6) Let $\Sigma = \Sigma_X$ or $\Sigma = \rho$. p142 = 4

The kth pair of pop canonical variables is $U_k = \underline{a}_k^T \underline{W}$ and $V_k = \underline{b}_k^T \underline{Y}$ and the kth pop canonical correlation $\rho_k = \text{corr}(U_k, V_k) = \sqrt{\lambda_k}$.

Sample analogs use S or S_U or R or R_U.

$\hat{U}_k = \hat{\underline{a}}_k^T \underline{W}$ $\hat{V}_k = \hat{\underline{b}}_k^T \underline{Y}$, $\hat{\rho}_k = \text{corr}(\hat{U}_k, \hat{V}_k) = \sqrt{\hat{\lambda}_k}$

7) Let $\hat{\Sigma} = S$ or R. p144 $(\hat{U}_1, \hat{V}_1)$ is the pair of linear combinations $(\hat{U}, \hat{V})$ having unit sample variances that maximizes $\text{corr}(\hat{U}, \hat{V})$, and the max is $\hat{\rho}_1$. The ith pair $(\hat{U}_i, \hat{V}_i)$ is the pair of linear combinations with unit sample variances that maximizes $\text{corr}(\hat{U}, \hat{V})$ among all choices uncorrelated with the previous $i-1$ canonical variable pairs.

want linearity in scatterplot matrix and DD plot.

8) want $n > 10p$ for classical CCA, $n > 20p$ for RCCA.

Kendall 1980 p69 interpretation difficulties mean few convincing CCA applications in the liter

9) Interpretation is hard. a) If $\underline{X} = (\underline{W}^T, \underline{Y}^T)^T$ and $\underline{W}$ ind $\perp \underline{Y}$, $\hat{\rho}_1$ is not close to 0 until sample size n is quite large.

b) (p455) corr $(W_i, \underline{a}_j)$ and corr $(Y_i, \underline{b}_j)$

are not proportional to a_{ij} and b_{ij}.

Computing corr $(W_i, \underline{a}_j)$, corr $(W_i, \underline{b}_j)$ for $i = 1, \ldots, m$

and corr $(Y_i, \underline{a}_j)$, corr $(Y_i, \underline{b}_j)$ for $i = 1, \ldots, s$ can help.

c) If some W_i variables are highly correlated and some Y_i variables are highly correlated, the canonical variables can be hard to interpret; multicollinearity.

ex] $\underset{\sim}{W} = \begin{pmatrix} \log \text{shell mass} = S \\ \log \text{muscle mass} = M \end{pmatrix}$

S and M are highly correlated.

$\hat{\underline{a}}_1 = \begin{pmatrix} .150 \\ .032 \end{pmatrix}$, but $\hat{\underline{c}} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$ with $c_1 \geq 0$, $c_2 \geq 0$

may produce $\hat{\underline{c}}^T \underline{W}$ that is highly correlated with $\hat{\underline{a}}_1^T \underline{W}$, in particular, $\hat{\underline{c}} = \begin{pmatrix} .032 \\ .150 \end{pmatrix}$ would work.

d) W_i's should have similar variances s_i^2

Y_j's " s_j^2.

Otherwise, components of $\hat{\underline{a}}_i$ and $\hat{\underline{b}}_i$ are hard to interpret.

10) 1st pair is "most important" MV 25

pairs with low $\hat{\rho}_k$ can be ignored.

11) a) want to know which w_i variables are most important for $\hat{a}_1$.

b) y_i

$\hat{b}_1$.

c) want to know which w_i variable most explains $\hat{b}_1^T \underline{Y}$ see ex)

y_i

$\hat{a}_1 \underline{w}$.

d) Are $\hat{a}_1^T \underline{w}$ and $\hat{b}_1^T \underline{Y}$ meaningful?

Often have a "wow factor"?

the client says "wow that makes sense,"

but often interpretation is difficult.

12) SAS proc cancorr output

canonical correlation ... other stuff

$\hat{\rho}_1$

$\vdots$

$\hat{\rho}_m$

standardized canonical coefficients for wname

$\underline{w}$ $\hat{a}_1 \ldots \hat{a}_m$

$\underline{Y}$ $\hat{b}_1 \ldots \hat{b}_m$ for yname

maybe $\hat{b}_{m+1} \ldots \hat{b}_q$

ex] Weight, waist, pulse | Chins Situps Jumps

$\underset{\sim}{W}$ = Phys measurements

$\underset{\sim}{Y}$ = exercises

Can corr
.7956
.2006
.0726

Waist most explains $\hat{b}_1' \underset{\sim}{Y}$

Chins most explains $\hat{a}_1' \underset{\sim}{W}$

	Phys 1
weight	−0.0314
waist	0.493
pulse	−0.0082
	$\hat{a}_1$

	Exer 1
chins	−0.0661
Situps	−0.0168
jumps	0.0140
	$\hat{b}_1$

bigger waist $\leftrightarrow$ fewer chinups

13) $\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{21} & \hat{\Sigma}_{22} \end{pmatrix}$ and elements of $\hat{\Sigma}_{12}$ measure pairwise associations

14) $\text{corr}(U,V) = \dfrac{\underset{\sim}{a}^T \Sigma_{12} \underset{\sim}{b}}{\sqrt{\underset{\sim}{a}^T \Sigma_{11} \underset{\sim}{a}} \sqrt{\underset{\sim}{b}^T \Sigma_{22} \underset{\sim}{b}}}$ if $\Sigma = \Sigma_x$ or $\Sigma = \rho$

$V(\underset{\sim}{a}^T \underset{\sim}{W}) = V(U) = \underset{\sim}{a}^T \Sigma_{11} \underset{\sim}{a} = 1$

$V(\underset{\sim}{b}^T \underset{\sim}{Y}) = V(V) = \underset{\sim}{b}^T \Sigma_{22} \underset{\sim}{b} = 1$

153 Theory for robust CCA is similar to that of robust PCA.

Discriminant Analysis ch8 (11)

1) p148 In supervised classification, there are $G \geq 2$ known groups and m cases to be classified. Each case is assigned to exactly one group based on its measurements $\underset{\sim}{W}_i$.

ex) patient suffering heart attack symptoms:
3 tests are done $\underset{\sim}{w} = (w_1, w_2, w_3)^T$
and person is classified as i) had a heart attack
or ii) did not have a heart attack.

ex) person applies for credit card
based on (salary, credit rating)^T.
person is classified as acceptable or
not acceptable.

2) Suppose there are $G \geq 2$ groups or
populations with pdf $f_j(\underset{\sim}{z})$ for
$j = 1, \ldots, G$ and $\underset{\sim}{z}$ is $p \times 1$. so if $\underset{p\times 1}{\underset{\sim}{X}}$
comes from pop j, then $\underset{\sim}{X}$ has pdf $f_j(\underset{\sim}{z})$.
Assume there is a random sample of
n_j cases $\underset{\sim}{x}_{1j}, \ldots, \underset{\sim}{x}_{n_j j}$ for each group.
(training data)
Let $(\overline{\underset{\sim}{x}}_j, S_j)$ be the sample mean and covariance
matrix for each group. Let $\underset{\sim}{W}_i$ be a new $p \times 1$
(test data)
random vector from 1 of the G groups, but
the group is unknown. Usually there are many $\underset{\sim}{W}_i$

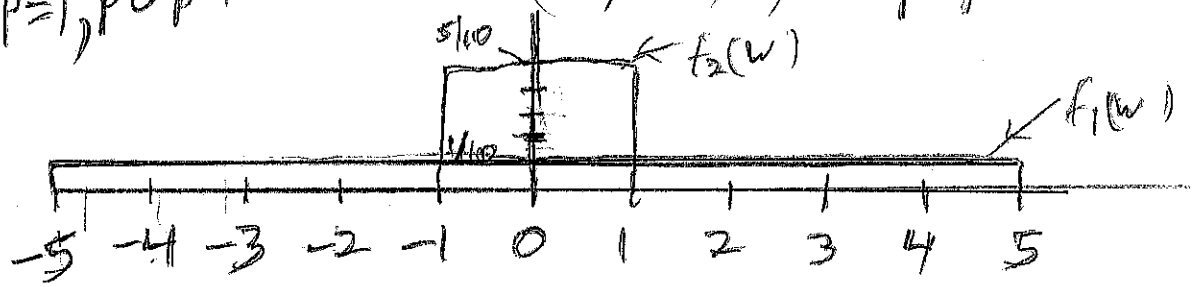
and discriminant analysis attempts to allocate the $\underset{\sim}{w}_i$ to the correct groups.

3] The maximum likelihood rule allocates case $\underset{\sim}{w}$ to group a if $\hat{f}_a(\underset{\sim}{w})$ maximizes $\hat{f}_j(\underset{\sim}{w})$ for $j = 1, \ldots, G$.

This rule is robust to nonnormality of the k groups, but it is hard to estimate $\hat{f}_j$ for $p > 1$.

4] E2 question given $p = 1$ and graph of $f_1, \ldots, f_G$, give the maximum likelihood rule.

ex] $p = 1$, pop 1 $\sim U(-5, 5)$, pop 2 $\sim U(-1, 1)$



ML rule: allocate w to pop 2 if $-1 < w < 1$

allocate w to pop 1 if

$-5 < w < -1$ or $1 < w < 5$.

5) For now, assume costs of correct and incorrect allocation are unknown or equal and that probabilities $p_a(\underline{w}_i)$ that $\underline{w}_i$ is in group a are unknown or equal $p_a(\underline{w}_i) = \frac{1}{G}$ for $a = 1, \ldots, G$.

6) If the G groups are assumed to have the same covariance matrix Σ_x, then the pooled covariance matrix estimator is

$$S_{pool} = \frac{1}{n-G} \sum_{j=1}^{G} (n_j - 1) S_j.$$

by assumption, can be used if n_j are not equal

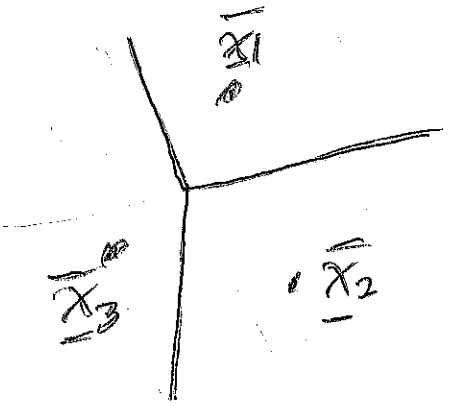
7) Assume the pop dispersion matrices Σ are equal: $\Sigma_j \equiv \Sigma$ for $j = 1, \ldots, k$. Let $\hat{\Sigma}_{pool}$ be an estimator of Σ. The linear discriminant rule is to allocate $\underline{w}$ to the group with the largest value of

$$d_j(\underline{w}) = \underline{\mu}_j^T \hat{\Sigma}_{pool}^{-1} \underline{w} - \frac{1}{2} \underline{\mu}_j^T \hat{\Sigma}_{pool}^{-1} \underline{\mu}_j$$
$$= \hat{\alpha}_j + \hat{\beta}_j^T \underline{w}.$$

Linear discriminant analysis LDA uses $\hat{\underline{\mu}}_j = \overline{\underline{x}}_j$ and $\hat{\Sigma}_{pool} = S_{pool}$.

8) LDA is the most used rule, is robust to nonnormality and somewhat robust to the assumption of equal pop covariance matrices.

LDA basically separates groups with G 27.5 hyperplanes.



9) The quadratic discriminant rule allocates $\underline{w}$ to the group with the largest value of $Q_j(\underline{w}) = -\frac{1}{2}\log(|\hat{\Sigma}_j|) - \frac{1}{2}(\underline{w}-\hat{\underline{\mu}}_j)^T \hat{\Sigma}_j^{-1}(\underline{w}-\hat{\underline{\mu}}_j)$.

Quadratic discriminant analysis QDA uses $(\hat{\underline{\mu}}_j, \hat{\Sigma}_j) = (\bar{\underline{x}}_j, S_j)$ and is somewhat robust to nonnormality.

10) The distance discriminant rule allocates $\underline{w}$ to the group with the smallest squared distance $D^2_{\underline{w}}(\hat{\underline{\mu}}_j, \hat{\Sigma}_j) = (\underline{w}-\underline{\mu}_j)^T \hat{\Sigma}_j^{-1}(\underline{w}-\hat{\underline{\mu}}_j)$ and is robust to nonnormality and robust to the assumption of equal Σ_j, but needs $n_j > 10p$ $j=1,\ldots,k$. QDA and LDA are usually better.

11) p149 Assume $G=2$ and that there is a group 0 and a group 1. Let $\underline{s}(\underline{w}) = P(\underline{w} \in \text{group 1})$. Let $\hat{\underline{s}}(\underline{w})$ be the logistic regression estimate of $\underline{s}(\underline{w})$. Then the estimated sufficient predictor $ESP = \hat{\alpha} + \hat{\beta}^T \underline{w}$ and $\hat{\underline{s}}(\underline{w}) = \frac{e^{ESP}}{1+e^{ESP}} =$

$\frac{\exp(\hat{\alpha} + \hat{\beta}^T \underline{w})}{1 + \exp(\hat{\alpha} + \hat{\beta}^T \underline{w})}$. The logistic regression discriminant rule allocates $\underline{w}$ to group 1 if $\hat{\underline{s}}(\underline{w}) \geq 0.5$ and allocates $\underline{w}$ to group 0 if $\hat{\underline{s}}(\underline{w}) < 0.5$. This rule is robust to nonnormality and to the assumption of equal Σ_j $j = 0, 1$.

$\hat{\underline{s}}(\underline{w}) \geq 0.5$ iff $ESP \geq 0$

$\hat{\underline{s}}(\underline{w}) < 0.5$ iff $ESP < 0$.

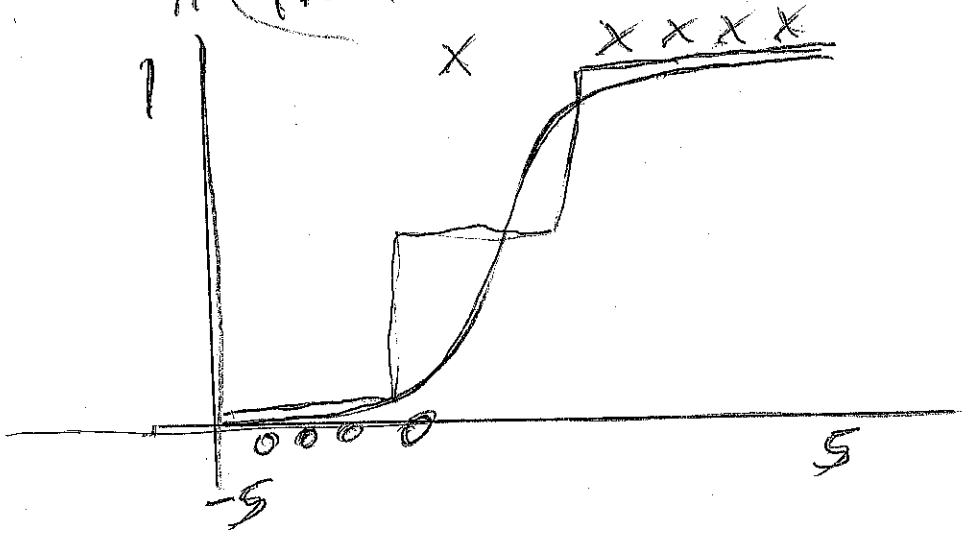
12) A response plot is a plot of ESP vs Y_i (or $\hat{\underline{s}}(\underline{w})$) with $\hat{\underline{s}}(\underline{x}_i) \equiv \hat{\underline{s}}(ESP)$ added as a visual aid. Also divide ESP into J slices with approx the same number of cases in each slice. Compute the

Y_i vertical axis

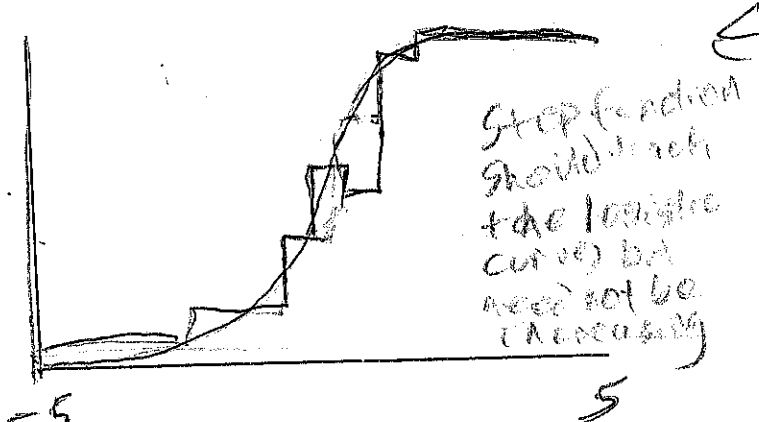
sample mean = sample prop in each slice;

$\hat{\pi}_s = \bar{Y}_s = \frac{\sum}{s} Y_i / m_s$ where m_s is the number of cases in slice s. Add the step function to the response plot. If n_0 and n_1 are the sample sizes of groups 0 and 1 and $n_i > 5p$, the logistic regression model is useful if the step function of slice proportions scatter fairly closely about the logistic curve $\hat{\pi}(ESP)$.

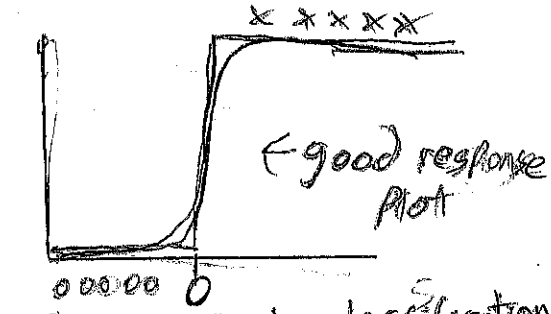
13] know for E2 Y_i Be able to tell a good LR response plot from a bad response plot.

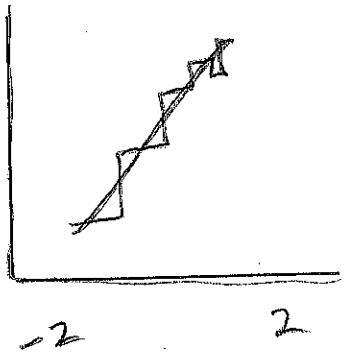


good response plot

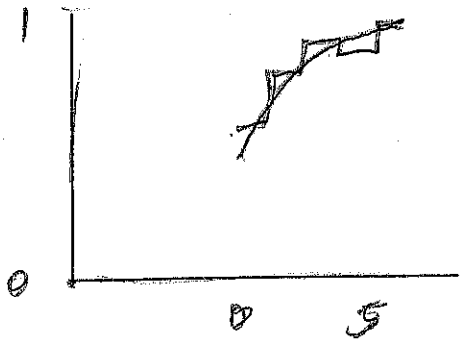


← good response plot

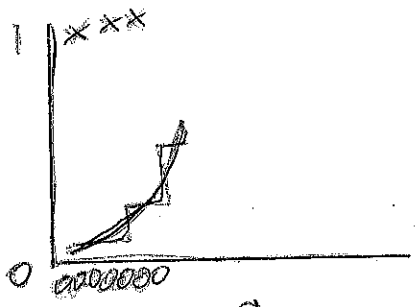




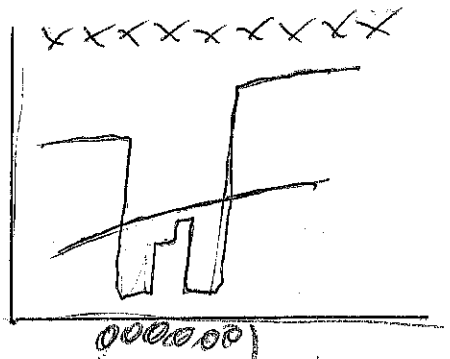
good

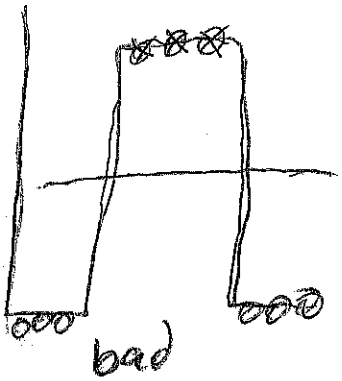


good



good





bad

-1.6 ↑ bad 1.1

step function

decreases dramatically, then increases dramatically

misclassification rate

[14] $\hat{\beta}$

ESP = 0 → 0.5 → 0.5

.25 or .75 → .25

$\hat{\beta}$ → $\min(\hat{\beta}, 1-\hat{\beta})$

ESP > 2 or ESP < -2 will have low error (misclassification) rate if LR model is good

[15] output

label	estimate	Std/Error	Est/SE	pvalue
constant	$\hat{\alpha}$			
x_1	$\hat{\beta}_1$			
⋮	⋮			
x_p	$\hat{\beta}_p$			

$\frac{e^{\hat{\alpha}}}{1+e^{\hat{\alpha}}} = \hat{\beta}(x=0)$ makes sense if $x=0$ does

$\alpha=0 \rightarrow \hat{\beta}(x=0) = \frac{1}{2}$ which rarely makes sense

ex] know for final

sex 0 f 1 m = group

label	Estimate
constant	−19.7762
circum	0.0244688
length	0.0371472

head measurements.

Let circum $= x_1 = 550$, length $= x_2 = 200$

a) Find ESP for $\underline{x}$.

$ESP = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 = -19.7762 + 0.0244688(550) + .0371472(200)$

$= 1.11108$

b) Is $\underline{x}$ classified in group 0 or group 1?

group 1 since $ESP > 0$

c) Find $\hat{p}(\underline{x})$. $= \dfrac{e^{ESP}}{1+e^{ESP}} = \dfrac{3.0376}{4.0376} = \boxed{0.7523}$

16] Training data: $\underline{x}_{ij}$, n_j, group j.

use a discriminant analysis method to classify the training data. If m_j of the n_j group j cases are correctly classified, then the apparent error rate for group j is $1 - \frac{m_j}{n_j}$. Suppose $m = \sum_{j=1}^{G} m_j$