Statistical learning for biological data

David Causeur
Department of Statistics and Computer Science
Agrocampus Ouest
IRMAR CNRS UMR 6625
http://www.agrocampus-ouest.fr/math/causeur/

Outline

Outline

Outline

Latent predicting variables

Let $x^* = (x_1^*, \ldots, x_p^*)'$ denote the profile of scaled predictors

then the best univariate regression model

$$Y = \beta_0 + \beta_j x_j^* + \varepsilon_j, \text{ where } \hat{\beta}_0 = \bar{Y} \text{ and } \hat{\beta}_j = s_{x_j^* y}$$

has the largest $R_j^2 = \hat{\beta}_j^2 / s_y^2$ or, equivalently, the largest $s_{x_j^* y}^2$.

Latent predicting variable: $t = \alpha_1 x_1^* + \ldots + \alpha_p x_p^*$ such that s_{ty}^2 is maximal among all possible linear combinations with $\alpha_1^2 + \ldots + \alpha_p^2 = 1$.

- Extraction of the latent variable using `R`

Latent predicting variables

Y is related to t by a linear regression model:

$$Y = b_0 + bt + \varepsilon$$

- Regression modeling using the latent variable in R

Latent predicting variables

Not all the explanatory information is concentrated in t:

- First 'deflate' the explanatory variables from t:

$$x_k = b_{0k} + b_{1k}t + e_k, \quad e_k : \text{ deflated } x_k \text{ from t}$$

- Then extract a 2nd latent variable from the scaled e_k

$$t_2 = \alpha_1^{(2)} e_1^* + \ldots + \alpha_p^{(2)} e_p^*$$

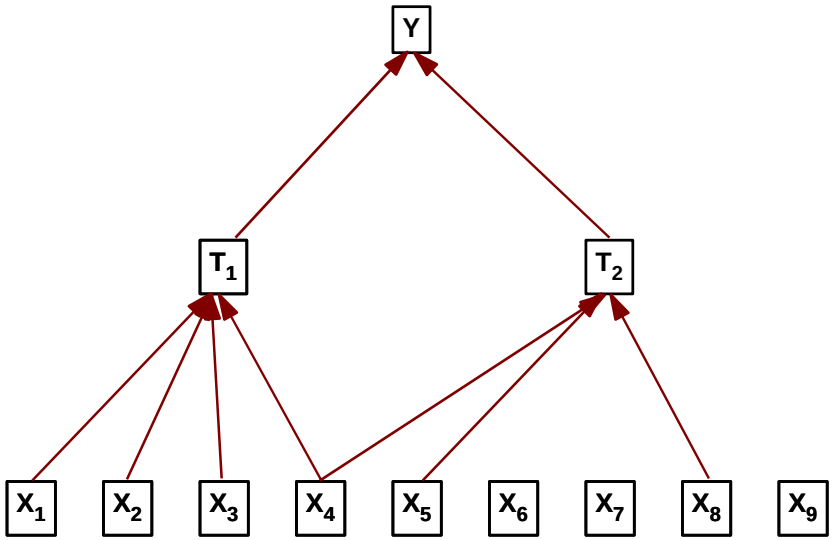
where $s^2_{t_2 y}$ is maximal among all possible linear combinations of e_k^* with $[\alpha_1^{(2)}]^2 + \ldots + [\alpha_p^{(2)}]^2 = 1$.

- Extraction of a 2nd latent variable using R

Latent predicting variables

Y is related to the LVs by a linear regression model:

$$Y = b_0 + b_1 t_1 + b_2 t_2 + \varepsilon$$



- Comparison of latent variable models using `R`

Dimensionality of a regression model

Dimensionality: Optimal number k of LVs needed to predict Y

- If $k = \min(n, p)$, then OLS fit of the full regression model
- If $k < \min(n, p)$, then Partial-Least-Squares (PLS)
- k can be determined by CV

▶ Choosing the number of PLS components using `R`

Dimensionality of a regression model

To sum up:

- PLS regression extracts latent explanatory variables;
- The number of LVs can be very limited w.r.t the number of variables;
- For the same number of LVs, PLS does better than Regression on Principal Components (PCR)

... PCR is also implemented in package `PLS` (function `pcr`)

The LDA framework

$Y \in \{y_1, \ldots, y_K\}$ is a K-class variable

Y has prior probabilities $\pi_k = \mathbb{P}(Y = y_k)$

X is a p-vector of explanatory variables

$$\begin{aligned} X = (X_1, \ldots, X_p)' &\sim \mathcal{N}(\mu_k; \Sigma), \text{ given } Y = y_k; \\ &\sim \pi_1 \mathcal{N}(\mu_1; \Sigma) + \ldots + \pi_K \mathcal{N}(\mu_K; \Sigma); \end{aligned}$$

where

- μ_k is the mean vector in class k,
- Σ is the within-class variance-covariance matrix.

Prediction based on posterior class probabilities

Y has posterior probabilities:

$$\mathbb{P}(Y = y_k \mid X = x) \quad = \quad \pi_k \frac{f(x; \mu_k; \Sigma)}{\pi_1 f(x; \mu_1; \Sigma) + \ldots + \pi_K f(x; \mu_K; \Sigma)};$$

where $f(.; \mu, \Sigma)$ is the density function of the multivariate normal distribution with mean μ and variance Σ.

Prediction based on the Bayes rule:

$$\hat{Y} = y_{k^\star} \text{ if } \mathbb{P}(Y = y_{k^\star} \mid X = x) = \max_{k=1,\ldots,K} \mathbb{P}(Y = y_k \mid X = x).$$

- Bayes prediction rule using R

Fisher's LDA score

Let us focus on the two-class prediction issue, with $p = 1$:

$$\begin{aligned}\log\frac{\mathbb{P}(Y = y_2 \mid X = x)}{\mathbb{P}(Y = y_1 \mid X = x)} &= \log\left(\frac{\pi_2}{\pi_1}\right) + \frac{\mu_2 - \mu_1}{\sigma^2}\left(x - \frac{\mu_1 + \mu_2}{2}\right), \\ &= L(x; \mu_1, \mu_2, \sigma), \quad \text{[Bayes linear classifier]}.\end{aligned}$$

Bayes prediction rule (two-class, $p = 1$):

$$\hat{Y} = y_2 \text{ if } \log\left(\frac{\pi_2}{\pi_1}\right) + \frac{\mu_2 - \mu_1}{\sigma^2}\left(x - \frac{\mu_1 + \mu_2}{2}\right) > 0.$$

Fisher's LDA score estimates Bayes Linear classifier:

$$\begin{aligned}\hat{L}(x) &= L(x; \bar{x}_1, \bar{x}_2, s), \\ &\propto \frac{\bar{x}_2 - \bar{x}_1}{s^2}\left(x - \frac{\bar{x}_1 + \bar{x}_2}{2}\right)\end{aligned}$$

Fisher's LDA score

Still with $K = 2$, but now $p > 1$, the same with matrix notations:

$$\hat{L}(x) \propto (\bar{x}_2 - \bar{x}_1)' W_x^{-1} \left(x - \frac{\bar{x}_1 + \bar{x}_2}{2} \right) = \hat{\beta}' \left(x - \frac{\bar{x}_1 + \bar{x}_2}{2} \right)$$

where W_x is the within-class variance-covariance matrix of x, $\hat{\beta} = W_x^{-1}(\bar{x}_2 - \bar{x}_1)'$.

Interestingly, $\hat{L}(x)$ is the linear score with largest ANOVA F-test statistic for the group mean comparison issue.

Therefore, now with $K > 2$ and $p > 1$: Fisher's LDA score is defined as the linear score with largest ANOVA F-test statistic for the group mean comparison issue.

- Fisher's LDA score using `R`

A geometrical viewpoint

Prediction based on the MAP class probability ($K = 2$):

$$\hat{Y} = y_2 \text{ if } \hat{\mathbb{P}}(Y = y_2 \mid X = x) \geq \hat{\mathbb{P}}(Y = y_1 \mid X = x).$$

Equivalently:

$$\hat{Y} = y_2 \text{ if } \Delta^2(L; \bar{L}_2) \geq \Delta^2(L; \bar{L}_1);$$

where

- $\bar{L}_1$ and $\bar{L}_2$ are the class means of the Fisher score $\hat{L}$
- $\Delta^2(L; \bar{L}) = (L - \bar{L}_k)^2 - 2 \log p_k$
- p_1 and p_2 are prior class probabilities

- Minimum distance prediction using `R`

Multiclass LDA

1st LD score: $L_1(x) = \beta_1^{(1)} x_1 + \ldots + \beta_p^{(1)} x_p$ where the $\beta_j^{(1)}$ are such that the F-statistic for the class comparison is as large as possible.

2nd LD score: $L_2(x) = \beta_1^{(2)} x_1 + \ldots + \beta_p^{(2)} x_p$ where the $\beta_j^{(2)}$ are such that:

- The sample covariance of L_2 and L_1 is zero;
- the F-statistic of L_2 for the class comparison is as large as possible under the restrictions above.

..... and so on until the $(K-1)$th LD score.

- K-class LDA using `R`

LDA in high-dimension

In high-dimension, W_x^{-1} does not exist

If Z stands for the dummy coding of Y, then LDA can be reformulated as a least-squares minimization issue:

- A penalization can be added to obtain a sparse LDA fitting algorithm;
  see package `sparseLDA`

- Considering Z as a profile of quantitative responses leads to a PLS approach
  see package `mixOmics` for PLS-DA or even sPLS-DA