Understanding Wasserstein t-SNE

University of Tübingen

Fynn Bachmann

24.09.2021

Contents

1 Introduction: Visualizing Structure in Datasets **3**
- 1.1 Examples and Motivation . . . 3
  - 1.1.1 Why visualize? . . . 3
- 1.2 Hierarchical Data . . . 3
  - 1.2.1 Loosing Information by taking the mean . . . 3

2 Theory: Methods and Distances **4**
- 2.1 Methods . . . 4
  - 2.1.1 PCA . . . 4
  - 2.1.2 t-SNE . . . 4
- 2.2 Wasserstein Distance . . . 4
  - 2.2.1 Exact Formulation . . . 4
  - 2.2.2 Special Case: Gaussians . . . 4
  - 2.2.3 Convex Interpolation Method . . . 4
- 2.3 Linear Programming . . . 4
  - 2.3.1 Scalability in Participants . . . 4
  - 2.3.2 Scalability in Features . . . 4

3 Analysis: Is there structure in Covariance? **5**
- 3.1 Synthetic Data . . . 6
  - 3.1.1 Hierarchical Gaussian Mixture . . . 6
  - 3.1.2 Proof of Concept . . . 6
- 3.2 German Election 2017/2021 . . . 6
  - 3.2.1 Finer Structure . . . 6
  - 3.2.2 Covariance Embeddings . . . 6
  - 3.2.3 Results . . . 6
- 3.3 European Value Study 2017-2020 . . . 6
  - 3.3.1 Embeddings . . . 6

3.3.2 Discrete Data 6
3.3.3 Comparison to Exact Wasserstein Embedding 6
3.4 Big5 Personality Traits Survey 6
3.4.1 Results 6

4 Outlook and Discussion **7**
4.1 Complexity Analysis 7
4.1.1 Improving the runtime of Exact Wasserstein 7
4.2 Finding more Use-Cases 7
4.2.1 Medical Data 7

Chapter 1

Introduction: Visualizing Structure in Datasets

1.1 Examples and Motivation

1.1.1 Why visualize?

1.2 Hierarchical Data

1.2.1 Loosing Information by taking the mean

Chapter 2

Theory: Methods and Distances

2.1 Methods

2.1.1 PCA

2.1.2 t-SNE

2.2 Wasserstein Distance

2.2.1 Exact Formulation

2.2.2 Special Case: Gaussians

2.2.3 Convex Interpolation Method

2.3 Linear Programming

2.3.1 Scalability in Participants

2.3.2 Scalability in Features

Chapter 3

Analysis: Is there structure in Covariance?

3.1 Synthetic Data

3.1.1 Hierarchical Gaussian Mixture

3.1.2 Proof of Concept

3.2 German Election 2017/2021

3.2.1 Finer Structure

3.2.2 Covariance Embeddings

3.2.3 Results

3.3 European Value Study 2017-2020

3.3.1 Embeddings

3.3.2 Discrete Data

3.3.3 Comparison to Exact Wasserstein Embedding

3.4 Big5 Personality Traits Survey

3.4.1 Results

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