Reconnaissance de chiffres et débruitage par généralisation de l'ACP aux RKHS

1 Introduction

L'objectif est d'illustrer l'utilisation de l'analyse en composante principale et de sa variante RKHS pour la reconnaissance automatique de chiffres. Ce problème est rencontré, par exemple, par les services postaux - tri automatique du courrier depuis les codes postaux - ou dans le domaine bancaire - paiement par chèques.

Pour cela, il vous est fourni des scripts et fonctions MATLAB. Le script `Reconnaissance_chiffre` vous permet notamment de lire des données (dictionnaire de chiffres manuscrits), puis de résoudre le problème de reconnaissance tel que décrit dans la suite de l'énoncé. Le sujet consiste à modéliser ce problème, puis à implanter les fonctions nécessaires à sa résolution. Dans toute la suite de l'énoncé, nous noterons $\mathbb{N}_n$ l'ensemble des entiers naturels de 1 à n.

Nous faisons l'hypothèse que nous avons déjà à notre disposition un grand volume de données de chiffres. Ces données ont été préalablement triées - une classe est associée à chaque chiffre - et mises sous un format permettant leur usage (calculs) sur un ordinateur. Dans notre cas, cette dernière étape consiste à transformer l'image d'un chiffre en un vecteur de taille 256 (image de 16 pixels par 16 pixels stockée sous forme vectorielle), dont les composantes représentent le niveau de gris des pixels. Dans notre cas, nous avons entre 72 et 104 images selon les chiffres : ceci conduit à des matrices de données ayant $m = 256$ lignes et $n \in \{72, 78, 84, 90, 104\}$ colonnes. Notre objectif est d'extraire suffisamment d'information de ces données pour être en mesure de reconnaitre, sans intervention humaine, les chiffres présents sur de nouvelles images. Ceci va se faire en deux étapes :

1. Réduction de la dimension du problème via la construction de sous-espaces représentatifs de chacune de ces classes via une analyse en composantes principales (ACP).
2. Reconnaissance du chiffre via des projections orthogonales sur ces sous-espaces représentatifs. Cette étape pourra s'accompagner d'une reconstruction de l'image du chiffre ainsi reconnu dans une optique de débruitage de celle-ci.

Nour proposons d'implanter cette stratégie pour l'espace de Hilbert $\mathbb{R}^m$ (à

savoir l'espace de nos données "images"), puis de généraliser cette approche aux espaces de Hilbert à noyau reproduisant (RKHS).

2 Stratégie 1 : ACP "classique"

La première partie porte sur l'implantation de cette approche : implantation de l'ACP et algorithme de classification/reconstruction.

2.1 Réduction de la dimension du problème par ACP

Le tableau des données $X \in \mathbb{R}^{m\times n}$ est constitué de n individus $\mathbb{R}^m$. Autrement dit, chaque colonne i de X correspond à une donnée $x_i \in \mathbb{R}^m$ (image d'un chiffre). On note $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ l'individu moyen. On appelle tableau centré des données la matrice

$$X^c = \big[(x_1 - \overline{x})\cdots(x_n - \overline{x})\big] = \big[x_1^c \cdots x_n^c\big]$$

La matrice de variance/covariance $\Sigma = \frac{1}{n}X^c\big(X^c\big)^\top \in \mathbb{R}^{m\times m}$ indique le niveau de corrélation entre les axes de la base dans laquelle les individus sont initialement exprimés. Puisque Σ est symétrique, elle est diagonalisable dans une base orthonormée. Ainsi,

$$\exists U, D \in \mathbb{R}^{m\times m} : \Sigma = UDU^\top$$

avec U une matrice orthogonale, et D matrice diagonale, telle que $d_1 \geq d_2 \geq ... \geq d_m \geq 0$.

Si l'on veut exprimer un individu x_i dans ce nouveau repère, on applique d'abord le changement d'origine : $x_i^c = x_i - \overline{x}$, suivi du changement de base, i.e. le produit par la matrice de passage de la base canonique vers la base principale : $c_i = U^\top x_i^c$.

Chacun des vecteurs propres $(u_i)_{i=1:m}$ (colonnes de U) est solution du problème d'optimisation suivant :

$$\begin{array}{rcll} u_i & = & \text{argmax} & w^T\Sigma w \\ & & w \in \mathbb{R}^m, w^T w = 1, w \perp \{u_1, \cdots u_{i-1}\} & \end{array} \tag{1}$$

Le sous-espace représentatif de la classe de chiffres est à chercher parmi les sous-espace de dimension k engendrés par les vecteurs $(u_i)_{i=1:m}$: le sous-espaces engendré par les k premiers vecteurs de la famille $(u_i)_{i=1:r}$ sera appelé sous-espace dominant associé à la matrice des données centrées X^c.

La valeur de k est déterminée afin d'obtenir une approximation dont la précision est définie a priori :

$$PrecApprox = 1 - \sqrt{\frac{d_k}{d_1}} \tag{2}$$

On cherchera donc le plus petit entier k permettant d'atteindre cette précision.

2.2 Classification et reconstruction

Notons $p \in \mathbb{N}_5$ le nombre de classes dont nous disposons (seules les données pour les chiffres allant de 1 à 5 sont disponibles) , $(U_i)_{i\in\mathbb{N}_p}$ les bases orthonormales de chacun des sous-espaces définis lors de la première étape, et C le vecteur représentant l'image du chiffre à déterminer (stockage sous forme vectoriel plutôt que matriciel).

La reconnaissance du chiffre est réalisé en déterminant à quelle classe il appartient. Pour cela, nous comparons les distance de C à chacun des sous-espaces représentatifs des différentes classes. Ceci est obtenu en calculant la norme de la composante de C dans l'orthogonal de l'espace engendré par les colonnes de chacune des matrices U_i :

$$d_i = \frac{\|(I - U_i U_i^T)C\|_2}{\|C\|_2}, \quad \forall i \in \mathbb{N}_p. \tag{3}$$

Nous choisissons alors la classe i_0 associée à la plus petite valeur des $(d_i)_{\mathbb{N}_p}$.

2.3 Travail à réaliser

— Développer les fonctions réalisant les étapes de réduction de la dimension du problème par ACP, de reconnaissance de chiffres et de débruitage de l'image des chiffres reconnus.
— Etudier la sensibilité des résultats par rapports aux différent paramètres expérimentaux (choix de $PrecApprox$, variance du bruit, etc..)

3 Stratégie 2 : généralisation aux RKHS

3.1 Principes

On se propose de reprendre cette approche sous l'angle du RKHS $\mathcal{H}$ de noyau reproduisant k le noyau linéaire, afin de généraliser aux RKHS quelconques.

Le problème d'optimisation (1) s'écrit alors :

$$\begin{aligned} f_i &= \operatorname{argmax} \sum_{j=1}^{n} f(x_j)^2 \\ & f \in \mathcal{H}, \|f\|_{\mathcal{H}^2} = 1, f \perp \{f_1, \cdots f_{i-1}\} \end{aligned} \tag{4}$$

D'après le théorème de représentation, $\exists \alpha^i \in \mathbb{R}^n$ tel que $f_i(\circ) = \sum_{j=1}^{n} \alpha_j^i k(\circ, x_j)$.

On obtient alors la séquence de problèmes d'optimisation :

$$\begin{aligned} \alpha_i &= \operatorname{argmax} \quad \alpha^T K^2 \alpha \\ & \alpha \in \mathbb{R}^n, \alpha^T K \alpha = 1, \alpha^T K \alpha_j = 0, \quad \forall j = 1 : i-1 \end{aligned} \tag{5}$$

Notons $(u_i)_{i=1:n} \in \mathbb{R}^n$ une base orthonormée de vecteur propre de K, associés aux valeurs propres $d_1 \geq d_2 \geq ... \geq d_n \geq 0$. En notant r le rang de la matrice du noyau K, on montre que

$$\forall i = 1 : r, \quad \alpha_i = \frac{1}{\sqrt{d_i}} u_i. \tag{6}$$

L'algorithme de PCA dans les RKHS s'écrit alors

1. Evaluer la matrice du noyau sur les données : cette étape devra tenir compte du centrage des données dans l'espace des *features* : $(\phi(x_i))_{i=1:n}$ sont centrées.
2. Calculer les couples propres (d_i, u_i) de K nécessaires pour atteindre le niveau de précision souhaité.
3. Calculer les vecteurs $\alpha_i = \frac{1}{\sqrt{d_i}} u_i$ associés.
4. Les composantes principales sont alors obtenues par

$$y_i(x) = \sum_{j=1}^{n} \alpha_j^i k(x, x_j).$$

La projection d'une *feature* $\phi(x)$ sur le sous-espace engendré par la famille de vecteurs de $\mathcal{H}$, associée à la décomposition en éléments propres de K, s'écrit alors

$$\Pi_k \phi(x) = \sum_{j=1}^{k} y_j(x) v_j,$$

avec

$$v_j = \sum_{i=1}^{n} \alpha_j^i \phi(x_i), \text{ et } k(x_i, x_j) = \phi(x_i)^T \phi(x_j).$$

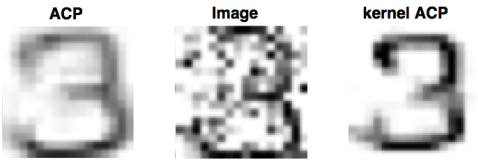

Figure 1 – Image bruitée du chiffre 3 : reconstruction par ACP et par "Kernel" ACP avec le noyau Gaussien.

La reconstruction se fait alors par résolution d'un problème aux moindres carrés dans l'espace des *features* :

$$\min \rho(z) = \|\Pi_k \phi(x) - \phi(z)\|^2_{\mathcal{H}},$$

ce qui est équivalent à minimiser

$$\rho(z) = k(z,z) - 2\sum_{j=1}^{k} \phi(x)^T v_j \sum_{l=1}^{n} \alpha_l^j k(x_l, z).$$

Dans le cas du noyau Gaussien, il est possible d'estimer z itérativement par la formule de récurrence :

$$z_{t+1} = \frac{\sum_{i=1}^{n} \gamma_i k(z, x_i) x_i}{\sum_{i=1}^{n} \gamma_i k(z, x_i)},$$

avec $\gamma_i = \sum_{j=1}^{k} y_j(x)\alpha_j^i$.

3.2 Travail à réaliser

— Développer les fonctions réalisant les étapes de réduction de la dimension du problème par ACP dans le RKHS, associé au noyau linéaire, et de reconnaissance de chiffres. Que constatez-vous à l'execution ?
— Généraliser l'approche aux noyau polynomial et Gaussien. Pour ce dernier, vous implanterez la reconstruction débruitée (cf Fig. 3.2).
— Etudier la sensibilité des résultats par rapports aux différent paramètres expérimentaux (choix de *PrecApprox*, variance du bruit, etc..)