INSA TOULOUSE Année 2009-2010
Département STPI
$3^{\text{ème}}$ année MIC

UV Des données aux modèles
Partie - Problèmes inverses

Mercredi 9 juin 2009 - 14 :00 à 15 :30
Les documents, calculettes et téléphones portables sont interdits.
Le barême entre crochets est donné à titre indicatif.

Dans tout l'examen, on utilise le produit scalaire canonique noté $\langle \cdot, \cdot \rangle$ et la norme associée notée $\| \cdot \|_2$.

Exercice 1 - Cours (10 points)

1. *Rappelez le théorème de décomposition en valeurs singulières [2 pts].*
   Voir cours. Ne pas oublier dans l'énoncé A est de rang k, il y a k valeurs non nulles sur la diagonale.
2. *Est-ce que la SVD est unique ? Si oui, expliquez pourquoi. Si non, donnez un contre-exemple [1,5 pts].*
   La SVD n'est pas unique. Par exemple $I = UIU^T$ pour toute matrice unitaire U.
3. *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice qui admet la décomposition SVD $A = U\Sigma V^T$. On admet que A est inversible. Donnez la SVD de A^{-1} [1,5 pt].*
   On a $A^{-1} = V\Sigma^{-1}U^T$ qui est une SVD de A^{-1}.
4. *On souhaite résoudre le système $Ax = b$ où $A \in \mathcal{M}_{n,n}(\mathbb{R})$. A quelle condition sur A, ce système est-t'il bien/mal posé [1 pt] ?*
   Il suffit que A soit inversible. Note : en pratique, pour la stabilité, il faut s'intéresser à la plus petite valeur singulière de A.
5. *On s'intéresse à la résolution de $Ax = b$ où $A \in \mathcal{M}_{m,n}(\mathbb{R})$, $b \in \mathbb{R}^n$ et $m > n$ (m lignes, n colonnes).*
   - *Expliquez à quelle condition ce système est bien/mal posé [1,5 pts].*
     Le système n'est jamais bien posé (plus d'équations que d'inconnues $\Rightarrow$ pas de solution en général).
   - *Ecrivez le système des équations normales à $Ax = b$ (les équations que satisfont les solutions au sens des moindres carrés) [1 pt].*
     Les équations normales sont $A^T(Ax - b) = 0$.
   - *A quelle condition sur A le système des équations normales est-il bien/mal posé [1,5 pts] ?*
     Il faut que $rang(A) = n$ ce qui signifie aussi que les n valeurs singulières de A soient non nulles.

****

Exercice 2 - Directions privilégiées de la SVD (3,5 pts)

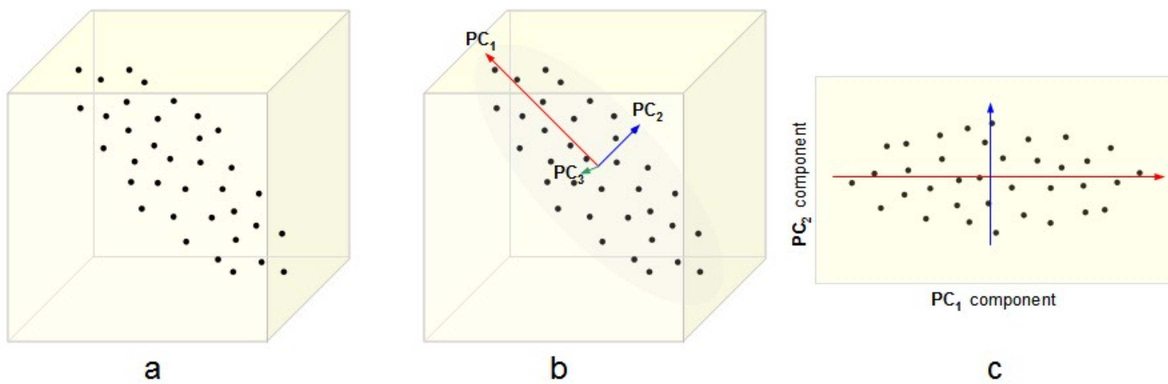
1. *Soit $A = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}$. Comment choisir $x \in \mathbb{R}^2$ tel que $\|x\|_2 = 1$ et $\|Ax\|_2$ soit maximal [1 pt] ?*
   Il suffit de choisir le vecteur colonne $x = (0; 1)$ ou $x = (0; -1)$.
2. *Soit $A = U\Sigma V^T$ avec $U = (u_1, u_2, ..., u_m)$ et $V = (v_1, v_2, ..., v_n)$. Comment choisir $x \in \mathbb{R}^n$ tel que $\|x\|_2 = 1$ et $\|Ax\|_2$ soit maximal (justifier) [2,5 pts] ?*

   Il faut pour que $\|Ax\|_2$ soit maximal, que $V^T x = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$. On écrit x dans la base des $\{v_i\}$. Ainsi, on se rend compte que la solution optimale est $x = v_1$.

****

Exercice 3 - Analyse en composantes principales (9 pts)

Cet exercice est une introduction à l'analyse en composantes principales.
On dispose d'un ensemble de n points $a_i \in \mathbb{R}^m$. Ces points peuvent être rangés dans une matrice $A = (a_1, a_2, ..., a_n)$. L'idée de l'analyse en composante principale consiste à représenter ces points dans un sous-espace vectoriel de dimension inférieure à m en gardant l'information essentielle. La figure suivante illustre ce principe.



On voit à gauche un nuage de points de $\mathbb{R}^3$. La visualisation de ce nuage est compliquée (c'est encore plus vrai en dimension 4 ou plus). Pour effectuer une visualisation en 2D, on cherche deux axes principaux au nuage notés PC1 et PC2 (PC=Principal Component), et on représente le nuage 3D en projetant les points sur ces deux axes principaux. Cette représentation contient l'information principale du nuage.
Le but de cet exercice est de trouver l'expression des axes principaux. La recherche du premier axe peut être formulée ainsi : on cherche $x \in \mathbb{R}^m$, tel que $\|x\|_2 = 1$ [1] et $a_i \approx \alpha_i x$ pour un certain α_i. De manière plus précise, on cherche :

$$\text{Trouver} \quad x \in \underset{x\in\mathbb{R}^m, \|x\|_2=1}{\arg\min} \min_{\alpha\in\mathbb{R}^n} \sum_{i=1}^{n} \|\alpha_i x - a_i\|_2^2 \tag{1}$$

1. *On note $f_i(\alpha_i) = \|\alpha_i x - a_i\|_2^2$. Calculez :*

$$\underset{\alpha_i\in\mathbb{R}}{\arg\min} \|\alpha_i x - a_i\|_2^2 \quad \textit{[2 pts].}$$

   On dérive f par rapport à α et on annule la dérivée. On trouve ainsi :

$$\alpha_i = \frac{\langle x, a_i\rangle}{\|x\|_2^2}$$

2. *En déduire que le problème (4) est équivalent à :*

$$\textit{Trouver} \quad x \in \underset{x\in\mathbb{R}^m, \|x\|_2=1}{\arg\min} \sum_{i=1}^{n} \|\langle x, a_i\rangle x - a_i\|_2^2 \quad \textit{[1 pt].} \tag{2}$$

   Il suffit de remplacer l'expression précédente dans (4) et de se rendre compte que $\|x\|_2 = 1$.

3. *En développant la somme des normes, montrez que le problème (2) est encore équivalent à :*

$$\textit{Trouver} \quad x \in \underset{x\in\mathbb{R}^m, \|x\|_2=1}{\arg\min} -\|A^T x\|_2^2 \quad \textit{[3 pts]} \tag{3}$$

   En développant la norme, on trouve :

$$\|\langle x, a_i\rangle x - a_i\|_2^2 = \|a_i\|_2^2 - \langle a_i, x\rangle^2.$$

[1] on choisit de le normaliser car il sert juste à définir un sous-espace vectoriel

On peut supprimer $\|a_i\|_2^2$ du problème de minimisation, car c'est une constante. On remarque pour finir que :

$$\sum_{i=1}^{n} \langle a_i, x\rangle^2 = \|A^T x\|_2^2.$$

4. *On admet que $A = U\Sigma V^T$ avec $U = (u_1, u_2, ..., u_n)$ et $V = (v_1, v_2, ..., v_n)$. En utilisant cette décomposition en valeurs singulières, trouvez la solution du problème (3) [3 pts].*

   On a $A^T = V\Sigma U^T$. En utilisant le fait que U et V soient des isométries et en faisant le changement de variable : $y = U^T x$, on se rend compte que le problème (3) est équivalent à :

$$\underset{y\in\mathbb{R}^m, \|y\|_2=1}{\arg\min} \quad -\|\Sigma y\|_2^2.$$

   La solution de ce problème est $y^* = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$. Or $x^* = Uy^*$, donc $x^* = u_1$

5. *On souhaite maintenant connaître le meilleur sous-espace de dimension k pour représenter nos données. Proposez un problème d'optimisation à résoudre pour trouver les k-vecteurs engendrant le sous-espace. Avez-vous une idée de la solution, c'est-à-dire pouvez-vous donner une expression explicite de ces k vecteurs [* pts] ?*

   On serait tenté de résoudre :

$$\text{Trouver} \quad X = [x_1, x_2, ..., x_k] \in \underset{X\in\mathcal{M}_{m,k}, \|x_i\|_2=1}{\arg\min} \ \min_{B\in\mathcal{M}_{k,n}} \sum_{i=1}^{n} \|BX - A\|_F^2 \tag{4}$$

   On peut imaginer en regardant le résultat précédent que la solution est donnée par : $x_1 = u_1$, $x_2 = u_2$,... $x_k = u_k$.

Note : vous reverrez ce type de représentation et l'utiliserez souvent dans les cours de statistique de 4ème année.

****