Index of 5-L

Practical Information

WELCOME (again) to all of you...

Webpage changes:

- added quiz on PCA and FA,
- added data sets used in factor analysis lecture.

Today's lecture:

- PCA and FA example/quiz: brief review,
- multivariate distances (Manly: Chapter 4(5)) of different types and with different uses, most importantly
  - Euclidean distance in cluster analysis,
  - Mahalanobis distance in discriminant analysis and for multivariate outlier detection,
  - data examples: dogs, sparrows, skulls,

  — some other topics not included here, in particular
  - Mantel randomization test to compare two distance/similarity matrices,
- discriminant analysis (Manly: Chapter 7(8), TF: Chapter 9)
  - includes linear discriminant analysis and a comparison with logistic regression,
  - data examples: sparrows, skulls, employment.

Multivariate Distances: Overview

Consider p-dimensional observations of the form $\mathbf{X} = (X_1\ X_2\ \ldots X_p)^t$, and populations represented by p-dimensional distributions with multivariate mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, assumed invertible. Definitions of distance:

- Euclidean distance between two observations $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$:

$$d_E(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}) = \sqrt{\sum_{i=1}^{p} \left(X_i^{(1)} - X_i^{(2)}\right)^2},$$

  the generalization of two/three-dimensional distance; other distances exist but this is the most intuitive and most commonly used one,

- Mahalanobis distance between an observation $\mathbf{X}$ and a population $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

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

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

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

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

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

Euclidian Distance: Dogs

Summary: 6 jaw measurements on craniums of prehistoric dogs in Thailand and 6 other possibly related species; interest is in quantifying distances between species.

First step: standardize the measurements/variables by subtracting their mean and dividing by their stand. dev.?

* necessary to give variables same weight in distances, but makes interpretation more difficult,
* variable ratios between stand. deviations (range ≈ 1–5) $\Rightarrow$ probably best to standardize.

Second step: calculate d_E between all pairs of observations (species), and collect in a distance matrix:[1]

Species	Species no. 1	2	3	4	5	6	7
1. Modern dog	–						
2. Golden jackal	1.91	–					
3. Chinese wolf	5.38	7.11	–				
4. Indian wolf	3.38	5.06	2.14	–			
5. Cuon	1.51	3.19	4.57	2.91	–		
6. Dingo	1.56	3.18	4.21	2.20	1.67	–	
7. Prehistoric dog	0.66	2.39	5.12	3.24	1.26	1.71	–

* in distance matrix, diagonal could be filled with 0s, and upper triangle could be filled as well (by symmetry),
* clearly "closest" relative to Prehistoric dog: Modern dog (0.66); same conclusion without standardization.

[1] Minitab: distance matrix from `Multivariate-Cluster Obs.-Storage` menu.

Mahalanobis Distance and Outliers

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

* compute sample mean $\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}}$ and sample covariance matrix $\hat{\boldsymbol{\Sigma}}$ (Lecture 2),
* for each observation i, compute $d_M(\mathbf{X}_i, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) \equiv d_i$ and declare observation as outlier if d_i is "large".

Two useful results:

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

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

  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".[3]

Sparrow data: Minitab outlier plot (in PCA menu) with $F_{.95}(p, n-p-1)$ "cutoff" for high Mahalanobis distance: most extreme obs. (no. 31) has low value of `len_beakhead` and highest value of `len_keelst`.

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

[3] Filzmoser et al. (2005), Multivariate outlier detection in exploration geochemistry, *Computers & Geosciences* **31**, 579-587.

Mahalanobis Distance: Skulls

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

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

Second step: for each pair (k, l) of populations $(k, l = 1, \ldots, 5)$, calculate d_M and $d_P(\hat{\boldsymbol{\mu}}^{(k)}, \hat{\boldsymbol{\mu}}^{(l)}, \mathbf{S})$ distances $\sim$ distance matrix.[4]

Distance matrices: Penrose (upper) and Mahalanobis (lower):

	Population no.				
Populations	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, fairly good agreement between Penrose and Mahalanobis distances.

Penrose vs. Mahalanobis distances for comparing populations:

- Penrose is simpler and does not require estimation of a full covariance/correlation matrix,
- Mahalanobis preferable for strongly correlated variables, and when sample size exceeds 100 (Manly).

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

Classification/Discrimination: Overview

Assume p-dimensional observations of the form $\mathbf{X} = (X_1\ X_2\ \ldots X_p)^t$ on observations for which a (true and perfect) classification into m groups exist; interest is in developing a *classification rule* based on $\mathbf{X}$ to "predict" group membership, in order to

- classify (probabilistically) future observations,
- obtain insight into structure involved in group membership (e.g., associations with $\mathbf{X}$-components).

The performance of the classification rule may be examined on "validation data" after it has been developed on "learning data".

Many approaches exist for such problems, e.g.[5]

* logistic or polytomous/multinomial regression models,
* (linear) discriminant (function) analysis (commonly LDA),
* classification trees,
* neural networks,
* support vector machines.

Contents in course — linear discriminant analysis, the classical method...

- starting with 2 groups (Sparrow data), despite the method not really being recommended with 2 groups[6], and comparing with logistic regression,
- discussing > 2 groups relatively briefly from examples.

[5] An excellent reference is: Hastie, Tibshirani & Friedman (2009), *The Elements of Statistical Learning*, 2nd ed., Springer; freely available at `http://statweb.stanford.edu/~tibs/ElemStatLearn/download.html`.

[6] Press & Wilson (1976), Choosing between logistic regression and discriminant analysis, *J. Amer. Statist. Assoc.* **73**, 699-705.

LDA and Logistic Classification: Sparrows

Summary: 5 measures on 49 sparrows, of which 21 (42.9%) survived.

Estimation − logistic regression and LDA:

Variable	logistic regression	LDA (groups) 0	1	diff
intercept	13.58	-1348.98	-1335.66	13.32
`len_total`	-0.163	5.692	5.537	-0.155
`ext_alar`	-0.028	8.025	7.998	-0.027
`len_beakhead`	-0.084	25.317	25.225	-0.093
`len_hum`	1.062	-38.637	-37.605	1.0325
`len_keelst`	0.072	-10.874	-10.804	0.069
prediction for $\mathbf{X} = (152, 240, 31, 18.3, 21)$	$\text{logit}(\hat{p}) = 0.575$ $\hat{p} = 0.64$	$d_M(\mathbf{X}, \boldsymbol{\mu}^{(0)}) = 7.040$ $\hat{p}_0 = 0.30$	$d_M(\mathbf{X}, \boldsymbol{\mu}^{(1)}) = 5.386$ $\hat{p}_1 = 0.70$	$Z = 0.827$ −

- note: similarities of estimated coefficients ($\hat{\beta}$ and Z) and predicted probabilities ($\hat{p}$ and $\hat{p}_1$) between logistic and LDA.

Performance − logistic classification and LDA:

Method		logist. reg		LDA		LDA cv[a]	
Group		$\hat{p} \geq .43$	$\hat{p} < .43$	1	0	1	0
`surv`	1	13	8	13	8	10	11
	0	8	20	9	19	16	12
% correct		67.4		65.3		44.9	

[a] LDA evaluated by leave-one-out cross-validation

- note: similar classification by logistic and LDA, but no predictive power (e.g., no significant predictors in logistic model).

LDA: How It Works

(1) Classification by Mahalanobis distance ("predictive LDA"),

(2) Discrimination by linear function(s) of X-variables ("descriptive").

— the corresp. methods involve different model/data assumptions.[7]

Classification by Mahalanobis distance based on group means $\hat{\boldsymbol{\mu}}^{(k)} = \bar{\mathbf{X}}^{(k)}$, for $k = 1, \ldots, m$ groups, and pooled variance matrix $\mathbf{S}$:

* assign a new observation $\mathbf{X}$ to the group (k) to which it has the smallest (squared) Mahalanobis distance $d_M(\mathbf{X}, \hat{\boldsymbol{\mu}}^{(k)}, \mathbf{S})$,[8]
* the classification probability for group k is

$$\hat{p}_k = \exp\left(-\tfrac{1}{2}\, d_M(\mathbf{X}, \hat{\boldsymbol{\mu}}^{(k)}\right) / \textstyle\sum_{g=1}^{m} \exp\left(-\tfrac{1}{2}\, d_M(\mathbf{X}, \hat{\boldsymbol{\mu}}^{(g)})\right),$$

valid under the i.i.d. MVN($\boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}$) assumption in each group (k).[9]

Linear (also canonical) discriminant function(s) are linear functions $Z = a_1 X_1 + \ldots + a_p X_p$ to separate the groups as well as possible:

- Z_1 has largest possible F = MSG/MSE in a 1-way ANOVA for Z_1; same for Z_2 subject to Z_2 being uncorrelated with Z_1 within groups; same for Z_3 (uncorrelated with Z_1 and Z_2), etc.,
- at most $\min(p, m-1)$ such variables may be found, but only the first of these may be significant (formal tests require MVN),
- Z's may be determined as eigenvectors for $\mathbf{W}^{-1}\mathbf{B}$, where $\mathbf{W}$ and $\mathbf{B}$ are within-group and between-group "variances"[10], and the eigenvectors and eigenvalues may be explored similarly to PCA.

[7] All methods described here assume equal within-group (co)variances in groups.

[8] Mahalanobis distances between the groups can also be computed (5L−5).

[9] Technical note: these are posterior probabilities with equal prior probabilities $p_{0k} = 1/m$; non-uniform prior probabilities may be included in the calculation.

[10] $\mathbf{W}$ is the MANOVA residual matrix, and if $\mathbf{T}$ is the total SSCP (sum of squares and cross-products) matrix, then $\mathbf{B} = \mathbf{T} - \mathbf{W}$.

Predictive and Descriptive LDA: Skulls

Summary of descriptive LDA results (based on Stata listing):[11] [12]

- eigenvalues for $\mathbf{W}^{-1}\mathbf{B}$: $0.425, 0.039, 0.016, 0.002$ — the first $\sim$ 88% variance,
- statistical significance of discriminants: only for Z_1,
- coefficients/loadings (standardized) for Z_1: 0.58, -0.17, -0.71, 0.26 — can be interpreted in terms of cranial shape,
- correlations/structure coefficients of Z_1 with variables: 0.62, -0.29, -0.73, 0.26 — also for interpretation,
- group means for Z_1: -0.80, -0.65, 0.05, 0.57, 0.83 $\sim$ clear temporal (increasing) trend.

Summary of predictive LDA results (ordinary/cross-val. agreements)[13]:

Classified		1	2	3	4	5	total	prop. corr.
True	1	12 (9)	8 (10)	4 (5)	4 (4)	2 (2)	30	0.40 (0.30)
	2	10 (11)	8 (7)	5 (5)	4 (4)	3 (3)	30	0.27 (0.23)
	3	4 (6)	4 (4)	15 (12)	2 (2)	5 (6)	30	0.50 (0.40)
	4	3 (3)	3 (3)	7 (7)	5 (5)	12 (12)	30	0.17 (0.17)
	5	2 (2)	4(4)	4 (4)	9 (10)	11 (10)	30	0.37 (0.33)
total		31 (31)	27 (28)	35 (33)	24 (25)	33 (33)	150	0.34 (0.29)

- a single linear discriminant explaining the majority of group differences was found,
- nevertheless, the predictive LDA performance was poor.

[11] Minitab gives only predictive LDA results.

[12] Some deviations from results in Manly; not clear why that is (round-off?).

[13] The same classification tables were obtained in Minitab and Stata, and in agreement with Manly's results. Minitab also provides linear discriminant functions, one per group, to determine group proximity.

Employment Example & Final Remarks

Data summary: 9 percentages on 30 European countries in four groups.

Summary of descriptive and predictive LDA results:[14]

- eigenvalues for $\mathbf{W}^{-1}\mathbf{B}$: $5.349, 0.570, 0.202$ — 1st $\sim 87\%$ variance,
- statistical significance of discriminants: only for Z_1,
- coefficients for Z_1: all positive, in the range $0.22 - 4.77$,
- correlations of Z_1 with variables: ranging from -0.23 to 0.50 (some close to zero),
- group means for Z_1: clear separation of Eastern countries by lower Z_1 mean,
- classifications quite decent: overall 0.73 (0.60), for individual groups ≥ 0.667 (except for "Other" (4 countries)),

Conclusion: single strong discriminant with good separation of Eastern countries from others, and quite decent classification.

Summary remarks on discriminant analysis:

- assumed variance homogeneity may be relaxed $\Rightarrow$ quadratic discriminant analysis (QDA),
- assumed normality critical for significance, less for general principles and classification rules; note: $\mathbf{W}^{-1}\mathbf{B}$ is sensitive to outliers,
- recommended to include both descriptive and predictive aspects in analysis and discussion (otherwise one might be better of choosing another method),
- recommended to always report cross-validated classification rates (perhaps in addition to simple rates).

[14] Manly recommends to eliminate the collinearity between variables; `TC` dropped.