

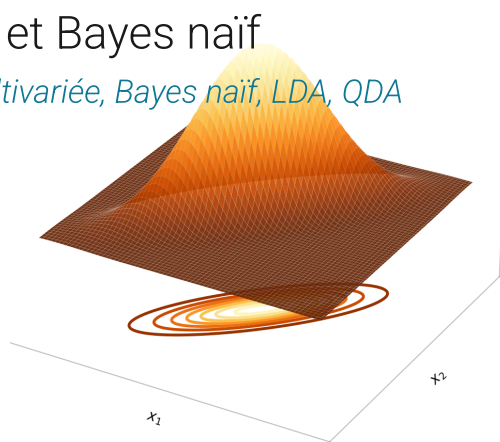
Modèles gaussiens et Bayes naïf

Séance 2 – Loi normale multivariée, Bayes naïf, LDA, QDA

Christophe Ambroise 

UEVE, UMR CNRS 8071

4 septembre 2026



1

La loi normale multivariée

2

Le point de départ

En classification, la règle de Bayes choisit la classe :

$$\arg \max_k P(Y = k \mid X = x) = \arg \max_k \frac{\pi_k f_k(x)}{\sum_{\ell} \pi_{\ell} f_{\ell}(x)}$$

Deux quantités à estimer. La proportion π_k se compte. La densité f_k se modélise.

L'approche paramétrique donne une forme à f_k , puis estime ses paramètres.

3

La gaussienne

Calculs explicites, géométrie claire, interprétation directe des paramètres.

Pour $x \in \mathbb{R}^d$, moyenne μ , covariance Σ définie positive :

$$f(x) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left(-\frac{1}{2} (x - \mu)^{\top} \Sigma^{-1} (x - \mu) \right)$$

$$(x - \mu)^{\top} \Sigma^{-1} (x - \mu) = \text{constante}$$

Équation d'un **ellipsoïde**. Isodensités : des ellipses.

4

Interprétation des paramètres

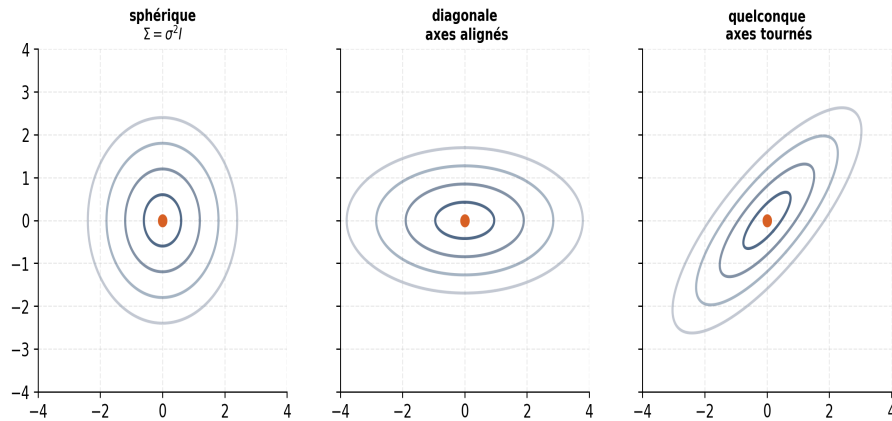


Figure 1

Trois contraintes sur Σ , trois géométries. Le panneau du milieu est celui du Bayes naïf, celui de droite celui de LDA et QDA.

Lire une matrice de covariance

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

Diagonale : variances, étalement selon chaque axe.

Hors-diagonale : covariance, inclinaison des ellipses.

$\sigma_{12} = 0$: axes alignés sur les variables. Pour une gaussienne, décorrélation et indépendance coïncident.

Décomposition spectrale

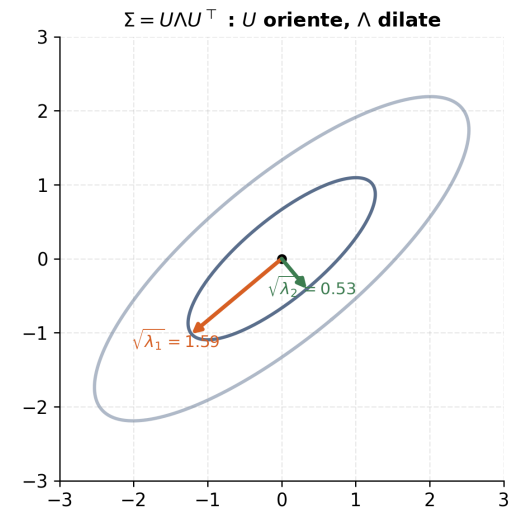


Figure 2

$\Sigma = U\Lambda U^\top$. Vecteurs propres : directions. Racines des valeurs propres : longueurs. $\det(\Sigma) = \prod_i \lambda_i$: volume de l'ellipsoïde.

7

Trois nombres, trois rôles

En dimension 2, une covariance se lit :

- une **orientation** : l'angle de U ;
- deux **allongements** : $\sqrt{\lambda_1}$ et $\sqrt{\lambda_2}$.

Trois paramètres libres pour Σ , contre deux pour μ .

En dimension d : d paramètres pour le centre, $d(d+1)/2$ pour la covariance.

8

La distance de Mahalanobis

L'exposant de la densité est une distance :

$$d_\Sigma^2(x, \mu) = (x - \mu)^\top \Sigma^{-1} (x - \mu)$$

Elle rapporte l'écart à la dispersion.

$\Sigma = I$ redonne la distance euclidienne.

9

Euclide ne suffit pas

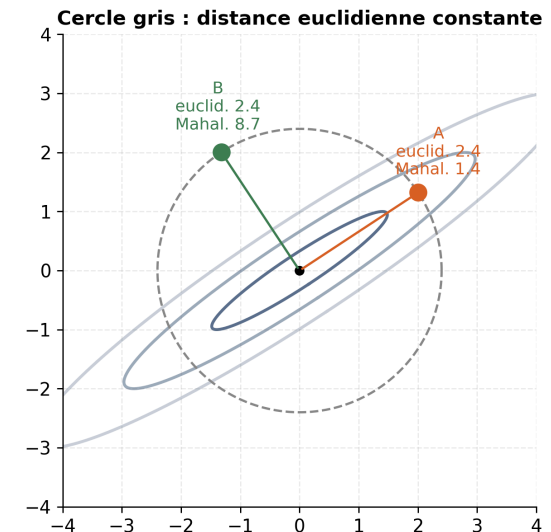


Figure 3

A et B sont à la même distance euclidienne du centre. B est nettement plus atypique.

10

Ce que la gaussienne fournit

Une densité dont le logarithme est une **forme quadratique** en x :

$$\log f_k(x) = -\frac{1}{2} d_{\Sigma_k}^2(x, \mu_k) - \frac{1}{2} \log |\Sigma_k| - \frac{d}{2} \log 2\pi$$

Comparer deux classes revient à comparer deux formes quadratiques. La géométrie de la frontière en découle.

11

Bayes naïf, revisité

12

Ce qui a été vu

La séance précédente contournait le fléau de la dimension par une hypothèse :

Conditionnellement à la classe, les variables sont indépendantes.

$$f_k(x) = \prod_{j=1}^d f_{kj}(x_j)$$

13

L'hypothèse naïve, traduite en covariance

Si chaque f_{kj} est gaussienne, le produit est une gaussienne de covariance **diagonale** :

$$\Sigma_k = \text{diag}(\sigma_{k1}^2, \dots, \sigma_{kd}^2), \quad |\Sigma_k| = \prod_{j=1}^d \sigma_{kj}^2.$$

Bayes naïf gaussien = une gaussienne par classe, aux **axes alignés sur les variables**. C'est le panneau du milieu de la première figure.

Aucune corrélation n'est modélisée — mais chaque classe garde ses propres étalements.

La fonction de décision

En reprenant $\delta_k(x) = \log \pi_k + \log f_k(x)$ avec Σ_k diagonale :

$$\delta_k(x) = \log \hat{\pi}_k - \sum_{j=1}^d \left[\frac{(x_j - \hat{\mu}_{kj})^2}{2 \hat{\sigma}_{kj}^2} + \log \hat{\sigma}_{kj} \right]$$

Le terme en x_j^2 dépend de k par $\hat{\sigma}_{kj}$: il ne s'élimine pas.

Les frontières du Bayes naïf gaussien sont **quadratiques**, pas linéaires.

Sur une variable d'Iris

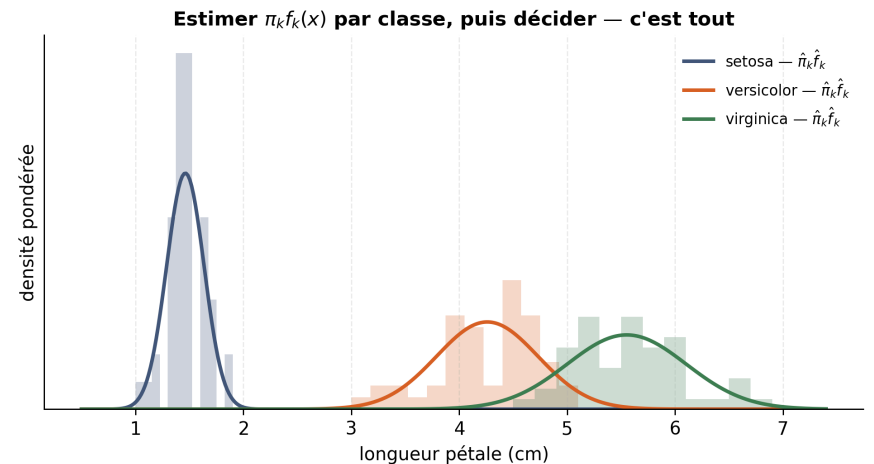


Figure 4

Une gaussienne ajustée par espèce, pondérée par $\hat{\pi}_k$. En dimension d , le modèle fait cela d fois, indépendamment.

16

Analyse discriminante linéaire

17

Le modèle

Une gaussienne par classe, de moyenne propre :

$$X \mid Y = k \sim \mathcal{N}(\mu_k, \Sigma_k)$$

Hypothèse d'homoscédasticité

$$\Sigma_k = \Sigma \quad \text{pour toute classe } k$$

Les classes diffèrent par leur position, pas par leur forme.

18

La règle de décision

Maximiser $\pi_k f_k(x)$ équivaut à maximiser $\log \pi_k + \log f_k(x)$:

$$\begin{aligned} \delta_k(x) &= \log \pi_k - \frac{1}{2}(x - \mu_k)^\top \Sigma^{-1}(x - \mu_k) \\ &= \log \pi_k - \frac{1}{2}x^\top \Sigma^{-1}x + x^\top \Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^\top \Sigma^{-1}\mu_k \end{aligned}$$

Le terme $x^\top \Sigma^{-1}x$ ne dépend pas de k : il disparaît de la comparaison. C'est là toute la différence avec le Bayes naïf.

19

Une fonction affine

$$\delta_k(x) = \underbrace{x^\top \Sigma^{-1} \mu_k}_{\text{linéaire en } x} - \underbrace{\frac{1}{2} \mu_k^\top \Sigma^{-1} \mu_k + \log \pi_k}_{\text{constante}}$$

La frontière $\delta_k(x) = \delta_\ell(x)$ est un **hyperplan**. Conséquence de l'homoscédasticité, pas une approximation.

20

Frontière explicite, cas binaire

$$\delta_1(x) - \delta_0(x) = x^\top \underbrace{\Sigma^{-1}(\mu_1 - \mu_0)}_w + \underbrace{\log \frac{\pi_1}{\pi_0} - \frac{1}{2}(\mu_1 - \mu_0)^\top \Sigma^{-1}(\mu_1 + \mu_0)}_b$$

Décision : $w^\top x + b > 0$.

Le vecteur w n'est pas $\mu_1 - \mu_0$, mais cet écart **corrigé par la covariance**.

21

Lien avec Mahalanobis

Si $\pi_0 = \pi_1$, la règle se réduit à :

$$d_\Sigma^2(x, \mu_1) < d_\Sigma^2(x, \mu_0)$$

Affecter x à la classe au centre le plus proche — au sens de Mahalanobis, pas d'Euclide.

Des proportions déséquilibrées déplacent la frontière vers la classe rare : la même mécanique qu'au seuil t^* de la séance précédente.

22

Estimation

Trois quantités, toutes explicites :

$$\hat{\pi}_k = \frac{n_k}{n}, \quad \hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

$$\hat{\Sigma} = \frac{1}{n - K} \sum_k \sum_{i:y_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^\top$$

La covariance est commune à toutes les classes. Chaque observation y contribue.

23

Analyse discriminante quadratique

24

Lever l'hypothèse

$$X | Y = k \sim \mathcal{N}(\mu_k, \Sigma_k)$$

Une matrice de covariance **par classe**, sans contrainte de forme.

Le terme $x^\top \Sigma_k^{-1} x$ dépend de k . Il ne s'élimine plus — comme pour le Bayes naïf, mais cette fois les axes peuvent tourner.

25

Une fonction quadratique

$$\delta_k(x) = \log \pi_k - \frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x - \mu_k)^\top \Sigma_k^{-1} (x - \mu_k)$$

La frontière est une **conique** : ellipse, parabole ou hyperbole selon les covariances.

Le terme $\log |\Sigma_k|$ pénalise les classes très dispersées.

26

Les deux modèles, côte à côte

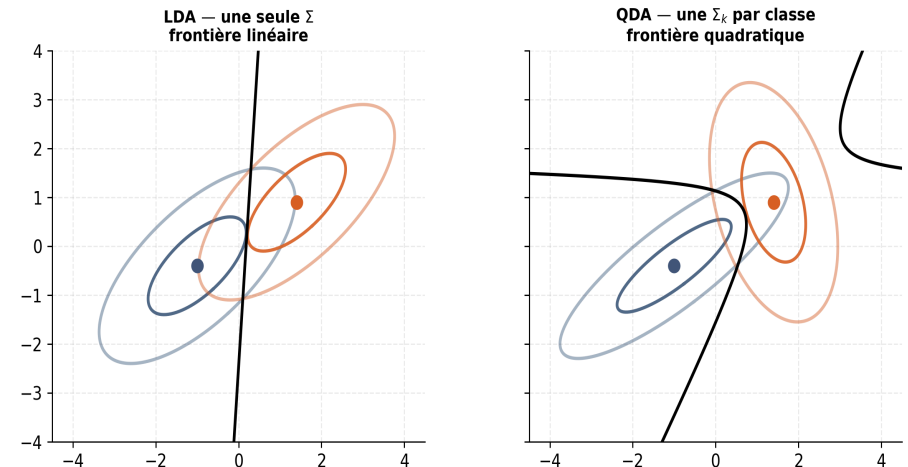


Figure 5

À gauche : ellipses identiques à translation près. À droite : chacune a sa forme.

27

Sur Iris

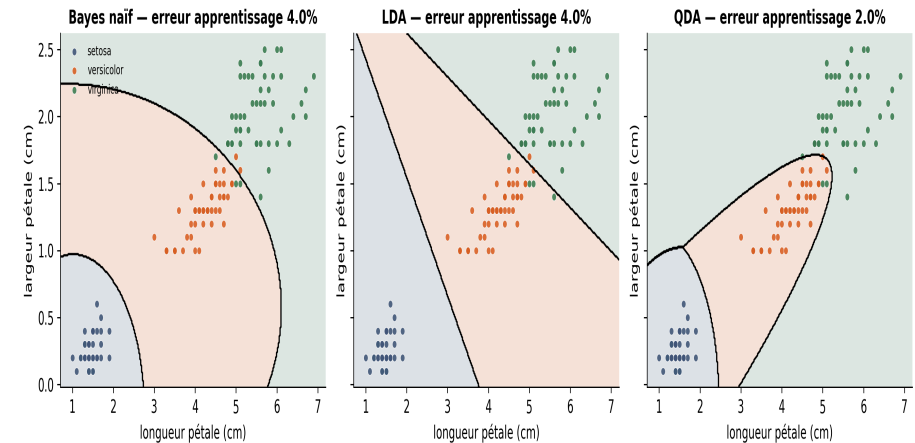


Figure 6

Frontières droites pour LDA, courbes pour les deux autres.

28

Ce que la figure montre

Sur les deux mesures de pétale, en validation croisée à dix plis :

Modèle	Erreur d'apprentissage	Erreur validée
Bayes naïf	4,0 %	4,0 %
LDA	4,0 %	4,0 %
QDA	2,0 %	2,7 %

QDA gagne ici : les trois espèces n'ont visiblement pas la même dispersion.

Sur les **quatre** variables, LDA passe à 2,0 % et rejoint QDA, tandis que Bayes naïf reste à 4,7 %. Le classement dépend

des variables retenues.

29

En scikit-learn

```
1 from sklearn.naive_bayes import GaussianNB
2 from sklearn.discriminant_analysis import (
3     LinearDiscriminantAnalysis, QuadraticDiscriminantAnalysis)
4
5 nb = GaussianNB().fit(X, y)
6 lda = LinearDiscriminantAnalysis(store_covariance=True).fit(X, y)
7 qda = QuadraticDiscriminantAnalysis(store_covariance=True).fit(X, y)
8
9 nb.var_          # les sigma_kj^2, une ligne par classe
10 lda.covariance_  # la Sigma commune
11 qda.covariance_  # la liste des Sigma_k
```

`store_covariance=True` conserve les matrices estimées. Sans lui, elles sont calculées puis jetées.

30

Le coût de la richesse

31

Compter les paramètres

Pour K classes en dimension d :

Modèle	Covariance	Paramètres de covariance
Sphérique	$\sigma_k^2 I$	K
Bayes naïf	Σ_k diagonale	Kd
LDA	Σ commune	$d(d+1)/2$
QDA	Σ_k pleine	$K d(d+1)/2$

Croissance **quadratique** en d pour les deux derniers.

32

L'explosion

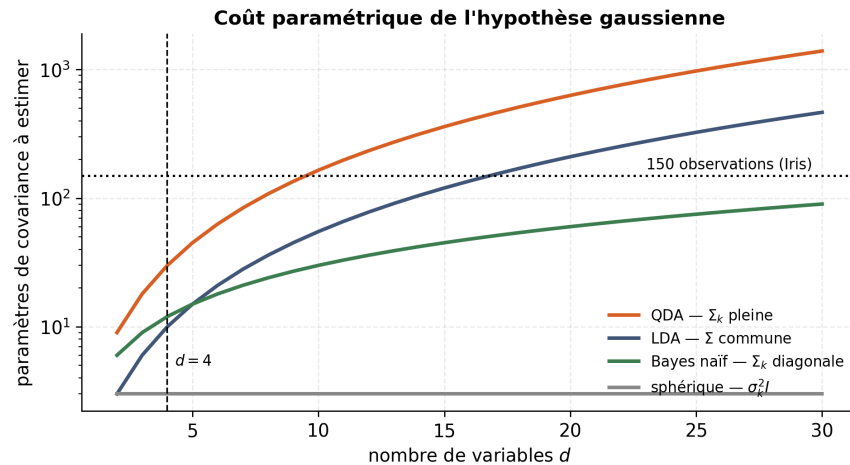


Figure 7

Échelle logarithmique. La ligne pointillée marque les 150 observations d'Iris.

Un exemple chiffré

Sur Iris, $d = 4$ variables et $K = 3$ classes :

	Covariance	Total avec μ_k et π_k
Bayes naïf	12	26
LDA	10	24
QDA	30	44

QDA estime 30 paramètres de covariance sur 150 observations, soit 10 par classe pour 50 fleurs. La marge est courte.

Le compromis, en clair

Plus de paramètres : moins de biais, plus de variance. Le dilemme de la séance précédente, transposé au choix de modèle.

QDA gagne toujours sur l'erreur d'apprentissage : strictement plus de liberté.

QDA perd souvent sur l'erreur de test : matrices mal estimées.

Modèles parcimonieux

Contraindre Σ_k sans l'imposer identique :

- **Diagonale** : variables indépendantes par classe. Exactement le Bayes naïf gaussien.
- **Sphérique** : une seule variance par classe. Frontières circulaires.
- **Régularisée** : $\hat{\Sigma}_k^{(\alpha)} = \alpha \hat{\Sigma}_k + (1 - \alpha) \hat{\Sigma}$, interpolation continue entre QDA et LDA.

α se choisit par validation croisée.

36

Bayes naïf, LDA, QDA

Trois points d'une même famille. Seule la contrainte sur Σ change.

	Corrélations	Formes par classe	Frontière
Bayes naïf	ignorées	oui	quadratique
LDA	modélisées	non	linéaire
QDA	modélisées	oui	quadratique

Aucun n'est meilleur dans l'absolu. Le choix dépend de n , de d et de la géométrie réelle des classes.

37

Un modèle faux, mais utile

L'indépendance conditionnelle est fautive sur Iris : au sein de *versicolor*, longueur et largeur de pétale sont corrélées à 0,79.

Bayes naïf s'en tire pourtant à 4 % d'erreur. Pour **classer**, il suffit que la bonne classe arrive en tête ; les probabilités estimées peuvent être mauvaises sans que la décision le soit.

Un modèle faux peut être utile. C'est même le cas général.

38

Remarques

Une variable qualitative n'a pas de densité gaussienne : la traiter comme telle n'a pas de sens.

Une variable très asymétrique respecte mal l'hypothèse. Une **transformation logarithmique** aide souvent.

Le choix du modèle contraint les données admissibles. L'inverse est aussi vrai.

39

Synthèse

1. La gaussienne réduit une densité à un centre et une géométrie.
2. La distance de Mahalanobis rapporte l'écart à la dispersion.
3. Bayes naïf, c'est Σ_k **diagonale** : pas de corrélations, mais une forme par classe, donc des frontières quadratiques.
4. Covariance commune : frontière linéaire. Covariance par classe et pleine : frontière quadratique.
5. Le coût paramétrique croît en d^2 ; le choix se tranche sur des données jamais vues.