An Introduction to Principal Component Analysis

Outline

- LEARNING OBJECTIVES:
  - You will know why we use dimensionality reduction.
  - You will know the PCA implementation process.
  - You will know at a high level what the PCA formulas are doing in a small sample setting.

- YOU WILL NOT:
  - Know the linear algebra that makes PCA work.
  - You will not be able to magically think in more than 3 dimensions.

Key Topics:

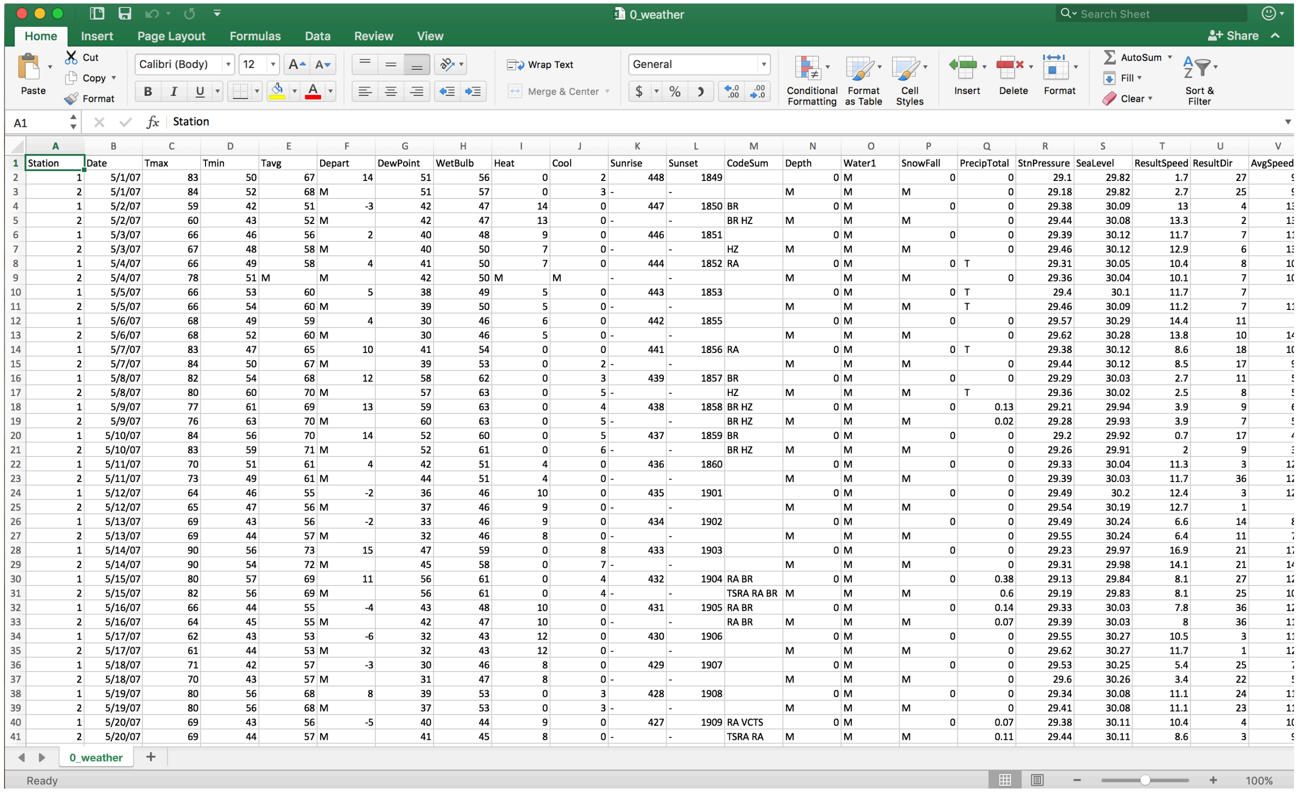
- What is dimensionality reduction?

- How do we use PCA for dimensionality reduction?

- How is are the PCA formulas working to do this?

Everyone has been here…

- You have all the columns below in your dataset. Which ones are you going to use to explain what is happening?



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	Station	Date	Tmax	Tmin	Tavg	Depart	DewPoint	WetBulb	Heat	Cool	Sunrise	Sunset	CodeSum	Depth	Water1	SnowFall	PrecipTotal	StnPressure	SeaLevel	ResultSpeed	ResultDir	AvgSpeed
2	1	5/1/07	83	50	67	14	51	56	0	2	448	1849		0	M	0	0	29.1	29.82	1.7	27	9
3	2	5/1/07	84	52	68	M	51	57	0	3	-	-		M	M	M	0	29.18	29.82	2.7	25	9
4	1	5/2/07	59	42	51	-3	42	47	14	0	447	1850	BR	0	M	0	0	29.38	30.09	13	4	13
5	2	5/2/07	60	43	52	M	42	47	13	0	-	-	BR HZ	M	M	M	0	29.44	30.08	13.3	2	13
6	1	5/3/07	66	46	56	2	40	48	9	0	446	1851		0	M	0	0	29.39	30.12	11.7	7	11
7	2	5/3/07	67	48	58	M	40	50	7	0	-	-	HZ	M	M	M	0	29.46	30.12	12.9	6	13
8	1	5/4/07	66	49	58	4	41	50	7	0	444	1852	RA	0	M	0	T	29.31	30.05	10.4	8	10
9	2	5/4/07	78	51	M	M	42	50	M	M	-	-		M	M	M	0	29.36	30.04	10.1	7	10
10	1	5/5/07	66	53	60	5	38	49	5	0	443	1853		0	M	0	T	29.4	30.1	11.7	7	
11	2	5/5/07	66	54	60	M	39	50	5	0	-	-		M	M	M	T	29.46	30.09	11.2	7	11
12	1	5/6/07	68	49	59	4	30	46	6	0	442	1855		0	M	0	0	29.57	30.29	14.4	11	
13	2	5/6/07	68	52	60	M	30	46	5	0	-	-		M	M	M	0	29.62	30.28	13.8	10	14
14	1	5/7/07	83	47	65	10	41	54	0	0	441	1856	RA	0	M	0	T	29.38	30.12	8.6	18	10
15	2	5/7/07	84	50	67	M	39	53	0	2	-	-		M	M	M	0	29.44	30.12	8.5	17	9
16	1	5/8/07	82	54	68	12	58	62	0	3	439	1857	BR	0	M	0	0	29.29	30.03	2.7	11	5
17	2	5/8/07	80	60	70	M	57	63	0	5	-	-	HZ	M	M	M	T	29.36	30.02	2.5	8	5
18	1	5/9/07	77	61	69	13	59	63	0	4	438	1858	BR HZ	0	M	0	0.13	29.21	29.94	3.9	9	6
19	2	5/9/07	76	63	70	M	60	63	0	5	-	-	BR HZ	M	M	M	0.02	29.28	29.93	3.9	7	5
20	1	5/10/07	84	56	70	14	52	60	0	5	437	1859	BR	0	M	0	0	29.2	29.92	0.7	17	4
21	2	5/10/07	83	59	71	M	52	61	0	6	-	-	BR HZ	M	M	M	0	29.26	29.91	2	9	3
22	1	5/11/07	70	51	61	4	42	51	4	0	436	1860		0	M	0	0	29.33	30.04	11.3	3	12
23	2	5/11/07	73	49	61	M	44	51	4	0	-	-		M	M	M	0	29.39	30.03	11.7	36	12
24	1	5/12/07	64	46	55	-2	36	46	10	0	435	1901		0	M	0	0	29.49	30.2	12.4	3	12
25	2	5/12/07	65	47	56	M	37	46	9	0	-	-		M	M	M	0	29.54	30.19	12.7	1	
26	1	5/13/07	69	43	56	-2	33	46	9	0	434	1902		0	M	0	0	29.49	30.24	6.6	14	8
27	2	5/13/07	69	44	57	M	32	46	8	0	-	-		M	M	M	0	29.55	30.24	6.4	11	7
28	1	5/14/07	90	56	73	15	47	59	0	8	433	1903		0	M	0	0	29.23	29.97	16.9	21	17
29	2	5/14/07	90	54	72	M	45	58	0	7	-	-		M	M	M	0	29.31	29.98	14.1	21	14
30	1	5/15/07	80	57	69	11	56	61	0	4	432	1904	RA BR	0	M	0	0.38	29.13	29.84	8.1	27	12
31	2	5/15/07	82	56	69	M	56	61	0	4	-	-	TSRA RA BR	M	M	M	0.6	29.19	29.83	8.1	25	10
32	1	5/16/07	66	44	55	-4	43	48	10	0	431	1905	RA BR	0	M	0	0.14	29.33	30.03	7.8	36	12
33	2	5/16/07	64	45	55	M	42	47	10	0	-	-	RA BR	M	M	M	0.07	29.39	30.03	8	36	11
34	1	5/17/07	62	43	53	-6	32	43	12	0	430	1906		0	M	0	0	29.55	30.27	10.5	3	11
35	2	5/17/07	61	44	53	M	32	43	12	0	-	-		M	M	M	0	29.62	30.27	11.7	1	12
36	1	5/18/07	71	42	57	-3	30	46	8	0	429	1907		0	M	0	0	29.53	30.25	5.4	25	7
37	2	5/18/07	70	43	57	M	31	47	8	0	-	-		M	M	M	0	29.6	30.26	3.4	22	5
38	1	5/19/07	80	56	68	8	39	53	0	3	428	1908		0	M	0	0	29.34	30.08	11.1	24	11
39	2	5/19/07	80	56	68	M	37	53	0	3	-	-		M	M	M	0	29.41	30.08	11.1	23	11
40	1	5/20/07	69	43	56	-5	40	44	9	0	427	1909	RA VCTS	0	M	0	0.07	29.38	30.11	10.4	4	10
41	2	5/20/07	69	44	57	M	41	45	8	0	-	-	TSRA RA	M	M	M	0.11	29.44	30.11	8.6	3	9

Key Topics:

- **What is dimensionality reduction?**

- *How do we use PCA for dimensionality reduction?*

- *How is are the PCA formulas working to do this?*

High-level definition:

- Dimensionality reduction is used to end up with fewer features in our data set that we can then analyze.

Key Topics:

- What is dimensionality reduction?

- *How do we use PCA for dimensionality reduction?*

- *How is are the PCA formulas working to do this?*

Two ways to do it.

- 1. FEATURE SELECTION: Select which columns to use and ignore the rest.
- 2. FEATURE EXTRACTION: Generalizes the columns into a set of new columns. Some of the new columns might be more useful than their original setup and some columns might be less useful. So we drop the ones that are less useful.

PCA is a type of feature extraction.

Key Topics:

- **What is dimensionality reduction?**

- *How do we use PCA for dimensionality reduction?*

- *How is are the PCA formulas working to do this?*

Breaking down what dimensionality reduction means for PCA.

- Looking at this dataset below… what are the “Dimensions”?

Blue	***Red***	***Yellow***
20%	57%	23%
60%	40%	0%
50%	0%	50%

Key Topics:

- What is dimensionality reduction?

- How do we use PCA for dimensionality reduction?

- How is are the PCA formulas working to do this?

How can you reduce these dimensions with feature extraction? … enter PCA.

- PCA is going to look at all our features and then create new ones.
- These new ones are going to be a little special though… the first one of the new features will explain more of the data than the BLUE feature we currently have.
    - *If you're trying to figure out HOW pca does this, we will have an example after this section.*

Key Topics:

- What is dimensionality reduction?

- How do we use PCA for dimensionality reduction?

- *How is are the PCA formulas working to do this?*

The oh so important process when using PCA.

- 1. You first need to take all your variables and “Standardize” them.
  - *This is so that when the PCA formula looks to make new features from all the ones we have it can make apples from apples.*
- 2. You then get to simply decide how many new variables to keep.

Wait… What’s the catch?

Key Topics:

- What is dimensionality reduction?

- How do we use PCA for dimensionality reduction?

- How is are the PCA formulas working to do this?

The catch.

- Each of the new features… or principle components are going to look very different.

Principle Component 1	***Principle Component 2***	***Principle Component 3***
70.1%	25.2%	4.5%
20.7%	32.4%	2.1%
45.3%	29.4%	3.4%

Key Topics:

- What is dimensionality reduction?

- How do we use PCA for dimensionality reduction?

- How is are the PCA formulas working to do this?

So your boss asks…

- Great predictive results! What was the most important thing that determined your findings?

Uh…...

Key Topics:

- What is dimensionality reduction?

- How do we use PCA for dimensionality reduction?

- *How is are the PCA formulas working to do this?*

The OTHER catch.

- Remember we said that we are dropping features… so... If we don’t know what these new features mean… how do we know we aren’t losing something really important?

Key Topics:

- What is dimensionality reduction?

- How do we use PCA for dimensionality reduction?

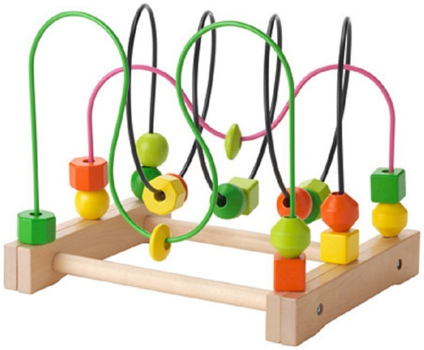
- How is are the PCA formulas working to do this?

PCA explained with spoons and my daughters toys.

- So remember we have three ”Dimensions”.

Blue ***Red*** ***Yellow***

- PCA will assign plot all of these data points in space that is equal to the number of features/dimensions we have. 3 features, so 3 dimensions.



Key Topics:

- What is dimensionality reduction?

- How do we use PCA for dimensionality reduction?

- How is are the PCA formulas working to do this?

Cool so we have our data plotted.

- PCA will now make principle components… or... Spoons. One spoon per dimension/feature.

Spoon 1

Spoon 2

Spoon 3

Key Topics:

- What is dimensionality reduction?

- How do we use PCA for dimensionality reduction?

- How is are the PCA formulas working to do this?

These spoons have rules…

- 1. The spoons have to be used to plot the data in the best way one at a time.
- 2. The spoons MUST be perpendicular to each other.
- *You can think of these as plotting a plane in our 3D space.*

Key Topics:

- What is dimensionality reduction?

- How do we use PCA for dimensionality reduction?

- How is are the PCA formulas working to do this?

Each spoon now says something about our data. In a new 3D plot.

Spoon 1

Captures 75% of the data spread.

Spoon 2

Captures an additional 22% of the data spread.

Spoon 3

Captures an additional 3% of the data spread.

Key Topics:

- What is dimensionality reduction?

- *How do we use PCA for dimensionality reduction?*

- How is are the PCA formulas working to do this?

So 3 features is nice… but what if I have 100 like with the beginning weather data?

Key Topics:

- *What is dimensionality reduction?*

- *How do we use PCA for dimensionality reduction?*

- How is are the PCA formulas working to do this?

So in the end… why you would want to use PCA.

1. ***If you DO NOT need to explain the data behind your results.***

2. ***Because PCA will give us the BEST features while using the least amount of columns.***

3. ***Because PCA makes cookies.***

That’s all folks.

Questions?