ETC3250/5250: Regularization

Semester 1, 2020

Professor Di Cook

Econometrics and Business Statistics
Monash University

Week 9 (a)

Penalised LDA

Recall: LDA involves the eigen decomposition of $\Sigma^{-1}\Sigma_B$, where

$$\Sigma_B = \frac{1}{K}\sum_{i=1}^{K}(\mu_i - \mu)(\mu_i - \mu)'$$

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_i)(x_i - \mu_i)'$$

The eigen-decomposition is an analytical solution to a sequential optimisation problem:

$$\underset{\beta_k}{\text{maximize}}\,\beta_k^T\hat{\Sigma}_B\beta_k$$

$$\text{subject to } \beta_k^T\hat{\Sigma}\beta_k \leq 1, \beta_k^T\hat{\Sigma}\beta_j = 0\ \forall i < k$$

Penalised LDA

The problem is inverting Σ^{-1}, fix it by

$$\underset{\beta_k}{\text{maximize}} \left(\beta_k^T \hat{\Sigma}_B \beta_k + \lambda_k \sum_{j=1}^{p} |\hat{\sigma}_j \beta_{kj}| \right)$$

$$\text{subject to } \beta_k^T \tilde{\Sigma} \beta_k \leq 1$$

where $\hat{\sigma}_j$ is the within-class standard deviation for variable j. This is penalised LDA, and see reference, and the R package.

PDA Index

Penalised LDA projection pursuit index. Available in the `tourr` package.

$$I_{PDA}(A, \lambda) = 1 - \frac{\left| A'\left\{ (1 - \lambda)\hat{\Sigma} + n\lambda I_p \right\} A \right|}{\left| A'\left\{ (1 - \lambda)(\hat{\Sigma}_B + \hat{\Sigma}) + n\lambda I_p \right\} A \right|}$$
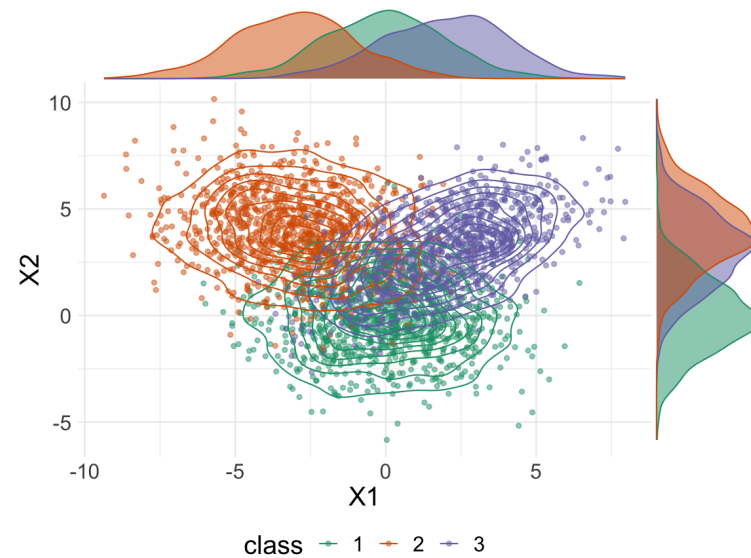
Optimising this function over $p \times d$ projection matrix A.

Diagonal Discriminant Analysis

📊 The simplest form of regularisation assumes that the features are independent within each class.
📊 Consider a *diagonal-covariance* LDA rule for classifying classes
📊 A special case of the naive-Bayes classifier

Discriminant Function

It can be shown that the discriminant score for a new observation $\mathbf{x}^*$ when the features are considered independent reduces to the following:

$$\delta_k(\mathbf{x}^*) = -\sum_{j=1}^{p} \frac{(x_j^* - \bar{x}_{kj})^2}{s_j^2} + 2\log \pi_k.$$

The classification rule is then

$$C(\mathbf{x}^*) = \ell \quad \text{if} \quad \delta_\ell(\mathbf{x}^*) = \max_k \delta_k(\mathbf{x}^*).$$

Filter features for prediction

To motivate the upcoming method, consider a binary classfication DLDA problem.

One way we could establish which of the features are driving prediction would be to perform a two-sample t-test

$$t_j = \frac{\bar{x}_{1j} - \bar{x}_{0j}}{s_j}$$

with the t statistic providing a measure of how significant the difference in class means for predictor j.
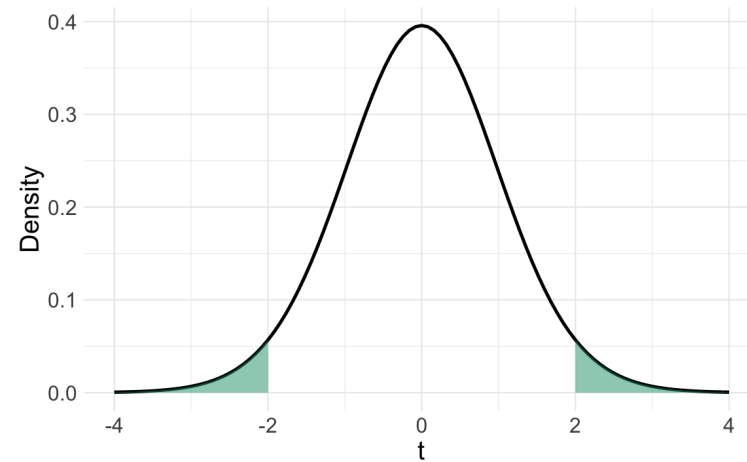
Filter features for prediction

Think about it: Using the t statistic - $t_j = \frac{\bar{x}_{1j} - \bar{x}_{0j}}{s_j}$ for all features, what is one way we can determine important features for prediction?

00:32

Filter features for prediction

Answer: Can consider filtering for features with $|t_j| > 2$, as this is deemed significant at the 5% level.

Note - further consideration can be given to the issue of *Multiple Testing*

Nearest Shrunken Centroids (NSC)

Now consider the following statistic,

$$d_{kj} = \frac{\bar{x}_{kj} - \bar{x}_j}{m_k(s_j + s_0)} \quad \text{with} \quad m_k^2 = \frac{1}{N_k} - \frac{1}{N}$$
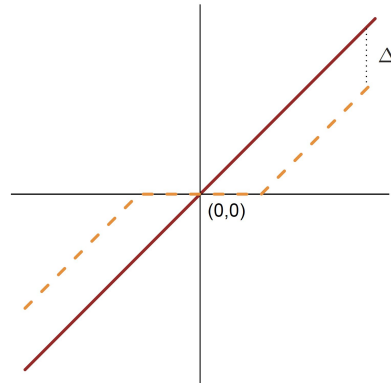
and s_0 a small value to protect d_{kj} from small expression values.

This statistic is a measure for how significant the difference between the class k mean for predictor j, and the overall mean for predictor j.

Soft Thresholding

Each d_{kj} is reduced by an amount Δ in absolute value, and is set to zero if its absolute value is less than zero.

$$d'_{kj} = \text{sign}(d_{kj})(|d_{kj}| - \Delta)_+,$$

Nearest Shrunken Centroids Classifier

The NSC uses either version of the statistic d'_{kj} to regularise by shinking the class means towards the overall mean for each predictor seperately as follows:

$$\bar{x}'_{kj} = \bar{x}_j + m_k(s_j + s_0)d'_{kj}$$

Unless a predictor has a significant difference to the overall mean for at least one class, it is useless for classification.

We then use the shrunken centroids $\bar{x}'_{kj}$ in place of $\bar{x}_{kj}$ in the DLDA discriminant function.

Alternative - penalised multiple hypothesis testing (multiDA)

Another approach to high dimensional DA involves formulating the problem as a multiple hypothesis testing problem, and asking the question - "What defines a discriminative feature?", and then choosing discriminative features through a penalised likelihood ratio test.

LRT – compare to the null

For $K = 3$ classes, there are $m = 5$ potential partitions of the data.

For all 5 hypotheses, compare the likelihood to the null. Pick the "partition" that is the most likely.
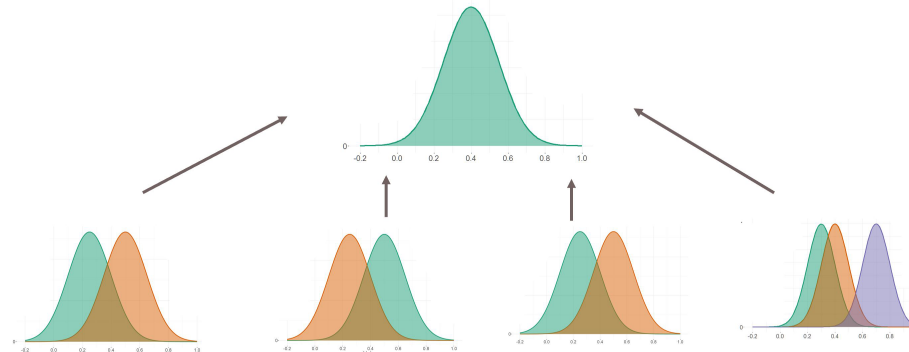
A penalised likelihood ratio test statistic

Two forms of penalisation can be considered:

📊 The BIC - useful when Positive Selection Rate is preferred to controlling False Discovery Rate (FDR).

$$\nu_m \log(n)$$

📊 The Extended BIC - useful for high dimensional data, penalising additionally on the number of features p.

$$\nu_m [\log(n) + 2 \log(p)]$$

(Note - $\nu_m = g_m - 1$ where g_m is the number of groupings considered in model m).

Recall - SRBCT cancer prediction

📊 The SRBCT dataset (Khan et al., 2001) looks at classifying 4 classes of different childhood tumours sharing similar visual features during routine histology.

📊 Data contains 83 microarray samples with 1586 features.

📊 We will revisit this data later on in the course to explore high dimensional DA. Now is that time
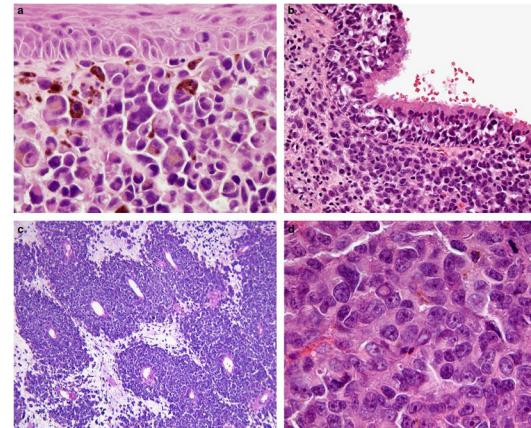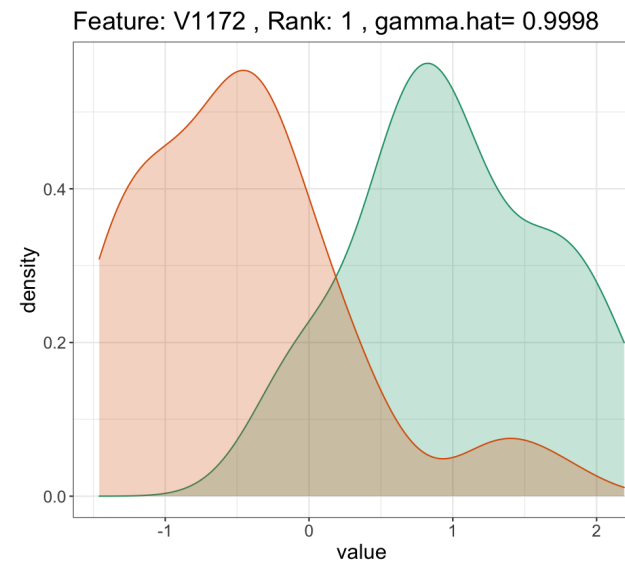

Source: Nature

multiDA in R

```r
#remotes::install_github("sarahromanes/multiDA")
library(multiDA)
res <- multiDA(y = SRBCT$y,
               X = SRBCT$X,
               penalty = "EBIC",
               equal.var = TRUE,
               set.options = "exhaustive")
```
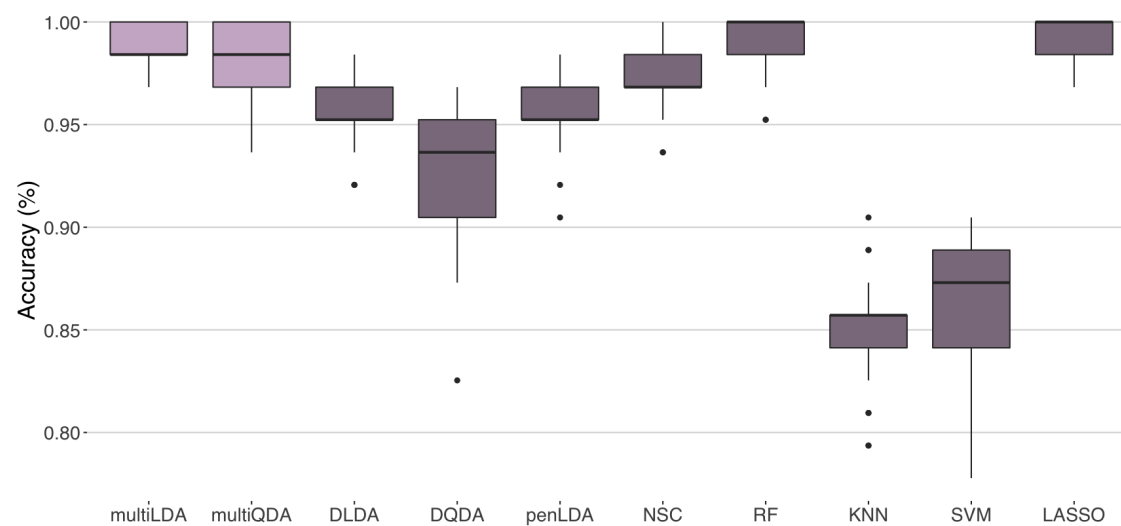
We can then examine the class groupings using the `plot()` method for `multiDA`:

```r
plot(res, ranks= 1)
```


Feature: V1172 , Rank: 1 , gamma.hat= 0.9998

Compare performance - 100 trial, 5 fold CV

👩‍💻 Made by a human with a computer

Slides at https://iml.numbat.space.

Code and data at https://github.com/numbats/iml.

Created using R Markdown with flair by **xaringan**, and **kunoichi** (female ninja) style.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.