

## Big Data

### TD 2 : Grande dimension en probabilités et statistiques BUT 3

Guillaume Metzler et Antoine Rolland

Institut de Communication (ICOM)


Université de Lyon, Université Lumière Lyon 2

Laboratoire ERIC UR 3083, Lyon, France

[guillaume.metzler@univ-lyon2.fr](mailto:guillaume.metzler@univ-lyon2.fr); [antoine.rolland@univ-lyon2.fr](mailto:antoine.rolland@univ-lyon2.fr)

On s'intéresse maintenant aux comportements et aux limites de la grande dimension en probabilités et statistiques. Pour cela, on prendra les exemples des intervalles de confiance, de la régression linéaire et de la loi gaussienne multivariée.

#### Exercice 1 : Intervalle de confiance et limite numérique

Les outils informatiques sont actuellement dotés d'une grande précision numérique, mais la précision n'est pour autant infinie comme elle pourrait l'être humain. Ainsi des logiciels comme  ou Python ne parviennent pas à distinguer la différence entre deux nombres réels est inférieure à  $10^{-16}$ . On se propose l'impact de cette précision numérique dans les intervalles de confiance.

1. Etant donné un  $n$  échantillon  $X_n$  de  $\mathcal{N}(\theta, 1)$ . Rappeler la formule de l'estimateur  $\hat{\theta}_n$  de  $\theta$  par la méthode de votre choix.

L'estimateur de la moyenne  $\hat{\theta}_n$  peut être obtenu par la méthode des moments ou encore par maximum de vraisemblance. La relation ainsi obtenue, classique, est

$$\hat{\theta}_n = \overline{X_n}.$$

2. Soit  $\alpha \in ]0, 1[$  une probabilité, rappeler la formule de la taille du demi-intervalle de confiance de niveau  $1 - \alpha$  de  $\theta$ , basée sur l'estimateur  $\hat{\theta}_n$ .

Nous sommes dans le cas où la variance de la distribution des données est connue. L'intervalle de confiance repose donc sur la loi normale centrée et réduite :

$$IC_{1-\alpha} = \left[ \hat{\theta}_n - z_{1-\alpha/2} \frac{1}{\sqrt{n}}, \hat{\theta}_n + z_{1-\alpha/2} \frac{1}{\sqrt{n}} \right].$$

Ainsi, la longueur de l'intervalle de confiance est égale à  $2 \frac{z_{1-\alpha/2}}{\sqrt{n}}$  ou encore  $\frac{z_{1-\alpha/2}}{\sqrt{n}}$  pour la longueur du demi-intervalle.

Pourquoi cela ne changera pas grand chose de considérer la loi de student ou la loi normale dans le cas où  $n$  est très grand ?

3. Déterminer à partir de quelle taille  $n$  de l'échantillon, la longueur de l'intervalle de confiance passe en dessous de cette précision numérique.

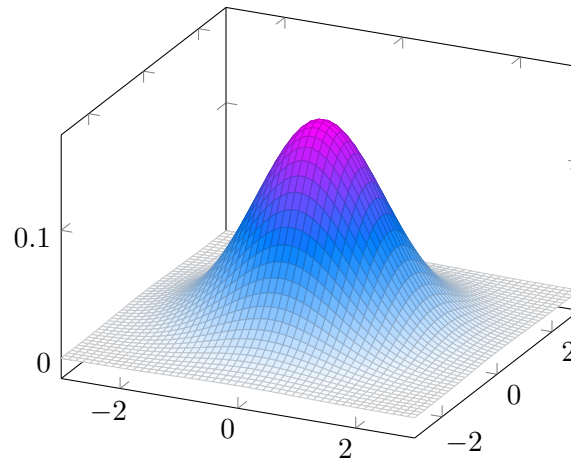
Une rapide application numérique nous donne

$$\frac{2}{\sqrt{n}} = 10^{-16} \iff n = (2 \cdot 10^{16})^2 = 4 \cdot 10^{32}$$

A titre de comparaison : le nombre de secondes par an est de  $10^8$ , le nombre de requête google par jour de  $10^{10}$ ... on est loin de  $10^{32}$  ! D'autres exemples ont également été donnés dans le TD précédent.

## Exercice 2 : Loi gaussienne en grande dimension

Dans cet exercice, on va montrer que, contrairement à ce que l'on pourrait penser, toute la masse d'une loi gaussienne multivariée se trouve dans les queues de distribution. Si cela n'est pas perceptible en faible dimension, où l'on aperçoit que toute la masse est concentrée autour de la moyenne, comme le montre le graphique ci-dessous, on montre que cette masse sera plus diffuse lorsque la dimension  $p$  augmente.



1. Donner la densité  $f_p$  d'une loi gaussienne multivariée de dimension  $p$  de moyenne nulle et de matrice de covariance égale à l'identité.

On commencera par rappeler la densité de la gaussienne centrée et réduite (en dimension 1) et on pourra essayer de généraliser en prenant garde au fait que les objets manipulés sont des vecteurs.

La densité  $f_p$  est donnée par

$$f_p(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}} e^{-\frac{1}{2}\|\mathbf{x}\|^2}$$

2. Evaluer cette densité en le vecteur nul et étudier sa limite lorsque  $p$  tend vers l'infini. Que constatez vous ?

On obtient  $f_p(\mathbf{0}) = \frac{1}{(2\pi)^{p/2}}$  dont la limite est 0 en  $p = \infty$ .

3. On va maintenant montrer que toute la masse se trouve dans les queues de la distribution, c'est-à-dire loin du "pic" de notre gaussienne.

Pour cela on considère l'ensemble  $B_{p,\delta}$ ,  $\delta > 0$ , défini par

$$B_{p,\delta} = \{\mathbf{x} \in \mathbb{R}^p \mid f_p(\mathbf{x}) \geq \delta f_p(\mathbf{0})\}.$$

Plus la valeur de  $\delta$  est faible, plus l'ensemble  $B_{p,\delta}$  est grand.

On va ensuite regarder la probabilité qu'un élément appartienne à cet ensemble

- (a) Montrer que l'on a

$$B_{p,\delta} = \{\mathbf{x} \in \mathbb{R}^p \mid \|\mathbf{x}\|^2 \leq 2 \ln(1/\delta)\}.$$

$$\begin{aligned} f_p(\mathbf{x}) &\geq \delta f_p(\mathbf{0}), \\ \iff \frac{1}{(2\pi)^{p/2}} e^{-\frac{1}{2}\|\mathbf{x}\|^2} &\geq \delta \frac{1}{(2\pi)^{p/2}}, \\ \iff e^{-\frac{1}{2}\|\mathbf{x}\|^2} &\geq \delta, \\ \iff -\frac{1}{2}\|\mathbf{x}\|^2 &\geq \ln(\delta), \\ \iff \|\mathbf{x}\|^2 &\leq -2 \ln(\delta). \end{aligned}$$

- (b)

- (c) A l'aide de l'inégalité de Markov et en utilisant le fait que  $\int_{\mathbf{x} \in \mathbb{R}^p} e^{-\|\mathbf{x}\|^2} d\mathbf{x} = \pi^{p/2}$ , montrer que si  $X$  est une variable aléatoire gaussienne multivariée de moyenne nulle et de covariance égale à  $I_p$ , on a

$$\mathbb{P}[X \in B_{p,\delta}] \leq \frac{1}{\delta \times 2^{p/2}}.$$

On rappelle que l'inégalité de Markov, nous dit que pour tout  $\varepsilon > 0$  et pour tout variable aléatoire  $X$  qui admet un moment d'ordre 1, nous avons

$$\mathbb{P}[X \geq \varepsilon] \leq \frac{\mathbb{E}[X]}{\varepsilon}.$$

Ainsi, nous avons

$$\begin{aligned} \mathbb{P}[X \in B_{p,\delta}] &= \mathbb{P}\left[e^{-\frac{1}{2}\|\mathbf{x}\|^2} \geq \delta\right], \\ &\downarrow \text{on applique l'inégalité de Markov} \\ &= \frac{1}{\delta} \mathbb{E}\left[e^{-\frac{1}{2}\|\mathbf{x}\|^2}\right], \\ &\downarrow \text{théorème de transfert} \\ &= \frac{1}{\delta} \int_{\mathbf{x} \in \mathbb{R}^p} e^{-\|\mathbf{x}\|^2} \times \frac{d\mathbf{x}}{(2\pi)^{p/2}}, \\ &\downarrow \text{on utilise l'indication sur la valeur de l'intégrale} \\ &= \frac{1}{2^{p/2}\delta}. \end{aligned}$$

Interpréter le résultat lorsque  $p$  tend vers  $\infty$ .

Cette probabilité, pour tout  $\delta > 0$ , converge vers 0 lorsque  $p$  tend vers  $+\infty$ . Ce qui montre bien que toute la masse se trouve dans les queues de la distribution de la gaussienne.

## Exercice 3 : Modèle linéaire et erreur d'estimation

En régression linéaire, nous cherchons à prédire les valeurs prises par une variable aléatoire réelle  $Y$ , dite *indépendante*, en fonctions de plusieurs variables aléatoires réelles  $X_1, \dots, X_p$  par la relation linéaire suivante

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon,$$

où  $\varepsilon$  désigne un bruit gaussien de moyenne nulle et de variance inconnue  $\sigma^2$  et  $\beta_0, \dots, \beta_p$  désigne les paramètres du modèle.

Pour estimer les paramètres du modèle, on dispose d'un échantillon  $S = \{(y_i, \mathbf{x}_i)\}_{i=1}^n$  où  $y_i \in \mathbb{R}$  et  $\mathbf{x}_i \in \mathbb{R}^{p+1}$ . On notera alors  $\mathbf{X}$  la matrice des données (ou matrice de *design*). Notre modèle linéaire peut alors se réécrire sous la forme

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

où  $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$  désigne le vecteur des paramètres du modèle et  $\boldsymbol{\varepsilon} \in \mathbb{R}^n$  désigne le vecteur aléatoire des erreurs dont les entrées sont *i.i.d.* selon une loi  $\mathcal{N}(\mathbf{0}, \sigma^2)$ .

1. Donner l'expression de l'estimateur  $\hat{\boldsymbol{\beta}}$  de  $\boldsymbol{\beta}$ . A quelle condition cet estimateur est bien défini ?

On a

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

Pour être défini il faut vérifier que la matrice  $\mathbf{X}^\top \mathbf{X}$  est bien inversible. Ce qui est le cas si les deux conditions suivantes sont réunies :

- $p \leq n$
- les colonnes de la matrice sont indépendantes, *i.e.* il n'existe aucune relation linéaire entre les colonnes de la matrice.

2. Montrer qu'il s'agit d'un estimateur sans biais et que sa variance est égale à  $\sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}$ .

**(i) Espérance de l'estimateur :  $\mathbb{E}[\hat{\boldsymbol{\beta}}]$**  On repart de la définition de l'estimateur et on garde l'esprit que  $X$  est **déterministe** (ce n'est pas une variable aléatoire), la seule partie aléatoire dans l'expression de  $\hat{\boldsymbol{\beta}}$  vient de la variable aléatoire  $Y$  via la variable aléatoire  $\boldsymbol{\varepsilon}$ .

$$\begin{aligned} \mathbb{E}[\hat{\boldsymbol{\beta}}] &= \mathbb{E} \left[ (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top Y \right], \\ &\quad \downarrow \text{seule } \mathbf{y} \text{ est aléatoire} \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbb{E}[\mathbf{y}], \\ &\quad \downarrow \text{par définition de } \mathbf{y} \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbb{E}[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}], \\ &\quad \downarrow \text{seule } \boldsymbol{\varepsilon} \text{ est aléatoire} \end{aligned}$$

$$\begin{aligned}
&= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X} \boldsymbol{\beta} + \mathbb{E}[\boldsymbol{\varepsilon}], \\
&\quad \downarrow \text{ car } \mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}, \text{ hypothèse sur les erreurs du modèle} \\
&= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X} \boldsymbol{\beta}, \\
\mathbb{E}[\hat{\boldsymbol{\beta}}] &= \boldsymbol{\beta}.
\end{aligned}$$

**(ii) Variance de l'estimateur**  $\text{Var}[\hat{\boldsymbol{\beta}}]$  On garde à l'esprit ce que nous avons utilisés précédemment, à savoir que seule  $\boldsymbol{\varepsilon}$ , *i.e.* les erreurs sont aléatoires ainsi que les propriétés de la variance.

$$\begin{aligned}
\text{Var}[\hat{\boldsymbol{\beta}}] &= \mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top], \\
&\quad \downarrow \text{ On repart de la définition de } \hat{\boldsymbol{\beta}} \\
&= \mathbb{E}\left[\left((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} - \boldsymbol{\beta}\right)\left((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} - \boldsymbol{\beta}\right)^\top\right], \\
&\quad \downarrow \text{ définition de } Y \\
&= \mathbb{E}\left[\left((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}) - \boldsymbol{\beta}\right)\left((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}) - \boldsymbol{\beta}\right)^\top\right], \\
&\quad \downarrow \text{ on développe et on simplifie les expressions} \\
&= \mathbb{E}\left[\left((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\varepsilon}\right)\left((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\varepsilon}\right)^\top\right], \\
&\quad \downarrow \text{ par définition de la transposition} \\
&= \mathbb{E}\left[\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top \mathbf{X} \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\right], \\
&\quad \downarrow \text{ seule la partie en } \boldsymbol{\varepsilon} \text{ est aléatoire} \\
&= \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbb{E}\left[\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top\right] \mathbf{X} \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}, \\
&\quad \downarrow \text{ or } \mathbb{E}\left[\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top\right] = \sigma^2 \mathbf{I}_n \\
&= \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{X} \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}, \\
&\quad \downarrow \text{ par simplification} \\
\text{Var}[\hat{\boldsymbol{\beta}}] &= \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}.
\end{aligned}$$

3. Dans cette question, on fait l'hypothèse que les colonnes de la matrice  $\mathbf{X}$  sont orthogonales.

On cherche à évaluer l'erreur quadratique moyenne d'estimation définie par

$$\mathbb{E}[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2].$$

(a) Montrer que l'on

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\varepsilon}.$$

Il suffit de reprendre la définition de l'estimateur  $\hat{\boldsymbol{\beta}}$ .

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top Y, \\
&= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}), \\
&= \boldsymbol{\beta} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\varepsilon},
\end{aligned}$$

(b) Montrer que si  $\mathbf{x}$  et  $\mathbf{x}'$  de  $\mathbb{R}^n$  sont deux vecteurs alors on

$$\langle \mathbf{x}, \mathbf{x}' \rangle = \text{Tr}(\mathbf{x}^\top \mathbf{x}') = \text{Tr}(\mathbf{x}' \mathbf{x}^\top),$$

où  $\langle \cdot, \cdot \rangle$  désigne le produit scalaire et  $\text{Tr}$  désigne la *trace*.

Il suffit de faire le calcul en prenant de vecteurs  $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$ .

- (c) En déduire que l'erreur quadratique moyenne d'estimation est égale à  $\sigma^2(p+1)$ . Commenter l'évolution de l'erreur d'estimation en fonction de la dimension du problème.

$$\begin{aligned} \mathbb{E}[\|\hat{\beta} - \beta\|^2] &= \mathbb{E}[\|\mathbf{X}^\top \mathbf{X}\|^{-1} \mathbf{X}^\top \varepsilon\|^2], \\ &= \mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \varepsilon)^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \varepsilon)], \\ &\quad \downarrow \text{on utilise l'indication} \\ &= \mathbb{E}[\text{Tr} \left( (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \varepsilon \varepsilon^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \right)], \\ &\quad \downarrow \text{On peut permuter trace et espérance} \\ &= \text{Tr} \left( \mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \varepsilon \varepsilon^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1}] \right), \\ &\quad \downarrow \text{seule } \varepsilon \text{ est aléatoire} \\ &= \text{Tr} \left( (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \underbrace{\mathbb{E}[\varepsilon \varepsilon^\top]}_{\text{variance de } \varepsilon} \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \right), \\ &\quad \downarrow \text{définition de la variance de } \varepsilon \\ &= \text{Tr} \left( (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \sigma^2 \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \right), \\ &= \sigma^2 \text{Tr} \left( (\mathbf{X}^\top \mathbf{X})^{-1} \right), \\ &\quad \downarrow \text{car les colonnes de } \mathbf{X} \text{ sont orthogonales} \\ &= (p+1)\sigma^2 \end{aligned}$$