High-Dimensional Data Visualisation via Sampling-Based Approaches

Effect of Perplexity at different levels of Sampling-Based Approach

Muhammad Arslan Bhatti
(M.A.Bhatti@student.tudelft.nl)

Backgroun

The **t-distributed Stochastic Neighbor Embedding (t-SNE)** algorithm is a popular method for visualizing high-dimensional data in two or three dimensions. It operates by minimizing the Kullback-Leibler (KL) divergence between two probability distributions: one defined over pairs of points in the original high-dimensional space, and the other over the corresponding points in the low-dimensional embedding space [1].

A key hyperparameter in t-SNE is **perplexity**, which determines the effective number of neighbors considered for each point during the computation of similarity distribution. Formally, for each data point x_i, t-SNE fits a Gaussian distribution centered at x_i and adjusts its variance such that the Shannon entropy of the resulting conditional similarity distribution matches a target perplexity value. This is done via binary search on to solve:

where is the Shannon entropy of a discrete distribution.

When t-SNE is applied to datasets exhibiting hierarchical structure, it presents two weaknesses:

i. t-SNE does not always preserve the global structure [2] because it inherently preserves local neighborhoods
ii. running t-SNE multiple times to tune perplexity becomes computationally expensive [3].

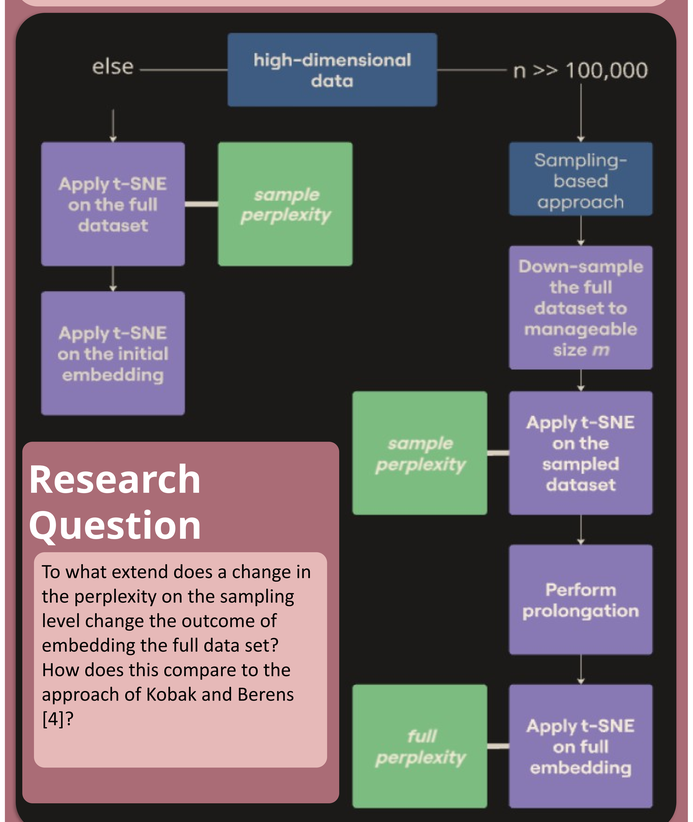
Kobak and Berens [3] propose the following:

iii. multi-scale similarities: for each data point , similarities to other points are computed using two Gaussian kernels with two different bandwidths and , corresponding to two perplexities:

where is the Euclidean distance between the points.

iv. For very large datasets, use sample-based approach as shown in Figure 1

Another study [4] proposes that the user can choose the sample perplexity, and then full perplexity is calculated by:



Research Question

To what extend does a change in the perplexity on the sampling level change the outcome of embedding the full data set? How does this compare to the approach of Kobak and Berens [4]?

Figure 1: Sample-based approach proposed by Kobak and Berens

Methodolog

To analyse our approach:

- create a grid of embeddings using various combinations of sample and full perplexities to systematically study their effects on the final embeddings.
- Analyse qualitatively by inspecting cluster separation, cluster fragmentation, structure preservation, and interpretability.
- Analyse quantitatively using the following metrics:
  1. KNN (k-nearest neighbor): Measures how well local neighborhoods are preserved by checking if each point's closest neighbors in the high dimensional space stay close together in the embedding as well.
  2. KNC (k-nearest class means): Assesses preservation of relationships between classes by comparing the nearest class centers before and after embedding.
  3. CPD ((Cross-Pairwise Distance Correlation): Evaluates global structure by measuring the correlation between all pairwise distances in the original and embedded spaces.

To compare with Kobak and Berens approach:

- We cover both cases for small datasets and for very large datasets.
- The perplexity settings are replicated; all other parameters are held constant.
- For each of perplexity values used in the multi-scale setting, we generate a matching embedding using our method with corresponding sample perplexity and full perplexity.
- We also include scaling-based full perplexities from Navigating t-SNE (2023) for further comparison.

Results

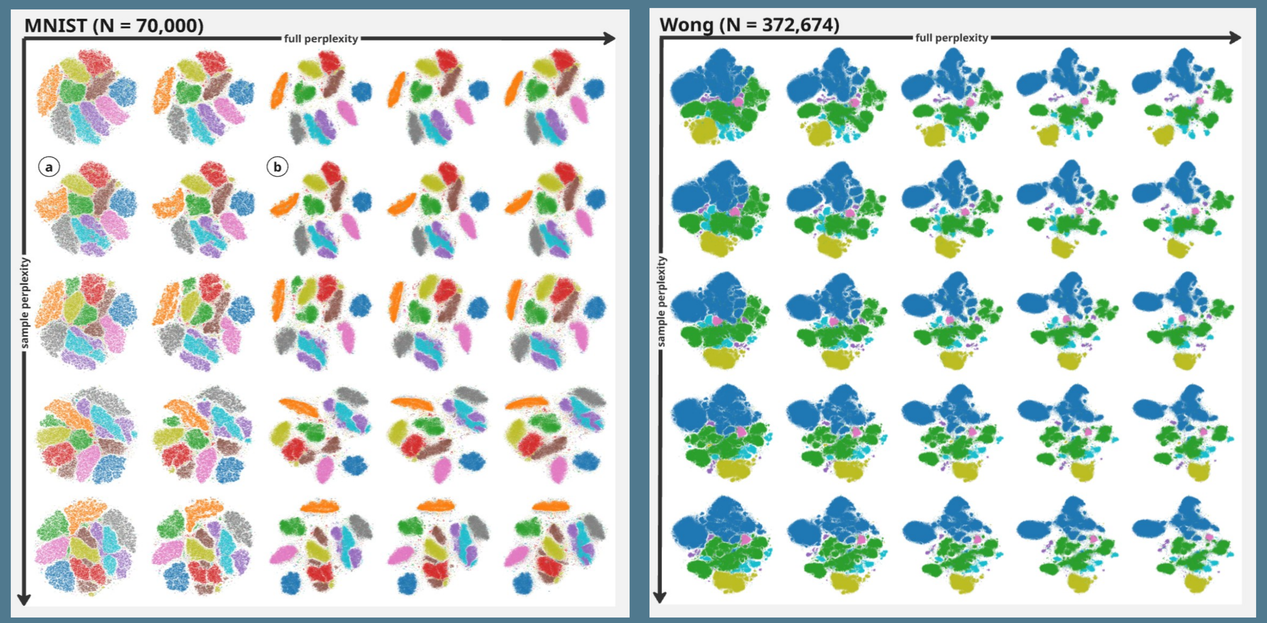
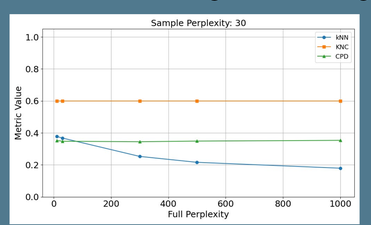


Figure 2: Embeddings for the MNIST dataset (left) and Wong dataset (right) produced by our approach



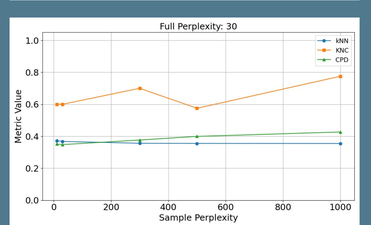
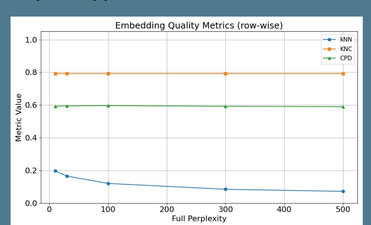


Figure 3: Metrics plotted for the MNIST dataset across rows (above) and across columns (below)

Dataset	Sample Perp	Full Perp	CPD
MNIST	30	30	0.990
	30	500	0.959
	30	1000	0.954
Tasic et al.	30	30	0.992
	30	100	0.977
	30	300	0.939

Dataset	Sample Perp	Full Perp	CPD
MNIST	300	500	0.861
	500	500	0.782
	1000	500	0.643
Tasic et al.	30	100	0.920
	100	100	0.397
	300	100	0.546

Figure 4:CPD values between embeddings for the MNIST and Tasic et al. datasets across rows (above) and across columns where the reference embedding is the first in the row or column



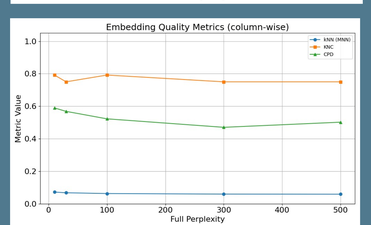


Figure 5: Metrics plotted for the Wong dataset across rows (above) and across columns (below)

Results

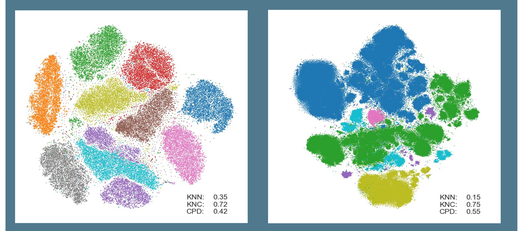


Figure 2: Embeddings for the MNIST (left) and Wong dataset (right) produced by Kobak and Berens

- Across rows (fixed sample perplexity), increasing full perplexity does not change the over all structure; however clusters start to overlap more. This can be seen with stable CPD values across rows, and the KNN value declining suggesting loss of local detail.
- Across columns (fixed full perplexity): varying sample perplexity causes significant structural changes. Clusters may fragment or shift positions. This can be seen by varying CPD and KNC values showing changes in global/mesoscopic structure. KNN remains stable indicating preserved local neighborhood.

- For MNIST, embedding by Kobak & Berens approach shows clear cluster separation with some overlap. Our embedding exhibits more cluster overlap; it also has lower KNN and CPD value. For Wong dataset, Our method shows fragmentation in the green cluster, resulting in lower KNN than Kobak & Berens.
- Strategy to compare based on previous results: Match global layout using embedding with low full perplexity (from first column), and then analyse refinement by increasing full perplexity and comparing quality metrics.
- Highlight (a) matches global layout at sample perplexity 30.
- Highlight (b) (scaled full perplexity) shows KNN drop to 0.25, lower than Kobak & Berens.

Conclusion

- Sample perplexity defines the global layout, while full perplexity controls local refinement in t-SNE embeddings.
- multi-scale method achieves better local structure preservation (higher KNN), our method is better suited for interpretable and customizable visualizations.

References

[1] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. Journal of machine learning research, 9(Nov):2579–2605, 2008.
[2] John A Lee, Diego H Peluffo-Ord´o˜nez, and MichelVerleysen. Multi-scale similarities in stochastic neighbour embedding: Reducing dimensionality while preserving both local and global structure.Neurocomputing, 169:246–261, 2015.
[3] Dmitry Kobak and Philipp Berens. The art ofusing t-sne for single-cell transcriptomics. NatureCommunications, 10(1):5416, 2019.
[4] Martin Skrodzki, Nicolas F. Chaves de Plaza, Thomas Hollt, Elmar Eisemann, and Klaus Hildebrandt.Navigating perplexity: A linear relationship with the data set size in t-sne embeddings, 2024.