Northeastern UNIVERSITY

Data Mining in Engineering

Dimension Reduction

Xuemin Jin

Why Dimension Reduction?

- It is likely that subsets of variables are highly correlated with each other.
- Highly correlated variables, or variables that are unrelated to the outcome of interest, can lead to overfitting.
- Large numbers of variables may cause computational problems.

Dimension Reduction Approaches:

- Incorporating domain knowledge to remove or combine categories.
- Using data summaries to detect information overlap between variables.
- Using data conversion techniques such as converting categorical variables into numerical variables.
- Employing automated reduction techniques (PCA…)
- Using data mining methods such as regression models and regression and classification trees.

Curse of Dimensionality

- The more variables ≠ the better model
- The more variables => the more sparse the data space
  - Chessboard has 64 squares (two dimensions)
  - Increase the dimension by 50% (three dimension), 512 cubes, a 800% increase.
  - Too much noise is added, more difficult to discern patterns and structures
- A key step is to reduce dimensionality with minimal sacrifice of accuracy. In AI, factor selection (extraction)

Using Domain Knowledge

- The first step: To make sure that the variables measured are reasonable for the task.
- Ask questions like:
  - Which variables are most important (or useless) for the task at hand?
  - Which variables are likely to contain much error?
  - Which variables will be available for measurement?...

Data Exploration: Summary Statistics

- Average
- Standard deviation
- Min
- Max
- Median
- Count…

Summary Statistics: Boston Housing

Right skewed

	Average	Median	Min	Max	Std	Count	Countblank
CRIM	3.61	0.26	0.01	88.98	8.60	506	0
ZN	11.36	0.00	0.00	100.00	23.32	506	0
INDUS	11.14	9.69	0.46	27.74	6.86	506	0
CHAS	0.07	0.00	0.00	1.00	0.25	506	0
NOX	0.55	0.54	0.39	0.87	0.12	506	0
RM	6.28	6.21	3.56	8.78	0.70	506	0
AGE	68.57	77.50	2.90	100.00	28.15	506	0
DIS	3.80	3.21	1.13	12.13	2.11	506	0
RAD	9.55	5.00	1.00	24.00	8.71	506	0
TAX	408.24	330.00	187.00	711.00	168.54	506	0
PTRATIO	18.46	19.05	12.60	22.00	2.16	506	0
B	356.67	391.44	0.32	396.90	91.29	506	0
LSTAT	12.65	11.36	1.73	37.97	7.14	506	0
MEDV	22.53	21.20	5.00	50.00	9.20	506	0

Correlations between Pairs of Variables

	PTRATIO	*B*	*LSTAT*	*MEDV*
PTRATIO	1			
B	-0.17738	1		
LSTAT	0.374044	-0.36609	1	
MEDV	-0.50779	0.333461	-0.73766	1

Summarize Using Pivot Tables

- Count and percentage are useful for summarizing categorical data.
- Boston housing example:
  - *471 neighborhoods border the Charles River (1)*
  - *35 neighborhoods do not (0)*

Count of MEDV	
CHAS	Total
0	471
1	35
Grand Total	506

Pivot Tables – cont.

- Averages are useful for summarizing grouped numerical data
- Boston housing example:
  - *Compare average home values in neighborhoods that border Charles River (1) and those that do not (0)*

Average of MEDV	
CHAS	Total
0	22.09
1	28.44
Grand Total	22.53

Pivot Tables – cont.

- Used for multiple variables.
- Boston housing example:
  - By # rooms and location
  - E.g., neighborhoods on the Charles with 6-7 rooms have average house value of 25.92 ($000)

Average of MEDV	CHAS		
RM	0	1	Grand Total
3-4	25.30		25.30
4-5	16.02		16.02
5-6	17.13	22.22	17.49
6-7	21.77	25.92	22.02
7-8	35.96	44.07	36.92
8-9	45.70	35.95	44.20
Grand Total	22.09	28.44	22.53

Correlation Analysis

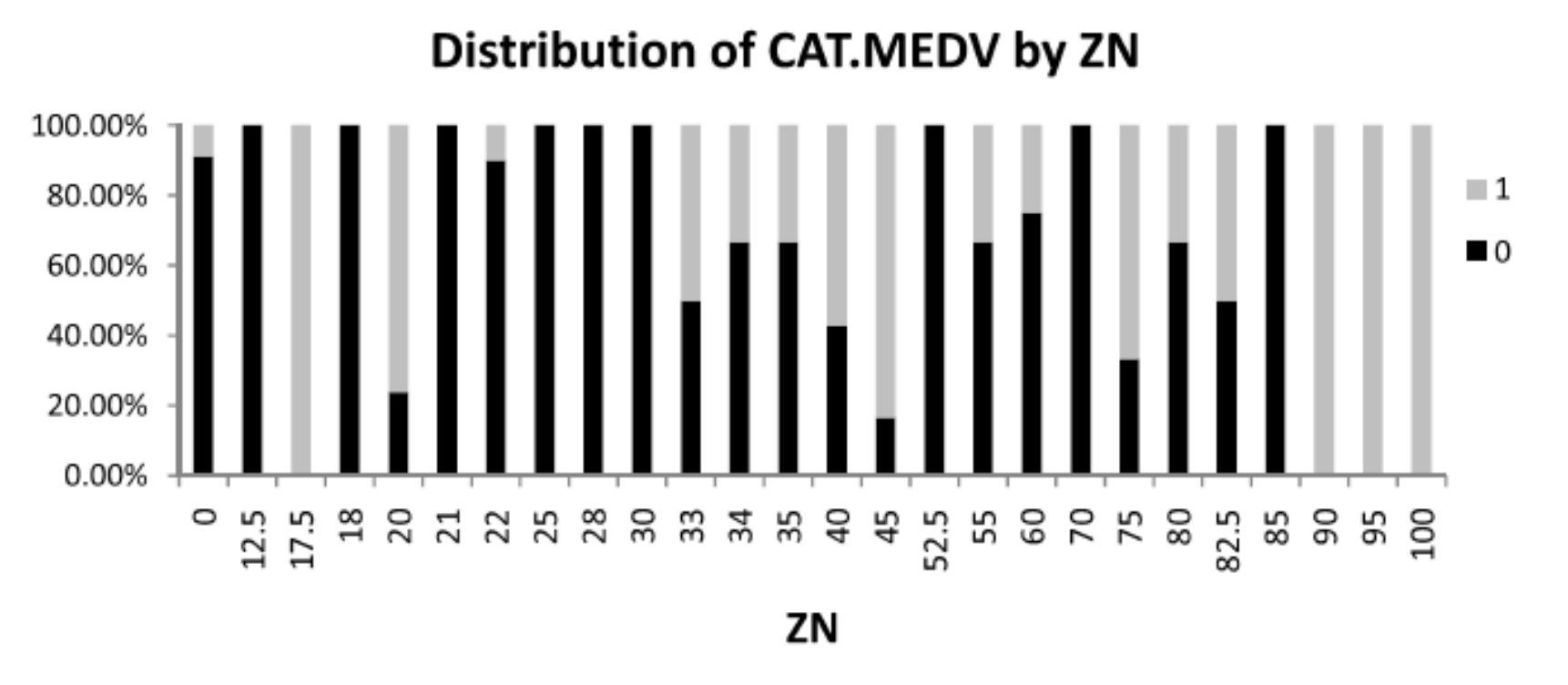
- Shows all the pairwise correlations between variables.
- Pairs that have a very strong correlation contain a lot of overlap in information.
- Removing variables that are strongly correlated to others is useful for avoiding **multicollinearity** problems.
- Correlation analysis is good for detecting duplications of variables.

	CRIM	*ZN*	*INDUS*	*CHAS*	*NOX*	*RM*
CRIM	1					
ZN	-0.20047	1				
INDUS	0.406583	-0.53383	1			
CHAS	-0.05589	-0.0427	0.062938	1		
NOX	0.420972	-0.5166	0.763651	0.091203	1	
RM	-0.21925	0.311991	-0.39168	0.091251	-0.30219	1

Reducing Categories

- A single categorical variable with *m* categories is typically transformed into *m-1* **dummy variables**.
- Each dummy variable takes the values 0 or 1.
  - *0 = "no" for the category*
  - *1 = "yes"*
- Problem: Can end up with too many variables
- Solution: Reduce by **combining categories** that are close to each other
- Use pivot tables to assess outcome variable sensitivity to the dummies
- Exception: Naïve Bayes can handle categorical variables without transforming them into dummies

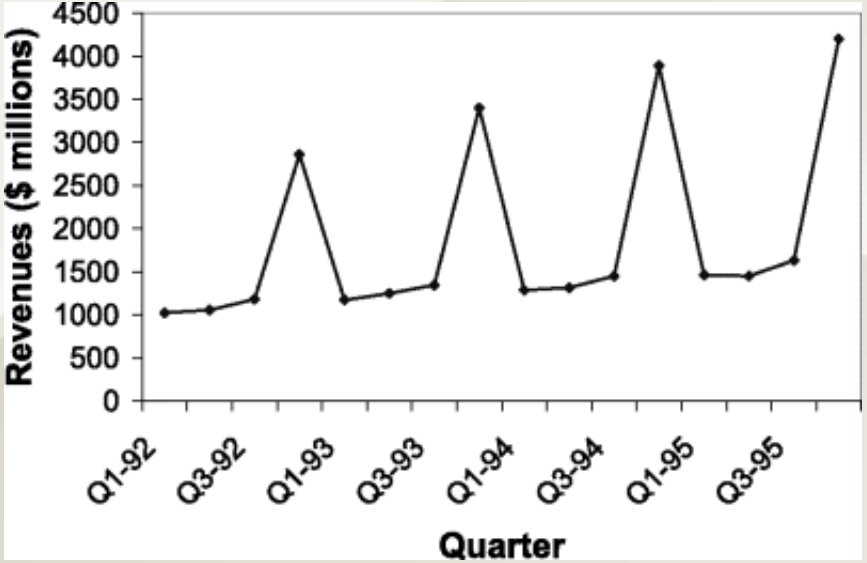
Combining Categories



- Many zoning categories are the same or similar with respect to CATMEDV

Combining Categories

- In a time series context categorical variable denote season (e.g., month or hour of day) that will serve as a predictor.
- Categories can be reduced by examining the time series plot and identifying similar periods.
- For example, the time plot shows the quarterly revenues of Toys "R" Us between 1992 and 1995. Only quarter 4 periods appear different, and therefore we can combine quarters 1–3 into a single category.



Converting Categorical Variable to Numerical

When categories in a categorical variable represent intervals.

E.g.:

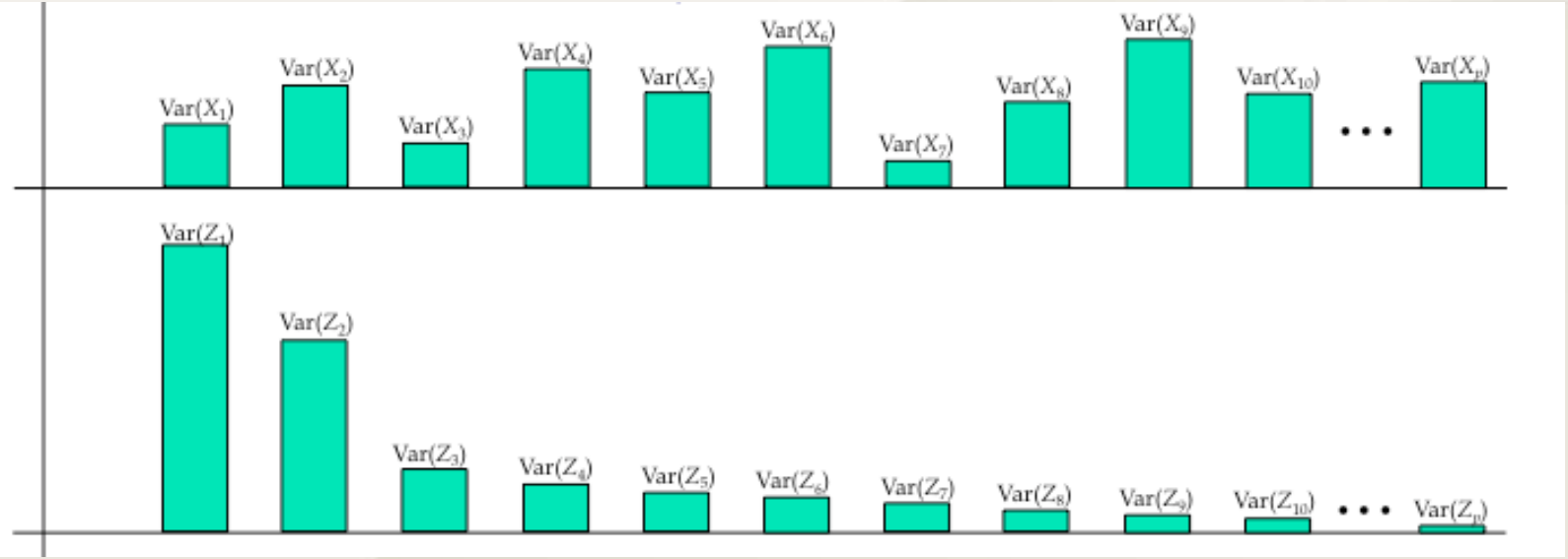
- Category 2 is the age interval 20-30
- Replace the categorical value (“2”) with the midinterval value (“25”)

Principal Components Analysis (PCA)

- PCA is a useful procedure for reducing the number of predictors in the model by analyzing the input variables.
- **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.

Principal Components Analysis

- PCA transforms original variables $X_1, X_2, \cdots, X_p$ to principal components $Z_1, Z_2, \cdots, Z_p$.
- $\text{Var}(X_1)+\text{Var}(X_2)+\cdots+\text{Var}(X_p) = \text{Var}(Z_1)+\text{Var}(Z_2)+\cdots+\text{Var}(Z_p)$
- $\text{Var}(Z_1) \geq \text{Var}(Z_2) \geq \cdots \geq \text{Var}(Z_p)$



Principal Components Analysis (PCA)

$$z_j(i) = a_{j1}(x_1(i) - \bar{X}_1) + a_{j2}(x_2(i) - \bar{X}_2) + \cdots + a_{jp}(x_p(i) - \bar{X}_p)$$

- $z_j(i)$ is the jth principal component score of the ith case
- $x_k(i)$ is the kth attribute of the ith case in the data
- a_{jk} is the weight for kth attribute to compute jth principal component
- $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, p$; $k = 1, 2, \ldots, p$

	X_1	X_2	…	X_k	…	X_p
1	$x_1(1)$	$x_2(1)$	…	$x_k(1)$	…	$x_p(1)$
2	$x_1(2)$	$x_2(2)$	…	$x_k(2)$	…	$x_p(2)$
…	…	…	…	…	…	…
i	$x_1(i)$	$x_2(i)$	…	$x_k(i)$	…	$x_p(i)$
…	…	…	…	…	…	…
n	$x_1(n)$	$x_2(n)$	…	$x_k(n)$	…	$x_p(n)$
	$\bar{X}_1$	$\bar{X}_2$		$\bar{X}_k$		$\bar{X}_p$

	Z_1	Z_2	…	Z_j	…	Z_p
1	$z_1(1)$	$z_2(1)$	…	$z_j(1)$	…	$z_p(1)$
2	$z_1(2)$	$z_2(2)$	…	$z_j(2)$	…	$z_p(2)$
…	…	…	…	…	…	…
i	$z_1(i)$	$z_2(i)$	…	$z_j(i)$	…	$z_p(i)$
…	…	…	…	…	…	…
n	$z_1(n)$	$z_2(n)$	…	$z_j(n)$	…	$z_p(n)$

Principal Components Analysis (PCA)

$$z_j(i) = a_{j1}(x_1(i) - \bar{X}_1) + a_{j2}(x_2(i) - \bar{X}_2) + \cdots + a_{jp}(x_p(i) - \bar{X}_p)$$

	X_1	X_2	...	X_k	...	X_p
1	$x_1(1)$	$x_2(1)$	...	$x_k(1)$	...	$x_p(1)$
2	$x_1(2)$	$x_2(2)$	...	$x_k(2)$	...	$x_p(2)$
...	...	...	...	...	...	...
i	$x_1(i)$	$x_2(i)$	...	$x_k(i)$	...	$x_p(i)$
...	...	...	...	...	...	...
n	$x_1(n)$	$x_2(n)$	...	$x_k(n)$	...	$x_p(n)$
	$\bar{X}_1$	$\bar{X}_2$		$\bar{X}_k$		$\bar{X}_p$

$$\begin{bmatrix} a_{11} & a_{21} & \cdots & a_{j1} & \cdots & a_{p1} \\ a_{12} & a_{22} & \cdots & a_{j2} & \cdots & a_{p2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{1k} & a_{2k} & \cdots & a_{jk} & \cdots & a_{pk} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ a_{1p} & a_{2p} & \cdots & a_{jp} & \cdots & a_{pp} \end{bmatrix}$$

	Z_1	Z_2	...	Z_j	...	Z_p
1	$z_1(1)$	$z_2(1)$	...	$z_j(1)$	...	$z_p(1)$
2	$z_1(2)$	$z_2(2)$	...	$z_j(2)$	...	$z_p(2)$
...	...	...	...	...	...	...
i	$z_1(i)$	$x_2(i)$	...	$z_j(i)$	...	$z_p(i)$
...	...	...	...	...	...	...
n	$z_1(n)$	$z_2(n)$	...	$z_j(n)$	...	$z_p(n)$

$$\mathrm{Var}(Z_1) \geq \mathrm{Var}(Z_2) \geq \cdots \mathrm{Var}(Z_j) \geq \cdots \mathrm{Var}(Z_p)$$

- PCA weights a_{jk} are computed using eigenvalue and eigenvector methods

Principal Components Analysis (PCA)

	X_1	X_2	…	X_k	…	X_p
1	$x_1(1)$	$x_2(1)$	…	$x_k(1)$	…	$x_p(1)$
2	$x_1(2)$	$x_2(2)$	…	$x_k(2)$	…	$x_p(2)$
…	…	…	…	…	…	…
i	$x_1(i)$	$x_2(i)$	…	$x_k(i)$	…	$x_p(i)$
…	…	…	…	…	…	…
n	$x_1(n)$	$x_2(n)$	…	$x_k(n)$	…	$x_p(n)$
	$\bar{X}_1$	$\bar{X}_2$		$\bar{X}_k$		$\bar{X}_p$

$$\begin{bmatrix} a_{11} & a_{21} & \cdots & a_{j1} & \cdots & a_{p1} \\ a_{12} & a_{22} & \cdots & a_{j2} & \cdots & a_{p2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{1k} & a_{2k} & \cdots & a_{jk} & \cdots & a_{pk} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ a_{1p} & a_{2p} & \cdots & a_{jp} & \cdots & a_{pp} \end{bmatrix}$$

	Z_1	Z_2	…	Z_j	…	Z_p
1	$z_1(1)$	$z_2(1)$	…	$z_j(1)$	…	$z_p(1)$
2	$z_1(2)$	$z_2(2)$	…	$z_j(2)$	…	$z_p(2)$
…	…	…	…	…	…	…
i	$z_1(i)$	$x_2(i)$	…	$z_j(i)$	…	$z_p(i)$
…	…	…	…	…	…	…
n	$z_1(n)$	$z_2(n)$	…	$z_j(n)$	…	$z_p(n)$

$$\text{Var}(Z_1) \geq \text{Var}(Z_2) \geq \cdots \text{Var}(Z_j) \geq \cdots \text{Var}(Z_p)$$

$$\text{Percentage of varience captured in } m \text{ principal components} = \left[\frac{\sum_{j=1}^{m} Var[Z_j]}{\sum_{i=1}^{p} Var[X_i]} \right] (100)$$

Example: Breakfast Cereals

Cereal Name	mfr	type	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins
100% Bran	N	C	70	4	1	130	10	5	6	280	25
100% Natural Bran	Q	C	120	3	5	15	2	8	8	135	0
All-Bran	K	C	70	4	1	260	9	7	5	320	25
All-Bran with Extra Fiber	K	C	50	4	0	140	14	8	0	330	25
Almond Delight	R	C	110	2	2	200	1	14	8		25
Apple Cinnamon Cheerios	G	C	110	2	2	180	1.5	10.5	10	70	25
Apple Jacks	K	C	110	2	0	125	1	11	14	30	25
Basic 4	G	C	130	3	2	210	2	18	8	100	25
Bran Chex	R	C	90	2	1	200	4	15	6	125	25
Bran Flakes	P	C	90	3	0	210	5	13	5	190	25
Cap'n'Crunch	Q	C	120	1	2	220	0	12	12	35	25
Cheerios	G	C	110	6	2	290	2	17	1	105	25
Cinnamon Toast Crunch	G	C	120	1	3	210	0	13	9	45	25
Clusters	G	C	110	3	2	140	2	13	7	105	25
Cocoa Puffs	G	C	110	1	1	180	0	12	13	55	25
Corn Chex	R	C	110	2	0	280	0	22	3	25	25
Corn Flakes	K	C	100	2	0	290	1	21	2	35	25
Corn Pops	K	C	110	1	0	90	1	13	12	20	25
Count Chocula	G	C	110	1	1	180	0	12	13	65	25
Cracklin' Oat Bran	K	C	110	3	3	140	4	10	7	160	25

Description of Variables

Variable	Description
mfr	Manufacturer of cereal (American Home Food Products, General Mills, Kellogg, etc.)
type	Cold or hot
calories	Calories per serving
protein	Grams of protein
fat	Grams of fat
sodium	Milligrams of sodium
fiber	Grams of dietary fiber
carbo	Grams of complex carbohydrates
sugars	Grams of sugars
potass	Milligrams of potassium
vitamins	Vitamins and minerals: 0, 25, or 100, indicating the typical percentage of FDA recommended
shelf	Display shelf (1, 2, or 3, counting from the floor)
weight	Weight in ounces of one serving
cups	Number of cups in one serving
rating	Rating of the cereal calculated by *Consumer Reports*

Calories and **Rating** for 77 cereals in the data set

Cereal	Calories	Rating
100% Bran	70	68.40297
100% Natural Bran	120	33.98368
All-Bran	70	59.42551
All-Bran with Extra Fiber	50	93.70491
Almond Delight	110	34.38484
Apple Cinnamon Cheerios	110	29.50954
Apple Jacks	110	33.17409
Basic 4	130	37.03856
Bran Chex	90	49.12025
Bran Flakes	90	53.31381
Cap'n'Crunch	120	18.04285
Cheerios	110	50.765
Cinnamon Toast Crunch	120	19.82357
Clusters	110	40.40021
Cocoa Puffs	110	22.73645
Corn Chex	110	41.44502
Corn Flakes	100	45.86332
Corn Pops	110	35.78279
Count Chocula	110	22.39651
Cracklin' Oat Bran	110	40.44877
Cream of Wheat (Quick)	100	64.53382
Crispix	110	46.89564
Crispy Wheat & Raisins	100	36.1762
Double Chex	100	44.33086
Froot Loops	110	32.20758
Frosted Flakes	110	31.43597
Frosted Mini-Wheats	100	58.34514
Fruit & Fibre Dates, Walnuts & Oats	120	40.91705
Fruitful Bran	120	41.01549
Fruity Pebbles	110	28.02577
Golden Crisp	100	35.25244
Golden Grahams	110	23.80404
Grape Nuts Flakes	100	52.0769
Grape-Nuts	110	53.37101
Great Grains Pecan	120	45.81172
Honey Graham Ohs	120	21.87129
Honey Nut Cheerios	110	31.07222
Honey-comb	110	28.74241
Just Right Crunchy Nuggets	110	36.52368
Just Right Fruit & Nut	140	36.471512
Kix	110	39.241114
Life	100	45.328074
Lucky Charms	110	26.734515
Maypo	100	54.850917
Muesli Raisins, Dates & Almonds	150	37.136863
Muesli Raisins, Peaches & Pecans	150	34.139765
Mueslix Crispy Blend	160	30.313351
Multi-Grain Cheerios	100	40.105965
Nut&Honey Crunch	120	29.924285
Nutri-Grain Almond-Raisin	140	40.69232
Nutri-grain Wheat	90	59.642837
Oatmeal Raisin Crisp	130	30.450843
Post Nat. Raisin Bran	120	37.840594
Product 19	100	41.50354
Puffed Rice	50	60.756112
Puffed Wheat	50	63.005645
Quaker Oat Squares	100	49.511874
Quaker Oatmeal	100	50.828392
Raisin Bran	120	39.259197
Raisin Nut Bran	100	39.7034
Raisin Squares	90	55.333142
Rice Chex	110	41.998933
Rice Krispies	110	40.560159
Shredded Wheat	80	68.235885
Shredded Wheat 'n'Bran	90	74.472949
Shredded Wheat spoon size	90	72.801787
Smacks	110	31.230054
Special K	110	53.131324
Strawberry Fruit Wheats	90	59.363993
Total Corn Flakes	110	38.839746
Total Raisin Bran	140	28.592785
Total Whole Grain	100	46.658844
Triples	110	39.106174
Trix	110	27.753301
Wheat Chex	100	49.787445
Wheaties	100	51.592193
Wheaties Honey Gold	110	36.187559

Consider Calories & Ratings

- Covariance matrix:

	Calories	Ratings
Calories	379.63	-189.68
Ratings	-189.68	197.32

- Correlation:

$$\frac{-189.68}{\sqrt{(379.63)(197.32)}} = -0.69$$

- Or, 69% variation in one variable is duplicated by similar variation in the other.

Consider Calories & Ratings

- **Covariance matrix:**

	calories	ratings
calories	379.63	-189.68
ratings	-189.68	197.32

- **Total variance (information)= sum of individual variances:**

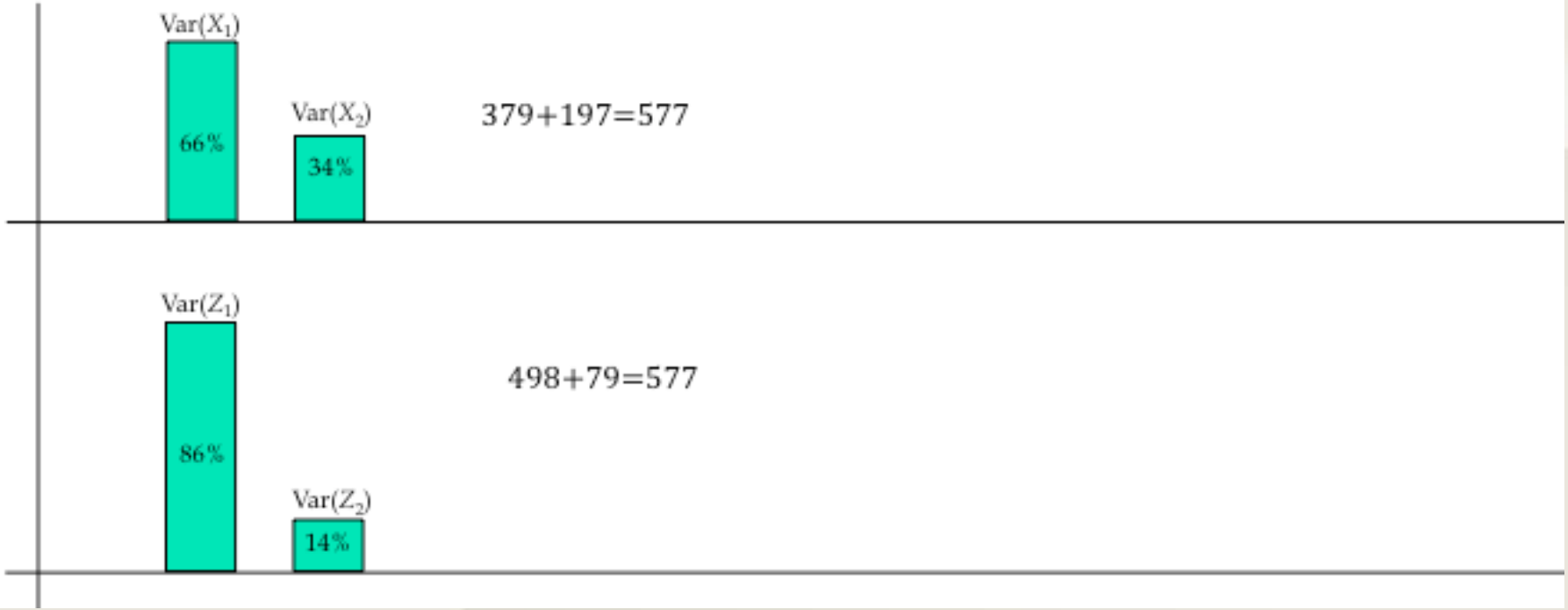
379.63+197.32=577

- **Calories accounts for**

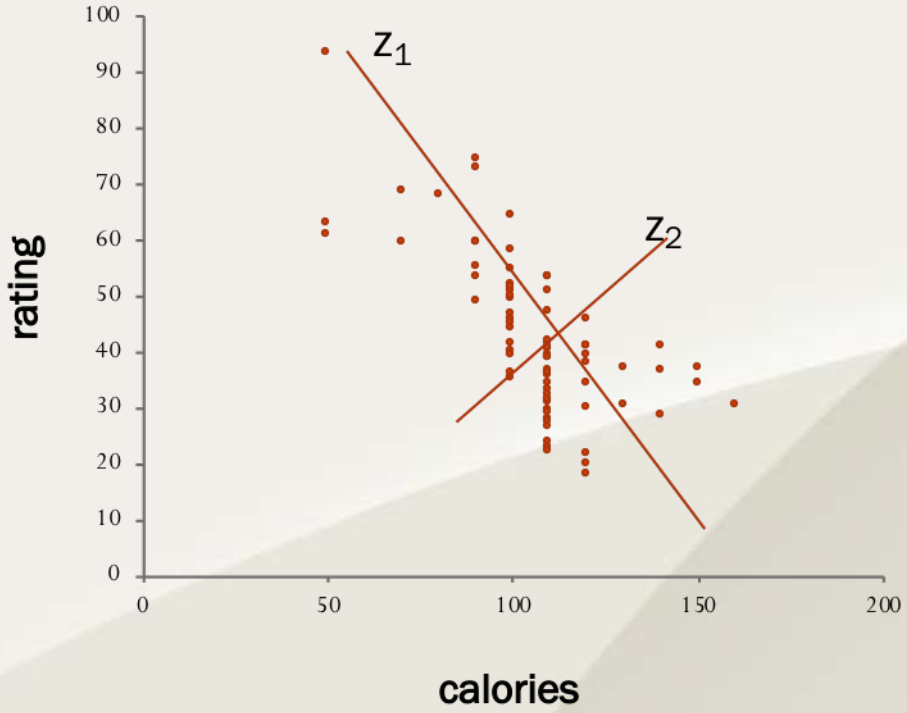
66%=379.63/577

Principal Components Analysis (PCA)

- X1 = Calories
- X2 = Rating



Principal Components



- Z_1 & Z_2 are two linear combinations.
- Z_1, or first principal component, is the direction in which the variability of the points is largest.
- Z_2, or second principal component, is perpendicular to Z_1 with second largest variability.

PCA Output: PCA Weights

- Top: **weights** a_{jk} to project original data onto Z_1 and Z_2

- a_{jk} is the weight of jth principle component for the kth variable, where j =1, 2, …, p and k=1, 2, …,p

a_{11} = -0.84705
a_{12} = 0.53151
a_{21} = 0.53151
a_{22} = 0.84705

	Components	
Variable	**1**	**2**
calories	-0.84705347	0.53150767
rating	0.53150767	0.84705347
Variance	498.0244751	78.932724
Variance%	86.31913757	13.68086338
Cum%	86.31913757	100
P-value	0	1

- Bottom: **reallocated variance** for new variables

Z_1 : 86% of total variance
Z_2 : 14% of total variance

Principal Component Scores

- Weights are used to compute principal component scores
- e.g.: Col. 1 scores are computed Z_1 scores using weights (-0.84705, 0.5315)
- For 100% Bran (with 70 calories and a rating of 68.4), Z1 score is

$$(-0.84705)(70-106.88)+(0.5315)(68.4-42.67)=44.92$$

subtracting the means

Row Id.	1	2
100% Bran	44.92152786	2.19717932
100% Natural Bran	-15.7252636	-0.38241446
All-Bran	40.14993668	-5.40721178
All-Bran with Extra Fiber	75.31076813	12.99912071
Almond Delight	-7.04150867	-5.35768652
Apple Cinnamon Cheerios	-9.63276863	-9.48732758
Apple Jacks	-7.68502998	-6.38325357
Basic 4	-22.57210541	7.52030993
Bran Chex	17.7315464	-3.50615811
Bran Flakes	19.96045494	0.04600986
Cap'n'Crunch	-24.19793701	-13.88514996
Cheerios	1.66467071	8.5171833
Cinnamon Toast Crunch	-23.25147057	-12.37678337
Clusters	-3.84429598	-0.26235023
Cocoa Puffs	-13.23272038	-15.2244997
Corn Chex	-3.28897071	0.62266076
Corn Flakes	7.5299263	-0.94987571

Properties of the Resulting Variables

New distribution of information:

- New variances = 498 (for Z_1) and 79 (for Z_2)
- Sum of variances for Z_1 and Z_2 = sum of variances for original variables *calories* and *ratings*
- New variable Z_1 has most of the total variance, might be used as proxy for both *calories* and *ratings*
- Z_1 and Z_2 have correlation of zero (no information overlap)

Generalization to p Variables ($X_1, X_2, \ldots, X_p$)

$X_1, X_2, X_3, \ldots X_p$, original p variables

$Z_1, Z_2, Z_3, \ldots Z_p$, principal components, which are weighted averages of original variables (after subtracting means)

a_{jk} is the weight of jth principle component for the kth variable, where $j = 1, 2, \ldots, p$, $k = 1, 2, \ldots, p$, and $i = 1, 2, \ldots, n$

$$z_j(i) = a_{j1}\left(x_1(i) - \bar{X}_1\right) + a_{j2}\left(x_2(i) - \bar{X}_2\right) + \cdots + a_{jk}\left(x_k(i) - \bar{X}_k\right) + \cdots + a_{jp}\left(x_p(i) - \bar{X}_p\right)$$

Generalization to p Variables ($X_1, X_2, \ldots, X_p$)

$$z_j(i) = a_{j1}\left(x_1(i) - \bar{X}_1\right) + a_{j2}\left(x_2(i) - \bar{X}_2\right) + \cdots + a_{jk}\left(x_k(i) - \bar{X}_k\right) + \cdots + a_{jp}\left(x_p(i) - \bar{X}_p\right)$$

- All pairs of principal components (Z_i and Z_j) variables have 0 correlation
- Order Z's by variance (var. of Z_1 largest, var. of Z_p smallest)
- Usually the first few Z_i variables contain most of the information, and so the rest can be dropped.

PCA Computed on a Full Dataset (without Normalization)

Variable	1	2	3	4	5	6
calories	0.07624155	-0.01066097	0.61074823	-0.61706442	0.45754826	0.12601775
protein	-0.00146212	0.00873588	0.00050506	0.0019389	0.05533375	0.10379469
fat	-0.00013779	0.00271266	0.01596125	-0.02595884	-0.01839438	-0.12500292
sodium	0.98165619	0.12513085	-0.14073193	-0.00293341	0.01588042	0.02245871
fiber	-0.00479783	0.03077993	-0.01684542	0.02145976	0.00872434	0.271184
carbo	0.01486445	-0.01731863	0.01272501	0.02175146	0.35580006	-0.56089228
sugars	0.00398314	-0.00013545	0.09870714	-0.11555841	-0.29906386	0.62323487
potass	-0.119053	0.98861349	0.03619435	-0.042696	-0.04644227	-0.05091622
vitamins	0.10149482	0.01598651	0.7074821	0.69835609	-0.02556211	0.01341988
shelf	-0.00093911	0.00443601	0.01267395	0.00574066	-0.00823057	-0.05412053
weight	0.0005016	0.00098829	0.00369807	-0.0026621	0.00318591	0.00817035
cups	0.00047302	-0.00160279	0.00060208	0.00095916	0.00280366	-0.01087413
rating	-0.07615706	0.07254035	-0.30776858	0.33866307	0.75365263	0.41805118

Variance	7204.161133	4833.050293	498.4260864	357.2174377	72.47863007	4.33980322
Variance%	55.52834702	37.25226212	3.84177661	2.75336623	0.55865192	0.0334504
Cum%	55.52834702	92.78060913	96.62238312	99.37575531	99.93440247	99.96785736

- First 6 principal components are shown.
- First 2 capture 93% of the total variation

First Two PCA Scores for 17 Cereals

Row Id.	1	2
100% Bran	44.92152786	2.19717932
100% Natural Bran	-15.7252636	-0.38241446
All-Bran	40.14993668	-5.40721178
All-Bran with Extra Fiber	75.31076813	12.99912071
Almond Delight	-7.04150867	-5.35768652
Apple Cinnamon Cheerios	-9.63276863	-9.48732758
Apple Jacks	-7.68502998	-6.38325357
Basic 4	-22.57210541	7.52030993
Bran Chex	17.7315464	-3.50615811
Bran Flakes	19.96045494	0.04600986
Cap'n'Crunch	-24.19793701	-13.88514996
Cheerios	1.66467071	8.5171833
Cinnamon Toast Crunch	-23.25147057	-12.37678337
Clusters	-3.84429598	-0.26235023
Cocoa Puffs	-13.23272038	-15.2244997
Corn Chex	-3.28897071	0.62266076
Corn Flakes	7.5299263	-0.94987571

Standardizing (Normalizing) the Data

- In these results, sodium dominates first PC.
- Just 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
  - ***Subtract mean first and Divide by std. deviation***
- **Standardization** is usually performed in PCA; otherwise measurement units affect results
- Note that using the **correlation matrix** means that you are operating on the standardized data.

PCA Computed on a Full Dataset (with Normalization Variables)

Variable	1	2	3	4	5	6
calories	0.32422706	0.36006299	0.13210163	0.30780381	0.08924425	-0.20683768
protein	-0.30220962	0.16462311	0.2609871	0.43252215	0.14542894	0.15786675
fat	0.05846959	0.34051308	-0.21144024	0.37964511	0.44644874	0.40349057
sodium	0.20198308	0.12548573	0.37701431	-0.16090299	-0.33231756	0.6789462
fiber	-0.43971062	0.21760374	0.07857864	-0.10126047	-0.24595702	0.06016004
carbo	0.17192839	-0.18648526	0.56368077	0.20293142	0.12910619	-0.25979191
sugars	0.25019819	0.3434512	-0.34577203	-0.10401795	-0.27725372	-0.20437138
potass	-0.3834067	0.32790738	0.08459517	0.00463834	-0.16622125	0.022951
vitamins	0.13955688	0.16689315	0.38407779	-0.52358848	0.21541923	0.03514972
shelf	-0.13469705	0.27544045	0.01791886	-0.4340663	0.59693497	-0.12134896
weight	0.07780685	0.43545634	0.27536476	0.10600897	-0.26767638	-0.38367996
cups	0.27874646	-0.24295618	0.14065795	0.08945525	0.06306333	0.06609894
rating	-0.45326898	-0.22710647	0.18307236	0.06392702	0.03328028	-0.16606605
Variance	3.59530377	3.16411042	1.86585701	1.09171081	0.96962351	0.72342771
Variance%	27.65618324	24.3393116	14.35274601	8.39777565	7.45864248	5.5648284
Cum%	27.65618324	51.99549484	66.34824371	74.74601746	82.20465851	87.76948547

- First 6 principal components are shown.
- First 6 capture 90% of the total variation

PCA in Classification and Prediction

- 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

Disadvantages of PCA base Dimension Reduction for Prediction/Classification

- One disadvantage of using a subset of principal components as predictors in a supervised task is that
  - We might lose predictive information that is nonlinear (e.g., a quadratic effect of a predictor on the outcome or an interaction between predictors).
  - This is because PCA produces linear transformations, thereby capturing linear relationships between the original variables.

Regression-based Dimension Reduction

- Multiple Linear Regression or Logistic Regression
- Use subset selection
- Algorithm chooses a subset of variables
- This procedure is integrated directly into the predictive task

Summary

- **Data summarization** is an important method for data exploration
- **Data summaries** include numerical metrics (average, median, etc.) and graphical summaries
- **Data reduction** is useful for compressing the information in the data into a smaller subset
  - Categorical variables can be reduced by combining similar categories
  - Principal components analysis transforms an original set of numerical data into a smaller set of weighted averages of the original data that contain most of the original information in less variables.