Classification non supervisée
Etude des étoiles

DUQUESNE Théo
GUEYE Taliesin
NAVARRE Victor
PINEAU Benjamin

DUQUESNE
Théo
GUEYE
Taliesin
NAVARRE
Victor
PINEAU
Benjamin

Sommaire

DUQUESNE
Théo
GUEYE
Taliesin
NAVARRE
Victor
PINEAU
Benjamin

Jeu de données

On dispose d'un jeu de données comprenant des informations sur des étoiles tel que :

- Température (K)
- Rayon
- Couleur
- Classe spectrale
- ...

Objectif : classifier les étoiles selon leur classe : Naine rouge, Supergiant, Hypergiant ...

Jeu de données

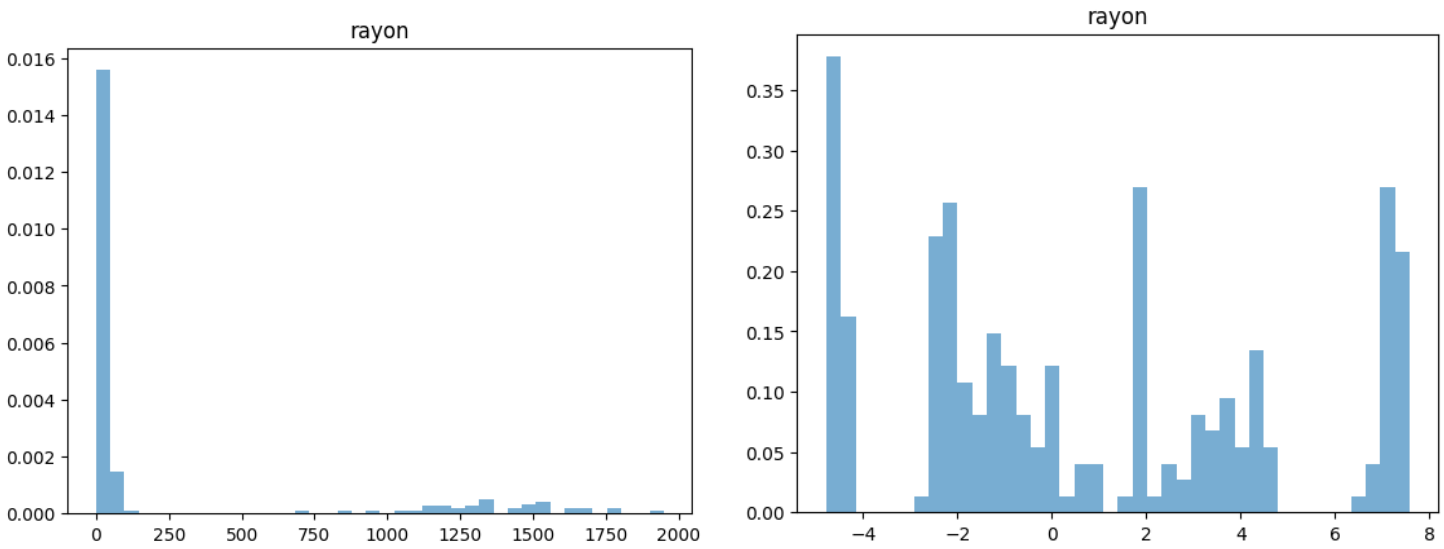
Comment faire ?
Classification non supervisée !

1. Le jeu de données contient les classes des étoiles
2. On supprime cette colonne lors de l'apprentissage : on essaie de classifier sans connaitre la classe
3. On vérifie le résultat en comparant avec la classe connue

DUQUESNE
Théo
GUEYE
Taliesin
NAVARRE
Victor
PINEAU
Benjamin

DUQUESNE
Théo
GUEYE
Taliesin
NAVARRE
Victor
PINEAU
Benjamin

Visualisation



(a) Echelle classique

(b) Echelle log

Figure: Histogramme selon le rayon

Visualisation

Objectif :

- Meilleure séparation des clusters
- Réduction de l'effet des valeurs extrêmes
- Le mélange gaussien suppose que chaque composante suit une distribution gaussienne

Variables qualitatives

Changement des variables catégorielles :

- Couleur $\longleftrightarrow$ 0, 1, 2 ...
- Classe spectrale M, G, O ... $\longleftrightarrow$ 0, 1, 2 ...
  Regroupement de certaines classes spectrales ? *M et K représentent le même type d'étoile*

DUQUESNE
Théo
GUEYE
Taliesin
NAVARRE
Victor
PINEAU
Benjamin

Variables qualitatives

Premiers résultats avec 6 clusters et une matrice de covariance '`full`' ($\Sigma_g = \lambda_g Q_g A_g Q_g^\top$):

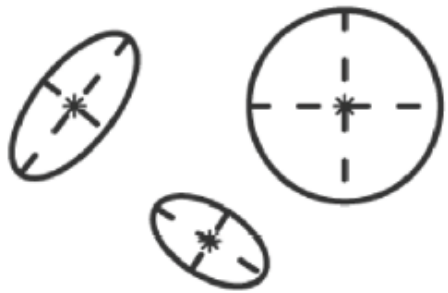


Figure: Matrice de Covariance Σ_g

$\Longrightarrow$ Résultats non convaincants

DUQUESNE
Théo
GUEYE
Taliesin
NAVARRE
Victor
PINEAU
Benjamin

Sélection du modèle

Décomposition Volume-Shape-Orientation de Σ_g:

$$\Sigma_g = \lambda_g Q_g A_g Q_g^\top$$

avec

- Q_g la matrice des vecteurs propores de Σ_g (orientation)
- A_g la matrice diagonale proportionnelle comosée des valeurs propres de Σ_g (forme)
- λ_g la constante de proportionalité associée (volume)

DUQUESNE
Théo
GUEYE
Taliesin
NAVARRE
Victor
PINEAU
Benjamin

Sélection du modèle

Choix du modèle : calcul du Bayesian Information Criterion (BIC)

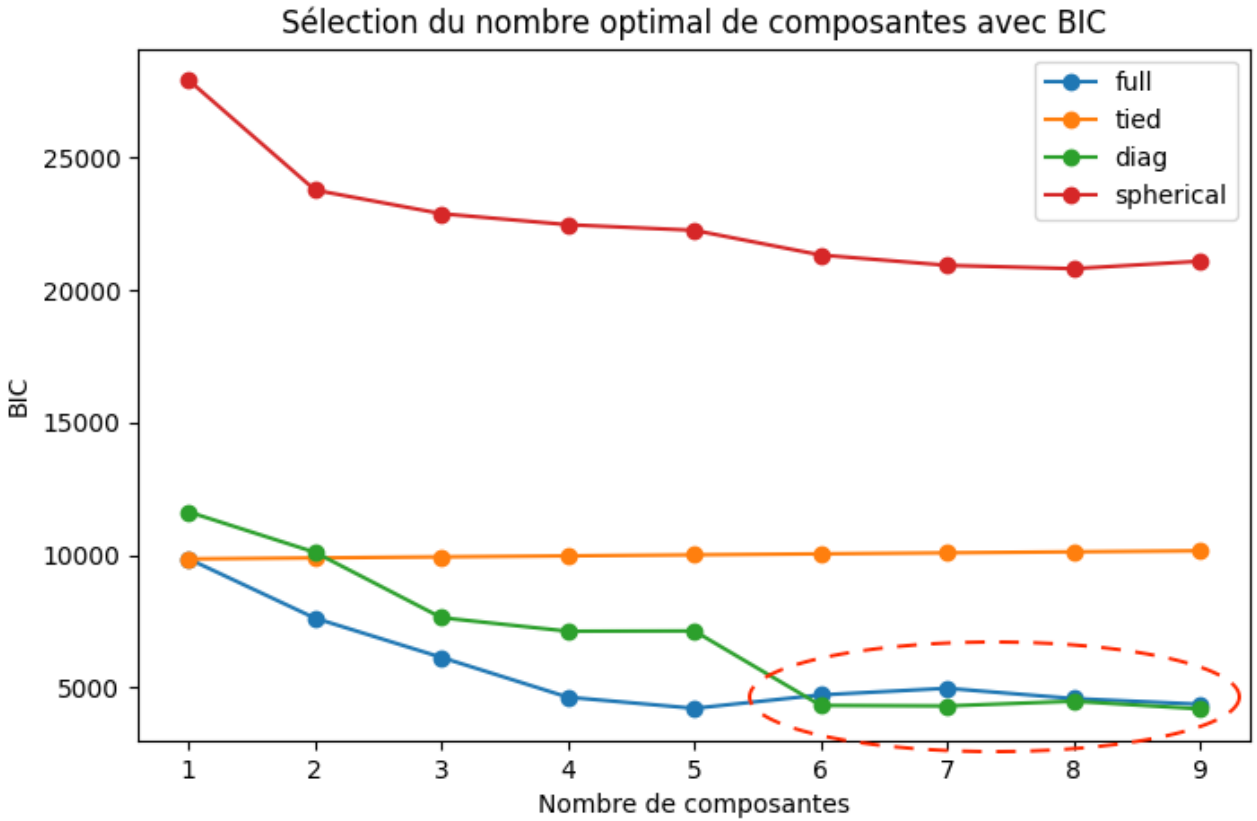


Figure: Calcul du BIC

DUQUESNE
Théo
GUEYE
Taliesin
NAVARRE
Victor
PINEAU
Benjamin

Plongement dans un espace plus grand

- On est bloqués par le fait que les variables soient catégorielles
- On aimerait éviter de mettre une relation d'ordre lors de la transformation en variables scalaires

Solution : Plonger dans un espace de dimension plus grand qui respecte ces critères

DUQUESNE
Théo
GUEYE
Taliesin
NAVARRE
Victor
PINEAU
Benjamin

Plongement dans un espace plus grand

Pour chaque catégorie, on crée une dimension par classe possible (on les note "is white", "is M" etc...)

Chacun des points va être emmené sur la coordonée 0 ou 1 suivant sa classe

Avantages : On se sépare de la relation d'ordre qu'on avait
On peut toujours espérer que la distance euclidienne ait une pertinence
Inconvénients : C'est artificiel, toujours pas continu, et on passe d'une dimension 8 à une dimension 25

DUQUESNE
Théo
GUEYE
Taliesin
NAVARRE
Victor
PINEAU
Benjamin

Modèle de Mixture Gaussienne

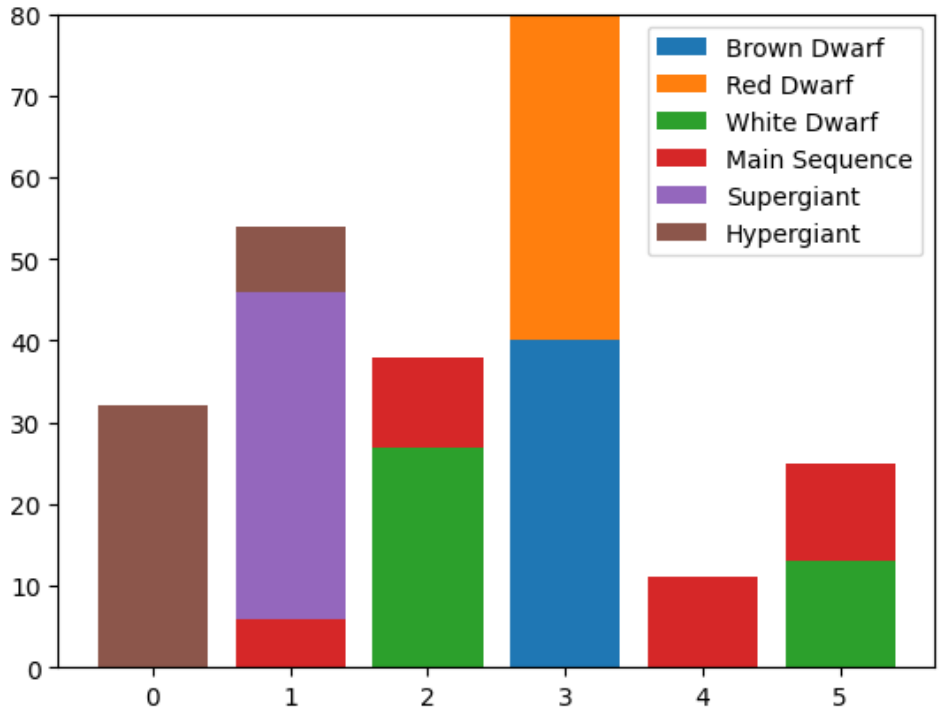


Figure: mixture gaussienne à 6 classes en dimension 25

DUQUESNE
Théo
GUEYE
Taliesin
NAVARRE
Victor
PINEAU
Benjamin

Modèle de Mixture Gaussienne

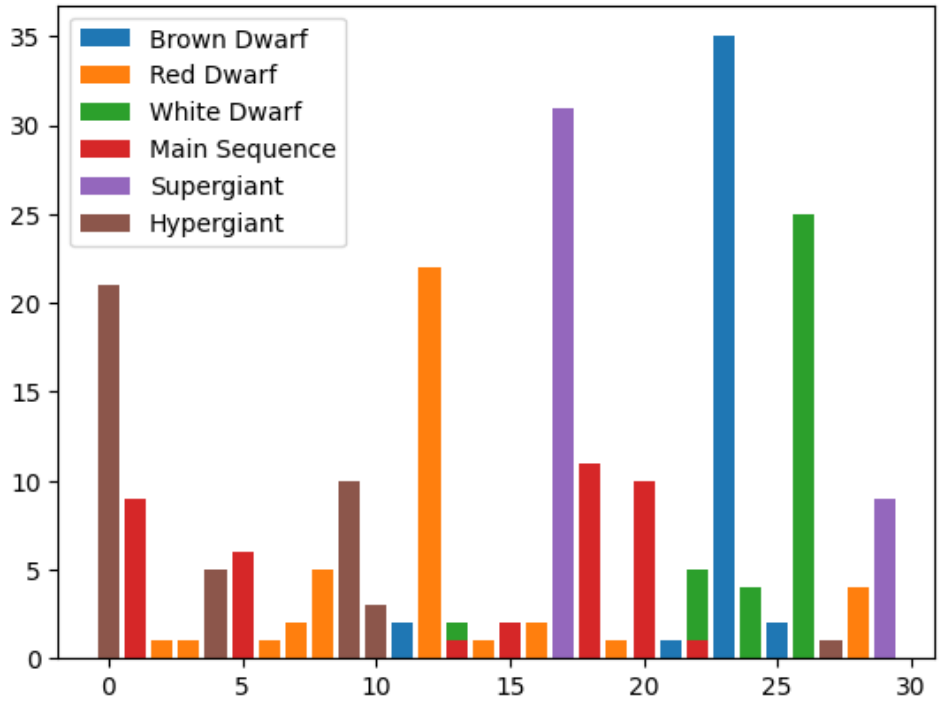


Figure: mixture gaussienne à 30 classes en dimension 25

Solution

Idée : On garde cette mixture gaussienne à plein de classes, et on élague à l'aide d'une métrique (un peu comme un arbre)

DUQUESNE
Théo
GUEYE
Taliesin
NAVARRE
Victor
PINEAU
Benjamin

Mesure par quantiles

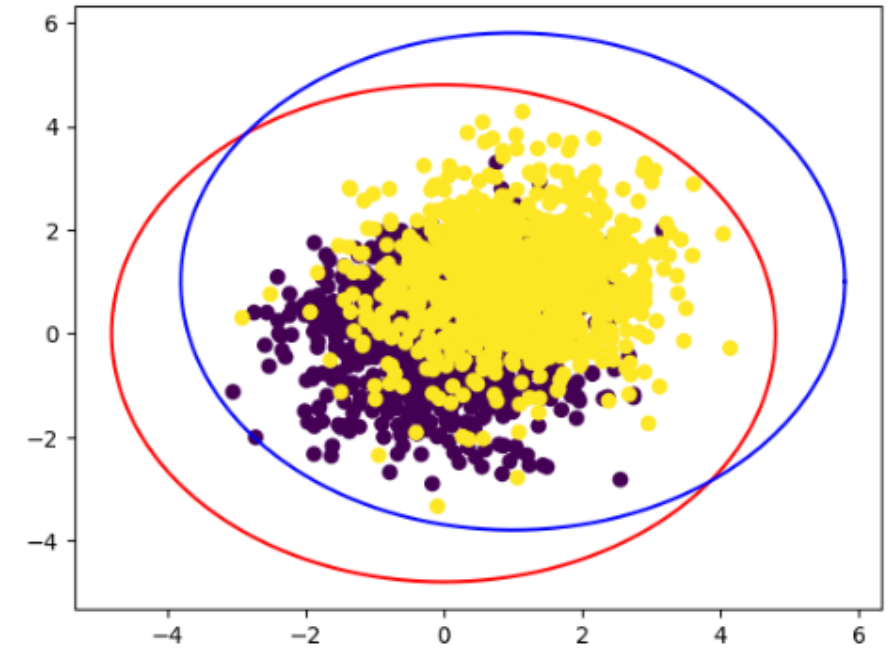


Figure: Illustration métrique

DUQUESNE
Théo
GUEYE
Taliesin
NAVARRE
Victor
PINEAU
Benjamin

Mesure par quantiles

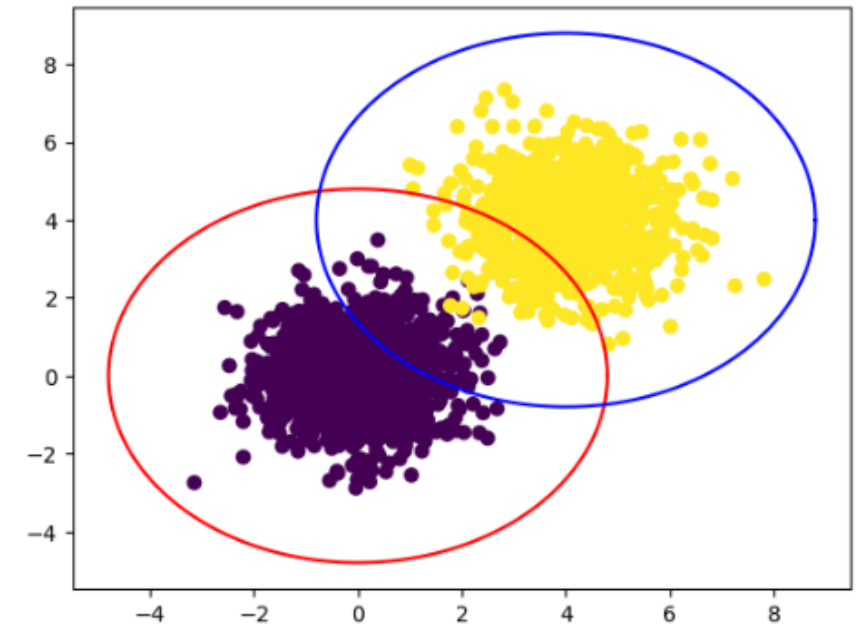


Figure: Illustration métrique

Mesure par quantiles

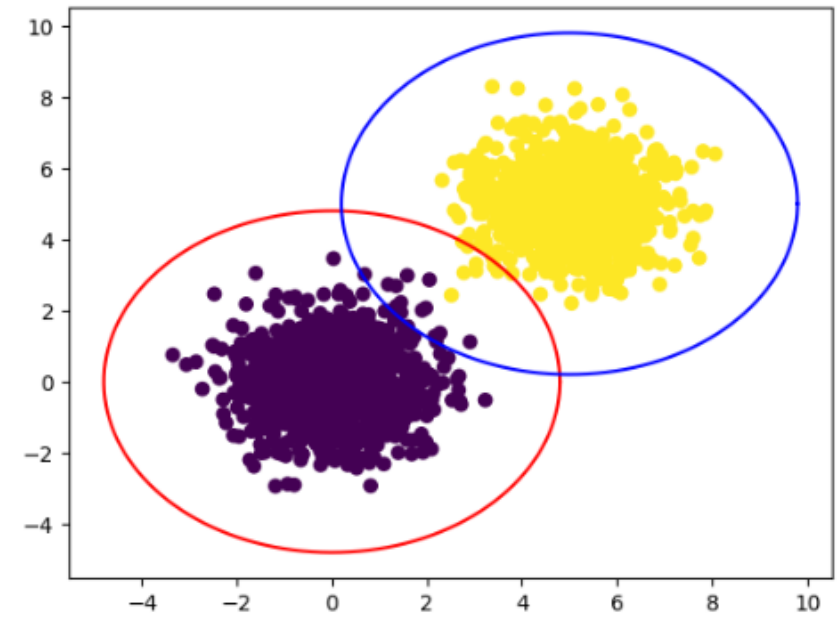


Figure: Illustration métrique

DUQUESNE
Théo
GUEYE
Taliesin
NAVARRE
Victor
PINEAU
Benjamin

Mesure par quantiles

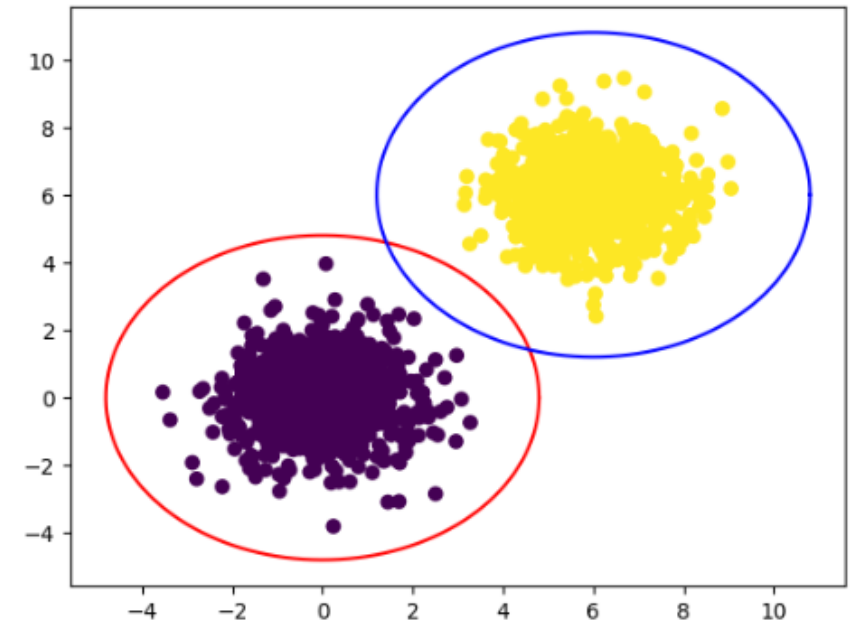


Figure: Illustration métrique

DUQUESNE
Théo
GUEYE
Taliesin
NAVARRE
Victor
PINEAU
Benjamin

Mesure par quantiles

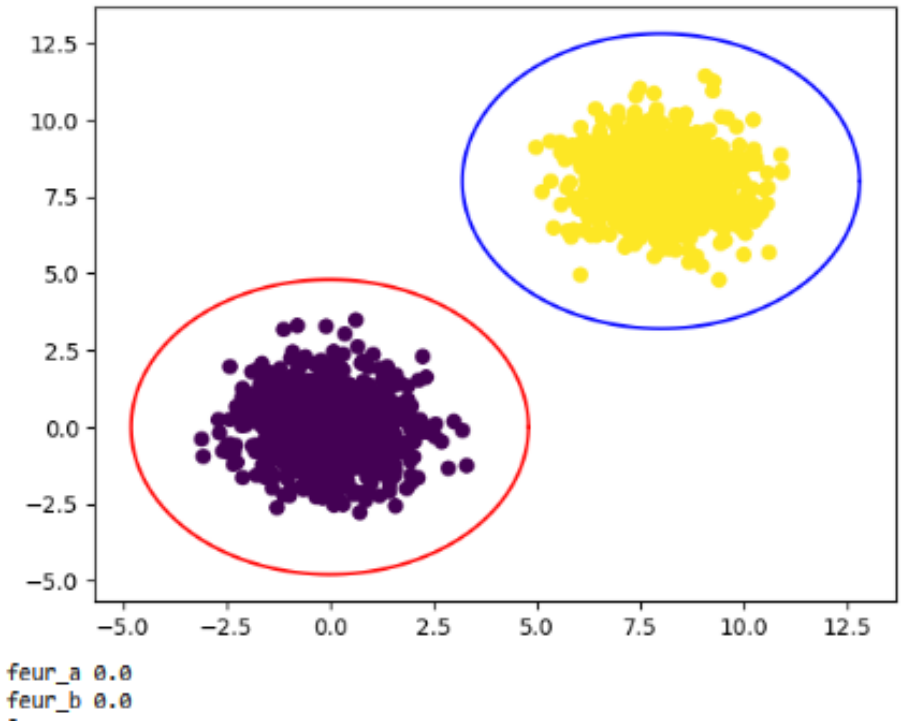


Figure: Illustration métrique

DUQUESNE
Théo
GUEYE
Taliesin
NAVARRE
Victor
PINEAU
Benjamin

Mesure par quantiles

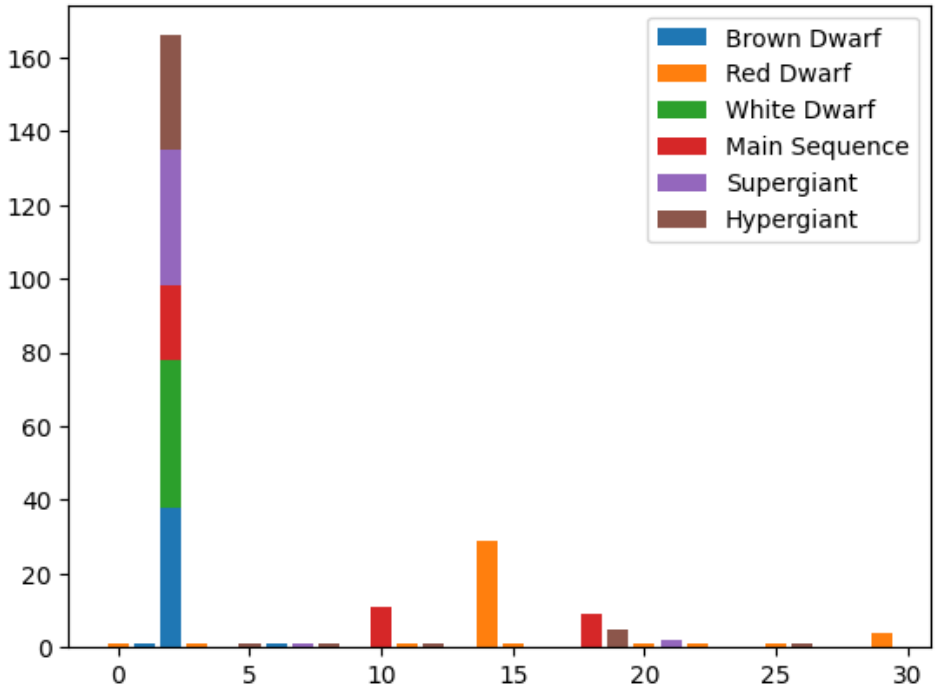


Figure: Résultat après élaguage

Pistes

Solutions ?

- Autre métrique
- Pénalisation
- Réduction de dimension
- Featurisation

Du continue au discret

Utilisation de variables qualitatives : `Star type`, `Star color` ou `Spectral Class`

Nous disposons d'un vecteur aléatoire multivarié $Y = (Y_1, ..., Y_d)$ dont les composantes sont des variables aléatoires discrètes avec $m_1, ...m_d$ niveaux

On passe du cas continu au cas discret $\Rightarrow$ on ne calcule plus la moyenne μ_g et la matrice de covariance Σ_g mais un vecteur de probabilité d'appartenance pour les variables catégorielles Y_j

DUQUESNE
Théo
GUEYE
Taliesin
NAVARRE
Victor
PINEAU
Benjamin

Du continue au discret

On passe de la densité de mélange continue

$$p(y) = \sum_{g=1}^{G} \tau_g f_g(y, \theta_g)$$

à la densité de mélange discrète

$$p(y) = \sum_{g=1}^{G} \tau_g \prod_{j=1}^{d} p_{g,j,y_j}$$

DUQUESNE
Théo
GUEYE
Taliesin
NAVARRE
Victor
PINEAU
Benjamin

Transformation des données

Certaines variables ne sont pas qualitatives : le rayon, la luminosité ...

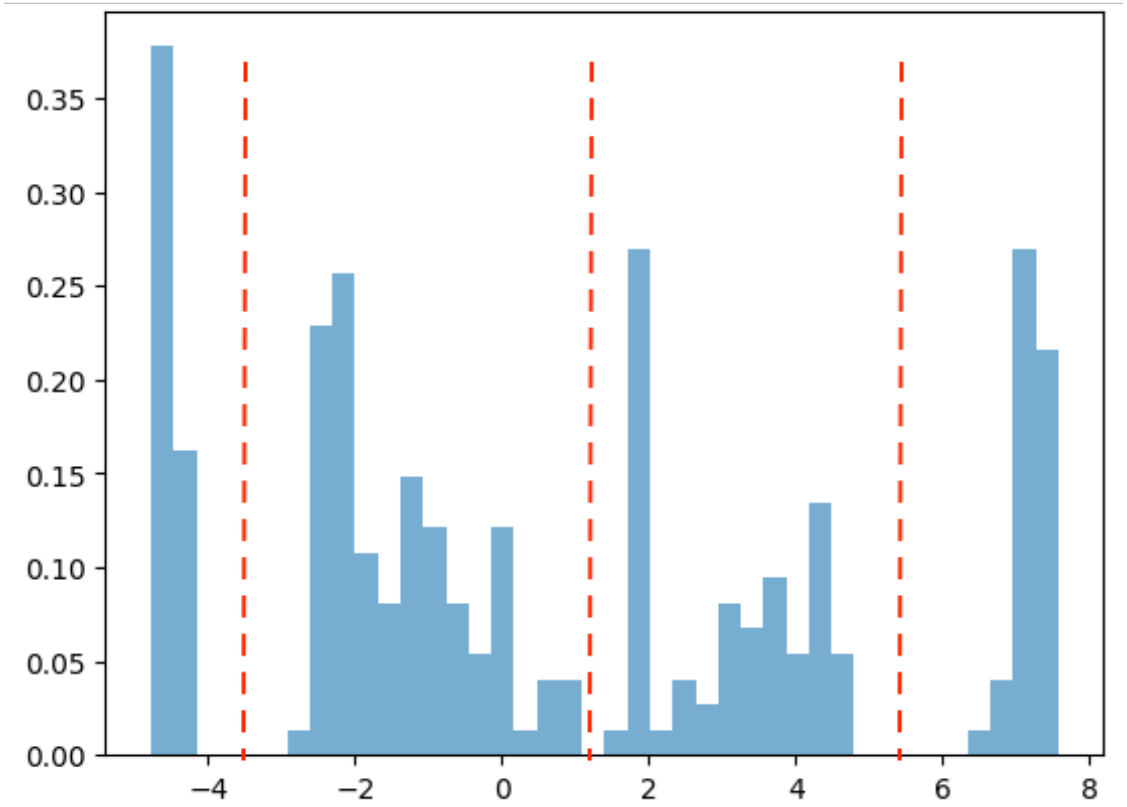


Figure: Histogramme selon le rayon (échelle log)

DUQUESNE
Théo
GUEYE
Taliesin
NAVARRE
Victor
PINEAU
Benjamin

Résultats avec StepMix

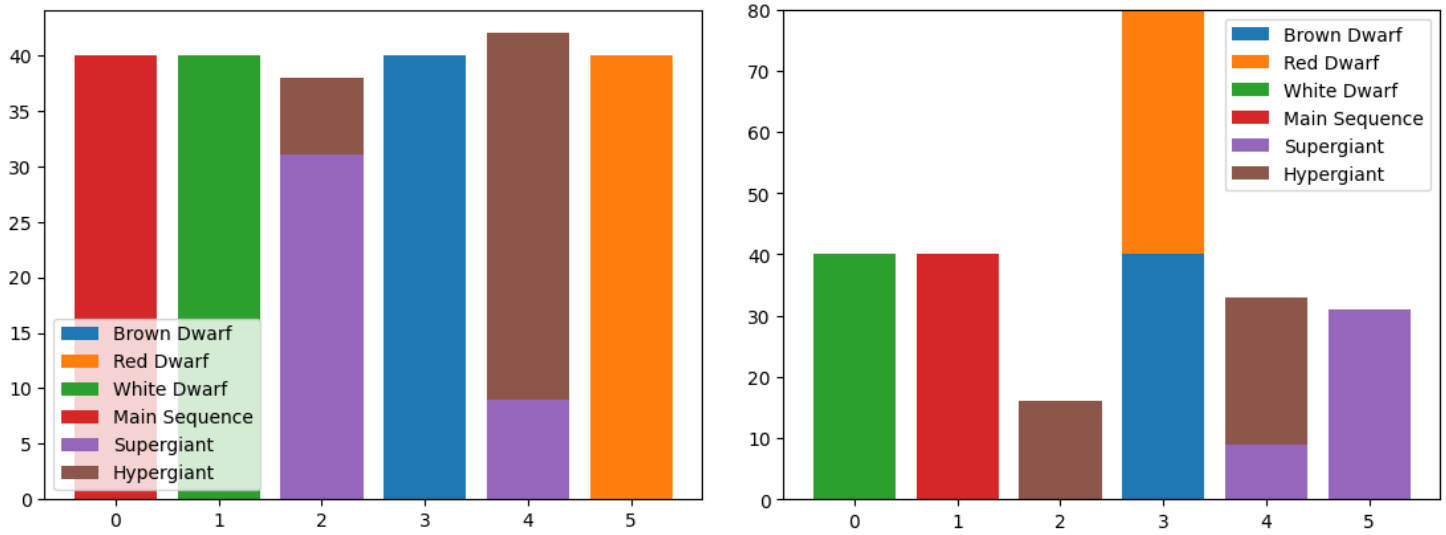


Figure: Instabilité de la méthode

DUQUESNE
Théo
GUEYE
Taliesin
NAVARRE
Victor
PINEAU
Benjamin

Distance de Wasserstein

Est il pertinent de modéliser les distributions par des distributions gaussiennes ? On calcule la distance de Wasserstein.

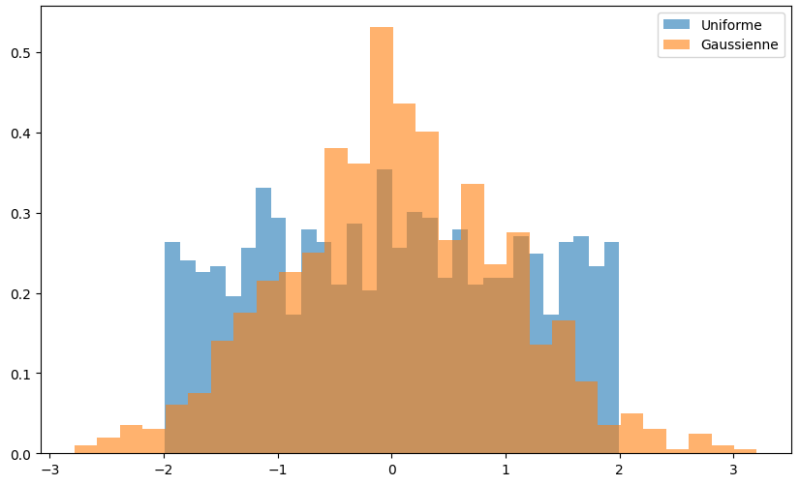


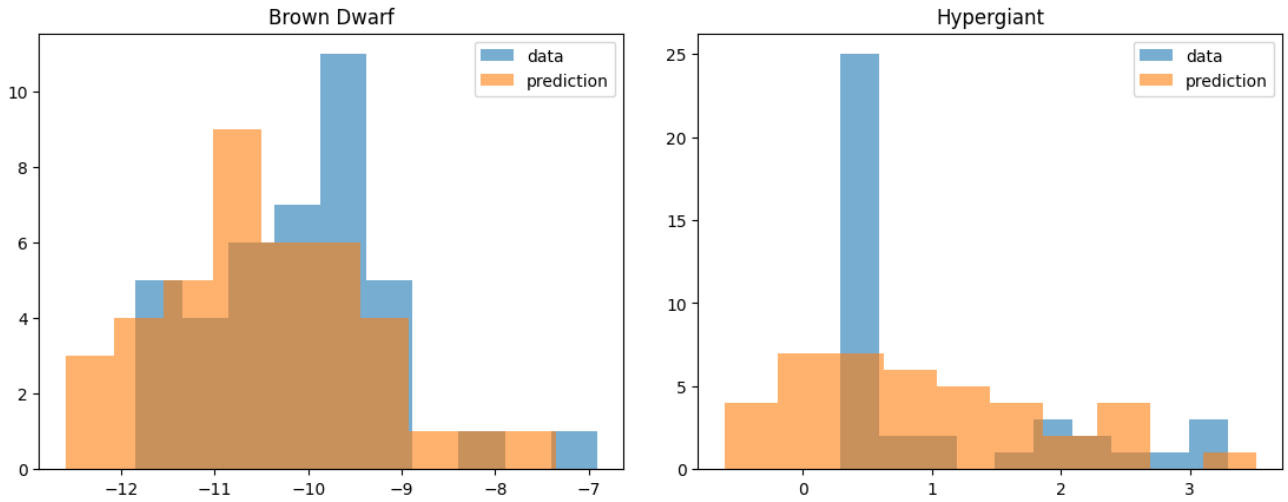
Figure: Comparaison des distributions : Uniforme vs Gaussienne

la distance de Wasserstein est $W = 0.24$

DUQUESNE
Théo
GUEYE
Taliesin
NAVARRE
Victor
PINEAU
Benjamin

Distance de Wasserstein pour le dataset

Distance de Wasserstein entre le dataset et la loi gaussienne $\mathcal{N}(\mu_g, \Sigma_g)$



(a) Brown Dwarf (b) Hypergiant

Figure: Pertinence de l'approche gaussienne

Type d'étoile	**Distance**
Brown Dwarf	0.1972
Red Dwarf	0.1343
White Dwarf	0.3039
Main Sequence	0.3690
Supergiant	0.3780
Hypergiant	0.3689

Table: Tableau des distances de Wasserstein