(Chapter 15-17) GA

Multivariate analysis I – hypothesis testing

- Multivariate data is data in which we have measures of >1 metric (typically many) for each of a number of exp/obs subjects (= many variables).
- We have dealt with cases where we use several variables/factors as predictors in mutiple regressions and other forms of general/generalized linear models (ANOVA, ANCOVA, ANODEV).
- In other cases, our set of variables are **not predictors:** a set of variables may collectively be our **response variables** (lecture I; *group identity known*) or we may be interested in detecting/describing some pattern of **covariation** across, or between sets of, variables **(**lecture II; *group identity unknown*).

1. Basic concepts (Covariance matrices; linear combiations of variables).
2. Principal components analysis
3. MANOVA and related methods

1. Basic concepts

Among a set of variables measured over multiple subjects, we can express the pattern of covariation in variance-covariance matrix (**C**):

$$\begin{array}{lllll}
 & X1 & X2 & X3 & X4 \\
X1 & V_{X1} & COV_{X1X2} & COV_{X1X3} & COV_{X1X4} \\
X2 & COV_{X2X1} & V_{X2} & COV_{X2X3} & COV_{X2X4} \\
X3 & COV_{X3X1} & COV_{X3X2} & V_{X3} & COV_{X3X4} \\
X4 & COV_{X4X1} & COV_{X4X2} & COV_{X4X3} & V_{X4}
\end{array}$$

Where the diagonal is the variance (s^2) of each variable and the cells are the covariance between pairs of variables. If these are standardized by dividing with the standard deviations of variables, we get the correlation matrix (**R**):

$$\begin{array}{lllll}
 & X1 & X2 & X3 & X4 \\
X1 & 1 & r_{X1X2} & r_{X1X3} & r_{X1X4} \\
X2 & r_{X2X1} & 1 & r_{X2X3} & r_{X2X4} \\
X3 & r_{X3X1} & r_{X3X2} & 1 & r_{X3X4} \\
X4 & r_{X4X1} & r_{X4X2} & r_{X4X3} & 1
\end{array}$$

C is used when we in different ways seek “patterns” of covariation in our data, by means of matrix algebra. Different multivariate methods use slightly different ways of seeking such patterns, but they are all based on this matrix.

- Our variables are often measured on different scales and have very different variances...if we want to compare their relative importance, this is often a problem.
- The solution is often to first **standardize the data**, to put all variables on a common scale → typically by, for each observation, subtracting the mean and dividing by the standard deviation (all variables will then have a mean of zero and a standard deviation of one):
  $$z = \frac{x - \mu}{\sigma},$$
  where x is a raw value to be standardized; μ is the mean and σ is the standard deviation of the variable in question.

Linear combinations of variables

- When we have many variables, the pattern of covariation among them become too complex to "grasp". We need methods to simplify things somehow...
- We do this by letting **C** guide us when extracting a **new set** of a more limited number of ***derived variables*** (*also called latent variables*) – these are called different things in different multivariate methods (e.g., discriminant functions, principal components, canonical variates, DCA axes, NMDS dimensions, etc, etc).
- These derived variables are all ***linear combinations of the original variables***, such that:
  $Z = (\text{constant} +)\ c_1X1 + c_2X2 + c_3X3 + c_4X4$

for the derived variable **Z**, where X1-4 are the original variables and the coefficients $c_{1\text{-}4}$ are weights given each variable. For example, we construct two new derived variables out of our original data matrix:

Observation#	**X1**	**X2**	**X3**	**X4**	$\mathbf{Z_1}$	$\mathbf{Z_2}$
1	*x*	*x*	*x*	*x*	*x*	*x*
2	*x*	*x*	*x*	*x*	*x*	*x*
3	*x*	*x*	*x*	*x*	*x*	*x*
.	.	.	.	.	.	.
.	.	.	.	.	.	.
n	*x*	*x*	*x*	*x*	*x*	*x*

where *x* are "actual numbers" and $Z_{1\text{-}2}$ are two new derived variables that are different linear combinations of X1 – X4.

- Different multivariate methods differ in ***how*** these linear combinations are constructed (the "aim" differs!), but they are all based on constructing such derived variables!
- Each new derived variable is associated with an **eigenvalue (λ)**; this value describes the amount of variance in the original data matrix that is explained (i.e., retained) by a given derived variable.
- Each new derived variable is also associated with an **eigenvector**; this is the set of coefficients that describes the loading that each original variable has on the derived variable – it can be seen as a "vector representation" of the derived variable that "cuts though" the space or the original variables:

	$\mathbf{Z_1}$	$\mathbf{Z_2}$
X1	c_{11}	c_{21}
X2	c_{12}	c_{22}
X3	c_{13}	c_{23}
X4	c_{14}	c_{24}

Dissimilarity / similarity

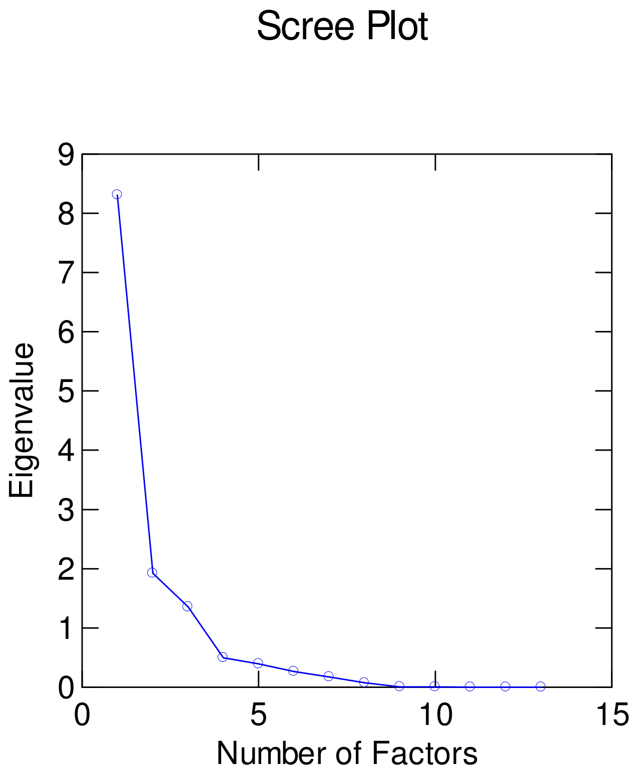
- We often ask: how different are two observations/subjects?
- If we have only measured one aspect of each subject (i.e., one variable), this is simply the difference between the two observations!
- If we have measured many variables and need **one *measure of dissimilarity***, we instead measure some form of **distance between the points representing the two observations** in a multivariate space described by the variables measured.
- The simplest is the **Euclidean distance** (the geometrical straight-line distance), but many other such measures of multivariate dissimilarity exists (*see page 409-415 in book and **lecture II***)!

<u>2. Principal component analysis (PCA)</u>

- It is common that we have "too many" variables to be able to make sense, or get an overview, of our data…
- We can then use a principal component analysis of our original $X_{1\text{-}p}$ variables to get a new set of $Z_{1\text{-}k}$ derived variables (where k << p), that are

called **principal components** (or factors) – very common and useful method.

- The orginal variables can be continuous, interval or ordinal variables or a mix of these!
- We typically retain only PC's with λ > 1 for further analysis/inspection (this rule of thumb can only be used when PCA is done of the correlation matrix though – see below); λ is often rescaled and expressed as "percent of variance in the original data matrix explained" by each PC.
- The PCs are extracted such that they **explain a maximum amount of variance in the orginal data** – often a few PC's can capture most of the variance among a large set of original variables! This is particularly true if many of the original variables are intercorrelated…**scree plots important and informative**!



- PC’s are **orthogonal** (=uncorrelated!) to one another by default – very useful property!
- The default is to use (1) “no rotation” and (2) base the PCA on the correlation matrix – **recommended** in most cases. Then, each variable influences the PCA equally. If you instead use the covariance matrix, variables influences the PCA in proportion to their absolute variance and if you allow some form of rotation (e.g. varimax) PCs will not be orthogonal.
- PC’s can be “understood” (i.e., interpreted) by looking at how each PC correlates with the original variables – these are called **principal component loadings** important to inspect – can be tested using resampling tests.

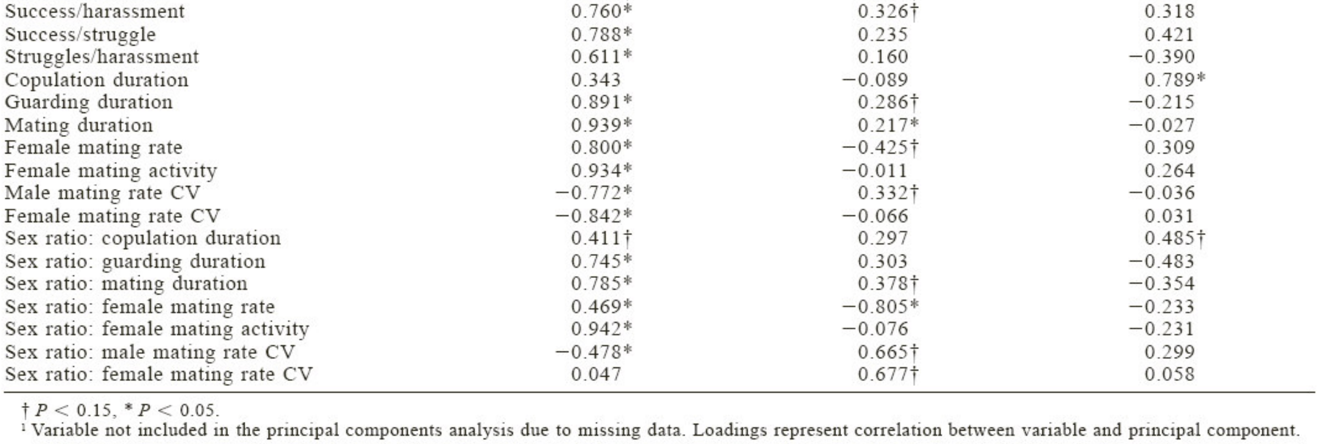


TABLE 4. Principal components analysis of mating behavior from 15 species of *Gerris*. The bottom seven (sex-ratio) variables were calculated as the difference in each behavior when tested at two sex ratios (see Materials and Methods). Collectively, the three principal components explain over 77% of the variance in the total dataset (50.77%, 15.37%, and 11.26%, respectively). Tests of significance of variable loadings represent frequency of loadings different in sign to the ones observed, among 100 bootstrap replicate analyses corrected for axis reversals (Mehlman et al. 1995).

Behaviors	Component loadings		
	1	2	3
Unsuccessful struggle duration	0.799*	−0.181	−0.046
Successful struggle duration[1]	0.733	0.413	0.297
Harassment index	0.287†	−0.625†	0.273
Success/harassment	0.760*	0.326†	0.318
Success/struggle	0.788*	0.235	0.421
Struggles/harassment	0.611*	0.160	−0.390
Copulation duration	0.343	−0.089	0.789*
Guarding duration	0.891*	0.286†	−0.215
Mating duration	0.939*	0.217*	−0.027
Female mating rate	0.800*	−0.425†	0.309
Female mating activity	0.934*	−0.011	0.264
Male mating rate CV	−0.772*	0.332†	−0.036
Female mating rate CV	−0.842*	−0.066	0.031
Sex ratio: copulation duration	0.411†	0.297	0.485†
Sex ratio: guarding duration	0.745*	0.303	−0.483
Sex ratio: mating duration	0.785*	0.378†	−0.354
Sex ratio: female mating rate	0.469*	−0.805*	−0.233
Sex ratio: female mating activity	0.942*	−0.076	−0.231
Sex ratio: male mating rate CV	−0.478*	0.665†	0.299
Sex ratio: female mating rate CV	0.047	0.677†	0.058

† $P < 0.15$, * $P < 0.05$.

[1] Variable not included in the principal components analysis due to missing data. Loadings represent correlation between variable and principal component.

- Very frequently used method in ecology and evolution; for example when we (1) wish to characterize a pattern of variation/covariation among a large set of variables, (2) wish to ordinate/compare sites/observations in fewer dimensions than our original variables (lecture 2), (3) have "too" many explanatory variables in multiple regression type models (to evade problems with collinearity) or (4) have many explanatory variables relative to our sample size. In (3) and (4), we first do a PCA and then use the resultant PC's as explanatory variables in our linear models.
- Will get back to PCA in lecture II…

3. MANOVA and related methods

- Sometimes we have **multiple response variables** in a design that would ordinarily have been analysed by ANOVA...
- To answer "What were the treatment effects on my response variables collectively?"; we then perform a **multivariate analysis of variance (MANOVA)** or, analogously, a multivariate analysis of covariance (MANCOVA) – can be one-way, two-way, etc., etc.
- First, we take the matrix of response variables and extract the linear combination Z of these (i.e., a derived variable) that maximizes the ratio of between-group and within-group variances of Z (a derived variable that is as much "affected" by our treatment as possible!).
- For one-way MANOVA, this derived variable Z is also called the **first discriminant function** (because it best "disciminates" between factor level groups).

- We then test for effects fo our treatment on *Z* using an "ANOVA-type" model, where our test statistic is typically **Wilk's lambda** (alternatively, Hotelling-Lawley trace or Pillai trace).
- If we have any effects; we then often use ordinary ANOVA's on single reponse variables to help us interpret our results.
- Example - MANCOVA on three response variables (A,B,X_0) using two factors and two continuous covariates:

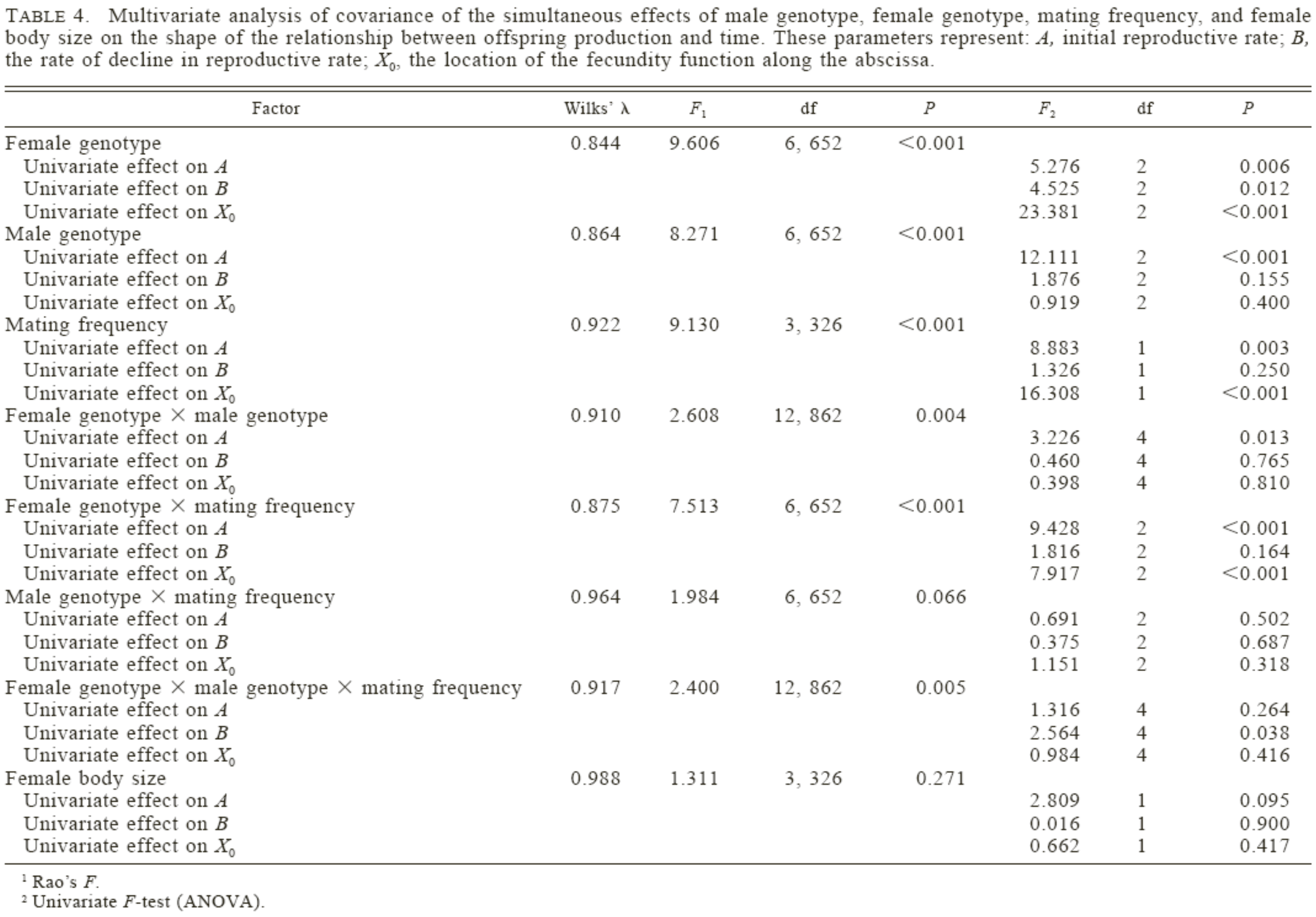


TABLE 4. Multivariate analysis of covariance of the simultaneous effects of male genotype, female genotype, mating frequency, and female body size on the shape of the relationship between offspring production and time. These parameters represent: *A*, initial reproductive rate; *B*, the rate of decline in reproductive rate; X_0, the location of the fecundity function along the abscissa.

Factor	Wilks' λ	F_1	df	*P*	F_2	df	*P*
Female genotype	0.844	9.606	6, 652	<0.001			
Univariate effect on *A*					5.276	2	0.006
Univariate effect on *B*					4.525	2	0.012
Univariate effect on X_0					23.381	2	<0.001
Male genotype	0.864	8.271	6, 652	<0.001			
Univariate effect on *A*					12.111	2	<0.001
Univariate effect on *B*					1.876	2	0.155
Univariate effect on X_0					0.919	2	0.400
Mating frequency	0.922	9.130	3, 326	<0.001			
Univariate effect on *A*					8.883	1	0.003
Univariate effect on *B*					1.326	1	0.250
Univariate effect on X_0					16.308	1	<0.001
Female genotype × male genotype	0.910	2.608	12, 862	0.004			
Univariate effect on *A*					3.226	4	0.013
Univariate effect on *B*					0.460	4	0.765
Univariate effect on X_0					0.398	4	0.810
Female genotype × mating frequency	0.875	7.513	6, 652	<0.001			
Univariate effect on *A*					9.428	2	<0.001
Univariate effect on *B*					1.816	2	0.164
Univariate effect on X_0					7.917	2	<0.001
Male genotype × mating frequency	0.964	1.984	6, 652	0.066			
Univariate effect on *A*					0.691	2	0.502
Univariate effect on *B*					0.375	2	0.687
Univariate effect on X_0					1.151	2	0.318
Female genotype × male genotype × mating frequency	0.917	2.400	12, 862	0.005			
Univariate effect on *A*					1.316	4	0.264
Univariate effect on *B*					2.564	4	0.038
Univariate effect on X_0					0.984	4	0.416
Female body size	0.988	1.311	3, 326	0.271			
Univariate effect on *A*					2.809	1	0.095
Univariate effect on *B*					0.016	1	0.900
Univariate effect on X_0					0.662	1	0.417

[1] Rao's *F*.
[2] Univariate *F*-test (ANOVA).

- Good to answer questions of **overall effects** – offers an “omnibus” test – but limited in terms of biological understandning and analytical resolution...
- A complemetrary approach sometimes used: first do a PCA of the reponse variables and then do ANOVA’s using PC’s as response variables. However, this has some serious limitations (MANOVA is generally preferable!), and is only really useful when one suspects a reponse in more than one multivariate “direction” or “dimension”.

Discriminant (function) analysis

- Analogous to a **one-way MANOVA**: is used to classify observations to groups along one dimension (i.e., one “factor”), based on linear combinations of variables, and assess the success of such classification. Loadings between orginal variables and DF’s – allows interpretaion of which variables contribute most to group differences. Also, multiple discriminant functions (orthogonal) can be extracted from any give data set.
- Ordination of objects in DF space allows visualization of patterns which can be useful (but see lecture II).
- Constrained to one-way designs (i.e., only one factor with k levels).

Example; 12 traits measured in replicated individuals from 18 populations (i.e., a single factor with 18 levels):

Classification matrix

Pop	%correct
1	27
2	0
3	40
4	42
5	50
7	56
8	0
9	22
10	50
11	38
12	23
13	50
14	100
15	33
16	50
17	50
18	27
Total	37

Eigenvalues (*=amount of variance between factor levels explained by the 10 first DFs)*

2.306967	0.469487	0.214942	0.168752	0.131747	0.121444

0.068914	0.054076	0.028249	0.007301

Canonical correlations (=*correlation between the discriminant function and the factor*)

0.835229	0.565235	0.420613	0.379982	0.341189	0.329079

0.253911	0.226498	0.165749	0.085138

Cumulative proportion of total dispersion *(=cumulative "importance" of DFs)*

0.645869	0.777309	0.837485	0.884730	0.921615	0.955615

0.974908	0.990047	0.997956	1.000000

```
      Wilks' lambda=      0.098
          Approx.F=      2.075  df= 160,     1061  p-tail=  0.0000

     Pillai's trace=      1.714
          Approx.F=      1.693  df= 160,     1310  p-tail=  0.0000

Lawley-Hotelling trace=      3.572

          Approx.F=      2.683  df= 160,     1202  p-tail=  0.0000
```

Canonical discriminant functions - standardized:

	1	2	3	4	5
RIGHTPADDLE	0.112836	1.527461	0.805605	0.19673	0.433967
LEFTPADDLE	0.309083	-0.57778	-0.07613	-0.25659	0.103518
BODYLENGTH	1.206225	2.122372	1.127032	-3.77088	0.868008
DORSAL1FIN	-0.2485	-1.48365	2.444111	0.153414	0.270195
DORSAL2FIN	-0.08018	0.084628	-2.19509	-0.97008	0.000327
TAILUPPERFI	0.268278	0.655319	-0.69472	-0.13649	-0.04053
TAILLOWERFI	0.012229	-0.78906	-0.43516	0.015047	-0.10085
ANALFINLEN	-0.33817	0.28846	-0.20447	0.810586	-1.60575
MEANPADDLE	.	.	.	.	.
SASPADDLE	.	.	.	.	.
ASPADDLE	0.204416	0.271626	-0.13659	0.194228	0.584736
CENTROID	-0.20731	-2.01509	-1.05039	3.910924	-0.58782

(*this shows the standardized discriminant coefficients, which are used to compare the relative importance of the independent variables to each DF - much as standardized betas are used to compare importance among explanatory variables in multiple regression!*)

Multivariate correlations

If we want to study and test for covariation between two sets, $X_{1\text{-}i}$ *and* $X_{j\text{-}m}$, of several variables, we use **canonical correlation or PLS analysis** (e.g., “are these 10 variables correlated with these 8 variables?”).

- A **canonical correlation** extracts one linear combination of set A and one for set B such that this pair of derived variables (=canonical variates) **correlates maximally with one another**. Several such pairs can be extracted (all orthogonal). **Generalized canonical correlation** analysis (gCCA), is a way of making sense of cross-correlations between more than two sets of variables.
- **Partial least-squares (PLS) analysis** is basically very similar to canonical correlations, but sets of derived variables not constrained to be orthogonal - very useful for multivariate predictions! (was invented by Herman Wold…used a lot in chemometrics, pioneered by his son Svante...).

Reading: Key chapters: 15.1-15.6, 16.1-16.4, 17.1-17.4, 18.1-18.4.
Course: A special PhD student course in multivariate methods for biologists is given each year at SLU (see http://info1.ma.slu.se/IMA/Utbildning/forskarutbildning_kurs1.html), typically during March – ***highly recommended*** for those who will work much with these types of methods – especially ordination techniques (lecture II).

Software: All "standard" statistic packages does all or most of these forms of multivariate analyses! For ordination and classification (lecture II), several special packages exists...