Index of 3-L

Page	Title
1	Practical information
2	Multivariate regression: the idea
3	Hotelling's multivariate t-test
4	Multivariate example: sparrows
5	Multivariate linear model
6	Multivariate example: skulls
7	Extra notes on multivariate regression/ANOVA
8	PCA: the idea
9	PCA: the steps
10	PCA example: sparrows
11	PCA example: employment
12	PCA: limitations and practical issues

Practical Information

WELCOME (again) to all of you...

Registration for graduate students for VHM 881 (Directed studies): currently at 5 students, we may need some more students to register ...

Webpage changes:

- datasets: added link to repository at UC Irvine.
- added link for Minitab projects from sessions.

Today's lecture:

- multivariate analysis of variance and regression (Manly: Chapter 3(4); TF: Chapter 7; VER2 Section 23.2.3), often referred to by the acronym MANOVA,[1]
  - data examples: sparrows, skulls,
- principal components analysis (Manly: Chapter 5(6), TF: Chapter 13 (with factor analysis)), often referred to by the acronym PCA,
  - data examples: sparrows, employment.

[1] Similarly, one may use MANCOVA as the multivariate equivalent of ANCOVA, but in our view this is largely redundant because ANCOVA should be considered as a topic under multiple regression.

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, e.g. with repeated measures data,[2]
- construct estimates/tests involving relations between multiple outcomes, espec. when outcomes are "related",[2]
- estimate and possibly test hypotheses about the covariance/correlation between multiple outcomes.

$\Rightarrow$ 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 subject drops out),
* difficult to generalize beyond assumptions of normality and independence across subjects.[3]

[2] Multivariate analysis of repeated measures data treats the series of observations over time on each subject as a multivariate outcome, see e.g. Davis (2000), *Statistical Methods for Repeated Measurements*, Chapters 3-4. In this context, time effects are expressed as hypotheses across the multiple measures per subject.

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

Hotelling's Multivariate T-Test

Consider two samples on $X_1, \ldots, X_p$ of sizes n_1 and n_2 assumed to follow MVN distributions with mean vectors $\mu^{(1)} = (\mu_1^{(1)}, \ldots, \mu_p^{(1)})^t$ and $\mu^{(2)}$, and covariance matrices $\mathbf{\Sigma}^{(1)}$ and $\mathbf{\Sigma}^{(2)}$, respectively. [4]

(1) estimate the μ's by sample means ($\bar{\mathbf{X}}^{(j)}$) and the $\mathbf{\Sigma}$'s by empirical covariance matrices ($\mathbf{S}^{(j)}$), for $j = 1, 2$ (as described on slide 2L−5),

(2) the hypothesis of interest is $\mathrm{H}_0 : \mu^{(1)} = \mu^{(2)}$ (i.e., equality of the means for all variables $X_1, \ldots, X_p$),

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

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

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

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

and evaluate its significance (after suitable scaling[5]) in an F-distribution (p, n_1+n_2-p-1).

Notes:

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

[4] 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.

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

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[6]:

Variable	non-survivors $\bar{X}$	s	survivors $\bar{X}$	s	means t (P)	var. (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.

[6] Manly describes this special test (a generalization of Levene's test; not available in standard software) to specifically test whether variances (for multiple outcomes) are larger in group than another; consult Section 4.6 (3.6) for details.

Multivariate Linear Model

In notation analogous to ordinary linear models: $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$,[7]

- model assumes *same mean structure* for all p measures,
- parameters $\mathbf{B}$ and $\mathbf{\Sigma}$ estimated by
  - $\hat{\mathbf{B}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y} \sim$ least squares estimate,
  - $\hat{\mathbf{\Sigma}} = \mathbf{S} = (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})^t(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})/(n-k)$ unbiased estimate,
- usual linear model methods for parameters for the p measures,
- formulate linear hypothesis H_0 about $\mathbf{B}$, 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 w. 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" (power/robustness), but Wilk's lambda considered to perform reasonably well in almost all situations.

[7] The specifics of the notation are:

- $\mathbf{Y}$ $(n \times p)$: p-dimensional outcome on n subjects,
- $\mathbf{X}$ $(n \times k)$: design matrix $\sim k$ "regression" parameters per outcome,
- $\mathbf{B}$ $(k \times p)$: parameter matrix,
- $\mathbf{E}$ $(n \times p)$: matrix of errors $\sim \mathrm{MVN}(0, I_n \otimes \Sigma)$, where $I_n \otimes \Sigma$ is a block-diagonal covariance matrix $\sim$ 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 period comparisons:

Variable	period mean 1	2	3	4	5	SE	F-test (P)
`breadth`	131.4	132.4	134.5	135.5	136.2	0.84	5.95 (.000)
`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 (.000)
`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 tests strongly significant ($P < .001$) — not too surprisingly,
- Box's M-test for variance homogeneity non-significant[8], same result for generalization of robust variance tests (Manly),
- error correlations between measures quite weak (largest 0.22) $\Rightarrow$ multivariate analysis less attractive.

[8] 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[9]

- 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[10] 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've just scratched the surface. . .

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

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

PCA: the Idea

PCA is a mathematical[11] procedure to represent the "information" contained in a number (p) of variables $X_1, \ldots, X_p$, with a given covariance/correlation structure, by a new set $Z_1, \ldots, Z_p$ of variables such that

- each Z_j is a linear combination of the X_k's,
- the variables $Z_1, \ldots, Z_p$ are independent (uncorrelated/orthogonal),
- the total "variance" explained by the Z_j's equals that explained by the X_j's, and the Z_j's can be ordered by decreasing "variance" explained.

First interpretations:

* the *principal components* $(Z_j) \sim$ independent "directions" among the (X_j),
* no loss of information by switching: $(X_j) \mapsto (Z_j)$,
* the first (few) components $Z_1, Z_2, \ldots$ that explain most of the "variance",
  - may have useful subject-matter interpretations,
  - may represent a useful data reduction.

[11] PCA is not based on a statistical model, and does not in itself involve statistical inference.

PCA: the Steps

Given a covariance matrix[12] $\mathbf{S}$ $(p \times p)$ for the multivariate vector $(X_1, \ldots, X_p)^t$, for each $j = 1, \ldots, p$:
Z_j is an eigenvector for the jth eigenvalue λ_j of $\mathbf{S}$, where

* eigenvalues ordered as: $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p \geq 0$, and $\sum_k \lambda_k = \mathrm{tr}(\mathbf{S})$,
* $\mathrm{Var}(Z_j) = \lambda_j$, (i.e., Z_j has variance λ_j)
* Z_j is expressed from the (X_k) as:

$$Z_j = \sum_k a_k^{(j)} X_k = a_1^{(j)} X_1 + a_2^{(j)} X_2 + \ldots + a_p^{(j)} X_p \,, \qquad (1)$$

where the coefficients (also loadings) $(a_1^{(j)}, \ldots, a_p^{(j)})$ are unique up to a sign change, and $\sum_k (a_k^{(j)})^2 = 1$.

Notes and interpretations:

- "small" loadings (e.g. $|a_k^{(j)} \leq 0.3|$) are often disregarded for interpretations of the principal components,
- by *standardizing* the (X_k) prior to PCA, all X_k are brought on same scale, and the decomposition is of the correlation matrix $\mathbf{R}$,[13]
- a *scree plot* of cumulative eigenvalues $\sum_{k \leq j} \lambda_k$ against component number (j) is a simple (subjective) visual tool to decide which components to explore further,
- with data on $X_1, \ldots, X_p$, the scores for the principal components are computed from equation (1) for each observation,
  * scores may e.g. be used as predictors in subsequent regression.

[12] Typically $\mathbf{S}$ would be the empirical covariance matrix based on (say) n observations of $(X_1, \ldots, X_p)^t$, but in principle the matrix could have another origin.

[13] Working with the correlation matrix is generally recommended, in particular if the (X_k) show differences in variance; note that $\mathrm{tr}(\mathbf{R}) = p$.

PCA Example: Sparrows

Summary: 5 measures on 49 sparrows; survival (0/1) not part of PCA.

- variances: range $(0.32, 25.7)$; correlations: range $(0.53, 0.77)$
  $\Rightarrow$ potential for reducing data dimension; should work with correlation matrix $\mathbf{R}$,
- table of eigenvalues and loadings for $\mathbf{R}$ (values from Minitab):

Parameter/ Variable		component number j				
		1	2	3	4	5
eigenvalue	λ_j	3.62	0.53	0.39	0.30	0.16
cumulative	$\sum_{k \le j} \lambda_k / 5$	0.72	0.83	0.91	0.97	1.00
`len_total`	X_1	0.45	0.05	0.69	-0.42	-0.37
`ext_alar`	X_2	0.46	-0.30	0.34	0.55	0.53
`len_beakhead`	X_3	0.45	-0.33	-0.45	-0.61	0.34
`len_hum`	X_4	0.47	-0.19	-0.41	0.39	-0.65
`len_keelst`	X_5	0.40	0.88	-0.18	0.07	0.19

  - 1st eigenvalue by far strongest, others "ignorable",
  - 1st component $\sim$ bird size (average of measures),
  - 2nd component $\sim$ shape feature (X_5 vs. X_2, X_3, X_4), mostly determined by X_5 (visible in loading plot),
- scores for components (e.g. 1st vs. 2nd) plotted against each other (score plot), with survival indicator:
  - no obvious mean differences between survivor groups,
  - indication of larger spread among non-survivors,
- use of scores as uncorrelated predictors in logistic regression for survival: no significant differences between groups.

PCA Example: Employment

Summary: 9 percentages on 30 countries; the country grouping is not part of PCA.

- variances: range $(0.39, 151)$; correlations: range $(-0.81, 0.47)$; negative correlations expected from percentages adding up to 100% $\Rightarrow$ should work with correlation matrix $\mathbf{R}$; one eigenvalue will be zero (but no need to exclude variables apriori),
- table of (first 5) eigenvalues and loadings (values from Minitab):

Parameter/ Variable		component number j				
		1	2	3	4	5
eigenvalue	λ_j	3.11	1.81	1.50	1.06	0.71
cumulative	$\sum_{k \le j} \lambda_k / 9$	0.35	0.55	0.71	0.83	0.91
AGR	X_1	-.51	.02	-.28	-.02	.02
MIN	X_2	-.38	-.00	.52	-.11	-.35
MAN	X_3	.25	-.43	-.50	-.06	.23
PS	X_4	.32	-.11	-.29	-.02	-.85
CON	X_5	.22	.24	.07	-.78	-.06
SER	X_6	.38	.41	.07	-.17	.27
FIN	X_7	.13	.53	-.10	.49	-.13
SPS	X_8	.43	-.06	.36	.32	.05
TC	X_9	.21	-.52	.41	.04	.02

  - 4 eigenvalues above 1: not clear how many to include (slowly decreasing scree plot),
  - several directions (e.g., 2-4) effectively only involve a few variables; however, 1st direction almost involves all variables,
  - score plots and loading plots for pairs of components probably necessary to understand the decomposition obtained.

PCA: Limitations and Practical Issues[14]

- sensitivity to problems with correlations:
  - outlying observations,
  - non-linear associations,
  - unnatural ranges arising from sampling (too narrow or too wide),
- sample size recommendations: "good" starting at 300 observations, under certain circumstances "acceptable" down to 50,
- normality: no specific assumption (beyond quantitative scale) involved for PCA, but correlations will be most meaningful for continuous variables with roughly symmetrical distributions,
- outliers among cases: assessment may be based on multivariate distances (Session 5),
- outliers among variables: a variable with low correlations with all other variables may be ignored in PCA,[15]
- known structure: "PCA will reveal the gross features of the data, which may already be known, and is often best applied to residuals after the known structure has been removed".[16]

[14] Based largely on TF: Section 13.3.2.

[15] In the extreme case of a variable being uncorrelated with all other variables, it will retain its own component in PCA.

[16] Venables & Ripley (2000), *Modern Applied Statistics with S-Plus*, 3rd ed.