RÉDUCTION DE DIMENSION

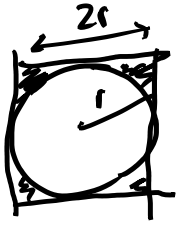
Idée : Passer d'une représentation en p variables à une représentation en d variables
$d \ll p$

1 Motivations

- Visualisation (en particulier $d=2$)
- Utiliser moins de ressources
  - espace de stockage
  - acquisition des données
  - tps de calcul des modèles

- Fléau de la dimension (curse of dimentionality)
  - Apprentissage supervisé marche mieux avec peu de variables
  - En grande dimension, toutes les distances sont grandes

$d = 2$



Proportion du carré en dehors du cercle $= 1 - \frac{\pi}{4}$

$d = 3$ Proportion du cube en dehors de la sphère $= 1 - \pi/6$

$d \to +\infty$ proportion $\to 1$

Les intuitions développées en $d=2$ ou $d=3$ ne sont pas toutes valides en grande dimension.

[2] Sélection vs extraction de variables

- Sélection → Eliminer certaines des variables
- Extraction → Créer de nouvelles variables
  - combiner les variables existantes en de nouvelle variables.

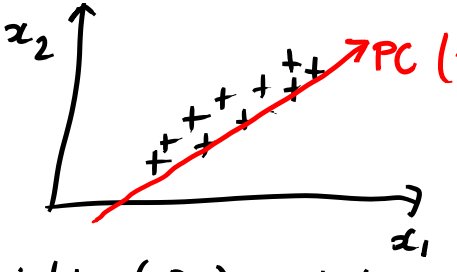
→ ne garder qu'une variable parmi un ensemble de variables corrélés

→ Lasso : éliminer les variables qui ne sont pas utilisées.

[3] Analyse en composantes principales (ACP / PCA)

$\mathcal{D} = \{ \vec{x}_1, \dots, \vec{x}_n \}$ $\vec{x}_i \in \mathbb{R}^p$

Idée : trouver un repère orthonormé sur lequel projeter $\mathcal{D}$ de sorte à maximiser la variance des données projetés



- les nouvelles variables (PCs) sont des combinaisons linéaires des variables initiales
- On connait facilement la proportion de variance expliquées par chaque composante
  → utile pour choisir le nombre de composantes.

[4] Positionnement multidimensionnel / MDS = Multidimensional Scaling

n observations $\vec{x_1}, \dots, \vec{x_n}$ ⟶ distances deux à deux $D_{i\ell} = d(\vec{x_i}, \vec{x_\ell})$

dissimilarité

Je cherche $\vec{z_1}, \dots, \vec{z_n}$ $\vec{z_i} \in \mathbb{R}^d$ $d < p$

tels que $\min \sum_{i,\ell} \left(\| \vec{z_i} - \vec{z_\ell} \|_2 - D_{i\ell} \right)^2$

Si $d(\vec{x_i}, \vec{x_\ell}) = \| \vec{x_i} - \vec{x_\ell} \|_2$ alors on retrouve l'ACP

[5] kPCA = kernel PCA

version à noyau de l'ACP ⇒ Les nouvelles variables sont des fonctions non-linéaires des variables initiales.

[6] t-SNE t-Student Neighborhood Embedding

Idée : modéliser la distribution des distances entre observations par une loi de Student.

très utilisé pour la visualisation

⚠ ne préserve que les structures locales

[7] UMAP