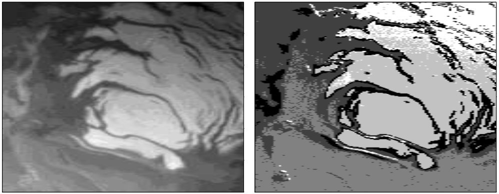
Classification des données de grande dimension : application à l'analyse d'images

J. Blanchet, C. Bouveyron, F. Forbes, S. Girard



LMC-IMAG & INRIA Rhône-Alpes

Grenoble, 31 mai 2006

Plan de l'exposé

1. Classification des données de grande dimension
2. Une famille de modèles gaussiens pour la grande dimension
3. Clustering des données de grande dimension : HDDC
4. Expérimentations et applications à l'analyse d'images

Plan de l'exposé

1. Classification des données de grande dimension
2. Une famille de modèles gaussiens pour la grande dimension
3. Clustering des données de grande dimension : HDDC
4. Expérimentations et applications à l'analyse d'images

Le modèle de mélange gaussien

On suppose classiquement que les données $x_1, ..., x_n$ sont les réalisations indépendantes d'un vecteur aléatoire $X \in \mathbb{R}^p$ dont la fonction de densité s'écrit :

$$f(x) = \sum_{i=1}^{k} \pi_i f_i(x),$$

où k est le nombre de classes, f_i la densité de la $i$ème composante du mélange et les π_i sont les proportions du mélange.

Généralement, les densités f_i sont supposées être celles de lois normales $\mathcal{N}(\mu_i, \Sigma_i)$. Le modèle de mélange devient alors :

$$f(x) = \sum_{i=1}^{k} \pi_i f(x, \theta_i),$$

où $\theta_i = \{\mu_i, \Sigma_i\}$.

La règle de Bayes

Le problème de la classification étant d'affecter une observation x à une classe parmi k classes, il faut donc construire une règle de décision δ :

$$\begin{array}{rcl} \delta \,:\, R^p & \rightarrow & \{1, ..., k\}, \\ x & \rightarrow & z. \end{array}$$

La règle de décision optimale δ^*, qui minimise le risque de mauvaise affectation, consiste à affecter l'observation x à la classe la plus probable *a posteriori* :

$$\delta^*(x) = \underset{i=1,...,k}{\operatorname{argmax}}\, P(Z = i | X = x, \theta).$$

Cette règle porte également le nom de *maximum a posteriori* (MAP).

La règle de Bayes et le modèle de mélange gaussien

Dans le cadre du modèle de mélange gaussien, la formule de Bayes permet d'écrire :

$$P(Z = i|X = x, \theta) = \frac{\pi_i f(x, \theta_i)}{f(x)},$$

et comme le dénominateur $f(x)$ est commun à toutes les classes, la règle du MAP s'écrit :

$$\delta^*(x) = \underset{i=1,\ldots,k}{\operatorname{argmax}} \, \pi_i f(x, \theta_i).$$

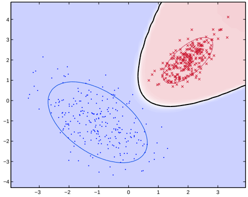
Si de plus on définit la fonction de coût K_i :

$$K_i(x) = -2 \log(\pi_i f(x, \theta_i)),$$

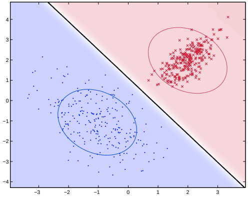
la règle du MAP s'écrit simplement :

$$\delta^*(x) = \underset{i=1,\ldots,k}{\operatorname{argmin}} \, K_i(x).$$

Modèles Full-GMM et Com-GMM



Full-GMM



Com-GMM

Modèle gaussien Full-GMM :

$$K_i(x) = (x - \mu_i)^t \Sigma_i^{-1} (x - \mu_i) + \log(\det \Sigma_i) - 2\log(\pi_i) + C^{te}.$$

- Cependant, ce modèle est pénalisé par l'estimation de nombreux paramètres quand la dimension p devient grande.

Modèle gaussien Com-GMM :

- ce modèle est un modèle plus parcimonieux que Full-GMM,
- qui suppose que $\Sigma_i = \Sigma$, pour tout $i = 1, ..., k$:

$$K_i(x) = \mu_i^t \Sigma^{-1} \mu_i - 2\mu_i^t \Sigma^{-1} x - 2\log(\pi_i) + C^{te}.$$

Fléau de la dimension en classification

Quand le nombre d'observations n devient petit devant la dimension p :

- les estimations des matrices de covariance sont mal conditionnées ou singulières,
- il est alors difficile ou impossible de les inverser,
- et la règle de décision en est d'autant perturbée.

Les solutions classiques pour pallier ces limitations sont :

- réduction de la dimension (ACP, ...),
- régularisation des estimateurs des matrices de covariance,
- utilisation de modèles parcimonieux.

Plan de l'exposé

1. Classification des données de grande dimension
2. Une famille de modèles gaussiens pour la grande dimension
3. Clustering des données de grande dimension : HDDC
4. Expérimentations et applications à l'analyse d'images

Le phénomène de l'espace vide

Le phénomène de "l'espace vide" met en évidence que :

- les espaces de grande dimension sont quasiment vides,
- les données vivent dans des sous-espaces de dimensions intrinsèques inférieures à la dimension de l'espace p.

Il est d'autre part naturel de penser que :

- les données de chaque classe vivent dans des sous-espaces différents,
- dont les dimensions intrinsèques peuvent être différentes.

Nous proposons donc une paramétrisation du mélange gaussien qui :

- prend en compte le fait que les données vivent dans des sous-espaces,
- permet de faire des hypothèses supplémentaires afin de réduire le nombre de paramètres à estimer.

Le modèle $[a_{ij}b_iQ_id_i]$

Nous nous plaçons dans le cadre du modèle de mélange gaussien :

$$f(x) = \sum_{i=1}^{k} \pi_i f(x, \theta_i), \text{ avec } f(x, \theta_i) \sim \mathcal{N}(\mu_i, \Sigma_i).$$

En se basant sur la décomposition spectrale de Σ_i, on peut écrire :

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

où :

- Q_i est la matrice orthogonale composée des vecteurs propres de Σ_i,
- Δ_i est une matrice diagonale contenant les valeurs propres de Σ_i.

Le modèle $[a_{ij}b_iQ_id_i]$

Nous supposons de plus que la matrice Δ_i a la forme suivante :

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

où $a_{ij} > b_i$, pour $j = 1, ..., d_i$.

Le modèle $[a_{ij}b_iQ_id_i]$ et ses sous-modèles

Ainsi, nous obtenons une re-paramétrisation du mélange gaussien :

- dont la complexité est contrôlée par les dimensions d_i des sous-espaces,
- que nous noterons $[a_{ij}b_iQ_id_i]$ dans la suite.

En forçant certains paramètres à être communs dans une même classe ou entre les classes :

- nous obtenons des modèles de plus en plus régularisés,
- qui vont du modèle gaussien le plus complet,
- aux modèles les plus parcimonieux.

Notre famille contient 28 modèles répartis de la façon suivante :

- 14 modèles à orientations libres,
- 12 modèles à orientations communes,
- 2 modèles à matrice de covariance communes.

Le modèle $[a_{ij}b_iQ_id_i]$ et ses sous-modèles

Modèle	Nombre de paramètres	Ordre asymptotique	Nb de prms $k=4$, $d=10$, $p=100$	Estimation par MV
$[a_{ij}b_iQ_id_i]$	$\rho+\bar\tau+2k+D$	kpd	4231	CF
$[a_{ij}bQ_id_i]$	$\rho+\bar\tau+k+D+1$	kpd	4228	CF
$[a_ib_iQ_id_i]$	$\rho+\bar\tau+3k$	kpd	4195	CF
$[ab_iQ_id_i]$	$\rho+\bar\tau+2k+1$	kpd	4192	CF
$[a_ibQ_id_i]$	$\rho+\bar\tau+2k+1$	kpd	4192	CF
$[abQ_id_i]$	$\rho+\bar\tau+k+2$	kpd	4189	CF
$[a_{ij}b_iQ_id]$	$\rho+k(\tau+d+1)+1$	kpd	4228	CF
$[a_jb_iQ_id]$	$\rho+k(\tau+1)+d+1$	kpd	4198	CF
$[a_{ij}bQ_id]$	$\rho+k(\tau+d)+2$	kpd	4225	CF
$[a_jbQ_id]$	$\rho+k\tau+d+2$	kpd	4195	CF
$[a_ib_iQ_id]$	$\rho+k(\tau+2)+1$	kpd	4192	CF
$[ab_iQ_id]$	$\rho+k(\tau+1)+2$	kpd	4189	CF
$[a_ibQ_id]$	$\rho+k(\tau+1)+2$	kpd	4189	CF
$[abQ_id]$	$\rho+k\tau+3$	kpd	4186	CF
$[a_{ij}b_iQd_i]$	$\rho+\tau+D+2k$	pd	1396	FG
$[a_{ij}bQd_i]$	$\rho+\tau+D+k+1$	pd	1393	FG
$[a_ib_iQd_i]$	$\rho+\tau+3k$	pd	1360	FG
$[a_ibQd_i]$	$\rho+\tau+2k+1$	pd	1357	FG
$[ab_iQd_i]$	$\rho+\tau+2k+1$	pd	1357	FG
$[abQd_i]$	$\rho+\tau+k+2$	pd	1354	FG
$[a_{ij}b_iQd]$	$\rho+\tau+kd+k+1$	pd	1393	FG
$[a_jb_iQd]$	$\rho+\tau+k+d+1$	pd	1363	FG
$[a_{ij}bQd]$	$\rho+\tau+kd+2$	pd	1390	FG
$[a_ib_iQd]$	$\rho+\tau+2k+1$	pd	1357	IP
$[ab_iQd]$	$\rho+\tau+k+2$	pd	1354	IP
$[a_ibQd]$	$\rho+\tau+k+2$	pd	1354	IP
$[a_jbQd]$	$\rho+\tau+d+2$	pd	1360	CF
$[abQd]$	$\rho+\tau+3$	pd	1351	CF
Full-GMM	$\rho+kp(p+1)/2$	$kp^2/2$	20603	CF
Com-GMM	$\rho+p(p+1)/2$	$p^2/2$	5453	CF
Diag-GMM	$\rho+kp$	$2kp$	803	CF
Sphe-GMM	$\rho+k$	kp	407	CF

Table 1. Propriétés de la famille du modèle $[a_{ij}b_iQ_id_i]$: $\rho=kp+k-1$, $\bar\tau=\sum_{i=1}^k d_i[p-(d_i+1)/2]$, $\tau=d[p-(d+1)/2]$ et $D=\sum_{i=1}^k d_i$.

Plan de l'exposé

1. Classification des données de grande dimension
2. Une famille de modèles gaussiens pour la grande dimension
3. Clustering des données de grande dimension : HDDC
4. Expérimentations et applications à l'analyse d'images

L'algorithme EM

Généralement, les paramètres du modèle sont estimés grâce à l'algorithme EM :

- Étape E : cette étape calcule à l'itération q les probabilités *a posteriori* $t_{ij}^{(q)} = P(Z = i | X = x_j, \theta^{(q)})$:

$$t_{ij}^{(q)} = \frac{\pi_i^{(q-1)} f(x_j, \theta_i^{(q-1)})}{\sum_{\ell=1}^{k} \pi_\ell^{(q-1)} f(x_j, \theta_\ell^{(q-1)})}.$$

- Étape M : cette étape calcule les estimateurs des paramètres θ_i en maximisant la vraisemblance conditionnelle :

$$\hat{\pi}_i^{(q)} = \frac{n_i^{(q)}}{n}, \quad \hat{\mu}_i^{(q)} = \frac{1}{n_i^{(q)}} \sum_{j=1}^{n} t_{ij}^{(q)} x_j,$$

$$\hat{\Sigma}_i^{(q)} = \frac{1}{n_i^{(q)}} \sum_{j=1}^{n} t_{ji}^{(q)} (x_j - \hat{\mu}_i^{(q)})(x_j - \hat{\mu}_i^{(q)})^t,$$

où $n_i^{(q)} = \sum_{j=1}^{n} t_{ij}^{(q)}$.

L'étape E de l'HDDC

L'étape E calcule à l'itération q les probabilités $t_{ij}^{(q)}$ grâce à la relation :

$$t_{ij}^{(q)} = 1/\sum_{\ell=1}^{k} \exp\left(\frac{1}{2}(K_i^{(q-1)}(x_j) - K_\ell^{(q-1)}(x_j))\right),$$

avec $K_i(x) = -2\log(\pi_i f(x,\theta_i))$ qui vaut :

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

et $\|.\|^2_{\Lambda_i}$ est la distance de Mahalanobis avec $\Lambda_i = \tilde{Q}_i \Delta_i \tilde{Q}_i{}^t$.

L'étape E de l'HDDC

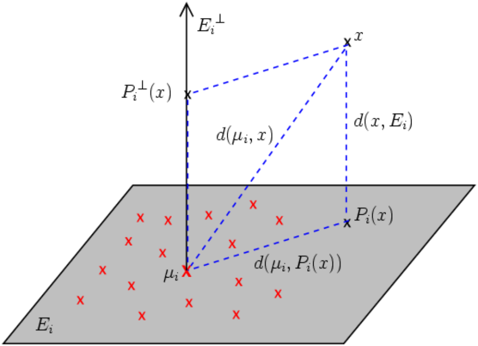


Fig. 1. Les sous-espaces $\mathbb{E}_i$et $\mathbb{E}_i^\perp$de la *i*ème composante.

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

L'étape M de l'HDDC

Les estimateurs du MV de paramètres du modèle $[a_{ij}b_iQ_id_i]$ sont explicites :

- Sous-espace $\mathbb{E}_i$: les d_i premières colonnes de Q_i sont estimées par les vecteurs propres associés aux d_i plus grandes valeurs propres λ_{ij} de $\hat{\Sigma}_i$.

- Estimateur de a_{ij} : les paramètres a_{ij} sont estimés par les d_i plus grandes valeurs propres λ_{ij} de $\hat{\Sigma}_i$.

- Estimateur de b_i : le paramètre b_i est estimé par :

$$\hat{b}_i = \frac{1}{(p - d_i)} \left(\mathrm{trace}(\hat{\Sigma}_i) - \sum_{j=1}^{d_i} \lambda_{ij} \right).$$

L'étape M de l'HDDC

Nous avons choisi d'estimer la dimension intrinsèque d_i de chaque classe grâce au *scree-test* de Cattell :

- méthode basée sur l'ébouli des valeurs propres de $\hat{\Sigma}_i$,
- qui analyse les différences entre les valeurs propres consécutives.

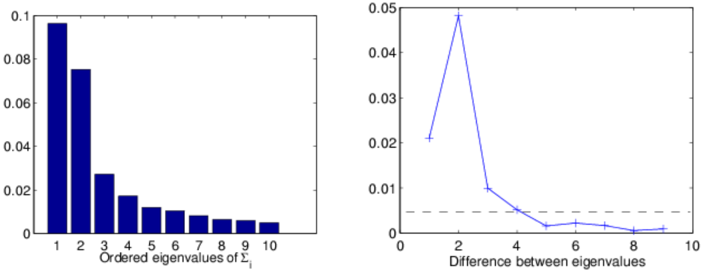


Fig. 2. Le scree-test de Cattell.

Plan de l'exposé

1. Classification des données de grande dimension
2. Une famille de modèles gaussiens pour la grande dimension
3. Clustering des données de grande dimension : HDDC
4. Expérimentations et applications à l'analyse d'images

Influence de la dimension

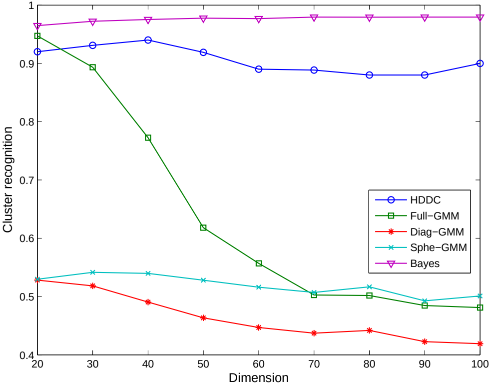


Fig. 3. Influence de la dimension des données sur le taux de classification correcte.

L'algorithme HDDC en action

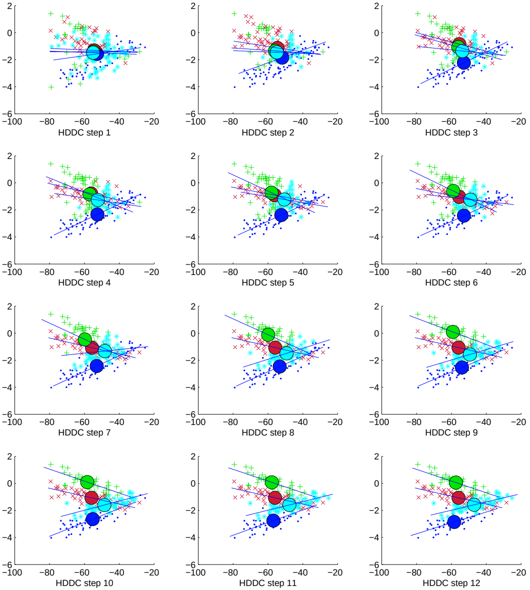


Fig. 4. Les étapes de l'algorithme EM sur les données "Crabes".

Application à la reconnaissance de textures

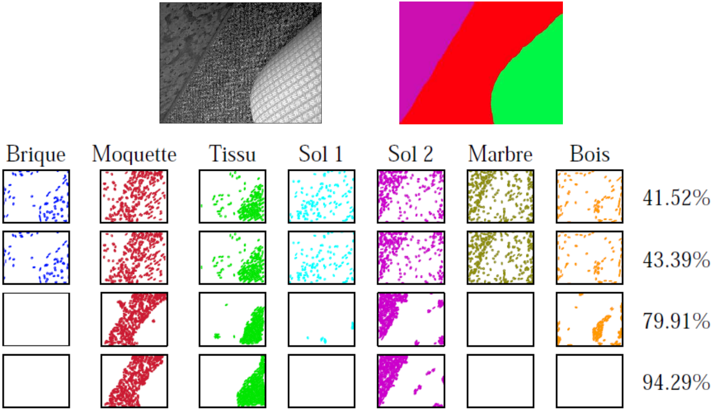


Fig. 4. Segmentation d'une image multi-texture : de haut en bas, Σ^{diag}, Σ^{hdim}, Σ^{diag} + champ de Markov caché et Σ^{hdim} + champ de Markov caché.

Application à la caractérisation du sol de Mars

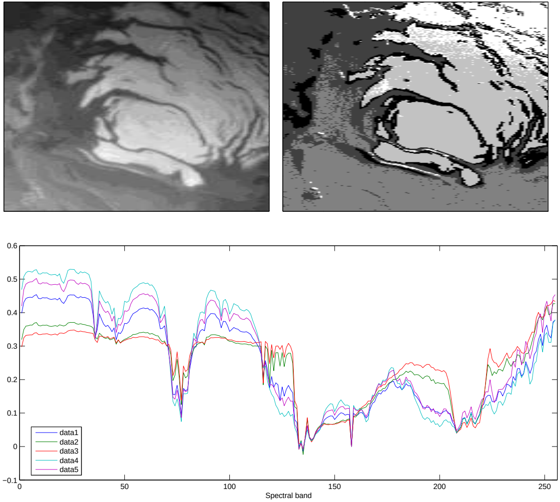


Fig. 5. Analyse de données spectrales de la planète Mars.