Machine Learning

January 18, 2017

1 La regression dans tous ses états

1.1 Introduction à la régression multiple

La régression muyltiple permet d'étudier la liaison entre une variable (à expliquer) Y et un ensemble de p variables explicatives $X_1, \ldots, X_p$.

Le modèle de la régression multiple est définie par

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \tag{1}$$

où les coefficients de régression β_j sont des paramètres fixe mais inconnus, et ϵ un terme aléatoire suivant une $\mathcal{N}(0, \sigma^2)$

1.1.1 Données et modèle statistique

On dispose de n observations des variables $X_1, \ldots, X_p$ Soit

$$X = \begin{bmatrix} y_1 & x_{11} & \ldots & x_{p1} \\ \ldots & \ldots & \ldots & \ldots \\ y_m & x_{1m} & \ldots & x_{pm} \end{bmatrix} \tag{2}$$

(Tableau d'observations)

La modèle statistique est définie comme

Pour chaque individu t, on considère que la valeur y_i prise par Y est une réalisation d'une variable aléatoire Y_i définie par

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_p X_{pi} + \epsilon_i \tag{3}$$

où σ_i est un terme aléatoire suivant une $\mathcal{N}(0, \sigma^2)$. Il faut supposer de plus que les $\epsilon_1, \ldots, \epsilon_m$ sont indépendantes les un les autres.

Sur l'exemple des automobiles, on considère que le prix Y_i de la voiture i suit une loi normale de moyenne

$$\mu_i = \beta_0 + \beta_1 PUISSANCE_i, \ldots, \beta_p LARGEUR_i + \epsilon. \tag{4}$$

1.1.2 Estimations des paramètres du modèle

À l'aide des n observations des variables $Y, X_1, \ldots, X_p$, nous allons chercher à estimer les paramètres $\beta_0, \ldots, \beta_p$ du modèle.

On cherche $\hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)$ tel que

$$\hat{\beta} = argmin||y - X\beta||_2^2 = argmin(y - X\beta)^T(y - X\beta) \tag{5}$$

Géométriquement, cas où $p = 1$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 \tag{6}$$

On cherche $\hat{\beta}_0$ et $\hat{\beta}_1$ tel que $\sum \epsilon_i^2$ soit minimal.

$$\frac{\partial L}{\partial \beta} = -X^T y - X^T y + 2X^T X\hat{\beta} = 0 \tag{7}$$

Ainsi

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{8}$$

remarque 1: Notons que $\hat{\beta}$ est de la forme Hy
remarque 2:

$$\hat{\beta} = (X^T X)^{-1} X^T y = (X^T X)(X^T X)^{-2} X^T y = X^T \hat{\alpha} \tag{9}$$

Cela signifie que $\hat{\beta}$ s'exprime comme la combinaison linéaire des individus

$$\hat{\beta} = \sum_{i=1}^{n} \hat{\alpha}_i x_i \tag{10}$$

Remarque 3 Posons

$$X = V\Sigma U^T \tag{11}$$

(Décomposition en valeurs singuliers)

$$V^T V = I \tag{12}$$

et Σ est diagonale et $U^T U = I$ Posons

$$\Lambda = \Sigma^T \Sigma \tag{13}$$

matrice diagonale des valeurs de $X^T X$. Ainsi

$$\begin{aligned} \hat{\beta}^{OLS} &= (X^T X)^{-1} X^T y = (U\Sigma V^T V \Sigma U^T)^{-} 1 U \Sigma V^T y \\ &= U\Sigma^{-1}\Sigma^{-1} U^T U \Sigma V^T y \\ &= U\Lambda^{-1}\Sigma V^T y \\ &= \sum_{j=1}^{P} \frac{v_j^T y}{\sqrt{\lambda_j}} \mu_j \end{aligned} \tag{14}$$

Interprétation géométrique de la régression multiple...
Pour un nouvel individu x

$$\hat{y}(x) = x^t \hat{\beta} \tag{15}$$

Remarque 4
En termes de prédiction, on peut aussi obtenir des expressions duales

$$\hat{y} = x^T \hat{\beta} = \sum_{j=1}^{n} \alpha_j x_j^T x \tag{16}$$

Cette expression duale a la particularité de ne dépendre que des produits scalaires entre observations.

1.1.3 Qualité des estimations

On souhaite évaluer la précision des estimations

$$\begin{aligned}\mathbb{E}[\hat{\beta}] &= \mathbb{E}[(X^TX)^{-1}X^TY] = (X^TX)^{-1}X^T\mathbb{E}[Y] \\ &= (X^TX)^{-1}X^T\mathbb{E}[X\beta + \epsilon] \\ &= (X^TX)^{-1}(X^TX)\beta \\ &= \beta\end{aligned} \tag{17}$$

$$\begin{aligned}Var(\hat{\beta}) &= var((X^TX)^{-1}X^TY) \\ &= (X^TX)^{-1}X^Tvar(Y)X(X^TX)^{-1} \\ &= \sigma^2(X^TX)^{-1}\end{aligned} \tag{18}$$

où

$$var(Y) = \sigma^2 I \tag{19}$$

Le MSE(Mean Square Error) d'un estimateur $\hat{\beta}$ d'un vecteur β est défini

$$\begin{aligned}MSE(\hat{\beta}) &= \mathbb{E}[tr(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T] \\ &= \mathbb{E}[(\hat{\beta} - \beta)^T(\hat{\beta} - \beta)] \\ &= \mathbb{E}[||\hat{\beta} - \beta||_2^2] \\ &= [\mathbb{E}[\hat{\beta} - \beta]^T][\mathbb{E}[\hat{\beta}] - \beta] + \mathbb{E}[(\hat{\beta} - \mathbb{E}[\hat{\beta}])(\hat{\beta} - \mathbb{E}[\hat{\beta}])] \\ &= biais^2(\hat{\beta}) + tr(var(\hat{\beta})) \\ &= \sigma^2 tr(X^TX)^{-1} \\ &= \sigma^2 \sum_{j=1}^{P} \frac{1}{\lambda_j}\end{aligned} \tag{20}$$

Si les données sont mal-conditionnées, alors $\Rightarrow$ petit valeurs de X^TX $\Rightarrow$ instabilité des coefficients de regression $\Rightarrow$ Explosion du MSE $\Rightarrow$ écart entre β et $\hat{\beta}$

Illustration de la multi-linéarité

Considérons le modèle suivant

$$Y = \beta_1 X_1 + \beta_2 X_2 + \Sigma \tag{21}$$

On suppose que les données sont standardisées

$$cor(X_1, X_2) = r_{12}, cor(X_j, Y) = r_{jy} \tag{22}$$

L'estimation des moindres carrées

$$\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2) \tag{23}$$

est donné par

$$(X^TX)\hat{\beta} = X^TY \tag{24}$$

Comme les données sont standardisées

$$\begin{pmatrix} 1 & r_{12} \\ r_{21} & 1 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} \sim \begin{pmatrix} r_{1y} \\ r_{2y} \end{pmatrix} \tag{25}$$

L'inverse de $X^T X$ est donnée par

$$C = (X^T X)^{-1} = \frac{1}{1 - r_{12}^2} \begin{pmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{pmatrix} \tag{26}$$

On rappelle que

$$var(\hat{\beta}) = \sigma^2 (X^T X)^{-1} \tag{27}$$

donc

$$var(\hat{\beta}_j) = \sigma^2 C_j \tag{28}$$

Alors, une forte corrélation entre X_1 et X_2 est indiquée par $|r_{12}| \to 1$. Ceci implique que $var(\hat{\beta}_j) \to +\infty$. Ainsi, MSE qui explose.

De maniére générale, on peut montrer que dans le cas de p variables explicatives, des elements diagonaux de $C = (X^T X)^{-1}$ sont égaux à

$$C_j = \frac{1}{1 - R_j^2} \tag{29}$$

où R_j^2 est le coefficient de determination entre X_j et les p autres variables. S'il existe une forte multi-colinéarité entre X_j et les $(p-1)$ autres variables. $R_j^2 \to 1$

Exemple II

(Voir photos)

1.2 Facteur de shrinkage

On rappelle que

$$\hat{\beta}^{OLS} = (X^T X)^{-1} X^T y = \sum_{j=1}^{P} \frac{v_j^T y}{\sqrt{\lambda_j}} \mu_j \sum_{j=1}^{P} z_j \tag{30}$$

Dans sa forme générale, on féfinit un estimateur générale de β par

$$\hat{\beta}^{shi} = \sum_{j=1}^{P} f(\lambda_j) z_j \tag{31}$$

$f(\lambda_j)$ est un facteur de shrinkage qui va jouer sur le MSE. Ici, on se concentre sur les facteurs de shrinkage indépendant de y. Malheureusement, on a

$$\hat{\beta}^{shi} = U \Sigma^{-1} D V^T y = H^{shi} y \tag{32}$$

avec

$$D = diag(f(\lambda_1), \ldots, f(\lambda_p)) \tag{33}$$

Étudions l'influence de (Photos)

$$\begin{aligned}\hat{\beta} =& argmin_{\beta\in\mathbb{R}^p}\{||y - X\beta|| + \lambda||\beta||_2^2\} \\ =&(X^TX + \lambda I_p)^{-1}X^Ty \\ =&\sum_{j=1}^{p}\frac{\lambda_j}{\lambda_j+\lambda}\frac{v_j^Ty}{\sqrt{\lambda_j}} \\ =&\sum_{j=1}^{p}f(\lambda_j)z_j\end{aligned} \tag{34}$$

$$\hat{y}_\lambda = X\hat{\beta}_\lambda^{RR} = V\Sigma U^TU(\Lambda+\lambda I)^{-1}\Sigma\Lambda y = V\Sigma(\Lambda+\lambda I)^{-1}\Sigma V^Ty \tag{35}$$

1.2.1 Choix de modèle (λ) + ridge path validation croisée

- Ensemble d'apprentissage permet de construire le modèle
- Ensemble de test permet d'évaluer la qualité du modèle

Nécessite de déterminer λ sur la base d'un critére objectif

Un modèle devrait predire efficacement des individus qui n'ont pas servi à sa construction

Pour évaluer la qualité du modèle, on utilise des stratégies de validation croisée.

K-fold cross validation

1. Partition du jeu de données en K parties de taille égale $T_1, \ldots, T_K$
2. Pour chaque $k = 1, \ldots, K$ construire $\hat{\beta}_\lambda^{-k}$ sans T_k
3. $\hat{y}_\lambda^{-k} = T_k\hat{\beta}_\lambda^{-k}$
4. Calcul d'erreurs en test

$$CV_{erreur}(\lambda, k) = \frac{1}{n_k}\sum(y_k - \hat{y}_\lambda^{-k})^2 \tag{36}$$

Ainsi, Taux d'erreur global(λ)

$$= \frac{1}{K}\sum_{k=1}^{K}CV_{erreur}(\lambda, k) \tag{37}$$

Ainsi,

$$\lambda^* = argmin(Tauxd'erreurglobal(\lambda)) \tag{38}$$

On appelle que

$$\hat{\beta}^{RR} = argmin\{||y - x||_2^2 + \lambda||\beta||_2^2\} = (X^T X + \lambda I)^{-1} X^T y \tag{39}$$

Proposition $\forall \lambda > 0$ on a

$$(X^T X + \lambda I_p)^{-1} X^T = X^T (X X^T + \lambda I_\lambda)^{-1} \tag{40}$$

Preuve

$$(X^T X + \lambda I_p)^{-1} X^T (X X^T + \lambda I_n) = X^T \tag{41}$$

$$(X^T X + \lambda I_p)^{-1} (X^T X + \lambda I_p) X^T = X^T \tag{42}$$

Il vient de cette proposition une formulation duale pour $\hat{\beta}^{RR}$

$$\begin{aligned} \hat{\beta}^{RR} =& X^T (X X^T + \lambda I_n)^{-1} y \\ =& X^T \hat{\alpha}^{RR} \end{aligned} \tag{43}$$

Remarque: $\hat{\beta}^{RR}$ s'exprime comme une combinaison linéaire des observations.
En termes de prédiction

$$\hat{y}_\lambda = X \hat{\beta}_\lambda^{RR} = X X^T \hat{\alpha}^{RR} \tag{44}$$

$$h(x) = \sum_{i=1}^{n} \alpha_i x_i^T x \tag{45}$$

Conclusion
On remarque que $\hat{\alpha}$ et $\hat{y}$ ne s'expriment qu'au (Photo)

2 Méthode à noyaux

Introduction Supposons que l'on recherche à résoudre un problème de régression non-linéaire.

Il est tout à fait possible d'utiliser une fonction de redescription $\Phi : \mathbb{R}^p \rightarrow \mathcal{F}$(espace de description). La solution du problème de régression linéaire dans l'espace de redescription est de la forme

$$\begin{aligned} h(x) =& < \beta, \Phi(x) > \\ =& \sum_{j=1}^{n} \beta_j \Phi_j(x) + \beta_0 \end{aligned} \tag{46}$$

avec $Phi(x) = \{\Phi_1(x), \Phi_2(x), \ldots\}$

En résolvant comme précédemment dans l'espace de redescription, devient la représentation duale.

$$h(x) = \sum_{i=1}^{n} \Phi(x_i)^T \Phi(x) \tag{47}$$

En rapprochant $h(x) = \beta^T \Phi(x)$ et $h(x) = \sum_{i=1}^{n} \alpha_i \Phi(n_i)^T \Phi(x)$ devient

$$\beta^T = \sum_{i=1}^{n} \alpha_i \Phi(x_i)^T \tag{48}$$

Commentaire 1: Comment trouver une fonction de redescription adéquate?

Commentaire 2: Coût calculatoire des produits scalaires dans un espace $\mathcal{F}$ dont la dimension peut être très grande.

Plutôt que de représenter les observations par un tableau individus x variables, on peut les représenter par une matrice de similarité de dimension $n \times n$

$$[K]_y = k(x_i, x_j) \tag{49}$$

$$[X] \sim [K] \tag{50}$$

$$(x_i)_{n \times p} \sim (k(x_i, x_n))_{n \times n} \tag{51}$$

Remarque: Mesure de similarité quelle que soit la nature et la complexité des objets.(images, graphs, séquences ADN,...) Remarque 2: La matrice K est toujours de dimension $n \times n \forall$ la taille de l'objet.

Dans ce cours nous nous restrienons àune classe particulières de fonction k.

Définition Fonction semi definie positive

Une fonction k semi definie positive sur l'ensemble X est une fonction $k : X, X \to \mathcal{R}$ symmétrique

$$\forall (x, x') \in X, k(x, x') = k(x', x) \tag{52}$$

et quisatisfait pour (Photo)

et $(a_1, \ldots, a_n) \in \mathcal{R}^n$

$$\sum_{i,j=1}^{n} a_i a_j k(x_i, k_j) >= 0 \tag{53}$$

En d'autres termes $K = (k(x_i, x_j))$ est semi definie positive.

Exemple 1: Le plus simple des noyaux semi definie positif

Soit $X \in \mathbb{R}^P$ et la fonction $k : \mathcal{X}^2 \to \mathbb{R}$ definie par

$$\forall (x, x') \in (\mathbb{R}^p)^2, k(x, x') =< x, x' >_{\mathbb{R}^p} \tag{54}$$

Symétrie

$$< x, x' >_{\mathbb{R}^p} =< x', x >_{\mathbb{R}^p} \tag{55}$$

Théorème de Moose-Aroszajin, 1950

(Photo)

Definition: Soit $\mathcal{H}$

(Photo)

2. La fonction k est une fonction noyau reproduisante

$$\forall f \in \mathcal{H}ona < f, k(n, .) >= f(x) \tag{56}$$

Le produit scalaire est alors défini comme suit:
Soit f et $g \in \mathcal{H}$ définies par

$$f(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x) \tag{57}$$

et

$$g(x) = \sum_{i=1}^{n} \beta_j k(x_j, x) \tag{58}$$

alors

$$\begin{aligned} < f, g > &= \sum_{j=1}^{n} \sum_{i=1}^{n} \alpha_i \beta_j k(x_i, x_j) \\ &= \sum_{i=1}^{n} \alpha_i g(x_i) \\ &= \sum_{j=1}^{n} \beta_j f(x_j) \end{aligned} \tag{59}$$

Representation theorem (Kimeldoif & Wabhia 1970)
Soit $\mathcal{H}$ un RKHS associé à un noyau k. Soit $\Omega[0, +\infty[\rightarrow \mathbb{R}$ une fonction strictement croissante Soit $\mathcal{X}$ un emsemble et $l(\mathcal{X} \times \mathbb{R}) \rightarrow \mathbb{R}$ une fonction de coût.

Alors toute solution du problèle d'optimisation suivant

$$\min_f J_f = l((x_1, y_1, f(x_1)), \ldots, (x_n, y_n, f(x_n))) + \lambda \Omega(||f||_{\mathcal{H}}) \tag{60}$$

admet une représentation de la forme $f^* = \sum_{i=1}^{n} \alpha_i k(x_i, .)$
Remarque 1

$$l((x_1, y_1, f(x_1)), \ldots, (x_n, y_n, f(x_n))) + \lambda \Omega(||f||_{\mathcal{H}}) \tag{61}$$

conduit à la regression rigide.

Remarque 2: Quelque soit la dimension de l'espace $\mathcal{H}$, la solution évolue dans $span\{k(x_i, .), 1 <= i <= n\}$ de dimension n connue, bien que H puisse être de dimension imporeante. $\Rightarrow$ Developpement d'algo thno efficace.

Preuve du théoreme du representant Supposons $f \in \mathcal{H}$ projectée sur $spank(x_i, .), 1 <= i <= n$ Soit f_s(la composante dans l'espace engendré) et $f_\perp$ la composante orthogonale

$$f = f_s + f_\perp \tag{62}$$

Ainsi

$$||f||_{\mathcal{H}}^2 = ||f_s||_{\mathcal{H}}^2 + ||f_\perp||_{\mathcal{H}}^2 >= ||f||_{\mathcal{H}}^2 \tag{63}$$

Puisque Ω est strictement croissante, on a

$$\Omega(||f||_{\mathcal{H}}^2) > \Omega(||f_s||_{\mathcal{H}}^2) \tag{64}$$

Cela signifie que Ω est minimisée pour des fonctions $f \in span\{k(x_i,.), 1 <= i <= n\}$

Par ailleurs, des propriétés reproduisantes de k on obtient

$$f(x_i) =< f, k(x_i,.) >=< f_s, k(x_i,.) >_{\mathcal{H}} + < f_\perp, k(x_i,.) >_{\mathcal{H}} =< f_s, k(x_i,.) >= f_s(x_i) \tag{65}$$

Par conséquent,

$$l((x_1, y_1, f(x_1)), \ldots, (x_n, y_n, f(x_n))) = ((x_1, y_1, f_s(x_1)), \ldots, (x_n, y_n, f_s(x_n))) \tag{66}$$

Donc l ne dépend que des composantes de $f \in span\{k(x_i,.), 1 <= i <= n\}$ et Ω est minimisée que si $f \in$ cet espace

Conclusion: $J(f)$ est minimisée que si $f \in span\{k(x_i,.), 1 <= i <= n\}$ et ;a solution est fonc de la forme

$$f^* = \sum_{i=1}^{n} \alpha_i k(x_i,.) \tag{67}$$

Kernel Rigide Regression

Un point d'une fonctionnelle

Soit $\mathcal{H}$ un RKHS de noyau k. On considère le problème suivant

$$f^* = argmin_{f\mathcal{H}}\{\frac{1}{2}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \frac{\lambda}{2}||f||^2_{\mathcal{H}}\} \tag{68}$$

= termed'attache aux données + terme de régularisation.

Par le théorème de représentant, on sait que

$$f^* \sum_{j=1}^{n} \alpha_j k(x_j,.) \tag{69}$$

(Photo)

Exercise: Montrer que

(Photo)

Kernel Ridige Regression

(Photo)

Franchement, le choix de la fonction Phi se ramene au choix d'une fonction k sdp.

Exemple demonstratif:

similarity: focntion noyau

$$k(x_i, x_j) = (1 + x_i^T x_j)^2 \tag{70}$$

On va construire explicitement l'espace induit par la fonction k.

$$\begin{aligned} k(x_i, x_j) =& 1 + x_{i1}^2 x_{j1}^2 + 2x_{i1}x_{j1}x_{i2}x_{j2} + x_{i2}^2 x_{j2}^2 + 2x_{i1}x_{j1}x_{i2}x_{j2} \\ =& [1 x_{i1}^2 \sqrt{2}x_{i1}x_{i2} x_{i2}^2 \sqrt{2}x_{i1}\sqrt{2}x_{i2}]^T [1 x_{j1}^2 \sqrt{2}x_{j1}x_{j2} x_{j2}^2 \sqrt{2}x_{j1}\sqrt{2}x_{j2}] \\ =& \Phi(x_i)^T \Phi(x_j) \end{aligned} \tag{71}$$

Astude du noyau

$$\hat{\alpha}^{RR} = (XX^T + \lambda I_n)^{-1} y \tag{72}$$

$$\hat{\alpha}^{RR} = (K + \lambda I_n)^{-1} y \tag{73}$$

Autres exemples de noyaux, Noyau polynomial

$$k(x, y) = (c + x_i^T x_j)^p \tag{74}$$

Noyau gaussien

$$exp(-\frac{||x_i - x_j||^2}{2\sigma^2}) \tag{75}$$

3 Analyse discriminante

Introduction Soit Y une variable à expliquer QUALITATIVE à K catégorie $y \in 1, \ldots, K$

Soient $X_1, \ldots, X_p$ les p variables explicatives.

Objectif 1: L'analyse discriminante descriptive: Proposer un système de représentation qui permet de discerner le plus possible les differents groupes d'individus. *Rightarrow* L'analyse factorielle discriminante

Objective 2: Construire une fonction de classement qui permet de prédire le groupe d'appartenance d'un individu à partir des valeurs prises par les variables explicatives. *Rightarrow* Analyse Discriminante Bayesienne

Soit n_j le nombre d'observation appartenant au groupe j. Soit n le nombre d'observations total.

$$g = (g_1, \ldots, g_p) \tag{76}$$

les centres de gravité du nuage global. noyaux des X_j calculé à partir de toute les observations.

$$m_k = (m_{k1}, \ldots, m_{kp}) \tag{77}$$

centre de gravité du nuage des individus de la class k.

$$V = \frac{1}{n} \sum_{i=1}^{n} (x_i - y)(x_i - y)^T \tag{78}$$

matrice de covariance des données complètes.

$$B = \frac{1}{n} \sum_{k=1}^{K} n_k (g_k - g)(g_k - g)^T \tag{79}$$

$$W - \frac{1}{n} \sum_{k=1}^{K} n_k V_k \tag{80}$$

avec

$$V_k = \frac{1}{n_k} \sum_{i \in \mathcal{C}_k} (x_i - m_k)(x_i - m_k)^T \tag{81}$$

qui est la matrice de covariance intraclasse.

On peut montrer(exercise) que(voir TP)

$$V = B + W \tag{82}$$

Construisons la variance globale du nuage de points.

$$\begin{aligned} \sigma^2 =& \frac{1}{n}\sum_{i=1}^{n}(x_i - y)(x_i - y)^T \\ =& \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{L}_k}(x_i - g)^*(x_i - g) \\ =& \frac{1}{n}\sum_{k=1}^{K} SS(k) \end{aligned} \tag{83}$$

$$\begin{aligned} SS(k) &= \sum_{i\in\mathcal{L}_k}(x_i - g)^T(x_i - g) \\ &= \sum_{i\in\mathcal{L}_k}(x_i - m_k + m_k - g)^T(x_i - m_k + m_k - g) \\ &= \sum_{i\in\mathcal{L}_k}(||x_i - m_k||^2 + ||m_k - g||^2 + 2(x_i - m_k)(m_k - g)) \\ &= \sum_{i\in\mathcal{L}_k}(||x_i - m_k||^2 + ||m_k - g||^2) \end{aligned} \tag{84}$$

Donc, la variance totale s''ecrit

$$\begin{aligned} \sigma^2 =& \frac{1}{n}[\sum_{k=1}^{K}\sum_{i\in\mathcal{L}_k}||x_i - m_k||^2 + \sum_{k=1}^{K} n_k||m_k - m||^2] \\ =& \frac{1}{n}\sum_{k=1}^{K} n_k \frac{1}{n_k}\sum_{i\in\mathcal{L}_k}||x_i - m_k||^2 + \sum_{k=1}^{K}||m_k - g||^2 \\ =& \frac{1}{n}\sum_{k=1}^{K} n_k[\frac{1}{n_k}\sum_{i\in\mathcal{L}_k}||x_i - m_k||^2 + ||m_k - g||^2] \end{aligned} \tag{85}$$

Si l'on definit une operation de projection $\Pi = vv^T$avec $v^Tv = 1$

$$\begin{aligned} \sigma^2_{within} =& \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{L}_k}(\Pi_{x_i} - \Pi_{m_k})^T(\Pi_{x_i} - \Pi_{m_k}) \\ =& \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{L}_k}(vv^Tx_i - vv^Tm_k)^T(vv^Tx_i - vv^Tm_k) \\ =& \frac{1}{n}\sum_{k=1}^{K}(x_i - m_k)^Tvv^Tvv^T(x_i - m_k) \end{aligned} \tag{86}$$

$$
\begin{aligned}
\sigma_w^2 =& \frac{1}{n}\sum_{k=1}^{K}\sum_{[} i \in \mathcal{L}_k](x_i - m_k)^T v v^T (x_i - m_k) \\
=& \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{L}_k} v^T (x_i - m_k)(x_i - m_k)^T v \\
=& v^T[\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{L}_k} (x_i - m_k)(x_i - m_k)^T]v \\
=& v^T[\frac{1}{n}\sum_{k=1}^{K} n_k \sum_{i\in\mathcal{L}_k} (x_i - m_k)(x_i - m_k)^T]v \\
=& v^T v
\end{aligned}
\tag{87}
$$

Pour la variance between projetée sur V

$$
\begin{aligned}
\sigma^2_{between}(v) =& \frac{1}{n}\sum_{k=1}^{K} n_k(\Pi m_k - \Pi y)^T(\Pi m_k - \Pi y) \\
=& v^T[\sum_{k=1}^{K} n_k \frac{1}{n}(m_k - y)(m_k - y)^T]v
\end{aligned}
\tag{88}
$$

Comme $V = B + W$

$$
\sigma^2_{total}(v) + v^T V v \tag{89}
$$

L'analyse discriminante factorielle est définie par le problème d'optimisation suivant

$$
\max(\frac{v^T B v}{v^T W v}) \tag{90}
$$

On remarque si v^* est solution alors αv^* est également solution

$$
\max(v^T B v),\ \text{sc}\ v^T W v = 1 \tag{91}
$$

Remarque: On aurait pu également considérer le probléme d'optimisation suivant

$$
\max(\frac{v^T B v}{v^T V v}) \Leftrightarrow \max v^T B v\ \text{sc}\ v^T V v = 1 \tag{92}
$$

Considerons le lagrangien associée à ce problème d'optimisation

$$
L = v^T B v - \lambda(v^T V v - 1) \tag{93}
$$

En annulant la dérivée de L par rapport à v, devient

$$
\frac{\partial L}{\partial v} = 2Bv - \lambda V v = 0 \tag{94}
$$

Ainsi,

$$V^{-1}Bv = \lambda v \tag{95}$$

La solution est obtenue en considérant v le premier vecteur propre.

Commentaire: Le nombre de vecteurs propres associé à des valeurs propres non nulles est égal à $K-1$ (du fait du rang de B)

3.1 Analyse Discriminante Bayesienne

On cherche à construire une règle de prédiction sur l'affectation d'un individu à l'un des groupes. On s'intéresse à estimer

$$p_k(x) = P(y = k|X_1 = x_1, \ldots, X_p = x_p) \tag{96}$$

On effectuera l'observation un groupe le plus probable. On utilise la formule de Bayesienne pour calculer $p_k(x)$ avec l'hypothèse que le $X = (X_1, \ldots, X_p) \sim \mathcal{N}(\mu_k, \Sigma_k)$ sur chaque population et en fonction des probabilis´es a priori $q_k = \mathbb{P}(y = k)$ d'appartenance aux differentes sous-population. La formule de Bayes permet d'obtenir la probabilité a posterori $p_k(x)$ en fonction de la densité $f_k(x)$ et des probabilité q_k..

$$p_k(x) = \mathbb{P}(Y = k|X = x) = \frac{f_h(x)q_h}{\sum_{k=1}^{K} f_h(x)q_k} \tag{97}$$

3.1.1 Cas de matrice de covariances homogène

Ici, on suppose que $X = (X_1, \ldots, X_p) \sim \mathcal{N}(\mu_h, \Sigma)$ sur la sous population $\mathcal{P}_k$ (Photo)

En remplacant la moyenne μ_h par leur estimation,

$$\hat{\mu}_h = m_h \tag{98}$$

$$\hat{q}_k = \frac{n_k}{n} \tag{99}$$

Et la matrice de covariance par son estimation

$$S = \frac{1}{n-k}\sum_{k=1}^{K} n_k V_k \tag{100}$$

On obtient des fonctions discriminantes linéaire $d_h(x)$ suivante

$$d_h(x) = \frac{1}{2}m_k^T S^{-1} m_k - h^T S^{-1} x + \ln(\hat{q}_h) \tag{101}$$

on en déduit une estimation des probabilités a posteriori

$$\hat{p}_h(x) = \frac{exp(d_h(x))}{\sum_{k=1}^{K} exp(d_k(x))} \tag{102}$$

Un nouvel individu sur lequel a été observé les valeurs $x = (x_1, \ldots, x_p)$ est affecté au groupe h pour lequel $d_h(x)$ est maximum.

3.1.2 Cas où les covariances sont hétérogène

On suppose maintenant que $X = (X_1, \ldots, X_p)$ est une loi multinomial $\mathcal{N}(\mu_h, \Sigma_h)$ sur chaque sous population $\mathcal{P}_h$. Avec un raisonnement analogue et enconsidérant

$$\hat{\mu} \tag{103}$$

On obtient des fonctions discriminantes quadratiques definie par

$$d_h(x) = -\frac{1}{2}(x - m_k)^T S_h^{-1}(x - m_h) \tag{104}$$

(Photo)

Pour chacun entre la version linéaire ou quadratique, on procède par cross validation.