Biostatistics VHM 802 Course
Winter 2021, Atlantic Veterinary College, PEI
Henrik Stryhn

Index of Lecture 11: MANOVA and Ordination (with a few typos fixed)

PRACTICAL INFORMATION

Today's lecture — classical statistical material on multivariate ANOVA and regression, and an assortment of other statistical inference tools,

- ∗ MANOVA (multivariate ANOVA); Manly 3/4, Chapter 4,
- ∗ distances involving populations; Manly 3/4, Chapter 5 (material previously skipped),
- ∗ simulation-based tests; for most parts not in the Manly text,
- ∗ some brief comments on ordination; Manly 3/4, Chapter 12,

— this will be the multivariate lecture most focused on ***P***-values!

Other news:

- ∘ Schedule: the coming weeks are mixed with shared VHM 802/812 sessions,
  - ∘ regular Lab session 11 for this lecture on Monday,
  - ∘ first lecture on clustering (≠ cluster analysis!): synchronous session this Friday (by Javier), 9−11 am, connect by link at the VHM 812 Moodle site,
  - ∘ 2 recorded (short) video lectures (by me) to be posted next week,
- ∘ Home assignment 5 all received, just one to go (deadline after Easter!) — now is the time to make progress on your projects. . .

MULTIVARIATE REGRESSION: THE IDEA

Models/procedures similar to univariate regression/ANOVA for multidimensional outcomes, in order to:

– simultaneously test hypotheses for several outcomes,
thereby potentially increasing power and reducing problems with multiple testing of hypotheses for each outcome separately,

– construct estimates/tests involving relations between multiple outcomes, especially when outcomes are “related” (e.g. with repeated measures data[1]),

– estimate and possibly test hypotheses about the covariance/correlation between multiple outcomes,

⇒ classical mathematical/statistical theory based on multivariate normal (MVN) distribution with exact test distributions, however also with the limitations:

∗ requires complete data (usually, if one measure is missing, the entire observation/row drops out),

∗ difficult to generalize beyond assumptions of normality and independence across subjects, and MANOVA methods only generalize standard ANOVA methods.[2]

---

[1] Multivariate analysis of repeated measures data treats the series of observations over time on each subject as a multivariate outcome.

[2] The multilevel software package MLwiN offers multivariate analysis by approximation methods that allow for missing values and hierarchical data structure.

HOTELLING'S MULTIVARIATE T^2-TEST

Two samples on $X_1, \ldots, X_p$ of sizes n_1 and n_2 assumed to follow MVN distributions with mean vectors $\mu^{(1)}$ and $\mu^{(2)}$, and covariance matrices $\Sigma^{(1)}$ and $\Sigma^{(2)}$, respectively.[3]

(1) estimate the μ's by sample means ($\bar{\mathrm{X}}^{(j)}$) and the Σ's by empirical covariance matrices ($\mathrm{S}^{(j)}$), for $j = 1, 2$ (as described on slide 8L−13),

(2) the hypothesis of interest is $\mathrm{H}_0 : \mu^{(1)} = \mu^{(2)}$ (i.e., equal means for all variables),

(3) assume additionally equal variances ($\Sigma^{(1)} = \Sigma^{(2)}$), and estimate the joint covariance matrix Σ as:

$$\hat{\Sigma} = [(n_1-1)\mathrm{S}^{(1)} + (n_2-1)\mathrm{S}^{(2)}]/(n_1+n_2-2)$$

(4) compute Hotelling's T^2-statistic as the quadratic form,

$$T^2 = \frac{n_1 n_2}{n_1+n_2} \left(\bar{\mathrm{X}}^{(1)} - \bar{\mathrm{X}}^{(2)}\right)^t \hat{\Sigma}^{-1} \left(\bar{\mathrm{X}}^{(1)} - \bar{\mathrm{X}}^{(2)}\right),$$

and assess significance (after scaling[4]) in an F-distribution (p, n_1+n_2-p-1).

- ∗ similar to two-sample t-statistic (squared!) with equal variances,
- ∗ some robustness to model assumptions can be assumed (as with univariate t-statistics).

---

[3] We formulate the test in a two-sample context, although the principle applies to many "one-dimensional" hypotheses, e.g. also to testing a specific value for μ in a single sample.

[4] The F-statistic equals: $F = T^2(n_1+n_2-p-1)/[p(n_1+n_2-2)]$.

MULTIVARIATE EXAMPLE: SPARROWS

Summary: 5 measures on 49 sparrows; interest is in comparing survivors and non-survivors.

- table of univariate t-tests for means and Levene's test for variances for each variable, and multivariate T^2 test for means and Van Valen's test for variances[5]:

Variable	Non-survivors $\bar{X}$	Non-survivors s	Survivors $\bar{X}$	Survivors s	Means t (P)	Variances (P)
`len_total`	158.4	3.88	157.4	3.32	1.02 (.32)	(.24)
`ext_alar`	241.6	5.71	241.0	4.18	0.40 (.69)	(.24)
`len_beakhead`	31.48	.853	31.43	.729	0.20 (.84)	(.42)
`len_hum`	18.45	.659	18.50	.420	-0.35 (.73)	(.062)
`len_keelst`	20.84	1.15	20.81	.758	0.11 (.91)	(.17)
multivariate					2.82 (.76)	(.045)

  - Hotelling's $T^2 = 2.82 \Rightarrow F = 0.52 \sim F(5, 43)$ — totally non-significant,
  - all measures more variable for non-survivors
    — has biological interpretation as "stabilizing selection" (Manly),

- error correlations between measures (adjusted for survivor effects): range $(0.53, 0.77)$, very similar to simple (unadjusted) correlations.

---

[5] Manly describes this special test (a generalization of Levene's test; not available in Minitab/Stata and with a manual calculation in R) to specifically test whether variances (for multiple outcomes) are larger in one group than another.

MULTIVARIATE LINEAR MODEL

In notation analogous to ordinary linear models: $\mathrm{Y} = \mathrm{XB} + \mathrm{E}$,[6]

- model assumes same mean structure for all p measures,
- parameters B and Σ estimated by
  - $\hat{\mathrm{B}} = (\mathrm{X}^t\mathrm{X})^{-1}\mathrm{X}^t\mathrm{Y}$ ∼ least squares estimate,
  - $\hat{\Sigma} = \mathrm{S} = (\mathrm{Y} - \mathrm{X}\hat{\mathrm{B}})^t(\mathrm{Y} - \mathrm{X}\hat{\mathrm{B}})/(n-k)$ unbiased estimate,
- usual linear model methods (from estimates and SEs) for parameters for the p measures,
- formulate linear hypothesis H_0 about B, e.g. corresponding to equality between subject groups or to relations between measures,
- test of H_0:
  - 4 different test statistics: Wilk's lambda (likelihood ratio test), Pillai's trace, Hotelling-Lawley trace, and Roy's root statistic,
  - "one-dimensional" cases: agreement with Hotelling's T^2,
  - except in simple cases, their distribution is known only approximately (by suitable F-distributions, usually indicated in the software),
  - no test generally "best" (with respect to power and robustness), but Wilk's lambda considered to perform reasonably well in almost all situations.

---

[6] The specifics of the notation are:
- Y ($n \times p$): p-dimensional outcome on n subjects,
- X ($n \times k$): design matrix ∼ k "regression" parameters per outcome,
- B ($k \times p$): parameter matrix,
- E ($n \times p$): matrix of errors ∼ $\mathrm{MVN}(0, I_n \otimes \Sigma)$, where $I_n \otimes \Sigma$ is a block-diagonal covariance matrix ∼ assumed independence between subjects.

Multivariate Example: Skulls

Summary: 4 measures on 30 skulls from each of 5 periods.

- table of univariate means, SEs and F-tests for comparisons across all periods:

	Mean for period						
Variable	1	2	3	4	5	SE	F-test (P)
`breadth`	131.4	132.4	134.5	135.5	136.2	0.84	5.95 ($< .001$)
`hght_bas`	133.6	132.7	133.8	132.3	130.3	0.88	2.45 (.049)
`len_bas`	99.17	99.07	96.03	94.53	93.50	0.90	8.31 ($< .001$)
`hght_nas`	50.53	50.23	50.57	51.97	51.37	0.58	1.51 (.203)

  - strong significance for some measures, and most period differences seem to be time-ordered.

- multivariate tests for periods: all multivariate tests strongly significant ($P < .001$) — not too surprisingly with the strong differences for two variables,

- Box's M-test for variance homogeneity non-significant ($P = 0.25$)[7], same result for generalization of robust variance tests (details in Manly),

- error correlations between measures are quite weak (largest 0.22) ⇒ multivariate analysis less attractive in this example.

---

[7] This test generalizes Bartlett's test for equality of univariate variances, which due to its strong sensitivity to normal distribution assumptions is not generally recommended; the same concern applies to Box's M-test.

EXTRA NOTES ON MULTIVARIATE REGRESSION/ANOVA[8]

- Advantages over univariate analyses:
  - greatest with highly negatively correlated outcomes or moderate correlations among outcomes in both directions,
  - minimal with strongly positively correlated or uncorrelated outcomes,
- design requirement/recommendation: more subjects than outcomes in every "cell" formed by predictors[9] in order to ensure estimability and adequate power,
- sensitivity to assumptions: some robustness to normality assumption has been shown, but procedures can be very sensitive to outliers; robustness to equal variance assumption exists for roughly balanced designs,
- exclusion of redundant outcomes (due to high collinearity) is recommended, or use of multivariate dimension-reduction techniques,
- tools for exploration of relations among outcomes exist, with issues similar to exploring specific relations between predictor categories (e.g. multiple testing),
- MANOVA designs can be as complex as ANOVA designs (of course) and will then require understanding of experimental design to be analyzed properly. . . ,
- this is a pretty large topic, and we have just scratched the surface. . .

---

[8] Based largely on TF: Sections 7.3 and 7.5.

[9] For example, the `sparrows` design had 2 cells, and the `skulls` design had 5 cells.

MORE MULTIVARIATE DISTANCES

We already (Session 9) discussed distances between points, i.e. p-dimensional observations of the form $\mathrm{X} = (X_1 \ X_2 \ \ldots X_p)$.

If a population is represented by p-dimensional distribution with multivariate mean vector μ and variance-covariance matrix Σ, we can define other distances:

- Mahalanobis distance between an observation X and a population (μ, Σ):

  $$\begin{aligned} d_M(\mathrm{X}, \mu, \Sigma) &= (\mathrm{X}-\mu)^t \Sigma^{-1} (\mathrm{X}-\mu) \\ &= \textstyle\sum_{i=1}^{p} \sum_{j=1}^{p} (X_i - \mu_i) v_{ij} (X_j - \mu_j), \end{aligned}$$

  where v_{ij} is the $(i,j)^{th}$ element of Σ^{-1}; thus, d_M is a quadratic form,

- distance between 2 populations: based on mean distance and assuming a common ("pooled") variance-covariance matrix Σ:
  - Mahalanobis distance between means: $d_M(\mu^{(1)}, \mu^{(2)}, \Sigma)$,
  - Penrose distance between means:

    $$d_P(\mu^{(1)}, \mu^{(2)}, \Sigma) = \frac{1}{p} \sum_{i=1}^{p} \frac{\left(\mu_i^{(1)} - \mu_i^{(2)}\right)^2}{\sigma_{ii}},$$

    i.e., the sum of mean squared differences weighted by the inverse variances, without accounting for correlations between the variables.[10]

---

[10] The formula for d_P is simpler and requires less information; Manly states d_M to be preferable to d_P for strongly correlated variables and when sample size exceeds 100.

MAHALANOBIS DISTANCE AND OUTLIERS

Idea: use Mahalanobis distance to detect multivariate outlying observation(s) among $\mathrm{X}_1, \ldots, \mathrm{X}_n$,

- compute sample mean $\hat{\mu} = \bar{\mathrm{X}}$ and sample covariance matrix $\hat{\Sigma}$,
- for each observation i, compute $d_M(\mathrm{X}_i, \hat{\mu}, \hat{\Sigma}) \equiv d_i$ and declare observation as outlier if d_i is "large".

Two useful results:

- if $\mathrm{X}_1, \ldots, \mathrm{X}_n$ are i.i.d. from $\mathrm{MVN}(\mu, \Sigma)$, then $d_i \approx \chi^2(p)$ distribution,[11]
- if the p variables of $\mathrm{X}_1, \ldots, \mathrm{X}_n$ are used as predictors in a linear regression, then:

$$d_i^2 = (n-1) \times (h_i - \tfrac{1}{n}), \quad \text{for } i = 1, \ldots, n$$

  where h_i is the leverage; thus, d_i gives essentially the same information as h_i.

Recommendation: may use Mahalanobis distance to screen for "suspected outliers", but more sophistication than a $\chi^2(p)$-percentile is required for a "rule".[12]

Example — Sparrow data: Minitab outlier plot (in PCA menu) based on Mahalanobis distance[13]: most extreme observation (number 31) has low value of `len_beakhead` and highest value of `len_keelst`.

---

[11] Based on this result, TF recommend the 99.9% percentile of $\chi^2(p)$ as outlier cutoff — "with caution".

[12] Filzmoser et al. (2005), Multivariate outlier detection in exploration geochemistry, *Comput. Geosci.* 31, 579−587.

[13] Cut-off for high values computed as $\sqrt{p\,F_{.95}(p, n-p-1)}$, with $p = 5$ and $n = 49$.

DISTANCES BETWEEN POPULATIONS: SKULLS

Summary: 4 measures on 30 skulls from each of 5 periods; interest is in quantifying distances between populations ∼ periods.

First step: estimate means $\hat{\boldsymbol{\mu}}^{(k)} = \bar{\mathrm{X}}^{(k)}$ and variances $\hat{\Sigma}^{(k)} = \boldsymbol{S}^{(k)}$ for each period $\boldsymbol{k} = 1, \ldots, 5$, and estimate joint variance matrix by pooling: $\boldsymbol{S} = \frac{1}{5}(\boldsymbol{S}^{(1)} + \ldots + \boldsymbol{S}^{(5)})$.

Second step: for each pair $(\boldsymbol{k}, \boldsymbol{l})$ of populations $(\boldsymbol{k}, \boldsymbol{l} = 1, \ldots, 5)$, calculate $\boldsymbol{d}_M$ and $\boldsymbol{d}_P(\hat{\boldsymbol{\mu}}^{(k)}, \hat{\boldsymbol{\mu}}^{(l)}, \mathrm{S})$ distances, and display in a distance matrix.[14]

Distance matrices:
Penrose (upper triangle),
Mahalanobis (lower triangle):

Populations	Population number 1	2	3	4	5
1. Early predynastic	–	0.023	0.216	0.493	0.736
2. Late predynastic	0.091	–	0.163	0.404	0.583
3. 12-13th dynasties	0.903	0.729	–	0.108	0.244
4. Ptolemaic	1.881	1.594	0.443	–	0.066
5. Roman	2.697	2.176	0.911	0.219	–

- closest distances between adjacent populations (as expected),
- despite on different scales (note: the $\boldsymbol{d}_P$ formula divides by $\boldsymbol{p} = 4$), fairly good agreement between Penrose and Mahalanobis distances (maybe not too surprisingly, as the 4 measures are little correlated).

---

[14] Minitab has no built-in function for $\boldsymbol{d}_P$, but the calculations are univariate summations. A Discriminant Analysis gives between-group Mahalanobis distances. MANOVA shows S (multiplied by df as the SSCP matrix) and stores the residuals.

SIMULATION-BASED STATISTICAL TESTS

Synthesis: Permutation or Randomization tests are non-parametric tests that use simulation to calculate P-values for usual statistical tests (with a null hypothesis H_0, an alternative hypothesis H_a, a significance level, errors of type I and II etc.).

Idea: Instead of relying on a (reference) distribution (exact or approximate), e.g. a t-distribution, that the test statistic follows under H_0 (i.e., assuming H_0 is true), that distribution is approximated by simulation, in total comprising the following steps:

- compute value of desired test statistic, say T_{obs},[15]
- simulate[15] a large number, say N, of pseudo-observations of T under H_0, say $T_1, \ldots, T_N$,
- calculate the P-value as the proportion among (T_i) that are as extreme as or more extreme than T_{obs}, e.g. for a one-tailed calculation: $P = \Pr(T_i \geq T_{\text{obs}})$,
- interpret P the usual way (with the usual caveats around statistical testing).

Example — Sparrow data: differences in means for `total_length` between survivors and non-survivors:

- two-sample t-test for $\mathrm{H}_0 : \mu_1 = \mu_2$: $t = 1.02$, $P = 0.315$ from $t(31)$-distribution,
- permutation test for "equal means": $T_{\text{obs}} = 1.048$ (mean difference), $P = 0.323$ from $N = 1000$ permutations (random divisions of 49 sparrows into 21×S and 28×NS).

---

[15] The choice of test statistic, and how the simulation is conducted, depend crucially on the data and on the nature of H_0. The resulting P-value is random, with standard error $\approx \sqrt{P(1-P)/N}$; usually N is at least 1000.

MANTEL TEST FOR MATRICES

Problem: How to quantify (and assess) similarity/associations between two matrices (of the same size)?

- Idea: compute correlations[16] between relevant matrix elements[17],
- use permutation test, where permutations are of row (columns) of one of the matrices[18] — the Mantel test.[19]

Example — Skull data: compare multivariate distances between periods (d_P or d_M) to actual (approximate) time distances[20]: a total of $5! = 120$ permutations of the rows, and the observed correlations (0.954 and 0.964) ∼ the upper 98.3% percentile among the permuted matrices. Thus P appears to be small.

Concerns[21] with the validity of the P-values from the Mantel test:

- the permutations may not realistically reflect the variation for unrelated matrices, when both matrices are structured,[22]
- a particular concern, when the matrices have a structure ∼ autocorrelation.

---

[16] Usually Pearson correlations, but there is no reason one could not use other measures of association.

[17] Which matrix elements are relevant, depend on the type of matrix (e.g., symmetrical matrix, correlation matrix).

[18] Under the null hypothesis of no association between the matrices, the rows (columns) of the matrices are not associated and one set of these can be permuted.

[19] In a partial Mantel test, a third variable/matrix Z is involved, and partial correlations after adjusting for Z are used.

[20] For example, 700 years between period 1 (Early predynastic, 4000 BC) and period 2 (Late predynastic, 3300 BC).

[21] Following Guillot & Rousset (2013), Dismantling the Mantel tests, *Meth. Ecol. Evol.* 4, 336−344.

[22] The problem is that the permutations of rows may break that structure, and hence produce unrealistic matrices.

ANALYSIS OF SIMILARITY (`ANOSIM`)

Problem: How to compare/assess known groups on the basis of distance/dissimilarity between individual points (but not data values behind them)? — may occur, if

- the data (and distances) are spatial,
- high-dimensional input data are expressed by dissimilarity measures,
- high-dimensional data are represented by a few coordinates of MDS, which are used to calculated distances.

Idea: compare dissimilarity in observation pairs within and between groups, and assess resulting statistic by a permutation test where points are permuted among groups.[23]

Steps of `ANOSIM` procedure:[24]

- compute the average rank dissimilarity within (r_W) and between (r_B) groups — note the use of ranks to reduce any scale dependence,
- the test statistic $R = 2(r_B - r_W)/[n(n-1)/2]$ takes values: $-1 \leq R \leq 1$, with large values $\sim$ more similarity (less dissimilarity) within groups,
- compute P as the proportion of permutations exceeding observed value (one-sided).

---

[23] Similar idea as a one-way ANOVA permutation test, where variation within and between groups are compared by the usual F-statistic, with a simulation-based reference distribution.

[24] For a one-way layout, as described in Clarke (1993), Non-parametric multivariate analyses of changes in community structure, *Austral. Ecology* 18, 117−143, which also sketches some (limited) extensions to two-way layouts.

Permutational Analysis of Variance (`PERMANOVA`)

Problem: How to separate effects of multiple factors, as in (M)ANOVA, again on the basis of distance/dissimilarity between individual points (as on previous slide)?

Idea: replace the sums of squared deviations (from means) in ANOVAs by sums of squared distances to obtain similar decompositions for different effects, then use formulae similar to ANOVA for test statistics as ratios between normalized squared distance terms, and assess resulting statistics with suitable permutation tests.[25]

Steps of `PERMANOVA` procedure:[26]

- compute the total sum of squared distances: $\mathrm{SS}_T = \frac{1}{n} \sum_{i<j=1}^{n} d_{ij}^2$, and obtain by restricting the within-group sums to same factor levels the decompositions,[27]

$$\begin{aligned} \text{one-way}: \ \mathrm{SS}_T &= \mathrm{SS}_G + \mathrm{SS}_E, \quad \text{(between-group and within-group sums)} \\ \text{two-way}: \ \mathrm{SS}_T &= \mathrm{SS}_A + \mathrm{SS}_B + \mathrm{SS}_{AB} + \mathrm{SS}_E, \quad \text{(also an interaction sum)} \end{aligned}$$

- define analogous test statistics as in univariate ANOVAs, e.g. in a one-way layout as $F = [\mathrm{SS}_G/\mathrm{DF}_G]/[\mathrm{SS}_E/\mathrm{DF}_E]$,
- compute P as the proportion of (suitably defined[25]) permutations, where F exceeds the observed value (one-sided).

---

[25] Permutation tests for higher-order ordinary ANOVAs are not so simple to conduct, see e.g. Manly (2006), *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 3rd ed., CRC Press.

[26] For one- and two-way layouts, as described in Anderson (2001), A new method for non-parametric multivariate analysis of variance, *Austral. Ecology* 26, 32 − 46.

[27] In order to enable the decompositions, the sums must be normalized appropriately by the design dimensions.

SIMULATION-BASED ANOVA EXAMPLE: SPARROWS

Reference comparisons between survivorship groups (slide 11L−4):

- `len_total` only: $\boldsymbol{t = 0.993}$, $\boldsymbol{P = 0.326}$ (equal variances, two-sided alternative),
- multivariate: $\boldsymbol{T^2 = 2.82 \sim F = 0.52}$, $\boldsymbol{P = 0.76}$ (two-sided alternative),
- standardization (for multivariate distances) does not affect results.

Analysis of similarity (`ANOSIM`) to compare survivorship groups:

- `len_total` only: $\boldsymbol{R = -0.006}$, $\boldsymbol{P = 0.47}$ for Euclidean and Manhattan ($\boldsymbol{L_1}$) distances,
- multivariate: $\boldsymbol{R = -0.038}$, $\boldsymbol{P = 0.93}$ for Euclidean distance; $\boldsymbol{R = -0.043}$, $\boldsymbol{P = 0.96}$ for Manhattan distance.

Permutational analysis of variance (`PERMANOVA`) to compare survivorship groups:

- `len_total` only: $\boldsymbol{F = 0.986 = t^2}$, $\boldsymbol{P = 0.30}$ for both distances,
- multivariate: $\boldsymbol{F = 0.317}$, $\boldsymbol{P = 0.74}$ for Euclidean distance; $\boldsymbol{F = 0.203}$, $\boldsymbol{P = 0.76}$ for Manhattan distance (note that $\boldsymbol{F \not\sim F}$-distribution).

Interpretations:

- no new insights by simulation-based ANOVAs (none of which are designed to utilize correlations), but `PERMANOVA` results very close (as expected),
- differences in dispersion between survivor groups not picked up.

ORDINATION REVISITED

Meaning/Definition according to Manly (Chapter 12):[28]

> [. . .] the process of producing a small number of variables that can be used to describe the relationship between a group of objects.

Methods typically referred to under ordination:

- principal coordinates analysis = classical (metric) MDS based on a distance matrix (and = PCA for Euclidean distance between variables; Sessions 9−10),
- modern MDS with a stress function (Session 9),
- correspondence analysis: a dimension-reduction technique akin to PCA and factor analysis but for categorical variables, apparently also popular with species abundance data (Manly 3/4, Section 12.5) and in different versions (e.g. "detrended correspondence analysis"),
- Manly also mentions discriminant analysis and canonical correlations (Session 12).

Distance measures (Session 9) are integral in all ordination; commonly used is the Bray-Curtis dissimilarity index (based on species counts).

---

[28] Other explanations exist, e.g. (from `http://ordination.okstate.edu/`): "The term 'ordination' derives from early attempts to order a group of objects, for example in time or along an environmental gradient. Nowadays the team is used more generally and refers to an 'ordering' in any number of dimensions (preferably few) that approximates some pattern of response of the set of objects. The usual objective of ordination is to help generate hypotheses about the relationship between the species composition at a site and the underlying environmental gradients."