Index of Lecture 13: Classification and Canonical correlation

Page	Title
1	Practical information
2	Project presentations
3	Project reports
4	Classification/discrimination: overview
5	Logistic regression as classification
6	Performance of discrimination
7	Linear discrimination analysis (LDA)
8	LDA illustration with 2 groups
9	Multinomial logistic classification
10	LDA (3 groups) example
11	Plots for 3-group LDA example
12	Summary remarks for classical methods
13	kth nearest neighbor (KNN) classification
14	KNN settings and examples
15	Reducing predictor variable dimension
16	Canonical correlation analysis: the idea
17	Canonical correlations example: butterflies
18	Canonical correlation analysis: the steps
19	Canonical correlations path diagram: butterflies
20	Canonical correlation analysis: limitations and practical issues

Practical Information

Today's lecture — wrap-up of multivariate methods with classical statistical material on linear discriminant analysis (LDA), classification, and canonical correlation analysis (CCA),

* LDA; Manly 3/4, Chapter 8,
* classification with logistic regression; Manly 3/4, Chapter 8 (but not much new),
* *K*-nearest neighbor classification; SL (James et al. (2013) text), Section 2.2,[1]
* CCA: Manly 3/4, Chapter 10; Tabachnick & Fidell, Chapter 12;

— this will be the LAST LECTURE (apart from final review before exam)!

Other news:

- Schedule as the course winds down:
  - last shared session with VHM 812 is the review this Friday, 9-11am,
  - no full lab session for this lecture, but some problems included on April 19th,
  - this Monday (April 12th): PRESENTATIONS (see next slide),
- last home assignment (#6) received, to be returned on Monday,
- the teaching evaluation survey closes on Friday — please fill it!

[1] The SL text also contains nice introductions to more radical novel regression techiques, e.g. support vector machines.

Project Presentations

- scheduled for April 12, 1-4pm, in the large computer lab (218S), but with extra arrangements for online presentation,
- approx. 12 min. overview of problem, data, statistical analysis and conclusions,
  - statistical models/methods must be explained!
  - conclusions must be presented, including estimated effects,
  - reduce biological introduction and discussion to the essentials. . . ,
- approx. 3 minutes informal discussion, involving
  - all course participants,
  - both biological and statistical issues,
- use Word, Powerpoint and Minitab/Stata/R (use my laptop or bring your own), as you like,
- any priorities on order? (otherwise mostly random),
- marking scheme:
  - no marks for presentation alone (only combined with report),
  - my main emphasis is on your understanding of what you did. . . ,
  - format and layout of presentation are of minor importance.

Project Reports

- recommended to aim for manuscript-like layout:
  introduction, material and methods (in particular, statistical methods), results, discussion/conclusion,
- remember, statistical methods must be described in more detail than you would do in an applied paper,
  - you need to document your analyses by suitable software listings or program files (e.g. a Stata do-file),
  - please attach a data set prepared for analysis,
- the statistical analysis often comprises several parts/methods (contrary to statistics reported in papers that are usually restricted to a single method),
- avoid your report being essentially a pile of annotated Minitab/Stata listings,
- listings may be put in an appendix (and could be numbered),
- probably 5-10 pages of text,
- marked (30% of course mark),
  - emphasis will be on: problem and data description, statistical models and their validation, statistical inference, conclusions and presentation of results.
- due date listed at course homepage & Moodle account.

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 g groups (or classes) 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 any structure involved in group membership (e.g., associations with $\mathbf{X}$-components).

Many approaches exist for such problems; one possible classification. . . :

* parametric, i.e. based on a specific (statistical) model assumptions, e.g. normality,
* partly parametric, i.e. based on assumed forms of probability of group membership, e.g. logistic or polytomous/multinomial regression models,
* distribution-free techniques, e.g. based on specific distance measures or training of a complex classifier.

Classical methods, such as linear discriminant analysis, have a long history, but are increasingly overtaken by computer-intensive methods with a more intuitive and less technical mathematical/statistical foundation.

Logistic Regression as Classification

Example: sparrow data, with 5 quantitative measurements on 49 sparrows after a storm, to predict survival (0/1).

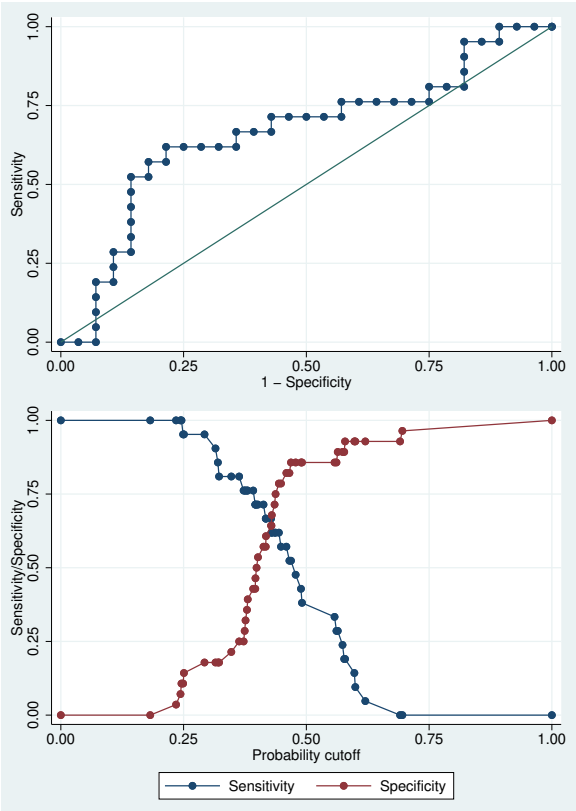
Logistic regression for survival:

Statistic	Effect/Predictor					
	intercept	length	alar	beak_head	humerous	keel_stern
estimate	13.58	-0.16	-0.03	-0.08	1.06	0.07
SE	15.86	0.14	0.11	0.63	1.02	0.42
P-value	–	0.24	0.79	0.89	0.30	0.86

- overall, non-significant model ($\chi^2(5) = 2.85$),
- classification table (at prob. cut-off 0.5) (later also termed confusion matrix):

true surv	predicted survival 0	1	total
0	24	4	28
1	14	7	21
total	38	11	49

- ROC curve: (ROC area = 0.66 — very low!),
- Se = 0.33, Sp = 0.86 (at cut-off 0.5), 63% correct,
- Se = 0.62, Sp = 0.64 (at cut-off 21/49 ≈ 0.43).



Performance of Discrimination

We should not measure the performance of a classification rule on the same data as it was developed,

- naturally an unrealistically good agreement between observed and predicted,
- difficult to assess degree of "overfitting": a too close adaptation of classifier to the data $\sim$ good data fit but potentially poor predictive performance for other data.

Alternatives for performance assessment ("reliability" in VER2, Section 15.9):

- split of the data into "training (or learning) data", and "validation data" on which the performance is measured,
  — guidelines for split not very specific, e.g. about proportions of data in the two parts and whether one or multiple randomized procedure(s) are required,
- (leave-one-out) cross-validation: training with full data minus one observation left out, which in turn becomes the validation data point; results are then summarized after iteratively leaving observations out in turn.

Typical summaries include:

* confusion matrix: a cross-tabulation of true and predicted class memberships,
* proportion correctly classified, possibly sensitivity and specificity, or involving misclassification costs (all computed from the confusion matrix).

Linear Discriminant Analysis (LDA)

Classical two-part multivariate methods[2] for data with known division into g groups:

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

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

(1): Classification (1) from group means $\hat{\mu}^{(j)} = \bar{\mathrm{X}}^{(j)}$ and the pooled variance matrix S:

* assign a new observation X to the group (j), to which it has the smallest (squared) Mahalanobis distance $d_M(\mathrm{X}, \hat{\mu}^{(j)}, \mathrm{S})$,
* the j^{th} group classification probability[3]: $\hat{p}_j = \exp\left(-\frac{1}{2}\, d_M(\mathrm{X}, \hat{\mu}^{(j)}\right) / \sum_{l=0}^{g-1} \exp\left(-\frac{1}{2}\, d_M(\mathrm{X}, \hat{\mu}^{(l)})\right)$, is valid under an i.i.d. MVN$(\mu^{(k)}, \Sigma)$ assumption for all groups $k = 1, \ldots, g$.

(2): Linear (also canonical) discriminant function(s) of the form: $Z = a_1 X_1 + \ldots + a_p X_p$ aim to separate the groups as well as possible, in the following sense:

- Z_1 has largest possible $F = \mathrm{MSG/MSE}$ 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 subject to ...

Explicit solution using matrix algebra expressions:

- at most $\min(p, g-1)$ such variables may be found, but they may not all be significant,[4]
- the Z's are determined as eigenvectors for $\mathrm{W}^{-1}\mathrm{B}$, where W and B are within-group and between-group "variances"[5], and the eigenvectors and eigenvalues may be explored similarly to PCA.

[2] The methods involve different model/data assumptions; both assume equal within-group (co)variances across groups.
[3] A posterior probability for equal prior probabilities $p_{0k} = 1/g$; non-uniform prior probabilities are also possible.
[4] Formal tests for the discriminant functions require the MVN assumption.
[5] W is the MANOVA residual matrix, and B = T−W, with T the total SSCP (sum of squares/cross-products) matrix.

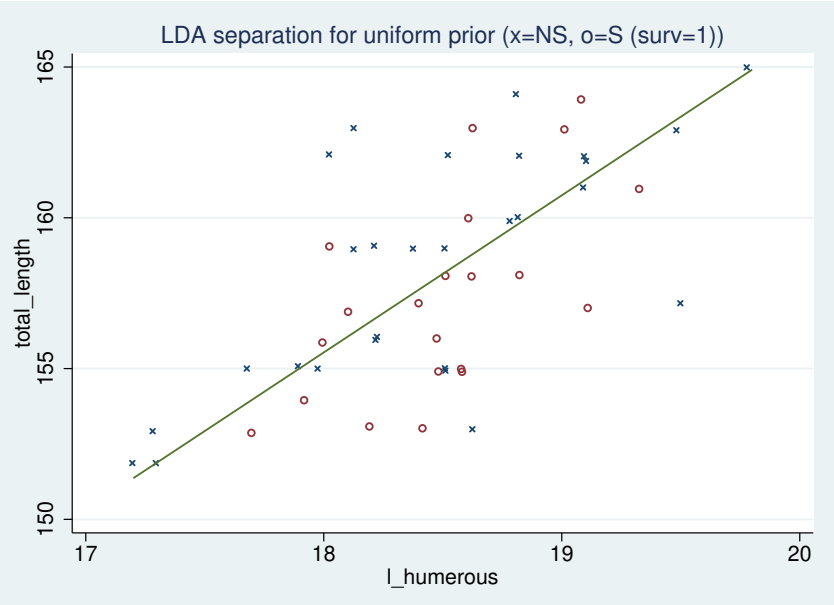
LDA Illustration with 2 Groups

Objective: illustrate LDA with 2 predictors and 2 groups:[6]

Sparrow data with predictors:
`total_length` and `l_humerous`;
estimated discriminant functions:[7]

$$\text{LDA}: \quad 0 = -0.357\,\texttt{tlen} + 1.859\,\texttt{lhum} + 22.033,$$
$$\text{logistic}: \quad 0 = -0.177\,\texttt{tlen} + 0.922\,\texttt{lhum} + 10.623.$$

note: the equations are almost multiples of each other!



Similar performance of LDA and logistic classifiers (results for data priors):

Method		LDA		LDA cv[a]		logistic		logistic cv[a]	
Group		0	1	0	1	0	1	0	1
surv	0	24	4	24	4	24	4	21	7
	1	14	7	14	7	14	7	16	5
% correct		63.3		63.3		63.3		53.1	

[a] evaluation by leave-one-out cross-validation

[6] With two groups LDA, may be considered as inferior to logistic classification (slide L13−12).
[7] Note that the LDA discriminant function matches the separation for equal prior probabilities.

Multinomial Logistic Classification

The multinomial multiple-category logistic regression model[8] with g groups (denoted as $0, \ldots, g-1$) assumes

$$\log \frac{\Pr(Y=j)}{\Pr(Y=0)} = \beta_0^{(j)} + \sum_1^p \beta_k^{(j)} X_k, \quad j = 1, \ldots, g-1.$$

For $g=2$, there is only the single logit equation for $j=1$; otherwise there are equations with separate parameters for all ratios relative to baseline ($g=0$).

- ML estimation of the parameters $\beta_0^{(j)}, \beta_1^{(j)}, \ldots, \beta_p^{(j)}$, $j = 1, \ldots, g-1$,
- given a predictor value $\mathrm{x} = (x_1, \ldots, x_p)$, we can compute estimates for the ratios $\Pr(Y=j|\mathrm{x})/\Pr(Y=0|\mathrm{x})$ and hence also all $\Pr(Y=0|\mathrm{x}), \ldots, \Pr(Y=g|\mathrm{x})$,
- rule: assign to group $0, \ldots, g-1$ with highest probability.

Adjustment for prior probabilities:

- estimates for $\beta_0^{(j)}$ are based on the observed proportions of groups $0, \ldots, g-1$, but can also be adjusted by known prior probabilities.

Example: `beef_ultra` data with 8 measures (5 quantitative, 3 binary) on 487 beef cattle, to predict carcass quality (AAA,AA,A); AAA is best.

- prediction for first row (#43): $\Pr(\mathrm{A})=0.165$, $\Pr(\mathrm{AA})=0.73$, $\Pr(\mathrm{AAA})=0.11$: $\rightarrow$ AA.

[8] Multinomial logistic regression is covered in Chapter 17 of VER/MER.

LDA (3 Groups) Example

Full analysis for `beef_ultra` data with data priors:[9]

- first eigenvalue (for $\mathbf{W}^{-1}\mathbf{B}$) with 95% of the variance,
- misclassification rate $\approx$ 36% (cross-validated), slightly lower than logistic classifier (37%), and both classificers struggle to identify grade A subjects;
  — confusion matrices:

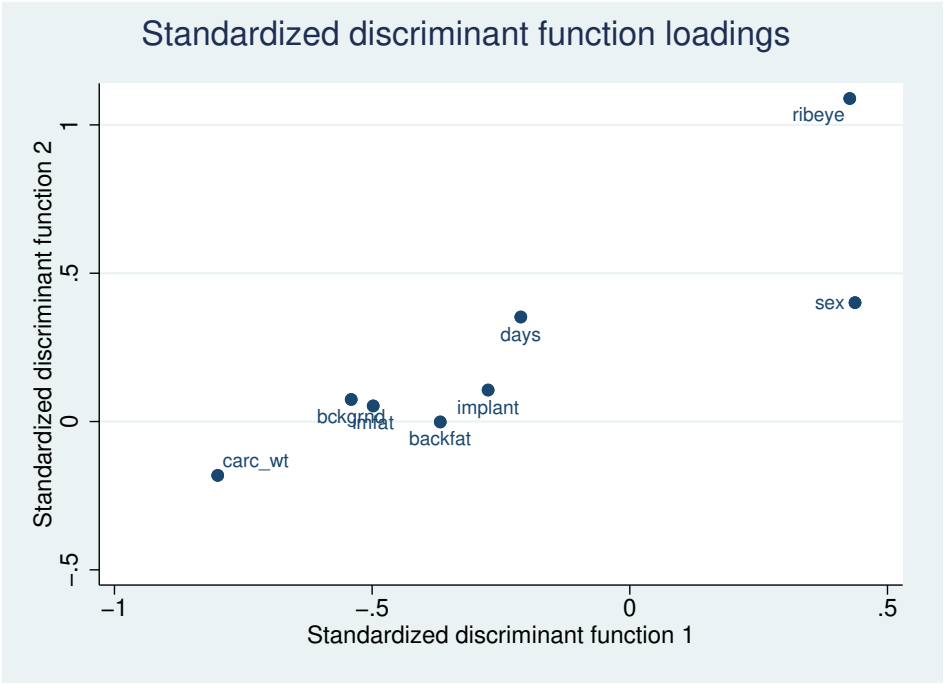
Method		LDA			logistic			
Group		AAA	AA	A	AAA	AA	A	total
`grade`	AAA	76	88	0	74	90	0	164
	AA	43	230	4	43	229	5	277
	A	3	39	4	3	40	3	46

- loading plot (next page): strongest influence on discriminant function by `ribeye`, `sex`, `carc_wt`,
- score plot (next page): no direct separation of groups, but visible association between first component and occurences of groups 1 and 2; third group seems not well captured at all.

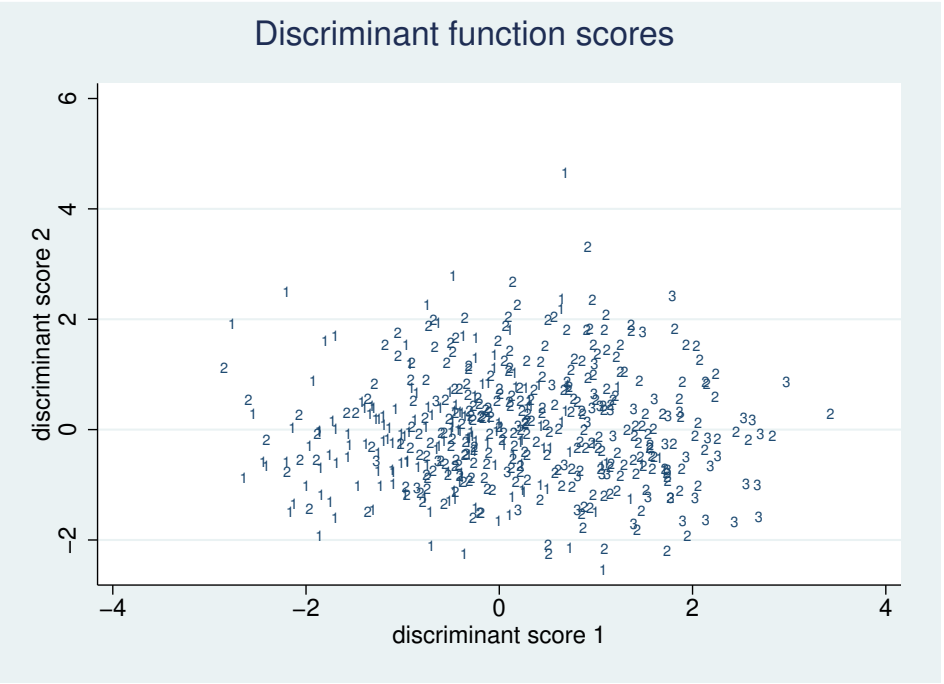
[9] All predictors treated as quantitative; there is really no other option with LDA.

Plots for 3-Group LDA Example

Loading plot:



Score plot:



Summary Remarks for Classical Methods

Main critical point for LDA is the reliance on normality assumptions,

- lack of ability to incorporate non-quantitative variables is a serious drawback,
- even for quantative variables, the performance can be substantially affected by non-normality and outliers,
- additional assumption made (for predictive LDA): equal variances across groups — can be relaxed by so-called quadratic discriminant analysis (QDA),
  - did not perform well for `beef_ultra` data: — cross-validated confusion matrix:

Group		AAA	AA	A
grade	AAA	93	70	1
	AA	77	179	21
	A	4	30	12

  - extension seems of limited interest in this case,
- for two groups, LDA is similar to logistic classification (which does not have the drawbacks above); some potential advantages of LDA have also been noted: [10]
  - with well-separated classes, logistic classifiers may have unstable parameter estimates,
  - for small n and approximately normal distributions, the extra assumptions may stabilize the procedure.

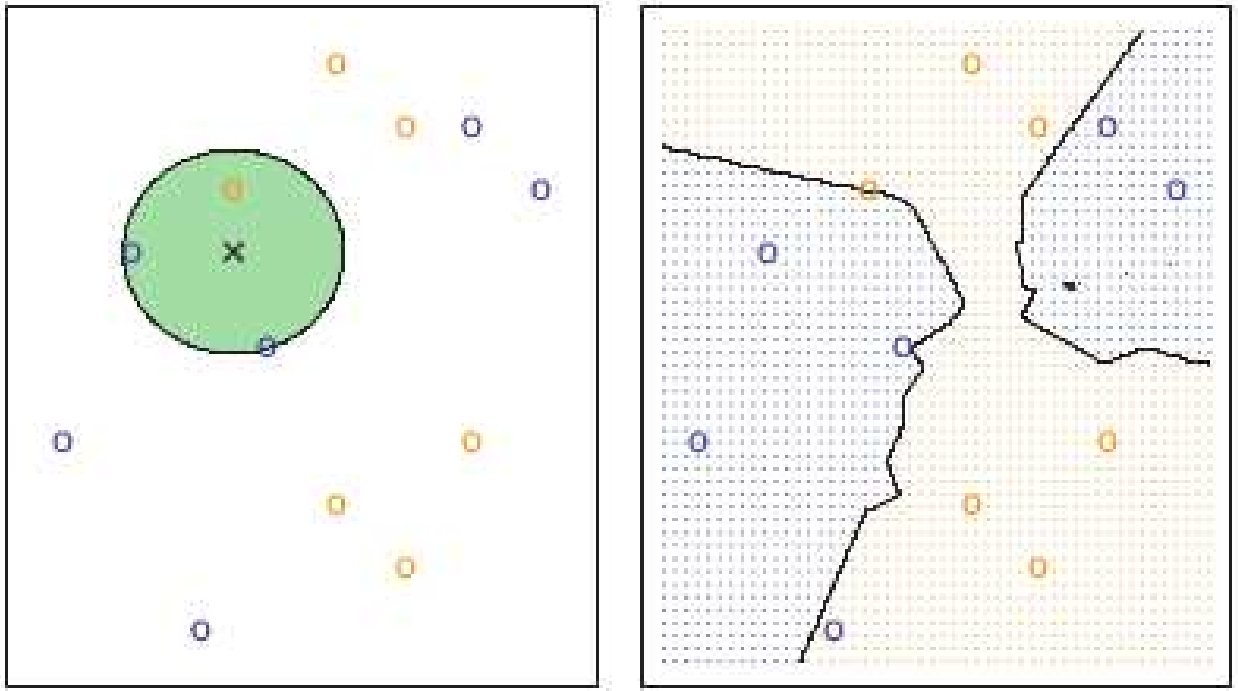
[10] For example, SL (James et al., 2013), Section 4.4.

Kth Nearest Neighbor (KNN) Classification

- a non-parametric discrimination algorithm dating back to the 1950s,
- its name comes from the idea of associating any new point (observation) x, for which a classification is desired, with its k nearest neighbours (say $\mathcal{N}_x$) in the training set,
- estimate group probabilities as the sample proportions in $\mathcal{N}_x$ — done!

Illustration
(Figure 2.14 of SL):

KNN with $k=3$,
6 blue and 6 orange
observations



left: prediction from new observation at x, right: full KNN decision boundary

KNN Settings and Examples

Settings/choices (suggestions from Stata manual):

- value of k[11]: with 2 groups in the data, choose k odd, in range $n^{.25} - n^{.375}$ for roughly equally-sized groups; or as $\sqrt{\cdot}$ of typical group size,
- distance measure: previous discussions of distance apply here as well.

Binary classification: `sparrow` data

- select $k = 3$ according to recommendations,
- misclassification rates with cross-validation $\approx$ same as with logistic/LDA classifiers,
- results clearly worse after standardization of variables.

Full 3-group classification: `beef_ultra` data with all predictors,

- at best, a slightly inferior performance to LDA and logistic classifiers; sample results (with $k = 7$, L_1-distance on standardized variables):

Group		AAA	AA	A
grade	AAA	69	94	1
	AA	56	219	2
	A	1	37	8

- quite commonly, some points are left unclassified by the algorithm (due to ties between groups).

[11] The role of k in the algorithm can be understood as to balance flexibility and stability against overfitting.

Reducing Predictor Variable Dimension

Considerations for a systematic approach to dealing with many predictors (p) relative to number of observations (n):

- for regression, dedicated approaches exist, e.g. ridge regression, partial least squares. . . ,
- principal components regression is simply the method of replacing the original predictors with the principal components derived from them, [12]

but in regression we often want to use our understanding of the predictors (epi!):

* maybe use the components/factors as guidance for selection of variables or for construction of new variables from existing ones,
* maybe useful to split the multivariate analysis into separate analyses for a number of strata where interpretations of components/factors are more manageable,
* maybe measures of agreement among variables within such strata (e.g., Cronbach's alpha) are useful.

What about classification? — a real issue?

- classifiers based on a large number of predictor variables may be numerically feasible and even more robust,
- for "black-box" predictive systems, misclassifications must be investigated.

[12] This raises the interesting question of whether we can select the most important components for analysis; it has been argued to be invalid: Joliffe (1982), *J. Royal Statist. Soc. C* 31, 300−303.

Canonical Correlation Analysis (CCA): the Idea

CCA is a mathematical[13] procedure to represent the correlation between two sets of (correlated) variables: $(X_1, \ldots, X_p)$ and $(Y_1, \ldots, Y_q)$ (where $q \leq p$) by two new sets of variables $(U_1, \ldots, U_q)$ and $(V_1, \ldots, V_q)$ such that:

- each U_i is a linear combination of the X_j's; each V_i is a linear comb. of the Y_j's,
- the variables $U_1, \ldots, U_q$ are uncorrelated/orthogonal; same for $V_1, \ldots, V_q$,
- the variables U_1 and V_1 have maximal correlation; the variables U_2 and V_2 have maximal correlation subject to being orthogonal to U_1 and V_1, resp.; and so forth.

First impressions/interpretations:

* the pairs (U_1, V_1), (U_2, V_2),...,(U_q, V_q) are unique (up to sign change) when standardized; these are called canonical variates, and $\sim$ independent "directions" among the (X_j) and (Y_j) with maximal correlation, and
  - may have useful subject-matter interpretations,
  - may represent useful data reductions,
* some similarity with PCA, but here we maximize correlation instead of variance,
* can be viewed as an extension of multiple regression to multiple Y_j's, because with only Y_1 the predictions $X\hat{\beta}$ have maximal correlation with the data Y_1.[14]

[13] CCA is not in itself based on a statistical model, but inference about the number of signficant components is.
[14] Although perhaps suggested by the notation, there is no direction X $\rightarrow$ Y (and of course no claim of causality).

Canonical Correlation Analysis Example: Butterflies

Data: 4 environmental and 6 genetic variables (Pgi gene frequencies) for 16 colonies of a butterfly in California and Oregon; interest is in describing relation between the genetic and environmental variables,

- complete collinearity among genetic variables ⇒ one variable must be dropped (type 1.30), and types 0.40 and 0.60 are combined (due to low frequencies),

⇒ 4 $\boldsymbol{X}$-variables (environmental) and 4 $\boldsymbol{Y}$-variables (genetic).

(For illustration only) Multiple linear regression for `freq_pgi08`:

- $\boldsymbol{R^2 = 0.415 = 0.6443^2}$ equals the (only) canonical correlation squared,
- $\boldsymbol{F = 1.95}$ test for overall significance equals test for (only) canonical variate.

Results of full CCA — canonical correlations: $\mathbf{0.862, 0.450, 0.386, 0.089}$,[15]

- no overall significance by multivariate tests ⇒ focus here on first component:

U_1	altitude	annprec	maxtemp	mintemp
stand. coef.	0.124	0.293	-0.468	-0.260
loading/corr.	0.922	0.771	-0.898	-0.919

V_1	pgi046	pgi08	pgi10	pgi116
stand. coef.	-0.548	-0.422	0.089	-0.826
loading/corr.	-0.384	-0.740	0.961	-0.475

- interpretations (details on next slide): $\boldsymbol{U_1}$ ∼ contrast altitude and precipitation vs. temperatures, $\boldsymbol{V_1}$ ∼ mostly a contrast between `freq_pgi10` and others,
- plot of scores for $\boldsymbol{V_1}$ vs. $\boldsymbol{U_1}$ shows one outlying observation (colony DP).

[15] Results agree roughly with Manly (recall that sign reversals are unimportant); same results in Stata and SAS.

Canonical Correlation Analysis: the Steps

The full covariance matrix S $(p+q) \times (p+q)$ for the combined set $(X_1, \ldots, X_p, Y_1, \ldots, Y_q)$ is partitioned (split) into submatrices:[16]

$$\mathrm{S} = \begin{pmatrix} \mathrm{A} & \mathrm{C} \\ \mathrm{C}^t & \mathrm{B} \end{pmatrix},$$

- let $\mathrm{R} = \mathrm{B}^{-1}\mathrm{C}^t\mathrm{A}^{-1}\mathrm{C}$ $(q \times q)$,
- the eigenvalues of R (proportions of variance explained) are the squared canonical correlations $r_1^2, \ldots, r_q^2$ (where $r_i = \mathrm{Corr}(U_i, V_i)$),
- the ith eigenvector (say $b^{(i)}$) gives the V_i coefficients: $V_i = b_1^{(i)}Y_1 + \ldots + b_q^{(i)}Y_q$,
- from $b^{(i)}$, also the U_i coefficients as: $U_i = a_1^{(i)}X_1 + \ldots + a_p^{(i)}X_p$ with $a^{(i)} = \mathrm{A}^{-1}\mathrm{C}\, b^{(i)}$,
- assuming MVN for $(X_1, \ldots, X_p, Y_1, \ldots, Y_q)$, tests can be computed for all canonical correlations/components.[17]

Interpretation of components[18] — two suggestions (described here for (U_1, V_1)), look at:

- loadings of X_j's on U_1 and of Y_j's on V_1,[19]
- or correlations between X_j's and U_1 and between Y_j's and V_1.

[16] Here A $(p \times p)$ and B $(q \times q)$ are the covariance matrices for X_j's and Y_j's, respectively, and C $(p \times q)$ is the matrix of correlations between X_j's and Y_j's.

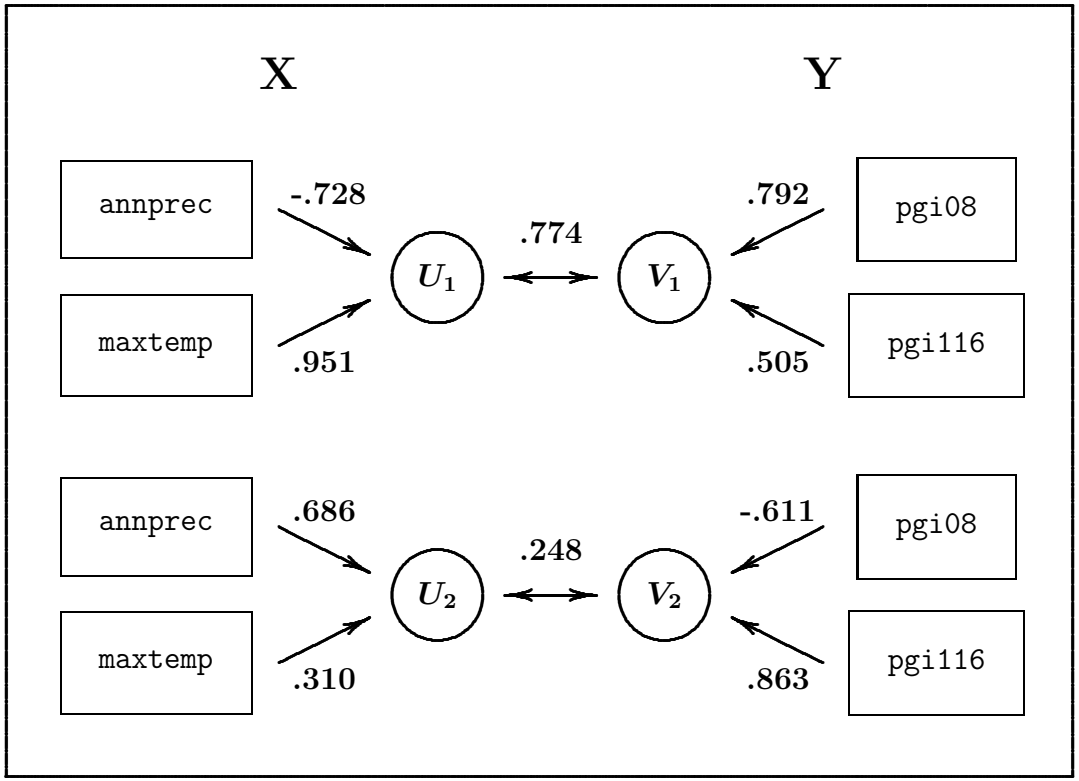
[17] Manly mentions Bartlett's (χ^2-) test for all components; Stata uses F-tests. Manly also dismisses tests for some components only as "not reliable", but Stata will gladly compute them ...

[18] Interpretation of components may differ between approaches if X_j's or Y_j's are highly collinear (Manly).

[19] Squared loadings averaged across the X_j's (or Y_j's) give the proportion of variance explained by corresponding variate; further multiplication with the squared correlations give the redundancies: proportions of variance explained for the other set of variables; details in TF.

CCA Path Diagram: Butterflies

Representation of loadings for simple CCA with 2 environmental variables (annprec,maxtemp) and 2 genetic variables (pfpgi08,pfpgi116):



Interpretations:

* eigenvalues: 0.60 ($= .774^2$) and 0.06 ($= .248^2$), also interpretable as proportions of overlapping variance between X and Y for the two pairs (U_1, V_1) and (U_2, V_2),
* among X, the canonical variable U_1 explains $(.728^2 + .951^2)/2 = .717$ and U_2 explains $(.686^2 + .310^2)/2 = .283$ of the variance,
* among Y, the canonical variable V_1 explains $(.792^2 + .505^2)/2 = .441$ and V_2 explains $(.611^2 + .863^2)/2 = .559$ of the variance,
* redundancy: among Y, the canonical variable U_1 explains $.717 \times .774^2 = .43$ and U_2 explains $.283 \times .248^2 = .02$ of the variance; analogously, among X, the canonical variable V_1 explains $.441 \times .774^2 = .26$ and V_2 explains $.559 \times .248^2 = .003$ of the variance.

CCA: Limitations and Practical Issues[20]

General comment: as indicated by the path diagrams, CCA has strong links to, and is perhaps best understood in the context of, structural equation models — a large and complicated topic that we do not aim to cover here.

Some issues/assumptions are similar to PCA — with some adaptations:

- normality is required for validity of test statistics, but CCA can be meaningfully applied to any variables for which variance and correlation make sense,
- assumed linearity enters in two ways: correlations measure linear relationships only, and by the linearly constructed variates,
- generally quite sensitive to minor changes in the data.

Interpretation of components and relations is not straightforward or automatic:

- the canonical variates maximize correlation, but not necessarily interpretability; rotations[21] are less common (and documented) than in factor analysis,
- results depend on both sets of variables: changing one set of variables will affect both sets of canonical variates — may be undesired or at least requires caution,
- highly collinear variables (in one or both sets) make the canonical variates difficult to interpret (Manly), and also using the correlations has its issues (Stata manual).

[20] Based largely on TF: Section 12.3.2.
[21] Stata offers varimax rotation.