Understanding Wasserstein t-SNE

University of Tübingen

Fynn Bachmann

24.09.2021

Contents

Chapter 1

Introduction: Visualizing Structure in Datasets

In the modern world data is collected on a daily basis
GB per day
leading to lots of interesting datasets.
Getting overview of the dataset first thing are visualization techniques.
However when trying to visualize structure

1.1 Examples and Motivation

Recent improvement in data visualization have led to methods like tSNE ... Thus a complex multidimensional dataset can be embedded onto the 2D plane and give away structure as in Figure
BAWU

1.1.1 Embedding Information

We can also encode other external information into an embedding, for example the average income leading tio Figure 2
BAWU

1.2 Hierarchical Data

In this thesis efforts are done to anaylse hierarchical data, that is . ...

We want to cluster or visualize the higher level using samples of the probability distribution as lower level

1.2.1 Loosing Information by taking the mean

In a later used dataset we find that correlation plays an interesting role as well, thus we lose information by collapsing the whole distribution into the mean.

Chapter 2

Theory: Methods and Distances

We have seen in t he introduction that it can be very valueable to visualize structure of datasets in as low dimensional space. However there exist a mutlitude of algorithms that put emphasis on different aspects. I will briefly give an overvierw to the

2.1 Dimension Reduction

The most basic algorithm has been known since ...

2.1.1 PCA

Principal Component Analzysis finds the axes of largest variation and projects the whole dataset onto these vectors.

2.1.2 t-SNE

A non-linear alternative in Dimension Reduction is called t Stochastic Neighbor Embedding. We define a distance measure such as Euclidean distance in the high dimensional space and t
RThe points are then moved along the gradient until it results in a local maximum where no point can be slighlty moved without resulting in a worse embeding than before
When interpreting t-SNE embedding is is important to keep in mind that the choice of distance measure puts emphasis on close points. Points the

are embedded far from each other don't need to be far away in the high dimensional dtaset.
Initilazation

2.2 Wasserstein Distance

We have learned that for a meaningful embedding it is necessary to have a distance measure in the high diomensional space. When dealing with hierarchical data this becomes non trivial as there is no default distacne measure for probability distribution. A standard way is to collapse the dsitrbition into the mean and then use the Euclidian distance of the means. But one can easily imagine that this technique can loose arbitryily much of the information. A standard metric in computer Science is called the Wasserstein metric and has been widely used to compare distibutions. Its downside is the complex compuation durance. However, for small datasets we have been able to counter this problem and computed Wasserstein distances. I will thus give a brief overview of the theory and then explain the computation.

2.2.1 Exact Formulation

The Wasserstein metrix is formally edfined by

2.2.2 Special Case: Gaussians

For multivariate normals there exists a closed form solution of the 2-Wasserstein metric. it reads
formalu
and the derivation can be found in Frechet distance blabla

2.2.3 Convex Interpolation Method

We first nore, that the first part of the sum is just the Euclidiean distance of the mean of the distribution. The Wasserstein distance can therfore be seen as an extension of the Euclidiean distance. As Landau etc showed, the second summand in euqation 1 is a proper meric on covarainces. We can therefore combine these two distances in any way other than

This leads to the convex generalization of the Wasserstein distance which ziels both other distances for $\lambda = 0$ or $\lambda = 1$ respectively.

2.3 Linear Programming

We have earlier stated that the Wasserstein distance is hard to compute for continuous distributions. However for discrete distributions it boils down to the linear program described in Equaton
LINEAR PROGRAMM

2.3.1 Scalability in Participants

If the metric space on which the distributions are defined is small, a linear programm can solve the wasserstein distance with unique optimal solution.

2.3.2 Scalability in Features

When dealing with larger spaces the linear program grows exponentially in the number of dimensions. A discrete space with 10 values yields a linear programm with 100 variables, n features yiled a linear programm with 10^{2n} variables. However, since every variable must be non negative and every row/column must sum to the respetive marginal distribution, each zero in a marginal distribution forces the whole colum/row to contain zeros as well. Thus we can exclude these variable from the ;linear programm and eventually arrive at an upper bound of $N \cdot M$ variables for marginal distributions with N and M samples respectively. This approach is indifferent to the number of features, since we can compute the pairwise Euclidean distance of the samples efficiently using vectorization.

Chapter 3

Analysis: Is there structure in Covariance?

We have seen in the Introduction that it can be useful to include difference in correlation into the distance of two distributions. In this section I will show a Proof of Concept and then apply the method to real-world data to show that in practice new insights are won by the inclusion of covariance into the analysis. While I first create synthetic data with a Hierarchical gaussian Mixture Model to show a proof of conccetp, I will later analyse the Germen Federal Election 2017 as well as the European Values Study.

3.1 Synthetic Data

I will define in the following what I call the Hierarchical Gaussian Mixture Model as a tool to understand Wasserstein t-SNE on snynthetic data before we apply the method to real world data. Since t-SNE is a cloustering method le's first define the clusters in our model, which the we aim to visualize with the method without providing the cluster membership of each datapoint.

3.1.1 Hierarchical Gaussian Mixture

A Hierarchical Gaussian Mixture is defined by K classes, from which we have N_k datapoints each. Each class is defined by a meana and Covariance matrix ...

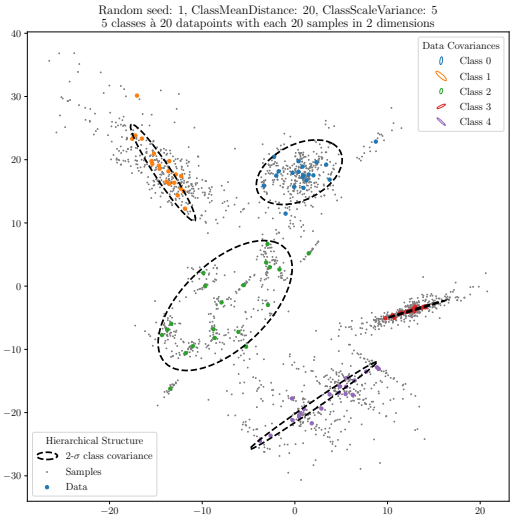

Figure 3.1: Example of a HGM

3.1.2 Proof of Concept

It is easy to image an example in which the information about the means is not enough to cluster the data, while a Wasserstein t-SNE approach succesgully does. In Figure **??** we can see a HGM that consists of 4 classes, two of which share the mean while two of them share the Scale matrix. The Euclidean embedding expectely doesn't cpature the structuire of the dataset, at the same time the sole information about Coavrarince isn't enough neither. The convex combination of both, the Wasserstein embedding, hoewver sepreates all 4 classes from ach other and can there be considered superioir to the other in this setting.

3.1.3 Distance of Covariances

In section 2 I introduced the interpolation of Covariance and Mean distances. However it is not intuitive how much contribution each of the two different aspect of distance have, inparticular, how much impact a different Covarince has to the Wasserstein distance. In the follwoing experiment I sampled $N = 100$ Covariance matrices from a Wishart distribution and computed the

Figure 3.2: Proof of Concept

distance to a reference Covariance Σ with the above described formula.
We can see that

3.2 German Election 2017/2021

The Federal German Election is divided into 299 voting districts, each of which consists of roughly 150-850 poll stations (voting by mail excluded). In the following analysis we shall find that certain parties correlate differently within these voting districts, i.e. that
For simplification we exclude any party from the nalysis that hasn't received a minimum share of 5 per cent of the total votes. Otherwise the correlation would be manipulated by parties that only get elected in a specific state.

3.2.1 Embeddings

The plain Euclidean t-SNE embedding as well as the Pure Covariance embedding of the GER dataset are givven in Figure xy. In the middle we see the interlpolation of $\lambda = 0.75$. Interestingly the middle one has 4 clusters wheras the outer ones only admit 3 clusters.

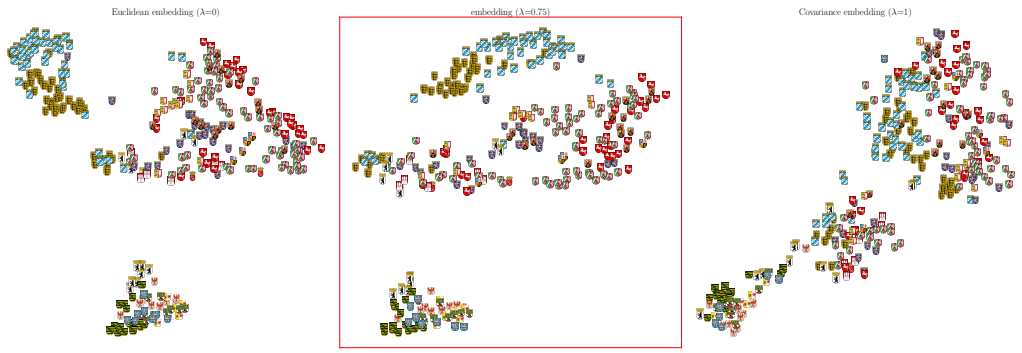

Figure 3.3: GER embedding

Gaussian Wasserstein

If we look at the legend in the appendix we see that the clusters share certain meta inforamtion: university cities, easetrn germany, southern Germany, western germany.

Exact Wasserstein

3.2.2 Covariance

It is interesting to analyse if the structure in covariance is due to different correlations or mainly due to variance of each individual feature. In Figure xy we differentiate these two causes and observe that both factores play a role here.

3.2.3 Correlation

3.3 European Value Study 2017-2020

3.3.1 Embeddings

3.3.2 Discrete Data

3.3.3 Comparison to Exact Wasserstein Embedding

Chapter 4

Outlook and Discussion

4.1 Complexity Analysis

4.1.1 Improving the runtime of Exact Wasserstein

4.2 Finding more Use-Cases

4.2.1 Medical Data