Mesures de similarités, dissimilarités et distances

RICM 5 - SID - A. DOUZAL

I. Les étapes préliminaires à une analyse de don

1) Analyse des données

Un ensemble de méthodes dont l'objectif essentiel est la mise en relief des relations existantes entre les objets, entre les paramètres qui les caractérisent et entre les objets et les paramètres"

Objet ⇔ item ⇔ sujet ⇔ Forme ⇔ individu ...
paramètre ⇔ variable ⇔ descripteur ..

2) Modélisation des individus / variables.

Ω : ensemble des individus observés

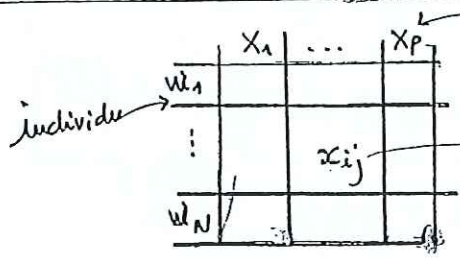
$\mathcal{O}$: espace des observations.

S : une structure algébrique définie sur $\mathcal{O}$

Une variable v : $\Omega \longrightarrow \mathcal{O}$ muni de la structure S.

Structure S \ Cardinal $\mathcal{O}$	Continu	Fini ou Dénombrable	
= #		CSP	Nominale
≤	Âge, Température	rang, ressemblance	Ordinale
≤ + ×	Revenu		Mesurable
	Quantitative	Qualitative ←	Variable

II Type de Tableaux de données à analyser.



→ observation de la variable X_j pour l'individu i

a) Tableaux quantitatifs

$\forall j \in \{1..p\}$ X_j est une variable quantitative.

b) Tableaux de contingence (fréquence)

$\forall i \in \{1..N\}, j \in \{1..p\}$ x_{ij} est le nb d'occurences ou la fréquence d'apparition de la variable X_j pour l'ind W_i.

exp : W_i : un utilisateur identifié par son @IP
X_j : un Thème spécifique d'un site
x_{ij} : la fréquence d'accès de l'utilisateur W_i au thème X_j du site

c) Tableaux de préférence

$j \in \{1..p\}$ X_j : les modalités d'une variable qualitative
x_{ij} : une note exprimant un ordre de préfér de la modalité X_j par l'individu W_i.

d) Tableaux binaires.

Origine : données d'enquêtes
X_j : une modalité d'1 variable qualitative.
x_{ij} : la valeur 0, ou 1 exprimant la prés ou non de la modalité X_j chez l'ind

III

e) Tableaux de proximité (individu × individu)

	w_1	...	w_N
w_1			
⋮		d_{ij}	
w_N			

→ mesure la proximité (similarité, Dissimilarité, distance, ...) entre w_i et w_j

f) Tableaux Hétérogène

Tableaux combinant des variables de # types.

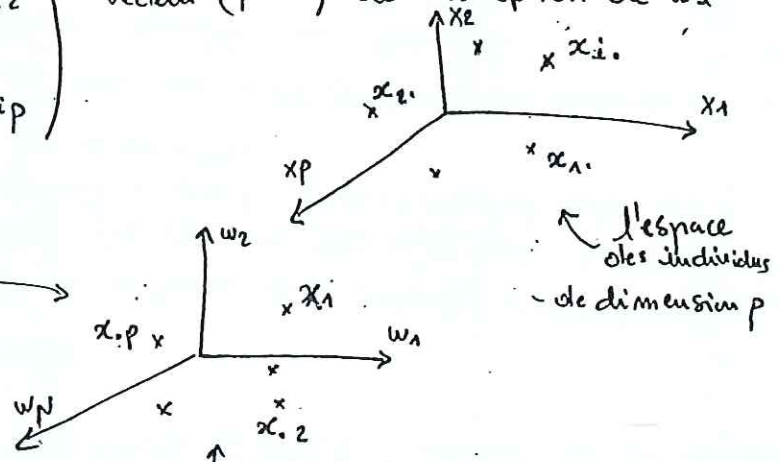
3) Résumés numériques et espaces associés

3.1 Notations :

$$X_{(N\times p)} = \begin{array}{c} \\ 1 \\ \vdots \\ N \end{array}\begin{array}{c} \begin{array}{ccc} 1 & \dots & p \end{array} \\ \begin{bmatrix} & x_{ij} & \end{bmatrix} \end{array}$$

$x_{i\cdot} = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$ vecteur $(p \times 1)$ de description de w_i

$x_{\cdot j} = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{Nj} \end{pmatrix}_{(N\times 1)}$



↖ l'espace des individus - de dimension p

↑ l'espace des variables de

IV

3.2 La matrice des poids

$$D_{(N\times N)} = \begin{pmatrix} p_1 & & 0 \\ & \ddots & \\ 0 & & p_N \end{pmatrix}$$

avec $\sum_{i=1}^{N} p_i = 1$

$p_i \geq 0$

p_i : poids associé à l'individu w_i

En général, on considère $D = \frac{1}{N} I$ (I : matrice identi

3.3 Centre de gravité g

$$g = 1'DX = (\bar{x}_{\cdot 1}, \dots, \bar{x}_{\cdot p})$$

$$1_{(N\times 1)} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}_{(p\times 1)(1\times p)}, \text{ où } \boxed{\bar{x}_{\cdot j} = \sum_{i=1}^{N} p_i x_{ij}}$$

3.4 Centrage et réduction des données

On note la matrice $Y_{(N\times p)}(y_{ij})$ / $\boxed{y_{ij} = x_{ij} - \bar{x}_{\cdot j}}$

la matrice centrée associée à X ($Y = X - 1g'$)

3.5 Matrice de variance - Covariance / Corrélation

$$V_{(p\times p)} = X'DX - g'g = Y'DY$$

$$\boxed{v_{jj'} = \sum_{i=1}^{N} p_i (x_{ij} - \bar{x}_{\cdot j})(x_{ij'} - \bar{x}_{\cdot j'})}$$

↖ covariance entre X_j et $X_{j'}$

On note $s_j = \sqrt{v_{jj}}$: écart-type associé à X_j

$$D_{1/s} = \begin{pmatrix} 1/s_1 & & 0 \\ & \ddots & \\ 0 & & 1/s_p \end{pmatrix} \quad (p \times p) \qquad D_{1/s^2} = \begin{pmatrix} 1/v_{11} & & 0 \\ & \ddots & \\ 0 & & 1/v_{pp} \end{pmatrix} \quad (p \times p)$$

La matrice $\underset{(N \times p)}{Z}$ centrée & réduite associée à X est :

$$\underset{(N \times p)}{Z} = Y D_{1/s} \qquad \boxed{z_{ij} = \frac{x_{ij} - \overline{x_{.j}}}{s_j}}$$

On note $\underset{(p \times p)}{R}$ la matrice des coefficients de corrélation linéaire entre les p variables :

$$\underset{(p \times p)}{R} = D_{1/s} V D_{1/s} = Z' D Z \quad , \quad \boxed{r_{jj'} = \sum_{i=1}^{N} p_i z_{ij} z_{ij'}}$$

4 Mesures de Similarité / Dissimilarité / Distance

4.1 Indice de similarité

est une application qui vérifie les trois propriétés :

1) s est une application $\Omega \times \Omega \longrightarrow \mathbb{R}^+$
2) s est symétrique : $\forall (w_i, w_j) \in \Omega \times \Omega \quad s(w_i, w_j) = s(w_j, w_i)$
3) $\forall (w_i, w_j) \in \Omega \times \Omega$ avec $w_i \neq w_j$
$$s(w_i, w_i) = s(w_j, w_j) > s(w_i, w_j)$$

4.2 Indice de dissimilarité

est une application qui satisfait aux conditions 1) et 2) d'1 indice de similarité et à :

$$s'(w_i, w_i) = 0 \quad \forall w_i \in \Omega.$$

4.3 Une distance

une distance est un indice de dissimilarité qui vér en plus les deux propriétés suivantes :

1) $s'(w_i, w_j) = 0 \iff w_i = w_j$
2) $s'(w_i, w_j) \leq s'(w_i, w_k) + s'(w_k, w_j) \quad \forall w_i, \in$
(Inégalité triangulaire)

Remarque : un indice de dissimilarité vérifiant uniq la propriété 1) est un indice de distanc Unique + la propriété 2) est un "écart".

4.4 Une ultramétrique

un indice de dissimilarité qui vérifie en plus l deux propriétés suivantes :

1) $s'(w_i, w_j) = 0 \iff w_i = w_j$
2) $s'(w_i, w_j) \leq \max\left(s'(w_i, w_k), s'(w_k, w_j)\right)$
$\forall (w_i, w_k, w_j) \in \Omega^3$

5 Mesures de proximité. Standardisation entre individus

5.1 Cas de variables quantitatives.

$x_{i.}$ $(p\times 1)$ $= \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$ donnant la description de w_i

5.1.1 distance de Minkowski

$$d(w_i, w_{i'}) = \left(\sum_{j=1}^{p} | x_{ij} - x_{i'j} |^r \right)^{1/r} \qquad r \geq 1$$

$r = 1$ (distance Manhattan, Taxicab, City block)

$$d(w_i, w_{i'}) = \sum_{j=1}^{p} | x_{ij} - x_{i'j} |$$

$r = 2$ (distance euclidienne)

$$d(w_i, w_{i'}) = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{i'j})^2}$$

$r \to \infty$ $\quad d(w_i, w_{i'}) = \max_{1 \leq j \leq p} (| x_{ij} - x_{i'j} |)$ (Chebychev)

5.1.2 Distance de Mahalanobis

distance souvent utilisée en analyse discriminante

$$d^2(w_i, w_{i'}) = (x_{i.} - x_{i'.})' V^{-1} (x_{i.} - x_{i'.})$$

5.1.3 Distance du χ^2 (principe d'équivalence distributionnel)

souvent utilisée sur des Tableaux de fréquence (Analyse factorielle des correspondances)

$$d^2(w_i, w_{i'}) = \sum_{j=1}^{p} \frac{1}{\dot{x}_j} \left(\frac{x_{ij}}{\dot{x}_i} - \frac{x_{i'j}}{\dot{x}_{i'}} \right)^2$$

$\dot{x}_j = \sum_{i=1}^{n} x_{ij}$ $\qquad$ $\dot{x}_i = \sum_{j=1}^{p} x_{ij}$

x_{ij} : mb d'occurence de la modalité x_j pour w_i.

5.2 Cas de variables binaires

	X_1	...	X_p
w_i	x_{i1}	...	x_{ip}
$w_{i'}$	$x_{i'1}$	...	$x_{i'p}$

$x_{ij} = 0 / 1$

w_i \ $w_{i'}$	1	0
1	a	b
0	c	d

a : mb de fois où $x_{ij} = x_{i'j} = 1$
b : " " $x_{ij} = 0$ et $x_{i'j}$
c : " " $x_{ij} = 1$ et $x_{i'j}$
d : " " $x_{ij} = 0 = x_{i'j}$

Mesures de ressemblances usuelles

Russel et Rao	$D_1(w_i, w_{i'}) = \frac{a}{a+b+c+d}$	$p = a+b+c+d$
Jaccard & Needham	$D_2(w_i, w_{i'}) = \frac{a}{a+b+c}$	
Dice	$D_3(w_i, w_{i'}) = \frac{2a}{2a+b+c}$	
Sokal & Sneath	$D_4(w_i, w_{i'}) = \frac{a}{a+2(b+c)}$	
Sokal & Michener	$D_5(w_i, w_{i'}) = \frac{a+d}{a+b+c+d}$	
Kulzinsky	$D_6(w_i, w_{i'}) = \frac{a}{b+c}$	
Roger & Tanimoto	$D_7(w_i, w_{i'}) = \frac{a+d}{a+d+2(b+c)}$	
Yule	$D_8(w_i, w_{i'}) = \frac{ad-bc}{ad+bc}$	

D_5, D_7, (D_8+1), (D_9+1) sont des mesures de similarités.

5.3 Cas de variables qualitatives nominales

On transforme les variables qualitatives en variables binaires par codage disjonctif complet, puis appliquer les mesures de ressemblances vue en 5.2

exp de codage disjonctif complet

	Couleur	Forme
w_i	rouge	Ellipsoïde
$w_{i'}$	Jaune	circulaire

couleur $\in$ {rouge, jaune, bleue}

Forme $\in$ {Ellipsoïde, circulaire, }

⇓ codage

	Rouge	jaune	bleue	Ellipsoïde	circulaire
w_i	1	0	0	1	0
$w_{i'}$	0	1	0	0	1

5.4 Cas de variables qualitatives ordinales.

Idem cas des variables qualitatives nominales.

6) Mesures de proximité entre variables

6.1 Cas de deux variables quantitatives

$$r_{jj'} = d^2(x_j, x_{j'}) = \sum_{i=1}^{N} p_i \left(\frac{x_{ij} - \bar{x}_{\cdot j}}{s_j}\right)\left(\frac{x_{ij'} - \bar{x}_{\cdot j'}}{s_{j'}}\right)$$

↑ coefficient de corrélation linéaire entre x_j $x_{j'}$

6.2 Cas de deux variables qualitatives nominales

Soit X_1, X_2 deux variables nominales à q, r modalités

$$\begin{array}{c|cccc} X_1 \backslash & 1 & 2 & \cdots & r \\ 1 & & & & \\ 2 & & & & \\ \vdots & & & & \\ q & & & n_{ij} & \end{array}$$

n_{ij} = nb d'individus ayant la modalité i de X_1 et j de X_2

$n_{\cdot j} = \sum_{i=1}^{q} n_{ij}$ $\quad$ $n_{i\cdot} = \sum_{j=1}^{r} n_{ij}$

6.3.2 Coefficients

On associe à chaque variable ordinale X_i une nouvelle var

$$r_i : \mathcal{R} \longrightarrow \{1 \dots n\} \qquad n = \text{Card}(\mathcal{R})$$

En présence d'ex-aequo on utilise le codage par rang Moyen.

exemple:

	x_j	$x_{j'}$
w_1	a	f
w_2	b	e
w_3	b	g
w_4	d	f

avec $a < b < c$ et $e < f < g$

	x_j	$x_{j'}$
w_1	1	2
w_2	2	1
w_3	3	4
w_4	4	3

→

	x_j	$x_{j'}$
w_1	1	2,5
w_2	2,5	1
w_3	2,5	4
w_4	4	2,5

codage par rang moyen

On note r_{ij} : le rang de w_i pour la variable x_j

$$\overline{r_{\cdot j}} = \frac{1}{n} \sum_{i=1}^{n} r_{ij}$$

le coefficient de corrélation des rangs de Spearman : R_S

$$R_S = \frac{\sum_{i=1}^{n} \left(r_{ij} - \bar{r}_j\right)\left(r_{ij'} - \bar{r}_{j'}\right)}{\sqrt{\sum_{i=1}^{n} \left(r_{ij} - \bar{r}_j\right)^2 \sum_{i=1}^{n} \left(r_{ij'} - \bar{r}_{j'}\right)^2}}$$

La mesures standard entre deux variables nominales est la mesure du χ^2,

$$\chi^2 = D_1(x_1, x_2) = n_{..} \sum_{i=1}^{q} \sum_{j=1}^{q} \left(n_{ij} - \frac{n_{i.} n_{.j}}{n_{..}} \right)^2 \times \frac{1}{n_{i.}}$$

6.3 Cas de deux variables qualitatives ordinales.

La distance entre deux variables qualitatives ordinales revient à mesurer la distance entre ordres. 3 trois mesures :

- les coefficients de corrélation des rangs de Kendall
- " " " de Spearman
- " " " de Guttman

6.3.1 Coefficients de corrélation des rangs de Kendall

Soit x_1, x_2 deux variables ordinales. On associe à chaque variable une nouvelle variable y_j définie comme suit

$$y_j : \mathcal{R} \times \mathcal{R} \longrightarrow \{-1, 0, 1\}$$

$$\begin{cases} y_j(w_i, w_{i'}) = -1 & \text{si } x_{i1} < x_{i2} \\ y_j(w_i, w_{i'}) = 0 & \text{si } x_{i1} = x_{i2} \\ y_j(w_i, w_{i'}) = +1 & \text{si } x_{i1} > x_{i2} \end{cases}$$

le rang de Kendall (Taux de Kendall) τ_{x_1, x_2} :

$$\boxed{\tau_{x_1, x_2} = \text{cor}(y_1, y_2)}$$

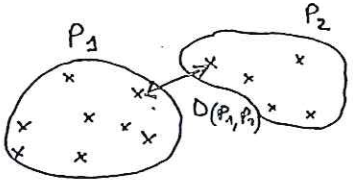
Remarque y_j porte sur $\frac{n(n-1)}{2}$ couples d'individus.

7) Mesures de proximité entre groupes

Soit P_1, P_2 deux populations d'individus de $\mathcal{R}$.
Il existe trois mesures classiques de ressemblance entre P_1, P_2

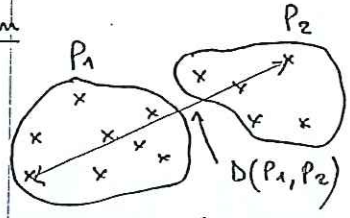
7.1 Distance du lien minimum

$$D(P_1, P_2) = \min_{\substack{w_i \in P_1 \\ w_{i'} \in P_2}} d(w_i, w_{i'})$$



7.2 Distance du lien maximum

$$D(P_1, P_2) = \max_{\substack{w_i \in P_1 \\ w_{i'} \in P_2}} d(w_i, w_{i'})$$



7.3 Distance des noyaux

Un sous ensemble N_j de P_j constitue un noyau de P_j à $n_i = \text{Card}(N_j)$ individus ssi il minimise l'expression :

$$\sum_{w \in N_j} \sum_{w_i \in P_j} d(w, w_i)$$

La distance des noyaux de P_1, P_2 :

$$D(P_1, P_2) = \sum_{\substack{w_i \in N_1 \\ w_{i'} \in N_2}} d(w_i, w_{i'})$$

Cas particuliers : - $\text{Card}(N_1) = \text{Card}(N_2)$
- espace d'observation est 1 espace vectoriel muni d'une distance quadratique

$D(P_1, P_2) = d(G_1, G_2)$ G_i : centre de gravité de

On se situe dans le cas où l'espace d'observation est 1 espace vectoriel muni d'une distance quadratique.

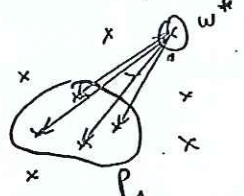
le barycentre de P_j : $G(P_j) = \frac{1}{p(P_j)} \sum_{w_i \in P_j} p(w_i)\, x_i$ ← vecteur de description de w_i

$p(w_i)$: poids de l'individu w_i

$$p(P_j) = \sum_{w_i \in P_j} p(w_i)$$

l'inertie de P_j par rapport à w^* de Ω est :

$$I_{w^*}(P_j) = \sum_{w_i \in P_j} p(w_i)\, d(w^*, w_i)$$



Cas particulier où $w^* = G(P_j)$, l'inertie de P_j est alors :

$$I(P_j) = \sum_{w_i \in P_j} p(w_i)\, d(G(P_j), w_i)$$

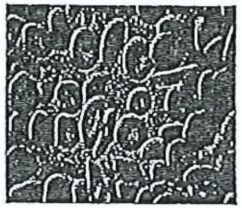
d'après le <u>Théorème de Huygens</u>(1), l'inertie $I_{w^*}(P_j)$ est minimale pour $w^* = G(P_j)$.

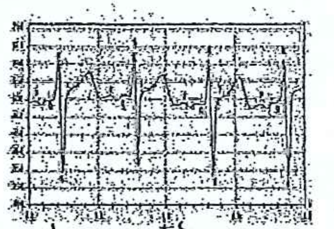
(1) $$\boxed{I_{w^*}(P_j) = I(P_j) + p(P_j)\, d(G(P_j), w^*)}$$

La mesure de proximité fondée sur l'inertie proposée par Ward (1963)

$$d(P_1, P_2) = I(P_1 \cup P_2) - I(P_1) - I(P_2) = \boxed{\frac{p(P_1)\, p(P_2)}{p(P_1) + p(P_2)}\, d(G(P_1), G(P_2))}$$

1) <u>Type de données à analyser</u>

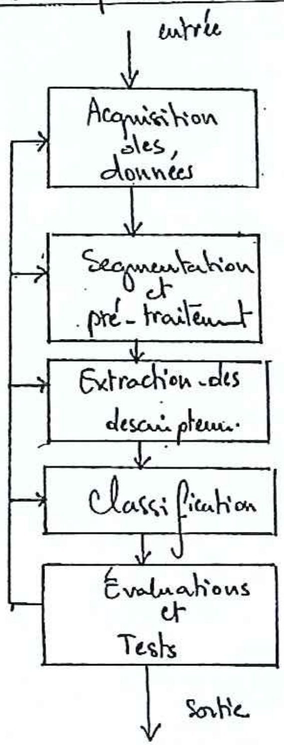




Déclaration par les parents	N	n	%	95 % IC
Toux nocturne	3 535	846	24.0	[22,8-25,2]
Toux provoquée par le froid	3 516	532	15.1	[14,1-16,1]
Toux à l'effort	3 533	372	10.6	[9,7-11,5]
Otite nocturne	3 533	475	12.1	[11,2-13,0]
Sifflements dans la poitrine	3 530	360	10.2	[9,3-11,1]
Sifflements au cours des 12 derniers mois	3 535	202	5.7	[5,0-6,4]
Bronchite sifflante	3 525	473	13.4	[12,4-14,4]
Bronchite sifflante au cours des 12 derniers mois	3 535	180	5.1	[4,5-5,7]
Asthme diagnostiqué déclaré	3 511	213	6.1	[5,4-6,8]

N : (effectif des [illegible]) varie en fonction des [illegible]

2) <u>Les composantes d'un système de reconnaissance de Forme.</u>



- choix de la technique de segmentation
- choix du modèle de pré-traitement des données
- choix des descripteurs
- choix du modèle de classification
- conception du classifieur par apprentissage
- Évaluation du classifieur.

2.1 <u>Acquisition des données</u>

- Constitution d'1 échantillon d'apprentissage et d'1 échantillon Test exhaustifs

- Segmentation
  - Extraire l'objet à analyser de son environnement (isoler les cellules, extraire les sous-segments QRS, ...)

- Pré-traitement
  - Élimination du bruit dû aux instruments de collecte (bruitage d'1 image, translation d'1 signal, déviation d'1 capteur)

2.3 Sélection des descripteurs caractérisant les objets à analyser

L'objectif est de choisir des descripteurs dont les mesures sont proches pour des objets d'1 même classe et loins pour des objets de ≠ classes

critères de choix:

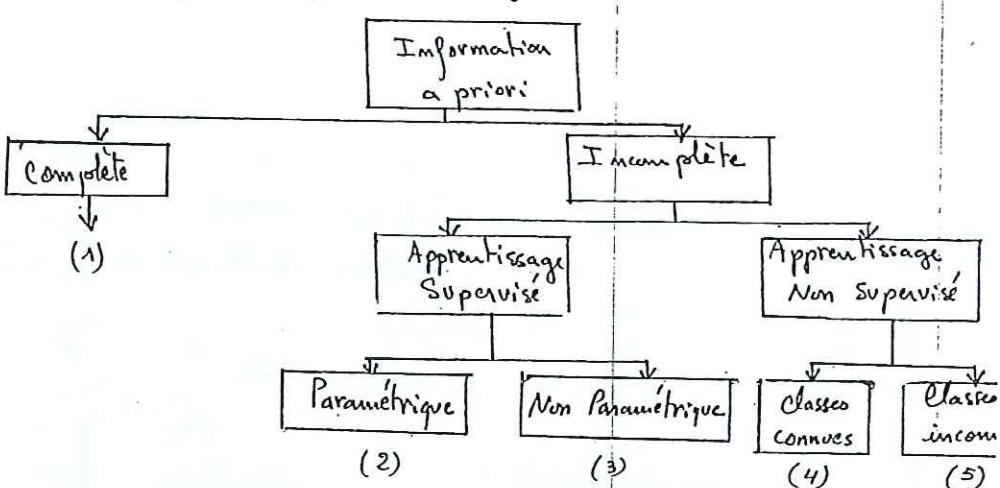
- Invariance en translation / Rotation
  (Exp: - position absolue d'1 cellule est non significative
  - l'instant de début d'1 cycle ...)
- Invariance par rapport à l'échelle
  (Exp: - Échelle d'acquisition de l'image)
- Invariance aux distortions des projections
  (Exp: - distance entre l'objectif et la caméra change)

...

2.4 Choix du modèle de classification.

- un ensemble (échantillon d'apprentissage) d'objets décrits dans un espace à K dimensions (K descripteurs)
- l'objectif consiste à déterminer les frontières séparatrices des différentes classes.

3) Les différentes approches classificatoires en RF Statistique



	Connu a priori	Inconnu	à Estimer	Méthodes Standards
(1)	$C(x)$: classe de x $P(C_j)$: proba de C_j $P(x/C_j)$	$P(C_j/x)$	$P(C_j/x)$	• Théorie de la décision bayésienne
(2)	$C(x)$ $P(C_j)$ $P(x/C_j) \sim L(\theta)$ $L(\theta)$: loi de la densité de proba de $P(x/C_j)$	θ $P(C_j/x)$	θ $P(C_j/x)$	• Maximum de Vraisemblance • Estimation Bayésienne
(3)	$C(x)$ $P(C_j)$	• fdp de $P(x/C_j)$ • $P(C_j/x)$ • $P(x/C_j)$	$P(C_j/x)$	• Estimation fdp - fenêtre de Parzen - KPPV
(4)	Nb classe $P(C_j)$ $P(x/C_j) \sim L(\theta)$	$C(x)$ θ $P(C_j/x)$	• θ $P(C_j/x)$	Modèles de densité mixtes
(5)		• Nb classe $P(C_j)$, $C(x)$ $P(x/C_j)$	Extraire les classe C_j $L(x)$	• classification - Hiérarchique - Partitionnement K-MEANS