DEEP DIRICHLET PROCESS MIXTURE MODELS

Naiqi Li*[1], Wenjie Li*[1], Yong Jiang[1,2], Shu-Tao Xia[1,2]

[1]Shenzhen International Graduate School, Tsinghua University

[2]PCL Research Center of Artificial Intelligence, Peng Cheng Laboratory

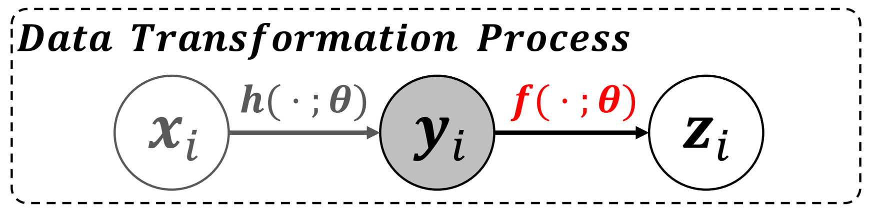
{ lnq18, liwj20}@mails.tsinghua.edu.cn {jiangy, xiast}@sz.tsinghua.edu.cn

http://github.com/naiqili/DDPM

Introduction

Problem and Preliminaries : Deep Clustering with Unknown K

- Learn representation with neural nets



- Clustering on the Transformed Space

$$p(\mathbf{z}_i|\boldsymbol{\mu}_k, \lambda_k) = \mathcal{N}(\mathbf{z}_i|\boldsymbol{\mu}_k, \lambda_k^{-1}\mathbf{I}) = (\frac{\lambda_k}{2\pi})^{\frac{d}{2}} \exp\left(-\frac{\lambda_k}{2}||\mathbf{z}_i - \boldsymbol{\mu}_k||^2\right)$$

- The Goal of clustering is to inference:

$$p(\mathbf{c}, \{\boldsymbol{\mu}_k\}_{k=1}^K, \{\lambda_k\}_{k=1}^K|\mathbf{Y};\boldsymbol{\theta},\boldsymbol{\Phi}). \quad \mathbf{c} = \{c_i \in \{1, \dots, K\}\}_{i=1}^N$$

- Challenges

1. The number of clusters K is unknown
2. A Neural Net for representation learning needs to be learned
3. the cluster information need to be computed at the same time

↘here follows the right bottom part

Robust Clustering Preferences Shown on Synthetic Data

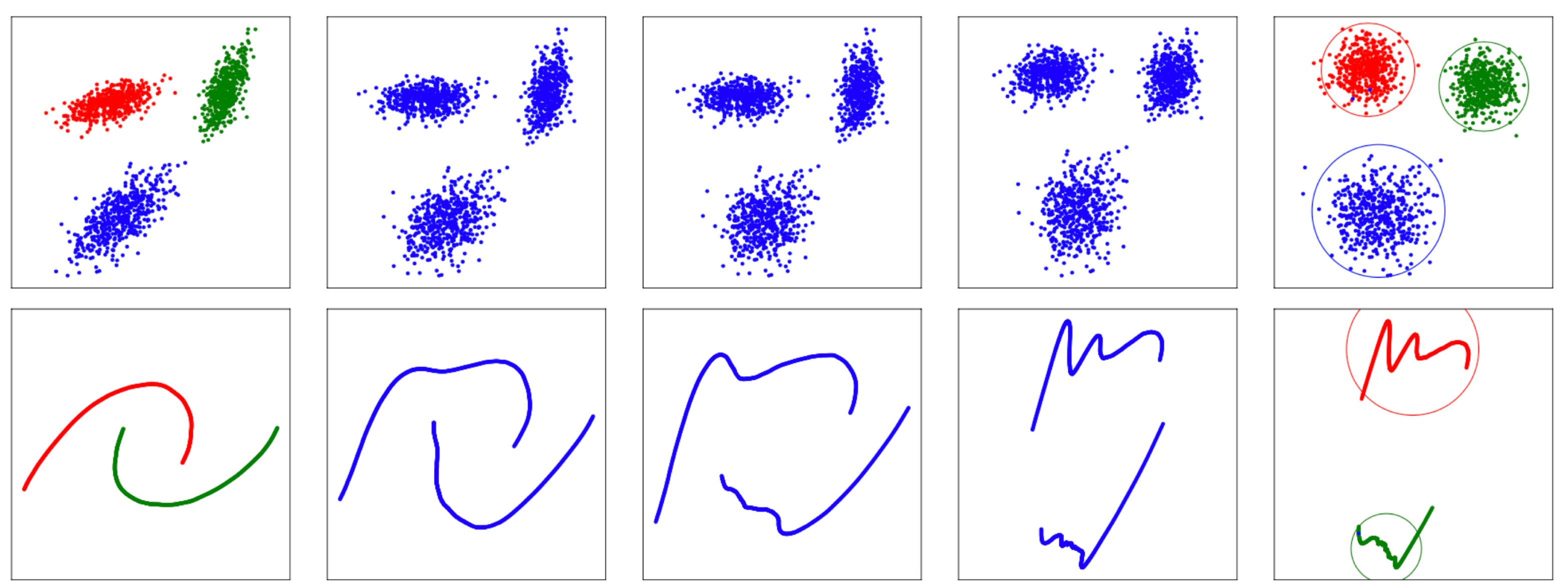
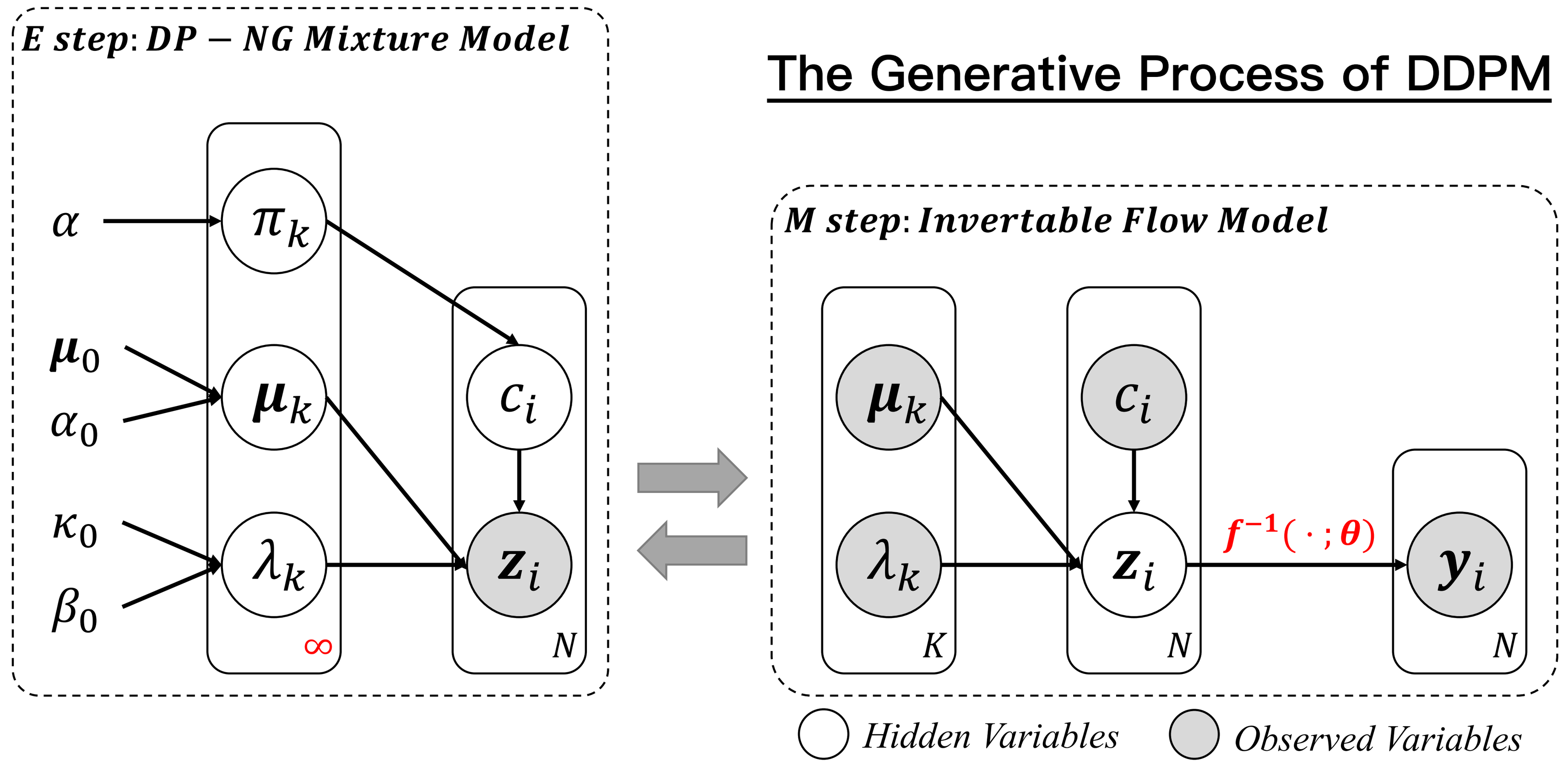


Figure 1: Demonstration of the clustering and representation learning process on two synthetic datasets. The leftmost figures are the ground truth clustering results. The middle figures show the latent representation learned by DDPM during the training. The rightmost figures show the final clustering results, with the circles denoting 2 standard deviations of the Gaussian distributions. We can see that DDPM is able to learn better representation during clustering. Particularly in the second example, the raw data representation is challenging for many centroid-based clustering methods, and the benefit of the new representation learned by DDPM is quite evident. Also note that the number of clusters is unknown in advance.

Methodology

Model Specification



Unified Parameter Estimation

By treating the cluster parameters and assignments as hidden variables, we maximize the complete data likelihood in the MC-EM framework

$$\text{E: } Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(old)}) = E_{\mathbf{H},\mathbf{c}|\mathbf{Y},\boldsymbol{\theta}^{(old)}}[\log p(\mathbf{H}, \mathbf{c}, \mathbf{Y}|\boldsymbol{\theta})]$$

$$\text{M: } \boldsymbol{\theta}^{(new)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(old)}) \quad \mathbf{H} = \{\{\boldsymbol{\mu}_k\}_{k=1}^K, \{\lambda_k\}_{k=1}^K\}$$

Gibbs Sampling

- The Conditionals of μ_k and λ_k:

$$p(\boldsymbol{\mu}_k, \lambda_k|\mathbf{H} \setminus \{\boldsymbol{\mu}_k, \lambda_k\}, \mathbf{c}, \mathbf{Z}^{(old)}) = NG(\boldsymbol{\mu}_k, \lambda_k|\boldsymbol{\mu}_n, \kappa_n, \alpha_n, \beta_n),$$

- The Conditional of c_i:

$$\log p(c_i = k|\mathbf{c}_{-i}, \mathbf{Z}^{(old)}, \mathbf{H}) =$$

$$\begin{cases} \log \frac{n_{-i,k}}{N-1+\alpha} + \log \mathcal{N}(\mathbf{z}_i|\boldsymbol{\mu}_k, \lambda_k^{-1}\mathbf{I}) + \text{const} \quad (\text{If } n_{-i,k} > 0) & \textit{for existing clusters} \\ \log \Gamma(\alpha_n') - \log \Gamma(\alpha_0) + \alpha_0 \log \beta_0 - \alpha_n' \log \beta_n' + \quad (\text{If } n_{-i,k} = 0) & \textit{for a new cluster} \\ \frac{1}{2}(\log \kappa_0 - \log \kappa_n') - \frac{nd}{2} \log 2\pi + \log \frac{\alpha}{N-1+\alpha} + \text{const}, & \end{cases}$$

Maximization

- The Gradient of the Q function:

$$\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + \lambda_s \frac{\partial}{\partial \boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(old)})$$

$$\approx \boldsymbol{\theta}_t + \frac{\lambda_s}{G} \sum_g \sum_i -\lambda_{c_i}^{(g)} (\mathbf{z}_i - \boldsymbol{\mu}_{c_i}^{(g)}) \frac{\partial f(\mathbf{y}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \leftarrow \textit{(the gradients of the flow model)}$$

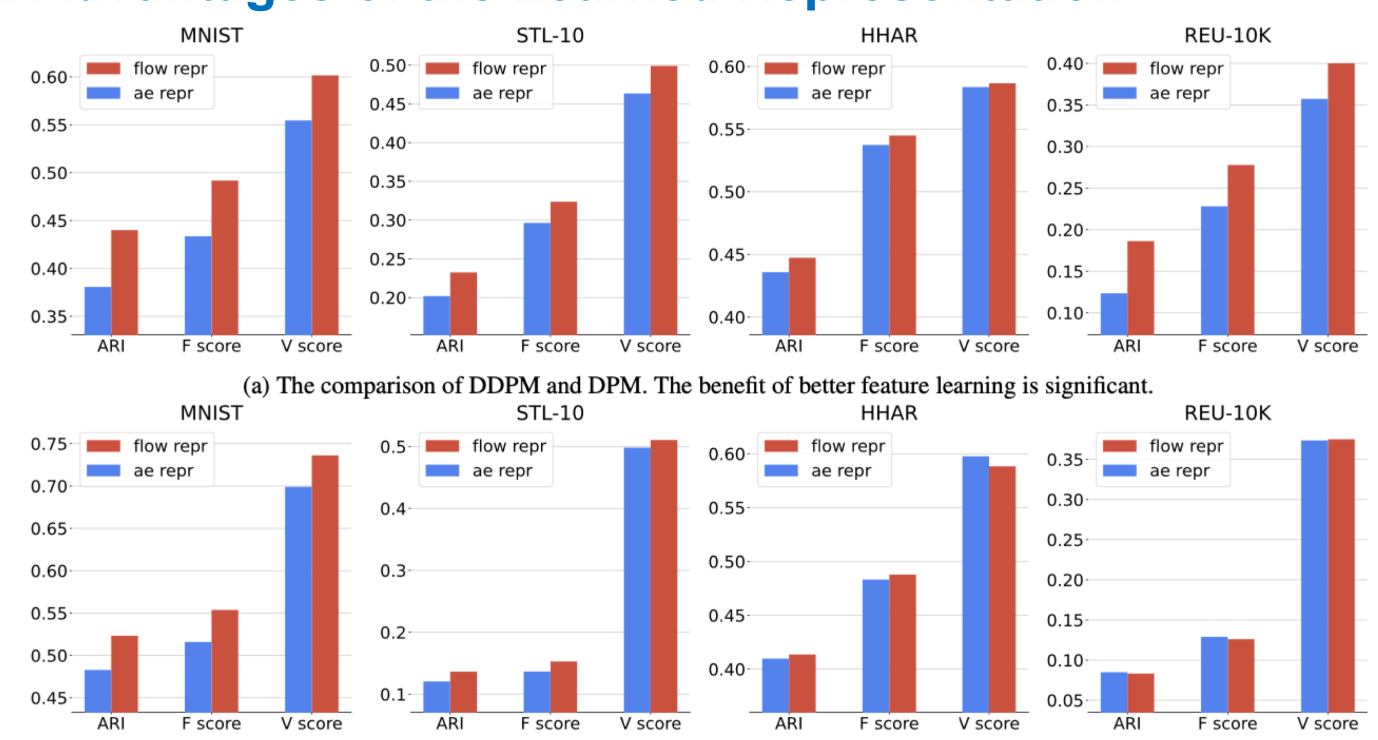
Experimental Study

K-agnostic Clustering Evaluation

Table 2: Performance comparison on real-world datasets.

Dataset	Methods	ARI	F score	V score
MNIST	G-means	0.1126	0.1255	0.5314
	DPM	0.3974	0.4511	0.5571
	DDPM	**0.4400**	**0.4917**	**0.6016**
HHAR	G-means	0.0904	0.1146	0.4358
	DPM	0.4342	0.5385	0.5761
	DDPM	**0.4473**	**0.5449**	**0.5865**
STL-10	G-means	0.2140	0.2512	0.4830
	DPM	0.2156	0.3073	0.4679
	DDPM	**0.2269**	**0.3193**	**0.4917**
REU-10K	G-means	0.0581	0.0933	0.3147
	DPM	0.1406	0.2365	0.3662
	DDPM	**0.1827**	**0.2756**	**0.3918**

Advantages of the Learned Representation



(a) The comparison of DDPM and DPM. The benefit of better feature learning is significant.

(b) The comparison of applying k-means to DDPM's learned feature and the raw feature.

Figure 2: The performance of clustering using the raw autoencoder features (ae repr) and DDPM's learned features (flow repr). (a) DDPM significantly outperforms DPM by learning better representation. (b) By using the learned features in the standard k-means, all metrics in almost all the datasets are improved, and the improvement in the MNIST dataset is particularly significant. This demonstrates DDPM's ability to learn better and transferable representation.

Generative Quality of the Learned Model

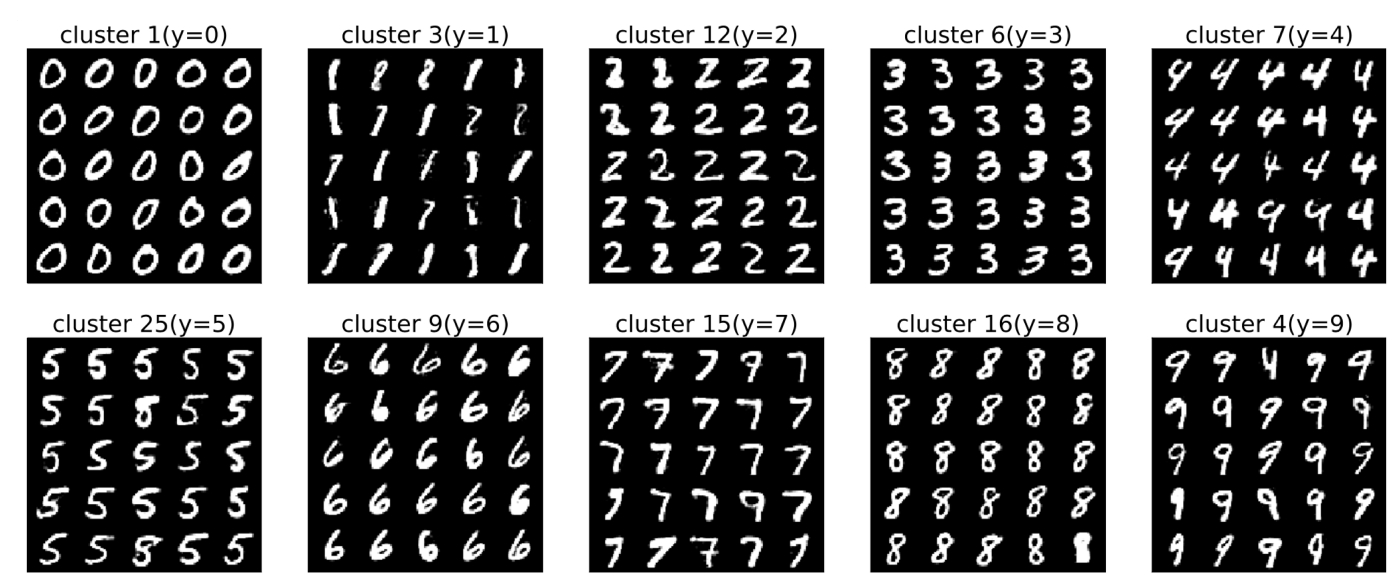


Figure 4: The generated handwritten digits in the MNIST dataset.