vMF-SNE: Embedding for Spherical Data

Mian Wang & Dong Wang*

CSLT/RIIT, Tsinghua University

wangdong99@mails.tsinghua.edu.cn

Presented by Zhiyuan Tang

ICASSP, 20-25 March 2016, Shanghai, China

OUTLINE

> > > > >

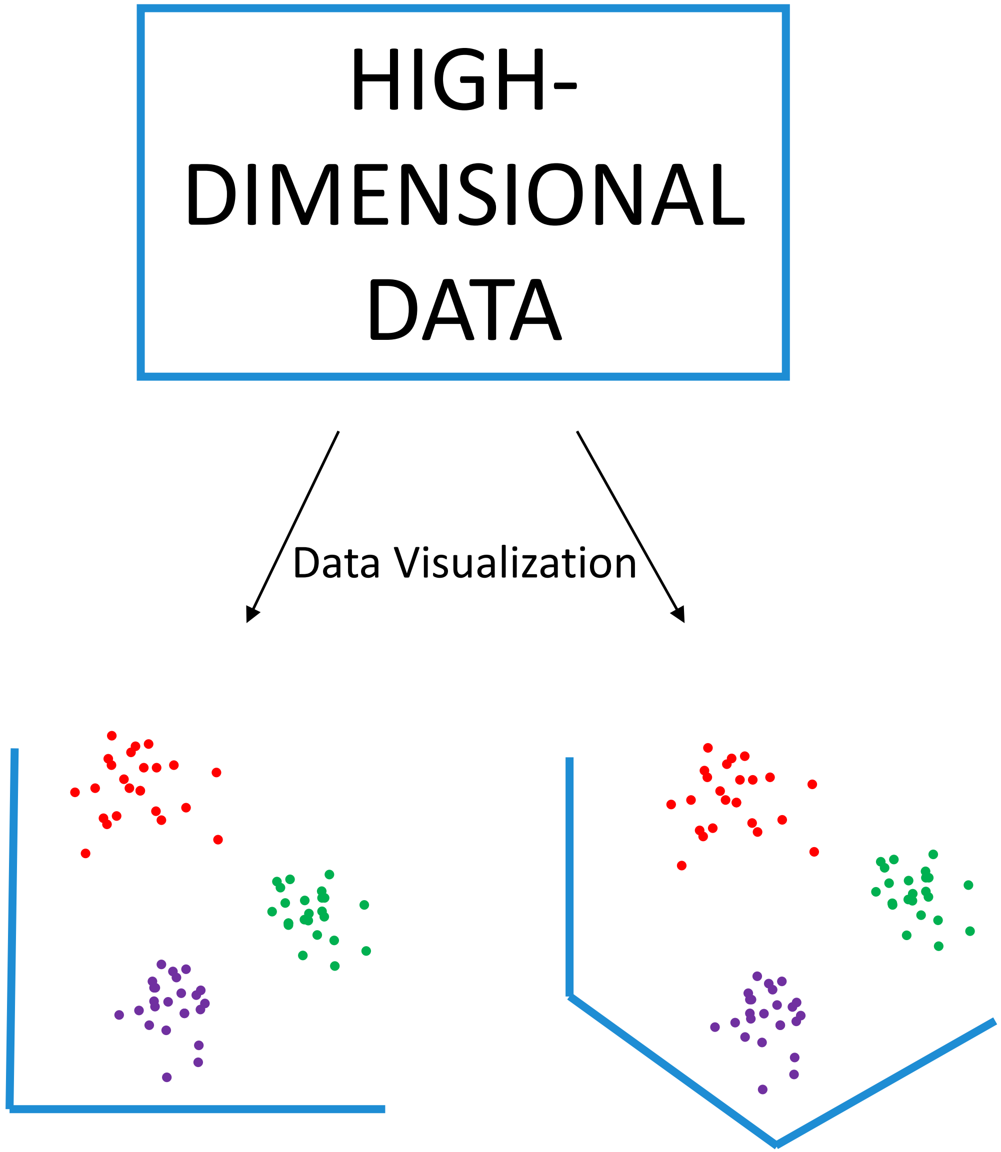
1 Introduction

2 From *t-SNE* to *vMF-SNE*

3 Experiments

4 Conclusions

Data Embedding



Data Embedding

- For data lying on or near a linear subspace

  Principal Component Analysis (PCA),
  Multi-Dimensional Scaling (MDS), etc.

- For data within non-linear manifolds

  - Derive the global non-linear structure from local proximity

    Isometric Feature Mapping (ISOMAP),
    self-organizing map (SOM),
    generative topographic mapping (GTM),
    local linear embedding (LLE),
    Stochastic Neighbor Embedding (SNE),
    UNI-SNE
    t-SNE

  - Derive the global non-linear structure involves kernel methods

    kernel PCA,
    colored maximum variance unfolding (CMVU), etc.

Data Embedding

- For data lying on or near a linear subspace

  Principal Component Analysis (PCA),
  Multi-Dimensional Scaling (MDS), etc.

- For data within non-linear manifolds

  - Derive the global non-linear structure from local proximity

    Isometric Feature Mapping (ISOMAP),
    self-organizing map (SOM),
    generative topographic mapping (GTM),
    local linear embedding (LLE),
    Stochastic Neighbor Embedding (SNE),
    UNI-SNE
    t-SNE

  - Derive the global non-linear structure involves kernel methods

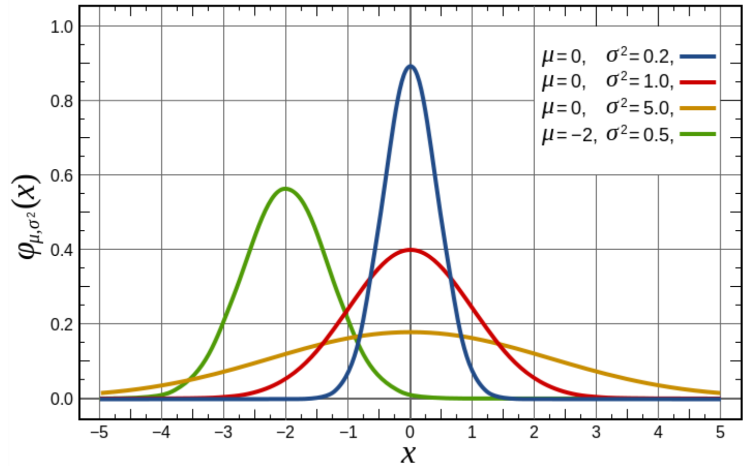
    kernel PCA,
    colored maximum variance unfolding (CMVU), etc.

Hinton, Geoffrey E., and Sam T. Roweis. "Stochastic neighbor embedding." Advances in neural information processing systems. 2002.

t-SNE

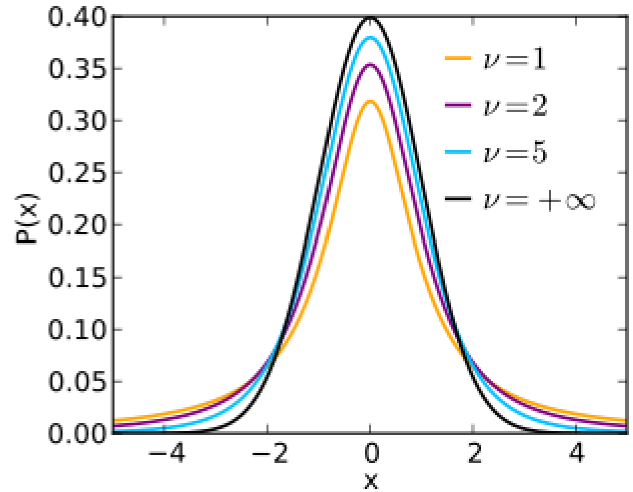
- Local Proximity in **original** data
  Gaussian distribution

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



- Local Proximity in **embedding** data
  Student t-distribution

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} \left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

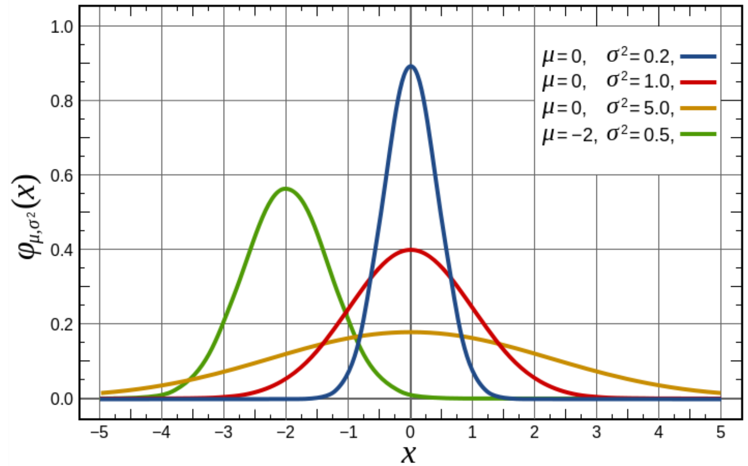


- **Optimization** (KL divergence, gradient descendant)

t-SNE

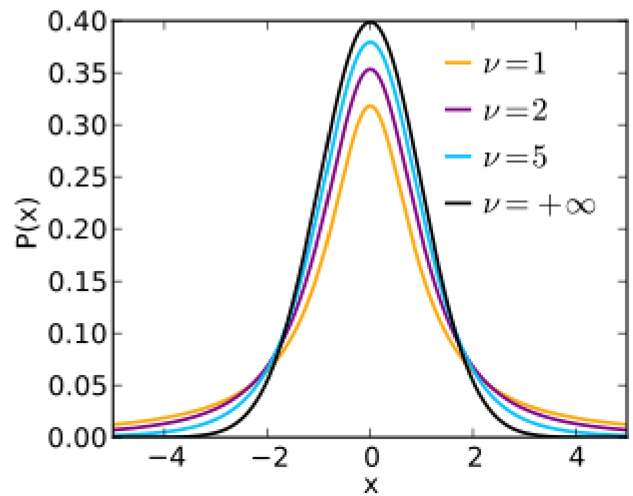
- Local Proximity in **original** data
  Gaussian distribution

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



- Local Proximity in **embedding** data
  Student t-distribution

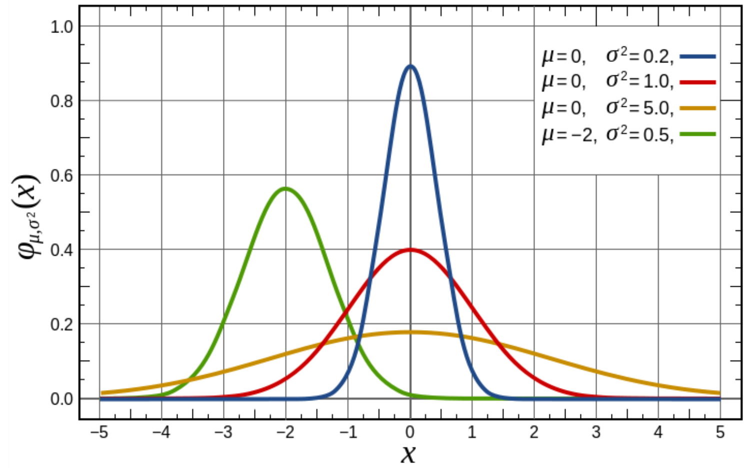
$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} \left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$



- **Optimization** (KL divergence, gradient descendant)

Afraid of data not Gaussian, such as spherical data!
topic vectors, i-vectors …

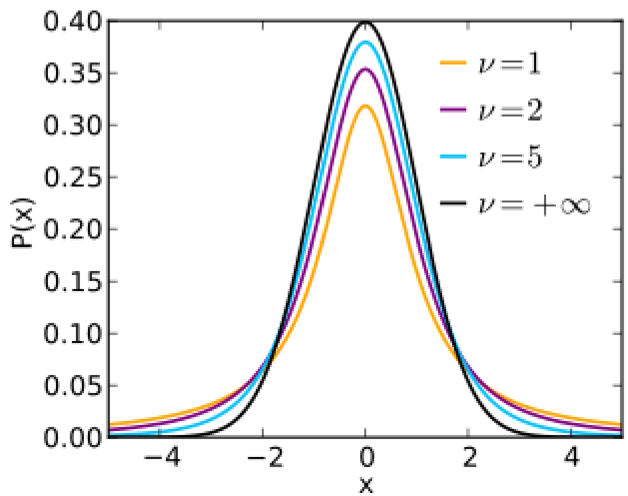
t-SNE



- Local Proximity in **original** data

  Gaussian distribution

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



- Local Proximity in **embedding** data

  Student t-distribution

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

- **Optimization** (KL divergence, gradient descendant)

This paper will handle spherical data, motivated by t-SNE!

OUTLINE

> > > > >

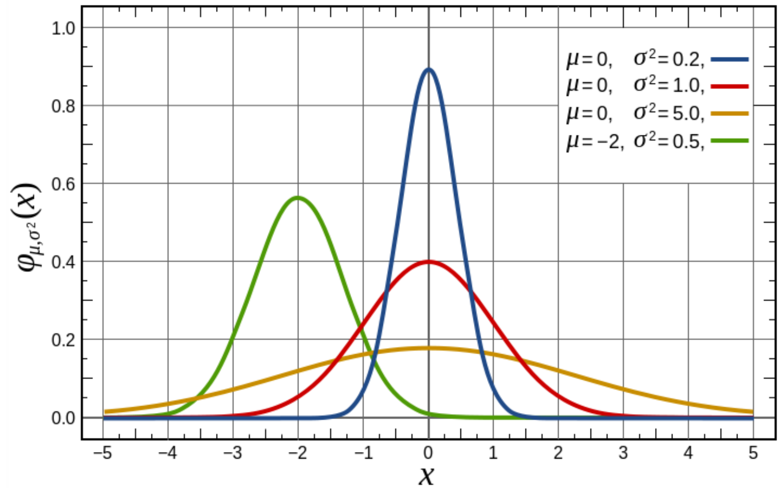
1 Introduction

2 From *t-SNE* to *vMF-SNE*

3 Experiments

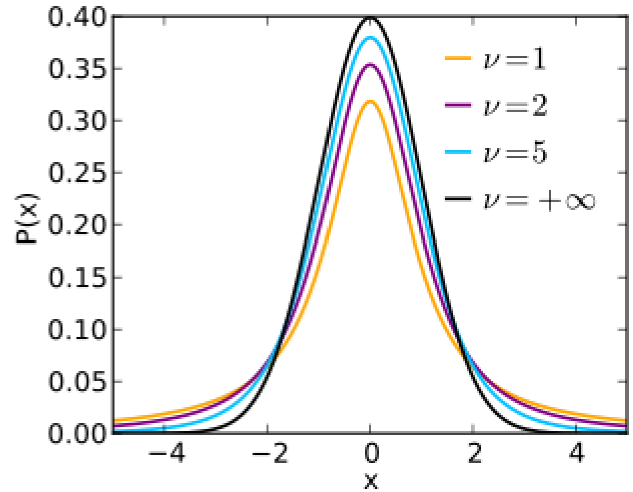
4 Conclusions

t-SNE



- Local Proximity in **original** data
  Gaussian distribution

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



- Local Proximity in **embedding** data
  Student t-distribution

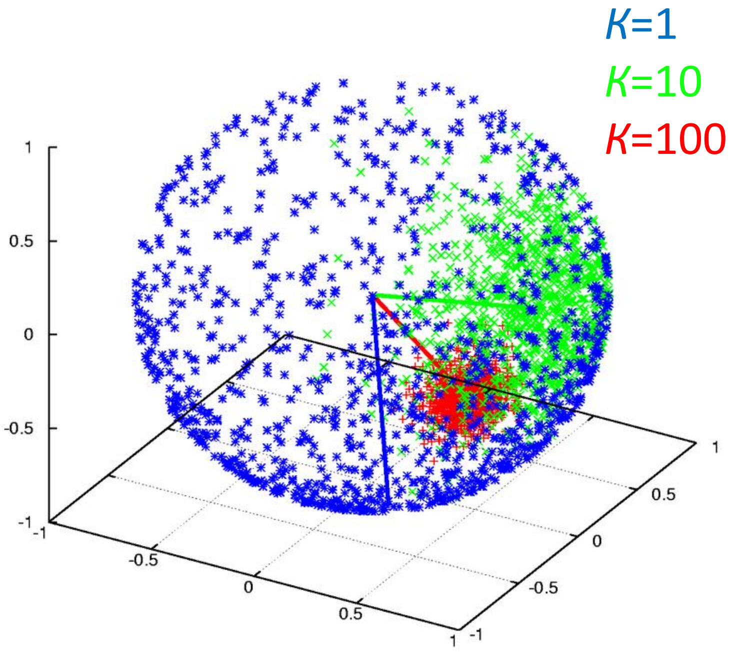
$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

- **Optimization** (KL divergence, gradient descendant)

vMF-SNE

- Local Proximity in **original** data
  vMF distribution

$$f_p(\mathbf{x};\boldsymbol{\mu},\kappa) = C_p(\kappa)\exp\left(\kappa\boldsymbol{\mu}^T\mathbf{x}\right)$$



- Local Proximity in **embedding** data
  vMF distribution

$$f_p(\mathbf{x};\boldsymbol{\mu},\kappa) = C_p(\kappa)\exp\left(\kappa\boldsymbol{\mu}^T\mathbf{x}\right)$$

- **Optimization** (KL divergence, gradient descendant)

vMF-SNE

Algorithm 1 vMF-SNE

Require:

Input:

$\{x_i; ||x_i|| = 1, i = 1, ..., N\}$: data to embed

$\mathcal{P}$: perplexity in the original space

κ: concentration parameter in the embedding space

T: number of iterations

η: learning rate

Output:

$\{y_i; ||y_i|| = 1, i = 1, ..., N\}$: data embeddings

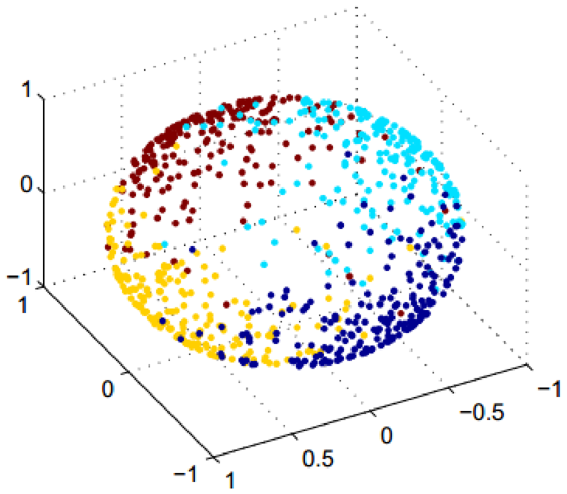
Procedure:

```
compute {κ_i} according to Eq. (9)
compute p_ij according to Eq. (4), and set p_ii = 0
randomly initialize {y_i}
for t = 1 to T do
    compute q_ij according to Eq. (5)
    for i = 1 to N do
        δ_i = ∂L̃/∂y_i according to Eq. (8)
        y_i = y_i + η δ_i
    end for
end for
```

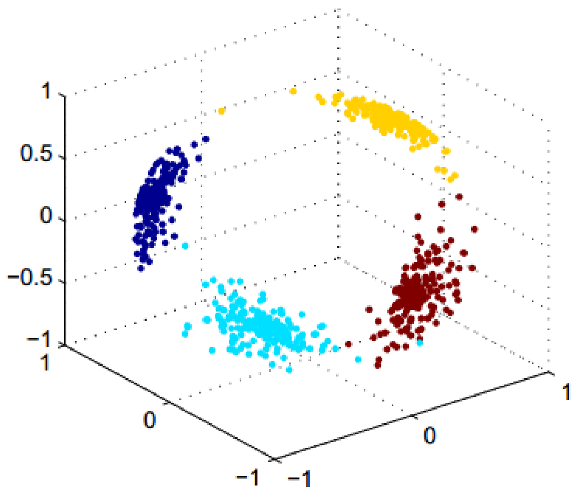
EM process

Visualization test

- vMF-SNE on simulation data



K = **15** for sampling data



K = **40** for sampling data

OUTLINE

> > > > >

1 Introduction

2 From *t-SNE* to *vMF-SNE*

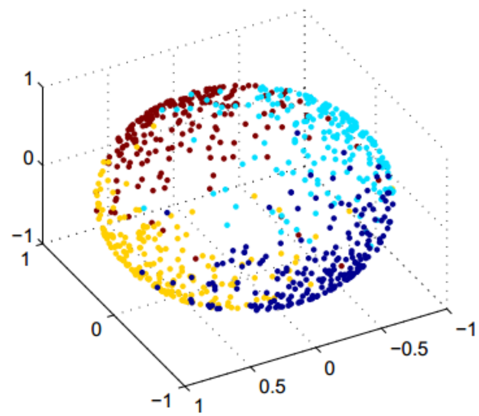
3 Experiments

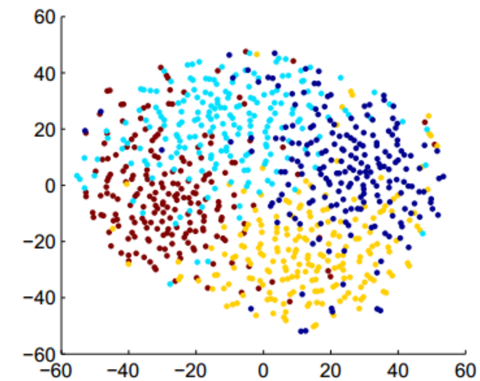
4 Conclusions

Visualization

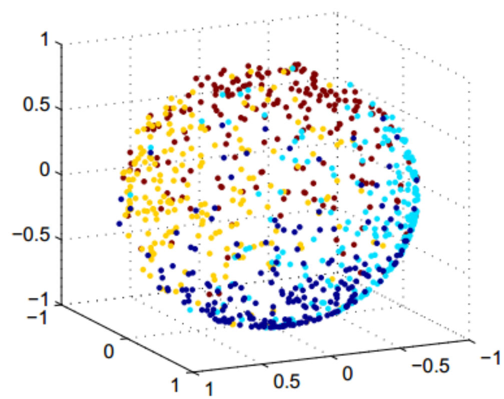
vMF-SNE vs. **t-SNE**

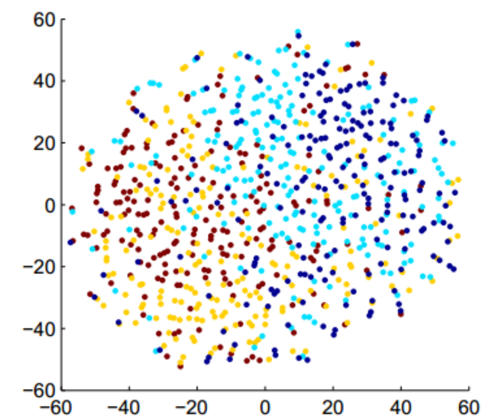
K=15 for sampling





K=10 for sampling





Entropy and accuracy

- vMF-SNE vs. t-SNE, **quantitative** criteria

  - Entropy

$$H(i) = \sum_{j=1}^{k} c(i,j) ln(c(i,j))$$

$c(i,j)$ is the proportion of the data points generated from the j-th cluster but are classified as the i-th cluster in the embedding space.

  - Accuracy

the proportion of the data that are correctly classified.

TABLE I: Results of Entropy and Accuracy

4 Clusters	Entropy		Accuracy	
κ	t-SNE	vMF-SNE	t-SNE	vMF-SNE
10	0.6556	0.5922	42%	64.13%
20	0.4725	0.4187	85.38%	92.63%
30	0.3804	0.3676	97.38%	98.5%
40	0.3485	0.3466	99.75%	99.95%
16 Clusters	Entropy		Accuracy	
10	0.3152	0.2975	15.5%	16.88%
20	0.2812	0.2608	38.25%	40.75%
30	0.2312	0.2383	68.25%	55.13%
40	0.1964	0.2187	91.25%	60.63%

OUTLINE

> > > > >

1 Introduction

2 From *t-SNE* to *vMF-SNE*

3 Experiments

4 Conclusions

Conclusions

- vMF-SNE assumes vMF distributions and cosine similarities with the original data and the embeddings.

- Compared with t-SNE, vMF-SNE is suitable for spherical data embedding.

- Future work involves studying long-tail vMF distributions to handle crowding data, as t-SNE does with the Student t-distribution.

- Tool for vMF-SNE from
  http://cslt.riit.tsinghua.edu.cn/resources.php?Public%20tools

Thanks a lot.