Index of Lecture 7: Introduction to multivariate analysis

Practical Information

Today's lecture — preliminaries for multivariate part of course:

- introduction to five datasets used throughout, to illustrate possible objectives of multivariate analysis (Chapter 1 of Manly[1], in short M.1),
- gentle introduction to matrix notation and algebra (Chapter M.2) — very helpful to understand methods even in a non-technical sense (our aim),
- some graphical tools to display multivariate data (Chapter M.3).

Aim of course coverage of multivariate methods, adapted from Manly's book:

> The purpose of this course is to introduce multivariate statistical methods to non-mathematicians. It is not intended to be a comprehensive course. Rather the intention is to keep the details to a minimum while serving as a practical guide that illustrates the possibilities of multivariate statistical methods.

Other news:

- home assignment #2 due tomorrow (only students not in VHM 812),
- need to discuss format/logistics of classes during the Canada Games period (starting with tomorrow's lab),
- time to start thinking about/planning your project.

[1] The 4th edition of the book is by Manly & Alberto (2016), but Manly was sole author of the earlier editions.

Textbook(s)

Major textbook:

(M) Manly BFJ, Alberto JAN (2016), *Multivariate Statistical Methods: A Primer*, 4th ed.; only addition from 3rd ed. (2004) is R code (even a brief introduction to R).[2]

Supplementary texts (some suggestions, based on availability):

(TF) Tabachnick BG, Fidell LS (2006/12/18), *Using Multivariate Statistics*, 5th/6th/7th ed.[3]; very thorough with well-worked examples, includes many topics not covered in course,

(SL) James G, Witten D, Hastie T, Tibshirani R (2013/21), *An Introduction to Statistical Learning, with Applications in R*, 1st/2nd ed.[4]; excellent introduction to classification methods (in a broad sense),

(ED) Everitt BS, Dunn G (2001), *Applied Multivariate Data Analysis*, 2nd ed.[5]; applied text focusing on classical methods.

Anticipated use of textbooks for course:

(1) lectures will follow M closely, including using the provided datasets for illustration of methods,

(2) other books or articles will be used, as needed and for methods not covered by these texts (e.g. PERMANOVA, SATSCAN) — feel free to contribute material of interest.

[2] The 1st edition from 1986 is at the Robertson Library; many parts are essentially unchanged (however, the 1st edition has no Chapter 3 on graphical methods).
[3] The 5th edition is at the Robertson Library.
[4] Freely downloadable from `https://www.statlearning.com/`.
[5] Available at Robertson Library as E-book; also Everitt (2005), *An R and S-Plus Companion to Multivariate Analysis*.

Multivariate Statistical Analysis

The terminology is not definitive, but in our usage...

- does not include methods such as multiple regression with a single ("univariate") outcome per subject but multiple predictor variables,[6]
- involves having multiple outcomes on each "subject", viewed together as a multi-dimensional set of values (but there may also be (multiple) predictors).

Overview of (main) methods in course (others may be added):

- multivariate regression and MANOVA: $\sim$ ordinary regression/ANOVA but for multivariate outcome,
- principal components and factor analysis: reduce the information in a set of variables to suitable new variables combined linearly,
- classification and discriminant analysis: separate existing groups in data based on suitable linear or non-linear combinations of the original variables,
- cluster analysis: identify groups in data based on suitable linear combinations of the original variables, or based on distances in a space for the observations,
- multidimensional scaling: create distances between observations along multiple directions.

[6] Also termed multivariable models, to distinguish them from multivariate models.

Example M1.1: Sparrows

Data: 5 morphological measurements on 49 moribund female sparrows, of which 21 subsequently survived, taken to a biological laboratory after a storm (in 1898!),

X_1 = total length (mm),
X_2 = alar extent (mm),
X_3 = length of beak and head (mm),
X_4 = length of humerus (mm),
X_5 = length of keel of sternum (mm),

Bird	X_1	X_2	X_3	X_4	X_5	survival
1	156	245	31.6	18.5	20.5	1
2	154	240	30.4	17.9	19.6	1
...			...			...
48	162	245	32.5	18.5	21.1	0
49	164	248	32.3	18.8	20.9	0

Possible objectives of analysis:

(a) describe relations between variables $X_1 - X_5$, within each of the survivor groups, and compare the two groups in terms of means, variability and relationships,

(b) construct some combined feature(s) from $X_1 - X_5$ that allow(s) accurate prediction of survival, $\sim$ index of fitness of the birds.

First thoughts:

- descriptive analysis for (a): means, standard deviations, correlations,
- (b) could be approached by multiple logistic regression,

$\Rightarrow$ multivariate regression, principal components, classification/discriminant analysis.

Example M1.2: Skulls

Data: 4 morphometric measurements on 150 Egyptian male skulls, 30 from each of 5 distinct time periods,[7]

X_1 = maximum breadth (mm),
X_2 = basibregmatic height (mm),
X_3 = basiolvelar length (mm),
X_4 = nasal heightlength of humerus (mm),

Skull	X_1	X_2	X_3	X_4	period
1	131	138	89	49	1
2	125	131	92	48	1
...			...		...
150	136	133	97	51	5

Possible objectives of analysis:

(a) describe relations between $X_1 - X_4$ within each of the periods, and compare the 5 groups in terms of means, variability and relationships,

(b) measure distances of distributions of $X_1 - X_4$ across the periods, and relate such distances to physical time,

(c) construct a combined (linear) function from $X_1 - X_4$ that in some sense describes development over time.

First thoughts (beyond Example M1.1):

- not obvious how to quantify the impact of time (e.g., considering the 5 periods as ordered groups does not directly involve time differences)...

$\Rightarrow$ multivariate regression, multivariate distance, classification/discriminant analysis.

[7] (1): early predynastic (4000 B.C.); (2): late predynastic (3300 B.C.); (3): 12-13th dynasties (1850 B.C.); (4): Ptolemaic (200 B.C.); (5): Roman (A.D. 150).

Example M1.3: Butterflies

Data: 4 environmental and 6 genetic variables (frequencies for Phosphoglucose-Isomerase (Pgi) genes) for 16 colonies of the butterfly *Euphydryas editha* in California and Oregon (one colony only),

X_1 = altitude (ft),
X_2 = annual precipitation (in),
X_3 = maximum temperature (°F),
X_4 = minimum temperature (°F),
Y_r = frequency (%) for Pgi gene type r; $r = 0.4, 0.6, 0.8, 1.0, 1.16, 1.3$.

Colony	X_1	X_2	X_3	X_4	$Y_{0.4}$	$Y_{0.6}$	$Y_{0.8}$	$Y_{1.0}$	$Y_{1.16}$	$Y_{1.3}$
SS	500	43	98	17	0	3	22	57	17	1
SB	808	20	92	32	0	16	20	38	13	13
...		...					...			
GL	10500	50	81	-12	0	3	1	92	4	0

Possible objectives of analysis:

(a) describe relation between the Pgi gene frequencies and environmental variables,

(b) is environmental similarity or physical proximity more important for genetic similarities? — the latter presumably representing effects of migration.

First thoughts:

- dataset seems quite small, and contains no grouping of interest,
- the grouping of interest is among variables, not observations,
- spatial information needed for (b),

⇒ multivariate distances, canonical correlations.

Example M1.4: Prehistoric dogs

Data: 6 mandible (lower jaw) morphometric measurements (all in mm) on craniums of prehistoric dogs in Thailand and 6 other possibly related species,

X_1 = breadth of mandible,
X_2 = height of mandible below 1st molar,
X_3 = length of 1st molar,
X_4 = breadth of 1st molar,
X_5 = length 1st molar $-$ 3rd molar,
X_6 = length 1st molar $-$ 4th premolar,

Species	X_1	X_2	X_3	X_4	X_5	X_6
Modern dog	9.7	21.0	19.4	7.7	32.0	36.5
Golden jackal	8.1	16.7	18.3	7.0	30.3	32.9
Chinese wolf	13.5	27.3	26.8	10.6	41.9	48.1
Indian wolf	11.5	24.3	24.5	9.3	40.0	44.6
Cuon	10.7	23.5	21.4	8.5	28.8	37.6
Dingo	9.6	22.6	21.1	8.3	34.4	43.1
Prehistoric dog	10.3	22.1	19.1	8.1	32.2	35.0

Possible objectives of analysis:

(a) describe relations between the prehistoric dog and other species, in particular quantify distances between species.

First thoughts:

- dataset seems quite small (e.g., lacks replication within species), and contains no grouping of interest.

$\Rightarrow$ multivariate distances, cluster analysis.

Example M1.5: Employment

Data: Percentages of workforce employment in 9 different sectors[8] of 30 European countries grouped into 4 political/economical zones[9], as of the 1990s (approximately),

X_r = employment (%) in sector r; $r = 1, \ldots, 9$,

Country	Political group	X_1 AGR	X_2 MIN	X_3 MAN	X_4 PS	X_5 CON	X_6 SER	X_7 FIN	X_8 SPS	X_9 TC
Belgium	EU	2.6	0.2	20.8	0.8	6.3	16.9	8.7	36.9	6.8
Denmark	EU	5.6	0.1	20.4	0.7	6.4	14.5	9.1	36.3	7.0
...	...					...				
Turkey	Other	44.8	0.9	15.3	0.2	5.2	12.4	2.4	14.5	4.4

Possible objectives of analysis:

(a) describe employment patterns, and if possible establish groups of countries with similar patterns; such groupings could then be compared with the political grouping provided.

First thoughts:

- primary interest probably not related to the grouping provided,
- it may pose problems that the employment percentages add up to 100%,

⇒ multivariate distances, principal components/factor analysis, classification/ discriminant analysis, cluster analysis.

[8] (1): AGR (agriculture); (2): MIN (mining); (3): MAN (manifacturing); (4): PS (power and water supplies), (5): CON (construction); (6): SER (services); (7): FIN (finances); (8): SPS (social and personal services); (9): TC (transport and communications).

[9] (1): EU (European union); 2: EFTA (European free trade area); 3: East (former Eastern Europe union); (4): Other.

Common Assumptions

Two common assumptions:

- independence between sets of measures taken on different "subjects" (whereas independence is never assumed for the multiple measures on the same subject),
  - can be violated by data possessing a particular structure, e.g. hierarchical,
  - generalizations to dependent data depend on the specific methods but are generally not straightforward ($\Rightarrow$ tempting to ignore and "interpret with caution"),
- normally distributed, interval-scale measures, more precisely that the p-dimensional set of measures, say $X_1, \ldots, X_p$, follows a MVN (multivariate normal) distribution with mean μ and variance-covariance matrix Σ, where
  - μ is the set of means for the components,
  - Σ contains the covariances between all pairs of components (see also 7L−13);

  this assumption includes a normal distribution for each component[10]; violations of normality can be dealt with in different ways,
  - transformation to "better" scale (e.g. for skewed distributions),
  - specialized methods, such as correspondence analysis for categorical data,
  - claim of robustness of methods towards non-normality (to be further discussed).

[10] A MVN also requires the conditional distributions of subsets of components given the others (or all linear combinations of the components) to be normal, but this extra condition is often ignored in practice; M1.3, TF (p. 78).

Matrix Definitions

An $m \times n$ matrix A is a collection of mn values organized in m rows and n columns:

$$\mathrm{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \vdots & & & \vdots & \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{pmatrix}$$

- a square matrix has equally many rows and columns: $m = n$,
- a row vector, say $\mathrm{r} = (r_1\ r_2\ \dots\ r_n)$, has only one row: $m = 1$,
- a column vector has only one column: $n = 1$,
- the transpose matrix A' or A^t results from interchanging the rows and columns (i.e., $a^t_{ij} = a_{ji}$),
  - a symmetric matrix A satisfies: $\mathrm{A}^t = \mathrm{A}$,
  - the transpose of a row vector is a column vector, and vice versa,
- a zero matrix 0 has all elements $= 0$; a one matrix 1 has all elements $= 1$,
- a square diagonal matrix D has only 0s outside the diagonal,[11]
- the unity matrix I is a diagonal matrix with 1s in the diagonal,[12]
- the trace of a square matrix is the sum of the diagonal elements,

$$\mathrm{tr}(\mathrm{A}) = a_{11} + a_{22} + \ldots + a_{nn}.$$

[11] The condition can be expressed that $d_{ij} = 0$ when $i \neq j$.
[12] The condition can be expressed that $i_{ij} = 1$ when $i = j$, and $i_{ij} = 0$ otherwise.

Matrix Operations

Let A and B be matrices, and let k be a (real-valued) number, often termed a scalar, so as to distinguish it from vectors and matrices.

- addition $(\mathrm{A}+\mathrm{B})$ and subtraction $(\mathrm{A}-\mathrm{B})$ of matrices occurs for each element (at its position in the matrix),
  - A and B must have the same dimensions,
  - e.g.: $(\mathrm{A}+\mathrm{B})_{11} = a_{11} + b_{11}$; $(\mathrm{A}-\mathrm{B})_{32} = a_{32} - b_{32}$,
- scalar multiplication by k also occurs for each element, e.g. $(k\mathrm{A})_{11} = k \cdot a_{11}$,
- matrix multiplication ($\mathrm{A}\cdot\mathrm{B}$ or A.B or $\mathrm{A}\times\mathrm{B}$, but most commonly just AB) is not defined by multiplying pairs of matrix elements,
  - requires compatible matrix dimensions: no. columns of A must equal no. of rows of B, say A $(m\times n)$ and B $(n\times p)$,
  - for A $(m\times n)$ and B $(n\times p)$, A.B has dimension $(m\times p)$,
  - the (i,k)th element of A.B is the sum

$$(\mathrm{A.B})_{ik} = \sum_j a_{ij}\, b_{jk} = a_{i1}b_{1k} + a_{i2}b_{2k} + \ldots + a_{in}b_{nk}\,,$$

  - it is not generally true that A.B = B.A; note that A.B and B.A are not even of the same dimensions, unless A, B are square matrices.

Matrix Inversion and Quadratic Form

Matrix inversion=the matrix version of division but more complicated than for numbers.

Definition: a square matrix $\mathbf{A}$ $(n\times n)$ has inverse $\mathbf{A}^{-1}$ $(n\times n)$ if the matrices satisfy:
$$\mathbf{A}.\mathbf{A}^{-1} = \mathbf{A}^{-1}.\mathbf{A} = \mathbf{I}.$$

Example: the 2×2 matrix inverse of $\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$

- exists exactly when the determinant[13] $\Delta = a_{11}a_{22} - a_{21}a_{12}$ is nonzero,
- can be computed as $\mathbf{A}^{-1} = \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix} /\Delta.$

General definitions and results for $\mathbf{A}$ $(n\times n)$:

- $\mathbf{A}$ has an inverse $\mathbf{A}^{-1}$ exactly when the determinant[13] $|\mathbf{A}|$ is non-zero,
- if $\mathbf{A}^{-1}$ does not exist, then $\mathbf{A}$ is called singular,
- if $\mathbf{A}^{-1}$ exists and $\mathbf{A}^{-1} = \mathbf{A}^t$, then $\mathbf{A}$ is called orthogonal.

A quadratic form Q for a symmetric matrix $\mathbf{A}$ $(n\times n)$ is a linear function of a column vector $\mathbf{x} = (x_1\ x_2\ \ldots\ x_n)^t$,
$$Q = Q(\mathbf{x}) = \mathbf{x}^t\mathbf{A}\mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j.$$

[13] The standard notation for the determinant of $\mathbf{A}$ is $|\mathbf{A}|$. Also for square matrices larger than (2×2), the determinant $|\mathbf{A}|$ has an explicit, but complex form; $|\mathbf{A}|$ is therefore best calculated using mathematical/statistical software.

Mean Vectors and Covariance Matrices

Multivariate normal (MVN) distribution for a p-dimensional column vector $\mathrm{X} = (X_1\ X_2\ \ldots X_p)^t \sim \mathrm{MVN}(\mu, \Sigma)$, where

* $\mu = (\mu_1\ \mu_2\ \ldots\ \mu_p)^t$ is a p-dimensional column vector of means (i.e., $\mathrm{E}X_j = \mu_j$),
* $\Sigma = (\sigma_{jk})_{jk}$ is a $(p \times p)$ positive definite[14] matrix of the pairwise covariances,[15]

has p-dimensional probability density function

$$f(\mathrm{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\tfrac{1}{2}(\mathrm{x}-\mu)^t\Sigma^{-1}(\mathrm{x}-\mu)\right)$$

Estimation of mean and covariance:
From i.i.d.[16] multivariate observations $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n$ we can estimate the mean vector and covariance by

* $\hat{\mu} = \bar{\mathrm{X}} = (\bar{X}_1\ \bar{X}_2 \ldots \bar{X}_p)^t$, the column vector of means for each variable,
* $\hat{\Sigma} = \mathrm{S} = (s_{jk})_{jk}$, the empirical covariance matrix with the $s_{jk} = r_{jk}s_j s_k$ defined in terms of the Pearson correlation coefficient (r_{jk}) between X_j and X_k and the standard deviations (s_j, s_k) for X_j and X_k, respectively.

[14] A symmetric matrix A is positive definite if: $Q(\mathrm{x}) = \mathrm{x}^t\mathrm{Ax} > 0$ for any non-zero x; see also next slide.
[15] Formally, $\sigma_{jk} = \mathrm{E}(X_jX_k) - (\mathrm{E}X_j)(\mathrm{E}X_k)$, and $\sigma_{jj} = \mathrm{Var}(X_j)$; also, the correlation matrix $\mathrm{R} = (\rho_{jk})_{jk}$ gives the off-diagonal correlations $\rho_{jk} = \sigma_{jk}/(\sigma_{jj}\sigma_{kk})$.
[16] Also here, i.i.d. stands for: independent and identically distributed observations.

Eigenvalues and Eigenvectors

For a symmetric matrix A ($n \times n$), a scalar λ and a column vector x ($n \times 1$), we consider the equation

$$\mathbf{A}\mathbf{x} = \lambda\,\mathbf{x} \quad \text{or} \quad (\mathbf{A} - \lambda\,\mathbf{I})\,\mathbf{x} = \mathbf{0},$$

which can also be represented as the set of equations:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= \lambda x_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= \lambda x_2, \\ \qquad\qquad\qquad\qquad\qquad\quad \vdots & \quad \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n &= \lambda x_n. \end{cases}$$

The possible solutions[17] reflect fundamental properties of A:

- the eigenvalues (or characteristic roots) λ, of which there can be at most n distinct values, all of which are > 0 for a positive definite matrix, and which satisfy

$$\mathrm{tr}(\mathbf{A}) = \lambda_1 + \ldots + \lambda_n,$$

- the eigenvectors x for each λ express specific "directions" in the matrix.

Applying these general matrix concepts to a covariance matrix (Σ) gives:

- $\mathrm{tr}(\Sigma) = \sigma_{11} + \ldots + \sigma_{pp} = \lambda_1 + \ldots + \lambda_n$,
  — i.e., the eigenvalues decompose the total variance. . . ,
- the eigenvectors correspond to specific directions of the variation in the data.

[17] Calculation of solutions λ and x to the eigenvalue problem is not straightforward to establish; TF work through a (2×2) example.

Graphing Multivariate Data

Fundamental challenge: getting beyond 2-d (or perhaps 3-d) scatterplots to display relationships between variables.

Extension to 3-d plots:

- ○ use 3rd axis (depths), and rotate points,
- ○ or create (interpolated/smooth) surfaces for 3rd variable,
- \+ both ideas useful, but not exactly grand extensions...

First idea: pairwise scatter plots ⇒ matrix or draftman's plot,

- \+ can show relations between a limited number of variables,
- − difficult to see similarity between observations.

Next idea: represent variables' values symbolically in displays involving multiple features, e.g. faces (Chernoff) or stars.

Another idea: position multiple measures per subject on x-axis at suitable (arbitrary) points and join values for same subject by line (profile plot), or position them as multiple bars beside each other (bar graph),

- \+ can illustrate grouping and distance between subjects,
- − requires measures on comparable scales, difficult to see relation between variables.

Final comment: one key purpose of multivariate methods is to reduce the dimension of information across many variables into a few ("index"; Manly) variables, e.g. for the purpose of facilitating graphical exploration.

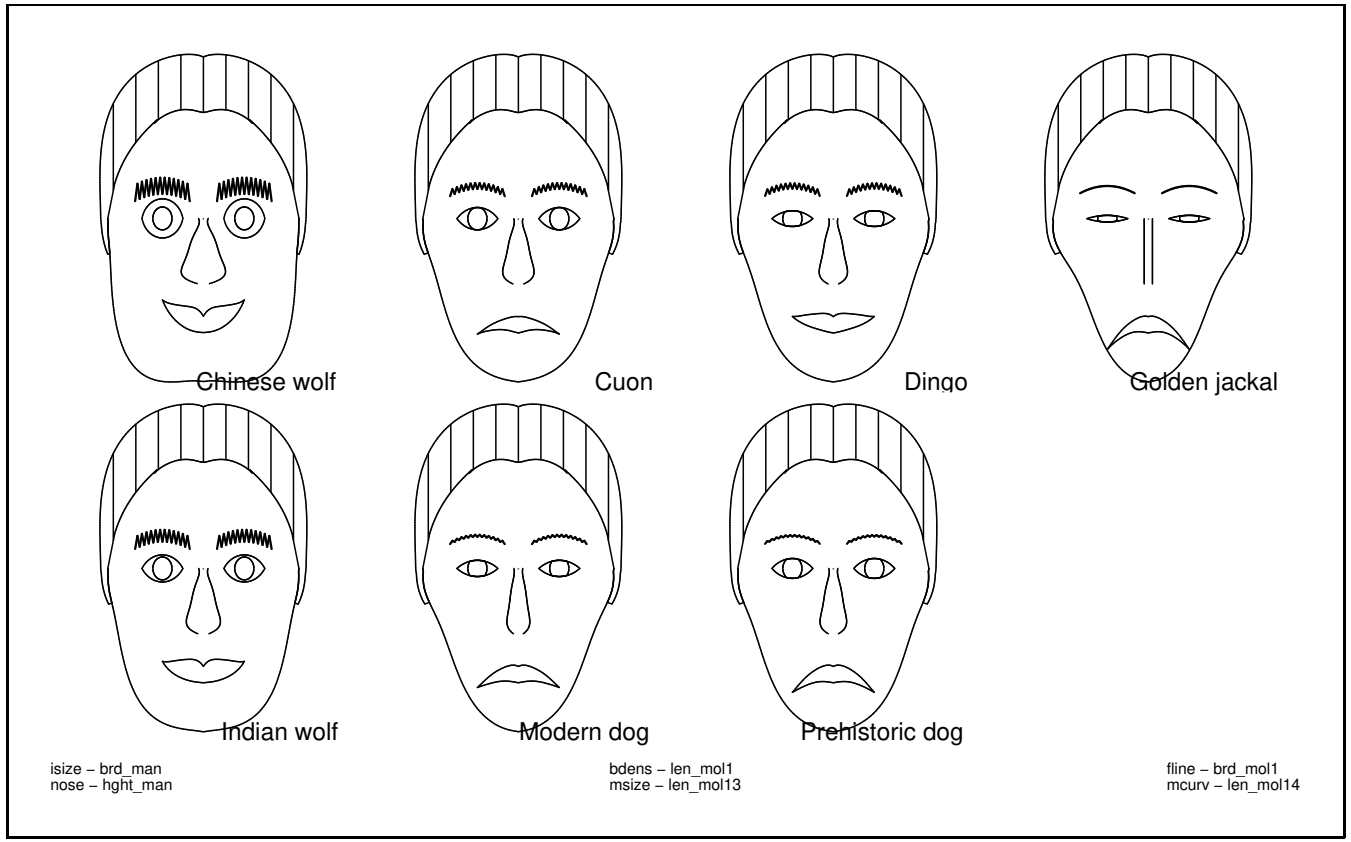
Sparrow: Matrix Plot

- all variables seem positively and approximately linearly associated, to similar degree (correlation range: $0.526 - 0.769$),
- no obvious differences between survivors and morts,
- no apparent strong outliers.

Dog: Chernoff plot

Idea[18]: use different facial features (up to 18) to represent variables,

- one selected feature per variable, scaled so that low/high values correspond to extreme features,
- some arbitrariness/flexibility in the allocation of features to variables,
- not clear that visual impression is not affected substantially by allocation of features to variables (other symbols, e.g. stars, may be more robust).



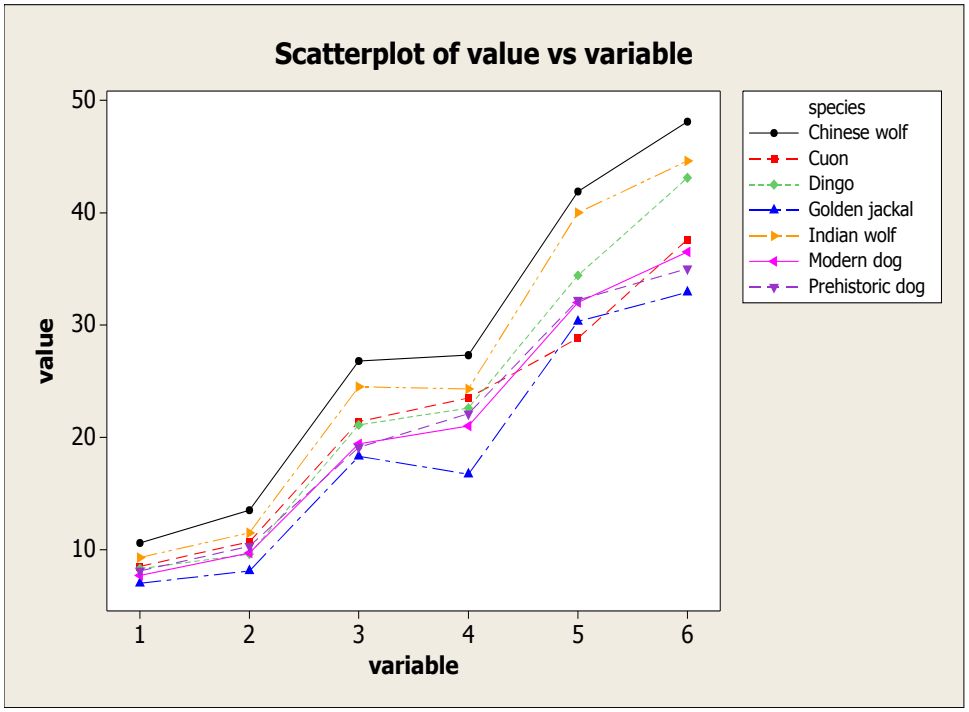
Interpretation:

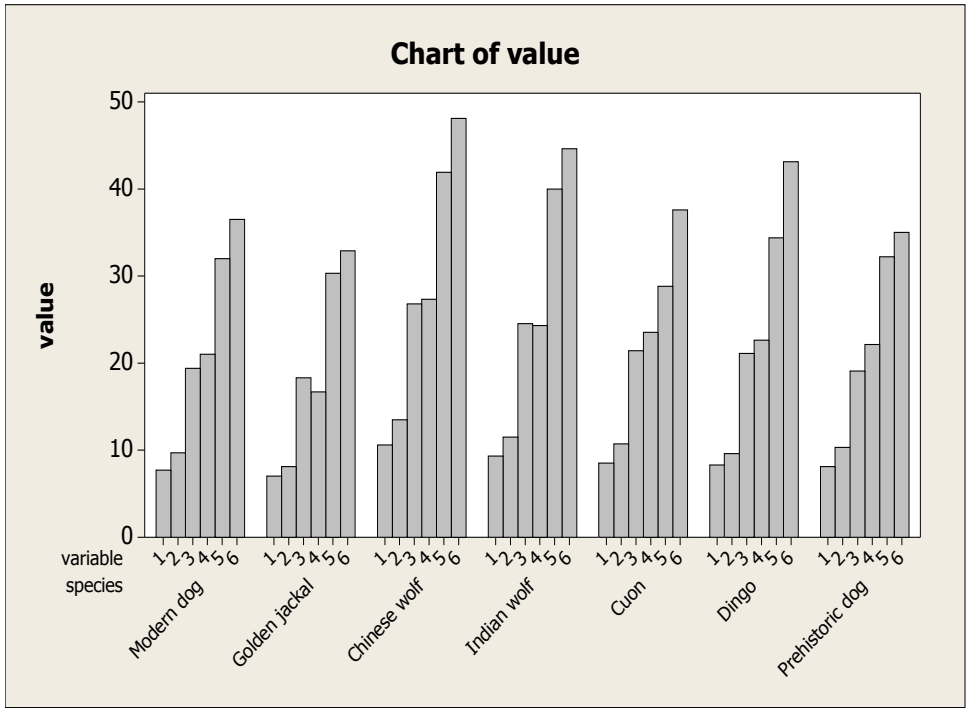
- the plot[19] shows closeness of prehistoric and modern dog, also their distance to the wolves.

[18] Attributed to Chernoff H (1973), *J. Amer. Statist. Assoc.* 68, 361-368.

[19] Done with Stata `chernoff` add-on package, using the features: `isize` (eye size), `nose` (nose line), `bdens` (brow density), `msize` (mouth size), `mcurv` (mouth curvature) and `fline` (face line).

Dog: Profile Plot and Bar Graph





Interpretation:

- both plots show closeness of prehistoric and modern dog, and their differences from the wolves (Chinese and Indian).

Structuring Multivariate Data

Two major formats of datasets:

- wide: usual format, one record (worksheet row) per subject, with multiple columns for the multivariate measurements as well as any extra (predictor) variables,[20]
- long: one record per measurement, with the multivariate measurements put together in a single column and the type of measurement indicated in another column, with one additional column containing subject id's, as well as (repeated values of) predictor variables in other columns,
  - most common data format for univariate data and also for multivariate data, when the multivariate measurements are different instances of the same variable, e.g. over time (i.e., repeated measures data).

How to switch between data formats? — software specific implementations:[21]

- Minitab: "stacking" columns for: wide → long; "unstacking" for reverse process,
- Stata: "reshape" command in versions "wide" and "long", indicating the targeted data format,

Why shift data format? — typically due to specific requirements of implementations of analytical or graphical procedures.

[20] All examples of the lecture showed the data in this format.
[21] Note also the related operation transpose (Stata: "xpose"), switching rows and columns in a worksheet.