




Nonparametric Bayesian Deep Visualization

Haruya Ishizuka @Bridgestone Corporation, Japan

Daichi Mochihashi @The Institute of Statistical Mathematics,
Japan

ECML-PKDD 2022  ECML
PKDD
2022
統計数理セミナー 2022-11-30

Benefits of the proposed NPDV

- Intuitive understanding of internal structures using **latent clusters**
- Using a Bayesian neural network with **very few parameters** thanks to Neural Network Gaussian processes



NPDV



t-SNE (2008)



PaCMAP (2021)

Introduction

High-dimensional data visualization

- Requires dimensionality reduction
- Will enhance knowledge discovery in various domains, e.g. RNA analysis

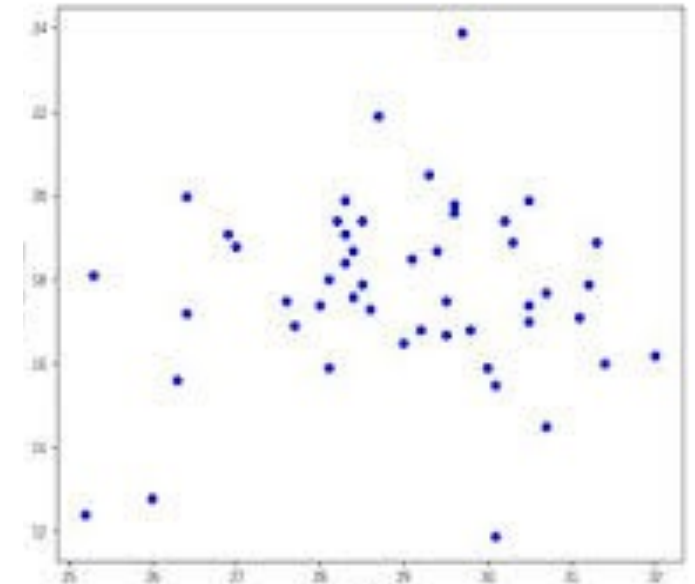
n	Y_1	Y_2	...	Y_D
1				
2				
...				
N				

Not plottable
when $D > 3$

Dimensionality
Reduction

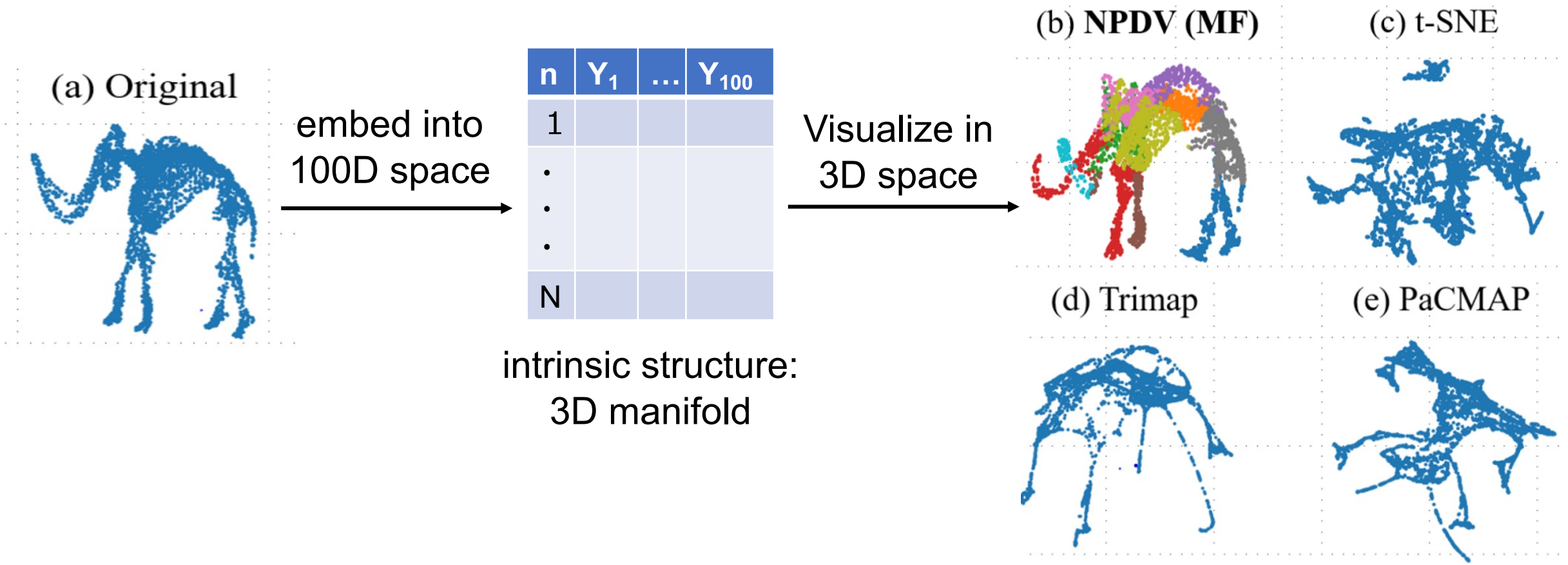
n	V_1	V_2
1		
2		
...		
N		

Plottable



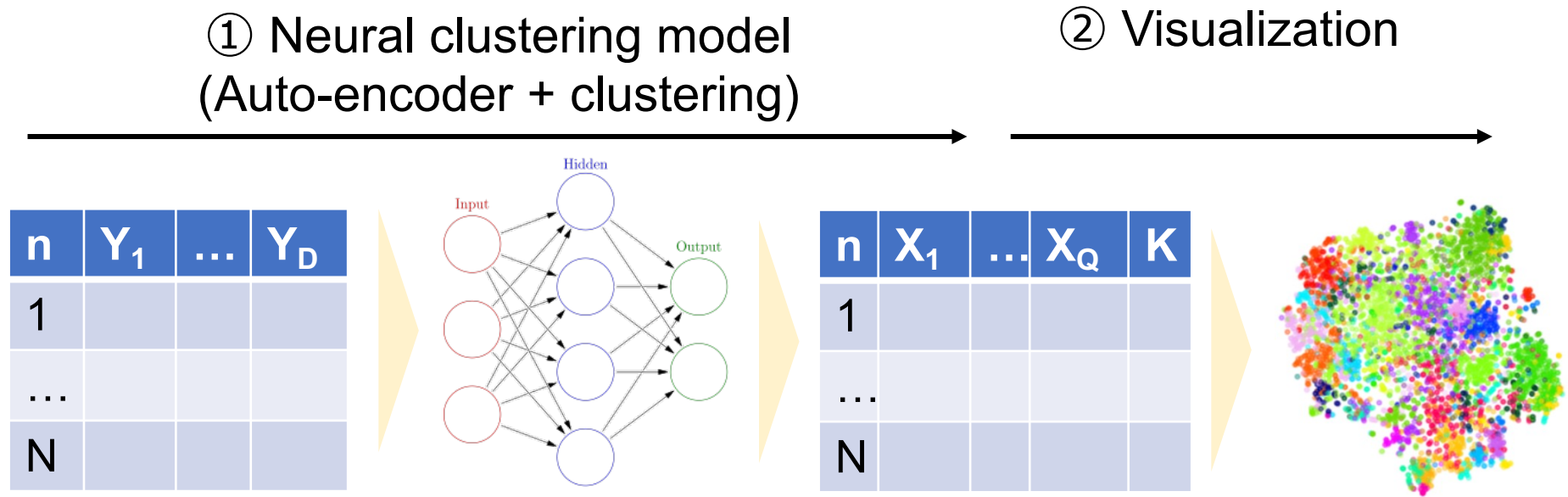
Issues of existing methods

1. Inaccurate visualization Y has an intrinsic structure on latent manifold
2. Low interpretability on possible clusters



Alternative approach: neural clustering models

- Estimate latent coordinates and clusters jointly



- **Cons.1: Large # of hyperparameters**
- **Cons.2: X only will not always be suitable for visualization**

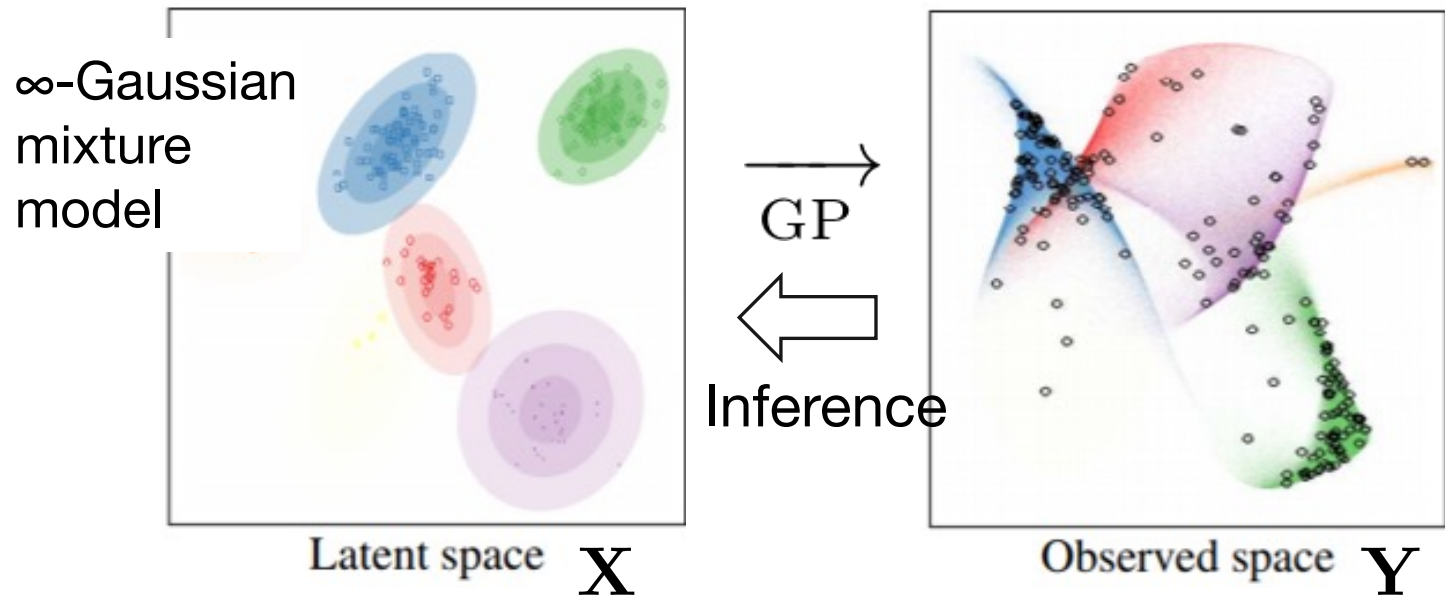
Neural network infinite Warped Mixture Model (NN-iWMM)

- Reduce some hyperparameters by Nonparametric Bayes techniques

×: to be optimized ✓: no need to be optimized

	Neural clustering	NN-iWMM	Components
# of layers	×	×	-
# of units	×	✓	ARD-NNGP
# of latent dims	×	✓	
# of clusters	×	✓	∞-GMM

Infinite Warped Mixture Model (iWMM) (Iwata+ 2013)

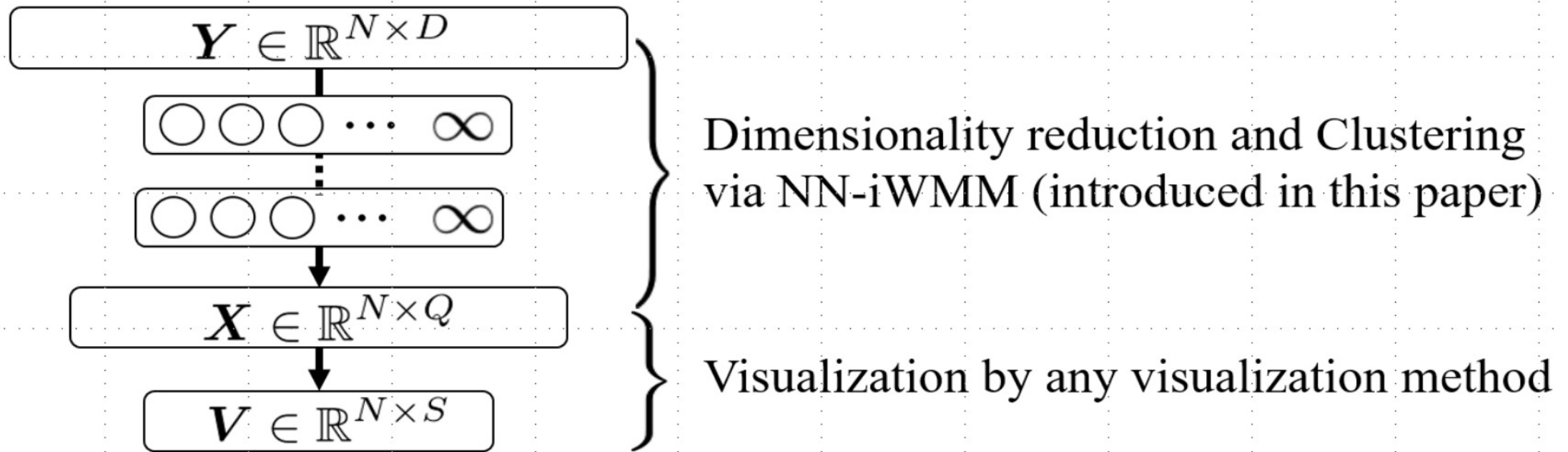


- Infinite Gaussian mixture in latent space
→ Map to observation space through Gaussian processes
- d 'th dimension of observation Y distributes according to GP on X :

$$Y_d \sim \text{GP}(\mu, K_{\mathbf{X}})$$

Nonparametric Bayesian Deep Visualization: Overview

- Integrate NN-iWMM and a visualization method into a Bayesian model
 - Infer \mathbf{X} so as to be optimal to estimate \mathbf{V}

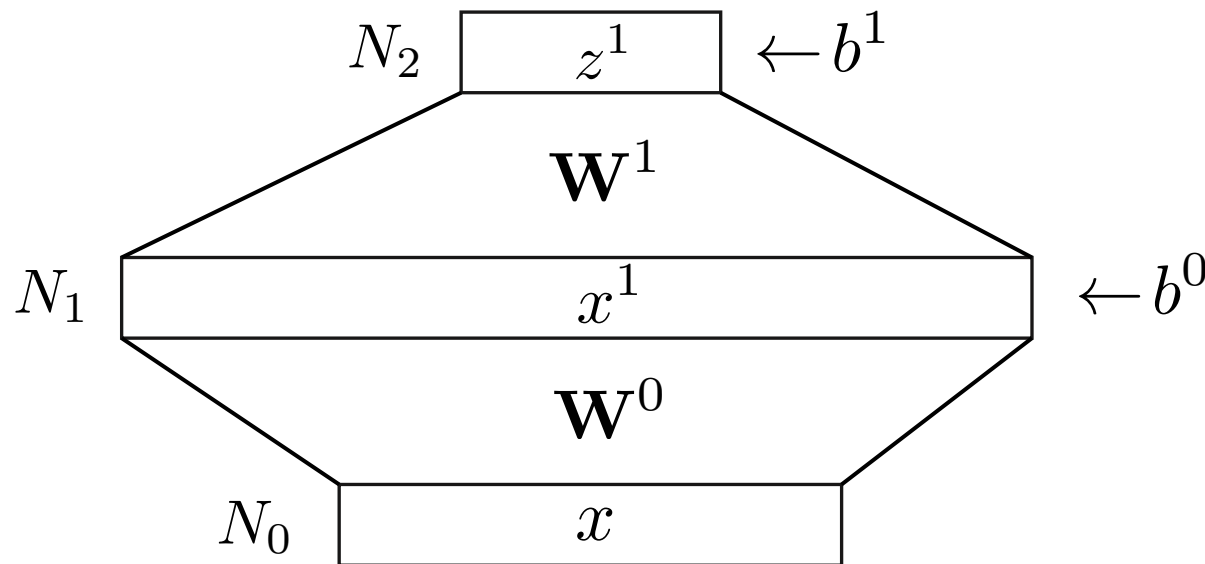


Proposed NPDV

Neural Network = Gaussian process (Neal 1994)

- Single hidden layer NN with inputs layer x
- i 'th output $z_i(x)$ is written as:

$$z_i^1(x) = b_i^1 + \sum_{j=1}^{N_1} W_{ij}^1 x_j^1(x), \quad x_j^1(x) = \phi \left(b_j^0 + \sum_{k=1}^K W_{jk}^0 x_k \right)$$



Connection

$$W_{ij}^\ell \sim \mathcal{N}(0, \sigma_w / N_\ell)$$

Bias

$$b_i^\ell \sim \mathcal{N}(0, \sigma_b)$$

Neural Network = Gaussian process (2)

$$z_i^1(x) = b_i^1 + \sum_{j=1}^{N_1} W_{ij}^1 x_j^1(x), \quad x_j^1(x) = \phi \left(b_j^0 + \sum_{k=1}^K W_{jk}^0 x_k \right)$$

- Sum of independent weights W_{ij}^ℓ and biases b_i^ℓ
→ Joint probability $p(z_1^1(x), z_2^1(x), \dots, z_{N_2}^1(x))$
has a **multivariate Gaussian distribution**
by Central Limit Theorem ... **Gaussian process**
- Mean: clearly 0
- Covariance is

$$\begin{aligned} K^1(x, x') &\equiv \mathbb{E} [z_i^1(x) z_i^1(x')] \\ &= \sigma_b^2 + \sigma_w^2 \mathbb{E} [x_i^1(x) x_i^1(x')] = \sigma_b^2 + \sigma_w^2 C(x, x') \end{aligned}$$

NNGP (Neural Network Gaussian Process) (Lee+ 2017)

- Assume $\ell-1$ layer output $z_j^{\ell-1}$ is GP:

$$z_i^\ell(x) = b_i^\ell + \sum_{j=1}^{N_\ell} W_{ij}^\ell x_j^\ell(x), \quad x_j^\ell(x) = \phi(z_j^{\ell-1}(x))$$

- ℓ -layer Mean is 0, variance is

$$\begin{aligned} K^\ell(x, x') &\equiv \mathbb{E} [z_i^\ell(x) z_i^\ell(x')] \\ &= \sigma_b^2 + \sigma_w^2 \mathbb{E}_{z_i^{\ell-1} \sim \text{GP}(0, K^{\ell-1})} [\phi(z_i^{\ell-1}(x)) \phi(z_i^{\ell-1}(x'))] \end{aligned}$$

- This expectation can be computed by:
 - (1) GP regression
 - (2) Numerical approximation
 - (3) **Analytical solution** for specific ϕ like ReLU

NNGP (Neural Network Gaussian Process) (2)

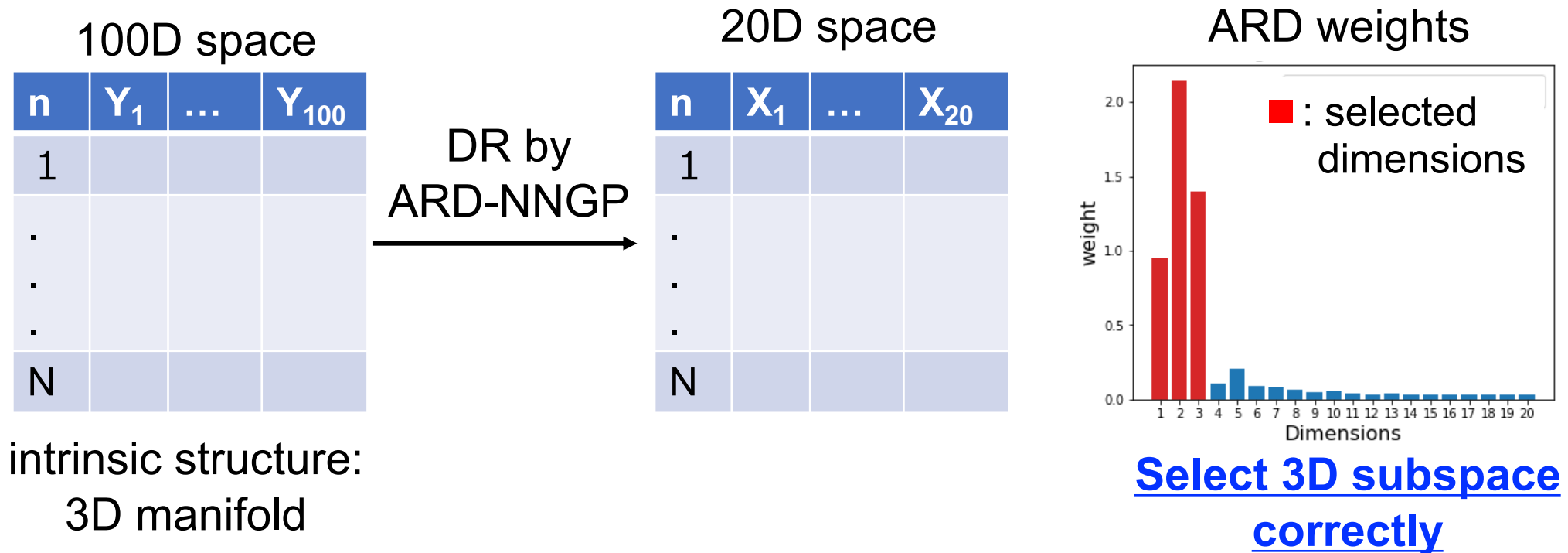
- When φ is ReLU (Cho&Saul 2009, Lee+ 2017) :

$$K^\ell(x, x') = \sigma_b^2 + \frac{\sigma_w^2}{2\pi} \sqrt{K^{\ell-1}(x, x)K^{\ell-1}(x', x')} \\ \times \left(\sin \theta_{x, x'}^{\ell-1} + (\pi - \theta_{x, x'}^{\ell-1}) \cos \theta_{x, x'}^{\ell-1} \right)$$
$$\theta_{x, x'}^\ell = \cos^{-1} \left(\frac{K^\ell(x, x')}{\sqrt{K^\ell(x, x)K^\ell(x', x')}} \right)$$

- Multi-layer NN is obtained **just by matrix multiplications!**

NN-iWMM - Component 1: ARD-NNGP

- NNGP [1] defines Gaussian processes equivalent to a ∞ -unit NN
- ARD [2] allows to determine the dimensionality automatically



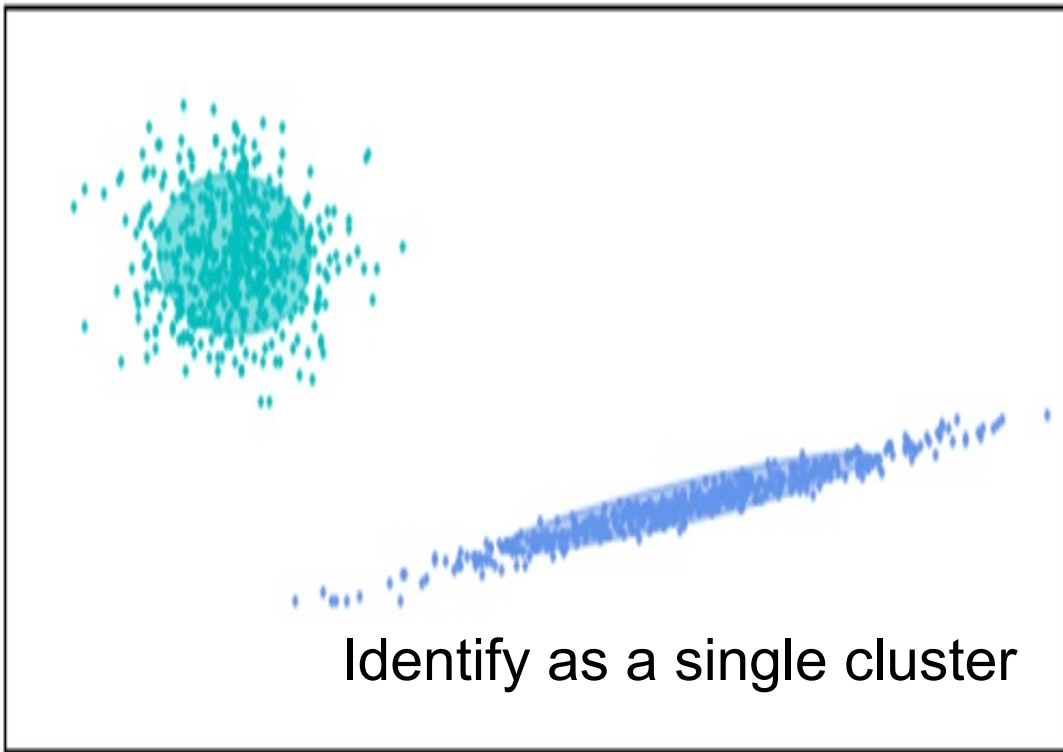
[1] Lee, J. +: Deep neural networks as gaussian processes, International Conference on Learning Representation 2018 48 478–487 (2018)

[2] Mackay, D.: Bayesian Non-Linear Modeling for the Prediction Competition, ASHRAE Transaction 100(2) 1053–1062 (1994)

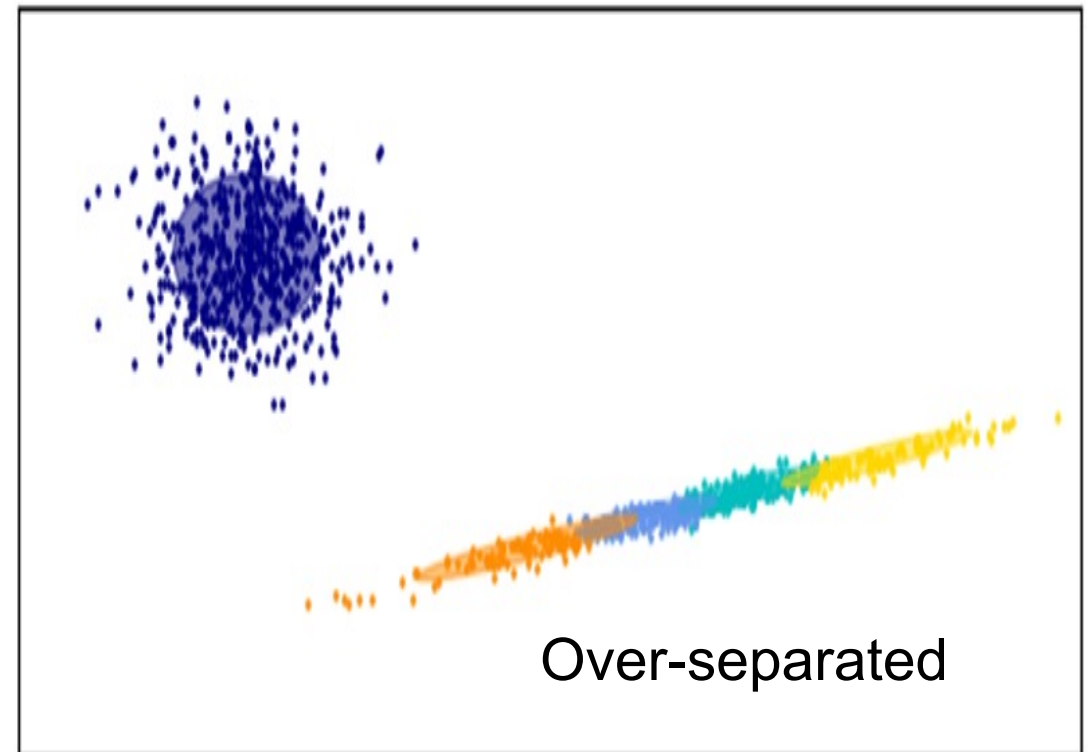
NN-iWMM - Component 2: ∞ -GMM

- Can infer the number of clusters unlike finite GMM

infinite-GMM

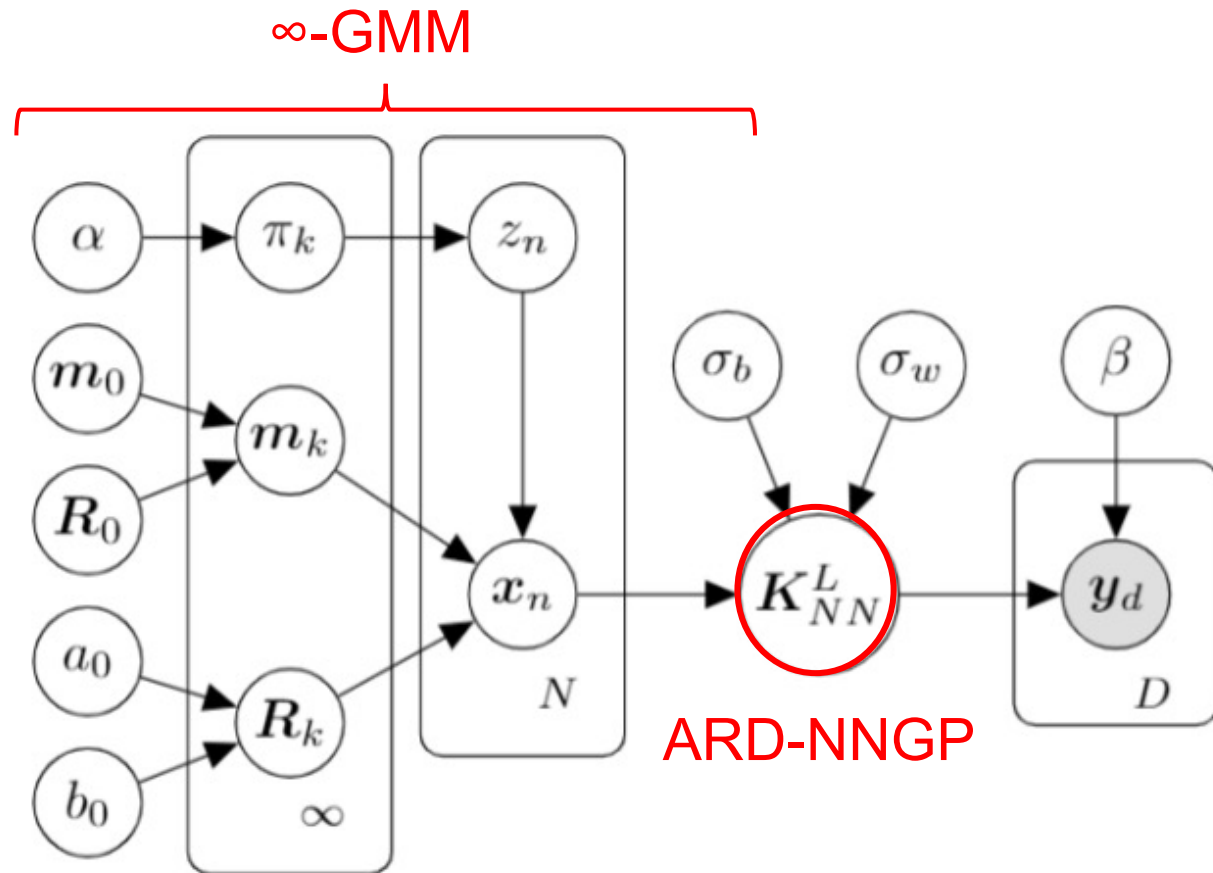


Finite GMM (K=5)



NN-iWMM: Generative process

- Incorporate ARD-NNGP and ∞ -GMM into a latent variable model



×: to be optimized

✓: no need to be optimized

	NN-iWMM
# of layers	×
# of units	✓
# of latent dims	✓
# of clusters	✓

NPDV: Probabilistic model

- The joint distribution of NPDV

$$q^*(\mathbf{X}, \mathbf{V}) \propto p(\mathbf{Y}|\mathbf{X})p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Z})p(\mathbf{Z}|\boldsymbol{\pi})p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) \textcircled{1}$$
$$\times \exp(-\lambda \mathcal{R}_{\text{DR}}(\mathbf{X}, \mathbf{V})) \textcircled{2}$$

① : The likelihood of NN-iWMM ② : Visualization loss

- Note: ② = Regularization term to infer the posterior of ①

- Two algorithms based on NPDV
 - NPDV (MF) : NN-iWMM + Matrix Factorization
 - NPDV (*t*-SNE) : NN-iWMM + *t*-SNE

NPDV: Training algorithm

- Employ variational inference to infer the posterior of NPDV
- Maximize the ELBO consisting of 4 terms:

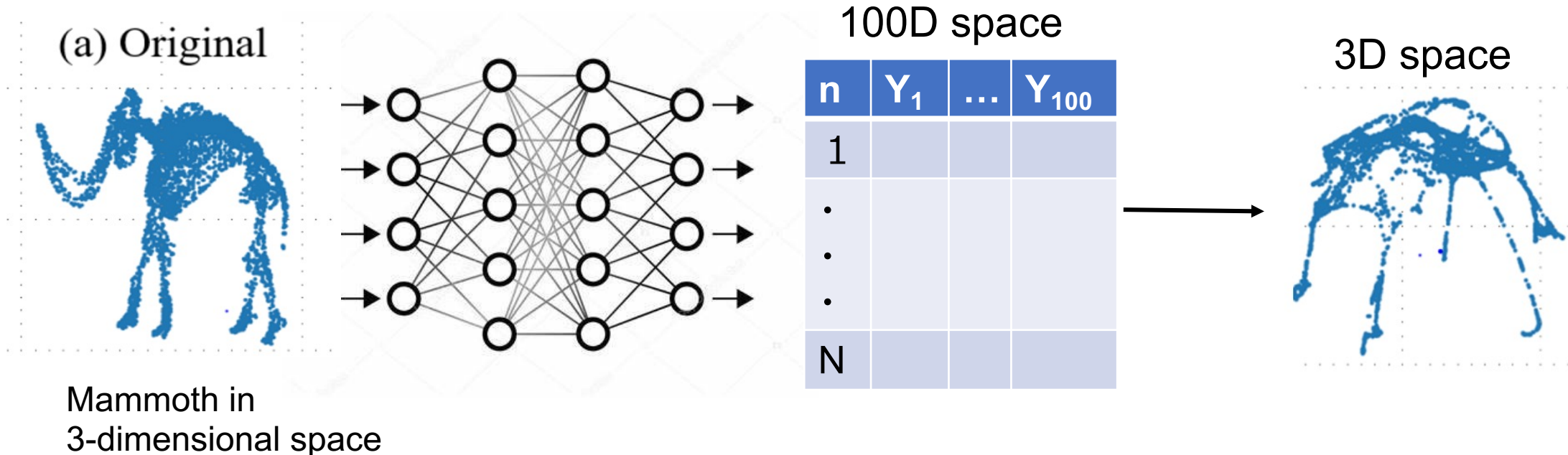
$$\begin{aligned}\mathcal{L} &= \mathbb{E}_{q(\mathbf{X})}[\log p(\mathbf{Y}|\mathbf{X})] - \mathbb{E}_{q(\mathbf{X})}[\log q(\mathbf{X})] \\ &+ \mathbb{E}_{q(\mathbf{X}, \mathbf{z}, \mathbf{m}, \Sigma, \phi)} \left[\log \frac{p(\mathbf{X}, \mathbf{z}, \{\mathbf{m}_k, \mathbf{r}_k, \pi_k\}_{k=1}^K)}{q(\mathbf{z}, \{\mathbf{m}_k, \mathbf{R}_k\}_k, \boldsymbol{\pi})} \right] - \lambda \mathbb{E}_{q(\mathbf{X})}[\text{KL}[\mathbf{p}^X \parallel \mathbf{p}^V]] \\ &= \underbrace{\mathcal{L}_1 + \mathcal{L}_2 - \lambda \mathcal{R}}_{(*)} + \underbrace{\mathcal{H}(q(\mathbf{X}))}_{\text{Gauss Entropy}}\end{aligned}$$

- Difficulty: (*) cannot be computed analytically
- ⇒ Approximate (*) using reparameterization trick

Experiments 1: Simulation study

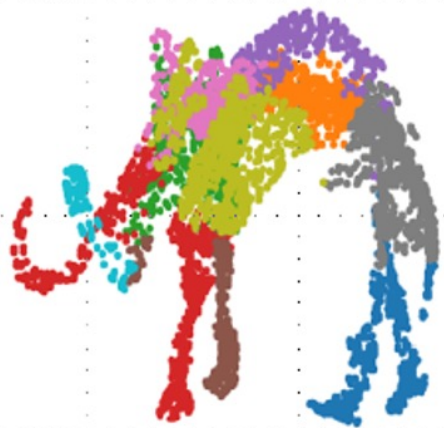
Data generation setup

- Generate 100-dimensional data through NN transformation
- Apply 4 methods including NPDV (MF) to recover the original data

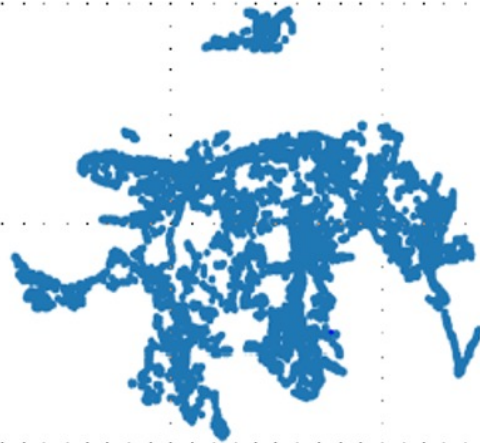


Visualization results

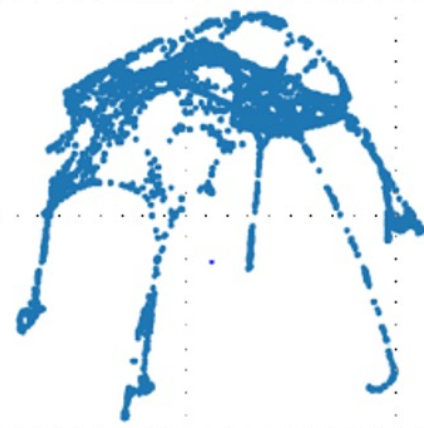
- NPDV (MF) could recover the original mammoth shape accurately
- Some body parts (horn, paw) was found through estimated clusters



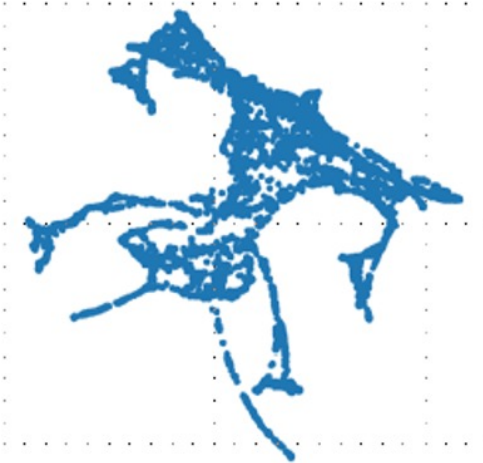
(a) NPDV (MF)



(b) *t*-SNE
(2008)



(c) Trimap
(2018)



(d) PaCMAP
(2021)

Experiment 2: Real data experiments @ 20-newsgroups

20-Newsgroups:

Dataset of ~20,000 documents from 20 newsgroups of USENET
(Lang 1995)

Experimental settings

- Dataset

# of labels	Input	D	Remarks
6	TF-IDF	1,000	20 labels are converted into 6 parent labels

- Methods

Remarks
-
-
Hyperparameters are tuned by Bayesian optimization
of layers is set to 6

t-SNE (2008)

-

PaCMAP (2021)

-

VSB-DVM[1] + t-SNE

Hyperparameters are tuned by Bayesian optimization

NPDV (t-SNE)

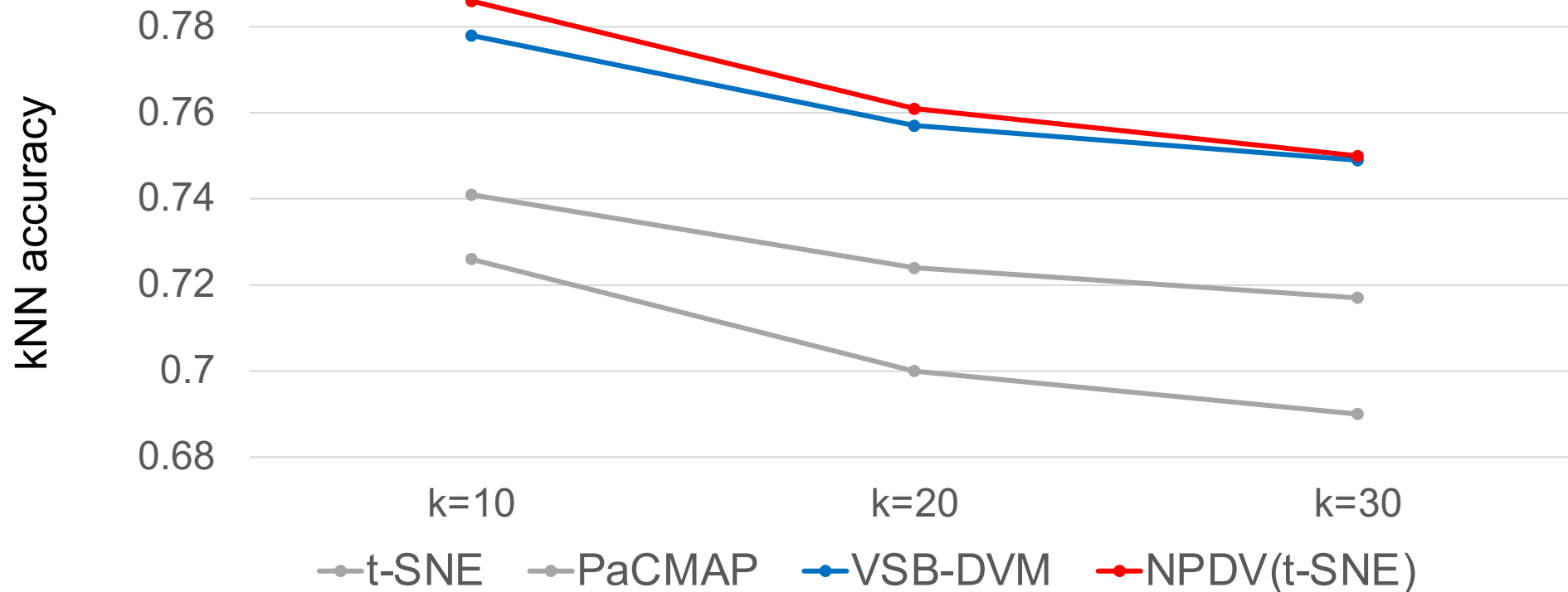
of layers is set to 6

- Performance measure: kNN classification accuracy

[1] Yang, X.+.: VSB-DVM: An end-to-end Bayesian nonparametric generalization of deep variational Mixture Model,

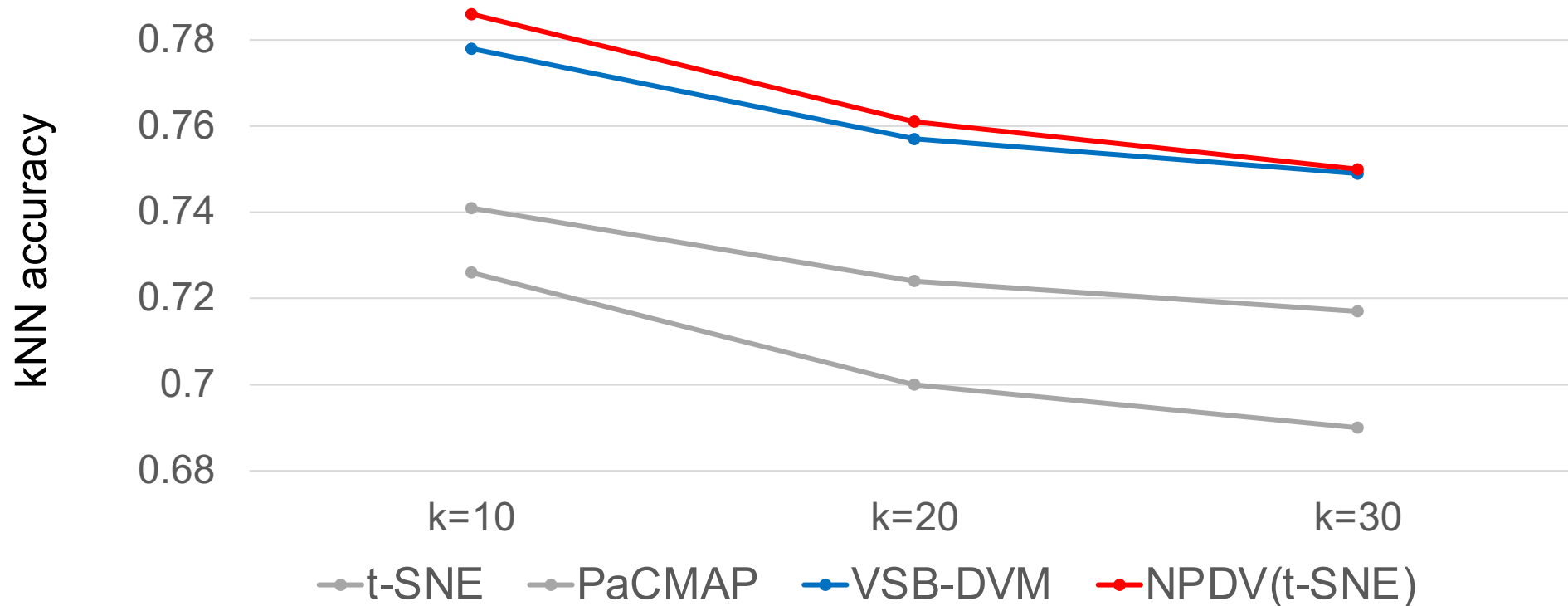
kNN classification accuracy

- NPDV shows comparable performance to a well-tuned VSB-DVM
- Less tuning time than VSB-DBM (NPDV : 8 hours vs. VSB-DMV: 6 days)



kNN classification accuracy

- NPDV shows comparable performance to a well-tuned VSB-DVM
- Less tuning time than VSB-DBM (NPDV : 8 hours vs. VSB-DMV: 6 days)



Achieve comparable acc. to a well-tuned NN with low tuning cost

Qualitative comparison @ 20 newsgroups

- Coloring provides intuitive understanding for the cluster structure
- NPDV shows better cluster separation than (b)



(a) NPDV(t-SNE)



(b) VSB-DVM



(c) t-SNE
(2008)

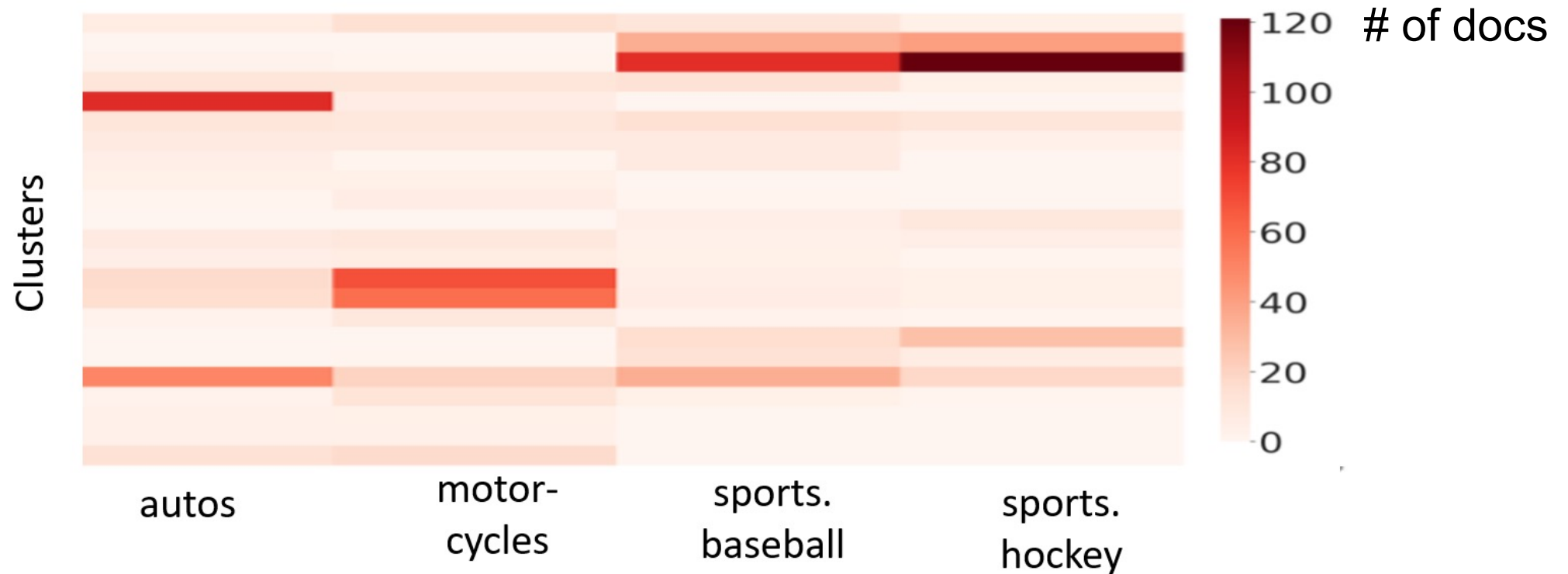


(d) PaCMAP
(2021)

Relation between estimated clusters and ground truth

- The clusters correlate with ground truth labels strongly
⇒ NPDV estimates the plausible clusters without label information

Cross table of estimated clusters & ground truth



Discovering clusters in Brown corpus

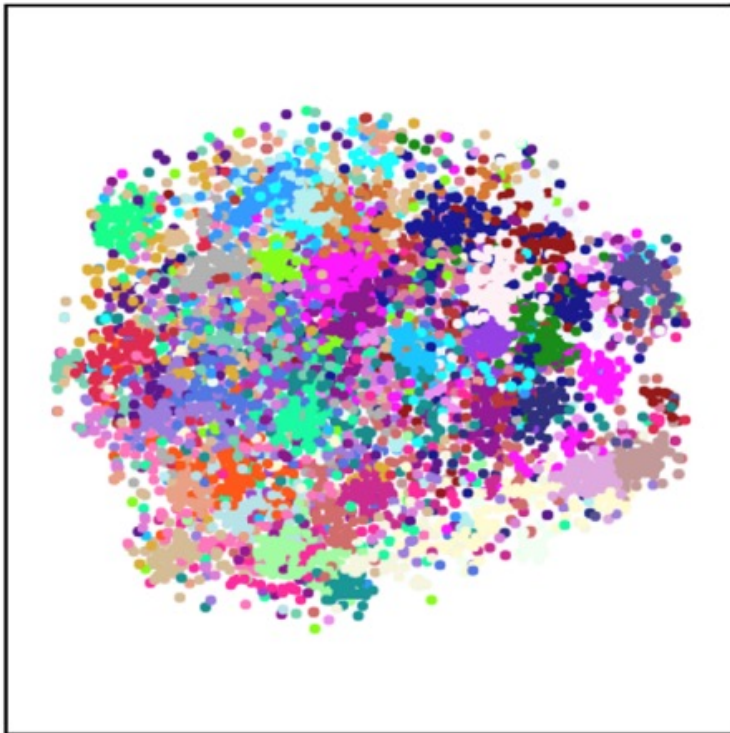
Brown corpus:

Balanced corpus of 1M words of English in various texts
(Kucela and Francis 1967)

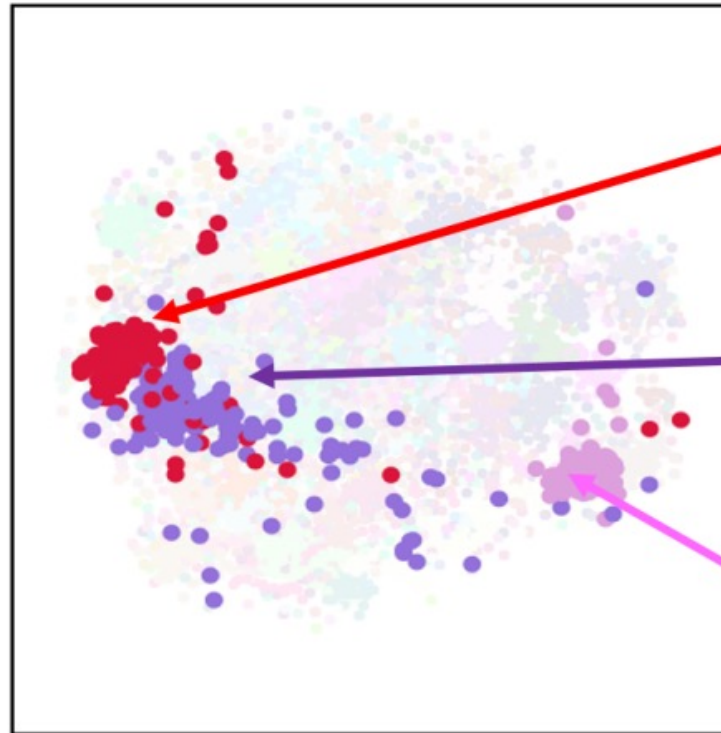
Latent cluster discovery @ Brown corpus

- Visualize plausible topics and their relationships through NPDV

all clusters



3 clusters



Religion

Pious christians rome receive...
Zen owes Chinese quietism Buddhism...

Social thought

Nationalism political principle epitomizes ...
Social civilizational factors rooted ...

Science

Bacteria formed typical activated sludge.
Spectra obtained temperature range ...

Summary

- NN-iWMM utilizes the power of neural clustering models with much less parameters
- NPDV integrates NN-iWMM and a visualization method to render latent coordinates suitable for visualization
- NPDV outperforms existing methods, and shows comparable accuracy to a well-tuned neural clustering models with much less tuning time.
- NPDV estimates plausible clusters without label information