STATISTICS 407

METHODS OF MULTIVARIATE ANALYSIS

TOPICS

- **Principal Component Analysis (PCA):** Reduce the ________________, summarize the sources of variation in the data, transform the data into a new data set where the variables are uncorrelated.
- **Factor Analysis (FA):** When its not possible to _________________ of interest directly, measure what's possible and create the variables of interest from the observed data.
- **Discriminant Analysis (classification, supervised learning):** Build a rule to _____________ or group id from observed training data.

TOPICS

- **Cluster Analysis (unsupervised learning):** Find similar groups of individuals, or ______________________ based on their similarity.
- **M(ultivariate)ANOVA:** Infer information about the ____________________ based on the sample means.
- **Multivariate Regression and Canonical Correlation Analysis:** ______________________ variables, and explore the association between the set of dependent variables and a set of explanatory variables.
- **PLUS** how to **plot** multivariate data, ____________.

A TAXONOMY OF TECHNIQUES

- **Variable-directed:** Quantifying the relationships between variables, eg ___________________________ ____ ____.

- **Individual-directed:** Summarizing relationships that exist between individuals, or experimental units, eg ____.

MULTIVARIATE DATA

- Example, nutritional information of chocolates (100g equivalent) from around the world:

Name	MFR	Country	Type	Calories	TotFat	Chol	Na	Fiber	Sugars
Fine Extra Dark	Ritter Sport	German	Dark	558	44.6	0.01	48.0	5.54	29.0
Classic Milk Chocolate Bar	Nestle	Switzerland	Milk	501	27.3	11.39	148.1	2.28	54.7
Rich Dark Chocolate Kisses	Hershey's	US	Dark	561	31.7	12.20	61.0	7.32	51.2
Dark Chocolate Bar	Choceur	Switzerland	Dark	558	39.5	34.88	34.9	4.65	39.5
Jet Milk Chocolate	Jet	Colombia	Milk	560	36.0	0.00	80.0	2.00	50.0
Dark Chocolate Bar	Guylian	Belgium	Dark	576	33.3	0.00	121.2	9.09	18.2
Rich Dark Chocolate Bar	Dove	US	Dark	515	32.5	13.55	0.0	8.13	46.1
Dark Chocolate Bar 86%	Poulain	France	Dark	640	50.0	0.00	8.0	10.00	16.0
Bliss Dark Chocolate	Hershey's	US	Dark	465	32.6	11.63	23.3	6.98	46.5
After Eight Dark Milk Chocolate Bar	Nestle	Switzerland	Dark	578	35.6	11.11	44.4	6.67	48.9

SOME MATH...

Matrix Notation

Data (n observations, p variables) has matrix form as follows:

$$\mathbf{X} = [\mathbf{X}_1 \; \mathbf{X}_2 \; \ldots \; \mathbf{X}_p]$$

$$= \begin{bmatrix} X_{11} & X_{12} & \ldots & X_{1p} \\ X_{21} & X_{22} & \ldots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \ldots & X_{np} \end{bmatrix}_{n \times p}$$

X_{ij} is the element in the i^{th} row and j^{th} column, that is i^{th} case and j^{th} variable.

SOME MATH...

Mean Vector, Variance-Covariance/Correlation Matrices

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{X}_1 \\ \vdots \\ \bar{X}_p \end{bmatrix}$$

$$\mathbf{S} = \begin{bmatrix} S_{11} & S_{12} & \dots & S_{1p} \\ S_{21} & S_{22} & \dots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \dots & S_{pp} \end{bmatrix}$$

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{bmatrix}$$

How do you calculate S_{11}, S_{12}, r_{12}?

MULTIVARIATE DATA

- For the chocolates data example, we'd calculate the mean vector, var-cov matrix, and correlation matrix for the _________________________.
- For the __________ variables we'd report counts, and proportions.
- The mean vector, var-cov and corr matrix might also be reported _____________ for each category of the categorical variables.

MULTIVARIATE DATA

- Chocolates data:
  n=10, p=6

$$\bar{\mathbf{X}} = \begin{bmatrix} 551.18 \\ 36.31 \\ 9.48 \\ 56.88 \\ 6.27 \\ 40.01 \end{bmatrix}$$

- What's the mean Calories?
- variance of Fiber?
- correlation between Sugars and Calories?
- Which variables have negative covariance?
- Which variable has the largest variance?
- Standard deviation of Chol?

$$\mathbf{S} = \begin{bmatrix} 2299.6 & 229.26 & -156.05 & -296.3 & 48.78 & -425.1 \\ 229.3 & 45.14 & -16.62 & -166.5 & 5.97 & -66.3 \\ -156.1 & -16.62 & 115.50 & -108.8 & -5.65 & 60.4 \\ -296.3 & -166.50 & -108.79 & 2275.2 & -61.30 & 90.8 \\ 48.8 & 5.97 & -5.65 & -61.3 & 7.16 & -23.6 \\ -425.1 & -66.28 & 60.36 & 90.8 & -23.60 & 196.9 \end{bmatrix}$$

$$\mathbf{R} = \begin{bmatrix} 1.000 & 0.712 & -0.303 & -0.130 & 0.380 & -0.632 \\ 0.712 & 1.000 & -0.230 & -0.520 & 0.332 & -0.703 \\ -0.303 & -0.230 & 1.000 & -0.212 & -0.197 & 0.400 \\ -0.130 & -0.520 & -0.212 & 1.000 & -0.480 & 0.136 \\ 0.380 & 0.332 & -0.197 & -0.480 & 1.000 & -0.629 \\ -0.632 & -0.703 & 0.400 & 0.136 & -0.629 & 1.000 \end{bmatrix}$$

DATA SUMMARY

Table 1: Summary statistics for nutritional information about chocolates from around the world, based on 10 observations.

	Calories	TotFat (g)	Cholesterol (mg)	Sodium (mg)	Fiber (g)	Sugars (g)
Mean	551.18	36.31	9.48	56.88	6.27	40.01
Std dev	47.95	6.72	10.75	47.7	2.68	14.03

It is also a good idea to include the minimum, maximum and median of each variable.

DATA SUMMARY

Table 2: Correlation between for nutrition variables collected on chocolates from around the world, based on 10 observations.

	Calories	TotFat	Cholesterol	Sodium	Fiber	Sugars
Calories	1	0.71	-0.3	-0.13	0.38	-0.63
TotFat	0.71	1	-0.23	-0.52	0.33	-0.7
Cholesterol	-0.3	-0.23	1	-0.21	-0.2	0.4
Sodium	-0.13	-0.52	-0.21	1	-0.48	0.14
Fiber	0.38	0.33	-0.2	-0.48	1	-0.63
Sugars	-0.63	-0.7	0.4	0.14	-0.63	1

MULTIVARIATE DATA

- Chocolates data - summary of categorical variables using counts. These are the ____________, or even the dependent variables that might be used for classifying observations.

```
        MFR           Country    Type
Hershey's:2   Switzerland:3   Dark:8
Nestle   :2   US         :3   Milk:2
Choceur  :1   Belgium    :1
Dove     :1   Colombia   :1
Guylian  :1   France     :1
Jet      :1   German     :1
```

MORE MATH...

- Linear combinations, and projections:

If

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{bmatrix}_{p\times 1}$$

then

$$\mathbf{X}\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 X_{11} + \alpha_2 X_{21} + \ldots + \alpha_p X_{p1} \\ \vdots \\ \alpha_1 X_{1n} + \alpha_2 X_{2n} + \ldots + \alpha_p X_{pn} \end{bmatrix}_{n\times 1}$$

If $\sqrt{\alpha_1^2 + \ldots + \alpha_p^2} = 1$ then $\boldsymbol{\alpha}$ is a projection vector, and $\mathbf{X}\boldsymbol{\alpha}$ is a projection of the data.

- Used in ____.

MORE MATH...

- Distance measures:

For two points (rows of the data matrix) $\mathbf{A} = (A_1\ A_2\ \ldots\ A_p)$ and $\mathbf{B} = (B_1\ B_2\ \ldots\ B_p)$, Euclidean distance is defined as

$$d(\mathbf{A},\mathbf{B}) = \sqrt{(\mathbf{A}-\mathbf{B})(\mathbf{A}-\mathbf{B})'} = \sqrt{(A_1-B_1)^2 + \ldots + (A_p-B_p)^2}$$

and *statistical distance* (or Mahalobis distance) is defined as

$$d(\mathbf{A},\mathbf{B}) = \sqrt{(\mathbf{A}-\mathbf{B})\mathbf{S}^{-1}(\mathbf{A}-\mathbf{B})'}.$$

Generally any distance measure can be defined, but it must satisfy (1) $d(\mathbf{A},\mathbf{B}) = d(\mathbf{B},\mathbf{A})$, (2) $d(\mathbf{A},\mathbf{B}) > 0$, if $\mathbf{A} \neq \mathbf{B}$, (3) $d(\mathbf{A},\mathbf{B}) = 0$, if $\mathbf{A} = \mathbf{B}$, (4) $d(\mathbf{A},\mathbf{B}) \leq d(\mathbf{A},\mathbf{C}) + d(\mathbf{C},\mathbf{B})$, for any intermediate point $\mathbf{C}$.

- Used in ____________________________.

MORE MATH...

- Scaling:

  The standardized data matrix is

$$\mathbf{Z} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & & z_{2p} \\ \vdots & \vdots & & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{bmatrix}$$

where $z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \quad i = 1, ..., n; j = 1, ..., p$. The standardized data has mean vector all zeros, and variances all equal to 1.

- Different from ____________! Doesn't change the correlation between variables. _____________ does remove correlation - more to come on this.