IFT 6390, Devoir 1

Adel Nabli, Myriam Laiymani

20 Septembre 2018

1 Petit exercice de probabilités

On se place sur l'ensemble de probabilité $\Omega = (i, C_i, T_i)$ où:

- i = identifiant de la femme se faisant tester
- $C_i = \mathbf{1}_{\{}i$ a un cancer$_{\}}$
- $T_i = \mathbf{1}_{\{}i$ a un test positif$_{\}}$

La tribu considéré est $\mathcal{P}(\Omega)$ l'ensemble des parties de Ω. On sait que:

- P(C) = 1.5 %
- P(T|C) = 87 %
- P(T|C^c) = 9.6 %

Or, on cherche P(C|T) la probabilité qu'une femme ait un cancer sachant que son test est positif. Par la loi de Bayes, on a:

$$\mathrm{P(C|T)} = \frac{P(T|C)P(C)}{P(T|C^c)P(C^c) + P(T|C)P(C)} = \frac{87\% \cdot 1.5\%}{9.6\% \cdot (1 - 1.5\%) + 87\% \cdot 1.5\%}$$

Ce qui donne P(C|T)$\approx$ 12.13 % et la bonne réponse est **la réponse E**.

2 Fléau de la dimensionalité et intuition géométrique en haute dimension

1. V $= c^d$ est le volume de l'hypercube de dimension d et de côté c.

2. Soit $\mathcal{C}$ l'hypercube. On a:

$$\forall x \in \mathbb{R}^d \; p(x) = \frac{1}{c^d} \mathbf{1}_{\{x \in \mathcal{C}\}}$$

En effet, on a bien $p : \mathbb{R}^d \to \mathbb{R}$ est une application borélienne positive telle que $P(\mathbb{R}^d) = 1$.

En considérant λ la mesure de Lebesgue sur $\mathbb{R}^d$, on a bien:

$$\begin{aligned} P(\mathbb{R}^d) &= \int_{\mathbb{R}^d} p(x)\lambda(dx) \\ &= \int_{\mathbb{R}^d} \frac{1}{c^d} \mathbf{1}_{\{x \in \mathcal{C}\}} \lambda(dx) \\ &= \frac{1}{c^d} \int_{\mathcal{C}} \lambda(dx) \\ &= \frac{1}{c^d} \lambda(\mathcal{C}) \end{aligned}$$

Or, on peut interpréter $\lambda(\mathcal{C})$ comme la mesure du volume d-dimensionnel de l'hypercube $\mathcal{C}$. Ainsi, cela donne bien $P(\mathbb{R}^d) = 1$.

3. On se place sur l'hypercube $\mathcal{C}$. Soit $\mathcal{B}$ la bordure du cube, on a $\mathcal{B} \subset \mathcal{C}$. On a donc:

$$P(\mathcal{B}) = \textstyle\int_{\mathcal{B}} p(x)\lambda(dx) = \frac{1}{c^d}\lambda(\mathcal{B})$$

Soit $\mathcal{C}' = \mathcal{C} \backslash \mathcal{B}$ l'hypercube plus petit. On a:

$$\lambda(\mathcal{B}) = \lambda(\mathcal{C}) - \lambda(\mathcal{C}') = c^d - (c - \frac{2 \cdot 3 \cdot c}{100})^d = c^d(1 - (\frac{47}{50})^d)$$

Ce qui donne alors:

$$P(\mathcal{B}) = 1 - (\frac{47}{50})^d \text{ et } P(\mathcal{C}') = 1 - P(\mathcal{B}) = (\frac{47}{50})^d$$

4. En faisant les applications numériques, on trouve:

   - Pour d=1: $P(\mathcal{B}) = 0.06$ et $P(\mathcal{C}') = 0.94$
   - Pour d=2: $P(\mathcal{B}) = 0.116$ et $P(\mathcal{C}') = 0.884$
   - Pour d=3: $P(\mathcal{B}) = 0.169$ et $P(\mathcal{C}') = 0.831$
   - Pour d=5: $P(\mathcal{B}) = 0.266$ et $P(\mathcal{C}') = 0.734$
   - Pour d=10: $P(\mathcal{B}) = 0.461$ et $P(\mathcal{C}') = 0.539$
   - Pour d=100: $P(\mathcal{B}) = 0.998$ et $P(\mathcal{C}') = 0.002$
   - Pour d=1000: $P(\mathcal{B}) \approx 1.0$ et $P(\mathcal{C}') \approx 0.0$

5. On a: $\lim_{d\to\infty} P(\mathcal{B}) = 1$ et $\lim_{d\to\infty} P(\mathcal{C}') = 0$.

   Ainsi, contrairement à notre intuition en petite dimension, le volume d-dimensionnel de la bordure de l'hypercube représente une part de plus en plus grande du volume totale de l'hybercube quand la dimension d augmente, à tel point que le "*volume intérieur*" de l'hypercube tant à représenter une part nulle du volume total quand d devient très grand.

3 Estimation de densité paramétrique Gaussienne, v.s. estimation de densité par fenêtres de Parzen

1. (a) On sait que la densité d'une gaussienne isotropique est, $\forall x \in \mathbb{R}^d$, $p(x) = \frac{1}{(2\pi)^{d/2}\sigma^d} \exp(-\frac{1}{2}\frac{\|x-\mu\|^2}{\sigma^2})$, avec les paramètre $\mu \in \mathbb{R}^d$ et $\sigma^2 \in \mathbb{R}$, qui sont, respectivement, la **moyenne** et la **variance**.

   (b) En exprimant en fonction de $D = \{x^{(1)}...x^{(n)}\}$ les estimateurs de maximum de vraissemblance $\hat{\mu}$ et $\hat{\sigma^2}$ de μ et σ^2, on obtient:

   $$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)} \text{ et } \hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x^{(i)} - \hat{\mu})(x^{(i)} - \hat{\mu})^T$$

   (c) L'estimateur $\hat{\mu}$ étant une moyenne sur toute les données, il faut faire un passage sur tout D qui contient n éléments. Or, chaque élément est de dimension d. Ainsi, le calcul de $\hat{\mu}$ se fait en $O(nd)$. Une fois $\hat{\mu}$ calculé, on peut s'ateler à $\hat{\sigma^2}$. Sachant que le produit scalaire de 2 vecteurs de dimension d se fait en $O(d)$, on a également que le calcul de $\hat{\sigma^2}$ est en $O(nd)$.

   (d) $\forall x \in \mathbb{R}^d$ test, on a $\hat{p}_{gauss\_isotrop}(x) = \frac{1}{(2\pi)^{d/2}\sqrt{\hat{\sigma^2}^d}} \exp(-\frac{1}{2}\frac{\|x-\hat{\mu}\|^2}{\hat{\sigma^2}})$

(e) Le calcul d'une norme au carré d'un vecteur de dimension d se faisant en $O(d)$, on en déduit que le calcul de cette prédiction pour tout nouveau point x est en $O(d)$.

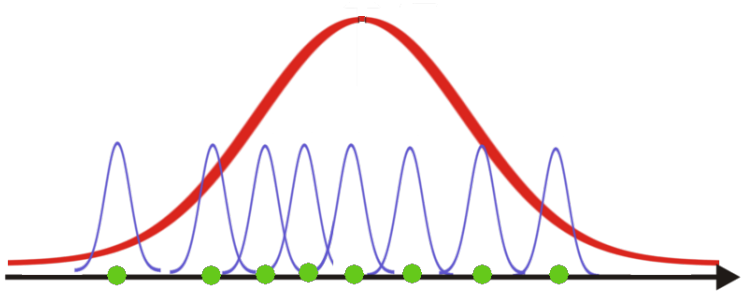
2. (a) Dans le cas de Parzen, l'écart type σ étant fixé, le seul paramètre à apprendre est donc le lieux des centroïdes des n gaussiennes isotropiques que l'on va sommer, ie le vecteur moyenne $\mu = (\mu_1 ... \mu_n)$. Or, dans le cas de Parzen, les centroïdes sont pris sur les points constituants le jeu de données d'apprentissage.
Ainsi la phase d'apprentissage consiste juste à dire $\mu = D$.

   (b) $\forall x \in \mathbb{R}^d$ test, on a $\hat{p}_{Parzen}(x) = \frac{1}{n} \frac{1}{(2\pi)^{d/2}\sigma^d} \sum_{i=1}^{n} \exp(-\frac{1}{2\sigma^2} d(x^{(i)}, x)^2)$

   (c) Le calcul de distance $d(x^{(i)}, x)$ se faisant en $O(d)$ $\forall x \in \mathbb{R}^d$ et $\forall i \in [\![1, n]\!]$, on en déduit que le calcul de la prédiction pour tout nouveau point x est en $O(nd)$.

3. (a) Le modèle le plus "*expressif*" est le modèle de Parzen. En effet, il "*colle*" beaucoup plus aux données d'apprentissage, générant une gaussienne pour chaque point de D tandis que le modèle gaussien est plus "*général*" et ne génère qu'une seule gaussienne qui est censé représenter la tendance globale de D. Ainsi, si la distribution réelle de D est non gaussienne, le modèle de Parzen aura une chance de le capturer (avec suffisamment de données et un bon choix de σ) tandis que cela sera impossible pour le modèle gaussien.

   (b) Dans le cas ou on a un σ **petit** et **des données trop espacées**, le modèle de Parzen se généralise mal à de nouvelles données et overfitt: les différentes gaussiennes sont tellement piquées et espacées qu'elles ne s'additionnent pas vraiment et la distribution résultante n'est qu'une collection de pics.



Sur cet exemple, la distribution rouge
est une distribution acceptable au vu des données fournies,
mais le modèle de Parzen n'arrive à produire que la distribution bleu.

(c) Dans le cas de Parzen, σ n'est pas estimé directement à partir des données comme dans le cas gaussien isotropique, mais σ est un "*à priori*" de l'utilisateur, ou représente sa volonté de modélisation. C'est une valeur qui n'est pas dérivée directement du jeu de données mais que le modélisateur doit fournir, d'où le mot "hyper-paramètre".

4. (a) Soit $X = (X_1, ...X_d)$ le vecteur aléatoire à valeur dans $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ (où $\mathcal{B}(\mathbb{R}^d)$ désigne la tribu borélienne sur $\mathbb{R}^d$) et admettant une densité gaussienne diagonale $p_{gauss-diag}$.

Les paramètres d'une telle densité sont le vecteur moyenne $\mu = \begin{pmatrix} \mu_1 \\ ... \\ \mu_d \end{pmatrix}$ et la matrice de covariance $\Sigma = \begin{pmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_d^2 \end{pmatrix}$.

X admet une densité si et seulement si la matrice de covariance a un déterminant non nul, à savoir $(\sigma_i)_{i \in [\![1,d]\!]}$ tous non nuls.

On a alors, $\forall\ x = \begin{pmatrix} x_1 \\ ... \\ x_d \end{pmatrix}$:

$$\begin{aligned} p_{gauss-diag}(x) &= \frac{1}{(2\pi)^{(d/2)}\sqrt{\det(\Sigma)}} \exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)) \\ &= \frac{1}{(2\pi)^{(d/2)}\sigma_1 \cdot ... \cdot \sigma_d} \exp(-\frac{1}{2}\sum_{j=1}^{d}\frac{1}{\sigma_j^2}(x_j-\mu_j)^2) \end{aligned}$$

(b) On veut montrer que les $(X_i)_{i\in[\![1,d]\!]}$, famille de variable aléatoires à valeur dans $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, sont indépendants.

C'est à dire, pour toute famille de boréliens $(A_i)_{i\in[\![1,d]\!]}$, on veut montrer que:

$$\forall J \subset [\![1,d]\!], P(X_j \in A_j; \forall j \in J) = \prod_{j\in J} P(X_j \in A_j)$$

Or, on sait que $X = (X_1, ...X_d)$ admet la densité de probabilité $p_{gauss-diag} : (x_1, ..., x_d) \to p_{gauss-diag}(x_1, ..., x_d)$.

Ainsi, on a:

$$\forall A \in \mathcal{B}(\mathbb{R}^d), P(X \in A) = \int_A p_{gauss-diag}(x_1, ..., x_d)dx_1...dx_d$$

En particulier, on peut considérer:

$$A^* = A_{j_1} \times ... \times A_{j_{|J|}} \times \mathbb{R}^{d-|J|}$$

Or, on sait que $\forall x \in \mathbb{R}^d$:

$$\begin{aligned} p_{gauss-diag}(x) &= \frac{1}{(2\pi)^{(d/2)}\sigma_1 \cdot ... \cdot \sigma_d} \exp(-\frac{1}{2}\sum_{j=1}^{d} \frac{1}{\sigma_j^2}(x_j - \mu_j)^2) \\ &= \frac{1}{(2\pi)^{(d/2)}\sigma_1 \cdot ... \cdot \sigma_d} \prod_{j=1}^{d} \exp(-\frac{1}{2}\frac{1}{\sigma_j^2}(x_j - \mu_j)^2) \\ &= \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi}\sigma_j} \exp(-\frac{(x_j - \mu_j)^2}{2\sigma_j^2}) \end{aligned}$$

Ainsi, on a:

$$P(X \in A^*) = \int_{A^*} \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi}\sigma_j} \exp(-\frac{(x_j - \mu_j)^2}{2\sigma_j^2})dx_1...dx_d$$

Ce qui donne bien $P(X_j \in A_j; \forall j \in J) = \prod_{j\in J} P(X_j \in A_j)$ et les $(X_i)_{i\in[\![1,d]\!]}$ sont indépendants.

(c) Apprendre les paramètres en minimisant le risque empirique sur l'ensemble D revient à chercher
$argmin_{\mu,\sigma^2} \frac{1}{|D|} \sum_{x\in D} cout(x) = argmin_{\mu,\sigma^2} \frac{1}{n} \sum_{x\in D} -\log(p(x)) =$
$argmin_{\mu,\sigma^2} \frac{1}{n} \sum_{i=1}^{n} \{\frac{d}{2}\log(2\pi) + \frac{1}{2}\log(\prod_{j=1}^{d} \sigma_j^2) + \frac{1}{2}\sum_{j=1}^{d} \frac{1}{\sigma_j^2}(x_{i,j} - \mu_j)^2\}$

(d) Nommons la fonction que l'on cherche à minimiser l. On a:

$$l(\mu,\sigma) = \frac{d}{2}\log(2\pi) + \sum_{j=1}^{d}\log(\sigma_j) + \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{d}\frac{1}{\sigma_j^2}(x_{i,j}-\mu_j)^2$$

Ainsi, la fonction l qui à (μ,σ) associe $l(\mu,\sigma)$ est définie et **différentiable** de $\mathbb{R}^d \times \mathbb{R}_+^{*d}$ dans $\mathbb{R}$. On peut donc dériver l par rapport aux μ_j et σ_j pour trouver les extrema locaux. On a alors:

$$\forall j \in [\![1,d]\!] \begin{cases} \dfrac{\partial l}{\partial \sigma_j} = \dfrac{1}{\sigma_j} - \dfrac{1}{n\sigma_j^3}\sum_{i=1}^{n}(x_{i,j}-\mu_j)^2 \\ \dfrac{\partial l}{\partial \mu_j} = \dfrac{\mu_j}{\sigma_j^2} - \dfrac{1}{n\sigma_j^2}\sum_{i=1}^{n}x_{i,j} \end{cases}$$

l étant convexe par rapport à μ, l'argument μ^* qui minimise l est celui trouvé en trouvant les μ_j^* annulant les $\dfrac{\partial l}{\partial \mu_j}$ et ainsi:

$$\mu^* = \begin{pmatrix} \mu_1^* \\ ... \\ \mu_d^* \end{pmatrix} \text{ avec } \forall j \in [\![1,d]\!], \mu_j^* = \frac{1}{n}\sum_{i=1}^{n}x_{i,j}$$

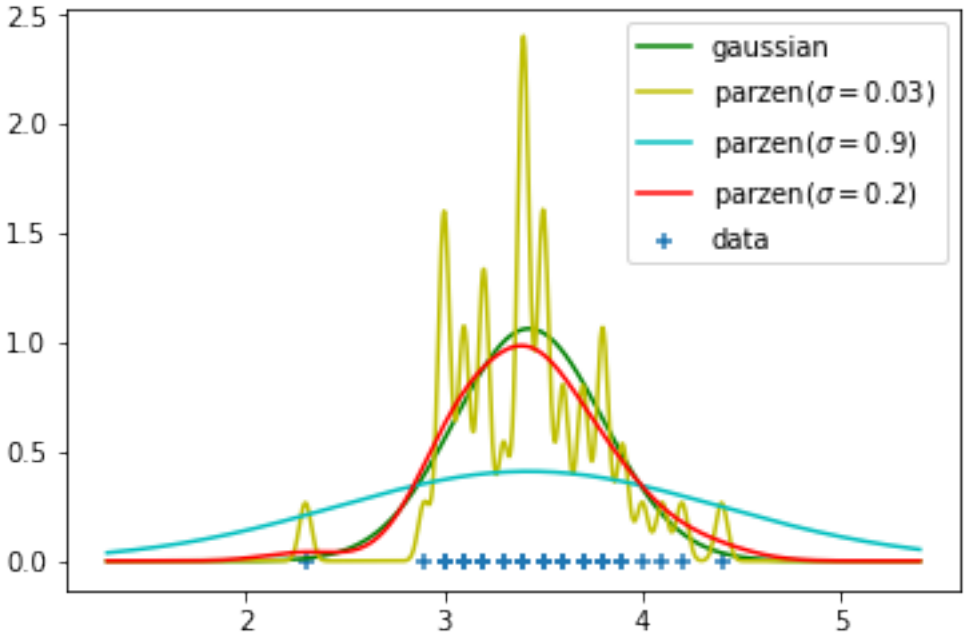
En annulant les $\dfrac{\partial l}{\partial \sigma_j}$, on trouve:

$$\sigma^{2*} = \begin{pmatrix} \sigma_1^{2*} \\ ... \\ \sigma_d^{2*} \end{pmatrix} \text{ avec } \forall j \in [\![1,d]\!], \sigma_j^{2*} = \frac{1}{n}\sum_{i=1}^{n}(x_{i,j}-\mu_j^*)^2$$

Or, l décroit sur $]0,\sigma^*]$ et croit sur $[\sigma^*,+\infty[$, ainsi σ^* est bien l'argument qui minimise l.

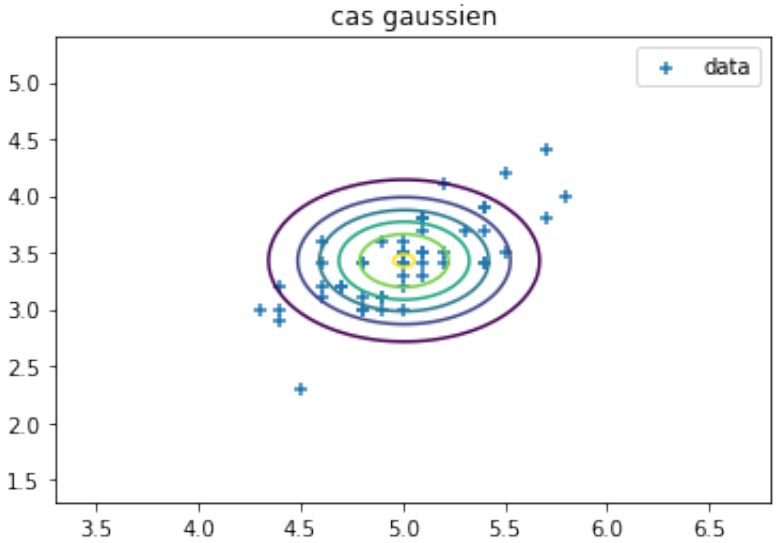
4 Partie pratique: estimation de densité

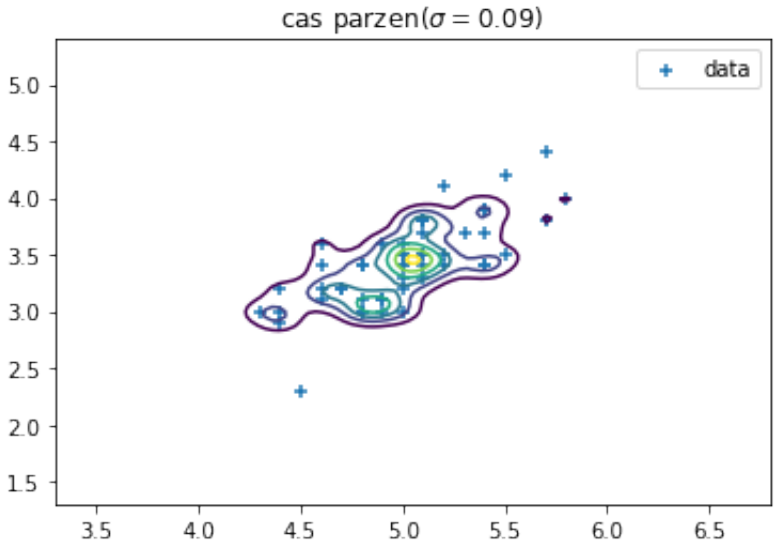
4.3

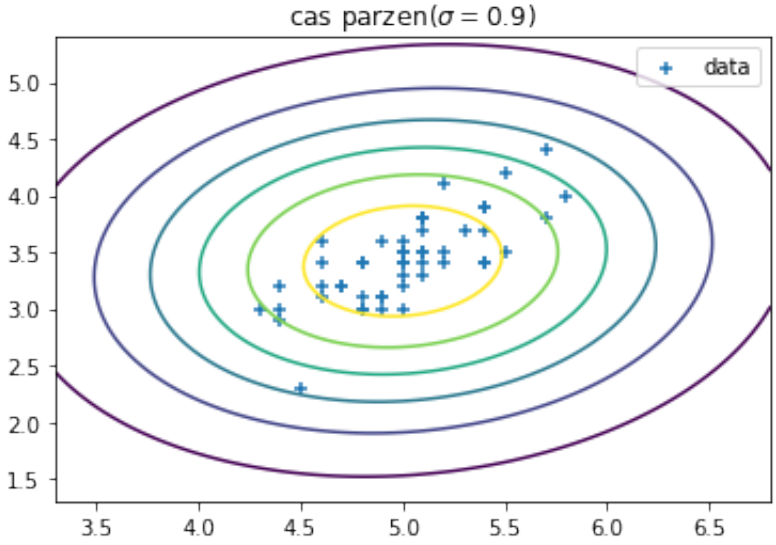


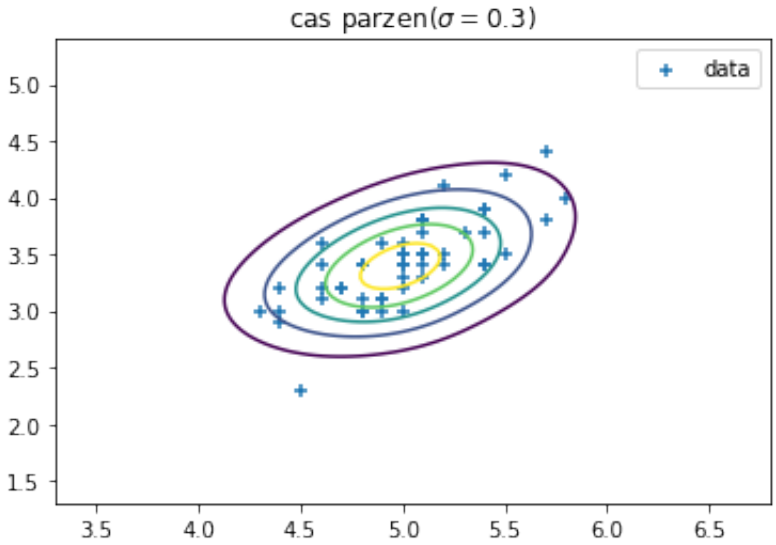
Pour choisir σ, nous avons "tâtonné" jusqu'à trouver une valeur qui semblait produire un modèle qui semble bien se généraliser tout en évitant d'être trop "large". Pour voir cela, nous avons utiliser le graphique produit par notre code.

4.4


cas gaussien
data
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
3.5
4.0
4.5
5.0
5.5
6.0
6.5


cas parzen($\sigma = 0.09$)
data
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
3.5
4.0
4.5
5.0
5.5
6.0
6.5





De la même manière que précédemment, on choisi notre σ en tâtonnant et grâce au graphique.