# Index of 1-L

# Introduction

## WELCOME to all of you...

Logistics:

- all info at webpage: `people.upei.ca/hstryhn/vhm881mult`,
- students should formally either take course for credit or audit,
  - course has been approved as VHM 881, but probably not set up at Registrar's Office yet,
- may use Moodle account (but has not been set up yet either).

Aim of course/reading group, 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.

Specifically, to provide

- basic understanding of (selected) multivariate techniques, and the differences between them,
- a first experience with their application to real data.

Software:

- no in-depth coverage of a single software package,
- simplicity of software use prioritized over flexibility or completeness: many illustrations will use the built-in menus of Minitab,

Discussions among participants may initiate explorations of more complex methodology and software.

# Textbook Material

Three major texts:

**M** Manly BFJ (2004), Multivariate Statistical Methods: A Primer, 3rd ed.;
1st edition available at UPEI library (currently with Jenny),

**TF** Tabachnick BG, Fidell LS (2006/12), Using Multivariate Statistics, 5th/6th ed.;
5th edition available at UPEI library (currently with Henrik), includes many topics not covered in course,

**S** Scharma S (1995): Applied Multivariate Techniques;
not at UPEI library, Sami has this book.

. . . many other books exist, both applied and mathematical/technical.

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 may be used as needed, based on discussion in group,

- feel free to contribute material you think could be of interest,
- feel free to suggest expansions of the coverage of methods initially scheduled. . .

# 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,[1]
- involves having multiple outcomes on each "subject", viewed together as a multi-dimensional set of values.

Multivariate models may also involve multiple predictors, and their results may be integrated in "univariate" analysis.

Overview of methods in course

- multivariate regression and MANOVA: ∼ 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,
- discriminant analysis: separate existing groups in data based on suitable linear combinations of the original variables,
- cluster analysis: identify groups in data based on suitable linear combinations of the original variables,
- canonical correlations: create highest correlation between linear combinations of variables in two groups of variables,
- multidimensional scaling: create distances between observations along multiple directions.

[1] Also called *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,

$$
\begin{aligned}
X_1 &= \text{total length } (mm),\\
X_2 &= \text{alar extent } (mm),\\
X_3 &= \text{length of beak and head } (mm),\\
X_4 &= \text{length of humerus } (mm),\\
X_5 &= \text{length of keel of sternum } (mm),\\
g &= \text{survival } (0/1),
\end{aligned}
$$

| 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, std.dev., correlations,
- (b) could be approached by multiple logistic regression,

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

## Example M1.2: Skulls

Data: 4 morphometric measurements on 150 Egyptian male skulls, 30 from each of 5 distinct time periods,[2]

$$
\begin{aligned}
X_1 &= \text{maximum breadth } (mm),\\
X_2 &= \text{basibregmatic height } (mm),\\
X_3 &= \text{basiolvelar length } (mm),\\
X_4 &= \text{nasal heightlength of humerus } (mm),\\
g &= \text{period } (1/2/3/4/5),
\end{aligned}
$$

| 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):

- 5 groups instead of 2, so logistic regression not an option here,
- not obvious how to quantify the impact of time,

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

[2] (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 on craniums of prehistoric dogs in Thailand and 6 other possibly related species,

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

| 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 |
| ... | | | ... | | | |
| 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, and contains no grouping of interest.

⇒ multivariate distances, cluster analysis.

## Example M1.5: Employment

Data: Percentages of workforce employment in 9 different sectors[3] of 30 European countries grouped into 4 political/economical zones[4], as of the 1990s (approximately),

$$X_i = \text{employment (\%) in sector } i;\ r = 1, \ldots, 9,$$
$$g = \text{political grouping (1/2/3/4)},$$

| 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%,

$\Rightarrow$ multivariate distances, principal components/factor analysis, discriminant analysis, cluster analysis.

[3] (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).

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

## Common Tools and 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 or repeated measures,
  - ∗ generalizations of methods to dependent data depend on the specific methods but are generally not straightforward (so it is tempting to ignore this issue and interpret the results "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 $\mu$ and variance-covariance matrix $\Sigma$, where
  - ∗ $\mu$ is the set of means for the components,
  - ∗ $\Sigma$ contains the covariances between all pairs of components.

  This assumption includes a normal distribution for each component[5]; 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 discussed throughout the course).

[5] 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; M (p. 15), TF (p. 78).