# Développement de modèles non supervisés pour l'obtention de représentations latentes interprétables d'images

Emma Jouffroy

université de BORDEAUX

THÈSE PRÉSENTÉE

POUR OBTENIR LE GRADE DE

DOCTEURE DE L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE SCIENCES ET PHYSIQUES
DE L'INGÉNIEUR

AUTOMATIQUE, PRODUCTIQUE, SIGNAL ET IMAGES, INGÉNIERIE COGNITIVE

Par **Emma Jouffroy**

---

# Développement de modèles non supervisés pour l'obtention de représentations latentes interprétables d'images

---

Sous la direction de **Audrey Giremus** et **Yannick Berthoumieu**

Soutenue le 28 mars 2024, devant le jury composé de :

| | | | |
|---|---|---|---|
| **Yoann Altmann** | Associate Professor | Heriot-Watt University | Rapporteur |
| **Thomas Oberlin** | Professeur Associé | ISAE-SUPAERO | Rapporteur |
| **Jérôme Idier** | Directeur de Recherche | Centrale-Nantes | Président du jury |
| **Jean-François Aujol** | Professeur des universités | Université de Bordeaux | Examinateur |
| **Audrey Giremus** | Professeure des universités | Université de Bordeaux | Directrice |
| **Yannick Berthoumieu** | Professeur des universités | Université de Bordeaux | Co-directeur |
| **Olivier Bach** | Ingénieur-chercheur | CEA-CESTA | Invité |

# Remerciements

Ces trois années de thèse ont été particulièrement riches et intenses, tant sur le côté personnel que professionnel. J'ai eu l'occasion d'acquérir des compétences solides sur un sujet qui me tient particulièrement à cœur depuis sept ans maintenant, de faire de nombreuses rencontres humaines toutes plus riches les unes que les autres et d'être épaulée avec rigueur et bienveillance durant l'ensemble de ce parcours.

Je souhaite tout particulièrement remercier Audrey et Yannick, mes directeurs de thèse. Merci pour votre écoute et vos conseils, votre suivi tout au long de ma thèse et votre bienveillance. Merci de n'avoir jamais rien dit pour les nombreux mails que j'ai pu vous envoyer les week-ends, jours fériés et vacances, pour le temps que vous avez accordé à me transmettre vos connaissances lors de nos nombreuses réunions, entourés de mille cubes, fourches et cafés. Je n'aurais pu souhaiter de meilleures directions que la vôtre.
Je tiens également à remercier profondément Olivier, mon tuteur de thèse du côté CEA. Merci d'avoir initié ce projet, qui m'aura mené jusqu'ici. Merci pour nos échanges passionnés, ton soutien inestimable et ta passion vibrante du sujet que tu m'as toujours transmise avec gentillesse et bienveillance.

Par ailleurs, je suis également reconnaissante envers les nombreuses personnes que j'ai eues l'occasion de rencontrer tant au laboratoire du CEA que celui de l'IMS. D'abord, Guillaume, qui m'a soutenue dès le départ et pour qui je n'ai jamais cessé d'avoir de tendres pensées. Je souhaite également remercier Christophe, pour ton implication et ton écoute qui m'auront été d'une aide précieuse pour venir à bout de ce projet. Merci à mes collègues côté CEA pour nos nombreux échanges et ces moments partagés sur le centre, mais également à l'extérieur. Je pense notamment à Raph, Alain, Olivier, Thierry, chaton, Maëlys et à toi, Thibault, qui est devenu si cher à mes yeux aujourd'hui. Sache que j'ai souvent appliqué le conseil que tu n'as cessé de me donner ces trois dernières années : "Si ça ne marche pas, débranche la prise et rebranche-la." Merci aussi à tous les doctorants, stagiaires et permanents de l'IMS avec qui les échanges ont toujours été d'une grande richesse intellectuelle. Merci notamment à Sara et Anaïs, qui font aujourd'hui, j'en suis sûre, de brillantes chercheuses. Merci à Marwane et Solène, de m'avoir accompagnée dans la mise en place de ces projets à la BEE Branche, et ces échanges qui ont toujours été si intéressants. Merci à Paul pour sa brillance intellectuelle, à Pedro pour sa douceur et nos rires à Paris, à Camille pour son écoute, sa bonté et nos jeux de société avec Aymeric. Merci enfin à Rémi et Mathieu pour nos échanges sur des sujets toujours plus farfelus.

Je voudrais remercier également mes amis, que je considère désormais comme ma deuxième famille. Tout d'abord, merci à tous mes copains de MMI qui sont toujours à mes côtés aujourd'hui, et qui m'ont tant apporté. Merci à Mathilde, ma sœur de toujours. Qui aurait cru, lorsque l'on révisait le brevet dans ton jardin, que l'on en serait là aujourd'hui ? Je te promets d'être présente pour chaque étape de ta vie, quoi qu'il arrive, d'être la meilleure marraine possible pour votre petit petite Aria (je l'initierai au surf et à la science promis !), ainsi que de continuer à essayer d'expliquer l'IA à Loïs et son intérêt pour

tout, permanent. Merci également à mes amis de l'ENSC, pour ne jamais avoir perdu confiance en moi, m'avoir remonté le moral et conseillé quand ça n'allait pas. Merci à toi, Julien, de m'avoir poussée lorsque je ne me sentais plus capable. Tes mots ont toujours résonné lorsque je ne croyais plus en moi. Merci à Mathilde, Raph, Mathilda, Maria, Azi, Vincent, Emma, MC & Cyril d'être tous autant des personnes exceptionnelles et d'avoir fait de cette période de ma vie un moment si important. Merci également et surtout à Nathan, Marie et Nina. Je suis la plus heureuse et chanceuse de vous avoir à mes côtés, de pouvoir avoir tant de discussions riches, tant de fous rires, tant de découvertes. Merci de ne m'avoir jamais jugée, toujours épaulée et toujours aimée. Finalement, un grand merci à mes amis du lycée, de la butte. Merci à vous, Fanny, Marie, Amandine, Elsa et Laurine pour votre douceur, votre gentillesse, votre folie. Merci pour nos rires, merci pour nos discussions, merci de rester proches de moi, même de si loin. Merci d'avoir apporté cette sororité si importante dans ma vie.

Désormais, je voudrais remercier ma famille, sans qui je ne serais jamais arrivée là où je suis actuellement. Merci à mes grands-parents, oncles et tantes et cousins et cousines. Merci surtout à papou, qui a tout fait pour que je puisse rentrer au CEA, et grâce à qui je suis ici aujourd'hui. Je suis persuadée que la pile de sciences et vie que tu avais chez toi et que je lisais petite n'y a pas été pour rien dans mon parcours académique. Merci à Séverine, Yanis et Guillaume de m'avoir accueilli dans leur incroyable famille, et d'avoir toujours été curieux et présents. Merci à Matisse et Aloys, mes frères. Je suis fière d'être votre grande sœur. Finalement, et surtout, merci à vous, Papa et Maman. Merci d'avoir été présents, de m'avoir épaulée et soutenue ces trois dernières années, qui parfois n'ont pas été faciles pour vous également. Merci de m'avoir toujours protégée lors des moments difficiles, pour que ça n'impacte pas ma thèse. Merci d'être ces parents si géniaux que vous êtes, d'avoir toujours cru en moi et d'avoir toujours été là. Je vous aime.

Pour terminer, je tiens à te remercier pour tout Maxime. Je n'aurai pas suffisamment de mots pour décrire l'importance que tu as joué dans la réalisation de cette thèse, ni dans ma vie en général. Merci de m'avoir aidé pendant nos études en signal, en tant que professeur particulier. Si on avait parlé à l'époque que je ferais une thèse sur le sujet et que tu deviendrais docteur en psychologie, on aurait sûrement cru marcher sur la tête. Merci pour ta passion sans faille dans tous les sujets scientifiques que tu as su me transmettre, pour nos discussions jusqu'à pas d'heure sur les documentaires (pour ne pas dire vidéos YouTube) que tu regardes, que cela soit en Deep Learning, en psychologie, sur les échecs ou la rénovation de maisons. Merci d'avoir toujours accepté de m'aider lorsque je bloquais, d'avoir toujours donné ton avis instructif sur mes productions sans jamais poser de jugement, merci d'avoir toujours cru en mes capacités et d'avoir toujours respecté mes choix, quoi qu'il arrive. Je n'y serais jamais arrivé sans toi. Tu es mon moteur, mon pilier. Je t'aime de tout mon cœur.

Merci à tous. Merci pour tout.

# Resumés

**Développement de modèles non supervisés pour l'obtention de représentations latentes interprétables d'images**

Le Laser Mégajoule (LMJ) est un instrument d'envergure qui simule des conditions de pression et de température similaires à celles des étoiles. Lors d'expérimentations, plusieurs diagnostics sont guidés dans la chambre d'expériences et il est essentiel qu'ils soient positionnés de manière précise. Afin de minimiser les risques liés à l'erreur humaine dans un tel contexte expérimental, la mise en place d'un système anti-collision automatisé est envisagée. Cela passe par la conception d'outils d'apprentissage automatique offrant des niveaux de décision fiables à partir de l'interprétation d'images issues de caméras positionnées dans la chambre. Nos travaux de recherche se concentrent sur des méthodes neuronales génératives probabilistes, en particulier les auto-encodeurs variationnels (VAEs). Le choix de cette classe de modèles est lié au fait qu'elle rende possible l'accès à un espace latent lié directement aux propriétés des objets constituant la scène observée. L'enjeu majeur est d'étudier la conception de modèles de réseaux profonds permettant effectivement d'accéder à une telle représentation pleinement informative et interprétable dans un objectif de fiabilité du système. Le formalisme probabiliste intrinsèque du VAE nous permet, si nous pouvons remonter à une telle représentation, d'accéder à une analyse d'incertitudes des informations encodées.

**Mots-clés :** apprentissage automatique non supervisé, auto-encodeur variationnel, méthodes bayésiennes, espace latent

---

**Development of unsupervised models for obtaining interpretable latent representations of images.**

The Laser Megajoule (LMJ) is a large research device that simulates pressure and temperature conditions similar to those found in stars. During experiments, diagnostics are guided into an experimental chamber for precise positioning. To minimize the risks associated with human error in such an experimental context, the automation of an anti-collision system is envisaged. This involves the design of machine learning tools offering reliable decision levels based on the interpretation of images from cameras positioned in the chamber. Our research focuses on probabilistic generative neural methods, in particular variational auto-encoders (VAEs). The choice of this class of models is linked to the fact that it potentially enables access to a latent space directly linked to the properties of the objects making up the observed scene. The major challenge is to study the design of deep network models that effectively enable access to such a fully informative and interpretable representation, with a view to system reliability. The probabilistic formalism intrinsic to VAE allows us, if we can trace back to such a representation, to access an analysis of the uncertainties of the encoded information.

**Keywords:** Unsupervised machine learning, variational autoencoder, Bayesian methods, latent space.

---

**Laboratoire de l'Intégration du Matériau au Système**
UMR 5218 Université de Bordeaux, 33000 Bordeaux, France.

# Table des Matières

**Liste des Tableaux** **1**

**Liste des Figures** **3**

**Liste des Symboles** **5**

**Liste des Acronymes** **6**

**Introduction** **7**

**1 Apprentissage non supervisé de représentations** **13**
1.1 Réseau de neurones et apprentissage profond . . . . . 14
1.1.1 Du neurone formel au réseau de neurones artificiel . . . . . 14
1.1.2 Optimisation des paramètres . . . . . 16
1.1.3 Procédures d'apprentissage . . . . . 17
1.1.4 Infrastructures logicielles . . . . . 20
1.2 Problématiques d'apprentissage . . . . . 20
1.2.1 Notations . . . . . 21
1.2.2 Apprentissage supervisé . . . . . 21
1.2.3 Apprentissage . . . . . 22
1.3 Architectures génératives . . . . . 24
1.3.1 Modèles implicites . . . . . 24
1.3.2 Modèles explicites . . . . . 26
1.4 Les auto-encodeurs variationnels . . . . . 29
1.4.1 La méthode d'inférence variationnelle . . . . . 30
1.4.2 Développement du modèle et apprentissage . . . . . 31
1.4.3 Synthèse . . . . . 35
1.5 Conclusion . . . . . 36

**2 Espace latent et désentrelacement** **37**
2.1 Le désentrelacement, considéré comme une "bonne" représentation . . . . . 37
2.2 Aperçu des différentes définitions . . . . . 39
2.3 Méthodes de désentrelacement basées sur le VAE . . . . . 40
2.3.1 Méthodes basées sur une pondération de l'ELBO . . . . . 42
2.3.2 Méthodes basées sur l'ajout d'un terme de régularisation . . . . . 44
2.3.3 Méthodes basées sur une modification de la distribution *a priori* . . 47
2.3.4 Synthèse . . . . . 49
2.4 Évaluation de la capacité de désentrelacement . . . . . 50
2.4.1 Métriques basées sur une transformation . . . . . 52
2.4.2 Métriques basées sur un prédicteur . . . . . 54
2.4.3 Métriques basées sur la théorie de l'information . . . . . 55
2.4.4 Synthèse . . . . . 56
2.4.5 Jeux de données . . . . . 57
2.5 Conclusion . . . . . 59

**3 Polarisation de l'espace latent et modèle bayésien hiérarchique 61**
3.1 Concept de variables actives et passives et d'espace polarisé . . . . . 62
3.1.1 Notations . . . . . 62
3.1.2 Comportement empirique . . . . . 62
3.1.3 Formalisation théorique . . . . . 66
3.2 Approche proposée pour une séparation de l'espace latent . . . . . 69
3.2.1 Modèle hiérarchique bayésien induisant la polarisation . . . . . 69
3.2.2 Mise en œuvre de l'apprentissage . . . . . 76
3.3 Expériences . . . . . 83
3.3.1 Conditions expérimentales . . . . . 83
3.3.2 Résultats qualitatifs . . . . . 84
3.3.3 Résultats quantitatifs . . . . . 86
3.4 Conclusion . . . . . 88

**4 La polarisation comme vecteur d'un meilleur désentrelacement 89**
4.1 Approche proposée utilisant la polarisation comme un biais inductif permettant le désentrelacement . . . . . 90
4.1.1 Ajout d'un terme de corrélation totale . . . . . 91
4.1.2 Ajout d'un terme de groupe-lasso . . . . . 93
4.2 Expériences . . . . . 95
4.2.1 Impact du terme de corrélation totale dans la capacité de désentrelacement . . . . . 95
4.2.2 Apport de la polarisation dans la mesure de désentrelacement . . . 101
4.2.3 Comparaison du TC-NGVAE avec les méthodes de l'état de l'art . . 103
4.3 Conclusion . . . . . 116

**Conclusion et perspectives 118**
4.3.1 Synthèse . . . . . 118
4.3.2 Discussion . . . . . 120
4.3.3 Directions futures . . . . . 120

**A Astuce du ratio de densités 122**

**B Distributions marginales de la loi Normale-Gamma 124**

**C Divergence de Kullback entre deux lois Normales-Gamma 127**

**D Estimateurs stochastiques 131**
D.1 Estimateur "Minibatch Weighted Sampling" . . . . . 131
D.2 Estimateur "Minibatch Stratified Sampling" . . . . . 132

**E Résultats complémentaires des expériences 134**
E.1 Analyse de l'apport de la corrélation totale pour le désentrelacement . . . . 134
E.2 Analyse de la capacité de désentrelacement de différentes approches . . . . 135
E.2.1 Sélection des hyperparamètres pour les méthodes de l'état de l'art . 135
E.2.2 Entraînement effectué sur Dsprites . . . . . 136
E.2.3 Entraînement effectué sur Smallnorb . . . . . 138
E.2.4 Entraînement effectué sur Cars3d . . . . . 140

# Liste des Tableaux

1.1 Propriétés des méthodes génératives de l'état de l'art vis-à-vis de notre besoin. . . . . 35

2.1 Synthèse des propriétés des modèles de désentrelacement de l'état de l'art. 49

3.1 Architectures d'encodeur et de décodeur considérées pour les expériences. . 84
3.2 Indicateurs de polarisation pour le NGVAE et deux approches de l'état de l'art. . . . . 86
3.3 Indicateurs de polarisation entre le NGVAE et deux approches de l'état de l'art, mesurés uniquement sur les variables actives. . . . . 87

4.1 Indicateurs de polarisation pour mesurer l'apport du terme de corrélation totale dans la fonction coût. . . . . 96
4.2 Différence quadratique moyenne des indicateurs de désentrelacement obtenus par le TC-NGVAE pour différentes dimensions d'espaces latents . . . . . 102
4.3 Ensemble d'hyperparamètres à disposition pour les modèles proposés par la librairie "Disentanglement-Lib". . . . . 104
4.4 Moyennes des indicateurs de désentrelacement obtenues pour le DIP-VAE-II 104
4.5 Ensemble des hyperparamètres considéré pour chaque modèle selon le jeu de données. . . . . 105
4.6 Indicateurs de désentrelacement pour des méthodes entraînées sur Dsprites. 106
4.7 Indicateurs de désentrelacement pour des méthodes entraînées sur Smallnorb. 109
4.8 Indicateurs de désentrelacement pour des méthodes entraînées sur Cars3d. 112

E.1 Moyennes des indicateurs de désentrelacement obtenues pour le Factor-VAE 136
E.2 Moyennes des indicateurs de désentrelacement obtenues pour le $\beta$-VAE . . 136
E.3 Moyennes des indicateurs de désentrelacement obtenues pour le $\beta$-TCVAE 136

# Liste des Figures


1 Diagnostic de mesure inséré au sein du centre chambre du LMJ . . . . . 7
2 Illustrations d'espaces latents entrelacés et désentrelacés . . . . . 9

1.1 Illustration et formulation de la sortie d'un neurone formel. . . . . 15
1.2 Représentation d'un réseau de neurones à trois couches cachées. . . . . 16
1.3 Illustration de la propagation arrière du gradient. . . . . 17
1.4 Schématisation d'un réseau de convolutions. . . . . 18
1.5 Illustration d'une couche de convolution. . . . . 19
1.6 Taxonomie des modèles génératifs non supervisés. . . . . 24
1.7 Représentation simplifiée d'un réseau antagoniste génératif. . . . . 25
1.8 Illustration d'un modèle de flux normalisés. . . . . 27
1.9 Schématisation d'une machine de Boltzmann restreinte. . . . . 28
1.10 Représentation graphique de l'inférence variationnelle amortie. . . . . 31
1.11 Illustration d'un auto-encodeur variationnel. . . . . 32

2.1 Représentation des facteurs génératifs du jeu de données Dsprites. . . . . 38
2.2 Schématisation de la compacité et de la modularité dans un espace latent. 40
2.3 Taxonomie des méthodes de désentrelacement. . . . . 41
2.4 Illustration du parcours de l'espace latent. . . . . 50
2.5 Taxonomie des métriques de désentrelacement . . . . . 52
2.6 Exemples d'images provenant de jeux de données communs. . . . . 58

3.1 Densité empirique des moyennes inférées par l'encodeur obtenues par le Factor-VAE entraîné sur Dsprites. . . . . 64
3.2 Densité empirique des variances inférées par l'encodeur obtenues par le Factor-VAE entraîné sur Dsprites. . . . . 65
3.3 Représentation graphique du Normal-Gamma Variationnal Auto-Encodeur 70
3.4 Distributions de la variance des variables latentes selon leur information. . 74
3.5 Illustration du Normal-Gamma Variationnal Auto-Encodeur. . . . . 74
3.6 Fonction permettant de définir la valeur des paramètres de la loi des variances du Normal-Gamma Variationnal Auto-Encodeur. . . . . 76
3.7 Résultats obtenus par le TC-NGVAE entraîné sur Dsprites, en excluant le facteur d'orientation. La KL-divergence "Reverse" est utilisée. . . . . 79
3.8 Valeurs prises pour la mesure de KL-divergence, selon le sens des distributions, pour différentes valeurs de moyennes. . . . . 81
3.9 Impact du sens des distributions dans la KL-divergence pour une loi bimodale et une loi unimodale. . . . . 81
3.10 Valeurs prises pour la KL-divergence "Forward", pour différentes valeurs de moyennes. . . . . 82
3.11 Images obtenues en sorties du décodeur du NGVAE entraîné sur Dsprites. . 84
3.12 Matrices de covariances empiriques de l'espace latent obtenues pour le NGVAE et deux approches de l'état de l'art. . . . . 85

3.13 Illustration de la convergence du NGVAE vers un espace polarisé, pour les 300 premières itérations. . . . . 85

4.1 Matrices de corrélation de Spearman entre les facteurs génératifs et les variables latentes du NGVAE et du TC-NGVAE, entraînés sur Dsprites en excluant le facteur d'orientation. . . . . 97
4.2 Matrice de covariance empirique et parcours de l'espace latent obtenus par le TC-NGVAE, entraîné sur Dsprites en excluant le facteur d'orientation. . 98
4.3 Illustration de la corrélation linéaire entre l'information de position et son encodage respectif obtenu par le TC-NGVAE, entraîné sur Dsprites en excluant le facteur d'orientation. . . . . 99
4.4 Illustration de l'ambiguïté du facteur génératif d'orientation dans le jeu de données Dsprites. . . . . 100
4.5 Illustration d'une dépendance entre les facteurs de forme et d'échelle dans le jeu de données Dsprites. . . . . 100
4.6 Matrices de corrélation de Spearman mesurées entre les facteurs génératifs et les variables latentes, pour des modèles entraînés sur Dsprites. . . . . 107
4.7 Résultats obtenus par le TC-NGVAE entraîné sur Dsprites. . . . . 108
4.8 Matrices de corrélation de Spearman mesurées entre les facteurs génératifs et les variables latentes, pour des modèles entraînés sur Smallnorb. . . . . 110
4.9 Résultats obtenus par le TC-NGVAE entraîné sur Smallnorb. . . . . 111
4.10 Matrices de corrélation de Spearman mesurées entre les facteurs génératifs et les variables latentes, pour des modèles entraînés sur Cars3d. . . . . 113
4.11 Résultats obtenus par le TC-NGVAE entraîné sur Cars3d. . . . . 114
4.12 Illustration d'une dépendance entre le facteur génératif de catégorie et celui d'élévation dans Cars3d. . . . . 115
4.13 Illustration de l'ambiguïté du facteur génératif de condition d'éclairage dans Smallnorb. . . . . 115

E.1 Matrices de corrélation de Spearman mesurées entre les variables latentes et les facteurs génératifs pour des modèles entraînés sur Dsprites, en excluant le facteur d'orientation. . . . . 134
E.2 Matrices de corrélation de Spearman mesurées entre les variables latentes et les facteurs génératifs, pour des modèles entraînés sur Dsprites. . . . . 137
E.3 Matrices de covariance empirique mesurées sur l'espace latent, pour des modèles entraînés sur Dsprites. . . . . 138
E.4 Matrices de corrélation de Spearman mesurées entre les variables latentes et les facteurs génératifs, pour des modèles entraînés sur Smallnorb. . . . . 139
E.5 Matrices de covariance empirique mesurées sur l'espace latent, pour des modèles entraînés sur Smallnorb. . . . . 139
E.6 Matrices de corrélation de Spearman mesurées entre les variables latentes et les facteurs génératifs, pour des modèles entraînés sur Cars3d. . . . . 140
E.7 Matrices de covariance empirique mesurées sur l'espace latent, pour des modèles sur Cars3d. . . . . 141

# Liste des symboles

## Nombres et Tableaux

| | |
|---|---|
| $a$ | Scalaire (réel ou entier) |
| $\mathbf{a}$ | Vecteur |
| $A$ | Matrice |
| $I$ | Matrice identité |

## Ensembles

| | |
|---|---|
| $\mathbb{A}$ | Un ensemble |
| $\mathbb{R}$ | Ensemble des réels |
| $\{0, ..., n\}$ | Ensemble des nombres entiers entre 0 et $n$ |
| $[a, b]$ | Intervalle des réels incluant $a$ et $b$ |
| $(a, b]$ | Intervalle des réels incluant $a$ mais excluant $b$ |

## Indexing

| | |
|---|---|
| $i$ | Elément $i$ du vecteur $\boldsymbol{a}$, avec l'index commençant à 1 |
| $\boldsymbol{a}_{-i}$ | Tous les éléments du vecteur $\boldsymbol{a}$, excepté l'élément $i$ |
| $a_{i \neq j}$ | $i$-ème élément du vecteur $\boldsymbol{a}$ qui est différent de l'élément $a_j$ |
| $A_{ij}$ | Elément $i, j$ de la matrice $A$ |

## Calculs

| | |
|---|---|
| $\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}$ | Dérivée partielle de $\boldsymbol{y}$ par rapport à $\boldsymbol{x}$ |
| $\partial_{\boldsymbol{xy}}$ | Gradient de $\boldsymbol{y}$ par rapport à $\boldsymbol{x}$ |
| $\frac{\partial f}{\partial \boldsymbol{x}}$ | Matrice jacobienne $J \in \mathbb{R}^{mn}$ pour $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ |
| $\int f(\boldsymbol{x}) d\boldsymbol{x}$ | Intégrale définie sur le domaine de $\boldsymbol{x}$ |

## Théorie de l'information

| | |
|---|---|
| $P(a)$ | Distribution de probabilité sur une variable discrète |

| | |
|---|---|
| $p(a)$ | Distribution de probabilité sur une variable continue |
| $a \sim P$ | Variable aléatoire $a$ suivant la distribution $P$ |
| $\mathbb{E}_{\boldsymbol{x} \sim P}[f(\boldsymbol{x})]$ | Espérance de $f(\boldsymbol{x})$ par rapport à $P(\boldsymbol{x})$ |
| $Var(f(\boldsymbol{x}))$ | Variance de $f(\boldsymbol{x})$ selon $P(\boldsymbol{x})$ |
| $Cov(f(\boldsymbol{x}), g(\boldsymbol{x}))$ | Covariance de $f(\boldsymbol{x})$ et $g(\boldsymbol{x})$ selon $P(\boldsymbol{x})$ |
| $H(\boldsymbol{x})$ | Entropie de Shannon de la variable aléatoire $\boldsymbol{x}$ |
| $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma)$ | Distribution gaussienne sur $\boldsymbol{x}$ avec $\boldsymbol{\mu}$ pour moyenne et $\Sigma$ pour covariance |
| $\mathcal{B}(\boldsymbol{x}; \boldsymbol{p})$ | Distribution de Bernoulli sur $\boldsymbol{x}$ avec $\boldsymbol{p}$ pour paramètre de probabilité |
| $\mathcal{G}(\boldsymbol{x}; \alpha, \beta)$ | Distribution gamma sur $\boldsymbol{x}$ avec $\alpha$ pour paramètre de forme et $\beta$ pour paramètre d'intensité |
| $\mathcal{NG}(\boldsymbol{x}, \boldsymbol{y}; \mu, \alpha, \beta)$ | Distribution Normale-Gamma sur $\boldsymbol{x}$ et $\boldsymbol{y}$ avec $\mu$ pour moyenne, $\alpha$ pour paramètre de forme et $\beta$ pour paramètre d'intensité |

## Fonctions

| | |
|---|---|
| $f(\boldsymbol{x}; \theta)$ | Une fonction de $\boldsymbol{x}$ parametrisée par $\theta$ |
| $\log(\boldsymbol{x})$ | Logarithme néperien de $\boldsymbol{x}$ |
| $\|\mathbf{x}\|_\mathbf{p}$ | Norme $L^p$ de $\mathbf{x}$ |
| $\|\mathbf{x}\|$ | Norme 2 de $\mathbf{x}$ |

## Jeux de données et distributions

| | |
|---|---|
| $p_{\text{data}}$ | Distribution des données générant l'ensemble d'entraînement |
| $\mathbf{x}^{(\mathbf{i})}$ | Le $i$-ème example du jeu de données |
| $\mathbf{y}^{(\mathbf{i})}$ | La $i$-ème cible associée à $\mathbf{x}^{(\mathbf{i})}$ pour l'apprentissage supervisé |

# Liste des Acronymes

**CNN**

Réseau de neurones à convolutions ("Convolutional Neural Network"). 18, 19, 29, 31, 35

**DCI**

Désentrelacement, Compacité, Informativité. 54–57

**ELBO**

Borne basse de l'*evidence* ("Evidence Lower Bound"). 31, 33, 34, 41, 42, 50, 59, 118

**GAN**

Réseaux antagoniste génératifs ("Generative Adversarial Network"). 25, 26

**KL**

Kullback-Leibler. 30, 33, 42–45, 61, 75–77, 80–82, 88, 93, 119

**LMJ**

Laser Mégajoule. 11, 118, 120, 121

**MCMC**

Chaînes de Markov par méthodes de Monte Carlo ("Monte Carlo Markov Chain"). 24–26, 28–30

**MIG**

Écart d'information mutuelle ("Mutual Information Gap"). 56, 57

**NGVAE**

Normal-Gamma Variational Auto-Encoder. 10, 11, 61, 69, 71, 73–79, 83–90, 95, 96, 101, 119, 134

**RBM**

Machines de Boltzmann restreintes ("Restricted Boltzmann Machine"). 26, 28–30, 35

**TC-NGVAE**

Total-Correlation Normal-Gamma Variational Auto-Encoder. 3, 11, 89–97, 99, 101–105, 107, 109–114, 116, 117, 119–121, 134–138

**VAE**

Auto-encodeur variationnel ("Variational Auto-Encoder"). 8–10, 13, 24, 26, 29–31, 33–37, 40–42, 44, 52, 59–71, 74, 76, 82, 88–93, 118, 121, 131

# Introduction

Au sein du Commissariat à l'Énergie Atomique (CEA) du Centre d'Études Scientifiques et Techniques d'Aquitaine (CESTA), se trouve le Laser Mégajoule (LMJ), un très grand instrument de recherche qui vise à simuler les conditions de température et de pression présentes au cœur d'une étoile. Les expérimentations qui y sont menées consistent à focaliser des faisceaux lasers sur une cible, dans ce qu'on appelle la chambre d'expériences. Les données d'intérêt sont capturées par des diagnostics plasma qui varient en forme et en proximité avec la cible, à l'image de celui illustré dans la Figure 1. Leurs positions, définies par les spécifications de l'expérience, sont contrôlées par des agents humains. Toutefois, cette méthode présente des risques, notamment en termes de collisions potentielles entre les diagnostics. En effet, l'ajustement manuel peut entraîner des imprécisions et des erreurs d'alignement, augmentant ainsi la probabilité de contacts accidentels entre les instruments lors de leur insertion. Un premier algorithme permettant de prévenir ces collisions a été mis en place, mais son utilisation est limitée, car il ne permet pas l'insertion de deux diagnostics dans un même périmètre. Pour surmonter ces défis, une approche envisagée consiste en son automatisation partielle. Le principe est de s'appuyer sur des images acquises durant les expériences pour guider les opérateurs en contrôle en leur fournissant la position spécifique de chacun des diagnostics. De plus, dans ce contexte nucléarisé, la fiabilité de l'algorithme est primordiale. À cet effet, il est souhaitable d'associer l'information de position à une mesure d'incertitude relative aux calculs de l'algorithme. C'est dans ce contexte général que s'inscrit cette thèse.

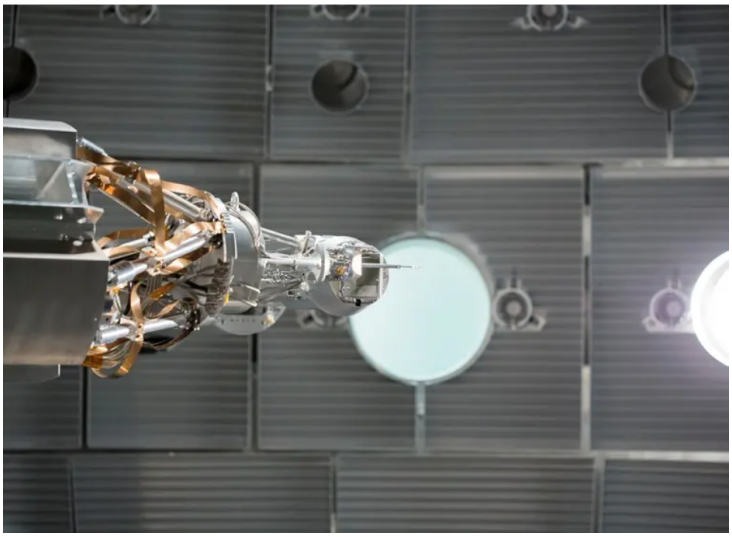


Fig 1: Diagnostic de mesure inséré au sein de la chambre d'expériences dans l'objectif de récupérer des données d'intérêt.

Les travaux de recherche exposés dans ce manuscrit se sont concentrés sur le premier besoin d'amélioration du système anti-collision, à savoir faire émerger les informations inhérentes à un objet présent dans l'image. Pour cela, des algorithmes reposant sur

des approches conventionnelles de traitement d'images ainsi que sur de l'apprentissage automatique peuvent être considérés. Ces derniers sont particulièrement adaptés à ce contexte, car ils permettent de découvrir des relations implicites dans les données. Leur mise en œuvre suppose d'apprendre une fonction paramétrique dépendant des données d'entrée par la minimisation d'une fonction coût. Dans le cadre de l'apprentissage supervisé, celle-ci est dépendante des données d'entrée et d'une étiquette qui leur est associée, représentant le résultat attendu. À l'inverse, dans l'apprentissage non supervisé, la fonction de coût dépend uniquement des données d'entrée. Notre recherche s'est particulièrement intéressée à ce dernier type d'apprentissage, car la construction d'une base de données étiquetées est coûteuse en termes de temps et de ressources. Parmi eux, les modèles probabilistes génératifs, dont l'objectif principal est d'apprendre la distribution sous-jacente des données d'entrées, représentent les meilleurs candidats. En effet, la méthode choisie doit pouvoir capturer des informations intrinsèques à une image. Elles peuvent résider dans des espaces à grandes dimensions, mais peuvent également être compressées dans des espaces de dimension réduite, facilitant ainsi la manipulation et l'interprétation des données encodées.

Dans le cadre de nos recherches, nous avons spécifiquement étudié les auto-encodeurs variationnels ("Variational Auto-Encoders") (VAEs), dont l'espace compressé des données est appelé espace latent. Contrairement à d'autres modèles de l'état de l'art aux propriétés similaires, la conception des VAEs permet de contraindre l'espace latent pour y extraire des informations pertinentes. En outre, ils offrent un cadre propice à l'évaluation d'incertitudes à deux niveaux distincts : au sein de l'espace latent, ainsi que dans l'espace des données via la modélisation d'une fonction de *vraisemblance* qui permet d'évaluer la capacité de génération de l'algorithme. En effet, ils sont composés d'un encodeur, qui infère les paramètres d'une distribution *a posteriori* pour le vecteur latent à partir des données d'entrée, et d'un décodeur, qui modélise la *vraisemblance* des données en fonction de ce vecteur latent. Après entraînement, celui-ci peut être mis à profit pour simuler des images présentant des propriétés similaires à celles de la base d'apprentissage. Dans le contexte du système anti-collision, on s'attend à ce qu'un VAE permette de faire émerger de façon non supervisée l'information de position d'un diagnostic, tout en facilitant une analyse des incertitudes.

Toutefois, le VAE, tel que proposé initialement par Kingma et Welling [Kingma et Welling, 2014], présente un espace latent difficilement interprétable, ce qui pose un défi pour notre contexte qui nécessite une représentation dans laquelle les variables sont compréhensibles. Pour surmonter cette difficulté, il est possible de régulariser les représentations obtenues au sein de l'espace latent pour faire émerger les propriétés attendues. L'étude de l'apprentissage de représentations est un champ de recherche particulièrement actif et complexe. Cette thèse, initialement axée sur l'examen de la fiabilité des informations encodées dans un espace latent, a évolué pour privilégier cet axe d'étude. Elle s'est tout d'abord attachée à proposer des analyses approfondies et un cadre théorique rigoureux pour traiter cette question. Divers types de représentations sont envisageables, mais un espace latent dit "désentrelacé" semble faire consensus au sein de la communauté scientifique.

Le désentrelacement repose sur l'hypothèse selon laquelle les données du monde réel sont générées à partir de facteurs génératifs indépendants, qui permettent de différencier les images les unes des autres. Un espace latent désentrelacé est donc un espace dans lequel les variables latentes sont indépendantes, et où la modification d'un facteur génératif dans les données se traduit par le changement d'une seule variable dans l'espace latent. Au

sein du système anti-collision, les images acquises peuvent être décrites à partir des facteurs comme la position du diagnostic, sa taille ou encore sa forme. L'encodage de ces facteurs physiquement interprétables se retrouve dans les variables latentes. La Figure 2a illustre un espace latent entrelacé, dans lequel la variable latente 1 est responsable de l'encodage de la taille des points, tandis que la variable latente 2 encode simultanément la couleur et la taille, illustrant une corrélation entre les deux variables. Les caractéristiques varient de manière interdépendante, ce qui souligne l'entrelacement de l'espace considéré. Une telle représentation pourrait rendre la manipulation des informations encodées moins aisée et restreindre leur interprétation. À l'inverse, la Figure 2b présente un espace latent désentrelacé, dans lequel les variations entre les variables latentes sont indépendantes : la variable latente 1 encode uniquement la taille, tandis que la variable 2 uniquement la couleur. Cette représentation facilite la manipulation des informations d'intérêt.

Actuellement, de nombreuses recherches se focalisent sur le développement de modèles basés sur un VAE qui possèdent ce type de propriété, constituant ainsi un domaine d'étude actif en apprentissage automatique.

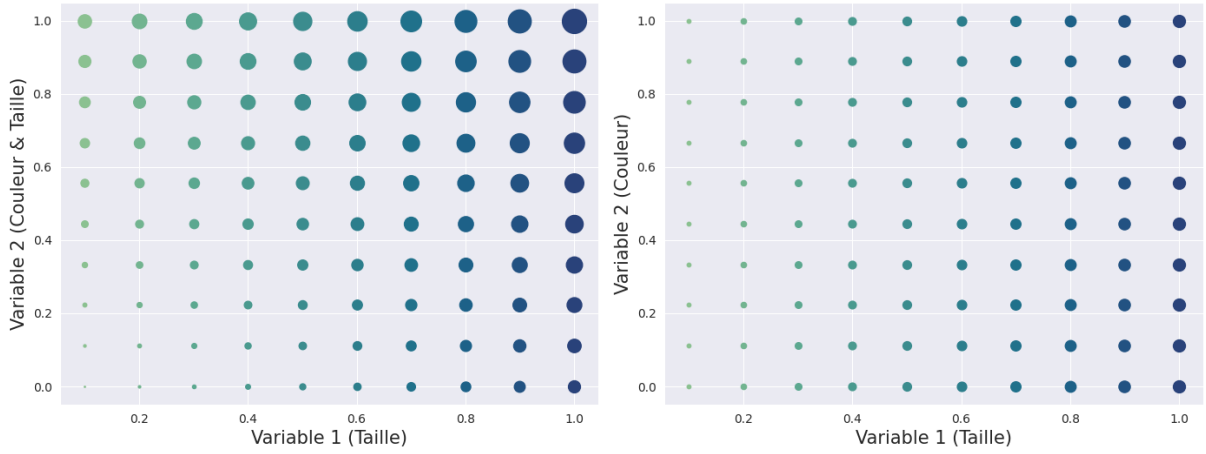



Fig 2: Illustrations d'espaces latents obtenus à partir d'images générées par deux facteurs génératifs : la taille d'un objet et sa couleur. (a) : Espace latent entrelacé. (b) : Espace latent désentrelacé.

Les approches de l'état de l'art présentent toutefois des limitations. Une contrainte majeure réside dans le compromis entre la qualité de la reconstruction des données en sortie du décodeur et les caractéristiques de l'espace latent. De plus, la détermination *a priori* de la dimension de l'espace latent est complexe et crucial pour la convergence vers un espace latent désentrelacé. Ce choix soulève deux défis majeurs. Lorsque le nombre de variables latentes est supérieur au nombre de facteurs génératifs, il est nécessaire que le modèle converge vers un espace latent dit polarisé, se traduisant par deux comportements distincts pour les variables. Un sous-ensemble de variables informatives encode l'ensemble des informations liées aux facteurs génératifs, tandis qu'un second sous-ensemble ne contient pas d'information sur les données d'entrées. Le premier objectif consiste à aligner le nombre de variables informatives sur le nombre de facteurs génératifs. Cette problématique reste peu abordée dans la littérature, ou est traitée de manière implicite, souvent au moyen de termes de régularisation pondérés par des coefficients définis de manière *ad hoc*. Le second objectif consiste à séparer les informations encodées dans les variables informatives, permettant l'obtention d'un espace désentrelacé.

Nos travaux ont ainsi visé à améliorer les limitations identifiées dans les approches de l'état de l'art, notamment en concentrant nos recherches sur les deux problématiques citées ci-dessus. Ce manuscrit constitue une synthèse de nos recherches, regroupant une analyse de l'état de l'art, la proposition de nouveaux algorithmes, ainsi que la présentation des résultats issus des expériences menées. Il s'articule de la façon suivante :

- Le premier chapitre établit les bases théoriques de l'apprentissage automatique. Cette section introductive contextualise les travaux menés et fournit un éclairage vis-à-vis des principes fondamentaux sur lesquels ils reposent. Un ensemble de modèles génératifs non supervisés est également présenté, permettant de souligner les avantages et inconvénients de chacun et justifiant l'intérêt que nous avons porté aux VAEs. La fin du chapitre est consacrée à exposer de façon détaillée les fondements théoriques de ce modèle et les méthodes mises en place pour son apprentissage.

- Le deuxième chapitre présente les diverses définitions du désentrelacement proposées dans la littérature. Les propriétés fondamentales de cette représentation y sont analysées et une revue de l'état de l'art des modèles basés sur un VAE visant à obtenir un espace latent désentrelacé est également présentée. Dans ce contexte, nous proposons une taxonomie de ces derniers, en nous appuyant sur les types de biais introduits par les auteurs pour favoriser le désentrelacement. Cette démarche permet d'identifier les avantages et inconvénients de chaque modèle. Un point notable réside dans le faible nombre de travaux portant sur la polarisation. Enfin, le chapitre se termine par une présentation des métriques utilisées pour évaluer la capacité de convergence des modèles vers un espace latent désentrelacé, ainsi que des jeux de données couramment utilisés par la communauté scientifique à des fins de validation. L'ensemble de ce chapitre fournit les justifications nécessaires à certains des choix de modélisation adoptés dans notre recherche, établissant ainsi un fondement méthodologique pour nos développements et expériences ultérieures.

- Le troisième chapitre du manuscrit se focalise sur la polarisation de l'espace latent. Cette propriété, qui représente une des problématiques cruciales dans l'objectif de désentrelacement, est dans un premier temps étudiée à travers des définitions et des hypothèses proposées, en complément d'une analyse empirique fondée sur les résultats d'un modèle existant. Une première contribution de nos travaux réside dans la formalisation théorique de ces comportements, ce qui enrichit notre compréhension du processus de désentrelacement dans les VAEs. Il est observé que très tôt dans la phase d'apprentissage, la distribution des paramètres inférés par l'encodeur pour la distribution *a posteriori* varie en fonction de l'information contenue dans la variable latente. Plus précisément, la variance des variables informatives se révèle être nettement inférieure à celle des variables non informatives. La seconde contribution est un nouvel algorithme tirant parti de ce constat, le Normal-Gamma Variational Auto-Encoder (NGVAE), dont certains résultats ont été présentés à la conférence EUSIPCO [Jouffroy *et al.*, 2022]. Ce modèle intègre une information *a priori* sur la distribution des variances, en utilisant un mélange de lois Inverse-Gamma, résultant en un modèle bayésien hiérarchique. Les paramètres de cette distribution dépendent de nouvelles variables issues de l'encodeur, destinées à évaluer la probabilité qu'une variable latente soit informative. Ces éléments sont conçus afin de favoriser la polarisation dans l'espace latent conformément aux théorèmes énoncés en début de chapitre. L'objectif est également de favoriser la convergence vers un

nombre de variables informatives similaires au nombre de facteurs génératifs. La dernière partie du chapitre présente une comparaison expérimentale du NGVAE avec deux autres modèles de l'état de l'art. Les résultats obtenus tendent à montrer que l'approche proposée répond aux propriétés de la polarisation, mais permet également d'encoder les facteurs génératifs dans le bon nombre de variables latentes, sans besoin d'hyperparamètres. Par ailleurs, le NGVAE montre une meilleure capacité à décorréler les variables latentes par rapport aux modèles comparés et semble moins sensible aux initialisations des paramètres pour converger vers un espace latent polarisé. Toutefois, les variables informatives relevées restent toujours indépendantes les unes aux autres.

- Le dernier chapitre du manuscrit traite de la seconde problématique liée au désentrelacement : induire cette caractéristique au sein des variables informatives d'un espace polarisé. Bien que le NGVAE converge vers un espace polarisé, il reste entrelacé. Pour surmonter cette limitation, nous introduisons des termes de régularisation supplémentaires à la fonction de coût originale, aboutissant à notre troisième contribution : un algorithme nommé Total-Correlation Normal-Gamma Variational Auto-Encoder (TC-NGVAE). Ces termes comprennent une mesure de corrélation totale dans l'espace latent, visant à favoriser l'indépendance statistique entre les variables. Un terme de groupe-lasso est également appliqué aux probabilités inférées par l'encodeur, forçant l'algorithme à encoder des facteurs génératifs similaires dans les mêmes dimensions de l'espace latent pour différentes images. Des expériences approfondies sont réalisées, permettant de comparer la capacité de désentrelacement du TC-NGVAE, avec cinq autres approches de l'état de l'art et sur trois jeux différents de données. L'analyse des résultats démontre que notre algorithme converge vers un espace latent plus désentrelacé que les autres modèles, particulièrement lorsque les facteurs génératifs des jeux de données sont indépendants les uns des autres. De plus, l'évaluation des différentes propriétés spécifiques au désentrelacement révèle que le TC-NGVAE est particulièrement efficace pour encoder un seul facteur génératif par variable latente, ce qui facilite grandement l'interprétation et la manipulation des représentations obtenues. Ce chapitre inclut également des analyses approfondies des jeux de données, qui révèlent des ambiguïtés ou des interdépendances entre plusieurs facteurs génératifs. Ces observations permettent d'expliquer certaines performances du TC-NGVAE qui sont similaires à l'état de l'art en termes de désentrelacement sur un jeu de données particulier. La soumission d'un article détaillant ces résultats au journal "IEEE Transactions on Neural Networks and Learning Systems" est en cours afin de valoriser nos travaux.

Nos recherches ont abouti au développement d'un algorithme d'apprentissage automatique non supervisé, permettant l'obtention de représentations compressées d'images d'entrées dont les propriétés favorisent l'interprétabilité et l'étude d'incertitudes. En effet, l'information de position est encodée linéairement et de façon distincte aux autres caractéristiques dans une seule variable latente. Cette contribution permet de répondre aux exigences initiales définies pour l'amélioration du système anti-collision du LMJ.
Dans une perspective plus large, nos travaux ont également contribué à une compréhension plus fine du phénomène de polarisation à travers la formalisation théorique de ses propriétés et la mise en lumière de son rôle crucial pour la convergence vers un espace désentrelacé. De plus, le TC-NGVAE se distingue des autres modèles dans sa meilleure capacité de désentrelacement. Finalement, l'analyse effectuée sur les jeux de données de l'état de l'art soulignent les interdépendances ou les ambiguïtés parmi les facteurs génératifs définis. Cela soulève des questions sur l'efficacité de certaines régularisations comme

la corrélation totale et pose la question de sa pertinence quant à l'utilisation de mesures supervisées afin d'évaluer le désentrelacement.

# Chapitre 1

# Apprentissage non supervisé de représentations

Les progrès croissants en puissance de calcul des dernières décennies ont permis le large déploiement de méthodes, parmi lesquelles figure l'apprentissage profond. À travers l'exploitation des architectures de réseaux de neurones profonds, cette branche de l'apprentissage automatique permet d'accéder à des représentations de haut niveau à partir de données non structurées, telles que des images, des textes ou des sons. Cette avancée majeure a engendré des progrès dans divers domaines, tels que la détection d'objets, la traduction automatique, la synthèse vocale, la recommandation de contenu ainsi que la représentation condensée de données à haute dimension, entre autres.

Dans ce chapitre, l'objectif est de présenter les fondements méthodologiques de l'apprentissage profond appliqué en particulier à la vision par ordinateur, qui se base sur l'analyse de données visuelles. En outre, cette section du manuscrit vise avant tout à offrir un éclairage particulier sur les principes fondamentaux de l'apprentissage profond sur lesquels reposent nos travaux, et ainsi de justifier certaines approches utilisées vis-à-vis de nos besoins. En effet, l'objectif principal de nos recherches consiste à développer un algorithme basé sur de l'apprentissage qui ne nécessite pas une labellisation préalable des données et qui permet d'obtenir une représentation de ces dernières dans un espace de dimension restreinte. Cette représentation devrait être interprétable afin de permettre une manipulation spécifique des variables encodées et notamment aboutir à une étude d'incertitudes en perspective de nos travaux. En reprenant les bases de l'apprentissage automatique dans ce chapitre, les défis particuliers liés à la manipulation de représentations latentes pourront ainsi être pleinement exploités dans le contexte de l'apprentissage automatique dans la suite du manuscrit.

Dans un premier temps, une introduction des concepts fondamentaux, incluant les réseaux neuronaux et les approches utilisées pour effectuer leur apprentissage, est réalisée. Par la suite, une explication des différents paradigmes de l'apprentissage profond est effectuée. Finalement, ce chapitre se conclue par le détail des modèles non supervisés pour la génération d'images, en mettant particulièrement en avant les auto-encodeurs variationnels ("Variational Auto-Encodeurs") (VAEs), qui constituent la base méthodologique de nos travaux. Toutefois, pour maintenir une cohérence et dans le souci de ne pas produire une revue exhaustive de l'ensemble des méthodes existantes, certains concepts que nous n'avons pas étudiés ne sont pas développés dans ce chapitre. Pour des informations plus approfondies, nous suggérons aux lecteurs de consulter les ouvrages proposés par Goodfellow et al. [Goodfellow *et al.*, 2016] et Bishop [Bishop, 2007].

## 1.1 Réseau de neurones et apprentissage profond

L'apprentissage fondé sur les réseaux de neurones trouve ses origines au milieu du vingtième siècle, en particulier dans les domaines de la psychologie et des neurosciences. Dans les années 1990, McCulloh et Pitts ont présenté une modélisation mathématique simplifiée d'un neurone biologique, dénommé "neurone formel" [McCulloch et Pitts, 1943]. Cette formalisation théorique, ayant pour objectif de créer une représentation schématique d'un élément constitutif du cerveau humain, permet de mieux comprendre certains fonctionnements sous-jacents des comportements cognitifs. Aujourd'hui, de nombreuses recherches et avancées, largement favorisées par les progrès des capacités de calcul, ont abouti à l'émergence de ce que nous appelons désormais l'apprentissage automatique profond de réseaux de neurones, basés sur des algorithmes composés d'une succession de milliers, voire de millions de neurones formels. Ces réseaux de neurones profonds possèdent la propriété générale d'être des approximateurs universels parcimonieux [Cybenko, 1989; Funahashi, 1989; Hornik *et al.*, 1989], suscitant ainsi un intérêt scientifique considérable.

Afin d'appréhender plus simplement les concepts complexes manipulés dans ce manuscrit de thèse, la suite du rapport vise à expliciter les fondements théoriques des réseaux neuronaux en détaillant le concept du neurone formel et son rôle au sein d'un réseau profond, la méthode d'apprentissage employée pour son optimisation, ainsi que les diverses bibliothèques logicielles adoptées par la communauté scientifique permettant de simplifier leur développement.

### 1.1.1 Du neurone formel au réseau de neurones artificiel

Le neurone formel, illustré dans la Figure 1.1[1], représente l'unité de calcul élémentaire au sein d'un réseau de neurones. Son objectif consiste à effectuer une transformation non linéaire à partir d'une donnée d'entrée à l'aide d'une fonction paramétrique.
Pour cela, il effectue une somme pondérée des données $\boldsymbol{x} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$ par un vecteur de paramètres $\boldsymbol{w} = (\boldsymbol{w}_1, ..., \boldsymbol{w}_n)$ appelé poids, caractérisant l'importance à attribuer à chaque entrée. Finalement, une fonction non linéaire sous-différentiable $\sigma(.)$, désignée sous le nom de fonction d'activation, est appliquée. La valeur résultante est considérée comme la sortie du neurone.

Mathématiquement, la sortie $a_j$ d'un neurone est calculée comme suit :

$$a_j = \sigma\left(\sum_{i=1}^{n} w_{j,i} \cdot x_i + b_j\right). \tag{1.1}$$

Dans cette formule, la sortie d'un neurone $a_j$ est désignée comme l'activation, et $b_j$ représente un terme variable appelé biais du neurone.

Comme mentionné précédemment, après avoir attribué des poids aux entrées, un neurone applique une fonction d'activation. L'introduction de cette dernière apporte la non-linéarité et la capacité à modéliser des relations complexes, éléments essentiels permettant l'apprentissage et la représentation efficace de relations non triviales dans les données. Les

[1] Certaines illustrations présentes dans ce manuscrit consistent en des modifications de figures proposées par [Neutelings, 2021] et [Stutz, 2020]. Les symboles $\star$ et $\diamond$ sont respectivement identifiés en légende des figures pour identifier leur provenance.

fonctions d'activation principalement utilisées regroupent les fonctions "Sigmoïde", "Softmax", "ReLu" ou encore "Tanh".

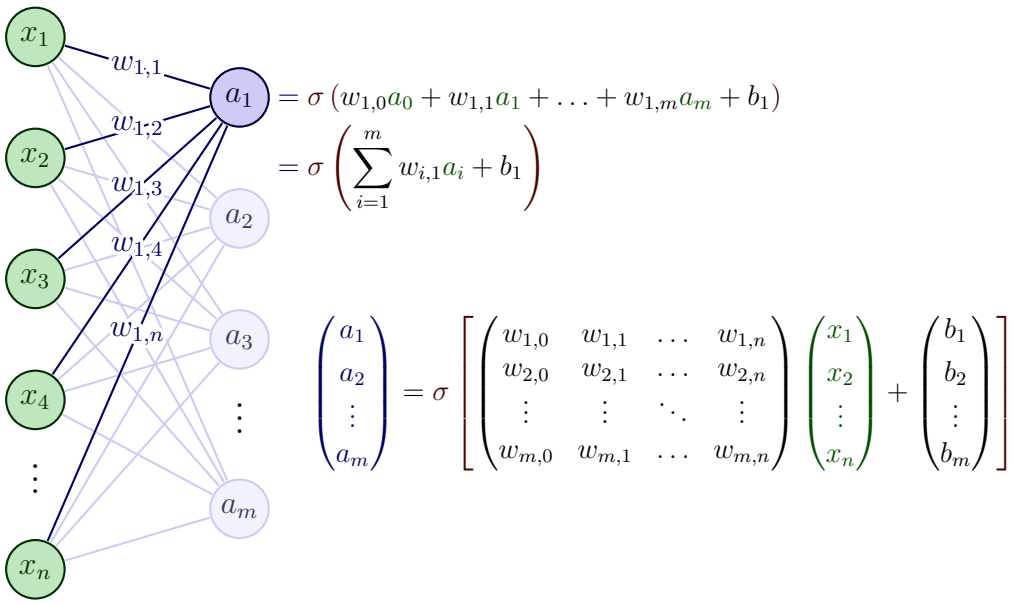



Fig 1.1: ($\star$) Illustration et formulation théorique de la sortie d'un neurone formel dénoté $a_1^{(1)}$, ainsi que la représentation matricielle du vecteur d'activations $\boldsymbol{a}^{(1)}$ des $m$ neurones de la première couche. En vertes sont représentées les activations de la couche 0, précédant le neurone d'intérêt, composée de $n$ neurones. La fonction d'activation quelconque est dénotée $\sigma(.)$, et les poids entre chaque neurone $w_{m,n}$.

Lorsque la tâche de classification est multiclasse, alors l'utilisation de la fonction Softmax est préférée, car elle opère sur un vecteur de scores en entrée pour générer une distribution de probabilités catégorielle. Cette dernière est définie par $f(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}}$, où $N$ représente le nombre de classes et $x_i$ le score attribué à la classe $i$. Par ailleurs, certaines de ces fonctions d'activation présentent des limitations, en particulier la fonction ReLu lorsque l'ensemble des sorties neuronales sont négatives. Par conséquent, diverses généralisations de cette fonction sont proposées. Par exemple, la fonction pReLu [He *et al.*, 2015] considère l'hyperparamètre $\alpha_i$ de la fonction ReLu comme une variable à apprendre par le réseau tandis que la fonction leaky ReLU [Maas *et al.*, 2013] fixe cette dernière à une petite valeur différente de 0.

Le concept de couche de neurones réunit un ensemble de neurones formels sous la forme d'un graphique acyclique orienté. On parle de couche d'entrée pour l'ensemble des neurones de la première couche, de couche de sortie pour les neurones de la dernière couche, et de couche cachée pour les couches intermédiaires. Un réseau de neurones profond, illustré dans la Figure 1.11, se distingue par la présence d'au moins deux couches cachées [Bengio *et al.*, 2009; Schmidhuber, 2015, 2007; Bengio et Delalleau, 2011]. Cette Figure illustre un réseau de neurones profond à trois couches cachées. Chaque neurone est lié à l'ensemble des neurones de la couche précédente et de la couche suivante par un paramètre de poids modélisé par un trait continu.

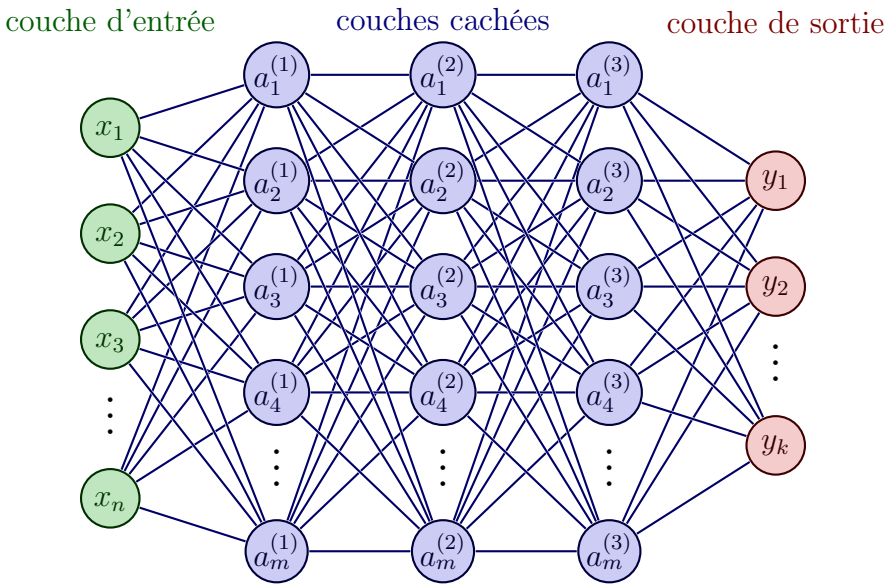



Fig 1.2: ($\star$) Illustration d'un réseau de neurones dense profond composé par trois couches cachées. En vert sont représentés les $n$ neurones de la couche d'entrée du réseau, en violet les $m$ neurones des trois couches cachées et en rouge les $k$ neurones de la couche de sortie.

### 1.1.2 Optimisation des paramètres

Une fois la structure du réseau définie, incluant le nombre de couches et de neurones par couche, l'optimisation de ses paramètres se fait lors de l'étape d'apprentissage. Cette dernière consiste à ajuster les paramètres du réseau dans l'objectif de minimiser une fonction coût ou erreur, qui dépend de la sortie $\boldsymbol{y}$ du réseau, des paramètres $\boldsymbol{\theta}$ et des entrées $\boldsymbol{x}$.
En d'autres termes, cela revient à trouver l'ensemble de paramètres $\boldsymbol{\theta} = (\boldsymbol{w}, \boldsymbol{b})$ pour lequel le gradient de la fonction d'erreur $\delta$ s'annule, tel que :

$$\nabla\delta(\boldsymbol{y}(\boldsymbol{x};\boldsymbol{\theta})) = 0. \tag{1.2}$$

En réalité, à mesure que le nombre de couches et de neurones dans le réseau croît, il devient souvent impossible de trouver une solution analytique pour résoudre l'équation (1.2). L'approche itérative des méthodes de descente de gradient est alors privilégiée. Elle consiste à ajuster les paramètres du réseau par le biais d'une série d'étapes visant à effectuer des modifications dans l'espace des poids et biais, en suivant la direction opposée du gradient. De cette façon, la mise à jour des poids est réalisée de la manière suivante :

$$\boldsymbol{\theta}^{\tau+1} = \boldsymbol{\theta}^{\tau} - \eta\nabla\delta(\boldsymbol{y}(\boldsymbol{x};\boldsymbol{\theta}^{\tau})). \tag{1.3}$$

Ici, $\tau$ représente l'itération en cours et $\eta > 0$ le coefficient d'apprentissage. Au cours de ce processus, l'étape consistant à calculer la sortie du réseau pour de nouvelles valeurs de paramètres $\boldsymbol{\theta}$, c'est-à-dire $\boldsymbol{y}(\boldsymbol{x};\boldsymbol{\theta})$, est appelée propagation avant. En revanche, la deuxième étape itérative qui implique le calcul de la dérivée de la fonction de coût par rapport aux nouveaux paramètres est appelée propagation arrière, et est illustrée dans la Figure 1.3.

La méthode la plus courante, l'algorithme de descente de gradient stochastique est qualifiée de "stochastique" car elle utilise un échantillon aléatoire des données, appelé "batch" ou lot de données pour estimer le gradient de la fonction de coût. Cela lui confère une rapidité supérieure par rapport à la descente de gradient classique qui nécessite l'utilisation de l'ensemble complet des données. Toutefois, cette approche peut introduire du bruit

dans les mises à jour des paramètres pouvant entraîner des fluctuations lors de la convergence de l'apprentissage. Pour atténuer ce phénomène, plusieurs variantes ont été proposées. C'est notamment le cas de Nesterov Momentum [Sutskever *et al.*, 2013] et Adam [Kingma et Ba, 2014].

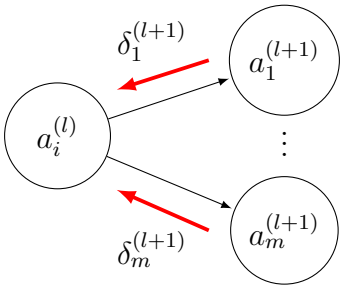



Fig 1.3: ($\diamond$) Illustration de la propagation arrière de l'erreur dans un neurone. Les flèches rouges représentent la propagation arrière de l'erreur $\delta$.

Par ailleurs, dans les réseaux profonds comportant un grand nombre de couches cachées, le gradient peut subir une diminution significative à mesure qu'il se propage. Par conséquent, les poids des couches profondes proches de l'entrée $\boldsymbol{x}$ peuvent recevoir des mises à jour très faibles, voire nulles, ce qui entraîne une stagnation de l'apprentissage. Pour résoudre ce problème d'évanouissement des gradients, différentes techniques sont disponibles, comme l'utilisation des fonctions d'activation à l'instar de ReLu, Leaky ReLu et tanH, qui ont démontré des performances satisfaisantes, ou encore l'ajout de liaisons résiduelles entre différentes couches de neurones qui sont notamment utilisées dans l'architecture du réseau "Residual Network" dit ResNet [He *et al.*, 2016].

En ce qui concerne les applications concrètes de ce projet, la fonction d'activation ReLu a été principalement employée en conjonction avec l'algorithme d'optimisation Adam.

### 1.1.3 Procédures d'apprentissage

Avant d'amorcer la description du processus d'apprentissage d'un réseau, il est important de définir certaines étapes cruciales pour garantir une bonne convergence. La section qui suit détaille les étapes successives de la procédure que nous avons adoptée avant d'initier l'entraînement des méthodes proposées.

La mise en place de chacune de ces étapes a été guidée par l'objectif de faciliter la convergence des modèles développés, mais également de conserver une certaine cohérence par rapport à d'autres approches de l'état de l'art. Cela permet de minimiser les biais lors de la comparaison des performances obtenues.

#### Collecte et préparation des données

La création d'un jeu de données approprié en vue de l'entraînement du réseau constitue une étape cruciale influençant considérablement les performances du modèle. Toutefois, la manipulation de représentations condensées des données obtenues par un modèle d'apprentissage automatique est un sujet largement étudié par la communauté scientifique. Ainsi, de nombreux jeux de données ont été créés dans l'objectif de comparer les performances obtenues. Nous effectuons une revue de ces derniers dans la section 2.4.5 du manuscrit. En choisissant de les mettre à profit pour effectuer l'apprentissage de nos

méthodes, nous n'avons pas eu à en créer de nouveaux, ni à remettre en œuvre les approches de l'état de l'art. Néanmoins, une étape significative a consisté à normaliser les images de ces jeux de données, de manière à les adapter à la plage de valeurs des fonctions d'activation, généralement située dans l'intervalle $[0, 1]$. Cette normalisation consiste à appliquer une transformation linéaire modifiant l'échelle des données sans changer leur distribution ou leur relation les unes par rapport aux autres. Elle vise notamment à prévenir d'éventuels problèmes liés à des écarts d'échelle importants entre les données.

Dans certains scénarios, il peut être nécessaire de diviser le jeu de données en trois ensembles distincts : l'ensemble d'entraînement, qui sert à ajuster les paramètres du modèle, l'ensemble de validation utilisé à ajuster les hyperparamètres du modèle, et finalement l'ensemble de test destiné à évaluer la capacité de généralisation du réseau une fois l'entraînement terminé sur des données non vues auparavant. Nous avons appliquée cette séparation des données pour le développement de certaines mesures s'appuyant sur la sortie d'un réseau de neurones dont l'objectif est de quantifier la capacité d'un modèle à obtenir une bonne représentation des données. Néanmoins, le cadre spécifique de l'entraînement des modèles proposés par nos approches n'a pas nécessité l'utilisation de ces trois sous-ensembles. En effet, les approches basées sur de l'apprentissage non supervisé, en particulier les méthodes génératives que nous détaillerons dans la suite de ce manuscrit, sont entraînées et testées sur le même ensemble de données. Par ailleurs, l'utilisation de ces sous-ensembles pourraient être envisagés en vue d'améliorer nos propositions, notamment pour une application spécifique à la mesure d'incertitudes.

**Définition de l'architecture d'un modèle**

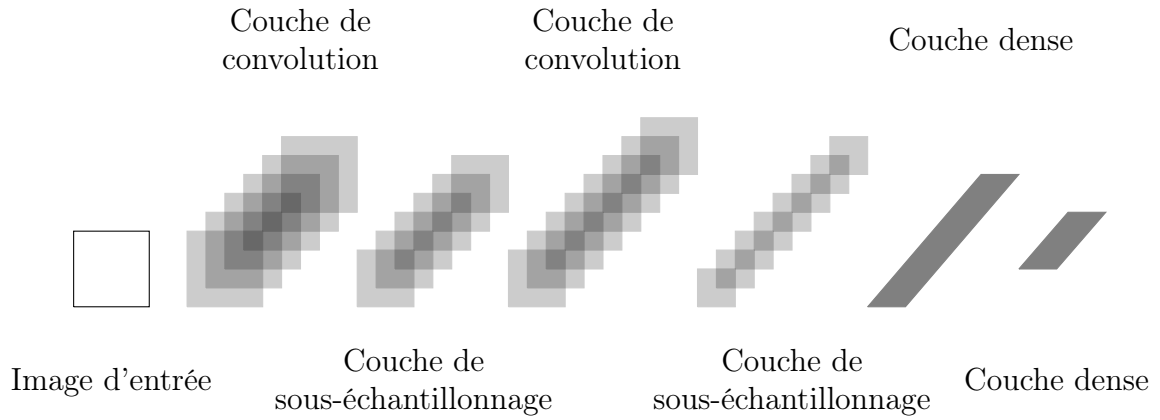



Fig 1.4: ($\diamond$) Illustration du réseau de convolution initialement introduit par Lecun et al. [LeCun *et al.*, 1989], composé d'une succession de couches de convolutions schématisées par une succession de carrés et couches denses schématisées par des rectangles pleins.

Une fois que le jeu de données a été créé et mis en forme, il est crucial de concevoir l'architecture du modèle de réseau de neurones à entraîner. Cela implique des choix tels que le nombre de couches cachées, les fonctions d'activation à utiliser, le nombre de neurones par couche et le type de couches à incorporer. Afin d'éviter d'introduire trop de biais lors de la comparaison des résultats obtenus par nos approches avec celles de l'état de l'art, l'architecture des réseaux mis en œuvre est principalement composée de réseaux spécifiques appelés Réseau de neurones à convolutions ("Convolutional Neural Network") (CNN). Ces derniers, pour lesquels un exemple est illustré dans la Figure 1.4, sont majoritairement composés de deux types majeurs de couches de neurones : les couches de

convolution et les couches denses.

Une couche de neurones dense établit une connexion entre chaque neurone de la couche précédente avec chaque neurone de la couche suivante, comme détaillé précédemment et illustré dans la Figure 1.1. Les couches de convolution, quant à elles, ont une fonction spécifique : elles sont conçues pour détecter des motifs locaux dans les images, tels que des contours, des textures et des formes. Ces couches fonctionnent en utilisant des neurones qui agissent comme des fenêtres de convolution, parcourant l'ensemble de l'image et appliquant des filtres discrets, également appelés noyaux de convolution. Ces filtres sont essentiellement des matrices de paramètres qui se déplacent sur l'entrée pour calculer des valeurs caractéristiques, tel qu'illustré dans la Figure 1.5.

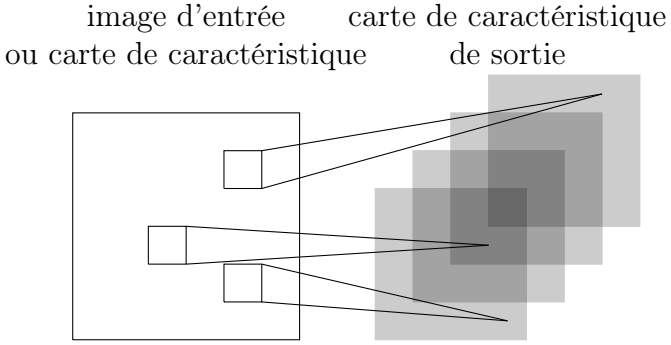



Fig 1.5: ($\diamond$) Illustration d'une couche de convolution. Le carré blanc représente l'image d'entrée de la couche. À l'intérieur de celle-ci, les trois carrés de taille inférieure représentent différents filtres de convolution qui vont chacun glisser sur l'image pour ressortir les cartes de caractéristiques représentées en gris.

Ainsi, les CNNs possèdent les propriétés adéquates pour construire et manipuler des représentations de données. De ce fait, les modèles présentés dans la section suivante, ainsi que les propositions d'amélioration décrites dans ce manuscrit, se reposent sur de telles architectures. Il est tout aussi crucial de définir soigneusement la fonction de coût, qui permettra d'ajuster les paramètres du modèle tout au long de l'entraînement. En effet, l'apport de nos recherches réside principalement dans la définition d'une erreur à nos objectifs.

### Initialisation des paramètres d'un modèle

Une fois le modèle défini, une étape critique consiste à initialiser l'ensemble des paramètres à apprendre du réseau, à savoir les poids et les biais des neurones [Glorot et Bengio, 2010]. Dans cet objectif, de nombreuses heuristiques basées sur des distributions de probabilité ont été développées. Les approches les plus couramment utilisées sont l'initialisation de Xavier [Glorot et Bengio, 2010], également appelée initialisation Glorot, et l'initialisation de He [He *et al.*, 2015].

Dans le cadre des modèles considérés, l'initialisation de Xavier Uniforme a été utilisée. Cette dernière est définie de la manière suivante :

$$\boldsymbol{\theta} \sim \mathcal{U}\left(-\sqrt{\frac{6}{n_{\text{in}} + n_{\text{out}}}}, +\sqrt{\frac{6}{n_{\text{in}} + n_{\text{out}}}}\right), \tag{1.4}$$

où, $n_{\text{in}}$ représente le nombre de neurones dans la couche précédente, $n_{\text{out}}$ est le nombre de neurones dans la couche suivante et $\mathcal{U}(\cdot, \cdot)$ la distribution Uniforme. En effet, plusieurs tests nous ont permis de vérifier de façon empirique qu'une meilleure convergence était obtenue avec cette initialisation.

**Entraînement d'un réseau de neurones**

Une fois les précédentes étapes menées à bien, la phase d'apprentissage visant à minimiser la fonction de coût à l'aide d'un algorithme d'optimisation basé sur la descente de gradient peut être entamée. Différents critères peuvent être employés pour déterminer quand mettre fin au processus d'entraînement. En effet, celui-ci peut être stoppé après un nombre préétabli d'itérations ou d'épochs (passages complets à travers l'ensemble de données), lorsque la diminution de la fonction de coût cesse d'être significative, ou encore lorsque le temps d'exécution maximal est atteint. Une combinaison de plusieurs critères peut également être utilisée pour déterminer le moment optimal.

Par ailleurs, le critère d'arrêt considéré par les méthodes avec lesquelles nous avons effectué des comparaisons de performances, comme explicité dans la section 2.3 du manuscrit, repose sur un nombre maximal d'épochs, défini de façon empirique. Ce critère ne fait pas sens vis-à-vis de nos algorithmes qui nécessitent davantage de temps pour converger. De ce fait, nous avons décidé de nous fonder sur la stabilité de la fonction coût pour arrêter l'apprentissage.

**Réglage des hyperparamètres**

Contrairement aux paramètres appris par le réseau, les hyperparamètres sont propres à la modélisation adoptée. Ces derniers sont généralement définis empiriquement lors de la première phase d'apprentissage, mais ne garantissent pas nécessairement la convergence du modèle vers une solution optimale. Ainsi, une étape préliminaire à l'entraînement d'un réseau consiste à trouver leur valeur optimale dans une pratique appelée hyperparamétrage.

La méthode traditionnelle d'hyperparamétrage sur laquelle nous nous sommes reposés consiste en une "recherche en grille", pour laquelle différentes combinaisons prédéfinies de valeurs sont évaluées dans l'objectif de déterminer la configuration optimale. Une alternative moins coûteuse consiste à effectuer une recherche aléatoire parmi les combinaisons d'hyperparamètres [Bergstra et Bengio, 2012], ce qui réduit le coût computationnel associé aux nombreuses possibilités.

### 1.1.4 Infrastructures logicielles

Plusieurs infrastructures logicielles populaires sont disponibles pour l'apprentissage automatique profond, offrant des bibliothèques qui simplifient le développement, l'entraînement et le déploiement de modèles de réseaux de neurones. Aujourd'hui, les plus utilisées sont **PyTorch** [Paszke *et al.*, 2019], développée par Facebook, **TensorFlow** [Abadi *et al.*, 2016], créée par Google, et plus récemment **Jax** [Bradbury *et al.*, 2021] également proposée par Google. L'interface de programmation d'application de haut niveau **Keras** [Chollet, 2023], compatible depuis peu avec ces trois librairies, permet une programmation simple, efficace et modulaire des réseaux de neurones profonds.

Dans le cadre de cette thèse, les librairies TensorFlow et Keras ont été utilisées.

## 1.2 Problématiques d'apprentissage

En apprentissage automatique profond, différentes approches d'apprentissage jouent un rôle essentiel en fonction du type de données disponibles et de la tâche à accomplir. Trois

grandes approches se démarquent, notamment l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage par renforcement. Plus récemment, les approches émergentes d'apprentissage semi-supervisé et d'apprentissage actif, des catégories à la frontière entre l'apprentissage supervisé et non supervisé, sont également considérées.

Dans la suite de ce chapitre, nous présentons brièvement le cadre de l'apprentissage supervisé, en mettant en avant ses deux principales finalités : la classification et la régression. Cette approche supervisée a notamment été utilisée dans nos travaux pour la mise en œuvre de mesures permettant la comparaison des performances obtenues par différents modèles, décrites dans la section 2.4. Ensuite, nous abordons l'approche de l'apprentissage non supervisé, suivi d'une exploration plus approfondie des modèles génératifs. Ces architectures constituent en effet la base principale de nos travaux.

### 1.2.1 Notations

Dans la suite de ce chapitre, nous notons $\boldsymbol{\theta} = \{\boldsymbol{w}, \boldsymbol{b}\}$ l'ensemble des paramètres du réseau, $f_{\boldsymbol{\theta}}(\cdot)$ la transformation appliquée par ce réseau et $\boldsymbol{y}(\boldsymbol{x}; \boldsymbol{\theta})$ sa sortie. Si, pour des données $\boldsymbol{x} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_N\}$, un ensemble de cibles $\boldsymbol{t} \in \mathbb{R}^N$ est connu, alors on considère un ensemble d'entraînement $D = \{(\boldsymbol{x}_1, t_1), (\boldsymbol{x}_2, t_2), ..., (\boldsymbol{x}_n, t_n)\}$ avec $t_i$ la cible associée à $\boldsymbol{x}_i$.
Lorsque aucune cible n'est associée aux données, on considère un second ensemble de données $D^* = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$ avec $\boldsymbol{x} \in \mathbb{R}^N$.

### 1.2.2 Apprentissage supervisé

L'apprentissage supervisé consiste à entraîner un modèle à partir d'un ensemble de données étiquetées $D$, c'est-à-dire un ensemble de données dans lequel chaque exemple est associé à une étiquette ou à une réponse connue. L'objectif de cette approche est d'apprendre à partir de ces exemples étiquetés afin de pouvoir faire des prédictions ou des classifications sur de nouvelles données non étiquetées. En d'autres termes, l'objectif de l'apprentissage supervisé consiste à apprendre $\boldsymbol{\theta}$ de manière à ce que $f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) = t_i$.

Les principales approches de l'apprentissage supervisé sont la classification et la régression, selon la nature de $t_i$.

#### Classification

La classification consiste à séparer les données en groupes prédéfinis en fonction de leurs caractéristiques. Un groupe est identifié par une étiquette de classe. L'objectif de cette approche est de créer un modèle capable d'associer automatiquement une donnée à son étiquette correspondante.
Lorsqu'il n'y a que deux étiquettes de classe possibles (c'est-à-dire $t = [-1, 1]$), on parle de classification binaire. En revanche, lorsque davantage de classes sont possibles, on parle de classification multiclasse. Dans le cadre de cette dernière, la couche de sortie du réseau de neurones utilisé est équipée d'une fonction d'activation Softmax, composée de $K$ neurones correspondant au nombre d'étiquettes de classes possibles, définie par :

$$P(y(\boldsymbol{x}; \boldsymbol{\theta}) = k) = \frac{e^{a_k}}{\sum_{i=1}^{K} e^{a_i}}, \tag{1.5}$$

où $a_k$ est l'activation associée au neurone de sortie pour la classe $k$, $\boldsymbol{y}(\boldsymbol{x}; \boldsymbol{\theta})$ la sortie du réseau de neurones, et $P(y(\boldsymbol{x}; \boldsymbol{\theta}) = k)$ la probabilité que l'entrée $\boldsymbol{x}$ appartienne à la classe $k$. Le modèle prédit alors que l'entrée appartient à la classe ayant la probabilité la plus élevée :

$$\hat{y}(\boldsymbol{x};\boldsymbol{\theta}) = \underset{k\in[0,K-1]}{\operatorname{argmax}} P\big(y(\boldsymbol{x};\boldsymbol{\theta}) = k\big). \tag{1.6}$$

De manière à apprendre les paramètres optimaux $\boldsymbol{\theta}^*$ permettant d'obtenir $\hat{y}(\boldsymbol{x}_i,\boldsymbol{\theta}) = t_i$, la fonction de coût à optimiser à l'aide de l'algorithme de descente de gradient est généralement la fonction d'entropie croisée, bien que d'autres fonctions coût puissent être considérées. Cette dernière, à minimiser, est définie par :

$$\begin{aligned} H(\boldsymbol{t},\boldsymbol{y}(\boldsymbol{x};\boldsymbol{\theta})) &= H(\boldsymbol{t}) + D_{KL}\big[\boldsymbol{t}||\boldsymbol{y}(\boldsymbol{x};\boldsymbol{\theta})\big] \\ &= -\sum_{i=1}^{K} t_i \log\big(P(y(\boldsymbol{x};\boldsymbol{\theta}) = i)\big). \end{aligned} \tag{1.7}$$

L'optimisation de cette fonction permet de réduire la différence de distribution entre les probabilités prédites pour la classe et la probabilité réelle de cette classe.

#### Régression

Contrairement à la classification où l'étiquette cible est une variable catégorielle, la régression vise à prédire une valeur continue, ce qui permet de compléter les valeurs manquantes d'un signal ou d'une image. De cette façon, contrairement à la classification multiclasse, un modèle de régression possède généralement un unique neurone en sortie. La fonction de coût couramment optimisée correspond à l'Erreur Quadratique Moyenne (EQM) entre la valeur cible et la prédiction du réseau, telle que :

$$EQM(\boldsymbol{t},\boldsymbol{y}(\boldsymbol{x};\boldsymbol{\theta})) = \frac{1}{N}\sum_{i=1}^{N}\big(t_i - y_i(\boldsymbol{x};\boldsymbol{\theta})\big)^2, \tag{1.8}$$

où $N$ représente le nombre d'images dans le jeu de données d'entraînement.

Bien que les approches d'apprentissage supervisées ne soient pas le principal sujet de cette thèse, des méthodes de classification et de régression ont été utilisées dans la section 2.4, notamment pour développer des mesures permettant d'évaluer et de comparer les performances des modèles proposés par rapport à ceux de l'état de l'art.

### 1.2.3 Apprentissage

Dans l'approche de l'apprentissage non supervisé, l'optimisation du modèle s'effectue sur un ensemble pour lequel aucune étiquette n'est associée aux données, $D^*$. L'objectif principal de ces méthodes est d'apprendre la loi de probabilité $p_{\text{data}}$ à partir de laquelle les données sont générées. Les approches majeures de l'apprentissage non supervisé sont représentées par les modèles de regroupement, la réduction de dimension ainsi que les modèles génératifs, sur lesquels nous nous sommes principalement appuyés et que nous décrivons plus en détail dans la suite de ce manuscrit.

#### Méthodes génératives neuronales probabilistes

Les méthodes génératives sont des approches probabilistes. On notera par la suite $p_{\text{data}}(\boldsymbol{x})$ la densité de probabilité des données, *i.e* l'*évidence* réelle, et $p_{\text{model}}(\boldsymbol{x};\boldsymbol{\theta})$ la densité de probabilité des données inférées par le modèle, l'*evidence* approchée.

L'idée principale de la modélisation générative est de construire une représentation $p_{model}(\boldsymbol{x};\boldsymbol{\theta})$ d'une distribution de probabilité complexe et souvent multimodale $p_{data}(\boldsymbol{x})$ avec $\boldsymbol{x} \in \mathbb{R}^N$,

où $N$ est souvent de grande dimension. L'objectif de ces approches consiste à apprendre les paramètres $\boldsymbol{\theta}$ d'un générateur $g$ formalisé sous la forme d'un réseau de neurones, de manière à ce qu'il puisse générer de nouvelles données qui ressemblent à celles fournies dans la base d'apprentissage, et ce, tout en suivant la distribution $p_{data}(\boldsymbol{x})$, tel que :

$$g_{\boldsymbol{\theta}} : \mathbb{R}^K \to \mathbb{R}^N, \quad \boldsymbol{z} \mapsto g_{\boldsymbol{\theta}}(\boldsymbol{z}). \tag{1.9}$$

En d'autres termes, pour chaque échantillon $\boldsymbol{x} \sim p_{data}(\boldsymbol{x})$, l'hypothèse émise est qu'il existe un vecteur $\boldsymbol{z}$, échantillonné depuis une distribution $p(\boldsymbol{z})$, tel que l'application du générateur $g_{\boldsymbol{\theta}}$ à ce vecteur produise une approximation de $\boldsymbol{x}$, c'est-à-dire $g_{\boldsymbol{\theta}}(\boldsymbol{z}) \approx \boldsymbol{x}$.

Étant donné que le vecteur $\boldsymbol{z}$ est généralement inconnu, on le désigne souvent comme la variable latente, et l'espace $\mathcal{Z}$ est appelé espace latent. Il est important de noter que la dimension de l'espace latent, notée $K$, peut différer de la dimension de l'espace des données et la plupart du temps $K < N$. Cette différence de dimensions permet au générateur de capturer des variations complexes et de générer des échantillons diversifiés dans l'espace des données.

En supposant que le générateur g soit connu, il est possible de générer de nouveaux exemples de données en échantillonnant $\boldsymbol{z} \sim p_{\boldsymbol{\theta}}(z)$ et en calculant $g_{\boldsymbol{\theta}}(\boldsymbol{z})$. De plus, le générateur peut également être utilisé pour calculer la *vraisemblance* ou l'*evidence* d'un échantillon particulier $\boldsymbol{x}$ par marginalisation :

$$p_{\text{model}}(\boldsymbol{x}; \boldsymbol{\theta}) = \int p(\boldsymbol{x}|\boldsymbol{z}) p_{\boldsymbol{\theta}}(\boldsymbol{z}) d\boldsymbol{z}, \tag{1.10}$$

où la distribution de *vraisemblance* $p(\boldsymbol{x}|\boldsymbol{z})$ mesure à quel point un vecteur en sortie de $g_{\boldsymbol{\theta}}(\boldsymbol{z})$ est proche de $\boldsymbol{x}$.

Ces méthodes génératives probabilistes possèdent des propriétés particulièrement intéressantes vis-à-vis de notre sujet de recherche. Dans un premier temps, elles s'appuient sur l'approche d'apprentissage non supervisé, permettant d'éviter la tâche de labellisation d'images, qui peut être relativement fastidieuse. De plus, elles s'appuient sur une représentation des données qui est probabiliste, ce qui permet d'obtenir des informations supplémentaires quant à la qualité de la représentation obtenue.

Une taxonomie proposée par Goodfellow, illustrée dans la Figure 2.5, propose une distinction des méthodes génératives basées sur la maximisation de la distribution de *vraisemblance marginale* également appelée *evidence*, à paramètres $\boldsymbol{\theta}$ fixés [Goodfellow, 2016]. Les premières approches considérées sont celles pour lesquelles la distribution d'*evidence* est dite "implicite", c'est-à-dire que l'apprentissage du modèle ne nécessite pas d'y avoir accès. Dans le second type de méthodes, l'*evidence* est à l'inverse considérée comme "explicite", auquel cas la fonction coût à minimiser dépend de cette distribution. Parmi ces dernières, une décomposition est à nouveau proposée, où l'on diffère les approches pour lesquelles l'expression réelle de l'*evidence* peut être calculée, et les approches pour lesquelles on présuppose un modèle sur sa distribution. On parle alors respectivement de forme "analytique" de l'*evidence*, ou de forme "approchée".
Historiquement, la génération de nouvelles données est l'objectif principal des modèles génératifs. On peut citer en exemple les modèles faisant l'actualité comme Dalle-E 2 [Ramesh *et al.*, 2022] ou encore ChatGPT [OpenAI]. Plus récemment, la manipulation de l'espace latent obtenu par ces méthodes est devenue un sujet de recherche à part entière. En effet, les représentations des données apprises par les modèles non supervisés conditionnent la capacité de génération du réseau, mais elles permettent également une

exploitation de ces représentations pour des tâches annexes telles que la détection de données hors distribution [An et Cho, 2015] ou encore la mesure d'incertitudes [Postels *et al.*, 2020; Alemi *et al.*, 2018] .

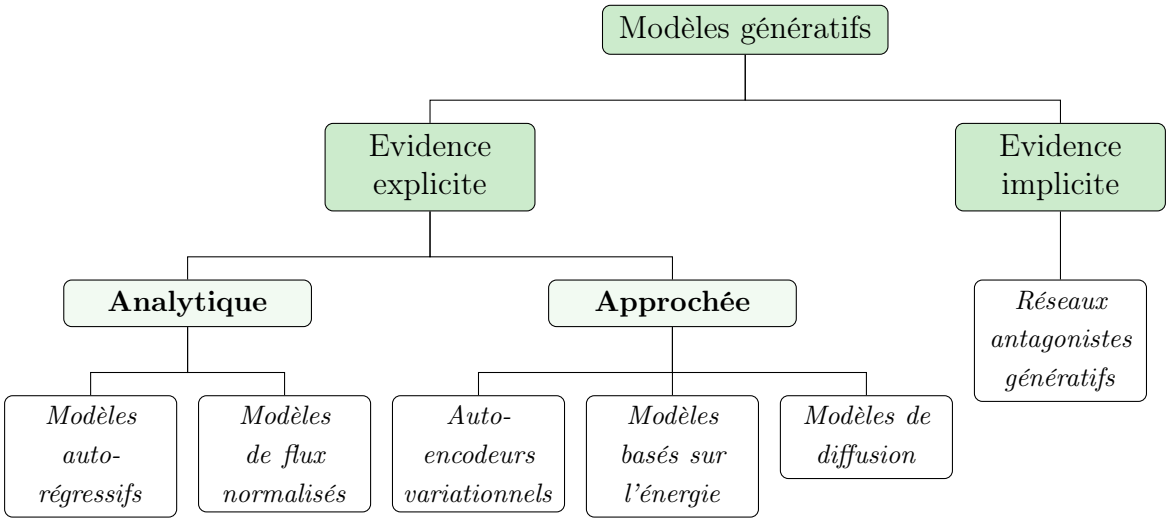



Fig 1.6: Taxonomie des modèles génératifs non supervisés proposée par [Goodfellow, 2016].

Les sections suivantes proposent une analyse de certaines de ces méthodes, notamment les modèles basés sur l'énergie, les modèles de flux normalisés et les réseaux antagonistes génératifs. Celle-ci permet de justifier en quoi les approches qui modélisent l'*evidence* de façon implicite, tout comme celles dans laquelle elle est définie de façon explicite, mais calculée de façon analytique, ne possèdent pas les caractéristiques permettant de traiter de façon efficace notre sujet de recherche. Finalement, les VAEs sont détaillés plus en profondeur.

## 1.3 Architectures génératives

Comme indiqué précédemment, le rôle principal des modèles génératifs probabiliste est d'estimer la distribution réelle des données $p_{\text{data}}(\boldsymbol{x})$. Cette approche implique la recherche des paramètres du modèle, dénotés $\hat{\boldsymbol{\theta}}$, qui maximisent l'*evidence* des données d'entraînement, sous l'hypothèse d'une loi $p_{\text{model}}(\boldsymbol{x};\boldsymbol{\theta})$ tel que

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i=1}^{N} p_{\text{model}}\big(\boldsymbol{x}^{(i)};\boldsymbol{\theta}\big). \tag{1.11}$$

Ces méthodes sont qualifiées de modèles à densité explicite lorsque leur modélisation permet de définir explicitement $p_{\text{model}}(\boldsymbol{x}^{(i)};\boldsymbol{\theta})$. À l'inverse, les méthodes qualifiées de modèles à densité implicite sont entraînées sans nécessiter la définition explicite de cette distribution.

### 1.3.1 Modèles implicites

Les modèles implicites ont la capacité d'être entraînés uniquement à partir de la distribution des données réelles $p_{\text{data}}(\boldsymbol{x})$ et pour lesquels la distribution d'*evidence* $p_{model}(\boldsymbol{x};\boldsymbol{\theta})$ n'est pas utilisée lors de l'apprentissage. Certaines de ces approches définissent des opérateurs de transition en utilisant des méthodes de Chaînes de Markov par méthodes de

Monte Carlo ("Monte Carlo Markov Chain") (MCMC). Elles consistent à générer une suite de données $\{\boldsymbol{x}_j\}_{j\gg 1}$ qui, après un temps de chauffe, sont distribuées selon la loi souhaitée. C'est notamment le cas des réseaux génératifs stochastiques [Bengio *et al.*, 2014]. Néanmoins, ces derniers ne sont pas toujours applicables à des espaces de données à haute dimension, ce qui en fait des méthodes peu utilisées en vision par ordinateur. Les Réseaux antagoniste génératifs ("Generative Adversarial Network") (GAN) [Goodfellow *et al.*, 2014] ont été proposés dans l'objectif de palier cette limitation.

### Réseaux antagonistes génératifs

L'objectif initial des GANs [Goodfellow *et al.*, 2014] est de générer des données $\boldsymbol{x}$ suivant la loi de $p_{data}$ à partir d'un échantillon latent $\boldsymbol{z} \sim p(\boldsymbol{z})$, où $p(\boldsymbol{z})$ est généralement définie comme une distribution normale centrée réduite.

Pour atteindre cela, l'architecture d'un GAN est composée de deux réseaux de neurones profonds : un générateur $G$ et un discriminateur $D$. Le générateur prend en entrée $\boldsymbol{z}$ et apprend à transformer ce vecteur latent en un échantillon qui ressemble aux données d'entraînement. Le discriminateur, quant à lui, prend en entrée à la fois des échantillons réels provenant de $p_{\text{data}}$ et des échantillons générés par le générateur $G(\boldsymbol{z})$. Son rôle est de distinguer les données réelles des échantillons générés. Une schématisation du modèle est proposée dans la Figure 1.7.

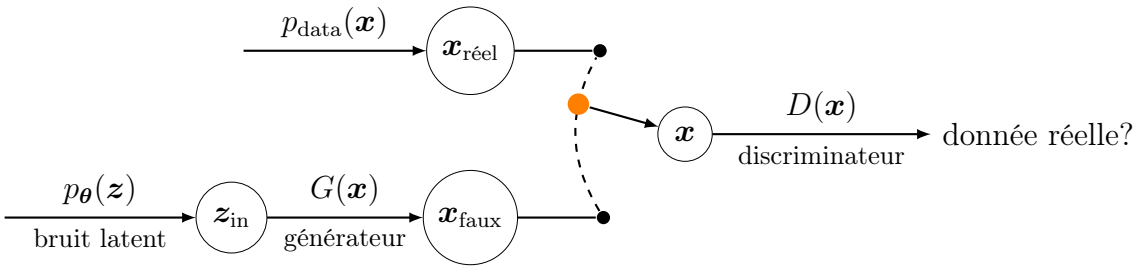



Fig 1.7: ($\star$) Illustration du GAN.

Le processus d'entraînement de ce modèle repose sur un jeu d'optimisation entre le générateur et le discriminateur. Le générateur s'efforce de simuler des échantillons de plus en plus réalistes pour tromper le discriminateur. La fonction de coût est définie comme suit :

$$G_{loss} = -\log\big(D(G(\boldsymbol{z}))\big). \tag{1.12}$$

Dans le même temps, le réseau du discriminateur cherche à améliorer sa précision pour différencier les données réelles des échantillons générés. Son objectif est alors de minimiser la fonction suivante :

$$D_{loss} = \log\big(D(\boldsymbol{x})\big) + \log\big(1 - D(G(\boldsymbol{z}))\big). \tag{1.13}$$

La fonction de coût finale consiste ainsi en un processus d'optimisation de type min-max entre les fonctions de coût du discriminateur et du générateur, formulée par :

$$\mathcal{L}_{gan} = \min_G \max_D \mathbb{E}_{p_{\text{data}}(\boldsymbol{x})}\big[\log D(\boldsymbol{x})\big] + \mathbb{E}_{p_z(\boldsymbol{z})}\big[\log\left(1 - D(G(\boldsymbol{z}))\right)\big]. \tag{1.14}$$

Depuis son apparition, le succès du GAN dans sa capacité à approcher la distribution des données réelles et ainsi à générer des échantillons très réalistes a suscité une abondante littérature sur des variantes, qui reposent chacune sur des critères d'optimisation

différents. Certains de ces travaux ont pour objectif d'aborder des problèmes d'instabilité souvent rencontrés lors de l'apprentissage du GAN. En effet, il est courant que le modèle n'apprenne qu'une sous-partie de la distribution des données d'entrées et reste coincé dans un espace limité. Ce problème, nommé mode d'effondrement, implique que le générateur produit toujours les mêmes sorties. Par ailleurs, la complexité de l'entraînement liée à l'optimisation de type min-max entraîne parfois un processus d'apprentissage lent et instable durant lequel les paramètres du modèle oscillent et présentent des problèmes de convergence. En outre, lorsque le réseau de discrimination arrive à discriminer les données du générateur, on observe une disparition du gradient, ralentissant fortement l'apprentissage. Parmi les approches visant à atténuer ces problèmes figurent les MMD-GANs [Li *et al.*, 2017], les f-GANs [Nowozin *et al.*, 2016] et les Wasserstein-GANs [Arjovsky *et al.*, 2017].

La plupart des travaux sur l'interprétation de l'espace latent des GANs se sont appuyées sur des approches non supervisées. C'est notamment le cas des modèles StyleGAN [Karras *et al.*, 2019] et InfoGAN [Chen *et al.*, 2016]. Cependant, à l'instar des méthodes génératives à densité implicite, l'espace latent du GAN n'est pas issu d'un encodage des données d'entrée et n'est donc pas interprétable. De cette façon, les caractéristiques de ces approches ne semblent pas appropriées vis-à-vis de nos sujets de recherche.

### 1.3.2 Modèles explicites

Les approches génératives explicites définissent directement la densité de probabilité sous forme paramétrique $p_{\text{model}}(\boldsymbol{x};\boldsymbol{\theta})$. Parmi elles, certaines méthodes permettent un calcul analytique de l'*evidence*. On peut citer notamment les modèles génératifs auto-régressifs tels que PixelRNN [So *et al.*, 2023], MADE [Germain *et al.*, 2015], NADE [Larochelle et Murray, 2011] et IAF [Kingma *et al.*, 2016] qui ont démontré d'excellentes capacités de génération d'images. Cependant, ces approches possèdent certaines limitations, notamment le coût calculatoire nécessaire à leur optimisation. De plus, la plupart de ces méthodes ne reposent pas directement sur un espace latent. Les modèles de flux normalisés permettent d'éviter ces contraintes en proposant une approche pour laquelle $p_{model}(\boldsymbol{x};\boldsymbol{\theta})$ est calculable directement sans nécessiter une capacité de calcul trop importante tout en se reposant sur un espace latent.

Afin de surmonter certains des inconvénients liés aux contraintes de conception des modèles à densité analytiques, d'autres approches ont également été développées. Elles fournissent toujours une densité de probabilité explicite, mais reposent sur une modélisation de l'*evidence* qui nécessite l'utilisation d'approximations. Ces approches peuvent être distinguées entre celles qui utilisent des approximations bayésiennes via des méthodes variationnelles et celles recourant aux méthodes de MCMC.

Les sections à venir consistent à présenter les modèles de flux normalisés et les Machines de Boltzmann restreintes ("Restricted Boltzmann Machine") (RBM), avant d'approfondir la compréhension des VAEs.

#### Densité analytique : les flux normalisés

Les modèles de flux normalisés [Müller *et al.*, 2019; Tabak et Turner, 2013] forment une catégorie de méthodes génératives basée sur la transformation d'une distribution simple, typiquement une gaussienne isotrope, en une distribution plus complexe. Cette transformation se fait au moyen d'une séquence de réseaux qui modélisent des transformations différentiables bijectives. Une illustration est proposée dans la Figure 1.8, où un vecteur

$\boldsymbol{z}$ est d'abord échantillonné selon la distribution $\mathcal{N}(0,1)$, puis sa distribution est modifiée à travers une composition de fonctions différentiables et bijectives, l'objectif étant d'approximer la distribution de $p_{data}(\boldsymbol{x})$.

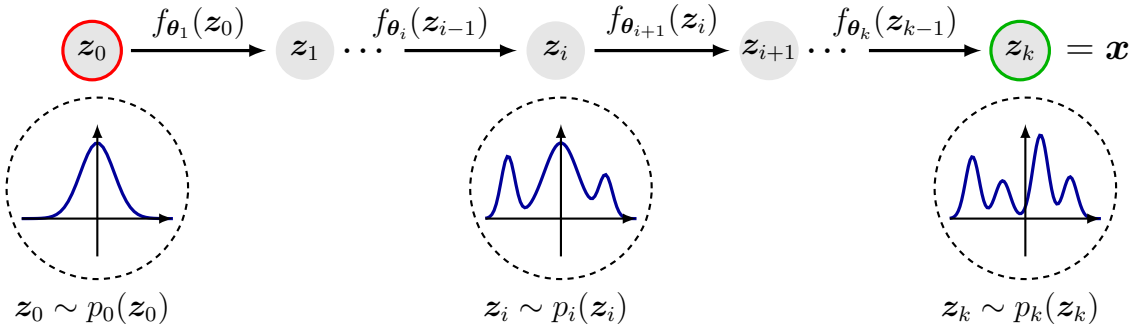



Fig 1.8: ($\star$) Représentation d'un modèle de flux normalisés.

Ces approches reposent sur la théorie du changement de variables. Plus précisément, elles considèrent deux variables aléatoires, $\boldsymbol{x}$ et $\boldsymbol{z}$, reliées par une transformation $f_{\boldsymbol{\theta}} : \mathbb{R}^n \to \mathbb{R}^n$, telles que $\boldsymbol{x} = f_{\boldsymbol{\theta}}(\boldsymbol{z})$ et $\boldsymbol{z} = f_{\boldsymbol{\theta}}^{-1}(\boldsymbol{x})$. Dans ce contexte, la distribution de *vraisemblance* s'exprime de la manière suivante :

$$\begin{aligned} p(\boldsymbol{x};\boldsymbol{\theta}) &= q_{\boldsymbol{\theta}}\big(f_{\boldsymbol{\theta}}^{-1}(\boldsymbol{x})\big)\Big|\det\left(\frac{\partial f_{\boldsymbol{\theta}}^{-1}(\boldsymbol{x})}{\partial \boldsymbol{x}}\right)\Big| \\ &= q_{\boldsymbol{\theta}}(\boldsymbol{z})\Big|\det\left(\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}}\right)\Big|. \end{aligned} \tag{1.15}$$

Le résultat obtenu dans l'équation (1.15) peut être appliqué à la jacobienne des fonctions inversibles en utilisant le théorème de la fonction inverse. Par conséquent, cette idée peut être étendue à une séquence de $K$ transformations $f_{\boldsymbol{\theta}_k}$, permettant ainsi de transformer une variable aléatoire initiale $\boldsymbol{z}_0 \sim q_0$ en une densité transformée $q_{\boldsymbol{\theta}_k}(\boldsymbol{z})$ :

$$\boldsymbol{z}_{\boldsymbol{\theta}_k} = f_{\boldsymbol{\theta}_k} \circ \ldots \circ f_{\boldsymbol{\theta}_2} \circ f_{\boldsymbol{\theta}_1}, \tag{1.16}$$

ce qui permet de mesurer :

$$\log q_k(\boldsymbol{z}_k) = \log q_0(\boldsymbol{z}_0) - \sum_{k=1}^{K} \log\Big|\det\left(\frac{\partial f_{\theta_k}^{-1}}{\partial \boldsymbol{z}_{k-1}}\right)\Big|. \tag{1.17}$$

Ces approches sont également optimisées à l'aide d'algorithmes de descente de gradient dans l'objectif d'optimiser les paramètres $\boldsymbol{\theta}$.

L'un des avantages majeur des méthodes de flux normalisés réside dans leur capacité à générer des échantillons de manière très efficace et à calculer la fonction de log-*vraisemblance* de manière analytique. Cependant, une limitation importante réside dans le fait que leur efficacité dépend de la complexité des transformations bijectives pour le calcul du jacobien. De plus, la bijectivité de la transformation conduit souvent à un espace latent de très haute dimension, rendant ainsi la manipulation et l'interprétation de cet espace difficile. Différentes approches ont été proposées pour surmonter ces limitations, notamment en ce qui concerne la manipulation de l'espace latent à partir de données d'entrées textuelles [Zhu *et al.*, 2022] ou d'images [Mahajan *et al.*, 2020]. Ces approches combinent les modèles de flux normalisés avec d'autres méthodes génératives, complexifiant davantage les architectures proposées. Du fait de l'ensemble de ces réserves, les méthodes génératives modélisant la distribution $p_{model}(\boldsymbol{x};\boldsymbol{\theta})$ de façon explicite et directement calculable ne nous ont pas paru être de bons candidats vis-à-vis de nos besoins.

### Densité approximée par MCMC : les machines de Boltzmann restreintes

Les Machines de Boltzmann restreintes ("Restricted Boltzmann Machine") (RBM), initialement appelés Harmoniums, sont un type de réseaux génératifs dont les origines remontent à un article de Smolensky [Smolensky *et al.*, 1986]. La théorie initiale des RBMs repose sur la deuxième loi de la thermodynamique, qui stipule qu'un système physique qui évolue dans le temps va approcher des conditions maximisant son entropie à condition que certaines contraintes restent inchangées, notamment son énergie. Cette approche permet l'utilisation de la distribution de Boltzmann, laquelle évalue la probabilité qu'un système occupe un état spécifique en fonction de l'énergie de cet état et de la température du système. Elle est définie de la façon suivante :

$$p_i \propto \exp\left(\frac{-E_i}{kT}\right), \tag{1.18}$$

où $p_i$ correspond à la probabilité du système d'être dans un état $i$, $E_i$ l'énergie de cet état et $kT$ une constante correspondant au produit entre la constante de Boltzmann $k$ et de la température thermodynamique $T$.

En se basant sur ces principes, l'objectif principal des RBMs consiste à définir les paramètres d'un réseau de neurones minimisant une énergie lorsqu'un état d'entrée est pertinent. Pour cela, ces réseaux sont définis comme un modèle de graphes probabilistes contenant une couche de variables observables dénotée $\boldsymbol{x} = \{x_1, ..., x_N\} \in \{0,1\}^N$ et une couche simple de variables latentes $\boldsymbol{h} = \{h_1, ..., h_M\} \in \{0,1\}^M$. Ce graphe bipartite restreint les connexions entre les variables de la couche visible, ainsi qu'entre les variables de la couche cachée. Cela permet à la fois une simplicité dans les calculs et dans leur compréhension, une propriété souhaitée par Smolenski.

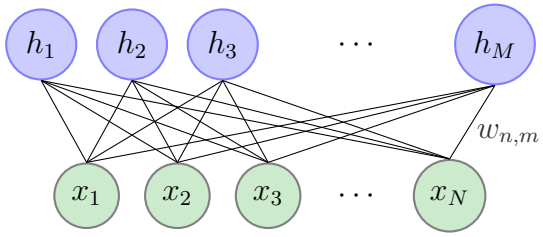



Fig 1.9: Illustration d'un RBM, composé de deux couches totalement connectées. La couche $\boldsymbol{x}$ est composée de $N$ neurones, et la couche $\boldsymbol{h}$ est composée de $M$ neurones.

Cette modélisation permet de définir l'énergie du système pour un couple $(\boldsymbol{x}, \boldsymbol{h})$ de la façon suivante :

$$E(\boldsymbol{x}, \boldsymbol{h}; \boldsymbol{\theta}) = -\left(\sum_{n,m} w_{n,m} x_n h_m + \sum_n b_n x_n + \sum_m c_m h_m\right), \tag{1.19}$$

où $w_{n,m}$ représente le poids entre le neurone d'entrée $n$ et le neurone caché $m$, $x_n$ est l'état du neurone visible $n$ , $h_m$ est l'état du neurone caché $m$. $b_n$ et $c_m$ sont les poids respectifs des neurones $x_n$ et $h_m$ et $\boldsymbol{\theta} = \{w_{n,m}, b_n, c_m\}$.

La probabilité jointe $p(\boldsymbol{x}, \boldsymbol{h}; \boldsymbol{\theta})$ du réseau est définie par l'utilisation d'une distribution de Boltzmann de la façon suivante :

$$p(\boldsymbol{x}, \boldsymbol{h}; \boldsymbol{\theta}) = \frac{\exp\big(-E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})\big)}{Z(\boldsymbol{\theta})}, \tag{1.20}$$

où la fonction de partition $Z(\boldsymbol{\theta}) \equiv \sum_{\boldsymbol{x},\boldsymbol{h}} \exp\big(-E(\boldsymbol{x},\boldsymbol{h};\boldsymbol{\theta})\big)$ correspond à la somme de l'ensemble des configurations possibles. Les probabilités marginales $p(\boldsymbol{x};\boldsymbol{\theta})$ et $p(\boldsymbol{h};\boldsymbol{\theta})$ pour les couches visibles et cachées sont obtenues en sommant respectivement les variables visibles et cachées, telles que :

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{h} p(\boldsymbol{x},\boldsymbol{h};\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \sum_{h} \exp(-E(\boldsymbol{x},\boldsymbol{h};\boldsymbol{\theta})), \tag{1.21}$$

$$p(\boldsymbol{h};\boldsymbol{\theta}) = \sum_{x} p(\boldsymbol{x},\boldsymbol{h};\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \sum_{x} \exp(-E(\boldsymbol{x},\boldsymbol{h};\boldsymbol{\theta})). \tag{1.22}$$

L'apprentissage d'un RBM consiste à ajuster $\boldsymbol{\theta}$ de telle sorte à ce que la probabilité marginale de la couche visible $p(\boldsymbol{x};\boldsymbol{\theta})$ soit aussi proche que possible de la distribution inconnue $p_{data}(\boldsymbol{x})$ qui génère l'ensemble des données d'entraînement. L'objectif consiste alors à minimiser la fonction de log *vraisemblance* à l'aide d'un algorithme de descente de gradient. On peut démontrer que le gradient de cette fonction s'exprime de la façon suivante :

$$\frac{\partial\big(-\log p(\boldsymbol{x})\big)}{\partial\boldsymbol{\theta}} = \mathbb{E}_{\boldsymbol{h}}\left[\frac{\partial E(\boldsymbol{x},\boldsymbol{h})}{\partial\boldsymbol{\theta}}|x\right] - \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\left[\frac{\partial E(\boldsymbol{x},\boldsymbol{h})}{\partial\boldsymbol{\theta}}\right]. \tag{1.23}$$

Le premier terme de l'équation (1.23) est appelé la phase positive et peut être calculé directement. En revanche, le second terme, appelé phase négative, requiert une approximation à l'aide de méthodes de MCMC, appelée divergence contrastive [Hinton, 2002].

En raison de leur simplicité, les RBMs peuvent constituer une architecture intéressante pour la manipulation de l'espace latent, représenté ici par la couche de neurones $\boldsymbol{h}$. Cependant, la nature binaire des neurones de la couche visible ne permet leur application qu'à des données binaires. Pour pallier ce problème, des extensions ont été proposées permettant de considérer des entrées continues. C'est le cas notamment des RBMs gaussiens [Cho *et al.*, 2011; Hinton et Salakhutdinov, 2006; Chen et Murray, 2003].

Toutefois, il a été précisé dans les sections précédentes l'intérêt d'utiliser une succession de couches de neurones convolutives et denses, à l'instar des architectures de CNN, lorsqu'il s'agit d'extraire des caractéristiques pertinentes à partir d'images. Malheureusement, l'extension des RBMs à des réseaux de neurones profonds présente des apprentissages instables et une mauvaise convergence.

Du fait de ces limitations, la section suivante vise à présenter un dernier modèle génératif qui utilise une approximation bayésienne via une méthode variationnelle. Cette modélisation définit ainsi un cadre propice à la manipulation de représentations.

## 1.4 Les auto-encodeurs variationnels

La méthode de l'Auto-encodeur variationnel ("Variational Auto-Encoder") (VAE) [Kingma et Welling, 2014; Rezende *et al.*, 2014] a pour objectif de modéliser la distribution sous-jacente des données, $p_{data}(\boldsymbol{x})$, à l'aide d'un premier réseau qui apprend à encoder les observations d'entrée dans un espace latent continu $p(\boldsymbol{z})$, dans lequel les facteurs génératifs sont représentés de manière compressée. Ce premier réseau est appelé modèle d'inférence ou plus communément encodeur. Une fois l'espace latent inféré, un second réseau, le réseau génératif ou décodeur, a pour objectif de générer de nouvelles données dont la distribution est la plus proche possible de la distribution $p_{data}$, à partir d'un échantillon $\boldsymbol{z}$.

Un élément clé de ce modèle réside dans l'utilisation de la technique d'inférence variationnelle dont le but est d'estimer la distribution inconnue de l'espace latent à partir des données d'entrée, c'est-à-dire $p(\boldsymbol{z}|\boldsymbol{x})$. Son intérêt pour les modèles de VAEs repose principalement sur sa rapidité et son efficacité calculatoire. En effet, elle repose sur une approximation analytique, ce qui permet d'obtenir des échantillons plus rapidement qu'en utilisant des méthodes MCMC, telle la divergence contrastive intervenant dans l'apprentissage des RBMs.

La suite de ce chapitre a pour but d'expliciter l'objectif et la mise en œuvre de la technique d'inférence variationnelle utilisée pour la modélisation des VAEs. Les démonstrations sont davantage détaillées, permettant une compréhension fine des concepts sur lesquels reposent les développements menés ultérieurement dans le manuscrit.

### 1.4.1 La méthode d'inférence variationnelle

L'inférence variationnelle est une méthode bayésienne approchée qui a pour but d'obtenir une estimation analytique de la densité de probabilité conditionnelle des variables latentes $\boldsymbol{z}$ étant donné les données observées $\boldsymbol{x}$, soit $p(\boldsymbol{z}|\boldsymbol{x})$. En utilisant la formule de Bayes, cette dernière peut être exprimée de la manière suivante :

$$p(\boldsymbol{z}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{p(\boldsymbol{x})}, \tag{1.24}$$

où $p(\boldsymbol{z}|\boldsymbol{x})$ est appelée la distribution *a posteriori*, $p(\boldsymbol{x}|\boldsymbol{z})$ représente la distribution de *vraisemblance* et la densité marginale des observations $p(\boldsymbol{x})$ est connue sous le nom d'*évidence*. Celle-ci comporte un grand nombre de composantes et est généralement difficile à calculer, nécessitant ainsi une inférence approximative. Les méthodes d'inférence variationnelle reposent sur l'idée de définir une famille $\mathcal{D}$ de distributions paramétriques pour les variables latentes. En résolvant un problème d'optimisation, on cherche l'élément $q(\boldsymbol{z})^*$ de cette famille qui minimise une mesure de similarité, souvent exprimée sous forme de divergence de Kullback-Leibler (KL) :

$$q(\boldsymbol{z})^* = \text{argmin}_{\boldsymbol{q}\in\mathcal{D}} D_{KL}\big[q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{x})\big]. \tag{1.25}$$

En utilisant la définition de la KL-divergence, il est possible de développer l'équation (1.25) comme suit :

$$\begin{aligned}
\int_{\boldsymbol{z}} q(\boldsymbol{z}) \log\left(\frac{q(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x})}\right) d\boldsymbol{z} &= -\int_z q(\boldsymbol{z}) \log\left(\frac{p(\boldsymbol{z}|\boldsymbol{x})}{q(\boldsymbol{z})}\right) d\boldsymbol{z} \\
&= -\int_{\boldsymbol{z}} q(\boldsymbol{z}) \left[\log\left(\frac{p(\boldsymbol{z},\boldsymbol{x})}{p(\boldsymbol{x})q(\boldsymbol{z})}\right)\right] d\boldsymbol{z} \\
&= -\int_{\boldsymbol{z}} q(\boldsymbol{z}) \left[\log\left(\frac{p(\boldsymbol{z},\boldsymbol{x})}{q(\boldsymbol{z})}\right)\right] d\boldsymbol{z} + \log p(\boldsymbol{x}).
\end{aligned} \tag{1.26}$$

En réécrivant les termes, nous obtenons finalement :

$$D_{KL}\big[q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{x})\big] = -\mathbb{E}_{q(\boldsymbol{z})}\left[\log\left(\frac{p(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})}\right)\right] + \log p(\boldsymbol{x}). \tag{1.27}$$

L'objectif initial est de trouver une approximation $q(\boldsymbol{z})$ qui soit proche de la vraie distribution *a posteriori*. Comme le terme $\log p(\boldsymbol{x})$ ne dépend pas de $q(\boldsymbol{z})$, trouver $q(\boldsymbol{z})^*$ revient à minimiser le premier terme de l'équation (1.27). En effet, en réorganisant (1.27), et avec la positivité de la KL-divergence, ce dernier représente une borne inférieure de l'*évidence*

et est pour cela dénommé Borne basse de l'*evidence* ("Evidence Lower Bound") (ELBO).

Plusieurs décompositions de l'ELBO peuvent être proposées, aboutissant à différentes interprétations, néanmoins le principe initial du VAE repose sur l'approche suivante :

$$\begin{aligned} ELBO &= \int_{\boldsymbol{z}} q(\boldsymbol{z}) \log \left( \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})} \right) d\boldsymbol{z} \\ &= \int_{\boldsymbol{z}} q(\boldsymbol{z}) \log \left( \frac{p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{q(\boldsymbol{z})} \right) d\boldsymbol{z} \\ &= \int_{\boldsymbol{z}} q(\boldsymbol{z}) \log \left( p(\boldsymbol{x}|\boldsymbol{z}) + \log \left( \frac{p(\boldsymbol{z})}{q(\boldsymbol{z})} \right) \right) d\boldsymbol{z} \\ &= E_{q(\boldsymbol{z})}\big[\log p(\boldsymbol{x}|\boldsymbol{z})\big] - D_{KL}\big[q(\boldsymbol{z})||p(\boldsymbol{z})\big]. \end{aligned} \tag{1.28}$$

Dans cette expression générale, le choix du modèle de distribution $q(\boldsymbol{z})$ est arbitraire. Cependant, dans le contexte de modèles de variables latentes qui nous intéresse, il est pertinent de rendre la modélisation dépendant explicitement des données $\boldsymbol{x}$, c'est-à-dire $q(\boldsymbol{z}|\boldsymbol{x})$. Cela nous permet d'aboutir à la formulation suivante :

$$ELBO = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\big[\log p(\boldsymbol{x}|\boldsymbol{z})\big] - D_{KL}\big[q(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})\big]. \tag{1.29}$$

Les distributions inconnues $q(\boldsymbol{z}|\boldsymbol{x})$ et $p(\boldsymbol{x}|\boldsymbol{z})$ dans l'équation (1.29) peuvent être modélisées à l'aide de deux fonctions paramétriques. Ainsi, le problème d'inférence se traduit par l'optimisation des paramètres de ces fonctions qui sont partagés par l'ensemble des données. Ce procédé est alors qualifié d'amorti et est illustré dans la Figure 1.10.

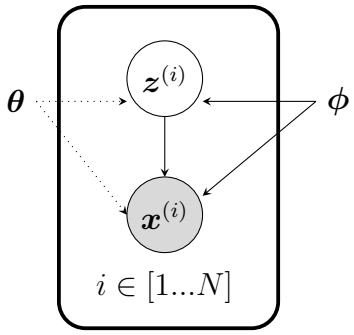



Fig 1.10: Représentation graphique de l'inférence variationnelle amortie, pour un jeu de données $\boldsymbol{x} \in \mathbb{R}^N$.

### 1.4.2 Développement du modèle et apprentissage

Lorsque les données d'entrées sont des images, les réseaux du VAE sont généralement modélisés par des CNNs permettant l'extraction de caractéristiques visuelles. Le décodeur a pour objectif de reconstruire les images données en entrée de l'encodeur à partir d'échantillons $\boldsymbol{z}$ de l'espace latent. La majeure partie du temps, l'architecture du décodeur est similaire à celle de l'encodeur à la différence de l'ordre des couches de neurones qui est inversée. Les paramètres de ces réseaux, notés respectivement $\boldsymbol{\phi}$ et $\boldsymbol{\theta}$, sont conjointement optimisés dans le but de maximiser l'ELBO. Cet apprentissage a pour objectif de faire émerger une représentation suffisamment bonne au sein de l'espace latent afin que le décodeur soit en capacité de générer des données suivant une distribution proche de $p_{data}(\boldsymbol{x})$.

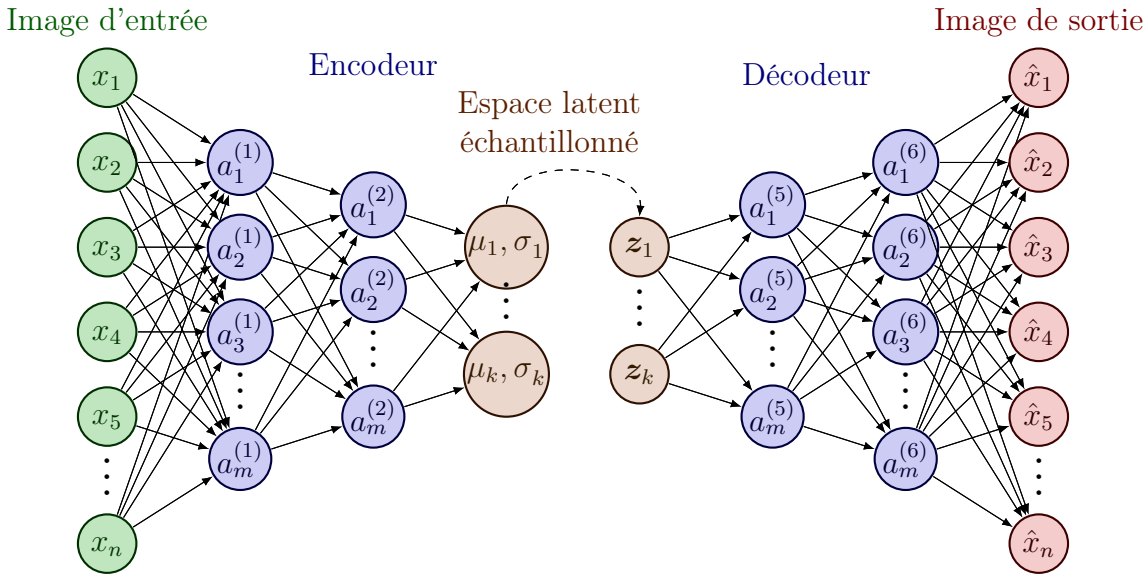



Fig 1.11: Illustration d'un auto-encodeur variationnel. Les images d'entrée et de sortie sont respectivement colorées en vert et rouge, les couches cachées en bleues et les couches latentes correspondant aux paramètres de la distribution *a posteriori* $\{\boldsymbol{\mu}(\boldsymbol{x};\boldsymbol{\phi}), \boldsymbol{\sigma}^2(\boldsymbol{x};\boldsymbol{\phi})\}$ et aux échantillons $\boldsymbol{z}$ en orange. Dans cette illustration, les couches cachées sont représentées par des couches denses, toutefois il est également possible de considérer des couches de convolution.

### L'encodeur

Le réseau d'encodeur vise à inférer les paramètres de la distribution *a posteriori* approximée, notée $q_\phi(\boldsymbol{z}|\boldsymbol{x})$, dont le choix est défini en amont de l'apprentissage. L'hypothèse classique consiste à modéliser cette dernière comme une distribution gaussienne, et à considérer les variables $\boldsymbol{z}$ comme indépendantes. Cela implique que l'encodeur a pour objectif d'apprendre les paramètres $\boldsymbol{\phi}$ de telle sorte à ce que l'ensemble $\{\boldsymbol{\mu}(\boldsymbol{x};\boldsymbol{\phi}), \boldsymbol{\sigma}^2(\boldsymbol{x};\boldsymbol{\phi})\} \in \mathbb{R}^K$, où $K$ correspond à la dimension de l'espace latent, représente les paramètres de la distribution approchée tels que :

$$q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) = \prod_{i=1}^{K} \mathcal{N}\big(z_i|\mu_i(\boldsymbol{x};\boldsymbol{\phi}), \sigma_i^2(\boldsymbol{x};\boldsymbol{\phi})\big). \tag{1.30}$$

Ainsi, il devient possible d'obtenir rapidement des variables depuis l'espace latent en échantillonnant directement à partir de la distribution *a posteriori* apprise.

### Le décodeur

Le modèle du décodeur a pour fonction de générer une image ressemblant étroitement à une observation réelle $\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x})$ à partir d'un échantillon de $p(\boldsymbol{z})$. Pour cela, son objectif est d'apprendre les paramètres de la distribution de *vraisemblance*, dont la nature est également définie avant apprentissage.
En fonction du type des données d'entrée, cette dernière est souvent modélisée à l'aide d'une distribution normale ou d'une distribution de Bernoulli telles que :

$$p_\theta(\boldsymbol{x}|\boldsymbol{z}) = \prod_{i=1}^{N} \mathcal{N}\big(\boldsymbol{x}_i; \mu_i(\boldsymbol{z};\boldsymbol{\theta}), \sigma_i^2(\boldsymbol{z})\big) \text{ ou } \prod_{z}^{N} \mathcal{B}\big(\boldsymbol{x}_i; p_i(\boldsymbol{z};\boldsymbol{\theta})\big). \tag{1.31}$$

Dans cette équation, $N$ correspond au nombre de pixels dans une image, ces derniers étant considérés comme indépendants. $\mathcal{B}(\boldsymbol{x}; p(\cdot))$ représente une distribution de Bernoulli

de paramètre $p$, et $p(\boldsymbol{z};\boldsymbol{\theta})$ correspond à la sortie du décodeur.

Il est important de souligner que lorsque la distribution est modélisée par une loi gaussienne, le paramètre de variance $\boldsymbol{\sigma}^2(\boldsymbol{z})$ est généralement traité comme un hyperparamètre fixé et connu *a priori*, permettant de garantir la stabilité du processus d'apprentissage. Bien que certains travaux suggèrent qu'il est préférable de laisser ce paramètre libre et de l'optimiser en même temps que $\boldsymbol{\theta}$) pour une meilleure modélisation générative [Yu, 2020; Rybkin *et al.*, 2021], nous avons choisi de fixer ce paramètre pour maintenir la robustesse de l'apprentissage. On peut également noter que la majorité des auteurs modélise la densité de *vraisemblance* à l'aide d'une distribution de Bernoulli, bien que dans le cas général, les pixels ne prennent pas des valeurs binaires. Cela évite d'avoir à déterminer $\boldsymbol{\sigma}^2(\boldsymbol{z})$, qui a un impact considérable sur l'apprentissage. Cette erreur de modélisation commune a donné lieu à des propositions d'amélioration [Loaiza-Ganem et Cunningham, 2019], mais reste toutefois largement appliquée.

**Mise en œuvre de l'apprentissage**

L'apprentissage du VAE consiste à optimiser conjointement l'ensemble de paramètres $\{\boldsymbol{\phi},\boldsymbol{\theta}\}$ dans l'objectif de maximiser l'ELBO, équation (1.29), ce qui revient à minimiser son opposé à travers une méthode de descente de gradient. Pour ce faire, il est nécessaire de calculer le gradient de la fonction de coût par rapport aux paramètres à optimiser, soit :

$$\nabla_{\phi,\theta}\left[\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\big[-\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\big] + D_{KL}\big[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})\big]\right]. \tag{1.32}$$

Le calcul du gradient par rapport à $\boldsymbol{\phi}$ pour le second terme de cette équation pose peu de problème, car la KL-divergence entre deux distributions gaussiennes peut être calculée de manière analytique.
En effet, lorsque $\boldsymbol{z}$ est une variable aléatoire suivant deux distributions de probabilité notées $p_1(\boldsymbol{z}|\boldsymbol{x})$ et $p_2(\boldsymbol{z})$, définies comme suit :

$$\begin{aligned} &p_1(\boldsymbol{z}|\boldsymbol{x}) : \boldsymbol{z} \sim \mathcal{N}\big(\boldsymbol{z};\mu_1(\boldsymbol{x}),\sigma_1^2(\boldsymbol{x})\big)\\ &p_2(\boldsymbol{z}) : \boldsymbol{z} \sim \mathcal{N}\big(\boldsymbol{z};\mu_2,\sigma_2^2\big), \end{aligned} \tag{1.33}$$

alors, la divergence de Kullback-Leibler entre ces deux distributions s'écrit :

$$D_{KL}\big[p_1(\boldsymbol{z}|\boldsymbol{x})||p_2(\boldsymbol{z})\big] = \frac{1}{2}\left[\frac{(\mu_2-\mu_1(\boldsymbol{x}))^2}{\sigma_2^2} + \frac{\sigma_1(\boldsymbol{x})^2}{\sigma_2^2} - \log\left(\frac{\sigma_1(\boldsymbol{x})^2}{\sigma_2^2}\right) - 1\right]. \tag{1.34}$$

L'hypothèse classique consiste à définir la distribution *a priori* $p(\boldsymbol{z})$ comme une distribution gaussienne avec une moyenne nulle et une covariance identité, ce qui revient à calculer le terme suivant :

$$D_{KL}\big[p_1(\boldsymbol{z}|\boldsymbol{x})||p_2(\boldsymbol{z})\big] = \frac{1}{2}\big[\mu_1(\boldsymbol{x})^2 + \sigma_1(\boldsymbol{x})^2 - \log\big(\sigma_1(\boldsymbol{x})^2\big) - 1\big]. \tag{1.35}$$

Ce terme peut être interprété comme une forme de régularisation qui guide la topologie de l'espace latent inféré. En effet, cette contrainte favorise un espace latent dont les variables sont proches de la distribution gaussienne isotrope, contribuant ainsi à l'apprentissage d'une représentation structurée des données.

De manière similaire, la fonction de log-*vraisemblance* d'une distribution gaussienne $\mathcal{N}(x;\mu,\sigma^2)$ peut être exprimée de la façon suivante :

$$\log p(\boldsymbol{x}|\boldsymbol{z}) = -\frac{1}{2}\log(2\pi) - \log\big(\sigma^2(\boldsymbol{z})\big) - \frac{1}{2\sigma^2(\boldsymbol{z})}\sum_{i=1}^{N}\left(x_i-\mu_i(\boldsymbol{z})\right)^2, \tag{1.36}$$

tandis que pour une distribution Bernoulli $\mathcal{B}(x;p)$ on obtient :

$$\log p(\boldsymbol{x}|\boldsymbol{z}) = \sum_{i=1}^{N} x_i \log\big(p_i(z)\big) + (1-x_i)\log\big(1-p_i(z)\big). \tag{1.37}$$

Ces termes sont conçus pour favoriser la capacité de reconstruction du décodeur. On observe ainsi un terme représentant la somme des erreurs quadratiques pour la modélisation de la *vraisemblance* par distribution gaussienne, tandis que c'est la fonction d'entropie croisée qui émerge lorsque la distribution en question suit une loi de Bernoulli.

Cependant, il est essentiel de noter qu'au sein de l'ELBO, le terme de log-*vraisemblance* inféré par le décodeur est en réalité une espérance par rapport à $\boldsymbol{z}$, dépendante des paramètres $\boldsymbol{\phi}$ de l'encodeur. Cette modélisation requiert l'emploi d'une technique particulière. L'astuce la plus couramment employée pour aborder le calcul du gradient de cette espérance, appelée l'astuce de reparamétrisation [Salimans et Knowles, 2013; Kingma et Welling, 2014; Rezende *et al.*, 2014], consiste à exprimer la variable stochastique $\boldsymbol{z}$ comme une fonction inversible d'une autre variable $\boldsymbol{\epsilon}$, c'est-à-dire $\boldsymbol{z} = \mathcal{T}(\boldsymbol{\epsilon};\boldsymbol{\phi})$, de manière que la distribution de la nouvelle variable aléatoire $q_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})$ ne dépende pas des paramètres $\boldsymbol{\phi}$. Sous ces hypothèses, l'espérance par rapport à $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ peut être réécrite comme $\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}[-\log p(\boldsymbol{x}|\boldsymbol{z})] = \mathbb{E}_{q_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})}[-\log p(\boldsymbol{x}|\mathcal{T}(\boldsymbol{\epsilon};\boldsymbol{\phi}))]$, et le gradient par rapport à $\boldsymbol{\phi}$ peut être déplacé à l'intérieur de l'espérance, ce qui donne :

$$-\nabla_{\boldsymbol{\phi},\boldsymbol{\theta}} ELBO = \mathbb{E}_{q_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})}\Big[\nabla_{\boldsymbol{\theta}} - \log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\nabla_{\boldsymbol{\phi}}\mathcal{T}(\boldsymbol{\epsilon};\boldsymbol{\phi})\Big] + \nabla_{\boldsymbol{\phi}} D_{KL}\big[q_{\phi}(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})\big]. \tag{1.38}$$

Le gradient peut ainsi être estimé à l'aide d'une approche de Monte-Carlo pour laquelle l'approche stochastique des algorithmes de descente de gradient permet de faire en sorte qu'un seul échantillon de $q_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})$ soit suffisant, et que le gradient présente une faible variance.

Lorsque la distribution variationnelle est gaussienne avec une moyenne $\mu(\boldsymbol{x};\boldsymbol{\phi})$ et une variance $\sigma^2(\boldsymbol{x};\boldsymbol{\phi})$, une reparamétrisation simple consiste à standardiser la variable aléatoire $\boldsymbol{z}$, c'est-à-dire :

$$\boldsymbol{z} = \mathcal{T}(\boldsymbol{\epsilon};\boldsymbol{\phi}) = \mu(\boldsymbol{x};\boldsymbol{\phi}) + \sigma(\boldsymbol{x};\boldsymbol{\phi}) \cdot \boldsymbol{\epsilon}, \tag{1.39}$$

où $\cdot$ représente le produit élément par élément, et $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbb{I})$. Cette reparamétrisation permet d'obtenir une expression différentiable de $\boldsymbol{z}$ par rapport aux paramètres $\boldsymbol{\phi}$, facilitant ainsi le calcul de la méthode d'apprentissage nécessaire à l'entraînement des modèles basés sur un VAE.

Pour conclure, les modèles de type VAE, contrairement aux autres méthodes génératives non supervisées, possèdent des caractéristiques intéressantes vis-à-vis de notre problématique de recherche. L'espace latent étant conditionné par les données d'entrées, ce qui permet d'encoder des caractéristiques pertinentes de ces dernières, est une caractéristique fondamentale vis-à-vis de notre besoin. De plus, la dimension de $\boldsymbol{z}$ étant généralement définie comme étant inférieure à celle des données d'entrées, et la facilité d'obtenir des variables latentes en échantillonnant directement depuis $q_{\boldsymbol{\phi}}(\boldsymbol{z}|x)$ permet l'obtention d'un espace pouvant être facilement manipulé. Ces propriétés permettent un cadre favorable à une analyse d'incertitude effectuée sur les représentations obtenues au sein de l'espace latent.

### 1.4.3 Synthèse

Pour rappel, notre besoin consiste à proposer un algorithme ayant la capacité d'obtenir une représentation des données d'entrées dans un espace à dimensions réduites, qui puisse être interprétable et manipulable dans un objectif futur d'effectuer une analyse d'incertitude sur les représentations obtenues. Pour cela, nous nous sommes concentrés sur les méthodes non supervisées dû à leur capacité à s'entraîner sur des données non étiquetées ainsi qu'à s'appuyer sur des approches probabilistes. Parmi ces modèles, l'analyse des méthodes génératives a permis de mettre en avant certaines limitations par rapport à notre besoin, dont la synthèse est proposée dans le tableau 1.1. En effet, les approches dites "implicites", pour lesquelles l'*evidence* $p_{model}(\boldsymbol{x};\boldsymbol{\theta})$ n'est pas définie, ne permettent pas un encodage des données d'entrée au sein de l'espace latent, et ainsi ne répondent pas à notre besoin. D'un autre côté, parmi les modèles dits "explicites", les architectures permettant une estimation analytique de cette distribution sont relativement coûteuses, et ne permettent généralement pas une représentation des données dans un espace à dimension réduites, ce qui rend la manipulation et l'interprétation de l'espace latent complexe. Finalement, les approches pour lesquelles la distribution $p_{model}(\boldsymbol{x};\boldsymbol{\theta})$ est approchée permettent la définition d'un espace latent conditionné par les données du jeu d'entraînement, dans un espace réduit favorisant sa manipulation. En outre, afin d'extraire des caractéristiques souhaitables à partir d'images, la succession de couches de neurones et notamment l'utilisation d'architectures de CNN est l'approche la plus pertinente. Cela restreint davantage le choix du modèle pour nos approches. Les RBMs, du fait des techniques sur lesquelles repose leur apprentissage, ne permettent pas l'utilisation de telles couches de neurones.

| Propriétés | **Méthodes implicites** | **Méthodes explicites** | | | | |
|---|---|---|---|---|---|---|
| | GAN | N.F | A.R | EBM | D.M | VAE |
| $\boldsymbol{z}$ conditionné par les entrées | × | ✓ | × | ✓ | ✓ | ✓ |
| Espace latent à dimensions réduites | ✓ | × | × | ✓ | × | ✓ |
| Extraction de caractéristiques pertinentes | × | ✓ | × | × | ✓ | ✓ |

Tableau 1.1: Propriétés des méthodes de l'état de l'art vis-à-vis de notre besoin. Dans le souci de simplifier la lecture, les architectures de flux normalisés sont dénotés par l'acronyme "N.F", les méthodes auto-régressives par "A.R" et les modèles de diffusion par "D.M".

Suite à l'ensemble des méthodes génératives présentées et des analyses effectuées, nous avons déterminé que les méthodes présentant les meilleures propriétés vis-à-vis de nos objectifs, à savoir un modèle génératif non supervisé présentant un espace de représentations accessible, sont les VAEs. En effet, leur modélisation répond à l'ensemble des propriétés citées dans le tableau 1.1.

## 1.5 Conclusion

Dans de ce premier chapitre, les éléments fondamentaux de l'apprentissage profond manipulés lors des études présentées dans ce manuscrit sont repris. Pour ce faire, le concept de neurone artificiel est examiné en détail avant d'être élargi au sein d'un réseau de neurones profond. Le processus d'apprentissage des paramètres à travers la méthode de propagation arrière du gradient est ensuite détaillé, puis les différentes infrastructures logicielles utilisées par la communauté scientifique pour une programmation efficace de l'apprentissage de ces modèles sont passées en revue.

Les paradigmes de l'apprentissage supervisé, à partir duquel se basent les mesures présentées dans la section 2.4, et non supervisé, à partir duquel est défini notre cadre de travail, sont également définis. Les modèles génératifs dont l'objectif est d'apprendre la distribution $p_{data}$ des données d'entrée sont particulièrement détaillés. En effet, des modèles possèdent des capacités intéressantes pour notre objectif de recherche, à savoir développer un algorithme basé sur de l'apprentissage qui ne nécessite pas une labellisation préalable et permettant d'obtenir une représentation interprétable des données d'entrée dans un espace à dimensions réduites. Après avoir effectué une revue des différentes approches, qui se différencient par le type de modélisation de la loi de *vraisemblance* des données générées, le choix de nous baser sur les méthodes de type VAEs est justifié. En effet, contrairement aux modèles dits implicites, l'approche du VAE présente une caractéristique essentielle vis-à-vis de notre besoin : ils possèdent un espace latent directement conditionné par les données d'entrée. Cette propriété leur permet de capturer de manière compacte les informations cruciales des données observables dans un espace restreint. Cette caractéristique permet ainsi la manipulation d'un espace latent composé de représentations significatives. De plus, contrairement aux autres approches génératives à densité explicite, la méthode d'inférence variationnelle sur laquelle se base la modélisation des VAEs permet un échantillonnage rapide depuis l'espace latent tout en préservant une capacité d'apprentissage efficace sur des images, sans nécessiter de ressources de calcul excessives.

L'ensemble des concepts explicités jusqu'alors permet de poursuivre sur le deuxième chapitre de ce manuscrit. L'objet de ce second chapitre est de faire un état de l'art des approches qui se basent sur un VAE, et qui ont pour objectif d'obtenir une représentation spécifique au sein de l'espace latent, à savoir le désentrelacement des variables qui le compose.

# Chapitre 2

# Espace latent et désentrelacement

L'objectif de ce chapitre est de se consacrer à l'étude du désentrelacement, une structure particulière de représentation au sein de l'espace latent.
Pour cela, les diverses définitions du concept décrites dans l'état de l'art sont examinées. Ensuite, une revue détaillée des méthodes reposant sur un VAE afin d'obtenir un espace latent désentrelacé est présentée, suivie d'une exploration des métriques permettant d'évaluer la qualité de ce dernier. Enfin, un panorama des différents jeux de données utilisés à des fins de validation est réalisé.

## 2.1 Le désentrelacement, considéré comme une "bonne" représentation

Un VAE projette les données d'entrée dans un espace de dimension restreinte, ce qui rend cette méthode attrayante pour la manipulation de représentations. Cependant, l'espace latent obtenu après apprentissage est composé la plupart du temps par une représentation désorganisée et entrelacée des données. Utiliser cet espace en tant que donnée d'entrée à des fins d'apprentissage peut entraîner des modèles non interprétables et potentiellement conduire à des prédictions fausses lors de son application à des tâches de plus haut niveau [Locatello *et al.*, 2019a], telles que la classification ou la régression. Ainsi, la recherche en apprentissage de représentations cherche actuellement à se rapprocher des aspects biologiques de l'intelligence humaine [Lake *et al.*, 2017] en améliorant la capacité des modèles à transférer des connaissances et à généraliser au-delà de la distribution des données d'entraînement.

Dans le domaine spécifique de la vision par ordinateur, la grande quantité de données utilisées lors du processus d'apprentissage peut conduire à l'obtention d'une certaine structure au sein de l'espace latent. La qualité des représentations obtenues est alors dépendante de la tâche spécifique pour laquelle elles ont été apprises. En effet, certains auteurs mentionnent que pour être considéré comme utile à une tâche ultérieure, un bon encodage doit répondre à un besoin plus spécifique défini selon l'utilisation que l'on souhaite en faire [Radford *et al.*, 2015; Tschannen *et al.*, 2018; Bengio *et al.*, 2013]. À partir de ce constat, différents paradigmes de structures de représentations sont présentés. Ces derniers se basent sur le postulat que les données sont générées à partir de facteurs génératifs $\boldsymbol{v} \in \mathbb{R}^M$, des facteurs physiquement interprétables permettant d'expliquer la variabilité entre deux images [Sepliarskaia *et al.*, 2019a]. La Figure 2.1 illustre ce concept sur le jeu de données Dsprites [Matthey *et al.*, 2017], généré à partir de cinq facteurs génératifs : la forme, l'échelle, l'orientation ainsi que la position de l'objet sur les axes x et y. Le facteur d'orientation a délibérément été omis, car il est mal défini selon la forme considérée. En

effet, appliquer une rotation de 0 ou de 180 degrés sur une forme carrée résulte en une image similaire. Dans ce contexte, le facteur génératif d'orientation ne permet pas de distinguer les deux images obtenues.

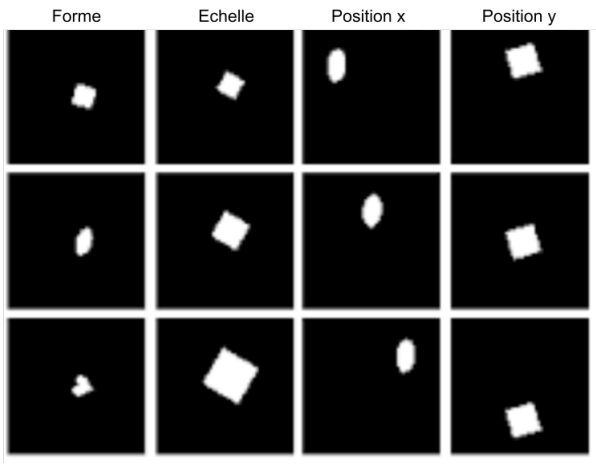



Fig 2.1: Exemples d'images provenant du jeu de données Dsprites [Matthey *et al.*, 2017], généré à partir de cinq facteurs génératifs dont quatre sont illustrés dans cette Figure.

Ces paradigmes définissent trois types de structures différentes qui peuvent être souhaitées pour l'espace latent inféré, selon ce que l'on cherche à faire ; on parle de structure regroupée, hiérarchique ou désentrelacée.

- La **structure regroupée** se traduit par des régions dans l'espace latent où les points de données partagent des similitudes ou des caractéristiques communes [Dupont, 2018; Makhzani *et al.*, 2015]. Ces regroupements peuvent être interprétés comme des représentations latentes significatives des différentes classes ou des catégories d'objets dans les données.

- La **structure hiérarchique** se réfère à une organisation en niveaux imbriqués des représentations latentes, où les niveaux supérieurs capturent des caractéristiques plus globales ou abstraites, tandis que les niveaux inférieurs représentent des détails plus fins ou spécifiques [Sønderby *et al.*, 2016a; Maaløe *et al.*, 2019a; Gulrajani *et al.*, 2016; Zhao *et al.*, 2017]. Cette approche repose sur l'idée que le monde peut être expliqué à travers une hiérarchie croissante de concepts abstraits.

- La **structure désentrelacée** se réfère à une représentation dans laquelle les facteurs génératifs sont clairement séparés et indépendants les uns des autres [Kim *et al.*, 2019b; Higgins *et al.*, 2016]. Dans ce contexte, le terme "désentrelacé" signifie que les différentes dimensions de l'espace latent permettent de représenter les variations spécifiques de chaque facteur génératif de manière isolée et non mêlée avec les autres facteurs.

Au cours de nos travaux de recherche, nous avons accordé une attention particulière aux espaces latents possédant une structure désentrelacée. En effet, ce type de représentation se rapproche davantage de la manière dont fonctionne la représentation mentale humaine,

notamment en adoptant le principe des représentations compositionnelles, qui est l'un des piliers de la capacité de généralisation [Montero *et al.*, 2020]. En outre, disposer d'une représentation compacte et interprétable des informations présentes dans un jeu de données [Bengio *et al.*, 2013] permet le développement de modèles d'apprentissage profonds explicables et contrôlables [Gilpin *et al.*, 2018].

## 2.2 Aperçu des différentes définitions

Actuellement, il n'existe pas de consensus scientifique quant à une définition du désentrelacement. Cependant, cette notion repose sur l'idée que les données du monde réel peuvent être décrites par des facteurs génératifs indépendants les uns des autres, lesquels peuvent être découverts par des modèles d'apprentissage non supervisés. En d'autres termes, le désentrelacement cherche à identifier et à isoler ces différents facteurs génératifs, permettant ainsi une meilleure compréhension des informations contenues dans les données [Carbonneau *et al.*, 2022].

La définition la plus partagée est qu'un espace latent désentrelacé est une représentation pour laquelle le changement d'un facteur génératif des données d'entrées implique le changement d'uniquement une variable de l'espace latent [Locatello *et al.*, 2019c; Eastwood et Williams, 2018; Ridgeway et Mozer, 2018; Pati et Lerch, 2021].

En se fondant sur cette propriété, Rudin et al. interprètent le désentrelacement comme la capacité d'un réseau de neurones à faire circuler l'information relative à une caractéristique des images vers un endroit spécifique de l'architecture, tandis que l'information liée à une autre caractéristique transite par un autre endroit [Rudin *et al.*, 2022]. En d'autres termes, le désentrelacement permet d'identifier et d'isoler les voies de propagation de l'information dans le réseau de neurones pour des caractéristiques similaires, ce qui facilite son interprétation.

La définition de Carbonneau et al. présente trois propriétés observables et désirables du désentrelacement : la qualité explicite de la représentation, la compacité des facteurs génératifs et la modularité des variables latentes [Carbonneau *et al.*, 2022]. Ces propriétés définissent les critères pris en compte pour évaluer et mesurer l'efficacité du désentrelacement dans de nombreux travaux.

- La **qualité explicite** d'un espace latent correspond à la quantité et à la qualité des informations capturées dans celui-ci. En effet, une bonne représentation doit décrire complètement les variables d'entrée. Une qualité explicite parfaite implique qu'une relation généralisable entre les facteurs génératifs et les variables latentes a été apprise, c'est-à-dire qu'il est possible de prédire les facteurs génératifs uniquement à partir des variables latentes. Dans l'étude menée par Ridgeway et Mozer, les auteurs spécifient que cette relation doit être linéaire [Ridgeway et Mozer, 2018], contrairement à celle effectuée par Eastwood et Williams où la relation peut être non linéaire [Eastwood et Williams, 2018].

- La propriété de **modularité**, illustrée dans la Figure 2.2b, désigne le fait qu'une variable latente représente un seul facteur génératif. Dans cette Figure, l'espace latent est peu modulaire, car $z^3$ encode des informations propres à plusieurs facteurs génératifs.

- La propriété de **compacité** désigne le fait qu'un facteur génératif soit encodé dans au plus une variable latente. L'avantage d'une représentation compacte réside dans le fait que les facteurs génératifs soient séparés de manière distincte dans les dimensions de l'espace latent, ce qui facilite leur interprétation et leur utilisation séparément dans le processus de génération ou d'inférence. Dans la Figure 2.2a, illustrant cette propriété, l'espace latent est peu compact, car le facteur génératif $v^1$ est représenté dans plusieurs variables latentes.

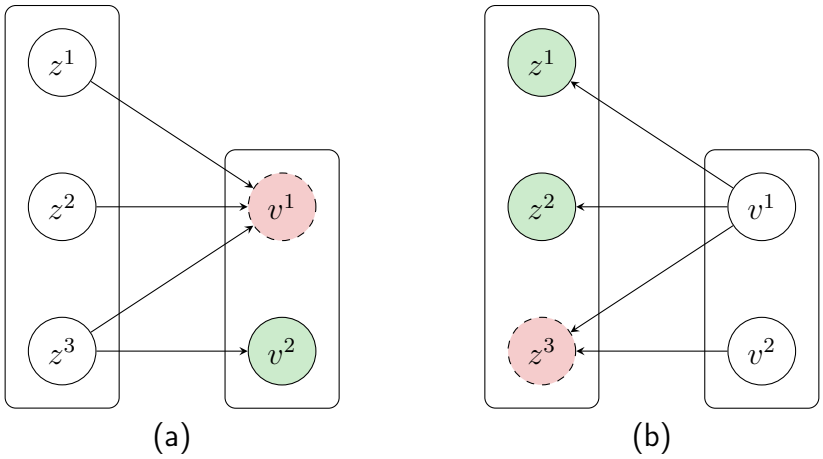



Fig 2.2: Schématisation d'un espace latent peu compact et peu modulaire. Les variables latentes sont dénotées par $\boldsymbol{z}$ et les facteurs génératifs par $\boldsymbol{v}$. (a) : Représentation de la capacité compacte d'un espace latent, caractérisé sur les facteurs génératifs. (b) : Représentation de la capacité modulaire d'un espace latent, caractérisé sur les variables latentes.

Une extension de cette définition est présentée par Locatello et al. Les auteurs proposent une décomposition sémantique approfondie du concept de la qualité explicite en distinguant les variables latentes "informatives" des variables latentes "simples" [Locatello *et al.*, 2020]. Dans l'étude effectuée par Zhu et al., le désentrelacement est modélisé à travers les hypothèse de constriction spatiale et de perception simple [Zhu *et al.*, 2021]. En effet, les auteurs soutiennent qu'une représentation est généralement interprétable s'il est possible de déterminer l'endroit où se trouvent les variations de l'image, et qu'une variable latente est déchiffrable si elle encode des variations simples. D'un autre côté, Suter et al. considèrent davantage le désentrelacement comme étant la propriété d'un procédé causal responsable de la génération des données [Suter *et al.*, 2019]. Cette approche repose sur la considération d'une structure causale sous-jacente à la génération des données, impliquant que certains facteurs génératifs puissent être encodés de façon dépendante. Enfin, Higgins et al. et Bouchacourt et al. proposent une définition du désentrelacement basée sur l'inclusion de groupes indépendants dans le sens topologique [Higgins *et al.*, 2018; Bouchacourt *et al.*, 2018]. Selon leur perspective, une représentation est désentrelacée si elle est cohérente avec les transformations qui caractérisent l'ensemble des données.

## 2.3 Méthodes de désentrelacement basées sur le VAE

Une étude présentée par les auteurs Locatello et al. a montré qu'obtenir un tel espace latent dans un VAE est impossible sans l'ajout de biais dans l'architecture et/ou le jeu de

données d'apprentissage [Locatello *et al.*, 2019b]. À partir de ce constat, de nombreuses approches proposent des modifications du VAE original proposé par Kingma et Welling [Kingma et Welling, 2014], que nous dénommons vanilla-VAE pour la suite du manuscrit.

Ces approches peuvent être classées en deux groupes distincts. D'une part, certaines ajoutent des biais induisant le désentrelacement de façon intrinsèque dans l'architecture des réseaux d'encodeur et de décodeur du VAE. C'est le cas du Spatial Broadcast Decoder [Watters *et al.*, 2019] ou du CoordConv Decoder [Liu *et al.*, 2018] qui permettent la dissociation entre les facteurs génératifs de position et les autres. En pavant les cartes de caractéristiques en sortie des couches de convolution de l'encodeur par des données spatiales, ces approches mettent à profit la propriété de l'invariance aux translations des réseaux de convolutions pour faire émerger l'information de position dans l'espace latent.

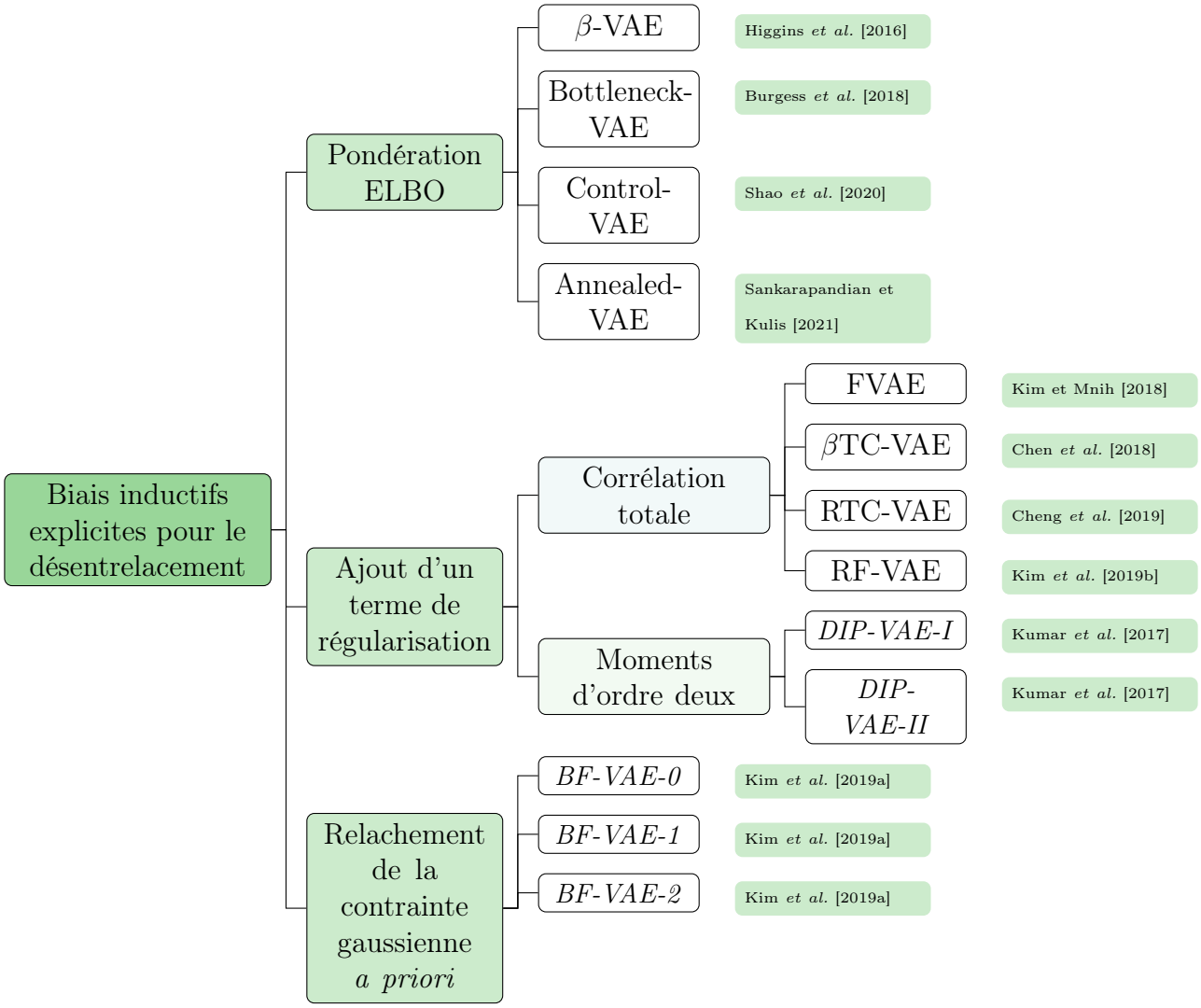



Fig 2.3: Taxonomie des méthodes de désentrelacement se basant sur des biais explicites. Ils peuvent être introduits à travers une pondération spécifique de la fonction coût, à travers une modélisation spécifique de la distribution *a priori* $p(z_k)$ selon la pertinence de la variable latente $z_k$.

D'autre part, des approches ajoutent des biais explicitement dans la fonction de coût, qui sont pris en compte lors de l'entraînement. Ce rapport se concentre sur ces dernières, dont nous proposons une décomposition en trois catégories distinctes : les méthodes basées sur une pondération de l' ELBO ; celles qui y ajoutent un terme de régularisation ; et celles

qui s'appuient sur une modification de la distribution *a priori*. Les modèles principaux de l'état de l'art, basés sur la taxonomie présentée dans la Figure 2.3, sont décrits dans la suite de ce chapitre.

### 2.3.1 Méthodes basées sur une pondération de l'ELBO

Dans le but d'encourager le désentrelacement de l'espace latent du vanilla-VAE, certaines méthodes visent à trouver un compromis entre la qualité de reconstruction de l'image en sortie du décodeur et la contrainte d'indépendance statistique des variables latentes. Pour cela, elles proposent d'ajuster de manière appropriée la pondération du terme de KL-divergence intervenant dans l'ELBO. Ces méthodes s'inspirent de l'interprétation du VAE basée sur la théorie de l'information, similaire à celles présentées dans [Tishby *et al.*, 2000; Chechik *et al.*, 2003; Alemi *et al.*, 2016].

Le modèle **Beta-VAE** [Higgins *et al.*, 2016] est une variante du vanilla-VAE dont l'objectif est d'encourager l'espace latent à être davantage factorisé aux dépens de la capacité d'encodage de l'information. De cette manière, les auteurs proposent d'aborder l'apprentissage comme une optimisation sous contrainte, de la manière suivante :

$$\max_{\boldsymbol{\phi},\boldsymbol{\theta}} = \mathbb{E}_{p_{data}(\boldsymbol{x})}\left[\mathbb{E}_{q_{\boldsymbol{\phi}(\boldsymbol{z}|\boldsymbol{x})}}[\log p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})]\right], \tag{2.1}$$

sujet à

$$D_{KL}[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})] < \boldsymbol{\epsilon}, \tag{2.2}$$

où $\epsilon$ est considérée comme étant la contrainte appliquée. En réécrivant l'équation (2.1) selon les conditions de Karush-Kuhn-Tucker (KKT), on obtient :

$$\mathcal{F}(\boldsymbol{\theta}, \boldsymbol{\phi}, \beta; \boldsymbol{x}, \boldsymbol{z}) = \mathbb{E}_{p_{data}(\boldsymbol{x})}\Big[\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\big[\log p_{\theta}(\boldsymbol{x}|\boldsymbol{z})\big] - \beta\big(D_{KL}\big[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})\big] - \epsilon\big)\Big]. \tag{2.3}$$

Dans cette approche, le multiplicateur KKT, $\beta$, agit comme un coefficient de régularisation sur la capacité d'encodage du réseau, en appliquant une contrainte d'indépendance implicite dû à la nature isotrope de l'*a priori* $p(\boldsymbol{z})$. Avec la condition de relâchement supplémentaire où $\beta, \epsilon > 0$, la fonction de coût à minimiser résultante devient :

$$\mathcal{F}(\boldsymbol{\theta}, \boldsymbol{\phi}, \beta; \boldsymbol{x}, \boldsymbol{z}) \leq \mathcal{L}_{\beta\text{vae}} = \mathbb{E}_{p_{data}(\boldsymbol{x})}\Big[\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\big[-\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\big] + \beta D_{KL}\big[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})\big]\Big]. \tag{2.4}$$

En particulier, lorsque $\beta$ est supérieur à 1, cela encourage davantage de variables latentes à avoir une moyenne inférée de 0 et une variance de 1, ce qui les fait converger vers la distribution *a priori*, les rendant non informatives. Ce phénomène est dénommé effondrement *a posteriori* [Lucas *et al.*, 2019a,b] et est particulièrement intéressante dans le cadre du désentrelacement, notamment pour les principes de modularité et de compacité. En effet, ces propriétés impliquent qu'un facteur génératif soit encodé dans au plus une variable latente et qu'une variable latente représente un unique facteur génératif. Dans le contexte dans lequel la dimension de l'espace latent est supérieure au nombre de facteurs génératifs, l'effondrement *a posteriori* permet de restreindre la dimension de $\boldsymbol{z}$ encodant l'information des données et favorise ainsi l'émergence des propriétés citées ci-dessus.

Les résultats de désentrelacement obtenus avec ce modèle se sont avérés bons en comparaison avec ceux du vanilla-VAE. Cependant, le $\beta$-VAE présente certaines limitations, notamment une moins bonne capacité de reconstruction. Cela est dû à un compromis qui pénalise la qualité de la reconstruction afin d'encourager le désentrelacement dans

les représentations latentes. En effet, plusieurs études ont démontré que le terme de KL-divergence, tel qu'il est écrit dans la fonction coût, peut faire apparaître un terme d'information mutuelle entre les données d'entrée et les variables latentes [Hoffman et Johnson, 2016] :

$$\begin{aligned}\mathbb{E}_{p_{data}(\boldsymbol{x})}[D_{KL}[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})]] &= D_{KL}[q_{\boldsymbol{\phi}}(\boldsymbol{z},\boldsymbol{x})||q_\phi(\boldsymbol{z})p_{data}(\boldsymbol{x})] \\ &\quad + D_{KL}[q_{\boldsymbol{\phi}}(\boldsymbol{z})||\prod_{j=1}^{K} q_\phi(\boldsymbol{z_j})] + \sum_{j=1}^{K} D_{KL}[q_{\boldsymbol{\phi}}(\boldsymbol{z_j})||p(z_j)] \\ &= I(\boldsymbol{z},\boldsymbol{x}) + TC(\boldsymbol{z}) + \sum_{j=1}^{K} D_{KL}[q_{\boldsymbol{\phi}}(\boldsymbol{z_j})||p(z_j)].\end{aligned} \tag{2.5}$$

Le premier terme représente une mesure d'information mutuelle entre les variables latentes $\boldsymbol{z}$ et les données d'entrée $\boldsymbol{x}$, tandis que le second terme correspond à la corrélation totale mesurée sur l'espace latent. Enfin, le dernier représente un terme de KL-divergence entre la distribution marginale des variables latentes par dimension, et la distribution *a priori*.

Afin de mieux concilier ces deux objectifs, de nombreuses variantes ont été proposées.

C'est le cas du **Bottleneck-VAE** [Burgess *et al.*, 2018]. La méthode définit une alternative consistant à conserver la contrainte appliquée sur le terme $D_{KL}$ telle que :

$$\mathcal{L}_{\text{Bottleneck-vae}} = \mathbb{E}_{p_{data}(\boldsymbol{x})}\Big[\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\big[-\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\big] + \gamma\big(D_{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})\right] - C\big)\Big]. \tag{2.6}$$

Pendant l'apprentissage, $\gamma$ est fixé à une valeur supérieure à 1, et $C$ est progressivement augmenté, ce qui permet d'accroître graduellement la capacité d'information stockée dans l'espace latent. Ce processus conduit à une amélioration de la reconstruction tout en préservant la capacité de désentrelacement.

Les auteurs Sankarapandian et Kulis se basent sur le constat que les critères ne constituent plus des bornes basses de l'*evidence* dès lors que $\beta > 1$ dans les modèles $\beta$-VAE et Bottleneck-VAE. Pour remédier à cela, ils proposent le modèle **Annealed-VAE** [Sankarapandian et Kulis, 2021], qui consiste à décroître progressivement la valeur de $\beta$ pendant l'apprentissage, passant de $\beta >> 1$ à $\beta = 1$. Cette approche permet d'optimiser une borne basse de l'*evidence* de manière plus efficace au cours des dernières étapes de l'apprentissage.

Enfin, Shao et al. proposent le modèle **ControlVAE** [Shao *et al.*, 2020] et traitent l'un des principaux défis des approches proposées jusqu'alors : trouver une pondération adéquate entre le terme de reconstruction et le terme de KL-divergence. Dans les méthodes préalablement décrites, cette pondération était déterminée de manière heuristique, en effectuant une recherche empirique des hyperparamètres permettant une meilleure convergence de la fonction coût. Pour résoudre ce problème, ils préconisent de modifier la valeur de $\beta$ pendant l'entraînement en utilisant une version alternative de l'algorithme "proportionnel, intégral, dérivé". [Åström et Hägglund, 2006]. Cela leur permet d'avoir un contrôle explicite sur la pondération du terme de KL-divergence, rendant ainsi le processus d'optimisation plus efficace et offrant une meilleure capacité de désentrelacement.

Cependant, trouver une pondération appropriée pour équilibrer la reconstruction et l'indépendance statistique des variables latentes reste un problème difficile. En conséquence, la pondération du terme de KL-divergence employé par ces méthodes impose la minimisation de ce terme d'information mutuelle, compromettant ainsi la propriété de qualité explicite de l'espace latent.

### 2.3.2 Méthodes basées sur l'ajout d'un terme de régularisation

La deuxième approche pour désentrelacer l'espace latent consiste à introduire un terme de régularisation dans la fonction de coût. Ces régularisations peuvent agir sur l'espace des pixels, comme dans l'architecture DAVA [Estermann et Wattenhofer, 2023], ou sur l'espace des variables latentes. Cette étude se concentre sur les régularisations appliquées à l'espace des variables latentes, telles que l'ajout d'un terme corrélation totale ou la manipulation de moments d'ordre deux des distributions d'intérêt.

#### Ajout d'un terme de corrélation totale

De nombreuses approches intègrent un terme de corrélation totale dans la fonction de coût pour quantifier le degré de factorisation de l'espace latent. Il est défini comme la KL-divergence entre la distribution jointe de l'espace latent et le produit de ses marginales, tel que :

$$TC(\boldsymbol{z}) = D_{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z})||\prod_{j=1}^{K} q_{\boldsymbol{\phi}}(z_j)\right] = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z})}\left[\log\left(\frac{q_{\boldsymbol{\phi}}(\boldsymbol{z})}{\prod_{j=1}^{K} q_{\boldsymbol{\phi}}(z_j)}\right)\right]. \tag{2.7}$$

De cette façon, une valeur faible de la corrélation totale indique un espace latent plus désentrelacé. Cette mesure est en accord avec les propriétés de modularité et de compacité du désentrelacement et a été utilisée dans les méthodes présentées ci-dessous.

Pour combiner le désentrelacement et la qualité de la reconstruction tout en évitant les problèmes liés à la pondération de la KL-divergence, le **factor-VAE** [Kim et Mnih, 2018] modifie la fonction coût en ajoutant un terme de corrélation totale. La formulation de la fonction coût résultante est définie telle que :

$$\mathcal{L}_{\text{fvae}} = \mathbb{E}_{p_{data}(\boldsymbol{x})}\Big[\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\big[-\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\big] + D_{KL}\big[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})\big]\Big] + \gamma D_{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z})||\prod_{j=1}^{K} q_\phi(z_j)\right]. \tag{2.8}$$

Cependant, en raison du grand nombre de composantes impliquées dans les mélanges de $q_{\boldsymbol{\phi}}(\boldsymbol{z})$ et $q_{\boldsymbol{\phi}}(z_j)$, le calcul direct du terme de corrélation totale devient impossible. Pour résoudre ce problème, les auteurs proposent deux astuces. Dans un premier temps, une technique basée sur la permutation des dimensions de l'espace latent obtenu à partir de différentes images permet d'obtenir des échantillons suivant la distribution de $q_{\boldsymbol{\phi}}(z_j)$. Dans un second temps, ils proposent d'utiliser l'astuce du ratio de densité [Sugiyama *et al.*, 2012; Rosca *et al.*, 2018]. Cette méthode consiste à estimer un rapport de densité incalculable en utilisant un discriminateur annexe. Plus précisément, un ratio $r(\boldsymbol{x}) = \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}$ peut être estimé à l'aide de données provenant de $p(\boldsymbol{x})$ et $q(\boldsymbol{x})$ et d'un classifieur probabiliste $D_\gamma(\boldsymbol{x})$, en utilisant l'estimateur $\hat{r}(\boldsymbol{x}) = \frac{D_\gamma(\boldsymbol{x})}{1-D_\gamma(\boldsymbol{x})}$. Le développement de cette astuce est proposé en Annexe A. Toutefois, cette approche implique l'apprentissage de paramètres $\gamma$ supplémentaires, ce qui complexifie l'architecture encodeur/décodeur du VAE.

Les auteurs Chen et al. proposent une approche similaire au Factor-VAE, mais sans nécessiter l'apprentissage de paramètres supplémentaires pour estimer le terme de corrélation totale. Ils présentent le modèle $\beta$**TC-VAE** [Chen *et al.*, 2018], qui repose sur la décomposition du terme de KL-divergence en trois composantes illustrée dans l'équation (2.5). Ainsi, pour éviter la minimisation du terme d'information mutuelle et l'utilisation d'un

discriminateur annexe, la fonction de coût suivante est proposée :

$$\begin{aligned}\mathcal{L}_{\beta\text{-TC}} = \mathbb{E}_{p_{data}(\boldsymbol{x})}\Big[\mathbb{E}_{q_\phi(\boldsymbol{z}|x)}\left[-\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right] + \alpha I(\boldsymbol{z},\boldsymbol{x})\Big] \\ + \beta D_{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z})||\prod_{j=1}^{K} q_{\boldsymbol{\phi}}(z_k)\right] + \gamma\sum_{j=1}^{K} D_{KL}\big[q_\phi(z_j)||p(z_j)\big].\end{aligned} \tag{2.9}$$

Dans cette approche, les valeurs de $\alpha$ et $\gamma$ sont fixées à 1 et l'hyperparamètre $\beta$ prend une valeur supérieure à 1. De plus, une méthode basée sur une estimation par Monte-Carlo, décrite dans la suite du manuscrit, est utilisée pour approcher les termes qui n'admettent pas une expression analytique.

Une étude de Locatello et al. indique qu'une valeur basse de la corrélation totale de la distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z})$ ne garantit pas nécessairement une faible corrélation totale sur la distribution des moyennes inférées par le décodeur $q(\boldsymbol{\mu}(\boldsymbol{z};\boldsymbol{\phi})|\boldsymbol{x})$ [Locatello *et al.*, 2019b]. Cela implique certaines limitations aux approches présentées ci-dessus. Ainsi, Cheng et al. proposent le modèle **RTC-VAE** [Cheng *et al.*, 2019], qui vise à améliorer le $\beta$TC-VAE. Ils préconisent alors l'ajout d'un terme de régularisation visant à minimiser la trace de la matrice de covariance de $\boldsymbol{z}$, aboutissant à la fonction de coût suivante :

$$\mathcal{L}_{\text{RTC}} = \mathcal{L}_{\beta\text{-TC}} + \eta\text{tr}\Big(\mathbb{E}_{p_{data}(\boldsymbol{x})}\Big[\text{Cov}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}[\boldsymbol{z}]\Big]\Big). \tag{2.10}$$

En minimisant ce terme de régularisation, les distributions des moyennes inférées et des variables latentes se rapprochent, offrant ainsi une solution au problème soulevé dans [Locatello *et al.*, 2019b]. Cette approche améliore le désentrelacement en assurant une meilleure concordance entre ces distributions.

Kim et al. proposent le modèle **Relevance Factor-VAE** [Kim *et al.*, 2019b], une approche hybride entre le $\beta$-VAE et le Factor-VAE. Les auteurs intègrent un terme de corrélation totale tout en traitant le compromis entre reconstruction et désentrelacement. De cette façon, ils suggèrent d'effectuer une séparation explicite de l'espace latent en deux sous-ensembles : un premier composé de variables latentes contenant de l'information des données d'entrée et un second contenant des variables latentes décorrélées de $\boldsymbol{x}$. Pour cela, ils introduisent une variable apprenable $r_j$, avec $j = [1, ..., K]$ associé à chaque variable latente et prenant la valeur 1 si elle est pertinente et la valeur 0 sinon. Cette variable est conjointement optimisée avec les paramètres de l'encodeur et du décodeur. La fonction de coût à minimiser est la suivante :

$$\begin{aligned}\mathcal{L}_{\text{RF-VAE}}(\{\boldsymbol{\theta},\boldsymbol{\phi}\},\boldsymbol{r}) = \mathbb{E}_{p_{data}(\boldsymbol{x})}\Big[\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\big[-\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\big] + \sum_{j=1}^{K}\lambda(r_j)D_{KL}\big[q_{\boldsymbol{\phi}}(\boldsymbol{z_j}|\boldsymbol{x})||p(\boldsymbol{z_j})\big]\Big] \\ + \gamma\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z})}\left[q_\phi(\boldsymbol{z}\cdot\boldsymbol{r})||\prod_{j=1}^{K} q_{\boldsymbol{\phi}}(z_j\cdot r_j)\right] + \eta_S||\boldsymbol{r}||_1 + \eta_H H(\boldsymbol{r}),\end{aligned} \tag{2.11}$$

avec $\boldsymbol{r} = (r_1, ..., r_k)$. La fonction $\lambda(r_j)$ est une fonction linéaire qui retourne des valeurs proches de 0 lorsque $r_j = 1$ et des valeurs proches de 1 sinon. Cela permet d'appliquer une contrainte différente sur le terme de KL-divergence selon que la variable latente est étiquetée par $r$ comme pertinente ou non. Les auteurs utilisent également l'astuce du ratio de log-densités à l'aide d'un discriminateur annexe, mais en fournissant en entrée du discriminateur des échantillons de $q_{\boldsymbol{\phi}}(\boldsymbol{z})$ pour lesquels les dimensions considérées comme

non pertinente sont éteintes grâce à $\boldsymbol{r}$. Finalement, le vecteur $\boldsymbol{r}$ est régularisé à l'aide une norme $L_1$ qui favorise l'émergence d'un nombre minimal de variables informatives, et par un terme d'entropie défini comme $H(\boldsymbol{r}) = -\sum_{j=1}^{K} r_j \log(r_j) + (1 - r_j)\log(1 - r_j)$. Les valeurs de pondérations $\eta_S$ et $\eta_H$ permettent de contrôler la force de régularisation appliquée à $r$ et ainsi de spécifier la dimension de l'espace latent encodant l'information de $\boldsymbol{x}$.

Ces approches présentent également des limitations. Le calcul de la corrélation totale peut être complexe, ralentissant ainsi l'entraînement et nécessitant plus de ressources. De plus, l'utilisation d'estimateurs basés sur l'astuce du ratio de densité peut introduire des biais difficiles à quantifier.

**Méthodes basées sur les moments d'ordre deux des distributions inférées**

Pour éviter d'approximer le terme de corrélation totale, le modèle **DIP-VAE** [Kumar *et al.*, 2017] propose une approche alternative pour favoriser le désentrelacement. Les auteurs se basent sur l'hypothèse que minimiser une distance entre la distribution jointe de l'espace latent $q_\phi(\boldsymbol{z})$ et la distribution *a priori* gaussienne isotrope $p(\boldsymbol{z})$ permet d'obtenir une représentation souhaitable. De cette façon, ils proposent de minimiser une distance appropriée $\mathcal{D}(q_{\boldsymbol{\phi}}(\boldsymbol{z}), p(\boldsymbol{z}))$ pour définir leur fonction de coût :

$$\mathcal{L}_{\text{Dipvae}} = \mathbb{E}_{p_{data}(\boldsymbol{x})}\Big[\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\big[-\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\big] + D_{KL}\big[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})\big]\Big] + \lambda\mathcal{D}\big(q_\phi(\boldsymbol{z}), p(\boldsymbol{z})\big). \tag{2.12}$$

Pour minimiser la distance $\mathcal{D}$, les auteurs adoptent une approche qui consiste à aligner les moments des deux distributions. Cependant, pénaliser les termes hors diagonale contraint conjointement la valeur des éléments diagonaux. Pour résoudre cette limitation, un second terme est ajouté en tant que compensation, résultant en la fonction de coût suivante :

$$\begin{aligned}\mathcal{L}_{\text{Dipvae}_{II}} = \mathbb{E}_{p_{data}(\boldsymbol{x})}\Big[\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\big[-\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\big] + D_{KL}\big[q_{\boldsymbol{\phi}}(\boldsymbol{z}|x)||p(\boldsymbol{z})\big]\Big] + \lambda_{od}\sum_{i\neq j}\big(\text{Cov}_{q_{\boldsymbol{\phi}}(\boldsymbol{z})}[\boldsymbol{z}]_{ij}^2\big) \\ + \lambda_d\sum_i\big(\big[\text{Cov}_{q_{\boldsymbol{\phi}}(\boldsymbol{z})}[\boldsymbol{z}]\big]_{ii}^2 - 1\big)^2,\end{aligned} \tag{2.13}$$

où $\lambda_d$ et $\lambda_{od}$ sont des hyperparamètres permettant de régler la force de régularisation des termes portant sur la matrice de covariance.
Toutefois, si le réseau d'encodeur infère les paramètres d'une distribution gaussienne $\Sigma(\boldsymbol{x};\boldsymbol{\phi})$ et $\boldsymbol{\mu}(\boldsymbol{x});\boldsymbol{\phi}$, alors la matrice de covariance de $\boldsymbol{z}$ est définie comme suit :

$$\text{Cov}_{q_{\boldsymbol{\phi}}(\boldsymbol{z})}[\boldsymbol{z}] = \mathbb{E}_{p_{data}(\boldsymbol{x})}\big[\Sigma(\boldsymbol{x};\boldsymbol{\phi})\big] + \text{Cov}_{p_{data}(\boldsymbol{x})}\big[\boldsymbol{\mu}(\boldsymbol{x};\boldsymbol{\phi})\big]. \tag{2.14}$$

En utilisant une matrice diagonale $\Sigma(\boldsymbol{x};\boldsymbol{\phi})$ pour représenter les covariances, les corrélations croisées entre les variables latentes proviennent du terme $\text{Cov}_{p_{data}(\boldsymbol{x})}[\mu(\boldsymbol{x};\boldsymbol{\phi})]$ dans l'équation (2.14), résultant en une seconde proposition de fonction de coût définie de la façon suivante :

$$\begin{aligned}\mathcal{L}_{\text{Dipvae}_I} = \mathbb{E}_{p_{data}(\boldsymbol{x})}\Big[\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\big[-\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\big] + D_{KL}\big[q_{\boldsymbol{\phi}}(\boldsymbol{z}|x)||p(\boldsymbol{z})\big]\Big] \\ \lambda_{od}\sum_{i\neq j}\big(\text{Cov}_{p_{data}(\boldsymbol{x})}[\mu(\boldsymbol{x};\boldsymbol{\phi})]_{ij}^2\big) + \lambda_d\sum_i\big(\big[\text{Cov}_{p_{data}(\boldsymbol{x})}[\mu(\boldsymbol{x};\boldsymbol{\phi})]\big]_{ii}^2 - 1\big)^2,\end{aligned} \tag{2.15}$$

Dans l'équation (2.13), la régularisation proposée ne porte pas directement sur les moyennes inférées par l'encodeur, mais sur les variables latentes, contrairement à l'approche proposée

dans l'équation (2.15). Cela implique indirectement que pour compenser la réduction des éléments diagonaux engendrée par ce terme, l'optimisation du DIP-VAE-II favorise une concentration de l'information dans un sous-espace de l'espace latent. Cette méthode est ainsi plus adaptée lorsque la taille définie pour l'espace latent est supérieure au nombre de facteurs génératifs.

En outre, trouver un équilibre approprié entre les deux hyperparamètres, $\lambda_d$ et $\lambda_{od}$, constitue un défi pour ces approches, car ils doivent se compenser efficacement : l'un contribue à faire tendre $\boldsymbol{\mu}(\boldsymbol{x};\boldsymbol{\phi})$ vers 0, tandis que l'autre contrebalance cela. Cette formulation peut rapidement amener le modèle à se retrouver coincé dans des minimums locaux lors du processus d'optimisation. Il est donc crucial de bien ajuster ces hyperparamètres pour obtenir un désentrelacement efficace de l'espace latent tout en maintenant une bonne qualité de reconstruction.

### 2.3.3 Méthodes basées sur une modification de la distribution *a priori*

La plupart des méthodes présentées jusqu'à présent modélisent l'*a priori* et l'*a posteriori* à travers des distributions unimodales, généralement gaussiennes. Cependant, ces choix de modélisations simples peuvent s'avérer restrictifs pour modéliser de façon adéquate la distribution réelle des variables latentes. En effet, le choix d'une distribution unimodale favorise que l'ensemble des variables latentes, celles comportant de l'information ou non, soient identiquement distribuées. Pour remédier à cela, certaines approches définissent des distributions *a priori* non gaussiennes [Van Den Oord *et al.*, 2017; Tomczak et Welling, 2018; Casale *et al.*, 2018; Razavi *et al.*, 2019; Bauer et Mnih, 2019; Mathieu *et al.*, 2019]. Dans l'article [Kim *et al.*, 2019a], les auteurs supposent que pour obtenir un espace latent bien désentrelacé, il est nécessaire de traiter de façon séparée les variables qui ne contiennent pas d'information des données de celles encodant leur variabilité.

Dans un premier temps, les auteurs proposent le modèle **BF-VAE-0**, qui a pour objectif de relâcher la contrainte gaussienne sur la distribution *a priori*. Pour cela, la variance de $p(\boldsymbol{z})$ devient une variable à apprendre et est optimisée conjointement avec les paramètres $\phi$ et $\theta$ de l'encodeur et du décodeur. Cela aboutit à la fonction de coût suivante :

$$\begin{aligned}\mathcal{L}_{\text{BF-VAE-0}}(\{\boldsymbol{\phi},\boldsymbol{\theta}\},\boldsymbol{\alpha}) = \mathbb{E}_{p_{data}(\boldsymbol{x})}\Bigg[\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\big[-\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\big] + \sum_{j=1}^{K} D_{KL}\big[q_{\boldsymbol{\phi}}(\boldsymbol{z_j}|\boldsymbol{x})||\mathcal{N}(\boldsymbol{z}_j;0,\boldsymbol{\alpha}_j^{-1})\big]\Bigg] \\ + \gamma\text{TC}(\boldsymbol{z}) + \eta\sum_{j=1}^{K}\left(\boldsymbol{\alpha}_j^{-1}-1\right)^2,\end{aligned} \tag{2.16}$$

où $\boldsymbol{\alpha}^{-1}$ représente la variable apprenable qui modélise la précision de la distribution normale *a priori*, et $\text{TC}(\boldsymbol{z})$ est l'estimation de la corrélation totale obtenue à l'aide de l'astuce du ratio de log-densités. Une norme $L_2$ régularise la variable de précision, ce qui permet faire en sorte que la plupart des $\alpha$ soient proches de 0 tout en contrôlant leur cardinalité à travers l'ajustement de l'hyperparamètre $\eta$.

Dans un second temps, les auteurs proposent de modéliser la loi *a priori* comme une loi Gamma afin de relâcher la contrainte gaussienne. Cette dernière est définie de la façon

suivante :

$$p(\boldsymbol{\alpha}) = \prod_{j=1}^{K} p(\alpha_j) = \prod_{j=1}^{K} \mathcal{G}(\alpha_j; a_j, b_j), \tag{2.17}$$

où $\mathcal{G}(\alpha; a, b)$ représente une distribution Gamma de paramètres $(a, b)$ et $K$ la dimension de l'espace latent. Ils proposent alors de se reposer sur un cadre bayésien variationnel pour définir leur fonction coût, modélisée de la façon suivante :

$$\begin{aligned}\mathcal{L}_{\text{BF-VAE-1}}(\{\boldsymbol{\phi}, \boldsymbol{\theta}\}, \boldsymbol{a}, \boldsymbol{b}, \hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}) = \mathbb{E}_{p_{data}(\boldsymbol{x})}\Big[\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\big[-\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\big] + \mathbb{E}_{q_{(\boldsymbol{\alpha})}}\big[D_{KL}\big[q_{\phi}(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z}|\boldsymbol{\alpha})\big]\Big]\Big] \\ + \frac{1}{N} D_{KL}\big[q(\boldsymbol{\alpha})||p(\boldsymbol{\alpha})\big] + \mathbb{E}_{q_{(\boldsymbol{\alpha})}}\left[D_{KL}\big[q(\boldsymbol{z}|\boldsymbol{\alpha})||\prod_{j=1}^{K} q(z_j|\alpha_j)\big]\right].\end{aligned} \tag{2.18}$$

Dans cette modélisation, l'optimisation est réalisée de manière conjointe sur les paramètres de l'encodeur et du décodeur $\boldsymbol{\phi}$ et $\boldsymbol{\theta}$, ainsi que sur les paramètres apprenables supplémentaires $\{a_j, b_j, \hat{a}_j, \hat{b}j\}_{j=1}^{K}$, à travers une optimisation par descente de gradient stochastique. Toutefois, le problème semble mal posé. En effet, les paramètres des distributions *a priori* et *a posteriori*, respectivement $(a, b)$ et $(\hat{a}, \hat{b})$ sont appris, mais ne sont pas liés aux données d'entrées, ce qui donne lieu à des écritures abusives et des comportements aberrants.

Afin de remédier à la limitation précédente, les auteurs Kim et al. proposent une troisième modélisation dans laquelle une variable supplémentaire $\boldsymbol{r}$ est apprise [Kim *et al.*, 2019a]. Cette dernière correspond à un curseur permettant d'affiner la loi *a priori* sur $\alpha$ en autorisant une distribution à queue plus ou moins lourde en fonction de la pertinence de la variable latente traitée. La fonction coût de cette approche est définie de la façon suivante :

$$\begin{aligned}\mathcal{L}_{\text{BF-VAE-2}}(\{\boldsymbol{\phi}, \boldsymbol{\theta}\}, \boldsymbol{r}, \hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}) = \mathbb{E}_{p_{data}(\boldsymbol{x})}\Big[\mathbb{E}_{q_{\phi}(\boldsymbol{z}|\boldsymbol{x})}\big[-\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\big]\Big] + \frac{1}{N} D_{KL}\big[q(\boldsymbol{\alpha})||p(\boldsymbol{\alpha}|\boldsymbol{r})\big] \\ + \mathbb{E}_{q(\boldsymbol{\alpha})}\Big[\mathbb{E}_{p_{data}(\boldsymbol{x})}\big[D_{KL}[q_{\boldsymbol{\phi}}(\boldsymbol{z}|x)||p(\boldsymbol{z}|\boldsymbol{\alpha}, \boldsymbol{r})]\big]\Big] \\ + \gamma D_{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z} \cdot \boldsymbol{r})||\prod_{j \in \boldsymbol{r}} q_{\phi}(z_j)\right] + \eta_S \|\boldsymbol{r}\|_1 + \eta_H H(\boldsymbol{r}).\end{aligned} \tag{2.19}$$

La distribution *a priori* $p(\boldsymbol{z}, \boldsymbol{\alpha}|\boldsymbol{r})$ est formulée comme un produit de distributions Normales–Inverse-Gamma $\mathcal{N}(z_k; 0, \alpha_k)\mathcal{G}^{-1}(a_k; \frac{1+2\epsilon}{r_k+\epsilon}, \frac{1+2\epsilon}{r_k+\epsilon} - 1)$ pour chaque variable latente $k$. Cette formulation permet d'ajuster la distribution $p(\boldsymbol{z}|\alpha)$ en fonction du facteur de pertinence $r_k$ associé à la variable latente $z_k$. Lorsque $r_k$ est élevé, la distribution *a priori* tend vers une distribution à queue lourde, tandis qu'elle se rapproche d'une distribution gaussienne avec une covariance identité lorsque $r_k$ est faible. Afin de favoriser davantage le désentrelacement de l'espace latent, un terme de corrélation totale est ajouté à la fonction de coût, estimé à l'aide de l'astuce du ratio de log-densités. Le produit scalaire $(z \cdot r)$ est utilisé pour renforcer spécifiquement l'indépendance des variables latentes dont la modélisation *a priori* est définie comme une distribution à queue plus lourde, encourageant davantage le désentrelacement de l'espace latent spécifiquement sur ces variables.

Le modèle **BF-VAE-2** se démarque des autres approches à la fois pour sa capacité à relacher la contrainte de normalité sur l'*a priori*, mais également à traiter de façon distincte des variables considérées comme plus ou moins pertinentes. Néanmoins, outre les problèmes de biais déjà relevés propres à l'estimation du terme de corrélation totale à l'aide d'un discriminateur annexe, d'autant plus importants lorsque certaines dimensions sont mises à 0, cette dernière modélisation possède d'autres limitations. Le facteur de pertinence $\boldsymbol{r}$ étant décorrélé des données d'entrées, ce dernier est uniquement contrôlé par les valeurs définies pour les hyperparamètres $\eta_S$ et $\eta_H$. De plus, le cadre bayésien variationnel utilisé pour modifier la fonction coût donne lieu à des écritures abusives et des comportements aberrants observés après apprentissage.

### 2.3.4 Synthèse

L'analyse des différentes approches de désentrelacement proposées dans l'état de l'art a permis de mettre en exergue leurs avantages et inconvénients, résumés dans le tableau 2.1.
On peut notamment citer la minimisation du terme d'information mutuelle entre les données d'entrées et les variables latentes, qui est un défaut important de certaines approches induisant une baisse de la propriété de qualité explicite. De plus, la nécessité d'utiliser un réseau de discrimination pour estimer le terme de corrélation totale ajoute un ensemble de paramètres supplémentaires à apprendre, complexifiant davantage l'architecture.
Toutefois, l'ajout d'un terme de régularisation apparaît être une approche largement utilisée et qui porte ses fruits. De plus, relâcher la contrainte de normalité sur la distribution *a priori* selon la pertinence de la variable latente considérée semble être une approche rarement utilisée et prometteuse.

| Propriétés | $\beta$-VAE | FVAE | TCVAE | C-VAE | A-VAE | RTC VAE | DIP VAE | BF VAE2 |
|---|---|---|---|---|---|---|---|---|
| $I(\boldsymbol{z}, \boldsymbol{x})$ non minimisée | × | × | × | × | × | × | × | ✓ |
| Relachement de la contrainte sur $p(\boldsymbol{z})$ | × | × | × | × | × | ✓ | × | ✓ |
| Terme de régularisation | × | ✓ | ✓ | × | × | ✓ | ✓ | ✓ |
| Pas de réseau annexe | ✓ | × | ✓ | ✓ | ✓ | × | ✓ | × |
| Borne basse de l'*evidence* minimisée | × | × | × | ✓ | × | × | × | × |
| Pas de nécéssité d'hyperparamétrage | × | × | × | × | × | × | × | × |

Tableau 2.1: Synthèse des propriétés intéréssantes de modélisation prises en compte dans les architectures de l'état de l'art pour un objectif de désentrelacement de l'espace latent. Afin de simplifier la compréhension du tableau, les acronymes "A-VAE" et "C-VAE" on été utilisé pour spécifier respectivement les modèles Annealed-VAE et Control-VAE.

Cette analyse permet de mettre en avant les points suivants :

- La fonction optimisée pour l'ensemble des méthodes dont l'objectif est d'obtenir une représentation désentrelacée au sein de l'espace latent ne représente plus une borne

basse de l'*evidence*. En effet, l'ajout de terme de régularisation ou de pondération sur le critère modifie l'objectif à optimiser.

- La dimension de l'espace latent est sensible et difficile à ajuster. Une hypothèse commune semble de définir le vecteur $\boldsymbol{z}$ de grande taille, et après apprentissage de trouver le meilleur hyperparamètre permettant de restreindre la capacité d'encodage de $\boldsymbol{z}$ afin qu'elle corresponde au nombre de facteurs génératifs.

- Le terme de divergence de Kullback dans la fonction coût, telle qu'elle est exprimée dans l'ELBO, fait émerger un terme d'information mutuelle entre $\boldsymbol{z}$ et $\boldsymbol{x}$. Appliquer la contrainte de régularisation trop forte sur ce terme impact négativement la capacité de reconstruction du modèle.

- Une modélisation spécifique de la variable latente selon sa pertinence repose sur une modification de la distribution *a priori*, ce qui implique que le biais intinsèque lié au désentrelacement n'est pas corrélé avec les données d'entrées et est donc totalement dépendant des hyperparamètres.

Les sections suivantes du manuscrit ont pour objectif d'effectuer une analyse des mesures utilisées pour la comparaison des capacités de désentrelacement entre les modèles. Par la suite, une revue des différents jeux de données utilisés pour évaluer le désentrelacement est réalisée.

## 2.4 Évaluation de la capacité de désentrelacement

Afin d'évaluer la capacité de désentrelacement des différentes approches présentées dans la section précédente, de nombreuses mesures sont proposées dans l'état de l'art. Initialement, les comparaisons de modèles se fondaient sur des approches subjectives et qualitatives du désentrelacement, notamment à travers l'analyse d'un parcours de l'espace latent, illustré dans la Figure 2.4.

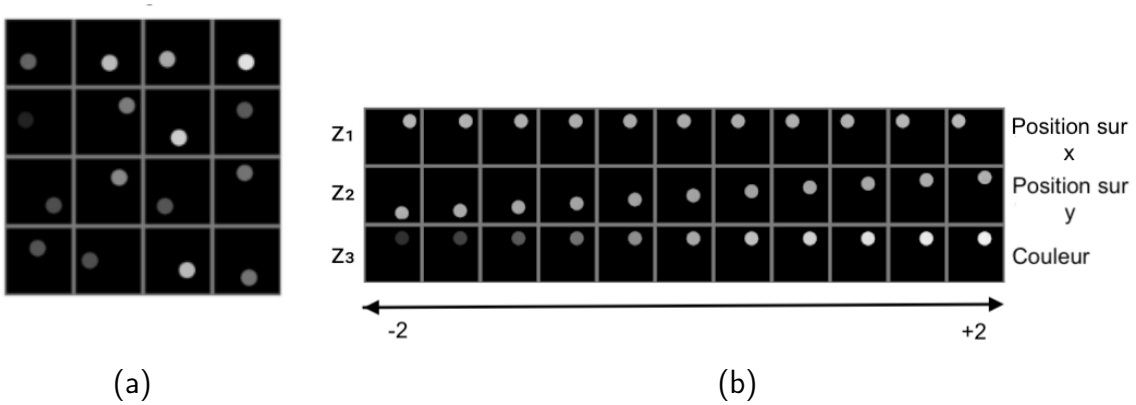



Fig 2.4: Illustration du parcours de l'espace latent sur un jeu de données simple. Figure 2 dans [Higgins *et al.*, 2018] (a) : Exemples d'images issues d'un jeu de données généré à partir de trois facteurs génératifs : le niveau de gris du cercle, sa position sur $x$ et sa position sur $y$. (b) : Illustration de parcours de l'espace latent permettant d'évaluer la capacité de désentrelacement d'un modèle ayant appris sur les images présentées dans (a).

Le parcours de l'espace latent consiste à parcourir la valeur de moyenne inférée pour chaque variable latente dans un certain intervalle, tout en maintenant les autres valeurs de moyennes constantes, avant de reconstruire l'image en passant l'espace latent modifié dans le décodeur. En comparant l'image initiale avec celle reconstruite, on peut déterminer si la variable latente contient de l'information, et dans le cas d'un espace désentrelacé, quel facteur génératif elle encode. Si les images sont similaires, alors la variable latente est considérée comme non informative.
Cette technique est illustrée dans la Figure 2.4b, où chaque ligne correspond à une dimension spécifique de l'espace latent, de la première à la troisième. Dans les colonnes sont représentées des images reconstruites par le décodeur lorsque l'on vient modifier la valeur de $\mu_k(\boldsymbol{x}; \boldsymbol{\phi})$ où $k$ correspond à la dimension spécifiée par la ligne. Cette modification est effectuée de façon linéaire en ajoutant un nombre compris dans $[-2\sigma, 2\sigma]$ à $\mu_k(\boldsymbol{x}; \boldsymbol{\phi})$. On remarque que la première dimension, dénotée par $z_1$, encode les informations propre au facteur génératif de la position sur $x$, $z_2$ encode le facteur génératif propre à la position sur $y$, tandis que $z_3$ encode le niveau de gris de l'objet. Cependant, cette mesure repose sur une évaluation subjective de la capacité de désentrelacement du modèle, ce qui a motivé la recherche de méthodes quantitatives plus fiables.

Parmi les approches quantitatives, certaines sont qualifiées de non supervisées, car elles ne nécessitent pas la connaissance des facteurs génératifs spécifiques à chaque image [Do et Tran, 2019; Duan *et al.*, 2019; Liu *et al.*, 2020]. Elles s'appuient sur des critères statistiques pour évaluer l'indépendance des variables latentes, la séparabilité des facteurs génératifs, ou d'autres propriétés du désentrelacement. Dans ce manuscrit, l'étude ne s'attarde pas sur ces dernières, car elles sont moins couramment utilisées dans l'état de l'art.

Les métriques qualifiées de supervisées, quant-à-elles, peuvent être classées en trois approches distinctes, en accord avec la taxonomie proposée par Carbonneau et al. [Carbonneau *et al.*, 2022], présentée dans la figure 2.5. La première nécessite une intervention directe sur l'espace latent pour mesurer des propriétés spécifiques telles que la compacité ou la modularité de l'espace latent. La deuxième repose sur l'utilisation d'un modèle de prédiction annexe pour évaluer la qualité du désentrelacement. Enfin, la troisième est basée sur la théorie de l'information, où des mesures d'information mutuelle ou d'entropie sont utilisées pour évaluer l'indépendance statistique entre les variables latentes et les facteurs génératifs. Chacune de ces approches possède ses avantages et ses limites, et le choix de la métrique dépend souvent du contexte spécifique de l'application et des propriétés désirées dans l'espace latent. C'est pourquoi également il est intéréssant de les calculer conjointement.

Dans la suite du rapport, certaines de ces mesures sont présentées, en mettant en évidence leurs avantages et leurs inconvénients. Les indices $i$ et $j$ sont respectivement utilisés pour désigner l'indice du facteur génératif, c'est-à-dire $v_i \in \{v_1, v_2, ..., v_M\}$ et l'indice de la variable latente, c'est-à-dire $z_j \in \{z_1, z_2, ..., z_K\}$.

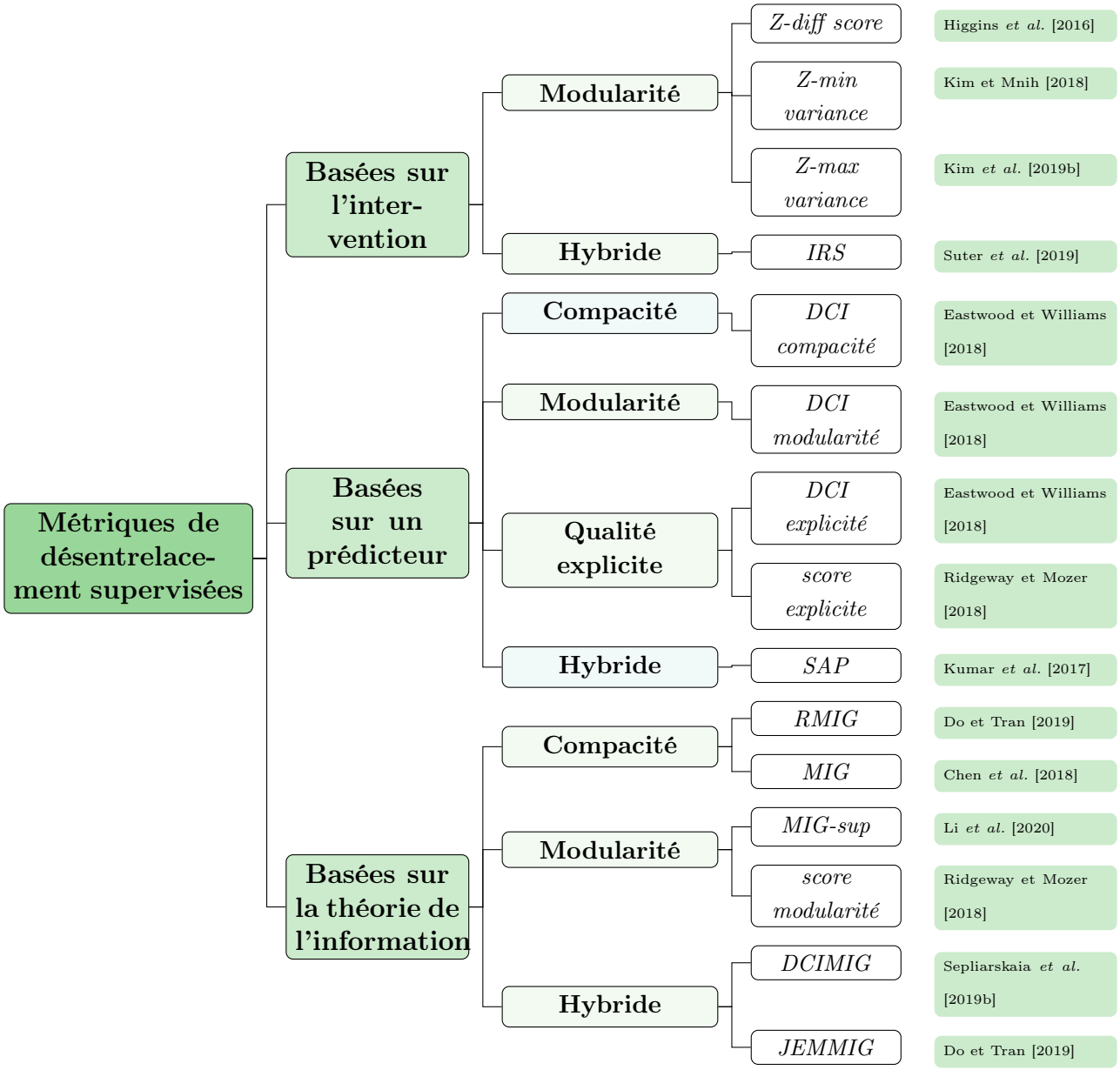



Fig 2.5: Taxonomie des métriques de désentrelacement proposée par [Carbonneau *et al.*, 2022]. Une métrique qualifiée d'"hybride" quantifie plusieurs propriétés du désentrelacement.

### 2.4.1 Métriques basées sur une transformation

Les métriques basées sur une transformation consistent à comparer des variables latentes en créant des sous-espaces de variables pour lesquels un ou plusieurs facteurs sont gardés constants. Certaines d'entre elles nécessitent l'entraînement d'un réseau en sortie du VAE permettant la prédiction des facteurs génératifs à partir des variables latentes, après avoir effectué la transformation dans l'espace d'intérêt. Des scores sont alors calculés en comparant les facteurs génératifs prédits par ce réseau supplémentaire à partir des variables latentes. Ces métriques fournissent une évaluation objective du désentrelacement en mesurant la capacité du modèle à séparer les facteurs génératifs et à les représenter de manière indépendante dans l'espace latent. Cela permet également d'évaluer la capacité de généralisation du modèle pour des tâches ultérieures. Cependant, il est important de noter que ces métriques ne fournissent qu'une évaluation partielle du désentrelacement, et qu'il est nécessaire de les compléter par d'autres métriques objectives pour avoir une

évaluation plus complète et robuste des performances du modèle.

Parmi ces métriques, celle du $\beta$-**VAE**, également appelé **z-diff score** [Higgins *et al.*, 2016], repose sur l'intuition des auteurs selon laquelle si un facteur génératif $v_i$ est fixé lors de la création des données d'entraînement, tandis que les autres facteurs sont échantillonnés de manière aléatoire, alors la valeur de la variable latente $z_j$ encodant le facteur génératif fixé fluctuera moins que les autres. Sur la base de ce postulat, les auteurs proposent l'entraînement d'un modèle de classification linéaire basé sur un ensemble d'entraînement composé de données et d'étiquettes $(\tilde{\boldsymbol{z}}, \boldsymbol{v})$, où $\tilde{\boldsymbol{z}}$ correspond à l'espace latent transformé.

La métrique **Factor-VAE** ou $z$**-min variance** [Kim et Mnih, 2018] repose sur la même hypothèse que précédemment. Toutefois, les auteurs proposent une amélioration de cette dernière pour éviter un dysfonctionnement où la justesse du classifieur linéaire est de 100% dès lors que $K-1$ sur $K$ variables latentes sont désentrelacées. Pour éviter ce problème, la construction de l'ensemble d'entraînement et le type de modèle de classification diffèrent. La construction de ce nouvel ensemble d'entraînement est explicité dans l'algorithme 2.1 ci-dessous.

**Algorithm 2.1** *Construction de l'ensemble d'entraînement pour la métrique Factor-VAE*

Calcul de l'écart-type empirique $s_k$ pour chaque dimension $k$ de l'espace latent
**for** $j \in [1, \ldots, M]$ **do** ▷ Boucle sur les facteurs génératifs
    Initialisation d'un tableau $variance[k]$
    **for** $i \in [1, \ldots, L]$ **do** ▷ Boucle sur les données
        Simulation de $\boldsymbol{x}_j^{(i)}$, où le facteur $j$ fixé et les autres varient
        Encodage de $\boldsymbol{x}_j^{(i)}$ pour obtenir sa représentation latente $\boldsymbol{z}_j^{(i)}$
        **for** $k \in [1, \ldots, K]$ **do** ▷ Boucle sur les dimensions de z
            Normalisation de $\boldsymbol{z}_j^{(i)}[k]$ par $s_k$
            Accumulation de la variance pour chaque dimension :
            $variance[k]$ += $(\boldsymbol{z}_j^{(i)}[k]/s_k)^2$
        **end for**
    **end for**
    Calcule de la variance empirique pour chaque dimension : $variance[k]/ = L$
    Récupération de l'indice de la dimension avec la plus faible variance, noté $k^*$
    Création d'une instance d'entraînement : $(k^*, j)$
**end for**
Rassemblement des instances d'entraînement pour former l'ensemble d'entraînement complet

Une fois l'ensemble d'entraînement défini, un classifieur par vote majoritaire est entraîné pour déterminer l'indice du facteur génératif $j$ qui produit le plus grand nombre de faibles variances pour la variable latente $k$. La mesure retournée correspond à la justesse du classifieur. L'utilisation de ce type de classifieur permet de se défaire du besoin d'hyperparamètre, normalement nécessaire à l'apprentissage d'un classifieur linéaire, comme c'est le cas pour la métrique du $\beta$-VAE, mais également d'apprendre des combinaisons non linéaires entre les variables latentes et les facteurs génératifs.

La métrique appelée **z-max variance** ou encore métrique **RF-VAE** [Kim *et al.*, 2019b] est similaire à la métrique précédente, mais avec un changement dans l'hypothèse de

départ. Les auteurs partent du postulat que si tous les facteurs génératifs sont fixés à l'exception d'un seul, alors la variable latente encodant le facteur génératif laissé libre devrait posséder une variance plus importante. Le **score de robustesse interventionnelle** (IRS) [Suter *et al.*, 2019] est une mesure qui diffère des métriques précédentes, car elle ne nécessite pas l'utilisation d'un réseau de classification. Cette approche se base sur un calcul de distance entre deux ensembles de variables latentes, avant et après modification des facteurs génératifs. L'intuition des auteurs pour cette métrique est que le changement d'un facteur génératif $v_i$ ne devrait pas avoir d'impact sur une variable latente encodant un autre facteur génératif $v_k$ avec $k \neq i$.

### 2.4.2 Métriques basées sur un prédicteur

Les métriques qui se basent sur un prédicteur utilisent les espaces latents sans y effectuer de modification afin d'entraîner des modèles de régression ou de classification annexes dont l'objectif est de prédire les facteurs génératifs $\boldsymbol{v}$ à partir des variables latentes $\boldsymbol{z}$. L'analyse de ces réseaux permet de vérifier l'utilité de chacune des variables latentes pour la prédiction des facteurs génératifs. Une variable latente qui possède une forte corrélation avec un facteur génératif donné est considérée comme explicite et utile pour le désentrelacement.

La mesure de ***Qualité explicite (DCI)*** [Eastwood et Williams, 2018] repose sur une distance normalisée entre les facteurs génératifs de l'ensemble de test et les facteurs prédits. Généralement, l'erreur quadratique moyenne est utilisée pour calculer cette distance. Une représentation est considérée comme non informative lorsque la distance obtenue est supérieure à l'erreur quadratique théorique entre deux variables indépendantes ($X$ et $Y$) distribuées uniformément sur l'intervalle $[0, 1]$, ce qui donne $E[(X - Y)^2] = \frac{1}{6}$. Ainsi, la qualité explicite est calculée comme suit :

$$DCI_I = \frac{1}{M} \sum_{i=1}^{M} 1 - 6 * \mathbb{E}\big[(v_i - \hat{v}_i)^2\big], \tag{2.20}$$

où $M$ est le nombre de facteurs génératifs, $v_i$ représente le facteur génératif réel de l'ensemble de test, et $\hat{v}_i$ est la prédiction du facteur génératif par le modèle à partir des variables latentes. Un score de qualité explicite élevé indique que les variables latentes du modèle contiennent des informations explicites sur les facteurs génératifs, ce qui est souhaitable pour un modèle de désentrelacement. En revanche, un score faible suggère que les variables latentes ne sont pas suffisamment informatives.

La mesure de ***compacité (DCI)*** vise à déterminer si un facteur génératif est principalement décrit par une seule variable latente. Son calcul dépend d'une matrice de probabilité de relation $p_{ij}$ obtenue comme suit :

$$p_{ij} = \frac{R_{ij}}{\sum_{k=1}^{K} R_{ik}}, \tag{2.21}$$

où $R_{ij}$ représente l'importance prédite du facteur $i$ dans la variable $j$ basé sur l'importance de Gini, et $K$ correspond à la dimension de l'espace latent. Ensuite, une mesure basée sur l'entropie quantifie à quel point un facteur génératif est présent dans toutes les variables latentes ou une seule, avant d'être normalisée sur l'ensemble des $M$ facteurs, ce qui donne :

$$C = \frac{1}{M} \sum_{i=1}^{M} \left( 1 + \sum_{j=1}^{K} p_{ij} \log_K p_{ij} \right). \tag{2.22}$$

Un score élevé de compacité indique que chaque facteur génératif est principalement représenté par une seule variable latente. En revanche, un score faible suggère que les facteurs génératifs sont distribués de manière diffuse sur plusieurs variables latentes, ce qui pourrait indiquer un problème de désentrelacement.

D'autre part, la mesure dite de ***désentrelacement (DCI)*** identifie de manière inverse si une variable latente représente uniquement un facteur génératif. À cette fin, la matrice de probabilité d'importance de la relation est calculée en faisant la moyenne de $R_{ij}$ sur tous les facteurs génératifs inférés :

$$p'_{ij} = \frac{R_{ij}}{\sum_{k=1}^{M} R_{kj}}. \tag{2.23}$$

Cela permet de calculer la mesure d'entropie suivante :

$$DCI_D = \sum_{j=1}^{K} \rho_j * \left(1 + \sum_{i=1}^{M} p'_{ij} \log_M p'_{ij}\right), \tag{2.24}$$

où $\rho_j$ agit comme un terme de pondération en mesurant l'importance globale relative pour chaque variable latente, et est calculé de la manière suivante : $\rho_j = \frac{\sum_{i=1}^{M} R_{ij}}{\sum_{k=1}^{K}\sum_{i=1}^{M} R_{ik}}$. Cette pondération permet de ne pas prendre en considération les variables latentes qui ne comportent aucune information des données d'entrée, c'est à dire qu'ils se sont effondrés vers la distribution *a priori*.

La métrique de **Score de prévisibilité des attributs** [Kumar *et al.*, 2017] a pour objectif de mesurer la qualité explicite et compacte de l'espace latent. Pour cela, la mesure attribue un score $S_{ij}$ à chaque paire de facteur génératif $v_i$ et variable latente $z_j$. La mesure finale consiste à calculer la différence entre les deux scores $S_{ij}$ les plus importants pour chaque facteur génératif $v_i$. Enfin, le **Score de la qualité explicite** [Ridgeway et Mozer, 2018] repose sur l'hypothèse que les facteurs génératifs sont discrets.

### 2.4.3 Métriques basées sur la théorie de l'information

Les métriques basées sur la théorie de l'information reposent sur une estimation de la mesure d'information mutuelle entre les facteurs génératifs et les variables latentes. L'information mutuelle permet de quantifier la quantité d'information partagée entre différentes variables. Elle est définie pour deux variables de la façon suivante :

$$I(X,Y) = \int_{\boldsymbol{x}}\int_{\boldsymbol{y}} p(\boldsymbol{x},\boldsymbol{y}) \log\left(\frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})}\right) d\boldsymbol{x}d\boldsymbol{y} = D_{KL}[p(\boldsymbol{x},\boldsymbol{y})||p(\boldsymbol{x})p(\boldsymbol{y})], \tag{2.25}$$

où X et Y sont deux variables aléatoires, $p(\boldsymbol{x},\boldsymbol{y})$ est la probabilité jointe de $X$ et $Y$ et $p(\boldsymbol{x})$ et $p(\boldsymbol{y})$ leurs probabilités marginales respectives.

Dans le contexte de mesure du désentrelacement, l'information mutuelle est utilisée pour évaluer la dépendance ou l'indépendance statistique entre les facteurs génératifs et les variables latentes. Une faible information mutuelle indique une indépendance, ce qui signifie que les variables latentes n'encodent pas d'information sur les facteurs génératifs. En revanche, une information mutuelle élevée indique une forte dépendance. Grâce à l'approche basée sur la théorie de l'information, les métriques associées offrent une évaluation plus globale de la qualité explicite de l'espace latent, en prenant en compte l'ensemble

des facteurs génératifs et des variables latentes, sans nécessiter de considérations particulières sur leur distribution ou leur nature.

La métrique d'**Écart d'information mutuelle ou "Mutual Information Gap"(MIG)** [Chen *et al.*, 2018] évalue la qualité de désentrelacement en mesurant la différence moyenne entre les deux valeurs les plus élevées d'information mutuelle pour chaque variable latente $z_j$ et chaque facteur génératif $v_i$. Cette mesure permet de quantifier la capacité de chaque variable latente à encoder de manière explicite un facteur génératif spécifique. Le score obtenu pour chaque facteur est ensuite normalisé par l'entropie de ce facteur spécifique. Cette normalisation permet de prendre en compte la distribution des valeurs du facteur génératif, afin de rendre la métrique indépendante de l'échelle de ce dernier et de permettre une comparaison équitable entre les différents facteurs :

$$\mathrm{MIG}(\boldsymbol{z}, \boldsymbol{v}) = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{H(v_i)} (I(z_{j^{(i)}}; v_i) - \underset{j \neq j^{(i)}}{\mathrm{argmax}} I(z_j, v_i)), \tag{2.26}$$

où $j^{(i)} = \mathrm{argmax} I(z_j, v_i)$, et $H(\cdot)$ le terme d'entropie. La métrique MIG offre une approche globale et objective pour mesurer à la fois la capacité explicite et la compacité de l'espace latent. Elle permet également d'identifier les variables latentes les plus informatives.

De nombreuses alternatives à cette métrique ont été proposées. Une approche similaire, le **RMIG**, est définie par les auteurs Do et Tran [Do et Tran, 2019]. La différence principale avec le MIG est que l'information mutuelle est calculée dans l'espace des données d'entrée plutôt que dans l'espace latent. Afin d'étendre cette approche à la propriété de modularité, la mesure **MIG-sup** [Li *et al.*, 2020] propose de quantifier l'information mutuelle dans l'espace des variables latentes de manière supervisée. Ridgeway et Mozer définissent une autre manière de mesurer la même propriété à travers le **score de modularité** [Ridgeway et Mozer, 2018]. Ce score consiste à identifier les facteurs génératifs possédant le maximum d'information mutuelle pour chaque variable latente, puis à les comparer avec les valeurs d'information mutuelle pour tous les autres facteurs. Les auteurs de **JEMMIG** [Do et Tran, 2019] abordent le problème de la non-mesure de la modularité dans la métrique MIG en y introduisant l'entropie conjointe des facteurs génératifs et de la meilleure variable latente. Finalement, la métrique **DCIMIG** [Sepliarskaia *et al.*, 2019b], inspirée des métriques DCI et MIG, calcule l'écart d'information mutuelle entre les facteurs génératifs et les variables latentes tout en se basant sur une matrice de score d'importance.

### 2.4.4 Synthèse

En ce qui concerne les travaux présentés dans les chapitres suivants du manuscrit, seules certaines mesures ont été utilisées dans le but de comparer les performances des modèles dans leur capacité à inférer un espace latent désentrelacé. Un parti pris consistait à considérer une mesure pour chacune des approches proposées par Carbonneau et al. , à savoir une mesure basée sur une transformation de l'espace latent, une mesure basée sur un réseau de prédiction, et une mesure basée sur la théorie de l'information [Carbonneau *et al.*, 2022]. Les trois les plus couramment utilisées dans l'état de l'art sont le z-diff score [Kim *et al.*, 2019a], le MIG [Chen *et al.*, 2018], et le SAP [Kumar *et al.*, 2017].

Cependant, l'étude effectuée par Carbonneau et al. a conduit à considérer, pour l'ensemble des comparaisons effectuées dans ce manuscrit, la mesure de z-min variance [Kim et Mnih,

2018], basée sur une transformation de l'espace latent, qui évite le mode de défaillance rencontré par le z-diff score. En ce qui concerne les approches basées sur la théorie de l'information, la mesure MIG est conservée car elle semble aussi efficace que ses successeurs. Finalement, la mesure basée sur un réseau de prédiction, DCI [Eastwood et Williams, 2018], en utilisant un algorithme de forêts aléatoires, se révèle être l'approche la plus prometteuse. En effet, à travers son cadre complet, elle permet de mesurer le désentrelacement pour chacune des propriétés de qualité explicite, de modularité et de compacité de l'espace latent. Contrairement aux autres approches, elle prend également en compte les variables latentes ayant effectué de l'effondrement *a posteriori*, un élément essentiel lorsque l'on observe une distinction dans l'espace latent entre différent type de variables.

### 2.4.5 Jeux de données

En complément des nombreuses mesures présentées ci-dessus, la communauté scientifique a proposé divers jeux de données composés d'images plus ou moins complexes. Ces jeux de données sont utilisés pour entraîner les modèles et évaluer la qualité d'encodage des représentations dans leur espace latent. Comme observé dans la section présentant les différentes métriques utilisées, la plupart d'entre elles sont qualifiées de supervisées, car elles nécessitent l'accès aux facteurs génératifs. Ainsi, différents jeux de données ont été créés de manière procédurale à partir des facteurs génératifs fournis pour chacune des images. Ces facteurs génératifs servent à contrôler les caractéristiques des images et permettent de disposer de données de référence pour évaluer la séparation des facteurs dans l'espace latent. Ils sont précieux pour la recherche sur le désentrelacement, car ils permettent d'effectuer des comparaisons objectives entre les modèles et d'évaluer leur capacité à capturer de manière explicite et compacte les facteurs sous-jacents à la génération des images. Certains de ces jeux de données sont conçus pour être simples, avec des facteurs génératifs clairs et bien définis, tandis que d'autres sont plus complexes, avec des facteurs génératifs plus nuancés et difficiles à distinguer. Cela permet d'évaluer la robustesse et la généralisation des modèles dans des scénarios réalistes.

Le jeu de données **Dsprites** [Matthey *et al.*, 2017], dont des exemples sont illustrés dans la Figure 2.6b, est constitué de formes en deux dimensions et comprend six facteurs génératifs : la couleur, la forme, l'échelle, la rotation, ainsi que les positions de la forme sur l'axe horizontal et vertical de l'image. Ces derniers sont définis de la façon suivante :

- **Couleur :** blanche,
- **Forme :** trois variables catégorielles permettant de construire des carrés, des ellipses ou des coeurs,
- **Échelle :** six valeurs linéairement espacées dans l'intervalle $[0.5, 1]$,
- **Orientation :** quarante valeurs linéairement espacées dans l'intervalle $[0, 2\pi]$,
- **Position** $x$ : trente-deux valeurs dans l'intervalle $[0, 1]$,
- **Position** $y$ **:** trente-deux valeurs dans l'intervalle $[0, 1]$.

Les images ont été crées de façon séquentielle en faisant varier chaque valeur génératif à la fois. Ce jeu de données est particulièrement intéressant car il propose un cadre clair pour évaluer la capacité des modèles à séparer ces facteurs génératifs dans leur espace latent.

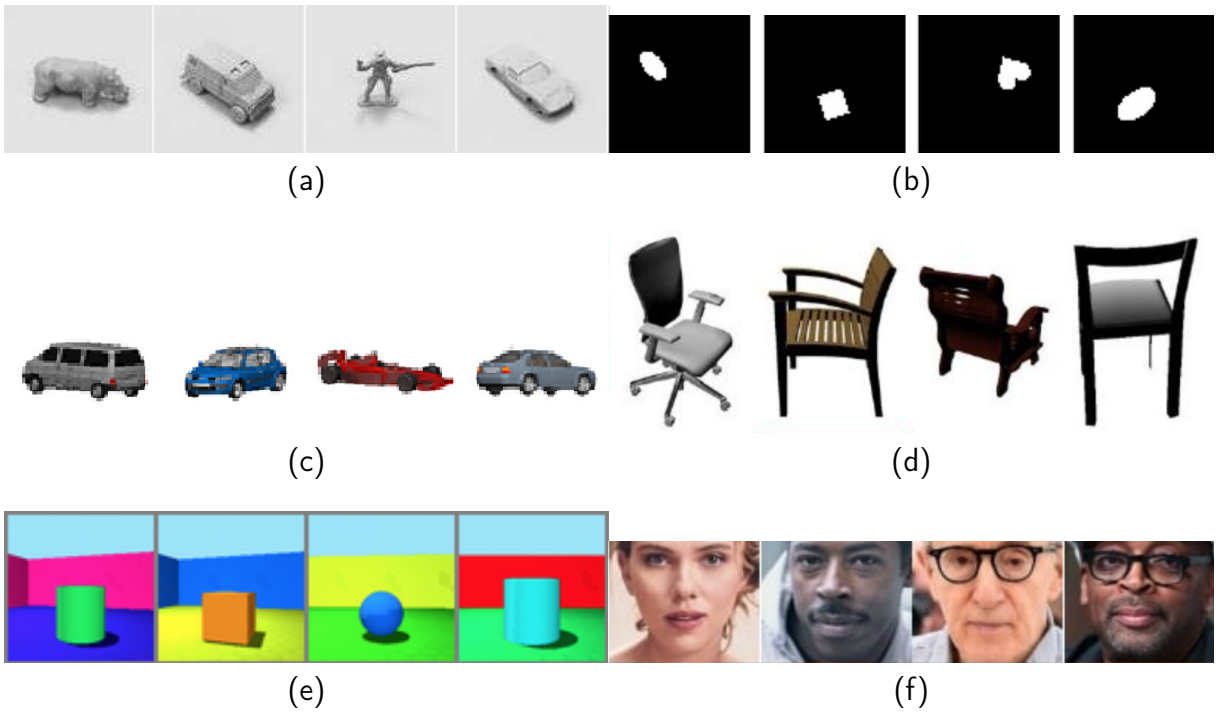



Fig 2.6: Exemple de données d'entraînement sur plusieurs ensembles de données. (a) : Smallnorb. (b) : Dsprites. (c) : Cars3d. (d) : 3DChairs. (e) : Shapes3D. (f) : CelebA.

Le jeu de données **SmallNORB** [LeCun *et al.*, 2004], dont des exemples sont illustrés dans la Figure 2.6a, est composé d'une collection d'images représentant cinquante jouets différents répartis en cinq grandes catégories génériques. Les objets sont capturés à travers les objectifs de deux caméras, en utilisant six conditions d'éclairage distinctes, neuf élévations et dix-huit orientations. Chaque objet est représenté sous différents angles d'éclairage et de vue, ce qui génère des images avec des variations significatives dans l'apparence des objets. Ces variations permettent de créer un jeu de données plus complexe et réaliste, mettant ainsi à l'épreuve la capacité des modèles de désentrelacement à séparer efficacement la représentation des facteurs génératifs au sein de l'espace latent. Ces facteurs ont été constitués de la façon suivante :

- **Catégorie des jouets :** cinq variables catégorielles désignant les jouets de type "animaux à quatre pattes", "humains", "avions", "camions" et "voitures".
- **Élévation :** neuf valeurs espacées de cinq degrés dans l'intervalle $[30, 70]$,
- **Orientation :** dix-huit valeurs espacées de vingt degrés dans l'intervalle $[0, 340]$,
- **Condition d'éclairage :** six variables catégorielles.

**Cars3D** [Reed *et al.*, 2015], dont des exemples sont illustrés dans la Figure 2.6c, est constitué de cent quatre-vingt-dix-neuf modèles de voitures, à partir desquels des rendus en couleur ont été générés pour différents angles de rotation. De plus, les images ont été capturés à partir de caméras situées à des élévations différentes. De façon plus exhaustive, ces facteurs génératifs sont définis de la façon suivante :

- **Type d'objet** : variable catégorielle correspondant à un des cent quatre-vingt-dix-neuf modèles de voitures,
- **Élévation :** variable catégorielle correspondant à une des caméras à partir de laquelle l'image a été capturée, parmi les quatre possédant des élévations différentes,

- **Orientation :** vingt-quatre valeurs d'angles différentes, espacées de quinze degrés.

La diversité des angles de rotation et des élévations de caméra permet de créer un jeu de données riche et complexe, où chaque voiture est représentée sous de nombreuses perspectives différentes. Cela permet d'évaluer la capacité des modèles de désentrelacement à séparer efficacement les facteurs génératifs, en tenant compte des variations d'apparence liées aux différents angles de vue et aux élévations. La couleur n'est pas considérée comme étant un facteur génératif dans le jeu de données, car chaque modèle de voiture est susceptible de posséder sa propre couleur.

Des jeux de données couramment utilisés dans les approches de l'état de l'art sont également **3Dshapes** [Kim et Mnih, 2018], illustré dans la Figure 2.6e, qui représente des formes en trois dimensions dans un environnement en couleur, **3DChairs** [Aubry *et al.*, 2014] composé de modèles de chaises selon différents points de vue comme illustré dans la Figure 2.6d ou encore **CelebA** [Liu *et al.*, 2015], illustré dans la Figure 2.6f, qui est un ensemble de données d'attributs de visage à grande échelle. Pour l'ensemble de ces jeux de données, les images sont également associées aux valeurs respectives de leurs facteurs génératifs.

Pour évaluer les performances des méthodes présentées dans la suite du manuscrit, les apprentissages ont été réalisés dans un premier temps sur le jeu de données simple Dsprites, où le nombre de facteurs génératifs est directement identifiable. Ensuite, il a également été souhaité d'effectuer ces apprentissages sur des données plus complexes, telles que Cars3d, qui contient des images en couleur, ainsi que Smallnorb, où l'identification des facteurs génératifs est plus complexe.

## 2.5 Conclusion

Ce chapitre a pour objectif de fournir une revue synthétique de l'état de l'art concernant le désentrelacement au sein des méthodes basées sur un VAE. Différents paradigmes de représentations possibles dans un espace latent sont présentés en se concentrant sur certaines définitions du désentrelacement proposées dans la communauté scientifique. Pour la suite du rapport, la définition de désentrelacement proposée par Carbonneau et al. [Carbonneau *et al.*, 2022] a été choisie. Elle identifie trois propriétés désirables du désentrelacement : la qualité explicite, la compacité et la modularité. Ces propriétés permettent de mesurer la capacité de désentrelacement des méthodes considérées de manière plus complète et consensuelle.

Une explication de certaines méthodes de désentrelacement basées sur un VAE a également été proposée en se basant sur une taxonomie définie. Dans cette dernière, les méthodes sont regroupées en trois catégories : celles basées sur une pondération de l'ELBO, celles ajoutant un terme de pondération à la fonction coût et celles distinguant la modélisation de la distribution *a priori* pour des variables latentes considérées comme pertinentes ou non. Les avantages et les inconvénients de chacune de ces approches sont analysés, faisant émerger le constat que la taille de l'espace latent est à la fois sensible et difficile à ajuster. De ce fait, l'ensemble des modélisations étudiées définissent un espace initial à grande dimension dont la capacité d'encodage est réduite à l'aide d'hyperparamètres pour répondre aux propriétés de désentrelacement. Leur valeur initiale est souvent définie lors du processus d'hyperparamétrage à l'aide d'une recherche de grille. De plus, aucune de ces approches n'introduit de biais inductif au sein de la distribution *a posteriori*, qui est

toujours modélisée comme une distribution gaussienne.

Enfin, une revue des différentes métriques de désentrelacement basée sur la taxonomie proposée par Carbonneau et al. [Carbonneau *et al.*, 2022] est réalisée avant de définir les propriétés des jeux de données utilisés dans l'état de l'art, pour lesquels les facteurs génératifs de chaque image sont fournis. Ces derniers permettent ainsi de calculer les métriques de désentrelacement de manière supervisée. Cette revue permet de mieux comprendre les défis et opportunités liés au désentrelacement des méthodes basées sur un VAE. De ce fait, elle a permis d'orienter nos travaux de recherche vers de nouveaux algorithmes prometteurs pour répondre à cette tâche de l'apprentissage non supervisé.

En conséquence, la suite de ce manuscrit vise à expliquer les méthodes que nous avons proposées durant nos recherches. Elles visent à éviter certaines limitations remarquées dans les approches de l'état de l'art dans l'objectif d'obtenir une représentation pertinente et désentrelacée au sein d'un espace latent.

# Chapitre 3

# Polarisation de l'espace latent et modèle bayésien hiérarchique

Le troisième chapitre de ce manuscrit présente la première contribution de la thèse : une variante du vanilla-VAE permettant de favoriser un comportement spécifique au sein des variables latentes. Elle consiste en une distinction entre, d'une part, les variables responsables de l'encodage des facteurs génératifs et, d'autre part, celles qui ne contiennent pas d'information. Une telle séparation, nommée "polarisation" dans l'état de l'art, peut être observée quand la dimension de l'espace latent est définie comme supérieure au nombre de facteurs génératifs. Il est alors nécessaire que seulement un sous-ensemble de variables latentes contienne de l'information afin de répondre aux propriétés de compacité et de modularité du désentrelacement. De ce fait, il fait sens qu'elles soient caractérisées par une distribution différente de celle des variables non informatives, ce qui pourrait permettre une adaptation automatique de la dimension de l'espace latent nécessaire pour encoder les génératifs sans ajouter d'hyperparamètres.

Dans la première partie de ce chapitre, les différentes définitions et propositions d'explication du comportement bimodal des variables latentes observé au sein d'un régime polarisé sont reprises. Ces comportements sont validés empiriquement à travers une analyse de la distribution des variables inférées par le réseau d'encodeur d'un modèle Factor-VAE ayant convergé vers un espace latent désentrelacé. À la suite de cette analyse, une formalisation théorique de chacun des comportements observés est proposée. Cette démarche permet de justifier une première approche développée au cours de nos travaux, à savoir un modèle bayésien hiérarchique nommé Normal-Gamma Variational Auto-Encoder (NG-VAE) [Jouffroy *et al.*, 2022]. Ce dernier est élaboré dans le but d'obtenir un espace latent polarisé de manière efficace.

La seconde partie de ce chapitre a pour objectif de décrire le NGVAE et d'expliciter certaines pratiques mises en œuvres pour l'apprentissage du modèle. Elles consistent en l'introduction d'une nouvelle méthode de reparamétrisation de gradient, la justification de l'inversion des distributions au sein du terme de KL-divergence dans la fonction coût et l'utilisation de la polarisation comme un biais inductif lors de l'apprentissage.

La fin du chapitre est consacrée à l'analyse de la capacité du modèle NGVAE à obtenir un espace latent polarisé, conformément à la définition que nous en proposons, par rapport à des méthodes de référence de l'état de l'art. Ces comparaisons reposent sur des observations empiriques propres au contexte de polarisation, ainsi que sur des mesures quantitatives relatives à la capacité d'encodage et à la décorrélation entre les variables latentes.

## 3.1 Concept de variables actives et passives et d'espace polarisé

Pour mieux appréhender le désentrelacement au sein d'un espace latent, certaines études se sont concentrées sur l'analyse de la capacité d'un modèle de type VAE à effectuer une séparation entre des variables contenant de l'information et d'autres n'en contenant pas. En effet, pour satisfaire les propriétés de compacité, de modularité et de qualité explicite du désentrelacement, il est pertinent de considérer que certaines des variables de l'espace latent n'encodent aucune information sur les données d'entrées dès lors que la dimension de l'espace latent est définie comme supérieure au nombre de facteurs génératifs. La capacité du VAE à désactiver certaines variables latentes tout en préservant l'information dans les autres est connue sous le nom de "polarisation" ou "effondrement sélectif *a posteriori*" [Rolinek *et al.*, 2019; Dai *et al.*, 2018, 2020; Bonheme et Grzes, 2021]. Les variables latentes corrélées aux données d'entrée sont usuellement appelées variables "actives", tandis que les autres convergeant vers une distribution *a priori* non informative sont appelées variables "passives". Des études effectuées par Dai et al. ont également permis de démontrer que cette polarisation était une condition nécessaire pour les VAEs afin d'atteindre une bonne capacité de reconstruction [Dai *et al.*, 2018, 2020].

Dans la suite de cette section, le comportement de la distribution des paramètres inférés par l'encodeur dans le cadre de l'obtention d'un régime polarisé est examiné de façon empirique. Pour cela, la définition initiale proposée dans l'article de référence [Rolinek *et al.*, 2019] ainsi que les extensions proposées par les auteurs Bonheme et Grzes [Bonheme et Grzes, 2021] sont analysées. À la suite de cette étude, une formalisation théorique permettant de justifier les comportements observés est proposée, un point manquant à l'état de l'art.

### 3.1.1 Notations

Pour ce chapitre, le vecteur latent est dénoté en gras par $\boldsymbol{z}$. Dans le contexte d'un régime polarisé, on note $\boldsymbol{z} \in \{\mathbb{Z}_a, \mathbb{Z}_p\}$, où $\mathbb{Z}_a$ et $\mathbb{Z}_p$ représentent respectivement les ensembles des variables actives et passives, avec $\mathbb{Z}_a \cap \mathbb{Z}_p = \emptyset$. L'exposant $(i)$ est employé pour indiquer l'indice du $i$-ème élément du jeu de données $\boldsymbol{x}^{(i)}$, et $k$ représente l'indice de la variable latente, de sorte que $z_k^{(i)}$ correspond à la $k$-ème composante de $\boldsymbol{z}$ obtenue après l'encodage de $\boldsymbol{x}^{(i)}$. $\mathcal{Z}_a$ correspond à l'ensemble des indices des variables actives, tandis que $\mathcal{Z}_p$ est utilisé pour l'ensemble des indices des variables passives. Finalement, les vecteurs de moyennes et de variances inférés par l'encodeur, qui dépendent des paramètres de biais et de poids du réseau rassemblés dans le vecteur $\boldsymbol{\phi}$, sont respectivement désignés de la façon suivante : $\boldsymbol{\mu}(\boldsymbol{x}; \boldsymbol{\phi})$ et $\boldsymbol{\sigma}^2(\boldsymbol{x}; \boldsymbol{\phi})$.

### 3.1.2 Comportement empirique

Dans l'article de référence sur le régime polarisé [Rolinek *et al.*, 2019], un VAE pour lequel l'espace latent est divisé en deux sous-espaces est considéré. Dans le premier, un ensemble initial de variables latentes est composé de bruit et ne contient donc aucune information, tandis que le second conserve l'information des données d'entrées de manière condensée. Pour une donnée spécifique $\boldsymbol{x}^{(i)}$, les auteurs définissent le régime polarisé de l'espace latent de la manière suivante :

**Définition 1** (régime polarisé sur un échantillon $\boldsymbol{x}^{(i)}$ [Rolinek *et al.*, 2019])**.** *Les paramètres d'un réseau $\boldsymbol{\phi}, \boldsymbol{\theta}$ induisent un régime polarisé si les variables latentes $\boldsymbol{z} \in \mathbb{R}^K$ peuvent être*

*partitionnées en deux sous-ensembles $\mathbb{Z}_a \cup \mathbb{Z}_p$ (ensemble de variables actives et passives), avec $\mathbb{Z}_a \cap \mathbb{Z}_p = \emptyset$, tels que, pour chaque échantillon $\boldsymbol{x}^{(i)}$ du jeu de données :*

*(a)* $|\boldsymbol{\mu}_j(\boldsymbol{x}^{(i)}; \boldsymbol{\phi})| \ll 1$ *et* $\boldsymbol{\sigma}_j^2(\boldsymbol{x}^{(i)}; \boldsymbol{\phi}) \approx 1$, $\forall j \in \mathcal{Z}_p$,

*(b)* $\boldsymbol{\sigma}_j^2(\boldsymbol{x}^{(i)}; \boldsymbol{\phi}) \ll 1$ $\forall j \in \mathcal{Z}_a$,

*(c) Le décodeur ignore les variables latentes passives, tel que*

$$\frac{\partial Dec_{\boldsymbol{\theta}}(\boldsymbol{z})}{\partial \boldsymbol{z}_j} = 0, \forall j \in \mathcal{Z}_p,$$

où $\text{Dec}_{\boldsymbol{\theta}}(\boldsymbol{z})$ indique le réseau du décodeur de paramètres $\boldsymbol{\theta}$. Cette définition repose sur l'observation que les variables passives ont convergé vers la distribution *a priori* $p(\boldsymbol{z}) = \mathcal{N}(0, 1)$. Elles contiennent alors du bruit et possèdent un gradient nul par rapport au vecteur de paramètres $\boldsymbol{\theta}$ du décodeur. Ainsi, elles ne participent pas à la reconstruction. Inversement, les variables actives possèdent davantage d'information vis-à-vis d'une donnée d'entrée $\boldsymbol{x}^{(i)}$, impliquant une valeur de leur variance plus petite.

En se reposant sur le fait que le décodeur tend à ignorer les variables latentes passives, des études effectuées par Dai et al. permettent de définir le régime polarisé comme un mécanisme visant à induire de la parcimonie dans la colonne de poids de la première couche du réseau du décodeur, de telle sorte que le réseau considère uniquement les variables actives pour la reconstruction [Dai *et al.*, 2018, 2020]. Les auteurs montrent que ce comportement peut être observé très tôt dans le processus d'apprentissage.

À partir de cette définition, Bonheme et Grzes proposent d'étendre le comportement des paramètres inférés par l'encodeur à l'ensemble des données d'entraînement $\boldsymbol{x}$ [Bonheme et Grzes, 2021]. Pour cela, ils émettent des hypothèses concernant le comportement des variables latentes actives et passives sur l'ensemble du jeu de données. Dans ce contexte, une première proposition est formulée de la façon suivante.

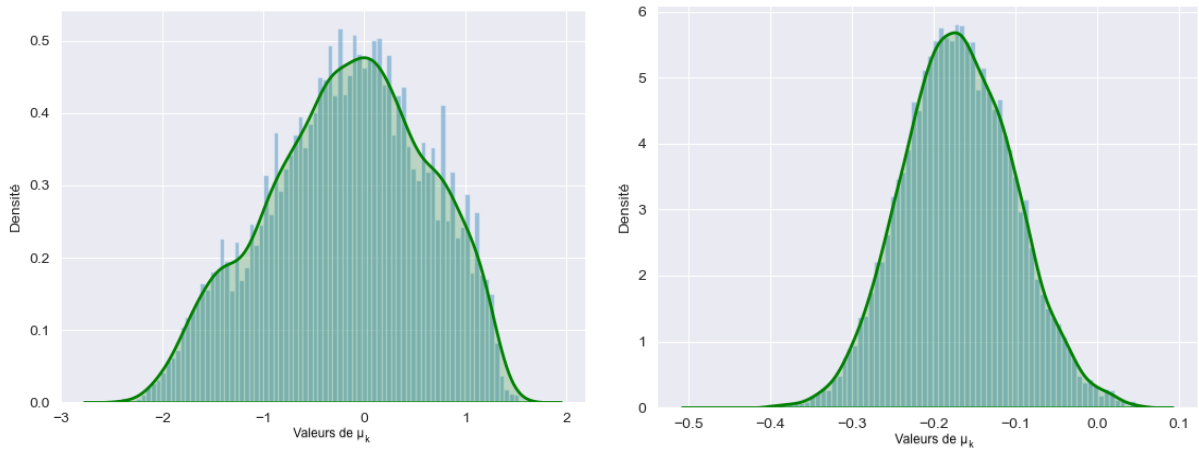
**Proposition 1** (régime polarisé de $\boldsymbol{\mu}(\boldsymbol{x}; \boldsymbol{\phi})$ sur l'ensemble des données d'entraînement [Bonheme et Grzes, 2021])**.** *Lorsqu'un VAE apprend un régime polarisé au sein de son espace latent, la moyenne inférée par l'encodeur $\boldsymbol{\mu}(\boldsymbol{x}; \boldsymbol{\phi})$ est composée d'un ensemble de variables passives et actives $\mathbb{Z}_p \cup \mathbb{Z}_a$, tel que, sur l'ensemble du jeu de données d'entraînement $\boldsymbol{x}$:*

*(a)* $|\mathbb{E}_{p_{data}(\boldsymbol{x})}\big[\mu_j(\boldsymbol{x}; \boldsymbol{\phi})\big]| \ll 1$ *et* $Var\big(\mu_j(\boldsymbol{x}; \boldsymbol{\phi})\big) \ll 1$, $\forall j \in \mathcal{Z}_p$,

*(b)* $Var\big(\mu_j(\boldsymbol{x}; \boldsymbol{\phi})\big) > Var\big(\mu_k(\boldsymbol{x}; \boldsymbol{\phi})\big)$, $\forall j \in \mathcal{Z}_a, \forall k \in \mathcal{Z}_p$,

*où $Var(.)$ dénote la variance calculée par rapport à $p_{data}$, et $|\cdot|$ la valeur absolue.*

Afin de visualiser le comportement de la distribution de $\boldsymbol{\mu}(\boldsymbol{x}; \boldsymbol{\phi})$, décrit dans la proposition 1, nous avons utilisé un FactorVAE [Kim et Mnih, 2018]. La dimension de l'espace latent a été fixée à 10 et le modèle est préalablement entraîné sur le jeu de données dsprites [Matthey *et al.*, 2017]. L'hyperparamètre $\gamma = 35$ est utilisé, car il est considéré comme la valeur optimale pour ce jeu de données par les auteurs. Ce modèle, déjà entraîné, a été récupéré de la bibliothèque "Disentanglement-Lib" développée par Locatello et al., qui met à disposition 12 600 modèles de VAEs pré-entraînés sur différents jeux de données [Locatello *et al.*, 2019b]. Cette ressource a grandement facilité notre analyse et la comparaison des différentes approches, permettant un gain de temps significatif.




(a) (b)

Fig 3.1: (a) : Densité empirique de $\mu_k(\boldsymbol{x};\boldsymbol{\phi})$, avec $k = 2$, correspondant à une dimension de $\boldsymbol{z}$ considérée comme active. (b) : Densité empirique de $\mu_k(\boldsymbol{x};\boldsymbol{\phi})$, avec $k = 5$, correspondant à une dimension de $\boldsymbol{z}$ considérée comme passive. Elle est calculée à partir de 10 000 images du jeu de données Dsprites, après l'entraînement d'un modèle de Factor-VAE pour un espace latent de taille 10.

La Figure 3.1 représente la distribution empirique de différentes composantes de $\boldsymbol{\mu}(\boldsymbol{x};\boldsymbol{\phi})$ pour 10 000 images. Nous nous sommes intéressés à deux dimensions spécifiques de l'espace latent, correspondant à une variable active et à une variable passive. Ce choix s'est basé sur la mesure introduite par Rolinek et al. , qui définit une variable $z_j$ active si $\sqrt{Var(\mu_j(\boldsymbol{x};\boldsymbol{\phi}))} > 0.5$ [Rolinek *et al.*, 2019].
L'analyse de ces deux distributions révèle, comme attendu, deux comportements distincts. La distribution de $\mu_2(\boldsymbol{x};\boldsymbol{\phi})$, variable active illustrée dans la Figure 3.1a, présente une variance nettement plus élevée que celle de $\mu_5(\boldsymbol{x};\boldsymbol{\phi})$, variable passive illustrée dans la Figure 3.1b. De plus, il est à noter que la variance et l'espérance des moyennes pour les variables passives sont nettement inférieures à 1.
Ces constatations confirment la proposition 1 avancée par les auteurs Bonheme et Grzes. Il est également intéressant de noter que la moyenne inférée pour une variable active sur l'ensemble du jeu de données ne suit pas une loi gaussienne.

Afin d'étendre leur analyse concernant le comportement de $\boldsymbol{\sigma}^2(\boldsymbol{x};\boldsymbol{\phi})$, les auteurs soulignent l'existence de variables mixtes. L'émergence de ces dernières repose sur l'hypothèse qu'une sous-partie des données d'entraînement n'est pas générée par l'ensemble des facteurs génératifs. Cela signifie que pour certaines images, une variable mixte suit le comportement d'une variable latente active tandis que pour d'autres images, elle est considérée comme passive. L'espace latent considéré devient alors $\boldsymbol{z} \in \{\mathbb{Z}_a, \mathbb{Z}_p, \mathbb{Z}_m\}$ avec $\mathbb{Z}_a \cap \mathbb{Z}_p \cap \mathbb{Z}_m = \emptyset$ et $\mathbb{Z}_m$ représente l'ensemble des variables mixtes. De cette façon, une seconde proposition est formulée.

**Proposition 2** (régime polarisé de $\boldsymbol{\sigma}^2(\boldsymbol{x};\boldsymbol{\phi})$ sur l'ensemble des données d'entraînement. [Bonheme et Grzes, 2021])**.** *Lorsqu'un VAE apprend un régime polarisé au sein de son espace latent, la variance inférée par l'encodeur* $\boldsymbol{\sigma}^2(\boldsymbol{x};\boldsymbol{\phi})$ *est composée d'un ensemble de variables passives, actives et mixtes* $\mathbb{Z}_p \cup \mathbb{Z}_a \cup \mathbb{Z}_m$ *tel que, sur l'ensemble du jeu de données* $\boldsymbol{x}$*:*

*(a)* $\mathbb{E}_{p_{data}(\boldsymbol{x})}\big[\boldsymbol{\sigma}_j^2(\boldsymbol{x};\boldsymbol{\phi})\big] \approx 1$ *et* $Var\big(\boldsymbol{\sigma}_j^2(\boldsymbol{x};\boldsymbol{\phi})\big) \ll 1$, $\forall j \in \mathcal{Z}_p$,

*(b)* $\mathbb{E}_{p_{data}(\boldsymbol{x})}\big[\boldsymbol{\sigma}_j^2(\boldsymbol{x};\boldsymbol{\phi})\big] \ll 1$ *et* $Var\big(\boldsymbol{\sigma}_j^2(\boldsymbol{x};\boldsymbol{\phi})\big) \ll 1$, $\forall j \in \mathcal{Z}_a$,

*(c)* $Var\big(\boldsymbol{\sigma}_j^2(\boldsymbol{x};\boldsymbol{\phi})\big) < Var\big(\boldsymbol{\sigma}_k^2(\boldsymbol{x};\boldsymbol{\phi})\big)$, $\forall j \notin \mathcal{Z}_m$, $\forall k \in \mathcal{Z}_m$.

*où* $Var(.)$ *dénote la variance calculée par rapport à* $p_{data}$.

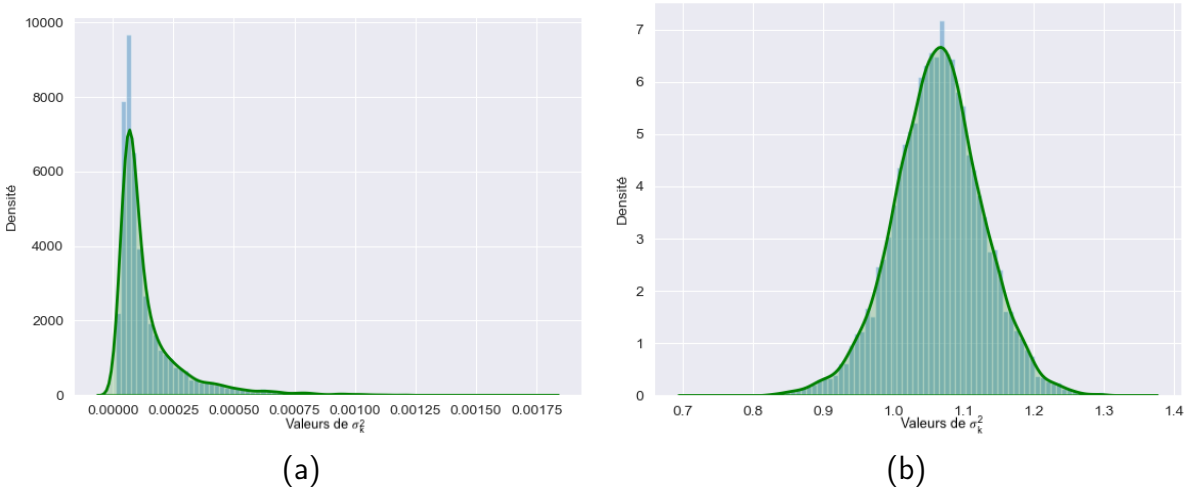



Fig 3.2: (a) : Densité empirique de $\sigma_k^2(\boldsymbol{x};\boldsymbol{\phi})$, avec $k = 2$, correspondant à une dimension de $\boldsymbol{z}$ considérée comme active. (b) : Densité empirique de $\sigma_k^2(\boldsymbol{x};\boldsymbol{\phi})$, avec $k = 5$, correspondant à une dimension de $\boldsymbol{z}$ considérée comme passive. Elle est calculée à partir de 10 000 images du jeu de données Dsprites, après l'entraînement d'un modèle de Factor-VAE pour un espace latent de taille 10.

De manière analogue à l'analyse des moyennes, nous avons illustré la distribution de $\sigma^2(\boldsymbol{x};\boldsymbol{\phi})$ pour une variable active et une variable passive dans la Figure 3.2. Les composantes de $\boldsymbol{z}$ représentées restent les mêmes que celles utilisées pour l'analyse de la moyenne. Nous n'avons pas pris en compte les variables mixtes. En effet, nous estimons que la définition de cet ensemble de variables repose sur l'interprétation initiale faite pour les facteurs génératifs. Bien qu'une étude supplémentaire soit nécessaire pour justifier nos propos, nous émettons l'hypothèse que l'émergence de telles variables observées par les auteurs Bonheme et Grzes repose sur la façon dont le jeu de données Dsprites a été créé. En effet, nous montrerons plus tard dans ce manuscrit que certains facteurs génératifs utilisés pour la génération des données sont corrélés entre eux, ce qui peut avoir une incidence quant à la capacité du modèle à encoder ces facteurs.
On peut également identifier deux comportements distincts en fonction de l'information contenue dans la variable considérée. En effet, la variance des variables passives semble suivre une distribution gaussienne centrée autour de 1. En revanche, pour une variable active, le réseau est davantage "confiant" dans les informations encodées. Dans ce sens, la distribution de $\boldsymbol{\sigma}_k^2(\boldsymbol{x};\boldsymbol{\phi})$ pour $k = 2$ est piquée et favorise des valeurs proches de zéro. Ces observations semblent confirmer la plupart des éléments énoncés dans les points (a) et (b) de la proposition 2.

En conclusion, la capacité des modèles de VAE à atteindre un régime polarisé au sein de leur espace latent revêt une importance fondamentale pour l'obtention d'un désentrelacement efficace. En effet, lorsque la dimension définie *a priori* pour l'espace latent dépasse le nombre de facteurs génératifs, il est souhaitable d'obtenir une telle séparation afin de satisfaire aux exigences de compacité. De plus, l'ensemble des méthodes de l'état de l'art visant à obtenir des variables latentes désentrelacées, comme nous l'avons précédemment défini dans ce manuscrit, ont toutes pour objectif intrinsèque d'atteindre un régime polarisé. On peut notamment citer le $\beta$-VAE [Higgins *et al.*, 2016], où l'augmentation

de $\beta$ favorise le nombre de variables latentes passives. De même, la modélisation du DIP-VAE [Kumar *et al.*, 2017] requiert la définition des paramètres $\lambda_d$ et $\lambda_{od}$, dont les valeurs s'équilibrent pour obtenir le nombre souhaité de variables actives. Le modèle de BF-VAE-2 [Kim *et al.*, 2019a], quant à lui, s'appuie sur la définition de l'ensemble d'hyperparamètres $\{\eta_1, \eta_2\}$ afin de définir la dimension de l'ensemble $\mathbb{Z}_p$.

Cependant, les analyses visant à mieux comprendre les causes de ce phénomène reposent sur des hypothèses ou des vérifications empiriques. Par conséquent, dans la suite de ce chapitre, nous proposons une formalisation théorique de ce comportement, en nous basant sur des principes de la théorie de l'information. À notre connaissance, une telle analyse n'a encore jamais été proposée dans la littérature.

### 3.1.3 Formalisation théorique

Afin d'approfondir la compréhension de la polarisation, nous proposons de formaliser les comportements observés au sein de ce régime au travers de démonstrations théoriques. Nous considérons un espace latent composé de variables latentes $\boldsymbol{z} \in \mathbb{R}^K$, avec $K > M$, où $M$ représente le nombre de facteurs génératifs présents dans le jeu de données. De plus, la polarisation entraîne l'émergence de deux sous-espaces distincts, à savoir $\boldsymbol{z} \in \{\mathbb{Z}_a, \mathbb{Z}_p\}$, avec $\mathbb{Z}_a \cap \mathbb{Z}_p = \emptyset$. Comme précédemment, $\mathbb{Z}_a$ désigne l'ensemble des variables actives et $\mathbb{Z}_p$ celui des variables passives, et les variables mixtes ne sont pas considérées.
Ces démonstrations reposent sur une proposition de séparation entre les variables latentes actives et passives fondée sur la théorie de l'information. Cette distinction est formulée de la manière suivante :

$$I\left(z_j^{(i)}, \boldsymbol{x}^{(i)}\right) > I\left(z_k^{(i)}, \boldsymbol{x}^{(i)}\right), \forall j \in \mathcal{Z}_a, \forall k \in \mathcal{Z}_p, \tag{3.1}$$

où l'on considère qu'une variable active possède une information mutuelle dénotée $I(\cdot, \cdot)$ avec l'image d'entrée plus importante qu'une variable passive.

Dans un premier temps, notre intérêt se porte sur la comparaison des valeurs de $\sigma^2(\boldsymbol{x}; \boldsymbol{\phi})$ dans le contexte du régime polarisé, pour une image spécifique du jeu de données $\boldsymbol{x}^{(i)}$. En reprenant l'hypothèse usuellement utilisée dans un VAE, nous considérons que la distribution *a priori* $p(z)$ est définie comme $\mathcal{N}(0, 1)$.

**Théorème 1** (Régime polarisé de $\sigma^2(\boldsymbol{x}^{(i)}; \boldsymbol{\phi})$, sur un échantillon $\boldsymbol{x}^{(i)}$)**.** *Dans le cadre d'un régime polarisé au sein de l'espace latent d'un VAE, où la distribution a priori est définie comme $p(\boldsymbol{z}) = \mathcal{N}(0, 1)$ et la distribution a posteriori comme $q_{\boldsymbol{\phi}}(z|\boldsymbol{x}^{(i)}) = \mathcal{N}(\mu(\boldsymbol{x}^{(i)}; \boldsymbol{\phi}), \sigma^2(\boldsymbol{x}^{(i)}; \boldsymbol{\phi}))$, la variance inférée pour une variable active est plus faible que celle inférée pour une variable passive :*

$$\sigma_j^2\left(\boldsymbol{x}^{(i)}; \boldsymbol{\phi}\right) < \sigma_k^2\left(\boldsymbol{x}^{(i)}; \boldsymbol{\phi}\right), \forall j \in \mathcal{Z}_a, \forall k \in \mathcal{Z}_p. \tag{3.2}$$

*Preuve.*

$$\begin{aligned} & I\big(\boldsymbol{z}_j, \boldsymbol{x}^{(i)}\big) > I\big(\boldsymbol{z}_k, \boldsymbol{x}^{(i)}\big) \\ & \Leftrightarrow H\big(\boldsymbol{z}_j\big) - H\big(\boldsymbol{z}_j|\boldsymbol{x}^{(i)}\big) > H\big(\boldsymbol{z}_k\big) - H\big(\boldsymbol{z}_k|\boldsymbol{x}^{(i)}\big). \end{aligned} \tag{3.3}$$

Or, la distribution *a priori* est similaire pour les variables actives et les variables passives. Cela implique que la mesure d'entropie de la distribution *a priori* d'une variable

active $H(z_j)$ est similaire à celle d'une variable passive $H(z_k)$, nous permettant de réécrire l'équation (3.3) de la façon suivante :

$$\begin{aligned} & I\big(z_j, \boldsymbol{x}^{(i)}\big) > I\big(z_k, \boldsymbol{x}^{(i)}\big) \\ \Leftrightarrow\ & H\big(z_j|\boldsymbol{x}^{(i)}\big) < H\big(z_k|\boldsymbol{x}^{(i)}\big) \end{aligned} \tag{3.4}$$

Par ailleurs, la définition de l'entropie conditionnelle :

$$H(z|\boldsymbol{x}^{(i)}) = \frac{1}{2}\log\big(2\pi\sigma^2\left(\boldsymbol{x}^{(i)};\boldsymbol{\phi}\right)\big) + \frac{1}{2\sigma^2\big(\boldsymbol{x}^{(i)};\boldsymbol{\phi}\big)}\mathbb{E}\big[\big(\boldsymbol{z} - \mu\left(\boldsymbol{x}^{(i)};\boldsymbol{\phi}\right)\big)^2\big], \tag{3.5}$$

où l'espérance est mesurée sur $p_{\text{data}}$, nous permet d'obtenir la relation suivante :

$$\begin{aligned} & I\big(z_j, \boldsymbol{x}^{(i)}\big) > I\big(z_k, \boldsymbol{x}^{(i)}\big) \\ \Leftrightarrow\ & \frac{1}{2}\log\left(2\pi\sigma_j^2\left(\boldsymbol{x}^{(i)};\boldsymbol{\phi}\right)\right) + \frac{1}{2} < \frac{1}{2}\log\left(2\pi\sigma_k^2\left(\boldsymbol{x}^{(i)};\boldsymbol{\phi}\right)\right) + \frac{1}{2} \\ \Leftrightarrow\ & \sigma_j^2\left(\boldsymbol{x}^{(i)};\boldsymbol{\phi}\right) < \sigma_k^2\left(\boldsymbol{x}^{(i)};\boldsymbol{\phi}\right), \forall j \in \mathcal{Z}_a, \forall k \in \mathcal{Z}_p. \end{aligned} \tag{3.6}$$

Ici, $I(\cdot,\cdot)$ désigne la mesure d'information mutuelle entre deux variables, et $H(\cdot)$ la mesure d'entropie. □

Ce théorème permet de fournir un cadre théorique pour une partie des points énoncés dans la définition 1 proposée par Rolinek et al. [Rolinek *et al.*, 2019]. L'analyse du comportement des variances peut être étendue à l'ensemble des données d'entraînement $\boldsymbol{x} \in \mathbb{R}^N$. L'espérance calculée sur $p_{data}$ peut être appliquée aux résultats du théorème 1, conduisant à une seconde formalisation.

**Théorème 2** (Régime polarisé de $\boldsymbol{\sigma}^2(\boldsymbol{x};\boldsymbol{\phi})$, sur l'ensemble des données d'entraînement $\boldsymbol{x}$)**.** *Dans le cadre d'un régime polarisé au sein de l'espace latent d'un VAE, où la distribution a priori est définie comme* $p(z) = \mathcal{N}(0,1)$ *et la distribution a posteriori comme* $q_{\boldsymbol{\phi}}(z|\boldsymbol{x}) = \mathcal{N}(\mu(\boldsymbol{x};\boldsymbol{\phi}), \sigma^2(\boldsymbol{x};\boldsymbol{\phi}))$*, l'espérance de la distribution des variances inférées pour les variables actives est inférieure à celle de la distribution des variances inférées pour les variables passives lorsqu'elle est calculée sur l'ensemble des données d'entraînement :*

$$\mathbb{E}_{p_{data}(\boldsymbol{x})}\big[\boldsymbol{\sigma}_j^2\left(\boldsymbol{x};\boldsymbol{\phi}\right)\big] < \mathbb{E}_{p_{data}(\boldsymbol{x})}\big[\boldsymbol{\sigma}_k^2\left(\boldsymbol{x};\boldsymbol{\phi}\right)\big], \forall j \in \mathcal{Z}_a, \forall k \in \mathcal{Z}_p \tag{3.7}$$

Ce théorème corrobore le point (a) énoncé dans la proposition 2 et est cohérent avec l'analyse des courbes illustrées dans la Figure 3.2.

Formaliser le comportement de $Var(\boldsymbol{\sigma}^2(\boldsymbol{x};\boldsymbol{\phi}))$ est un sujet plus complexe. En effet, il dépend de la robustesse du réseau, ce qui rend difficile une justification théorique. Selon les images présentes dans le jeu de données, certaines peuvent posséder des caractéristiques moins représentées, engendrant une certaine dispersion dans les variances inférées. Toutefois, en examinant la Figure 3.2, on peut constater une différence d'écart-type en fonction de l'information contenue dans la variable associée. Cette observation peut s'expliquer par le fait que lorsque le modèle encode une variable active, sa variance inférée par l'encodeur est faible pour une image et reste constante sur le jeu de données. Néanmoins, ces dernières peuvent être modélisées à travers une distribution à queue lourde favorisant des valeurs proches de zéro, impliquant une dispersion plus importante et contredisant le point (b) énoncé dans la proposition 2 par Bonheme et Grzes[Bonheme et Grzes, 2021].

Notre analyse a également porté sur le comportement observé pour les distributions empiriques de $\boldsymbol{\mu}(\boldsymbol{x};\boldsymbol{\phi})$. En effet, l'espérance de la distribution des moyennes inférées pour des variables actives tend à être plus élevée que celle inférée pour variables passives, comme le démontre le théorème suivant.

**Théorème 3** (Espérance de $\boldsymbol{\mu}(\boldsymbol{x};\boldsymbol{\phi})$ sur l'ensemble des données d'entraînement $\boldsymbol{x}$, au sein d'un régime polarisé)**.** *Dans le cadre d'un régime polarisé au sein de l'espace latent d'un VAE, où la distribution a priori est modélisée telle que* $p(z) = \mathcal{N}(0,1)$ *et la distribution a posteriori telle que* $q_{\boldsymbol{\phi}}(z|\boldsymbol{x}) = \mathcal{N}(\mu(\boldsymbol{x};\boldsymbol{\phi}), \sigma^2(\boldsymbol{x};\boldsymbol{\phi}))$*, le moment d'ordre deux de la distribution des moyennes inférées pour les variables actives est supérieur à celui de distribution des moyennes inférées pour les variables passives :*

$$\mathbb{E}_{p_{data}(\boldsymbol{x})}\left[\mu_j^2\left(\boldsymbol{x},\boldsymbol{\phi}\right)\right] > \mathbb{E}_{p_{data}(\boldsymbol{x})}\left[\mu_k^2\left(\boldsymbol{x},\boldsymbol{\phi}\right)\right], \forall j \in \mathcal{Z}_a, \forall k \in \mathcal{Z}_p. \tag{3.8}$$

*Preuve.* Dans le contexte de la modélisation usuelle du VAE, la distribution *a priori* est similaire pour une variable active et une variable passive. De cette façon, on obtient :

$$\begin{aligned} &Var(z_j) = Var(z_k) = 1 \\ &\Rightarrow \sigma_j^2 = \sigma_k^2 \\ &\Rightarrow \mathbb{E}_{p_{data}(\boldsymbol{x})}\left[\sigma_j^2\left(\boldsymbol{x};\boldsymbol{\phi}\right)\right] + \mathbb{E}_{p_{data}(\boldsymbol{x})}\left[\mu_j^2\left(\boldsymbol{x};\boldsymbol{\phi}\right)\right] = \mathbb{E}_{p_{data}(\boldsymbol{x})}[\sigma_k^2(\boldsymbol{x};\boldsymbol{\phi})] + \mathbb{E}_{p_{data}(\boldsymbol{x})}[\mu_k^2(\boldsymbol{x};\boldsymbol{\phi})], \end{aligned} \tag{3.9}$$

pour $z_j \in \mathcal{Z}_a, z_k \in \mathcal{Z}_p$ et $Var(\cdot)$ est mesurée sur $p_{data}$. Par ailleurs, le théorème 2 nous informe que l'espérance des variances inférées sur l'ensemble du jeu de données pour les variables actives est plus petite que pour les variables passives, de sorte que

$$\mathbb{E}_{p_{data}(\boldsymbol{x})}\left[\sigma_j^2\left(\boldsymbol{x};\boldsymbol{\phi}\right)\right] < \mathbb{E}_{p_{data}(\boldsymbol{x})}\left[\sigma_k^2\left(\boldsymbol{x};\boldsymbol{\phi}\right)\right]. \tag{3.10}$$

La combinaison de ces résultats aboutit à l'équation suivante :

$$\mathbb{E}_{p_{data}(\boldsymbol{x})}\left[\mu_j^2\left(\boldsymbol{x};\boldsymbol{\phi}\right)\right] > \mathbb{E}_{p_{data}(\boldsymbol{x})}\left[\mu_k^2\left(\boldsymbol{x};\boldsymbol{\phi}\right)\right], \forall j \in \mathcal{Z}_a, \forall k \in \mathcal{Z}_p. \tag{3.11}$$

□

Ce théorème permet également de corroborer certains points de la proposition 1, et est en adéquation avec l'analyse des courbes représentées dans la Figure 3.1.

De plus, comme indiqué dans le point (b) de la proposition 2, la variance de la distribution des moyennes actives est supérieure à celle de la distribution des moyennes passives. Ce résultat est démontré dans le théorème suivant.

**Théorème 4** (Variance de $\boldsymbol{\mu}(\boldsymbol{x};\boldsymbol{\phi})$ sur l'ensemble des données d'entraînement $\boldsymbol{x}$, au sein d'un régime polarisé)**.** *Dans le cadre d'un régime polarisé au sein de l'espace latent d'un VAE, où la distribution a priori est modélisée telle que* $p(z) = \mathcal{N}(0,1)$ *et la distribution a posteriori telle que* $q_{\boldsymbol{\phi}}(z|\boldsymbol{x}) = \mathcal{N}(\mu(\boldsymbol{x};\boldsymbol{\phi}), \sigma^2(\boldsymbol{x};\boldsymbol{\phi}))$*, la variance de la distribution des moyennes inférées pour les variables actives est supérieure à celle de la distribution inférée pour les variables passives :*

$$Var\left(\mu_j\left(\boldsymbol{x},\boldsymbol{\phi}\right)\right) > Var\left(\mu_k\left(\boldsymbol{x},\boldsymbol{\phi}\right)\right), \forall j \in \mathcal{Z}_a, \forall k \in \mathcal{Z}_p. \tag{3.12}$$

*Preuve.* À l'instar de la preuve du théorème 2, $Var(z_j) = Var(z_k)$. De plus, la variance peut être décomposée de sorte que

$$Var(\boldsymbol{z}) = \mathbb{E}_{p_{data}(\boldsymbol{x})}\left[Var\left(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})\right)\right] + Var\left(\mathbb{E}_{p_{data}(\boldsymbol{x})}[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})]\right), \tag{3.13}$$

ce qui nous permet d'obtenir les équivalences suivantes :

$$\begin{aligned} &\sigma_j^2 = \sigma_k^2 \\ &\Leftrightarrow \mathbb{E}_{p_{data}(\boldsymbol{x})}\big[\sigma_j^2(\boldsymbol{x};\boldsymbol{\phi})\big] + Var\big(\mu_j(\boldsymbol{x};\boldsymbol{\phi})\big) = \mathbb{E}_{p_{data}(\boldsymbol{x})}\big[\sigma_k^2(\boldsymbol{x};\boldsymbol{\phi})\big] + Var\big(\mu_k(\boldsymbol{x};\boldsymbol{\phi})\big). \end{aligned} \tag{3.14}$$

Par ailleurs, d'après le théorème 2, l'espérance des variances inférées sur l'ensemble du jeu de données est plus petite pour les variables actives que pour les variables passives, de sorte que

$$\mathbb{E}_{p_{data}(\boldsymbol{x})}\big[\sigma_j^2\left(\boldsymbol{x};\boldsymbol{\phi}\right)\big] < \mathbb{E}_{p_{data}(\boldsymbol{x})}\big[\sigma_k^2(\boldsymbol{x};\boldsymbol{\phi})\big]. \tag{3.15}$$

La combinaison de ces résultats aboutit à l'équation finale

$$Var\big(\mu_j\left(\boldsymbol{x};\boldsymbol{\phi}\right)\big) > Var\big(\mu_k\left(\boldsymbol{x};\boldsymbol{\phi}\right)\big), \forall j \in \mathcal{Z}_a, \forall k \in \mathcal{Z}_p. \tag{3.16}$$

□

Ces théorèmes contribuent à établir formellement le comportement bimodal constaté dans un espace latent polarisé. Par conséquent, cette étude offre une meilleure appréhension de la manière dont un VAE converge vers un espace latent désentrelacé.
Ainsi, en nous appuyant sur les résultats obtenus, nous avons développé une première approche visant à favoriser un espace latent bien polarisé, dénommé le Normal-Gamma Variational Auto-Encoder (NGVAE). Ce dernier est détaillé dans la section suivante.

## 3.2 Approche proposée pour une séparation de l'espace latent

La propriété de compacité du désentrelacement, qui mesure le fait qu'un facteur génératif soit encodé dans au plus une variable latente, revêt une grande importance, notamment lorsqu'il s'agit de s'intéresser à un facteur génératif spécifique. Par ailleurs, le comportement bimodal du régime polarisé peut être favorisé en introduisant un biais inductif dans le VAE. C'est l'idée qui est développée dans cette section.

### 3.2.1 Modèle hiérarchique bayésien induisant la polarisation

Deux types de comportements peuvent être observés dans le cadre d'un régime polarisé, comme démontré dans le théorème 1. D'une part, les variables porteuses d'information sont associées à des variances apprises très faibles. D'autre part, les autres variables subissent un phénomène d'effondrement *a posteriori* et présentent ainsi des variances plus élevées. Il est justifié d'introduire des informations *a priori* sur les variances pour encourager cette séparation de l'espace latent. Partant de ce constat, nous présentons un nouveau modèle de VAE, dans lequel nous imposons à l'architecture de concentrer l'information dans un nombre limité de variables latentes et de ne pas utiliser les autres variables, contribuant à renforcer la propriété de compacité et à favoriser un ajustement automatique du nombre minimal de variables latentes nécessaire à un encodage désentrelacé des facteurs génératifs.

Dans cet objectif, nous proposons le modèle NGVAE illustré dans la Figure 3.3. La modélisation proposée étend celle du Vanilla-VAE en introduisant des variables auxiliaires dans l'espace latent pour établir un modèle bayésien hiérarchique. Elles permettent de considérer des variances qui ne sont plus déterministes. Au lieu de cela, nous leur

attribuons des distributions dont les paramètres dépendent des probabilités que les variables latentes correspondantes soient actives. Ces dernières sont obtenues en tant que sorties supplémentaires de l'encodeur et permettent de diviser directement l'espace latent en deux sous-ensembles. Pour des raisons de simplicité dans les calculs, nous simulons les inverses des variances des variables latentes au lieu des variances elles-mêmes. Ces inverses variances sont regroupées dans un vecteur $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_K)$, où $K$ représente la dimension de l'espace latent, de sorte qu'une variable latente soit représentée par un couple $(z_k, \lambda_k)$.

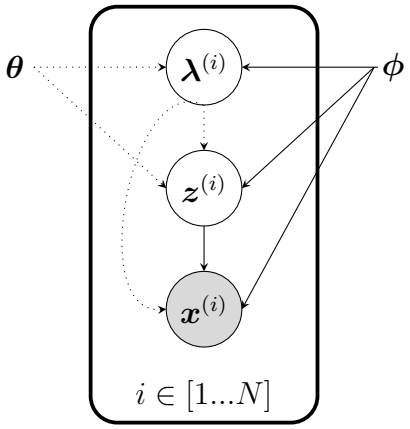



Fig 3.3: Représentation graphique du NGVAE pour un jeu de données $\boldsymbol{x} \in \mathbb{R}^N$. Les flèches en pointillées correspondent au processus d'inférence et les flèches solides représentent le processus de génération.

**L'encodeur:** Notre modèle se distingue par la distribution *a posteriori* que nous avons choisie pour les couples $(z_k, \lambda_k)$. Cette distribution est définie comme un mélange de lois Normales-Gamma, de la façon suivante :

$$\begin{aligned} q_{\boldsymbol{\phi}}\big(z_k, \lambda_k|\boldsymbol{x}\big) =& \pi_k\big(\boldsymbol{x}, \boldsymbol{\phi}\big)\mathcal{NG}\big(z_k, \lambda_k; \mu_k(\boldsymbol{x}, \boldsymbol{\phi}), \alpha_a, \beta_a\big) \\ &+ \big(1 - \pi_k(\boldsymbol{x}, \boldsymbol{\phi})\big)\mathcal{NG}\big(z_k, \lambda_k; \mu_k(\boldsymbol{x}, \boldsymbol{\phi}), \alpha_p, \beta_p\big). \end{aligned} \tag{3.17}$$

Dans cette distribution, les moyennes $\{\mu_k(\boldsymbol{x}, \boldsymbol{\phi})\}_{k=1,...,K}$ et les probabilités $\{\pi_k(\boldsymbol{x}, \boldsymbol{\phi})\}_{k=1,...,K}$ sont toutes deux fournies par l'encodeur. Nous avons choisi la modélisation des variables latentes à travers une distribution Normale-Gamma, car elle correspond à l'*a priori* conjugué pour les paramètres $(\boldsymbol{\mu}, \boldsymbol{\lambda})$ d'une distribution gaussienne. De plus, la distribution marginale de la loi sur $\boldsymbol{z}$ correspond à une distribution de Student non généralisée, qui présente une queue lourde. La démonstration est proposée en Annexe B. En effet, comme nous l'avons soulevé dans le chapitre précédent, la distribution gaussienne usuelle n'est pas suffisamment complexe pour représenter la distribution réelle sous-jacente des variables latentes [Kim *et al.*, 2019a; Tomczak et Welling, 2018]. Cette approche nous permet ainsi une modélisation plus riche que la loi gaussienne usuellement utilisée dans les approches de VAE.

Les distributions *a priori* sont désormais définies pour les paires $(\boldsymbol{z}, \boldsymbol{\lambda})$. Ces dernières sont également représentées sous la forme de lois *a priori* Normales-Gamma, avec une moyenne nulle et un ensemble de paramètres $(\alpha_{prior}, \beta_{prior})$ similaire aux paramètres $(\alpha_p, \beta_p)$ définis pour une distribution Gamma utilisée pour les variables passives.

En ce qui concerne les ensembles d'hyperparamètres $(\alpha_a,\ \beta_a)$ et $(\alpha_p,\ \beta_p)$, définis respectivement pour les distributions Gamma des inverses variances actives et passives, nous les avons choisis de manière à favoriser la polarisation au sein du modèle hiérarchique bayésien.

**Théorème 5** (Impact des paramètres $\alpha_a$ et $\alpha_p$ sur la polarisation de l'espace latent d'un VAE hiérarchique basé sur une distribution Normale-Gamma)**.** *Dans le contexte d'un modèle de VAE bayésien hiérarchique, où la distribution a priori est définie comme* $p(z, \lambda) = \mathcal{NG}(z, \lambda; \mu_{prior}, \alpha_{prior}, \beta_{prior})$ *et la distribution a posteriori, pour laquelle le paramètre de forme est fixé tel que* $\alpha = a$*, définie par* $q_{\boldsymbol{\phi}}(z, \lambda|\boldsymbol{x}) = \pi\mathcal{NG}(z, \lambda; \mu, \alpha = a, \beta_a) + (1 - \pi)\mathcal{NG}(z, \lambda; \mu, \alpha = a, \beta_p)$*, si* $\beta_a < \beta_p$ *alors la mesure d'information mutuelle entre les données d'entrées et les variables considérées comme actives est supérieure à l'information mutuelle entre les données d'entrées et les variables considérées comme passives :*

$$I\big(z_j; \boldsymbol{x}\big) > I\big(z_k; \boldsymbol{x}\big), \forall j \in \mathcal{Z}_a, \forall k \in \mathcal{Z}_p. \tag{3.18}$$

*Preuve.* La mesure d'information mutuelle entre deux variables aléatoires $z$ et $x$ s'écrit :

$$I\big(z; \boldsymbol{x}\big) = H\big(z\big) - H\big(z|\boldsymbol{x}\big), \tag{3.19}$$

où $H(\cdot)$ mesure l'entropie et $H(\cdot|\boldsymbol{x})$ l'entropie conditionnelle définie par $\mathbb{E}_{p(x)}[H(\cdot|X = x)]$. D'après la modélisation adoptée dans le NGVAE, on a :

$$\begin{aligned}
p\big(z_a\big) &= p\big(z_p\big), \\
p(\lambda_a) &= p(\lambda_p) \\
p(z_a|\lambda_a) &= p(z_p|\lambda_p), \\
q_{\boldsymbol{\phi}}(z_a, \lambda_a|\boldsymbol{x}) &= \mathcal{N}\big(z_a; \mu(\boldsymbol{x}, \boldsymbol{\phi}), \lambda_a^{-1}\big)\mathcal{G}\big(\lambda_a; \alpha_a, \beta_a\big), \\
q_{\boldsymbol{\phi}}(z_p, \lambda_p|\boldsymbol{x}) &= \mathcal{N}\big(z_p; \mu(\boldsymbol{x}, \boldsymbol{\phi}), \lambda_p^{-1}\big)\mathcal{G}\big(\lambda_p; \alpha_p, \beta_p\big),
\end{aligned} \tag{3.20}$$

avec $\mathcal{G}(\cdot)$ la distribution Gamma, $(z_a, \lambda_a) \in \mathbb{Z}_a^{'}$ l'ensemble des variables actives, $(z_p, \lambda_p) \in \mathbb{Z}_p^{'}$ l'ensemble des variables passives, $(\alpha_a, \beta_a)$ l'ensemble des hyperparamètres définis pour la distribution Gamma active et $(\alpha_p, \beta_p)$ l'ensemble des hyperparamètres définis pour la distribution Gamma passive.

D'après l'équation (3.19), l'information mutuelle des variables latentes actives s'écrit :

$$I\big(z_a; \boldsymbol{x}\big) = H(z_a) - \mathbb{E}_{p_{data}(\boldsymbol{x})}\big[H(z_a|X = \boldsymbol{x})\big]. \tag{3.21}$$

De la même façon, cette grandeur s'écrit :

$$I\big(z_p; \boldsymbol{x}\big) = H(z_p) - \mathbb{E}_{p_{data}(\boldsymbol{x})}\big[H(z_p|X = \boldsymbol{x})\big], \tag{3.22}$$

pour des variables passives.
Par ailleurs, les hypothèses du modèle définies dans l'équation (3.20) concernant les distributions *a priori* des variables actives et passives impliquent la relation suivante :

$$\begin{aligned}
&p(z_a) = p(z_p) \\
&\Leftrightarrow H(z_a) = H(z_p),
\end{aligned} \tag{3.23}$$

de telle sorte que la comparaison des grandeurs $I(z_a; \boldsymbol{x})$ et $I(z_p; \boldsymbol{x})$ se résume en une alayse des termes situés à droite dans les équations (3.21) et (3.22).

Considérons dans un premier temps la distribution $q_\phi(z|\boldsymbol{x})$, avec $q_{\boldsymbol{\phi}}(z|\lambda,\boldsymbol{x}) = \mathcal{N}(z;\mu,\lambda^{-1})$ et $q_{\boldsymbol{\phi}}(\lambda|\boldsymbol{x}) = \mathcal{G}(\lambda;\alpha,\beta)$. Par souci de compréhension de lecture, notons $\mu(\boldsymbol{x};\boldsymbol{\phi}) \triangleq \mu$.

$$\begin{aligned}
q_{\boldsymbol{\phi}}(z|\boldsymbol{x}) &= \int_\lambda \mathcal{N}(z;\mu,\lambda^{-1})\mathcal{G}(\lambda;\alpha,\beta)d\lambda \\
&= \int_\lambda \sqrt{\frac{\lambda}{2\pi}} \exp\left[-\frac{1}{2}(z-\mu)^2\lambda\right] \cdot \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{(\alpha-1)} \exp\big[-\beta\lambda\big]\mathbb{1}_{[0,+\infty[}(\lambda)d\lambda \\
&= \sqrt{\frac{\frac{\alpha}{\beta}}{2\alpha\pi}} \cdot \frac{\Gamma(\frac{2\alpha+1}{2})}{\Gamma(\alpha)} \cdot \left(1 + \frac{(z-\mu)^2}{2\alpha \cdot \left(\sqrt{\frac{\beta}{\alpha}}\right)^2}\right)^{-\frac{2\alpha+1}{2}} \\
&= \sqrt{\frac{\alpha}{\beta}} T\left(\frac{z-\mu}{\sqrt{\frac{\beta}{\alpha}}}, 2\alpha\right),
\end{aligned} \tag{3.24}$$

où l'on dénote par $T(\cdot, 2\alpha)$ une distribution de Student à $2\alpha$ degrés de libertés. Ce résultat est démontré dans l'Annexe B. Considérons désormais $H(z|\boldsymbol{x}) = \mathbb{E}_{p_{data}(\boldsymbol{x})}[H(z|X=\boldsymbol{x})]$, avec $H(z|X=\boldsymbol{x}) = -\int_{\boldsymbol{z}} q_{\boldsymbol{\phi}}(z|\boldsymbol{x}) \log q_{\boldsymbol{\phi}}(z|\boldsymbol{x})dz$. D'après sa définition, l'entropie de $z$ s'écrit :

$$H(z|X=x) = -\int_z \left[\sqrt{\frac{\alpha}{\beta}} T\left(\frac{z-\mu}{\sqrt{\frac{\beta}{\alpha}}}, 2\alpha\right) \log\left(\sqrt{\frac{\alpha}{\beta}} T\left(\frac{z-\mu}{\sqrt{\frac{\beta}{\alpha}}}, 2\alpha\right)\right)\right] dz. \tag{3.25}$$

Ainsi, en appliquant le changement de variables $y = \frac{(z-\mu)}{\sqrt{\frac{\beta}{\alpha}}}$ dans l'intégrale, cela s'écrit :

$$\begin{aligned}
H(z|X=x) &= -\int_y \left[T(y;2\alpha)\log\left(\sqrt{\frac{\alpha}{\beta}}T(y;2\alpha)\right)\right] dy \\
&= -\int_{\boldsymbol{y}} \left[T(y;2\alpha)\left[\log\left(\sqrt{\frac{\alpha}{\beta}}\right) + \log(T(y;2\alpha))\right]\right] dy \\
&= -\log\left(\sqrt{\frac{\alpha}{\beta}}\right)\int_y \Big[T(y;2\alpha)\Big]dy - \int_y \Big[T(y;2\alpha)\log\left(T(y;2\alpha)\right)\Big]dy \\
&= -\frac{1}{2}\log\left(\frac{\alpha}{\beta}\right) + H\big(y\big), \text{ avec } y \sim T(y,2\alpha).
\end{aligned} \tag{3.26}$$

En considérant la propriété suivante :

$$y \sim T(y;\nu) \Rightarrow H(y) = \log\left(\sqrt{\nu}B\left(\frac{1}{2},\frac{\nu}{2}\right)\right) + \frac{\nu+1}{2}\left[\psi\left(\frac{\nu+1}{2}\right) - \psi\left(\frac{\nu}{2}\right)\right], \tag{3.27}$$

où $T(\cdot;\nu)$ est une distribution de Student à $\nu$ degrés de liberté et $B$ la fonction Beta, l'équation (3.26) s'écrit :

$$\begin{aligned}
H(z|X=x) &= -\frac{1}{2}\log\left(\frac{\alpha}{\beta}\right) + \log\left(\sqrt{2\alpha}B\left(\frac{1}{2},\frac{2\alpha}{2}\right)\right) + \frac{2\alpha+1}{2}\left[\psi\left(\frac{2\alpha+1}{2}\right) - \psi\left(\frac{2\alpha}{2}\right)\right] \\
&= -\frac{1}{2}\log\left(\frac{\alpha}{\beta}\right) + \log\left(\sqrt{2\alpha}\frac{\sqrt{\pi}\Gamma(\alpha)}{\Gamma(\alpha+\frac{1}{2})})\right) + \frac{2\alpha+1}{2}\left[\psi\left(\frac{2\alpha+1}{2}\right) - \psi\left(\frac{2\alpha}{2}\right)\right] \\
&= \frac{1}{2}\log(\beta) + \frac{1}{2}\log(2) + \frac{1}{2}\log(\pi) + \log\big(\Gamma(\alpha)\big) - \log\left(\Gamma\left(\alpha+\frac{1}{2}\right)\right) \\
&+ \left(\alpha+\frac{1}{2}\right)\left[\psi\left(\alpha+\frac{1}{2}\right) - \psi\left(\alpha+\frac{1}{2}\right)\right].
\end{aligned} \tag{3.28}$$

Comme $H(z|X = x)$ est indépendant de $\boldsymbol{x}$ puisque selon nos hypothèses initiales $\alpha$ ne dépend pas de $\boldsymbol{x}$, alors $\mathbb{E}_{p_{data}(\boldsymbol{x})}\big[H(z|X = \boldsymbol{x})\big] = H(z|X = \boldsymbol{x})$.

Finalement, afin de comparer $I(z; \boldsymbol{x})$ dans le cadre d'un régime polarisé pour des variables passives et actives, il suffit d'analyser les variations de la fonction suivante :

$$\begin{aligned} f(a,b) = \frac{1}{2}\log(b) + \frac{1}{2}\log(2) + \frac{1}{2}\log(\pi) + \log\big(\Gamma(a)\big) - \log\left(\Gamma\left(a+\frac{1}{2}\right)\right) \\ + \frac{2a+1}{2}\left[\psi\left(\alpha+\frac{1}{2}\right) - \psi\left(a+\frac{1}{2}\right)\right]. \end{aligned} \tag{3.29}$$

Étudions le signe de $\frac{\partial f(a,b)}{\partial b}$:

$$\frac{\partial f(a,b)}{\partial b} = \frac{1}{2b} > 0,\ \forall a, b \in \mathbb{R}^{*+}. \tag{3.30}$$

Ainsi, pour $a$ fixé, $f(a,b)$ est croissante sur l'ensemble $\mathbb{R}^{*+}$. On peut en conclure la relation suivante :

$$\begin{aligned} &\beta_a < \beta_p \\ &\Rightarrow f(\alpha = a, \beta_a) < f(\alpha = a, \beta_p) \\ &\Rightarrow H(z_a|X = \boldsymbol{x}; a, \beta_a) < H(z_p|X = \boldsymbol{x}; a, \beta_p) \\ &\Rightarrow -H(z_a|X = \boldsymbol{x}; a, \beta_a) > -H(z_p|X = \boldsymbol{x}; a, \beta_p) \\ &\Rightarrow I(z_a; \boldsymbol{x}) > I(z_p; \boldsymbol{x}). \end{aligned} \tag{3.31}$$

□

Pour nos modélisations, nous avons choisi spécifiquement les valeurs ($\alpha_a$, $\beta_a$) et ($\alpha_p$, $\beta_p$) correspondant aux lois Inverse-Gamma représentées dans la Figure 3.4 pour la variance. De cette façon, leur définition vise à favoriser respectivement des valeurs proches de un ou de zéro pour $\mathbb{E}_{p_{data}(\boldsymbol{x})}[\lambda_k^{-1}] = \frac{\alpha}{\beta}$, comme indiqué dans le théorème 2. Par ailleurs, afin de favoriser l'information mutuelle avec les données d'entrée pour les variables actives, nous avons défini $\beta_a < \beta_p$ en cohérence avec le théorème 5, résultant aux valeurs suivantes : $(\alpha_a, \beta_a) = (1.0009, 0.009)$ et $(\alpha_p, \beta_p) = (1002, 1001)$.

**Le décodeur :** Le modèle génératif du NGVAE vise à définir les paramètres de la distribution de *vraisemblance* $p_\theta(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{\lambda})$. Lorsque les données d'entrée sont des images en couleur, une approche appropriée consiste à émettre l'hypothèse d'indépendance des pixels et à définir cette distribution de *vraisemblance* comme un produit de lois normales, c'est-à-dire $p_{\boldsymbol{\theta}}(\boldsymbol{x}_i|\boldsymbol{z}, \boldsymbol{\lambda}) = \mathcal{N}(\boldsymbol{x}_i|m_i(\boldsymbol{z}, \boldsymbol{\lambda}), \sigma^2; \boldsymbol{\theta})$, où l'indice $i$ correspond à l'indice du pixel d'entrée, le paramètre de moyenne $\boldsymbol{m}_i(\boldsymbol{z}, \boldsymbol{\lambda})$, correspond à la sortie du réseau décodeur, tandis que l'écart type, $\boldsymbol{\sigma}$, est un hyperparamètre fixé avant apprentissage.

Cependant, comme mentionné dans la section 1.4.2, une approche courante consiste à modéliser cette distribution comme un produit de distributions de Bernoulli, c'est-à-dire $p_{\boldsymbol{\theta}}(\boldsymbol{x}_i|\boldsymbol{z}, \boldsymbol{\lambda}) = \mathcal{B}(x_i|p_i(\boldsymbol{z}, \boldsymbol{\lambda}; \boldsymbol{\theta}))$, où le paramètre de probabilité $p_i(\boldsymbol{z}, \boldsymbol{\lambda}; \boldsymbol{\theta})$ correspond à la sortie du décodeur pour un pixel $i$. Cette méthode ne tient pas compte de la nature des pixels d'entrée, ce qui permet d'éviter de définir l'hyperparamètre $\sigma^2$. Afin de minimiser tout biais potentiel lors de l'évaluation et de la comparaison de notre approche avec les modèles de l'état de l'art, nous avons choisi de modéliser la distribution de *vraisemblance* sous forme d'une loi de Bernoulli.

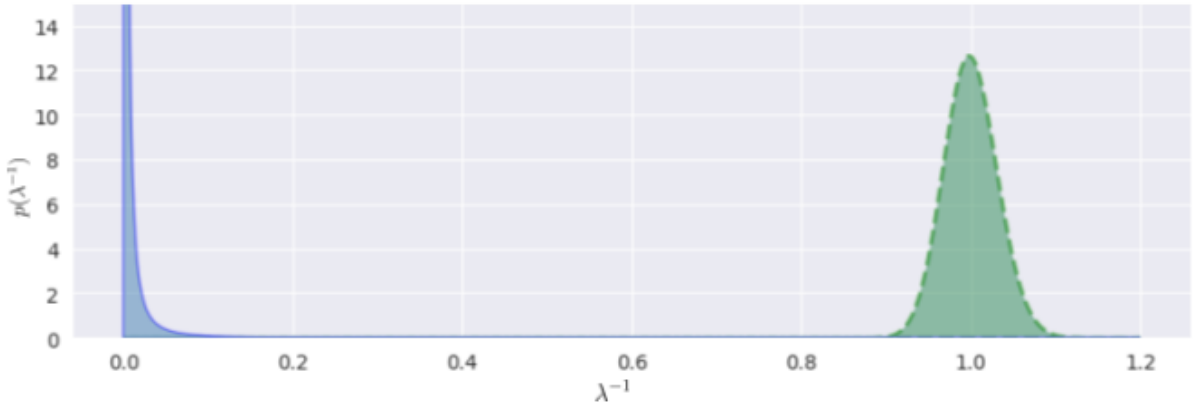



Fig 3.4: Distributions de la variance des variables latentes en fonction de l'information contenue dans ces variables. La courbe bleue représente la densité de probabilité d'une loi Inverse-Gamma de paramètres $(\alpha_a, \beta_a)$, qui modélise la variance des variables actives. La courbe verte représente la densité de probabilité d'une loi Inverse-Gamma avec les paramètres $(\alpha_p, \beta_p)$, qui modélise la variance des variables passives.

L'architecture globale du NGVAE, composée des réseaux encodeur et décodeur, est illustrée dans la Figure 3.5. Les couches constituant le modèle d'encodage et le modèle de décodage sont schématisées dans cette Figure comme une succession de couches de neurones denses. Toutefois, il est également pertinent de considérer une succession de couches de convolution dans le contexte dans lequel les données d'entrées sont des images.

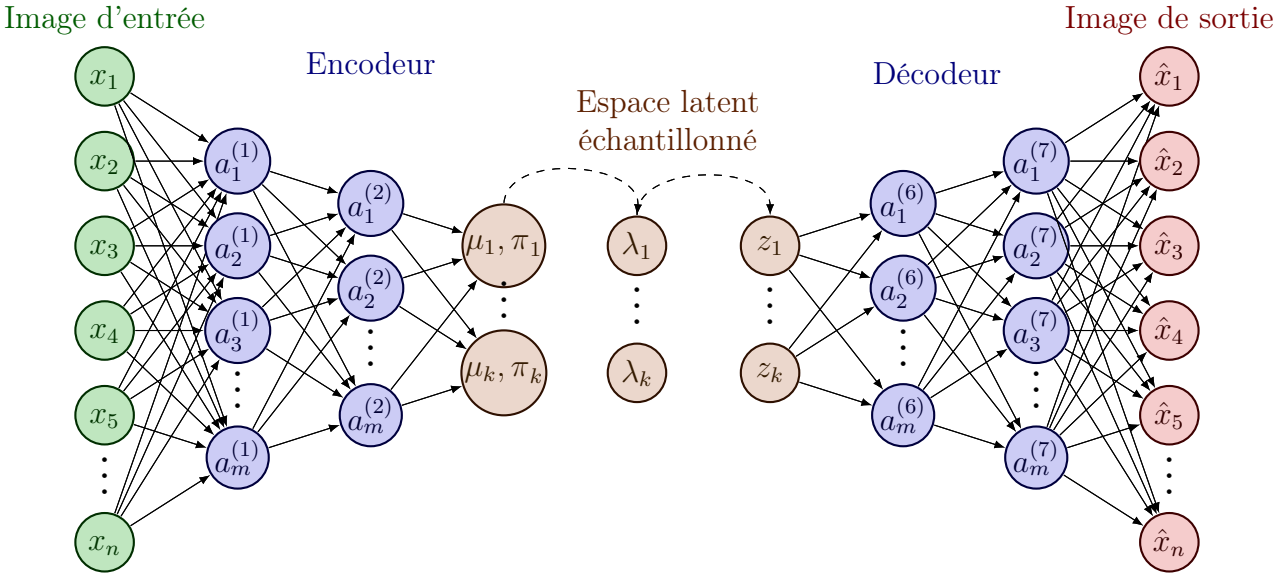



Fig 3.5: Illustration du Normal-Gamma Variationnal Auto-Encodeur.

**La fonction objectif :** À l'instar du VAE classique, pour l'entraînement du modèle NGVAE, le réglage de l'ensemble des paramètres variationnels $(\boldsymbol{\phi}, \boldsymbol{\theta})$ s'effectue à l'aide d'un algorithme de descente de gradient stochastique qui minimise une fonction de coût. L'enrichissement de l'espace latent et la modélisation de la distribution *a posteriori* sous forme d'une loi Normale-Gamma conduisent à la fonction suivante :

$$\mathcal{L}_{NGVAE}(\boldsymbol{\phi}, \boldsymbol{\theta}) = \mathbb{E}_{p_{data}(\boldsymbol{x})}\left[\mathbb{E}_{q_\phi(\boldsymbol{z}, \boldsymbol{\lambda}|\boldsymbol{x})}\left[-\log p_\theta(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{\lambda})\right] + \sum_{k=1}^{K} D_{KL}\left[q_\phi(z_k, \lambda_k|\boldsymbol{x})||p(z_k, \lambda_k)\right]\right]. \tag{3.32}$$

Il convient de noter que le choix de modélisation pour la distribution *a priori* $p(z_k, \lambda_k)$ comme une loi non informative permet de contraindre le terme de KL-divergence de l'équation (3.32) à agir comme une régularisation, incitant ainsi l'encodeur à inférer le nombre minimal de variables actives, ce qui correspond au comportement attendu pour la polarisation.

Lorsque la fonction de *vraisemblance* $p_\theta(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{\lambda})$ est modélisée comme une distribution de Bernoulli, le terme à l'intérieur de l'espérance dans l'équation (3.32) peut être formulé de la façon suivante, en émettant l'hypothèse d'indépendance des pixels :

$$\begin{aligned} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{\lambda}) &= \log \prod_i p_{\boldsymbol{\theta}}(x_i|\boldsymbol{z}, \boldsymbol{\lambda}) \\ &= \sum_i \log\big(p_i^{x_i}(1-p_i)^{(1-x_i)}\big) \\ &= \sum_i x_i \log(p_i) + (1-x_i)\log(1-p_i), \end{aligned} \tag{3.33}$$

où, $x_i$ représente la valeur du pixel $i$ dans l'image d'entrée, et $p_i \triangleq p_i(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{\lambda})$ est la sortie du décodeur. Il est à noter que cette fonction équivaut à la fonction d'entropie croisée entre l'image d'entrée fournie à l'encodeur et l'image reconstruite par le décodeur.

Toutefois, le deuxième terme de KL-divergence dans l'équation (3.32) est difficilement calculable lorsque l'on considère une loi de mélange avec une loi simple. Afin d'éviter cette limitation, une approche consiste à ne conserver qu'une des deux lois du mélange. Or, l'utilisation d'une fonction seuil dépendante de la valeur de $\pi_k(\boldsymbol{x}; \boldsymbol{\phi})$ dans l'objectif d'attribuer les hyperparamètres $(\alpha(\pi_k), \beta(\pi_k))$ aux variables latentes $(z_k, \lambda_k^{-1})$ introduirait des problèmes de différentiabilité, une condition nécessaire pour l'utilisation des algorithmes de descente de gradients utilisés pour l'apprentissage. Afin d'éviter cela, nous suggérons d'attribuer à chaque paire de variables latentes des hyperparamètres $(\alpha(\pi_k), \beta(\pi_k))$ qui varient de manière continue en fonction de $\pi_k$, avec $\pi_k \triangleq \pi_k(\boldsymbol{x}; \boldsymbol{\phi})$ , de la façon suivante :

$$\begin{aligned} \alpha(\pi_k) =& f_r(\pi_k) \\ =& \frac{\alpha_p - \alpha_a}{1 + e^{-r(\pi_k - 0.5)}} + \alpha_a. \end{aligned} \tag{3.34}$$

Une fonction analogue à (3.34) est prise en compte pour $\beta(\pi_k)$. Le paramètre $r$ est utilisé pour ajuster la pente de la sigmoïde $f_r(\pi_k)$ de manière à ce qu'elle induise un seuillage, comme illustré dans la Figure 3.6. Au sein du NGVAE, $r$ prend la valeur 50.

Ce choix de modélisation permet de calculer analytiquement le second terme de l'équation (3.32).

En effet, la KL-divergence entre deux lois Normales-Gamma $p(z, \lambda) = \mathcal{N}(z; \mu_1, \lambda)^{-1})\mathcal{G}(\lambda; \alpha_1, \beta_1)$ et $q_{\boldsymbol{\phi}}(z, \lambda) = \mathcal{N}(z; \mu_2, \lambda^{-1})\mathcal{G}(\lambda; \alpha_2, \beta_2)$ possède la forme analytique suivante :

$$\begin{aligned} D_{KL}[p(z,\lambda)||q_{\boldsymbol{\phi}}(z,\lambda)] = \frac{1}{2}\frac{\alpha_1}{\beta_1}(\mu_2 - \mu_1)^2 + \alpha_2 \log\left(\frac{\beta_1}{\beta_2}\right) - \log\left(\frac{\Gamma(\alpha_1)}{\Gamma(\alpha_2)}\right) \\ + (\alpha_1 - \alpha_2)\psi(\alpha_1) - (\beta_1 - \beta_2)\frac{\alpha_1}{\beta_1}. \end{aligned} \tag{3.35}$$

Ce résultat est démontré dans l'Annexe C.

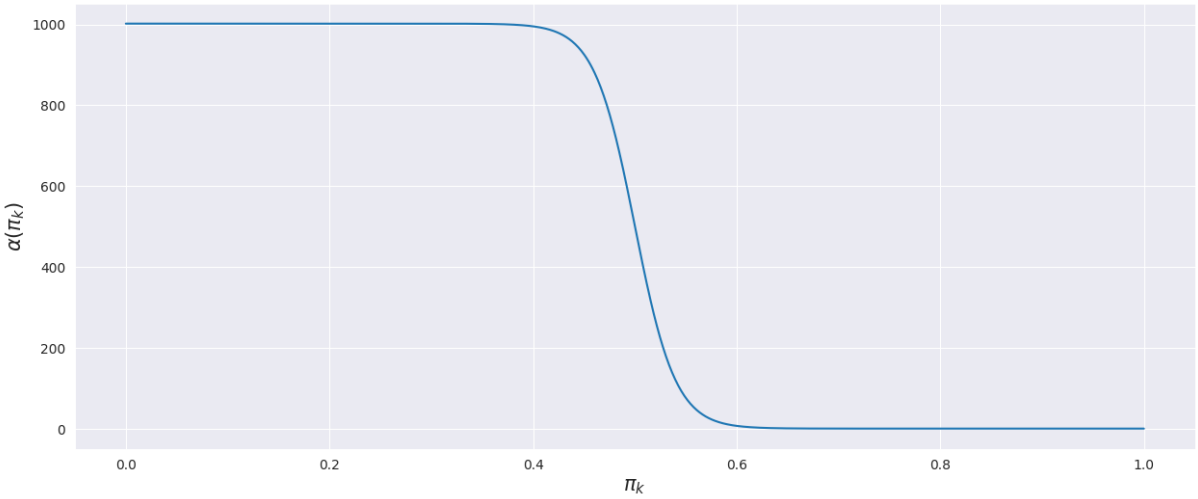



Fig 3.6: Illustration de la fonction $\alpha(\pi_k)$ définie dans l'équation (3.34), pour $\pi_k = [0, \ldots, 1]$

Dans les prochaines sections de ce chapitre, un nouvel algorithme de descente de gradient applicable au NGVAE est détaillé. Par la suite, une limitation est observée liée à la définition de la KL-divergence entre deux lois Normales-Gamma, justifiant l'inversion de l'ordre des distributions dans ce terme. Certains termes de régularisations apportés aux probabilités inférées par l'encodeur sont également explicités. Finalement, des comparaisons permettant d'évaluer la capacité de polarisation de l'approche proposée par rapport à d'autres modèles de l'état de l'art est effectuée.

### 3.2.2 Mise en œuvre de l'apprentissage

Le processus d'apprentissage du NGVAE nécessite l'utilisation d'un algorithme de descente de gradient sur des fonctions impliquant des espérances par rapport à des distributions dépendant des paramètres variationnels $(\boldsymbol{\phi}, \boldsymbol{\theta})$. Cette dépendance rend impossible le recours à des approximations classiquement utilisées dans la méthode d'optimisation stochastique du VAE. Lorsque l'on s'écarte de l'hypothèse gaussienne pour la distribution *a posteriori*, l'astuce standard de reparamétrisation proposée par Kingma et Welling [Kingma et Welling, 2014], ne peut pas être directement appliquée. Nous en avons donc développé une nouvelle, que nous expliquons dans la suite de ce chapitre.

De plus, la forme analytique de la KL-divergence entre deux distributions Normales-Gamma décrite dans l'équation (3.35) implique la pondération de $\boldsymbol{\mu}(\boldsymbol{x}, \boldsymbol{\phi})$ par le ratio entre les paramètres $(\alpha, \beta)$. Notre approche bimodale pose un problème dans ce contexte, et nous expliquons comment venir à bout de cette limitation.

Finalement, afin de favoriser le choix d'un unique mode de la distribution *a posteriori*, nous expliquons la régularisation adoptée pour les termes de probabilité inférés par l'encodeur avant d'expliquer comment ils peuvent être utilisés pour favoriser la polarisation comme un biais inductif pendant l'apprentissage.

#### Reparamétrisation de la distribution Normale-Gamma

Lors de l'entraînement du NGVAE, un algorithme de descente de gradient stochastique est utilisé. Cependant, il est essentiel de noter que sa mise en œuvre requiert de calculer des gradients d'espérances, lesquels ne peuvent pas être directement mesurés pour des distributions liées aux paramètres variationnels. Afin de résoudre ce problème, nous avons

élaboré une nouvelle technique de reparamétrisation spécifiquement adaptée au modèle proposé. Pour mieux comprendre cette méthode, commençons par examiner le calcul du gradient nécessaire à l'optimisation du NGVAE :

$$\boldsymbol{\nabla}_{\boldsymbol{\phi},\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\phi},\boldsymbol{\theta}) = \nabla_{\boldsymbol{\phi},\boldsymbol{\theta}}\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z},\boldsymbol{\lambda}|\boldsymbol{x})}\big[-\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\big] + \nabla_{\boldsymbol{\phi}} D_{KL}\big[q_{\boldsymbol{\phi}}(\boldsymbol{z},\boldsymbol{\lambda}|\boldsymbol{x})||p(\boldsymbol{z},\boldsymbol{\lambda})\big]. \tag{3.36}$$

La mesure de KL-divergence entre deux distributions Normale-Gamma pouvant être calculé analytiquement comme explicité dans l'équation (3.35), le second terme de (3.36) ne pose donc pas de difficultés. Cependant, l'espérance sur la distribution de *vraisemblance* est plus problématique et nécessite une reparamétrisation spécifique sur l'ensemble de vecteurs $(\boldsymbol{z},\boldsymbol{\lambda})$. Dans cet objectif, nous proposons une approche qui permet de remplacer le calcul du gradient de l'espérance par l'espérance du gradient. Cette nouvelle méthode d'optimisation est basée sur un algorithme de descente de gradient et peut de ce fait être appliquée aux paramètres variationnels $(\boldsymbol{\phi},\boldsymbol{\theta})$.

En désignant par $\boldsymbol{v} = (\boldsymbol{\phi},\boldsymbol{\theta})$ les paramètres variationnels et $f_{\boldsymbol{v}}(\boldsymbol{z},\boldsymbol{\lambda}) = -\log p_\theta(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{\lambda})$, nous proposons de considérer le changement de variable suivant :

$$\begin{aligned} \boldsymbol{z} &= T_2(\boldsymbol{\epsilon}_1,\boldsymbol{\epsilon_2};\boldsymbol{v}) = \boldsymbol{m}(\boldsymbol{v}) + \big(T_1(\boldsymbol{\epsilon_1};\boldsymbol{v})\big)^{-\frac{1}{2}}\boldsymbol{\epsilon_2} \\ \boldsymbol{\lambda} &= T_1(\boldsymbol{\epsilon_1};\boldsymbol{v}), \end{aligned} \tag{3.37}$$

où le vecteur $\boldsymbol{m}(\boldsymbol{v})$ stocke les moyennes calculées par le décodeur et $\boldsymbol{\epsilon}_2 \sim \mathcal{N}(0,\mathrm{I})$. Quant à la définition de $T_1$, nous nous basons sur l'approche proposée par Ruiz et al. [Ruiz *et al.*, 2016] de sorte que la $k$-ième composante du vecteur $\boldsymbol{\epsilon}_1$ satisfasse :

$$\boldsymbol{\epsilon}_{1,k} = \frac{\log(\lambda_k) - \Psi\big(\alpha(\pi_k)\big) + \log\big(\beta(\pi_k)\big)}{\sqrt{\Psi_1\big(\alpha(\pi_k)\big)}}, \tag{3.38}$$

où $\lambda_k$, $\alpha(\pi_k)$ et $\beta(\pi_k)$ dépendent implicitement de $\boldsymbol{v}$. Il est à noter que cette reparamétrisation a l'avantage de définir un vecteur $\boldsymbol{\epsilon}_1$ ne dépendant que faiblement des paramètres $\boldsymbol{v}$. En réécrivant le premier terme de l'équation (3.36) tout en prenant en compte le changement de variables, nous obtenons :

$$\begin{aligned} &\nabla_{\boldsymbol{\phi},\boldsymbol{\theta}}\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z},\boldsymbol{\lambda}|\boldsymbol{x})}\big[-\log p_\theta(\boldsymbol{x}|\boldsymbol{\lambda},\boldsymbol{z})\big] \\ &= \nabla_{\boldsymbol{v}}\mathbb{E}_{q_{\boldsymbol{v}}(\boldsymbol{z},\boldsymbol{\lambda}|\boldsymbol{x})}[f_{\boldsymbol{v}}(\boldsymbol{z},\boldsymbol{\lambda})] \\ &= \nabla_{\boldsymbol{v}}\left[\int_{\boldsymbol{z}}\int_{\boldsymbol{\lambda}} q_{\boldsymbol{v}}(\boldsymbol{z}|\boldsymbol{\lambda},\boldsymbol{x}) q_{\boldsymbol{v}}(\boldsymbol{\lambda}|\boldsymbol{x}) f_{\boldsymbol{v}}(\boldsymbol{z},\boldsymbol{\lambda}) d\boldsymbol{z} d\boldsymbol{\lambda}\right] \\ &= \nabla_{\boldsymbol{v}}\left[\int_{\boldsymbol{\epsilon}_1}\int_{\boldsymbol{\epsilon}_2} q_{\boldsymbol{v}}(T_1(\boldsymbol{\epsilon}_1;\boldsymbol{v})|\boldsymbol{x}) f_{\boldsymbol{v}}(T_2(\boldsymbol{\epsilon}_1,\boldsymbol{\epsilon_2};\boldsymbol{v}),T_1(\boldsymbol{\epsilon}_1;\boldsymbol{v})) q_{\boldsymbol{v}}(T_2(\boldsymbol{\epsilon_1},\boldsymbol{\epsilon_2};\boldsymbol{v})|T_1(\boldsymbol{\epsilon}_1;\boldsymbol{v}),\boldsymbol{x}) d\boldsymbol{\epsilon}_2 |J| d\boldsymbol{\epsilon}_1\right], \end{aligned} \tag{3.39}$$

avec $|J| = \begin{bmatrix} \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{\epsilon}_1} & \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{\epsilon}_2} \\ \frac{\partial \boldsymbol{\lambda}}{\partial \boldsymbol{\epsilon}_1} & \frac{\partial \boldsymbol{\lambda}}{\partial \boldsymbol{\epsilon}_2} \end{bmatrix}$.

Posons désormais la loi jointe de $(\boldsymbol{\epsilon}_1,\boldsymbol{\epsilon}_2)$ :

$$q(\boldsymbol{\epsilon}_1,\boldsymbol{\epsilon}_2;\boldsymbol{v}) = q_{\boldsymbol{v}}\big(T_2(\boldsymbol{\epsilon}_1,\boldsymbol{\epsilon}_2;\boldsymbol{v})|T_1(\boldsymbol{\epsilon}_1;\boldsymbol{v}),\boldsymbol{x}\big) q_{\boldsymbol{v}}\big(T_1(\boldsymbol{\epsilon}_1;\boldsymbol{v})|\boldsymbol{x}\big)|J|, \tag{3.40}$$

ce qui nous permet d'écrire :

$$\begin{aligned}
&\nabla_{\boldsymbol{\phi},\boldsymbol{\theta}} \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z},\boldsymbol{\lambda}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] \\
&= \nabla_{\boldsymbol{v}} \left[ \int_{\boldsymbol{\epsilon}_1} \int_{\boldsymbol{\epsilon}_2} g(\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2; \boldsymbol{v}) q(\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2; \boldsymbol{v}) d\boldsymbol{\epsilon}_2 d\boldsymbol{\epsilon}_1 \right] \\
&= \int_{\boldsymbol{\epsilon}_1} \int_{\boldsymbol{\epsilon}_2} \nabla_{\boldsymbol{v}} q(\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2; \boldsymbol{v}) g(\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2; \boldsymbol{v}) d\boldsymbol{\epsilon}_2 d\boldsymbol{\epsilon}_1 + \int_{\boldsymbol{\epsilon}_1} \int_{\boldsymbol{\epsilon}_2} g(\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2; \boldsymbol{v}) \nabla_{\boldsymbol{v}} q(\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2; \boldsymbol{v}) d\boldsymbol{\epsilon}_2 d\boldsymbol{\epsilon}_1.
\end{aligned} \tag{3.41}$$

Finalement, la technique de la fonction score, également appelée astuce de log-dérivation ou REINFORCE [Glynn, 1990; Williams, 1992], est appliquée. En utilisant la propriété $\nabla_{\boldsymbol{\phi}} \log p_{\boldsymbol{\phi}}(\boldsymbol{x}) = \frac{\nabla_{\boldsymbol{\phi}} p_{\boldsymbol{\phi}}(\boldsymbol{x})}{p_{\boldsymbol{\phi}}(\boldsymbol{x})}$, le gradient du second terme de l'équation (3.41) peut être réécrit sous la forme d'une espérance, qui peut alors être estimée à l'aide d'une approche de Monte-Carlo. Dans ce contexte, l'équation (3.41) devient :

$$\begin{aligned}
&\nabla_{\boldsymbol{\phi},\boldsymbol{\theta}} \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z},\boldsymbol{\lambda}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] \\
&= \int_{\boldsymbol{\epsilon}_1} \int_{\boldsymbol{\epsilon}_2} \nabla_{\boldsymbol{v}} q(\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2; \boldsymbol{v}) g(\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2; \boldsymbol{v}) d\boldsymbol{\epsilon}_2 d\boldsymbol{\epsilon}_1 + \int_{\boldsymbol{\epsilon}_1} \int_{\boldsymbol{\epsilon}_2} g(\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2; \boldsymbol{v}) q_{\boldsymbol{\epsilon}_2}(\boldsymbol{\epsilon}_2) \nabla_{\boldsymbol{v}} \log q(\boldsymbol{\epsilon}_1; \boldsymbol{v}) d\boldsymbol{\epsilon}_2 d\boldsymbol{\epsilon}_1 \\
&= \mathbb{E}_{q(\boldsymbol{\epsilon_1},\boldsymbol{\epsilon_2};\boldsymbol{v})}[\nabla_{\boldsymbol{v}} g(\boldsymbol{\epsilon_1}, \boldsymbol{\epsilon_2}; \boldsymbol{v})] + \mathbb{E}_{q(\boldsymbol{\epsilon_1},\boldsymbol{\epsilon_2};\boldsymbol{v})} \left[ g(\boldsymbol{\epsilon_1}, \boldsymbol{\epsilon_2}; \boldsymbol{v}) \nabla_{\boldsymbol{v}} \log q(\boldsymbol{\epsilon_1}; \boldsymbol{v}) \right].
\end{aligned} \tag{3.42}$$

Initialement, nous souhaitions utiliser cette technique de rétro propagation pour entraîner notre modèle. Cependant, le coût computationnel associé au calcul du jacobien présent dans le second terme de l'équation (3.42) s'est avéré extrêmement élevé. De plus, la librairie TensorFlow applique une technique de différenciation dès lors qu'une variable est échantillonnée à partir d'une distribution Gamma. Le niveau de permissivité de cette librairie ne nous a pas permis d'annuler cette reparamétrisation, impactant ainsi la mise en œuvre de notre approche. Par conséquent, nous avons choisi de poursuivre nos expérimentations en utilisant la technique développée dans TensorFlow. Cette méthode s'appuie sur les travaux proposés par les auteurs Figurnov et al. et permet de reparamétrer de manière indépendante une variable suivant une loi Normale et une variable suivant une loi Gamma [Figurnov *et al.*, 2018].

En théorie, la reparamétrisation d'une variable $\boldsymbol{z} \sim \mathcal{N}(z; \mu, \lambda^{-1})$ suivie d'une variable $\lambda \sim \mathcal{G}(\lambda; \alpha, \beta)$ diffère de la reparamétrisation d'une paire de variables $(z, \lambda) \sim \mathcal{NG}(z, \lambda; \mu, \alpha, \beta)$. Cependant, nous avons effectué plusieurs tests sur une fonction simple $f(x, y)$ avec $(x, y) \sim \mathcal{NG}(z, \lambda; \mu, \alpha, \beta)$ et avons constaté que $\nabla_\mu[\mathbb{E}[f(x, y)]] \approx \mathbb{E}[\nabla_\mu[f(x, y)]]$, où le second terme est calculé à l'aide de la reparamétrisation proposée par la bibliothèque TensorFlow. Des tests similaires ont été réalisés pour les paramètres $\alpha$ et $\beta$. Nous avons également mesuré la variance de l'espérance des gradients calculés sur mille échantillons et avons vérifié que cette variance était proche de zéro. Ces expériences, combinées à la rapidité de l'algorithme et la complexité d'annuler la reparamétrisation de TensorFlow pour uniquement utiliser la nôtre, ont justifié l'utilisation de la bibliothèque TensorFlow pour la méthode de reparamétrisation du NGVAE dans la suite de nos expériences.

### Inversion des termes dans la divergence Kullback-Leibler

Lorsque le modèle proposé est entraîné à partir de la fonction coût définie dans l'équation (3.32), une contradiction entre les résultats obtenus et les théorèmes avancés concernant la polarisation est observée. Cette contradiction est particulièrement visible en analysant la matrice de covariance illustrée dans la Figure 3.7a. Cette matrice a été calculée après

l'entraînement du NGVAE sur le jeu de données Dsprites, où seuls les facteurs génératifs, tels que la forme, la taille, la position sur l'axe x et la position sur l'axe y, ont été conservés. Elle a été calculée sur l'ensemble des variables latentes $\boldsymbol{z}$ du jeu de données d'entraînement.

Nous observons que toutes les dimensions des variables latentes sont considérées comme passives, ce qui signifie que $\pi_k(\boldsymbol{x}, \boldsymbol{\phi}) = 0$ pour chaque dimension $k$ de $\boldsymbol{z}$. Cela ne correspond pas au comportement attendu lorsque l'espace latent est polarisé. De plus, nous pourrions nous attendre à ce que les valeurs diagonales de la matrice de covariance soient faibles, ce qui n'est pas le cas. En examinant la première dimension de $\boldsymbol{z}$, nous constatons que $\mathbb{E}_{p_{data}(\boldsymbol{x})}[\boldsymbol{\sigma}_1^2(\boldsymbol{x}; \boldsymbol{\phi})] \approx 1$, comme le montre la Figure 3.7b. Cela correspond au comportement attendu pour une variable passive. Cependant, les valeurs élevées observées sur la diagonale de la matrice de covariance peuvent s'expliquer par le fait que $Var(\boldsymbol{\mu}_1(\boldsymbol{x}; \boldsymbol{\phi})) \gg 0$, comme illustré dans la Figure 3.7c. Cette observation va à l'encontre du théorème 4 et de la proposition 1 présentés par les auteurs [Bonheme et Grzes, 2021]. En effet, une variable étiquetée passive ne devrait contenir aucune information et dans ce sens, on ne devrait pas observer de variation dans sa moyenne.

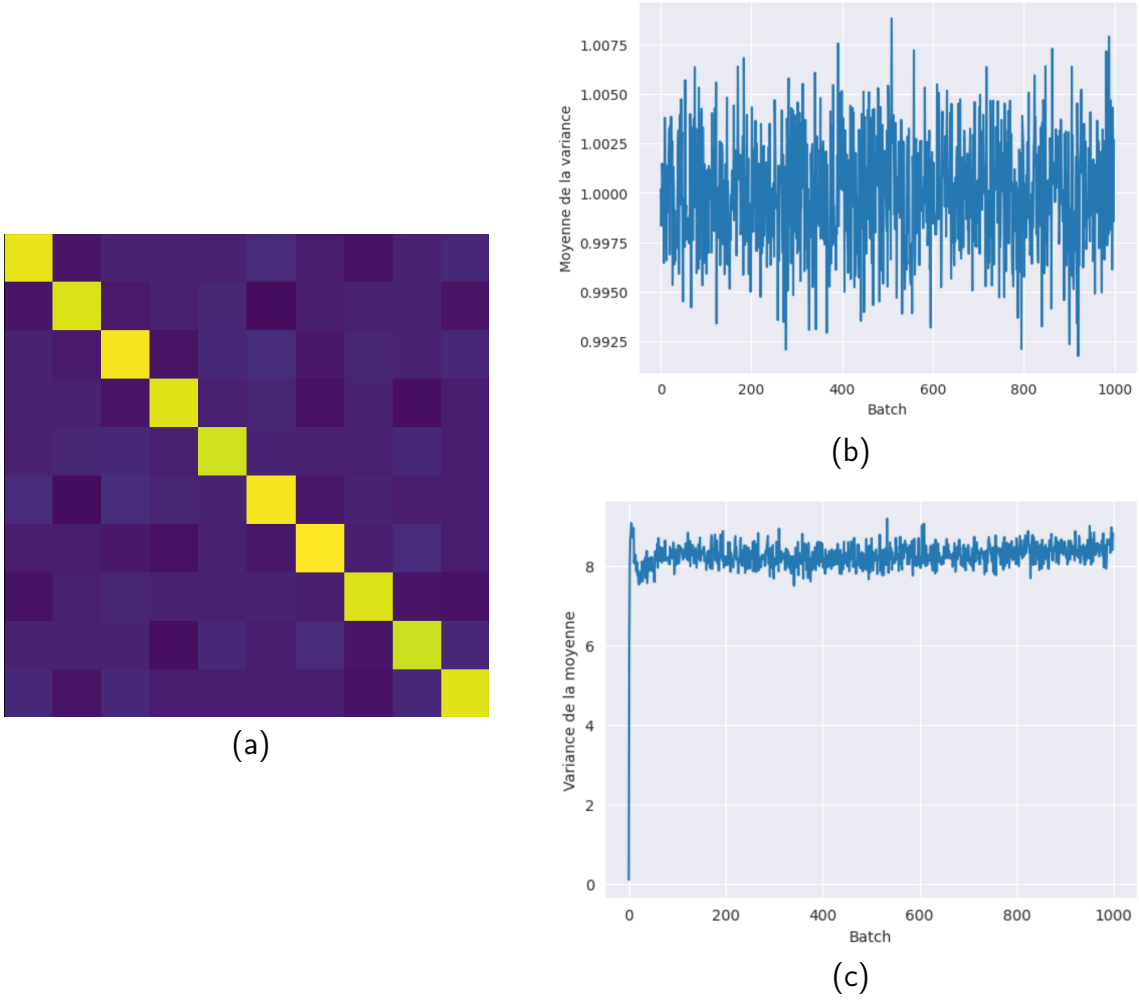



Fig 3.7: Résultats obtenus après optimisation de la fonction coût décrite dans l'équation (3.32) du modèle NGVAE entraîné sur Dsprites. (a) : $Cov(\boldsymbol{z})$ mesurée sur l'ensemble des données d'entraînement. (b) : $\mathbb{E}_{p_{data}(\boldsymbol{x})}(\boldsymbol{\sigma}_k^2(\boldsymbol{x}; \boldsymbol{\phi}))$ pour $k = 1$, dimension de l'espace latent labelisé par $\pi_k = 0$ comme passive. (c) : $Var(\boldsymbol{\mu}_k(\boldsymbol{x}; \boldsymbol{\phi}))$ pour $k = 1$, dimension de l'espace latent labellisé par $\pi_k = 0$ comme passive.

Ce comportement peut s'expliquer principalement pour deux raisons.

Tout d'abord, la définition de la KL-divergence entre deux distributions Normales-Gamma empêche les moyennes inférées pour les variables actives de prendre des valeurs élevées, contrairement aux moyennes inférées pour des variables passives. En effet, dans le contexte du choix entre deux modes à l'aide d'un seuillage pour la distribution *a posteriori* et en supposant l'indépendance des variables latentes, la forme analytique définie dans l'équation (3.35), peut être décomposée en deux termes. Dans ce contexte, on considère $\boldsymbol{z} \in \{\mathbb{Z}_a^{'}, \mathbb{Z}_p^{'}\}$ où $\mathbb{Z}_a^{'}$ correspond à l'ensemble des couples $(z_a, \lambda_a)$ considérés actifs, et $\mathbb{Z}_p^{'}$ à l'ensemble des couples $(z_a, \lambda_a)$ passifs, avec $\mathbb{Z}_a^{'} \cap \mathbb{Z}_p^{'} = \emptyset$. On obtient ainsi la décomposition suivante :

$$
\begin{aligned}
&D_{KL}[q_\phi(\boldsymbol{z}, \boldsymbol{\lambda}) || p(\boldsymbol{z}, \boldsymbol{\lambda})] \\
&= D_{KL}[q_{\boldsymbol{\phi}}(\boldsymbol{z}_a, \boldsymbol{\lambda}_a) || p(\boldsymbol{z}_a, \boldsymbol{\lambda}_a)] + D_{KL}[q_{\boldsymbol{\phi}}(\boldsymbol{z}_p, \boldsymbol{\lambda}_p) || p(\boldsymbol{z}_p, \boldsymbol{\lambda}_p)] \\
&= \sum_{j \in \mathcal{Z}_a^{'}} \left[ \frac{1}{2} \boxed{\frac{\alpha_a}{\beta_a} (\mu_{prior} - \mu_j)^2} + \alpha_{prior} \log \frac{\beta_a}{\beta_{prior}} - \log \frac{\Gamma(\alpha_a)}{\Gamma(\alpha_{prior})} \right. \\
&\left. + (\alpha_a - \alpha_{prior}) \psi(\alpha_a) - (\beta_a - \beta_{prior}) \frac{\alpha_a}{\beta_a} \right] \\
&+ \sum_{k \in \mathcal{Z}_p^{'}} \left[ \frac{1}{2} \boxed{\frac{\alpha_p}{\beta_p} (\mu_{prior} - \mu_k)^2} + \alpha_{prior} \log \frac{\beta_p}{\beta_{prior}} - \log \frac{\Gamma(\alpha_k)}{\Gamma(\alpha_{prior})} \right. \\
&\left. + (\alpha_p - \alpha_{prior}) \psi(\alpha_p) - (\beta_p - \beta_{prior}) \frac{\alpha_p}{\beta_p} \right],
\end{aligned}
\tag{3.43}
$$

où l'on note respectivement $(\alpha_a, \beta_a)$ et $(\alpha_p, \beta_p)$ les hyperparamètres des distributions Gamma pour des variables considérées comme actives et passives et $\mathcal{Z}_a^{'}$ et $\mathcal{Z}_p^{'}$ les ensembles respectifs des indices de ces variables. La distribution *a priori* est définie par $p(\boldsymbol{z}, \boldsymbol{\lambda}) = \mathcal{NG}(\boldsymbol{z}, \boldsymbol{\lambda}; \mu_{prior}, \alpha_{prior}, \beta_{prior})$.
Dans cette formulation, les termes encadrés posent un problème. En effet, on observe que les moyennes inférées par l'encodeur $\mu_j$ et $\mu_k$ sont respectivement pondérées par le ratio $\frac{\alpha_a}{\beta_a}$ et $\frac{\alpha_p}{\beta_p}$. Or, la relation

$$
\boldsymbol{\lambda} \sim G(\alpha, \beta) \Rightarrow \mathbb{E}[\boldsymbol{\lambda}^{-1}] = \frac{\alpha}{\beta} \tag{3.44}
$$

implique que pour répondre au théorème 2, $\frac{\alpha_a}{\beta_a} \gg \frac{\alpha_p}{\beta_p}$, avec $\frac{\alpha_a}{\beta_a} \gg 1$ et $\frac{\alpha_p}{\beta_p} \approx 1$ dans notre modélisation. Les moyennes inférées pour les variables actives sont ainsi contraintes à prendre de faibles valeurs, pratiquement nulles, tandis que les moyennes inférées pour les variables passives ne sont pas soumises à de telles contraintes.

Cette dynamique peut être observée dans la Figure 3.8, où les valeurs de $D_{KL}[q_{\boldsymbol{\phi}}(\boldsymbol{z}_a, \boldsymbol{\lambda}_a) || p(\boldsymbol{z}, \boldsymbol{\lambda})]$ sont représentées en vertes, tandis que les valeurs de $D_{KL}[q_{\boldsymbol{\phi}}(\boldsymbol{z}_p, \boldsymbol{\lambda}_p) || p(\boldsymbol{z}, \boldsymbol{\lambda})]$ sont en bleues, avec les moyennes inférées variant dans l'intervalle $[-20, 20]$. On constate en effet une disparité significative entre les deux courbes, ce qui a un impact négatif sur l'objectif de polarisation souhaité lors de l'apprentissage, favorisant ainsi l'émergence de variables passives.

De plus, la formulation du terme de KL-divergence pose un deuxième problème pour notre approche. Pour rappel, elle est définie par :

$$
D_{KL}[q_{\boldsymbol{\phi}}(\boldsymbol{z}, \boldsymbol{\lambda}) || p(\boldsymbol{z}, \boldsymbol{\lambda})] = \int_{\boldsymbol{z}} \int_{\boldsymbol{\lambda}} q_{\boldsymbol{\phi}}(\boldsymbol{z}, \boldsymbol{\lambda}) \log \frac{q_{\boldsymbol{\phi}}(\boldsymbol{z}, \boldsymbol{\lambda})}{p(\boldsymbol{z}, \boldsymbol{\lambda})} d\boldsymbol{z} d\boldsymbol{\lambda}. \tag{3.45}
$$

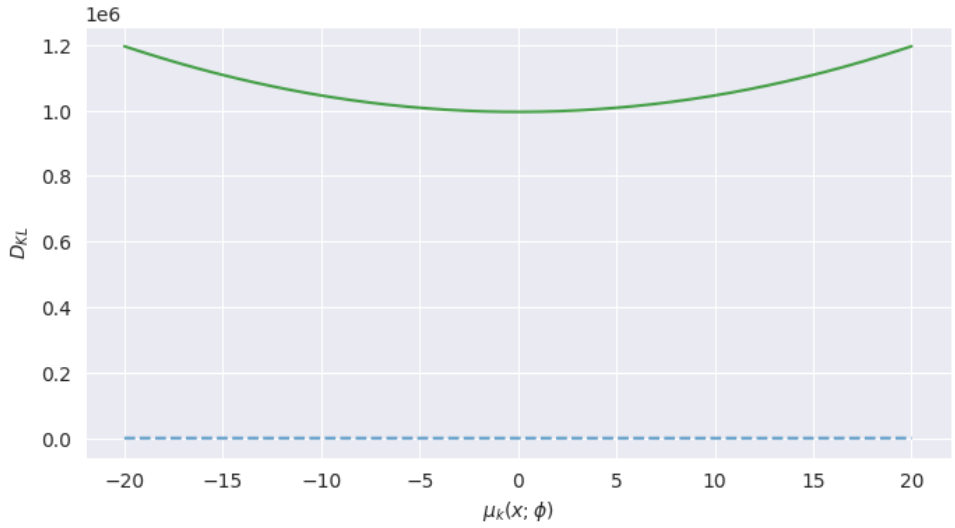



Fig 3.8: Illustration des valeurs prises par $D_{KL}[q_{\boldsymbol{\phi}}(\boldsymbol{z}_a, \boldsymbol{\lambda}_a)||p(\boldsymbol{z}, \boldsymbol{\lambda})]$ en vert et traits pleins et des valeurs prises par $D_{KL}[q_{\boldsymbol{\phi}}(\boldsymbol{z}_p, \boldsymbol{\lambda}_p)||p(\boldsymbol{z}, \boldsymbol{\lambda})]$ en bleu et pointillés, pour des valeurs de moyennes dans l'intervalle $[-20, \ldots, 20]$

Lorsque l'espérance est appliquée à la distribution *a posteriori*, la mesure de la KL-divergence explose dès lors qu'une variable est considérée comme active, comme représenté dans la Figure 3.9a. Ce problème est moins marqué si l'espérance est calculée par rapport à la loi *a priori*, car les régions où les variables actives présentent la plus forte densité de probabilité sont associées à une faible densité *a priori*. Comme l'objectif de l'apprentissage consiste à minimiser ce terme de KL, cela pousse le modèle à étiqueter l'ensemble des variables comme étant passives, ce qui a un impact négatif sur notre objectif de polarisation.

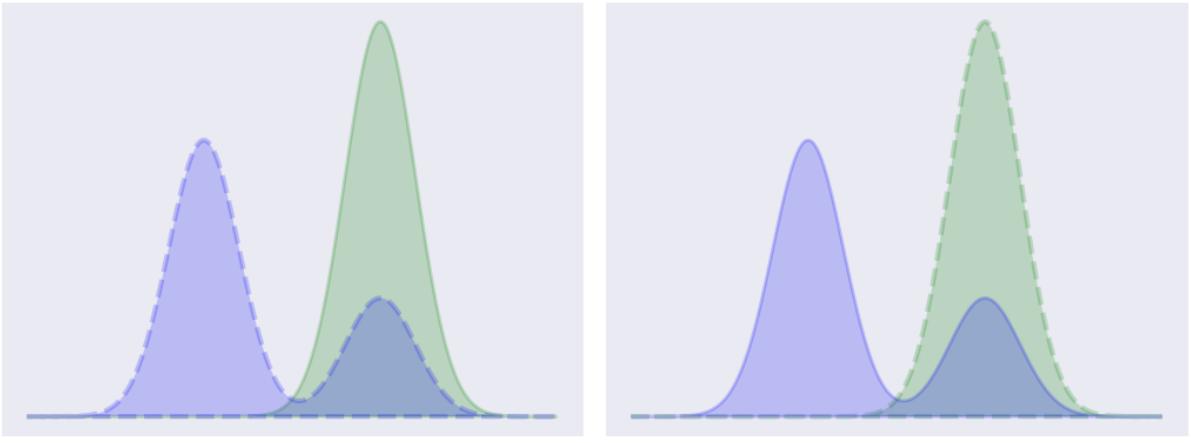



Fig 3.9: Illustration de l'impact du sens des termes de la mesure de divergence de Kullback-Leibler entre une loi bimodale en bleue, dénotée $q_{\boldsymbol{\phi}}(\boldsymbol{z}, \boldsymbol{\lambda})$, et une loi unimodale en vert, dénotée $p(\boldsymbol{z}, \boldsymbol{\lambda})$, sur le premier mode représentant la distribution des variables actives. La distribution selon laquelle est effectuée l'espérance dans la définition de la KL est illustrée par des pointillés. (a) : $D_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}, \boldsymbol{\lambda})||p(\boldsymbol{z}, \boldsymbol{\lambda}))$. (b) : $D_{KL}(p(\boldsymbol{z}, \boldsymbol{\lambda}))||q_{\boldsymbol{\phi}}(\boldsymbol{z}, \boldsymbol{\lambda}))$

Afin d'éviter les deux problèmes mentionnés ci-dessus, nous proposons d'inverser l'ordre des termes au sein de la divergence de Kullback-Leibler, pour obtenir ce qui est couram-

ment appelé la "Forward" KL-divergence, définie comme suit :

$$D_{KL}[p(\boldsymbol{z},\boldsymbol{\lambda})||q_{\boldsymbol{\phi}}(\boldsymbol{z},\boldsymbol{\lambda})] = \int_{\boldsymbol{z}}\int_{\boldsymbol{\lambda}} p(\boldsymbol{z},\boldsymbol{\lambda}) \log \frac{p(\boldsymbol{z},\boldsymbol{\lambda})}{q_{\boldsymbol{\phi}}(\boldsymbol{z},\boldsymbol{\lambda})} d\boldsymbol{z} d\boldsymbol{\lambda}. \tag{3.46}$$

Dans ce contexte, l'espérance est appliquée à la distribution *a priori*, qui est définie comme étant unimodale. Ainsi, comme illustré dans la Figure 3.10, les moyennes en sortie de l'encodeur sont pondérées par la même valeur, qu'elles soient étiquetées comme actives ou passives. Cela se traduit par deux courbes de variations similaires. De plus, l'espérance portée sur une distribution unimodale, correspondant au mode passif, comme illustré dans la Figure 3.9a, permet d'être plus tolérant envers le modèle lorsqu'il s'agit d'inférer des variables actives.

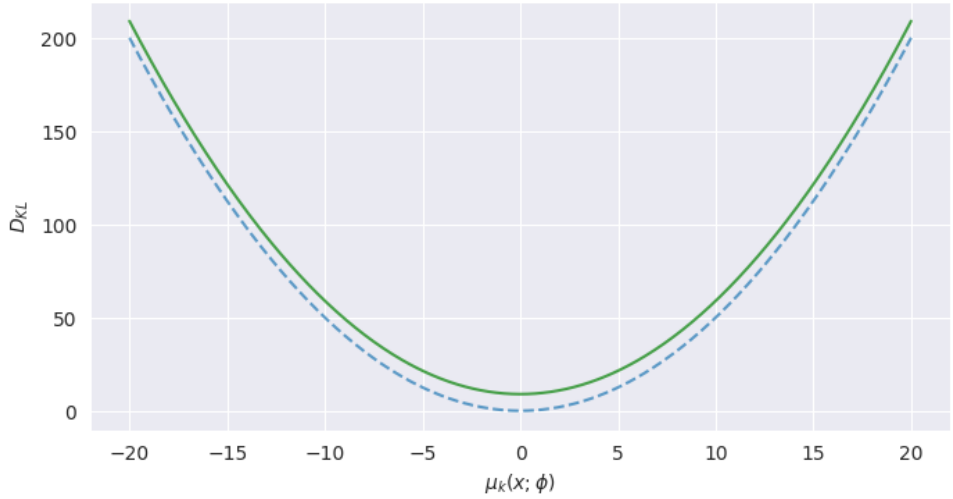



Fig 3.10: Illustration des valeurs prises par $D_{KL}[p(\boldsymbol{z},\boldsymbol{\lambda})||q_{\boldsymbol{\phi}}(\boldsymbol{z}_a,\boldsymbol{\lambda}_a)]$ en vert et traits pleins et des valeurs prises par $D_{KL}[p(\boldsymbol{z},\boldsymbol{\lambda})||q_{\boldsymbol{\phi}}(\boldsymbol{z}_p,\boldsymbol{\lambda}_p)]$ en bleu et pointillés, pour des valeurs de moyennes dans l'intervalle $[-20,\ldots,20]$

En effet, comme montré lors de l'explication du $\beta$-VAE dans la section 2.3.1, le modèle du VAE peut être considéré à travers une optimisation sous contrainte. Dans ce contexte, il est possible de considérer n'importe quelle mesure de divergence entre la distribution *a posteriori* et *a priori*.

### Régularisation des termes de probabilité

Afin de s'assurer que le réseau encodeur converge vers des variables actives ou passives, il est nécessaire que les probabilités inférées ne stagnent pas autour de 0.5. Dans cet objectif, un terme d'entropie est ajouté sur ces probabilités. Avec $\boldsymbol{\pi}(\boldsymbol{x};\boldsymbol{\phi}) \triangleq \boldsymbol{\pi}$, cette régularisation est définie par $H(\boldsymbol{\pi}) = -\sum^{K} \pi_k \log(\pi_k)$. Sa minimisation favorise l'émergence de variables $\pi$ possédant des valeurs extrêmes, proches de zéro ou proches de un.

Pour terminer, la polarisation est introduite comme un biais inductif au sein de l'apprentissage en effectuant une pondération terme à terme des variables latentes $\boldsymbol{z}$ fournies au réseau de décodeur par la variable $\boldsymbol{\pi}(\boldsymbol{x};\boldsymbol{\phi})$. Cela a pour objectif de favoriser la parcimonie dans la colonne de poids de la première couche du décodeur, et faire en sorte que le modèle concentre l'information uniquement dans les variables actives. Cette approche permet de conforter la convergence du modèle vers un espace latent où le point (c) de la définition de la polarisation proposée par Rolinek et al. [Rolinek *et al.*, 2019] est respecté.

Pour l'ensemble de ces raisons, la fonction coût à optimiser pour l'apprentissage du NGVAE est définie de la façon suivante :

$$\begin{aligned}\mathcal{L}_{NGVAE}(\boldsymbol{\phi},\boldsymbol{\theta}) =& \mathbb{E}_{p_{data}(\boldsymbol{x})}\Big[\mathbb{E}_{q_\phi(\boldsymbol{z},\boldsymbol{\lambda}|\boldsymbol{x})}[-\log p_\theta(\boldsymbol{x}|\boldsymbol{z}\cdot\boldsymbol{\pi}(\boldsymbol{x};\boldsymbol{\phi}),\boldsymbol{\lambda})] \\ &+\sum_{k=1}^{K} D_{KL}\left[p(z_k,\lambda_k)||q_\phi(z_k,\lambda_k|\boldsymbol{x})\right] + \eta H(\boldsymbol{\pi}(\boldsymbol{x},\boldsymbol{\phi}))].\end{aligned} \tag{3.47}$$

où $\eta$ dénote un hyperparamètre régularisant la valeur d'entropie.

La suite de ce chapitre consiste à vérifier la capacité du NGVAE à converger vers un bon espace latent polarisé, et de le comparer à certaines méthodes de l'état de l'art à travers différentes expérimentations.

## 3.3 Expériences

Dans le but d'évaluer la capacité du modèle proposé à atteindre un état polarisé dans son espace latent, des expériences visant à comparer de manière quantitative et qualitative les résultats obtenus après l'entraînement du NGVAE ont été menées.

Dans les sections suivantes du manuscrit, les conditions expérimentales dans lesquelles ces tests ont été réalisés sont détaillées, notamment les hyperparamètres spécifiques à chaque modèle testé et l'architecture des réseaux encodeurs/décodeurs utilisée. Ensuite, une analyse des indicateurs de performances calculés pour les différents apprentissages est effectuée avant d'aboutir à la conclusion de ce chapitre.

### 3.3.1 Conditions expérimentales

Pour évaluer la capacité du modèle proposé à partitionner l'espace latent, différents tests ont été réalisés à partir de modèles entraînés à partir du jeu de données Dsprites [Matthey *et al.*, 2017]. Dans cette base de données, nous avons conservé uniquement les images générées à partir de l'ensemble des facteurs génératifs suivants : la forme, la taille, la position sur l'axe x et la position sur l'axe y, car le facteur génératif d'orientation est mal défini résultant à des ambiguïtés pour les formes carrées et ovales.

Une comparaison entre les résultats obtenus pour le NGVAE et différents modèles de l'état de l'art, notamment un vanilla-VAE [Kingma et Welling, 2014] et un $\beta$-VAE [Higgins *et al.*, 2016] avec $\beta = 4$ est réalisée. La valeur de $\beta$ a été déterminée selon la métrique proposée dans l'article de Higgins et al., en fonction de la taille de l'espace latent fixée et de la dimension des données d'entrée [Higgins *et al.*, 2016].

Pour l'ensemble des modèles développés pour cette étude, les architectures des encodeurs et des décodeurs sont similaires et sont présentées dans le tableau 3.1. La dimension de l'espace latent a été fixée à 10.

Tous ces modèles ont été développés à l'aide de la bibliothèque TensorFlow et ont été entraînés à partir de cinq initialisations différentes des poids et des biais, en utilisant l'initialisation de Xavier uniforme. La taille du batch a été fixée à 128, et l'algorithme d'optimisation Adam a été utilisé avec un taux d'apprentissage $\alpha = 10^{-3}$. Enfin, chaque méthode a été entraînée sur mille cycles d'apprentissage.

| Encodeur |
|---|
| Entrée 64x64 images binaires |
| 4x4 conv, 32 ReLu, Stride 2 |
| 4x4 conv, 32 ReLu, Stride 2 |
| 4x4 conv, 64 ReLu, Stride 2 |
| 4x4 conv, 64 ReLu, Stride 2 |
| FC 256, FC 2x10 |

(a)

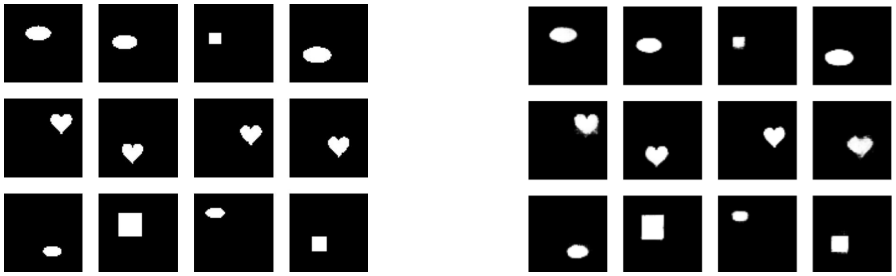
| Décodeur |
|---|
| Entrée $\in \mathbb{R}^{10}$ |
| FC 256, ReLu |
| FC 4x4x64, ReLu |
| 4x4 deconv, 64 ReLu, Stride 2 |
| 4x4 deconv, 64 ReLu, Stride 2 |
| 4x4 deconv, 32 ReLu, Stride 2 |
| 4x4 deconv, 32 ReLu, Stride 2 |
| 3x3 deconv, 1 ReLu, Stride 1 |

(b)

Tableau 3.1: Architectures des réseaux d'encodeur et de décodeur pour chacun des modèles testés sur le jeu de données Dsprites.

### 3.3.2 Résultats qualitatifs

Tout d'abord, le résultat des observations empiriques permet de confirmer que l'entraînement du NGVAE conduit, comme attendu, à un état polarisé dans l'espace latent, tout en vérifiant la validité des théorèmes sur lesquels nous nous sommes basés.

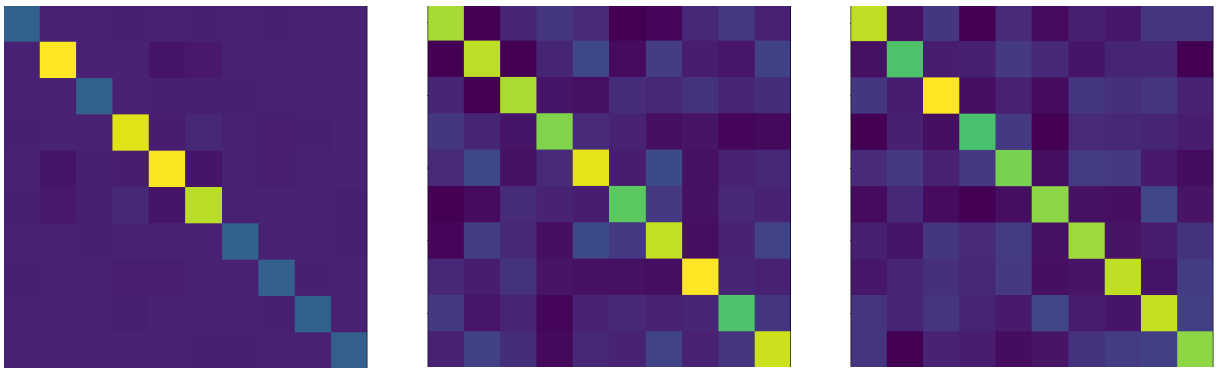


(a) (b)

Fig 3.11: Capacité de reconstruction du NGVAE. (a) : Échantillon d'images d'entrées. (b) : Échantillon d'images reconstruites.

La capacité de reconstruction du modèle, illustrée par certaines images générées en sortie du décodeur et représentées dans la Figure 4.2, permet d'obtenir des échantillons qui suivent la distribution de $p_{data}(\boldsymbol{x})$. En effet, l'ensemble des images présentées dans la Figure 3.11b correspond aux résultats produits en sortie du décodeur lorsque le modèle est alimenté avec les images de la Figure 3.11a. Cela permet de vérifier empiriquement que le NGVAE possède une capacité de reconstruction de qualité.

De plus, nous avons calculé les matrices de covariance $Cov(\boldsymbol{z})$ une fois l'apprentissage terminé pour chacune des méthodes comparée. Ces matrices sont illustrées dans la Figure 3.12. L'analyse de ces dernières nous permet à la fois de visualiser les dimensions de l'espace latent influencées par la diversité du jeu de données et de déterminer le degré de corrélation entre ces variables. Les résultats confirment notre hypothèse initiale : les modèles Vanilla-VAE et $\beta$-VAE, illustrés respectivement dans les Figures 3.12b et 3.12c, montrent peu de signes de réelle polarisation dans l'espace latent et présentent des variables $\boldsymbol{z}$ corrélées entre elles. En revanche, nous observons sur la Figure 3.12a que quatre dimensions de l'espace latent se démarquent davantage que les six autres pour le NGVAE. Ce comportement, contrairement à la matrice de covariance représentée dans la Figure

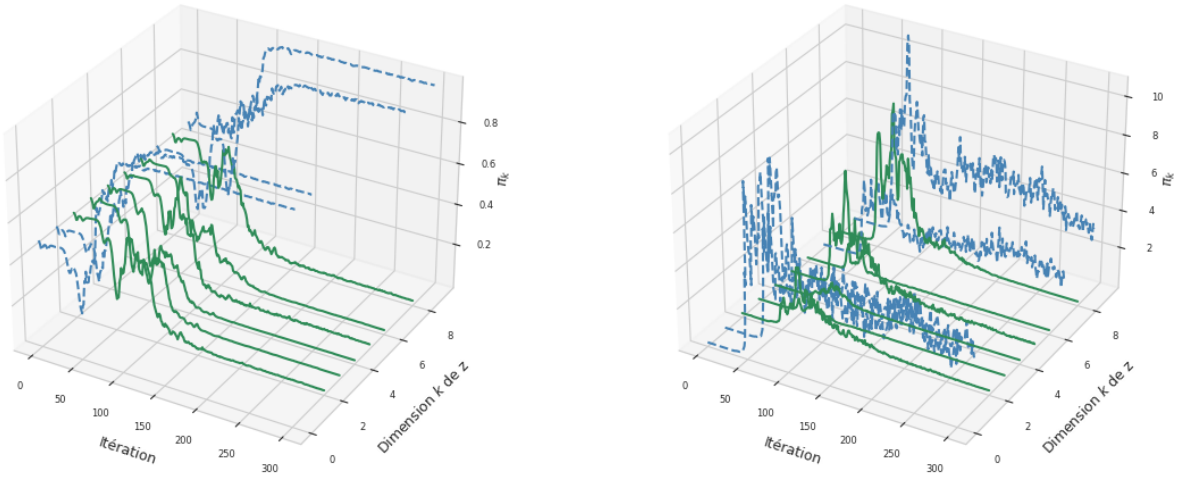


(a) (b) (c)

Fig 3.12: Matrices de covariances empiriques mesurée sur l'espace latent des modèles comparés après entraînement sur le jeu de données Dsprites. (a) : $Cov(\boldsymbol{z}, \boldsymbol{\lambda})$ NGVAE. (b) : $Cov(\boldsymbol{z})$ vanilla-VAE. (c) : $Cov(\boldsymbol{z})$ $\beta$-VAE

3.7a, suggère que le modèle a convergé vers un état polarisé dans son espace latent avec peu de corrélation entre les variables latentes.

Afin de confirmer cette polarisation, nous avons analysé le comportement des deux variables $\mu_k(\boldsymbol{x}, \boldsymbol{\phi})$ et $\pi_k(\boldsymbol{x}; \boldsymbol{\phi})$ obtenues en sortie de l'encodeur pour chacune des dimensions de l'espace latent. Dans cet objectif, les valeurs obtenues pour chacune de ces variables sont tracées dans la Figure 3.13 pour les 300 premières itérations de l'apprentissage, où une itération correspond à une application de descente de gradient. Afin de simplifier la lecture, nous notons dans la suite de l'analyse $\mu_k(\boldsymbol{x}, \boldsymbol{\phi}) \triangleq \mu_k$ et $\pi_k(\boldsymbol{x}; \boldsymbol{\phi}) \triangleq \pi_k$.




(a) (b)

Fig 3.13: Résultats obtenus durant les 300 premières itérations du NGVAE sur le jeu de données Dsprites. Les valeurs relevées pour les dimensions actives sont affichées en bleues et celles pour les dimensions passives en vertes. (a) : Valeur de $\mathbb{E}_{p_{data}(\boldsymbol{x})}[\pi_k(\boldsymbol{x}; \boldsymbol{\phi})]$ pour la $k$-ième dimension de l'espace latent relevée pour chaque itération. (b) : Valeur de $Var(\mu_k(\boldsymbol{x}; \boldsymbol{\phi}))$ pour la $k$-ième dimension de l'espace latent, calculée sur $p_{data}$ et relevée pour chaque itération.

Les graphiques présentés dans cette figure représentent, sur l'axe des abscisses, les itérations considérées, avec la dimension $K$ de l'espace latent fixée à 10. Les variables actives

sont indiquées en bleu, tandis que les variables passives sont en vert, dénotées respectivement par les indices $k_a$ et $k_p$ dans la suite de l'analyse. Il est pertinent de noter que nous avons choisi de visualiser les variations du paramètre de probabilité issu de l'encodeur plutôt que celui de la variance. En effet, notre approche implique une relation directe entre ces deux paramètres, permettant de considérer que $\mathbb{E}_{pdata(\boldsymbol{x})}[\pi_{k_a}] \approx 1 \Rightarrow \mathbb{E}_{p_{data}(\boldsymbol{x})}[\sigma^2_{k_a}] \approx 0$ et inversement $\mathbb{E}_{pdata(\boldsymbol{x})}[\pi_{k_p}] \approx 0 \Rightarrow \mathbb{E}_{p_{data}(\boldsymbol{x})}[\sigma^2_{k_p}] \approx 1$.

Une première observation significative dans cette figure concerne la rapidité avec laquelle le modèle apprend quelles variables doivent contenir de l'information et lesquelles seront considérées comme passives. En effet, dès les 200 premières itérations, un comportement distinct émerge entre les quatre variables actives et les six variables passives.

On note également dans la Figure 3.13b que $Var(\boldsymbol{\mu}_{k_a}) \gg Var(\boldsymbol{\mu}_{k_p})$. Ce comportement est en accord avec la proposition 1 des auteurs Bonheme et Grzes [Bonheme et Grzes, 2021] et, par extension, avec notre théorème 4.

Enfin, le comportement de $\boldsymbol{\sigma}^2$, en raison de la modélisation du NGVAE et des valeurs de paramètres définies pour $(\alpha_a, \beta_a)$ lorsque $\boldsymbol{\pi}_{k_a} = 1$ et $(\alpha_p, \beta_p)$ lorsque $\boldsymbol{\pi}_{k_p} = 0$, est en accord avec la proposition 2 de Bonheme et Grzes [Bonheme et Grzes, 2021] et notre théorème 2, comme le montre la Figure 3.13a. En outre, le nombre de variables actives correspond étroitement au nombre de facteurs génératifs définis pour le jeu de données sur lequel le modèle a été entraîné.

L'ensemble des résultats obtenus jusqu'à présent semble indiquer que l'entraînement d'un modèle NGVAE permet d'atteindre un état polarisé dans son espace latent. Dans cet état, le nombre de variables actives correspond au nombre de facteurs génératifs utilisés pour générer les données d'entraînement. De plus, cet entraînement conduit également à un espace latent où les variables semblent être moins corrélées les unes aux autres par rapport aux autres approches testées.

### 3.3.3 Résultats quantitatifs

Une fois la confirmation que l'espace latent du NGVAE était polarisé conformément à nos attentes, une comparaison quantitative de la capacité de polarisation avec différents modèles de l'état de l'art, entraînés à partir de cinq initialisations différentes, a été réalisée. Les résultats de cette comparaison sont résumés dans le tableau 3.2. Dans ce dernier, la moyenne obtenue et l'écart type, lorsqu'il est supérieur à $10^{-2}$, sont reportés pour chaque métrique. Le meilleur résultat est dénoté en gras.

| Modèle | card($\boldsymbol{z}_a$) | $\mathbb{E}[Var(\boldsymbol{\mu})]$ | decorr | z-min variance |
|---|---|---|---|---|
| NGVAE | $\mathbf{3,8 \pm 0.45}$ | $\mathbf{1.68 \pm 0.2}$ | $\mathbf{0.008}$ | $0.59 \pm 0.06$ |
| vanilla-VAE | $6,4 \pm 0.55$ | $0.65 \pm 0.05$ | $0.05 \pm 0.04$ | $0.62 \pm 0.11$ |
| $\beta$-VAE | $4,5 \pm 0.55$ | $0.44 \pm 0.02$ | $0.03 \pm 0.02$ | $\mathbf{0,74 \pm 0.09}$ |

Tableau 3.2: Indicateurs de polarisation mesurés sur l'ensemble de l'espace latent.

Dans cette étude comparative, nous avons relevé plusieurs mesures pour chaque modèle

entraîné, notamment le nombre de variables actives, correspondant à la cardinalité de $\boldsymbol{\pi}$ lorsque $\pi > 0.5$ pour le NGVAE. Pour les autres modèles, nous avons utilisé la mesure proposée par Rolinek et al. [Rolinek *et al.*, 2019]. Nous dénotons cette mesure $card(\boldsymbol{z}_a)$. Dans un état bien polarisé, le nombre de variables actives doit correspondre au nombre de facteurs génératifs. Nous avons également introduit deux nouvelles métriques permettant de mesurer la capacité de polarisation des méthodes. La première, $\mathbb{E}[Var(\boldsymbol{\mu})]$ où l'on note $\boldsymbol{\mu} \triangleq \boldsymbol{\mu}(\boldsymbol{x}; \boldsymbol{\phi})$, a pour objectif de mesurer la capacité de polarisation de la méthode. En effet, plus les valeurs obtenues pour cette métrique sont élevées, plus le modèle encode de l'information. La seconde métrique, dénotée "decorr" dans le tableau 3.2, est basée sur la matrice de covariance empirique des variables latentes et mesure le rapport entre la valeur absolue des termes hors diagonaux et la somme de la valeur absolue de l'ensemble des termes de la matrice. L'analyse de cette mesure permet de fournir une information quant à l'impact des termes hors diagonaux sur la matrice de covariance. Dans ce sens, plus cette mesure est faible, plus le modèle possède la capacité de converger vers un espace latent composé de variables décorrélées les unes aux autres. Enfin, nous avons inclus la mesure de désentrelacement "z-min variance" définie dans l'article de Kim et al. [Kim *et al.*, 2019a], afin d'avoir un aperçu de la capacité de désentrelacement pour chacun des modèles.

Les résultats confirment nos observations précédentes. Parmi les quatre modèles, le NGVAE obtient l'espace latent le plus décorrélé et le mieux polarisé. Il est également le seul modèle à identifier explicitement les variables actives à travers la valeur de $\boldsymbol{\pi}(\boldsymbol{x}; \boldsymbol{\phi})$, et à ne pas nécessiter le réglage d'hyperparamètres comme le fait le $\beta$-VAE. De plus, le nombre de variables actives correspond bien au nombre de facteurs génératifs. À l'inverse, les modèles Vanilla-VAE et $\beta$-VAE avec $\beta = 4$ tendent à converger vers un espace latent davantage corrélé, où le nombre de variables latentes actives est davantage sensible à l'initialisation des paramètres apprenables, et ainsi la polarisation moins bonne. Toutefois, à la différence des autres méthodes, le NGVAE ne parvient pas à réaliser un désentrelacement efficace d'après la mesure de z-diff. Cela est cohérent, car une bonne polarisation obtenue au sein de l'espace latent ne garantit pas nécessairement un désentrelacement efficace.

Afin de compléter cette analyse, nous avons comparé les résultats de ces mesures en ne considérant uniquement les variables actives. Afin de filtrer ces variables, nous avons conservé celles pour lesquelles $\pi_k(\boldsymbol{x}; \boldsymbol{\phi}) > 0.5$ concernant notre modèle, et celles pour lesquelles $\sqrt{Var(\boldsymbol{\mu}(\boldsymbol{x}; \boldsymbol{\phi}))} > 0.5$ pour les autres approches. Les résultats sont relevés dans le Tableau 3.3. De la même façon que précédemment, la moyenne obtenue et l'écart type, lorsqu'il est supérieur à $10^{-2}$, sont reportés et le meilleur résultat est dénoté en gras.

| Modèle | $\mathbb{E}[Var(\boldsymbol{\mu}_a)]$ | decorr | z-min variance |
|---|---|---|---|
| NGVAE | $\mathbf{4.53 \pm 1.34}$ | $\mathbf{0.009}$ | $0.50 \pm 0.04$ |
| vanilla-VAE | $1.01 \pm 0.04$ | $0.05 \pm 0.03$ | $0.59 \pm 0.12$ |
| $\beta$-VAE | $0.96 \pm 0.02$ | $0.04 \pm 0.04$ | $\mathbf{0,63 \pm 0.11}$ |

Tableau 3.3: Indicateurs de polarisation mesurés uniquement sur les variables actives.

Le modèle proposé présente deux avantages significatifs par rapport aux autres méthodes évaluées. Premièrement, l'analyse de $\mathbb{E}[Var(\boldsymbol{\mu}_a)]$ révèle une augmentation notable lorsqu'elle est mesurée exclusivement sur les variables actives que lorsqu'elle l'est sur

l'ensemble de l'espace latent. Cette augmentation est nettement plus prononcée que celle observée avec les autres approches testées. Cela suggère que, conformément aux attentes, les variables passives du NGVAE n'encodent aucune information, se distinguant ainsi des autres approches.

De plus, la corrélation mesurée exclusivement sur ces variables démontre que le modèle proposé converge vers une représentation dans laquelle les variables porteuses d'information sont complètement décorrélées les unes des autres, ce qui diffère des résultats obtenus avec les autres modèles. Ces observations viennent étayer les hypothèses formulées lors de l'analyse visuelle des matrices de covariance présentées dans la Figure 3.12.

## 3.4 Conclusion

Dans ce chapitre, nous exposons les diverses définitions et propositions issues de l'état de l'art en ce qui concerne la polarisation de l'espace latent, un phénomène observé lorsque l'apprentissage d'un modèle basé sur un VAE converge vers le désentrelacement de ses variables latentes. Nous proposons une formalisation du comportement des variables lorsque le modèle converge vers cet état polarisé. Cette analyse a conduit à la formulation de cinq nouveaux théorèmes.

Ensuite, nous introduisons une nouvelle architecture basée sur un VAE qui permet de déterminer automatiquement le nombre de variables latentes actives nécessaires pour représenter le nombre de facteurs génératifs utilisés lors de la génération des données d'entraînement. En nous appuyant sur une définition de la polarisation basée sur la théorie de l'information, notre modélisation consiste à rendre stochastiques les variances des distributions *a posteriori* et à leur attribuer un modèle de mélange favorisant les valeurs faibles ou élevées. Nous présentons également une technique de reparamétrisation pour l'optimisation d'une fonction de coût impliquant une variable échantillonnée selon une loi Normale-Gamma. L'inversion des termes au sein de la KL-divergence dans le contexte de la modélisation adoptée par notre approche qui favorise un mode pour la distribution *a posteriori* et également justifiée. Nous expliquons ensuite comment la polarisation est utilisée comme un biais inductif durant l'apprentissage.

Finalement, nous comparons les performances concernant l'obtention d'un régime polarisé du NGVAE à celles de différentes approches de l'état de l'art, ce qui nous permet de conclure que notre approche permet d'obtenir une meilleure polarisation. De plus, dans ces expériences, le nombre de variables latentes actives du NGVAE correspond étroitement au nombre de facteurs génératifs, sans nécessiter d'ajustement des hyperparamètres dépendants de l'ensemble de données.

Cependant, il est important de noter que le NGVAE ne parvient pas à converger vers un espace latent totalement désentrelacé. En effet, obtenir une bonne polarisation ne garantit pas automatiquement un désentrelacement, bien que cette dernière soit une condition nécessaire pour répondre aux propriétés de compacité et de modularité. En effet, les variables définies comme actives peuvent toujours présenter des corrélations entre elles. Afin de remédier à ces inconvénients, la suite de ce manuscrit vise à apporter des améliorations au modèle NGVAE dans le but de parvenir à un espace latent désentrelacé, ainsi qu'à présenter les résultats de comparaisons réalisées sur différents ensembles de données et modèles de l'état de l'art.

# Chapitre 4

# La polarisation comme vecteur d'un meilleur désentrelacement

Dans ce quatrième et dernier chapitre, les concepts et outils précédemment abordés sont mis à profit afin de proposer une nouvelle architecture reposant sur un VAE en vue de converger vers un espace latent désentrelacé et polarisé. Cette démarche consiste à étendre le modèle NGVAE présenté précédemment afin de répondre aux objectifs de désentrelacement.

Les résultats obtenus pour la mesure de "z-min variance" [Kim *et al.*, 2019a], évaluant la capacité d'obtenir un espace latent désentrelacé, sont inférieurs pour le NGVAE à ceux obtenus pour les autres modèles utilisés pour les comparaisons. Cette faiblesse peut s'expliquer par le fait que les variables actives encodent simultanément plusieurs facteurs génératifs. La stratégie adoptée dans nos travaux pour éviter cette limitation consiste à contraindre l'indépendance entre ces variables à travers l'introduction d'un terme de corrélation totale au sein de la fonction coût.

En complément, il est possible que la probabilité en sortie de l'encodeur, qui mesure le caractère actif d'une variable, entraîne une instabilité. En effet, cette probabilité ne dépend que de l'image en entrée et non pas de l'ensemble du jeu de données. Cela pourrait conduire le modèle à converger vers des comportements ambigus, dans lesquels un même facteur génératif est encodé dans deux dimensions distinctes de l'espace latent pour deux images différentes. Afin d'éviter l'émergence de ce comportement, l'incorporation d'un terme de groupe-lasso est envisagée. Il vise à contraindre le modèle à inférer les variables actives au même emplacement pour l'ensemble des images du jeu de données.

Ces réflexions nous ont conduits à développer une extension du NGVAE baptisée le Total-Correlation Normal-Gamma Variational Auto-Encoder (TC-NGVAE).

Dans la première section de ce chapitre, les diverses approches envisagées pour améliorer le NGVAE sont détaillées. Ces dernières comprennent l'intégration des termes de corrélation totale terme et de groupe-lasso au sein de la fonction de coût. La seconde partie du chapitre se concentre sur une série d'expériences visant à démontrer les capacités du TC-NGVAE à générer un espace latent désentrelacé. Dans un premier temps, l'apport de la modification de la fonction coût est analysée à travers une comparaison des résultats obtenus entre le NGVAE et la nouvelle approche. La suite vise à mettre en avant l'intérêt de la polarisation comme vecteur d'un meilleur désentrelacement à travers une comparaison des résultats obtenus pour des espaces latents à différentes dimensions. À la fin du chapitre, une étude approfondie intégrant différents jeux de données et plusieurs mesures

de désentrelacement permet de situer le TC-NGVAE par rapport à l'état de l'art et de mettre en évidence ses avantages et limitations. Une analyse des trois jeux de données considérés permet également d'expliquer certains résultats obtenus. En effet, l'existence de corrélations entre les facteurs génératifs empêche un désentrelacement total de l'espace latent.

## 4.1 Approche proposée utilisant la polarisation comme un biais inductif permettant le désentrelacement

Tel qu'explicité précédemment dans le manuscrit, lorsqu'un modèle basé sur un VAE converge vers un espace latent désentrelacé, il manifeste une polarisation au sein de cet espace. Ce comportement, illustré dans la section 3.1, a été utilisé pour modéliser un biais inductif dans la structuration du NGVAE. Ainsi, l'encodeur infère à la fois une moyenne $\mu_k(\boldsymbol{x};\boldsymbol{\phi})$ et un paramètre de probabilité $\pi_k(\boldsymbol{x};\boldsymbol{\phi})$. Ce dernier paramètre détermine l'activation ou la passivité d'une variable, contribuant à définir l'ensemble de paramètres $(\alpha_k(\pi(\boldsymbol{x};\boldsymbol{\phi})), \beta_k(\pi(\boldsymbol{x};\boldsymbol{\phi})))$ pour la distribution Gamma des inverses-variances associées aux variables latentes, où $k$ correspond à la dimension définie pour l'espace latent. L'ajustement des paramètres du modèle se déroule via l'optimisation de la fonction de coût détaillée dans l'équation (3.47).

L'analyse des résultats obtenus après entraînement sur Dsprites, effectuée dans la section 3.3, a révélé les atouts du modèle. Ils résident notamment en sa capacité à réaliser une polarisation efficace, à déterminer le nombre optimal de variables latentes pour l'encodage des facteurs génératifs, et à différencier les types de variables latentes en les classifiant comme actives ou passives en fonction de la valeur de $\pi_k(\boldsymbol{x};\boldsymbol{\phi})$. Cependant, ce modèle ne parvient pas à bien séparer les informations contenues dans les variables identifiées comme actives. Par conséquent, l'objectif principal réside dans l'amélioration de sa fonction de coût afin d'obtenir une représentation désentrelacée.

Dans cette perspective, la revue exhaustive des différentes méthodes présentées dans le chapitre 2.3 nous conduit à concevoir diverses approches. Celles-ci comprennent l'application d'une pondération sur le terme de KL-divergence pour induire l'indépendance des variables latentes en les contraignant à converger vers une distribution *a priori* non informative. Néanmoins, cette stratégie n'est pas viable dans le cadre du NGVAE car elle risquerait de compromettre la capacité de polarisation du modèle, déjà instaurée par la modélisation Normale-Gamma des variables latentes. Par conséquent, cela risquerait de restreindre le modèle à inférer uniquement des variables passives. De plus, comme démontré dans l'équation (2.5), cette pondération impacte la capacité de reconstruction du décodeur en incitant à minimiser le terme d'information mutuelle entre les variables latentes et les données d'entrée.

Une seconde approche envisagée implique l'ajout d'un terme de régularisation à la fonction de coût. Elle pourrait s'inspirer du modèle DIP-VAE en cherchant à aligner les moments d'ordre deux de la distribution jointe *a posteriori* $q_{\boldsymbol{\phi}}(\boldsymbol{z},\boldsymbol{\lambda})$ avec ceux de la distribution *a priori* $p(\boldsymbol{z},\boldsymbol{\lambda})$ [Kumar *et al.*, 2017]. Cependant, cette méthode exige la manipulation de deux termes qui se compensent mutuellement. Trouver un équilibre adéquat lors de l'apprentissage requiert le réglage précis de deux hyperparamètres pour garantir une compensation efficace entre ces termes. Cette approche présente des limitations, car une mauvaise paramétrisation pourrait entraîner le modèle dans des minima locaux lors de l'apprentissage.

En considération de ces points, l'approche envisagée réside dans l'introduction d'un terme de corrélation totale mesuré sur les variables $(z_k, \lambda_k)$ pour imposer leur indépendance et conduire à un espace latent plus désentrelacé.

### 4.1.1 Ajout d'un terme de corrélation totale

L'introduction d'un terme de corrélation totale évalué au sein de l'espace latent a été intégrée dans plusieurs modèles de l'état de l'art [Chen *et al.*, 2018; Kim et Mnih, 2018; Kim *et al.*, 2019a]. Cette démarche découle naturellement de la recherche d'une forme de désentrelacement par l'obtention d'une corrélation totale minimale. En théorie de l'information, la corrélation totale représente une forme généralisée de l'information mutuelle, mesurant la ressemblance entre la distribution jointe de variables aléatoires et le produit de leurs distributions marginales. Ainsi, une corrélation totale élevée indique une forte dépendance entre ces variables. Sa définition pour un vecteur aléatoire $\boldsymbol{z}$ de dimension $K$ est la suivante :

$$TC(\boldsymbol{z}) = D_{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z})||\prod_k q_{\boldsymbol{\phi}}(z_k)\right] = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z})}\left[\log \frac{q_{\boldsymbol{\phi}}(\boldsymbol{z})}{\prod_k q_{\boldsymbol{\phi}}(z_k)}\right]. \tag{4.1}$$

Au sein d'un modèle basé sur un VAE avec un espace latent de dimension $K$, la quantification de la corrélation totale sur cet espace $\boldsymbol{z}$ implique d'estimer les distributions $q_{\boldsymbol{\phi}}(\boldsymbol{z})$ et $\prod_k q_{\boldsymbol{\phi}}(z_k)$, qui ne sont pas directement calculables. Afin d'aborder cette problématique, diverses approches ont été proposées dans la littérature. Par exemple, l'approche naïve de Monte Carlo présente intrinsèquement un problème de sous-estimation de cette mesure. Pour pallier cette limitation, les travaux de Kim et al. suggèrent l'utilisation de la méthode du ratio de densité [Sugiyama *et al.*, 2012; Rosca *et al.*, 2018] en incorporant un réseau de discrimination dans l'apprentissage du modèle [Kim et Mnih, 2018; Kim *et al.*, 2019a]. De cette façon, le terme de corrélation totale est estimé de la manière suivante :

$$TC(\boldsymbol{z}) \approx \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z})}\left[\frac{D_{\boldsymbol{\psi}}(\boldsymbol{z})}{1 - D_{\boldsymbol{\psi}}(\boldsymbol{z})}\right], \tag{4.2}$$

où $D_{\boldsymbol{\psi}}$ correspond au réseau de discrimination de paramètres $\boldsymbol{\psi}$ entraîné à classifier des échantillons $\boldsymbol{z}$ provenant de la distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z})$ ou $\prod_k q_{\boldsymbol{\phi}}(z_k)$. La démonstration de cette astuce est développée dans l'Annexe A. Toutefois, cette approche nécessite l'apprentissage d'un ensemble de paramètres supplémentaires $\boldsymbol{\psi}$, pouvant aboutir à une complexité accrue de l'entraînement et des instabilités, à la manière du GAN.

L'ensemble de ces considérations nous a amené à préférer une des méthodes définies par les auteurs Chen et al., basées sur des approches de Monte-Carlo : le "Minibatch Stratified Sampling" (MSS) et le "Minibatch Weighted Sampling" (MWS) [Chen *et al.*, 2018]. Dans le contexte du TC-NGVAE, l'objectif de leur utilisation consiste à estimer la corrélation totale mesurée sur le vecteur joint $(\boldsymbol{z}, \boldsymbol{\lambda})$, à savoir :

$$TC(\boldsymbol{z}, \boldsymbol{\lambda}) = D_{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}, \boldsymbol{\lambda})||\prod_k q_{\boldsymbol{\phi}}(z_k, \lambda_k)\right] = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}, \boldsymbol{\lambda})}\left[\log \frac{q_{\boldsymbol{\phi}}(\boldsymbol{z}, \boldsymbol{\lambda})}{\prod_k q_{\boldsymbol{\phi}}(z_k, \lambda_k)}\right]. \tag{4.3}$$

#### Estimation par approche de Monte-Carlo

Les estimateurs stochastiques MSS et MWS présentent comme avantage principal de ne pas introduire de paramètres supplémentaires, ce qui favorise la stabilité de l'apprentissage.

Dans l'étude menée par Chen et al. [Chen *et al.*, 2018], la définition du MWS est énoncée de la manière suivante :

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z})}[\log q_{\boldsymbol{\phi}}(\boldsymbol{z})] \approx \frac{1}{M}\sum_{i=1}^{M}\left[\log \frac{1}{NM}\sum_{j=1}^{M} q_{\boldsymbol{\phi}}(\boldsymbol{z}(\boldsymbol{x}^{(i)})|\boldsymbol{x}_j)\right]. \tag{4.4}$$

Dans cette équation, $M$ représente la taille d'un sous-ensemble d'images du jeu de données, $N$ correspond au nombre total d'images, et $\boldsymbol{z}(\boldsymbol{x}^{(i)})$ est échantillonnée selon $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}^{(i)})$. Cette expression est davantage détaillée en Annexe D.2. Une méthode semblable peut être employée pour estimer $\log \prod_k q_{\boldsymbol{\phi}}(z_k)$.
Toutefois, cet estimateur est biaisé. Afin de corriger cela, les auteurs proposent un second estimateur, le MSS.

Pour ces raisons, le terme de corrélation totale ajouté à la fonction coût du TC-NGVAE, est estimé à l'aide de cette seconde approche, étendue à une distribution Normale-Gamma. Pour cela, un ensemble d'images, $B_M = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$, est utilisé pour estimer $q_{\boldsymbol{\phi}}(\boldsymbol{z}, \boldsymbol{\lambda})$ pour un couple $(\boldsymbol{z}, \boldsymbol{\lambda})$. Ce dernier est originellement échantillonné depuis $q_{\phi}(\boldsymbol{z}, \boldsymbol{\lambda}|\boldsymbol{x}^*)$ pour une image spécifique $\boldsymbol{x}^*$. $p(B_M)$ est alors définie comme une distribution Uniforme. Afin d'échantillonner depuis $p(B_M)$, $M$ indices sont tirés aléatoirement depuis $\{1, \ldots, N\}$ sans remplacement, où $N$ correspond au nombre total d'images du jeu de données. Dans ce contexte, on peut obtenir l'expression suivante :

$$\begin{aligned} q_{\boldsymbol{\phi}}(\boldsymbol{z}, \boldsymbol{\lambda}) &= \mathbb{E}_{p(B_M)}\left[\frac{1}{M}\sum_{m=1}^{M} q_{\boldsymbol{\phi}}(\boldsymbol{z}, \boldsymbol{\lambda}|\boldsymbol{x}_m)\right] \\ &= \frac{M}{N}\mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} q_{\boldsymbol{\phi}}(\boldsymbol{z}, \boldsymbol{\lambda}|\boldsymbol{x}_m)\Big|\boldsymbol{x}^{(*)} \in B_M\right] + \frac{N-M}{N}\mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} q_{\boldsymbol{\phi}}(\boldsymbol{z}, \boldsymbol{\lambda}|\boldsymbol{x}_m)\Big|\boldsymbol{x}^{(*)} \notin B_M\right]. \end{aligned} \tag{4.5}$$

Cette grandeur peut être estimée en échantillonnant un ensemble de $M+1$ images, dans lequel $\boldsymbol{x}^*$ est un élément, et en considérant $\hat{B}_M = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ un ensemble d'éléments différents de $\boldsymbol{x}^*$. Dans ce contexte, la première espérance de l'équation (4.5) peut être estimée en utilisant $\{\boldsymbol{x}^*\} \cup \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{M-1}\}$ et la seconde en utilisant $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$. L'estimateur, noté $f((\boldsymbol{z}, \boldsymbol{\lambda}), \boldsymbol{x}^*, \hat{B}_M)$, s'écrit alors :

$$f((\boldsymbol{z}, \boldsymbol{\lambda}), \boldsymbol{x}^*, \hat{B}_M) = \frac{1}{N} q_{\boldsymbol{\phi}}(\boldsymbol{z}, \boldsymbol{\lambda}|\boldsymbol{x}^*) + \frac{1}{M}\sum_{m=1}^{M-1} q_{\boldsymbol{\phi}}(\boldsymbol{z}, \boldsymbol{\lambda}|\boldsymbol{x}_m) + \frac{N-M}{NM} q_{\boldsymbol{\phi}}(\boldsymbol{z}, \boldsymbol{\lambda}|\boldsymbol{x}_M). \tag{4.6}$$

La démonstration, davantage développée de cet estimateur, est disponible en Annexe D.1. Une méthodologie similaire peut être employée pour estimer le terme $\log \prod_k q_{\boldsymbol{\phi}}(\boldsymbol{z}_k, \boldsymbol{\lambda}_k)$, également requis pour le calcul de $TC(\boldsymbol{z}, \boldsymbol{\lambda})$.

Dans un premier temps, il a été considéré de minimiser ce terme uniquement sur les variables étiquetées comme actives par les probabilités en sortie de l'encodeur, à la manière du BF-VAE-2. Cependant, une telle approche n'est pas forcément appropriée étant donné que, dans le contexte de polarisation induit par la modélisation, $TC(\boldsymbol{z}, \boldsymbol{\lambda}) = TC(\boldsymbol{z}_a, \boldsymbol{\lambda}_a)$. En effet, lorsque $(\boldsymbol{z}, \boldsymbol{\lambda}) = ((\boldsymbol{z}_a, \boldsymbol{\lambda}_a), (\boldsymbol{z}_p, \boldsymbol{\lambda}_p))$, où les ensembles $(\boldsymbol{z}_a, \boldsymbol{\lambda}_a)$ et $(\boldsymbol{z}_p, \boldsymbol{\lambda}_p)$ représentent respectivement les variables actives et passives, la mesure de corrélation totale sur les variables passives est nulle. Ce résultat est énoncé dans le théorème suivant.

**Théorème 6** (Valeur de la corrélation totale mesurée sur un ensemble de variables latentes $(\boldsymbol{z}, \boldsymbol{\lambda})$ dans le contexte d'un VAE polarisé.)**.** *Dans le cadre d'un VAE bayésien*

*hiérarchique polarisé où les variables latentes* $(\boldsymbol{z}, \boldsymbol{\lambda}) = ((\boldsymbol{z}_a, \boldsymbol{\lambda}_a), (\boldsymbol{z}_p, \boldsymbol{\lambda}_p))$*, si les variables passives* $(\boldsymbol{z}_p, \boldsymbol{\lambda}_p)$ *ne portent pas d'information, c'est-à-dire que* $q_{\boldsymbol{\phi}}(\boldsymbol{z}_p, \boldsymbol{\lambda}_p|\boldsymbol{x}) = q_{\boldsymbol{\phi}}(\boldsymbol{z}_p, \boldsymbol{\lambda}_p)$*, alors la mesure de la corrélation totale sur l'ensemble des variables latentes équivaut à la mesure de la corrélation totale exclusivement sur les variables actives :*

$$TC(\boldsymbol{z}, \boldsymbol{\lambda}) = TC(\boldsymbol{z}_a, \boldsymbol{\lambda}_a). \tag{4.7}$$

*Preuve.* Dans le cadre d'un modèle basé sur un VAE bayésien hiérarchique caractérisé par une configuration polarisée dans son espace latent, avec $(\boldsymbol{z}, \boldsymbol{\lambda}) = ((\boldsymbol{z}_a, \boldsymbol{\lambda}_a), (\boldsymbol{z}_p, \boldsymbol{\lambda}_p))$, où les ensembles $(\boldsymbol{z}_a, \boldsymbol{\lambda}_a)$ et $(\boldsymbol{z}_p, \boldsymbol{\lambda}_p)$ représentent respectivement les variables actives et passives, et en supposant l'indépendance entre ces variables latentes, on peut formuler :

$$\begin{aligned} TC(\boldsymbol{z}, \boldsymbol{\lambda}) &= D_{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}, \boldsymbol{\lambda}) || \prod_{k=1}^{K} q_{\boldsymbol{\phi}}(z_k, \lambda_k)\right] \\ &= D_{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}, \boldsymbol{\lambda}) || \prod_{j \in \mathcal{Z}_a} q_{\boldsymbol{\phi}}(z_j, \lambda_k) \prod_{k \in \mathcal{Z}_p} q_{\boldsymbol{\phi}}(z_k, \lambda_k)\right]. \end{aligned} \tag{4.8}$$

De plus, en prenant en compte l'indépendance des variables passives, et la définition de la KL-divergence, on peut écrire :

$$\begin{aligned} TC(\boldsymbol{z}, \boldsymbol{\lambda}) &= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}, \boldsymbol{\lambda})}\left[\log \frac{q_{\boldsymbol{\phi}}(\boldsymbol{z}_a, \boldsymbol{\lambda}_a) \prod_{k \in \mathcal{Z}_p} q_{\boldsymbol{\phi}}(z_k, \lambda_k)}{\prod_{j \in \mathcal{Z}_a} q_{\boldsymbol{\phi}}(z_j, \lambda_j) \prod_{k \in \mathcal{Z}_p} q_{\boldsymbol{\phi}}(z_k, \lambda_k)}\right] \\ &= D_{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}_a, \boldsymbol{\lambda}_a) || \prod_{j \in \mathcal{Z}_a} q_{\boldsymbol{\phi}}(z_j, \lambda_j)\right] \\ &= TC(\boldsymbol{z}_a, \boldsymbol{\lambda}_a). \end{aligned} \tag{4.9}$$

□

L'ensemble des considérations précédentes permet l'obtention d'une première fonction coût concernant la nouvelle approche TC-NGVAE, définie de la façon suivante :

$$\begin{aligned} \mathcal{L}_{TC-NGVAE}(\boldsymbol{\phi}, \boldsymbol{\theta}) = \mathbb{E}_{p_{data}(\boldsymbol{x})}\Big[&\mathbb{E}_{q_\phi(\boldsymbol{z}, \boldsymbol{\lambda}|\boldsymbol{x})}[-\log p_\theta(\boldsymbol{x}|\boldsymbol{z} \cdot \boldsymbol{\pi}(\boldsymbol{x}; \boldsymbol{\phi}), \boldsymbol{\lambda})] \\ &+ \sum_{k=1}^{K} D_{KL}\left[p(z_k, \lambda_k) || q_\phi(z_k, \lambda_k|\boldsymbol{x})\right] + \beta TC(\boldsymbol{z}, \boldsymbol{\lambda}) + \eta H(\boldsymbol{\pi}(\boldsymbol{x}; \boldsymbol{\phi}))\Big]. \end{aligned} \tag{4.10}$$

La section suivante vise à justifier le besoin d'un terme de régularisation supplémentaire sur les probabilités obtenues en sortie de l'encodeur.

### 4.1.2 Ajout d'un terme de groupe-lasso

L'estimation de la corrélation totale, présentée dans l'équation (4.6), conduit à définir la fonction de coût pour optimiser les paramètres $(\boldsymbol{\phi}, \boldsymbol{\theta})$ du TC-NGVAE, telle qu'indiquée dans l'équation (4.10). Cependant, cette approche présente toujours une limitation. En effet, le vecteur de probabilités $\boldsymbol{\pi}(\boldsymbol{x}; \boldsymbol{\phi})$ est exclusivement conditionné par l'image d'entrée fournie à l'encodeur. Théoriquement, cela pourrait le conduire à attribuer le statut de variable active à une dimension particulière de l'espace latent pour une image $\boldsymbol{x}^{(i)}$ fournie en entrée, tandis que cette même dimension pourrait être considérée comme passive pour

une image $\boldsymbol{x}^{(j)}$, avec $i \neq j$. Or, ce comportement n'est pas souhaitable dans le cadre de l'obtention d'un espace latent interprétable.

L'intégration d'un terme de régularisation sur le vecteur de probabilités $\boldsymbol{\pi}(\boldsymbol{x};\boldsymbol{\phi})$ peut être considéré afin de pallier cette limitation. Il viserait à favoriser que les composantes nulles du vecteur de probabilité soient aux mêmes indices quelle que soit l'image d'entrée, tout en conservant uniquement les composantes d'intérêt. Afin de répondre à ce second point, le terme de régularisation lasso, qui correspond à la somme des valeurs absolues des composantes [Tibshirani, 1996], peut être considéré. Il s'écrit dans notre contexte :

$$L(\boldsymbol{\pi}(\boldsymbol{x};\boldsymbol{\phi})) = \sum_{j=1}^{K} |\pi(\boldsymbol{x};\boldsymbol{\phi})_j|, \tag{4.11}$$

où $\pi(\boldsymbol{x};\boldsymbol{\phi})_j$ dénote la probabilité inférée par l'encodeur pour une dimension $j$ de l'espace latent. Toutefois, cette régularisation n'implique pas de dépendance entre les images d'entrées et les probabilités et donc ne répond pas à notre objectif initial. Cependant, elle peut être étendue en un second terme de régularisation, dénoté groupe-lasso, dans lequel la pénalité n'est pas appliquée de façon individuelle à chaque composante, mais à un ensemble de variables [Yuan et Lin, 2006]. Pour cela, les variables sont regroupées en fonction de leur relation et une pénalité est appliquée à la norme euclidienne de chaque groupe. Dans le contexte du TC-NGVAE, le terme de groupe-lasso s'écrit :

$$GL(\boldsymbol{\pi}(\boldsymbol{x};\boldsymbol{\phi})) = \frac{1}{K}\sum_{j=1}^{K}\left(\sqrt{\frac{1}{M}\sum_{i=1}^{M}\boldsymbol{\pi}(\boldsymbol{x};\boldsymbol{\phi})_{ij}^2}\right). \tag{4.12}$$

Dans cette équation, $K$ désigne la dimension de l'espace latent et $M$ correspond au nombre de données utilisées pour former le sous-ensemble d'images servant à l'optimisation de la fonction de coût. En regroupant les probabilités par dimension de l'espace latent pour l'ensemble des images, cette régularisation favorise la parcimonie au niveau du groupe et donc tend à conserver uniquement les composantes qui sont pertinentes pour l'ensemble du jeu de données. Il est intéressant de souligner que, contrairement au terme lasso défini dans l'équation (4.11), le terme groupe-lasso est différentiable en tout point, permettant ainsi l'optimisation par descente de gradient.

L'intégration du groupe-lasso dans la fonction définie par l'équation (4.10), pour entraîner le TC-NGVAE, résulte à la forme suivante :

$$\begin{aligned}\mathcal{L}_{TC-NGVAE}(\boldsymbol{\phi},\boldsymbol{\theta}) =& \mathbb{E}_{p_{data}(\boldsymbol{x})}\Big[\mathbb{E}_{q_\phi(\boldsymbol{z},\boldsymbol{\lambda}|\boldsymbol{x})}[-\log p_\theta(\boldsymbol{x}|\boldsymbol{z}\cdot\boldsymbol{\pi}(\boldsymbol{x};\boldsymbol{\phi}),\boldsymbol{\lambda})] \\ &+\sum_{k=1}^{K} KL\left[p(z_k,\lambda_k)||q_\phi(z_k,\lambda_k|\boldsymbol{x})\right] + \beta TC(\boldsymbol{z},\boldsymbol{\lambda}) \\ &+\eta_1 H(\boldsymbol{\pi}(\boldsymbol{x},\boldsymbol{\phi})) + \eta_2 GL(\boldsymbol{\pi}(\boldsymbol{x};\boldsymbol{\phi}))\Big].\end{aligned} \tag{4.13}$$

Dans cette expression, $\beta$, $\eta_1$ et $\eta_2$ représentent les hyperparamètres.

L'ensemble des termes correspondant au TC-NGVAE étant définis, la section suivante de ce chapitre est dédiée aux expériences effectuées pour évaluer la capacité du TC-NGVAE à obtenir un espace latent désentrelacé et polarisé, et le confronter à l'état de l'art.

## 4.2 Expériences

Afin d'évaluer la capacité du TC-NGVAE à converger vers un espace latent désentrelacé, des expériences approfondies sont menées pour comparer les performances de notre modèle avec le NGVAE ainsi qu'avec d'autres méthodes de l'état de l'art.

Cette analyse s'articule autour de trois phases distinctes. Tout d'abord, une comparaison est établie avec les performances du NGVAE dans des conditions expérimentales identiques, mettant en lumière l'impact de l'ajout du terme de corrélation totale à la fonction de coût pour le désentrelacement. Certains résultats obtenus sont étayés par une analyse du jeu de données Dsprites mettant en avant des limitations dans les définitions classiquement utilisées pour caractériser les facteurs génératifs.

Par la suite, la deuxième expérience réalisée a pour objectif de comparer les résultats obtenus par le TC-NGVAE en faisant varier la dimension de $\boldsymbol{z}$. Ces tests permettent de mettre en évidence l'apport de la polarisation dans la capacité de désentrelacement.

Finalement, des tests sont conduits dans le but de comparer la capacité du TC-NGVAE à converger vers un espace latent désentrelacé par rapport à d'autres méthodes de l'état de l'art. Pour cela, les modèles considérés, à savoir le $\beta$-VAE [Higgins *et al.*, 2016], le Factor-VAE [Kim et Mnih, 2018], le DIP-VAE-II [Kumar *et al.*, 2017], le $\beta$-TC-VAE [Chen *et al.*, 2018] et le BF-VAE-2 [Kim *et al.*, 2019a], sont d'abord entraînés sur les jeux de données Dsprites, Smallnorb et Cars3d. Ensuite, les métriques préalablement identifiées dans la section 2.4 manuscrit sont utilisées pour évaluer les propriétés d'un espace latent désentrelacé, à savoir sa compacité, sa modularité et sa qualité explicite.

Pour chacune de ces expériences, les conditions expérimentales sont détaillées. Elles incluent l'architecture des réseaux de neurones pour l'encodeur et le décodeur ainsi que les hyperparamètres associés à chaque modèle. Les jeux de données sur lesquels les modèles ont été entraînés sont spécifiés, ainsi que les indicateurs de performance utilisés pour la comparaison.

### 4.2.1 Impact du terme de corrélation totale dans la capacité de désentrelacement

Pour évaluer l'impact du terme de corrélation totale sur la capacité du modèle à obtenir un espace latent désentrelacé, les performances des modèles NGVAE et TC-NGVAE sont comparées.

#### Conditions expérimentales

Dans le cadre de ces comparaisons, les conditions expérimentales analogues à celles appliquées pour l'évaluation du modèle NGVAE, exposées dans la section 3.3.1, sont reproduites. Elles impliquent l'entraînement du nouveau modèle TC-NGVAE sur le jeu de données Dsprites, caractérisé par quatre facteurs génératifs : la forme, la taille, la position sur l'axe x et la position sur l'axe y. Le facteur d'orientation est exclu, car il est défini de manière ambiguë.

Les paramètres du TC-NGVAE, tels que $\beta$, $\eta_1$ et $\eta_2$, ont été déterminés via une recherche systématique sur une grille, conduisant aux valeurs $\beta = 6$, $\eta_1 = 1$ et $\eta_2 = 0.5$. Les architectures des réseaux encodeur et décodeur correspondent à celles spécifiées dans

le Tableau 3.1. Les paramètres liés à la taille du batch, au taux d'apprentissage et à l'initialisation des poids et biais ont été maintenus identiques pour tous les modèles élaborés, conformément à leur définition dans la section 3.3.1.

**Résultats obtenus**

Les métriques associées à la polarisation sont évaluées, ainsi que la mesure du désentrelacement via la métrique "z-min variance" [Kim *et al.*, 2019a]. Leurs résultats sont reportés dans le tableau 4.1.

| **Modèle** | **card**($\boldsymbol{z}_a$) | $\mathbb{E}[Var(\boldsymbol{\mu})]$ | **decorr** | **z-min** |
|---|---|---|---|---|
| NGVAE | $\mathbf{3,8 \pm 0.45}$ | $1.68 \pm 0.2$ | $0.008 \pm 0.004$ | $0.59 \pm 0.06$ |
| TC-NGVAE | 3 | $\mathbf{2,99 \pm 0.5}$ | $\mathbf{0,005 \pm 0.005}$ | $\mathbf{0.73} \pm 0.01$ |

Tableau 4.1: Indicateurs de polarisation pour mesurer l'apport du terme de corrélation totale dans la fonction coût.

Une première analyse du tableau montre, conformément aux attentes, que la modification de la fonction coût a amélioré la capacité du TC-NGVAE à obtenir un espace latent mieux désentrelacé par rapport au NGVAE. La métrique "z-min variance" est ainsi significativement plus élevée. Ce résultat s'explique par le fait que la corrélation totale favorise l'indépendance statistique des variables latentes, déjà polarisées grâce au modèle de mélange défini sur la distribution *a posteriori*. Cette propriété contribue à l'encodage de facteurs génératifs indépendants dans des variables latentes distinctes, améliorant ainsi la capacité de désentrelacement.

Une meilleure décorrélation des variables latentes, mesurée à partir de la matrice de covariance, ainsi qu'une plus grande quantité d'informations encodées dans la moyenne avec une variance plus élevée par rapport au NGVAE, sont également observées. Ces améliorations peuvent également être attribuées au terme de corrélation totale.

Toutefois, chaque apprentissage du TC-NGVAE converge vers trois variables actives au sein de l'espace latent, contrairement au NGVAE, pour lequel davantage de variables sont actives. Un tel comportement peut s'expliquer par la présence de dépendances statistiques complexes entre certains facteurs génératifs, tels que la forme et l'échelle dans le jeu de données Dsprites. Cela incite l'apprentissage à concentrer l'information dans trois variables latentes. Une analyse démontrant ces résultats est effectuée à la suite de cette section.

Une observation est également réalisée via les matrices de corrélation de Spearman entre les variables latentes et les facteurs génératifs, représentées dans la Figure 4.1. La corrélation de Spearman mesure une relation de monotonie entre deux variables à partir du rang des données. Elle offre ainsi un avantage sur les méthodes de corrélation traditionnelles, car elle permet de détecter des relations non linéaires entre les variables. De ce fait, elle est également adaptée au TC-NGVAE, dont la modélisation implique des lois non gaussiennes.

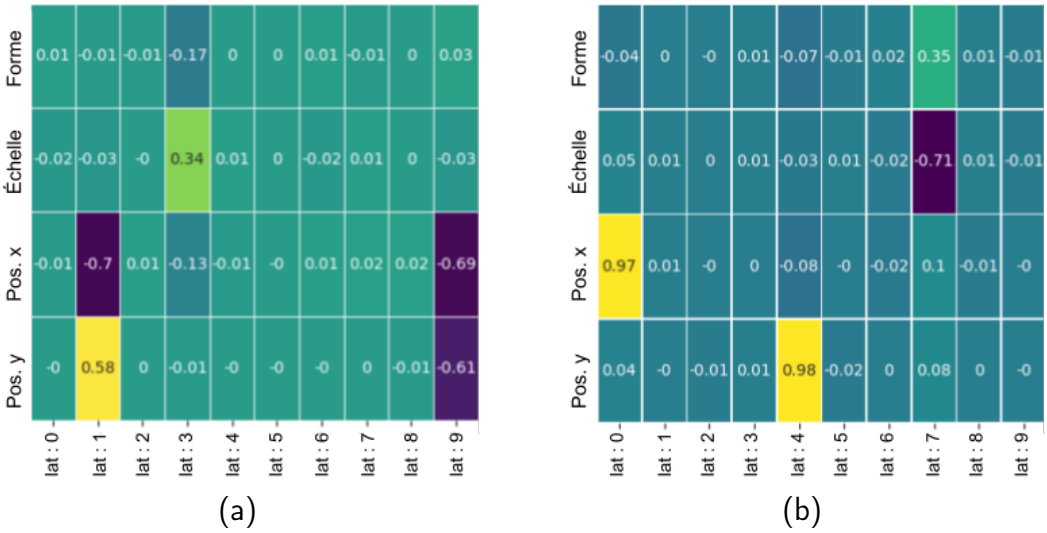



Fig 4.1: Matrices de corrélation de Spearman entre les facteurs génératifs et les variables latentes. Chaque ligne correspond à un facteur génératif du jeu de données Dsprites sans le facteur d'orientation et chaque colonne à une variable de l'espace latent. (a) : NGVAE. (b) : TC-NGVAE.

L'examen des résultats présentés dans la Figure 4.1 permet de mettre en évidence deux observations majeures :

1. Les facteurs correspondant à position de l'objet sur l'axe $x$ et à la position sur l'axe $y$ sont conjointement encodés dans les variables latentes 1 et 9 du NGVAE, tels qu'illustrés dans la Figure 4.1a. En revanche, dans la matrice du TC-NGVAE illustrée dans la Figure 4.1b, ces deux facteurs sont uniquement représentés dans les variables latentes 0 et 4 respectivement, sans corrélation entre elles. Pour ce qui est de l'encodage des facteurs génératifs de forme et d'échelle, le TC-NGVAE les encode conjointement dans la variable latente 7, une représentation favorisée par l'optimisation du terme de corrélation totale lorsque ces deux facteurs génératifs présentent une dépendance statistique.

2. Le degré de dépendance entre les facteurs génératifs et les variables latentes est plus élevée pour le TC-NGVAE que pour le NGVAE, particulièrement pour les deux facteurs génératifs de position. Cette augmentation peut être expliquée par la capacité accrue du TC-NGVAE à capturer plus d'informations du jeu de données, comme le suggère le résultat de la mesure $\mathbb{E}[Var(\boldsymbol{\mu})]$, répertoriée dans le tableau 4.1.

Pour une comparaison plus approfondie, les matrices de Spearman mesurées sur l'espace latent des modèles Vanilla-VAE et $\beta$-VAE entraînés sur le même jeu de données sont présentées en Annexe E.1.

La Figure 4.2a illustre la matrice de covariance de $\boldsymbol{z}$ du TC-NGVAE et met en évidence la polarisation vers laquelle l'espace latent a convergé. Les variables latentes encodant les facteurs génératifs se distinguent nettement le long de la diagonale, avec des valeurs extrêmement faibles hors diagonale, indiquant une forte décorrélation entre elles.

Un parcours de l'espace latent a également été effectué sur les moyennes encodées $\boldsymbol{\mu}(\boldsymbol{z}; \boldsymbol{\phi})$, comme présenté dans la Figure 4.2b. Cette approche qualitative implique la transformation du vecteur latent obtenu pour une image donnée. Elle permet d'observer l'impact de

celle-ci en analysant l'image obtenue en sortie du décodeur. La transformation appliquée consiste à considérer chaque variable latente indépendamment et à la modifier linéairement en ajoutant une valeur généralement comprise dans l'intervalle $[-3\sigma, +3\sigma]$ autour de sa moyenne, tout en maintenant les autres variables à leur valeur initiale. Dans la Figure 4.2b, chaque colonne représente une dimension de l'espace latent, tandis que chaque ligne correspond à une valeur différente ajoutée à la moyenne. Pour plus de clarté, les images décodées à partir des dimensions représentant les variables actives ont été regroupées à gauche et encadrées en rouge. Ces observations révèlent que la variable latente 7, en première colonne, encode effectivement la forme et l'échelle, tandis que les dimensions 4 et 0, correspondant respectivement à la deuxième et troisième colonne, capturent les variations de l'objet sur les axes $y$ et $x$.

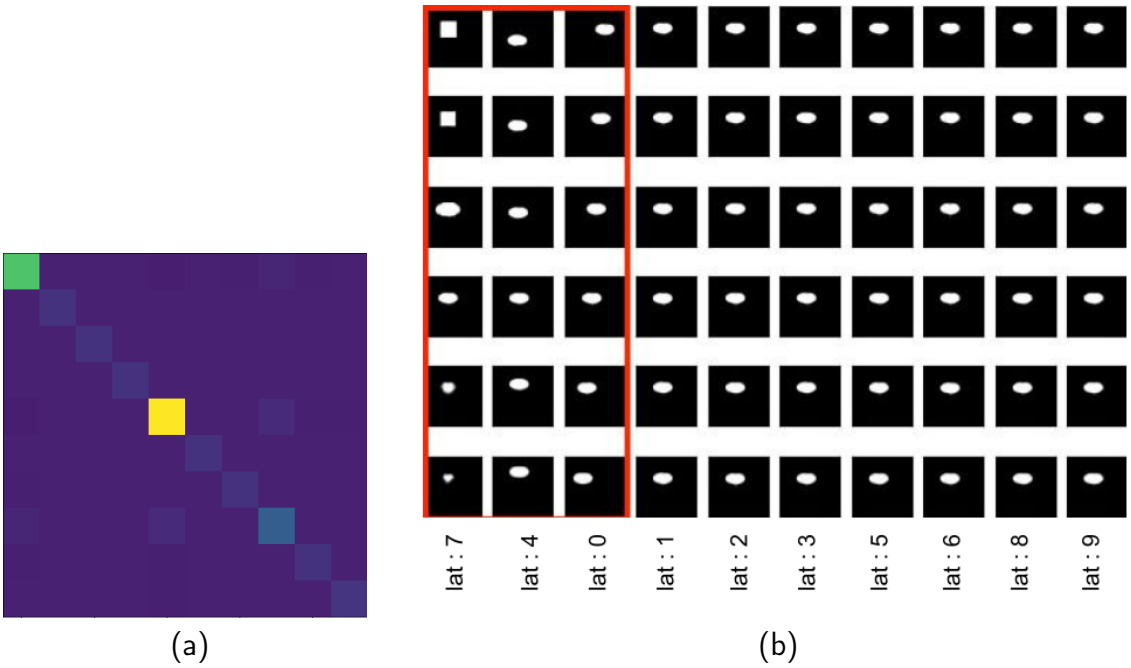



Fig 4.2: (a) : Matrice de covariance empirique de $\boldsymbol{z}$ du TC-NGVAE. (b) : Parcours de l'espace latent du TC-NGVAE après apprentissage sur Dsprites, en excluant le facteur génératif d'orientation.

Afin d'évaluer la corrélation entre l'encodage de la position de l'objet et le facteur génératif correspondant, une expérimentation supplémentaire a été menée. Elle implique une comparaison entre la localisation du barycentre de l'objet dans une image initiale et celui mesuré dans la même image après reconstruction du modèle. La reconstruction se fait après une transformation linéaire appliquée sélectivement à une dimension de l'espace latent, tandis que les autres sont constantes. L'effet de cette intervention sur $\boldsymbol{z}$ est analysé à travers les axes $x$ et $y$, respectivement dans les Figures 4.3a et 4.3b. Dans ces dernières, chaque courbe colorée représente le déplacement du barycentre en fonction de la dimension de la variable latente modifiée.

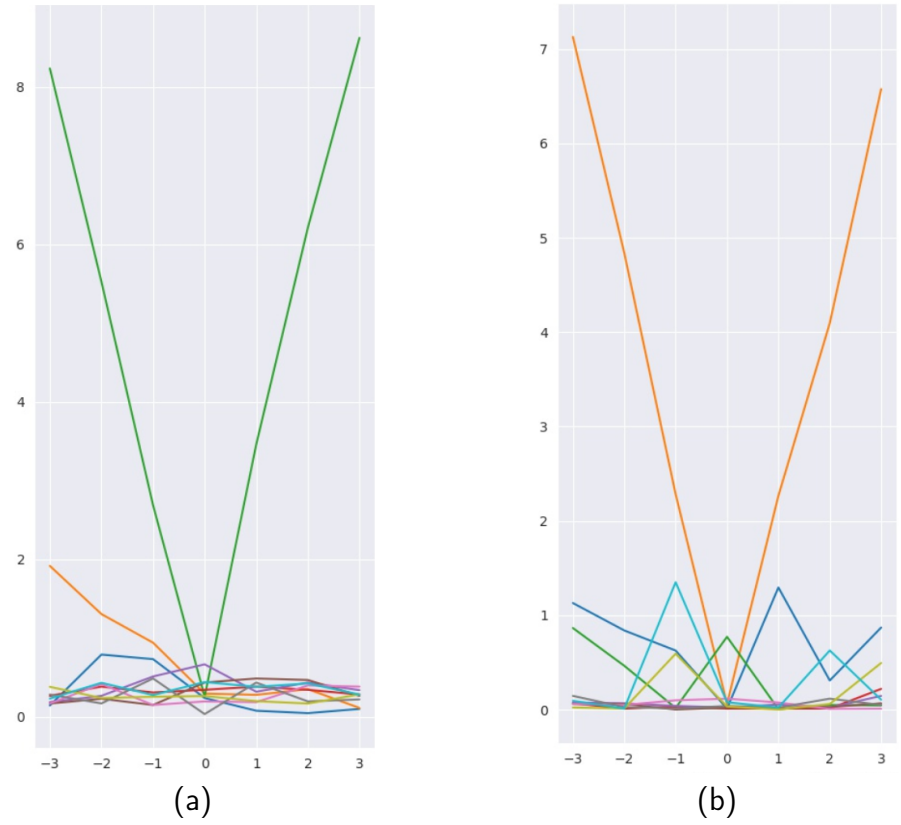



Fig 4.3: Différence de barycentres entre l'objet d'une image initiale et celui de la même image reconstruite lors d'un parcours de la moyenne dans l'espace latent. (a) : Différence mesurée sur l'axe $x$. (b) : Différence mesurée sur l'axe $y$.

L'analyse de cette Figure traduit deux résultats principaux :

- Dans les Figures 4.3a et 4.3b, la modification d'une unique dimension de $\boldsymbol{z}$ impacte le barycentre de l'objet dans l'image reconstruite. Cela implique que le facteur génératif de position est bien encodé dans une unique variable, respectivement pour l'axe $x$ et l'axe $y$.
- Il existe une relation linéaire entre le facteur génératif de la position de l'objet et l'encodage obtenu pour sa représentation sur les axes verticaux et horizontaux.

Cette découverte est importante dans notre contexte, car elle suggère que la représentation de l'objet dans l'image obtenue par le TC-NGVAE soit facilement manipulable.

La section suivante a pour objectif d'effectuer une analyse des facteurs génératifs utilisés pour la création du jeu de données Dsprites. Elle vise à expliquer l'ambiguïté observée dans le facteur génératif d'orientation ainsi qu'à mettre en exergue une dépendance entre les facteurs de forme et d'échelle, conduisant à leur encodage conjoint dans une même variable latente.

### Analyse du jeu de données Dsprites

La création du jeu de données Dsprites [Matthey *et al.*, 2017] présente des ambiguïtés dans la définition du facteur génératif d'orientation, ce qui a amené à ne pas le considérer dans l'analyse précédente. Ce facteur est caractérisé par quarante valeurs réparties de manière linéaire dans l'intervalle $[0, 2\pi]$. Or, des valeurs différentes, tous facteurs identiques par

ailleurs, peuvent conduire à la même image, particulièrement pour les formes ovales et carrées. Ce résultat est illustré dans la Figure 4.4.


Fig 4.4: Images provenant du jeu de données Dsprites comportant différents facteurs génératifs d'orientation. (a) : ovale. (b) : carré.

La Figure 4.4a représente deux images du jeu de données qui ont été sélectionnées chacune avec des facteurs génératifs similaires pour la forme carrée, les positions sur $x$ et $y$, ainsi que l'échelle. Dans l'image de gauche de cette même Figure, le facteur génératif d'orientation est fixé à 0, tandis que dans l'image de droite, il est établi à $2\pi$. De manière similaire, la Figure 4.4b illustre deux images ayant des formes ovales avec des facteurs génératifs identiques, à l'exception du facteur d'orientation fixé à 0 pour la première image et à $\pi$ pour la seconde. Il est notable que dans chacune de ces Figures, les images sont similaires, indiquant qu'il est impossible de déduire leur facteur génératif d'orientation. Ainsi, pour certaines images, le modèle n'est pas en mesure d'encoder correctement l'orientation telle qu'elle a été définie lors de la création du jeu de données, justifiant sa non prise en compte lors de l'étude précédente. Par ailleurs, l'ajout d'un terme de corrélation totale à la fonction de coût favorise l'encodage des facteurs génératifs, *a priori* indépendants, dans des variables latentes distinctes. Cependant, la Figure 4.5 démontre que les facteurs génératifs de forme et d'échelle présentent une certaine dépendance entre eux.

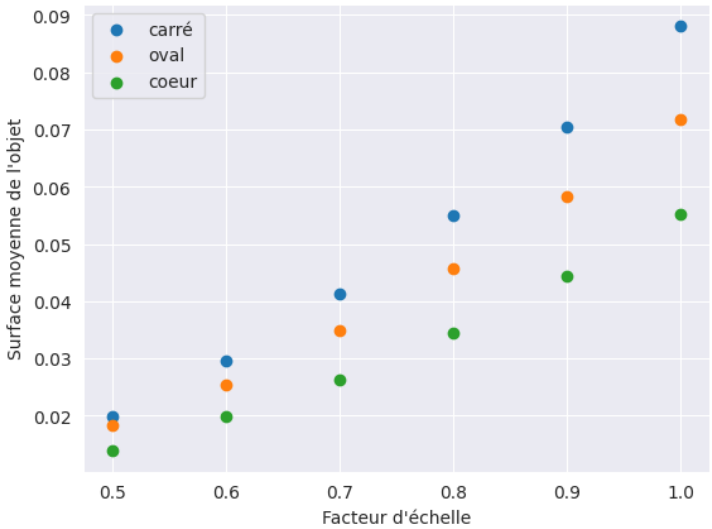



Fig 4.5: Forme d'un objet de Dsprites, identifié par rapport à sa surface dans l'image, pour différentes valeurs d'échelle.

Dans ce graphique sont représentées les surfaces moyennes des objets en nombre de pixels en fonction de la valeur du facteur d'échelle. L'ensemble des autres facteurs génératifs sont fixés et similaires. On observe la possibilité d'effectuer une séparation linéaire entre les différentes formes. Ce résultat suggère une certaine dépendance entre les facteurs génératifs de forme et d'échelle, expliquant ainsi la tendance des représentations latentes

obtenues par le TC-NGVAE, pour lequel l'espace latent est contraint par le terme de corrélation totale, à encoder les facteurs génératifs de forme et d'échelle dans une même variable.

**Synthèse**

L'analyse effectuée met en avant l'impact significatif de l'ajout d'un terme de corrélation totale au sein de la fonction coût en termes de polarisation et de désentrelacement. De plus, la comparaison des matrices de corrélation de Spearman entre les facteurs génératifs et les variables latentes permet de mettre en évidence la capacité du modèle à converger vers une représentation dont les propriétés de compacité, de modularité et de qualité explicite sont largement supérieures à celles du NGVAE. Une exploration des résultats du TC-NGVAE est menée via l'analyse de la matrice de covariance empirique mesurée sur $\boldsymbol{z}$ et à travers l'interprétation d'un parcours de l'espace latent.

Parallèlement, une relation linéaire entre les facteurs génératifs de position sur les axes $x$ et $y$ et leurs encodages respectifs est mise en avant. En effet, les variations du barycentre des objets, entre une image initiale et l'image similaire reconstruite après une transformation spécifique dans l'espace latent, mettent en avant cette linéarité. Finalement, l'exclusion du facteur génératif d'orientation lors de l'apprentissage des modèles et l'observation d'un regroupement des facteurs génératifs de forme et d'échelle dans une même variable latente sont justifiés par une analyse du jeu de données Dsprites.

Ces observations peuvent être résumées à travers les points suivants :

- L'introduction du terme de corrélation totale favorise l'encodage d'informations et l'indépendance statistique entre les variables latentes.
- Les facteurs génératifs de position sur les axes $x$ et $y$ sont linéairement corrélés avec leurs encodages respectifs.
- La mesure de corrélation totale entraîne un encodage conjoint des facteurs génératifs de forme et d'échelle dans une même variable latente. Ce résultat s'explique par l'analyse de Dsprites, qui met en avant une dépendance entre ces facteurs génératifs.

La suite de ce chapitre se concentre à vérifier l'apport de la polarisation dans la mesure de désentrelacement.

### 4.2.2 Apport de la polarisation dans la mesure de désentrelacement

Afin de mesurer l'apport du biais inductif de la polarisation dans l'objectif de désentrelacement, la seconde série d'expériences réalisée vise à comparer les résultats obtenus en termes de désentrelacement au sein de l'espace latent, pour un TC-NGVAE configuré avec 10 variables latentes et le même modèle, mais avec 20 variables latentes. Ainsi est évaluée la sensibilité de la capacité de désentrelacement à la dimension de l'espace latent. Elle s'avère délicate à ajuster, et de nombreuses méthodes de l'état de l'art définissent leurs hyperparamètres en fonction du nombre de variables latentes. Par exemple, l'approche visant à déterminer la valeur optimale de $\beta$ selon les auteurs du $\beta$-VAE dépend de la taille de l'espace latent. De manière similaire, les hyperparamètres du DIP-VAE-II ou du BF-VAE-2 sont intrinsèquement liés à la dimension de $\boldsymbol{z}$.

L'hypothèse testée est que notre modèle, conçu pour restreindre le nombre de variables actives en fonction des données, reste insensible à la dimension prédéterminée de l'espace latent. Pour vérifier ceci, nous analysons l'erreur quadratique moyenne des différentes mesures de désentrelacement pour des dimensions fixées à 10 et à 20, et ce sur chaque jeu de données. En effet, lorsque le nombre de variables actives est similaire pour les deux modèles, une faible valeur obtenue pour cette erreur quadratique indique que les capacités de désentrelacement des modèles testés sont similaires. Hormis la dimension de l'espace latent, les conditions expérimentales définies pour l'apprentissage des modèles restent constantes. Elles incluent l'architecture de l'encodeur et du décodeur, le nombre de cycles d'apprentissage, les hyperparamètres, la méthode de descente de gradient, et le coefficient d'apprentissage. Ces détails sont explicités plus avant dans le manuscrit.

| **Jeux de données** | **MIG** | **DCI Désent.** | **DCI Compacité** | **DCI Qualité explicite** | **z-min** |
|---|---|---|---|---|---|
| Dsprites | 0.04 | 0.03 | 0.02 | 0.01 | 0.02 |
| Smallnorb | 0.06 | 0.02 | 0 | 0 | 0.08 |
| Cars3d | 0.01 | 0.05 | 0.09 | 0.03 | 0.08 |

Tableau 4.2: Différence quadratique moyenne des mesures de désentrelacement obtenues par le TC-NGVAE entre deux espaces latents à 10 et à 20 dimensions.

Le tableau 4.2 présente l'erreur quadratique pour chaque résultat obtenu entre deux apprentissages effectués sur 10 et sur 20 variables latentes. Elle est calculée en moyennant les résultats obtenus pour cinq initialisations différentes des paramètres de l'encodeur et du décodeur.

Concernant les jeux de données Dsprites et Smallnorb, les modèles ont chacun convergé vers le même nombre de variables actives. Dans ce contexte, les valeurs relevées dans le tableau 4.2 indiquent que les deux modèles possèdent la même capacité de désentrelacement. Les légères différences constatées s'expliquent par l'ajout de paramètres de biais et de poids au sein de la couche finale de l'encodeur, pour le modèle dont la dimension de $\boldsymbol{z}$ est fixée à 20. Ces résultats confirment l'hypothèse initiale selon laquelle la dimension de l'espace latent n'affecte pas la capacité de désentrelacement du TC-NGVAE, pour les jeux de données Dsprites et Smallnorb, sous des conditions expérimentales équivalentes.

Les apprentissages réalisés sur le jeu de données Cars3d présentent à l'inverse des résultats moins satisfaisants. En effet, le nombre moyen de variables actives inférées diffère entre les deux modèles ; il est respectivement de 4 et de 6 pour une dimension de $\boldsymbol{z}$ fixée à 10 et à 20. Ce résultat ne permet pas d'analyser convenablement l'erreur quadratique calculée sur les différentes mesures, ces dernières étant sensibles au nombre de variables actives de l'espace latent. Toutefois, une partie de la suite du manuscrit est dédiée à présenter les résultats obtenus pour ce jeu de données entre le TC-NGVAE et d'autres approches de l'état de l'art, ainsi qu'à analyser les différentes images qui le constituent. Cette étude permet d'expliquer les résultats moins probants obtenus sur ce jeu de données, et pourquoi le modèle affiche une certaine difficulté à converger vers le bon nombre de variables actives.

La section suivante vise à proposer une comparaison des résultats obtenus par notre modèle par rapport à ceux de l'état de l'art sur leur capacité à converger vers un espace

latent désentrelacé.

### 4.2.3 Comparaison du TC-NGVAE avec les méthodes de l'état de l'art

La suite de ce chapitre se concentre spécifiquement sur l'évaluation de la capacité du TC-NGVAE à converger vers un espace latent désentrelacé. Une comparaison approfondie des performances obtenues par notre modèle par rapport à d'autres approches de l'état de l'art est réalisée. Les métriques MIG [Chen *et al.*, 2018], "Désentrelacement", "Compacité" et "Qualité explicite" de la mesure DCI [Eastwood et Williams, 2018], ainsi que la mesure "z-min variance" [Kim et Mnih, 2018], sont utilisées comme des indicateurs de désentrelacement, comme explicité dans la section 2.4.

La comparaison est effectuée sur les modèles de référence $\beta$-VAE [Higgins *et al.*, 2016], Factor-VAE [Kim et Mnih, 2018], DIP-VAE-II [Kumar *et al.*, 2017], $\beta$-TC-VAE [Chen *et al.*, 2018] et BF-VAE-2 [Kim *et al.*, 2019a], réputés pour leurs performances en désentrelacement. Chacune des méthodes est présentée dans la section 2.3 et entraînée sur les jeux de données Dsprites, Cars3d et Smallnorb. Toutefois, par souci de clarté, la comparaison du TC-NGVAE sera restreinte à deux modèles, différents pour chaque jeu de données. L'ensemble des résultats obtenu pour tous les modèles est disponible en Annexe E.

En premier lieu, les conditions expérimentales adoptées pour cette analyse sont précisées. Une attention particulière est portée à l'explication de la méthodologie adoptée afin de déterminer les hyperparamètres optimaux des méthodes de l'état de l'art. Ensuite, les performances obtenues sur les différents jeux de données sont comparées. Finalement, une analyse des jeux de données Cars3d et Smallnorb est réalisée dans l'objectif d'expliquer certains résultats obtenus.

#### Conditions expérimentales

La bibliothèque "Disentanglement-Lib" [Locatello *et al.*, 2019b] met à disposition des ensembles de paramètres pré-entraînés pour les encodeurs et décodeurs de différents modèles. Ces derniers sont ajustés à partir de cinquante initialisations distinctes sur des jeux de données de l'état de l'art, incluant ceux étudiés dans cette analyse. Les configurations d'hyperparamètres pour chaque méthode sont récapitulées dans le tableau 4.3.

Une rapide analyse du tableau 4.3 montre que la librairie met à disposition des résultats obtenus pour six configurations différentes de chaque jeu d'hyperparamètres. De ce fait, l'étape initiale pour les expériences à venir a consisté à déterminer les valeurs d'hyperparamètres pour chaque modèle conduisant à la meilleure convergence moyenne sur l'ensemble des métriques étudiées. Ce choix a été effectué en prenant en compte les cinquante initialisations, et est illustré pour la méthode du DIP-VAE-II dans le tableau 4.4.

Les résultats complets pour les autres modèles de l'état de l'art sont disponibles en Annexe E.2.1.

Afin de déterminer les valeurs optimales d'hyperparamètres pour le modèle BF-VAE-2, dont les résultats ne sont pas disponibles en ligne, ainsi que pour notre proposition, une recherche sur grille a été effectuée. À nouveau, les valeurs d'hyperparamètres retenues

| **Modèles** | **Hyperparamètre** | **Valeurs** |
|---|---|---|
| $\beta$-VAE | $\beta$ | $[1, 2, 4, 6, 8, 16]$ |
| Factor-VAE | $\gamma$ | $[10, 20, 30, 40, 50, 100]$ |
| DIP-VAE-II | $\lambda_{od}$ | $[1, 2, 5, 10, 20, 50]$ |
| | $\lambda_d$ | $\lambda_{od}$ |
| $\beta$-TC-VAE | $\beta$ | $[1, 2, 4, 6, 8, 10]$ |

Tableau 4.3: Ensemble d'hyperparamètres à disposition pour les modèles proposés par la librairie "Disentanglement-Lib".

sont celles ayant engendré les meilleures performances sur les métriques de désentrelacement. Un récapitulatif des configurations adoptées pour chaque modèle et chaque jeu de données est disponible dans le tableau 4.5. Le temps à disposition ne nous a pas permis de réaliser le même nombre de réalisations que celui effectué pour les modèles proposés par "Disentanglement-Lib", car cela aurait nécessité plusieurs semaines, voire mois d'entraînement.

| **Valeurs de $\lambda_{od}$** | **Valeur moyenne** | | |
|---|---|---|---|
| | Dsprites | Cars3d | Smallnorb |
| 1 | 0.249 | 0.409 | 0.348 |
| 2 | 0.249 | 0.405 | 0.354 |
| 5 | 0.255 | 0.395 | 0.359 |
| **10** | 0.263 | **0.438** | 0.364 |
| **20** | 0.269 | 0.431 | **0.372** |
| **50** | **0.281** | 0.400 | 0.360 |

Tableau 4.4: Moyennes des mesures de désentrelacement obtenues pour le DIP-VAE-II calculées sur cinquante initialisations, selon différentes valeurs de $\lambda_{od}$ avec $\lambda_d = \lambda_{od}$.

Dans le but de minimiser les biais dans l'analyse des résultats obtenus, les architectures de l'encodeur et du décodeur sont similaires pour tous les modèles comme détaillé dans le tableau 3.1 du manuscrit. Les paramètres de poids et de biais sont initialisés selon la méthode Xavier Uniforme [Glorot et Bengio, 2010]. Les batchs sont composés de 64 images, et l'algorithme d'optimisation utilisé est Adam, paramétré avec un coefficient d'apprentissage de $1e^{-4}$. La dimension de $\boldsymbol{z}$ a été fixée à 10 et la fonction de vraisemblance $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ modélisée comme une loi de Bernoulli.

Contrairement aux méthodes de l'état de l'art dont l'apprentissage est interrompu après 300000 étapes d'optimisation, celui du TC-NGVAE a été arrêté lorsque la fonction coût s'est stabilisée, c'est-à-dire lorsque aucune fluctuation significative n'est observée au cours des dix dernières epochs. Cette approche a été choisie, car le TC-NGVAE présente un niveau de stochasticité plus élevé du fait de la probabilisation de la variance, ce qui im-

plique plus de temps pour converger.

| **Modèle** | **Valeurs des hyperparamètres** | | |
|---|---|---|---|
| | Dsprites | Cars3d | Smallnorb |
| $\beta$-VAE ($\beta$) | 16 | 16 | 16 |
| Factor-VAE ($\gamma$) | 100 | 30 | 100 |
| DIP-VAE-II ($\lambda_d = \lambda_{od}$) | 50 | 20 | 10 |
| $\beta$-TC-VAE ($\beta$) | 10 | 10 | 1 |
| BF-VAE-2 ($\eta_1, \eta_2$) | (0.1, 0.1) | (0.1, 0.1) | (0.1, 0.5) |
| TC-NGVAE ($\beta, \eta$) | (6, 1) | (3, 1) | (3, 1) |

Tableau 4.5: Ensemble des hyperparamètres considéré pour chaque modèle selon le jeu de données.

Finalement, les valeurs des hyperparamètres utilisées pour les mesures de désentrelacement reposant sur des algorithmes supervisés ont été conservées identiques à celles de la "Disentanglement-Lib".

**Résultats obtenus**

Cette partie du manuscrit vise à analyser les performances du TC-NGVAE et des autres approches de l'état de l'art, sur le jeu de données Dsprites premièrement, puis sur Smallnorb, avant de terminer sur Cars3d. Pour chaque jeu de données, les mesures de désentrelacement sont relevées dans un tableau dans lequel les valeurs les plus élevées sont mises en évidence en gras.

**Dsprites :** Les indicateurs de désentrelacement relevés dans le tableau 4.6 démontrent que le TC-NGVAE surpasse les méthodes les plus récentes en ce qui concerne les mesures "DCI - Désentrelacement" et "DCI - Compacité", qui indiquent une représentation plus compacte et plus modulaire vis-à-vis des autres approches, dans lequel un facteur génératif est encodé dans moins de variables latentes et réciproquement. C'est également le cas pour la mesure MIG, ce qui indique que l'information mutuelle entre un facteur génératif et une variable latente est en moyenne supérieure aux autres approches. Les valeurs de "z-min variance" suggèrent que notre approche n'est pas meilleure pour cette mesure, mais reste comparable. Finalement, la réduction de la variance des résultats suggère que le TC-NGVAE est moins sensible à l'initialisation des paramètres, un point qui représente une limitation classique des approches génératives.

Ce tableau révèle également un phénomène singulier dans la mesure de la "z-min variance" spécifique au modèle BF-VAE-2, identifiée dans le tableau avec un astérisque en rouge. Les résultats obtenus présentent en effet une différence importance avec ceux des autres approches. Cette disparité peut s'expliquer à travers trois points :

- la modélisation du BF-VAE-2,
- la définition de la métrique "z-min variance",

- et le choix effectué visant à minimiser les biais lors de la comparaison des résultats dans cette analyse.

| Modèle | MIG | DCI Désent. | DCI Compacité | DCI Qualité explicite | z-min |
|---|---|---|---|---|---|
| $\beta$-VAE | $0.27 \pm 0.05$ | $0.28 \pm 0.07$ | $0.43 \pm 0.075$ | $0.37 \pm 0.02$ | $0.73 \pm 0.08$ |
| Factor-VAE | $0.27 \pm 0.04$ | $0.36 \pm 0.06$ | $0.41 \pm 0.06$ | $0.47 \pm 0.04$ | $\mathbf{0.81 \pm 0.02}$ |
| DIP-VAE-II | $0.09 \pm 0.04$ | $0.15 \pm 0.04$ | $0.25 \pm 0.03$ | $0.37 \pm 0.04$ | $0.65 \pm 0.05$ |
| $\beta$-TC-VAE | $0.20 \pm 0.05$ | $0.36 \pm 0.04$ | $0.40 \pm 0.03$ | $0.44 \pm 0.02$ | $0.78 \pm 0.05$ |
| BF-VAE-2 | $0.32 \pm 0.14$ | $0.45 \pm 0.14$ | $0.52 \pm 0.14$ | $\mathbf{0.68 \pm 0.02}$ | $0.26^{*} \pm 0.04$ |
| TC-NGVAE | $\mathbf{0.45 \pm 0.03}$ | $\mathbf{0.49 \pm 0.01}$ | **0.77** | 0.67 | $0.72 \pm 0.01$ |

Tableau 4.6: Indicateurs de désentrelacement pour des méthodes entraînées sur Dsprites.

La métrique en question repose sur l'hypothèse selon laquelle, lorsque l'on fixe un facteur génératif dans un ensemble d'images, sa représentation dans l'espace latent varie moins que les autres facteurs. Or la modélisation du BF-VAE-2 tend à converger vers un espace dans lequel les variances inférées par l'encodeur pour les variables étiquetées "non pertinentes" tendent à dépasser largement 1. Ce phénomène conduit à des situations dans lesquelles la variance des variables latentes actives est nettement inférieure à celle des variables passives, compromettant ainsi l'hypothèse sous-tendant la mesure "z-min variance" tel que l'on peut l'observer dans le tableau. Pour remédier à cette problématique, une approche consisterait à ne pas prendre en considération les variables non pertinentes dans le calcul de la mesure "z-min variance". Cependant, cette option introduirait un biais significatif dans l'analyse. En effet, la métrique est sensible au nombre de variables en entrée, et les modèles n'atteignent pas tous le même nombre de variables actives. Par conséquent, nous avons délibérément choisi de maintenir toutes les variables de l'espace latent, dont la dimension est fixée à 10 dans cette analyse.

Par ailleurs, les matrices de corrélation de Spearman entre les facteurs génératifs et les variables latentes pour le $\beta$-VAE, le $\beta$-TCVAE, ainsi que pour la méthode proposée permettent d'enrichir cette analyse. Ces dernières sont illustrées dans la Figure 4.6.
L'observation de ces matrices permet d'expliquer certains comportements relevés dans le tableau 4.6. Si l'on considère la Figure 4.6a, dédiée aux résultats du $\beta$-VAE, le facteur génératif associé à la position sur l'axe $x$ présente des corrélations avec deux variables latentes spécifiques, la 2 et la 7. Cette corrélation impacte la propriété de compacité attendue pour le désentrelacement et explique un score relativement faible dans la mesure de "DCI - Compacité" relevée dans le Tableau 4.6. De plus, le niveau des corrélations entre les facteurs génératifs et les variables latentes semble relativement faible, indiquant une perte d'information. Cette observation semble corroborer les résultats obtenus pour la mesure "DCI - qualité explicite".

La Figure 4.6b, qui illustre les corrélations pour le modèle $\beta$-TCVAE, montre une distinction notable avec les observations précédentes. En effet, les facteurs génératifs de position sont chacun représenté dans une unique variable latente. De façon similaire, les

niveaux de corrélations illustrés dans cette matrice sont supérieurs à ceux observés pour le $\beta$-VAE. Ces résultats permettent d'expliquer les différences relevées dans le tableau 4.6 concernant les mesures "DCI - Désentrelacement" et "DCI - Qualité explicite".

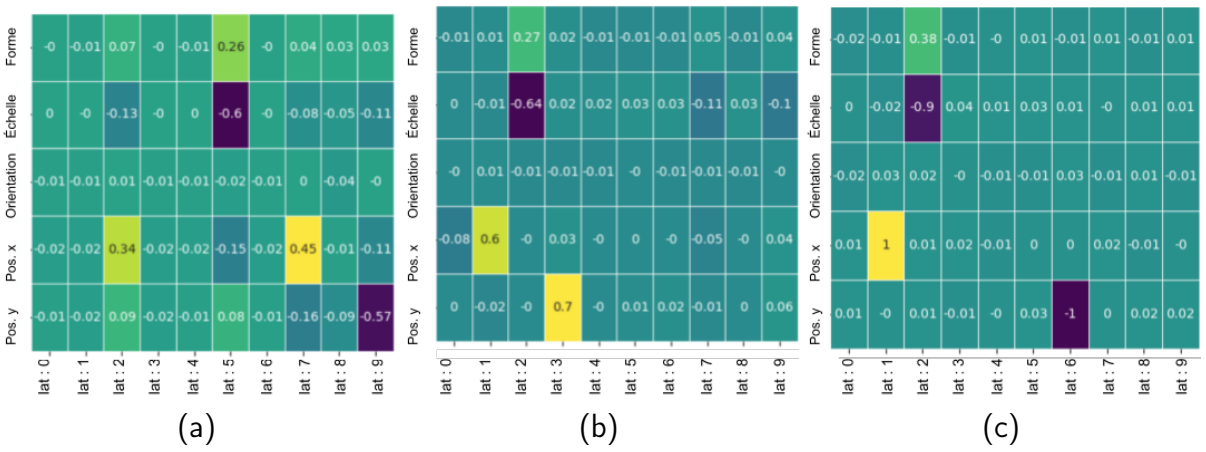



Fig 4.6: Matrices de corrélation de Spearman mesurées entre les facteurs génératifs et les variables latentes. (a) : $\beta$-VAE. (b) : $\beta$-TC-VAE. (c) : TC-NGVAE, le modèle proposé.

Quant à notre approche, dont les résultats sont représentés dans la Figure 4.6c, deux différences peuvent être observées. Premièrement, les corrélations sont plus élevées que celles des autres méthodes. Pour les facteurs génératifs représentant la position sur l'axe $x$ et sur l'axe $y$, elles sont égales à 1. Ce résultat implique une corrélation maximale entre les facteurs génératifs et leurs représentations respectives, et permet d'expliquer les écarts significatifs dans les mesures "DCI - Qualité explicite" et "MIG" relevées dans le Tableau 4.6. Par ailleurs, ces facteurs génératifs sont corrélés à une seule variable latente, justifiant la grande différence dans la mesure "DCI - Compacité".

Deux comportements communs se dégagent des résultats :

- les facteurs génératifs de forme et d'échelle sont systématiquement corrélés à une même et seule variable,
- le facteur d'orientation semble ne présenter aucune corrélation et donc ne pas être encodé dans l'espace latent.

Ces observations concordent avec l'analyse antérieure du jeu de données Dsprites, soulignant une ambiguïté dans la définition du facteur d'orientation et une dépendance entre les facteurs de forme et d'échelle.

En conclusion, un exemple d'images générées par le décodeur du TC-NGVAE, la matrice de covariance $Cov(\boldsymbol{z})$ et un parcours de l'espace latent issus de l'entraînement du TC-NGVAE sont représentés dans la Figure 4.7. Les images, reportées dans la Figure 4.7a, mettent en avant la capacité de génération de notre modèle. Elles ont été générées par le décodeur à partir d'un vecteur échantillonné depuis une distribution Normale de moyenne nulle et de variance dix, pour lequel les variables passives ont été éteintes. La variance choisie est cohérente avec la nature à queue lourde de la distribution *a posteriori* utilisée dans notre modèle.

La matrice de covariance, illustrée dans la Figure 4.7b, met en lumière trois variables latentes, exhibant une forte variance pour les moyennes correspondantes, traduisant leur

encodage des variations présentes dans les données, contrairement aux autres variables. Elle confirme également l'absence de corrélation entre les variables latentes, comme suggéré précédemment.

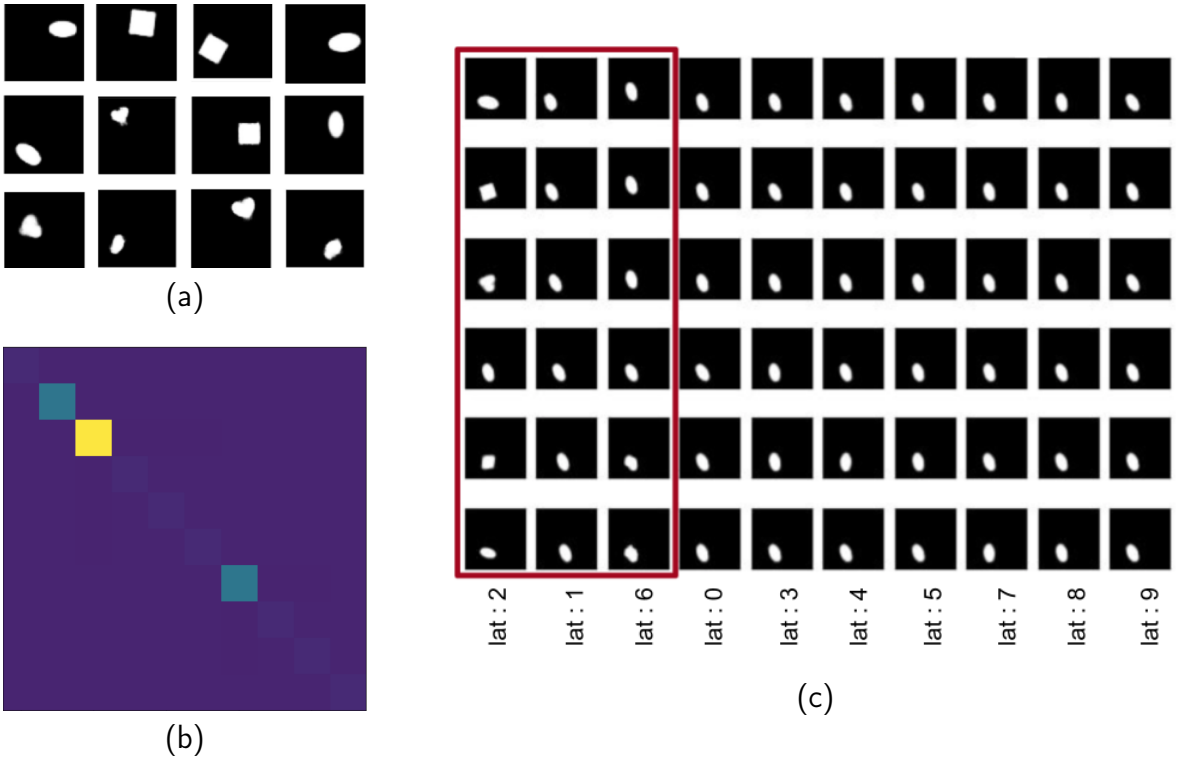



Fig 4.7: (a) : Images générées par le décodeur. (b) : Matrice de covariance empirique de $\boldsymbol{z}$. (c) : Parcours de l'espace latent après apprentissage sur Dsprites en considérant l'ensemble des facteurs génératifs.

En ce qui concerne le parcours de l'espace latent, illustré dans la Figure 4.7c dans laquelle les variables actives sont encadrées en rouge, la deuxième et la troisième colonne représentent respectivement les dimensions de $\boldsymbol{z}$ encodant les positions sur les axes $y$ et $x$. La première colonne correspond à une variable représentant à la fois la forme et l'échelle de l'objet. Par ailleurs, les images reconstruites dans cette colonne montrent également des variations dans l'orientation des objets, bien qu'aucune corrélation n'ait été relevée entre le facteur génératif d'orientation et les variables latentes dans la Figure 4.6c. Cela suggère que l'information d'orientation est présente dans $\boldsymbol{z}$ mais n'est pas interprétable comme un facteur génératif, en raison de l'ambiguïté dans sa définition. De plus, cette information semble être encodée dans la même variable que la forme et l'échelle, probablement due à une dépendance entre ces trois facteurs génératifs dans le jeu de données.

En conclusion, l'analyse des résultats obtenus par notre approche, lorsqu'elle est entraînée sur Dsprites en considérant l'ensemble des facteurs génératifs, peut être résumée à travers les points suivants :

- Le modèle a convergé vers un espace latent plus compact, modulaire et informatif que les autres approches de l'état de l'art, résultant en un espace latent bien mieux désentrelacé.
- Sa capacité de désentrelacement est moins sensible à l'initialisation des paramètres comparativement aux autres approches testées.

- La dépendance des facteurs génératifs observée au sein du jeu de données implique la représentation de ces informations dans une même variable latente, favorisée par le terme de corrélation totale.

Ces résultats démontrent l'amélioration significative de notre modèle à converger vers un espace latent désentrelacé par rapport aux autres approches, pour un jeu de données simple.

La suite des analyses se concentre sur les jeux de données plus complexes, à savoir Smallnorb dans un premier temps, puis Cars3d.

**Smallnorb** :

| Modèle | MIG | DCI Désent. | DCI Compacité | DCI Qualité explicite | z-min |
|---|---|---|---|---|---|
| $\beta$-VAE | 0.20 | **0.32** | 0.46(1) | 0.56 | $0.61 \pm 0.01$ |
| Factor-VAE | **0.27** | 0.30 | $0.46 \pm 0.01$ | 0.54 | $0.60 \pm 0.01$ |
| DIP-VAE-II | 0.26 | $0.23 \pm 0.01$ | 0.43 | 0.54 | $0.60 \pm 0.02$ |
| $\beta$-TC-VAE | 0.26 | 0.31 | 0.40 | 0.59 | **0.63** |
| BF-VAE-2 | $0.24 \pm 0.01$ | 0.29 | $0.48 \pm 0.02$ | $0.58 \pm 0.01$ | $0.31^{*} \pm 0.02$ |
| TC-NGVAE | $0.25 \pm 0.01$ | 0.28 | **0.66** | **0.71** | $0.58 \pm 0.02$ |

Tableau 4.7: Indicateurs de désentrelacement pour des méthodes entraînées sur Smallnorb.

L'observation des mesures de désentrelacement relevées dans le tableau 4.7 montre que le TC-NGVAE n'obtient pas de résultats supérieurs concernant les métriques MIG, "DCI - Désentrelacement" et "z-min variance". Toutefois, ces valeurs restent dans la fourchette moyenne. Cela signifie que le modèle ne surpasse pas les autres approches dans sa capacité à inférer une représentation dans laquelle une variable latente représente un unique facteur génératif. Cependant, une amélioration significative est observée pour les mesures "DCI - Compacité" et "DCI - Qualité explicite". Ils soulignent la meilleure capacité de notre approche à obtenir une représentation latente où les facteurs génératifs sont encodés dans au plus une variable, tout en étant suffisamment informatifs. Les matrices de corrélation de Spearman entre les facteurs génératifs et les variables latentes pour les modèles du BF-VAE-2, du Factor-VAE et de notre approche sont représentées dans la Figure 4.8.

Ces matrices, illustrées respectivement dans les Figures 4.8a et 4.8b liées aux apprentissages du BF-VAE-2 et du Factor-VAE, révèlent des similarités. Dans un premier temps, on constate que les facteurs de catégorie et d'élévation sont encodés dans plusieurs variables latentes. On observe en effet que la catégorie est corrélée avec au moins trois variables latentes pour les deux méthodes, tandis que l'élévation dans au moins deux. De plus, ces deux matrices ne révèlent que peu ou pas de corrélation entre l'orientation et les variables latentes.

En contraste, les résultats issus de notre approche, illustrés dans la Figure 4.8c, indiquent que seulement deux variables latentes sont corrélées avec le facteur génératif de forme, et une seule avec le facteur représentant l'élévation. Ces différences expliquent l'écart de résultats obtenu pour la mesure "DCI - Compacité", qui évalue la capacité du modèle à obtenir un espace latent dans lequel chaque facteur génératif est encodé dans au plus une variable latente. Par ailleurs, le niveau de corrélation entre les variables latentes et les facteurs génératifs, relativement plus élevé pour notre approche, permet d'expliquer la différence observée pour la mesure "DCI - Qualité explicite" en faveur du TC-NGVAE.

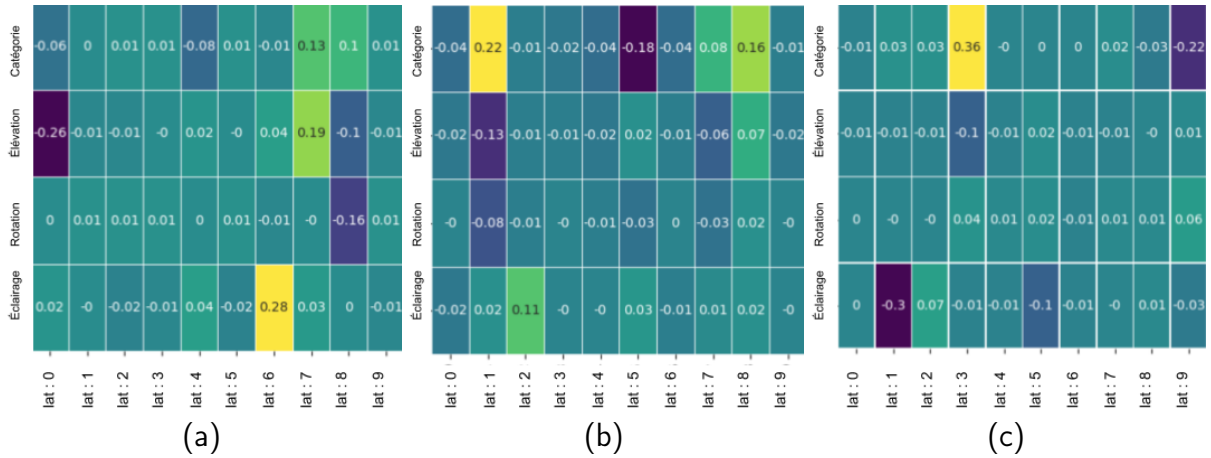



Fig 4.8: Matrices de corrélation de Spearman mesurées entre les facteurs génératifs et les variables latentes. En jaune sont affichées les corrélations fortement positives et en bleu foncé les corrélations fortement négatives. (a) : BF-VAE-2. (b) : Factor-VAE. (c) : TC-NGVAE, le modèle proposé.

Un point commun aux trois méthodes réside dans la représentation des facteurs génératifs de forme et d'élévation dans une même variable latente. On observe également l'absence de corrélation entre le facteur d'orientation et l'espace latent. Ces résultats peuvent s'expliquer par les caractéristiques intrinsèques du jeu de données. Des exemples de facteurs génératifs mal définis sont ainsi illustrés dans la section 4.2.3 du manuscrit.

Finalement, un exemple d'images générées par le décodeur, la matrice de covariance $Cov(\boldsymbol{z})$ ainsi qu'un parcours de l'espace latent, sont présentés dans la Figure 4.9.

Les images relevées dans la Figure 4.9a représentent des données obtenues en sortie du décodeur lorsqu'on lui fournit en entrée un vecteur échantillonné depuis une distribution Normale de moyenne nulle et de variance trois, pour lequel les variables passives sont éteintes. On observe une grande variabilité dans les résultats obtenus, ce qui démontre la capacité de notre approche à apprendre la loi $p_{data}$. La Figure 4.9b illustre la matrice de covariance mesurée sur l'espace latent du TC-NGVAE et met en évidence les variables latentes encodant la variabilité du jeu de données, spécifiquement les variables 1, 3 et 9. Ces résultats corroborent l'analyse effectuée jusqu'alors.

Les informations encodées dans les variables latentes 1, 3 et 9 sont visibles dans le parcours de l'espace latent présenté dans la Figure 4.9c, dans laquelle les variables actives sont encadrées en rouge. La première colonne illustre l'impact d'une modification de la moyenne inférée pour la variable 1, qui encode des informations liées à la condition d'éclairage. La troisième colonne, liée aux modifications effectuées pour la variable 9,

semble être la seule à encoder la catégorie des jouets. On y observe en effet des figurines humaines, des voitures et des camions. Ces résultats sont en accord avec l'analyse de la matrice de Spearman, relevée dans la Figure 4.8c. Concernant les facteurs de rotation et de condition d'éclairage, ces derniers sont définis de façon ambigüe, et sont abordés dans la section 4.2.3 du manuscrit. En effet, la modification apportée à la variable latente 3, et illustrée dans la deuxième colonne, impacte la rotation de l'objet dans l'image reconstruite. À l'instar de Dsprites, cette opposition avec l'analyse de Spearman, qui met en avant l'absence de corrélation entre le facteur de rotation et l'espace latent, est due à la définition adoptée par les auteurs pour le facteur de rotation. Finalement, chaque variable active semble encoder plus ou moins d'information liée à l'éclairage. En effet, ce facteur est intrinsèquement lié à la valeur moyenne du niveau de gris de l'image, impacté par la forme de l'objet et son élévation, ce qui permet d'expliquer les résultats observés.

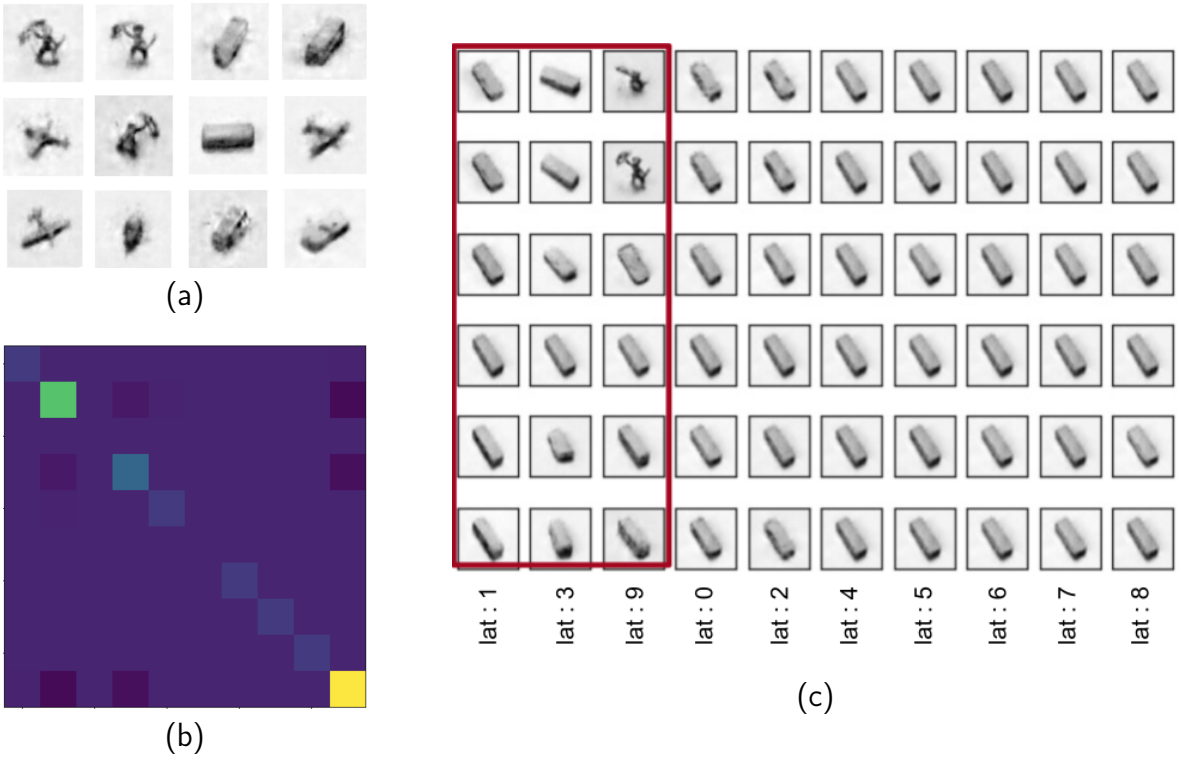



Fig 4.9: (a) : Images générées par le décodeur du TC-NGVAE. (b) : Matrice de covariance empirique de ($\boldsymbol{z}$) du TC-NGVAE. (b) : Parcours de l'espace latent du TC-NGVAE après apprentissage sur Smallnorb.

Les résultats majeurs de l'analyse du TC-NGVAE entraîné sur le jeu de données Smallnorb peuvent être résumés à travers les points suivants :

- Le TC-NGVAE converge vers un espace latent désentrelacé.
- L'approche proposée présente une capacité supérieure aux autres méthodes à représenter les facteurs génératifs dans un espace latent dans lequel on retrouve les propriétés de compacité et de qualité explicite.
- Les ambiguïtés observées dans la matrice de corrélation de Spearman mesurée entre les variables latentes et les facteurs génératifs, ainsi que l'interprétation du parcours de l'espace latent, peuvent être expliquées par la complexité inhérente du jeu de données.

Ces résultats permettent de mettre en évidence l'intérêt d'utiliser notre approche lorsqu'il s'agit d'obtenir une représentation désentrelacée à partir d'un jeu complexe de données pour lequel les facteurs génératifs restent tout de même suffisamment bien définis.

**Cars3d** : Contrairement aux entraînements effectués sur les jeux de données précédents, les indicateurs de désentrelacement relevés dans le tableau 4.8 pour le TC-NGVAE montrent des résultats supérieurs à l'état de l'art sur aucune mesure, voir légèrement inférieurs concernant la mesure "DCI - Qualité explicite". Les métriques restent en moyenne comparables à celles obtenues pour les autres méthodes, indiquant que lorsqu'elle est entraînée sur ce jeu de données, notre approche ne converge pas vers un espace latent mieux désentrelacé, que ce soit en termes de modularité, de compacité ou de qualité explicite. La suite de l'analyse confirme et explique ce comportement.

| Modèle | MIG | DCI Désent. | DCI Compacité | DCI Qualité explicite | z-min |
|---|---|---|---|---|---|
| $\beta$-VAE | $0.12 \pm 0.01$ | $0.35 \pm 0.03$ | $\mathbf{0.34 \pm 0.02}$ | $0.49 \pm 0.02$ | $0.89 \pm 0.03$ |
| Factor-VAE | $0.09 \pm 0.02$ | $0.25 \pm 0.03$ | 0.26 | $0.56 \pm 0.01$ | $0.91 \pm 0.02$ |
| DIP-VAE-II | $0.04 \pm 0.02$ | $0.18 \pm 0.07$ | $0.17 \pm 0.04$ | $0.56 \pm 0.03$ | $0.91 \pm 0.02$ |
| $\beta$-TC-VAE | $\mathbf{0.14 \pm 0.02}$ | $\mathbf{0.38 \pm 0.07}$ | $0.32 \pm 0.03$ | $0.57 \pm 0.03$ | $\mathbf{0.92 \pm 0.02}$ |
| BF-VAE-2 | $0.11 \pm 0.03$ | $0.25 \pm 0.07$ | $0.24 \pm 0.03$ | $\mathbf{0.66 \pm 0.03}$ | $0.36^{*} \pm 0.04$ |
| TC-NGVAE | $0.11 \pm 0.02$ | $0.23 \pm 0.13$ | $0.32 \pm 0.05$ | $0.48 \pm 0.03$ | $0.88 \pm 0.06$ |

Tableau 4.8: Indicateurs de désentrelacement pour des méthodes entraînées sur Cars3d.

Les matrices de corrélation de Spearman, mesurées entre les facteurs génératifs et les variables latentes pour la méthode du DIP-VAE-II, du $\beta$-TC VAE et du TC-NGVAE, sont reportées dans la Figure 4.10.

L'observation de cette Figure, en comparaison avec les résultats obtenus dans le tableau 4.8, met en avant certains points intéressants. Premièrement, la Figure 4.10a, qui représente les résultats obtenus pour le DIP-VAE-II, montre que l'espace latent obtenu converge vers une représentation dans laquelle les facteurs génératifs sont encodés dans plusieurs variables latentes. En effet, l'élévation et la rotation sont corrélées chacune dans trois dimensions de $\boldsymbol{z}$. De plus, une même variable latente encode plusieurs facteurs génératifs, tel qu'on peut l'observer pour la variable 8. Ces résultats permettent d'expliquer la mauvaise qualité des mesures de "DCI - Désentrelacement" qui quantifie la capacité de modularité de $\boldsymbol{z}$ et "DCI - Compacité".

À l'inverse, la matrice illustrée dans la Figure 4.10c, qui représente les résultats obtenus après entraînement de notre approche, montre que l'élévation ne semble corrélée qu'avec une seule variable, tandis que la rotation avec seulement deux. De ce fait, les métriques de compacité et de modularité sont meilleures. On pourrait s'attendre à des résultats similaires concernant le $\beta$-TC-VAE dans la Figure 4.10b et notre approche, mais on remarque

toutefois une capacité de représentation compacte similaire, et de modularité bien moindre pour le TC-NGVAE. De plus, les résultats inhérents à la mesure "z-min variance" et "DCI - Qualité explicite" obtenus par les deux approches de l'état de l'art sont également meilleurs que ceux de la méthode proposée. Cette constatation peut s'expliquer par deux points :

- La modélisation de notre approche favorise un nombre minimal de variables latentes actives, par rapport au nombre de facteurs génératifs. Toutefois, il semblerait que pour obtenir un espace latent suffisamment informatif pour ce jeu de données, le nombre de variables latentes devrait être largement supérieur au nombre de facteurs génératifs définis, comme c'est le cas pour le DIP-VAE-II et le $\beta$-TC-VAE, ce qui permet d'expliquer les meilleurs résultats obtenus en termes de qualité explicite pour ces deux modèles.

- Les indicateurs de modularité et de compacité obtenues par notre approche montrent des variations supérieures à l'état de l'art sur les cinq initialisations testées, ce qui explique que pour la matrice relevée dans la Figure 4.10c, propre à une initialisation spécifique, la modularité semble meilleure que celle obtenue par le $\beta$-TC-VAE bien qu'en moyenne, elle ne le soit pas. Ces résultats peuvent être expliqués par la nature des facteurs génératifs définis pour ce jeu de données. Une analyse de ces derniers effectuée dans la suite du manuscrit permet de mettre en évidence une dépendance forte entre l'ensemble des facteurs génératifs définis.

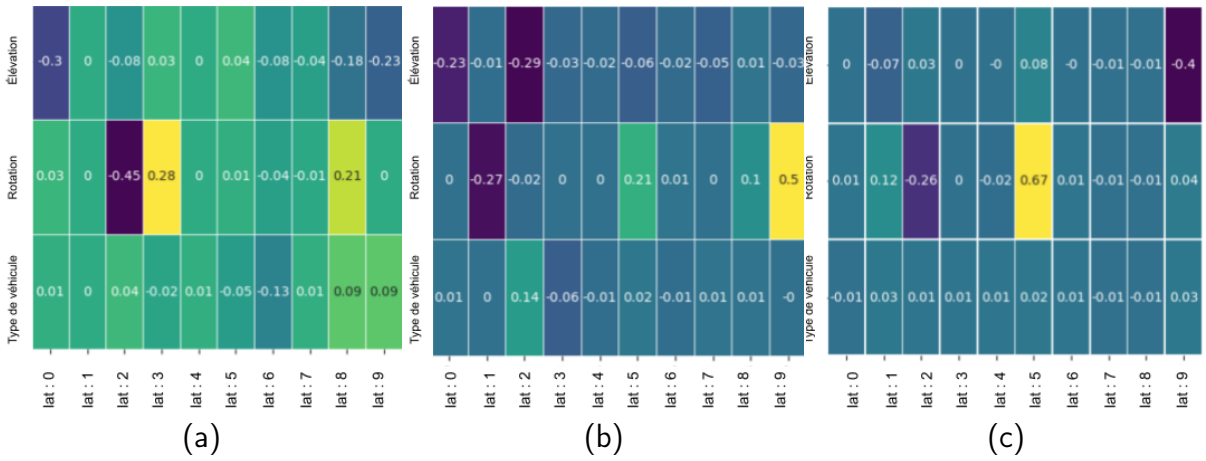



Fig 4.10: Matrices de corrélation de Spearman mesurées entre les facteurs génératifs et les variables latentes. (a) : DIP-VAE. (b) : $\beta$-TC VAE. (c) : TC-NGVAE, le modèle proposé.

L'ensemble des résultats obtenus pour ce jeu de données est disponible en Annexe E.2.4.

Finalement, les images générées par le décodeur, la matrice de covariance de $\boldsymbol{z}$ ainsi qu'un parcours de l'espace latent obtenu par le TC-NGVAE sont présentés dans la Figure 4.11. Les images relevées dans la Figure 4.11a sont obtenues en sortie du décodeur lorsqu'il prend en entrée un vecteur échantillonné depuis une loi Normale de moyenne nulle et variance identité, pour lequel les variables passives sont éteintes. Différents modèles de véhicules, d'élévation, de rotation et de couleur différentes, peuvent être observés. La valeur de la variance définie pour la distribution à échantillonner, plus faible que pour les autres jeux de données, s'explique par les performances moins élevées de notre approche sur Cars3d. Toutefois, la variabilité observée dans les images générées démontre

la capacité du TC-NGVAE à apprendre la distribution de $p_{data}$. Nous observons dans la Figure 4.11b une matrice de covariance mettant en évidence la capacité de polarisation du modèle, en faisant émerger quatre variables latentes actives et très peu de corrélation entre les variables. Comme attendu, le parcours de l'espace latent présenté dans la Figure 4.11c ne permet pas de discerner convenablement des facteurs génératifs spécifiques, mais fait clairement apparaître un encodage d'informations.

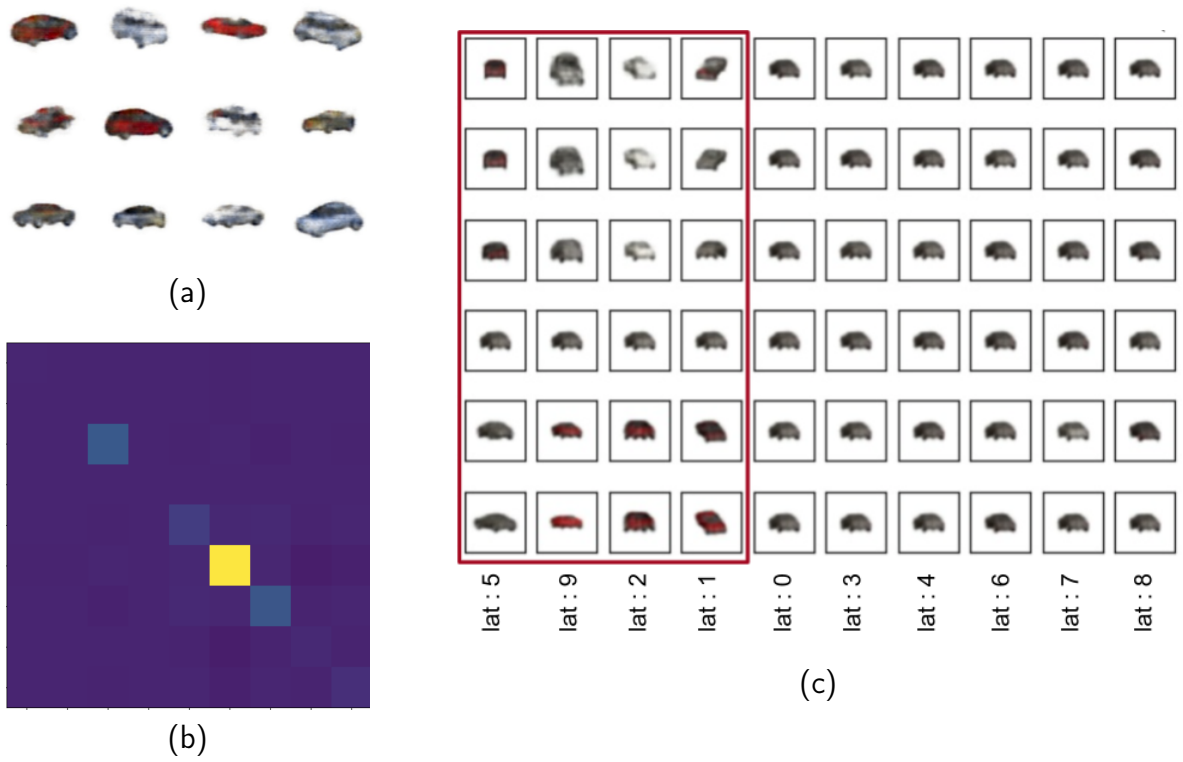



Fig 4.11: (a) : Images générées par le décodeur du TC-NGVAE. (b) : Matrice de covariance empirique de $\boldsymbol{z}$ du TC-NGVAE. (b) : Parcours de l'espace latent du TC-NGVAE après apprentissage sur Cars3d.

La section suivante a pour objectif d'illustrer des incohérences contenues dans les jeux de données Dsprites et Cars3d permettant ainsi de clarifier certains résultats.

### Analyse des jeux de données Cars3d et Smallnorb

Dans cette partie du manuscrit, une analyse des jeux de données Cars3d et Smallnorb est effectuée. L'objectif est d'apporter des éclaircissements sur les résultats observés précédemment.

Concernant le jeu de données Cars3d, les observations ont porté sur les facteurs génératifs correspondant au modèle de voiture et à la valeur d'élévation, dont les résultats sont illustrés dans la Figure 4.12.

Les images des cinq premiers modèles de voitures en fonction des valeurs d'élévation sont présentées dans la Figure 4.12a. Le graphique 4.12b affiche, quant à lui, la valeur de la position en $y$ du pixel le plus bas appartenant au véhicule pour chaque image. Il apparaît qu'il est possible, pour certains modèles, de déduire la catégorie de la voiture à partir de la valeur du pixel le plus bas, qui est influencé par l'élévation. Ainsi, il existe une certaine dépendance entre le facteur génératif de catégorie de voiture et celui de l'élévation. Des

analyses similaires pourraient être menées en considérant les modèles de voitures et le facteur génératif d'orientation. Il est fortement probable que selon la catégorie, la taille de la voiture impacte la position du pixel le plus à gauche dans l'image.

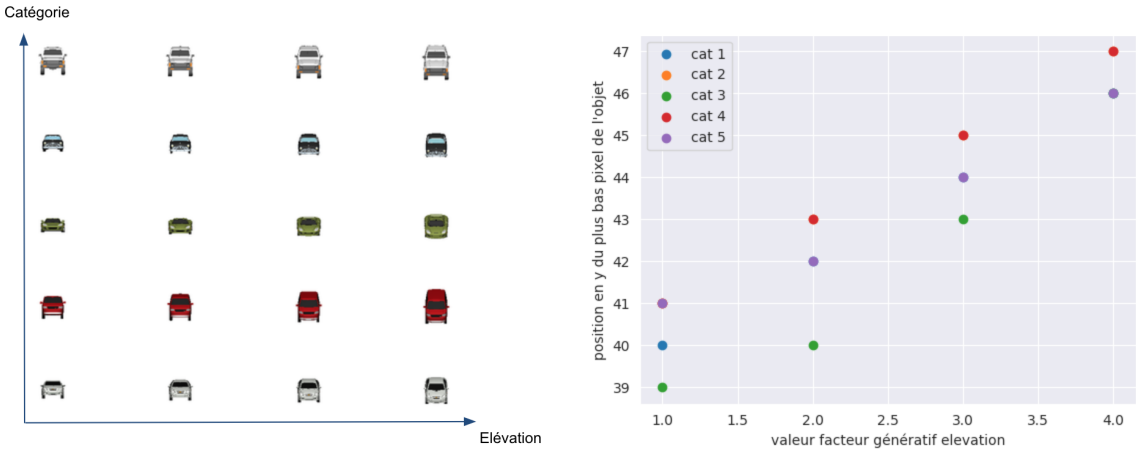



Fig 4.12: (a) : Variation du modèle de voiture en fonction de l'élévation. (b) : Valeur du pixel le plus bas de chaque véhicule en fonction du facteur génératif d'élévation.

Les trois facteurs génératifs définis par les auteurs pour ce jeu de données semblent ainsi tous montrer une certaine interdépendance, ce qui permet d'expliquer les résultats en retrait quant à la capacité de désentrelacement obtenue pour l'ensemble des méthodes testées comparativement aux autres jeux de données. Cela met également en avant une limitation de notre approche qui ne possède pas la capacité de converger vers une représentation mieux désentrelacée que l'état de l'art lorsque l'ensemble des facteurs génératifs est défini de façon trop ambigüe et s'éloigne de l'hypothèse d'indépendance entre ces derniers.

Concernant le jeu de données Smallnorb, l'analyse a porté sur les facteurs génératifs liés à la catégorie du jouet et à la condition d'éclairage. Les résultats obtenus sont présentés dans la Figure 4.13.

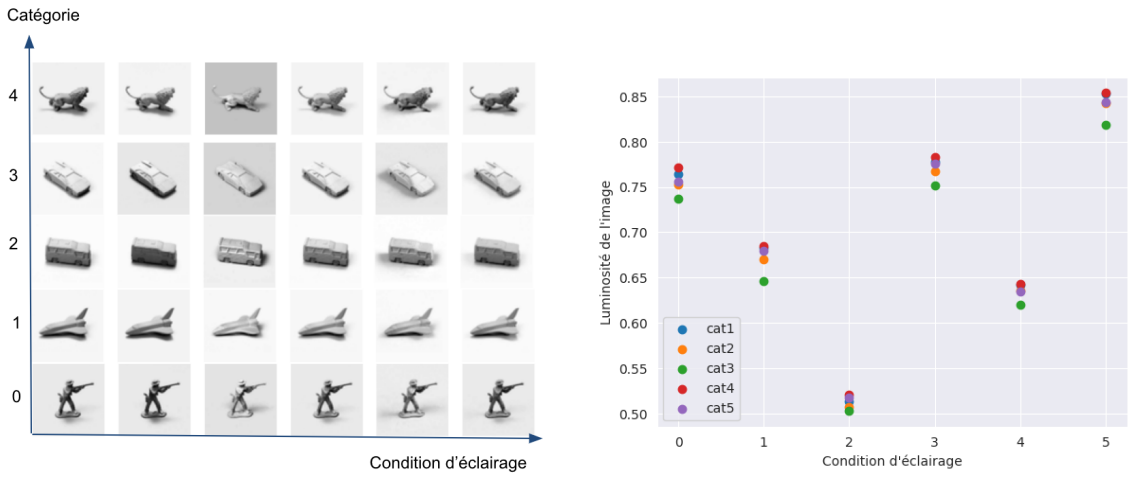



Fig 4.13: (a) : Variation de la catégorie d'objet en fonction de la condition d'éclairage. (b) : Luminosité de l'image en fonction de la condition d'éclairage.

Cette Figure présente les images du jeu de données pour les quatre premières catégories de jouets selon les six conditions d'éclairage. Il est notable que la luminosité de l'image, qui se définit ici comme l'intensité moyenne des pixels, varie en fonction de la catégorie du jouet, même sous une condition d'éclairage constante. Cela est illustré dans la troisième colonne de la Figure 4.13a, où l'on constate une différence d'intensité entre une voiture et un camion. Une analyse complémentaire permet de mettre en évidence une relation étroite entre cette condition d'éclairage et le niveau moyen de luminosité dans l'image, tel qu'illustré dans la Figure 4.13b. Elle contribue à expliquer pourquoi un camion, qui présente plus de pixels foncés, se retrouve dans une image plus lumineuse qu'une voiture à condition d'éclairage similaire. Cependant, il est important de noter que le niveau moyen de luminosité d'une image est également impacté par le facteur génératif d'élévation, qui joue un rôle sur l'intensité des ombres portées et donc la luminosité globale. Ces constatations permettent de clarifier certaines incohérences rencontrées lors de l'analyse des résultats du TC-NGVAE lorsqu'il est entraîné sur le jeu de données Smallnorb, notamment dans la mise en perspective du parcours de l'espace latent et de la matrice de Spearman entre les facteurs génératifs et les variables latentes.

L'examen de ces jeux de données, associé à l'analyse du jeu de données Dsprites réalisée dans la section 4.2.1 du manuscrit, permettent d'éclaircir des comportements observés lors de l'analyse de l'approche proposée. En effet, la modélisation adoptée repose sur une définition du désentrelacement selon laquelle les facteurs génératifs sont indépendants entre eux, ce qui n'est pas le cas pour tous ceux des jeux de données considérés. De plus, cela met en lumière certaines limitations lors de la comparaison des performances obtenues par les méthodes de l'état de l'art, lorsqu'elle est effectuée à partir de métriques supervisées par les facteurs génératifs.

## 4.3 Conclusion

Ce chapitre propose une extension du modèle NGVAE, baptisée TC-NGVAE, qui intègre un terme de groupe-lasso sur les probabilités en sortie de l'encodeur ainsi qu'un terme de corrélation totale estimé par une méthode de Monte-Carlo non biaisée. Une première étude comparative entre les deux modèles révèle l'impact positif de l'ajout de ces termes dans la fonction de coût. Le nouveau modèle proposé présente une meilleure capacité d'encodage, caractérisée par une variance des moyennes inférées significativement supérieure, une plus forte décorrélation des variables de l'espace latent et une amélioration de la capacité de désentrelacement mesurée par la métrique "z-min variance". Les matrices de corrélation de Spearman entre les facteurs génératifs et les variables latentes soulignent les propriétés supérieures de compacité et de modularité du TC-NGVAE, confirmant les résultats observés. De plus, la représentation latente obtenue par cette méthode lors de l'entraînement sur Dsprites a démontré une corrélation linéaire entre deux variables latentes et l'information de position respectivement encodée dans chacune d'elle, un atout majeur pour effectuer une manipulation *a posteriori* de cette caractéristique.

Enfin, une analyse comparative entre les résultats de désentrelacement obtenus pour des espaces latents de dimensions respectives de 10 et 20 montre que le TC-NGVAE, contrairement à l'ensemble des autres approches de l'état de l'art considérées, n'est que peu sensible à la dimension spécifiée pour $\boldsymbol{z}$. Ces résultats soulignent l'intérêt d'utiliser la polarisation comme un biais inductif dans la convergence vers un espace latent désentrelacé.

La seconde partie du chapitre consiste en une évaluation approfondie des performances

du modèle proposé comparativement à d'autres méthodes de désentrelacement de l'état de l'art. Elle s'appuie sur cinq mesures usuelles testées pour trois jeux de données à partir de cinq initialisations différentes. Le TC-NGVAE surpasse notablement les autres approches sur des jeux de données simples comme Dsprites, et démontre également une bonne insensibilité à l'initialisation des paramètres, particulièrement importante dans le contexte stochastique des méthodes génératives. Sur des jeux de données plus complexes, tels que Smallnorb, le TC-NGVAE converge vers un espace latent bien plus compact et explicite que les autres approches, tandis que ses performances pour les autres mesures restent comparables. Pour Cars3d, les performances du TC-NGVAE sont équivalentes à celles des autres méthodes. Les valeurs peu concluantes obtenues peuvent être expliquées par la complexité du jeu de données et la définition des facteurs génératifs.

Finalement, des analyses sont également menées sur chaque jeu de données, mettant en lumière des dépendances entre des paires de facteurs génératifs et des ambiguïtés dans leur définition. Ces observations expliquent certaines contradictions relevées dans les performances de désentrelacement et soulignent la limite à l'utilisation de métriques supervisées pour comparer différentes approches. Elles mettent également en avant certaines limitations dans l'utilisation du terme de corrélation totale au sein de la fonction de coût, car fondé sur l'hypothèse d'indépendance totale entre les facteurs génératifs.

L'étude conduite dans son ensemble met en évidence les avantages du TC-NGVAE dans la représentation compacte des données, tout en soulignant ses limites selon le jeu de données considéré. Malgré celles-ci, le TC-NGVAE répond efficacement au besoin initial d'obtenir une représentation interprétable et manipulable des données d'entrée dans un espace à dimensions réduites, tout en étant robuste à l'initialisation des paramètres et à la dimension de l'espace latent.

# Conclusion et perspectives

Ce chapitre résume les résultats obtenus et discute des verrous scientifiques rencontrés, avant d'explorer des pistes pour de futures recherches.

### 4.3.1 Synthèse

Dans le cadre de l'amélioration du système anti-collision du LMJ, l'objectif principal de nos travaux a consisté à développer un algorithme fondé sur l'apprentissage automatique non supervisé. Ce dernier vise à obtenir une représentation interprétable et manipulable de caractéristiques inhérentes à un objet présent dans les images acquises dans la chambre d'expérience, qui favorise une analyse d'incertitudes *a posteriori*. Une analyse des différents modèles de l'état de l'art a orienté nos recherches vers les méthodes génératives neuronales probabilistes, en particulier les VAEs. Ils se distinguent par leur capacité à projeter les données d'entrée dans un espace latent de dimension réduite, facilitant l'extraction de caractéristiques significatives pour l'apprentissage de représentations. De plus, la construction d'une fonction de *vraisemblance* des données d'entrée par rapport à cet espace latent favorise l'analyse d'incertitudes.

Un aspect crucial des travaux menés est la manipulation des représentations encodées dans l'espace latent. Il est en particulier souhaitable que celles-ci soient désentrelacées. Dans le contexte de l'amélioration du système anti-collision, cette propriété permettrait la manipulation de certaines informations encodées, telles que la position des diagnostics dans l'image. Elle repose sur l'hypothèse que les données réelles sont générées à partir de facteurs indépendants, appelés facteurs génératifs. Bien qu'il n'existe pas une unique définition du désentrelacement, un consensus émerge au sein de l'état de l'art, établissant qu'un espace latent désentrelacé est défini par une représentation dans laquelle la modification d'un facteur génératif des données d'entrée affecte une seule variable de l'espace latent. Dans ces conditions, trois propriétés sont souhaitables : la qualité explicite de l'espace latent qui mesure sa capacité à encoder suffisamment d'information pour décrire l'image d'entrée, sa modularité désignant le fait qu'une variable latente représente un unique facteur génératif, et sa compacité indiquant qu'un facteur génératif est encodé dans au maximum une variable latente. Bien que la pertinence de ces propriétés soit toujours discutée, nous nous sommes orientés sur ces dernières afin d'établir un cadre rigoureux pour nos recherches.

De nombreuses méthodes de l'état de l'art visent à modifier le VAE, tel que proposé dans [Kingma et Welling, 2014], en introduisant des biais inductifs favorisant la convergence du modèle vers un espace latent désentrelacé. Parmi elles, trois approches se distinguent : la première ajoute une pondération sur le terme de régularisation de l'ELBO pour encourager l'indépendance statistique des variables latentes, la seconde ajoute un terme de régularisation directement à la fonction coût afin d'induire celle-ci et la troisième, en parallèle de l'ajout d'un terme de régularisation à la fonction coût, modifie la distribution *a priori* des variables latentes afin de favoriser une distribution plus complexe. Toutefois,

l'ensemble de ces méthodes est sensible au réglage des hyperparamètres, et particulièrement à celui lié à la dimension de l'espace latent. En effet, ce dernier est crucial au modèle pour converger vers un espace latent désentrelacé.

Nos travaux ont conduit au développement d'une première proposition, le NGVAE, dont l'objectif est de converger vers un espace latent polarisé où le nombre de variables actives correspond au nombre de facteurs génératifs du jeu de données, tout en réduisant la sensibilité de l'apprentissage à la dimension définie pour le vecteur latent. Ce modèle utilise des variables auxiliaires dans l'espace latent pour établir un modèle bayésien hiérarchique, dans lequel les variances des distributions *a posteriori* deviennent des variables aléatoires. Leur distribution, définie par un mélange de lois inverses-Gamma, dépend d'une probabilité inférée par l'encodeur, qui décrit l'information contenue dans chaque variable. Les paramètres de cette loi sont définis suite à une analyse rigoureuse de la polarisation. Pour favoriser la convergence vers un espace partitionné comme attendu, une inversion de l'ordre des distributions défini dans le terme de KL-divergence de la fonction coût est nécessaire. En effet, étant donné l'expression de la KL entre deux distributions Normales-Gamma, selon la loi prise comme référence, elle tend à forcer le modèle à n'inférer que des variables non informatives. Finalement, une régularisation est ajoutée sur les termes de probabilités en sortie de l'encodeur, les encourageant à prendre des valeurs extrêmes à l'aide d'une mesure d'entropie, ce qui permet de les utiliser comme un biais inductif lors de l'apprentissage. En effet, elles sont mises à profit pour masquer les variables latentes non informatives fournies au décodeur, ce qui permet d'augmenter la parcimonie des paramètres de la première couche du réseau et favorise en retour que l'information ne soit contenue que dans un sous-ensemble de l'espace latent. Les analyses réalisées mettent en avant la capacité du NGVAE à bien converger vers un espace polarisé, dont le nombre de variables actives correspond bien au nombre de facteurs génératifs. Par ailleurs, l'information qu'elles contiennent est supérieure à celle obtenue par d'autres méthodes de l'état de l'art, et ces variables sont davantage décorrélées les unes aux autres, conformément à nos attentes. Finalement, à conditions expérimentales similaires, des tests effectués ont montré que le NGVAE était moins sensible que les autres modèles à la dimension de l'espace latent dans sa capacité de polarisation. Toutefois, ses performances en désentrelacement restent limitées : bien qu'encodées uniquement dans un sous-ensemble de l'espace latent de taille souhaitable, les représentations des facteurs génératifs restent entrelacées.

Pour améliorer les performances de notre approche, nous avons proposé un second modèle, le TC-NGVAE, qui consiste en une extension du précédent. Elle réside dans un premier temps en l'ajout dans la fonction coût d'un terme de régularisation groupe-lasso appliqué aux probabilités inférées par l'encodeur. Cela permet au modèle d'encoder les facteurs génératifs dans les mêmes dimensions de l'espace latent pour différentes images. Dans un second temps, une régularisation favorisant l'indépendance statistique des variables latentes à l'aide d'un terme de corrélation totale est également intégrée. Ne possédant pas de forme analytique, cette dernière est estimée par une approche de Monte-Carlo non biaisée, le "Minibatch-Stratified Sampling", étendu à une distribution Normale-Gamma. Afin de comparer la capacité de désentrelacement du TC-NGVAE avec d'autres modèles de l'état de l'art, différents tests ont été réalisés. Le choix de ces autres modèles s'est basé sur une analyse des mesures de désentrelacement mise à disposition dans la librairie "Disentanglement-Lib" [Locatello *et al.*, 2019b]. L'ensemble des résultats obtenus a permis de mettre en exergue l'excellente capacité de notre approche à décorréler les variables latentes les unes aux autres, comparativement aux méthodes testées. Cela a également montré sa très bonne capacité à converger vers un espace latent bien plus com-

pact et explicite lorsque les facteurs génératifs considérés au sein du jeu de données sont bien indépendants entre eux. L'obtention d'une telle représentation permet d'effectuer une meilleure manipulation des facteurs encodés, et d'envisager par la suite une analyse d'incertitudes. Toutefois, sur un jeu de données, le TC-NGVAE ne se démarque pas de l'état de l'art. Une analyse des jeux de données pour lesquels ce comportement est observé met en évidence deux propriétés qui expliquent ce comportement. Dans un premier temps, certains facteurs sont définis de façon ambigüe, ce qui ne permet pas au modèle l'apprentissage d'un encodage pertinent. Enfin, certaines paires de facteurs possèdent des interdépendances, ce qui va à l'encontre de l'hypothèse initiale sur laquelle se base la définition d'un facteur génératif. De ce fait, la modélisation adoptée pour le TC-NGVAE n'a pas la capacité d'encoder ces facteurs dans différentes variables latentes, et ainsi d'obtenir de meilleurs résultats.

### 4.3.2 Discussion

Le désentrelacement est un concept difficile à manipuler, d'autant plus qu'il n'existe pas de consensus scientifique quant à une définition formelle. L'analyse des propriétés obtenues pour une telle représentation doit être effectuée en tenant compte des manipulations *a posteriori* réalisées sur l'espace latent. De plus, chacune des définitions qui y sont associées donne lieu à des métriques différentes, qui possèdent parfois le même objectif.
Il est alors nécessaire de comprendre les avantages et limitations de chacune, et de définir précisément les propriétés à mesurer, afin de délinéer un cadre clair pour l'analyse des capacités de désentrelacement.

Par ailleurs, ces métriques sont pour la plupart supervisées par les facteurs génératifs utilisés pour la création des jeux de données. Cela peut être un inconvénient lorsqu'il s'agit de mesurer la capacité de désentrelacement d'un modèle entraîné sur des données réelles. Dans ce contexte, la valeur des facteurs génératifs n'est généralement pas connue, ce qui rend difficile l'évaluation de la représentation obtenue. De plus, certains facteurs génératifs définis ne suivent pas l'hypothèse sur laquelle repose leur définition, à savoir une indépendance statistique totale entre eux. Dans ce contexte, de mauvais résultats obtenus pour certaines mesures n'indiquent pas forcément un mauvais désentrelacement. Lors du développement de modèles visant à obtenir un espace latent désentrelacé, ainsi que durant leur analyse, l'ensemble de ces considérations est également à prendre en compte.

De ce fait, l'algorithme qui sera utilisé pour l'amélioration du système anti-collision du LMJ devra être entraîné à partir d'un jeu de données d'images qui prend en considération la sensibilité du modèle aux données d'entrées. Ainsi, l'utilisation du TC-NGVAE pour la manipulation et l'étude d'incertitudes des informations encodées est plus pertinente que les autres approches, à condition que les facteurs génératifs soient bien représentés dans la base de données, et qu'ils soient indépendants entre eux.

### 4.3.3 Directions futures

À la lumière des résultats obtenus, de nombreuses perspectives peuvent être envisagées.

Concernant le TC-NGVAE, la modélisation aujourd'hui adoptée pourrait encore être améliorée. En effet, la distribution *a posteriori* telle qu'elle est définie par un mélange de lois Normales-Gamma dont la probabilité prend des valeurs extrêmes, permet de consid-

érer uniquement deux modes. Il serait intéressant de parcourir l'espace des distributions de façon continue en définissant un chemin dans la variété de l'espace des paramètres. Finalement, les limitations relevées quant à la présence de paires de facteurs génératifs interdépendants remettent en cause la nécessité de converger vers une représentation dans laquelle un facteur génératif n'est encodé que dans une unique variable latente. Dans l'objectif d'améliorer les résultats obtenus sur des jeux de données complexes, il pourrait être envisagé de forcer l'espace latent du TC-NGVAE à encoder les facteurs génératifs par blocs de variables latentes, où la dimension de chaque bloc correspond au nombre minimal de variables nécessaires pour encoder correctement chaque facteur, et dans lequel les blocs sont indépendants entre eux.

Quant à l'intégration de l'algorithme au sein du système anti-collision, différentes pistes peuvent également être envisagées. La position des objets dans la chambre d'expérience du LMJ étant un paramètre d'intérêt, un biais inductif relativement à cette dernière pourrait être intégré de façon implicite à la fonction coût, à la manière du "Spatial Broadcast Decoder" [Watters *et al.*, 2019]. Cette approche, associée à la capacité de polarisation et de désentrelacement du TC-NGVAE, pourrait améliorer la capacité de notre modèle à faire émerger dans l'espace latent un encodage linéairement dépendant à l'information de position. Il serait également souhaitable que le TC-NGVAE puisse retourner des informations relatives à différents objets présents dans l'image. En effet, plusieurs diagnostics peuvent être insérés en même temps dans le centre chambre, ce qui requiert un traitement séparé de différentes régions dans l'image. Dans ce but, des modèles existants effectuent une première étape, dite d'attention, dont l'objectif est de segmenter les objets, avant d'obtenir une représentation pour chacun d'entre eux. C'est notamment le cas de MoNet [Burgess *et al.*, 2019], IODINE [Greff *et al.*, 2019] ou encore SPACE [Lin *et al.*, 2020], qui ont pour point commun de s'appuyer sur un VAE pour obtenir la représentation de chacun des objets segmentés. Dans ce contexte, le TC-NGVAE pourrait tout à fait être utilisé, et ainsi avoir la capacité d'inférer un espace latent désentrelacé pour chaque diagnostic.

Un retour à l'opérateur de l'information de position en temps réel pourrait également être envisagé. Cela supposerait l'extension du TC-NGVAE à un traitement séquentiel d'images. Cette propriété pourrait être appliquée à notre algorithme à travers la prise en considération d'une information temporelle, à l'instar des VAEs dynamiques [Girin *et al.*, 2020].

Enfin, concernant l'analyse d'incertitudes effectuées sur l'encodage des facteurs, la représentation obtenue par notre modèle est une première base solide pour effectuer des mesures statistiques. Toutefois, en plus de ces facteurs, il pourrait être intéressant d'avoir des retours quant à d'autres informations de plus bas niveau, relatives par exemple à la luminosité ou au bruit dans l'image. À cette fin, l'espace latent du TC-NGVAE pourrait être étendu en une hiérarchie de couches de variables latentes, à l'image du Ladder-VAE [Sønderby *et al.*, 2016b] ou encore de BIVA [Maaløe *et al.*, 2019b]. En effet, les études effectuées par les auteurs de ces modèles montrent que les informations bas niveau sont davantage encodées dans les premières couches de l'espace latent, tandis que celles relatives aux facteurs génératifs se retrouvent plutôt dans les couches les plus hautes. Une telle architecture aurait le potentiel de devenir un outil de prise de décision robuste pour manipuler la position des différents diagnostics, tout en fournissant une mesure d'incertitudes relative à leur position encodée, mais également aux bruits inhérents aux capteurs. Elle contribuerait ainsi à réduire les risques liés aux interventions humaines et notamment les potentielles collisions.

# Annexe A

# Astuce du ratio de densités

L'estimation de ratio de densité est couramment utilisée en apprentissage automatique lorsque deux ratios impliquent des intégrales difficilement calculables. Dans cet objectif, une méthode nécessitant un réseau de classification peut être envisagée. Cette annexe a pour objectif de démontrer l'estimation d'un ratio de densité à l'aide d'une telle approche et repose sur la démonstration présentée par [Sugiyama *et al.*, 2010]. Dans ce contexte, un réseau de classification binaire $D_\psi$ est entraîné pour distinguer des échantillons provenant des deux distributions.

Soit $D_p = \{\boldsymbol{x}_p^{(i)}\}_{i=1}^{n_p}$ et $D_q = \{\boldsymbol{x}_q^{(i)}\}_{i=1}^{n_q}$ des échantillons de deux distributions, respectivement $p(\boldsymbol{x})$ et $q(\boldsymbol{x})$. Dans ce contexte, le ratio $r = \frac{p(\boldsymbol{x})}{q(x)}$ peut être estimé par :

$$r(\boldsymbol{x}) \approx \hat{r}(\boldsymbol{x}) = \frac{D_{\boldsymbol{\psi}}(\boldsymbol{x})}{1 - D_{\boldsymbol{\psi}}(\boldsymbol{x})}, \tag{A.1}$$

où $D_\psi$ correspond au réseau de classification binaire entraîné pour classifier des données provenant de $p(\boldsymbol{x})$ ou $q(\boldsymbol{x})$.

*Preuve.* Assignons les labels $y = 1$ aux échantillons de $D_p$ et $y = 0$ aux autres. Cela permet d'écrire :

$$\begin{aligned} p(\boldsymbol{x}) &= p(\boldsymbol{x}|y = 1) \\ q(\boldsymbol{x}) &= p(\boldsymbol{x}|y = 0). \end{aligned} \tag{A.2}$$

Posons désormais $n = n_p + n_q$ formant le jeu de données $\{(\boldsymbol{x}_k, y_k)_{k=1}^n\}$ où :

$$\begin{aligned} (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) &= (\boldsymbol{x}_p^{(1)}, \ldots, \boldsymbol{x}_p^{n_p}, \boldsymbol{x}_q^{(1)}, \ldots, \boldsymbol{x}_q^{n_q}) \\ (y_1, \ldots, y_n) &= (1, \ldots, 1, 0, \ldots, 0). \end{aligned} \tag{A.3}$$

Appliquons désormais le théorème de Bayes

$$p(\boldsymbol{x}|y) = \frac{p(y|\boldsymbol{x})p(\boldsymbol{x})}{p(y)}, \tag{A.4}$$

nous permettant d'écrire de ratio des densités, dénoté $r(\boldsymbol{x})$, en termes de classes de probabilités :

$$
\begin{aligned}
r(\boldsymbol{x}) &= \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \\
&= \frac{p(\boldsymbol{x}|y=1)}{p(\boldsymbol{x}|y=0)} \\
&= \left[\frac{p(y=1|\boldsymbol{x})p(\boldsymbol{x})}{p(y=1)}\right]\left[\frac{p(y=0|x)p(\boldsymbol{x})}{p(y=0)}\right]^{-1} \\
&= \frac{p(y=0)}{p(y=1)}\frac{p(y=1|\boldsymbol{x})}{p(y=0|\boldsymbol{x})}.
\end{aligned}
\tag{A.5}
$$

Le ratio des densités *a priori* peut être estimé par le ratio de la taille des échantillons, de la manière suivante :

$$
\frac{p(y=0)}{p(y=1)} \approx \frac{n_q}{n_p+n_q}\frac{n_p+n_q}{n_p} = \frac{n_q}{n_p}. \tag{A.6}
$$

La distribution *a posteriori* $p(y|\boldsymbol{x})$ est estimée en entraînant un réseau de classification probabiliste binaire à reconnaître des échantillons venant de $D_p$ ou de $D_q$. Ainsi, un estimateur du ratio $\hat{r}(\boldsymbol{x})$ peut être construit à partir d'un estimateur de cette distribution *a posteriori*, noté $\hat{p}(y|\boldsymbol{x})$. Notons :

$$
\hat{r}(\boldsymbol{x}) = \frac{n_q}{n_p}\frac{\hat{p}(y=1|\boldsymbol{x})}{\hat{p}(y=0|\boldsymbol{x})} = \frac{n_q}{n_p}\frac{\hat{p}(y=1|\boldsymbol{x})}{1-\hat{p}(y=1|\boldsymbol{x})}. \tag{A.7}
$$

Dans un objectif de simplification, estimons que $n_p = n_q$, et notons $D(\boldsymbol{x}) = \hat{p}(y=1|\boldsymbol{x})$. On peut ainsi écrire :

$$
\begin{aligned}
\hat{r}(\boldsymbol{x}) &= \frac{D(\boldsymbol{x})}{1-D(\boldsymbol{x})} \\
&= \exp\left[\log\frac{D(\boldsymbol{x})}{1-D(\boldsymbol{x})}\right] \\
&= \exp(\sigma^{-1}(D(\boldsymbol{x}))),
\end{aligned}
\tag{A.8}
$$

où $\sigma$ représente la fonction sigmoïde. On constate que l'estimateur obtenu équivaut à la sortie d'un réseau de classification binaire. □

# Annexe B

# Distributions marginales de la loi Normale-Gamma

La démonstration ci-dessous, portant sur la distribution marginale d'une loi Normale-Gamma, est extraite du livre "The book of statistical proofs" [Soch *et al.*, 2024]. Considérons le couple de variables ($\boldsymbol{z}$, $\boldsymbol{\lambda}$), échantillonné depuis une loi Normale-Gamma comme suit :

$$(\boldsymbol{z}, \boldsymbol{\lambda}) \sim \mathcal{NG}(\boldsymbol{z}, \boldsymbol{\lambda}; \boldsymbol{\mu}, \alpha, \beta). \tag{B.1}$$

Notons $\boldsymbol{\mu}$ le paramètre de moyenne, $\alpha$ le paramètre d'intensité et $\beta$ le paramètre d'échelle. Dans ce contexte, la distribution marginale de l'inverse-variance, $\boldsymbol{\lambda}$, est une distribution Gamma, telle que :

$$\boldsymbol{\lambda} \sim \mathcal{G}(\boldsymbol{\lambda}; \alpha, \beta). \tag{B.2}$$

La distribution marginale de la variable $\boldsymbol{z}$, quant à elle, est une loi de Student multipliée par le facteur $\sqrt{\frac{\alpha}{\beta}}$. Elle est définie de la façon suivante :

$$\boldsymbol{z} \sim \sqrt{\frac{\alpha}{\beta}} T\left(\frac{\boldsymbol{z} - \boldsymbol{\mu}}{\sqrt{\frac{\beta}{\alpha}}}; 2\alpha\right). \tag{B.3}$$

*Preuve.* La réécriture de la fonction de densité de la loi Normale-Gamma peut être présentée de la manière suivante :

$$\begin{aligned} p(\boldsymbol{z}, \boldsymbol{\lambda}) &= p(\boldsymbol{z}|\boldsymbol{\lambda}) p(\boldsymbol{\lambda}) \\ p(\boldsymbol{z}|\boldsymbol{\lambda}) &= \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}, \boldsymbol{\lambda}^{-1}) \\ p(\boldsymbol{\lambda}) &= \mathcal{G}(\boldsymbol{\lambda}; \alpha, \beta). \end{aligned} \tag{B.4}$$

En appliquant la loi de la probabilité totale avec cette formulation, il est possible de dériver la distribution marginale de $\boldsymbol{\lambda}$ comme suit :

$$\begin{aligned} p(\boldsymbol{\lambda}) &= \int p(\boldsymbol{z}, \boldsymbol{\lambda}) d\boldsymbol{z} \\ &= \int \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}, \boldsymbol{\lambda}^{-1}) \mathcal{G}(\boldsymbol{\lambda}; \alpha, \beta) d\boldsymbol{z} \\ &= \mathcal{G}(\boldsymbol{\lambda}; \alpha, \beta) \int \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}, \boldsymbol{\lambda}^{-1}) d\boldsymbol{z} \\ &= \mathcal{G}(\boldsymbol{\lambda}; \alpha, \beta). \end{aligned} \tag{B.5}$$

Ce résultat correspond à la fonction de densité de la distribution Gamma, caractérisée par un paramètre d'intensité $\alpha$ et un paramètre d'échelle $\beta$.

De la même manière, en appliquant la loi de la probabilité totale, la distribution marginale de $\boldsymbol{z}$ est obtenue de la façon suivante :

$$
\begin{aligned}
p(\boldsymbol{z}) &= \int p(\boldsymbol{z}, \boldsymbol{\lambda}) d\boldsymbol{\lambda} \\
&= \int \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}, \boldsymbol{\lambda}^{-1}) \mathcal{G}(\boldsymbol{\lambda}; \alpha, \beta) d\boldsymbol{\lambda} \\
&= \int \sqrt{\frac{\boldsymbol{\lambda}}{2\pi}} \exp\left[-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^2 \boldsymbol{\lambda}\right] \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \boldsymbol{\lambda}^{(\alpha-1)} \exp[-\beta\boldsymbol{\lambda}] d\boldsymbol{\lambda} \\
&= \int \sqrt{\frac{1}{2\pi}} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \boldsymbol{\lambda}^{(\alpha-\frac{1}{2})} \cdot \exp\left[-\left(\beta + \frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^2 \boldsymbol{\lambda}\right)\right] d\boldsymbol{\lambda} \\
&= \int \sqrt{\frac{1}{2\pi}} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha+\frac{1}{2})}{(\beta + \frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^2)^{(\alpha+\frac{1}{2})}} \cdot \mathcal{G}\left(\boldsymbol{\lambda}; \alpha + \frac{1}{2}, \beta + \frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^2\right) d\boldsymbol{\lambda} \\
&= \sqrt{\frac{1}{2\pi}} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha+\frac{1}{2})}{(\beta + \frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^2)^{(\alpha+\frac{1}{2})}} \cdot \int \mathcal{G}\left(\boldsymbol{\lambda}; \alpha + \frac{1}{2}, \beta + \frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^2\right) d\boldsymbol{\lambda} \\
&= \sqrt{\frac{1}{2\pi}} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha+\frac{1}{2})}{(\beta + \frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^2)^{(\alpha+\frac{1}{2})}} \\
&= \frac{1}{\pi^{\frac{1}{2}}} \cdot \frac{\Gamma(\alpha+\frac{1}{2})}{\Gamma(\alpha)} \cdot \beta^\alpha \cdot \left(\beta + \frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^2\right)^{-\left(\alpha+\frac{1}{2}\right)} \\
&= \frac{1}{\pi^{\frac{1}{2}}} \cdot \frac{\Gamma(\alpha+\frac{1}{2})}{\Gamma(\alpha)} \cdot \left(\frac{1}{\beta}\right)^{-\alpha} \cdot \left(\beta + \frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^2\right)^{-\alpha} \cdot 2^{-\frac{1}{2}} \cdot \left(\beta + \frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^2\right)^{-\frac{1}{2}} \\
&= \frac{1}{\pi^{\frac{1}{2}}} \cdot \frac{\Gamma(\alpha+\frac{1}{2})}{\Gamma(\alpha)} \cdot \left(1 + \frac{1}{2\beta}(\boldsymbol{z}-\boldsymbol{\mu})^2\right)^{-\alpha} \cdot \left(2\beta + (\boldsymbol{z}-\boldsymbol{\mu})^2\right)^{-\frac{1}{2}} \\
&= \frac{1}{\pi^{\frac{1}{2}}} \cdot \frac{\Gamma(\alpha+\frac{1}{2})}{\Gamma(\alpha)} \cdot \left(\frac{1}{2\alpha}\right)^{-\alpha} \cdot \left(2\alpha + \frac{\alpha}{\beta}(\boldsymbol{z}-\boldsymbol{\mu})^2\right)^{-\alpha} \cdot \left(\frac{\beta}{\alpha}\right)^{-\frac{1}{2}} \cdot \left(2\alpha + \frac{\alpha}{\beta}(\boldsymbol{z}-\boldsymbol{\mu})^2\right)^{-\frac{1}{2}} \\
&= \frac{\sqrt{\frac{\alpha}{\beta}}}{(2\alpha)^{-\alpha}\pi^{\frac{1}{2}}} \cdot \frac{\Gamma(\alpha+\frac{1}{2})}{\Gamma(\alpha)} \cdot \left(2\alpha + \frac{\alpha}{\beta}(\boldsymbol{z}-\boldsymbol{\mu})^2\right)^{-\alpha} \cdot \left(2\alpha + \frac{\alpha}{\beta}(\boldsymbol{z}-\boldsymbol{\mu})^2\right)^{-\frac{1}{2}} \\
&= \frac{\sqrt{\frac{\alpha}{\beta}}}{(2\alpha)^{-\alpha}\pi^{\frac{1}{2}}} \cdot \frac{\Gamma(\alpha+\frac{1}{2})}{\Gamma(\alpha)} \cdot (2\alpha)^{-\alpha} \cdot \left(1 + \frac{1}{2\beta}(\boldsymbol{z}-\boldsymbol{\mu})^2\right)^{-\alpha} \cdot (2\alpha)^{(-\frac{1}{2})} \cdot \left(1 + \frac{1}{2\beta}(\boldsymbol{z}-\boldsymbol{\mu})^2\right)^{-\frac{1}{2}} \\
&= \sqrt{\frac{\frac{\alpha}{\beta}}{(2\alpha)^{\frac{1}{2}}\pi^{\frac{1}{2}}}} \cdot \frac{\Gamma(\alpha+\frac{1}{2})}{\Gamma(\alpha)} \cdot \left(1 + \frac{(\boldsymbol{z}-\boldsymbol{\mu})^2}{2\alpha \cdot \left(\sqrt{\frac{\beta}{\alpha}}\right)^2}\right)^{-(\alpha+\frac{1}{2})} \\
&= \sqrt{\frac{\frac{\alpha}{\beta}}{2\alpha\pi}} \cdot \frac{\Gamma(\alpha+\frac{1}{2})}{\Gamma(\alpha)} \cdot \left(1 + \frac{(\boldsymbol{z}-\boldsymbol{\mu})^2}{2\alpha \cdot \left(\sqrt{\frac{\beta}{\alpha}}\right)^2}\right)^{-(\alpha+\frac{1}{2})} \\
&= \sqrt{\frac{\alpha}{\beta}} \mathcal{T}\left(\frac{\boldsymbol{z}-\boldsymbol{\mu}}{\sqrt{\frac{\beta}{\alpha}}}; 2\alpha\right).
\end{aligned}
\tag{B.6}
$$

Dans cette équation, $\mathcal{T}(\cdot; 2\alpha)$ représente une distribution de Student de $2\alpha$ degrés de liberté, et $\Gamma(\cdot)$ la fonction gamma. □

# Annexe C

# Divergence de Kullback entre deux lois Normales-Gamma

La démonstration qui suit, concernant la divergence de Kullback-Leibler entre deux lois Normales-Gammas, est issue du livre "The book of statistical proofs" [Soch *et al.*, 2024]. Considérons les deux distributions Normales multivariées suivantes :

$$\begin{aligned} &p(\boldsymbol{x}; \boldsymbol{\mu}_1, \Sigma_1) \\ &q(\boldsymbol{x}; \boldsymbol{\mu}_2, \Sigma_2), \end{aligned} \tag{C.1}$$

de moyennes respectives $\boldsymbol{\mu}_1$ et $\boldsymbol{\mu}_2$ et de matrices de covariance $\Sigma_1$ et $\Sigma_2$. Dans ce contexte, la divergence de Kullback-Leibler entre $p$ et $q$ s'exprime de la façon suivante :

$$D_{KL}[p(\boldsymbol{x}; \boldsymbol{\mu}_1, \Sigma_1)||q(\boldsymbol{x}; \boldsymbol{\mu}_2, \Sigma_2)] = \frac{1}{2}\left[(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T \Sigma_2^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + tr(\Sigma_2^{-1}\Sigma_1) - \log\frac{|\Sigma_1|}{|\Sigma_2|} - n\right], \tag{C.2}$$

où $tr$ correspond à l'opérateur de trace, et $n$ la dimension des distributions.

*Preuve :* D'après les définitions de la divergence de Kullback-Leibler et de la densité de probabilité d'une loi Normale, on peut écrire :

$$\begin{aligned} &D_{KL}[p(\boldsymbol{x}; \boldsymbol{\mu}_1, \Sigma_1)||q(\boldsymbol{x}; \boldsymbol{\mu}_2, \Sigma_2)] \\ &= \mathbb{E}_{p(\boldsymbol{x})}\left[\log\left(\frac{\frac{1}{\sqrt{(2\pi)^n|\Sigma_1|}} \cdot \exp[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_1)^T\Sigma_1^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1)]}{\frac{1}{\sqrt{(2\pi)^n|\Sigma_2|}} \cdot \exp[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_2)^T\Sigma_2^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2)]}\right)\right] \\ &= \frac{1}{2}\mathbb{E}_{p(\boldsymbol{x})}\left[\log\frac{|\Sigma_2|}{|\Sigma_1|} - (\boldsymbol{x} - \boldsymbol{\mu}_1)^T\Sigma_1^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1) + (\boldsymbol{x} - \boldsymbol{\mu}_2)^T\Sigma_2^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2)\right]. \end{aligned} \tag{C.3}$$

L'application des propriétés de la trace et de l'espérance nous permet d'obtenir le résultat suivant :

$$\begin{aligned} &D_{KL}[p(\boldsymbol{x}; \boldsymbol{\mu}_1, \Sigma_1)||q(\boldsymbol{x}; \boldsymbol{\mu}_2, \Sigma_2)] \\ &= \frac{1}{2}\left[\log\frac{|\Sigma_2|}{|\Sigma_1|} - tr\left(\Sigma_1^{-1}\mathbb{E}_{p(\boldsymbol{x})}\left[(\boldsymbol{x} - \boldsymbol{\mu}_1)(\boldsymbol{x} - \boldsymbol{\mu}_1)^T\right]\right) + tr\left(\Sigma_2^{-1}\mathbb{E}_{p(\boldsymbol{x})}\left[(\boldsymbol{x} - \boldsymbol{\mu}_2)(\boldsymbol{x} - \boldsymbol{\mu}_2)^T\right]\right)\right] \\ &= \frac{1}{2}\left[\log\frac{|\Sigma_2|}{|\Sigma_1|} - tr\left(\Sigma_1^{-1}\mathbb{E}_{p(\boldsymbol{x})}\left[(\boldsymbol{x} - \boldsymbol{\mu}_1)(\boldsymbol{x} - \boldsymbol{\mu}_1)^T\right]\right) + tr\left(\Sigma_2^{-1}\mathbb{E}_{p(\boldsymbol{x})}\left[\boldsymbol{x}\boldsymbol{x}^T - 2\boldsymbol{\mu}_2\boldsymbol{x}^T + \boldsymbol{\mu}_2\boldsymbol{\mu}_2^T\right]\right)\right] \\ &= \frac{1}{2}\Bigg[\log\frac{|\Sigma_2|}{|\Sigma_1|} - tr\left(\Sigma_1^{-1}\mathbb{E}_{p(\boldsymbol{x})}\left[(\boldsymbol{x} - \boldsymbol{\mu}_1)(\boldsymbol{x} - \boldsymbol{\mu}_1)^T\right]\right) + tr\Bigg(\Sigma_2^{-1}\Big(\mathbb{E}_{p(\boldsymbol{x})}[\boldsymbol{x}\boldsymbol{x}^T] - \mathbb{E}_{p(\boldsymbol{x})}[2\boldsymbol{\mu}_2\boldsymbol{x}^T] \\ &+ \mathbb{E}_{p(\boldsymbol{x})}[\boldsymbol{\mu}_2\boldsymbol{\mu}_2^T]\Big)\Bigg)\Bigg]. \end{aligned} \tag{C.4}$$

Par ailleurs, l'espérance de la forme linéaire d'une loi Normale multivariée et l'espérance de sa forme quadratique, définies par :

$$\begin{aligned} \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma) &\Rightarrow \mathbb{E}_{p(\boldsymbol{x})}[A\boldsymbol{x}] = A\boldsymbol{\mu} \\ &\Rightarrow \mathbb{E}_{p(\boldsymbol{x})}[\boldsymbol{x}^T A\boldsymbol{x}] = \boldsymbol{\mu}^T A\boldsymbol{\mu} + tr(A\Sigma), \end{aligned} \tag{C.5}$$

permettent de réécrire le terme de divergence de Kullback-Leibler de l'équation (C.4) de la manière suivante :

$$\begin{aligned} &D_{KL}[p(\boldsymbol{x}; \boldsymbol{\mu}_1, \Sigma_1)||q(\boldsymbol{x}; \boldsymbol{\mu}_2, \Sigma_2)] \\ &= \frac{1}{2}\left[\log\frac{|\Sigma_2|}{|\Sigma_1|} - tr(\Sigma_1^{-1}\Sigma_1) + tr\Big(\Sigma_2^{-1}(\Sigma_1 + \boldsymbol{\mu}_1\boldsymbol{\mu}_1^T - 2\boldsymbol{\mu_2}\boldsymbol{\mu}_1^T + \boldsymbol{\mu}_2\boldsymbol{\mu}_2^T)\Big)\right] \\ &= \frac{1}{2}\left[\log\frac{|\Sigma_2|}{|\Sigma_1|} - tr(\mathbb{I}_n) + tr(\Sigma_2^{-1}\Sigma_1) + tr\Big(\Sigma_2^{-1}\left(\boldsymbol{\mu}_1\boldsymbol{\mu}_1^T - 2\boldsymbol{\mu_2}\boldsymbol{\mu}_1^T + \boldsymbol{\mu}_2\boldsymbol{\mu}_2^T\right)\Big)\right] \\ &= \frac{1}{2}\left[\log\frac{|\Sigma_2|}{|\Sigma_1|} - n + tr(\Sigma_2^{-1}\Sigma_1) + tr\left(\boldsymbol{\mu}_1^T\Sigma_2^{-1}\boldsymbol{\mu}_1 - 2\boldsymbol{\mu_1}^T\Sigma_2^{-1}\boldsymbol{\mu}_2 + \boldsymbol{\mu}_2^T\Sigma_2^{-1}\boldsymbol{\mu}_2\right)\right] \\ &= \frac{1}{2}\left[\log\frac{|\Sigma_2|}{|\Sigma_1|} - n + tr(\Sigma_2^{-1}\Sigma_1) + (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T\Sigma_2^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\right]. \end{aligned} \tag{C.6}$$

En réarrangeant les termes, on obtient finalement :

$$D_{KL}[p(\boldsymbol{x}; \boldsymbol{\mu}_1, \Sigma_1)||q(\boldsymbol{x}; \boldsymbol{\mu}_2, \Sigma_2)] = \frac{1}{2}\left[(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T\Sigma_2^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + tr(\Sigma_2^{-1}\Sigma_1) - \log\frac{|\Sigma_1|}{|\Sigma_2|} - n\right]. \tag{C.7}$$

□

Considérons désormais deux distributions Dammas univariées, définies par :

$$\begin{aligned} &p(\boldsymbol{x}; \alpha_1, \beta_1) \\ &q(\boldsymbol{x}; \alpha_2, \beta_2), \end{aligned} \tag{C.8}$$

dans lesquelles $\alpha_1$ et $\alpha_2$ représentent respectivement les paramètres d'intensité de $p$ et $q$, et $\beta_1$ et $\beta_2$ leurs paramètres respectifs d'échelle. Dans ce contexte, la divergence de Kullback-Leibler entre $p$ et $q$ s'écrit de la façon suivante :

$$D_{KL}[p(\boldsymbol{x}; \alpha_1, \beta_1)||q(\boldsymbol{x}; \alpha_2, \beta_2)] = \alpha_2\log\frac{\beta_1}{\beta_2} - \log\frac{\Gamma(\alpha_1)}{\Gamma(\alpha_2)} + (\alpha_1 - \alpha_2)\psi(\alpha_2) - (\beta_1 - \beta_2)\frac{\alpha_1}{\beta_1}, \tag{C.9}$$

où $\Gamma(\cdot)$ correspond à la fonction gamma, et $\psi(\cdot)$ à la fonction digamma.

*Preuve.* D'après les définitions de la divergence de Kullback-Leibler et de la densité de probabilité d'une loi Gamma, on peut écrire :

$$\begin{aligned} &D_{KL}[p(\boldsymbol{x}; \alpha_1, \beta_1)||q(\boldsymbol{x}; \alpha_2, \beta_2)] \\ &= \mathbb{E}_{p(\boldsymbol{x})}\left[\log\frac{\frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1)}x^{(\alpha_1-1)}\exp[-\beta_1\boldsymbol{x}]}{\frac{\beta_2^{\alpha_2}}{\Gamma(\alpha_2)}x^{(\alpha_2-1)}\exp[-\beta_2\boldsymbol{x}]}\right] \\ &= \mathbb{E}_{p(\boldsymbol{x})}\left[\log\left(\frac{\beta_1^{\alpha_1}}{\beta_2^{\alpha_2}}\cdot\frac{\Gamma(\alpha_2)}{\Gamma(\alpha_1)}\cdot\boldsymbol{x}^{(\alpha_1-\alpha_2)}\exp[-(\beta_1 - \beta_2)\boldsymbol{x}]\right)\right] \\ &= \mathbb{E}_{p(\boldsymbol{x})}\left[\alpha_1\log\beta_1 - \alpha_2\log\beta_2 - \log\Gamma(\alpha_1) + \log\Gamma(\alpha_2) + (\alpha_1 - \alpha_2)\log\boldsymbol{x} - (\beta_1 - \beta_2)\boldsymbol{x}\right]. \end{aligned} \tag{C.10}$$

Par ailleurs, la définition de la moyenne d'une loi Gamma, et de l'espérance du logarithme de cette loi, définies par :

$$\begin{aligned} \boldsymbol{x} \sim \mathcal{G}(\boldsymbol{x}; \alpha, \beta) &\Rightarrow \mathbb{E}_{p(\boldsymbol{x})}[\boldsymbol{x}] = \frac{\alpha}{\beta} \\ &\Rightarrow \mathbb{E}_{p(\boldsymbol{x})}[\log(\boldsymbol{x})] = \psi(\alpha) - \log \beta, \end{aligned} \tag{C.11}$$

permettent de réécrire le terme de divergence de Kullback-Leibler, défini dans l'équation (C.10), de la façon suivante :

$$\begin{aligned} &D_{KL}[p(\boldsymbol{x}; \alpha_1, \beta_1)||q(\boldsymbol{x}; \alpha_2, \beta_2)] \\ &= \alpha_1 \log \beta_1 - \alpha_2 \log \beta_2 - \log \Gamma(\alpha_1) + \log \Gamma(\alpha_2) + (\alpha_1 - \alpha_2)\left(\psi(\alpha_1) - \log \beta_1\right) - (\beta_1 - \beta_1)\frac{\alpha_1}{\beta_1} \\ &= \alpha_2 \log \beta_1 - \alpha_2 \log \beta_2 - \log \Gamma(\alpha_1) + \log \Gamma(\alpha_2) + (\alpha_1 - \alpha_2)\psi(\alpha_1) - (\beta_1 - \beta_2)\frac{\alpha_1}{\beta_1}. \end{aligned} \tag{C.12}$$

En combinant les logarithmes, on obtient finalement :

$$D_{KL}[p(\boldsymbol{x}; \alpha_1, \beta_1)||q(\boldsymbol{x}; \alpha_2, \beta_2)] = \alpha_2 \log \frac{\beta_1}{\beta_2} - \log \frac{\Gamma(\alpha_1)}{\Gamma(\alpha_2)} + (\alpha_1 - \alpha_2)\psi(\alpha_2) - (\beta_1 - \beta_2)\frac{\alpha_1}{\beta_1}. \tag{C.13}$$

Dans cette équation, $\Gamma(\cdot)$ correspond à la fonction gamma, et $\psi(\cdot)$ à la fonction digamma. □

Considérons maintenant les distributions multivariées normales-gamma suivantes :

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{y}) &= p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_1, (y\Lambda_1)^{-1})\mathcal{G}(\boldsymbol{y}; \alpha_1, \beta_1) \\ q(\boldsymbol{x}, \boldsymbol{y}) &= q(\boldsymbol{x}|\boldsymbol{y})q(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_2, (y\Lambda_2)^{-1})\mathcal{G}(\boldsymbol{y}; \alpha_2, \beta_2). \end{aligned} \tag{C.14}$$

Dans ce contexte, la divergence de Kullback-Leibler entre $p$ et $q$ s'écrit telle que :

$$\begin{aligned} D_{KL}[p(\boldsymbol{x}, \boldsymbol{y})||q(\boldsymbol{x}, \boldsymbol{y})] = &\frac{1}{2}\frac{\alpha_1}{\beta_1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu_1})^T \Lambda_2 (\boldsymbol{\mu_2} - \boldsymbol{\mu_1}) + \frac{1}{2} tr(\Lambda_2 \Lambda_1^{-1}) - \frac{1}{2}\log \frac{|\Sigma_1|}{|\Sigma_1|} - \frac{n}{2} \\ &+ \alpha_2 \log \frac{\beta_1}{\beta_2} - \log \frac{\Gamma(\alpha_1)}{\Gamma(\alpha_2)} + (\alpha_1 - \alpha_2)\psi(\alpha_1) - (\beta_1 - \beta_2)\frac{\alpha_1}{\beta_1}. \end{aligned} \tag{C.15}$$

*Preuve.* En exploitant la loi de la probabilité conditionnelle, la divergence de Kullback-Leibler entre les distributions Normales-Gamma peut être calculée comme suit :

$$\begin{aligned} D_{KL}[p(\boldsymbol{x}, \boldsymbol{y})||q(\boldsymbol{x}, \boldsymbol{y})] &= \int\int p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y}) \log \frac{p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y})}{q(\boldsymbol{x}|\boldsymbol{y})q(\boldsymbol{y})} d\boldsymbol{x}d\boldsymbol{y} \\ &= \int\int p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y}) \log \frac{p(\boldsymbol{x}|\boldsymbol{y})}{q(\boldsymbol{x}|\boldsymbol{y})} d\boldsymbol{x}d\boldsymbol{y} + \int\int p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y}) \log \frac{p(\boldsymbol{y})}{q(\boldsymbol{y})} d\boldsymbol{x}d\boldsymbol{y} \\ &= \int p(\boldsymbol{y}) \int p(\boldsymbol{x}|\boldsymbol{y}) \log \frac{p(\boldsymbol{x}|\boldsymbol{y})}{q(\boldsymbol{x}|\boldsymbol{y})} d\boldsymbol{x}d\boldsymbol{y} + \int p(\boldsymbol{y}) \log \frac{p(\boldsymbol{y})}{q(\boldsymbol{y})} \int p(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{x}d\boldsymbol{y} \\ &= \mathbb{E}_{p(\boldsymbol{y})}[D_{KL}[p(\boldsymbol{x}|\boldsymbol{y})||q(\boldsymbol{x}|\boldsymbol{y})]] + D_{KL}[p(\boldsymbol{y})||q(\boldsymbol{y})]. \end{aligned} \tag{C.16}$$

Autrement dit, la divergence de Kullback-Leibler entre deux distributions Normales-Gamma est la somme de deux composantes : la divergence de Kullback-Leibler multivariée gaussienne de $\boldsymbol{x}$ conditionnée par $\boldsymbol{y}$, avec $\boldsymbol{y}$ pris sous son espérance, et la divergence de

Kullback-Leibler univariée Gamma pour $\boldsymbol{y}$.

En réécrivant le premier terme de l'équation (C.16) par la définition de la divergence de Kullback-Leibler entre deux distributions Gammas définie dans l'équation (C.2), on obtient :

$$\begin{aligned}
&\mathbb{E}_{p(\boldsymbol{y})}[D_{KL}[p(\boldsymbol{x}|\boldsymbol{y})||q(\boldsymbol{x}|\boldsymbol{y})]] \\
&= \mathbb{E}_{p(\boldsymbol{y})}\left[\frac{1}{2}\left((\boldsymbol{\mu}_1-\boldsymbol{\mu}_1)^T\Sigma_2^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)+tr(\Sigma_2^{-1}\Sigma_1)-\log\frac{|\Sigma_1|}{|\Sigma_2|}-n\right)\right] \\
&= \mathbb{E}_{p(\boldsymbol{y})}\left[\frac{\boldsymbol{y}}{2}(\boldsymbol{\mu_2}-\boldsymbol{\mu}_1)^T\Lambda_2(\boldsymbol{\mu_2}-\boldsymbol{\mu_1})+\frac{1}{2}tr(\Lambda_2\Lambda_1^{-1})-\frac{1}{2}\log\frac{|\Lambda_2|}{|\Lambda_1|}-\frac{n}{2}\right].
\end{aligned} \tag{C.17}$$

De plus, l'espérance d'une variable suivant une loi Gamma, définie par :

$$\boldsymbol{y}\sim\mathcal{G}(\boldsymbol{y};\alpha,\beta)\Rightarrow\mathbb{E}_{p(\boldsymbol{y})}[\boldsymbol{y}]=\frac{\alpha}{\beta}, \tag{C.18}$$

permet de réécrire l'équation (C.17) de la façon suivante :

$$\mathbb{E}_{p(\boldsymbol{y})}[D_{KL}[p(\boldsymbol{x}|\boldsymbol{y})||q(\boldsymbol{x}|\boldsymbol{y})]]=\frac{1}{2}\frac{\alpha_1}{\beta_1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^T\Lambda_2(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)+\frac{1}{2}tr(\Lambda_2\Lambda_1^{-1})-\frac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|}-\frac{n}{2}. \tag{C.19}$$

Finalement, en remplaçant les premier et second terme de l'équation (C.16), respectivement par les résultats obtenus dans les équations (C.19) et (C.9), la divergence de Kullback-Leibler entre deux distributions Normales-Gamma s'écrit telle que :

$$\begin{aligned}
D_{KL}[p(\boldsymbol{x},\boldsymbol{y})||q(\boldsymbol{x},\boldsymbol{y})] = &\frac{1}{2}\frac{\alpha_1}{\beta_1}(\boldsymbol{\mu}_2-\boldsymbol{\mu_1})^T\Lambda_2(\boldsymbol{\mu_2}-\boldsymbol{\mu_1})+\frac{1}{2}tr(\Lambda_2\Lambda_1^{-1})-\frac{1}{2}\log\frac{|\Sigma_1|}{|\Sigma_1|}-\frac{n}{2} \\
&+\alpha_2\log\frac{\beta_1}{\beta_2}-\log\frac{\Gamma(\alpha_1)}{\Gamma(\alpha_2)}+(\alpha_1-\alpha_2)\psi(\alpha_1)-(\beta_1-\beta_2)\frac{\alpha_1}{\beta_1}.
\end{aligned} \tag{C.20}$$

□

# Annexe D

# Estimateurs stochastiques

Considérons le couple $(\boldsymbol{z}, \boldsymbol{\lambda})$ suivant une distribution Normale-Gamma notée $p(\boldsymbol{z}, \boldsymbol{\lambda})$, tel que :

$$(\boldsymbol{z}, \boldsymbol{\lambda}) \sim \mathcal{NG}(\boldsymbol{z}, \boldsymbol{\lambda}; \boldsymbol{\mu}, \alpha, \beta). \tag{D.1}$$

Notons $\boldsymbol{\mu}$ le paramètre de moyenne, $\alpha$ le paramètre d'intensité et $\beta$ le paramètre d'échelle. Dans ce contexte, la mesure de corrélation totale $TC(\boldsymbol{z}, \boldsymbol{\lambda})$ peut être calculée de la façon suivante :

$$TC(\boldsymbol{z}, \boldsymbol{\lambda}) = D_{KL}\left[p(\boldsymbol{z}, \boldsymbol{\lambda}) || \prod_{k=1}^{K} p(z_k, \lambda_k)\right] = \mathbb{E}_{p(\boldsymbol{z}, \boldsymbol{\lambda})}\left[\log \frac{p(\boldsymbol{z}, \boldsymbol{\lambda})}{\prod_{k=1}^{K} p(z_k, \lambda_k)}\right], \tag{D.2}$$

où $K$ correspond au nombre de composantes de la distribution considérée. Toutefois, $p(\boldsymbol{z}, \boldsymbol{\lambda})$ et $\prod_k p(z_k, \lambda_k)$ sont difficilement calculables dès lors que $K$ prend une forte valeur. Pour remédier à ce problème, les auteurs Chen et al. proposent d'estimer ces grandeurs dans le cadre d'une distribution Normale, à l'aide de deux estimateurs stochastiques distincts ; le "Minibatch Weighted Sampling", ou le "Minibatch Stratified Sampling" [Chen *et al.*, 2018]. Ces derniers peuvent être étendus, pour le contexte étudié, à une distribution Normale-Gamma.

## D.1 Estimateur "Minibatch Weighted Sampling"

Dans le contexte pour lequel le couple $(\boldsymbol{z}, \boldsymbol{\lambda})$ correspond aux sorties du réseau encodeur d'un VAE hiérarchique, alors la grandeur $p(\boldsymbol{z}, \boldsymbol{\lambda})$ présente dans l'équation (D.2) peut être estimée à l'aide du "Minibatch Weighted Sampling". Ce dernier est défini par :

$$\mathbb{E}_{p(\boldsymbol{z}, \boldsymbol{\lambda})}[\log p(\boldsymbol{z}, \boldsymbol{\lambda})] \approx \frac{1}{M} \sum_{i=1}^{M} \left[\log \sum_{j=1}^{M} p\left(\boldsymbol{z}(\boldsymbol{x}_i), \boldsymbol{\lambda}(\boldsymbol{x}_i) | \boldsymbol{x}_j\right) - \log(NM)\right]. \tag{D.3}$$

Pour cette équation, un ensemble de $N$ images, défini par $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, est considéré comme ensemble d'entraînement. $M$ correspond au nombre d'images utilisées comme sous-ensemble de données pour chaque itération. Finalement, $(\boldsymbol{z}(\boldsymbol{x}_i), \boldsymbol{\lambda}(\boldsymbol{x}_i))$ correspondent aux sorties de l'encodeur qui prend l'image $\boldsymbol{x}_i$ en entrée, tel que :

$$(\boldsymbol{z}(\boldsymbol{x}_i), \boldsymbol{\lambda}, \boldsymbol{x}_i) \sim p(\boldsymbol{z}, \boldsymbol{\lambda} | \boldsymbol{x}_i). \tag{D.4}$$

*Preuve.* Considérons le sous-ensemble $\mathcal{B}_M = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ constitué par $M$ images identiquement et indépendamment échantillonnées depuis une distribution uniforme $p(\boldsymbol{x})$, composée par $N$ éléments. Dans ce contexte, $p(\mathcal{B}_M)$ est défini par $\left(\frac{1}{N}\right)^M$. Dénotons également $r(\mathcal{B}_M | \boldsymbol{x}^*)$ la probabilité d'un sous-ensemble échantillonné, dans lequel un élément est fixé à $\boldsymbol{x}^*$ et les autres sont échantillonnés identiquement et indépendamment

depuis $p(\boldsymbol{x})$. Ainsi, $r(\mathcal{B}_M|\boldsymbol{x}^*) = \left(\frac{1}{N}\right)^{M-1}$. Ces considérations permettent d'obtenir les développements suivants :

$$\begin{aligned}
\mathbb{E}_{p(\boldsymbol{z},\boldsymbol{\lambda})}[\log p(\boldsymbol{z},\boldsymbol{\lambda})] &= \mathbb{E}_{p(\boldsymbol{z},\boldsymbol{\lambda},\boldsymbol{x})}\left[\log \mathbb{E}_{p(\mathcal{B}_M)}\left[\frac{1}{M}\sum_{m=1}^{M} p(\boldsymbol{z},\boldsymbol{\lambda}|\boldsymbol{x}_m)\right]\right] \\
&\geq \mathbb{E}_{p(\boldsymbol{z},\boldsymbol{\lambda},\boldsymbol{x})}\left[\log \mathbb{E}_{r(\mathcal{B}_M|\boldsymbol{x}^*)}\left[\frac{p(\mathcal{B}_M)}{r(\mathcal{B}_M|\boldsymbol{x}^*)}\frac{1}{M}\sum_{m=1}^{M} p(\boldsymbol{z},\boldsymbol{\lambda}|\boldsymbol{x}_m)\right]\right] \\
&= \mathbb{E}_{p(\boldsymbol{z},\boldsymbol{\lambda},\boldsymbol{x})}\left[\log \mathbb{E}_{r(\mathcal{B}_M|\boldsymbol{x}^*)}\left[\frac{1}{NM}\sum_{m=1}^{M} p(\boldsymbol{z},\boldsymbol{\lambda}|\boldsymbol{x}_m)\right]\right].
\end{aligned} \tag{D.5}$$

L'inégalité provient du fait que le support de $r$ est un sous-ensemble de celui de $p$. Dans ce contexte, lorsque $M$ échantillons sont utilisés pour définir le sous-ensemble d'entraînement, l'estimateur prend la forme suivante :

$$\mathbb{E}_{p(\boldsymbol{z},\boldsymbol{\lambda})}[\log p(\boldsymbol{z},\boldsymbol{\lambda})] \approx \frac{1}{M}\sum_{i=1}^{M}\left[\log \sum_{j=1}^{M} p\left(\boldsymbol{z}(\boldsymbol{x}_i),\boldsymbol{\lambda}(\boldsymbol{x}_i)|\boldsymbol{x}_j\right) - \log(NM)\right]. \tag{D.6}$$

□

Une dérivation similaire peut être considérée pour estimer $\prod_k p(z_k,\lambda_k)$. Toutefois, il est important de noter que cet estimateur est biaisé. En effet, l'inégalité de Jensen stipule que $\mathbb{E}_{p(\boldsymbol{x})}[\log p(\boldsymbol{z},\boldsymbol{\lambda}|\boldsymbol{x})] \leq \log \mathbb{E}_{p(\boldsymbol{x})}[p(\boldsymbol{z},\boldsymbol{\lambda}|\boldsymbol{x})]$. Pour remédier à cette limitation, l'estimateur non biaisé "Minibatch Stratified Sampling" peut être considéré pour mesurer $TC(\boldsymbol{z},\boldsymbol{\lambda})$.

## D.2 Estimateur "Minibatch Stratified Sampling"

Afin d'éviter les limitations rencontrées par "Minibatch Weighted Sampling", la grandeur $p(\boldsymbol{z},\boldsymbol{\lambda})$ peut être calculée à l'aide d'un second estimateur, le "Minibatch Stratified Sampling". Ce dernier, noté $f\Big((\boldsymbol{z},\boldsymbol{\lambda}),\boldsymbol{x}^*,\hat{\mathcal{B}}_M\Big)$, est défini de la façon suivante :

$$f\Big((\boldsymbol{z},\boldsymbol{\lambda}),\boldsymbol{x}^*,\hat{\mathcal{B}}_M\Big) = \frac{1}{N}p(\boldsymbol{z},\boldsymbol{\lambda}|\boldsymbol{x}^*) + \frac{1}{M}\sum_{m=1}^{M-1} p(\boldsymbol{z},\boldsymbol{\lambda}|\boldsymbol{x}_m) + \frac{N-M}{NM}p(\boldsymbol{z},\boldsymbol{\lambda}|\boldsymbol{x}_M). \tag{D.7}$$

Cet estimateur découle d'un développement de $p(\boldsymbol{z},\boldsymbol{\lambda})$, en utilisant l'image $\boldsymbol{x}^*$ pour déterminer les valeurs de $(\boldsymbol{z},\boldsymbol{\lambda})$ de sorte que $(\boldsymbol{z},\boldsymbol{\lambda})$ suit la distribution $p(\boldsymbol{z},\boldsymbol{\lambda};\boldsymbol{x}^*)$. Ici, $N$ représente le nombre total d'observations dans le jeu de données, tandis que $M$ désigne le nombre d'images sélectionnées pour le sous-ensemble utilisé à chaque itération. L'objectif de la démonstration suivante est d'expliciter davantage le calcul de $f\Big((\boldsymbol{z},\boldsymbol{\lambda}),\boldsymbol{x}^*,\hat{\mathcal{B}}_M\Big)$.

*Preuve.* Considérons un ensemble d'images, $\mathcal{B}_M = \{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_M\}$, est utilisé pour estimer $p(\boldsymbol{z},\boldsymbol{\lambda})$ pour un couple $(\boldsymbol{z},\boldsymbol{\lambda})$. Ce dernier est originellement échantillonné depuis $p(\boldsymbol{z},\boldsymbol{\lambda}|\boldsymbol{x}^*)$ pour une image spécifique $\boldsymbol{x}^*$. $p(\mathcal{B}_M)$ est alors définie comme une distribution Uniforme. Afin d'échantillonner depuis $p(\mathcal{B}_M)$, $M$ indices sont tirés aléatoirement depuis $\{1,\ldots,N\}$ sans remplacement, où $N$ correspond au nombre total d'images du jeu de

données. Dans ce contexte, on peut obtenir l'expression suivante :

$$\begin{aligned}
p(\boldsymbol{z}, \boldsymbol{\lambda}) &= \mathbb{E}_{p(\boldsymbol{x})}[p(\boldsymbol{z}, \boldsymbol{\lambda}|\boldsymbol{x})] \\
&= \mathbb{E}_{p(\mathcal{B}_M)}\left[\frac{1}{M}\sum_{m=1}^{M} p(\boldsymbol{z}, \boldsymbol{\lambda}|\boldsymbol{x}_m)\right] \\
&= \mathbb{P}(\boldsymbol{x}^* \in \mathcal{B}_M)\mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} p(\boldsymbol{z}, \boldsymbol{\lambda}|\boldsymbol{x}_m)\middle|\boldsymbol{x}^* \in \mathcal{B}_M\right] + \mathbb{P}(\boldsymbol{x}^* \notin \mathcal{B}_M)\mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} p(\boldsymbol{z}, \boldsymbol{\lambda}|\boldsymbol{x}_m)\middle|\boldsymbol{x}^* \notin \mathcal{B}_M\right] \\
&= \frac{M}{N}\mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} p(\boldsymbol{z}, \boldsymbol{\lambda}|\boldsymbol{x}_m)\middle|\boldsymbol{x}^* \in \mathcal{B}_M\right] + \frac{N-M}{N}\mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} p(\boldsymbol{z}, \boldsymbol{\lambda}|\boldsymbol{x}_m)\middle|\boldsymbol{x}^* \notin \mathcal{B}_M\right].
\end{aligned} \tag{D.8}$$

Cette grandeur peut être estimée en échantillonnant un ensemble de $M+1$ images, dans lequel $\boldsymbol{x}^*$ est un élément, et en considérant $\hat{B}_M = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ un ensemble d'éléments différents de $\boldsymbol{x}^*$. Dans ce contexte, la première espérance de l'équation (4.5) peut être estimée en utilisant $\{\boldsymbol{x}^*\} \cup \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{M-1}\}$ et la seconde en utilisant $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$. L'estimateur, noté $f\Big((\boldsymbol{z}, \boldsymbol{\lambda}), \boldsymbol{x}^*, \hat{\mathcal{B}}_M\Big)$, s'écrit alors :

$$f\Big((\boldsymbol{z}, \boldsymbol{\lambda}), \boldsymbol{x}^*, \hat{\mathcal{B}}_M\Big) = \frac{1}{N}p(\boldsymbol{z}, \boldsymbol{\lambda}|\boldsymbol{x}^*) + \frac{1}{M}\sum_{m=1}^{M-1} p(\boldsymbol{z}, \boldsymbol{\lambda}|\boldsymbol{x}_m) + \frac{N-M}{NM}p(\boldsymbol{z}, \boldsymbol{\lambda}|\boldsymbol{x}_M). \tag{D.9}$$

□

Une dérivation similaire peut être considérée pour estimer $\prod_k p(z_k, \lambda_k)$. Cet estimateur est non biaisé, c'est-à dire que $p(\boldsymbol{z}, \boldsymbol{\lambda}) = \mathbb{E}\left[f\Big((\boldsymbol{z}, \boldsymbol{\lambda}), \boldsymbol{x}^*, \hat{\mathcal{B}}_M\Big)\right]$, et exact si $M = N$.

# Annexe E

# Résultats complémentaires des expériences

Cette annexe propose de compléter les résultats issus de la comparaison des méthodes de désentrelacement, dont l'étude est menée dans le chapitre 4. Elle considère les méthodes développées lors de nos recherches, ainsi que celles de l'état de l'art, pour lesquelles les résultats sont omis dans le texte principal. L'annexe se divise en deux parties principales : la première approfondit les résultats concernant le modèle NGVAE, tandis que la seconde explique les critères utilisés pour choisir les modèles de l'état de l'art et présente les résultats obtenus avec ces méthodes, spécifiquement pour chaque jeu de données examiné.

## E.1 Analyse de l'apport de la corrélation totale pour le désentrelacement

Dans la section 4.2.1 du manuscrit, l'efficacité de l'ajour d'un terme de corrélation totale pour parvenir à un espace latent désentrelacé est étudiée. Dans cet objectif, les matrices de corrélation de Spearman sont relevées, permettant d'analyser la relation entre les variables latentes et les facteurs génératifs du jeu de données Dsprites, pour lequel le facteur d'orientation a été exclu. Cette analyse a été menée à la fois sur le NGVAE et le TC-NGVAE. Ces résultats peuvent être été mis en parallèle avec ceux obtenus en utilisant le Vanilla-VAE [Kingma et Welling, 2014] et le $\beta$-VAE [Higgins *et al.*, 2016], pour lequel $\beta = 4$. Concernant ces deux dernières méthodes, leurs résultats sont détaillés dans la Figure E.1.

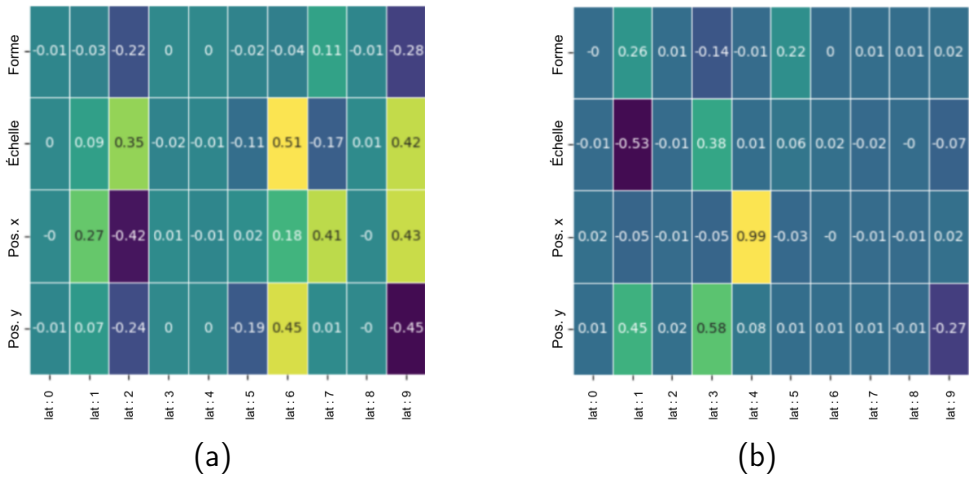



Fig E.1: Matrices de Spearman entre les variables latentes et les facteurs génératifs. (a) : Vanilla-VAE. (b) : $\beta$-VAE, avec $\beta = 4$.

En comparaison des résultats obtenus avec les modèles proposés, on observe que le Vanilla-VAE et le $\beta$-VAE tendent à répartir l'information de manière plus diffuse, tel qu'illustré dans les Figures E.1a et E.1b. Cette dispersion affecte la modularité de l'espace latent. En outre, pour ces deux méthodes, il semble qu'une seule variable latente encode plusieurs facteurs génératifs, ce qui nuit également à qualité compacte de $\boldsymbol{z}$. Ces constatations viennent renforcer l'efficacité de l'intégration du terme de corrélation totale dans le TC-NGVAE, favorisant une meilleure convergence vers un espace latent désentrelacé.

## E.2 Analyse de la capacité de désentrelacement de différentes approches

Dans un premier temps, les résultats qui ont orienté notre choix de modèles à comparer avec le TC-NGVAE, en particulier ceux non détaillés dans le texte principal, sont détaillés. Pour chacun des jeux de données utilisés dans nos expériences, les résultats des trois méthodes non incluses dans le manuscrit principal sont ensuite présentés.

### E.2.1 Sélection des hyperparamètres pour les méthodes de l'état de l'art

Rappelons que, dans le cadre de la comparaison du TC-NGVAE avec d'autres méthodes de l'état de l'art, les hyperparamètres du BF-VAE-2 [Kim *et al.*, 2019a] ont été définis après avoir effectué une recherche sur une grille. Concernant les autres méthodes, le Factor-VAE [Kim et Mnih, 2018], le DIP-VAE-II [Kumar *et al.*, 2017], le $\beta$-VAE [Higgins *et al.*, 2016] et le $\beta$-TC-VAE [Chen *et al.*, 2018], les modèles ont été obtenus en utilisant la bibliothèque "Disentanglement-Lib" [Locatello *et al.*, 2019b]. Cette dernière fournit des résultats pour diverses mesures de désentrelacement, à savoir les métriques MIG [Chen *et al.*, 2018], "Désentrelacement", "Compacité" et "Qualité explicite" de la mesure DCI [Eastwood et Williams, 2018], ainsi que la mesure "z-min variance" [Kim et Mnih, 2018], basés sur cinquante initialisations différentes et un ensemble de valeurs d'hyperparamètres.

Pour chaque méthode analysée, les hyperparamètres ont été ajustés afin d'atteindre la valeur moyenne la plus élevée des mesures de désentrelacement, déterminée à partir de cinquante initialisations. Les résultats moyens obtenus pour le DIP-VAE-II ont été inclus dans le texte principal. Les tableaux qui suivent détaillent les valeurs moyennes obtenues à partir de chaque hyperparamètre, pour les trois méthodes non explicitées dans le manuscrit.

Les résultats pour le Factor-VAE, présentés dans le tableau E.1, montrent que l'hyperparamètre optimal est $\lambda = 100$ lors de l'entraînement sur les jeux de données Dsprites et Cars3d, et $\lambda = 30$ pour un entraînement sur Smallnorb.

Les résultats obtenus pour le $\beta$-VAE, affichés dans le tableau E.2, indiquent que les meilleurs résultats obtenus résultent pour une valeur de $\beta = 16$ pour chacun des jeux de données considérés.
Enfin, pour le $\beta$-TCVAE, dont les résultats sont relevés dans le tableau E.3, le paramètre $\beta$ a été fixé à 10 pour les entraînements sur Dsprites et Smallnorb, et à 1 pour l'entraînement sur Cars3d.

| **Valeurs de $\lambda$** | **Valeur moyenne** | | |
|---|---|---|---|
| | Dsprites | Cars3d | Smallnorb |
| 10 | 0.357 | 0.440 | 0.399 |
| 20 | 0.368 | 0.441 | 0.419 |
| 30 | 0.374 | 0.438 | **0.423** |
| 40 | 0.371 | 0.438 | 0.417 |
| 50 | 0.380 | 0.440 | 0.417 |
| 100 | **0.395** | **0.442** | 0.399 |

Tableau E.1: Moyennes des indicateurs de désentrelacement obtenues pour le Factor-VAE calculée sur cinquante initialisations, selon différentes valeurs de $\lambda$.

| **Valeurs de $\beta$** | **Valeur moyenne** | | |
|---|---|---|---|
| | Dsprites | Cars3d | Smallnorb |
| 1 | 0.280 | 0.340 | 0.419 |
| 2 | 0.293 | 0.358 | 0.382 |
| 4 | 0.300 | 0.391 | 0.367 |
| 6 | 0.315 | 0.409 | 0.350 |
| 8 | 0.329 | 0.426 | 0.306 |
| **16** | **0.401** | **0.438** | **0.427** |

Tableau E.2: Moyennes des indicateurs de désentrelacement obtenues pour le $\beta$-VAE, calculée sur cinquante initialisations, selon différentes valeurs de $\beta$.

| **Valeurs de $\beta$** | **Valeur moyenne** | | |
|---|---|---|---|
| | Dsprites | Cars3d | Smallnorb |
| 1 | 0.280 | **0.437** | 0.339 |
| 2 | 0.361 | 0.431 | 0.382 |
| 4 | 0.408 | 0.402 | 0.429 |
| 6 | 0.415 | 0.389 | 0.454 |
| 8 | 0.417 | 0.380 | 0.475 |
| 10 | **0.418** | 0.375 | **0.483** |

Tableau E.3: Moyennes des indicateurs de désentrelacement obtenues pour le $\beta$-TCVAE, calculée sur cinquante initialisations, selon différentes valeurs de $\beta$.

### E.2.2 Entraînement effectué sur Dsprites

L'évaluation du TC-NGVAE à converger vers un espace latent désentrelacé par rapport aux autres méthodes de l'état de l'art, lorsque ces dernières sont entraînées sur le jeu de

données Dsprites, a été menée dans la section 4.2.3 du manuscrit. Cette analyse s'est principalement focalisée sur les performances du $\beta$-VAE et du $\beta$-TCVAE.

Pour compléter cette étude, les Figures ci-dessous affichent les résultats des méthodes non incluses dans l'analyse principale. Elles représentent les matrices de corrélation de Spearman, illustrant les relations entre les variables latentes et les facteurs génératifs pour le BF-VAE-2, le DIP-VAE-II et le Factor-VAE.

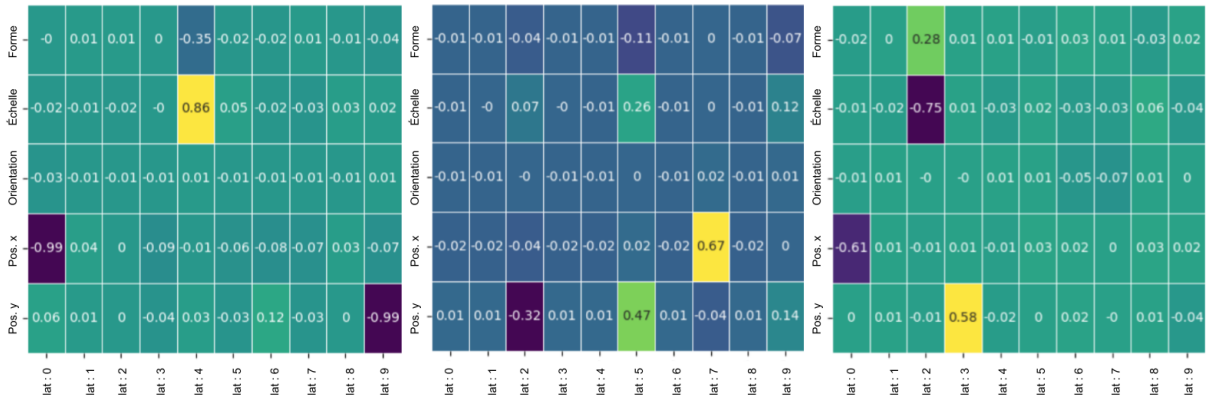



Fig E.2: Matrices de Spearman mesurées entre les variables latentes et les facteurs génératifs. (a) : Bf-VAE-2. (b) : DIP-VAE-II. (c) : Factor-VAE.

En parallèle avec l'analyse présentée dans le chapitre 4, cette étude souligne une caractéristique importante des trois méthodes examinées : l'absence de corrélation entre le facteur génératif d'orientation et les variables latentes. Cette observation met en lumière l'ambiguïté dans la définition de ce facteur génératif, relevée dans le manuscrit. En outre, on relève que les corrélations entre les facteurs génératifs et les variables latentes, relevées dans les Figures E.2, sont inférieures à celles obtenues pour notre méthode. Cette observation est confirmée par les scores plus élevés obtenus pour la métrique "DCI - Qualité Explicite" pour notre approche. Ces résultats mettent en évidence certaines caractéristiques des méthodes de l'état de l'art, tout en soulignant la capacité supérieure du TC-NGVAE à aboutir à un espace latent désentrelacé, caractérisé par sa modularité, sa compacité et sa qualité explicite.

Enfin, les matrices de covariance empiriques, calculées sur l'espace latent pour toutes les méthodes considérées, y compris le TC-NGVAE, sont présentées dans la Figure E.3.

L'analyse de la Figure E.3 met en évidence une différence notable par rapport au TC-NGVAE, dont les résultats sont détaillés dans la Figure E.3f. En effet, plusieurs corrélations demeurent pour les méthodes de l'état de l'art. Par ailleurs, chacune de ces méthodes présente une tendance à la polarisation, bien que certains résultats peu probants soient observés.

En prenant l'exemple de la variable latente 8 du DIP-VAE-II, ne montre pas de corrélation avec les facteurs génératifs telle que relevé dans la Figure E.2b. Toutefois, une forte valeur sur la diagonale de la matrice de covariance est visible dans la Figure E.3d. Cette observation suggère une variation de cette variable à travers différentes images du jeu de données, contredisant certains des principes de polarisation évoqués dans le manuscrit.

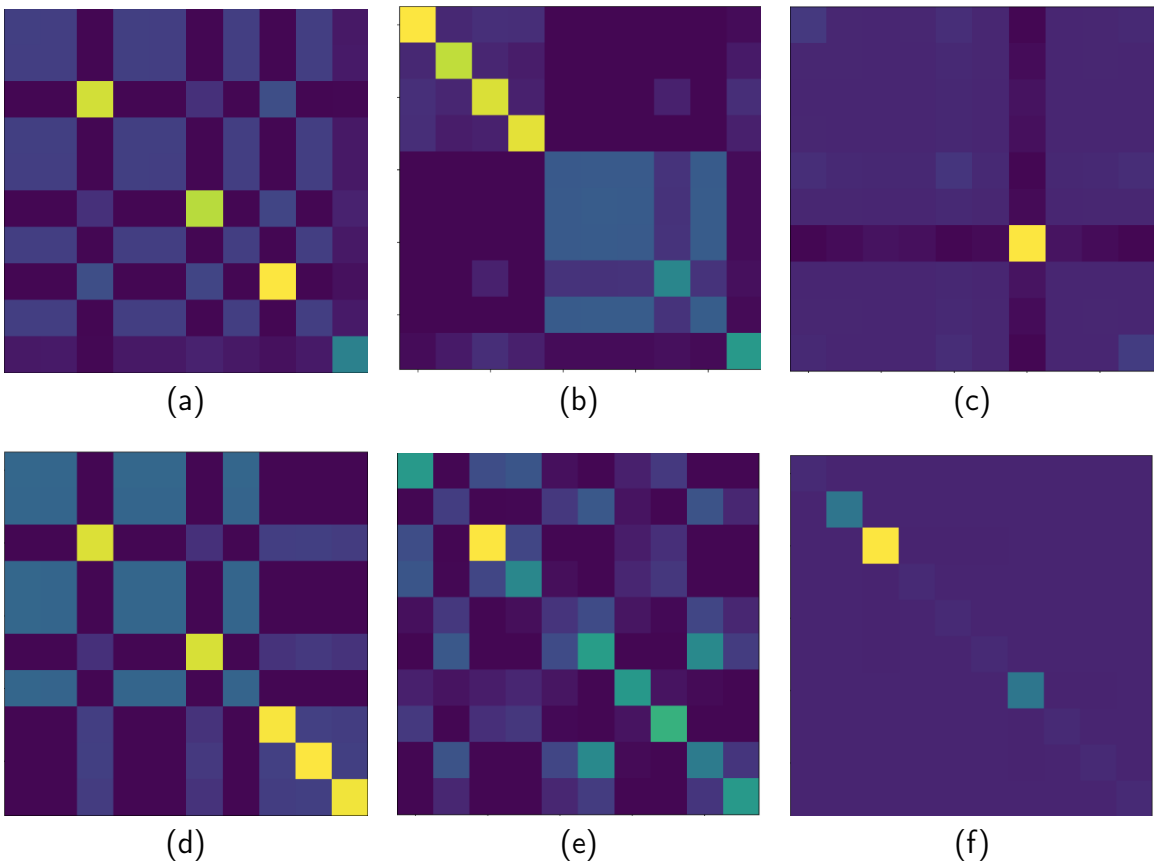



Fig E.3: Matrices de covariance empirique mesurées sur un espace latent de taille 10. (a) : $\beta$-VAE. (b) : $\beta$-TCVAE. (c) : BF-VAE-2. (d) : DIP-VAE-II. (e) : Factor-VAE. (f) : TCNGVAE.

De plus, la matrice de covariance du BF-VAE-2, illustrée dans la Figure E.3c, présente des résultats peu communs. Ils peuvent être interprétés comme des conséquences de la modélisation spécifique adoptée par cette méthode.

### E.2.3 Entraînement effectué sur Smallnorb

L'analyse menée sur le jeu de données Smallnorb s'est principalement concentrée sur l'évaluation des modèles du BF-VAE-2 et du Factor-VAE par rapport au TC-NGVAE. La section suivante est ainsi consacrée aux résultats obtenus par le $\beta$-VAE, le $\beta$-TCVAE, et le DIP-VAE-II. Les matrices de Spearman entre les variables latentes et les facteurs génératifs, sont présentées dans la Figure E.4.
Le facteur génératif d'orientation présente peu de corrélation avec les variables latentes dans ces méthodes, soulignant l'ambiguïté relevée dans l'analyse de Smallnorb effectuée dans le chapitre 4 du manuscrit. En outre, les facteurs génératifs sont dispersés sur plusieurs variables latentes, ce qui explique la meilleure capacité de notre méthode à converger vers un espace latent plus compact. Concernant la qualité explicite de l'espace latent, les corrélations entre les facteurs génératifs et les variables latentes s'avèrent également plus marquées pour le TC-NGVAE.

Enfin, les matrices de covariance empirique calculées sur l'espace latent pour toutes les méthodes considérées et entraînées sur Smallnorb sont présentées dans la Figure E.5.
Davantage de corrélations entre les variables passives sont observées pour ces méthodes, en comparaison avec le TCNGVAE, comme le montre la Figure E.5f. De plus, la distinction

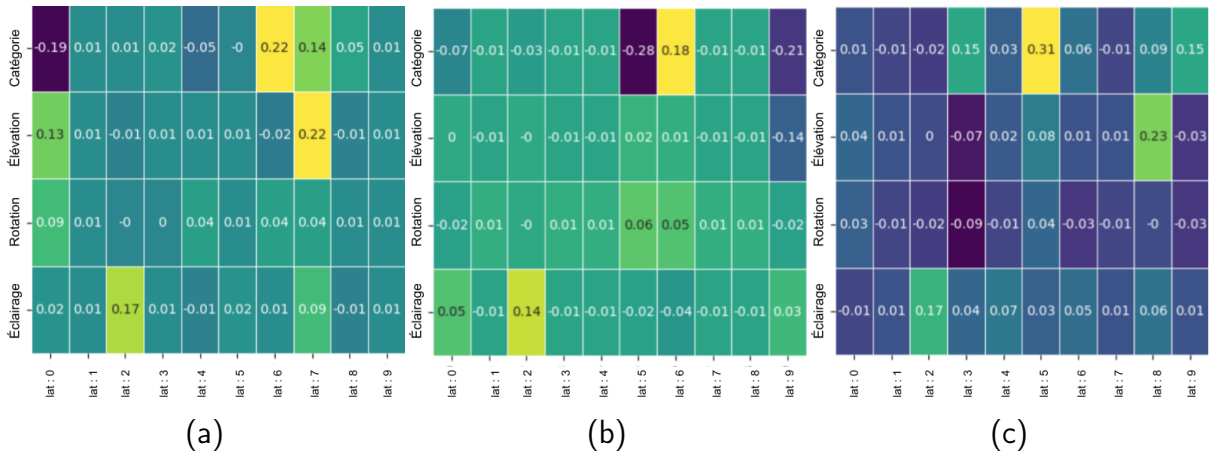



Fig E.4: Matrices de Spearman mesurées entre les variables latentes et les facteurs génératifs. (a) : $\beta$-VAE. (b) : DIP-VAE-II. (c) : $\beta$-TCVAE.

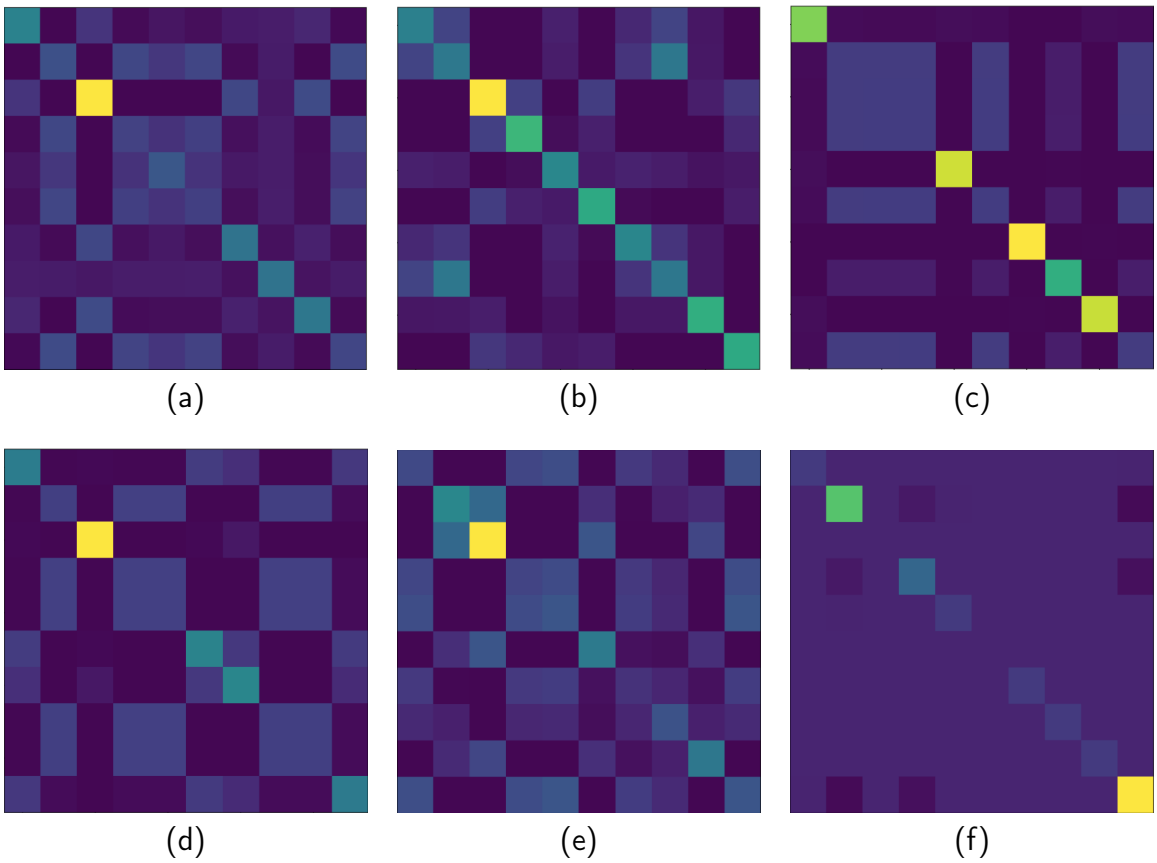



Fig E.5: Matrices de covariance empirique mesurées sur un espace latent de taille 10. (a) : $\beta$-VAE. (b) : $\beta$-TCVAE. (c) : BF-VAE-2. (d) : DIP-VAE-II. (e) : Factor-VAE. (f) : TCNGVAE.

moins nette entre les variables actives et passives dans ces matrices contraste avec l'analyse des données obtenues pour Dsprites. Cette différence peut être attribuée à la complexité supérieure du jeu de données Smallnorb.

### E.2.4 Entraînement effectué sur Cars3d

Pour conclure cette annexe, les matrices de Spearman entre les variables latentes et les facteurs génératifs sont relevées dans la Figure E.6. Elles concernent les modèles BF-VAE-2, $\beta$-TCVAE et Factor-VAE, suite à un entraînement sur le jeu de données Cars3d.

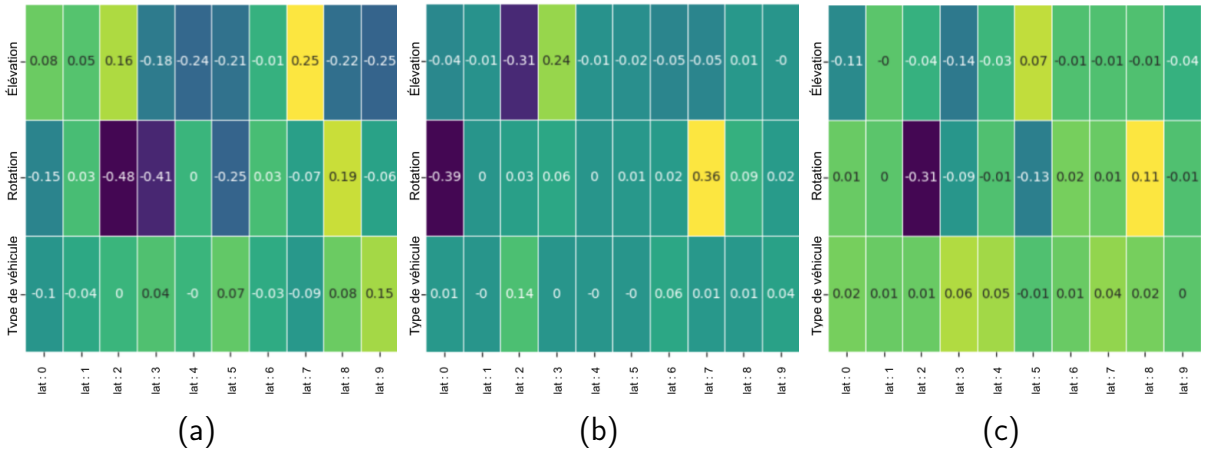



Fig E.6: Matrices de Spearman mesurées entre les variables latentes et les facteurs génératifs. (a) : BF-VAE-2. (b) : $\beta$-VAE. (c) : Factor-VAE.

Leur analyse révèle que, de manière similaire aux modèles étudiés dans le manuscrit, les trois méthodes de l'état de l'art examinées ne montrent pas de corrélations significatives entre les variables latentes et le facteur génératif lié à la catégorie de voiture. Cela est dû à l'ambiguïté inhérente aux facteurs génératifs utilisés pour constituer le jeu de données, un point que nous avons détaillé dans le manuscrit.

Pour terminer, les matrices de covariance empiriques calculées sur l'espace latent pour l'ensemble des méthodes considérées, sont illustrées dans la Figure E.7.
Les résultats obtenus pour le TC-NGVAE, illustrés dans la Figure E.7f, montrent une meilleure capacité de décorrélation entre les variables latentes qui se manifeste par les valeurs faibles ou nulles dans les éléments hors diagonale des matrices, ainsi qu'une polarisation efficace entre les variables actives et passives.

Cependant, il est important de prendre en considération les difficultés de convergence observées pour l'ensemble des méthodes sur ce jeu de données lors de l'évaluation des résultats. Une particularité de Cars3d réside dans l'indépendance des facteurs génératifs les uns par rapport aux autres. Cette caractéristique peut influence significativement la performance des différentes méthodes et doit être prise en compte lors de l'interprétation des résultats obtenus.

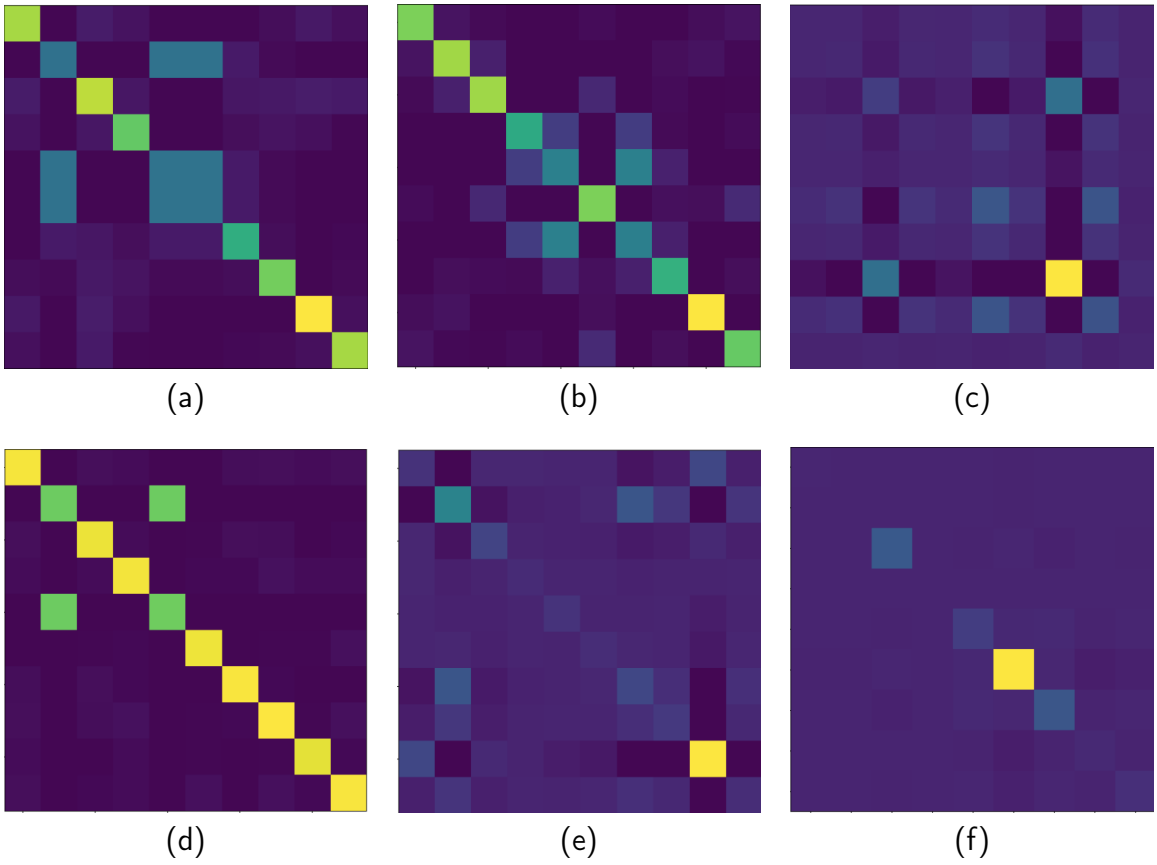



Fig E.7: Matrices de covariance empirique mesurées sur un espace latent de taille 10. (a) : $\beta$-VAE. (b) : $\beta$-TCVAE. (c) : BF-VAE-2. (d) : DIP-VAE-II. (e) : Factor-VAE. (f) : TC-NGVAE.

# Bibliographie


Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard *et al.* : Tensorflow: a system for large-scale machine learning. *In 12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pages 265–283, 2016.

Alexander A Alemi, Ian Fischer et Joshua V Dillon : Uncertainty in the Variational Information Bottleneck. *arXiv preprint arXiv:1807.00906*, 2018.

Alexander A Alemi, Ian Fischer, Joshua V Dillon et Kevin Murphy : Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Jinwon An et Sungzoon Cho : Variational autoencoder based anomaly detection using reconstruction probability. *Special lecture on IE*, 2(1):1–18, 2015.

Martin Arjovsky, Soumith Chintala et Léon Bottou : Wasserstein generative adversarial networks. *In International conference on machine learning*, pages 214–223. PMLR, 2017.

Karl Johan Åström et Tore Hägglund : *Advanced PID control*. ISA-The Instrumentation, Systems and Automation Society, 2006.

Mathieu Aubry, Daniel Maturana, Alexei A Efros, Bryan C Russell et Josef Sivic : Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3762–3769, 2014.

Matthias Bauer et Andriy Mnih : Resampled priors for variational autoencoders. *In The 22nd International Conference on Artificial Intelligence and Statistics*, pages 66–75. PMLR, 2019.

Yoshua Bengio *et al.* : Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.

Yoshua Bengio, Aaron Courville et Pascal Vincent : Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Yoshua Bengio et Olivier Delalleau : On the expressive power of deep architectures. *In International conference on algorithmic learning theory*, pages 18–36. Springer, 2011.

Yoshua Bengio, Eric Laufer, Guillaume Alain et Jason Yosinski : Deep generative stochastic networks trainable by backprop. *In International Conference on Machine Learning*, pages 226–234. PMLR, 2014.

James Bergstra et Yoshua Bengio : Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.

Christopher M Bishop : *Pattern recognition and machine learning (information science and statistics)*. Springer New York, 2007.

Lisa Bonheme et Marek Grzes : Be more active! understanding the differences between mean and sampled representations of variational autoencoders, 2021.

Diane Bouchacourt, Ryota Tomioka et Sebastian Nowozin : Multi-level variational autoencoder: Learning disentangled representations from grouped observations. *In Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne *et al.* : Jax: Autograd and xla. *Astrophysics Source Code Library*, pages ascl–2111, 2021.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins et Alexander Lerchner : Understanding disentangling in *beta*-vae. *arXiv preprint arXiv:1804.03599*, 2018.

Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick et Alexander Lerchner : Monet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019.

Marc-André Carbonneau, Julian Zaidi, Jonathan Boilard et Ghyslain Gagnon : Measuring disentanglement: A review of metrics. *IEEE transactions on neural networks and learning systems*, 2022.

Francesco Paolo Casale, Adrian Dalca, Luca Saglietti, Jennifer Listgarten et Nicolo Fusi : Gaussian process prior variational autoencoders. *Advances in neural information processing systems*, 31, 2018.

Gal Chechik, Amir Globerson, Naftali Tishby et Yair Weiss : Information bottleneck for gaussian variables. *Advances in Neural Information Processing Systems*, 16, 2003.

Hsin Chen et Alan F Murray : Continuous restricted boltzmann machine with an implementable training algorithm. *IEE Proceedings-Vision, Image and Signal Processing*, 150(3):153–158, 2003.

Ricky TQ Chen, Xuechen Li, Roger B Grosse et David K Duvenaud : Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever et Pieter Abbeel : Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016.

Ze Cheng, Juncheng B Li, Chenxu Wang, Jixuan Gu, Hao Xu, Xinjian Li et Florian Metze : rtc-vae: harnessing the peculiarity of total correlation in learning disentangled representations. 2019.

KyungHyun Cho, Alexander Ilin et Tapani Raiko : Improved learning of gaussian-bernoulli restricted boltzmann machines. *In Artificial Neural Networks and Machine Learning–ICANN 2011: 21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14-17, 2011, Proceedings, Part I 21*, pages 10–17. Springer, 2011.

François Chollet : Keras: Deep Learning for humans, août 2023. URL `https://github.com/keras-team/keras`. original-date: 2015-03-28T00:35:42Z.

George Cybenko : Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Bin Dai, Yu Wang, John Aston, Gang Hua et David Wipf : Connections with robust pca and the role of emergent sparsity in variational autoencoder models. *The Journal of Machine Learning Research*, 19(1):1573–1614, 2018.

Bin Dai, Ziyu Wang et David Wipf : The usual suspects? reassessing blame for vae posterior collapse. *In International conference on machine learning*, pages 2313–2322. PMLR, 2020.

Kien Do et Truyen Tran : Theory and evaluation metrics for learning disentangled representations, 2019.

Sunny Duan, Loic Matthey, Andre Saraiva, Nicholas Watters, Christopher P Burgess, Alexander Lerchner et Irina Higgins : Unsupervised model selection for variational disentangled representation learning. 2019.

Emilien Dupont : Learning disentangled joint continuous and discrete representations. *Advances in neural information processing systems*, 31, 2018.

Cian Eastwood et Christopher KI Williams : A framework for the quantitative evaluation of disentangled representations. *In International conference on learning representations*, 2018.

Benjamin Estermann et Roger Wattenhofer : Dava: Disentangling adversarial variational autoencoder. *arXiv preprint arXiv:2303.01384*, 2023.

Mikhail Figurnov, Shakir Mohamed et Andriy Mnih : Implicit reparameterization gradients. *Advances in neural information processing systems*, 31, 2018.

Ken-Ichi Funahashi : On the approximate realization of continuous mappings by neural networks. *Neural networks*, 2(3):183–192, 1989.

Mathieu Germain, Karol Gregor, Iain Murray et Hugo Larochelle : Made: Masked autoencoder for distribution estimation. pages 881–889, 2015.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter et Lalana Kagal : Explaining explanations: An overview of interpretability of machine learning. *In 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

Laurent Girin, Simon Leglaive, Xiaoyu Bie, Julien Diard, Thomas Hueber et Xavier Alameda-Pineda : Dynamical variational autoencoders: A comprehensive review. *arXiv preprint arXiv:2008.12595*, 2020.

Xavier Glorot et Yoshua Bengio : Understanding the difficulty of training deep feedforward neural networks. *In Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

Peter W Glynn : Likelihood ratio gradient estimation for stochastic systems. *Communications of the ACM*, 33(10):75–84, 1990.

Ian Goodfellow : Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

Ian Goodfellow, Yoshua Bengio et Aaron Courville : *Deep learning*. MIT press, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville et Yoshua Bengio : Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick et Alexander Lerchner : Multi-object representation learning with iterative variational inference. *In International Conference on Machine Learning*, pages 2424–2433. PMLR, 2019.

Ishaan Gulrajani, Kundan Kumar, Faruk Ahmed, Adrien Ali Taiga, Francesco Visin, David Vazquez et Aaron Courville : Pixelvae: A latent variable model for natural images. *arXiv preprint arXiv:1611.05013*, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren et Jian Sun : Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *In Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren et Jian Sun : Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende et Alexander Lerchner : Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed et Alexander Lerchner : beta-vae: Learning basic visual concepts with a constrained variational framework. *In International conference on learning representations*, 2016.

Geoffrey E Hinton : Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

Geoffrey E Hinton et Ruslan R Salakhutdinov : Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

Matthew D Hoffman et Matthew J Johnson : ELBO surgery: yet another way to carve up the variational evidence lower bound. *NIPS Workshop on Advances in Approximate Bayesian Inference*, 2016.

Kurt Hornik, Maxwell Stinchcombe et Halbert White : Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Emma Jouffroy, Audrey Giremus, Yannick Berthoumieu, Olivier Bach et Alain Hugget : Automatic selection of latent variables in variational auto-encoders. *In 2022 30th European Signal Processing Conference (EUSIPCO)*, pages 1407–1411. IEEE, 2022.

Tero Karras, Samuli Laine et Timo Aila : A style-based generator architecture for generative adversarial networks. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

Hyunjik Kim et Andriy Mnih : Disentangling by factorising. *In International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018.

Minyoung Kim, Yuting Wang, Pritish Sahu et Vladimir Pavlovic : Bayes-factor-vae: Hierarchical bayesian deep auto-encoder models for factor disentanglement. *In Proceedings of the IEEE/CVF international conference on computer vision*, pages 2979–2987, 2019a.

Minyoung Kim, Yuting Wang, Pritish Sahu et Vladimir Pavlovic : Relevance factor vae: Learning and identifying disentangled factors. *arXiv preprint arXiv:1902.01568*, 2019b.

Diederik P Kingma et Jimmy Ba : Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P. Kingma et Max Welling : Auto-Encoding Variational Bayes. *In 2nd International Conference on Learning Representations, ICLR, Conference Track Proceedings*, 2014.

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever et Max Welling : Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29, 2016.

Abhishek Kumar, Prasanna Sattigeri et Avinash Balakrishnan : Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*, 2017.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum et Samuel J Gershman : Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017.

Hugo Larochelle et Iain Murray : The neural autoregressive distribution estimator. *In Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 29–37. JMLR Workshop and Conference Proceedings, 2011.

Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, Richard Howard, Wayne Hubbard et Lawrence Jackel : Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2, 1989.

Yann LeCun, Fu Jie Huang et Leon Bottou : Learning methods for generic object recognition with invariance to pose and lighting. 2:II–104, 2004.

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang et Barnabás Póczos : Mmd gan: Towards deeper understanding of moment matching network. *Advances in neural information processing systems*, 30, 2017.

Zhiyuan Li, Jaideep Vitthal Murkute, Prashnna Kumar Gyawali et Linwei Wang : Progressive learning and disentanglement of hierarchical representations. *arXiv preprint arXiv:2002.10549*, 2020.

Zhixuan Lin, Yi-Fu Wu, Skand Vishwanath Peri, Weihao Sun, Gautam Singh, Fei Deng, Jindong Jiang et Sungjin Ahn : Space: Unsupervised object-oriented scene representation via spatial attention and decomposition. *arXiv preprint arXiv:2001.02407*, 2020.

Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev et Jason Yosinski : An intriguing failing of convolutional neural networks and the coordconv solution. *Advances in neural information processing systems*, 31, 2018.

Xiao Liu, Spyridon Thermos, Gabriele Valvano, Agisilaos Chartsias, Alison O'Neil et Sotirios A Tsaftaris : Measuring the biases and effectiveness of content-style disentanglement. *arXiv preprint arXiv:2008.12378*, 2020.

Ziwei Liu, Ping Luo, Xiaogang Wang et Xiaoou Tang : Deep learning face attributes in the wild. *In Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

Gabriel Loaiza-Ganem et John P Cunningham : The continuous bernoulli: fixing a pervasive error in variational autoencoders, 2019.

Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf et Olivier Bachem : On the fairness of disentangled representations. *Advances in neural information processing systems*, 32, 2019a.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf et Olivier Bachem : Challenging common assumptions in the unsupervised learning of disentangled representations. *In international conference on machine learning*, pages 4114–4124. PMLR, 2019b.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem et Michael Tschannen : Weakly-supervised disentanglement without compromises. *In International Conference on Machine Learning*, pages 6348–6359. PMLR, 2020.

Francesco Locatello, Michael Tschannen, Stefan Bauer, Gunnar Rätsch, Bernhard Schölkopf et Olivier Bachem : Disentangling factors of variation using few labels. *arXiv preprint arXiv:1905.01258*, 2019c.

James Lucas, George Tucker, Roger Grosse et Mohammad Norouzi : Understanding posterior collapse in generative latent variable models. 2019a.

James Lucas, George Tucker, Roger B Grosse et Mohammad Norouzi : Don't blame the elbo! a linear vae perspective on posterior collapse. *Advances in Neural Information Processing Systems*, 32, 2019b.

Lars Maaløe, Marco Fraccaro, Valentin Liévin et Ole Winther : Biva: A very deep hierarchy of latent variables for generative modeling. *Advances in neural information processing systems*, 32, 2019a.

Lars Maaløe, Marco Fraccaro, Valentin Liévin et Ole Winther : Biva: A very deep hierarchy of latent variables for generative modeling. *Advances in neural information processing systems*, 32, 2019b.

Andrew L Maas, Awni Y Hannun, Andrew Y Ng *et al.* : Rectifier nonlinearities improve neural network acoustic models. *In Proc. icml*, volume 30, page 3. Atlanta, GA, 2013.

Shweta Mahajan, Iryna Gurevych et Stefan Roth : Latent normalizing flows for many-to-many cross-domain mappings. *arXiv preprint arXiv:2002.06661*, 2020.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow et Brendan Frey : Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

Emile Mathieu, Tom Rainforth, Nana Siddharth et Yee Whye Teh : Disentangling disentanglement in variational autoencoders. *In International conference on machine learning*, pages 4402–4412. PMLR, 2019.

Loic Matthey, Irina Higgins, Demis Hassabis et Alexander Lerchner : dsprites: Disentanglement testing sprites dataset, 2017.

Warren S McCulloch et Walter Pitts : A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5:115–133, 1943.

Milton Llera Montero, Casimir JH Ludwig, Rui Ponte Costa, Gaurav Malhotra et Jeffrey Bowers : The role of disentanglement in generalisation. *In International Conference on Learning Representations*, 2020.

Thomas Müller, Brian McWilliams, Fabrice Rousselle, Markus Gross et Jan Novák : Neural importance sampling. *ACM Transactions on Graphics (ToG)*, 38 (5):1–19, 2019.

Izaak Neutelings : Neural networks, 2021. URL `https://tikz.net/neural_networks/`.

Sebastian Nowozin, Botond Cseke et Ryota Tomioka : f-gan: Training generative neural samplers using variational divergence minimization. *Advances in neural information processing systems*, 29, 2016.

OpenAI : Introducing ChatGPT. URL `https://openai.com/blog/chatgpt`.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga *et al.* : Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Ashis Pati et Alexander Lerch : Attribute-based regularization of latent spaces for variational auto-encoders. *Neural Computing and Applications*, 33:4429–4444, 2021.

Janis Postels, Hermann Blum, Yannick Strümpler, Cesar Cadena, Roland Siegwart, Luc Van Gool et Federico Tombari : The hidden uncertainty in a neural networks activations. *arXiv preprint arXiv:2012.03082*, 2020.

Alec Radford, Luke Metz et Soumith Chintala : Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu et Mark Chen : Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Ali Razavi, Aaron Van den Oord et Oriol Vinyals : Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

Scott E Reed, Yi Zhang, Yuting Zhang et Honglak Lee : Deep visual analogy-making. *Advances in neural information processing systems*, 28, 2015.

Danilo Jimenez Rezende, Shakir Mohamed et Daan Wierstra : Stochastic backpropagation and approximate inference in deep generative models. *In International conference on machine learning*, pages 1278–1286. PMLR, 2014.

Karl Ridgeway et Michael C Mozer : Learning deep disentangled embeddings with the f-statistic loss. *Advances in neural information processing systems*, 31, 2018.

Michal Rolinek, Dominik Zietlow et Georg Martius : Variational autoencoders pursue pca directions (by accident). *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12406–12415, 2019.

Mihaela Rosca, Balaji Lakshminarayanan et Shakir Mohamed : Distribution matching in variational inference. *arXiv preprint arXiv:1802.06847*, 2018.

Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova et Chudi Zhong : Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, 16:1–85, 2022.

Francisco R Ruiz, Titsias RC AUEB, David Blei *et al.* : The generalized reparameterization gradient. *Advances in neural information processing systems*, 29, 2016.

Oleh Rybkin, Kostas Daniilidis et Sergey Levine : Simple and effective vae training with calibrated decoders. *In International conference on machine learning*, pages 9179–9189. PMLR, 2021.

Tim Salimans et David A Knowles : Fixed-form variational posterior approximation through stochastic linear regression. 2013.

Sivaramakrishnan Sankarapandian et Brian Kulis : $\beta$-annealed variational autoencoder for glitches. *arXiv preprint arXiv:2107.10667*, 2021.

Jürgen Schmidhuber : New millennium ai and the convergence of history. *In Challenges for computational intelligence*, pages 15–35. Springer, 2007.

Jürgen Schmidhuber : Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

Anna Sepliarskaia, Julia Kiseleva et Maarten de Rijke : How to not measure disentanglement. *arXiv preprint arXiv:1910.05587*, 2019a.

Anna Sepliarskaia, Julia Kiseleva, Maarten de Rijke *et al.* : Evaluating disentangled representations. *arXiv preprint arXiv:1910.05587*, 2019b.

Huajie Shao, Shuochao Yao, Dachun Sun, Aston Zhang, Shengzhong Liu, Dongxin Liu, Jun Wang et Tarek Abdelzaher : Controlvae: Controllable variational autoencoder. *In International Conference on Machine Learning*, pages 8655–8664. PMLR, 2020.

Paul Smolensky *et al.* : Information processing in dynamical systems: Foundations of harmony theory. 1986.

Haley M So, Laurie Bose, Piotr Dudek et Gordon Wetzstein : Pixelrnn: In-pixel recurrent neural networks for end-to-end-optimized perception with neural sensors. *arXiv preprint arXiv:2304.05440*, 2023.

Joram Soch, The Book of Statistical Proofs, Maja, Pietro Monticone, Thomas J. Faulkenberry, Alex Kipnis, Kenneth Petrykowski, Carsten Allefeld, Heiner Atze, Adam Knapp, Ciarán D. McInerney, Lo4ding00 et amvosk : *StatProofBook/StatProofBook.github.io: StatProofBook 2023*. Zenodo, janvier 2024. URL https://doi.org/10.5281/zenodo.10495684.

Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby et Ole Winther : Ladder variational autoencoders. *Advances in neural information processing systems*, 29, 2016a.

Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby et Ole Winther : Ladder variational autoencoders. *Advances in neural information processing systems*, 29, 2016b.

David Stutz : Illustrating (Convolutional) Neural Networks in LaTeX with TikZ, 2020. URL https://davidstutz.de/illustrating-convolutional-neural-networks-in-latex-with-tikz/.

Masashi Sugiyama, Taiji Suzuki et Takafumi Kanamori : Density ratio estimation: A comprehensive review (statistical experiment and its related topics). , 1703:10–31, 2010.

Masashi Sugiyama, Taiji Suzuki et Takafumi Kanamori : *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

Raphael Suter, Djordje Miladinovic, Bernhard Schölkopf et Stefan Bauer : Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. *In International Conference on Machine Learning*, pages 6056–6065. PMLR, 2019.

Ilya Sutskever, James Martens, George Dahl et Geoffrey Hinton : On the importance of initialization and momentum in deep learning. *In International conference on machine learning*, pages 1139–1147. PMLR, 2013.

Esteban G Tabak et Cristina V Turner : A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013.

Robert Tibshirani : Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Naftali Tishby, Fernando C Pereira et William Bialek : The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Jakub Tomczak et Max Welling : Vae with a vampprior. *In International Conference on Artificial Intelligence and Statistics*, pages 1214–1223. PMLR, 2018.

Michael Tschannen, Olivier Bachem et Mario Lucic : Recent advances in autoencoder-based representation learning. *arXiv preprint arXiv:1812.05069*, 2018.

Aaron Van Den Oord, Oriol Vinyals *et al.* : Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Nicholas Watters, Loic Matthey, Christopher P Burgess et Alexander Lerchner : Spatial broadcast decoder: A simple architecture for learning disentangled representations in vaes. *arXiv preprint arXiv:1901.07017*, 2019.

Ronald J Williams : Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992.

Ronald Yu : A tutorial on vaes: From bayes' rule to lossless compression. *arXiv preprint arXiv:2006.10273*, 2020.

Ming Yuan et Yi Lin : Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68 (1):49–67, 2006.

Shengjia Zhao, Jiaming Song et Stefano Ermon : Learning hierarchical features from deep generative models. *In International Conference on Machine Learning*, pages 4091–4099. PMLR, 2017.

Kangchen Zhu, Zhiliang Tian, Ruifeng Luo et Xiaoguang Mao : Styleflow: Disentangle latent representations via normalizing flow for unsupervised text style transfer. *arXiv preprint arXiv:2212.09670*, 2022.

Xinqi Zhu, Chang Xu et Dacheng Tao : Where and what? examining interpretable disentangled representations. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5861–5870, 2021.