

Hyperspherical Variational Auto-Encoders

Reproduction & Analyse Critique

Mouhssine Rifaki Raphaël Rubrice

Master MVA - Computational Statistics

7 janvier 2026

Rappel : Auto-Encodeurs Variationnels

Objectif VAE : Maximiser l'ELBO

$$\mathcal{L}(\phi, \psi) = \underbrace{\mathbb{E}_{q_\psi(z|x)} [\log p_\phi(x|z)]}_{\text{Reconstruction}} - \underbrace{\text{KL}(q_\psi(z|x) \| p(z))}_{\text{Régularisation}}$$

Choix standard :

- Prior : $p(z) = \mathcal{N}(0, I)$
- Postérieur : $q_\psi(z|x) = \mathcal{N}(\mu(x), \sigma^2(x)I)$

Reparamétrisation : $z = \mu + \sigma \odot \epsilon$, $\epsilon \sim \mathcal{N}(0, I)$

Problèmes du Prior Gaussien

Basses dimensions : Gravité vers l'origine

- Prior gaussien attire tous les points vers 0.
- Empêche la séparation naturelle des clusters.

Hautes dimensions : Effet bulle de savon

- Masse concentrée sur une coquille sphérique.
- Distance euclidienne perd son sens.

Distribution de Von Mises-Fisher

Définition : Distribution sur l'hypersphère $S^{m-1} \subset \mathbb{R}^m$

$$q(z|\mu, \kappa) = C_m(\kappa) \exp(\kappa \mu^\top z), \quad \|z\|_2 = 1$$

où $C_m(\kappa) = \frac{\kappa^{m/2-1}}{(2\pi)^{m/2} I_{m/2-1}(\kappa)}$

Paramètres :

- $\mu \in S^m$: Direction moyenne
- $\kappa \geq 0$: Concentration ($\kappa = 0 \Rightarrow$ Uniforme)

Avantages : Prior uniforme (pas d'attraction origine), κ appris par échantillon.

Inconvénients : $C_m(\kappa)$ implique la fonction de Bessel (instabilité), Nouvelle reparamétrisation préservant les gradients à définir.

Divergence KL & Échantillonnage

Divergence KL :

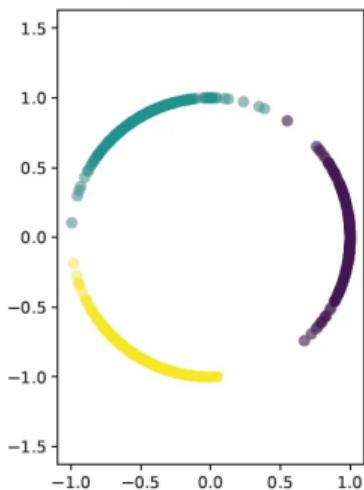
$$\text{KL}(\text{vMF}(\mu, \kappa) \| \mathcal{U}(S^{m-1})) = \kappa \frac{I_{m/2}(\kappa)}{I_{m/2-1}(\kappa)} + \log C_m(\kappa) + \text{cst}$$

Échantillonnage (Ulrich 1984) :

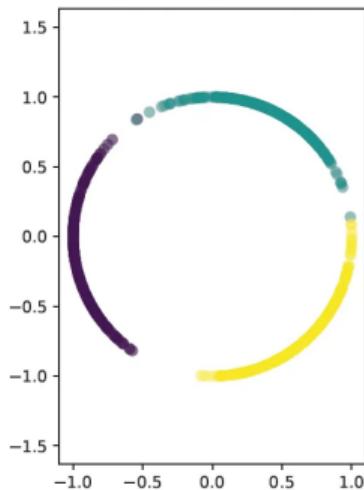
- ① Échantillonner $\omega \sim g(\omega | \kappa, m)$ par acceptation-rejet (1D)
- ② Échantillonner $v \sim \mathcal{U}(S^{m-2})$
- ③ Construire $z' = (\omega, \sqrt{1 - \omega^2} v^\top)^\top$
- ④ Transformation de Householder : $z = U(\mu)z'$

Rejet en 1D seulement \Rightarrow pas de malédiction de la dimensionnalité

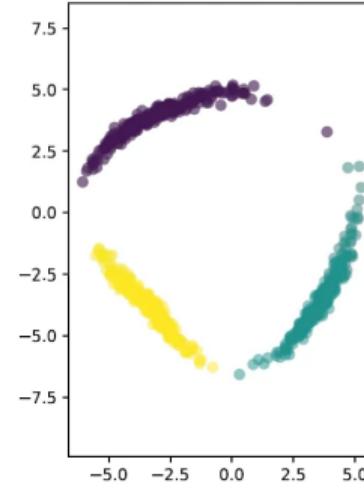
Exp. 1 : Reconstruction de 3 vMF sur S^1 , plongé dans \mathbb{R}^{100}



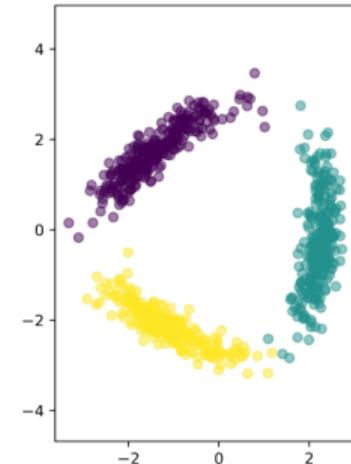
Données originales



\mathcal{S} -VAE : préservé



\mathcal{N} -VAE ($\beta_{KL}=0.1$) :
déformé

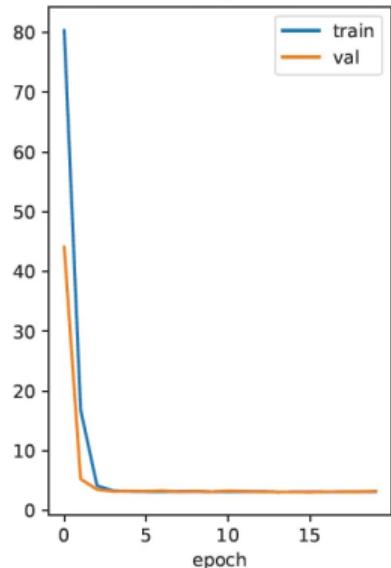


\mathcal{N} -VAE ($\beta_{KL}=1$) :
davantage déformé

Modèle	Train LL	Val LL
\mathcal{S} -VAE	-94.82	-94.85
\mathcal{N} -VAE ($\beta_{KL}=0.1$)	-107.73	-107.82
\mathcal{N} -VAE ($\beta_{KL}=1$)	-97.94	-97.97

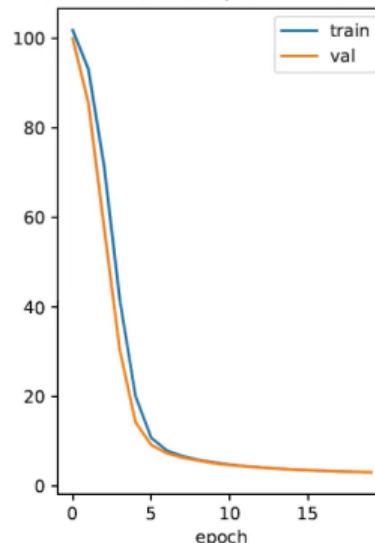
Exp. 1 : Courbes d'Apprentissage

S-VAE training & validation loss
Train LL: -94.8237, Val LL: -94.8456



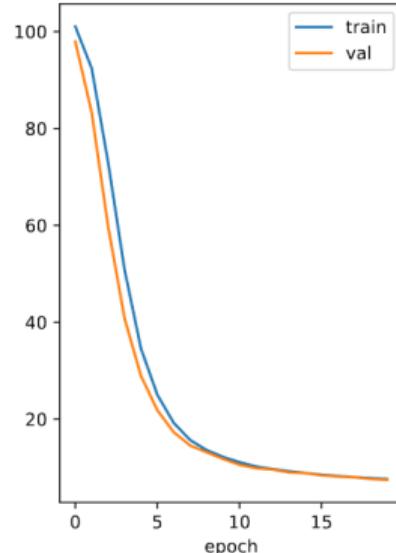
S-VAE : convergence rapide

N-VAE training & validation loss
Train LL: -107.7333, Val LL: -107.8241



N-VAE : convergence lente ($\beta_{KL}=0.1$)

N-VAE training & validation loss
Train LL: -97.9366, Val LL: -97.9718



N-VAE : convergence lente ($\beta_{KL}=1$)

S-VAE récupère la structure circulaire avec +13 nats de log-vraisemblance

Exp. 2 : Évaluation Non-Supervisée sur MNIST (Table 1)

Métriques : LL (IWAE, 500 samples), $\mathcal{L}[q]$ (ELBO), RE (Reconstruction), KL

Protocole : 5 runs, early stopping (patience=50), warmup 100 epochs, 10% données.

d	\mathcal{N} -VAE				\mathcal{S} -VAE			
	LL	$\mathcal{L}[q]$	RE	KL	LL	$\mathcal{L}[q]$	RE	KL
2	-137.42 \pm .12	-138.91 \pm .08	-131.52 \pm .24	7.39 \pm .18	-134.18 \pm .94	-135.47 \pm .01	-128.06 \pm .15	7.41 \pm .21
5	-112.03 \pm .58	-114.67 \pm .52	-101.89 \pm .61	12.78 \pm .19	-110.27 \pm .43	-112.94 \pm .47	-99.31 \pm .52	13.63 \pm .14
10	-95.51 \pm .67	-99.83 \pm .58	-80.17 \pm .71	19.66 \pm .23	-94.72 \pm .54	-99.24 \pm .61	-78.29 \pm .68	20.95 \pm .17
20	-90.34\pm.41	-96.18\pm.38	-72.51 \pm .63	23.67 \pm .42	-90.89 \pm .52	-97.73 \pm .55	-68.94\pm.71	28.79 \pm .34
40	-90.21\pm.48	-96.37\pm.44	-72.38 \pm .79	23.99 \pm .61	-92.14 \pm .57	-102.89 \pm .62	-68.91\pm.93	33.98 \pm .58

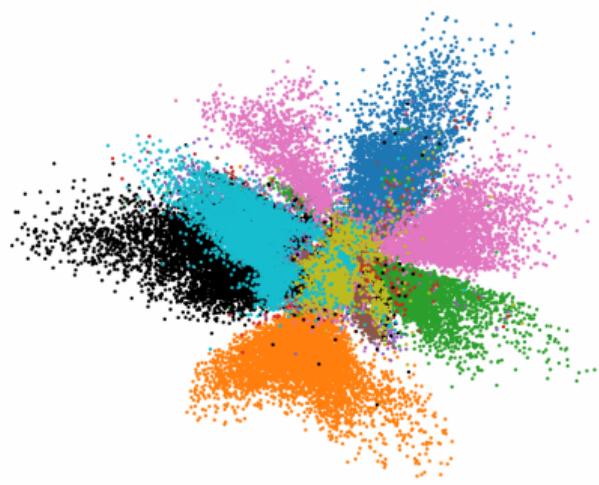
Observations

- $d \leq 10$: \mathcal{S} -VAE meilleur en LL
- $d \geq 20$: \mathcal{N} -VAE rattrape (+ de params)
- RE : \mathcal{S} -VAE toujours meilleur

Interprétation

Prior uniforme \Rightarrow pas de "gravité origine"
 \Rightarrow meilleure utilisation de l'espace latent
 \Rightarrow reconstruction plus fidèle

Exp. 3 : Visualisation du dataset MNIST



N-VAE: Espace latent dans \mathbb{R}^2



S -VAE: Espace latent dans S^2 (projection de Hammer)

Les classes superposées dans l'espace latent du \mathcal{N} -VAE sont mieux séparées dans celui du S -VAE (cohérent avec le papier).

Exp. 4 : Classification Semi-Supervisée (M1)

Protocole : K-NN sur l'espace latent (MNIST, 5 runs, 10% du train set).

Les valeurs indiquent l'accuracy (%). * : différence significative ($p < 0.01$).

Dim (d)	100 Labels		600 Labels		1000 Labels	
	\mathcal{N} -VAE	\mathcal{S} -VAE	\mathcal{N} -VAE	\mathcal{S} -VAE	\mathcal{N} -VAE	\mathcal{S} -VAE
2	57.2 ± 3.6	50.8 ± 4.9	57.7 ± 2.7	47.6 ± 6.2	56.1 ± 1.7	41.4 ± 12.8
5	71.8* ± 2.1	66.9 ± 1.3	79.9 ± 1.6	77.4 ± 1.3	80.2* ± 0.9	78.4 ± 0.7
10	70.3 ± 2.2	72.0 ± 1.6	83.7 ± 0.7	84.9 ± 0.3	86.7 ± 1.3	86.4 ± 0.5
20	66.9 ± 2.5	69.0 ± 1.3	85.4 ± 0.3	87.2* ± 0.5	87.7 ± 0.6	89.3* ± 0.4
40	65.8 ± 0.9	69.7* ± 2.0	83.1 ± 0.4	87.4* ± 0.4	85.7 ± 0.5	89.0* ± 0.4

- $d \leq 5$: \mathcal{N} -VAE meilleur (contraire aux résultats des auteurs)
- $10 \leq d \leq 20$: \mathcal{S} -VAE avantage clair (concordant avec les résultats des auteurs)
- $d > 20$: \mathcal{S} -VAE avantage clair (contraire aux résultats des auteurs)

Comparaison M1 : Résultats Originaux vs Reproduction

Table 2 du Papier (Davidson et al.)
(Entrainement sur 100% MNIST vs 10% pour nous)

Dim (d)	100 Labels		600 Labels		1000 Labels	
	\mathcal{N} -VAE	\mathcal{S} -VAE	\mathcal{N} -VAE	\mathcal{S} -VAE	\mathcal{N} -VAE	\mathcal{S} -VAE
2	72.6	77.9*	80.8	84.9*	81.7	85.6*
5	81.8	87.5*	90.9	92.8*	92.0	93.4*
10	75.7	80.6*	88.4	91.2*	90.2	92.8*
20	71.3	72.8*	88.3	89.1*	90.1	91.1*
40	72.3*	67.7	88.0	87.4	90.3	90.4

Divergences Méthodologiques

- **Données** : 60k (Eux) vs 6k (Nous).
- **Robustesse** : 20 runs (Eux) vs 5 runs (Nous).

Analyse de la Reproduction

- **Validé** : \mathcal{S} -VAE domine sur $10 \leq d \leq 20$.
- **Different** : Nos résultats favorisent \mathcal{S} -VAE à $d = 40$ et \mathcal{N} -VAE à $d \leq 5$

Exp. 5 : Classification Semi-Supervisée (Architecture M1+M2)

Comparaison des configurations Latent (z_1) + Latent (z_2).

Dimensions		Configuration Modèle		
z_1	z_2	$\mathcal{N} + \mathcal{N}$	$\mathcal{S} + \mathcal{S}$	$\mathcal{S} + \mathcal{N}$
5	5	61.5*	63.0*	65.3*
5	10	60.9	62.7	66.1*
5	50	62.5	62.4	66.6*
10	5	66.9*	70.1*	70.9*
10	10	68.7*	70.6*	71.3*
10	50	69.4*	70.3*	69.7
50	5	69.9	73.3*	73.4*
50	10	68.7	72.7*	73.1*
50	50	76.3*	71.9*	74.0*

Tendances Observées

- Performance \propto Dimension de z_1 .
- Optimum atteint pour $z_1 = 50$ (comme dans le papier).

Stratégie $\mathcal{S} + \mathcal{N}$

L'hybridation est la stratégie la plus robuste (Meilleur score ou équivalent dans **8/9** cas).

Comparaison avec Résultats Originaux

Papier (Table 3) :

Latent Dimensions		Model Configuration		
Dim (z1)	Dim (z2)	$\mathcal{N} + \mathcal{N}$	$\mathcal{S} + \mathcal{S}$	$\mathcal{S} + \mathcal{N}$
5	5	90.0	94.0*	93.8
5	10	90.7	94.1	94.8*
5	50	90.7	92.7	93.0*
10	5	90.7	91.7	94.0*
10	10	92.2	96.0*	95.9*
10	50	92.9	95.1	95.7*
50	5	92.0	91.7	95.8*
50	10	93.0	95.8	97.1*
50	50	93.2	94.2	97.4*

Écarts :

- Même que précédemment

Conclusion :

- Nos résultats confirment que $\mathcal{S} + \mathcal{N}$ est le plus intéressant (Meilleur ou comparable partout)
- Reproduction complète infaisable (coût computationnel)

Analyse des Divergences avec la Référence

1. Régime "Low-Data"

Passage 60k → 6k images ⇒ Faible densité de données ⇒ modèle plus dépendant de la géométrie imposée par le Prior.

- **\mathcal{N} -VAE** : Prior Gaussien (informatif) → Risque accru de *posterior collapse*.
- **\mathcal{S} -VAE** : Prior Uniforme (non-informatif) → Plus robuste, explique les observations à $d = 40$?

2. Instabilité de l'Estimateur ($N = 5$)

- **Variance élevée** : Écarts-types importants observés (ex: ± 12.8 en $d = 2$).
- **Conséquence** : Certaines inversions de performance par rapport au papier (surtout $d \leq 5$) pourraient être du bruit statistique

3. Biais d'Hyperparamètres

- Les schémas d'optimisation (KL warm-up, LR) sont calibrés pour le dataset complet ⇒ Inadaptés au régime 6k ? Nouvelle dynamique de convergence de κ (vMF) vs σ (Gaussienne).

Ce qui est Nouveau

Contributions :

- ① κ appris (vs fixé dans Guu et al. 2018)
- ② Extension reparamétrisation par rejet (Lemme 2)
- ③ Prior uniforme sur hypersphère

Points intéressants :

- Design espace latent conscient de la géométrie
- Meilleure séparabilité des clusters en basse dim.
- Adapté aux données directionnelles
- Amélioration prédiction de liens (graphes)

Limitations

Faiblesses :

- Difficulté d'estimation lorsque $\kappa \gg d$
- Collapse haute dimension : Aire $\mathcal{S}^{m-1} \rightarrow 0$ quand $m \rightarrow \infty$
- vMF moins expressif (1 param κ) que gaussienne diagonale
- Coût et stabilité : gradients des fonctions de Bessel, échantillonnage par rejet

Manquant :

- Pas d'analyse théorique de quand \mathcal{S} -VAE est meilleur
- Absence d'expériences sur des données directionnelles réelles (ex : directions du vent, structures protéiques) pourtant citées en motivation
- Manque de discussion plus poussée sur l'implémentation stable de \mathcal{S} -VAE (κ grand, contraintes sur le modèle)

Difficultés reproduction :

- $\sim 1\text{h/run}$ (training set complet) $\times 10\text{-}20$ runs \times configs

Bilan

Papier :

- VAE hypersphérique avec distribution vMF
- κ appris, avantages en basse dimension

Notre reproduction :

- Récupération variété circulaire : OK
- Table 1 (évaluation non-supervisée) : OK
- Tendances M1 pour $10 \leq d \leq 20$: OK
- Tendances M1+M2: OK
- Valeurs absolues : différentes (contraintes calcul)

Leçon : La géométrie de l'espace latent compte.

Pour aller plus loin

Papier

Davidson et al., *Hyperspherical Variational Auto-Encoders*
arxiv.org/abs/1804.00891

Code

github.com/blackswan-advitamaeternam/HVAE