Principal Component Analysis

EDP 619 Week 10

Dr. Abhik Roy

Survey Research Methods

Welcome!

There are a lot of things going on behind the scenes when using PCAs and this is just a very brief introduction without any audio. I have tried to minimize the jargon and complexity, though some items may not be as clear as others. If you have questions, please feel free to reach out.

Additionally you may notice the following icons in the footnotes. These contain links to external sites that provide extra materials that may be of interest to you.

HTML — Extra Content

PDF — Paper Material

Rmd — Guided Example

R Script

Online Video

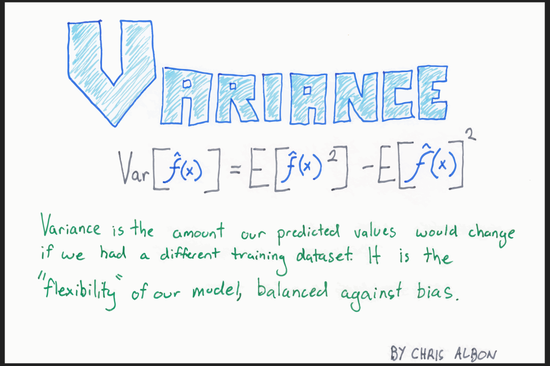
Prerequisites

This slideshow assumes that you have a basic understanding of variance and correlations. For a refresher, please take a look at both reviews below

Prerequisites

This slideshow assumes that you have a basic understanding of variance and correlations. For a refresher, please take a look at both reviews below

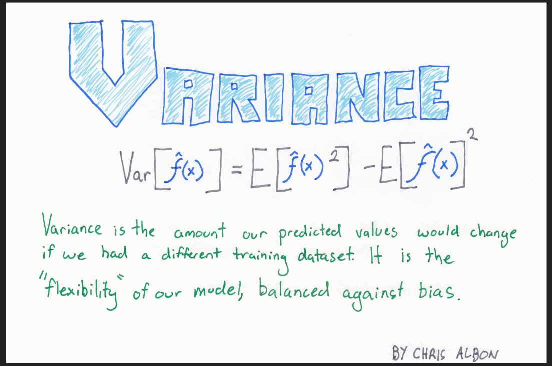
Variance is essentially a measure of the spread between points in a data set. Specifically it tells us how far each data point in a set is from the mean and by proxy from every other data point in that set.



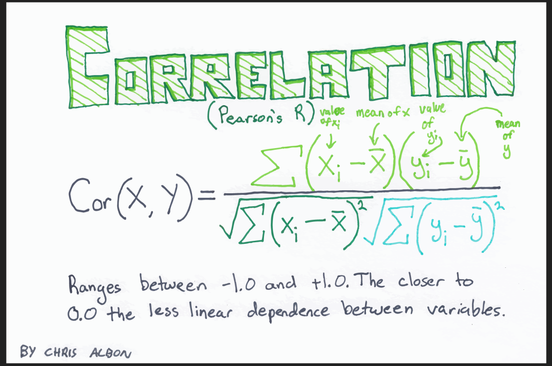
Prerequisites

This slideshow assumes that you have a basic understanding of variance and correlations. For a refresher, please take a look at both reviews below

Variance is essentially a measure of the spread between points in a data set. Specifically it tells us how far each data point in a set is from the mean and by proxy from every other data point in that set.



Correlation gives you an idea of the strength or weakness of the relationship between two variables. In a survey where each item is set to measure a single construct, these are essentially the applicable questions.



More Review

If you would like a deeper dive on either area, tale a look at the videos below

More Review

If you would like a deeper dive on either area, tale a look at the videos below

Variance

More Review

If you would like a deeper dive on either area, tale a look at the videos below

Variance

Correlation

From Basic to Better

From Basic to Better

correlations are great but they don't...

From Basic to Better

correlations are great but they don't...

1. tell you how every question is related to every other question
2. differentiate between relevant data and noise

From Basic to Better

correlations are great but they don't...

1. tell you how every question is related to every other question
2. differentiate between relevant data and noise

...enter a method called ***Principle Component Analysis***

Principle Component Analysis (PCA)

Steps in a Nutshell

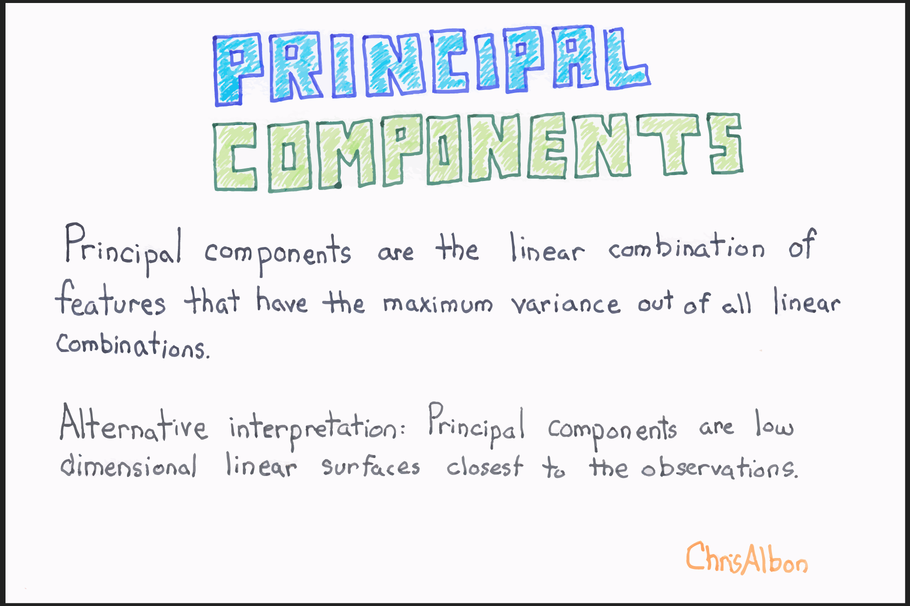
Steps in a Nutshell

The basic idea of a PCA can be broken into two steps

Steps in a Nutshell

The basic idea of a PCA can be broken into two steps

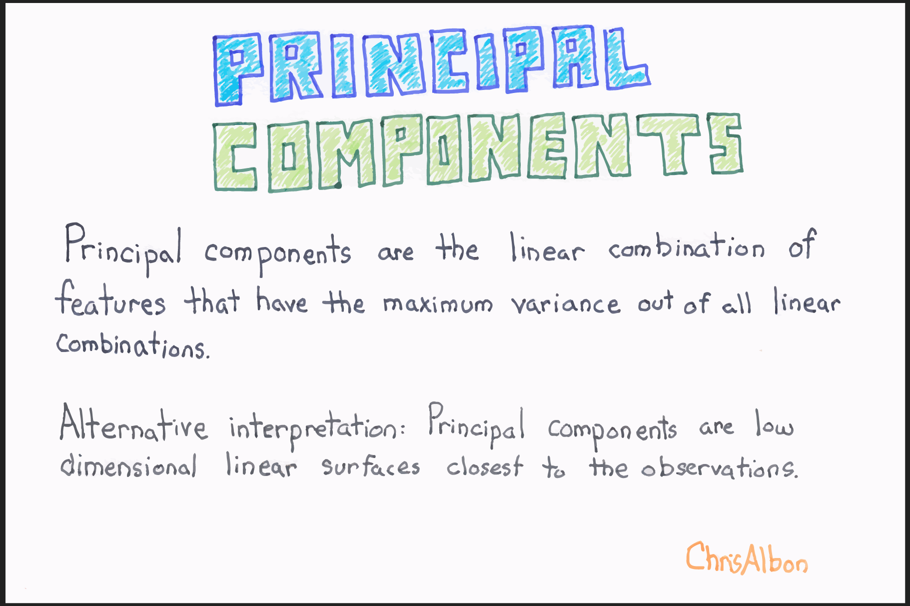
Locate the directions, or *components*, in a data set with high variance



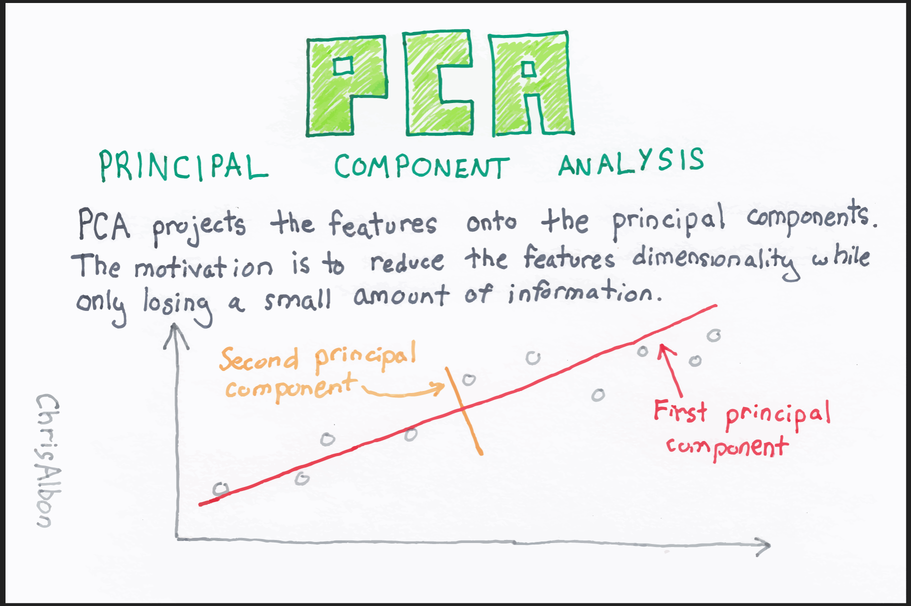
Steps in a Nutshell

The basic idea of a PCA can be broken into two steps

Locate the directions, or *components*, in a data set with high variance



Find a limited number of components with high variance that in aggregate can explain most of the overall variance in the data



Reducing Complexity

Reducing Complexity

First an overview of some terms:

Reducing Complexity

First an overview of some terms:

Dimensionality - The number of input variables, or *features* in a dataset. In a spreadsheet, you can think of these as the column names.

id	read	write	math	science	socst
70	57	52	41	47	57
121	68	59	53	63	61
86	44	33	54	58	31
141	63	44	47	53	56
172	47	52	57	53	61
113	44	52	51	63	61
50	50	59	42	53	61
11	34	46	45	39	36
84	63	57	54	58	51
48	57	55	52	50	51
75	60	46	51	53	61
60	57	65	51	63	61
95	73	60	71	61	71

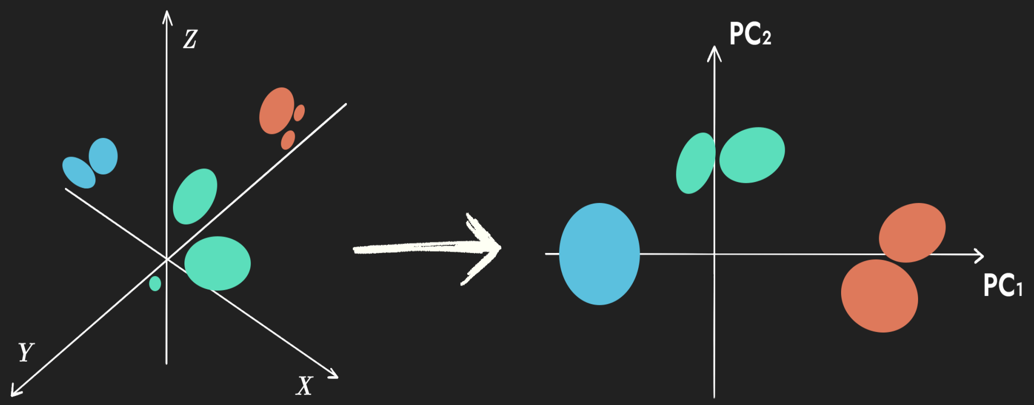
Reducing Complexity

First an overview of some terms:

Dimensionality - The number of input variables, or *features* in a dataset. In a spreadsheet, you can think of these as the column names.

id	read	write	math	science	socst
70	57	52	41	47	57
121	68	59	53	63	61
86	44	33	54	58	31
141	63	44	47	53	56
172	47	52	57	53	61
113	44	52	51	63	61
50	50	59	42	53	61
11	34	46	45	39	36
84	63	57	54	58	51
48	57	55	52	50	51
75	60	46	51	53	61
60	57	65	51	63	61
95	73	60	71	61	71

Dimensionality Reduction - Statistical techniques used to reduce the number of input variables.



The Problem with Dimensions

Curse of Dimensionality - In brief terms, this refers to a few aspects

The Problem with Dimensions

Curse of Dimensionality - In brief terms, this refers to a few aspects

- *statistical*. the error rate increases as the number of features increases

The Problem with Dimensions

Curse of Dimensionality - In brief terms, this refers to a few aspects

- *statistical*. the error rate increases as the number of features increases
- *computational*. algorithms are harder to design and and exponentially take more time to run in high dimensions

The Problem with Dimensions

Survey Research Methods

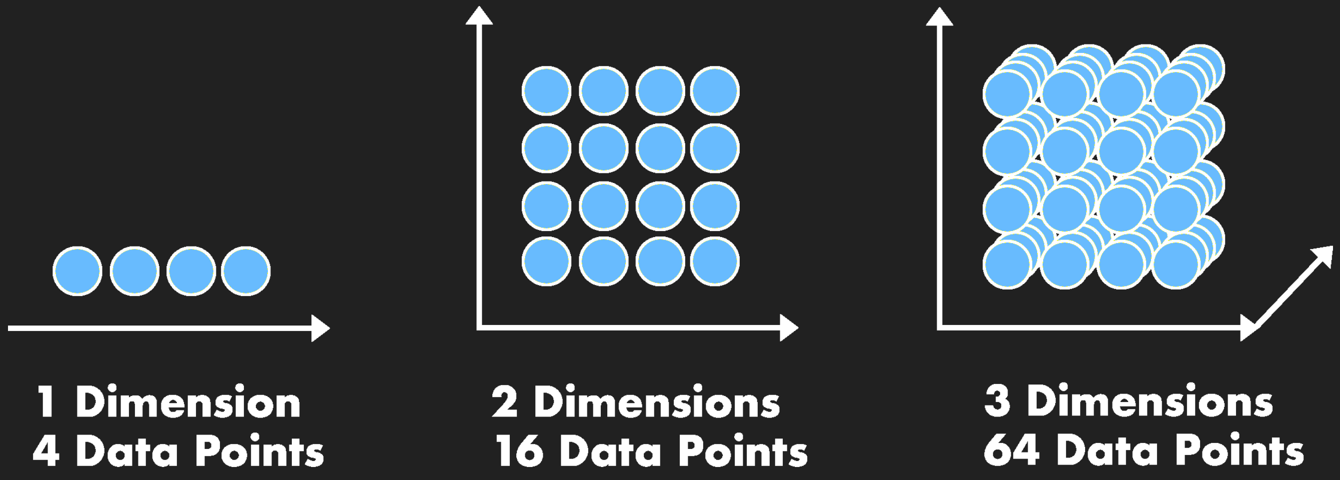
Curse of Dimensionality - In brief terms, this refers to a few aspects

- *statistical*. the error rate increases as the number of features increases
- *computational*. algorithms are harder to design and and exponentially take more time to run in high dimensions
- *practical*. higher number of dimensions theoretically allow more information to be stored, but in reality it rarely helps due to the higher possibility of noise and redundancy in real-world data

The Problem with Dimensions

Curse of Dimensionality - In brief terms, this refers to a few aspects

- *statistical*. the error rate increases as the number of features increases
- *computational*. algorithms are harder to design and and exponentially take more time to run in high dimensions
- *practical*. higher number of dimensions theoretically allow more information to be stored, but in reality it rarely helps due to the higher possibility of noise and redundancy in real-world data



Fundamentals of What PCAs Can and Cannot Do

PCAs are one of the most traditional methods used for dimension reduction.

Fundamentals of What PCAs Can and Cannot Do

PCAs are one of the most traditional methods used for dimension reduction.

Primary benefit

It transforms the data into the most informative space, thereby allowing the use of lesser dimensions which retain needed information from the data while shedding much of the noise

Fundamentals of What PCAs Can and Cannot Do

PCAs are one of the most traditional methods used for dimension reduction.

Primary benefit

It transforms the data into the most informative space, thereby allowing the use of lesser dimensions which retain needed information from the data while shedding much of the noise

Primary drawback

It assumes linearity so any nonlinear relationship in a given data set is lost possibly causing loss in accuracy and the ability to estimate the likelihood of causality.

Fundamentals of What PCAs Can and Cannot Do

PCAs are one of the most traditional methods used for dimension reduction.

Primary benefit

It transforms the data into the most informative space, thereby allowing the use of lesser dimensions which retain needed information from the data while shedding much of the noise

Primary drawback

It assumes linearity so any nonlinear relationship in a given data set is lost possibly causing loss in accuracy and the ability to estimate the likelihood of causality.

Note as with most other procedures: *what you gain in efficiency, you lose in precision*. In a nutshell, there is no known perfect method that can both get rid of all of the noise and leave only relevant information. However with an ever growing machine learning library of approaches, we could get pretty close well within your lifetime!

How Do PCAs Work?

Survey Research Methods

How Do PCAs Work?

Before moving on please note that this is a nutshell explanation of the steps and avoids the mathematics[1]. If you are interested in a more nuanced introduction coupled with the mathematics, watch this amazing lecture by Josh Starmer from StatQuest[2].

[1] PDF [2]

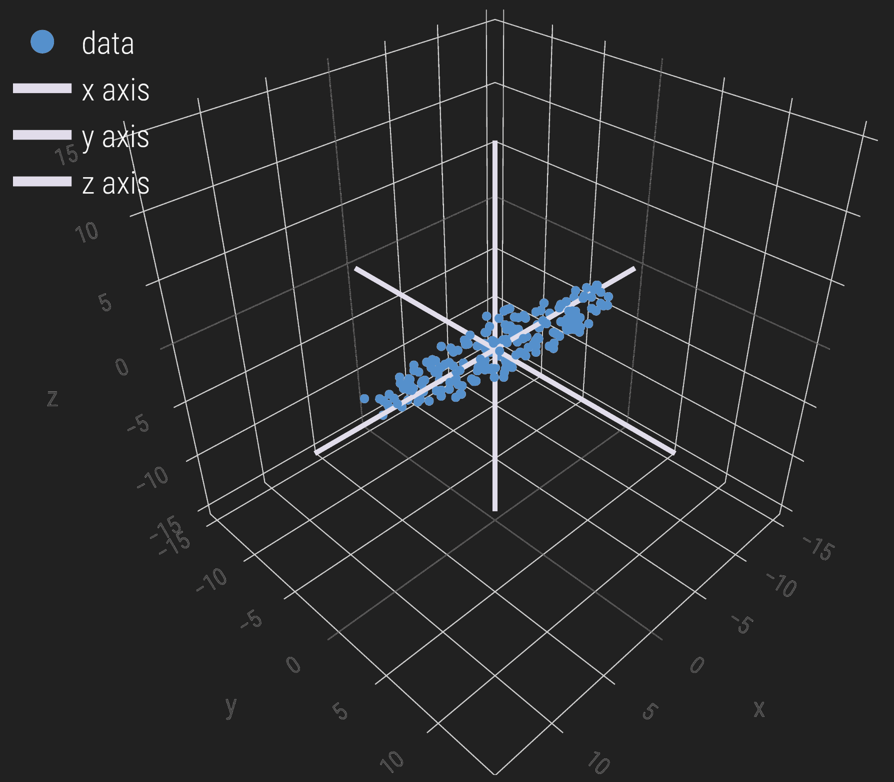
OK Now Really How Do PCAs Work?

OK Now Really How Do PCAs Work?

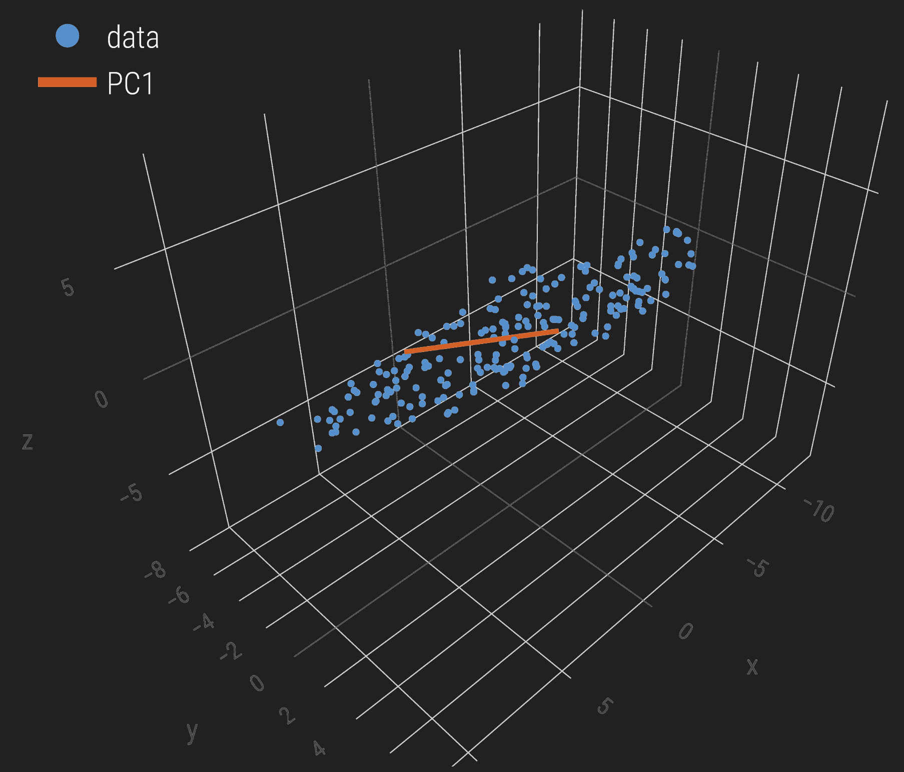
Let's look at a data set with 205 points randomly scattered in three-dimensions. Keep in mind that as you move along, the *PCA is carving out new dimensions which you will be able to see and interact with.*

When applying a PCA, it locates the...

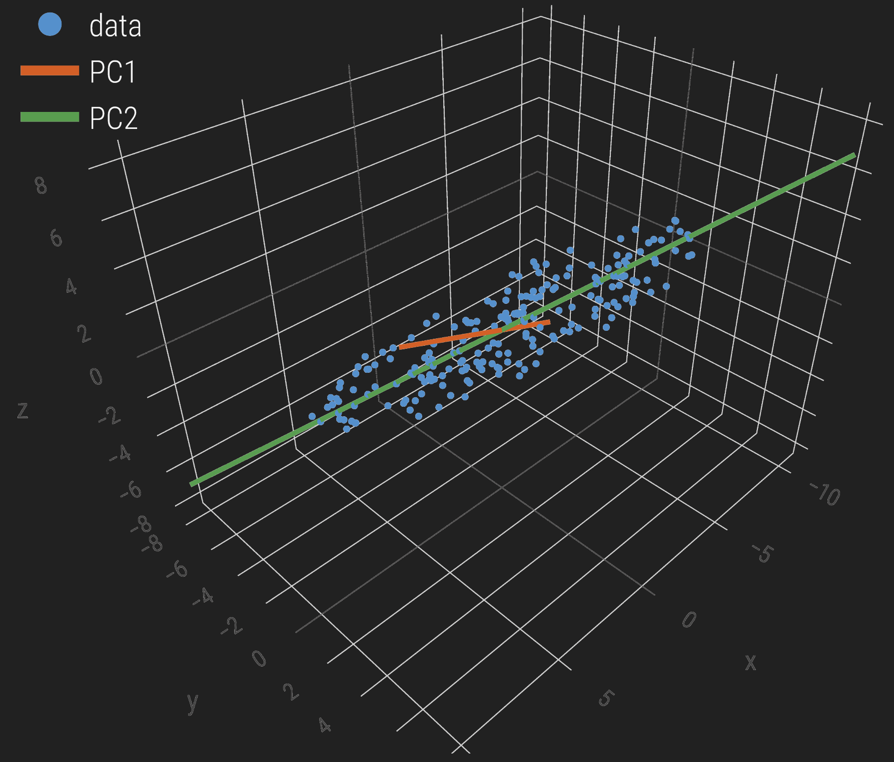
1. center point of data in multi-dimensional space



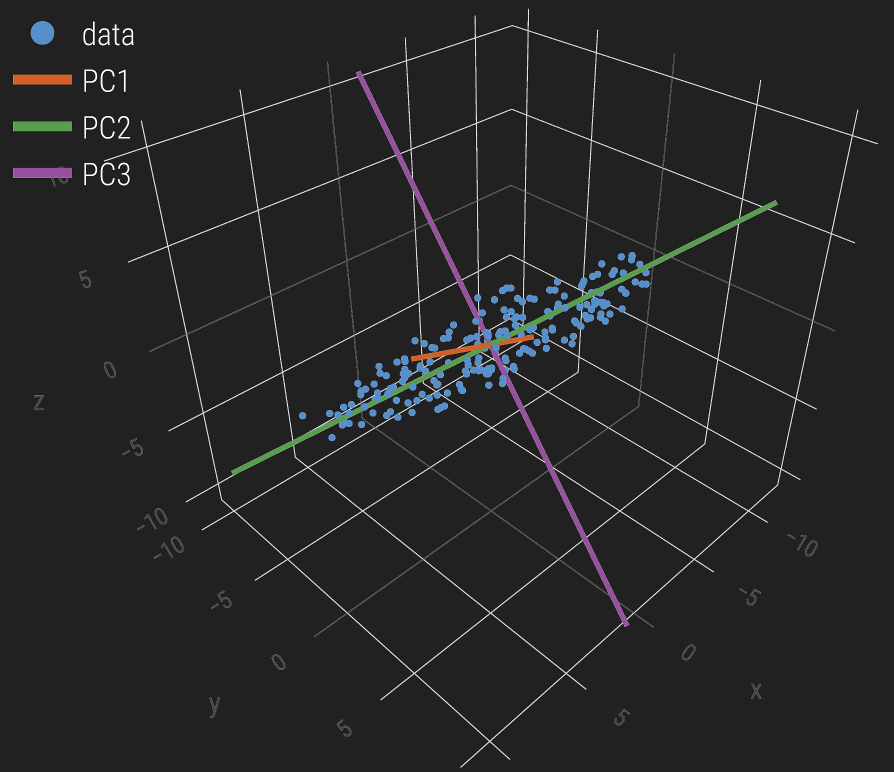
2. direction with the greatest variance. This is called the **1st component**



3. direction that is perpendicular, or *orthogonal* to the 1st component with the greatest variance. This is called the **2nd compone**



4. direction that is perpendicular, or *orthogonal* to the 1st and 2nd component with the greatest variance. This is called the **3rd component**.



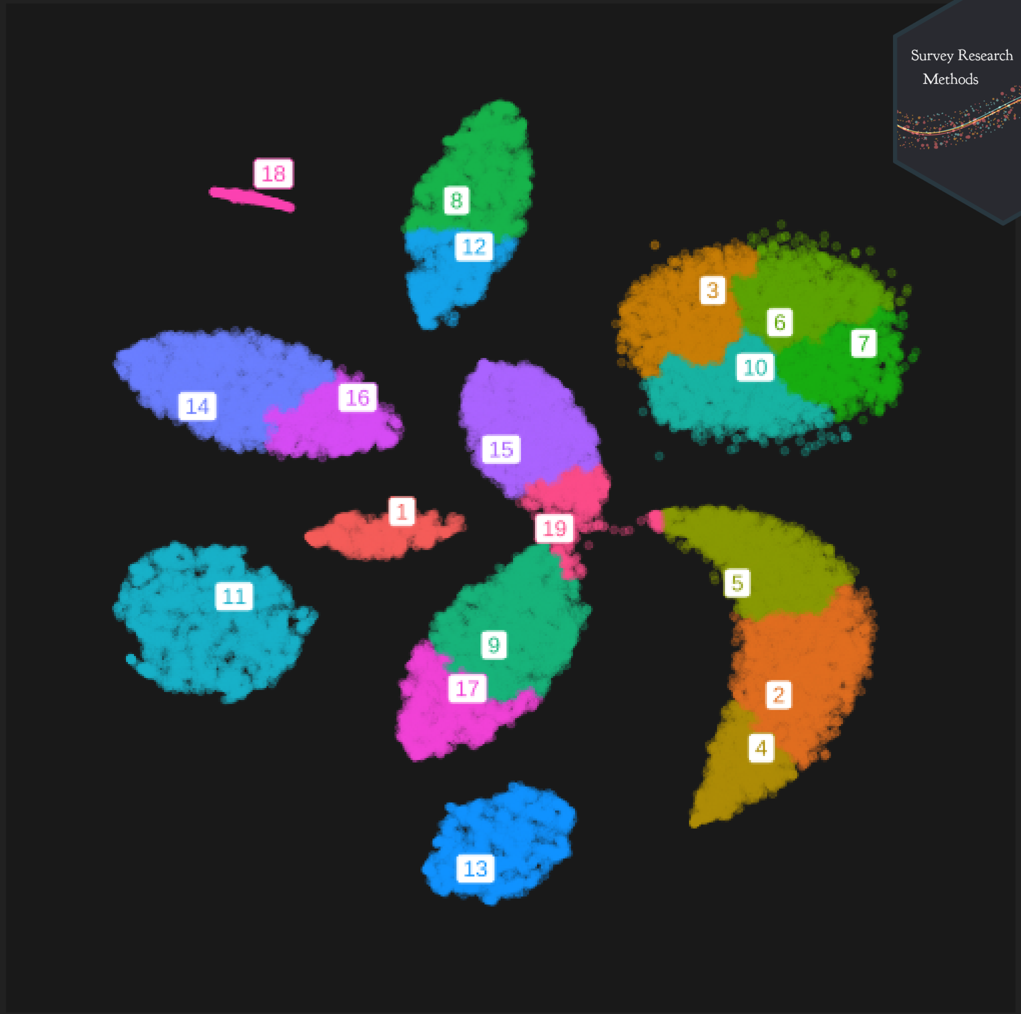
> and it keeps going like this for as many dimensions as we have in a data set...

> and it keeps going like this for as many dimensions as we have in a data set...
>
> so you can probably imagine that big data sets with hundreds or thousands of columns and rows can take quite a bit of time...

Survey Research Methods

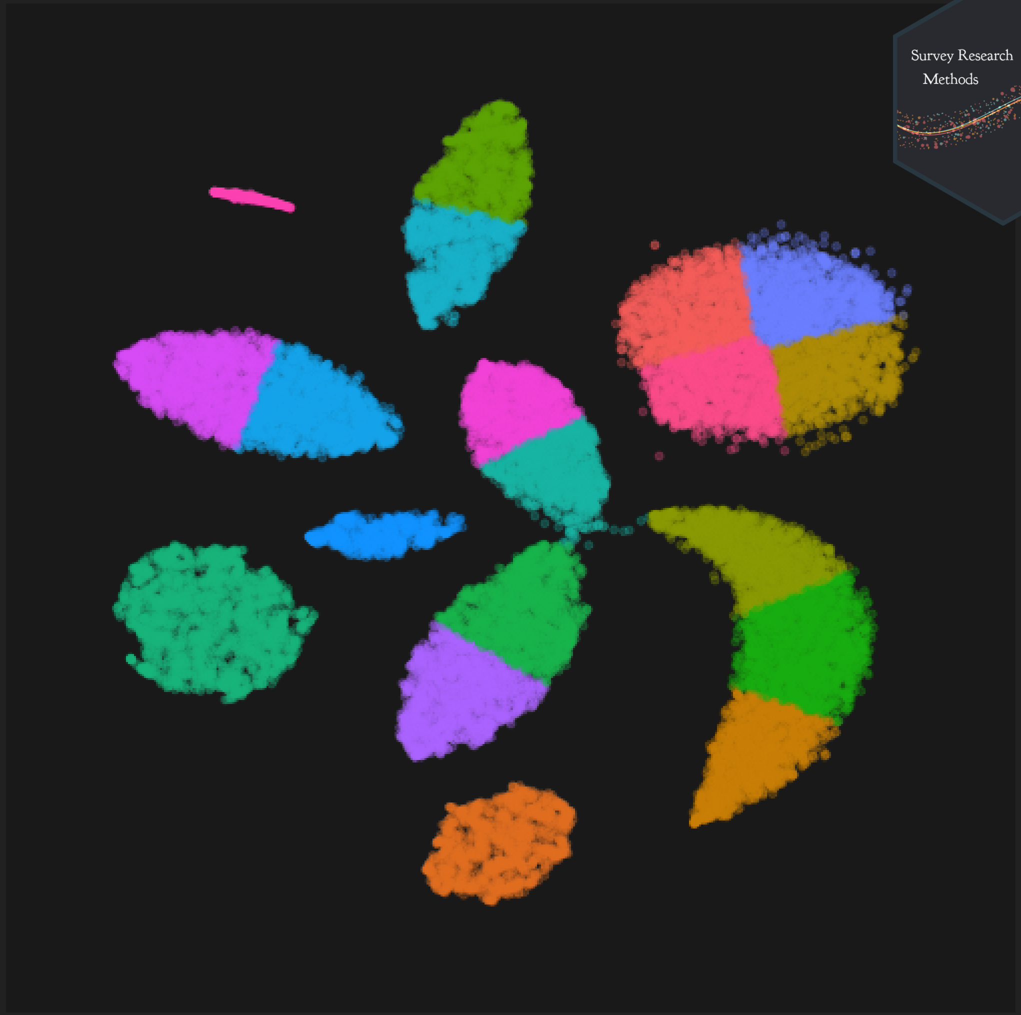
> and it keeps going like this for as many dimensions as we have in a data set...
>
> so you can probably imagine that big data sets with hundreds or thousands of columns and rows can take quite a bit of time...
>
> but there are many other methods of reducing dimensions like

Hierarchical Clustering[3]



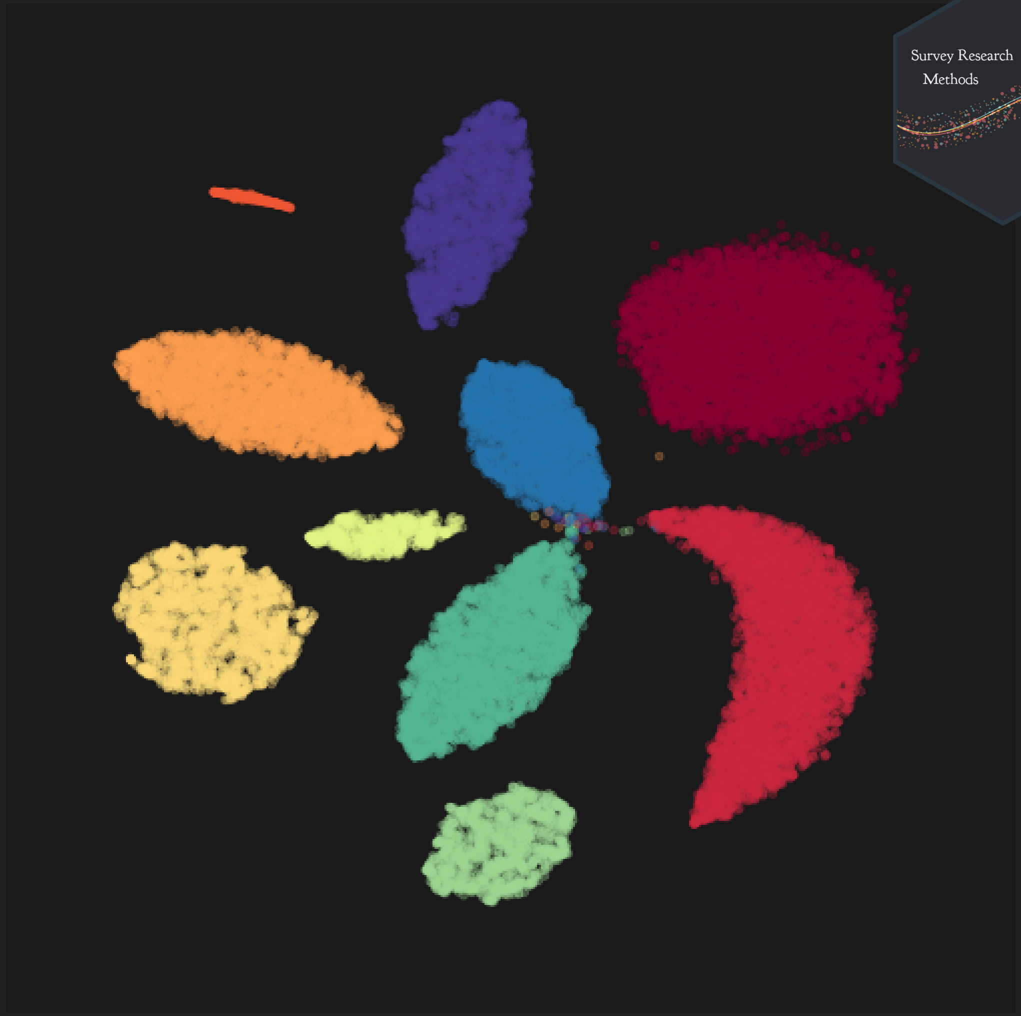
[3]

K-means Clustering[4]



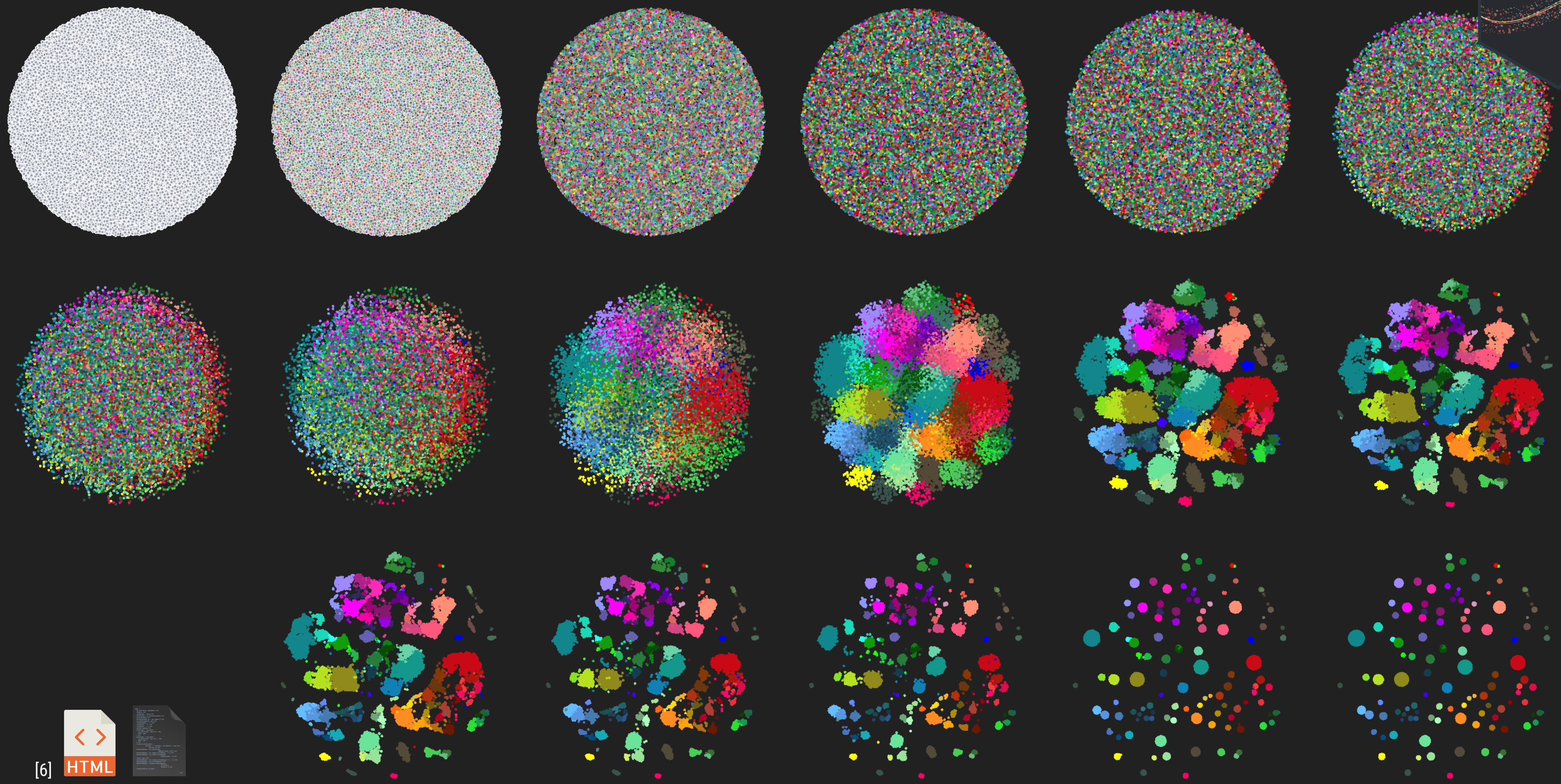
[4]

t-Distributed Stochastic Neighbor Embedding (t-SNE)[5]



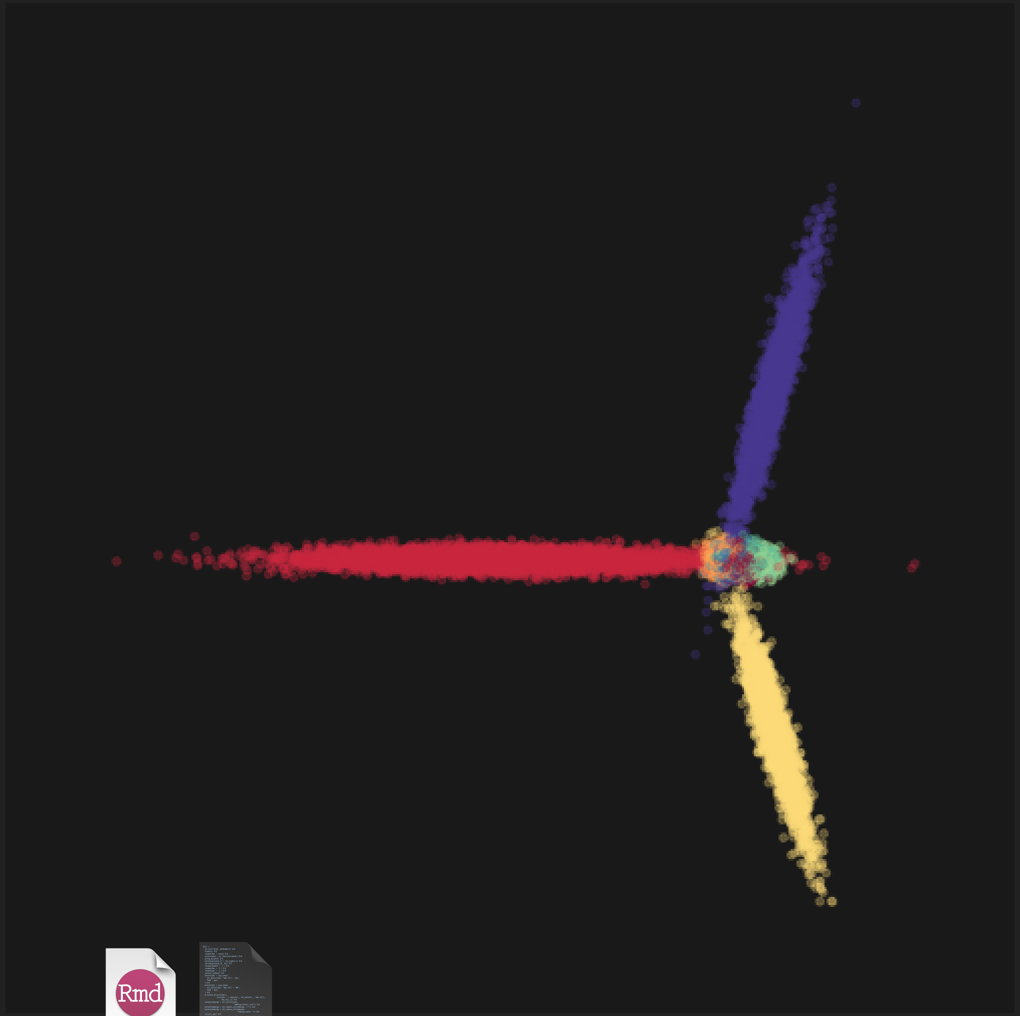
[5]

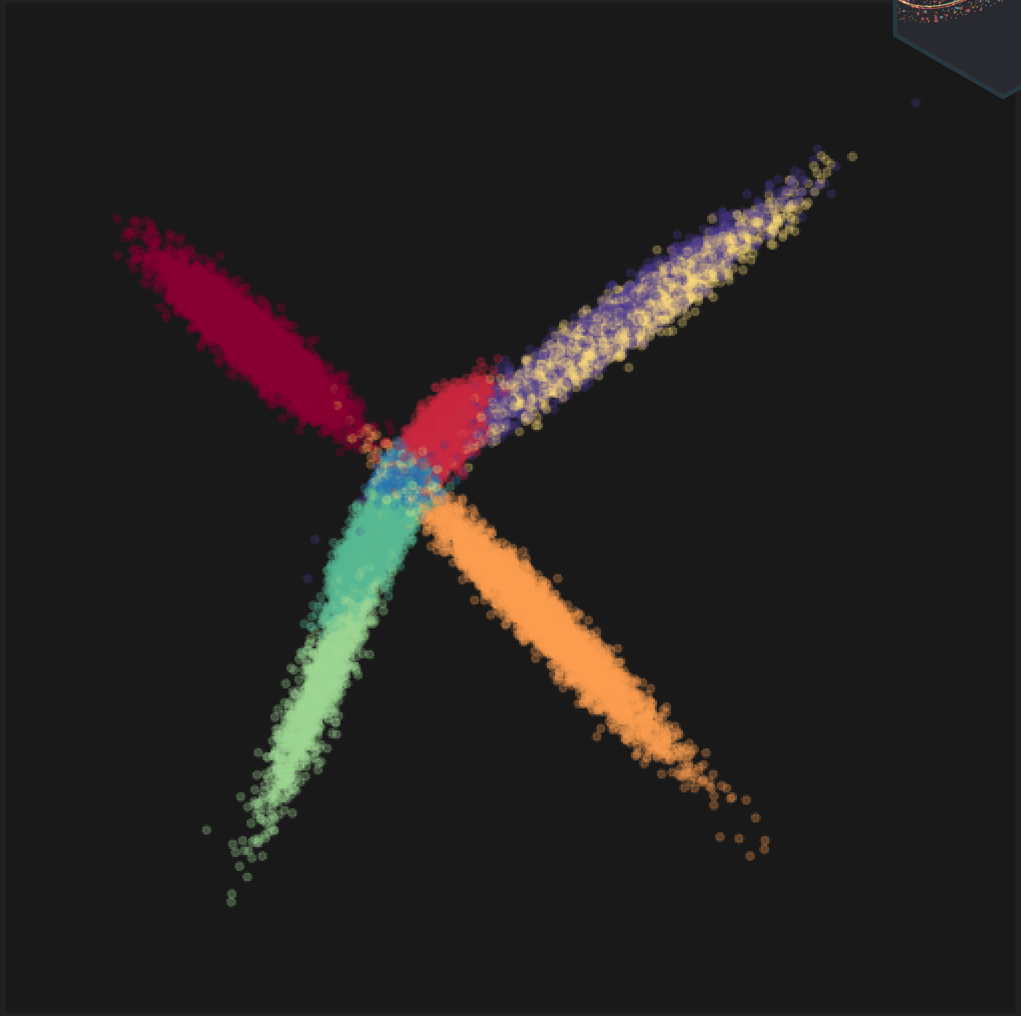
Below are steps within the the t-SNE process showing a complex data set, data reduction, and then clustering[6]



[6] HTML

And just for fun here are two pca rotations of the example data set[7]





[7]

Surveys and PCAs

In general, using a PCA in survey data analysis helps you to understand

Surveys and PCAs

In general, using a PCA in survey data analysis helps you to understand

- how each item is similar to all others and the strength of that relationship

Surveys and PCAs

In general, using a PCA in survey data analysis helps you to understand

- how each item is similar to all others and the strength of that relationship
- which items are should likely be kept or removed

But Wait There's More!

Again this is just the tip of the iceberg. To really see the power of PCAs, take a look at machine learning. This is just one of many ways to deal with classification and dimensionality. Here are a couple resources. At this time, its good just to ignore the coding and to simply get a basic idea of each.

[8] If you cannot see the entire page, please load the site in a private window. Directions on how to do this are provided for

But Wait There's More!

Again this is just the tip of the iceberg. To really see the power of PCAs, take a look at machine learning. This is just one of many ways to deal with classification and dimensionality. Here are a couple resources. At this time, its good just to ignore the coding and to simply get a basic idea of each.

- 11 Dimensionality reduction techniques you should know in 2021[8]

[8] If you cannot see the entire page, please load the site in a private window. Directions on how to do this are provided for

But Wait There's More!

Again this is just the tip of the iceberg. To really see the power of PCAs, take a look at machine learning. This is just one of many ways to deal with classification and dimensionality. Here are a couple resources. At this time, its good just to ignore the coding and to simply get a basic idea of each.

- 11 Dimensionality reduction techniques you should know in 2021[8]
- Understanding Dimension Reduction and Principal Component Analysis in R for Data Science[8]

[8] If you cannot see the entire page, please load the site in a private window. Directions on how to do this are provided for

But Wait There's More!

Again this is just the tip of the iceberg. To really see the power of PCAs, take a look at machine learning. This is just one of many ways to deal with classification and dimensionality. Here are a couple resources. At this time, its good just to ignore the coding and to simply get a basic idea of each.

- 11 Dimensionality reduction techniques you should know in 2021[8]
- Understanding Dimension Reduction and Principal Component Analysis in R for Data Science[8]
- Workshop: Dimension reduction with R

[8] If you cannot see the entire page, please load the site in a private window. Directions on how to do this are provided for

Thats it!

If you have any questions, please reach out

Thats it!

If you have any questions, please reach out

This work is licensed under a
Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License