Introduction to Simple Correspondence Analysis

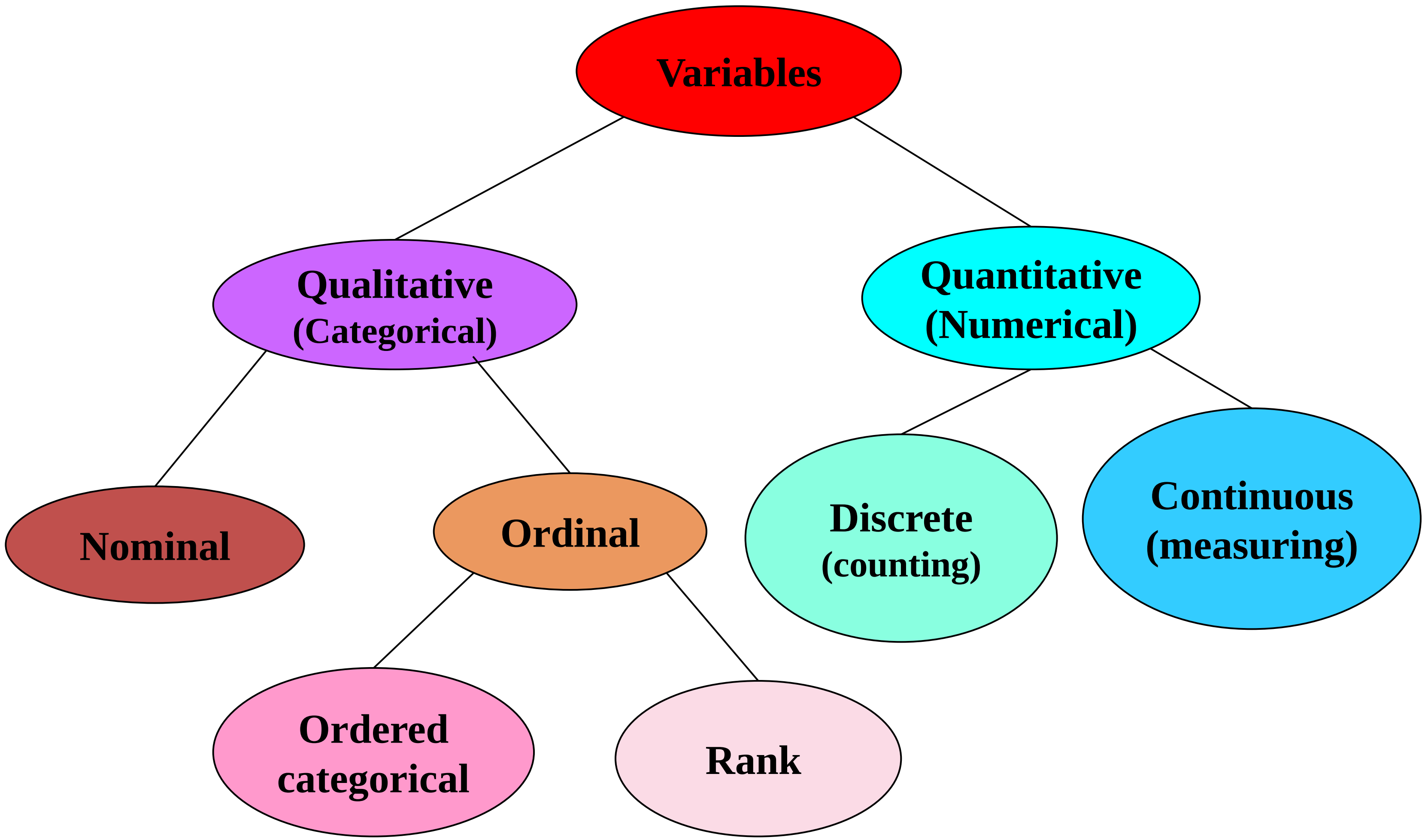
Prof. Kirtee K. Kamalja
Department of Statistics
School of Mathematical Sciences
K.B.C. North Maharashtra University, Jalgaon

Types of Variables

- Qualitative or Categorical Variables / Attributes
  - Nominal Variables
  - Ordinal Variables
- Quantitative Variables
  - Discrete Variables
  - Continuous Variables

2

Types of Variables



Categorical variables

Note: The topic is related to categorical data.

Categorical Variable: A categorical variable has a measurement scale consisting of a set of categories. That is, variables which record a response as a set of categories are termed categorical.

Nominal Variable	Ordinal Variable
• Do not have natural ordering • The order of listing the categories are irrelevant • Gender, religious affiliation, favorite type of music etc.	• Have natural ordering • Distances between categories are unknown • Economic status, patients condition etc.

Categorical variables

- **Nominal variables**: Variables having categories without a natural ordering are called nominal.
- **Example:**

1. Gender is a categorical variable having two categories (male and female) and there is no intrinsic ordering to the categories.
2. Hair color is also a categorical variable having a number of categories (blonde, brown, brunette, red, etc.)

Categorical variables...

Ordinal variables: If the categorical variables do have natural ordering, then that variable is called as an ordinal variable.

Example:

i) A variable, economic status, with three categories (low, medium and high),

ii) Social class (upper, middle, lower)

iii) Patient condition (Good, Fair, Serious, Critical).

Categorical data representation tool: Contingency Table

Contingency Table:

- The term first time is used by Karl Pearson (1904).
- Contingency tables show frequencies produced by cross-classifying observations according to categorical variables.
- It is used to represent and display the relationships between two or more categorical variables.

Cont…

- It is a type of table in a matrix format that displays the (multivariate) frequency ditribution (cross tabulation or cross tab) of the categorical variables.

Example: Contingency Table

Let's consider a real-life example related to **survey data** on people's exercise habits and whether they experience stress. The two variables are:

1. **Exercise habit**: Regular or Not Regular
2. **Stress level**: High or Low

	High Stress	Low Stress	Total
Regular Exercise	10	40	50
No Regular Exercise	30	20	50
Total	40	60	100

Example: Contingency Table

To study the relationship between hair colour and eye colour in a German population, an anthropologist observed a sample of 6,800 men, with the results shown.

Hair color and eye color

		Hair color			
		Brown	Black	Fair	Red
Eye	Brown	438	288	115	16
Color	Grey or Green	1,387	746	946	53
	Blue	807	189	1,768	47

This is a 3x4 contingency table.

General Contingency Table

Attribute A/B	B_1	B_2	…	B_n	Column totals
A_1	O_{11}	O_{12}		O_{1n}	$O_{1.}$
A_2	O_{21}	O_{22}		O_{2n}	$O_{2.}$
:					
A_m	O_{m1}	O_{m2}		O_{mn}	$O_{n.}$
Row totals	$O_{.1}$	$O_{.2}$		$O_{.n}$	

Here O_{ij} = Observed frequencies for attribute (A_i, B_j)

What is to be done with Contingency Table?

The focus of interest in a contingency table is

- the dependence or association between the column variable and the row variable
- For example: between treatment and response

What is to be done with Contingency Table?

Migraine Headache Patients who suffered from moderate to severe migraine headache took part in a double-blind clinical trial to assess an experimental surgery. A group of 75 patients were randomly assigned to receive either the real surgery on migraine trigger sites ($n = 49$) or a sham surgery ($n = 26$) in which an incision was made but no further procedure was performed. The surgeons hoped that patients would experience "a substantial reduction* in migraine headaches," which we will label as "success." Table 10.1.1 shows the results of the experiment.[1]

Table 10.1.1 Response to migraine surgery

		Surgery	
		Real	Sham
Substantial reduction in migraine headaches?	Success	41	15
	No success	8	11
	Total	49	26

columns represent : treatments
rows represent : responses.

Testing the Hypothesis:
Whether There is an Association or Not

H_o : The two variables are independent
H_a : The two variables are associated

The data is in the form of 2x2 contingency table.

Chi-square Test

The test statistics for testing association between the two attributes A and B is as follows.

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Where E_{ij} the expected frequencies are given by,

$$E_{ij} = \frac{O_{i.} \times O_{.j}}{O_{..}}$$

Rejection criteria:

Reject H_0 if

$$\chi^2 \geq \chi^2_{(m-1)(n-1)}(1-\alpha)$$

Example : Chi-square Test of Independence:

Attribute A: Expenditure on beauty products/month

Attribute B: Marital status

H_o : Both the factors viz. expenditure and marital status are independent.

Vs

H_1 : Both the factors viz. expenditure and marital status are dependent.

Number of respondents= 5012

Chisq. Test of Independence: Example...

	MARITAL_STATUS		
EXPENDITURE	Married	Not Married	TOTAL
Rs. 500 - 1000/-	1118	881	1999
Rs. 1001 - Rs.1500/-	1242	804	2046
Rs. 1501 - Rs. 2000/-	496	352	848
Rs. 2001 - Rs. 2501/-	66	53	119
TOTAL	2922	2090	5012

χ^2 Test of Independence: Example…

Chi-square	df	p_value
9.896	3	0.0193

Decision: By applying Chi- Square test of independence we reject H_0 at 5% LOS.

Conclusion: The Factors expenditure on beauty products and marital status are dependent (associated).

Correspondence Analysis and PCA

- Correspondence analysis (CA) is a generalized principal component analysis tailored for the analysis of qualitative data.
- Correspondence Analysis (CA) is an adaptation of PCA for categorical data.
- This means that CA extracts the important information from categorical variables in a way that can be interpreted geometrically.
- A key difference is that the data analyzed in CA is not a covariance/correlation matrix as in PCA.
- Instead, CA analyzes a contigency table between two categorical variables. Prior to running CA, the counts in a contingency table are transformed to instead reflect probabilities.

Introduction to Correspondence Analysis

- **What is Correspondence Analysis (CA)?**

  It is the graphical tool for checking the pattern of association between the categorical variables.

- **Need of CA:** In scientific investigations including sensory evaluation, Market research and customer satisfaction evaluations etc, questionnaires and surveys results in large number of responses with limited answer categories.

- **Objectives of CA:** Checking the pattern of association between the categorical variables.

  - The association among row and column categories
  - Association between both row and column categories

Introduction to CA

Objective: To study thoroughly the *symmetric or two-way association* between the two or more nominal/ordinal CVs cross-classified in a CT.

Concept:

- Correspondence analysis (CA) is a popular multivariate statistical technique.
- *CA visualizes graphically the symmetric association between the different categories of CVs* by representing high dimensional data as a points on a low-dimensional Euclidean space (especially two-dimensional).
- CA visualizes the association between CVs through two-dimensional plot known as Biplot.
- *Biplot is simply a generalisation of scatter plot* (which is used for the visualization of relationship between continuous variables).

Types of Correspondence Analysis

Depending upon the number of categorical variables, we can carry out

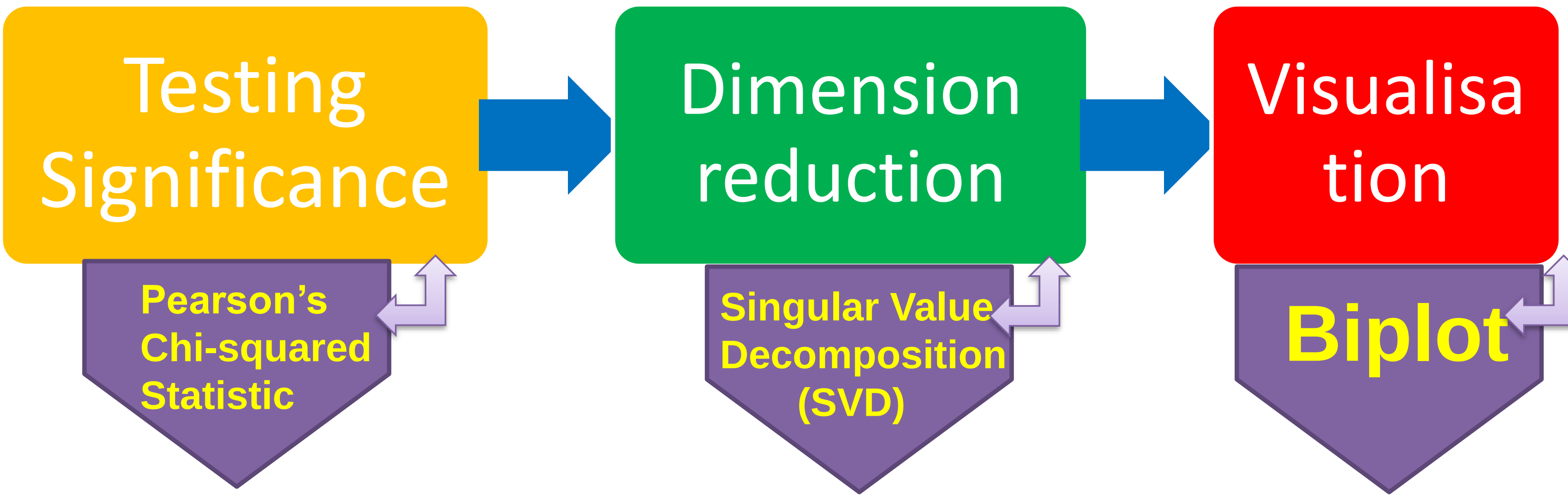
- **Simple Correspondence Analysis (SCA):** SCA involves two categorical variables and the graphical display of the corresponding two-way contingency table.

  SCA is carried out on contingency table.

- **Multiple Correspondence Analysis (MCA):** MCA is an extension of SCA to the case of three or more categorical variables.

  MCA is carried out on an indicator or *Burt Matrix* with cases as rows and categories of variables as column.

Three phases of CA



Three phases of CA…

Tools used in CA:

- *Pearson's Chi-squared Statistic:*

For testing the significance of two-way/symmetric association between variables.

- *Singular value decomposition*

Dimension reduction tool for visualization of high-dimensional data in low especially two-dimensional space.

- *Biplot*

For the visualization of association and better understanding of the hidden patterns of association among the categories of the variables.

Singular value decomposition

Some preliminary concepts related to SVD

- *Singular values:* Singular values of a rectangular matrix A are defined as the **square root of eigenvalues of the matrix AA' or $A'A$**.

- *Singular vectors:* Singular vectors of any matrix A are **eigenvectors of the matrices AA' or $A'A$**

  Left singular vectors of A: Singular vectors of AA'
  Right singular vector of A: Singular vectors of $A'A$

- *Weighted Norm of matrix/array:* The weighted norm of a matrix $A = ((aij))$ is defined as the square root of the sum of squares of its elements multiplied by the weights $w_{i.}$ and $w_{.j}$

$$||A||_w = \sqrt{\sum\sum w_i w_j a_{ij}^2}$$

Singular value decomposition of a matrix

- SVD is applicable to any rectangular matrix .
- SVD is the *generalisation of eigen decomposition (*which is applicable to only squares symmetric matrices) .
- The main idea of SVD is to decompose a rectangular matrix into three simple matrices; *two orthogonal matrices* and *one diagonal matrix.*
- That is, SVD decomposes any rectangular matrix into the product of three matrices, *two orthogonal matrices of left and right singular vectors*, and a *diagonal matrix of singular values.*

SVD…

- The SVD of any rectangular matrix A of size $m \times n$ is,

$$A = P\Delta Q'$$

where,

P : Orthonormal matrix of eigenvectors of AA' ($P'P = I_m$)

Q : Orthonormal matrix of eigenvectors of $A'A$ ($Q'Q = I_n$).

Δ : the diagonal matrix of the singular values and

$\Delta = \Lambda^{1/2}$ with Λ being the diagonal matrix of the eigenvalues of matrix $A'A$ and AA' .

- The columns of P are known as left singular vectors of A.
- The columns of are known as right singular vectors A.

Biplot in CA

- **What is Biplot** (‘$\boldsymbol{bi}$’ + ‘$\boldsymbol{plots}$’)**? :** Type of exploratory graph which is a generalization of the simple two-variable scatter plot.

- **What it depicts?:** The association between the categorical variables in the contingency table.

- **Introducer:** Grabiel (1971)

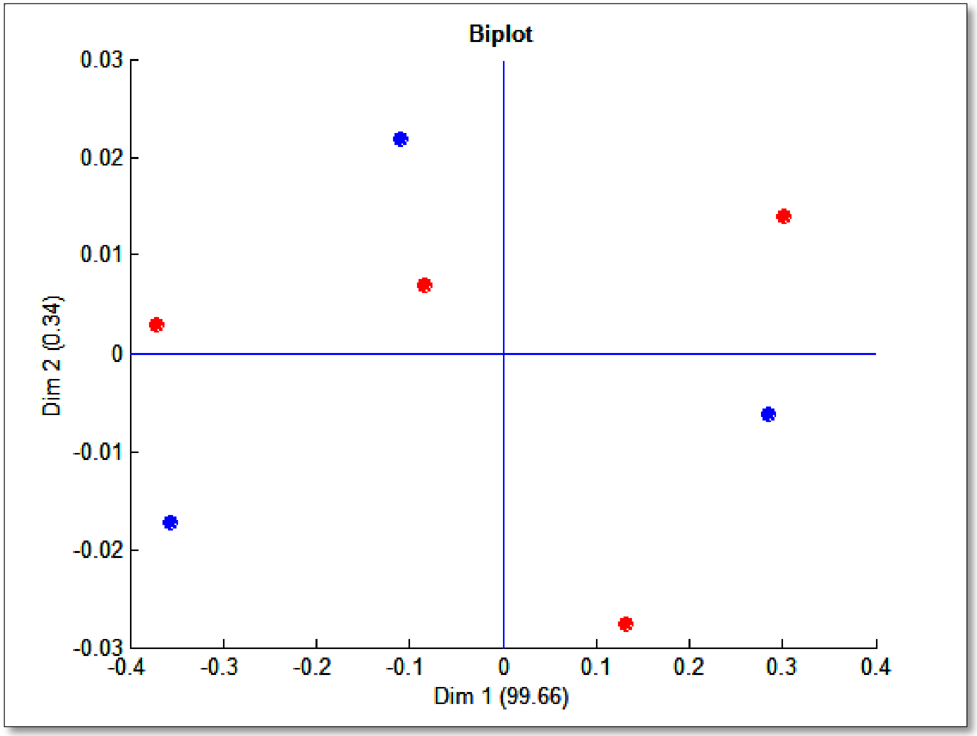
Bi: Both rows and columns

1 Biplot >>>2 Plots

Mathematical Definition of Biplot

Biplots are defined as the decomposition of target matrix into the product of two matrices, called as left and right matrices.

Software in which biplots are obtained: GGE biplot, Minitab, SPSS



Interpretations from Biplot

The Length of Biplot Vector:

- More the points apart from origin better its discriminating ability.
- The short length of Biplot vector shows that it is not related to the any parameters. Lack of variation or not well represented in the biplots.

The Cosine angle between the two vectors:

- **Acute angle:** Positive Correlation
- **Obtuse Angle:** Negative Correlation
- **Right Angle:** No Correlation

Example: Car rating on 10 parameters

Attribute 1: Car

Attribute 2 : Comfort/performance of Car

Categories of Attribute 1: 12 makes of Cars

Eldorado_GMC, Civic Honda, DL Volve etc.

Categories of Attribute 2: 10 parameters

MPG, Reliable, Ride, comfort, Visual

Response **:** Rating on 5 point scale given by a judge

Objective : Rank to the cars based on ratings of all 10 parameters.

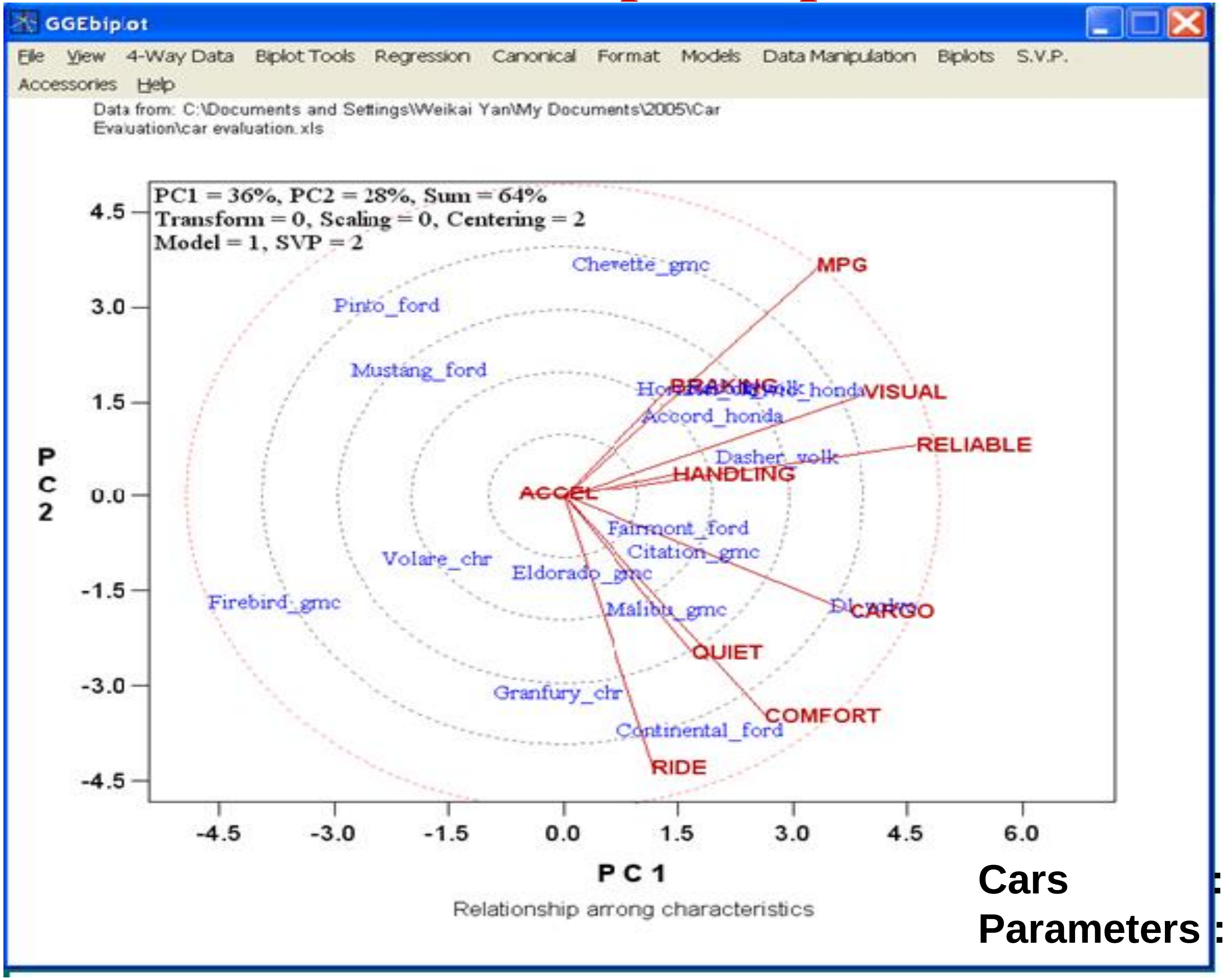
Example: Biplot

The following table gives the preference ratings (to all 10 parameters) for automobiles (car) manufactured in 1980 from the one judge.

Objective: Rank to the cars based on ratings of all 10 parameters.

	A	B	C	D	E	F	G	H	I	J	K
1	**Model**	MPG	Reliable	Accel	Braking	Handling	Ride	Visual	Comfort	Quiet	Cargo
2	ELDORADO_GMC	3	2	3	4	5	4	3	5	3	3
3	CHEVETTE_GMC	5	3	3	5	4	2	5	2	2	3
4	CITATION_GMC	4	1	5	5	5	5	5	5	2	5
5	MALIBU_GMC	3	3	3	3	4	4	4	5	4	4
6	FAIRMONT_FORD	3	3	2	4	3	4	5	4	3	4
7	MUSTANG_FORD	3	2	4	4	3	2	3	2	2	2
8	PINTO_FORD	4	1	3	4	3	1	3	2	2	2
9	ACCORD_honda	5	5	5	4	5	3	3	4	3	3
10	CIVIC_honda	5	5	4	5	4	3	5	4	3	4
11	CONTINENTAL_FORD	2	4	5	3	3	5	3	5	5	5
12	GRANFURY_CHR	2	1	3	4	3	5	3	5	3	5
13	HORIZON_CHR	4	3	4	5	5	3	5	2	3	5
14	VOLARE_CHR	2	1	5	3	3	3	3	4	2	4
15	FIREBIRD_GMC	1	1	5	3	5	5	1	2	3	1
16	DASHER_VOLK	5	3	5	5	5	4	5	4	3	5
17	RABBIT_VOLK	5	4	5	4	5	3	5	4	2	4
18	DL_VOLVO	4	5	2	4	5	5	5	5	5	5

Example: Biplot…



Cars : Blue
Parameters : Red

Results : Rank Car based on ‘MPG’ Parameter

- **Best car** : DL_Volvo
- **Poorest car** : Firebird_GMC
- **Ford Continental Car Best in** : Ride, Comfort
- **Ford Continental Car Poorest in** : MPG
- **Car which is High in MPG** : Civic_Honda, Chevette_GMC
- **Car which is Low in MPG** : Firebird _GMC

Notations used for a Contingency Table/CA

Consider, N be a $I \times J$ contingency table with two categorical variables A and B having I and J attribute categories respectively.

$\boldsymbol{n_{ij}}$: Observed frequencies associated with categorical variables (A_i, B_j)

$\boldsymbol{n_{i.}}$: i^{th} row total

$\boldsymbol{n_{.j}}$: j^{th} column total

$\boldsymbol{n = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}}$: Grand total

Attribute category for A/B	1	2	…	J	Row Total
1	n_{11}	n_{12}	$\cdots$	n_{1j}	$n_{1.}$
2	n_{21}	n_{22}	$\cdots$	n_{2j}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
I	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$n_{i.}$
Column Total	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.c}$	Grand Total(n)

Computation of Simple CA

Consider, N be a $I \times J$ contingency table with two categorical variables A and B having I and J categories respectively.

$\boldsymbol{N} = \left(\left(\boldsymbol{n_{ij}}\right)\right)$: $I \times J$ Contingency table

$\boldsymbol{P} = \left(\left(\boldsymbol{p_{ij}}\right)\right)$: $I \times J$ Correspondence matrix

$\boldsymbol{S} = \left(\left(\boldsymbol{s_{ij}}\right)\right)$: $I \times J$ matrix of standardized residuals

where, $s_{ij} = \frac{p_{ij} - r_i c_j}{\sqrt{r_i c_j}}$ $\quad$ $p_{ij} = \frac{n_{ij}}{n}$

$$p_{i.} = r_i = \frac{n_{i.}}{n} : \text{Row mass}$$

$$p_{.j} = c_j = \frac{n_{.j}}{n} : \text{Column mass}$$

$$\underline{r} = [r_1 \quad r_2 \quad \cdots \quad r_I]' \qquad D_r = diag(\underline{rr})$$

$$\underline{c} = [c_1 \quad c_2 \quad \cdots \quad c_J]' \qquad D_c = diag(\underline{c})$$

Computation of Simple CA…

- $N \equiv \left(\left(n_{ij}\right)\right)$ Contingency table of dimension $m_1 \times m_2$
- $P \equiv \left(\left(p_{ij}\right)\right) = \frac{\left(\left(n_{ij}\right)\right)}{n}$ Correspondence Matrix(where $n = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$)
- $r_i = \frac{n_{i.}}{n}$ Row mass (where $n_{i}. = \sum_{i=1}^{I} n_{ij}$)
- $c_j = \frac{n_{.j}}{n}$ Column mass (where $n_{.j} = \sum_{j=1}^{J} n_{ij}$)
- $D_r = diag(r_1, r_2, \dots, r_I)$
- $D_c = diag(c_1, c_2, \dots, c_J)$
- U and V = Unitary matrices ($UU^* = I, VV^* = I$)
- Σ = Rectangular diagonal matrix of singular values

Computation of Simple CA…

- Pearson's chi-square statistic is

$$\chi^2 = n\phi^2 = n\sum\sum\left(\frac{p_{ij}-r_ic_j}{\sqrt{r_ic_j}}\right)^2 = n\sum\sum s_{ij}^2$$

- Total inertia of CA = ϕ^2
- SVD of matrix of standardized residuals (S) is:

$$S = U\Sigma V'$$

where,

U is OM of left and right singular vectors of S such that $U'U = I$,

V is OM of right singular vectors of S such that $V'V = I$.

Σ is the diagonal matrix of singular values $\sigma_1, \sigma_2, \dots, \sigma_r$,

r = rank(S).

Computation of Simple CA…

- The coordinates for visualizing the association through biplot are:

Row coordinates : $F = D_r^{-1/2} U\Sigma$ ($I \times r$ matrix)

Column coordinates: G= $D_c^{-1/2}$VΣ (J× r matrix)

- Choose those two dimensions out of r (for representing on biplot) which are contributing maximum towards the total inertia.
- Plot row and column coordinates over the first two columns of matrices F and G on biplot.

Algorithm for computation of Simple CA

Step-wise algorithm to perform CA

1. Calculate the matrix S of standardized residuals of order $I \times J$ with $s_{ij} = \frac{p_{ij} - r_i c_j}{\sqrt{r_i c_j}}$ as,

$$S = D_r^{-\frac{1}{2}}(P - \underline{rc}')D_c^{-\frac{1}{2}}$$

2. Perform SVD on S as: $S = U\Sigma V'$

   where, r = rank(S) and $\Sigma = diag(\sigma_1, \sigma_2, \dots, \sigma_r, 0, \dots, 0)$

   and $\sigma_1 \geq \sigma_2, \geq \cdots, \geq \sigma_r$ are non-negative singular values of S

3. Calculate the total inertia of the data matrix as:

$$Intertia = \phi^2 = \sum \sum s_{ij}^2$$

   The Chi-square statistic is calculated as: $\chi^2 = n\phi^2$

Algorithm for computation of Simple CA

4) Obtain the biplot coordinates as:

Row coordinates : $F = D_r^{-1/2} U\Sigma$ ($I \times r$ matrix)

Column coordinates: G$= D_c^{-1/2}$VΣ (J$\times$ r matrix)

- The columns of F and G matrices are referred as the principal axes, or dimensions, of the biplot.
- For exploring the association between two CVs, the joint map of row and column coordinates (along the columns of F and G) is obtained.
- Plot row and column coordinates over the first two columns of matrices F and G on biplot.

Example

Suppose you collected data on the smoking habits of different employees in a company. The following data set is presented in Greenacre (1984, p. 55);

	Smoking Category				
Staff Group	(1) None	(2) Light	(3) Med.	(4) Heavy	Row Totals
(1) Senior Managers	4	2	3	2	11
(2) Junior Managers	4	3	7	4	18
(3) Senior Employees	25	10	12	4	51
(4) Junior Employees	18	24	33	13	88
(5) Secretaries	10	6	7	2	25
Column Totals	61	45	62	25	193

R code for Simple CA

```
#Topic: Correspondence Analysis for two-way CT
library(matlib); library(Matrix); library(readxl);
#---------------------------------------------------------------------------------------------
#Exporting Minitab data
N=read_excel("D:/Desktop/Ph.D Work/KKK Ma'am/CAData.xlsx", range = "C4:G14")
N=as.matrix(N);
rownames(N)=c("Geology", "Biochemistry",         "Chemistry",              "Zoology",    "Physics",
              "Engineering",            "Microbiology",            "Botany",     "Statistics",  "Mathematics");
#------------------------------------------------------------------------------------------------
N=matrix(c(21, 241, 251, 17, 54, 40, 10, 74, 65, 6, 11, 8, 11, 79, 108), nrow=5, ncol=3, byrow=TRUE)
dimnames(N)=list( marital_status = c('Married', 'Widowed', 'Divorced', 'Seperated', 'Never married'),
Attitude_about_life=c('Dull', 'Routine', 'Exciting'))
#---------------Performing CA Manually-------------------------------
n=sum(N);n;   P=N/n;
I=dim(N)[1]; J=dim(N)[2];
rm=apply(P, 1, sum); cm=apply(P, 2, sum)
DI=diag(rm); DJ=diag(cm);
#Matrix of Standardised residuals
S=inv(DI^(0.5))%*%((P-rm%*%t(cm)))%*%inv(DJ^(0.5)); S;
#Pearson's Chi-squared statistic
chisq=chisq.test(N);  chisq;
```

R code…

```
#SVD of Matrix of Standardised residuals
r=rankMatrix(S)[1];
svd_S=svd(S);
U=svd_S$u;  #t(U)%*%U=I
V=svd_S$v;  #t(V)%*%V=I
l=svd_S$d;
Lambda=diag(svd_S$d);

#Verification S=U*Lambda*t(V)
U%*%Lambda%*%t(V);
Phi2=sum(Lambda^2); chisq=n*Phi2;

#Contribution to Inertia
Axis=c(1:length(l), "Total");
Inertia=c(l^2, sum(l^2));
Prop=round(Inertia/sum(l^2),4);
cumulative=c(cumsum(Prop[-length(Prop)]), "-");
data.frame(Axis, Inertia, Prop, cumulative);
```

R code…

```
#Standard Coordinates
F=inv(DI^(0.5))%*%U;
G=inv(DJ^(0.5))%*%V;
#Principal Coordinates
F1=inv(DI^(0.5))%*%U%*%Lambda   #Row coordinates
G1=inv(DJ^(0.5))%*%V%*%Lambda;  #Column coordinates
Catlabs=c(rownames(N), colnames(N));

#Correspondence plot using row and column principal coordinates
x1=F1[,1]; y1=F1[, 2];  #Row coordinates
x2=G1[, 1]; y2=G1[, 2]; #Column coordinates
plot(x1, y1,
    xlab=paste0("Axis1: ", Prop[1]*100, "%"),
    ylab=paste0("Axis2: ", Prop[2]*100, "%"),
    main="Correspondence Plot", pch=19,
    xlim=c(min(c(x1,x2))-0.01, max(c(x1, x2))+0.01),
    ylim=c(min(c(y1,y2))-0.01, max(c(y1, y2))+0.01));
points(x2, y2, col="red", cex=0.8, pch=19);
text(x1+0.02, y1+0.02, labels=rownames(N));
text(x2+0.01, y2+0.01, labels=colnames(N));
abline(h=0);abline(v=0);
```

R code…

```
#--------------------Row isometric biplot------------------------------------
#x1=F1[,1]; y1=F1[, 2];  #Row coordinates
#x2=G[, 1]; y2=G[, 2]; #Column coordinates
#plot(x1, y1,
#     xlab=paste0("Axis1: ", Prop[1]*100, "%"),
#     ylab=paste0("Axis2: ", Prop[2]*100, "%"),
#     main="Row Isometric biplot", pch=19,
#     xlim=c(-2, 2),
#     ylim=c(-2, 2));
#points(x2, y2, col="red", cex=0.8, pch=19);
#text(x1+0.02, y1+0.02, labels=rownames(N));
#text(x2+0.01, y2+0.01, labels=colnames(N));
#abline(h=0);abline(v=0);

#---------------Performing CA using CAvariants Package--------------------------
#install.packages("CAvariants")
library(CAvariants);
CA=CAvariants(N, catype = "CA", alpha=0.05 )
CA;
plot(CA, plottype = "biplot", biptype="row", scaleplot=1.5);
```

Thank you