

# Recommandation - Rapport partiel

Ken Chanseau–Saint-Germain & Vincent Vidal

18 janvier 2014

## Table des matières

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Approximation basique</b>	<b>2</b>
2.1	Par moyenne . . . . .	2
2.2	Par minimisation de problème convexe . . . . .	2
<b>3</b>	<b>Approximation par voisinage</b>	<b>3</b>
3.1	Plus proche voisin . . . . .	3
3.2	$k$ plus proches voisins . . . . .	3
3.3	Voisinage . . . . .	3
<b>4</b>	<b>Décomposition en valeurs singulières</b>	<b>4</b>

## Notation

On notera pour  $p > 0$  un réel,  $X$  un vecteur et  $M$  une matrice quelconque :

$$\|X\|_p = \left( \sum_i |x_i|^p \right)^{\frac{1}{p}} \quad \|M\|_p = \left( \sum_{i,j} |m_{i,j}|^p \right)^{\frac{1}{p}}$$

$$\|M\|_p = \sup_{\|x\|_p=1} \|Mx\|_p$$

Et on posera  $\|X\|_0$  le nombre de composantes non nulles de  $X$ .

## 1 Introduction

On se donne ici  $n$  personnes donnant des notes à  $m$  objets.

On notera  $a_{i,j}$  la note de l'individu  $i$  sur l'objet  $j$  ainsi que  $A = (a_{i,j})_{\substack{1 \leq i \leq n \\ 1 \leq j \leq m}}$  la matrice des notes.

On suppose ici que l'on a accès qu'à une matrice incomplète  $B$  obtenu en annulant certaines composantes de  $A$ . Le but est alors de trouver une bonne approximation  $\hat{A}$  de  $A$  à partir de  $B$ .

On prendra comme mesure d'approximation, l'erreur moyenne suivante :

$$\text{RMSE}(\hat{A}) = \|A - \hat{A}\|_2 = \sqrt{\sum_{i,j} (a_{i,j} - \hat{a}_{i,j})^2}$$

## 2 Approximation basique

### 2.1 Par moyenne

On considère ici l'approximation suivante :

$$\hat{a}_{i,j} = m_{\hat{A}} + p_i + o_j$$

Avec  $m_{\hat{A}}$  la moyenne des valeurs de  $\hat{A}$ ,  $p_i$  la moyenne des notes recentrées qu'a donné la personne  $i$  et  $o_j$  la moyenne des notes recentrées qu'a obtenu l'objet  $j$ . C'est à dire :

$$m_{\hat{A}} = \frac{\sum_{i,j} b_{i,j}}{\|B\|_0} \quad p_i = \frac{\sum_j b_{i,j} - m_{\hat{A}}}{\|B^T \mathbf{e}_i\|_0} \quad o_j = \frac{\sum_i b_{i,j} - m_{\hat{A}}}{\|B \mathbf{e}_j\|_0}$$

---

**Algorithm 1:** Recommandations naïves par moyenne

---

**Data:** La matrice incomplète  $B$ .

**Result:** La matrice d'approximation  $\hat{A}$ .

$m \leftarrow \text{mean}(B)$ ;

**for**  $i \in \llbracket 1, n \rrbracket$  **do**

$p[i] \leftarrow 0$ ;

$card \leftarrow 0$ ;

**for**  $j \in \llbracket 1, m \rrbracket$  **do**

$p[i] \leftarrow p[i] + B[i, j]$ ;

**if**  $b[i, j] \neq 0$  **then**

$card \leftarrow card + 1$ ;

$p[i] \leftarrow p[i] / card - m$ ;

On calcul de même les  $o[j]$ ;

**for**  $i \in \llbracket 1, n \rrbracket$  **do**

**for**  $j \in \llbracket 1, m \rrbracket$  **do**

$\hat{A}[i, j] \leftarrow m + p[i] + o[j]$ ;

**return**  $\hat{A}$ ;

---

## 2.2 Par minimisation de problème convexe

De manière plus précise, on peut résoudre le problème suivant de minimisation convexe pénalisé sur les  $p_i$  et  $o_j$  pour trouver des  $p_i$  et  $o_j$  plus adaptés :

$$\mathcal{P}(p, o) = \sum_{i,j} (a_{i,j} - m_{\hat{A}} + p_i + o_j)^2 + \lambda (\|o\|_1 + \|p\|_1)$$

Problème que l'on peut écrire :

$$\mathcal{P}(x) = \|Mx + b\|_2^2 + \lambda \|x\|_1$$

Il suffit alors de résoudre ce problème, pour un  $\lambda$  donné et de calculer l'approximation.

On utilisera ici un algorithme présenté dans [Mal08] :

---

### Algorithm 2: Minimisation convexe

---

**Data:**  $x, \lambda$ , le nombre d'itération  $N$  et un paramètre  $\mu$ .

**Result:** une meilleur approximation  $x$ .

**for**  $k \in \llbracket 1, N \rrbracket$  **do**

$x \leftarrow x - \mu M^T (Mx - b);$

    /\* Seuillage doux sur chaque composante

\*/

**for**  $i \in \llbracket 1, n \rrbracket$  **do**

$x[i] \leftarrow \text{signe}(x[i]) \cdot \max(|x[i]| - \mu\lambda, 0);$

**return**  $x;$

---

## 3 Approximation par voisinage

Nous noterons par la suite  $P_i$  la  $i$ -ième colonne de  $B^T$ , représentant la personne  $i$  et  $O_j$  la  $j$ -ième colonne de  $B$ , représentant l'objet  $j$ .

### 3.1 Plus proche voisin

La méthode précédente suppose qu'il n'existe qu'un seul type de personne et d'objet : un objet est bien ou mauvais dans l'ensemble et une personne donne de manière générale des notes bonnes ou des notes mauvaises. Ceci apparaît comme très naïf.

On considère ici une méthode différente. Pour une personne  $i$  et un objet  $j$  dont l'information est manquante, nous allons chercher parmi les personnes ayant noté  $j$  celui qui se rapproche le plus  $i$  ainsi que l'objet ayant été noté par  $i$  celui qui se rapproche le plus de  $j$ . En notant  $i_0$  et  $j_0$  respectivement la personne et l'objet trouvé, on peut alors prendre comme approximation la moyenne des deux notes :  $\hat{a}_{i,j} = \frac{1}{2} (a_{i_0,j} + a_{i,j_0})$ .

Il reste à définir comment sélectionner le plus proche voisin. Si on cherche le plus proche voisin d'un vecteur  $X$  de support  $S$  parmi un ensemble de vecteur  $E$ , on considèrera  $E|_S$  les vecteurs de  $E$  restreints au support  $S$  et on prendra alors le plus proche voisin pour la norme euclidienne sur  $\mathbb{R}^{\text{Card}(S)}$ .

### 3.2 $k$ plus proches voisins

### 3.3 Voisinage

On ne cherche plus ici le plus proche voisin, mais une combinaison de la note donnée par les personnes ayant noté l'objet en question, et ce en fonction de leur ressemblance avec la personne initiale.

Il convient de donner une mesure de ressemblance entre deux personnes  $i_1$  et  $i_2$ , que l'on notera  $s(i_1, i_2)$ . Nous prendrons ici la valeur du cosinus de l'angle entre les deux personnes :

$$s(i_1, i_2) = \max\left(\frac{P_{i_1}^T \cdot P_{i_2}}{\|P_{i_1}\|_2 \|P_{i_2}\|_2}, 0\right)$$

On prendra alors comme approximation :

$$\hat{a}_{i,j} = \frac{\sum_{i_0 \neq i} (s(i_0, i) \cdot b_{i_0,j})}{\sum_{i_0 \neq i} s(i_0, i)}$$

## 4 Décomposition en valeurs singulières

Le but est ici de faire apparaître  $k$  modèles de personnes, dont toutes les autres seront des combinaisons linéaires.

Autrement dit, on souhaite faire apparaître la matrice  $\hat{A}$  de la manière suivante :

$$\hat{A} = P \cdot O$$

avec  $P \in \mathcal{M}_{n,k}(\mathbb{R})$  de colonnes libres que l'on peut prendre orthonormées et  $O \in \mathcal{M}_{k,m}(\mathbb{R})$ .

On peut facilement accéder à une telle décomposition à l'aide de la décomposition en valeurs singulières que l'on tronque à l'ordre  $k$ .

## Références

[Mal08] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 2008.