PCA : Principal Component Analysis
Explication intuitive

Réda Arab

1 Objectif

L'objectif est de passer d'un espace de features de dimension p à un espace de dimension $k < p$.

Cela peut être à des fins de **visualisation** ($k = 2, 3$), de **compression**, **computationnelles** (moins de complexité) ou encore pour éviter l'**overfitting** (mais ce n'est pas "une bonne pratique" ; il vaut mieux régulariser en partie car le terme de 'fitting' est inclus).

Utilisation importante : obtenir des ***nouvelles features décorrélées.***

Principe : On va projeter nos données $x_1, x_2, ..., x_n \in \mathbb{R}^p$ dans un espace de dimension $k < p$ telle que la *variance totale* résultante (c'est-à-dire la *dispersion* de nos données) soit maximisée .

Notation : On ecrit $X \in \mathbb{R}^{n*p}$ avec n le nombre de données, p le nombre de features.

$$X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix} = \begin{pmatrix} X_1 & X_2 & \dots & X_p \end{pmatrix}$$

En pratique on *centre et reduit* les données au préalable (cela simplifie les calculs et évite qu'une feature contienne toute la variance comme on le verra plus tard).

$$x \to \frac{x - \hat{x}}{\sigma_x}$$

2 Rappel de Maths préliminaires

2.1 SVD : Singular Value Decomposition

On pourra se référer à l'article Wikipedia (les schémas sont intéressants)
`https://en.wikipedia.org/wiki/Singular_value_decomposition`

Enoncé : Pour X une matrice réelle de $\mathbb{R}^{n*p}$, on peut écrire $X = UDV^T$ où U et V deux matrices orthogonales de $\mathbb{R}^{n*n}$ et $\mathbb{R}^{p*p}$ et D une matrice avec des termes diagonaux $D_{11} \geq D_{22} \geq \ldots \geq D_{mm}$ où $m = min(n, p)$ et les autres éléments de D étant nuls.

Par exemple, si $n > p$ on aura D de la forme :

$$D = \begin{pmatrix} D_{11} & 0 & \ldots & \ldots \\ 0 & D_{22} & & \\ \vdots & & \ddots & \\ \vdots & & & D_{mm} \\ 0 & \ldots & \ldots & 0 \end{pmatrix}$$

On a alors :

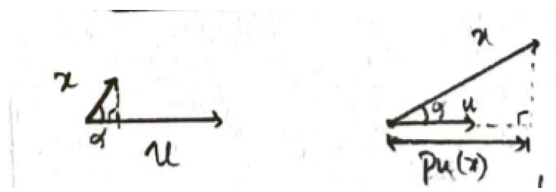
$$X^T X = (UDV^T)^T (UDV^T) = V D^T D V^T$$

$$\boxed{\mathrm{X}^T X = V \Lambda V^T}$$

$\Lambda = D^T D$ qui contient **les valeurs propres** de $X^T X$ (valeurs singulières au carrées) qui sont réelles et positives ou nulles.

V est la matrice qui contient les **vecteurs propres** qui forment une base orthonormée de $\mathbb{R}^p$.

2.2 Produit scalaire et projection en 2D



On a alors : $cos(\alpha) = \frac{p_u(x)}{||x||_2}$ où $p_u(x)$ est la projection orthogonale de x sur la droite avec vecteur directeur u.

Donc $< x, u >= ||x||_2 \cdot ||u||_2 \cdot cos(\alpha) = ||u||_2 \cdot p_u(x)$.

Si $||u||_2 = 1$, on a alors $\boxed{< x, u >= p_u(x)}$

2.3 Caractérisation projection orthogonale sur un plan dans $\mathbb{R}^n, n \geq 3$

Rappel : Pour P un plan, on a $\Pi_P(x) = \underset{y \in P}{\text{argmin}} ||y - x||_2^2$ où $\Pi_P(x)$ est la projection orthogonale de x sur P.

Prenons une base orthonormée de P : (v_1, v_2). $\forall y \in P, y = \lambda_1 v_1 + \lambda_2 v_2$. Donc :

$||y - x||_2^2 = ||x||_2^2 - 2 < x, y > + ||y||_2^2$

$||y - x||_2^2 = ||x||_2^2 - 2\lambda_1 < x, v_1 > - 2\lambda_2 < x, v_2 > + \lambda_1^2 + \lambda_2^2$

Considérons la fonction $\phi(\lambda_1, \lambda_2) = -2\lambda_1 < x, v_1 > - 2\lambda_2 < x, v_2 > + \lambda_1^2 + \lambda_2^2$.

$\frac{\partial \phi}{\partial \lambda_i} = -2 < x, v_i > + 2\lambda_i \quad i = 1, 2$
$\frac{\partial^2 \phi}{\partial \lambda_i^2} = 2 \quad i = 1, 2$
$\frac{\partial^2 \phi}{\partial \lambda_1 \partial \lambda_2} = 0$

On a donc $\frac{\partial^2 \phi}{\partial \lambda^2} \succ 0$ (matrice Hessienne positive définie strictement) donc la fonction est strictement convexe. Le minimum est atteint pour :

$$\frac{\partial \phi}{\partial \lambda} = 0 \quad \text{i.e}$$

$$\boxed{\lambda_i =< x, v_i >} \quad i = 1, 2$$

$$\text{Et} \quad \boxed{\Pi_p(x) =< x, v_1 > v_1 + < x, v_2 > v_2}$$

Généralisation

On généralise facilement à une projection sur un espace E_k de dimension k en prenant $(v_1, ..., v_k)$ une base orthonormée de E_k. On obtient alors :

$$\boxed{\Pi_{E_k}(x) = \sum_{i=1}^{k} < x, v_i > v_i = \sum_{i=1}^{k} (x^T v_i) v_i}$$

3 PCA - Introduction

La **variance totale** est définie comme :

$$\boxed{\frac{1}{n} \sum_{i=1}^{n} ||x_i - \hat{x}||_2^2 = \sum_{i=1}^{n} ||x_i||_2^2 = \sum_{i=1}^{p} ||X_i||_2^2}$$

car les données sont centrées. Le facteur $\frac{1}{n}$ (ou $\frac{1}{n-1}$) n'est pas important pour la suite.

$\rightarrow$ On cherche donc un espace de dimension k tel que la projection des x_i sur cet espace donne des données (avec des nouvelles features associées) qui gardent le maximum de variance totale possible.

On aura une nouvelle matrice des données Y telle que $Y = \begin{pmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_n^T \end{pmatrix}$ avec $y_i \in \mathbb{R}^k$ telle que $\sum_{i=1}^{n} ||y_i||_2^2$ est "maximale" dans le sens défini ci-dessus.

Exemple

Voyons un example quand on projette nos données d'une dimension p vers une dimension 1. Sur les schémas d'une dimension 2 à 1.

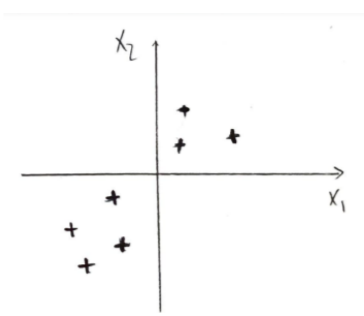


FIGURE 1 – Données en 2D

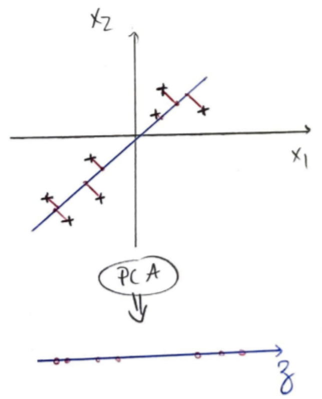


FIGURE 2 – Nouvelle feature avec une variance large

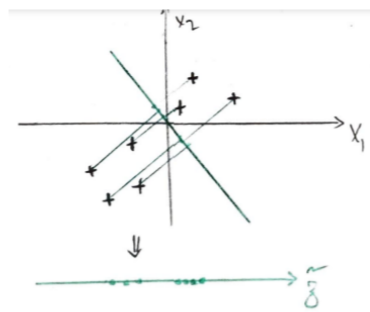


FIGURE 3 – Nouvelle feature avec une variance faible

Prenons le cas où l'on a des données avec 2 features qu'on veut compresser en une seule. On recherche donc un vecteur u (on peut se restreindre à $||u||_2 = 1$ car on cherche une direction) tel que les données projetées sur u ($x_i^T u, \quad i = 1, \ldots, n$) gardent une variance maximale.

i.e $Xu = \begin{pmatrix} x_1^T u \\ x_2^T u \\ \vdots \\ x_n^T u \end{pmatrix}$ et on a vu que $x_i^T u = p_u(x_i)$ pour u de norme 1.

On cherche donc $u \in \mathbb{R}^2$ tel que : $\boxed{u = \underset{v \in \mathbb{R}^2, ||v||_2 = 1}{\text{argmax}} ||Xv||_2^2}$ (maximise la variance totale).

Ou plus généralement dans un espace avec p features : $\boxed{u = \underset{v \in \mathbb{R}^p, ||v||_2 = 1}{\text{argmax}} ||Xv||_2^2}$

On a $||Xv||_2^2 = v^T X^T X v = v^T V D^T D V^T v$ avec $||v||_2 = 1$.

Ecrivons $a = V^T v$. On a $||a||_2 = ||V^T v||_2 = ||v||_2 = 1$ (car V orthogonale).

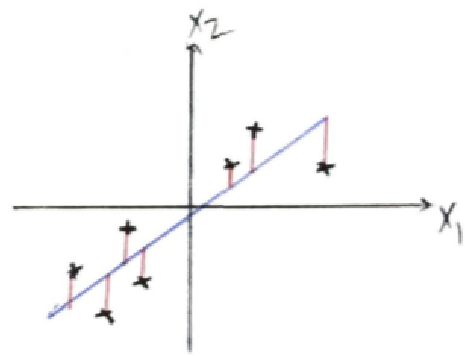
$\rightarrow ||Xv||_2^2 = a^T D^T D a = \sum_{i=1}^m a_i^2 D_{ii}^2 \leq D_{11}^2 \sum_{i=1}^m a_i^2 = D_{11}^2$

Ainsi $a = (1, 0, \ldots, 0)^T$ maximise $||Xv||_2^2$ i.e $\boxed{v = Va = V_1}$ qui est **le vecteur propre associé à** D_{11}^2. La direction qui maximise $||Xv||_2^2$ est donc V_1 et on a $||XV_1||_2^2 = ||D_{11}U_1||_2^2 = D_{11}^2$

Remarques

— Les données projetées $x_i^T v$ sont centrées ($\sum_{i=1}^n x_i^T v = (\sum_{i=1}^n x_i^T)v = 0$), donc la variance totale est bien $\sum_{i=1}^n (x_i^T v)^2 = ||Xv||_2^2$

— Ne pas confondre PCA et régression linéaire.



Ce qu'on cherche à minimiser (les traits rouges) est différent.

4 Formulation du problème

Notre problème peut se résumer à :

$$\boxed{\underset{(v_1,v_2,\ldots,v_k)\in\mathbb{R}^p}{\text{argmax}} \sum_{i=1}^{k}(||Xv_i||_2^2) \quad \text{tels que} \quad v_i^T v_j = \delta_{i,j}}$$

En effet, on a, pour un espace E_k de dimension k et une base orthonormée $(v_1, \ldots, v_k)$:

$\Pi_{E_k}(x_i) = \sum_{i=1}^{k}(x_i^T v_l)v_l \quad$ et $\quad \sum_{i=1}^{n} ||\Pi_{E_k}(x_i)||_2^2 = \sum_{i=1}^{n}\sum_{l=1}^{k}(x_i^T v_l)^2$

Dans ce nouvel espace, on peut réécrire **les cordonnées des x_i projetées dans la base orthonormée** :

$$Y = \begin{pmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_n^T \end{pmatrix} = \begin{pmatrix} x_1^T v_1 & x_1^T v_2 & \ldots & x_1^T v_k \\ \vdots & \vdots & \ldots & \vdots \\ x_n^T v_1 & x_n^T v_2 & \ldots & x_n^T v_k \end{pmatrix} = \begin{pmatrix} Xv_1 & Xv_2 & \ldots & Xv_k \end{pmatrix}$$

La variance totale est alors : $\sum_{i=1}^{n}\sum_{l=1}^{k}(x_i^T v_l)^2 = \sum_{i=1}^{k} ||Xv_l||_2^2$

5 Résolution

On peut résoudre le problème pas à pas, *direction par direction* (le problème étant séparable, c'est une somme).

On commence par V_1 pour la première feature (*cf PCA -Introduction*) puis on cherche un vecteur $v^{(2)}$ tel que $||Xv^{(2)}||_2^2$ est maximale avec $||v^{(2)}||_2 = 1$ et $(v^{(2)})^T V_1 = 0$.

On peut montrer facilement que $\boxed{v^{(2)} = V_2}$.

Récursivement, on cherche à obtenir pour chaque $j = 2, \ldots, k$, le vecteur $v^{(j)}$ défini ci-dessous :

$$\boxed{\begin{array}{c} v^{(j)} \text{ maximise } \|Xv\|_2 \text{ over } v \in \mathbb{R}^p \text{ avec les contraintes} \\ \|v\|_2 = 1 \text{ and } (v^{(l)})^T v = 0 \text{ for all } l < j \end{array}}$$

On obtient $\boxed{(V_1, \ldots, V_k)}$, **les vecteurs propres de $X^T X$** associés aux valeurs propres $D_{11}^2 \geq D_{22}^2 \geq \cdots \geq D_{kk}^2 \geq 0$.

On obtient alors :

$$Y = \begin{pmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_n^T \end{pmatrix} = \begin{pmatrix} XV_1 & XV_2 & \dots & XV_k \end{pmatrix} = \begin{pmatrix} D_{11}U_1 & D_{22}U_2 & \dots & D_{kk}U_k \end{pmatrix}$$

La variance totale de nos données projetées est alors : $\boxed{\sum_{l=1}^{k} ||D_{ll}U_l||_2^2 = \sum_{l=1}^{k} D_{ll}^2}$

La variance totale initiale, avant projection, est : $\boxed{\sum_{l=1}^{m} D_{ll}^2}$

Choix de k

On peut ainsi définir une façon de choisir k (nombre de nouvelles features).

$$\boxed{\min\{k \mid \frac{\sum_{l=1}^{k} D_{ll}^2}{\sum_{l=1}^{m} D_{ll}^2} \geq a\} \quad \text{avec}\, a = 0.80, 0.90\, \text{ou}\, 0.95\, \text{par exemple}}$$

Remarques finales

— Les nouvelles données sont centrées comme on a vu précédemment.

— Les nouvelles features crées sont **décorrélées** :

$$i \neq j, (XV_i)^T(XV_j) = V_i^T V \Lambda V^T V_j = 0.$$

— Les features créées perdent en interprétabilité, explicabilité.

— L'idée générale de cette méthode de réduction de dimension : **garder le plus de dispersion possible entre les données**.