Apprentissage statistique : modélisation descriptive et introduction aux réseaux de neurones (RCP208)

Données manquantes

Michel Crucianu
(prenom.nom@cnam.fr)
http ://cedric.cnam.fr/vertigo/Cours/ml/

Département Informatique
Conservatoire National des Arts & Métiers, Paris, France

17 novembre 2021

Plan du cours

Données manquantes

- Pour certaines observations, les valeurs de certaines variables manquent
  - Ex. sondages : absence de réponse à certaines questions, oublis de transcription, etc.
  - Ex. vidéosurveillance : un capteur fonctionne de façon intermittente, le réseau de transmission présente des pannes, etc.
- Que faire si certaines données manquent ?
  - Solution simple (souvent « par défaut »…) : supprimer les observations qui présentent des valeurs manquantes pour certaines variables

  $\Rightarrow$ au mieux, diminution des performances du modèle ; au pire, fort biais de modélisation et donc modèles inopérants

  …suivant les raisons pour lesquelles des données manquent !

  $\rightarrow$ comprendre pourquoi des données manquent

  $\rightarrow$ estimer les données manquantes, si possible compléter la collecte de données

Nature de l'absence de certaines données

1. Manquant de façon complètement aléatoire (*missing completely at random*, MCAR) : probabilité d'absence identique pour toute observation (cas peu fréquent)
   - par ex. chaque participant à un sondage décide de répondre à la question 1 en lançant un dé et en refusant de répondre si la face 1 apparaît
2. Manquant de façon aléatoire (*missing at random*, MAR) : la probabilité d'absence dépend de variables observées
   - par ex. un participant au sondage a plus de chances de ne pas répondre à la question 2 s'il a donné une certaine réponse (bien enregistrée) à la question 1
3. Manquant de façon non aléatoire (*missing not at random*, MNAR) : la probabilité d'absence dépend de variables non observées
   - Dépendance de variables non observés : par ex. la non réponse à une question dépend de la catégorie socioprofessionnelle et le sondage n'inclut aucune question sur la catégorie socioprofessionnelle (ou des variables permettant de la prédire)
   - Dépendance de la variable à valeurs manquantes : par ex., une des réponses à une question du sondage est souvent évitée (car non assumée)

Nature de l'absence de certaines données (2)

- Il n'est pas possible de savoir, uniquement à partir des données, si des données manquent de façon aléatoire ou non (MAR ou MNAR)
  - → Chercher d'autres travaux sur le même problème (caractérisé par les données manquantes) qui peuvent informer sur ce mécanisme
  - → Inclure dans l'étude un maximum de variables qui pourraient expliquer les données manquantes, afin de rendre MAR bien plus probable (dans les variables observées, on en trouve qui expliquent les variables à données manquantes) que MNAR
- Lorsque les données manquent de façon non aléatoire (cas MNAR), un modèle pour ce manque est indisponible
  - → Possible de compléter la collecte (observer de nouvelles variables, explicatives pour le manque de données) afin de passer de MNAR à MAR ?

Caractérisation de l'absence de données

- Données : tableau de n observations décrites par d variables ($\rightarrow$ matrice $\mathbf{X}$)

Observation	X_1	X_2	...	X_d
o_1	...	...	...	...
...	...	...	...	...
o_n	...	...	...	...

- Matrice indicatrice des données manquantes : $\mathbf{M} = (m_{ij})$, $m_{ij} = 1$ si x_{ij} est manquante et $m_{ij} = 0$ sinon
- On note par $\mathbf{X}_o$ les données observées et par $\mathbf{X}_m$ les données manquantes
- Il est alors possible de caractériser les différentes situations mentionnées :
  1. MCAR : $p(\mathbf{M}|\mathbf{X}) = p(\mathbf{M})$ (indépendance des données)
  2. MAR : $p(\mathbf{M}|\mathbf{X}) = p(\mathbf{M}|\mathbf{X}_o)$
  3. MNAR : pas de réduction possible ou $p(\mathbf{M}|\mathbf{X})$ ne donne pas d'indication utile

Plan du cours

Solution 1 : suppression des observations à données manquantes

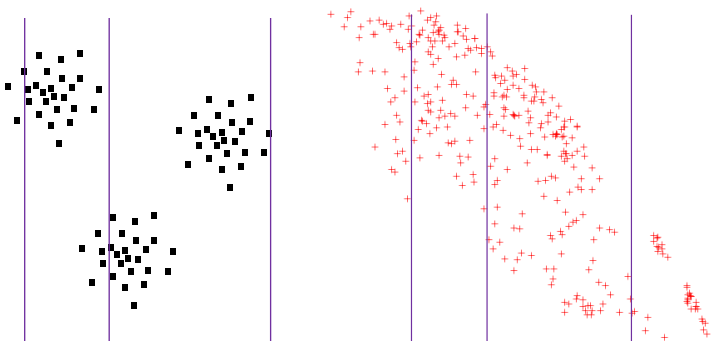
- Solution souvent employée par défaut
- Quel est son impact suivant la nature de l'absence de données ?
  - MCAR :
    - Ignorer les observations à données manquantes = conserver un échantillon aléatoire des observations
    - Ne biaise pas les résultats, mais peut réduire la qualité du modèle résultant (surtout en présence de classes déséquilibrées) s'il y a beaucoup d'observations à données manquantes
  - MAR :
    - Ignorer les observations à données manquantes biaise les résultats
    - Préférable de compléter les observations à données manquantes (imputer les données manquantes)
  - MNAR :
    - Ignorer les observations à données manquantes biaise les résultats
    - L'imputation des données manquantes est plus difficile à justifier, surtout à partir des autres variables observées, vu que l'absence dépend de variable(s) non observée(s)

Solution 2 : imputation des données manquantes

- Compléter les observations à données manquantes peut améliorer la précision du modèle résultant ou réduire son biais
- Nécessaire de comparer le modèle obtenu par imputation avec celui obtenu (si cela est possible) par suppression des observations à données manquantes
- Diverses méthodes d'imputation :
  1. Par une valeur unique (moyenne, médiane, etc.)
  2. Par le centre du groupe
  3. A partir des k plus proches voisins
  4. Par une moyenne partielle
  5. Par décomposition en valeurs singulières
  6. Autres méthodes

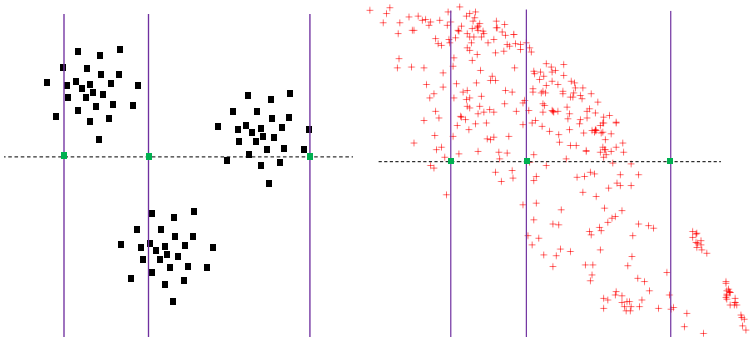
Imputation par une valeur unique

- Solution souvent employée par défaut, parfois dans la lecture même des fichiers de données
- La valeur utilisée peut être
  - Indépendante des données, par ex. 0 : à éviter, sauf si la variable correspondante est de moyenne nulle (voir cas suivant)
  - Représentative de la distribution de la variable concernée :
    - La valeur la plus fréquente : si la variable (quantitative ou nominale) peut prendre peu de valeurs différentes
    - La moyenne : peu robuste à la présence de quelques valeurs extrêmes
    - La médiane : meilleure robustesse



Imputation par une valeur unique

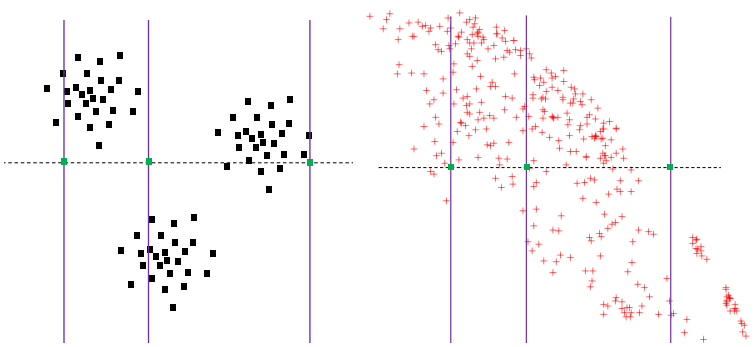
- Solution souvent employée par défaut, parfois dans la lecture même des fichiers de données
- La valeur utilisée peut être
  - Indépendante des données, par ex. 0 : à éviter, sauf si la variable correspondante est de moyenne nulle (voir cas suivant)
  - Représentative de la distribution de la variable concernée :
    - La valeur la plus fréquente : si la variable (quantitative ou nominale) peut prendre peu de valeurs différentes
    - La moyenne : peu robuste à la présence de quelques valeurs extrêmes
    - La médiane : meilleure robustesse



Solutions | Imputation des données manquantes 8 / 16

Imputation par une valeur unique

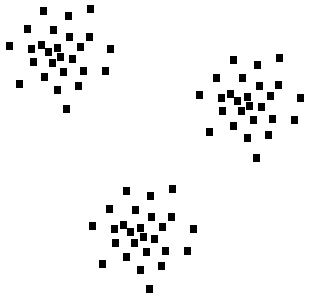
- Solution souvent employée par défaut, parfois dans la lecture même des fichiers de données
- La valeur utilisée peut être
  - Indépendante des données, par ex. 0 : à éviter, sauf si la variable correspondante est de moyenne nulle (voir cas suivant)
  - Représentative de la distribution de la variable concernée :
    - La valeur la plus fréquente : si la variable (quantitative ou nominale) peut prendre peu de valeurs différentes
    - La moyenne : peu robuste à la présence de quelques valeurs extrêmes
    - La médiane : meilleure robustesse



- Noter que remplacer les valeurs manquantes d'une variable par une valeur unique peut réduire très fortement la variance estimée pour cette variable

Imputation par le centre du groupe

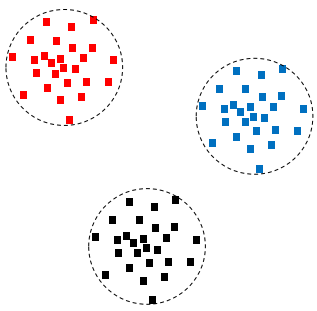
- Hypothèse : des regroupements naturels sont présents dans les données
  → utiliser les centres des groupes pour imputer les valeurs manquantes
- Comment procéder :



Solutions Imputation des données manquantes 9 / 16

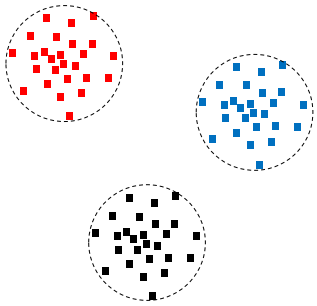
Imputation par le centre du groupe

- Hypothèse : des regroupements naturels sont présents dans les données
  → utiliser les centres des groupes pour imputer les valeurs manquantes
- Comment procéder :
  1. Appliquer un algorithme de classification automatique aux observations complètes (sans données manquantes)



Imputation par le centre du groupe

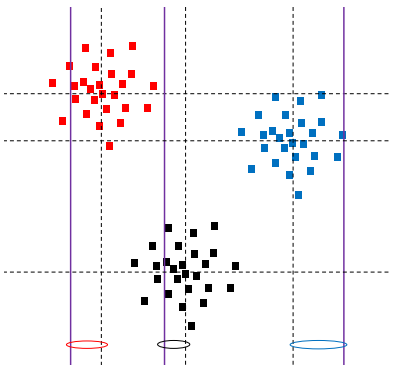
- Hypothèse : des regroupements naturels sont présents dans les données
  → utiliser les centres des groupes pour imputer les valeurs manquantes
- Comment procéder :
  1. Appliquer un algorithme de classification automatique aux observations complètes (sans données manquantes)
  2. Pour chaque observation à données manquantes



Solutions | Imputation des données manquantes 9 / 16

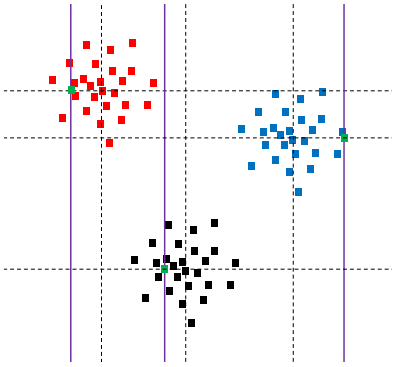
Imputation par le centre du groupe

- Hypothèse : des regroupements naturels sont présents dans les données
  → utiliser les centres des groupes pour imputer les valeurs manquantes
- Comment procéder :
  1. Appliquer un algorithme de classification automatique aux observations complètes (sans données manquantes)
  2. Pour chaque observation à données manquantes
     1. Calculer sa distance (partielle, utilisant seulement les variables renseignées) au centre de chaque groupe



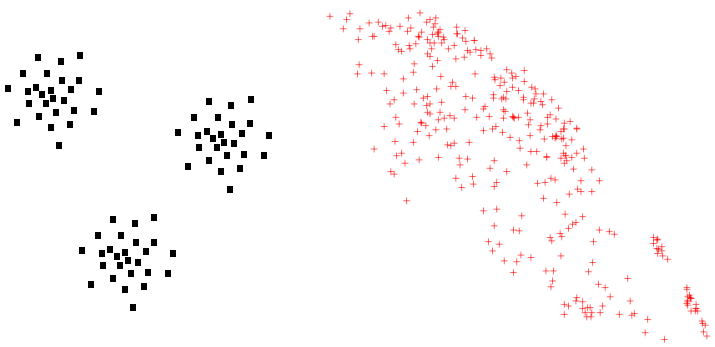
Imputation par le centre du groupe

- Hypothèse : des regroupements naturels sont présents dans les données
  → utiliser les centres des groupes pour imputer les valeurs manquantes
- Comment procéder :
  1. Appliquer un algorithme de classification automatique aux observations complètes (sans données manquantes)
  2. Pour chaque observation à données manquantes
     1. Calculer sa distance (partielle, utilisant seulement les variables renseignées) au centre de chaque groupe
     2. Donner à chaque variable non renseignée la valeur que prend la même variable pour le centre le plus proche



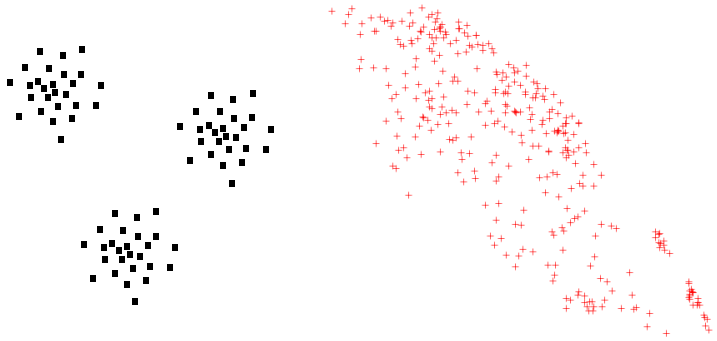
Imputation à partir des k plus proches voisins

- Hypothèse : pour une observation à données manquantes, les observations complètes les plus proches (distances utilisant seulement les variables renseignées) sont plus représentatives que les autres
- Comment procéder :



Imputation à partir des k plus proches voisins

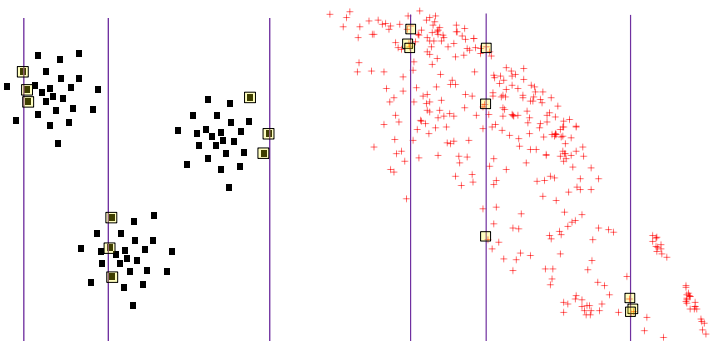
- Hypothèse : pour une observation à données manquantes, les observations complètes les plus proches (distances utilisant seulement les variables renseignées) sont plus représentatives que les autres
- Comment procéder :
  - Pour chaque observation à données manquantes



Solutions | Imputation des données manquantes 10 / 16

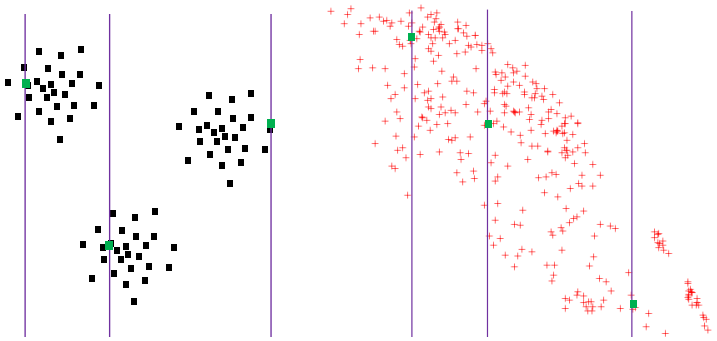
Imputation à partir des k plus proches voisins

- Hypothèse : pour une observation à données manquantes, les observations complètes les plus proches (distances utilisant seulement les variables renseignées) sont plus représentatives que les autres
- Comment procéder :
  - Pour chaque observation à données manquantes
    1. Trouver ses k plus proches voisins complets (distances utilisant seulement les variables renseignées !)



Imputation à partir des k plus proches voisins

- Hypothèse : pour une observation à données manquantes, les observations complètes les plus proches (distances utilisant seulement les variables renseignées) sont plus représentatives que les autres
- Comment procéder :
  - Pour chaque observation à données manquantes
    1. Trouver ses k plus proches voisins complets (distances utilisant seulement les variables renseignées !)
    2. Donner à chaque variable non renseignée la moyenne des valeurs que prend la même variable pour ces k voisins



Imputation par une moyenne partielle

- Considérons le cas de données décrites aussi par une variable nominale « classe »
- Pour une observation appartenant à une classe, estimer la valeur manquante d'une variable à partir de la moyenne (ou de la médiane) de cette variable limitée à la classe

  → utilisable seulement si la classe est connue (renseignée) pour l'observation
- L'imputation par le centre du groupe et l'imputation à partir des k plus proches voisins sont aussi des cas particuliers de cette approche
- Les moyennes employées sont en général non pondérées, mais des pondérations inversement proportionnelles aux distances peuvent éventuellement être employées

Imputation par décomposition en valeurs singulières

- Hypothèse : une décomposition en valeurs singulières (SVD) avec une réduction de rang fournit une bonne approximation des données
- Soit $\mathbf{X}$ la matrice $n \times d$ de données, $\mathbf{X}^c$ sa restriction aux observations (lignes) sans données manquantes et $\mathbf{X}^m$ la restriction aux lignes avec données manquantes
- Rappel décomposition en valeurs singulières, de rang réduit k, appliquée à $\mathbf{X}^c$: $\mathbf{X}^c_k = \mathbf{U}_k \mathbf{D}_k \mathbf{V}_k^T$
  - $\mathbf{U}_k$ $(n \times k)$ a pour colonnes les vecteurs propres de $\mathbf{X}^c(\mathbf{X}^c)^T$ correspondant aux k plus grandes valeurs propres
  - $\mathbf{D}_k$ $(k \times k)$ est diagonale avec sur la diagonale ces k valeurs propres en ordre décroissant
  - $\mathbf{V}_k$ $(d \times k)$ a pour colonnes les vecteurs propres de $(\mathbf{X}^c)^T\mathbf{X}^c$ correspondant aux k plus grandes valeurs propres
- $\mathbf{X}^c_k$ est la meilleure approximation de rang k de $\mathbf{X}^c$ (au sens des moindres carrés : $\mathbf{X}^c_k = \arg\min_{\mathbf{A} \text{ de rang } k} \|\mathbf{X}^c - \mathbf{A}\|^2$)

Imputation par décomposition en valeurs singulières (2)

- Comment procéder (voir par ex. [1]) :
  1. Appliquer la SVD de rang $k < d$ à $\mathbf{X}^c \Rightarrow \mathbf{X}^c_k = \mathbf{U}_k \mathbf{D}_k \mathbf{V}_k^T$

Imputation par décomposition en valeurs singulières (2)

- Comment procéder (voir par ex. [1]) :
  1. Appliquer la SVD de rang $k < d$ à $\mathbf{X}^c \Rightarrow \mathbf{X}^c_k = \mathbf{U}_k \mathbf{D}_k \mathbf{V}_k^T$
  2. Pour chaque observation $\mathbf{x}^m$ (transposée d'une ligne de $\mathbf{X}^m$) à données manquantes

Solutions | Imputation des données manquantes | 13 / 16

Imputation par décomposition en valeurs singulières (2)

- Comment procéder (voir par ex. [1]) :
  1. Appliquer la SVD de rang $k < d$ à $\mathbf{X}^c \Rightarrow \mathbf{X}^c_k = \mathbf{U}_k \mathbf{D}_k \mathbf{V}_k^T$
  2. Pour chaque observation $\mathbf{x}^m$ (transposée d'une ligne de $\mathbf{X}^m$) à données manquantes
     1. Soit $\mathbf{V}_k^*$ la matrice réduite obtenue de $\mathbf{V}_k$ en éliminant les lignes d'indices correspondant aux données manquantes de $\mathbf{x}^m$ et $\mathbf{V}_k^{(*)}$ ce qui reste de $\mathbf{V}_k$ une fois $\mathbf{V}_k^*$ extrait
     2. Soit $\mathbf{x}^{m*}$ le vecteur réduit obtenu de $\mathbf{x}^m$ en éliminant les données manquantes et $\mathbf{x}^{m(*)}$ la partie de $\mathbf{x}^m$ correspondant aux données manquantes

Imputation par décomposition en valeurs singulières (2)

- Comment procéder (voir par ex. [1]) :
  1. Appliquer la SVD de rang $k < d$ à $\mathbf{X}^c \Rightarrow \mathbf{X}_k^c = \mathbf{U}_k \mathbf{D}_k \mathbf{V}_k^T$
  2. Pour chaque observation $\mathbf{x}^m$ (transposée d'une ligne de $\mathbf{X}^m$) à données manquantes
     1. Soit $\mathbf{V}_k^*$ la matrice réduite obtenue de $\mathbf{V}_k$ en éliminant les lignes d'indices correspondant aux données manquantes de $\mathbf{x}^m$ et $\mathbf{V}_k^{(*)}$ ce qui reste de $\mathbf{V}_k$ une fois $\mathbf{V}_k^*$ extrait
     2. Soit $\mathbf{x}^{m*}$ le vecteur réduit obtenu de $\mathbf{x}^m$ en éliminant les données manquantes et $\mathbf{x}^{m(*)}$ la partie de $\mathbf{x}^m$ correspondant aux données manquantes
     3. $\mathbf{x}^{m(*)}$ est estimé par $\hat{\mathbf{x}}^{m(*)} = \mathbf{V}_k^{(*)} \left((\mathbf{V}_k^*)^T \mathbf{V}_k^* \right)^{-1} (\mathbf{V}_k^*)^T \mathbf{x}^{m*}$

Solutions | Imputation des données manquantes 13 / 16

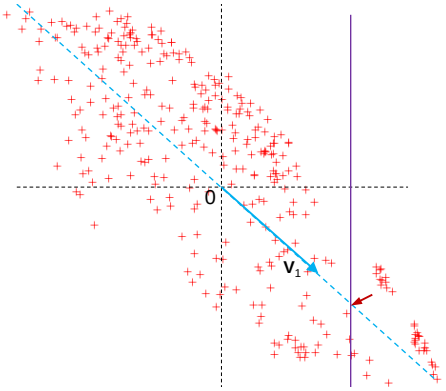
Imputation par décomposition en valeurs singulières (2)

- Comment procéder (voir par ex. [1]) :
  1. Appliquer la SVD de rang $k < d$ à $\mathbf{X}^c \Rightarrow \mathbf{X}_k^c = \mathbf{U}_k \mathbf{D}_k \mathbf{V}_k^T$
  2. Pour chaque observation $\mathbf{x}^m$ (transposée d'une ligne de $\mathbf{X}^m$) à données manquantes
     1. Soit $\mathbf{V}_k^*$ la matrice réduite obtenue de $\mathbf{V}_k$ en éliminant les lignes d'indices correspondant aux données manquantes de $\mathbf{x}^m$ et $\mathbf{V}_k^{(*)}$ ce qui reste de $\mathbf{V}_k$ une fois $\mathbf{V}_k^*$ extrait
     2. Soit $\mathbf{x}^{m*}$ le vecteur réduit obtenu de $\mathbf{x}^m$ en éliminant les données manquantes et $\mathbf{x}^{m(*)}$ la partie de $\mathbf{x}^m$ correspondant aux données manquantes
     3. $\mathbf{x}^{m(*)}$ est estimé par $\hat{\mathbf{x}}^{m(*)} = \mathbf{V}_k^{(*)} \left((\mathbf{V}_k^*)^T \mathbf{V}_k^* \right)^{-1} (\mathbf{V}_k^*)^T \mathbf{x}^{m*}$

- Choix du paramètre (rang k ici, nombre de voisins pour k plus proches voisins, etc.)
  1. Générer aléatoirement des matrices indicatrices des données manquantes (par rapport aux observations complètes de $\mathbf{X}^c$)
  2. Imputer ces données manquantes pour différentes valeurs du paramètre et calculer à chaque fois l'erreur moyenne
  3. Choisir la valeur pour laquelle l'erreur est minimale

Imputation par SVD : exemple simple 2D

- Observations complètes = points rouges ; observation à donnée manquante (l'ordonnée manque) = trait vertical (à la position de l'abscisse, connue)

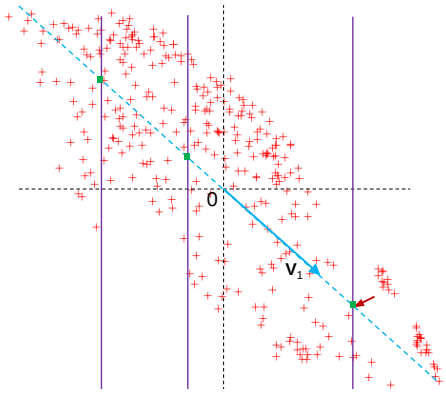


- $k = 1$, donc $\mathbf{V}_k$ est ici 2×1, on obtient par SVD : $\mathbf{v}_1 = \begin{pmatrix} 0,77 \\ -0,63 \end{pmatrix}$
- Soit l'observation $\mathbf{x}^m = \begin{pmatrix} 1,5 \\ ? \end{pmatrix}$ (la seconde composante manque)
- On obtient alors $\hat{\mathbf{x}}^{m(*)} = -1,23$, correspondant à l'ordonnée du point où le trait vertical intersecte le support du vecteur $\mathbf{v}_1$

Solutions | Imputation des données manquantes | 14 / 16

Imputation par SVD : exemple simple 2D

- Observations complètes = points rouges ; observation à donnée manquante (l'ordonnée manque) = trait vertical (à la position de l'abscisse, connue)



- $k = 1$, donc $\mathbf{V}_k$ est ici 2×1, on obtient par SVD : $\mathbf{v}_1 = \begin{pmatrix} 0,77 \\ -0,63 \end{pmatrix}$
- Soit l'observation $\mathbf{x}^m = \begin{pmatrix} 1,5 \\ ? \end{pmatrix}$ (la seconde composante manque)
- On obtient alors $\hat{\mathbf{x}}^{m(*)} = -1,23$, correspondant à l'ordonnée du point où le trait vertical intersecte le support du vecteur $\mathbf{v}_1$

Autres méthodes

- Approches itératives de type EM : à chaque itération,
  1. Grâce aux précédentes estimations des données manquantes, amélioration du modèle permettant d'imputer les données manquantes
  2. Grâce au nouveau modèle, amélioration des estimations des données manquantes

  - Appliquées avec diverses méthodes d'estimation par régression des valeurs manquantes : SVD [1], régression linéaire [1], arbres de régression, forêts aléatoires (*MissForest* [4]), etc.
- Imputation multiple :
  - Imputer plusieurs fois les données manquantes, analyser/modéliser chaque ensemble de données complétées, puis intégrer les résultats de ces différentes analyses/modèles
  - Différentes techniques sont employées pour imputer plusieurs fois les données manquantes, comme le tirage du modèle d'imputation suivant sa distribution *a posteriori* et ensuite l'imputation avec ce modèle
  - Voir par ex. [3], [2], [5]

Références I

T. Hastie, R. Tibshirani, G. Sherlock, M. Eisen, P. Brown, and D. Botstein.

Imputing missing data for gene expression arrays.

Technical report, Division of Biostatistics, Stanford University, September 1999.

J. Honaker, G. King, and M. Blackwell.

Amelia II. A program for missing data.

Journal of Statistical Software, 45 :1–47, 2011.

D. B. Rubin.

Multiple Imputation for Nonresponse in Surveys.

Wiley, 1987.

D. J. Stekhoven and P. Buehlmann.

Missforest - non-parametric missing value imputation for mixed-type data.

Bioinformatics, 28(1) :112–118, 2012.

S. van Buuren.

Flexible Imputation of Missing Data.

Chapman & Hall/CRC Press, 2012.