Multivariate Statistical Methods for Big Data Analysis and Process Improvement

Lecture 2: Intro to PCA

Dr. Brandon Corbett

Course notes for ChE 765/SEP 767, McMaster University

Getting started: Visualizing (big) data

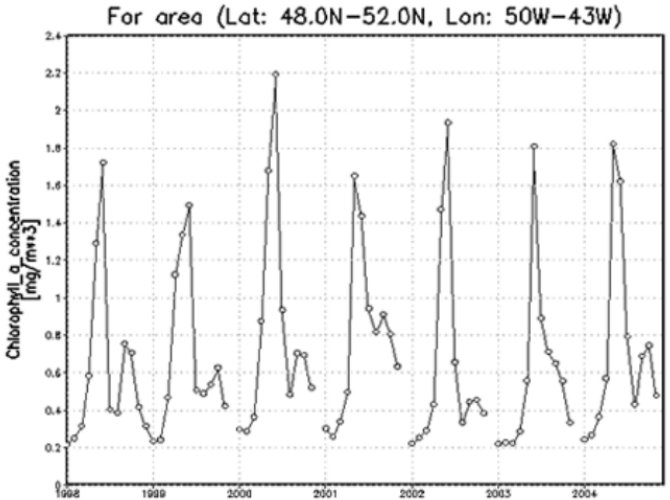
Is data learning visualization important?

Principle

Your first step with a new dataset should always be visualization (graphing it)!

- Human eyes and brains are very good at pattern recognition
- Learning how to make good plots enables understanding
- Time spent plotting data can avoid major headaches troubleshooting models
- Failure to understand what is in your data *will* cause bad outcomes
  - Garbage in = garbage out

Illustrative example

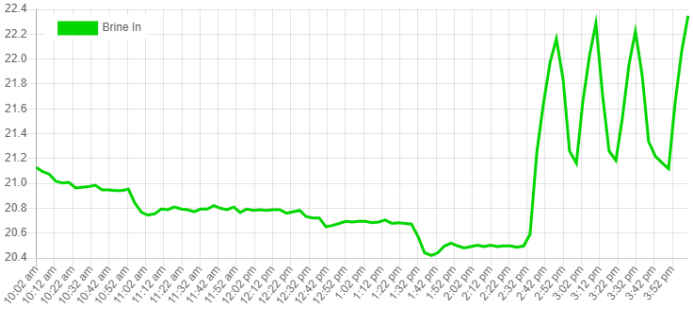


Types of plots

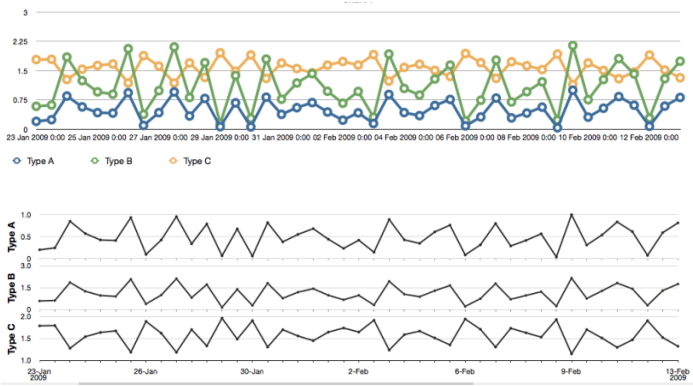
- **Time-series**
  - Univariate plot
  - Detect patterns in single variable
- **Bar plots**
  - Univariate plot
  - Compare categories for a single variable
- **Scatter plots**
  - Bivariate
  - Analyze relationship between two variables

Time-series

- Univariate because it only shows one sequence of data
- 2-dimensional plot:
  - Horizontal axis: time or some other logical order
  - Vertical axis: Data values



Time-series

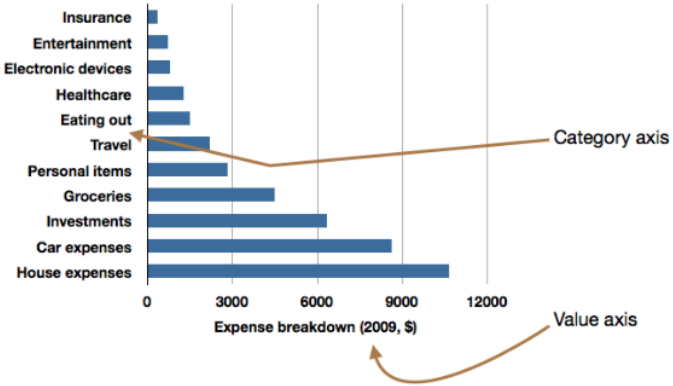


Making good time-series plots

Separate axes for each variable, use "minimal ink", don't use excel defaults!

Bar plots

- Has a *category axis* and a *value axis*
- Univariate plot because only shows one value axis
- Interpretation does not change when category axis is reordered
  - **Always sort category axis by values!**



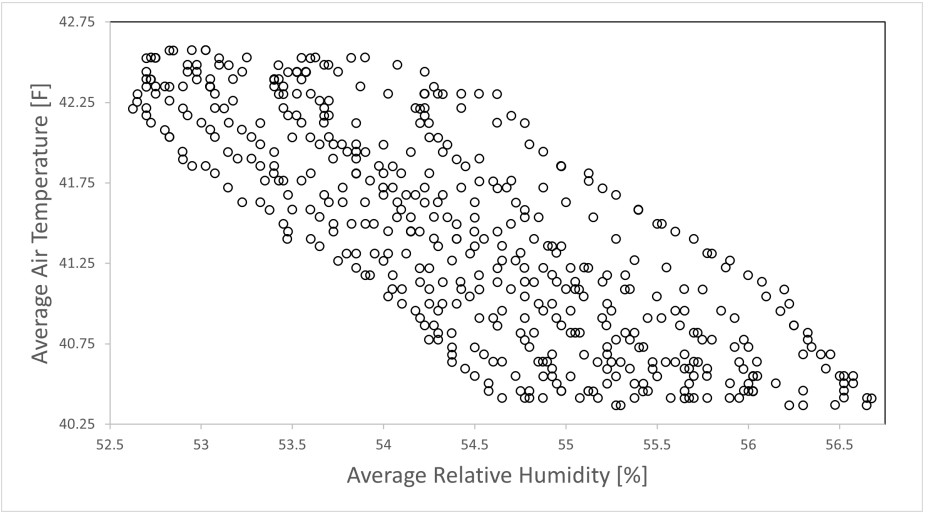
Scatter plot

- Bivariate: two variables
- **Goal:** understand relationship between two variables
- Each point is found by the intersection of values on each axis

Caution

Scatter plots implicitly ask viewer to draw a **causal** relationship between the two variables

Scatter plot



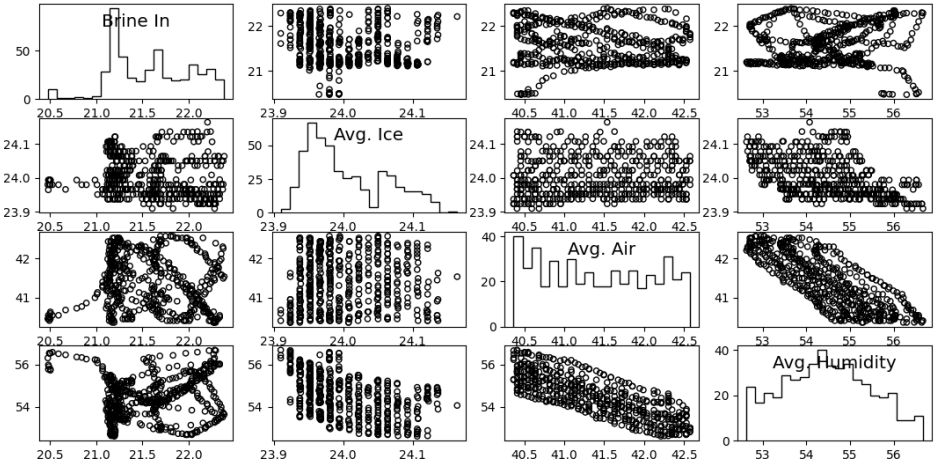
Avg. Humidity	Avg. Air
55.8	40.4
54.5	40.7
56.0	40.8
55.0	40.5
56.5	40.4
54.5	40.6
54.1	41.6
56.5	40.6
53.0	41.9
⋮	⋮

First attempt at multivariate analysis...

What if I want to analyze the relationships between 3+ variables?

Can this be done with scatter plots?

Scatter plot matrix



Scatter plot matrix code...

```
import matplotlib.pyplot as plt
import pandas as  pd

if __name__ == "__main__":
    data = pd.read_csv("L2_IntroToPCA/IceSenseDataRaw.csv")
    cols = data.columns
    for row_number, row_var in enumerate(cols):
        for column_number, col_var in enumerate(cols):
            if(row_number!= column_number):
                # Create scatter plots for off diagonal subplots
                plt.subplot( len(cols), len(cols), row_number*len(cols) + column_number +1)
                plt.scatter(data[col_var], data[row_var],
                            s=20, facecolors='none', edgecolors='k')
            else:
                # Put histograms on diagonal subplots
                plt.subplot( len(cols), len(cols), row_number*len(cols) + column_number +1)
                plt.hist(data[row_var], bins =20, histtype='step', color='k')

                # label each diagonal plot
                xlims = plt.xlim()
                ylims= plt.ylim()
                plt.text(0.5*xlims[0] + 0.5*xlims[1],
                         0.1*ylims[0] + 0.9*ylims[1],
                         row_var,
                         ha="center", va="top",fontsize ='x-large')
    plt.show()
```

Scatter plot matrix

Let's try again with a slightly larger dataset...

Observation ID	GrainCoffee	InstantCoffee	Tea	Sweetener	Biscuits	PowderSoup	TinSoup	Potato	FrozenFish	FrozenVeg	Apples	Oranges	TinnedFruit	Jam	Garlic	Butter	Margarine	OliveOil	Yogurt	CrispBread
Germany	90	49	88	19	57	51	19	21	27	21	81	75	44	71	22	91	85	74	30	26
Italy	82	10	60	2	55	41	3	2	4	2	67	71	9	46	80	66	24	94	5	18
France	88	42	63	4	76	53	11	23	11	5	87	84	40	45	88	94	47	36	57	3
Holland	96	62	98	32	62	67	43	7	14	14	83	89	61	81	15	31	97	13	53	15
Belgium	94	38	48	11	74	37	23	9	13	12	76	76	42	57	29	84	80	83	20	5
Luxembourg	97	61	86	28	79	73	12	7	26	23	85	94	83	20	91	94	94	84	31	24
England	27	86	99	22	91	55	76	17	20	24	76	68	89	91	11	95	94	57	11	28
Portugal	72	26	77	2	22	34	1	5	20	3	22	51	8	16	89	65	78	92	6	9
Austria	55	31	61	15	29	33	1	5	15	11	49	42	14	41	51	51	72	28	13	11
Switzerland	73	72	85	25	31	69	10	17	19	15	79	70	46	61	64	82	48	61	48	30
Sweden	97	13	93	31		43	43	39	54	45	56	78	53	75	9	68	32	48	2	93
Denmark	96	17	92	35	66	32	17	11	51	42	81	72	50	64	11	92	91	30	11	34
Norway	92	17	83	13	62	51	4	17	30	15	61	72	34	51	11	63	94	28	2	62
Finland	98	12	84	20	64	27	10	8	18	12	50	57	22	37	15	96	94	17		64
Spain	70	40	40		62	43	2	14	23	7	59	77	30	38	86	44	51	91	16	13
Ireland	30	52	99	11	80	75	18	2	5	3	57	52	46	89	5	97	25	31	3	9

How many scatter plots do I need?

Aside on data analysis

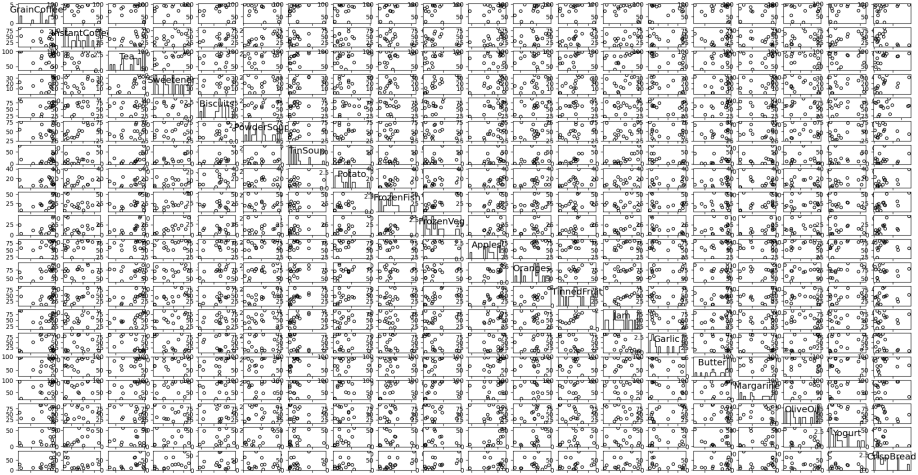
Principle

Don't say **"I will do data analysis"**... that is meaningless

We **always** analyze data to answer a question!

In this case, we will ask "Do European countries located close together have similar food consumption habits?"

Scatter plot matrix



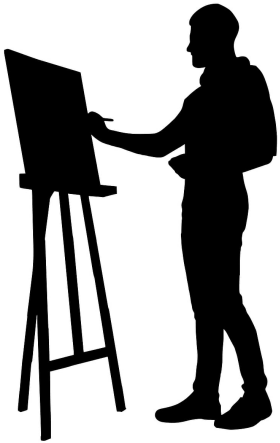
Brainstorm exercise

What should we do now?

How would you solve this problem?

An exercise in art

Wanted: volunteers with "art" skills to draw on the board...



Conclusions

How can we represent high dimensional data in a reduced dimensional space?

Introduction to Principal Component Analysis (PCA)

Overview

PCA provides an **optimal** low dimensional representation of a single table of data, $\mathbf{X}$

- In PCA, summary variables, $\mathbf{t_a}$, are used to summarize the full set of variables

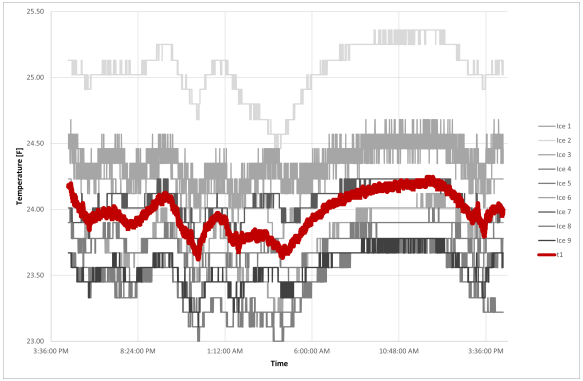
What is PCA?

Ice 1	Ice 2	Ice 3	Ice 4	Ice 5	Ice 6	Ice 7	Ice 8	Ice 9
23.67	25.02	24.12	23.33	23.45	24.35	24.01	23.67	23.45
23.67	25.02	24.23	23.45	23.45	24.35	24.01	23.67	23.56
23.56	25.02	24.12	23.33	23.45	24.35	24.12	23.78	23.45
23.67	25.02	24.12	23.33	23.45	24.23	24.01	23.67	23.45
23.67	25.02	24.12	23.33	23.45	24.23	24.12	23.78	23.56
23.78	25.02	24.23	23.45	23.45	24.35	24.01	23.67	23.56
23.67	25.02	24.12	23.33	23.45	24.12	24.01	23.67	23.45
23.67	25.02	24.12	23.33	23.45	24.23	24.01	23.67	23.45
23.67	25.13	24.12	23.33	23.45	24.35	24.12	23.67	23.45
23.78	25.02	24.23	23.45	23.56	24.35	24.01	23.67	23.56
23.67	25.02	24.12	23.45	23.45	24.35	24.01	23.67	23.45
23.56	25.13	24.12	23.33	23.45	24.35	24.12	23.67	23.45
23.67	25.02	24.12	23.33	23.45	24.23	24.01	23.67	23.45
23.67	25.02	24.23	23.45	23.45	24.23	24.01	23.67	23.45
23.78	25.13	24.23	23.45	23.56	24.23	24.01	23.67	23.56
23.67	25.02	24.12	23.33	23.45	24.35	24.01	23.67	23.45
23.78	25.13	24.35	23.56	23.56	24.35	24.12	23.78	23.67
23.56	25.02	24.12	23.33	23.56	24.23	24.12	23.78	23.56

How would you summarize this data in one variable?

What is PCA?

Ice 1	Ice 2	Ice 3	Ice 4	Ice 5	Ice 6	Ice 7	Ice 8	Ice 9	t1
23.67	25.02	24.12	23.33	23.45	24.35	24.01	23.67	23.45	23.90
23.67	25.02	24.23	23.45	23.45	24.35	24.01	23.67	23.56	23.93
23.56	25.02	24.12	23.33	23.45	24.35	24.12	23.78	23.45	23.91
23.67	25.02	24.12	23.33	23.45	24.23	24.01	23.67	23.45	23.88
23.67	25.02	24.12	23.33	23.45	24.23	24.12	23.78	23.56	23.92
23.78	25.02	24.23	23.45	23.45	24.35	24.01	23.67	23.56	23.95
23.67	25.02	24.12	23.33	23.45	24.12	24.01	23.67	23.45	23.87
23.67	25.02	24.12	23.33	23.45	24.23	24.01	23.67	23.45	23.88
23.67	25.13	24.12	23.33	23.45	24.35	24.12	23.67	23.45	23.92
23.78	25.02	24.23	23.45	23.56	24.35	24.01	23.67	23.56	23.96
23.67	25.02	24.12	23.45	23.45	24.35	24.01	23.67	23.45	23.91
23.56	25.13	24.12	23.33	23.45	24.35	24.12	23.67	23.45	23.91
23.67	25.02	24.12	23.33	23.45	24.23	24.01	23.67	23.45	23.88
23.67	25.02	24.23	23.45	23.45	24.23	24.01	23.67	23.45	23.91
23.78	25.13	24.23	23.45	23.56	24.23	24.01	23.67	23.56	23.96
23.67	25.02	24.12	23.33	23.45	24.35	24.01	23.67	23.45	23.90
23.78	25.13	24.35	23.56	23.56	24.35	24.12	23.78	23.67	24.03
23.56	25.02	24.12	23.33	23.56	24.23	24.12	23.78	23.56	23.92



What if you knew one sensor was not as good as the others?

What is PCA?

First understanding of PCA

- **PCA** provides an **optimal** low dimensional representation of a single table of data, $\mathbf{X}$
- Summary values, $\mathbf{t_a}$, are used to summarize the full set of variables for each observation
- Summary values, $\mathbf{t_a}$, are **weighted averages** calculated using **optimal weights**

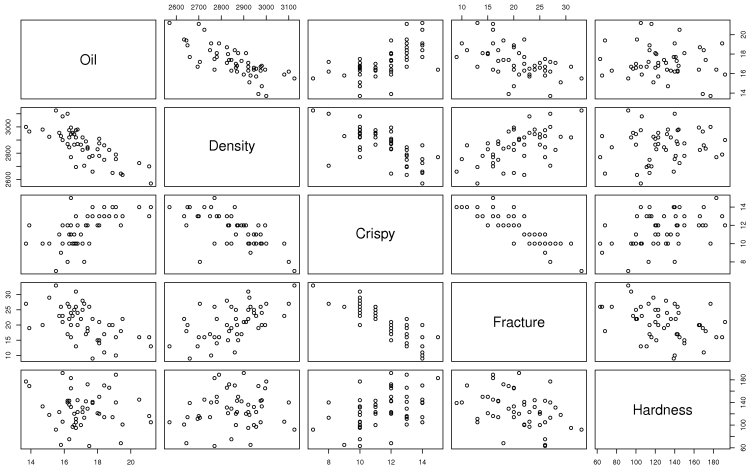
What is PCA?

How does this relate to the drawing exercise?

Let's look at a geometric interpretation of PCA... but first, basic data pre-processing

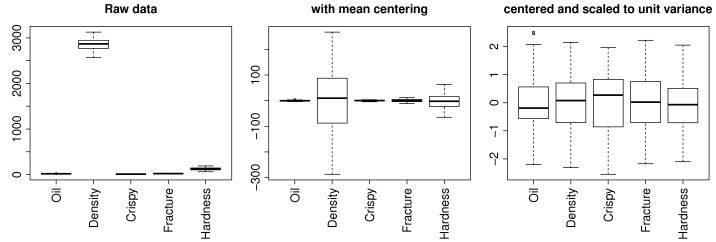
Preprocessing by example

Raw data:



Preprocessing by example

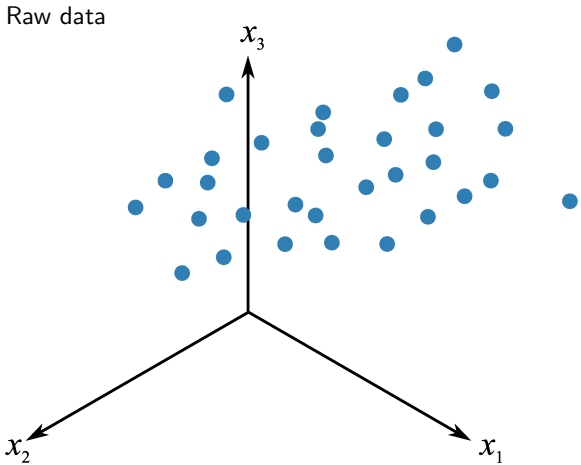
Center and scale the raw data



- Centering: $\mathbf{x}_{k,\text{center}} = \mathbf{x}_{k,\text{raw}} - \text{mean}\left(\mathbf{x}_{k,\text{raw}}\right)$
- Scaling: $\mathbf{x}_k = \dfrac{\mathbf{x}_{k,\text{center}}}{\text{standard deviation}\left(\mathbf{x}_{k,\text{center}}\right)}$
- Does *not change* relationships between variables

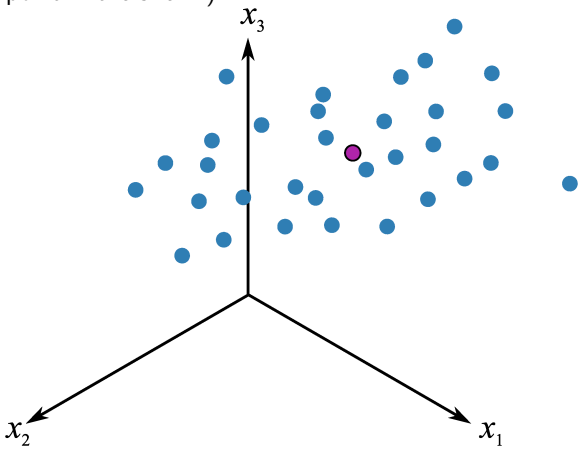
 12

Geometric explanation of preprocessing



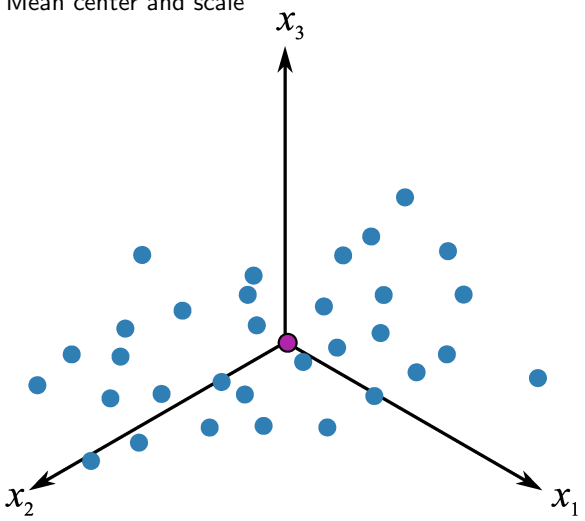
Geometric explanation of preprocessing

Calculate the mean of each variable (creates a "new" reference point in the swarm)

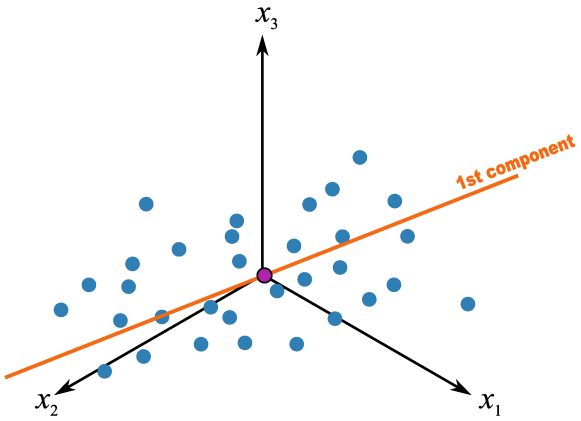


Geometric explanation of preprocessing

Mean center and scale

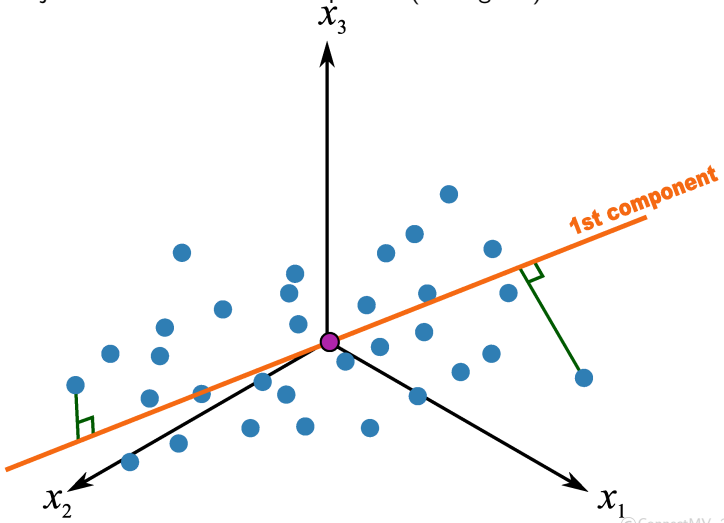


Geometric explanation of PCA



Geometric explanation of PCA

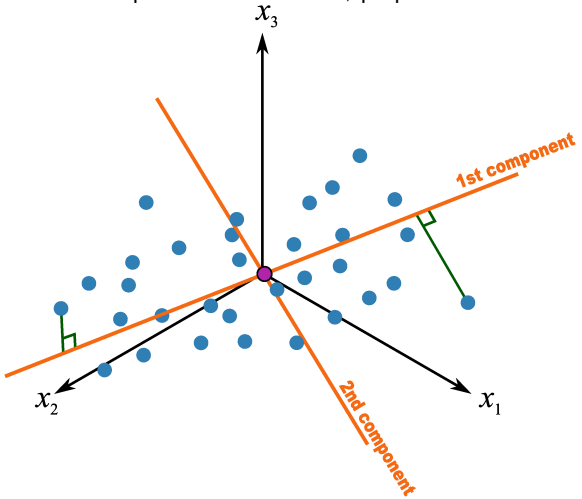
Project observations onto component (90 degrees)



 17

Geometric explanation of PCA

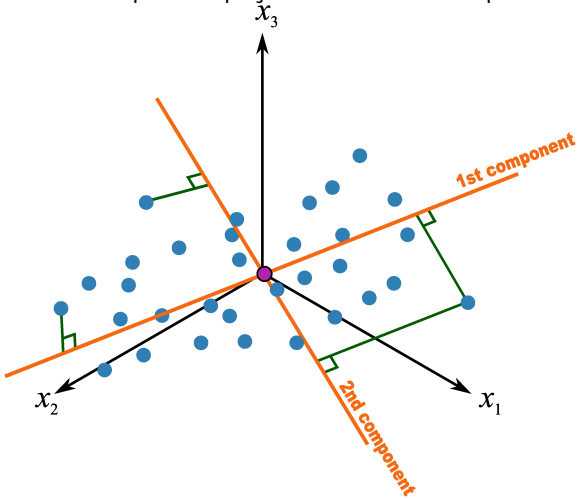
Second component: best-fit line; perpendicular to 1st component



 18

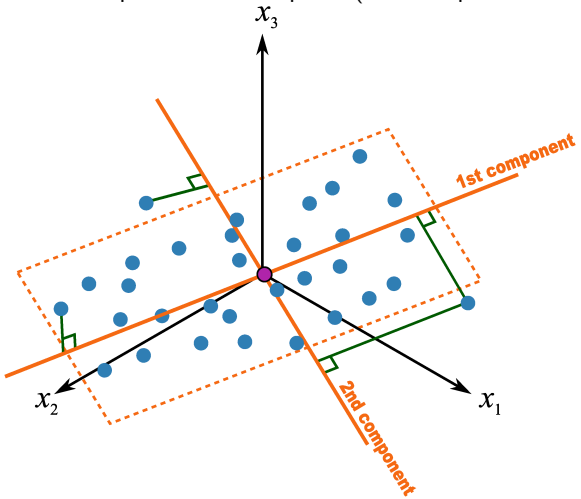
Geometric explanation of PCA

Second component: project onto second component



Geometric explanation of PCA

The 2 components form a plane (2-D subspace inside a 3D space)



Geometric explanation of PCA

What have we done here?

Broken **X** down into 2 parts:

- projected points "*on the plane*"
- residual distance "*off the plane*"

Mathematical derivation for PCA

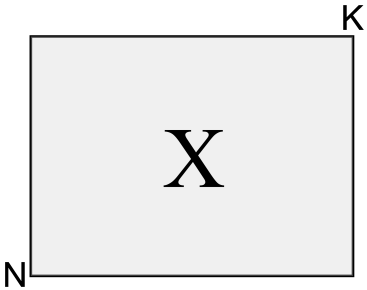
Time to break out some math!

Course notation

On board...

Data sources

- PCA considers a single data table (matrix)
- We will call it **X**

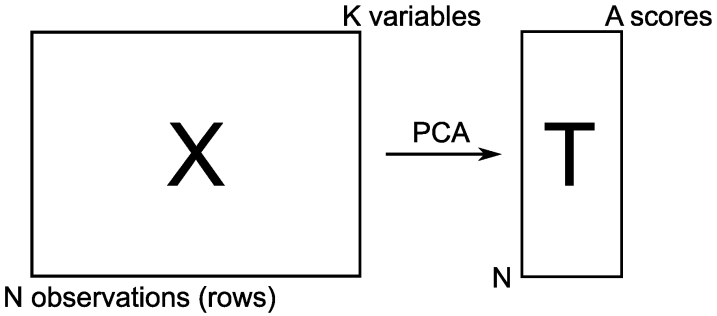


- N observations
- K variables
- What goes in the columns of **X**?
- What goes in the rows?

What is PCA (Principal Components Analysis)?

Mathematical objective

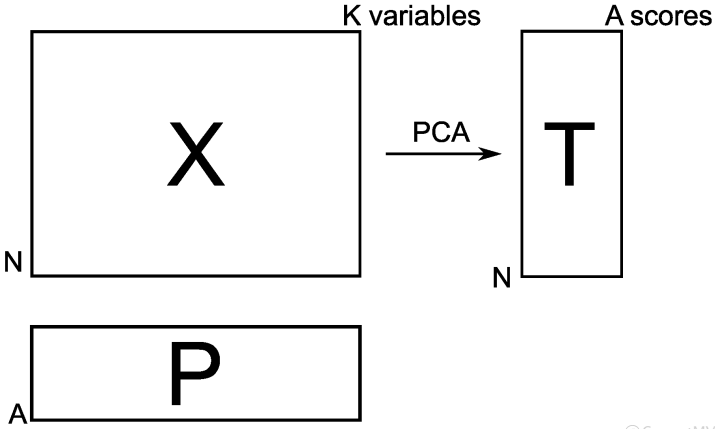
PCA: find me the best summary of my data, **X**, with the fewest number of summary variables, called scores, **T**.



Objectives for this class

PCA model will calculate from **X**:

- scores: **T**
- loadings: **P**

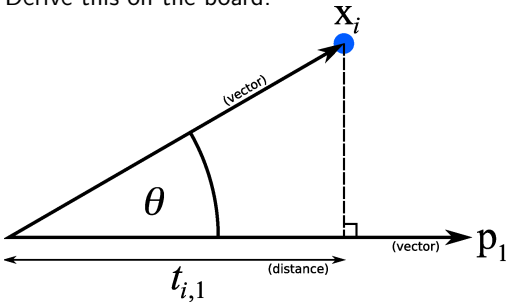


Mathematical derivation of PCA

On board...

Mathematical derivation for PCA

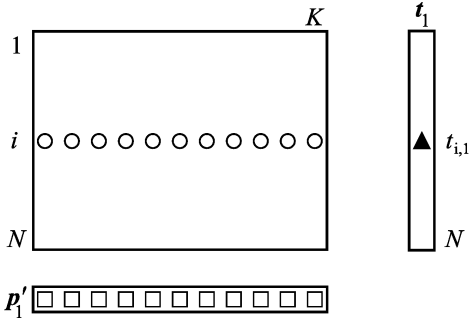
Derive this on the board:



$$\cos\theta = \frac{\text{adjacent length}}{\text{hypotenuse}} = \frac{t_{i,1}}{\|\mathbf{x}_i\|} \quad \text{and also } \cos\theta = \frac{\mathbf{x}_i^T\mathbf{p}_1}{\|\mathbf{x}_i\|\|\mathbf{p}_1\|}$$

$$\begin{aligned} \frac{t_{i,1}}{\|\mathbf{x}_i\|} &= \frac{\mathbf{x}_i^T\mathbf{p}_1}{\|\mathbf{x}_i\|\|\mathbf{p}_1\|} \\ t_{i,1} &= \mathbf{x}_i^T\mathbf{p}_1 \\ (1\times 1) &= (1\times K)(K\times 1) \end{aligned}$$

Mathematical derivation for PCA



$$
\begin{aligned}
t_{i,1} &= \mathbf{x}_i^T \mathbf{p}_1 \\
&= x_{i,1}p_{1,1} + x_{i,2}p_{2,1} + \ldots + x_{i,k}p_{k,1} + \ldots + x_{i,K}p_{K,1}
\end{aligned}
$$

- K individual terms add up (i.e. linear combination) to give t_1
- Stack $t_{i,1}$ values from N rows: $\mathbf{T} = \mathbf{XP}$

Interpreting $t_i = \mathbf{x}_i^T \mathbf{p}_1$

Given that:

- values in $\mathbf{x}_i^T$ are centred and scaled, and
- entries in $\mathbf{p}$ are between -1 and $+1$

using

$$= x_{i,1}p_{1,1} + x_{i,2}p_{2,1} + \ldots + x_{i,k}p_{k,1} + \ldots + x_{i,K}p_{K,1}$$

how would you

- get a large positive value of $t_{i,1}$?
- get a large negative value of $t_{i,1}$?
- get a value of $t_{i,1} \approx 0$?

- What can you say about observation (row) 13 and 22 if $t_{13,1} \approx t_{22,1}$?

PCA in one slide!

Summary of PCA

- **PCA** provides an **optimal** low dimensional representation of a single table of data, $\mathbf{X}$
- Summary values, $\mathbf{t_a}$, are **weighted averages** calculated using **optimal weights**
- Loading values, $\mathbf{p_a}$, are **weights** used to calculate **weighted averages**

Plots

- **Score plot:** How are observations related?
- **Loadings plot:** How are variables related?
- **R-squared:** How well do summary variables describe original data?
- **Hotellings T Squared:** How extreme are observations?
- **SPE:** How consistent are observations?

Let's see an example

Back to the foods example

Observation ID	GrainCoffee	InstantCoffee	Tea	Sweetener	Biscuits	PowderSoup	TinSoup	Potato	FrozenFish	FrozenVeg	Apples	Oranges	TinnedFruit	Jam	Garlic	Butter	Margarine	OliveOil	Yogurt	CrispBread
Germany	90	49	88	19	57	51	19	21	27	21	81	75	44	71	22	91	85	74	30	26
Italy	82	10	60	2	55	41	3	2	4	2	67	71	9	46	80	66	24	94	5	18
France	88	42	63	4	76	53	11	23	11	5	87	84	40	45	88	94	47	36	57	3
Holland	96	62	98	32	62	67	43	7	14	14	83	89	61	81	15	31	97	13	53	15
Belgium	94	38	48	11	74	37	23	9	13	12	76	76	42	57	29	84	80	83	20	5
Luxembourg	97	61	86	28	79	73	12	7	26	23	85	94	83	20	91	94	94	84	31	24
England	27	86	99	22	91	55	76	17	20	24	76	68	89	91	11	95	94	57	11	28
Portugal	72	26	77	2	22	34	1	5	20	3	22	51	8	16	89	65	78	92	6	9
Austria	55	31	61	15	29	33	1	5	15	11	49	42	14	41	51	51	72	28	13	11
Switzerland	73	72	85	25	31	69	10	17	19	15	79	70	46	61	64	82	48	61	48	30
Sweden	97	13	93	31		43	43	39	54	45	56	78	53	75	9	68	32	48	2	93
Denmark	96	17	92	35	66	32	17	11	51	42	81	72	50	64	11	92	91	30	11	34
Norway	92	17	83	13	62	51	4	17	30	15	61	72	34	51	11	63	94	28	2	62
Finland	98	12	84	20	64	27	10	8	18	12	50	57	22	37	15	96	94	17		64
Spain	70	40	40		62	43	2	14	23	7	59	77	30	38	86	44	51	91	16	13
Ireland	30	52	99	11	80	75	18	2	5	3	57	52	46	89	5	97	25	31	3	9

How can PCA help?

Demo in ProMV

Plan from here for PCA

- **Next two weeks**
  - How do I know my low dimensional space described my data? (Error)
  - How do I pick the number of components?
  - How do I calculate which direction components should point? (optimal weights)

- **After that**
  - Applications to real problems in industry...