Data Mining and Text Mining

Rei Sanchez-Arias, Ph.D.
rsanchezarias@floridapoly.edu

Principal Component Analysis (PCA)

Goal: Reduce a set of numerical variables.

The idea: Remove the overlap of *information* between these variable. ["*Information*" is measured by the sum of the variances of the variables.]

Final product: A smaller number of numerical variables that contain most of the information

How does PCA do this?

Create **new variables** that are **linear combinations** of the original variables (i.e., they are weighted averages of the original variables).

These linear combinations are **uncorrelated** (no information overlap), and only a few of them contain most of the original information.

The new variables are called **principal components**.

Quick Example: Breakfast Cereals

name	mfr	type	calories	protein	…	rating
100%_Bran	N	C	70	4	…	68
100%_Natural_Bran	Q	C	120	3	…	34
All-Bran	K	C	70	4	…	59
All-Bran_with_Extra_Fiber	K	C	50	4	…	94
Almond_Delight	R	C	110	2	…	34
Apple_Cinnamon_Cheerios	G	C	110	2	…	30
Apple_Jacks	K	C	110	2	…	33
Basic_4	G	C	130	3	…	37
Bran_Chex	R	C	90	2	…	49
Bran_Flakes	P	C	90	3	…	53
Cap'n'Crunch	Q	C	120	1	…	18
Cheerios	G	C	110	6	…	51
Cinnamon_Toast_Crunch	G	C	120	1	…	20

Name: name of cereal
mfr: manufacturer
type: cold or hot
calories: calories per serving
protein: grams

fat: grams
sodium: mg.
fiber: grams
weight: oz. 1 serving
cups: in one serving
rating: consumer reports

carbo: grams complex carbohydrates
sugars: grams
potass: mg.
vitamins: % FDA rec
shelf: display shelf

Calories and Ratings Covariance Matrix

	calories	rating
calories	379.63	-189.68
rating	-189.68	197.32

$$C = \begin{pmatrix} \sigma(x,x) & \sigma(x,y) \\ \sigma(y,x) & \sigma(y,y) \end{pmatrix}$$

$$\sigma(x,y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Total **variance** (="information") is the sum of individual variances:
379.63 + 197.32

Calories accounts for 379.63/577 = 66%

If we want to make do with just calories, we lose 34% of the variation

Linear Combinations

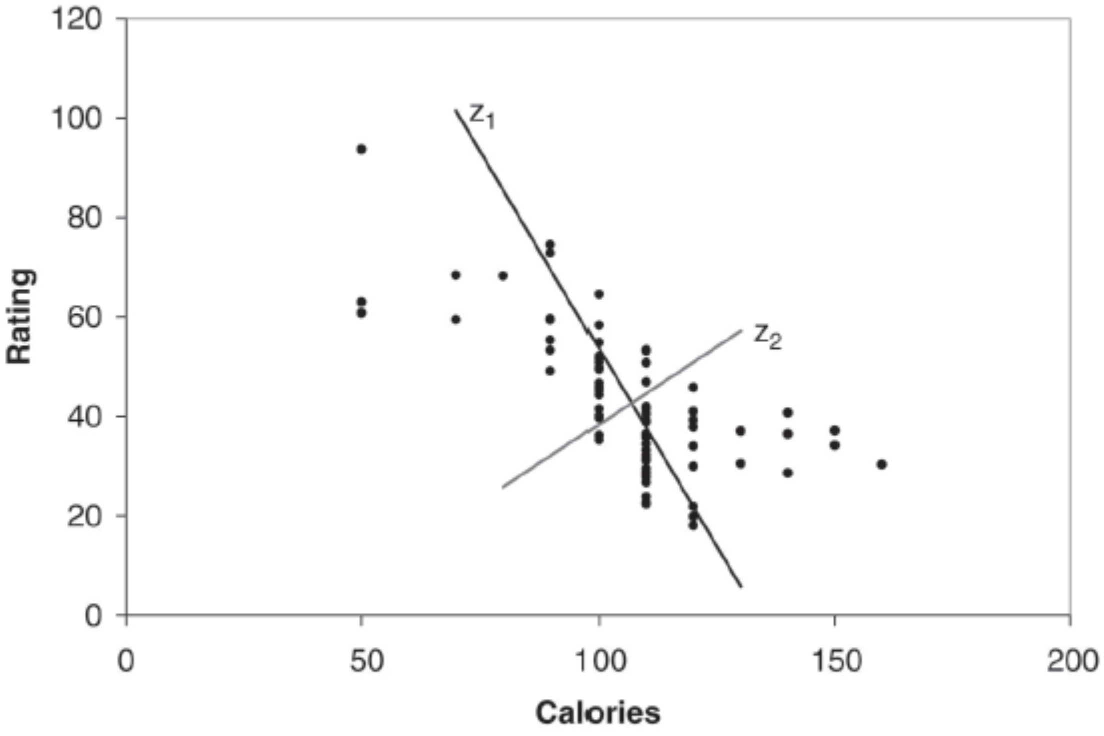
Using **linear combinations** to redistribute the variability in a more polarized way:

Z_1 and Z_2 are two linear combinations. Z_1 has the *highest* variation (*spread of values*)

Z_2 has the *lowest* variation

In general: $X_1, X_2, X_3, \ldots X_p$, original p variables

- Generate $Z_1, Z_2, Z_3, \ldots, Z_p$ weighted averages of original variables
- All pairs of Z variables have 0 **correlation**
- Order Z's by variance (Z_1 largest, Z_p smallest)
- Usually the first few Z variables contain *most of the information*, and so the rest can be dropped.



Standardization

Using standardization, we **center** the features at mean 0 and standard deviation 1.
In the cereals dataset example, without standardization, sodium will dominate the first principal component (PC), simply because of the way it is measured (mg): its scale is greater than almost all other variables

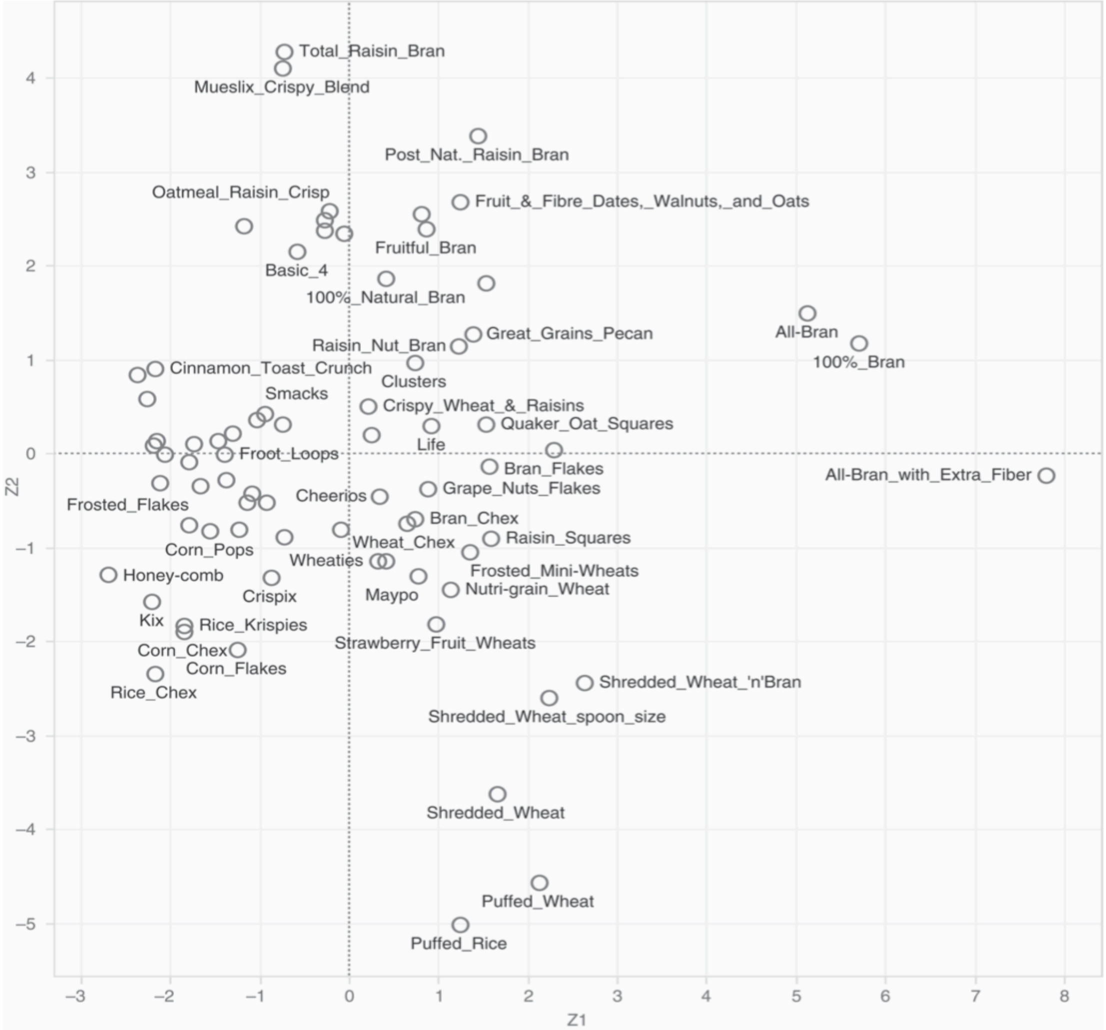
Hence its variance will be a dominant component of the total variance

- Standardize each variable to remove scale effect

- Standardization is usually performed in PCA; otherwise measurement units affect results

First Two Principal Components

Using (standardized) cereals dataset and 13 different attributes



One can see that PC1 measures the balance between: (1) calories and cups vs. (2) protein, fiber, potassium, and consumer rating. PC2 is most affected by the weight of a serving.

Ideally, one finds a way to “labeling” the principal components in a similar fashion to learn about the **structure of the data**.

PCA in Classification/Prediction

Sometimes one can consider using PCA as part of a predictive model development.

- Apply PCA to training data
- Decide how many PC's to use
- Use variable weights in those PC's with validation/new data
- This creates a new **reduced** set of predictors in validation/new data