CSLT TECHNICAL REPORT-20150009 [Thursday 23rd April, 2015]

VMF-SNE: Embedding for Spherical Data

Mian Wang[1,3] and Dong Wang[1,2*]

*Correspondence: wang-dong99@cslt.riit.tsinghua.edu.cn
[1]CSLT, RIIT, Tsinghua University, 100084 Beijing, China
Full list of author information is available at the end of the article

Abstract

T-SNE is a well-known approach to embedding high-dimensional data and has been widely used in data visualization. The basic assumption of t-SNE is that the data are non-constrained in the Euclidean space and the neighbouring proximity can be modeled by Gaussian distributions. This assumption does not hold for a wide range of data types in practical applications, for instance spherical data for which the neighbouring proximity is better modelled by the von Mises-Fisher (vMF) distribution instead of the Gaussian. This letter present a vMF-SNE embedding algorithm to embed spherical data. An iterative process is derived to produce an efficient embedding. The results on a simulation data set demonstrated that vMF-SNE produces better embeddings than t-SNE for spherical data.

Keywords:
visualization; high-dimensional unit data; T-SNE; vMF; vMF-SNE

1 Introduction

High-dimensional data embedding is a challenging task in machine learning and is important for many applications particularly data visualization. Principally, data embedding involves projecting high-dimensional data to a low-dimensional (often 2 or 3) space where the major structure (distribution) of the data in the original space is mostly preserved. Therefore data embedding can be regarded as a special task of data dimension reduction, with the objective function set to be preserving the structure of the data. .

Various traditional dimension reduction approaches can be used to perform data embedding, e.g., the principal component analysis(PCA) [11] and the multi-dimensional scaling(MDS) [4]. PCA finds low-dimensional embeddings that preserve the data covariance as much as possible. Classical MDS finds embeddings that preserve inter-sample distances, which is equivalent to PCA if the distance is Euclidean. Both the PCA and MDS are simple to implement and efficient in computation, and are guaranteed to discover the true structure of data lying on or near a linear subspace of the high-dimensional input space. The shortage is that they are ineffective for data within non-linear manifolds.

A multitude of non-linear embedding approaches have been proposed to deal with non-linear manifolds. The first approach is to derive the global non-linear structure from local proximity. For example, ISOMAP extends MDS by calculating similarities of distant pairs based on similarities of neighbouring pairs [21, 22]. The self-organizing map (SOM) or Kohonen net extends PCA and derives the global non-linearity by simply ignoring distant pairs [13]. The same idea triggers the generative topographic mapping (GTM) [3], where the embedding problem is cast to a

Bayesian inference with an EM procedure. The local linear embedding(LLE) follows the same idea but formulates the embedding as a local-structure learning based on linear prediction [17]. Another approach to deriving the global non-linear structure involves various kernel learning methods, e.g., the semi-definite embedding based on kernel PCA [24] and the colored maximum variance unfolding (CMVU) [19].

A major problem of the above non-linear embedding methods is that most of them are not formulated in a probabilistic way, which leads to potential problems in generalizability and tractability. The stochastic neighbor embedding(SNE) [10] attempts to solve the problem. It models neighbouring proximity of data in both the original and embedding space by Gaussian distributions, and the embedding processing minimizes the kullback-leibler(KL) divergence of the distributions in the original space and the embedding space.

A potential drawback of SNE is the 'crowding problem', i.e., the data samples tend to be crowded together in the embedding space van2008visualizing. A UNI-SNE approach was proposed to deal with the problem, which introduces a symmetric cost function and a smooth model when computing similarities between the images of data in the embedding space [5]. With the problem in concern, [23] proposed t-SNE, which also uses a symmetric cost function, but employs a Student t-distribution rather a Gaussian distribution when computing similarities between images (embeddings). T-SNE has shown clear superiority over other embedding methods particularly for data that lie within several different, but related, low-dimensional manifolds.

Although highly effective in general, t-SNE is weak in embedding data that are not Gaussian. For example, there are many applications where the data are distributed on a hyper-sphere, such as the topic vectors in document processing [16] and the normalized i-vectors in speaker recognition [6]. These spherical data are naturally modelled by the von Mises–Fisher (vMF) distribution rather than the Gaussian [9, 8, 15], hence unsuitable to be embedded by t-SNE. This paper will present a vMF-SNE algorithm to embed spherical data, based on the SNE philosophy. Specifically, the Gaussian distribution and the Student t-distribution used by t-SNE in the original and the embedding space respectively are all replaced by vMF distributions, and an EM-based iterative process is derived to conduct the embedding. The experimental results on simulation data show that vMF-SNE produces better embeddings for spherical data.

The rest of the paper is organized as follows. Section 2 describes the related work, and Section 3 presents the vMF-SNE method. The simulation experiment is presented in 4, and the paper is concluded in Section 5.

2 Related work

This work belongs to the extensively studied area of dimension reduction and data embedding. Among the rich methods, the linear embedding approaches are simple and fast, including PCA, MDS and their extensions [11, 4]. Non-linear embedding approaches involve ISOMAP [21, 22], GTM [3]. LLE [17], Sammon mapping [18], curvilinear components analysis(CCA) [7], maximum variance unfolding(MUV) [24, 19], and Laplacian eigenmaps [2]. SNE and its extensions [10, 5] formulate the embedding problem as an optimization task in the probabilistic framework, which is the foundation of this research. Particularly, t-SNE [23] solves the

data crowding problem and has established the state-of-the-art in data embedding and visualization. Our work is motivated by t-SNE, and is designed specifically to embed spherical data which are not suitable to be processed by t-SNE. A more related work is the parametric embedding (PE) [12], which embeds vectors of posterior probabilities, thus sharing a similar goal as our proposal: both attempt to embed data in a constrained space though the constrains are different (ℓ-1 in PE and ℓ-2 in vMF-SNE).

Probably the most relevant work is the spherical semantic embedding (SSE) [14]. In the SSE approach, document vectors and topic vectors are constrained on a unit sphere and are assumed to follow the vMF distribution. The topic model and the embedding model are then jointly optimized in a generative model framework by maximum likelihood. However, SSE infers local similarities between data samples (document vectors in [14]) using a pre-defined latent structure (topic vectors), which is difficult to be generalized to other tasks as the latent structure in most scenarios is not available. Additionally, the cost function of SSE is the likelihood, while vMF-SNE uses the symmetric Kullback-Leibler(KL) divergence.

3 vMF-distributed stochastic neighbouring embedding

3.1 Problems of t-SNE

Let $\{x_i\}$ denotes the data set in the high-dimensional space, and $\{y_i\}$ denotes the corresponding embeddings, or images. The t-SNE algorithm measures the pairwise similarities in the high-dimension space as the joint distribution of x_i and x_j which is assumed to be Gaussian, formulated by the following:

$$p_{ij} = \frac{e^{-||x_i - x_j||^2/2\sigma^2}}{\sum_{m \neq n} e^{-||x_m - x_n||^2/2\sigma^2}}. \tag{1}$$

In the embedding space, the joint probability of y_i and y_j is modelled by a Student t-distribution with one degree of freedom, given by:

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{m \neq n} (1 + ||y_m - y_n||^2)^{-1}}. \tag{2}$$

The cost function of the embedding is the KL divergence between $p_{i,j}$ and $q_{i,j}$, formulated by:

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}}.$$

A gradient descendant approach has been devised to conduct the optimization, which is fairly efficient. Additionally, the symmetric form of Eq. (1) and the long-tail property of the Student t-distribution alleviates the crowding problem of the original SNE and other embedding approaches.

It should be highlighted the two assumptions that t-SNE holds: the joint probabilities of the original data samples and the embeddings follow a Gaussian distribution

and a Student t-distribution, respectively. This is general fine in most scenarios, however fordata that are confined in a non-linear subspace, this assumption is potentially invalid and the t-SNE embedding is no longer optimal. This paper focuses on spherical data embedding, for which the t-SNE tends to fail as the Gaussian distribution assumed by t-SNE can hardly model spherical data, and the Euclidean distance associated with Gaussian distributions is not appropriate to measure similarities on a hyper-sphere. A new embedding algorithm is proposed, which shares the same embedding framework as in t-SNE, but uses a more appropriate distribution form and a more suitable similarity measure to model spherical data.

3.2 vMF-SNE

It has been shown that the vMF distribution is a better choice than the Gaussian in modelling spherical data, and the associated cosine distance is better than the Euclidean distance in measuring data similarities in a hyper-spherical space, for instance, in tasks such as spherical data clustering [20, 1]. Therefore, the vMF-SNE presented here assumes vMF distributions in both the original and the embedding space.

Mathematically, the probability density function of the vMF distribution on the (d-1)-dimensional sphere in R^d is given by:

$$f_d(x;\mu,\kappa) = C_d(\kappa)e^{\kappa\mu^T x}$$

where $||x|| = ||\mu|| = 1$, $\kappa > 0$ and μ are parameters of the distribution and $C_d(\kappa)$ is the normalization constant. Note that the vMF distribution implies the cosine distance. As in t-SNE, the symmetric distance is used in both the original and embedding space. In the original space, define the conditional probability of x_j given x_i as:

$$p_{j|i} = \frac{f_d(x_j;x_i,\kappa_i)}{\sum_{m\neq i} f_d(x_m;x_i,\kappa_i)},$$

the joint distribution p_{ij} is defined as follows:

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2}. \tag{3}$$

In the embedding space, a simpler form of joint distribution is chosen by setting the concentration parameter k_i the same for all y_i. This choice follows t-SNE, and the rationale is that the distribution $p_{j|i}$ in the original space needs to be adjusted according to the data scattering around x_i, however doing so in the embedding space will cause complexity in computation, as we will see shortly. The joint distribution q_{ij} with this simplification is given by:

$$q_{ij} = \frac{e^{\kappa y_i^T y_j}}{\sum_{m\neq n} e^{\kappa y_m^T y_n}}. \tag{4}$$

As in t-SNE, the KL divergence between the two distributions is used as the cost function:

$$\mathcal{L} = \sum_i \sum_j p_{ij} ln \frac{p_{ij}}{q_{ij}} \tag{5}$$

By gradient descendant, minimizing $\mathcal{L}$ with respect to $\{y_i\}$ leads to the optimal embedding. The gradients will be derived in the following section.

3.3 Gradient derivation

Note that

$$\mathcal{L} = \sum_{i,j} p_{ij} ln(p_{ij}) - \sum_{i,j} p_{ij} ln(q_{ij}).$$

Since the first item on the right hand side of the equation is in dependent of the embedding, minimizing $\mathcal{L}$ reduces to maximizing the following cost function:

$$\tilde{\mathcal{L}} = \sum_{i,j} p_{ij} ln(q_{ij}).$$

Define $Z = \sum_{m \neq n} e^{k y_m^T y_n}$, we have:

$$\tilde{\mathcal{L}} = \kappa \sum_{i,j} p_{ij} y_i^T y_j - lnZ$$

where $\sum_{i,j} p_{ij} = 1$ has been employed. The gradient of $\tilde{\mathcal{L}}$ with respect to the embedding y_k is then derived as:

$$\begin{aligned} \frac{\partial \tilde{\mathcal{L}}}{\partial y_k} &= 2\kappa \sum_i p_{ik} y_i - \frac{1}{Z} \frac{\partial lnZ}{\partial y_k} && (6) \\ &= 2\kappa \sum_i p_{ik} y_i - \frac{2\kappa}{Z} \{ \sum_i e^{\kappa y_i^T y_k} y_i \} && (7) \\ &= 2\kappa \sum_i (p_{ik} - q_{ik}) y_i && (8) \end{aligned}$$

This is a rather simple form and the computation is efficient. Note that this simplicity is partly due to the identical κ in the embedding space.

Algorithm 1 illustrates the vMF-SNE process. Notice that in the original data space, κ_i is required. Following [23], κ_i is set to a value that makes the perplexity $\mathcal{P}_i$ is equal to a pre-defined value $\mathcal{P}$, formulated by:

$$\mathcal{P}_i = 2^{H(p_{j|i})} \tag{9}$$

where $H(\cdot)$ is information entropy defined by:

$$H(p_{j|i}) = -\sum_j p_{j|i} log_2(p_{j|i})$$

where $p_{j|i}$ has been defined in Eq. (3). As mentioned in [23], making the perplexity associated to each data point the same value normalizes the data scattering and so benefits outliers and crowding areas.

Algorithm 1 vMF-SNE

Input & Output:
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
Implementation:
compute $\{\kappa_i\}$ according to Eq. (9)
compute p_{ij} according to Eq. (3), and set $p_{ii} = 0$
randomly initialize $\{y_i\}$
for $t = 1$ to T **do**
  compute q_{ij} according to Eq. (4)
  for $i = 1$ to N **do**
    $\delta_i = \frac{\partial \mathcal{L}}{\partial y_i}$ according to Eq. (8)
    $y_i = y_i + \eta \delta_i$
  end for
end for

4 Experiment

To evaluate the proposed vMF-SNE, it is employed to visualize both spherical data and Gaussian data and is compared with the traditional t-SNE. Since visualization is not a quantitative evaluation, an entropy-based criterion is proposed to compare the two embedding approaches.

4.1 Data simulation

The experiments are based on simulation data. The basic idea is to sample k clusters of data samples and examine if the cluster structure can be preserved after embedding. The sampling is straightforward for Gaussian data; for spherical data following the vMF distribution, it deserves some discussion.

The sampling process starts from the centers of the k clusters, i.e., $\{\mu_i; ||\mu_i|| = 1, i = 1, ..., k\}$. This is attained by sampling each dimension of μ_i independently according to a Gaussian $N(0, 1)$, and then place an ℓ-2 normalization to respect $||\mu_i|| = 1$. This process can be repeated to obtain every μ_i independently. However, a different approach is adopted in this study: firstly sample the first center μ_1, and then a new μ_i is derived by randomly selecting a subset of the dimensions of μ_1 and then flipping the sign of the values on these dimensions. By this way, the centers $\{\mu_i\}$ are ensured to be separated on the hyper-sphere, which generates a clear cluster structure associated with the data.

Once the cluster centers are generated, it is easy to sample the data points for each cluster following the vMF distribution. A toolkit provided by Arindam Banerjee and

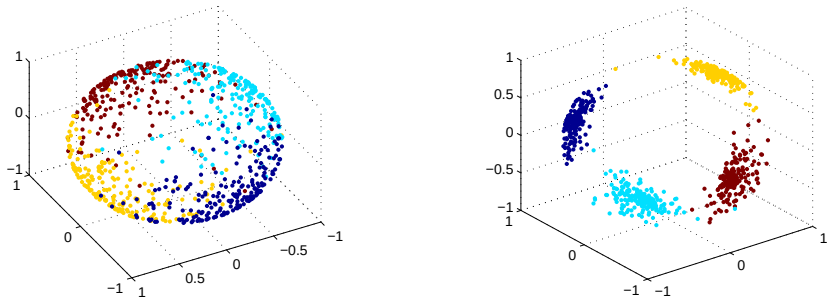

Figure 1 The 3-dimensional embedding with vMF-SNE, with data generated following a vMF distribution by setting $\kappa = 15$ (left) and $\kappa = 40$ (right). The original dimension is 50, and there are 4 clusters, each of which is represented by a particular color.

Suvrit Sra was adopted to conduct the vMF sampling[1]. In this work, the dimension of the data point is set to 50, and the number of clusters varies from 4 to 16. For each cluster 200 data points are sampled. The concentration parameter κ used in the sampling also varies, in order to investigate the performance of the embedding approaches in different overlapping conditions.

4.2 Visualization test

The first experiment visualizes the spherical data with vMF-SNE. The perplexity $\mathcal{P}$ is set to 40, and the value of κ in the embedding space is fixed to 2 (see Algorithm 1). The data are generated following vMF distributions by setting the scattering parameter κ to different values. Fig. 1 presents the embedding results on 3-dimensional spheres with vMF-SNE, where the two pictures show the results with κ=15 and κ=40 respectively. Note that the κ here is used in data sampling, neither the κ used to model the original data (which is computed from $\mathcal{P}$ for each data point) nor the κ used to model the embedding data (which has been fixed to 2). It can be seen that vMF-SNE indeed preserves the cluster structure of the data in the embedding space, and not surprisingly, data generated with a larger κ are more separated in the embedding space.

For comparison, the same data are embedded with t-SNE in 2-dimensional space. The tool provided by Laurens van der Maaten is used to conduct the embdding[2], where the perplexity is set to 40. The comparative results are shown in Fig. 2 and Fig. 3 for data generated by setting κ=15 and κ=10 respectively. It can be observed that when κ is large (Fig. 2), both vMF-SNE and t-SNE perform well and the cluster structure is clearly preserved. However when κ is small (Fig. 3), vMF-SNE shows clear superiority. This suggests that t-SNE is capable to model spherical data if the structure is clear, even if the underling distribution is non-Gaussian; however in the case where the structure is less discernable in the high-dimensional space, t-SNE tends to mess the boundary while vMF-SNE still works well.

[1] http://suvrit.de/work/soft/movmf
[2] http://lvdmaaten.github.io/tsne/

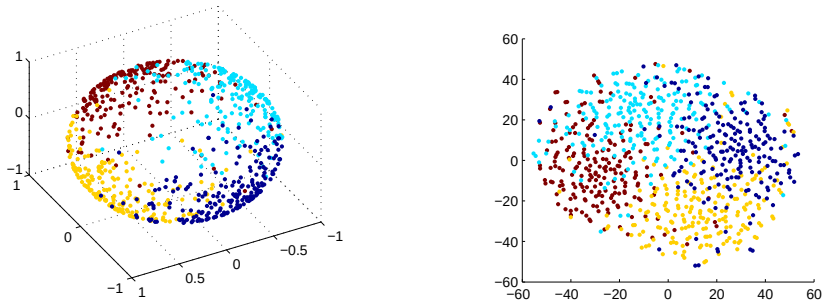

Figure 2 The 3-dimensional embedding with vMF-SNE (left) and 2-dimensional embedding with t-SNE (right). The data was generated following a vMF distribution by setting $\kappa = 15$.

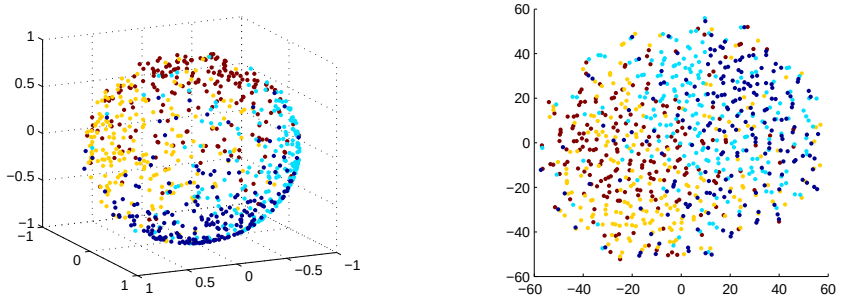

Figure 3 The 3-dimensional embedding with vMF-SNE (left) and 2-dimensional embedding with t-SNE (right). The data was generated following a vMF distribution by setting $\kappa = 10$.

4.3 Entropy and accuracy test

Visualization test is not quantitative. For further investigation, we propose to use the clustering accuracy and entropy as the criteria to measure the quality of the embedding. This is achieved by first finding the images of the cluster centers, denoted by $\{\hat{\mu}_i\}$, and cluster the data by finding their nearest $\{\hat{\mu}_i\}$ in the embedding space. The classification accuracy is computed as the proportion of the data that are correctly assigned to their clusters in the original space. The entropy of the i-th cluster is then derived by

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

where $c(i,j)$ is the proportion of the data points generated from the j-th cluster but are assigned to i-th cluster according to the nearest-neighbour rule in the embedding space. The entropy of the entire dataset is computed as the average of $H(i)$ over all the clusters. The results are presented in Table 1. It can be observed that in the case of 4 clusters, vMF-SNE achieves lower entropy and better accuracy than t-SNE when κ is small. If κ is large, both the two methods can achieve good performance, for the reason that we have discussed.

Table 1 Embedding on vMF data

4 Clusters	Entropy		Accuracy	
κ	t-SNE	vMF-SNE	t-SNE	vMF-SNE
10	4.7891	3.9174	48.38%	64.39%
20	2.0208	1.1727	86.5%	93.25%
30	0.2672	0.1339	98.75%	99.38%
40	0.063	0.0315	99.75%	99.88%
16 Clusters	Entropy		Accuracy	
10	37.3485	31.7905	16.25%	17%
20	29.3969	23.4429	35%	41.13%
30	14.5074	15.9642	75.13%	57.63%
40	10.2464	12.7097	80.88%	54%

Table 2 Embedding on Gaussian Data

4 Clusters		Entropy		Accuracy	
α	σ	t-SNE	vMF-SNE	t-SNE	vMF-SNE
8	6	1.1778	1.1147	91.5%	92%
8	6.5	1.5508	1.2595	85.5%	91%
9	6	0.7753	0.5982	94.25%	96.5%
9	6.5	1.1802	0.9499	90%	93.75%
10	6	0.6438	0.5297	96.75%	97.75%
10	6.5	0.8238	0.6006	94.5%	96.5%
16 Clusters		Entropy		Accuracy	
8	6	16.9689	18.6505	49%	44%
8	6.5	19.4723	20.2595	47.25%	42%
9	6	16.0618	18.1033	52.75%	45.5%
9	6.5	16.9288	18.5997	51.75%	44.25%
10	6	12.3825	13.1265	61.5%	53.25%
10	6.5	13.3391	16.5243	60.5%	49%

In the case of 16 clusters, it is observed that vMF-SNE outperforms t-SNE with small κ values (large overlaps). This seems an interesting property and demonstrates that using the matched distribution (vMF) is helpful to improve embedding for overlapped data. However, with *kappa* increases, vMF-SNE can not research a performance as good as obtained by t-SNE. A possible reason is that the large number of clusters leads to data crowding which can be better addressed with the long-tail Student t-distribution used by t-SNE. Nevertheless, this requires further investigation.

Another interesting investigation is to examine the performance of t-SNE and vmf-SNE on Gaussian data. The data generation process is similar to the one used for generating the vMF data. Firstly sample a center vector μ_1 following a Gaussian distribution and normalize its vector length to α. After that, the rest cluster centers $\{\mu_i\}$ are produced by flipping the sign of the values of a subset of dimensions of μ_1. Once the cluster centers are obtained, the data points of each cluster are generated by an isotropic Gaussian $N(\mu_i, \sigma)$, where σ controls the within-class variance.

The results are presented in Table 2. The interesting observation is that when the number of data clusters are small, vMF-SNE is still better than t-SNE, and if the number of clusters increase, t-SNE is superior. Again, we conjecture the discrepancy is caused by the capability of the long-tail property of the Student t-distribution used in t-SNE, which is capable of modeling crowding data.

5 Conclusions

We propose a vMF-SNE algorithm for embedding high-dimensional spherical data. Compared with the widely used t-SNE, vMF-SNE assumes vMF distributions and cosine similarities of the original data and the embeddings, hence suitable for spherical data embedding. The experiments on a simulation dataset demonstrated that

the proposed approach works fairly well. Further work involves studying long-tail vMF distributions to handle crowding data, as t-SNE does with the Student t-distribution.

Acknowledgement

This research was supported by the National Science Foundation of China (NSFC) under the project No. 61371136, and the MESTDC PhD Foundation Project No. 20130002120011. It was also supported by Sinovoice and Huilan Ltd.

Author details
[1]CSLT, RIIT, Tsinghua University, 100084 Beijing, China. [2]TNList, Tsinghua University, 100084 Beijing, China.
[3]Beijing University of Posts and Telecommunications, 100084 Beijing, China.

References

1. A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra, "Clustering on the unit hypersphere using von mises-fisher distributions," in *Journal of Machine Learning Research*, 2005, pp. 1345–1382.
2. M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering." in *NIPS*, vol. 14, 2001, pp. 585–591.
3. C. M. Bishop, M. Svensén, and C. K. Williams, "Gtm: The generative topographic mapping," *Neural computation*, vol. 10, no. 1, pp. 215–234, 1998.
4. I. Borg and P. J. Groenen, *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.
5. J. Cook, I. Sutskever, A. Mnih, and G. E. Hinton, "Visualizing similarity data with a mixture of maps," in *International Conference on Artificial Intelligence and Statistics*, 2007, pp. 67–74.
6. N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.
7. P. Demartines and J. Hérault, "Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets," *Neural Networks, IEEE Transactions on*, vol. 8, no. 1, pp. 148–154, 1997.
8. I. S. Dhillon and S. Sra, "Modeling data using directional distributions," Citeseer, Tech. Rep., 2003.
9. N. I. Fisher, *Statistical analysis of circular data*. Cambridge University Press, 1995.
10. G. E. Hinton and S. T. Roweis, "Stochastic neighbor embedding," in *Advances in neural information processing systems*, 2002, pp. 833–840.
11. H. Hotelling, "Analysis of a complex of statistical variables into principal components." *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.
12. T. Iwata, K. Saito, N. Ueda, S. Stromsten, T. Griffiths, and J. Tenenbaum, "Parametric embedding for class visualization," *Neural Computation*, vol. 19, no. 9, pp. 2536–2556, 2007.
13. T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 43, no. 1, pp. 59–69, 1982. [Online]. Available: http://dx.doi.org/10.1007/BF00337288
14. T. Le and H. W. Lauw, "Semantic visualization for spherical representation," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 1007–1016.
15. K. V. Mardia and P. E. Jupp, *Directional statistics*. John Wiley & Sons, 2009, vol. 494.
16. J. Reisinger, A. Waters, B. Silverthorn, and R. J. Mooney, "Spherical topic models," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 903–910.
17. S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
18. J. W. Sammon, "A nonlinear mapping for data structure analysis," *IEEE Transactions on computers*, vol. 18, no. 5, pp. 401–409, 1969.
19. L. Song, A. Gretton, K. M. Borgwardt, and A. J. Smola, "Colored maximum variance unfolding," in *Advances in neural information processing systems*, 2007, pp. 1385–1392.
20. A. Strehl, J. Ghosh, and R. Mooney, "Impact of similarity measures on web-page clustering," in *Workshop on Artificial Intelligence for Web Search (AAAI 2000)*, 2000, pp. 58–64.
21. J. B. Tenenbaum, "Mapping a manifold of perceptual observations," *Advances in neural information processing systems*, pp. 682–688, 1998.
22. J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.
23. L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 2579-2605, p. 85, 2008.
24. K. Q. Weinberger, F. Sha, and L. K. Saul, "Learning a kernel matrix for nonlinear dimensionality reduction," in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 106.