Team: Group Li

Course Project
19/12/2023

Team Project on the course “Numerical Linear Algebra”

Reduction of Dimensionality through Active Subspaces approach

David Li

Ignat Melnikov

Kamil Garifullin
Viktoriia Zinkovich

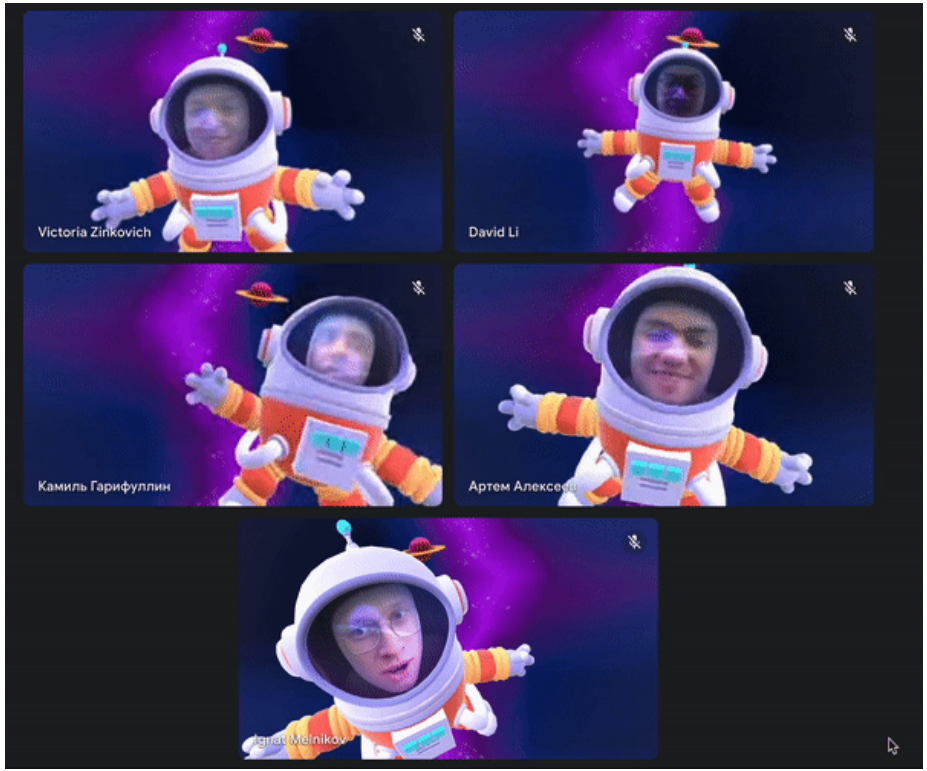
Artem Alekseev

Team: Group Li

Viktoriia Zinkovich
Data Science, MS-1

Kamil Garifullin
Data Science, MS-1

Ignat Melnikov
Data Science, MS-1



David Li
Data Science, MS-1

Artem Alekseev
Data Science, MS-1

Problems

Problems

Problems

Problems

Motivation for our research

Problems

Problems

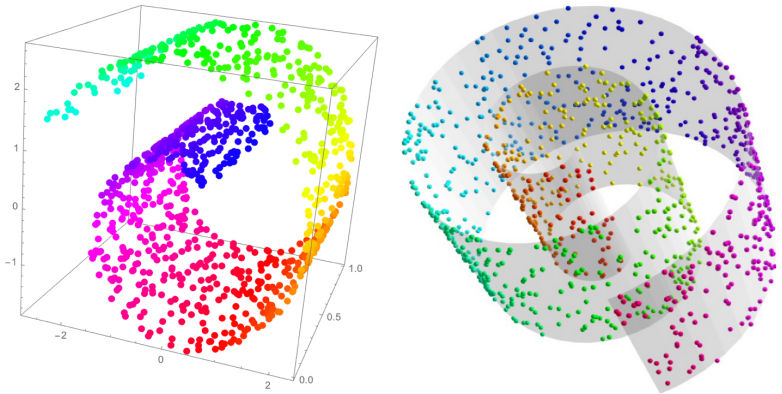
Problems

Problems

Problem

Manifold hypothesis

Many high-dimensional data sets that occur in the real world actually lie along low-dimensional manifolds



But **how to find** the dimensionality of smaller manifold?

[1] **The Manifold Hypothesis for Gradient-Based Explanations**. Anonymous authors.

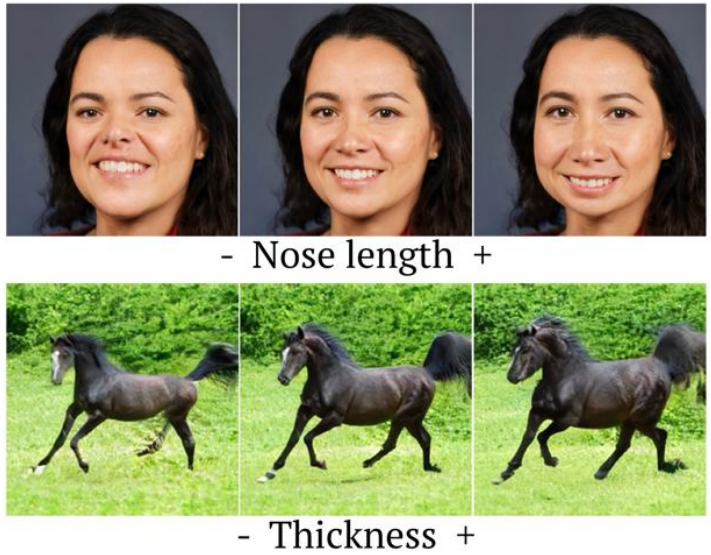
Motivation

Why do we need to find **smaller dimensionality** of the manifold?

Speed up neural networks

Compress the data

Understand the internal structure of the data



[2] **Navigating the GAN Parameter Space for Semantic Image Editin**. A. Cherepkov, A. Voynov, and A. Babenko

Methods

Methods

Methods

Methods

Theoretical methods used in the following work

Methods

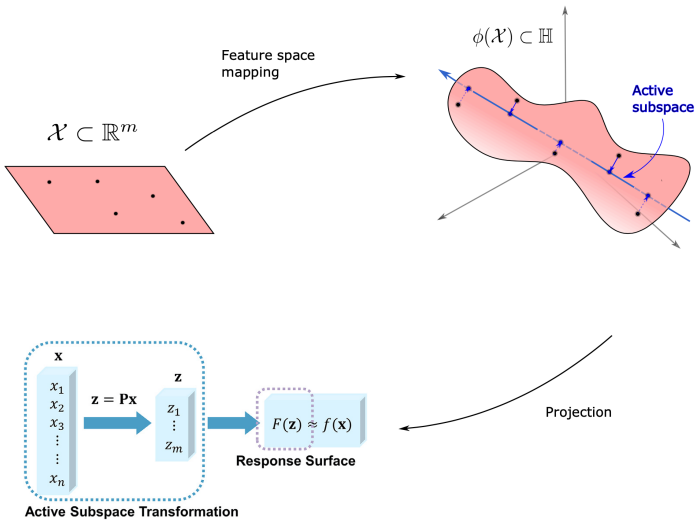
Methods

Methods

Methods

Active Subspaces

Idea: Want to **approximate `f(x)`** to find the dimension of manifold space



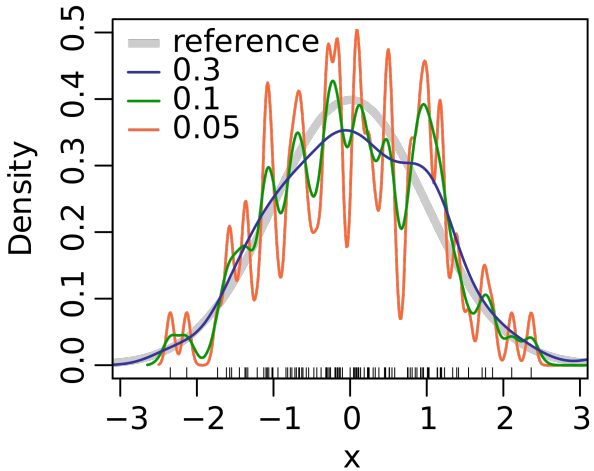
1. Draw samples $\{x_i\}_{i=1}^m$ from χ
2. Compute $\nabla f(x_i)$
3. Compute SVD of $G = \frac{1}{\sqrt{m}}\left[\nabla f(x_1) \quad \dots \quad f(x_m)\right] \approx U\Sigma V^T$
4. Estimate the rank r of $G \approx U_r \Sigma_r V_r^T$

[3] **Active Subspace Methods in theory and practice**. Paul G. Constantine, E. Dow, and Q. Wang.

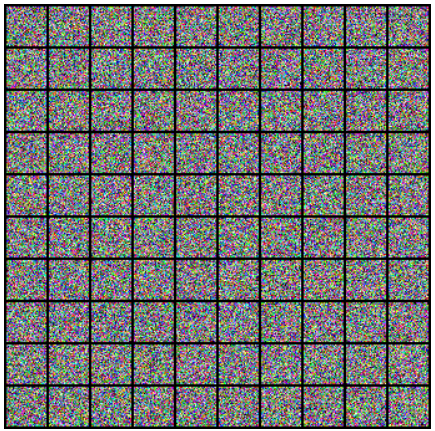
How to find `f(x)`

Good approximation of `f(x)` is **probability density function**

Deterministic ways
(kernel density estimation, KDE)



Neural methods
diffusion probabilistic models



[4] **Active Subspace Methods in theory and practice**. Paul G. Constantine, E. Dow, and Q. Wang.

Kernel Density Estimation

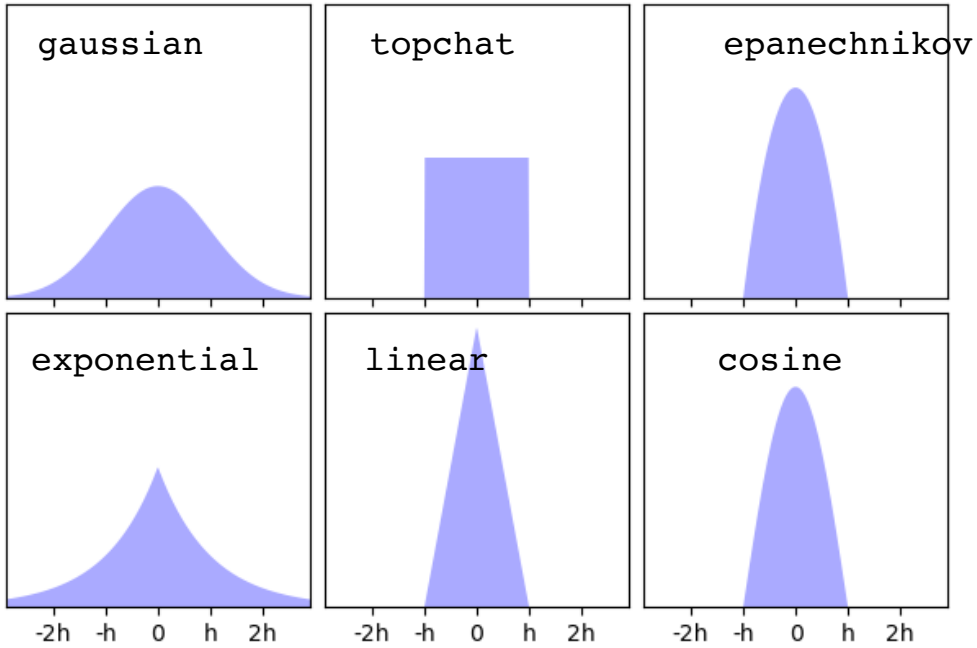
```
class KDEProbabilityDensityFunction:

    def __init__(self, data, kernel='gaussian')
    ...

    def _fit_kde(self):
    ...

    def value(self, x):
    ...

    def grad(self, x, epsilon=1e-5):
    ...
```

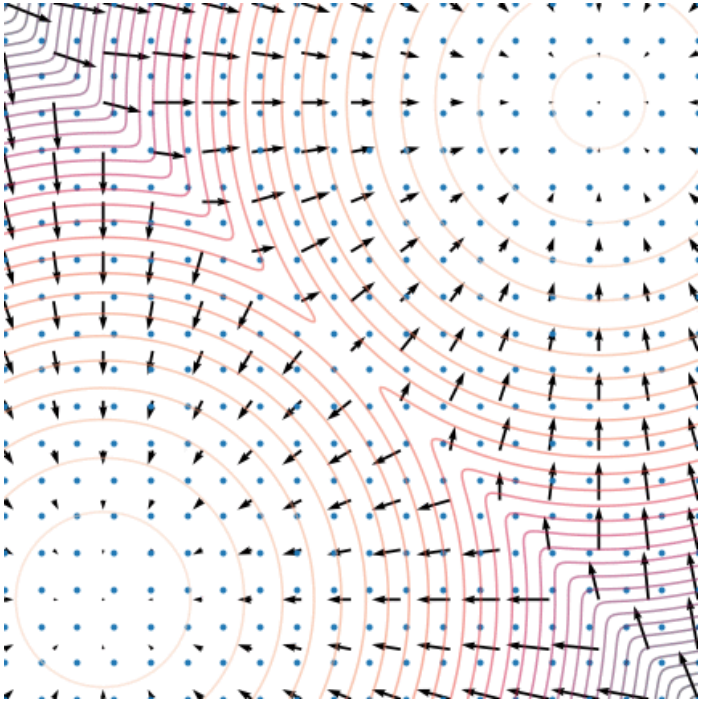


Diffusion Model

Models to approximate **score-function**

$$\mathbf{s}_\theta(\mathbf{x}) \approx \nabla_{\mathbf{x}} \log p(\mathbf{x})$$

If we take small enough time, then we can approximate true score function with high accuracy



[5] **Generative Modeling by Estimating Gradients of the Data Distribution**. Yang Song.

Evaluation

Evaluation

Evaluation

Evaluation

How to understand that the approximation of `f(x)` is good?

Evaluation

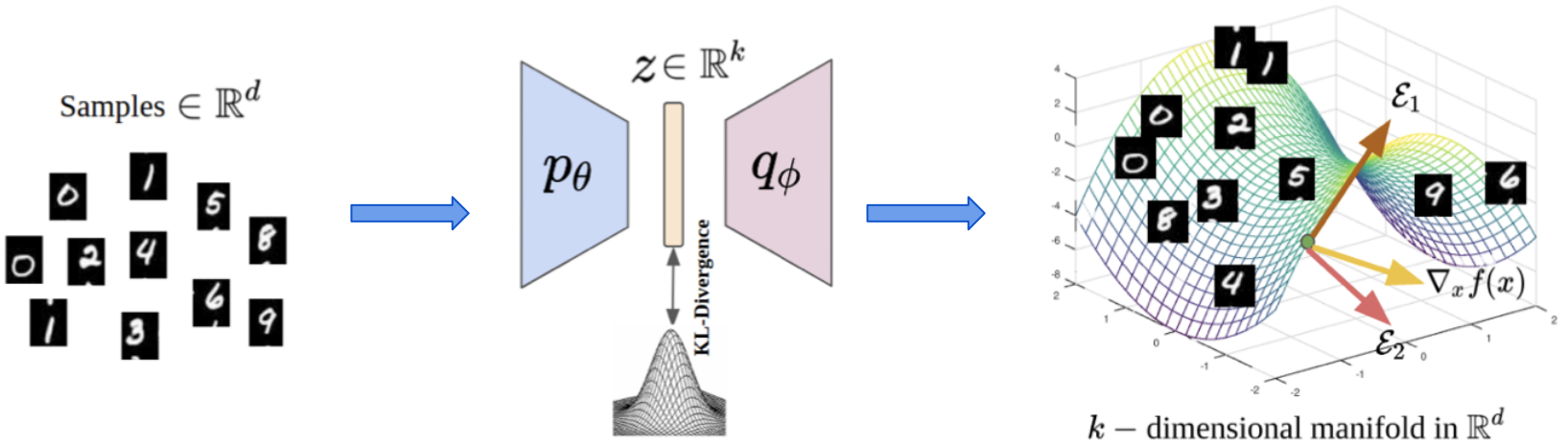
Evaluation

Evaluation

Evaluation

Visualization & VAE

Motivation to create synthetic datasets: using simple point distributions to visualise how well our method works



[1] **The Manifold Hypothesis for Gradient-Based Explanations**. Anonymous authors.

Experiments

Experiments

Experiments

Experiments

Most interesting part, u know:)

Experiments

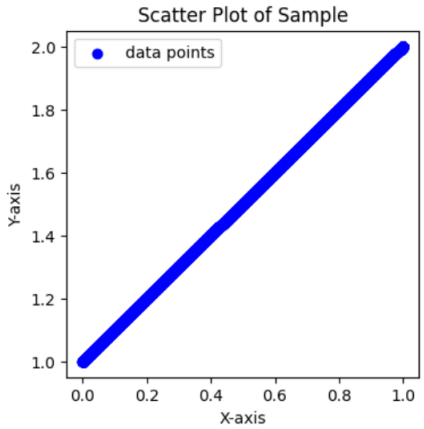
Experiments

Experiments

Experiments

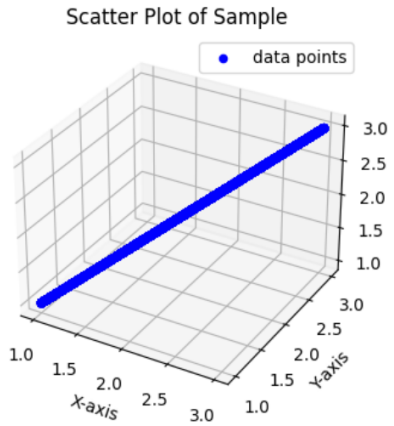
Datasets

```
generator = LinearGenerator(dim_of_space, dim_of_manifold)
```



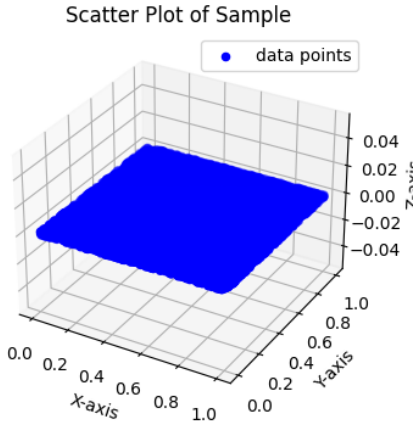
N = 2 m = 1

line in 2D



N = 3 m = 1

line in 3D



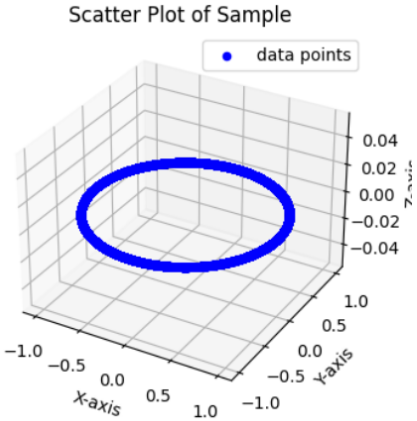
N = 3 m = 2

plane in 3D

+ m = 10 hyperplane in N = 30 space

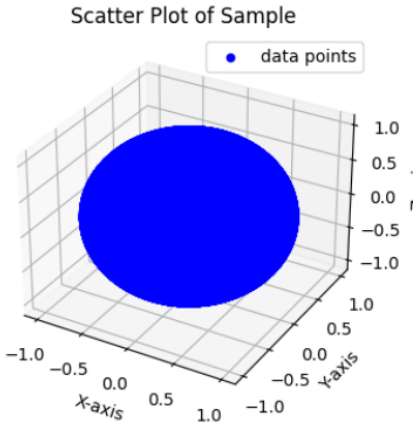
Datasets

```
generator = SphericalGenerator(dim_of_space, dim_of_manifold)
```



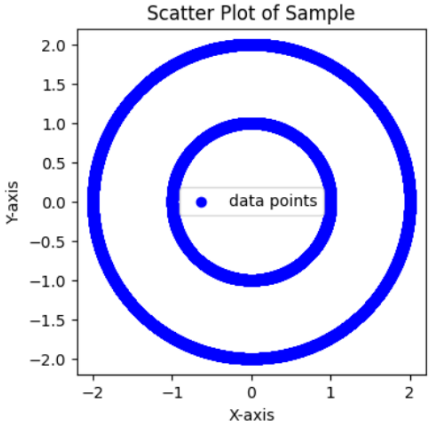
N = 3 m = 1

circle in 3D



N = 3 m = 2

sphere in 3D



N = 2 m = 1

nested in 2D

Datasets: MNIST (Zeros)

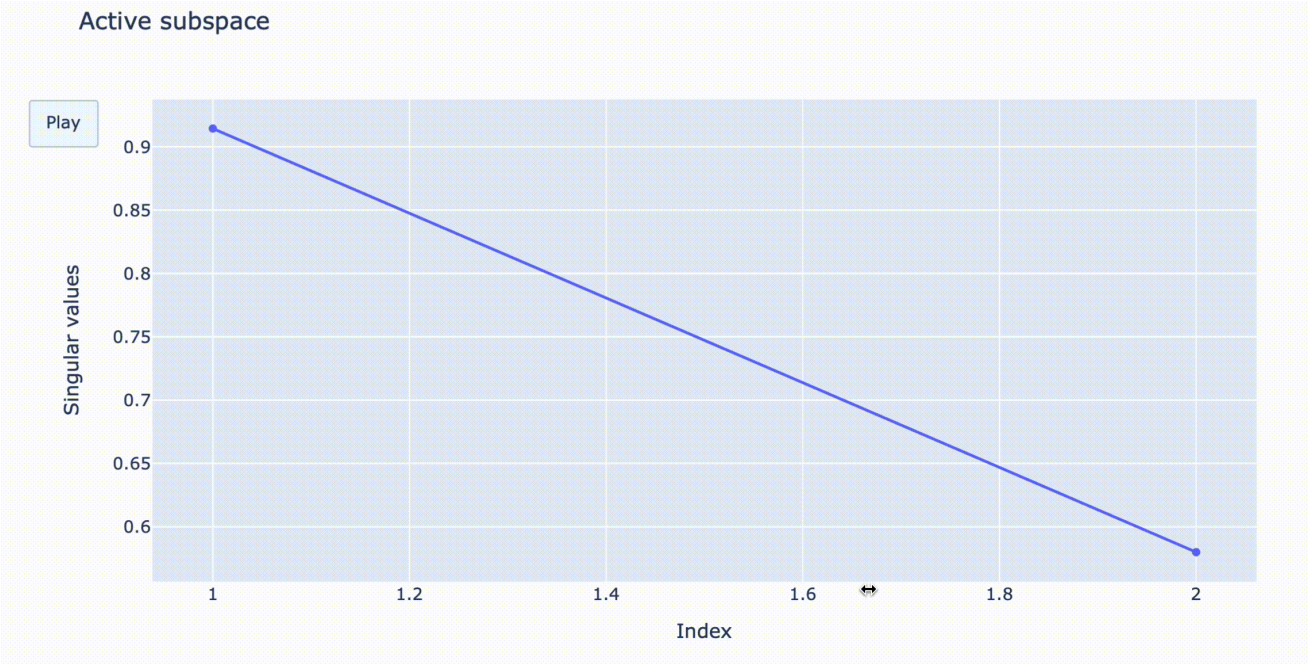
Different numbers lie in different manifolds, so only **zeros are considered** within the project

Results

	Manifold dim	Space dim	n_samples	KDE	Diffusion
Line 2D	1	2	10**5	1	-
Line 3D	1	3		1	1
Plane	2	3		2	2
Sphere	2	3		2	2
Circle	1	3		2	-
Nested	1	2		1	1
Highdim	10	30		10	-

Monte Carlo sample size

The **larger** the sample size, the **better** `manifold dimension` is approximated

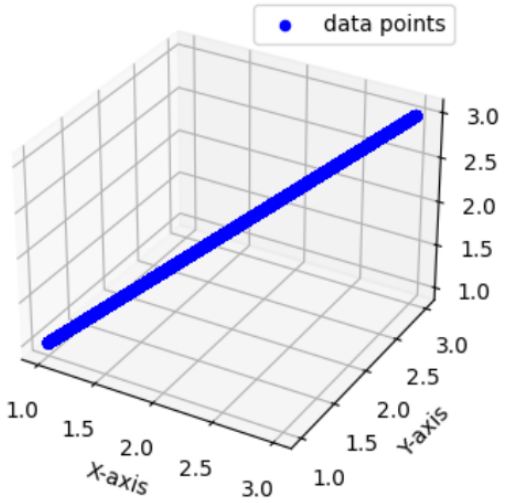


Hyperplane: `N = 30  m = 10`

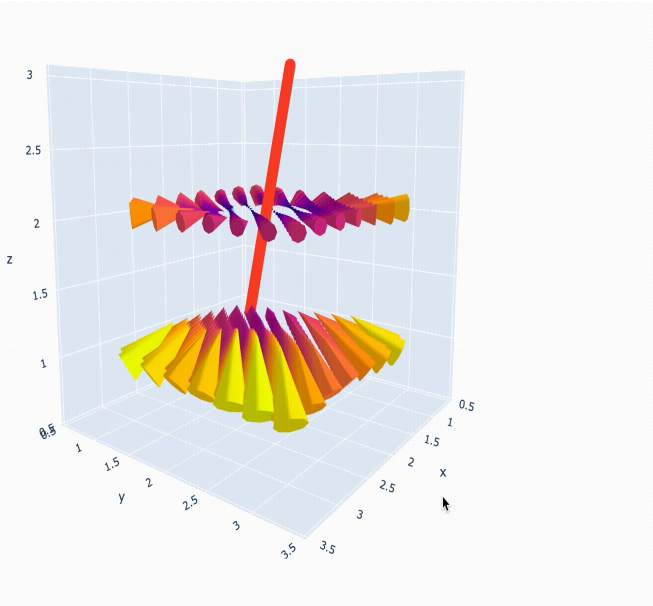
Results: Diffusion

line in 3D: `N = 3  m = 1`

True Distribution



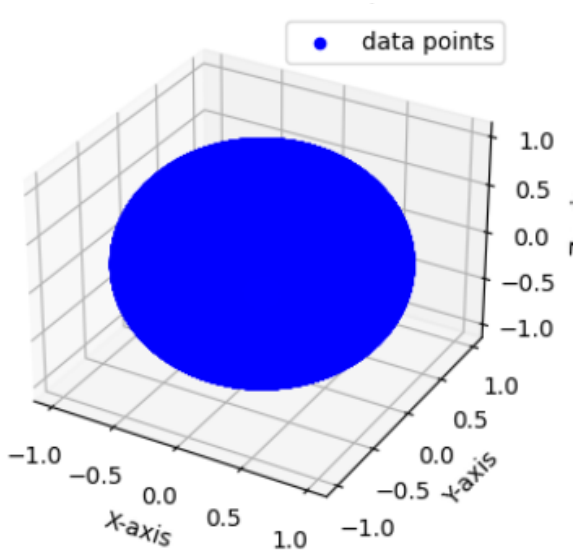
Score function



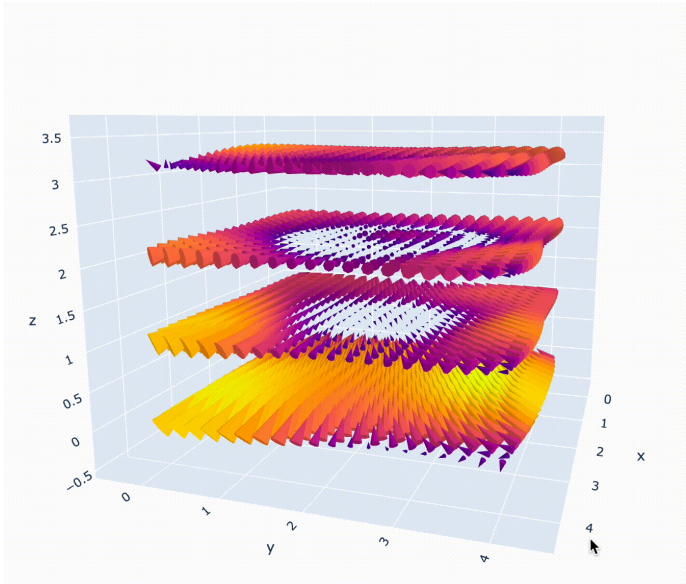
Results: Diffusion

line in 3D: `N = 3  m = 2`

True Distribution

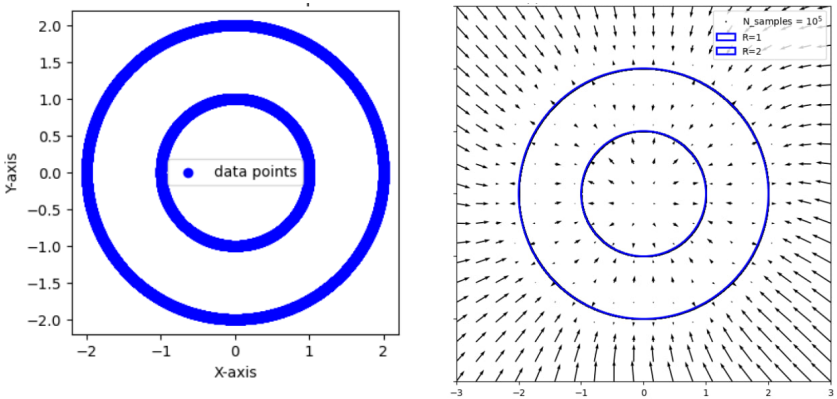


Score function



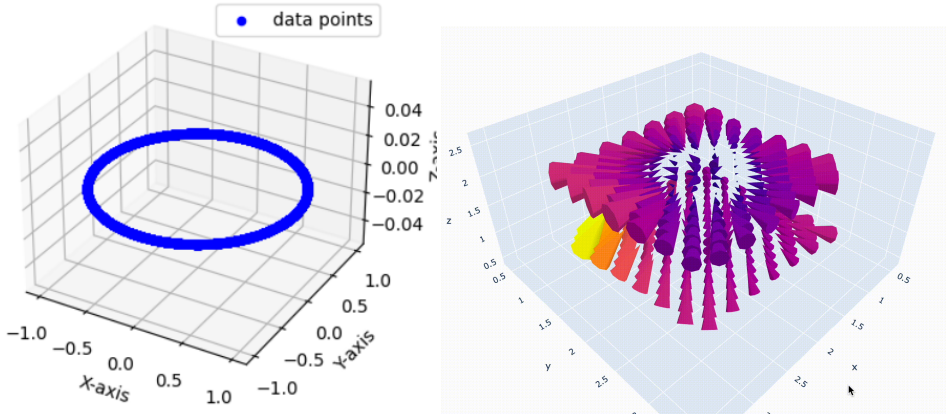
Results: Diffusion

Other examples of trained models



N = 2 m = 1

nested in 2D

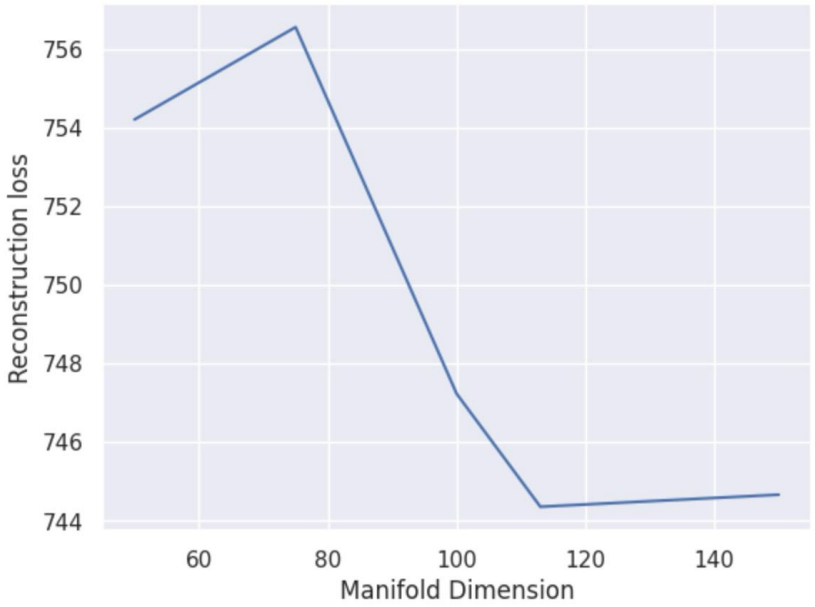


N = 3 m = 1

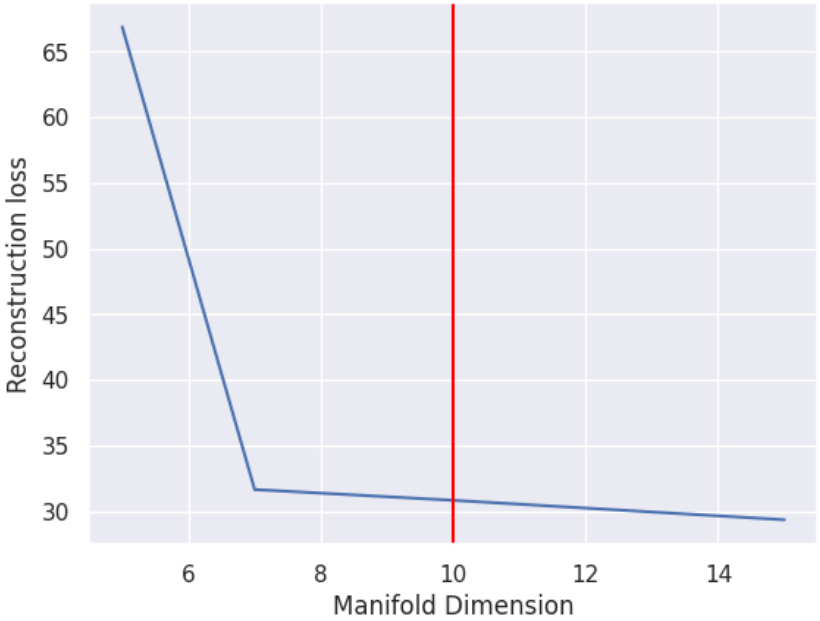
circle in 3D

Results: VAE

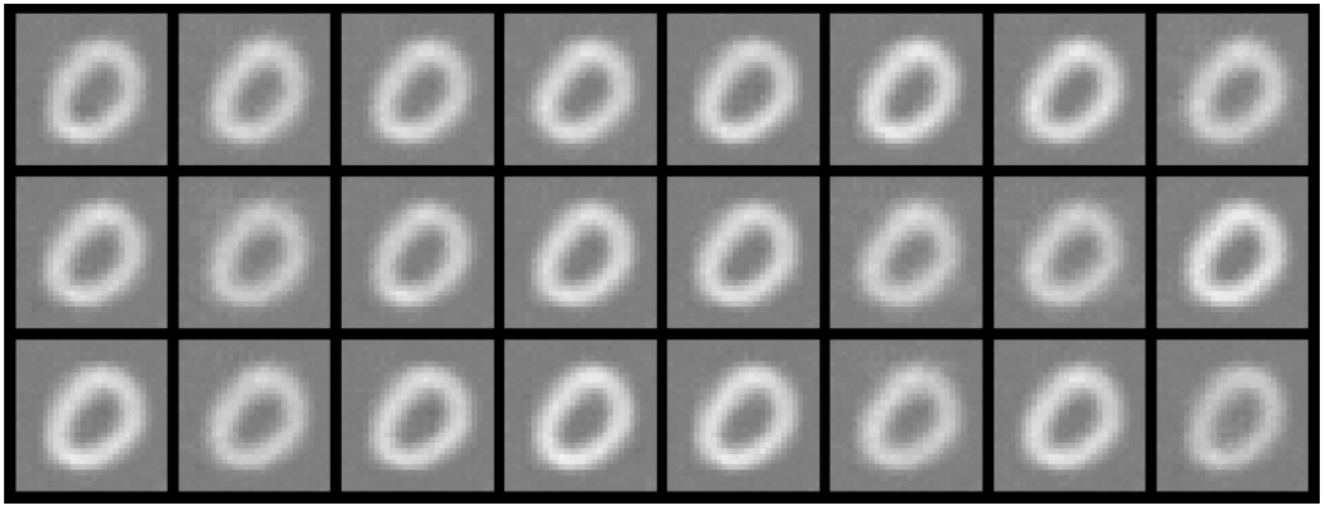
MNIST: Zeros



`m = 10` hyperplane in `N = 30` space



Results: VAE



Manifold Dim = 50

Further Research

Further Research

Further Research

Further Research

What we planned to do but didn’t:(

Further Research

Further Research

Further Research

Further Research

Further research

Trajectory research

Forward diffusion

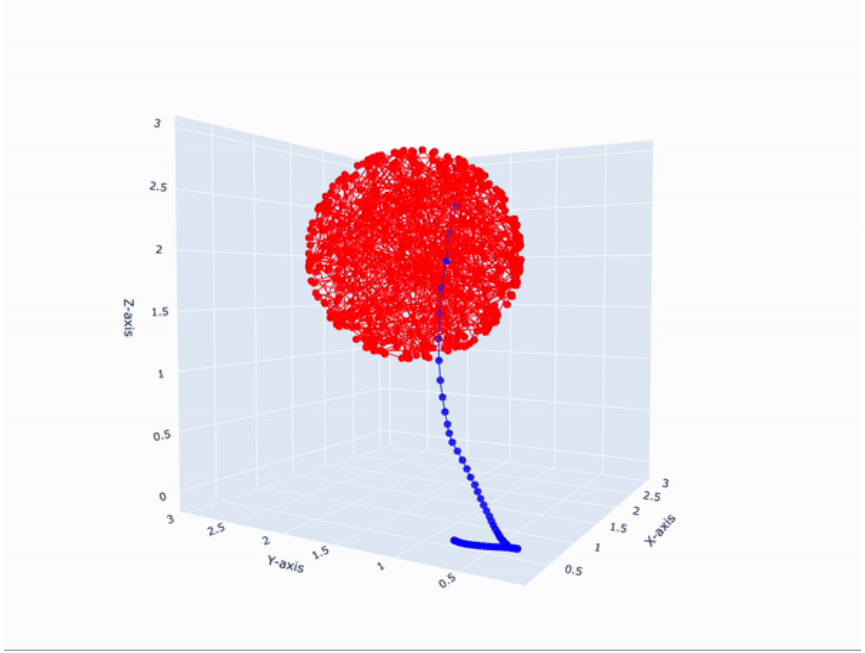
$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}$$

Backward diffusion

$$\mathrm{d}\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x})]\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}$$

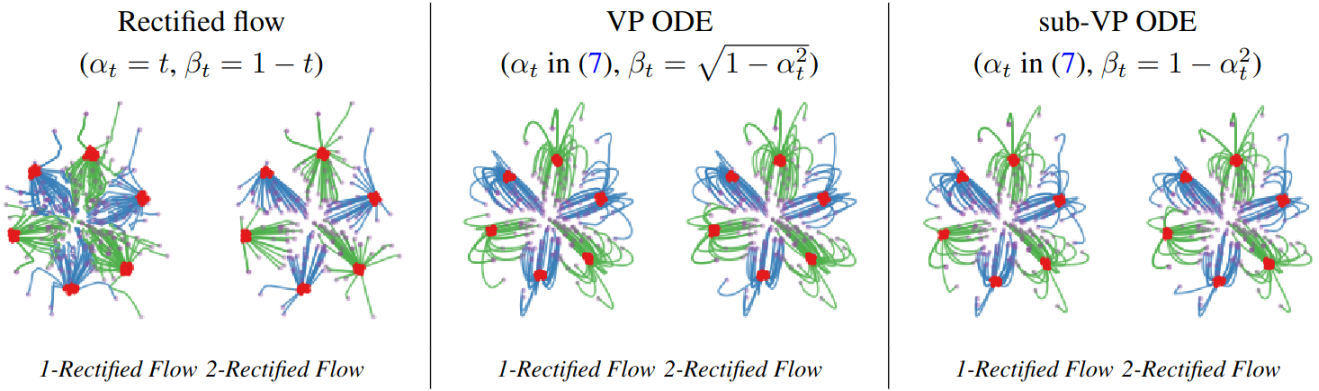
ODE diffusion

$$\mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - \frac{1}{2} g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x})\right]\mathrm{d}t$$



Further research

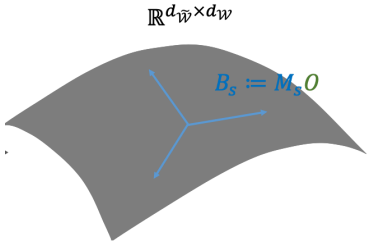
Flow matching
(Rectified flow),
Schrödinger bridge



Hypothesis:
Any generative model can be approximated easy projection function

[7] **Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow.** Xingchao Liu and et.al

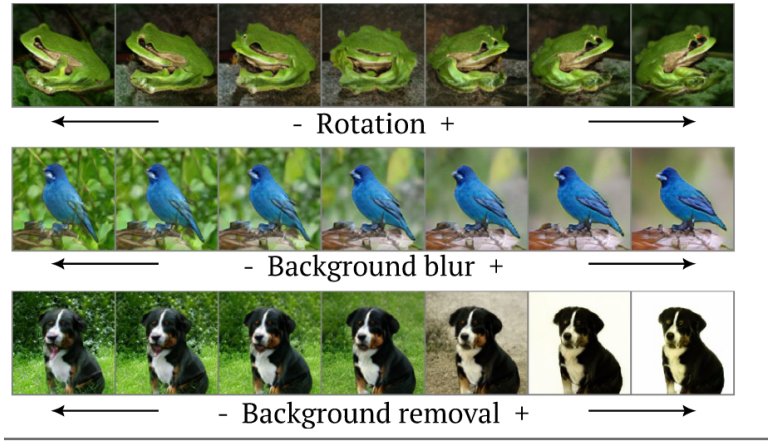
Further research



Tradeoff:
Low dimension = High loss
High dimension = Complex basis

Golden middle?

The ideally disentangled latent space in GAN involves the global representation of latent space with semantic attribute coordinates. In other words, considering that this disentangled latent space is a vector space, there exists the global semantic basis where each basis component describes one attribute of generated images.



[7] **Unsupervised Discovery of Interpretable Directions in the GAN Latent Space.** Andrey Voynov and Artem Babenko.

Conclusion

Conclusion

Conclusion

Conclusion

Let’s recap what we have done

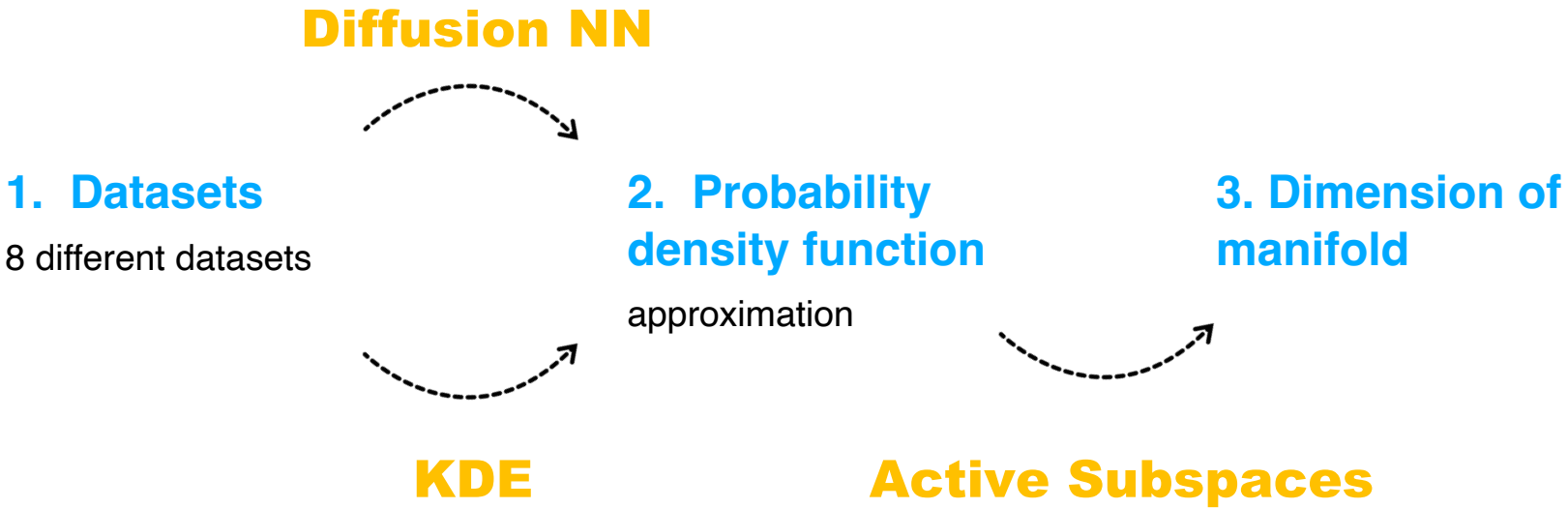
Conclusion

Conclusion

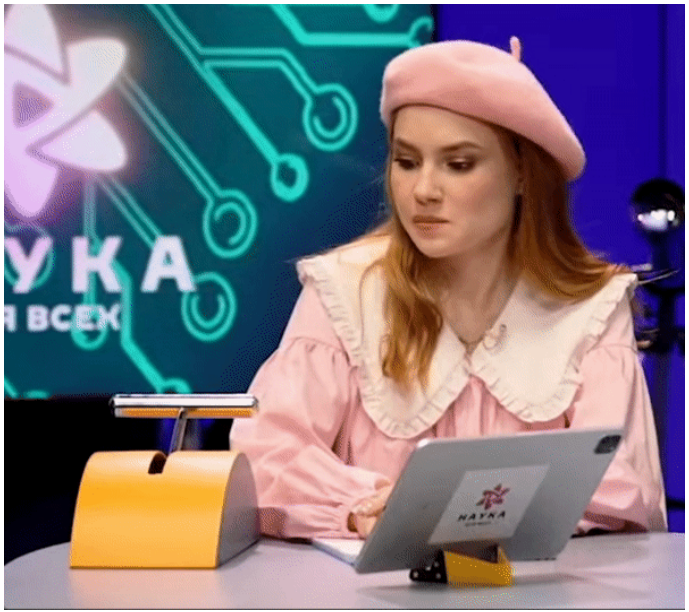
Conclusion

Conclusion

Concept


Diffusion NN
1. Datasets
8 different datasets
2. Probability density function
approximation
3. Dimension of manifold
KDE
Active Subspaces

Questions?



David Li
David.Li@skoltech.ru
Data Science

Artem Alekseev
Artem.Alekseev@skoltech.ru
Data Science

Ignat Melnikov
Ignat.Melnikov@skoltech.ru
Data Science

Viktoriia Zinkovich
Viktoriia.Zinkovich@skoltech.ru
Data Science

Kamil Garifullin
Kamil.Garifullin@skoltech.ru
Data Science

Code is available at Github!

[10] https://vk.com/video-221518001_456239055