Verifying the Union of Manifolds Hypothesis for Image Data

Final Project Report – Pattern Recognition

National Technical University of Athens | Electrical and Computer Engineering

Team 7

Nikolaos Giannakakis el10191@mail.ntua.gr
Panagiotis Karampinas el16170@mail.ntua.gr
Dimitrios Neroutsos el18934@mail.ntua.gr
Stylianos Zarifis el20435@mail.ntua.gr
Marios Rozos el20051@mail.ntua.gr

All authors contributed equally to this work.

February 27, 2024

Abstract

In recent years, the remarkable success of deep machine learning techniques is evident, despite the large dimensions of the data. One way to interpret this success is the manifold hypothesis according to which the data lie in spaces of much lower dimensions. Brown et al. [1] extended this hypothesis by considering that data from different classes lie in disjoint spaces of different dimensions and attempted to confirm this idea experimentally on image data. In this work, we attempt to confirm two basic experiments from [1], interpret them, and examine extensions for the further confirmation of this hypothesis. Our code is available at `https://github.com/SteliosZarifis/patrec_project`.

1 Introduction

According to the Manifold Hypothesis (MH) [2], high-dimensional data often lie on some unknown manifold of lower dimension, which is called the intrinsic dimension. The existence of such low-dimensional structure is experimentally confirmed by Pope et al. [3] by calculating estimators for the intrinsic dimension of common image datasets. In Brown et al. [1], the original hypothesis is extended and the Union of Manifolds Hypothesis (UoMH) is formulated, according to which high-dimensional data do not lie on a single manifold, but on a disjoint union of manifolds of different intrinsic dimensions, and the hypothesis is experimentally confirmed on known image datasets. The experimental confirmation of the UoMH includes two requirements: confirming that the various classes lie on different manifolds and that these manifolds have different intrinsic dimensions.

2 Basic Concepts

We consider $D = \{x_i\}_{i=1}^n$, where x_i are i.i.d. samples from some distribution P^* of a high-dimensional space $X = \mathbb{R}^D$.

2.1 Pushforward DGMs

A Pushforward DGM is a Deep Generative Model (DGM) whose samples X are produced as follows:

$$Z \sim P_Z \text{and} X = G(Z), \tag{1}$$

where P_Z is a distribution on some latent space Z and $G : Z \to X$ a neural network. Some known DGMs are Pushforward DGMs, such as: Variational Autoencoders (VAEs), Normalizing Flows (NFs), Generative Adversarial Networks (GANs), Wasserstein Autoencoders (WAEs), etc.

2.2 Estimation of Intrinsic Dimension

For the estimation of the intrinsic dimensions of the various classes, we follow [3], in which the estimator of Levina & Bickel with the correction of MacKay & Ghahramani is used:

$$\hat{d}_k := \left(\frac{1}{n(k-1)} \sum_{i=1}^{n} \sum_{j=1}^{k-1} \log \frac{T_k(x_i)}{T_j(x_i)} \right)^{-1}, \tag{2}$$

where $T_j(x)$ is the Euclidean distance of x from the j-th neighbor in $D \setminus \{x\}$, and k is a hyperparameter.

2.3 Non-Connectedness

In [1] it is proven that:

Proposition 1 (Connectedness preservation)**.** *Let Z and X be topological spaces and $G : Z \to X$ a continuous mapping. Suppose Z and X are measurable spaces equipped with their respective Borel σ-algebras, and let P_Z be a probability measure on Z such that* $\text{supp}(P_Z)$ *is connected and* $P_Z(\text{supp}(P_Z)) = 1$*. Then the support of the pushforward measure $G_\# P_Z$ is connected.*

The significance of Proposition 1 is the following: if the support of the data is not connected, i.e., if the UoMH holds, the pushforward models cannot effectively model the data. We will exploit this to test whether the support of the data distribution is disconnected.

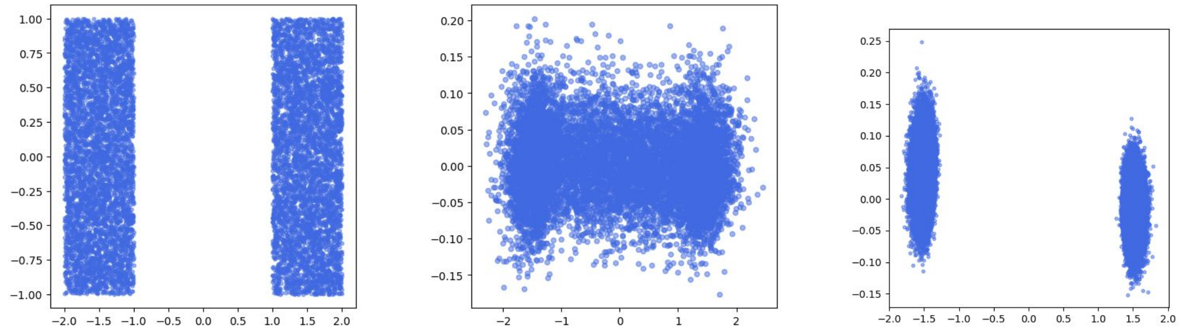
3 Experimental Verification

3.1 Non-Connectedness

To show experimentally that $\text{supp}(P^*)$ is non-connected, we use the disconnected DGMs introduced in [1], in which instead of training a single pushforward model (P_Z, G) on D, we split the dataset into classes $D = \bigsqcup_{\ell=1}^{L} D_\ell$, where D_ℓ is the dataset in the ℓ-th class, and then train a DGM $(P_Z^{(\ell)}, G_\ell)$ on each D_ℓ. More details about the algorithm are in Appendix A.

Based on Proposition 1, Brown et al. claim that if we observe some improvement of the disconnected DGMs over the usual DGMs, then the most likely explanation is that $\text{supp}(P^*)$ is non-connected [1]. This would mean that the data do not lie on a single manifold, but on different, disjoint ones.

A first experiment to see how VAEs fail when the data lie on some non-connected manifold is to generate a synthetic dataset, which consists of samples from two non-overlapping rectangles.



(a) Ground Truth

(b) VAE

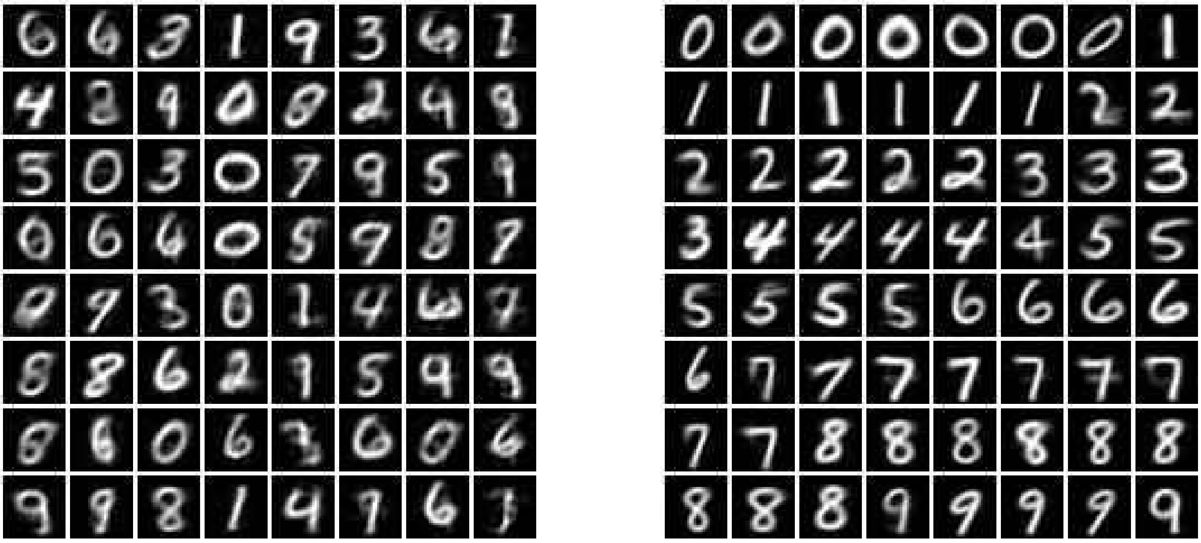
(c) D-VAE

Figure 1: Samples from a synthetic dataset and samples generated by VAE and D-VAE.

We observe that the VAE fails to produce a non-connected support for the data, unlike the D-VAE.

We train both standard and disconnected DGMs on four benchmark datasets and calculate the resulting FID score. We follow the same process a total of three times for each dataset and calculate the mean and standard deviation of the FID score. The DGMs we used are VAEs and D-VAEs (disconnected VAEs) on classes of the datasets. The datasets we used are: MNIST, FMNIST, SVHN and CIFAR-10, while the hyperparameters we used are from [4]. The tables with the FID score measurements from [1] and ours are shown in Tables 1 to 4.

It is worth noting that a smaller FID score equates to better results. We also provide images generated by the VAEs and D-VAEs we trained on the various datasets (see Figures 2 to 5).



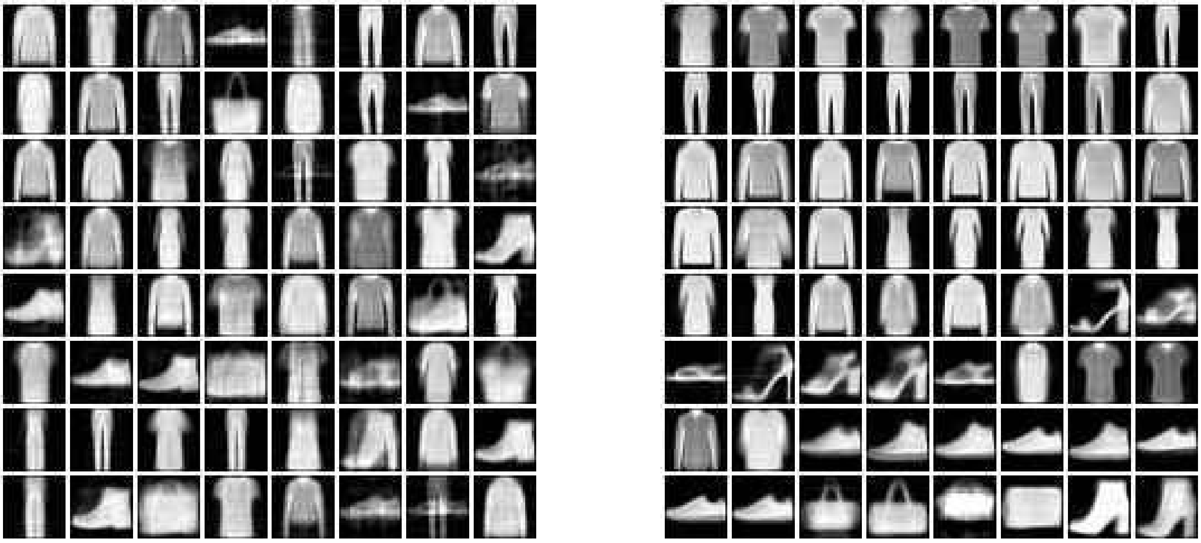
(a) VAE

(b) D-VAE

Figure 2: Generation of images from MNIST

As can be seen from both the above tables and the images we generated, in general D-VAEs yield better results than VAEs, which agrees with the conclusions of [1]. However, we observe that the results on CIFAR-10, which is the largest of the four datasets, are not satisfactory, while in this particular case the difference between VAE and D-VAE is relatively small.

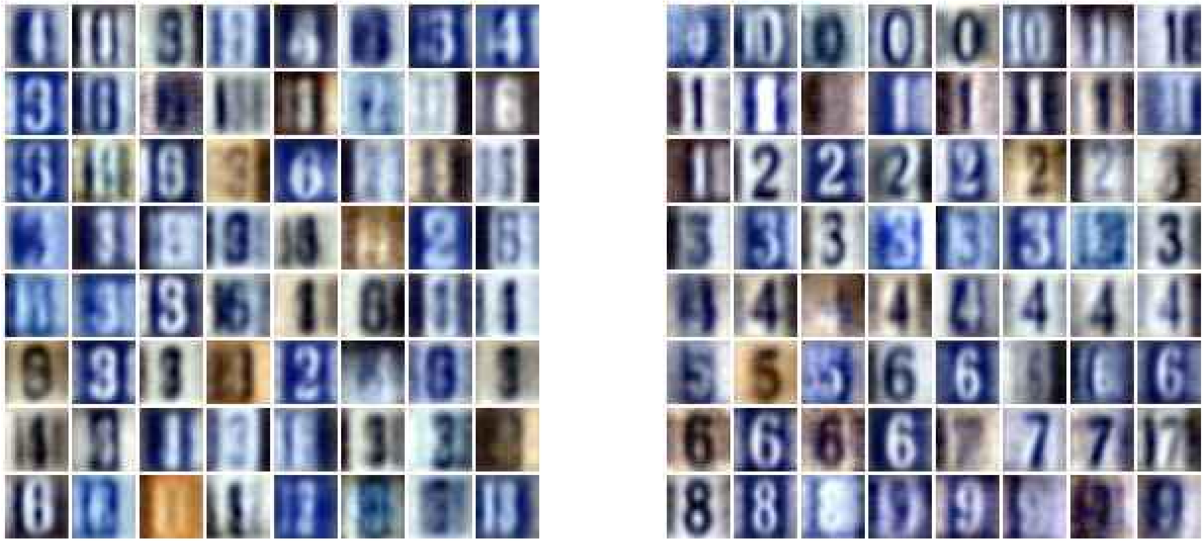
Next, we examine the performance of WAEs (Wasserstein Autoencoders, [5]) and their disconnected version, D-WAEs. The results from [1] and ours are shown in Tables 1 to 4:



(a) VAE

(b) D-VAE

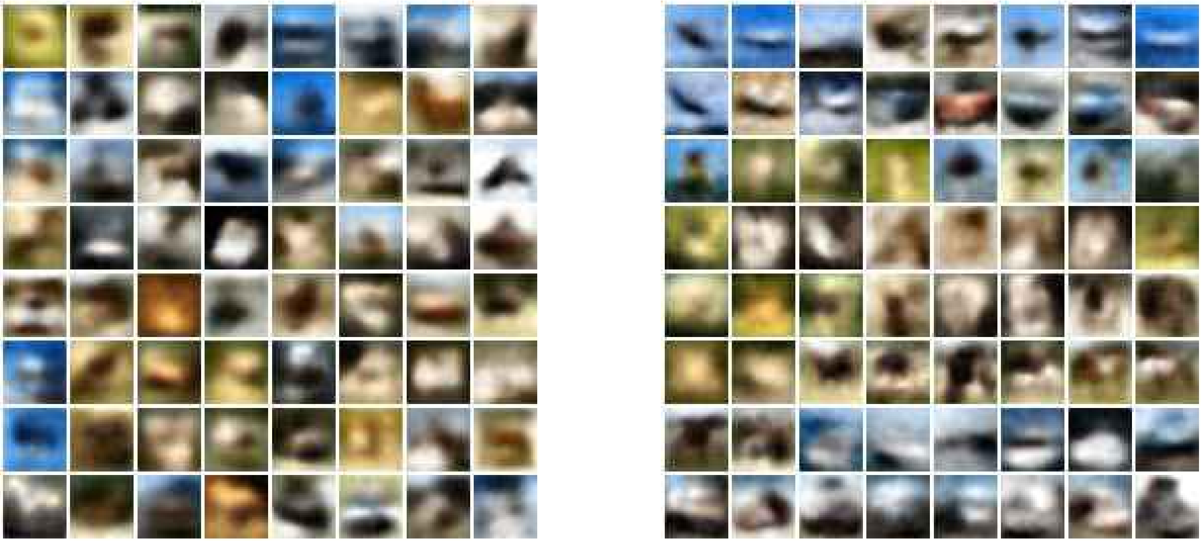
Figure 3: Generation of images from FMNIST



(a) VAE

(b) D-VAE

Figure 4: Generation of images from SVHN



(a) VAE

(b) D-VAE

Figure 5: Generation of images from CIFAR-10

Table 1: FID scores for VAE and D-VAE from [1]

Model	MNIST	FMNIST	SVHN	CIFAR-10
VAE	110.7 ± 1.7	100.1 ± 1.7	93.2 ± 0.2	213.8 ± 1.3
D-VAE (random)	155.4 ± 1.3	125.6 ± 0.6	156.6 ± 0.7	232.3 ± 0.8
D-VAE (classes)	81.5 ± 0.7	87.7 ± 1.0	86.4 ± 2.0	202.4 ± 0.6

Table 2: FID scores for VAE and D-VAE from us

Model	MNIST	FMNIST	SVHN	CIFAR-10
VAE	145.23 ± 1.94	127.12 ± 1.18	95.89 ± 2.35	212.36 ± 1.98
D-VAE (classes)	77.99 ± 0.35	82.58 ± 0.30	84.59 ± 0.89	203.16 ± 0.57

Table 3: FID scores for WAE and D-WAE from [1]

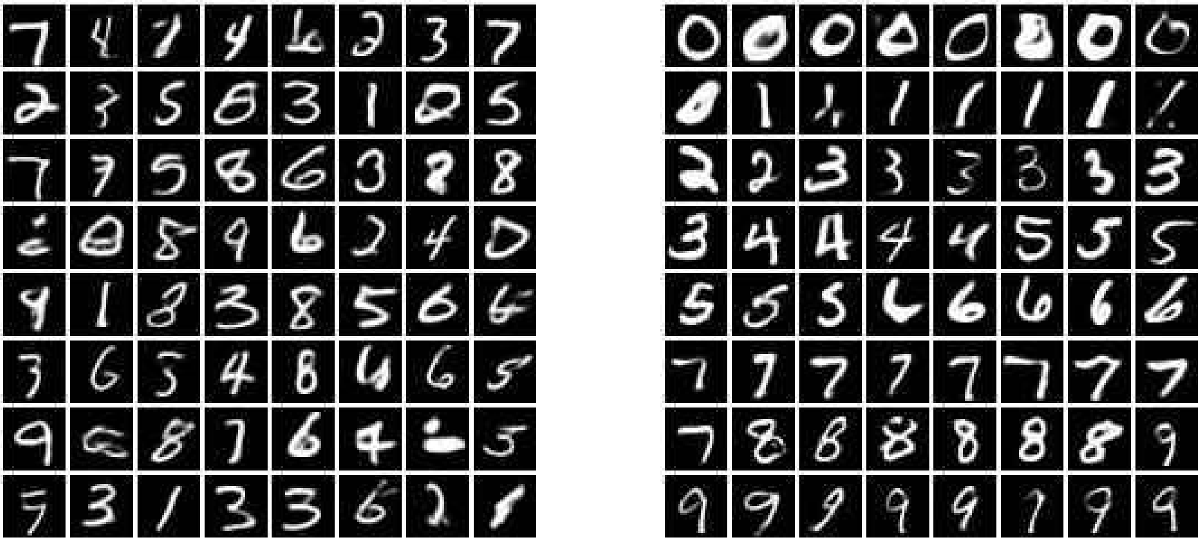
Model	MNIST	FMNIST	SVHN	CIFAR-10
WAE	19.1 ± 2.1	51.2 ± 2.4	89.8 ± 12.2	146.7 ± 1.2
D-WAE (random)	23.7 ± 2.8	61.1 ± 3.6	77.6 ± 1.7	154.7 ± 0.5
D-WAE (classes)	13.5 ± 0.2	36.8 ± 1.4	83.7 ± 15.2	133.3 ± 0.5

Table 4: FID scores for WAE and D-WAE from us

Model	MNIST	FMNIST	SVHN	CIFAR-10
WAE	14.48 ± 1.41	46.44 ± 2.71	74.84 ± 18.39[1]	145.88 ± 2.44
D-WAE (classes)	13.63 ± 0.81	42.24 ± 9.75	62.14 ± 5.73	134.75 ± 1.53

[1] Only two runs were performed for this experiment (SVHN, WAE).

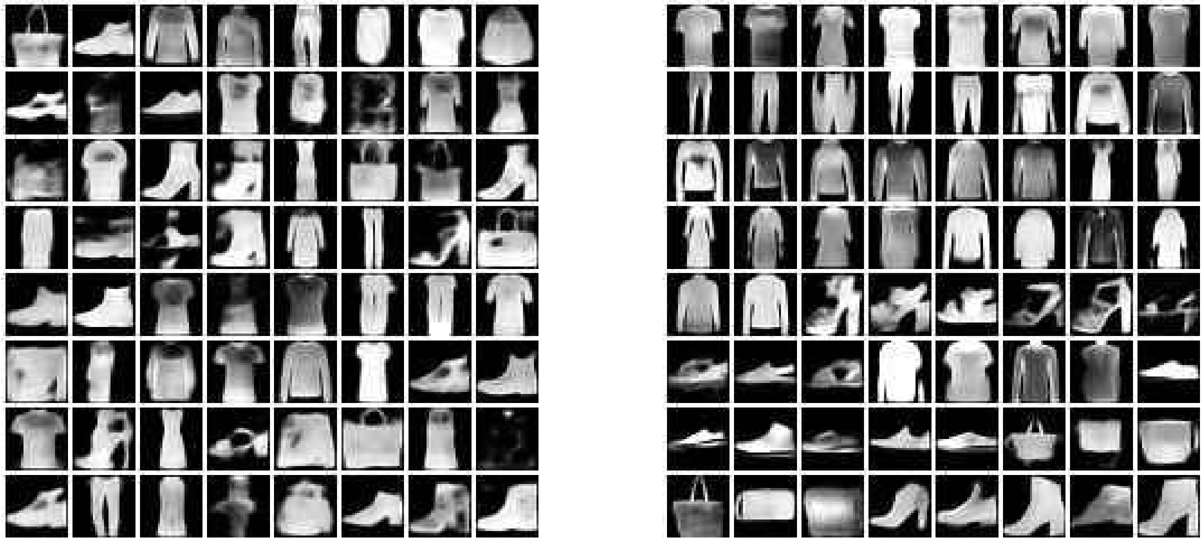
The results we arrive at are similar to those in [1] and in every dataset the D-WAEs appear to perform better than the corresponding WAEs, thus supporting that supp(P^*) is indeed non-connected for these datasets. We also provide images generated by the WAEs and D-WAEs we trained on the various datasets (see Figures 6 to 9).



(a) WAE

(b) D-WAE

Figure 6: Generation of images from MNIST



(a) WAE

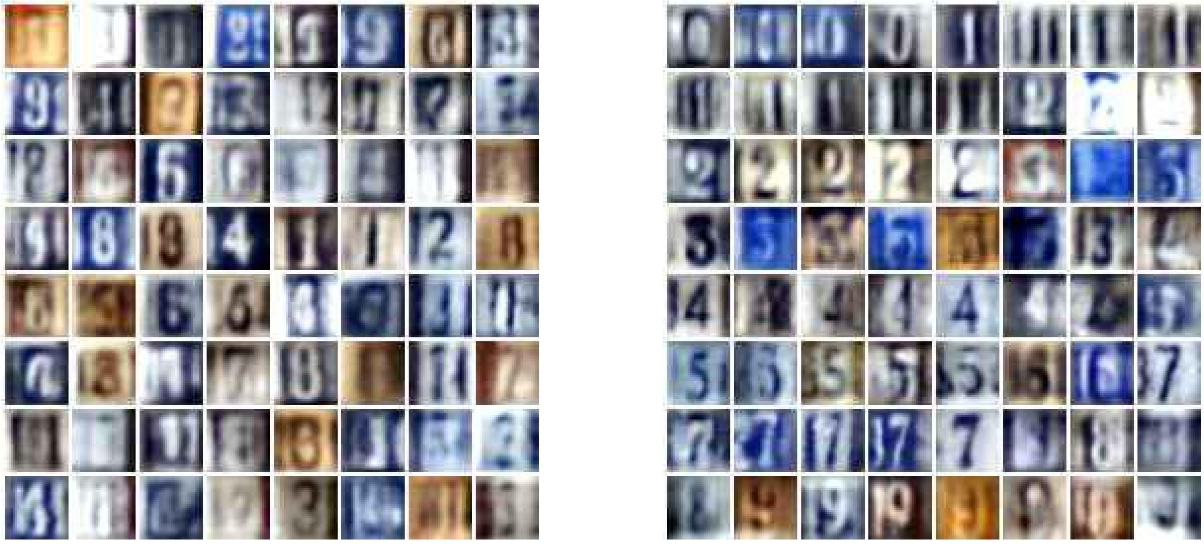
(b) D-WAE

Figure 7: Generation of images from FMNIST

3.2 Different Intrinsic Dimensions

As in [1], we want to confirm that the different classes have different intrinsic dimensions. To do this, we calculate the estimator $\hat{d}_k$ of the intrinsic dimension for each of the different classes for various values of the hyperparameter k, based on [6]. The datasets we use are MNIST, FMNIST, SVHN and CIFAR-10.

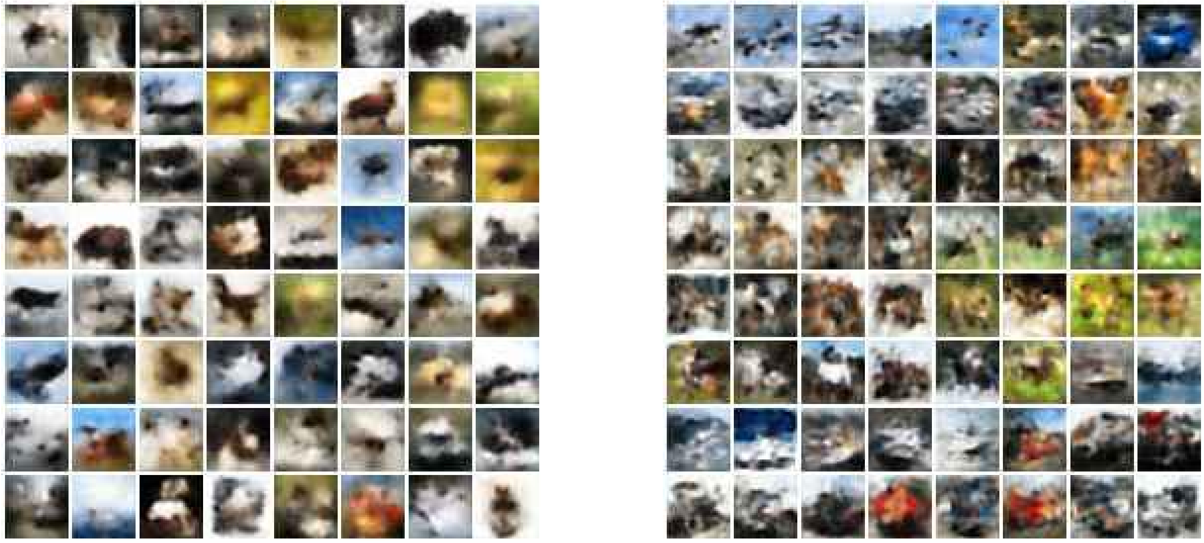
In Appendix B we provide the diagrams from the estimates of the intrinsic dimensions for the classes of each of the image datasets we used for various values of the hyperparameter k. From the diagram of Fihure 10 and the diagrams in Appendix B we observe that the results



(a) WAE

(b) D-WAE

Figure 8: Generation of images from SVHN



(a) WAE

(b) D-WAE

Figure 9: Generation of images from CIFAR-10

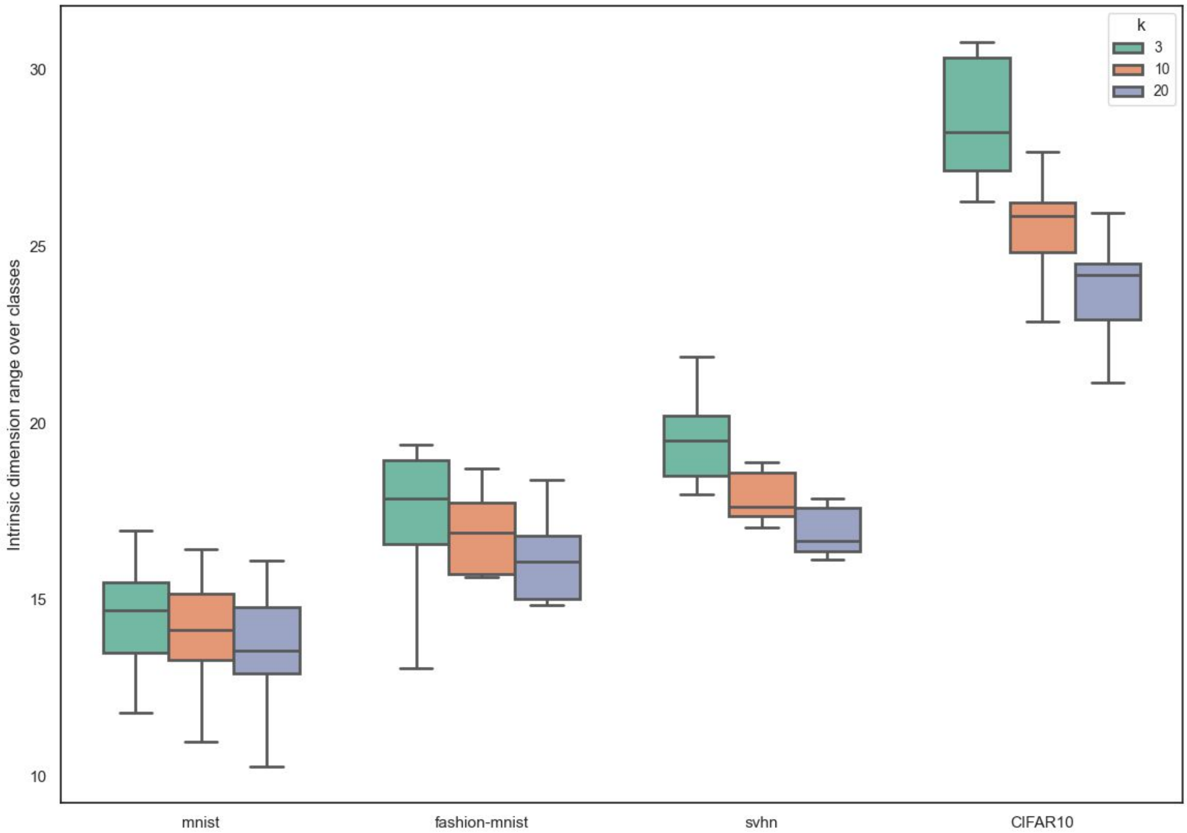


Figure 10: Boxplots with the dispersion of the estimation of the intrinsic dimension of the classes for various datasets and various values of k.

remain consistent for various values of k, so the conclusions we reach are not due to some specific hyperparameter value. Also, for each dataset - with the exception of SVHN - we observe great variety in the values of the estimated intrinsic dimensions, which confirms the validity of the UoMH on image datasets.

4 Application of the UoMH in Classification Problems

According to Pope et al. [3] datasets with larger intrinsic dimension are more difficult to classify. Brown et al. [1] then attempted to exploit the UoMH to improve classification by training the ResNet-18 model in two different ways: initially using cross entropy loss and then using cross entropy loss with weights for each class proportional to the intrinsic dimension. In this way, more emphasis is given to classes with larger intrinsic dimension, which are more difficult to classify. The accuracy results for ResNet-18 on CIFAR-100 from [1] (5 runs) and from us (1 run, due to limited resources) are shown in Table 5.

Table 5: Test accuracy of ResNet-18 on CIFAR-100 for 5 executions from [1] and for 1 execution from us.

Weights	Test accuracy (Brown et al.)	Test accuracy (Us)
Standard	61.38% ± 0.17%	49.13%
Proportional to the intrinsic dimension	61.77% ± 0.20%	47.35%

We also provide a diagram depicting the accuracy of each class with respect to the estimate of the intrinsic dimension, as well as the resulting regression line, in Figure 11.

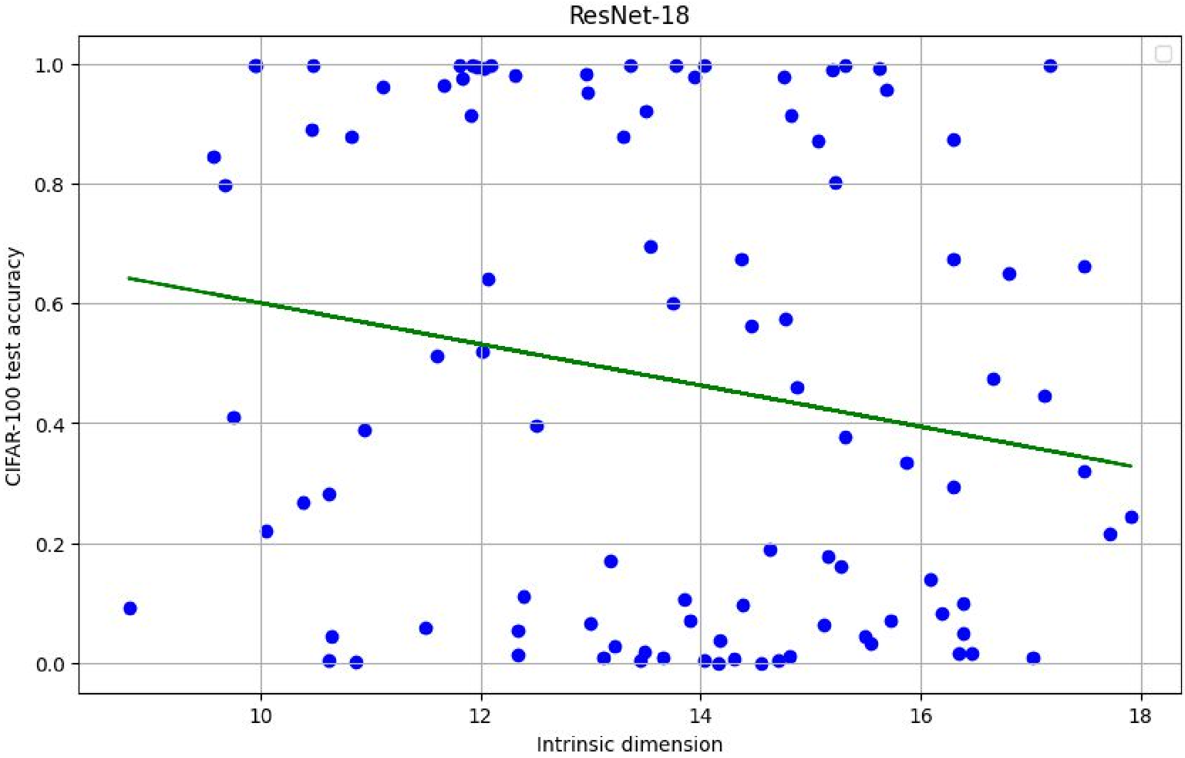


Figure 11: Diagram of intrinsic dimensions - test accuracy for ResNet-18 on CIFAR100 with weighted cross entropy loss. For the model we have correlation coefficient -0.192 with p-value 0.0554.

From the regression line, it follows that the correlation coefficient is -0.192 with p-value 0.0554, so there is some correlation between accuracy and intrinsic dimensions. Brown et al. got a very small improvement, which however resulted simply from appropriate modification of the cross entropy loss. On the contrary, we had smaller accuracy percentages in both cases, with the standard cross entropy loss giving better results. However, this is because we trained the model for fewer runs and epochs, due to large limitations in computational resources.

5 Clustering and D-VAE

The algorithm for training disconnected DGMs in [1] includes clustering into L classes, so that each DGM corresponding to a class is trained. In the previous experiments we assumed that for the training algorithm of the disconnected DGMs the classes coincide with the categories into which each dataset is divided. However, we want to investigate what happens in the case where L is smaller than the number of classes in the dataset and if we will gain some advantage in performance. This could be interpreted as some classes of data having similarities between them, so the manifolds of the corresponding classes are connected. The motivation for this is the following example in the MNIST dataset for the digits 1 and 7: they are often written in a similar way resulting in the difference sometimes not being distinguished even by humans.

To examine this case, we train D-VAEs for different values of the parameter L which corresponds to the number of classes into which the dataset will be divided. We examine the MNIST dataset with the k-means algorithm and the AudioMNIST dataset with k-means and agglomerative clustering. In AudioMNIST, due to the small size of the dataset, we were unable to achieve further group separation with agglomerative clustering. The results are shown in

Table 6: FID scores for D-VAE with different values of the parameter L for the MNIST and the AudioMNIST.

L	MNIST k-means	AudioMNIST k-means	AudioMNIST agglomerative
3	103.02	77.58	78.75
4	95.68	78.13	72.36
5	89.46	79.49	78.44
6	86.36	79.75	75.98
7	82.11	81.70	
8	79.86	80.34	
9	78.29	74.18	
10	78.22	73.22	

Table 7: FID scores for D-VAE for the best L for the MNIST and the AudioMNIST.

Dataset	Clustering algorithm	FID Score
MNIST	k-means	77.22 ± 0.61
AudioMNIST	k-means	74.09 ± 2.68
AudioMNIST	agglomerative	74.13 ± 0.87

Table 6.

Then for each of the models we perform 3 runs with the best L and the results are shown in Table 7.

In [1] they perform a similar experiment for values of L 7 to 15 and observe that they did not find a large difference in training, without giving any interpretation. In the case of MNIST we observe that for small values of L we have large FID and thus poor performance, so we are close to the case where we have one VAE. We also observe that we have the best result for $L = 10$, but we have quite close performances for $L = 7, 8, 9$. One explanation for this is that indeed the manifolds of some classes are connected to each other, so the corresponding support is connected.

6 Investigation of the UoMH in audio datasets

An important extension we investigate is to examine whether the UoMH holds for audio data. We used AudioMNIST [7], which includes 30000 recordings of digit utterances (0-9) in English with 50 repetitions for each digit from 60 different speakers. Due to limited resources we used a subset of the dataset that includes 14 speakers and 7000 different recordings with a sampling frequency of 22050 Hz.

From this we conclude that the volume of data in raw audio form - even with the subset of the dataset - is quite large and not easily manageable. For this we generated the Mel Spectrograms of the audio files in size 100×64 via the PyTorch library, which we also used for training VAE and D-VAE. The extraction of the Mel Spectrograms was based on the characteristics of a specific vocoder (Vocos, [8]), so that there is the possibility of qualitative control of the samples produced by the DGMs. During the training of the DGMs, min-max normalization of the data to $[0, 1]$ is performed. It is also noted that even with the reduction of the dimensions of the samples, each sample of this dataset is about 6 times larger than the samples of the image datasets (the largest 36×36) studied in the above parts of the work. For this reason, the dimensions of the models were increased. Therefore we use two models whose characteristics are shown in Table 8.

Table 8: Information about the parameters of the two models we examined for the AudioMNIST

Model	Large	Small
Encoder	CNN	CNN
Encoder hidden channels	[512, 256, 128, 64]	[256, 128, 64, 32]
Encoder kernel size	[3, 3, 3, 3]	[3, 3, 3, 3]
Encoder stride	[1,1,1,1]	[1,1,1,1]
Latent Dimension	128	64
Decoder	T-CNN	T-CNN
Decoder hidden channels	[64, 128, 256, 512]	[32, 64, 128, 256]
Decoder kernel size	[3, 3, 3, 3]	[3, 3, 3, 3]
Decoder stride	[1,1,1,1]	[1,1,1,1]

The pipeline we followed is shown in Figure 12:

For the large model we studied two cases regarding the training of D-VAEs by dividing into classes: with respect to digits $0, \dots, 9$ (Digit D-VAE) and with respect to speakers (Speakers D-VAE). The results are shown collectively in Table 9. We note that the values resulted from 3 runs of each model.

Table 9: FID scores for various VAE and D-VAE in AudioMNIST

Model	FID Score
Digits D-VAE (large)	83.87 ± 0.86
Speakers D-VAE (large)	79.84 ± 0.39
VAE(large)	69.11 ± 3.57
Speaker D-VAE (small)	79.12 ±1.02
VAE(small)	62.14 ± 1.47

The most fundamental observation is that the VAEs in both the small and large model have better performance than the corresponding D-VAEs, suggesting that the UoMH does not hold for audio datasets and that the supp(P^*) of the data is connected. One explanation for this is that voice as modality is inherently more connected and the representations of its characteristics are not so sparse. A second observation is that the Speakers D-VAE gave slightly better result than the Digits D-VAE, which indicates the importance of the appropriate categorization of the data into classes, so that the D-VAE has better performance.

7 Conclusions - Future Extensions

In this work, based on the work of Brown et al. [1], we performed various experiments that confirm the validity of the Union of Manifolds Hypothesis for image datasets. We also dealt with an application of the UoMH in classification problems, and while we did not get sufficiently encouraging results, more substantial ways need to be investigated in which the difference in intrinsic dimensions between classes can be exploited for classification.

At the same time, we built upon the experiments concerning the disconnectedness in [1] and examined the results when the training of D-DGMs is done in clusters with smaller number than the number of different classes. For these results we concluded that some manifolds are connected between them and therefore it may interest us to do some clustering before the training of DGMs. However, more and larger datasets need to be examined to lead to more substantial conclusions.

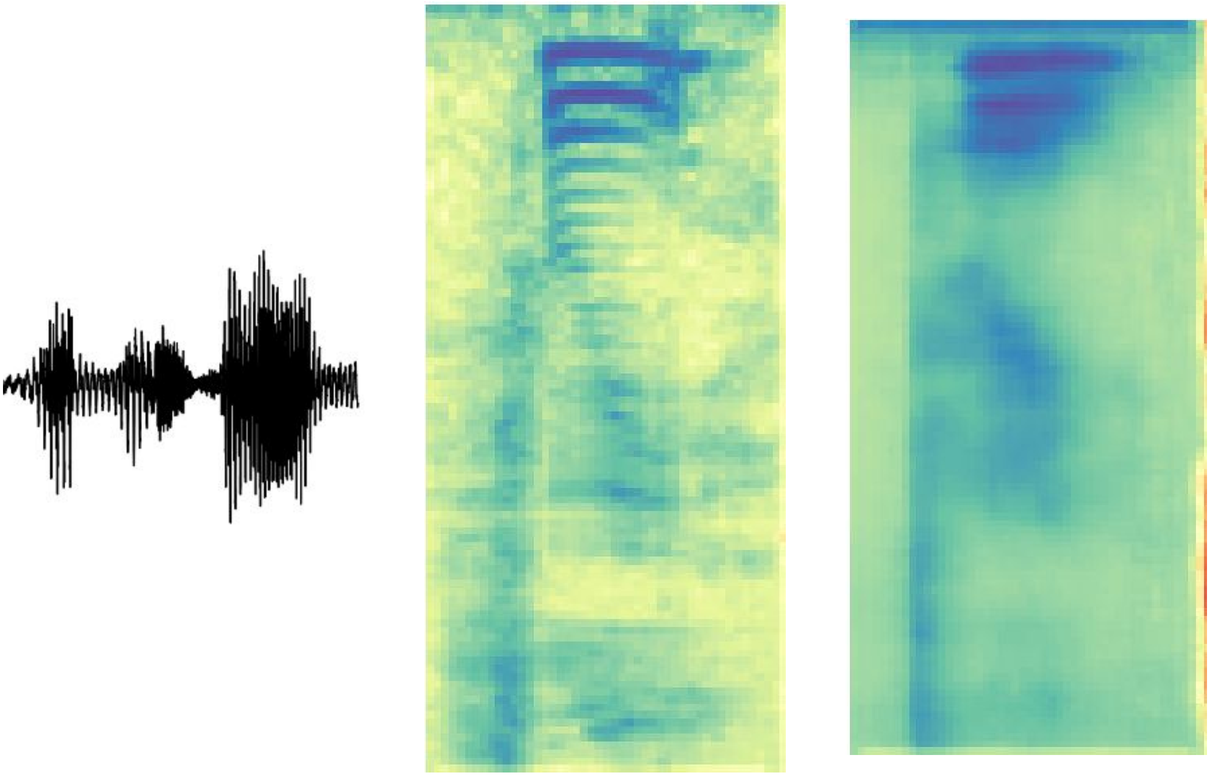
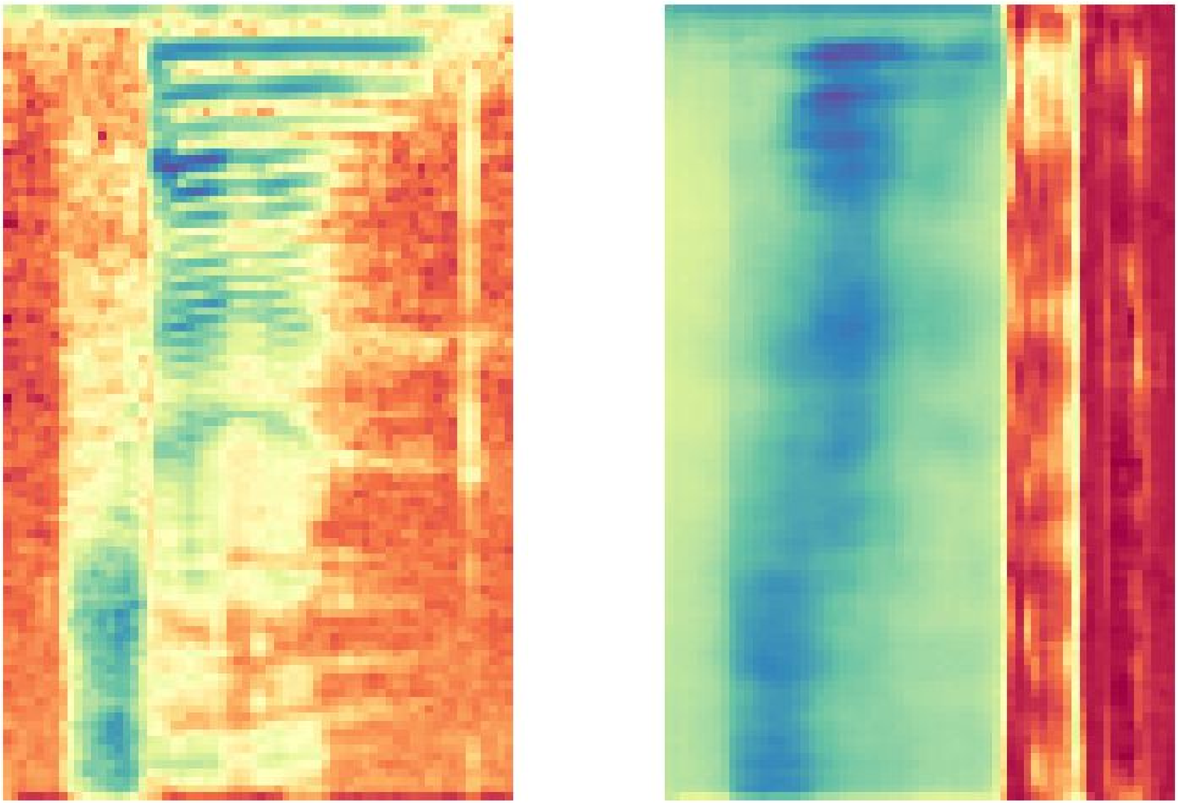


Figure 12: The pipeline includes: (a) loading of the AudioMNIST, (b) Mel Frequency Feature Extraction, so that we convert the sound to spectrogram of size (100 × 64), (c) training of some DGM / disconnected DGM, (d) production of spectrograms, and (e) use of vocoder for production of sound.

Another aspect we explored was whether the UoMH holds in audio datasets. We dealt with AudioMNIST and saw that the classic DGMs achieve better performances, concluding that it does not hold, thus showing that in sound the supports of the data are more connected. It is important, however, to investigate more datasets, but also different kinds of sound, e.g. music datasets. Also of great interest is the investigation of the validity of the UoMH for the modality of text, where we expect more encouraging results, as the text has quite sparse latent representation.



(a) Spectrogram from utterance of the digit 7 in AudioMNIST

(b) Spectrogram that resulted from D-VAE for utterance of the digit 7

Figure 13: Spectrograms of utterances from the AudioMNIST and from D-VAE for the digit 7

References

References

[1] B. C. A. Brown, A. L. Caterini, B. L. Ross, J. C. Cresswell, and G. Loaiza-Ganem, “Verifying the union of manifolds hypothesis for image data,” in *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[2] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[3] P. Pope, C. Zhu, A. Abdelkader, M. Goldblum, and T. Goldstein, “The intrinsic dimension of images and its impact on learning,” in *International Conference on Learning Representations (ICLR)*, 2021.

[4] L. A. Labs, “Union of manifolds hypothesis code repository,” 2023. Accessed: 2025-11-17.

[5] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf, “Wasserstein auto-encoders,” in *International Conference on Learning Representations (ICLR)*, 2018.

[6] P. Pope, “Intrinsic dimension estimation code,” 2021. Accessed: 2025-11-17.

[7] S. Becker, J. Vielhaben, M. Ackermann, K.-R. Müller, S. Lapuschkin, and W. Samek, “Audiomnist: Exploring explainable artificial intelligence for audio analysis on a simple benchmark,” *Journal of the Franklin Institute*, vol. 361, no. 1, pp. 418–428, 2024.

[8] H. Siuzdak, “Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis,” 2023. arXiv:2306.00814.

A Appendix: Algorithms for Disconnected DGMs

Algorithms 1 and 2 for training and sampling for the disconnected DGMs, as exactly mentioned in [1]. For more information see Appendix B.2 of [1].

Algorithm 1 Training of disconnected DGMs

```
Require: clustering_algorithm(·), $D$
Ensure: $\{P_Z^{(\ell)}, G_\ell\}_{\ell=1}^L$
$D_1, \ldots, D_L \leftarrow$ clustering_algorithm($D$)
for $\ell = 1$ to $L$ do
    Potentially initialize $P_Z^{(\ell)}$ and $G_\ell$
    Train $G_\ell$ and potentially $P_Z^{(\ell)}$ on $D_\ell$
end for
```

Algorithm 2 Sampling of disconnected DGMs

```
Require: $m$, trained disconnected DGM $\{P_Z^{(\ell)}, G_\ell\}_{\ell=1}^L$, and corresponding cluster sizes
    $|D_1|, \ldots, |D_L|$.
Ensure: $S$
$S \leftarrow \emptyset$
$(m_1, \ldots, m_L) \sim \text{Multinomial}\left(m, \left(\frac{|D_1|}{\sum_{\ell'=1}^L |D_{\ell'}|}, \ldots, \frac{|D_L|}{\sum_{\ell'=1}^L |D_{\ell'}|}\right)\right)$
for $\ell = 1$ to $L$ do
    for $t = 1$ to $m_\ell$ do
        $Z \sim P_Z^{(\ell)}$
        $X = G_\ell(Z)$
        $S \leftarrow S \cup \{X\}$
    end for
end for
```

B Appendix: Estimates of Intrinsic Dimensions for the Classes of Various Datasets

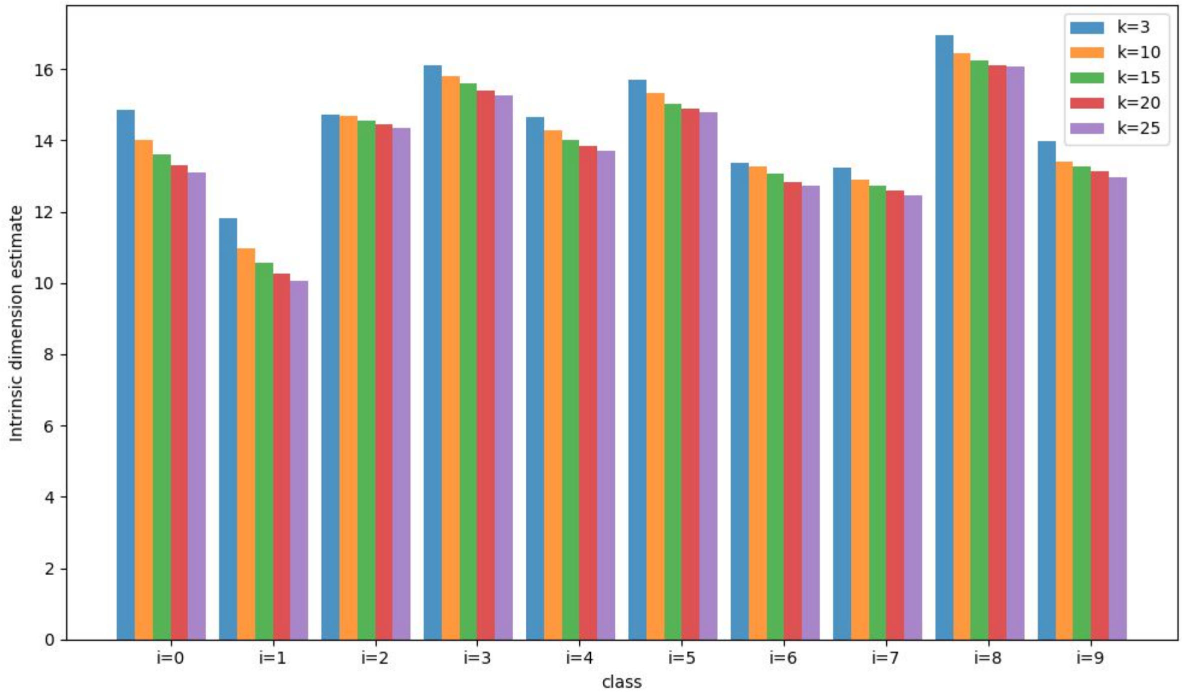


Figure 14: Estimates of intrinsic dimensions for the classes of the MNIST for various values of k.

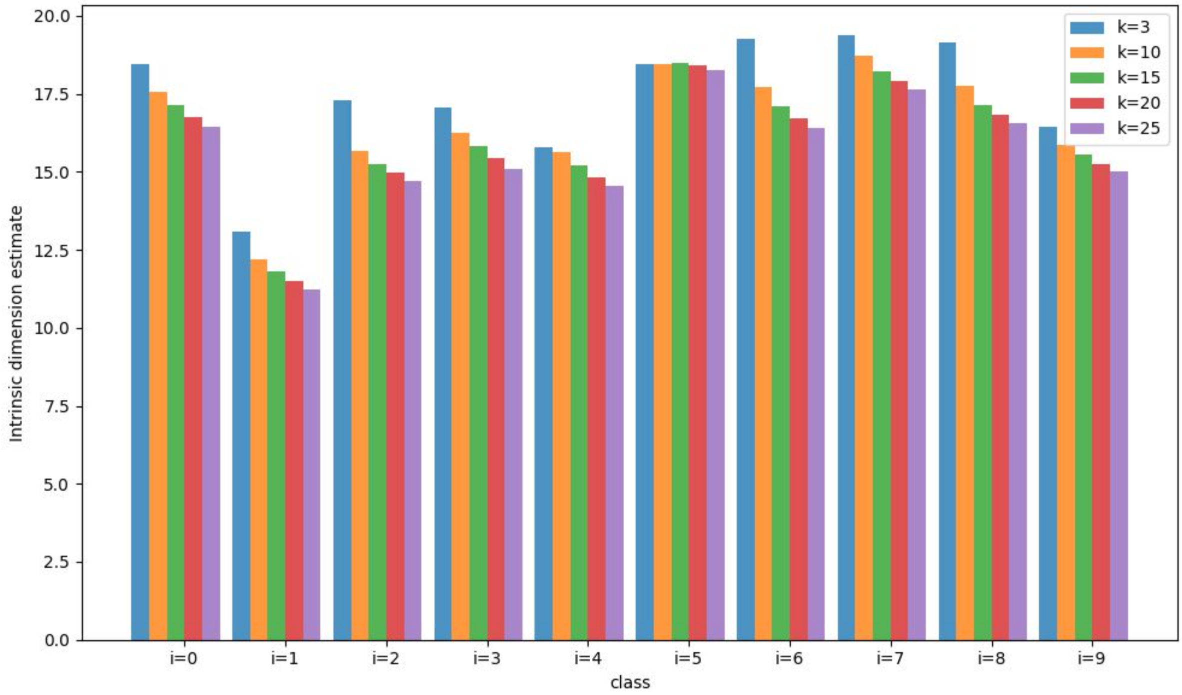


Figure 15: Estimates of intrinsic dimensions for the classes of the FMNIST for various values of k.

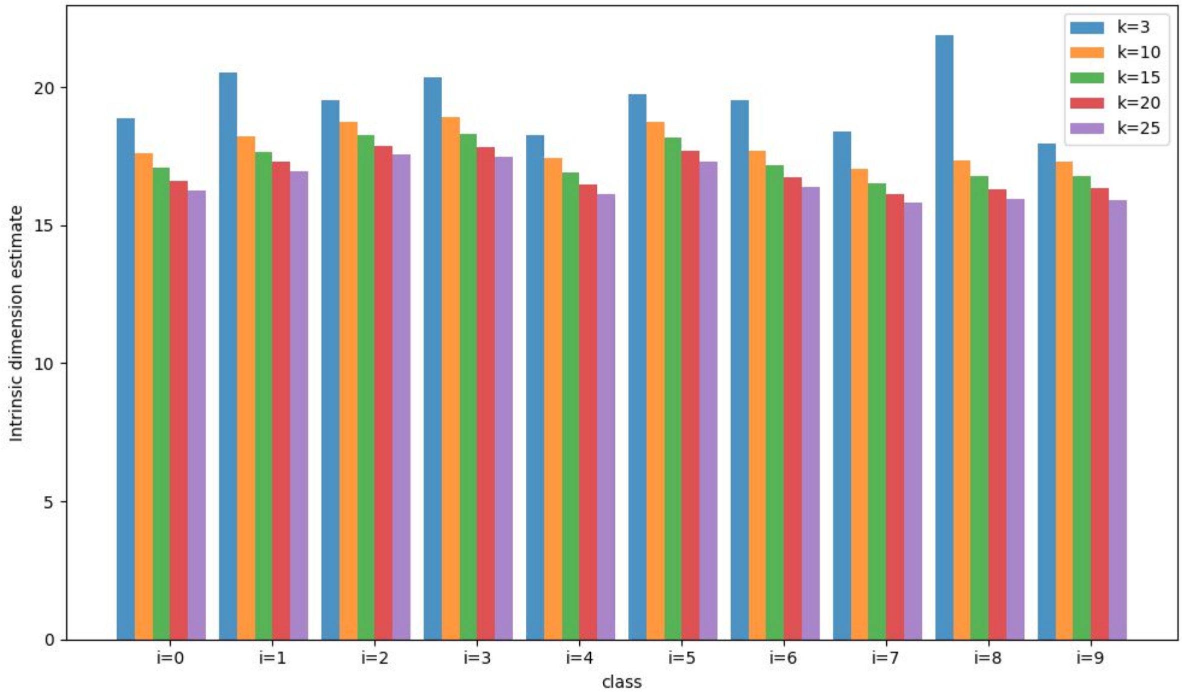


Figure 16: Estimates of intrinsic dimensions for the classes of the SVHN for various values of k.

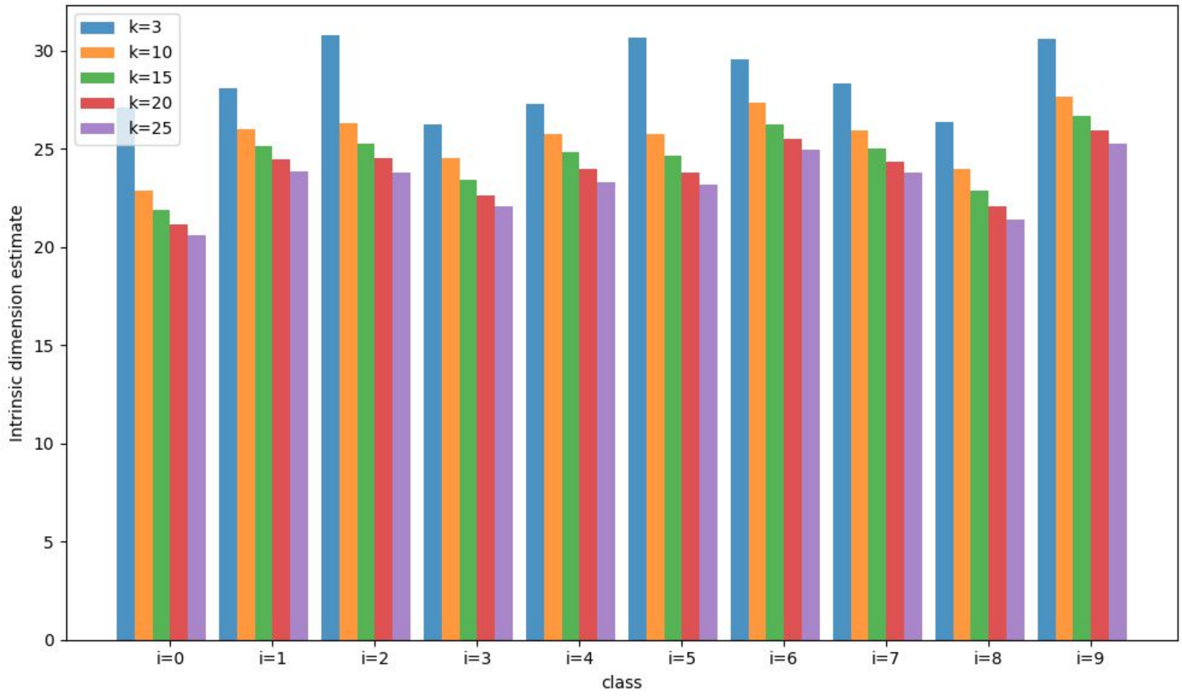


Figure 17: Estimates of intrinsic dimensions for the classes of the CIFAR-10 for various values of k.