Comparative study on uMap and t-SNE

Author: Deniz Geçer
Supervisor: Stéphane Marchand-Maillet

Plan

- Problems
- Questions
- State of the art
- Proposed solution
- Experiments
- Results
- Conclusion

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département d'informatique

Problems

Curse of dimensionality

- Huge quantity of data
- Overfitting
- Equidistance of points
- Sparsity
- High correlation

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département d'informatique

Questions

- How is the reliability of the reduced data with respect to the original one measured?
- On which criteria the comparison between multiple reduction techniques is made
- What is the relation between the strength of the reduction and the loss of information?
- How is the stability of reduction techniques analyzed?

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département d'informatique

State of the art

Evaluation metrics and methodology

(Stephen France, 2007)

- Stress as a metric (MDS)
- global distance conservation

$$STRESS = \sqrt{\frac{\sum_i \sum_j \left(d_{ij} - \hat{d}_{ij}\right)^2}{\sum_i \sum_j d_{ij}^2}}$$

(Rasa Karabauskaite, 2008)

- Topology preservation measures
- Spearman's rho $\rho_{Sp} = 1 - \frac{6 \sum_{i=1}^{T} (r_{\mathbf{X}}(i) - r_{\mathbf{Y}}(i))^2}{T^3 - T}$
- Konig's Measure $\mathrm{KM} = \frac{1}{3k_1 \times m} \sum_{i=1}^{m} \sum_{j=1}^{k_1} \mathrm{KM}_{ij}.$
- Mean Relative Rank Errors (MRRE)

$$\mathrm{MRRE}(\mathbf{X} \to \mathbf{Y}) = \frac{1}{C} \sum_{i=1}^{m} \sum_{j \in \mathrm{N}_K(X_i)} \frac{|\overline{\overline{r}}_{\mathbf{X}}(i,j) - \overline{\overline{r}}_{\mathbf{Y}}(i,j)|}{\overline{\overline{r}}_{\mathbf{X}}(i,j)},$$

$$\mathrm{MRRE}(\mathbf{Y} \to \mathbf{X}) = \frac{1}{C} \sum_{i=1}^{m} \sum_{j \in \mathrm{N}_K(Y_i)} \frac{|\overline{\overline{r}}_{\mathbf{X}}(i,j) - \overline{\overline{r}}_{\mathbf{Y}}(i,j)|}{\overline{\overline{r}}_{\mathbf{Y}}(i,j)},$$

State of the art

Evaluation metrics and methodology

(Francisco J. García-Fernández, 2013)

- Stability comparaison methodology
- Impact of parameters on embedding
- Impact of data on embedding (Increase number of points)
- Fixing the initialization technique and minimizing randomness

(Antonio Gracia, 2014)

- Methodology to compare the reduction in terms of loss of quality
- Use and comparison of all previous metrics

$$Quality\ Loss = (1 - quality\ value)$$

State of the art

Evaluation metrics and methodology

(Olga Kurasova, 2018)

- Comparison of reduction technique
- Clustered data
- Non clustered data
- Real-world data

(Ruizhi Xiang, 2021)

- Aggregation of all metrics to evaluate overall stability
- Sensitivity of methods to Hyperparameters

Proposed solution

Datasets:

- Swiss roll
- Clustered data
- Non clustered data

Evaluation metrics:

- Shepard diagram
- Configuration score
- Stress
- Spearman's rho
- Konig's measure
- Mean relative rank error
- Loss

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département d'informatique

Experiments

Experiment 1

- Add white noise as a 4th dimension with different std to the Swiss Roll dataset
- Perform reduction to 3D and 2D using 4 different initialization methods, 4 different std and 4 different perplexity/k
- Observe loss function, shepard diagram and plot

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département d'informatique

Experiments

Experiment 2

- Run of 16 different configurations on each dataset with each technique
- 4 different initialization methods and 4 different perplexity/k value
- Compare the 16 configurations using a defined configuration score

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département d'informatique

Experiments

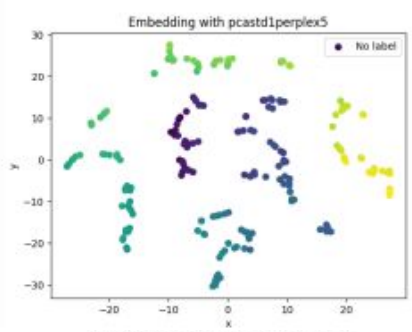
Experiment 3

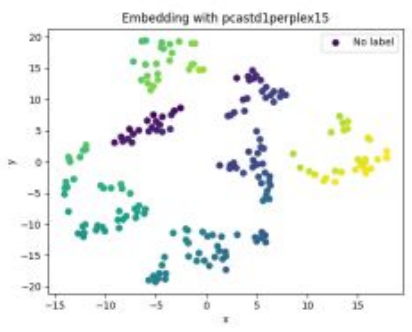
- Choose a specific configuration
- Perform 2D reduction on Swiss-Roll dataset with different perplexity/k value
- Perform 2D reduction on clustered and non-clustered data with different features size
- Evaluate the impact of perplexity/k and dimensionality using
    - Stress
    - Spearman's rho
    - Konig's measure
    - Mean relative rank error

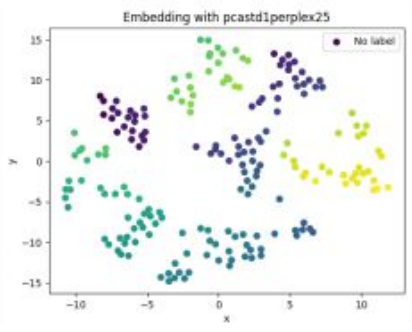
UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département d'informatique

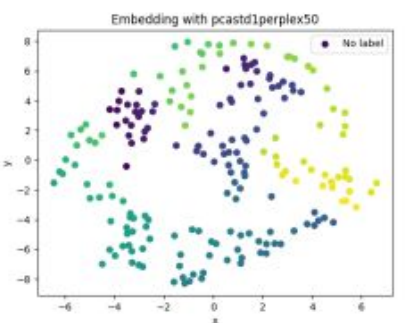
Experiment 1

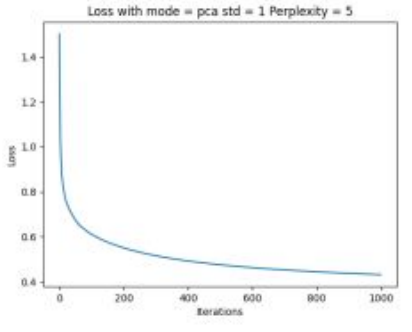
Results

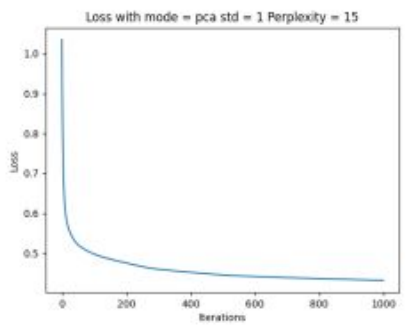


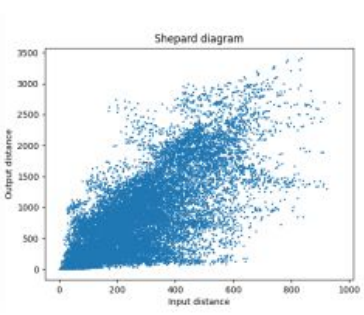


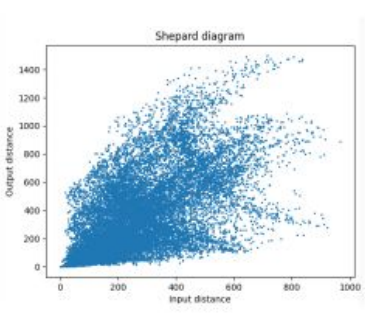






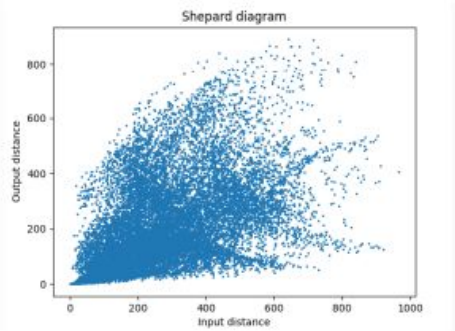


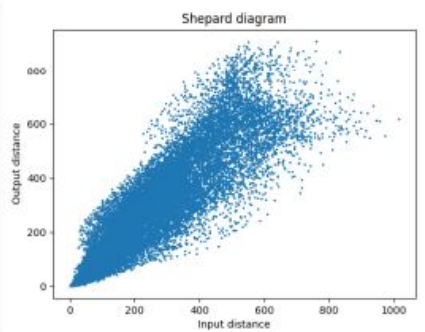


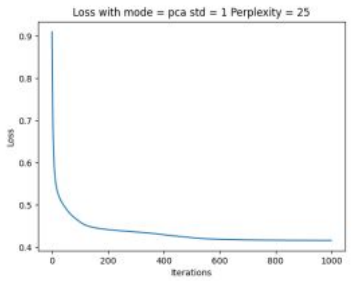


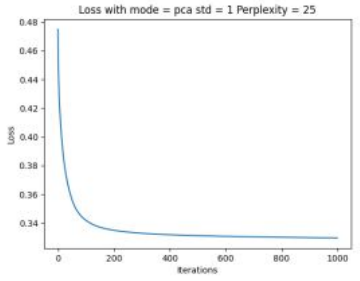
Experiment 1

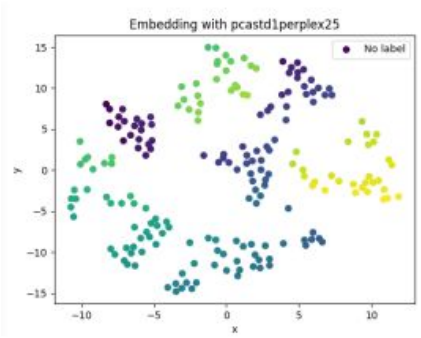
Results

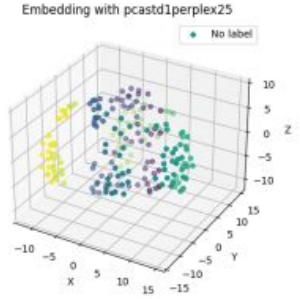












Experiment 2

Results

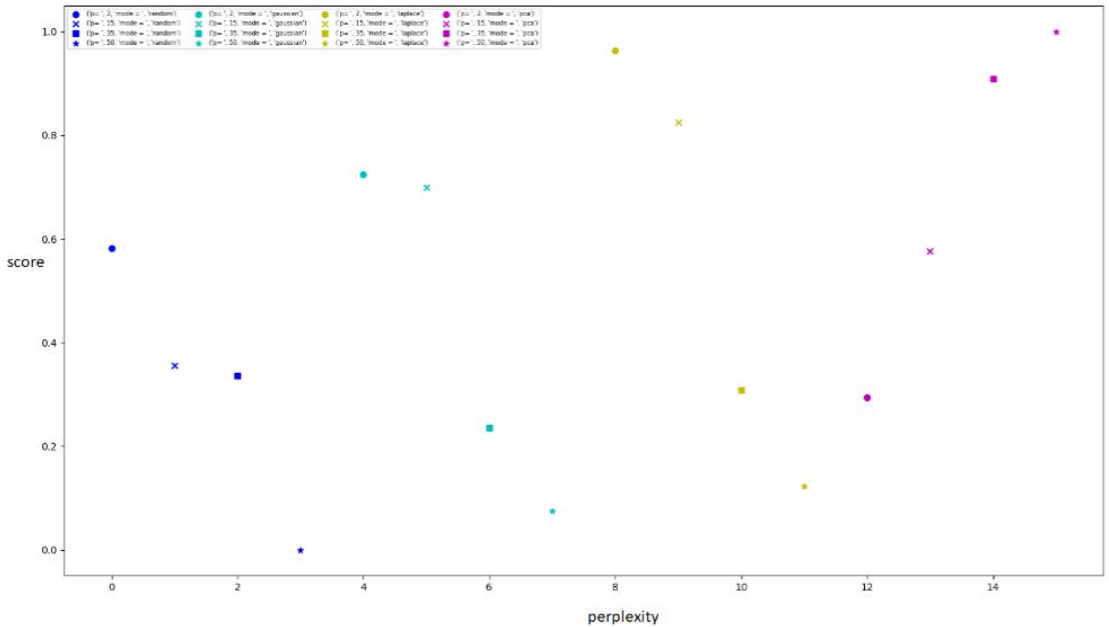


Figure 29: Configuration score - Swiss Roll - t-SNE

Experiment 2

Results

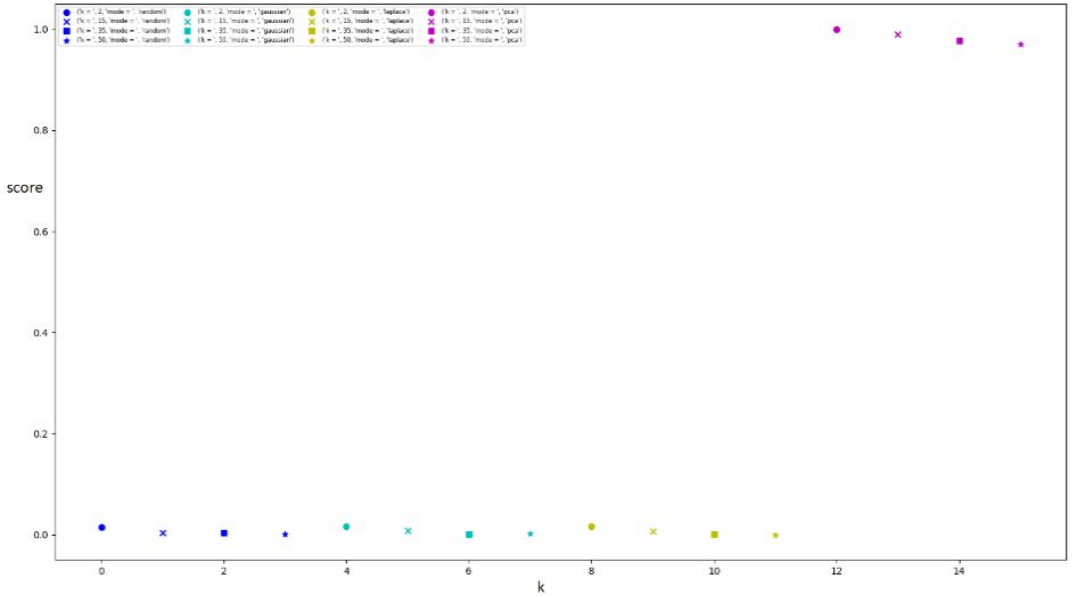


Figure 32: Configuration score - Swiss Roll - uMap

Experiment 3

Results

Swiss-roll

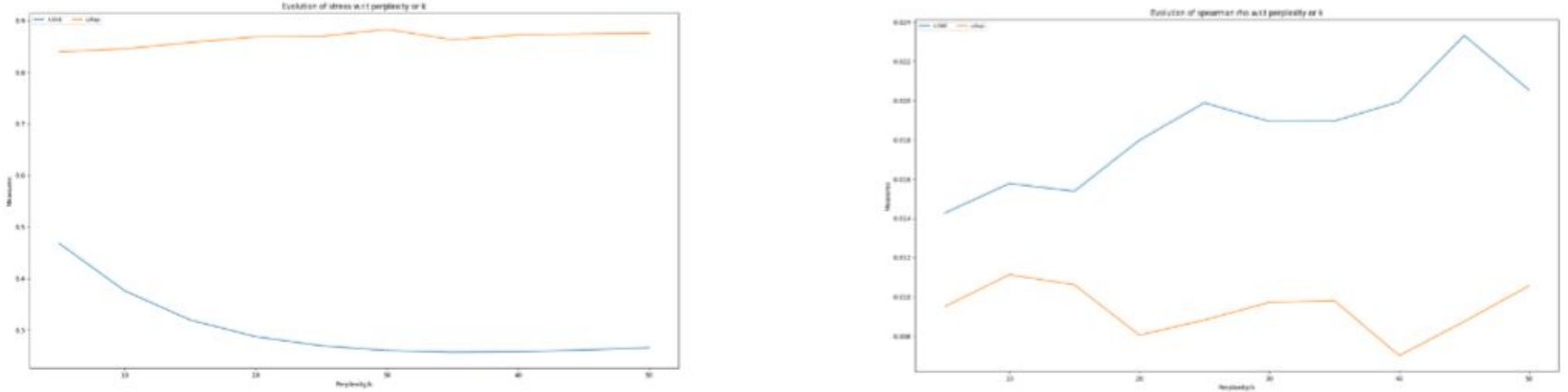


Figure 34: Stress and Spearman's Rho - Swiss Roll

Experiment 3

Results

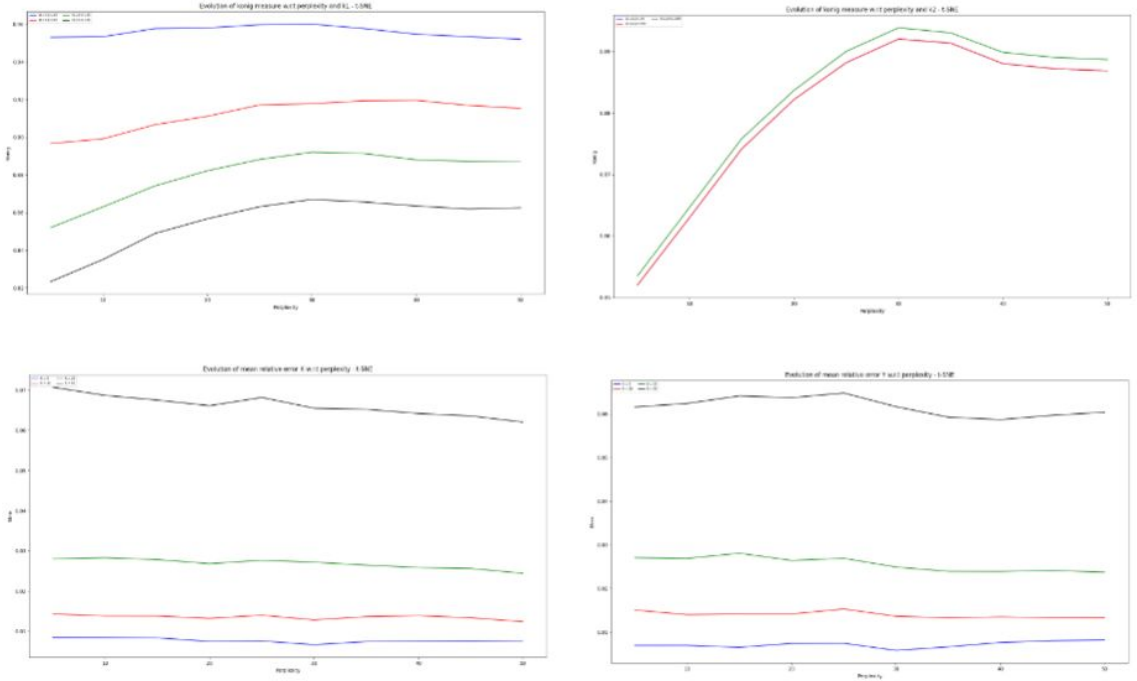


Figure 37: Konig, Mrre - Swiss Roll - t-SNE

Experiment 3

Results

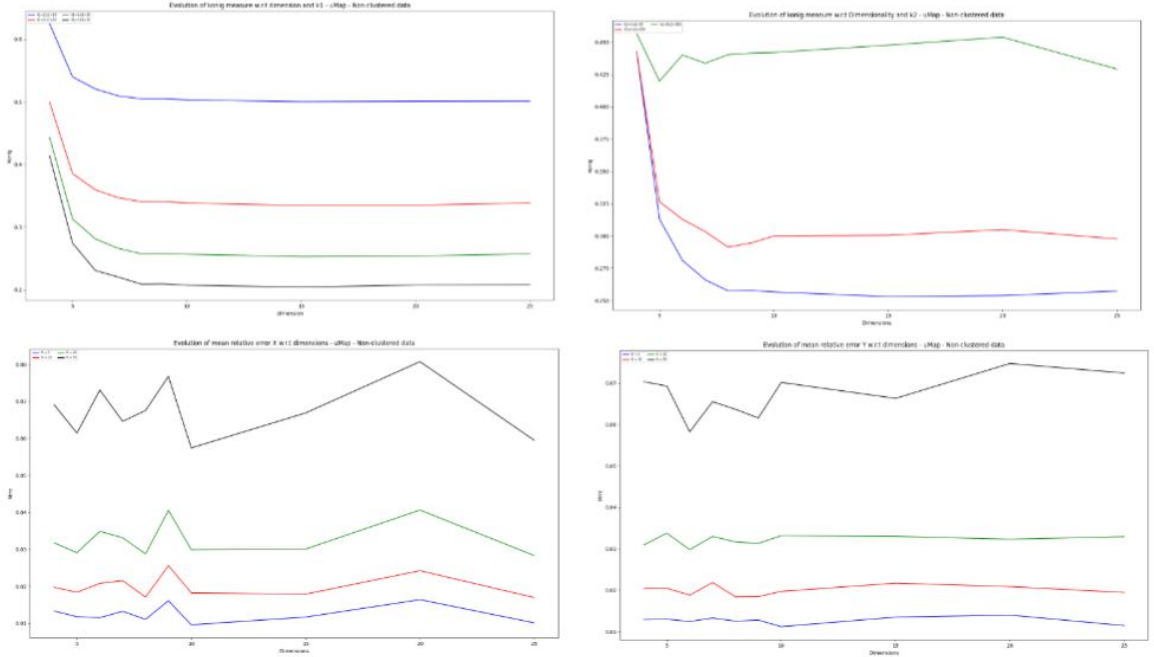


Figure 42: Konig and Mrre - Non Cluster - uMap

Conclusion

- uMap global preservation
- t-SNE local preservation
- Light impact of hyperparameters perplexity/k on local and global metrics
- Robustness to noise and initialization methods
- Main loss on first dimensions
- Good rank preservation in the close neighborhoods from both methods
- Normalize distances
- Deepen research on uMap: min_dist parameter and different distance metric

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département d'informatique

Thank you for your attention