

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; 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.

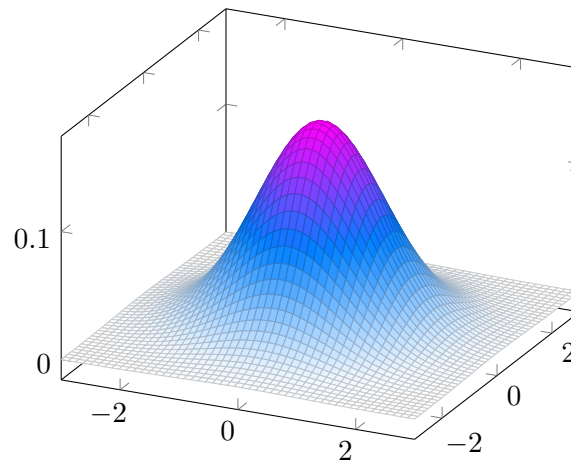
Exercice 1 : Intervalle de confiance et limite numérique

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 $\mathcal{N}(\theta, 1)$. Donner un estimateur $\hat{\theta}_n$ de θ par la méthode de votre choix.
2. Soit $\alpha \in]0, 1[$ une probabilité, déterminer un intervalle de confiance de niveau $1 - \alpha$ de θ , basée sur l'estimateur $\hat{\theta}_n$. Quelle est la longueur de cet intervalle de confiance ?
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.

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

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

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

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

Interpréter le résultat lorsque p tend vers ∞ .

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 ε désigne un bruit gaussien de moyenne nulle et de variance inconnue σ^2 et β_0, \dots, β_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 ?
2. Montrer qu'il s'agit d'un estimateur sans biais et que sa variance est égale à $\sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}$.
3. 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 Tr désigne la *trace*.

4. 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

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

- (b) 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.