# Topological degree as a discrete diagnostic for disentanglement, with applications to the ΔVAE*

Mahefa Ratsisetraina Ravelonanosy[1,*], Vlado Menkovski[1,2] and Jacobus W. Portegies[1,2]

[1]Department of Mathematics and Computer Science, Eindhoven University of Technology, 5612 AZ Eindhoven
[2]EAISI, Eindhoven University of Technology, 5612 AZ Eindhoven

**Abstract**

We investigate the ability of Diffusion Variational Autoencoder (ΔVAE) with unit sphere $\mathcal{S}^2$ as latent space to capture topological and geometrical structure and disentangle latent factors in datasets. For this, we introduce a new diagnostic of disentanglement: namely the topological degree of the encoder, which is a map from the data manifold to the latent space. We derive and implement an algorithm that computes this degree, and we use it to compute the degree of the encoder of models that result from the training procedure. Our experimental results show that the ΔVAE achieves relatively small LSBD scores, and that regardless of the degree after initialization, the degree of the encoder after training becomes $-1$ or $+1$, which implies that the resulting encoder is at least homotopic to a homeomorphism.

**Keywords**

Disentangled representation, Homeomorphic autoencoding, Topological degree., Variational Autoencoder,

## 1. Introduction

A data representation is often desired to capture or "disentangle" the explanatory factors of the dataset [1]. Although there is still no agreed definition for disentanglement, mathematical definitions and measures do exist, such as Linear Symmetry Based Disentanglement (LSBD) [2] and the LSBD score [3].

Besides formal definitions of disentanglement, there are desired characteristics for disentangled latent factors, for instance that nearby points in the dataspace should correspond to nearby points in the latent space representation. This could lead to a requirement that the encoder should be a homeomorphism [4, 5].

Given that an agreed definition of disentanglement does not yet exist, we consider it desirable to develop a wide range of diagnostics that are somehow related to the intuitive concept of disentanglement. Moreover, in practice it can be difficult to test whether a given encoder is a homeomorphism. Therefore, we introduce the topological degree as a discrete diagnostic for disentanglement. A homeomorphic encoder always has degree $+1$ or $-1$, whereas an encoder with degree $\pm 1$ is at least homotopic to a homeomorphism.

To achieve a homeomorphic encoder, or to get an encoder with degree $\pm 1$, one needs to choose a latent space that matches the topology of the dataset, otherwise one will encounter the manifold mismatch problem [6].

In order to have a wider range of latent spaces and solve the manifold mismatch problem [6], Pérez Rey et al [7] developed the Diffusion Variational Autoencoder (ΔVAE) that allows for any closed Riemannian manifold as latent space.

We immediately apply the degree as a diagnostic for disentanglement in an evaluation of the ΔVAE, in which we test the ΔVAE with a spherical latent space on data which naturally has a spherical latent structure.

*Discovery Science - Late Breaking Contributions 2024*
*This work was supported by NWO GROOT project UNRAVEL, OCENW.GROOT.2019.044.
*Corresponding author.
m.r.ravelonanosy@tue.nl (M. R. Ravelonanosy); v.menkovski@tue.nl (V. Menkovski); j.w.portegies@tue.nl (J. W. Portegies)

It can be challenging to train a ΔVAE, and we wondered whether this was due to the initialization and topological obstructions, see also [8, 4]. Indeed, if training corresponds to a continuous deformation of the encoder and decoder, if the degree would not be initialized at 1 or $-1$, the encoder would have no chance to reach suitable disentanglement.

Our experiments show that regardless of the initial model weight, the topological degree of the encoder can change to become eventually constant equal to $+1$ or $-1$, after some epochs of the training process. We perform the same experiments for the $\mathcal{S}$-VAE [6] and compare the results. The code that we used in the experiments can be found at https://gitlab.tue.nl/diffusion-vae/degree.

## 2. Related work

The VAE [9, 10] and its extensions are among the most used models when it comes to learning disentangled representations [11, 12, 13]. Some VAE extensions propose the use of more complex prior distribution other than the Gaussian in order to better match the distribution of the latent code [14, 15, 16, 17]. Some extensions propose independence of each latent dimension by modifying the VAE loss function [11, 13]. Other extensions use more geometric approaches to make the latent space itself match the geometry of the dataset [18, 6, 19, 4, 20, 7].

Intuitions and some aspects of disentangled representation are presented in [1, 21, 22], while overviews of several disentanglement metrics are given in [23] and [24]. Disentanglement is originally assessed with visual inspections and performance on downstream tasks [23]. Efforts have been devoted to propose metrics to evaluate different aspects of disentanglement [8, 25, 26, 27, 3]. The disentanglement metrics derived in these works do not check geometric aspects of disentanglement such as homeomorphism and topological degree according to the original mathematical definitions of these aspects. The degree *was* mentioned as a topological obstruction to homeomorphic autoencoding in [8].

## 3. Topological degree as a diagnostic for disentanglement

The encoder of a VAE can be considered as a continuous map from a dataspace $\mathcal{X} \subseteq \mathbb{R}^n$ to the latent space $Z$. Intuitively speaking, its topological degree is the number of times that the encoder wraps the data manifold around the latent space, counted in such a way that positive cancels negative orientation (cf. [28, Page 134] and [29, Page 27]). Encoder degree unequal to 1 or $-1$ indicates that the encoder cannot be a homeomorphism [29, Page 51].

**Computing the topological degree of the encoder** Although general methods exist [30], we developed and implemented a basic algorithm targeted to the case at hand of computing the degree of a map between spheres. We triangulate the two spheres, and create a "rounding" of the original map that maps vertices to vertices, edges to collections of edges and faces to collections of faces. We finally count how many times the faces in the target sphere are covered, taking into account orientation. Using tools from homology theory we can prove that the algorithm gives the correct result cf. [31].

## 4. Experiments

We train the ΔVAE with $\mathcal{S}^2$ as latent space, using a second-order expansion of the heat kernel [32]. We use a dataset of spherical harmonics as a proxy for a more natural dataset of axisymmetric pictures on the unit sphere, which naturally has the topology of $\mathcal{S}^2$ [33, Page 88] [34, 35]. We include a semisupervised LSBD-loss as in [3] and a semisupervised LSBD loss for the decoder. Also, we evaluate the LSBD score outlined in [3] with the group $SO(3)$. The representation of the data given by the models is then good if the corresponding LSBD score is small. Furthermore, we compute the distance distortion (DD) metric as given in [7], and the log-likelihood estimate as in [36]; for further details see also [7]. We compare the result with $\mathcal{S}$-VAE. The numerical results of the experiments are presented in Table 1.

**Evolution of the degree during the training** We conducted more experiments for spherical harmonics of degree $L = 7, 5, 3$ with ΔVAE in order to get insight into the evolution of the degree during

**Table 1**
Results for training the ΔVAE and the *S*-VAE. The "degree" column reports how often the absolute value of the degree equaled 1 after training. Each model was trained 5 times.

| Model | LL | ELBO | KL | RE | DD | Degree | LSBD |
|---|---|---|---|---|---|---|---|
| **Spherical harmonics of degree** $L = 5$ | | | | | | | |
| ΔVAE | $-15.92_{\pm 0.03}$ | $6.74_{\pm 0.01}$ | $6.70_{\pm 0.00}$ | $0.01_{\pm 0.00}$ | $0.05_{\pm 0.01}$ | 5 out of 5 | $0.01_{\pm 0.01}$ |
| *S*-VAE | $-0.22_{\pm 0.00}$ | $8.41_{\pm 0.08}$ | $8.39_{\pm 0.08}$ | $0.02_{\pm 0.03}$ | $0.00_{\pm 0.00}$ | 5 out of 5 | $0.00_{\pm 0.00}$ |
| **Spherical harmonics of degree** $L = 7$ | | | | | | | |
| ΔVAE | $-19.72_{\pm 0.03}$ | $6.96_{\pm 0.03}$ | $6.70_{\pm 0.00}$ | $0.27_{\pm 0.03}$ | $0.05_{\pm 0.02}$ | 5 out of 5 | $0.12_{\pm 0.06}$ |
| *S*-VAE | $-0.26_{\pm 0.00}$ | $8.33_{\pm 0.03}$ | $7.90_{\pm 0.03}$ | $0.43_{\pm 0.03}$ | $0.09_{\pm 0.01}$ | 5 out of 5 | $0.20_{\pm 0.03}$ |
| **Spherical harmonics of degree** $L = 9$ | | | | | | | |
| ΔVAE | $-23.32_{\pm 0.02}$ | $6.82_{\pm 0.00}$ | $6.69_{\pm 0.00}$ | $0.13_{\pm 0.00}$ | $0.01_{\pm 0.01}$ | 5 out of 5 | $0.01_{\pm 0.00}$ |
| *S*-VAE | $-0.31_{\pm 0.00}$ | $8.23_{\pm 0.04}$ | $7.57_{\pm 0.31}$ | $0.66_{\pm 0.02}$ | $0.12_{\pm 0.01}$ | 1 out of 5 | $0.29_{\pm 0.02}$ |
| **Spherical harmonics of degree** $L = 11$ | | | | | | | |
| ΔVAE | $-33.06_{\pm 0.05}$ | $12.90_{\pm 0.04}$ | $12.69_{\pm 0.00}$ | $0.18_{\pm 0.04}$ | $0.05_{\pm 0.02}$ | 5 out of 5 | $0.09_{\pm 0.04}$ |
| *S*-VAE | $-0.36_{\pm 0.00}$ | $9.07_{\pm 0.10}$ | $8.33_{\pm 0.11}$ | $0.73_{\pm 0.03}$ | $0.15_{\pm 0.03}$ | 1 out of 5 | $0.35_{\pm 0.05}$ |

the training. We performed 5 experiments where the absolute value of the degree before training was not 1, whereas the absolute value of the degree after all training was 1. In particular, even though we share the opinion that topological obstructions might hamper training [8, 4], for the ΔVAE the obstruction to the degree can be overcome.

## 5. Discussion

We derive a second order expansion of the heat kernel on the unit sphere $\mathcal{S}^2$ by using the theoretical result of [32], and use it as approximation in the ΔVAE loss function. Though the effect of such higher order approximation in the performance of ΔVAE is not studied yet.

Our algorithm for degree computation could be generalized to higher dimensional sphere $\mathcal{S}^d$ with $d > 2$, but due to the curse of dimensionality, practical computation is most likely only feasible in very low dimensions: for a $d$-dimensional manifold and a discretization length $\delta$, the number of faces needed in the triangulation scales as $\delta^{-d}$.

The amount of semisupervision is relatively high in our experiments. For lower degree spherical harmonics ($L = 1, 3, 5$), the amount of semisupervision can be reduced drastically, although we have not yet performed a systematic study.

## 6. Conclusion

We evaluate to what extent the ΔVAE can capture topological properties or disentangle generating factors, as measured by the LSBD score, and as expressed by a new discrete diagnostic for disentanglement: the degree of the encoder. We use the encoder degree as a means to gain more insight in the training behavior.

First, we obtain relatively small LSBD scores, which expresses that the ΔVAE indeed can capture or disentangle the latent rotational factor relatively well. In comparison with the *S*-VAE, we find that the *S*-VAE typically obtains better log-likelihood scores, while the reconstruction error and LSBD score are a bit better for the ΔVAE.

Secondly, we implemented an algorithm for computing the topological degree of the encoder and find that even though the encoder is typically initialized with degree 0, this degree can change and after training the encoder indeed has degree of $\pm 1$, which means that the encoder is at least homotopic to a homeomorphism and that the learned spherical representation preserves the topological structure of the dataset at least up to a homotopy. In particular, we find that the sphere in latent space is completely

covered by the image of the data manifold.

## References


[1] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE transactions on pattern analysis and machine intelligence 35 (2013) 1798–1828.

[2] I. Higgins, D. Amos, D. Pfau, S. Racaniere, L. Matthey, D. Rezende, A. Lerchner, Towards a definition of disentangled representations, arXiv preprint arXiv:1812.02230 (2018).

[3] L. Tonnaer, L. A. P. Rey, V. Menkovski, M. Holenderski, J. Portegies, Quantifying and learning linear symmetry-based disentanglement, in: International Conference on Machine Learning, PMLR, 2022, pp. 21584–21608.

[4] L. Falorsi, P. De Haan, T. R. Davidson, N. De Cao, M. Weiler, P. Forré, T. S. Cohen, Explorations in homeomorphic variational auto-encoding, ICML18 Workshop on Theoretical Foundations and Applications of Deep Generative Models (2018).

[5] P. de Haan, L. Falorsi, Topological constraints on homeomorphic auto-encoding, NeurIPS 2018 workshop on Integration of Deep Learning Theories (2018).

[6] T. R. Davidson, L. Falorsi, N. De Cao, T. Kipf, J. M. Tomczak, Hyperspherical variational auto-encoders, 34th Conference on Uncertainty in Artificial Intelligence (UAI-18) (2018).

[7] L. A. Perez Rey, V. Menkovski, J. Portegies, Diffusion variational autoencoders, in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, 2020. doi:10.24963/ijcai.2020/375.

[8] B. Esmaeili, R. Walters, H. Zimmermann, J.-W. van de Meent, Topological obstructions and how to avoid them, Advances in Neural Information Processing Systems 36 (2024).

[9] D. P. Kingma, M. Welling, Auto-encoding Variational Bayes, in: Y. Bengio, Y. LeCun (Eds.), 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings, 2014. URL: http://arxiv.org/abs/1312.6114.

[10] D. J. Rezende, S. Mohamed, D. Wierstra, Stochastic backpropagation and approximate inference in deep generative models, in: International conference on machine learning, PMLR, 2014, pp. 1278–1286.

[11] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, A. Lerchner, Understanding disentangling in beta-VAE, Learning Disentangled Representations: from Perception to Control Workshop, 2017 (2017).

[12] J. Cha, J. Thiyagalingam, Orthogonality-enforced latent space in autoencoders: an approach to learning disentangled representations, in: International Conference on Machine Learning, PMLR, 2023, pp. 3913–3948.

[13] H. Kim, A. Mnih, Disentangling by factorising, in: International conference on machine learning, PMLR, 2018, pp. 2649–2658.

[14] M. D. Hoffman, M. J. Johnson, Elbo surgery: yet another way to carve up the variational evidence lower bound, in: Workshop in Advances in Approximate Bayesian Inference, NIPS, volume 1, 2016.

[15] A. Klushyn, N. Chen, R. Kurle, B. Cseke, P. van der Smagt, Learning hierarchical priors in vaes, Advances in neural information processing systems 32 (2019).

[16] J. Tomczak, M. Welling, Vae with a vampprior, in: International conference on artificial intelligence and statistics, PMLR, 2018, pp. 1214–1223.

[17] C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, O. Winther, Ladder variational autoencoders, Advances in neural information processing systems 29 (2016).

[18] C. Chadebec, E. Thibeau-Sutre, N. Burgos, S. Allassonnière, Data augmentation in high dimensional low sample size setting using a geometry-based variational autoencoder, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (2022) 2879–2896.

[19] Z. Ding, Y. Xu, W. Xu, G. Parmar, Y. Yang, M. Welling, Z. Tu, Guided variational autoencoder for

disentanglement learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 7920–7929.
[20] I. Huh, J. M. Choe, Y. KIM, D. Kim, et al., Isometric quotient variational auto-encoders for structure-preserving representation learning, Advances in Neural Information Processing Systems 36 (2024).
[21] K. Do, T. Tran, Theory and evaluation metrics for learning disentangled representations, International Conference on Learning Representations, ICLR 2020 (2020).
[22] S. Van Steenkiste, F. Locatello, J. Schmidhuber, O. Bachem, Are disentangled representations helpful for abstract visual reasoning?, Advances in neural information processing systems 32 (2019).
[23] M.-A. Carbonneau, J. Zaidi, J. Boilard, G. Gagnon, Measuring disentanglement: A review of metrics, IEEE transactions on neural networks and learning systems (2022).
[24] A. Sepliarskaia, J. Kiseleva, M. de Rijke, How not to measure disentanglement, in: ICML Workshop on Theoretic Foundation, Criticism, and Application Trend of Explainable AI, 2021.
[25] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, A. Lerchner, beta-VAE: Learning basic visual concepts with a constrained variational framework., ICLR (Poster) 3 (2017).
[26] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, O. Bachem, Challenging common assumptions in the unsupervised learning of disentangled representations, in: international conference on machine learning, PMLR, 2019, pp. 4114–4124.
[27] R. Suter, D. Miladinovic, B. Schölkopf, S. Bauer, Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness, in: International Conference on Machine Learning, PMLR, 2019, pp. 6056–6065.
[28] A. Hatcher, Algebraic topology, Cambridge University Press, 2005.
[29] J. Milnor, Topology from the differentiable viewpoint, univ, Press of Virginia, Charlottesville 1990 (1965).
[30] T. Kaczynski, K. M. Mischaikow, M. Mrozek, Computational homology, volume 157, Springer, 2004.
[31] M. R. Ravelonanosy, V. Menkovski, J. W. Portegies, Topological degree as a discrete diagnostic for disentanglement, with applications to the ΔVAE, https://arxiv.org/abs/2409.01303 (2024).
[32] V. Menkovski, J. W. Portegies, M. R. Ravelonanosy, Small time asymptotics of the entropy of the heat kernel on a riemannian manifold, Applied and Computational Harmonic Analysis 71 (2024). URL: https://www.sciencedirect.com/science/article/pii/S1063520324000198.
[33] T. Bröcker, T. tom Dieck, Representations of compact lie groups, Graduate Texts in Mathematics (1985).
[34] M. A. Blanco, M. Flórez, M. Bermejo, Evaluation of the rotation matrices in the basis of real spherical harmonics, Journal of Molecular structure: THEOCHEM 419 (1997) 19–27.
[35] S. Harmonics, Claus mulier, Lecture Notes in Mathematics (LNM)) 17 (1966).
[36] Y. Burda, R. B. Grosse, R. Salakhutdinov, Importance weighted autoencoders, in: Y. Bengio, Y. LeCun (Eds.), 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings, 2016. URL: http://arxiv.org/abs/1509.00519.