Information Visualization

High-Dimensional Data

Dr. David Koop





Schedule

- Today: High-Dimensional Data Lecture
- Tuesday, Oct. 26: No Class
- Thursday, Oct. 28: High-Dimensional Data Critique Due





What techniques might you use for high-dimensional data visualization?







High-Dimensional Data Visualization Techniques

- Scatterplot Matrix (SPLOM)
- Parallel Coordinates Plot (PCP)
- Heatmap
- Interactive Elements:
 - Brushing (Linked Highlighting)
 - Tooltips





Scatterplot Matrix

• Each pair of quantitative attributes has its own plot



D. Koop, CSCI 627/490, Fall 2020









Parallel Coordinate Plots

- Use multiple parallel axes, one for each dimension
- Each data item is encoded as a line mark
- Positive and negative correlation can be seen in these plots...
- ... but ordering becomes important

D. Koop, CSCI 627/490, Fall 2020





[Munzner (ill. Maguire), 2014]



Northern Illinois University



Brushing



D. Koop, CSCI 627/490, Fall 2020









Our High-Dimensional Data Focus

- Projection Understanding
- Tours







Dimensionality Reduction

- individual attribute
- Example: Understanding the language in a collection of books
 - Count the occurrence of each non-common word in each book
 - (e.g. "western")
 - Don't want to have to manually determine such rules
- techniques

D. Koop, CSCI 628, Fall 2021

• Attribute Aggregation: Use fewer attributes (dimensions) to represent items • Combine attributes in a way that is more instructive than examining each

- Huge set of features (attributes), want to represent each with an aggregate feature (e.g. high use of "cowboy", lower use of "city") that allows clustering

Techniques: Principle Component Analysis, Multidimensional Scaling family of











Principle Component Analysis (PCA)

original data space



D. Koop, CSCI 628, Fall 2021



PC 1









PCA



[Principle Component Analysis Explained, Explained Visually, V. Powell & L. Lehe, 2015]







17 dimensions to 2

Alcoholic drinks Beverages Carcase meat Cereals Cheese Confectionery Fats and oils Fish Fresh fruit Fresh potatoes Fresh Veg Other meat Other Veg Processed potatoes Processed Veg Soft drinks Sugars

England	N Ireland	Scotland	
375	135	458	
57	47	53	
245	267	242	
1472	1494	1462	
105	66	103	
54	41	62	
193	209	184	
147	93	122	
<mark>1</mark> 102	674	957	
720	1033	566	
253	143	171	
685	586	750	
488	355	418	
198	187	220	
360	334	337	
<mark>137</mark> 4	1506	1572	
156	139	147	

Wales

England

Scotland

D. Koop, CSCI 628, Fall 2021



[Principle Component Analysis Explained, Explained Visually, V. Powell & L. Lehe, 2015]





Northern Illinois University



12

Non-linear Dimensionality Reduction



original data space \mathcal{X}

D. Koop, CSCI 628, Fall 2021

component space Z

Dimensionality Reduction in Visualization

D. Koop, CSCI 628, Fall 2021

[Glimmer, Ingram et al., 2009]

Northern Illinois University

Tasks in Understanding High-Dim. Data

Probing Projections

Probing Projection Goals

- Examining the Projection
- Exploring the Data
- Design Goals:
 - Show and correct approximation errors
 - Allow for multi-level comparisons
 - Spatial orientation
 - Consistent design
- Allow grouping of samples
 - Selections
 - Classes
 - Clusters

Tooltips with statistics

Austria • United States

United Kingdom

Israel

Luxembourg

Portugal

- Educational attainment $35 -2.4 \sigma_{\text{Slove}}$
- Employees working ve... 9.31 -0.034σ
 - Life expectancy 80.8 +0.39 σ
 - Life satisfaction 5.2
 - Self-reported health 50
- Student skills 488 -0.20 σ
- Time devoted to leisur... 14.95 +0.13 σ
 - Years in education 17.8 $+0.31 \sigma$

correct distances

D. Koop, CSCI 628, Fall 2021

18

Comparing Two Groups

South America 3 samples Northern Europe 9 samples

Educational attainment 50 77 Employees working ve... 18 6.2 Life expectancy 75 81 Life satisfaction 7.1 7.4 Self-reported health 65 77 Student skills 420 500 Time devoted to leisur... 14 15 Years in education 16 19

D. Koop, CSCI 628, Fall 2021

y 19

Heatmap from Dimension Hover

	PROJECTION Edit projection Display dendrogram Display errors 🗸 Display labels								
	SELECTIONS								
	+								
	new selection select samples								
	CLUSTERING 💿								
	Clusters: 5 clusters								
			(w)		$\langle v_{i}\rangle$	74			
	Cluster 1 10 sample	1 es	Cluster 2 9 samples	Cluster 3 4 samples	Cluster 4 4 samples	Cluster 5 9 samples			
	DIMENSIONS								
ssian Federati	Educational attainment				32	94			
	2.4	Employees working very long hours			0.17	43.29			
		Life expectancy			69	82.8			
		Life satisfaction Self-reported health			4.7	7.8			
					30	90			
	Student skills Time devoted to leisure and personal care				402	542			
				13.42	16.06				
	1	Years in education			14.1	19.7			

Showing Error via Sample-centric Halos

Showing Projection Errors

White: higher levels of similarity Gray: lower levels of similarity

User Study & Results

- Types of Questions:
 - How would you try to characterize the type X?
 - In what way are X and Y different in their properties?
 - Are the projections of X and Y correct or do they deviate? How do you interpret this?
 - Can you discover which parts of the cluster combinations are A, B, and C?
- Discussion:
 - Learnability: need more effective mechanisms for grasping the concepts behind dimensionality reduction
 - Manipulation: What happens with results? _
 - Large data: What about text corpora?

Different Projections

D. Koop, CSCI 628, Fall 2021

Northern Illinois University 24

Leads to Different Conclusions

D. Koop, CSCI 628, Fall 2021

[D. Cook et al., 2008]

Tours help explore projections

D. Koop, CSCI 628, Fall 2021

(Non–)linear association

Projection 1

Northern Illinois University 26

<u>Going beyond 2D and 3D to visualise higher</u> <u>dimensions, for ordination, clustering & other models</u>

D. Cook

Toward Comparing DNNs with UMAP Tour

M. Li and C. Scheidegger

