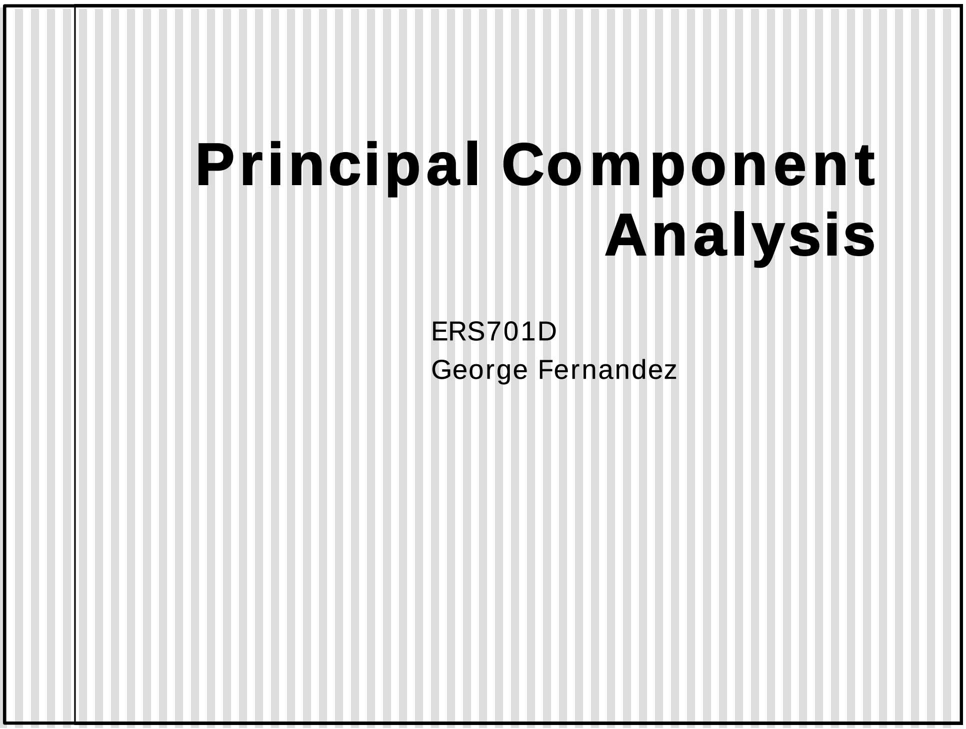
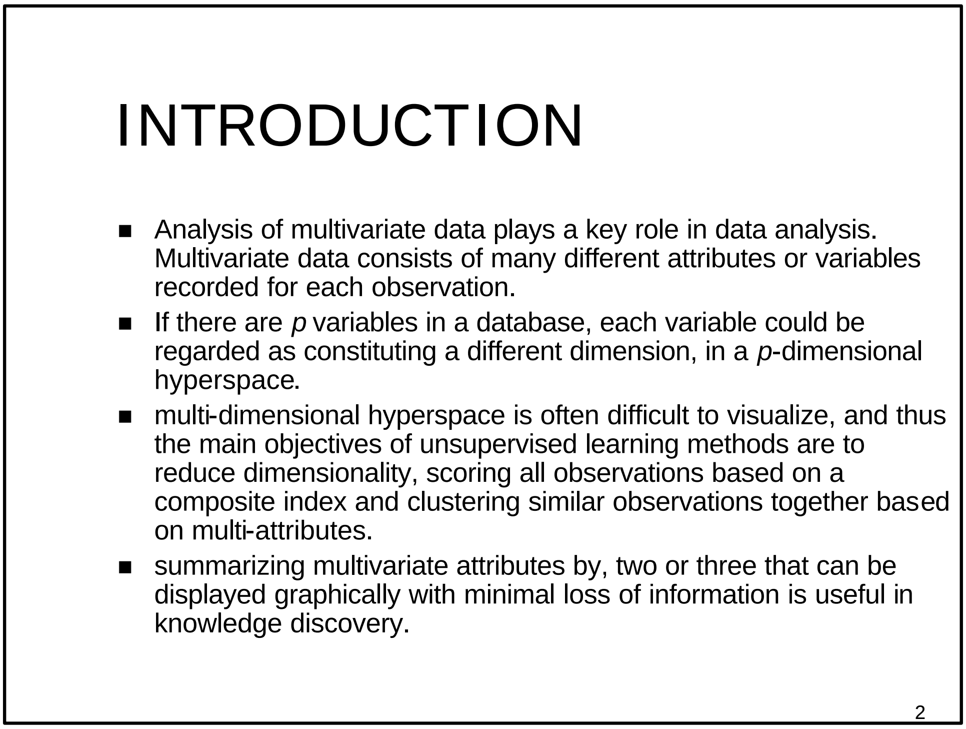


Principal Component Analysis

ERS701D
George Fernandez



INTRODUCTION

- Analysis of multivariate data plays a key role in data analysis. Multivariate data consists of many different attributes or variables recorded for each observation.
- If there are *p* variables in a database, each variable could be regarded as constituting a different dimension, in a *p*-dimensional hyperspace.
- multi-dimensional hyperspace is often difficult to visualize, and thus the main objectives of unsupervised learning methods are to reduce dimensionality, scoring all observations based on a composite index and clustering similar observations together based on multi-attributes.
- summarizing multivariate attributes by, two or three that can be displayed graphically with minimal loss of information is useful in knowledge discovery.

PRINCIPAL COMPONENT ANALYSIS

- **Because it is hard to visualize multi-dimensional space, principal components analysis (PCA), a popular multivariate technique, is mainly used to reduce the dimensionality of *p* multi-attributes to two or three dimensions.**
- **PCA summarizes the variation in a correlated multi-attribute to a set of uncorrelated components, each of which is a particular linear combination of the original variables.**
- **The extracted uncorrelated components are called principal components (PC) and are estimated from the eigenvectors of the covariance or correlation matrix of the original variables.**
- **Therefore, the objective of PCA is to achieve parsimony and reduce dimensionality by extracting the smallest number components that account for most of the variation in the original multivariate data and to summarize the data with little loss of information.**

PRINCIPAL COMPONENT ANALYSIS

- **In PCA, uncorrelated PC's are extracted by linear transformations of the original variables so that the first few PC's contain most of the variations in the original dataset.**
- **These PCs are extracted in decreasing order of importance so that the first PC accounts for as much of the variation as possible and each successive component accounts for a little less.**
- **Following PCA, analyst tries to interpret the first few principal components in terms of the original variables, and thereby have a greater understanding of the data.**
- **To reproduce the total system variability of the original *p* variables, we need all *p* PCs. However, if the first few PCs account for a large proportion of the variability (80-90%), we have achieved our objective of dimension reduction.**
- **Because the first principal component accounts for the co-variation shared by all attributes, this may be a better estimate than simple or weighted averages of the original variables. Thus, PCA can be useful when there is a severe high-degree of correlation present in the multi-attributes.**

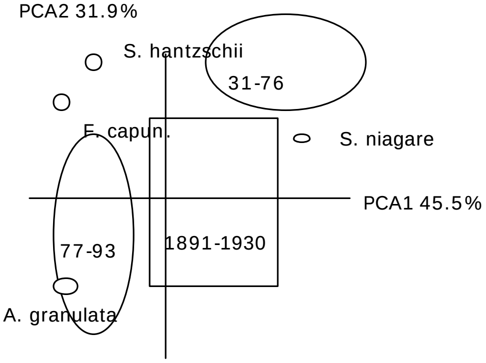
PRINCIPAL COMPONENT ANALYSIS

- In PCA, the extractions of PC can be made using either original multivariate datasets or using the covariance or the correlation matrix if the original dataset is not available.
- In deriving PC, the correlation matrix is commonly used when different variables in the dataset are measured using different units (annual income, educational level, numbers of cars owned per family) or if different variables have different variances.
- Using the correlation matrix is equivalent to standardizing the variables to zero mean and unit standard deviation.

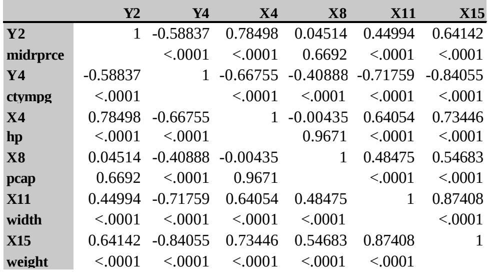
Applications of PCA analysis

- Hall R.I, Leavitt P.R,Quinlan R., Dixit A.S, Smol, J.P 1999 Effects of agriculture, urbanization, and climate on water quality in the northern Great plains. Limnol. Oceanogr. 44(3, part 2) 739-756
- PCA of sample scores and selected species from (a) diatom percent abundances, (b) fossil pigment concentrations, and Chironomid percent abundances.

Applications of PCA analysis



Correlation matrix



	Y2	Y4	X4	X8	X11	X15
Y2	1	-0.58837	0.78498	0.04514	0.44994	0.64142
midrprce		<.0001	<.0001	0.6692	<.0001	<.0001
Y4	-0.58837	1	-0.66755	-0.40888	-0.71759	-0.84055
ctympg	<.0001		<.0001	<.0001	<.0001	<.0001
X4	0.78498	-0.66755	1	-0.00435	0.64054	0.73446
hp	<.0001	<.0001		0.9671	<.0001	<.0001
X8	0.04514	-0.40888	-0.00435	1	0.48475	0.54683
pcap	0.6692	<.0001	0.9671		<.0001	<.0001
X11	0.44994	-0.71759	0.64054	0.48475	1	0.87408
width	<.0001	<.0001	<.0001	<.0001		<.0001
X15	0.64142	-0.84055	0.73446	0.54683	0.87408	1
weight	<.0001	<.0001	<.0001	<.0001	<.0001	

Eigen value / vector decomposition of a correlation matrics

$$\begin{bmatrix} 1 & r12 & r13 \\ r21 & 1 & r23 \\ r32 & r31 & 1 \end{bmatrix} = \sum_{j=1}^{p} l_j a_j a_{j'}$$

PCA TERMINOLOGY

- ***Eigenvalues*** measure the amount of the variation explained by each PC and will be largest for the first PC and smaller for the subsequent PCs.
  - An eigenvalue greater than 1 indicates that PCs account for more variance than accounted by one of the original variables in standardized data. This is commonly used as a cutoff point for which PCs are retained.
- ***Eigenvectors*** provides the weights to compute the uncorrelated PC, which are the linear combination of the centered standardized or centered un-standardized original variables.

Eigenvectors and eigen values

Eigenvectors			
		1	2
Y2	midrprce	0.3771	-0.442
Y4	ctympg	-0.4475	-0.0523
X4	hp	0.4181	-0.426
X8	pcap	0.2297	0.7522
X11	width	0.4394	0.1947
X15	weight	0.4867	0.1296

Eigenvalues of the Correlation Matrix: Total = 6				
	Eigenvalue	Difference	Proportion	Cumulative
1	3.9481323	2.7240979	0.658	0.658
2	1.2240344	0.8439396	0.204	0.862
3	0.3800949	0.1123044	0.0633	0.9254
4	0.2677905	0.1444804	0.0446	0.97
5	0.1233101	0.0666722	0.0206	0.9906
6	0.0566379		0.0094	1

PCA

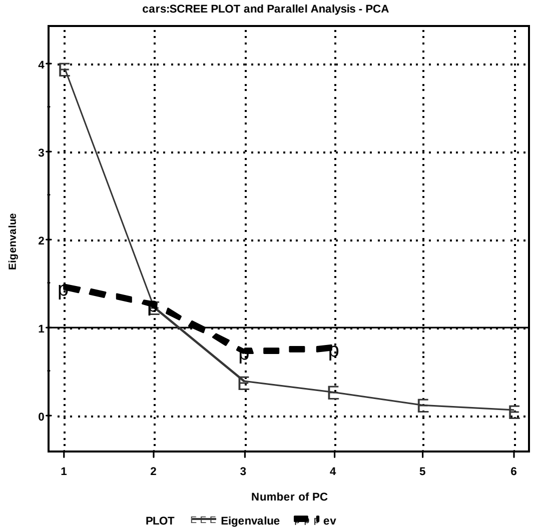
$$PC_1 = \sum_{1}^{p} a_{1j} X_j$$

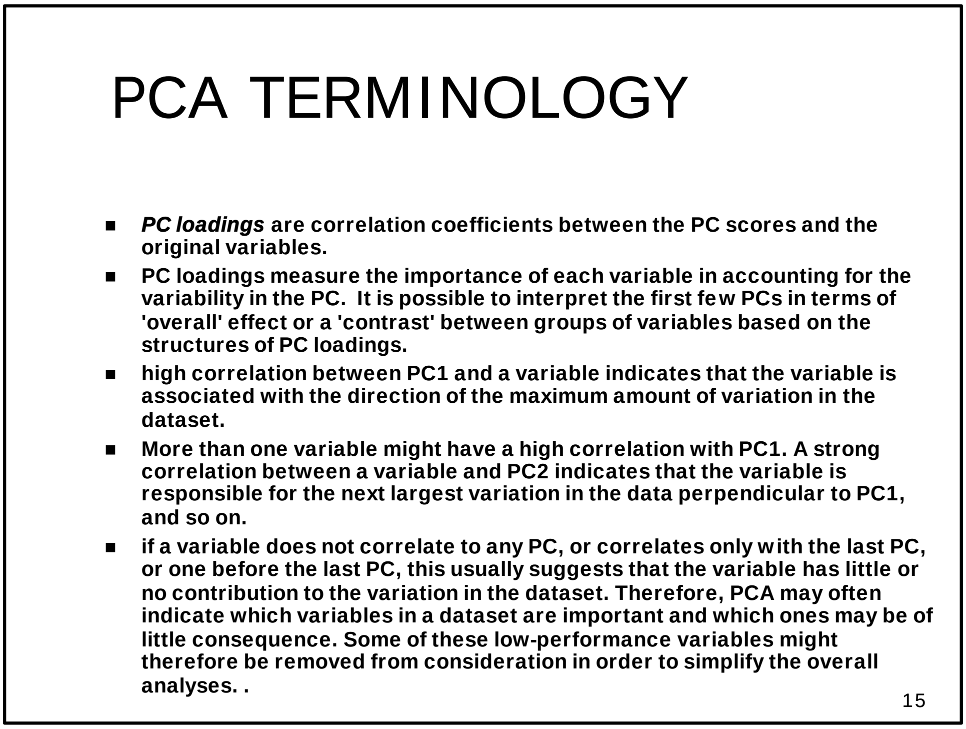
PC- principal compnent

a1j= Linear coefficient –eigen vectors

Estimating the Number of PC

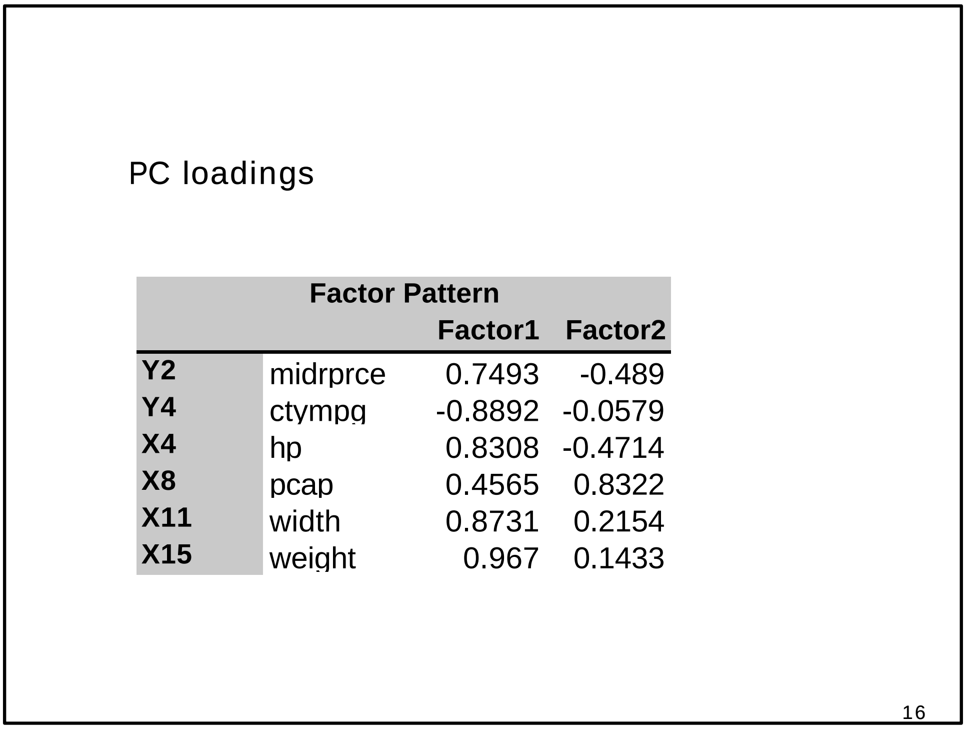
- ***Scree Test***: Plotting the eigenvalues against the corresponding PC produces a scree plot that illustrates the rate of change in the magnitude of the eigenvalues for the PC. The rate of decline tends to be fast first then levels off. The **'**elbow**'**, or the point at which the curve bends, is considered to indicate the maximum number of PC to extract. One less PC than the number at the elbow might be appropriate if you are concerned about getting an overly defined solution.





PCA TERMINOLOGY

- ***PC loadings*** **are correlation coefficients between the PC scores and the original variables.**
- **PC loadings measure the importance of each variable in accounting for the variability in the PC. It is possible to interpret the first few PCs in terms of 'overall' effect or a 'contrast' between groups of variables based on the structures of PC loadings.**
- **high correlation between PC1 and a variable indicates that the variable is associated with the direction of the maximum amount of variation in the dataset.**
- **More than one variable might have a high correlation with PC1. A strong correlation between a variable and PC2 indicates that the variable is responsible for the next largest variation in the data perpendicular to PC1, and so on.**
- **if a variable does not correlate to any PC, or correlates only with the last PC, or one before the last PC, this usually suggests that the variable has little or no contribution to the variation in the dataset. Therefore, PCA may often indicate which variables in a dataset are important and which ones may be of little consequence. Some of these low-performance variables might therefore be removed from consideration in order to simplify the overall analyses. .**



PC loadings

Factor Pattern			
		Factor1	**Factor2**
Y2	midrprce	0.7493	-0.489
Y4	ctympg	-0.8892	-0.0579
X4	hp	0.8308	-0.4714
X8	pcap	0.4565	0.8322
X11	width	0.8731	0.2154
X15	weight	0.967	0.1433

PC Scores

- ***PC scores*** **are the derived composite scores computed for each observation based on the eigenvectors for each PC.**
- **The means of PC scores are equal to zero, as these are the linear combination of the centered variables.**
- **These uncorrelated PC scores can be used in subsequent analyses, to check for multivariate normality, to detect multivariate outliers, or as a remedial measure in regression analysis with severe multi-colliniarity .**

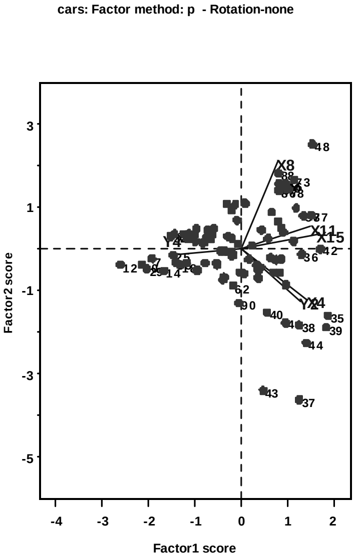
PC Scores

Obs	ID	Factor1	Factor2
1	39	-2.60832	-0.36804
2	82	-2.15032	-0.37408
3	81	-2.04321	-0.46095
4	32	-1.9308	-0.22925
5	42	-1.67883	-0.51745
6	85	-1.53146	0.31244
7	74	-1.46923	-0.13646
8	44	-1.4465	0.38442
9	87	-1.41928	-0.30308
10	47	-1.32023	-0.37438
11	23	-1.23447	0.37825
12	40	-1.1852	-0.31955
13	30	-1.16499	0.25891
14	62	-1.14394	0.38416
15	56	-1.03193	0.24003
16	80	-1.01518	0.35356
83	19	1.24816	-3.61769
84	28	1.25067	-1.81951
85	11	1.30762	-0.11547
86	31	1.34511	0.81154
87	57	1.40631	-2.25287
88	7	1.50488	0.81694
89	16	1.53423	2.52361
90	52	1.71018	0.00869
91	48	1.82993	-1.87121
92	10	1.87482	-1.58474

BI-PLOT DISPLAY OF PCA

- **Bi-plot** display is a visualization technique for investigating the inter-relationships between the observations and variables in multivariate data.
- To display a bi-plot, the data should be considered as a matrix, in which the column represents the variable space while the row represents the observational space.
- The term bi-plot means it is a plot of two dimensions with the observation and variable spaces plotted simultaneously.
- In PCA, relationships between PC scores and PCA loadings associated with any two PCs can be illustrated in a bi-plot display.

Bi-Plot display


cars: Factor method: p - Rotation-none
Factor2 score
Factor1 score
X8
X11
X15
Y4
Y2
X4