

# Classification des données de haute dimension : applications en vision par ordinateur

Charles Bouveyron

*LMC-IMAG & INRIA Rhône-Alpes*

*Travail conjoint avec S. Girard et C. Schmid*

*Journées de l'ACI Masse de Données – Bordeaux – Novembre 2005*

- Les données modernes sont généralement de grande dimension :
  - Imagerie : vision par ordinateur, OCR, IRM, ...
  - Biologie : puces ADN, ...
  - Fouille de données : données textuelles, ...
- La classification en grande dimension est un problème difficile :
  - la plupart des méthodes souffrent du *fléau de la dimension*,
  - les groupes sont souvent "cachés" dans des sous-espaces de faible dimension (phénomène de *l'espace vide*).

- Classification :
  - classification supervisée (analyse discriminante),
  - classification non-supervisée (clustering).
- Le problème de la classification supervisée :
  - affecter une observation  $x \in \mathbb{R}^p$  à une classe parmi  $k$  classes  $C_1, \dots, C_k$  connues *a priori*,
  - en se basant sur un jeu d'apprentissage  $\mathcal{A} = \{(x_1, c_1), \dots, (x_n, c_n)\}$ .
- Le problème de la classification non-supervisée :
  - organiser les données  $\{x_1, \dots, x_n\} \in \mathbb{R}^p$  en  $k$  groupes homogènes.

- 1 Un modèle gaussien pour les données de grande dimension
- 2 Classification des données de haute dimension (HDDA, HDDC)
- 3 Estimation des paramètres
- 4 Application à la reconnaissance optique de caractères
- 5 Application à la reconnaissance d'objets
- 6 Conclusion

- Le problème :
  - fléau de la dimension : la taille du jeu d'apprentissage est trop faible comparée à la dimension de l'espace,
  - phénomène de l'espace vide : les groupes sont "cachés" dans des sous-espaces de faible dimension.
- Les solutions existantes :
  - réduction de dimension (ACP, sélection de variables, ...),
  - utilisation de modèles gaussiens parcimonieux ou SVM.
- L'idée principale du nouveau modèle :
  - les données d'une même classe vivent dans un sous-espace de faible dimension,
  - les données de différentes classes peuvent vivre dans différents sous-espaces.

- Nous supposons que les densités conditionnelles des classes sont normales  $\mathcal{N}(\mu_i, \Sigma_i)$ ,
- On définit  $Q_i$  comme étant la matrice orthogonale composée des vecteurs propres de  $\Sigma_i$ ,
- Alors, la matrice de covariance  $\Delta_i$  de la classe  $C_i$  est définie dans l'espace propre de  $\Sigma_i$  par :

$$\Delta_i = Q_i^t \Sigma_i Q_i.$$

- Nous supposons de plus que  $\Delta_i$  est composée de 2 blocs :

$$\Delta_i = \left( \begin{array}{ccc|ccc} \boxed{\begin{array}{ccc} a_{i1} & & 0 \\ & \ddots & \\ 0 & & a_{id_i} \end{array}} & & & \mathbf{0} & & \\ & & & & & \\ & & & & & \\ \mathbf{0} & & & \boxed{\begin{array}{ccc} b_i & & 0 \\ & \ddots & \\ 0 & & b_i \end{array}} & & \\ & & & & & \end{array} \right) \left. \begin{array}{l} \vphantom{\Delta_i} \\ \vphantom{\Delta_i} \\ \vphantom{\Delta_i} \end{array} \right\} d_i$$
  
$$\left. \begin{array}{l} \vphantom{\Delta_i} \\ \vphantom{\Delta_i} \end{array} \right\} (p - d_i)$$

avec  $a_i > b_i, \forall j = 1, \dots, d_i$ .

- Soit  $\mathbb{E}_i$  l'espace affine engendré par les vecteurs propres associés aux  $d_i$  premières valeurs propres  $a_{ij}$  et tel que  $\mu_i \in \mathbb{E}_i$ ,
- On définit également  $\mathbb{E}_i^\perp$  l'espace complémentaire de  $\mathbb{E}_i$ ,  $P_i$  et  $P_i^\perp$  les opérateurs de projection sur  $\mathbb{E}_i$  et  $\mathbb{E}_i^\perp$ .
  
- Nous noterons  $[a_{ij}b_iQ_id_i]$  ce modèle.
  
- Nous pouvons aussi contraindre le modèle  $[a_{ij}b_iQ_id_i]$  :
  - si  $\forall j = 1, \dots, d_i, a_{ij} = a_i$ , alors on obtient le sous-modèle  $[a_ib_iQ_id_i]$ ,
  - si de plus  $\forall i = 1, \dots, k, b_i = b$ , alors on obtient le sous-modèle  $[a_ibQ_id_i]$ .

- La règle de décision optimale  $\delta^*$ , dite règle de Bayes :

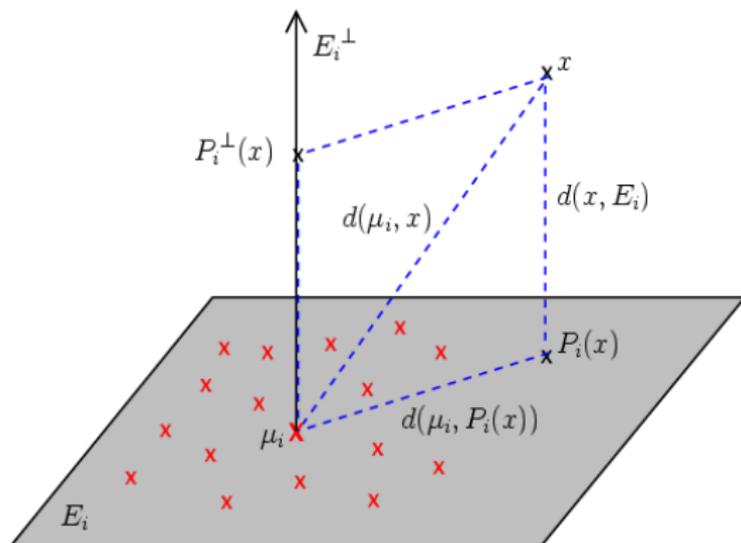
$$\delta^* : x \in C_{i^*}, \text{ si } i^* = \underset{i=1, \dots, k}{\operatorname{argmin}} \{-2 \log(\pi_i f_i(x))\}$$

- Pour le modèle  $[a_{ij} b_i Q_i d_i]$ , la règle de Bayes consiste à affecter  $x$  à la classe  $C_{i^*}$  minimisant la fonction de coût  $K_i(x)$  :

$$\begin{aligned} K_i(x) = & \|\mu_i - P_i(x)\|_{\Lambda_i}^2 + \frac{1}{b_i} \|x - P_i(x)\|^2 \\ & + \sum_{j=1}^{d_i} \log(a_{ij}) + (p - d_i) \log(b_i) - 2 \log(\pi_i). \end{aligned}$$

où  $\|\cdot\|_{\Lambda_i}^2$  est la distance de Mahalanobis associée à la matrice  $\Lambda_i = \tilde{Q}_i \Delta_i \tilde{Q}_i^t$ .

# Analyse discriminante de haute dimension



Dans le cas du modèle contraint  $[a_i b_i Q_i d_i]$  :

$$K_i(x) = \frac{1}{a_i} \|\mu_i - P_i(x)\|^2 + \frac{1}{b_i} \|x - P_i(x)\|^2 + d_i \log(a_i) + (p - d_i) \log(b_i) - 2 \log(\pi_i)$$

- Le modèle  $[a_{ij}b_iQ_id_i]$  étant génératif, il peut être également utilisé en classification non-supervisée.
- Les formules de ré-actualisation de EM deviennent :
  - Étape E :

$$t_{ij}^{(q)} = 1 / \sum_{l=1}^k \left[ \frac{1}{2} \exp \left( K_i^{(q-1)}(x_j) - K_l^{(q-1)}(x_j) \right) \right].$$

- Étape M : estimation classique des  $\pi_i^{(q)}$ ,  $\mu_i^{(q)}$  et  $\Sigma_i^{(q)}$  ; les paramètres  $a_i$ ,  $b_i$  et  $Q_i$  s'estiment à l'étape  $q$  de la même façon qu'en analyse discriminante.

- Les paramètres  $\pi_i$ ,  $\mu_i$  et  $\Sigma_i$  sont estimés classiquement,
- En supposant que les  $d_i$  sont connus,
  - les  $d_i$  premières colonnes de  $\hat{Q}_i$  sont les vecteurs propres associés aux  $d_i$  plus grandes valeurs propres de  $\hat{\Sigma}_i$ ,
  - $\hat{a}_{ij}$  est la  $j$ ème plus grande valeur propre de  $\hat{\Sigma}_i$  :

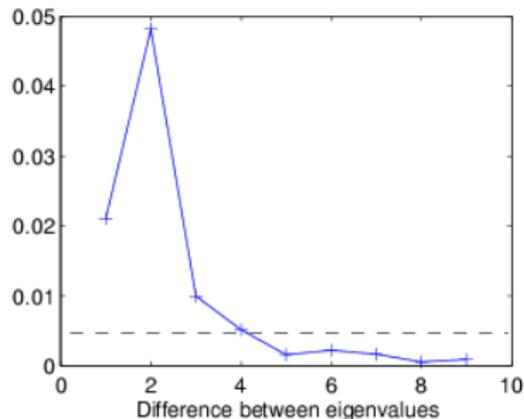
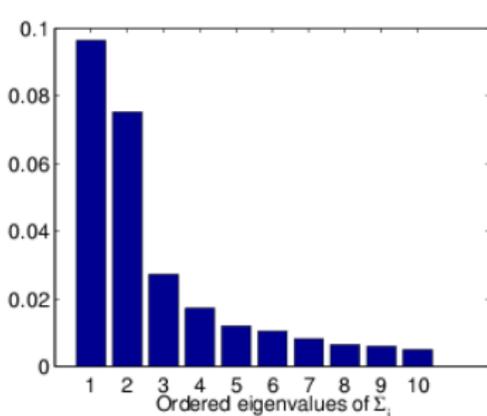
$$\hat{a}_{ij} = \lambda_{ij},$$

- $\hat{b}_i$  est la moyenne des  $(p - d_i)$  dernières valeurs propres de  $\hat{\Sigma}_i$  :

$$\begin{aligned}\hat{b}_i &= \frac{1}{(p - d_i)} \sum_{l=d_i+1}^p \lambda_{il}, \\ &= \frac{1}{(p - d_i)} \left[ \text{tr}(\hat{\Sigma}_i) - \sum_{l=1}^{d_i} \lambda_{il} \right].\end{aligned}$$

# Estimation des dimensions intrinsèques

- Notre approche est basée sur les valeurs propres de la matrice de covariance  $\Sigma_i$ ,
- Nous utilisons le *scree-test* de Cattell [2] :



Le *scree-test* de Cattell.

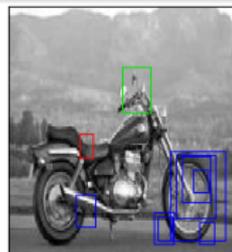


Exemples de caractères de la base USPS.

- Base USPS, fournie par la poste américaine, contient :
  - 2007 images pour l'apprentissage,
  - 7291 images pour le test.
- Résultats numériques :

<i>Méthode</i>	<i>Taux de reconnaissance</i>
HDDA [ $a_i b_i Q_i d_i$ ]	0.955
HDDA [ $a_i b Q_i d_i$ ]	<b>0.959</b>
LDA ( $d = 256$ )	0.746
FDA ( $d = 9$ )	0.902
SVM (linear)	0.943
Humain	<b>0.975</b>

# La reconnaissance d'objets

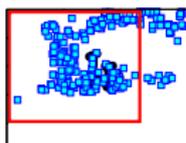


- Détection des points d'intérêts :
  - on utilise l'opérateur d'Harris-Laplace,
  - qui détecte les points saillants de l'image.
- Description des points d'intérêts :
  - on utilise l'opérateur Sift,
  - qui décrit chaque point d'intérêt dans un espace de grande dimension ( $d = 128$ ).
- Ainsi, chaque image est représentée par  $m$  descripteurs en dimension 128.

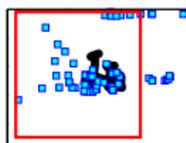
- Nous avons évalué notre méthode HDDC sur la base "Pascal test2" [3] :
  - qui contient 4 catégories d'images,
  - 684 images d'apprentissage et 956 images de test.
- Nous considérons deux types d'images pour l'apprentissage :
  - positive : l'objet considéré est présent dans l'image,
  - négative : l'objet considéré n'est pas présent dans l'image.
- Apprentissage des classes discriminantes de l'objet :
  - l'HDDC organise les données en  $k$  classes homogènes,
  - calcul du pouvoir discriminant de chacune des  $k$  classes :

$$R_i = P(C_i \in O|C_i).$$

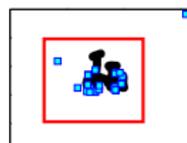
# Reconnaissance : localisation d'objets



All detections



$P(x_j \in O) > 0.5$



$P(x_j \in O) > 0.65$



$P(x_j \in O) > 0.7$

- Localisation d'un objet :

- l'HDDA fournit  $P(x \in C_i|x)$  pour les  $k$  classes,
- on calcule la probabilité que le descripteur  $x$  appartienne à l'objet :

$$P(x \in O|x) = \sum_{i=1}^k R_i P(x \in C_i|x).$$

# Reconnaissance : classification d'images



Vélos



Personnes

- Classification d'images :

- on veut décider si une image  $I$  contient ou non un objet donné,
- on calcule le score  $S$  de l'image :

$$S = \frac{1}{m} \sum_{x_j \in I} \sum_{i=1}^k R_i P(x_j \in C_i | x_j),$$

où  $m = \text{Card}(x_j \in I)$ .

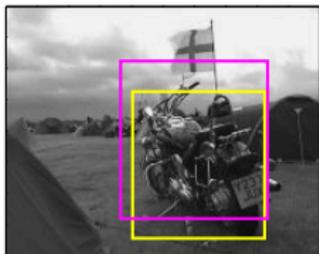
# Résultats : localisation d'objets

Méthode	Moto	Vélo	Humain	Voiture	Moy.
$[a_{ij}b_iQ_id_i]$	0.305	0.169	0.061	0.154	0.172
$[a_{ij}bQ_id_i]$	0.316	0.165	0.091	0.151	0.181
$[a_ib_iQ_id_i]$	0.315	<b>0.172</b>	<b>0.091</b>	0.155	<b>0.183</b>
$[a_ibQ_id_i]$	0.307	0.169	0.091	0.136	0.175
$[abQ_id_i]$	0.316	0.168	0.046	0.158	0.172
Pascal 2005	<b>0.341</b>	0.113	0.021	<b>0.304</b>	<i>0.112</i>

Résultats de localisation supervisée pour les 4 catégories d'objets du challenge Pascal [3].

# Résultats : localisation d'objets

- Résultats de localisation : 1 objet par image



Moto

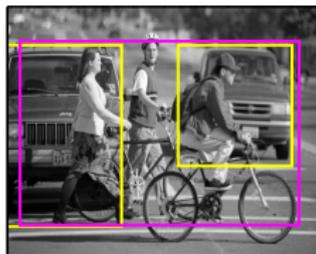


Voiture

- Résultats de localisation : plusieurs objets par image



Humain



Voiture

# Résultats : classification d'images

Méthode	Moto	Vélo	Humain	Voiture	Moy.
$[a_{ij}b_iQ_id_i]$	0.738	0.713	0.707	0.669	0.707
$[a_{ij}bQ_id_i]$	0.733	0.713	<b>0.728</b>	0.673	0.712
$[a_ib_iQ_id_i]$	0.743	0.717	0.715	0.662	0.709
$[a_ibQ_id_i]$	0.738	0.710	0.707	0.666	0.705
$[abQ_id_i]$	0.733	<b>0.728</b>	0.711	0.684	0.714
Pascal 2005	<b>0.798</b>	0.728	0.719	<b>0.720</b>	<b>0.741</b>

Résultats de classification d'images pour les 4 catégories d'objets du challenge Pascal [3].

- Les principaux avantages du modèle  $[a_{ij}b_iQ_id_i]$  sont :
  - bonnes performances sans réduction préalable de la dimension des données,
  - bonnes performances avec des jeux d'apprentissage de petites tailles,
  - flexible car il s'adapte aux différents types de données,
  - aussi rapide que les autres méthodes génératives,
  - peut être utilisé aussi bien en classification supervisée que non-supervisée.
- Nous avons appliqué avec succès le modèle  $[a_{ij}b_iQ_id_i]$  :
  - à la reconnaissance de caractères écrits grâce à l'HDDA.
  - à la reconnaissance d'objets dans des images naturelles grâce à l'HDDC.

- Bibliographie



C. Bouveyron, S. Girard, and C. Schmid.  
Analyse Discriminante de haute dimension.  
Rapport de recherche 5470, INRIA, January 2005.



R. B. Catell.  
The scree test for the number of factors.  
*Multivariate Behavioral Research*, 1 :140–161, 1966.



F. d'Alche Buc, I. Dagan, and J. Quinero, editors.  
*The 2005 Pascal visual object classes challenge*.  
Proceedings of the first PASCAL Challenges Workshop. Springer, 2006.



D. Lowe.  
Distinctive image features from scale-invariant keypoints.  
*International Journal of Computer Vision*, 60(2) :91–110, 2004.



K. Mikolajczyk and C. Schmid.  
Scale and affine invariant interest point detectors.  
*International Journal of Computer Vision*, 60(1) :63–86, 2004.



D. Scott and J. Thompson.  
Probability density estimation in higher dimensions.  
In *Proceedings of the Fifteenth Symposium on the Interface*, North Holland-Elsevier  
Science Publishers, pages 173–179, 1983.

- Page web du projet : <http://lear.inrialpes.fr/movistar/>