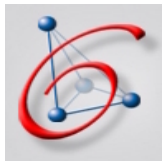
Master 2 Informatique

Probabilistic Learning and Data Analysis

Faicel Chamroukhi
Maître de Conférences
UTLN, LSIS UMR CNRS 7296




email: chamroukhi@univ-tln.fr
web: chamroukhi.univ-tln.fr

Overview

Estimation de paramètres

Estimation de paramètres

Pour étudier et caractériser un phénomène physique, naturel ou autre $\Rightarrow$ e.g., adoption d'un modèle probabiliste paramétrique représenté par une fonction de densité de probabilité $f(x;\theta)$ (ou une fonction de masse de probabilité $P(x;\theta)$ dans le cas discret).

$\Rightarrow$ L'explication de ce phénomène nécessite l'estimation de ce modèle probabiliste à partir des données que l'on a observées (l'échantillon que l'on a à notre disposition).

$\Rightarrow$ Ceci consiste donc à estimer le(s) paramètre(s) θ de ce modèle à partir des données observées $(x_1, \ldots, x_n)$ (i.i.d dans le cadre de ce cours.)

Estimation de paramètres

Définition d'un estimateur

Le problème d'estimation de paramètres est donc celui de déterminer une fonction appropriée des données $(x_1, \ldots, x_n)$, que nous noterons $h(x_1, \ldots, x_n)$ qui donne la "meilleure" estimation de θ au sens de critères d'optimalité que nous verrons.
Nous avons donc

$$\hat{\theta} = h(x_1, \ldots, x_n)$$

et plus généralement, sous forme de variable aléatoire (car en effet pour des nouvelles réalisation des X_i, la valeur de $\hat{\theta}$ change) :

$$\hat{\Theta} = h(X_1, \ldots, X_n).$$

Cette statistique à déterminer s'appelle *un estimateur*.

Critères de qualité pour les estimateurs

- Ce sont des critères selon lesquels la qualité d'une estimation peut être évaluée.
- Ces critères définissent en général des propriétés souhaitables pour un estimateur et fournissent un guide par lequel la qualité d'un estimateur peut être comparée à celle d'un autre.

$\Rightarrow$ Notre objectif est de déterminer un estimateur $\hat{\Theta} = h(X_1, \ldots, X_n)$ de θ.

$\Rightarrow$ Des propriétés comme la moyenne, la variance ou la distribution fournissent une mesure de qualité pour cet estimateur.

Critères de qualité pour les estimateurs

- Ce sont des critères selon lesquels la qualité d'une estimation peut être évaluée.
- Ces critères définissent en général des propriétés souhaitables pour un estimateur et fournissent un guide par lequel la qualité d'un estimateur peut être comparée à celle d'un autre.

$\Rightarrow$ Notre objectif est de déterminer un estimateur $\hat{\Theta} = h(X_1, \ldots, X_n)$ de θ.

$\Rightarrow$ Des propriétés comme la moyenne, la variance ou la distribution fournissent une mesure de qualité pour cet estimateur.

Estimateur vs Estimation

Une fois nous avons observé un échantillon de valeurs $(x_1, \ldots, x_n)$, la valeur de l'estimateur $\hat{\theta} = h(x_1, \ldots, x_n)$ qui est une valeur numérique, est appelé *estimation* du paramètre θ.

Absence de biais

Définition : **Absence de biais**

Un estimateur $\hat{\Theta}$ de θ est dis *sans biais* si

$$\mathbb{E}[\hat{\Theta}] = \theta, \tag{1}$$

$\Rightarrow$ en moyenne, on espère que $\hat{\Theta}$ est égal à la valeur du vrai paramètre θ.

Absence de biais

Définition : **Absence de biais**

Un estimateur $\hat{\Theta}$ de θ est dis *sans biais* si

$$\mathbb{E}[\hat{\Theta}] = \theta, \tag{1}$$

$\Rightarrow$ en moyenne, on espère que $\hat{\Theta}$ est égal à la valeur du vrai paramètre θ.

⚠ Remarque : Il est naturel que, si $\hat{\Theta}$ est à qualifier comme un bon estimateur de θ , non seulement sa moyenne doit être très proche du vrai paramètre θ mais aussi il faudrait qu'il y ait une grande probabilité que toute valeur $\hat{\theta}$ soit très proche de θ.

$\Rightarrow$ Cela revient à sélectionner un estimateur de façon à ce que non seulement il soit sans biais mais aussi sa variance soit la plus petite possible.

Variance minimale

Définition **Variance minimale**

Soit $\hat{\Theta}$ un estimateur sans biais de θ. Il est dit à variance minimale pour θ si, pour tout autre estimateur sans biais $\Theta^\star$ de θ, à partir du même échantillon, on a :

$$\text{var}(\hat{\Theta}) \leq \text{var}(\Theta^\star). \tag{2}$$

Variance minimale

Définition **Variance minimale**

Soit $\hat{\Theta}$ un estimateur sans biais de θ. Il est dit à variance minimale pour θ si, pour tout autre estimateur sans biais $\Theta^{\star}$ de θ, à partir du même échantillon, on a :

$$\text{var}(\hat{\Theta}) \leq \text{var}(\Theta^{\star}). \tag{2}$$

$\Rightarrow$ Étant donné deux estimateurs sans biais pour un paramètre donné, celui ayant la variance plus faible est préférable, car une plus petite variance implique que les estimations ont tendance à être plus proche de sa moyenne qui est la valeur du vrai paramètre.

Variance minimale

Définition **Variance minimale**

Soit $\hat{\Theta}$ un estimateur sans biais de θ. Il est dit à variance minimale pour θ si, pour tout autre estimateur sans biais $\Theta^\star$ de θ, à partir du même échantillon, on a :

$$\text{var}(\hat{\Theta}) \leq \text{var}(\Theta^\star). \tag{2}$$

$\Rightarrow$ Étant donné deux estimateurs sans biais pour un paramètre donné, celui ayant la variance plus faible est préférable, car une plus petite variance implique que les estimations ont tendance à être plus proche de sa moyenne qui est la valeur du vrai paramètre.

$\Rightarrow$ La question qui se pose donc est, étant donné un échantillon à partir duquel on construit plusieurs estimateurs sans biais, le quel parmi tout ces estimateurs qui a la variance minimale ? $\Rightarrow$ Thèorème : Borne de Cramer-Rao

Variance minimale : Borne de Cramer-Rao

Theorem (Borne de Cramer-Rao (Cramer-Rao Lower Bound (CRLB)))

Soit $(X_1, \ldots, X_n)$ *un échantillon de v.a issues d'une population de densité* $f(x;\theta)$ *où* θ *est le paramètre inconnu, et soit* $\hat{\Theta} = h(X_1, \ldots, X_n)$ *un estimateur sans biais pour* θ*. La variance de* $\hat{\Theta}$ *satisfait l'inégalité suivante*

$$\text{var}(\hat{\Theta}) \geq \left[n\mathbb{E}\left(\frac{\partial \ln f(X;\theta)}{\partial \theta} \right)^2 \right]^{-1} \tag{3}$$

⚠ Remarque : si l'espérance et la dérivée existent. Un résultat analogue avec $p(X;\theta)$ en remplaçant $f(X;\theta)$ est obtenue lorsque X est discrète.

$\Rightarrow$ Cette inéquation fournit donc une borne inférieure de la variance de n'importe quel estimateur sans biais.

Variance minimale : Borne de Cramer-Rao

Information de Fisher

La quantité $n\mathbb{E}\left(\frac{\partial \ln f(X;\theta)}{\partial \theta}\right)^2$ s'appelle l'*information de Fisher* contenue dans un échantillon de taille n et se note $\mathcal{I}_n(\theta)$.

Variance minimale : Borne de Cramer-Rao

Information de Fisher

La quantité $n\mathbb{E}\left(\frac{\partial \ln f(X;\theta)}{\partial \theta}\right)^2$ s'appelle l'*information de Fisher* contenue dans un échantillon de taille n et se note $\mathcal{I}_n(\theta)$.

Borne de Cramer-Rao et Information de FIsher

La borne inférieure de Cramér-Rao alors se définie aussi par :

$$\text{var}(\hat{\Theta}) \geq \frac{1}{\mathcal{I}_n(\theta)} \tag{4}$$

et énonce donc que l'inverse de l'information de Fisher, $\mathcal{I}_n(\theta)$, d'un paramètre θ, est une borne inférieure de la variance d'un estimateur sans biais de ce paramètre.

⚠ Remarque : En anglais, la borne inférieure de Cramér-Rao s'appelle **Cramér-Rao Lower Bound** abréviée par CRLB.

Variance minimale : Borne de Cramér-Rao

Deuxième forme opérationnelle. Si le modèle est régulier, l'espérance $\left[\mathbb{E}\left(\frac{\partial \ln f(X;\theta)}{\partial\theta}\right)^2\right]^{-1}$ dans (3) est équivalente à $-\left[\mathbb{E}\left(\frac{\partial^2 \ln f(X;\theta)}{\partial^2\theta}\right)\right]^{-1}$.

$\Rightarrow$ L'inégalité de Cramér-Rao peut également être alors mise sous la forme :

CRLB : deuxième forme opérationnelle

$$\text{var}(\hat{\Theta}) \geq -\left[n\mathbb{E}\left(\frac{\partial^2 \ln f(X;\theta)}{\partial^2\theta}\right)\right]^{-1}. \tag{5}$$

$\Rightarrow$ Cette expression alternative souvent offre des avantages de point de vue calcul.

⚠ Remarque : : ces résultats concernent le cas d'un seul paramètre θ

$\Rightarrow$ Le résultat peut être facilement étendu au cas de plusieurs paramètres.

Variance minimale : Borne de Cramér-Rao

Cas de plusieurs paramètres : Soit $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_m)^T$ $(m \leq n)$ le vecteur des paramètres inconnus du modèle (la densité) $f(x; \theta_1, \cdots, \theta_m)$ pour lequel on cherche un estimateur $\hat{\boldsymbol{\Theta}} = (\hat{\Theta}_1, \cdots, \hat{\Theta}_m)^T$.

Variance minimale : Borne de Cramér-Rao

Cas de plusieurs paramètres : Soit $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_m)^T$ $(m \leq n)$ le vecteur des paramètres inconnus du modèle (la densité) $f(x; \theta_1, \cdots, \theta_m)$ pour lequel on cherche un estimateur $\hat{\boldsymbol{\Theta}} = (\hat{\Theta}_1, \cdots, \hat{\Theta}_m)^T$.

Borne de Cramér-Rao pour un vecteur paramètre

L'inégalité de Cramér-Rao, pour le cas de paramètres multiples, est de la forme

$$\text{cov}(\hat{\boldsymbol{\Theta}}) \geq \frac{\Lambda^{-1}}{n}, \tag{6}$$

ou le terme général de la matrice d'information de Fisher Λ est donné par :

$$\Lambda_{ij} = \Lambda(\theta_i, \theta_j) = \mathbb{E}\left[\left(\frac{\partial \ln f(X; \boldsymbol{\theta})}{\partial \theta_i}\right)\left(\frac{\partial \ln f(X; \boldsymbol{\theta})}{\partial \theta_j}\right)\right], \; i, j = 1, \cdots, m. \tag{7}$$

$\Rightarrow$ On a donc remplacé l'information de Fisher par la matrice Λ qui est la *matrice d'information de Fisher*.

Efficacité d'un estimateur

Définition : **Efficacité d'un estimateur**

Étant donné un estimateur sans biais $\hat{\Theta}$ de θ, le rapport de sa CRLB par sa variance est appelé l'**efficacité** de $\hat{\Theta}$

$$e(\hat{\Theta}) = \frac{\text{CRLB pour var}(\hat{\Theta})}{\text{var}(\hat{\Theta})} \tag{8}$$

$\Rightarrow$ L'efficacité d'un estimateur sans biais est ainsi inférieure ou égale à 1.

Estimateur efficace

Un estimateur sans biais ayant une efficacité égale à 1 est dit **efficace**.

Efficacité d'un estimateur

Définition : **Efficacité d'un estimateur**

Étant donné un estimateur sans biais $\hat{\Theta}$ de θ, le rapport de sa CRLB par sa variance est appelé l'**efficacité** de $\hat{\Theta}$

$$e(\hat{\Theta}) = \frac{\text{CRLB pour var}(\hat{\Theta})}{\text{var}(\hat{\Theta})} \tag{8}$$

$\Rightarrow$ L'efficacité d'un estimateur sans biais est ainsi inférieure ou égale à 1.

Estimateur efficace

Un estimateur sans biais ayant une efficacité égale à 1 est dit **efficace**.

On souhaite aussi, en augmentant la taille de l'échantillon, pouvoir diminuer l'erreur d'estimation $\Rightarrow$ on parle de convergence

Consistance (ou convergence) d'un estimateur

Définition : **Consistance (ou convergence) d'un estimateur**

Un estimateur $\hat{\Theta}$ est dit **consistant** (on dit aussi convergent) pour θ si,

$$\lim_{n \to \infty} \mathbb{P}(|\hat{\Theta} - \theta| > \epsilon) = 0 \qquad \forall \epsilon > 0. \tag{9}$$

$\Rightarrow$ Convergence en probabilité

$\Rightarrow$ La probabilité de s'éloigner de la vraie valeur du paramètre de plus de ϵ tend vers 0 quand la taille de l'échantillon n augmente.

$\Rightarrow$ L'estimateur donc converge vers la valeur à estimer quand la taille de l'échantillon tends vers l'infini (asymptotiquement).

Consistance (ou convergence) d'un estimateur

Propriété

Un estimateur sans biais et de variance asymptotiquement nulle est convergent.
Soit $\hat{\Theta}$ un estimateur pour θ sur un échantillon de taille n. Alors, si

$$\lim_{n\to\infty} \mathbb{E}[\hat{\Theta}] = \theta, \quad \text{et} \quad \lim_{n\to\infty} \text{var}[\hat{\Theta}] = 0, \tag{10}$$

l'estimateur $\hat{\Theta}$ est dit *consistant* pour θ.

Suffisance d'un estimateur

Statistique suffisante (exhaustive)

Soit X un vecteur $i.i.d.$ de taille n. Soit θ un paramètre de la loi de probabilité des X_i. Une statistique $T(X)$ est dite exhaustive pour le paramètre θ (on dit aussi suffisante) si la probabilité conditionnelle d'observer X sachant $T(X)$ est indépendante de θ :

$$\mathbb{P}(X = x | T(X) = s, \theta) = \mathbb{P}(X = x | T(X) = s), \tag{11}$$

En pratique on se sert peu de cette formule pour montrer qu'une statistique est exhaustive et on préfère utiliser le critère de factorisation suivant (appelé critère de Fisher-Neyman) :

Statistique suffisante (exhaustive) et critère de Fisher-Neyman

Soit $f_\theta(x)$ la densité de probabilité du vecteur aléatoire X. Une statistique S est exhaustive si et seulement s'il existe deux fonctions u et v telles que :

$$f_\theta(x) = u(x)\, v(\theta, T(x))$$

Suffisance d'un estimateur

Définition : **Estimateur suffisant**

Soit $(X_1, X_2, \cdots, X_n)$ un échantillon *i.i.d.* de X de distribution à paramètre θ. Si $Y = T(X_1, X_2, \cdots, X_n)$ est une statistique telle que, pour toute autre statistique $Z = u(X_1, X_2, \cdots, X_n)$, la distribution conditionnelle de Z, étant donné $Y = y$, ne dépend pas de θ, ç.à.d

$$\mathbb{P}(Z = z | Y = y, \theta) = \mathbb{P}(Z = z | Y = y)$$

alors Y est appelée une **statistique exhaustive (suffisante)** pour θ. Si l'on a également $\mathbb{E}[Y] = \theta$, alors Y est dit un **estimateur suffisant** pour θ.

$\Rightarrow$ la définition de la suffisance dit que, si Y est une statistique suffisante pour θ, toute l'information de l'échantillon concernant θ est contenue dans Y.

Méthodes d'estimation I

Il existe plusieurs méthodes d'estimation de paramètres, notamment

1. **estimation ponctuelle** comme la méthode des moments, la méthode du maximum de vraisemblance,
2. estimation bayésienne : la méthode du maximum a posteriori
3. ou la méthode d'**estimation par intervalle**

Méthode du maximum de vraisemblance (Maximum Likelihood)

Fonction de vraisemblance

Définition : **Fonction de vraisemblance**

Soit $f(x;\theta)$ la densité de probabilité d'une v.a X à densité où θ est le paramètre (vrai paramètre) à estimer (Nous prenons ici le cas simple d'un seul paramètre). Soit $x = (x_1, \ldots, x_n)$ un échantillon d'observations des variables aléatoires $(X_1, \ldots, X_n)$. La *vraisemblance* du paramètre θ pour l'échantillon x est donnée par la densité jointe de x et se note ainsi :

$$L(\theta; x) = L(\theta; x_1, \ldots, x_n) = f(x_1, \ldots, x_n; \theta). \tag{12}$$

Fonction de vraisemblance

Définition : **Fonction de vraisemblance**

Soit $f(x;\theta)$ la densité de probabilité d'une v.a X à densité où θ est le paramètre (vrai paramètre) à estimer (Nous prenons ici le cas simple d'un seul paramètre). Soit $x = (x_1, \ldots, x_n)$ un échantillon d'observations des variables aléatoires $(X_1, \ldots, X_n)$. La *vraisemblance* du paramètre θ pour l'échantillon x est donnée par la densité jointe de x et se note ainsi :

$$L(\theta; x) = L(\theta; x_1, \ldots, x_n) = f(x_1, \ldots, x_n; \theta). \tag{12}$$

Vraisemblance pour un échantillon *i.i.d.*

Pour le cas *i.i.d.*, la fonction de vraisemblance est donnée par

$$\begin{aligned} L(\theta; x) = L(\theta; x_1, \ldots, x_n) &= f(x_1;\theta) f(x_2;\theta) \cdots f(x_n;\theta) \\ &= \prod_{i=1}^{n} f(x_i;\theta). \end{aligned} \tag{13}$$

Fonction de vraisemblance

Dans le cas où les X_i sont des v.a discrètes, on a

$$\begin{aligned} L(\theta; x) &= P(x_1; \theta) P(x_2; \theta) \cdots P(x_n; \theta) \\ &= \prod_{i=1}^{n} P(x_i; \theta). \end{aligned} \tag{14}$$

⚠ Remarque : On peut aussi rencontrer la notation $L(\theta)$ de la vraisemblance de θ au lieu de $L(\theta; x_1, \ldots, x_n)$ (notamment dans ce cours :)).

$\rightarrow$ pour des valeurs d'échantillon données, la fonction de vraisemblance est seulement fonction du paramètre θ.

Maximum de vraisemblance

Définition : **Définition du maximum de vraisemblance**

Maximum de vraisemblance. L'estimation de θ par la méthode du maximum de vraisemblance consiste à choisir, comme estimation de θ, la valeur de θ qui maximise la fonction de vraisemblance $L(\theta)$.

En effet, en choisissant une valeur de θ qui maximise L (ou $\ln L$), cela revient à dire que, parmi les valeurs possible de θ, nous prenons la valeur qui rend le plus probable que possible l'évènement que les les valeurs de l'échantillon observé $(x_1, \ldots, x_n)$ viennent de la population de densité $f(x; \theta)$.

Maximum de vraisemblance

⚠ Remarque :

Bien que la plupart des vraisemblances soient différentiables, les solutions de l'équation de vraisemblance (??) ne s'expriment pas toujours par des formes analytiques.

⇒ On a souvent recours à des méthodes d'optimisations numériques pour identifier les maxima de la fonction de vraisemblance (par exemple comme en régression Logistique, mélange de densités, modèles de Markov cachés, etc)

⇒ par exemple la montée de gradient, l'algorithme de Newton Raphson, l'algorithme EM, etc.

Propriétés du maximum de vraisemblance

Soit $\hat{\theta}$ la valeur de l'estimateur du maximum de vraisemblance $\hat{\Theta}$ de θ estimée à partir de l'échantillon $(x_1, \cdots, x_n)$ de taille n

Convergence

L'estimateur obtenu par la méthode du maximum de vraisemblance est convergent : $\lim_{n\to\infty} \mathbb{P}(|\hat{\Theta}_n - \theta| > \epsilon) = 0 \qquad \forall\, \epsilon > 0.$

Absence de biais et efficacité asymptotiques

Quand n tend vers l'infini on a :

$$\lim_{n\to\infty} \mathbb{E}[\hat{\Theta}] = \theta \quad : \text{asymptotiquement sans biais} \tag{15}$$

$$\lim_{n\to\infty} \text{var}[\hat{\Theta}] = \frac{1}{n\mathbb{E}\left[\left(\frac{\partial f(X;\theta)}{\partial \theta}\right)^2\right]} = \frac{1}{\mathcal{I}_n(\theta)} = \text{CRLB} \ : \text{asymptotiquement efficac}$$

Des résultats analogues sont obtenus lorsque X est une v.a. discrète.

Propriétés du maximum de vraisemblance

Normalité asymptotique

La distribution de $\hat{\Theta}$ tend vers une distribution normale lorsque n devient grand. L'EMV est donc *asymptotiquement normal*.

$$\sqrt{n}\,(\hat{\Theta} - \theta) \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}\left(0, \mathcal{I}_n(\theta)^{-1}\right). \tag{16}$$

Invariance

On peut montrer que, si $\hat{\Theta}$ est l'EMV de θ, alors l'EMV d'une fonction bijective différentiable de θ, soit $g(\theta)$, est $g(\hat{\Theta})$.

$\Rightarrow$ Cette importante propriété d'invariance implique que, par exemple, si $\hat{\sigma}$ est l'EMV de l'écart type σ pour une distribution donnée, alors l'EMV de la variance σ^2 est $\hat{\sigma}^2$.

Méthode des Moindres Carrées (Least Squares)

Méthode des Moindres Carrées I

La méthode des MC consiste à estimer les paramètres d'un modèle en minimisant les écarts quadratiques entre les données observées, d'une part, et leurs valeurs attendues, d'autre part

Très utilisée notamment en régression où l'on cherche à expliquer la variation d'une variable de sortie (expliquée) Y, par la variation d'une variable d'entrée (explicative, covariable) X

Compte tenu de la valeur de X, la meilleure prédiction de Y (en termes d'erreur quadratique) est l'espérance $f(X)$ de Y sachant X.

On dit que Y est une fonction de X plus un bruit (erreur) :

$$Y = f(X) + E \tag{17}$$

f est appelée la fonction de régression, et E est un bruit souvent supposé d'espérance nulle.

Méthode des Moindres Carrées II

L'estimateur des MC a des propriétés optimales d'absence de biais, de variance minimale (sous certaines conditions)

Critère des moindres carrés

Soit le modèle

$$Y_i = f(X_i) + E_i \tag{18}$$

La fonction f est à estimer à partir d'un échantillon des couples de covariables X_i et leur réponses Y_i : $((x_1, y_1), ..., (x_n, y_n))$

Cette estimation est effectuée en minimisant la somme des écarts (erreurs) quadratiques

Définition : **Critères des moindres carrés**

L'erreur quadratique est donnée par la somme des carrés des résidus (Residual Sum of Squares (RSS)) :

$$\text{RSS} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2 . \tag{19}$$

Moindres Carrés

Erreur quadratique dans le cas d'une fonction paramétrique

Soit $f(x;\theta)$ une fonction de paramètre θ à estimer. La somme des écarts qudratique dans ce cas est donnée par

$$RSS(\theta) = \sum_{i=1}^{n} (y_i - f(x_i;\theta))^2 . \tag{20}$$

Moindres Carrés

Erreur quadratique dans le cas d'une fonction paramétrique

Soit $f(x;\theta)$ une fonction de paramètre θ à estimer. La somme des écarts qudratique dans ce cas est donnée par

$$RSS(\theta) = \sum_{i=1}^{n} (y_i - f(x_i;\theta))^2 . \tag{20}$$

Définition : **Définition de l'estimateur des moindres carrés**

L'estimation de θ par la méthode des moindres carrés consiste à choisir, comme estimation de θ, la valeur de θ qui minimise la fonction RSS(θ).

$$\hat{\theta} = \arg\min_{\theta} \text{RSS}(\theta) \tag{21}$$

En effet, en choisissant une valeur de θ qui minimise RSS(θ), cela revient à dire que, parmi les valeurs possible de θ, nous prenons la valeur qui correspond à une erreur minimale que les réponses y s'écartent de $f(x;\theta)$ pour l'échantillon observé $((x_1, y_1), \ldots, (x_n, y_n))$.

Moindres Carrés

⚠ Remarque :

Bien que la plupart des critères d'EQ soient différentiables, la minimisation du critère des MC ne s'effectue pas toujours de façon analytique

⇒ On a souvent recours à des méthodes d'optimisations numériques (par exemple comme en réseau de neurones, etc)

⇒ la descente de gradient, l'algorithme de Newton Raphson, etc

Dans le cas où la fonction d'erreur est convexe, l'estimateur du Moindres Carrés fournit le minimum global. Cependant, dans beaucoup de problèmes réels, la fonction d'erreur n'est pas convexe et l'on a un minimum local ; atteindre le minimum global n'est pas toujours garanti

Des procédures algorithmiques existent (plusieurs initialisations, etc) et peuvent permettre d'atteindre un "bon" minimum local

Moindres Carrés : cas de paramètres multiples I

Dans le cas d'un paramètre multiple $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$,le critère d'erreur est donné par

$$\text{RSS}(\boldsymbol{\theta}) = \text{RSS}(\theta_1, \ldots, \theta_m)$$

Les estimateurs de MC de $\theta_j, j = 1, \cdots, m$, sont obtenus en résolvant simultanément le système d'équations suivant

$$\frac{\partial \text{RSS}(\boldsymbol{\theta})}{\partial \theta_j}|_{\theta_j = \hat{\theta}_j} = 0 \quad \text{pour} \, j = 1, \ldots, m \tag{22}$$

Propriétés de l'estimateur des moindres carrés

Soit $\hat{\theta}$ la valeur de l'estimateur des Moindres Carrés $\hat{\Theta}$ de θ estimée à partir de l'échantillon $((x_1, y_1), \cdots, (x_1, y_n))$ de taille n

Absence de biais

Si l'on suppose que les erreurs (le bruit) sont d'espérance nulle ($\mathbb{E}[E_i] = 0$), l'estimateur des MC est sans bias

Propriétés de l'estimateur des moindres carrés

Soit $\hat{\theta}$ la valeur de l'estimateur des Moindres Carrés $\hat{\Theta}$ de θ estimée à partir de l'échantillon $((x_1, y_1), \cdots, (x_1, y_n))$ de taille n

Absence de biais

Si l'on suppose que les erreurs (le bruit) sont d'espérance nulle ($\mathbb{E}[E_i] = 0$), l'estimateur des MC est sans bias

variance minimale

Si les erreurs sont d'espérance nulle ($\mathbb{E}[E_i] = 0$) et homoscédastiques décorrélées ($\mathbb{E}[E_i^T E_i] = \sigma^2 \mathbf{I}$) L'EMC est alors à variance minimale

$\Rightarrow$ efficace et est donc le meilleur estimateur sans biais

Ces propriétés sont valables quelle que soit la distribution des erreurs.

Propriétés de l'estimateur des moindres carrés

Soit $\hat{\theta}$ la valeur de l'estimateur des Moindres Carrés $\hat{\Theta}$ de θ estimée à partir de l'échantillon $((x_1, y_1), \cdots, (x_1, y_n))$ de taille n

Absence de biais

Si l'on suppose que les erreurs (le bruit) sont d'espérance nulle ($\mathbb{E}[E_i] = 0$), l'estimateur des MC est sans bias

variance minimale

Si les erreurs sont d'espérance nulle ($\mathbb{E}[E_i] = 0$) et homoscédastiques décorrélées ($\mathbb{E}[E_i^T E_i] = \sigma^2 \mathbf{I}$) L'EMC est alors à variance minimale

$\Rightarrow$ efficace et est donc le meilleur estimateur sans biais

Ces propriétés sont valables quelle que soit la distribution des erreurs.

Si en plus on fait l'hypothèse de normalité sur les erreurs ($e_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$) :

Normalité

La distribution de $\hat{\Theta}$ est normale centrée sur le vrai paramètre θ

Régression linéaire

Régression linéaire I

Une situation qu'on rencontre couramment est celle dans laquelle une variable aléatoire Y est fonction d'une ou plusieurs variables indépendantes (déterministes) $(x_1, \ldots, x_m)$.

Par exemple le prix d'un logement (Y) est une fonction de sa localisation (x_1) et de son âge (x_2) ;

la durée de vie d'un composant électronique (Y) peut être liée à la température (x_1), la pression (x_2), etc ; la vitesse d'un automobiliste (Y) en fonction du temps t, etc.

Notons que les variables indépendantes est aussi appelées **variables explicatives** car à travers elles on cherche à expliquer les variables Y qui sont dites **expliquées** [1]

Régression linéaire II

L'objectif est donc d'estimer "la relation" entre Y et les variables indépendantes $(x_1, \ldots, x_m)$ étant donnés un échantillon des couples $((\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_n, Y_n))$ des variables

$(Y_1, \ldots, Y_n)$ de la variable Y et les valeurs associées des variables explicatives $(\mathbf{x}_1, \ldots, \mathbf{x}_m)$, telle que x_j, $j = 1, \ldots, m$ pour chaque valeurs observée de Y_i, $i = 1 \ldots, n$.

1. En informatique et en particulier en machine learning (apprentissage), on trouve aussi l'appellation entrées/sorties pour respectivement variables explicatives et expliquées.

Le modèle linéaire simple I

prenons le cas simple où l'on suppose que Y ne dépend que d'une seule variable explicative x et que cette relation est supposée linéaire. en d'autres termes on a la relation suivante

$$Y = \beta_0 + \beta_1 x + \epsilon, \tag{23}$$

avec $(\beta_0, \beta_1) \in \mathbb{R}^2$ sont les deux paramètres de la droite de régression, β_0 appelé *ordonnée à l'origine (intercept)* et β_1 *pente (slope)*. Ce sont les **coefficients de régression**

ϵ est une variable aléatoire représentant un résidu (erreur de mesure).

En effet, ce modèle suppose que la variable expliquée que l'on observée résulte du vrai modèle (ici le modèle linéaire représentée par la droite) et un bruit (de mesure par exemple) ou tout autre type d'erreur.

Le modèle linéaire simple I

Le bruit est généralement supposé d'espérance nulle et de variance σ^2 et décorrélé ($\text{cov}(\epsilon_i, \epsilon_j)_{i \neq j} = 0$).

Dans ce cas σ^2 devient également un paramètre du modèle et est donc aussi à estimer.

Dans le cadre de ce cours, on va supposer que ce bruit est en plus Gaussien. Il en découle donc qu'il est indépendant (les ϵ_i dont i.i.d).

Les deux paramètres (β_0, β_1) sont inconnus et donc à estimer.

Cette estimation sera effectuée à partir d'un échantillon de couples $((x_1, Y_1), \ldots, (x_n, Y_n))$ [2].

2. Ici nous utilisons la notations (x_i, Y_i) vu que x est déterministe mais cela ne change rien au modèle si X est aléatoire.

Le modèle linéaire simple I

Le modèle s'écrit donc sous la forme

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{24}$$

où ϵ_i est le bruit associée à la $i$ème variable aléatoire. Étant donnés donc une réalisation (valeur) y_i de chaque Y_i et le résidu associé (réalisation de la variable aléatoire représentant le bruit) que nous notons e_i on obtient alors :

$$y_i = \beta_0 + \beta_1 x_i + e_i, \tag{25}$$

Le modèle de régression linéaire est souvent estimé par la méthode des moindres carrés

mais il existe aussi de nombreuses autres méthodes pour estimer ce modèle (par exemple par maximum de vraisemblance ou encore par inférence bayésienne (maximum a posteriori)).

Le modèle linéaire simple : Estimation par moindres carrés

La méthode des moindres carrés est une approche d'estimation ponctuelle des paramètres de régression (β_0, β_1).

fournit les estimations $(\hat{\beta}_0, \hat{\beta}_1)$ qui minimisent la somme des écarts quadratiques entre les valeurs observées y_i et l'espérance $\beta_0 + \beta_1 x_i$ du modèle de Y_i.

D'après (25), l'écart entre la valeur d'une observation et l'espérance du modèle est

$$e_i = y_i - (\beta_0 + \beta_1 x_i).$$

La somme des carrés des résidus est donc donnée par

$$\text{RSS}(\beta_0, \beta_1) = Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2. \qquad (26)$$

RSS(β_0, β_1) est quadratique $\Rightarrow$ minimisation analytique

Régression linéaire : Formulation vectorielle

maintenant nous reformulons le modèle que nous venons de voir sous forme de vecteurs-matrices.

Comme nous allons le voir, les résultats sous la forme matricielle sont obtenus à partir de calculs simples.

Cela permettra aussi de généraliser le modèle linéaire simple à des modèles généraux notamment la régression multiple.

Soit $(y_1, \ldots, y_n)$ l'ensemble des valeurs observées de la variable dépendante Y et $(x_1, \ldots, x_n)$ l'ensemble des valeurs observées de la variable explicative x.

⚠ Remarque : L'ensemble des couples $\big((x_1, y_1), \ldots, (x_n, y_n)\big)$ s'appelle aussi *ensemble d'apprentissage*. Car c'est l'ensemble de données à partir duquel on va estimer notre modèle (donc *apprendre le modèle*) pour pouvoir ensuite prédire la valeur de Y pour une nouvelle valeur de x

Régression linéaire : Formulation vectorielle

Selon le modèle de régression linéaire simple on a

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + e_1, \\ y_2 &= \beta_0 + \beta_1 x_2 + e_2, \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_i + e_n \cdot \end{aligned} \tag{27}$$

Régression linéaire : Formulation vectorielle

Selon le modèle de régression linéaire simple on a

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + e_1, \\ y_2 &= \beta_0 + \beta_1 x_2 + e_2, \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_i + e_n \ . \end{aligned} \tag{27}$$

Notons par

- $\mathbf{y} = (y_1, \ldots, y_n)^T$ le vecteur des valeurs d'observations de Y,
- $\mathbf{e} = (e_1, \ldots, e_n)^T$ le vecteur des valeurs des résidus
- $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ le vecteur des coefficients de régressions à estimer
- $\mathbf{X}$ la *matrice de régression* (matrice de *design* ou de *Vendermonde*)

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix},$$

Régression linéaire : Formulation vectorielle

le modèle (27) s'écrit donc sous la forme matricielle suivante :

Régression linéaire : Formulation vectorielle

$$y = \mathbf{X}\beta + \mathbf{e} \cdot \tag{28}$$

La somme des écarts quadratiques $\sum_{i=1}^{n} e_i^2$ est maintenant donnée par :

Régression linéaire : Formulation vectorielle

$$\text{RSS}(\beta) = \mathbf{e}^T\mathbf{e} = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) \tag{29}$$

Régression linéaire : Formulation vectorielle

L'estimation $\hat{\boldsymbol{\beta}}$ par moindres carrés de $\boldsymbol{\beta}$ s'obtient en minimisant (29) qui est une fonction quadratique en $\boldsymbol{\beta}$.

On a :

- $\text{RSS}(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$
- $\frac{\partial \text{RSS}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$
- $\frac{\partial \text{RSS}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = \mathbf{0} \Rightarrow -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0}$

$\Rightarrow$ les équations normales

$$\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}\,. \tag{30}$$

Régression linéaire : Formulation vectorielle

$$\begin{aligned} \mathbf{X}^T\mathbf{X}\hat{\beta} &= \mathbf{X}^T\mathbf{y} \\ (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ \hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \end{aligned}$$

$\Rightarrow$ Estimateur des moindres carrés :

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\,. \tag{31}$$

Notez que l'inverse $(\mathbf{X}^T\mathbf{X})^{-1}$ existe si les données comportent au moins deux valeurs distinctes de x_i.

Bibliography I