Mathematical Statistics
Stockholm University
Master Thesis **2020:12**
http://www.math.su.se

A Survey of Stochastic Neighbor Embedding for Dimension Reduction and Data Visualization

Tobias Wängberg*

September 2020

Abstract

During recent years machine learning and its applications to data science have opened many new opportunities in science, owing to increased computing power and higher availability of data, which has transformed fields such as biology, economics and social science. Along the many great opportunities that data driven science encompasses, new challenges emerge in how to correctly interpret and extract relevant information from complex data. One such challenge is that data today is often high dimensional, where, for example, a cell may be characterised by expression of hundreds of genes or a high resolution image by hundreds of pixels. This makes interpretation of the data difficult, and the need to reduce the dimension without losing critical information becomes an important task. In particular, the t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm has become a popular tool for visualising high dimensional data during recent years due to its capability of creating compelling 2D visualisations. Its inner mathematical workings are however poorly understood, and it can therefore be difficult to interpret the result of the dimension reduction using this algorithm. To this end, this thesis will explore the statistical properties of t-SNE, highlight its possible artifacts and benchmark it with test cases to illustrate its strengths and weaknesses.

Keywords: unsupervised learning, data visualisation, dimension reduction, t-SNE, evaluation, benchmarking

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: tobias@math.su.se. Supervisor: Chun-Biu Li.