.473 ± 0.008	0.001 ± 0.000 ✗	0.197 ± 0.004	0.109 ± 0.003	0.001 ± 0.000 ✗

Table. Word frequency reconstruction on Wikitext-2. ✓ does not reject the test, ✗ reject.

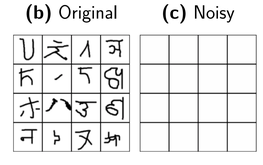
Image Denoising Application with SymPareto Decoder

- ParetoVAE remains robust on high-dimensional corrupted images.

(a) CelebA.



(b) Original (c) Noisy



(d) ParetoVAE (e) The others

Figure. Denoising results on CelebA and Omniglot. (a) shows CelebA; (b)–(e) show Omniglot.

*Equal contribution

†won.j@snu.ac.kr