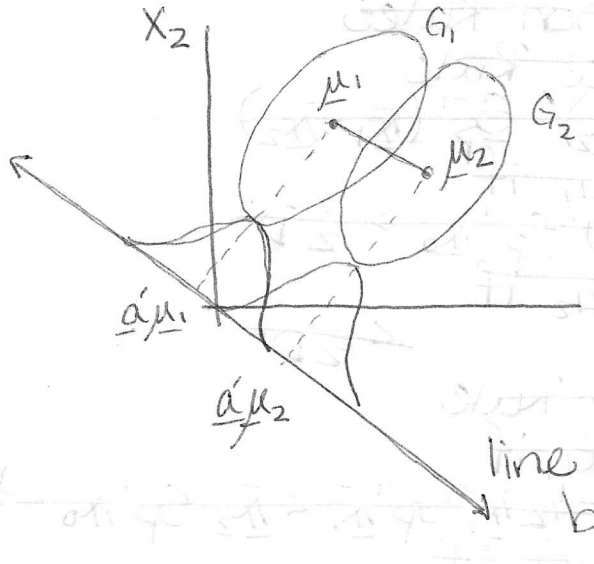
Discriminant Analysis



predefined groups
① describe how the groups differ
② predict group membership for a new datapoint

maximize standardized distance between univariate means

multivariate normality ~~not~~ assumed
sample size by group does not need to be equal
* assumption: the var-cov matrix is equal between groups (Bartlett's test)

Steps (2 groups)

① separate the data by group
② calculate means of variables by group
   └and var-cov matrices
③ $S_{pooled} = [(n_1-1)\underline{S}_1 + (n_2-1)\underline{S}_2] / (n_1+n_2-2)$
④ $\underline{a} = S_p^{-1}(\bar{X}_1 - \bar{X}_2)$
⑤ $y = \sum_{i=1}^{q} a_i X_i$ calculate for each entity
⑥ determine $\bar{y}$ for each group
⑦ new datapoint ($\underline{X}_0$)
   ⓐ calculate y_0
   ⓑ group ID determined by closeness to $\bar{y}$.

Classification Errors

Actual Group	Predicted Group 1	Predicted Group 2
1	n_{11}	n_{12} → mistake
2	n_{21} ← mistake	n_{22}

Methods of Assessment

① resubstitution: apply rule to the data used to construct the rule; good for large samples

② holdout data: keep half the data out of the analysis and apply the rule on those data; "training" vs. "test" samples;

* ③ cross-validation: "leave-one-out" can be computer intensive

Changing the Classification Rules

Ⓐ Prior knowledge of population proportions (p_1 and p_2)

assign x_0 to G_1 if

$$y_0 > \frac{1}{2}(\bar{y}_1 + \bar{y}_2) + \ln(p_2/p_1)$$

assign x_0 to G_2 if

$$y_0 < \frac{1}{2}(\bar{y}_1 + \bar{y}_2) + \ln(p_2/p_1)$$

Ⓑ Different costs of misclassification

Let $C(2|1)$ = cost of misclass. in group 2, when really in group 1 ...

$$R(2|1) = C(2|1)/C(1|2)$$

$$p_1^* = \frac{p_1 R(2|1)}{p_1 R(2|1) + p_2} \qquad p_2^* = \frac{p_2}{p_1 R(2|1) + p_2}$$

© Unequal Covariance Matrices
quadratic discriminant function
Is also a modified classification rule,
but we won't go in to the details.

More than two groups

- The program deals with this for you.
- $\underline{\underline{S}}_p = \dfrac{(n_1-1)\underline{\underline{S}}_1 + (n_2-1)\underline{\underline{S}}_2 + \ldots + (n_k-1)\underline{\underline{S}}_k}{n-k}$
- Picking the shortest Mahalanobis distance

$$d_i = (\underline{x}_0 - \underline{\bar{x}}_i)' \, S_p^{-1} \, (\underline{x}_0 - \underline{\bar{x}}_i)$$