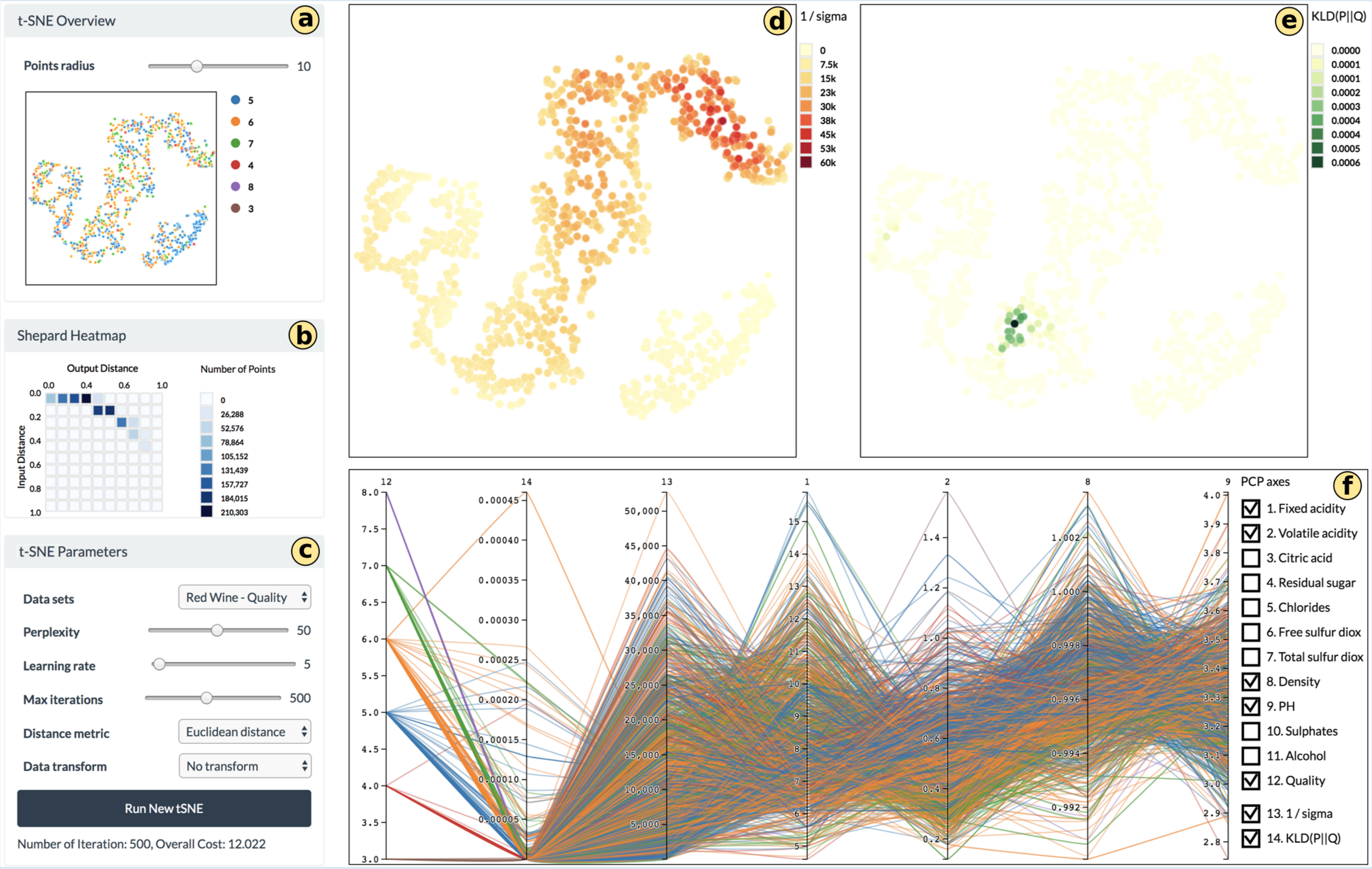
t-viSNE: A Visual Inspector for the Exploration of t-SNE

Angelos Chatzimparmpas*
Linnaeus University, Sweden

Rafael M. Martins
Linnaeus University, Sweden

Andreas Kerren
Linnaeus University, Sweden

The use of t-Distributed Stochastic Neighborhood Embedding (t-SNE) for the visualization of multidimensional data has proven to be a popular approach, with applications published in a wide range of domains. Despite their usefulness, t-SNE plots can sometimes be hard to interpret or even misleading, which hurts the trustworthiness of the results. By opening the black box of the algorithm and showing insights into its behavior through visualization, we may learn how to use it in a more effective way. In this work, we present t-viSNE, a visual inspection tool that enables users to explore anomalies and assess the quality of t-SNE results by bringing forward aspects of the algorithm that would normally be lost after the dimensionality reduction process is finished.



t-SNE Overview

Points radius 10

t-SNE Parameters

Data sets: Red Wine - Quality

Perplexity: 50

Learning rate: 5

Max iterations: 500

Distance metric: Euclidean distance

Data transform: No transform

Run New tSNE

Number of Iteration: 500, Overall Cost: 12.022

PCP axes

- ☑ 1. Fixed acidity
- ☑ 2. Volatile acidity
- □ 3. Citric acid
- □ 4. Residual sugar
- □ 5. Chlorides
- □ 6. Free sulfur diox
- □ 7. Total sulfur diox
- ☑ 8. Density
- ☑ 9. PH
- □ 10. Sulphates
- □ 11. Alcohol
- ☑ 12. Quality
- ☑ 13. 1 / sigma
- ☑ 14. KLD(P||Q)

Inspection of t-SNE results with our system, t-viSNE: (a) overview of the results with data-specific labels encoded with categorical colors; (b) the Shepard Heatmap of all pairwise distances; (c) t-SNE parameters and input data; (d) scatterplot showing density of neighborhoods in the original high-dimensional space; (e) scatterplot showing the final cost (Kullback-Leibler Divergence) of each point; (f) interactive parallel coordinates plot (PCP) of data features, density of neighborhoods, and cost for every point.

Research Question

How can we take advantage of the hidden internal workings of t-SNE which, when visualized, may provide important insights about the characteristics of the multidimensional data set?

Remaining Open Challenges

- Taking advantage of users' previous knowledge and adapting to each user type.
- Explore and interpret even more machine learning models.
- Steering the analysis by human-computer interaction.
- Evaluation of the systems.

We illustrate our tool with a data set of red wine samples from the north of Portugal:

1. From the overview, we can observe that the labels are not well-separated by the t-SNE layout and are mostly randomly distributed throughout the plot.
2. There is apparently no drastic change in density anywhere in the overview. However, the Sigma Plot shows that there is a gradient of increasing density that follows the layout from left to right.
3. The remaining KLD values are very low in most of the plot, except for a hot spot in the middle. Comparing the hot spot to the overview, there is no apparent correlation with the label distribution.

*Contact: angelos.chatzimparmpas@lnu.se

Linnæus University

IEEE VIS '18, 21–26 October 2018, Berlin, Germany

ISOVIS http://cs.lnu.se/isovis/