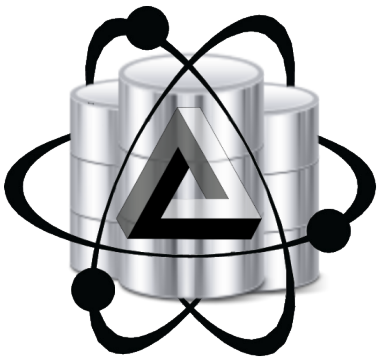
Fondements théorétiques en science des données

Analyse en composantes principales

STT 3795

Guy Wolf
guy.wolf@umontreal.ca

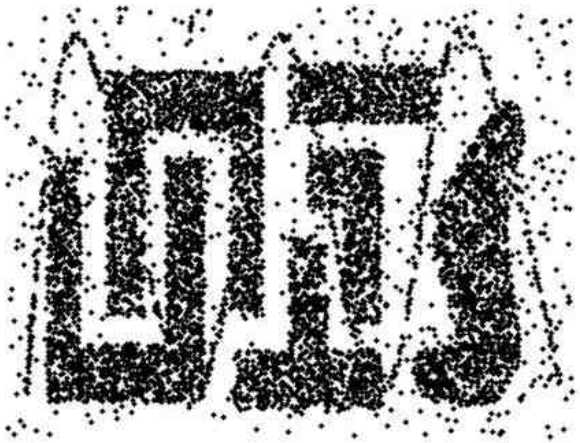
Université de Montréal
Hiver 2020



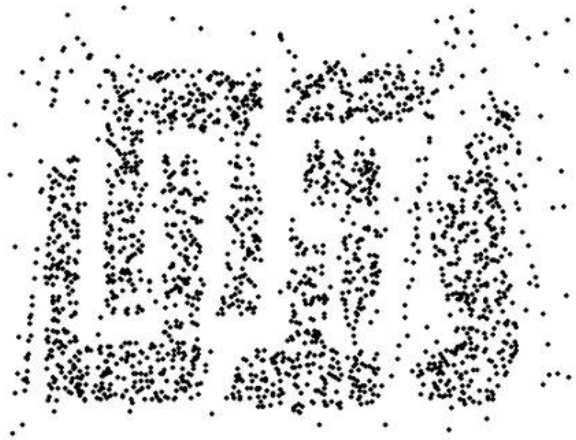
Échantillonnage

Sélectionnez un sous-ensemble de points de données représentatifs au lieu de traiter l'ensemble des données.

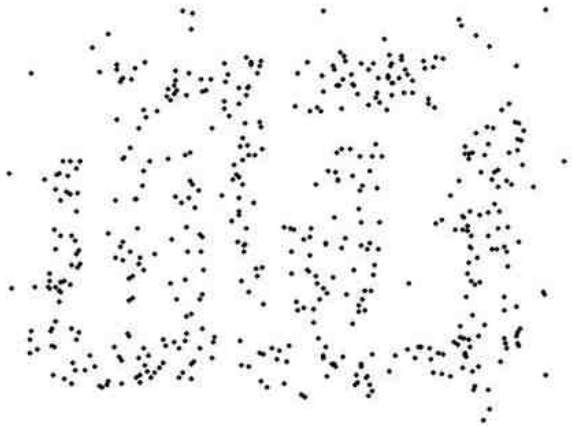
Un sous-ensemble échantillonné n'est utile que si son analyse donne les mêmes motifs, résultats, conclusions, etc. que celui de l'ensemble des données.



8000 points



2000 points



500 points

Échantillonnage

Sélectionnez un sous-ensemble de points de données représentatifs au lieu de traiter l'ensemble des données.

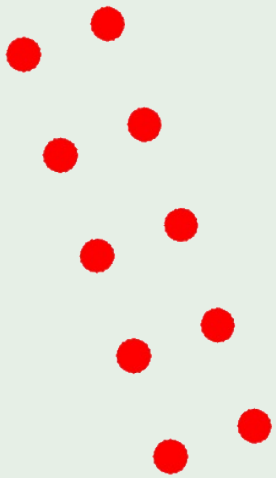
Approches fréquentes d'échantillonnage

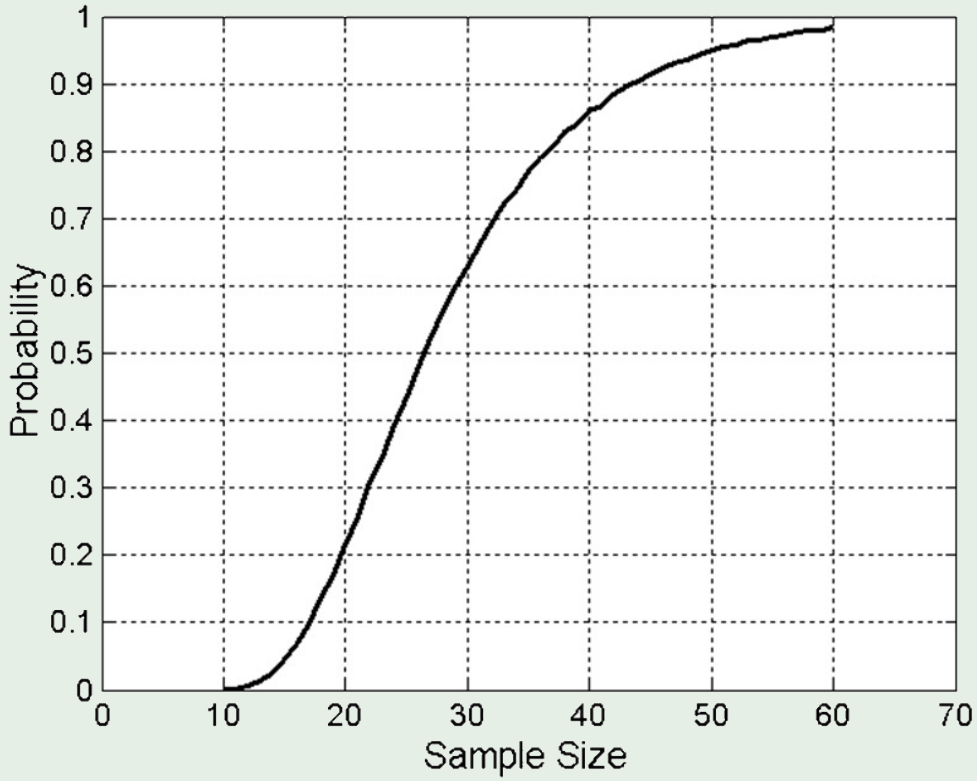
- Au hasard: probabilité égale de sélectionner chaque élément particulier.
- Sans remplacement : sélection itérative & supprimer des éléments.
- Avec remplacement : les articles sélectionnés restent dans la population.
- Stratifié : tirer des échantillons aléatoires de chaque partition.

Le choix d'une taille d'échantillon suffisante est souvent crucial pour un échantillonnage efficace.

Exemple

Choisissez suffisamment d'échantillons pour garantir qu'au moins un représentant est sélectionné dans chaque groupe distinct des données.





Au lieu d'échantillonner des points de données représentatifs, on peut obtenir des données grossières en agrégeant des attributs ou des points de données.

Agrégation

Combinaison de plusieurs attributs à une seule caractéristique, ou de plusieurs points de données en une seule observation.

Exemples

- Transformer les recettes mensuelles en recettes annuelles
- Analyser les quartiers plutôt que les maisons

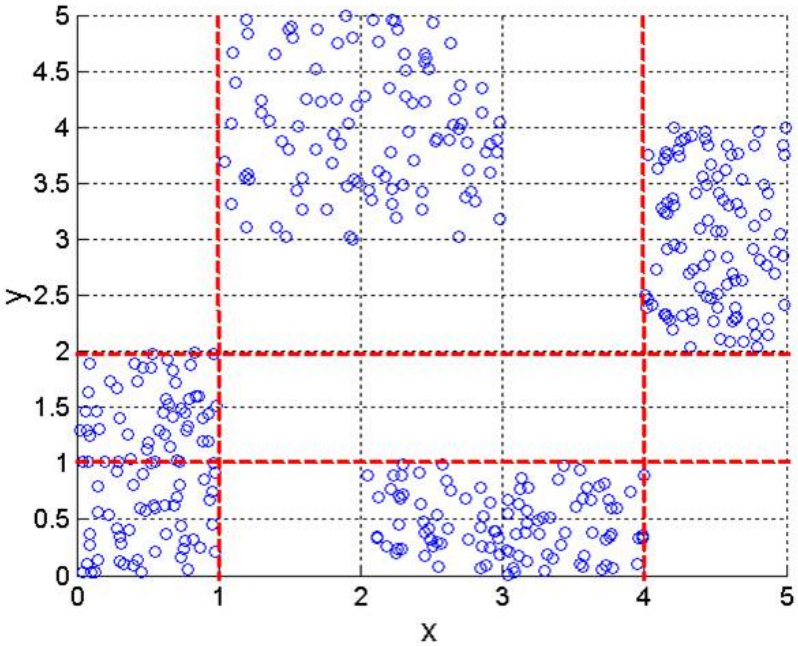
Il est parfois pratique de transformer l'ensemble des données en attributs nominaux (ou ordinaux).

Discrétisation

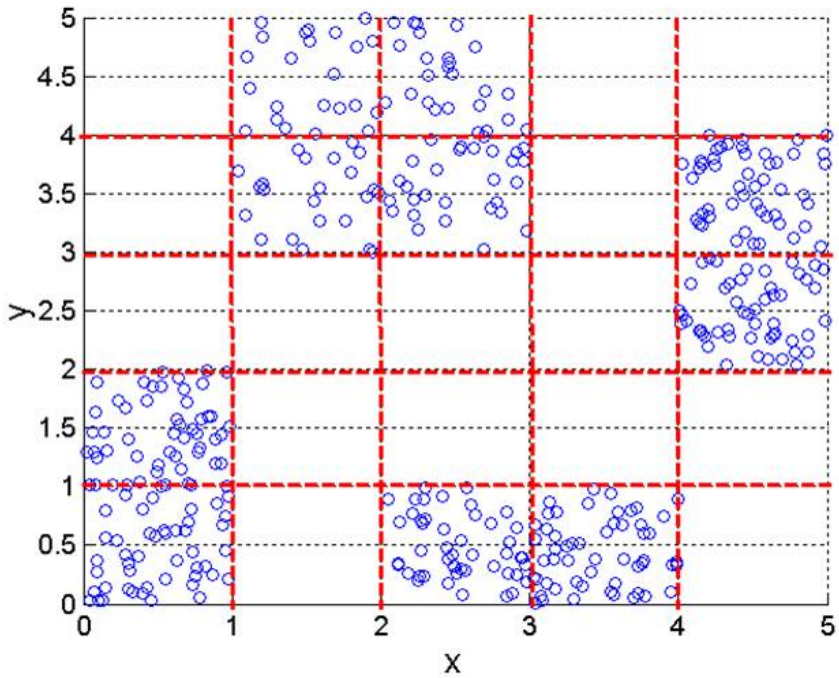
Transformation d'attributs continus (ou ayant une portée infinie) en attributs discrets ayant une portée finie.

La discrétisation peut être effectuée de manière supervisée (p.ex., en utilisant des étiquettes de classe) ou non supervisée (p.ex., en utilisant le regroupement).

Discrétisation supervisée basée sur la minimisation de l'impureté:

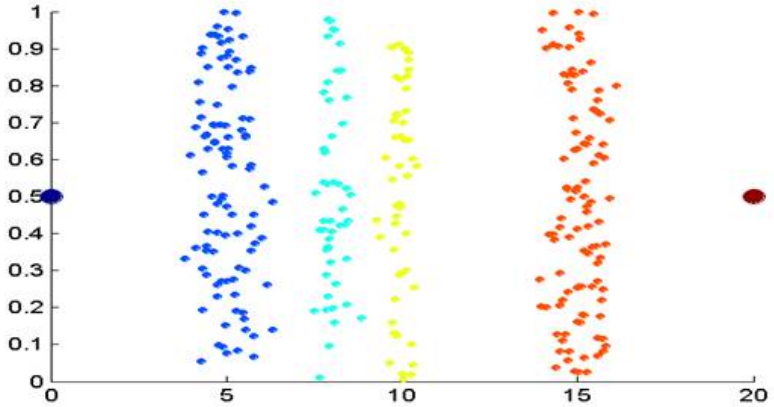


3 values per axis

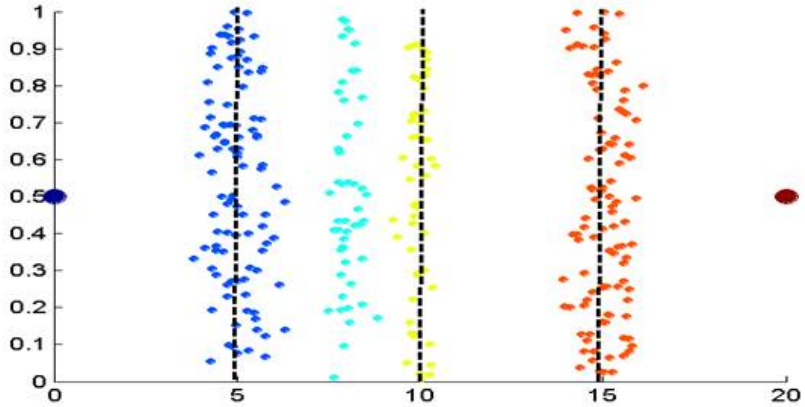


5 values per axis

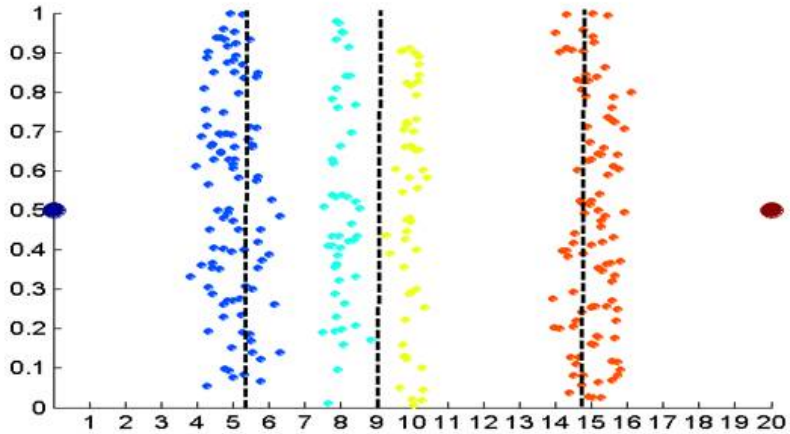
Discrétisation non supervisée:



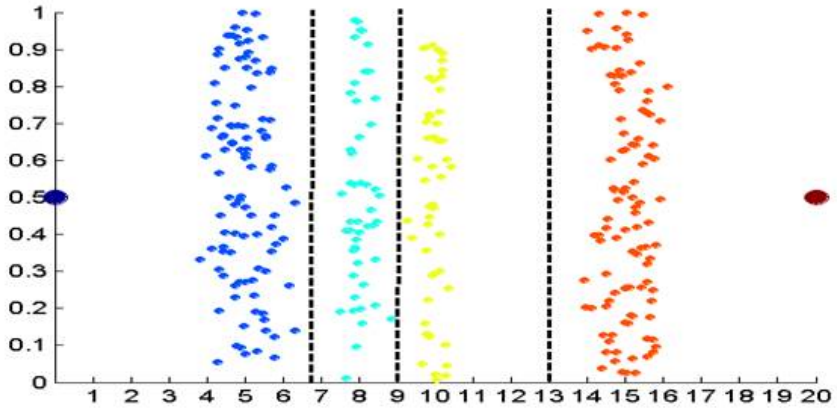
Data



Equal interval width



Equal frequency



K-means

Prétraitement pour simplifier les données

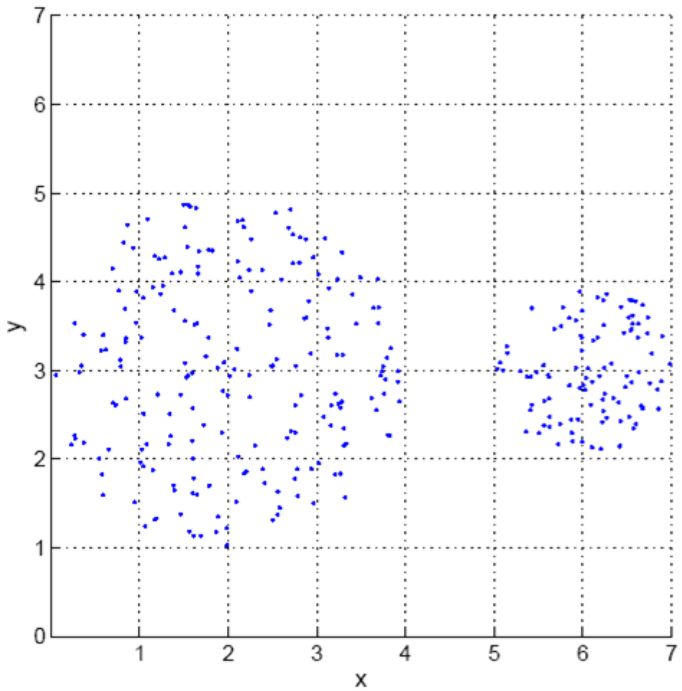
Estimation de la densité

La transformation des attributs des valeurs brutes en densités peut être utilisée pour grossir les données et amener leurs caractéristiques à des échelles comparables entre zéro et un.

Prétraitement pour simplifier les données

Estimation de la densité

La transformation des attributs des valeurs brutes en densités peut être utilisée pour grossir les données et amener leurs caractéristiques à des échelles comparables entre zéro et un.



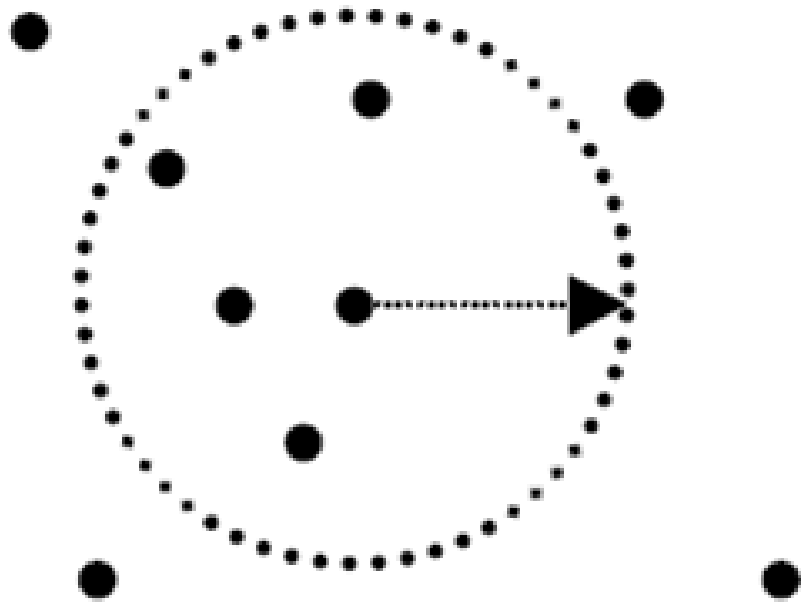
0	0	0	0	0	0	0
0	0	0	0	0	0	0
4	17	18	6	0	0	0
14	14	13	13	0	18	27
11	18	10	21	0	24	31
3	20	14	4	0	0	0
0	0	0	0	0	0	0

Estimation de la densité par des cellules

Prétraitement pour simplifier les données

Estimation de la densité

La transformation des attributs des valeurs brutes en densités peut être utilisée pour grossir les données et amener leurs caractéristiques à des échelles comparables entre zéro et un.



Estimation de la densité par des centres

La dimensionnalité des données est généralement déterminée par le nombre d'attributs qui représentent chaque point de données.

Malédiction de la dimensionnalité

Un terme général pour divers phénomènes qui surviennent lors de l'analyse et du traitement de données en haute dimension.

- Thème commun - la signification statistique est difficile, peu pratique, ou même impossible à obtenir en raison de la rareté des données dans les hautes dimensions
- Provoque de mauvaises performances des méthodes statistiques classiques par rapport aux données de faible dimension

Solution courante - réduire la dimension des données dans le cadre de leur (pré)traitement.

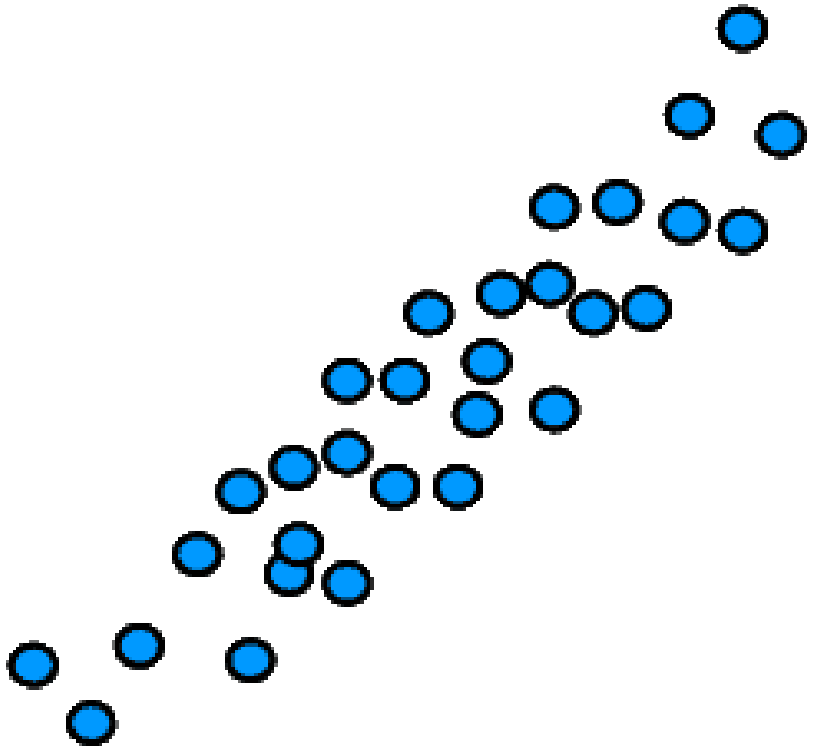
Il existe plusieurs approches pour représenter les données en dimensions inférieures, qui peuvent généralement être divisées en deux types:

Approches de réduction de la dimensionnalité

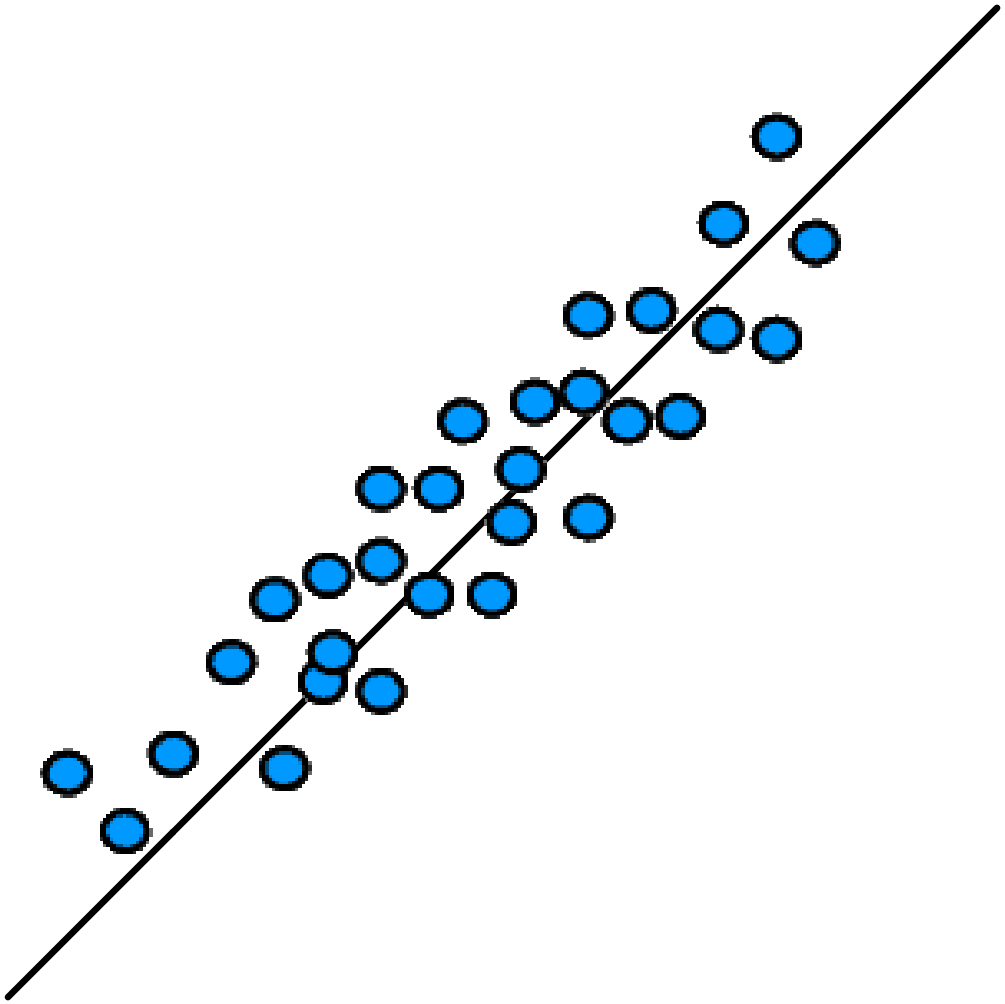
- Sélection/pondération des caractéristiques - sélectionner un sous-ensemble de caractéristiques existantes et ne les utiliser que dans l'analyse, tout en leur attribuant éventuellement des pondérations d'importance pour éliminer les informations redondantes
- Extraction/construction de caractéristiques - créer de nouvelles caractéristiques en extrayant les informations pertinentes des attributs de données

PCA et MDS sont deux des méthodes de réduction de la dimensionnalité les plus courantes dans l'analyse des données, mais il en existe beaucoup d'autres.

Analyse en composantes principales



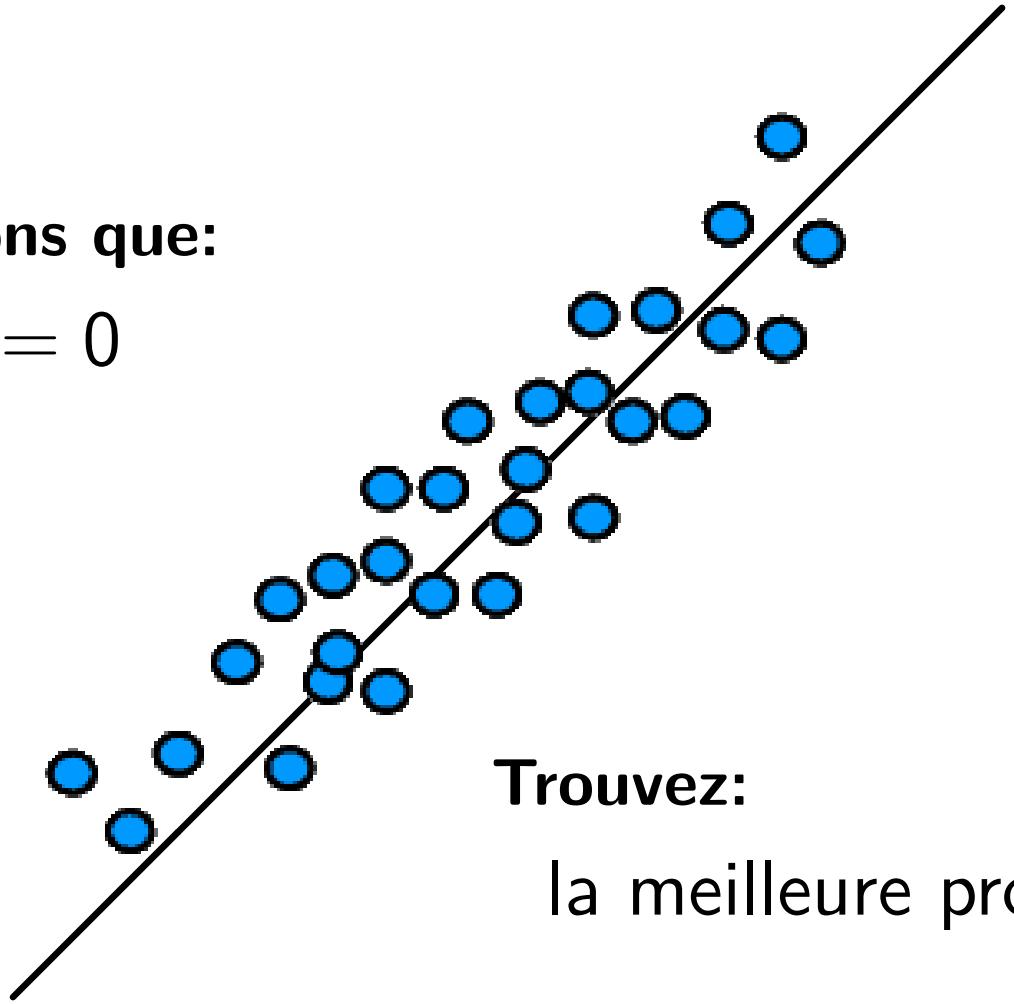
Analyse en composantes principales



Analyse en composantes principales

Supposons que:

$$avg = 0$$

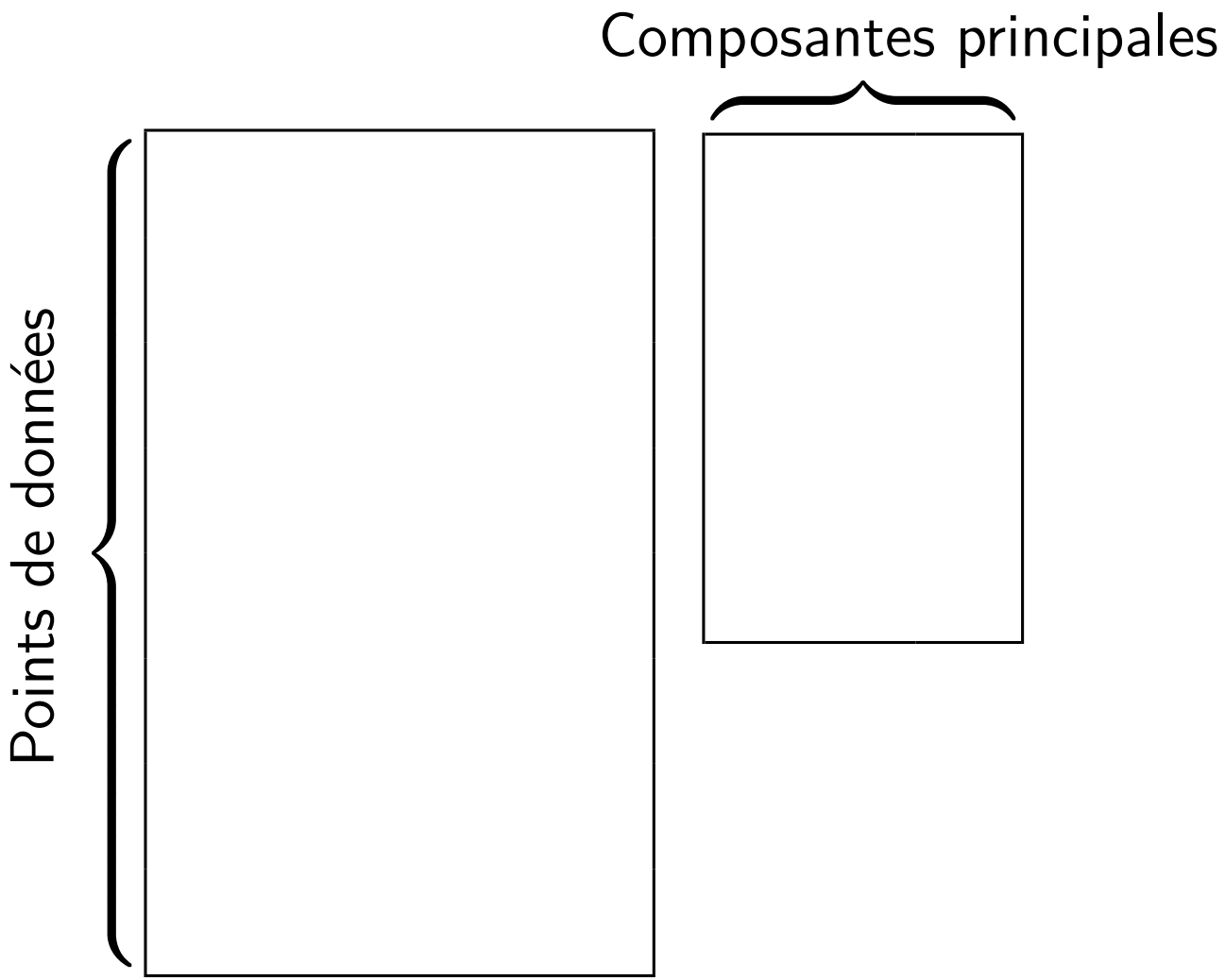


Trouvez:

la meilleure proj. en k-dim.

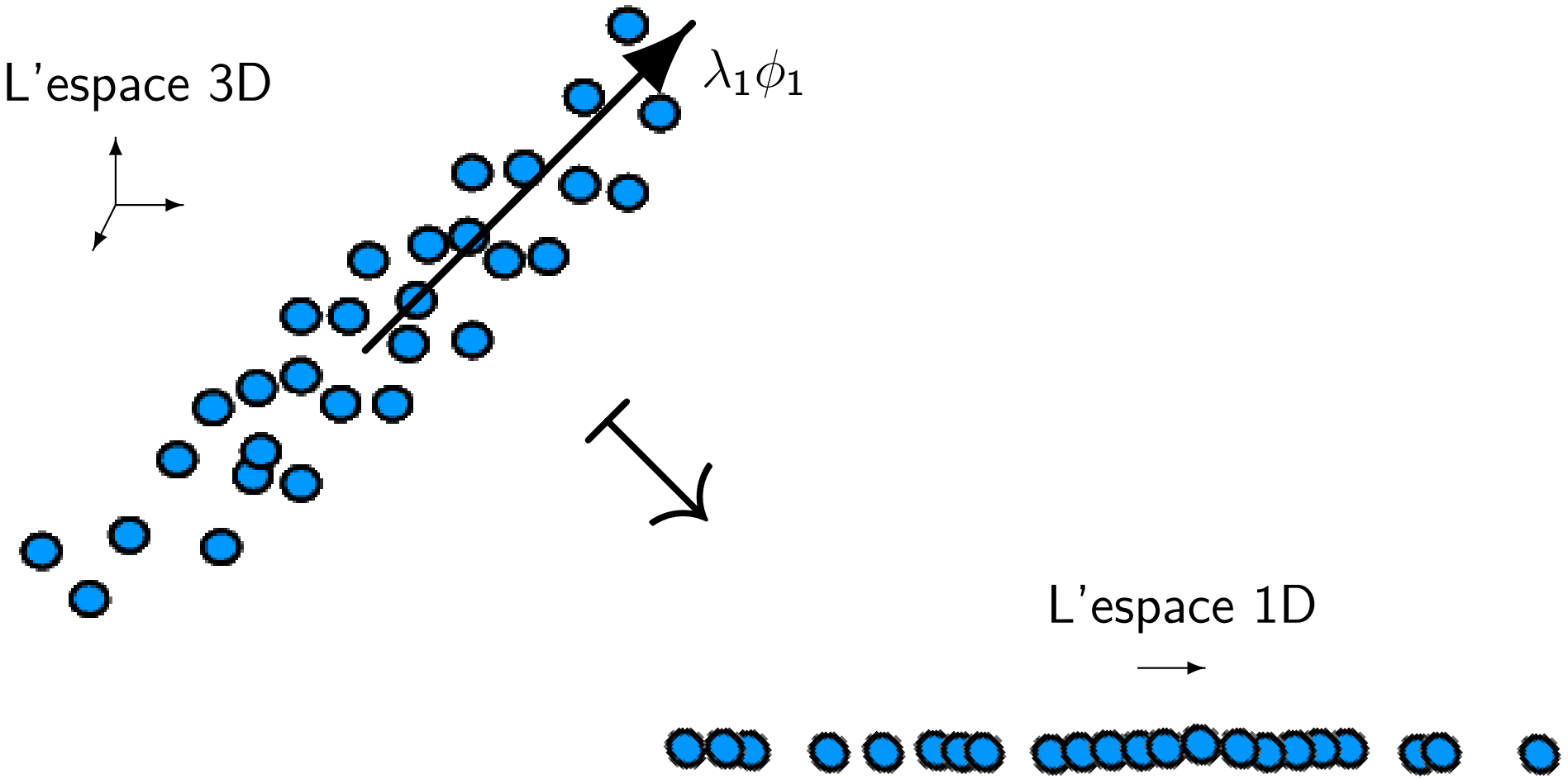
Analyse en composantes principales

Projection sur les composantes principales:



Analyse en composantes principales

Projection sur les composantes principales:



Analyse en composantes principales

Quelle serait la meilleure projection ?

Trouvez un sous-espace $S \subseteq \mathbb{R}^n$ t.q. $\dim(S) = k$ et les données sont bien approximées par $\hat{x} = \text{proj}_S\, x$.

Analyse en composantes principales

Quelle serait la meilleure projection ?

Trouvez un sous-espace $S \subseteq \mathbb{R}^n$ t.q. $\dim(S) = k$ et les données sont bien approximées par $\hat{x} = \text{proj}_S\, x$.

$$\Downarrow$$

Trouvez un sous-espace $S \subseteq \mathbb{R}^n$ t.q. $S = \text{span}\{u_1, \ldots, u_k\}$ et $\|x - \hat{x}\|$ sont minimales pour les points de données x où $\hat{x} = \text{proj}_S\, x$.

Analyse en composantes principales

Quelle serait la meilleure projection ?

Trouvez un sous-espace $S \subseteq \mathbb{R}^n$ t.q. $\dim(S) = k$ et les données sont bien approximées par $\hat{x} = \text{proj}_S\, x$.

$$\Downarrow$$

Trouvez un sous-espace $S \subseteq \mathbb{R}^n$ t.q. $S = \text{span}\{u_1, \ldots, u_k\}$ et $\|x - \hat{x}\|$ sont minimales pour les points de données x où $\hat{x} = \text{proj}_S\, x$.

$$\Downarrow$$

Trouvez k vecteurs $u_1, \ldots, u_k$ t.q. $N^{-1} \sum_{i=1}^{N} \|x_i - \hat{x}_i\|^2$ est minimale avec $\hat{x} = \text{proj}_{\text{span}\{u_1,\ldots,u_k\}}\, x$.

Analyse en composantes principales

Quelle serait la meilleure projection ?

Trouvez un sous-espace $S \subseteq \mathbb{R}^n$ t.q. $\dim(S) = k$ et les données sont bien approximées par $\hat{x} = \text{proj}_S\, x$.

$$\Downarrow$$

Trouvez un sous-espace $S \subseteq \mathbb{R}^n$ t.q. $S = \text{span}\{u_1, \ldots, u_k\}$ et $\|x - \hat{x}\|$ sont minimales pour les points de données x où $\hat{x} = \text{proj}_S\, x$.

$$\Downarrow$$

Trouvez k vecteurs $u_1, \ldots, u_k$ t.q. $N^{-1} \sum_{i=1}^{N} \|x_i - \hat{x}_i\|^2$ est minimale avec $\hat{x} = \text{proj}_{\text{span}\{u_1,\ldots,u_k\}}\, x$.

Comment trouver ces vecteurs $u_1, \ldots, u_k$?

Analyse en composantes principales

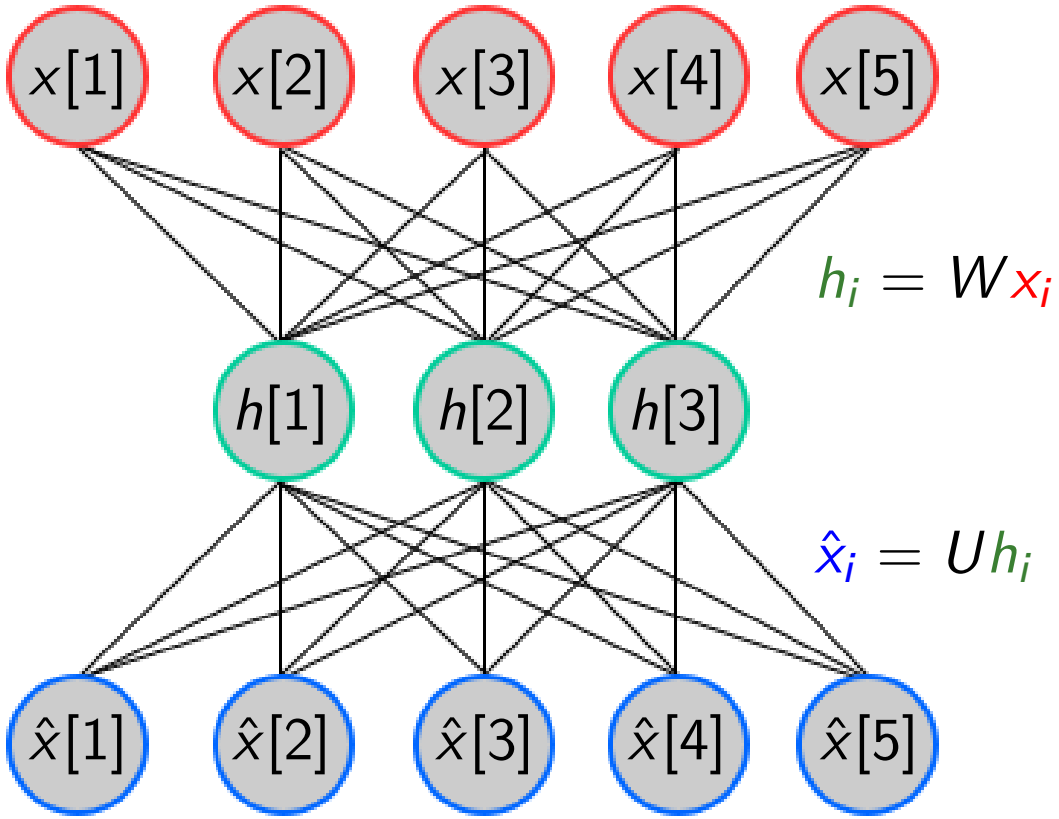
Autoencoder

Minimisez $N^{-1} \sum_{i=1}^{n} \|x_i - \hat{x}_i\|^2$ s.t. $\hat{x} = \text{proj}_{\text{span}\{u_1,\ldots,u_k\}} x$

Couche d'entrée:

Couche cachée:

Couche de sortie:



$h_i = W x_i$

$\hat{x}_i = U h_i$

$$\underset{W \in \mathbb{R}^{k \times n}, U \in \mathbb{R}^{n \times k}}{\arg\min} \sum_{i=1}^{N} \|x_i - UWx_i\|^2$$

Analyse en composantes principales

Minimisation des erreurs de reconstruction

Il suffit de considérer les vecteurs orthonormaux $u_1, \ldots, u_k$ (c-à-d, $\|u_i\| = 1$, $\langle u_i, u_j \rangle = 0$ pour $i \neq j$) qui forment la base du sous-espace. On peut ensuite étendre cet ensemble pour former une base $u_1, \ldots, u_n$ pour la totalité de $\mathbb{R}^n$.

Alors, on peut écrire $x = \sum_{j=1}^{n} \langle x, u_j \rangle u_j = \sum_{j=1}^{n} u_j u_j^T x$ et $\text{proj}_{\text{span}\{u_1,\ldots,u_k\}} x = \sum_{j=1}^{k} u_j u_j^T x$.

On considère maintenant l'erreur de reconstruction $N^{-1} \sum_{i=1}^{N} \|x_i - \hat{x}_i\|^2$.

Analyse en composantes principales

Minimisation des erreurs de reconstruction

D'abord, notez que

$$x - \hat{x} = \sum_{j=1}^{n} u_j u_j^T x - \sum_{j=1}^{k} u_j u_j^T x = \sum_{j=k+1}^{n} u_j u_j^T x$$

$$\Downarrow$$

$$\|x - \hat{x}\|^2 = \sum_{q=1}^{n} \Big(\sum_{j=k+1}^{n} u_j[q] u_j^T x \Big)^2$$

$$= \sum_{j=k+1}^{n} \sum_{j'=k+1}^{n} \Big(\sum_{q=1}^{n} u_j[q] u_{j'}[q] \Big) (u_j^T x)(u_{j'}^T x)$$

$$= \sum_{j=k+1}^{n} (u_j^T x)^2 = \sum_{j=1}^{n} (u_j^T x)^2 - \sum_{j=1}^{k} (u_j^T x)^2 = \|x\|^2 - \|\hat{x}\|^2$$

$$\Downarrow$$

La minimisation d'erreur de reconstruction équivaut à la maximisation de $N^{-1} \sum_{i=1}^{N} \|\hat{x}_i\|^2 = \sum_{j=1}^{k} N^{-1} \sum_{i=1}^{N} (u_j^T x_i)^2 = \sum_{j=1}^{k} \text{variance}(u_j^T x)$

Analyse en composantes principales

Maximisation de variance

Trouvez une direction qui maximise la variance des données projetées.

Analyse en composantes principales

Maximisation de variance

Trouvez une direction qui maximise la variance des données projetées.

$$\Downarrow$$

Trouvez un vecteur unitaire $u \in \mathbb{R}^n$ qui maximise
variance$(u^T x) = u^T \Sigma u$, où Σ est la matrice de covariance.

Analyse en composantes principales

Maximisation de variance

Trouvez une direction qui maximise la variance des données projetées.

$$\Downarrow$$

Trouvez un vecteur unitaire $u \in \mathbb{R}^n$ qui maximise:

$$\begin{aligned}\text{variance}(u^T x) &= N^{-1} \sum_{i=1}^{N} (u^T x_i)^2 = N^{-1} \sum_{i=1}^{N} (u^T x_i)(x_i^T u) \\ &= u^T \left(N^{-1} \sum_{i=1}^{N} x_i x_i^T \right) u = u^T \Sigma u\end{aligned}$$

où Σ est la matrice de covariance.

Analyse en composantes principales

Maximisation de variance

Trouvez une direction qui maximise la variance des données projetées.

$$\Downarrow$$

Trouvez un vecteur unitaire $u \in \mathbb{R}^n$ qui maximise
variance$(u^T x) = u^T \Sigma u$, où Σ est la matrice de covariance.

Analyse en composantes principales

Maximisation de variance

Trouvez une direction qui maximise la variance des données projetées.

$$\Downarrow$$

Trouvez un vecteur unitaire $u \in \mathbb{R}^n$ qui maximise
variance$(u^T x) = u^T \Sigma u$, où Σ est la matrice de covariance.

$$\Downarrow$$

Résoudre le problème de maximisation:

$$\begin{array}{ll} \text{maximize} & u^T \Sigma u \\ \text{s.t.} & \|u\| = 1 \end{array}$$

Résoudre le problème de maximisation:

$$\begin{aligned} \text{maximize} \quad & u^T \Sigma u \\ \text{s.t.} \quad & \|u\| = 1 \end{aligned}$$

On applique la méthode des multiplicateurs de Lagrange:

$$\begin{aligned} f(u, \alpha) &= u^T \Sigma u + \alpha(1 - u^T u) \\ \nabla_u f(u, \alpha) &= 2(\Sigma u - \alpha u) \\ \nabla_u f(u, \alpha) = 0 &\Rightarrow \Sigma u = \alpha u \end{aligned}$$

Par conséquent, u est un vecteur propre de Σ avec une valeur propre de α, qui doit être la valeur propre maximale pour maximiser $u^T \Sigma u = \alpha$.

De même, une deuxième direction est trouvée via:

$$\begin{aligned} \text{maximize} \quad & u_2^T \Sigma u_2 \\ \text{s.t.} \quad & \|u_2\| = 1 \\ & \langle u_2, u_1 \rangle = 0 \end{aligned}$$

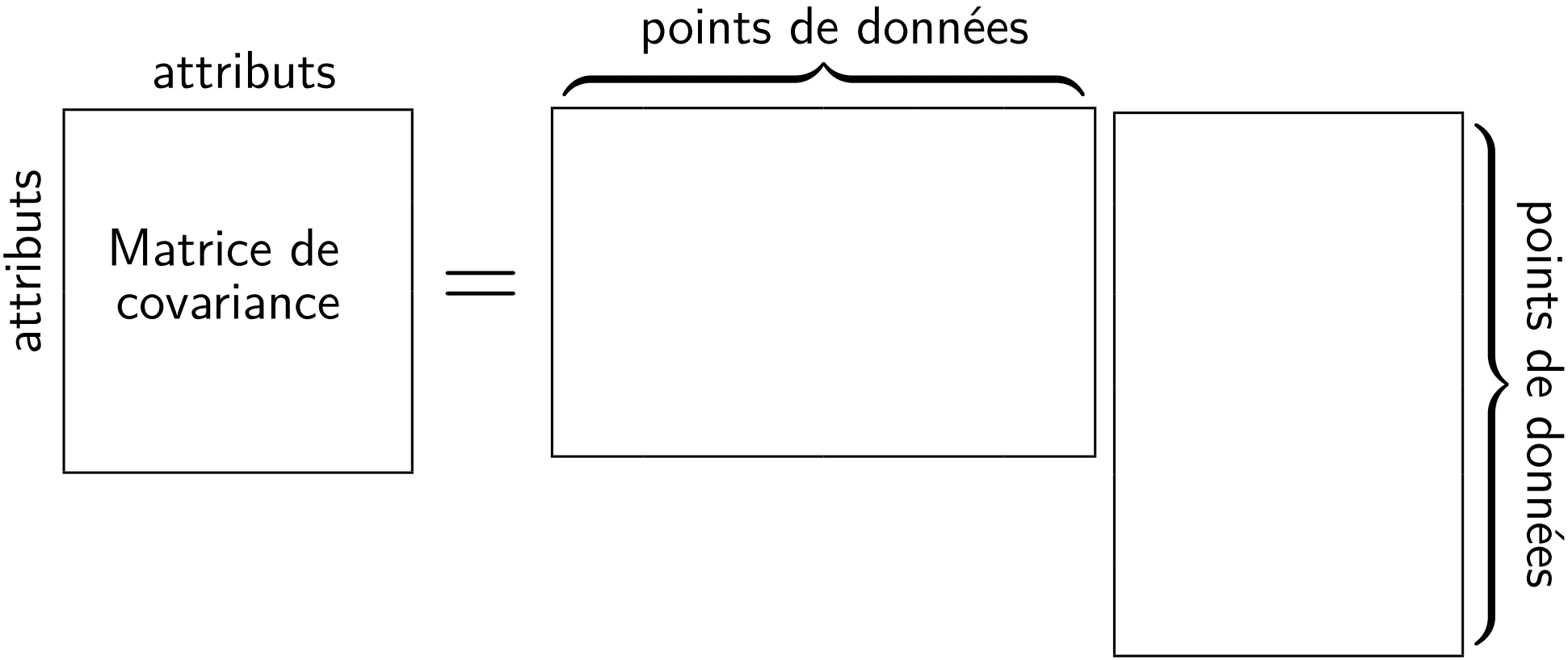
En appliquant les multiplicateurs de Lagrange:

$$\begin{aligned} f(u_2, \alpha, \beta) &= u_2^T \Sigma u_2 + \alpha(1 - u_2^T u_2) - \beta u_2^T u_1 \\ \nabla_{u_2} f(u_2, \alpha, \beta) &= 2(\Sigma u_2 - \alpha u_2) - \beta u_1 \\ \nabla_{u_2} f(u_2, \alpha, \beta) = 0 &\Rightarrow \beta = -\langle u_1, \nabla_{u_2} f(u_2, \alpha, \beta) \rangle = 0 \\ &\Rightarrow \Sigma u_2 = \alpha u_2 \end{aligned}$$

Ainsi, u_2 est un vecteur propre de Σ avec la deuxième plus grande valeur propre.

Analyse en composantes principales

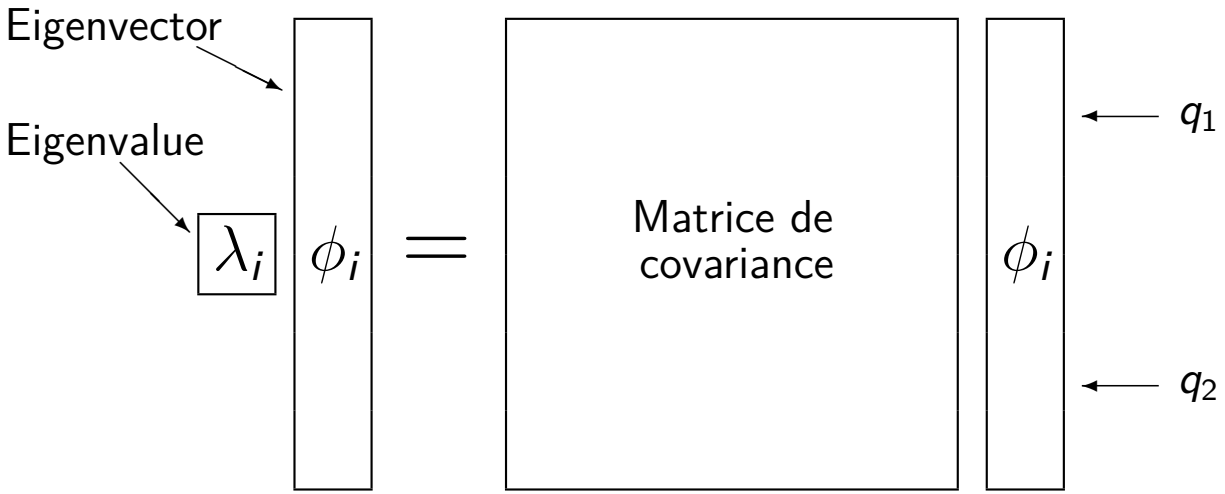
Eigendecomposition et SVD



$$cov(q_1, q_2) \triangleq \sum_i x_i[q_1] \cdot x_i[q_2]$$

Analyse en composantes principales

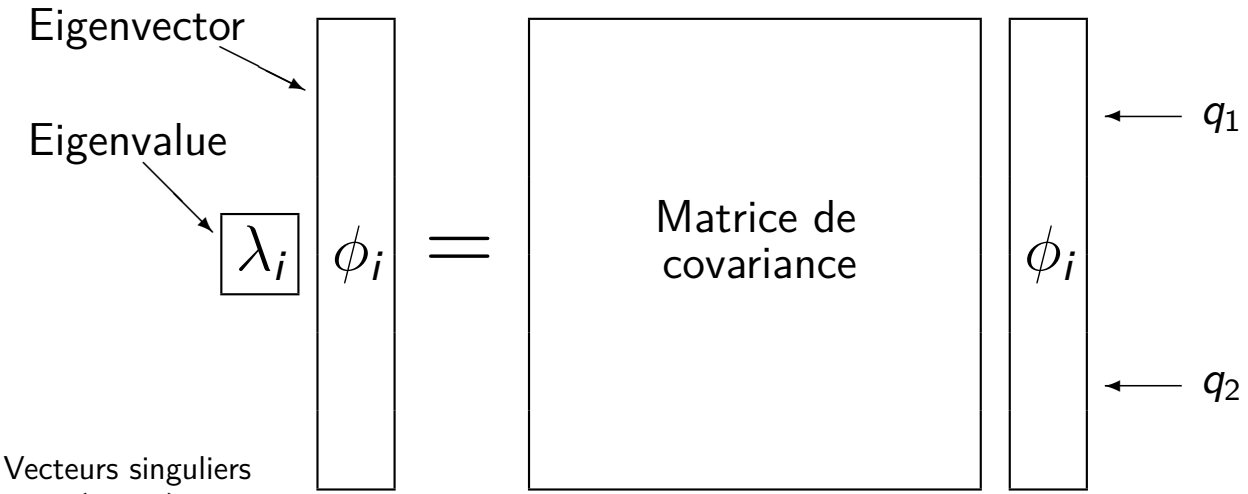
Eigendecomposition et SVD

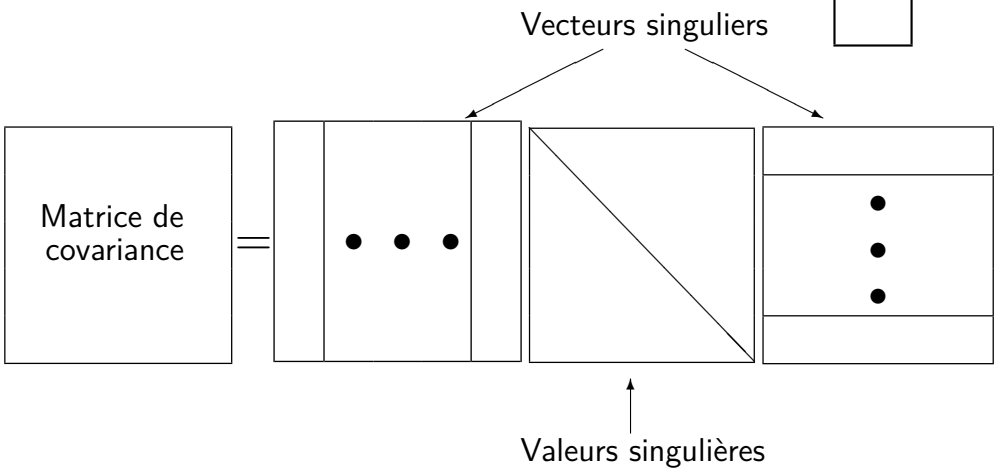


Analyse en composantes principales

Eigendecomposition et SVD

Théorème spectral s'applique au mat. de cov.:





SVD «Singular Value Decomposition»

Théorème spectral: $cov(q_1, q_2) = \sum_i \lambda_i \, \phi_i[q_1] \, \phi_i[q_2]$

Analyse en composantes principales

Décomposition en valeurs singulières

Toute matrice $M \in \mathbb{R}^{n \times k}$ peut être décomposée en $U, S, V \hookleftarrow \text{SVD}(M)$ comme

$$M = U\,S\,V^T$$

U : $n \times n$ orthogonale

S : $n \times k$ diagonale

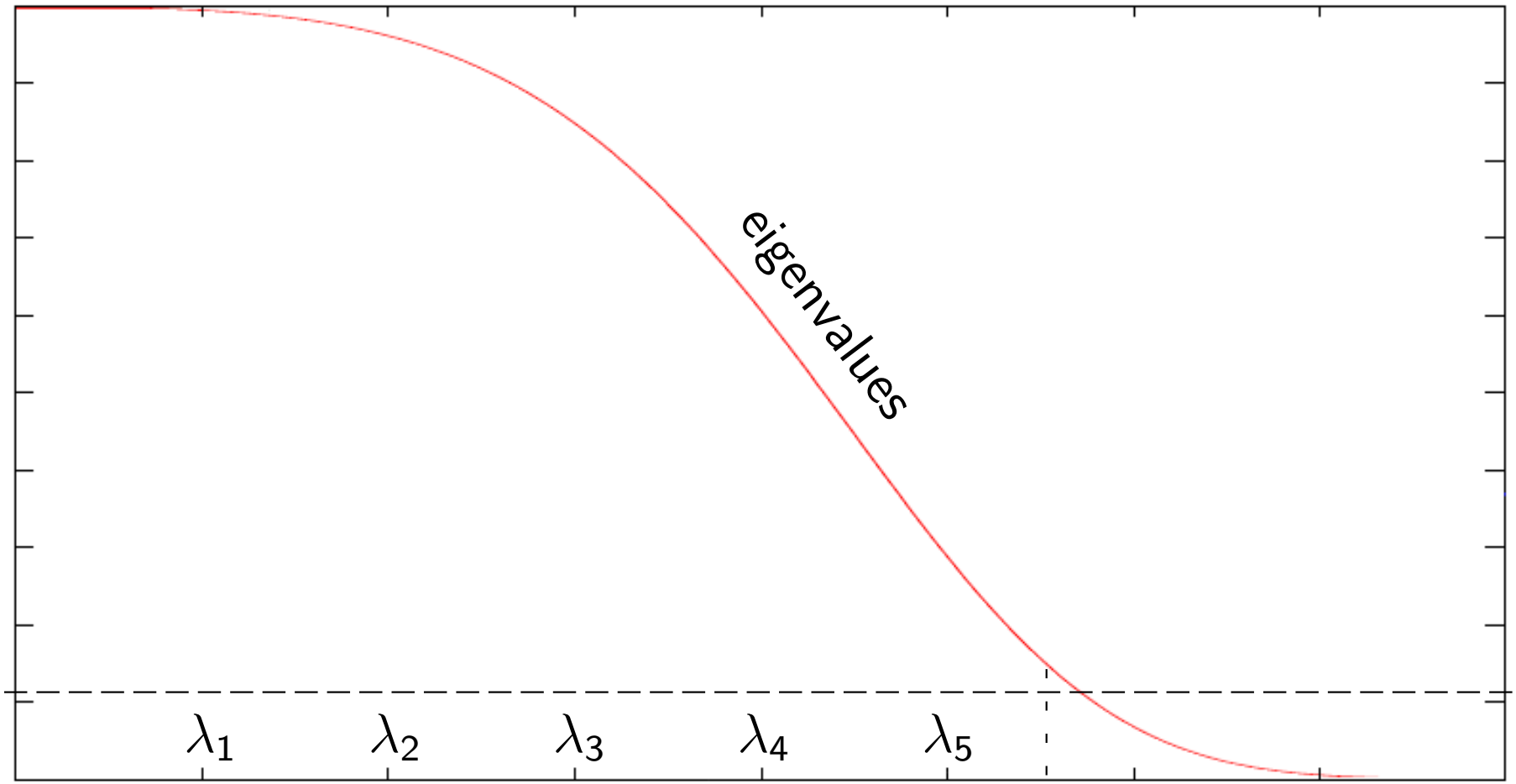
V : $k \times k$ orthogonale

- Les valeurs singulières en S sont la racine carrée des valeurs propres (non négatives) de MM^T et M^TM.
- Les vecteurs singuliers dans (les colonnes de) U sont les vecteurs propres de MM^T.
- Les vecteurs singuliers dans (les colonnes de) V sont les vecteurs propres de M^TM.

Preuve & plus de détails de SVD peuvent être trouvés sur Wikipedia.

Analyse en composantes principales

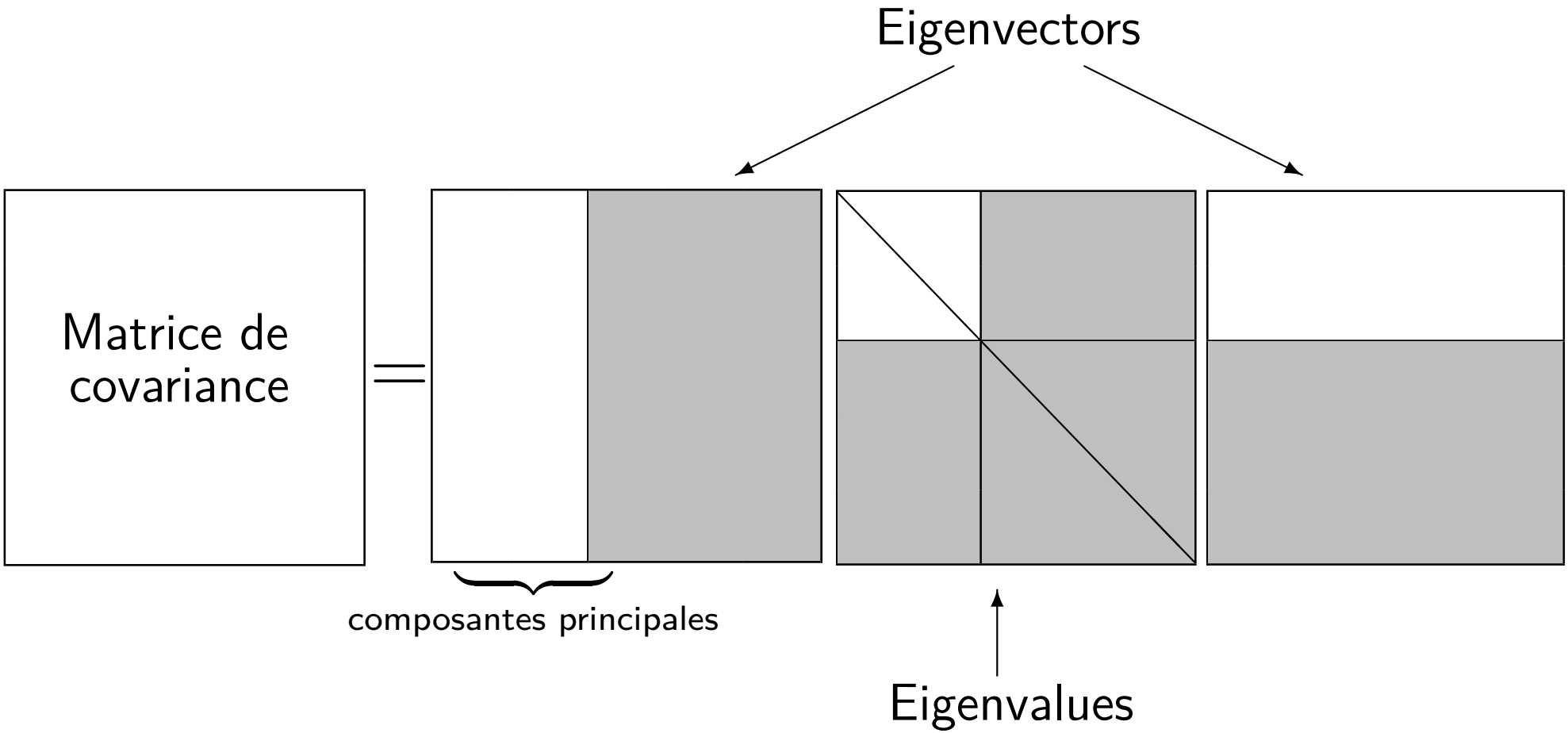
Décomposition en valeurs singulières



Le spectre décroissant de cov. révèle la (faible) dimensionnalité.

Analyse en composantes principales

Décomposition en valeurs singulières



La matrice de covariance peut être approximée par une SVD tronquée.

Analyse en composantes principales

Exemple trivial

Considérons le cas simple de points de données qui sont tous sur la même ligne à haute dimension

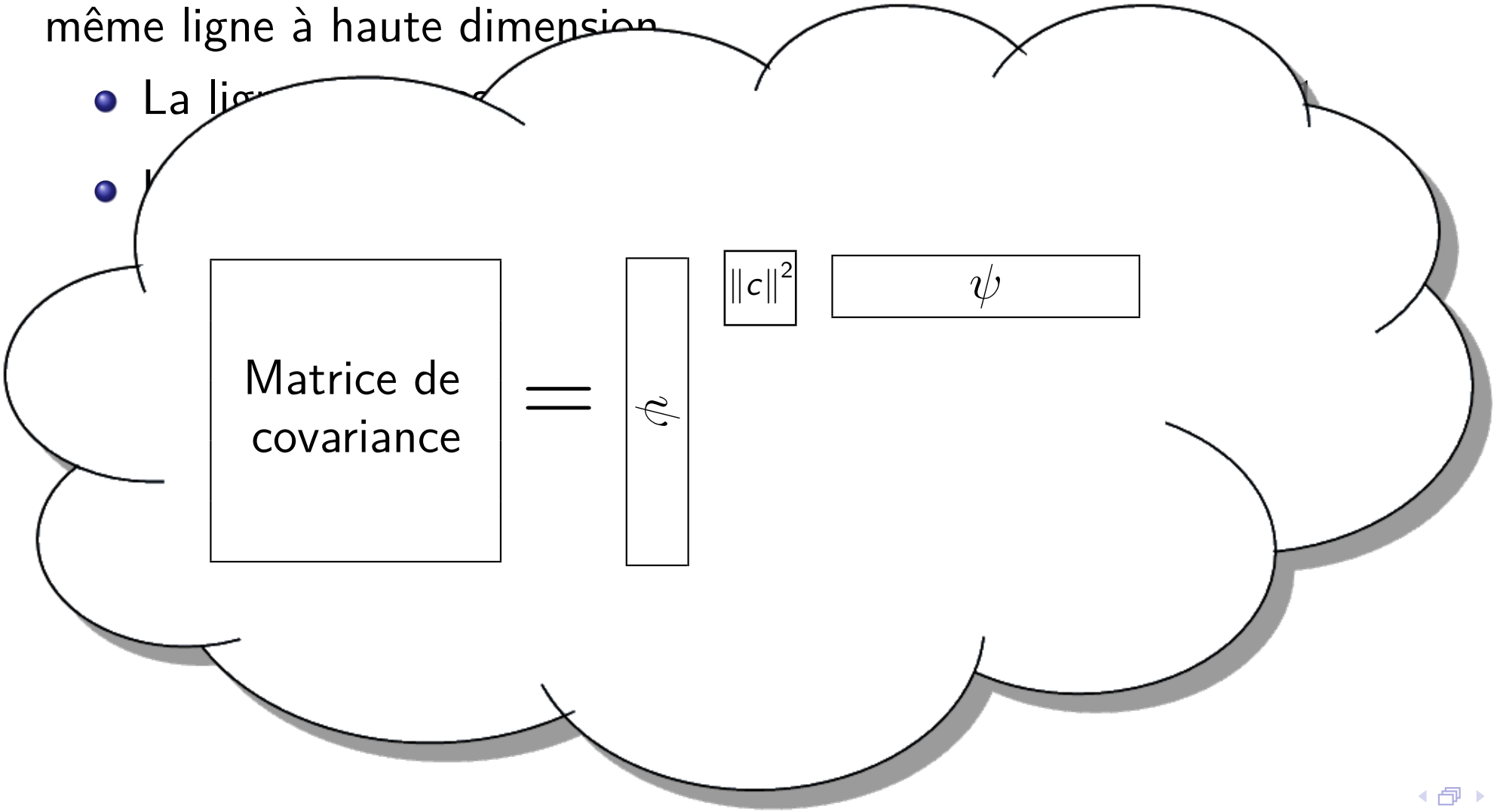
- La ligne droite est définie par un vecteur unitaire $\left\|\vec{\psi}\right\| = 1$
- Les points sur la ligne sont définis en multipliant $\vec{\psi}$ par des scalaires
- Les points peuvent être formulés comme suit : $x_i = c_i\vec{\psi}$
- Covariance: $\text{cov}(t_1, t_2) = \sum_i x_i[t_1]x_i[t_2] = \sum_i c_i\vec{\psi}[t_1]c_i\vec{\psi}[t_2] = (\sum_i c_i^2)\vec{\psi}[t_1]\vec{\psi}[t_2] = \|\vec{c}\|^2\,\vec{\psi}(t_1)\vec{\psi}(t_2) \qquad \vec{c} \triangleq (c_1, c_2, \ldots)$

Analyse en composantes principales

Exemple trivial

Considérons le cas simple de points de données qui sont tous sur la même ligne à haute dimension

- La lig
- I



Analyse en composantes principales

Exemple trivial

Considérons le cas simple de points de données qui sont tous sur la même ligne à haute dimension

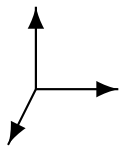
- La ligne droite est définie par un vecteur unitaire $\left\|\vec{\psi}\right\| = 1$
- Les points sur la ligne sont définis en multipliant $\vec{\psi}$ par des scalaires
- Les points peuvent être formulés comme suit : $x_i = c_i\vec{\psi}$
- Covariance: $\text{cov}(t_1, t_2) = \sum_i x_i[t_1]x_i[t_2] = \sum_i c_i\vec{\psi}[t_1]c_i\vec{\psi}[t_2] = (\sum_i c_i^2)\vec{\psi}[t_1]\vec{\psi}[t_2] = \|\vec{c}\|^2 \vec{\psi}(t_1)\vec{\psi}(t_2) \qquad \vec{c} \triangleq (c_1, c_2, \ldots)$

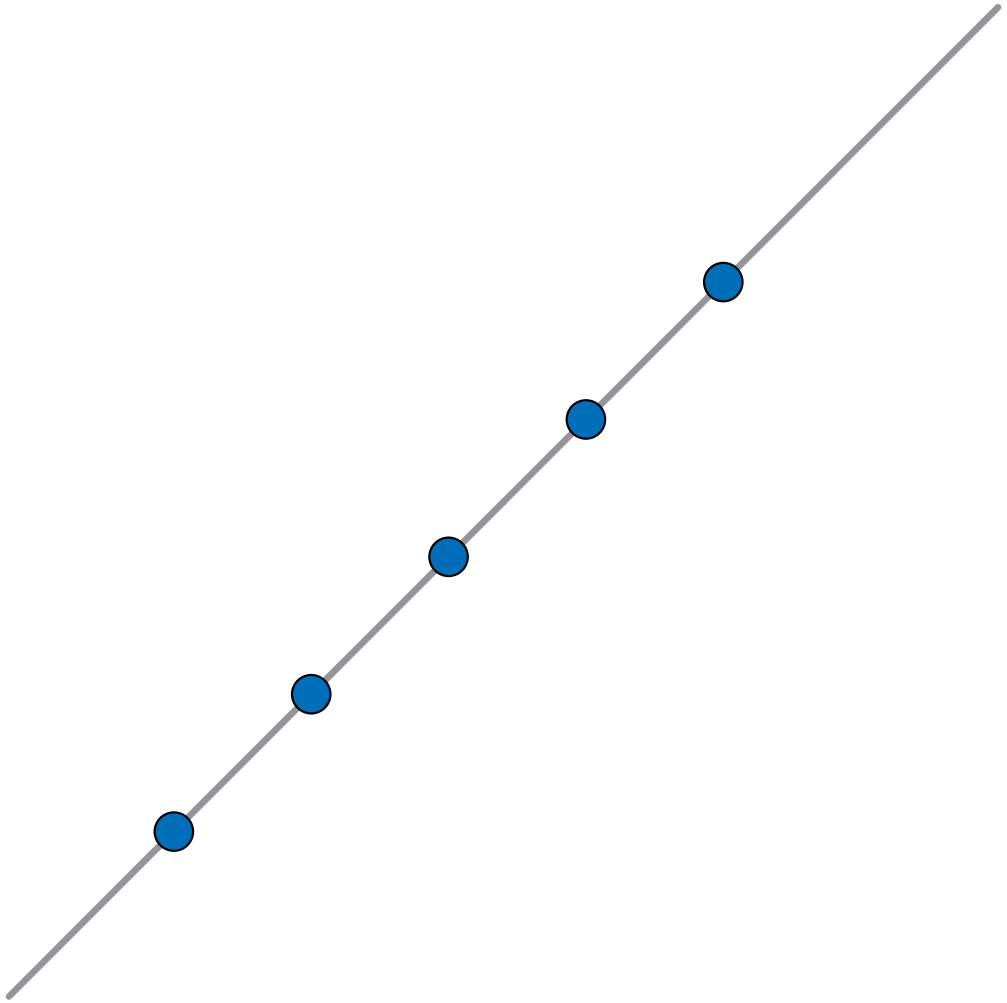
La matrice de covariance a une seule valeur propre $\|\vec{c}\|^2$ et un seul vecteur propre $\vec{\psi}$, qui définit la direction principale des vecteurs de points de données

Analyse en composantes principales

Exemple trivial

3D space

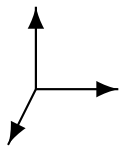


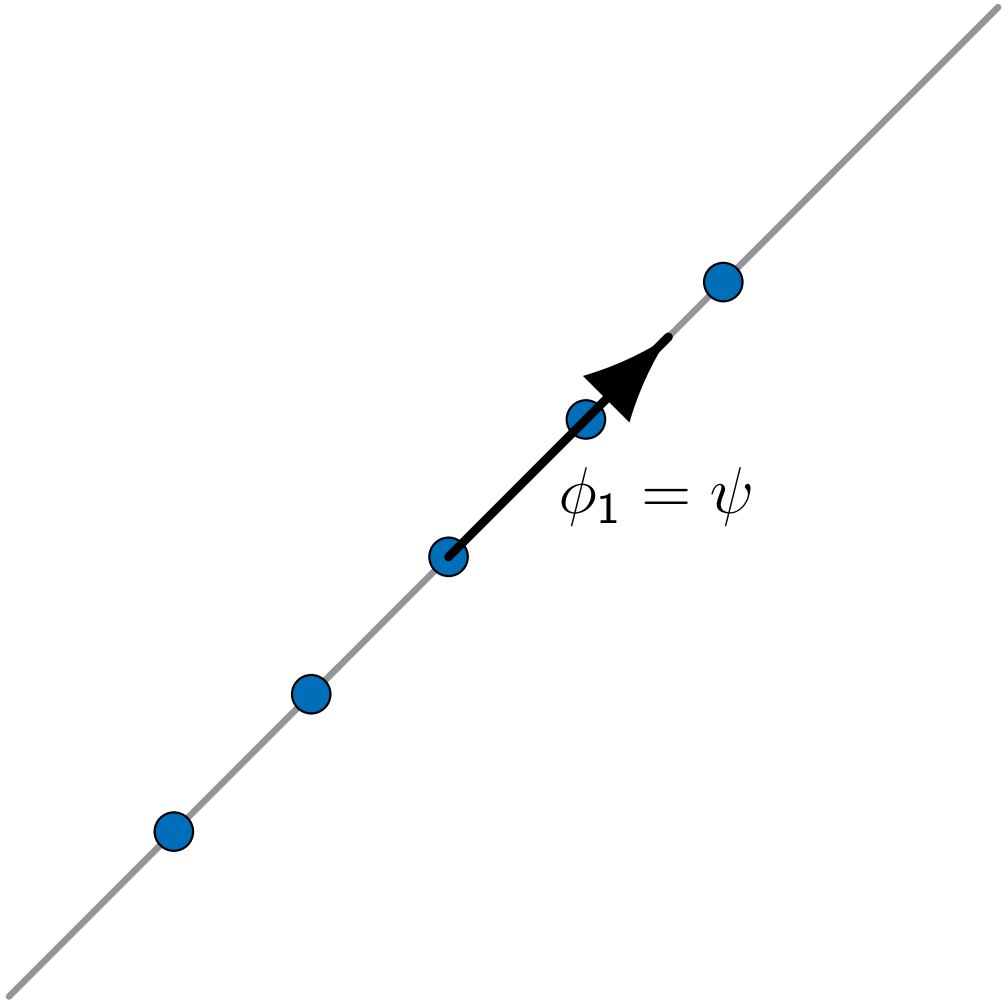


Analyse en composantes principales

Exemple trivial

3D space

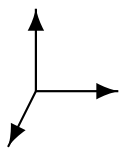


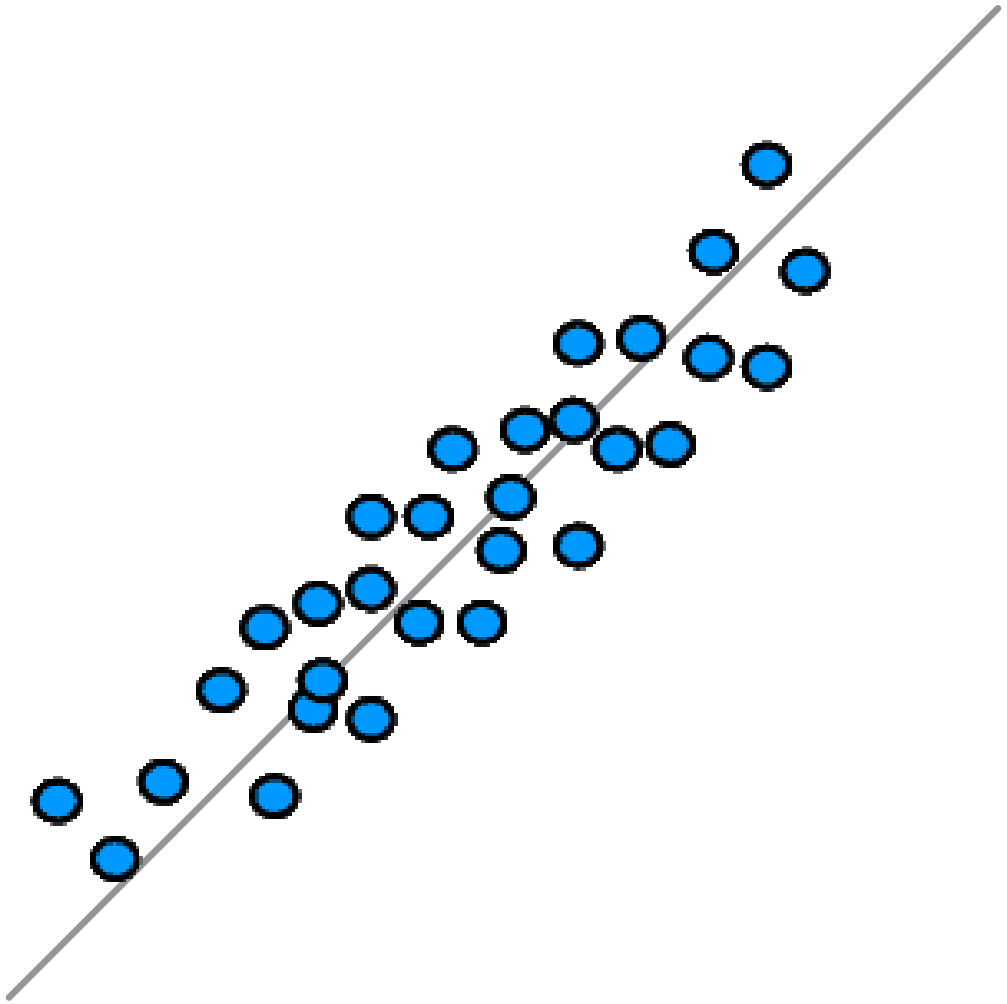


Analyse en composantes principales

Exemple trivial

3D space

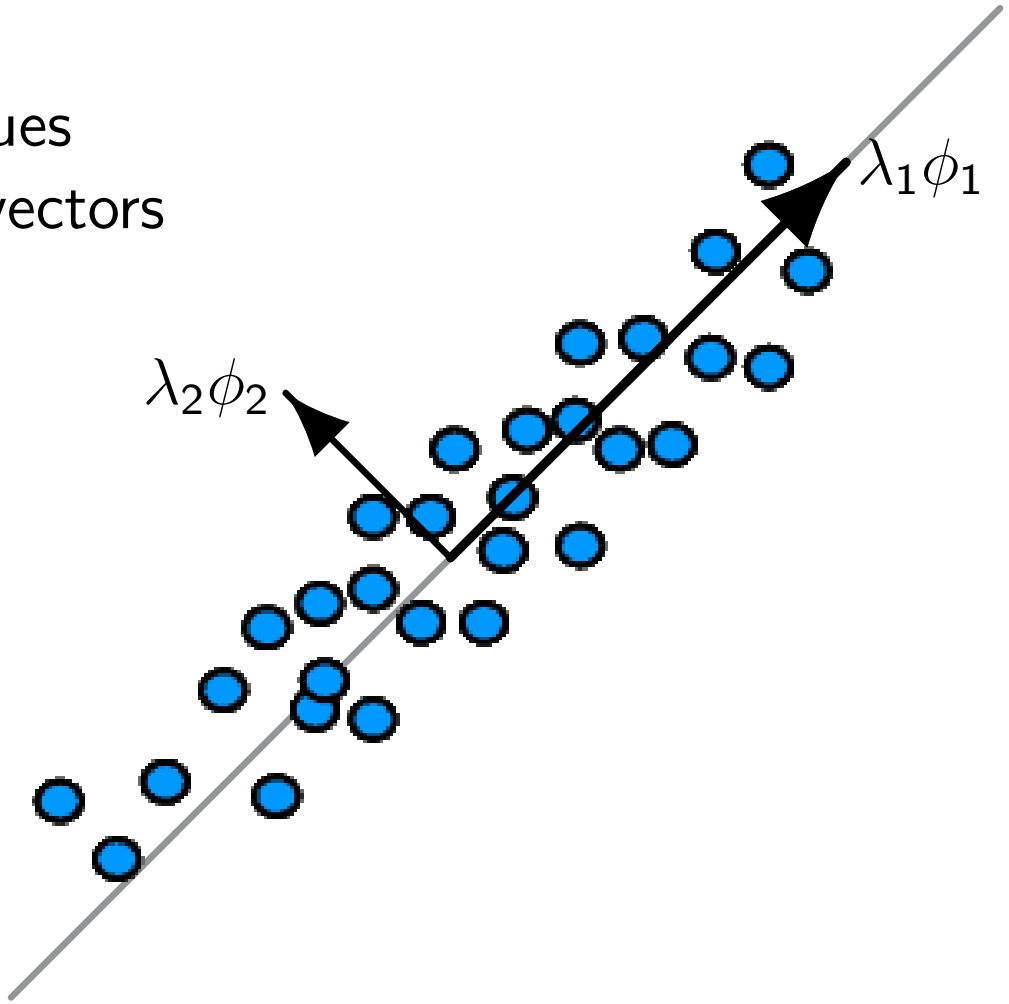




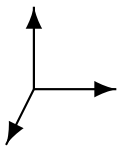
Analyse en composantes principales

Exemple trivial

Length: eigenvalues
Direction: eigenvectors



3D space



composantes principales $\Rightarrow$ directions de var. max.

Analyse en composantes principales

Algorithme de PCA:

1. Centrage
2. Covariance
3. SVD (ou eigendecomposition)
4. Projection

Méthode alternative: Multi-Dimensional Scaling (MDS) - préserver les distances/produits-scalaires avec un ensemble minimal de coordonnées.

Récapitulation

Les étapes de prétraitement sont cruciales pour préparer les données à une analyse significative.

Réduction de la dimensionnalité linéaire pour atténuer la malédiction de la dimensionnalité.

Analyse en composantes principales (PCA) est une approche standard de réduction de la dimensionnalité:

- Basé sur la projection de données sur les principaux vecteurs propres de la matrice de covariance.
- Minimise l'erreur de reconstruction par le projecion,
- De même, trouve un sous-espace qui maximise la variance capturée,
- En pratique, la SVD est utilisée à la place de l'eigendecomposition.