Classification with several populations

Model

- $g \geq 2$ populations: $\pi_1, \pi_2, \ldots \pi_g$
- π_i: $f_i(\mathbf{x}), \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \ i = 1, \ldots, g$
- MVN assumption: $\mathbf{X}|\pi_i \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \ i = 1, \ldots, g$

Sample space

- p-dimensional multivariate observations, $\mathbf{x} = [x_1, \ldots, x_p]^T \in R^p$
- training data available for each population $\rightarrow$ $\hat{f}_i(\mathbf{x}), \hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Sigma}}_i, \ i = 1, \ldots, g$

Classification

- splitting of sample space $R^p = \bigcup_{i=1}^{g} R_i, \quad R_i \cap R_k = \emptyset, \ i \neq k$
- decision rule for new observations $d: \mathbf{x}_0 \in R^p \rightarrow d \in \{d_1, d_2, \ldots, d_g\}$
  where $d(\mathbf{x}_0) = d_i \quad \text{for} \quad \mathbf{x}_0 \in R_i$

Misclassifications: Confusion matrix

Classify as →	d_1	d_2	...	d_g
True π_1	$P(1\|1)$	$P(2\|1)$	...	$P(g\|1)$
↓ π_2	$P(1\|2)$	$P(2\|2)$	...	$P(g\|2)$
...	...	...	...	...
π_g	$P(1\|g)$	$P(2\|g)$	...	$P(g\|g)$

where $P(k|i) = \int_{R_k} f_i(\mathbf{x}) d\,\mathbf{x}, \quad \mathbf{x} \in R^p$

Misclassifications: Cost matrix

Classify as →	d_1	d_2	...	d_g
True π_1	0	$c(2\|1)$	...	$c(g\|1)$
↓ π_2	$c(1\|2)$	0	...	$c(g\|2)$
...	...	...	...	...
π_g	$c(1\|g)$	$c(2\|g)$	...	0

where $c(k|i) \geq 0, \ i \neq k$

Priors

- $p_i = P(\pi_i), \ i = 1, \ldots, g$
  where $\sum_{i=1}^{g} p_i = 1$

Summarizing several-population discriminant analysis procedures

Training data

- collect "ground truth" data
- define or estimate prior probabilities, $p_i = P(\pi_i),\ i = 1, \ldots, g$
- define misclassification costs, $c(k|i) \geq 0,\ i, k = 1, \ldots, g,\ i \neq k$
- estimate parameters for $f(\mathbf{x}|\pi_i)$, i.e. $\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Sigma}}_i,\ i = 1, \ldots, g$
- MVN model control
- test $H_0: \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_g$ >>> LDA case for accept / QDA case for reject
- test $H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_g$ >>> significant separation of populations for reject
- specify/calculate sample based discriminant functions for LDA or QDA
- calculate confusion matrix for training data using the discriminant function
- calculate estimated classification error rates, APER and/or AER_{hat}

New data

- observe new data $\mathbf{x}_0$
- classify $\mathbf{x}_0$ as member of one of the populations 1,2,...,g using the calculated discriminant functions for the training data
- calculate posterior probabilities using
  - $P(\pi_i|\mathbf{x}_0) \propto p_i\, f(\mathbf{x}_0|\pi_i),\ i = 1, \ldots, g$ (posterior$_i$ $\propto$ prior$_i$·likelihood$_i$)
    where $\sum_{i=1}^{g} P(\pi_i|\mathbf{x}_0) = 1$ (normalization to actual probabilities)