Probabilités IV

STEP, MINES ParisTech*

8 décembre 2019 (`#ba30add`)

Table des matières

*Ce document est un des produits du projet `boisgera/CDIS`, initié par la collaboration de (S)ébastien Boisgérault (CAOR), (T)homas Romary et (E)milie Chautru (GEOSCIENCES), (P)auline Bernard (CAS), avec la contribution de Gabriel Stoltz (Ecole des Ponts ParisTech, CERMICS).

Nous nous sommes intéressé jusqu'à présent à l'étude de (suites de) variables aléatoires indépendantes. En pratique cependant, on rencontre souvent des variables dépendantes les unes des autres. Dans le cas de la météo, les variables température, vitesse du vent et pression en fournissent un exemple. Nous allons nous attacher dans ce chapitre à décrire les **lois conditionnelles** qui vont nous permettre de résumer l'information apportée par une variable (ou vecteur) sur une autre et nous intéresser en particulier à l'**espérance conditionnelle** qui nous indiquera le comportement moyen d'une variable conditionnellement à une autre. Ce dernier cas pose le cadre probabiliste d'un des problèmes fondamentaux en apprentissage statistique : l'apprentissage supervisé, où on dispose d'un ensemble de réalisations d'une variable dont on cherche à prédire le comportement à partir d'un ensemble de variables dites explicatives (ou prédicteurs).

Lois conditionnelles dans un couple

Soient deux variables aléatoire X et Y définies sur le même espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$. Dans le cas où X et Y sont indépendantes, on a vu que pour tous boréliens de $\mathbb{R}$ B_1 et B_2, on a

$$\mathbb{P}(X \in B_1, Y \in B_2) = \mathbb{P}(X \in B_1)\mathbb{P}(Y \in B_2) = \mathbb{P}_X(B_1)\mathbb{P}_Y(B_2) = \int_{B_1} \mathbb{P}_Y(B_2)\mathbb{P}_X(dx),$$

où on a utilisé la représentation intégrale $\mathbb{P}_X(B_1) = \int_{B_1} \mathbb{P}_X(dx)$ et le théorème de Fubini (les mesures de probabiltés sont finies, donc σ-finies et positives).

Du fait de l'indépendance, on a aussi $\mathbb{P}_Y(B_2) = \mathbb{P}(Y \in B_2) = \mathbb{P}(Y \in B_2 | X \in B_1) = \mathbb{P}_Y(B_2 | X \in B_1)$ ce qui exprime que pour tout borélien B_1, la loi conditionnelle de Y sachant $X \in B_1$ est identique à la loi de Y.

Dans le cas général, on va chercher une égalité de la forme

$$\mathbb{P}(X \in B_1, Y \in B_2) = \mathbb{P}_X(B_1)\mathbb{P}_Y(B_2 | X \in B_1) = \int_{B_1} \mathbb{P}_{Y|X=x}(B_2)\mathbb{P}_X(dx)$$

et s'intéresser à caractériser la *loi conditionnelle de Y sachant $X = x$*, que l'on notera donc $\mathbb{P}_{Y|X=x}$.

De même, pour toute application $g : \mathbb{R}^2 \to \mathbb{R}$ borélienne telle que $g(X, Y)$ admette une espérance (relativement à la loi du couple $\mathbb{P}_{X,Y}$), on voudrait écrire :

$$\mathbb{E}(g(X,Y)) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} g(x,y)\mathbb{P}_{Y|X=x}(dy) \right) \mathbb{P}_X(dx)$$

Pour bien fixer les idées, on va décrire spécifiquement les cas où X est discrète puis où le couple (X, Y) admet une densité avant d'aborder le cas général.

Cas où X est discrète

Dans ce paragraphe, on suppose que la variable aléatoire réelle X est discrète, c'est-à-dire que l'ensemble $X(\Omega) \subset \mathbb{R}$ des valeurs x_k prises par X est au plus dénombrable.

On peut imposer que $\forall x \in X(\Omega)$ on ait $\mathbb{P}(X = x) > 0$, quitte à modifier X sur un ensemble de probabilité nulle. On va ainsi pouvoir utiliser la définition d'une probabilité conditionnelle pour des événements de la forme $\{X = x\}$. Ceci permet d'écrire pour tous boréliens B_1 et B_2 de $\mathbb{R}$:

$$\begin{aligned}
\mathbb{P}(X \in B_1, Y \in B_2) &= \sum_{x \in X(\Omega) \cap B_1} \mathbb{P}(X = x, Y \in B_2) \\
&= \sum_{x \in X(\Omega) \cap B_1} \mathbb{P}(X = x)\mathbb{P}(Y \in B_2 | X = x) \\
&= \int_{B_1} \mathbb{P}(Y \in B_2 | X = x)\mathbb{P}_X(dx)
\end{aligned}$$

puisque $\mathbb{P}_X = \sum_{x \in X(\Omega)} \mathbb{P}(X = x)\delta_x$. On obtient ainsi l'écriture souhaitée en posant

$$\mathbb{P}_{Y|X=x}(B_2) = \mathbb{P}(Y \in B_2 | X = x), \;\; \forall x \in X(\Omega), \forall B_2 \in \mathcal{B}(\mathbb{R}).$$

Remarque

$\mathbb{P}_{Y|X=x}$ ainsi définie est simplement la probabilité sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ image par Y de la probabilité conditionnelle $\mathbb{P}(\cdot | X = x)$ définie sur $(\Omega, \mathcal{A})$, autrement dit, la **loi de Y relative à** $\mathbb{P}(\cdot | X = x)$ et non à $\mathbb{P}$.

La formule ci-dessus s'écrit $\mathbb{P}_{X,Y}(B_1 \times B_2) = \int_{B_1} \mathbb{P}(Y \in B_2 | X = x)\mathbb{P}_X(dx)$, où $\mathbb{P}_{X,Y}$ est la loi du couple. Elle se généralise à tout borélien de $B \subset \mathbb{R}^2$ de la manière suivante :

$$\begin{aligned}
\mathbb{P}_{X,Y}(B) = \mathbb{P}((X, Y) \in B) &= \sum_{x \in X(\Omega) \cap B_1} \mathbb{P}(X = x, (x, Y) \in B) \\
&= \sum_{x \in X(\Omega) \cap B_1} \mathbb{P}(X = x)\mathbb{P}((x, Y) \in B | X = x) \\
&= \sum_{x \in X(\Omega) \cap B_1} \mathbb{P}(X = x)\mathbb{P}_{Y|X=x}(B_x),
\end{aligned}$$

où $B_x = \{y \in \mathbb{R}, (x, y) \in B\}$. Ainsi, pour tout B borélien de $\mathbb{R}^2$,

$$\mathbb{E}(1_B(X, Y)) = \int_{\mathbb{R}^2} 1_B(x, y)\mathbb{P}_{X,Y}(dxdy) = \mathbb{P}_{X,Y}(B) = \int_{\mathbb{R}} \left(\int_R 1_B(x, y)\mathbb{P}_Y(dy | X = x) \right) \mathbb{P}_X(dx)$$

Par linéarité de l'espérance (héritée de celle de l'intégrale), on peut ainsi exprimer l'espérance d'une fonction étagée. Pour avoir le résultat pour une fonction positive, on exprime celle-ci comme limite simple d'une suite croissante de fonctions étagées, et on applique le théorème de convergence monotone. Enfin, on applique cette construction à g_+ et g_- pour une fonction g de signe quelconque $\mathbb{P}_{X,Y}$-intégrable. En d'autres termes, on reprend le procédé de construction de l'intégrale de Lebesgue. On obtient ainsi la formule souhaitée :

$$\mathbb{E}(g(X,Y)) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} g(x,y)\mathbb{P}_Y(dy|X=x) \right) \mathbb{P}_X(dx).$$

Exemple

Soit $X \geq 0$ une variable aléatoire à valeurs dans $\mathbb{N}$ et Y une variable aléatoire réelle positive telle que la loi du couple $\mathbb{P}_{X,Y}$ vérifie pour tout $n \in \mathbb{N}$ et tout borélien B_2 de $\mathbb{R}$:

$$\mathbb{P}_{X,Y}(\{n\} \times B_2) = (1-\alpha)\alpha^n \int_{B_2 \cap \mathbb{R}_+^\star} e^{-t}\frac{t^n}{n!}dt, \quad 0 < \alpha < 1$$

$\mathbb{P}_{X,Y}$ est bien une probabilité sur $\mathbb{R}^2$ puisque par convergence monotone :

$$\begin{aligned}
\mathbb{P}_{X,Y}(\mathbb{R}^2) &= \mathbb{P}_{X,Y}(\mathbb{N} \times \mathbb{R}) \\
&= \sum_{n \in \mathbb{N}} \mathbb{P}_{X,Y}(\{n\} \times \mathbb{R}) \\
&= \sum_{n \in \mathbb{N}} (1-\alpha)\alpha^n \int_{\mathbb{R}_+^\star} e^{-t}\frac{t^n}{n!}dt \\
&= (1-\alpha) \int_{\mathbb{R}_+^\star} e^{-t} \sum_{n \in \mathbb{N}} \frac{\alpha t^n}{n!}dt \\
&= (1-\alpha) \int_{\mathbb{R}_+^\star} e^{-(1-\alpha)t}dt = 1
\end{aligned}$$

où on aura reconnu la loi exponentielle de paramètre $(1-\alpha)$.

Comme $\forall n \in \mathbb{N}$,

$$\int_{\mathbb{R}_+^\star} e^{-t}\frac{t^n}{n!}dt = \int_{\mathbb{R}_+^\star} e^{-t}\frac{t^{(n-1)}}{(n-1)!}dt = \ldots = \int_{\mathbb{R}_+^\star} e^{-t}dt = 1$$

par intégration par parties itérée, la loi marginale de X s'écrit :

$$\forall n \in \mathbb{N}, \mathbb{P}(X=n) = \mathbb{P}_{X,Y}(\{n\} \times \mathbb{R}_+^\star) = (1-\alpha)\alpha^n,$$

loi géométrique de paramètre $(1-\alpha)$. On en déduit la loi conditionnelle de Y sachant $X = x$:

$$\mathbb{P}_{Y|X=x}(B_2) = \mathbb{P}(Y \in B_2 | X=x) = \frac{\mathbb{P}_{X,Y}(\{n\} \times B_2)}{\mathbb{P}(X=n)} = \int_{B_2 \cap \mathbb{R}_+^\star} e^{-t}\frac{t^n}{n!}dt$$

et $\mathbb{P}_{Y|X=x}$ est la donc la loi gamma de paramètre $(n+1,1)$.

Densités conditionnelles

On suppose maintenant que le couple (X, Y) admet une densité $f_{X,Y}$ (par rapport à la mesure de Borel-Lebesgue). On note $f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x,y)dy$ la (respectivement $f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x,y)dx$) la loi marginale de X (resp. de Y). On s'intéresse à caractériser la densité de la variable Y connaissant la valeur prise par la variable X, c'est la *densité conditionnelle* de Y sachant $\{X = x\}$:

Proposition

La formule suivante définit une densité sur $\mathbb{R}$, pour tout $x \in \mathbb{R}$ tel que $f_X(x) > 0$:

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}.$$

Cette fonction s'appelle la *densité conditionnelle de Y sachant* $\{X = x\}$. La probabilité conditionnelle de Y sachant $\{X = x\}$ s'écrit ainsi $\mathbb{P}_{Y|X=x} = f_{Y|X=x}(y)dy$, où dy représente la mesure de Borel-Lebesgue.

La preuve est immédiate puisque $f_{Y|X=x}$ est une fonction positive d'intégrale 1.

L'interprétation de cette définition est la suivante : la fonction $f(Y|X = x)$ est la densité de la "loi conditionnelle de Y sachant que $X = x$". Bien sûr, nous avons $\mathbb{P}(X = x) = 0$ puisque X admet une densité, donc la phrase ci-dessus n'a pas réellement de sens, mais elle se justifie heuristiquement ainsi : dx et dy étant de "petits" accroissements des variables x et y et lorsque f et f_X sont continues et strictement positives respectivement en (x, y) et x :

$$\begin{aligned} f_X(x)dx &\approx \mathbb{P}(X \in [x, x + dx]) \\ f_{X,Y}(x,y)dxdy &\approx \mathbb{P}(X \in [x, x + dx], Y \in [y, y + dy]) \end{aligned}$$

Par suite

$$\begin{aligned} f_{Y|X=x}(y)dy &\approx \frac{\mathbb{P}(X \in [x, x + dx], Y \in [y, y + dy])}{\mathbb{P}(X \in [x, x + dx])} \\ &\approx \mathbb{P}(Y \in [y, y + dy] | X \in [x, x + dx]) \end{aligned}$$

On a alors le résultat suivant qui résout le problème posé en introduction :

Proposition

Pour toute fonction $g : \mathbb{R}^2 \to \mathbb{R}$ telle que $g(X, Y)$ admette une espérance, on a :

$$\mathbb{E}(g(X,Y)) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} g(x,y) f_{Y|X=x}(y)dy \right) f_X(x)dx,$$

dont on déduit, en prenant $g = 1_{B_1 \times B_2}$, que :

$$\mathbb{P}(X \in B_1, Y \in B_2) = \int_{B_1} \left(\int_{B_2} f_{Y|X=x}(y)dy \right) f_X(x)dx.$$

Démonstration On a

$$\begin{aligned}
\mathbb{E}(g(X,Y)) &= \int_{\mathbb{R}^2} g(x,y) f_{X,Y}(x,y) dy dx \\
&= \int_{\mathbb{R}^2} g(x,y) f_{Y|X=x}(y) f_X(x) dy dx \\
&= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} g(x,y) f_{Y|X=x}(y) dy \right) f_X(x) dx,
\end{aligned}$$

les calculs étant licites par application du théorème de Fubini et du fait que l'application $x \mapsto \int_{\mathbb{R}} g(x,y) f_{Y|X=x}(y) dy$ est définie pour $X > 0$, soit presque partout relativement à la mesure $\mathbb{P}_x = f(x)dx$. ■

Cas général TODO : ajouter ref -> Neveu ?

On peut établir le résultat suivant que l'on admettra :

Théorème

Soit un couple (X,Y) de variables aléatoires réelles de loi jointe $\mathbb{P}_{X,Y}$, il existe une famille $\mathbb{P}_{Y|X=x}$ de probabilités sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, unique à une égalité $\mathbb{P}_X$-presque sûrement près[1], qui vérifie pour tous B_1, B_2, boréliens de $\mathbb{R}$:

$$\mathbb{P}_{X,Y}(B_1 \times B_2) = \int_{B_1} \left(\int_{B_2} \mathbb{P}_{Y|X=x}(dy) \right) \mathbb{P}_X(dx)$$

Ces probabilités sont appelées *lois conditionnelles* de Y sachant $X = x$. On a de plus pour toute application $g : \mathbb{R}^2 \to \mathbb{R}$ telle que $g(X,Y)$ admette une espérance :

$$\mathbb{E}(g(X,Y)) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} g(x,y) \mathbb{P}_{Y|X=x}(dy) \right) \mathbb{P}_X(dx).$$

Remarques

- Ce résultat peut être interprété comme un **théorème de Fubini conditionnel**, dans le sens où il permet une intégration séquentielle, mais où la mesure de probabilité du couple (X,Y) s'exprime comme un produit de mesures dont l'un des termes dépend de la variable d'intégration de l'autre.
- Fréquemment, dans les applications, la famille des lois conditionnelles est une donnée du modèle considéré, et leur existence ne pose donc pas de problème !
- On retrouve les cas vus précédemment en notant que pour tout borélien B_1 de $\mathbb{R}$ on a $\mathbb{P}_x(B_1) = \int_{B_1} \mathbb{P}_X(dx) = \sum_{x \in \mathcal{B}_1} \mathbb{P}(X = x)$ lorsque X est discrète, et que pour tous boréliens B_1 et B_2 de $\mathbb{R}$ on a $\mathbb{P}_X(B_1) = \int_{B_1} f_X(x) dx$ et $\mathbb{P}_{X,Y}(B_1 \times B_2) = \int_{B_1 \times B_2} f_{X,Y}(x,y) dx dy$.
- Dans tout ce qui précède, les rôles de X et Y peuvent évidemment être inversés.

1. c'est-à-dire qu'on peut définir ces probabilités de la manière qu'on souhaite pour les boréliens B tels que $\mathbb{P}_X(B) = 0$.

Conséquences

Le théorème précédent a deux conséquences majeures. Il fournit d'une part un moyen efficace d'identifier la loi marginale de Y connaissant la loi marginale de X et la loi de Y sachant X. En effet, en notant que pour tout borélien B de $\mathbb{R}$, $\mathbb{P}_Y(B) = \mathbb{P}_{X,Y}(\mathbb{R} \times B)$ en appliquant ce théorème, on a la proposition suivante :

Proposition

— La loi marginale $\mathbb{P}_Y$ de Y s'exprime comme la moyenne des lois conditionnelles $\mathbb{P}_{Y|X=x}$ pondérée par la loi de X. Pour tout B borélien de $\mathbb{R}$

$$\mathbb{P}_Y(B) = \int_{\mathbb{R}} \left(\int_{B_2} \mathbb{P}_{Y|X=x}(dy) \right) \mathbb{P}_X(dx) = \int_{\mathbb{R}} \mathbb{P}_{Y|X=x}(B) \mathbb{P}_X(dx)$$

— Dans le cas où X est discrète (à valeurs dans I dénombrable), on retrouve une expression de la formule des probabilités totales et composées :

$$\mathbb{P}_Y(B) = \mathbb{P}(Y \in B) = \sum_{x \in I} \mathbb{P}(Y \in B | X = x) \mathbb{P}(X = x)$$

— Dans le cas où le couple (X, Y) admet une densité, puisque l'on a $f_{X,Y}(x, y) = f_{Y|X=x}(y) f_X(x)$, on obtient l'expression suivante pour la loi marginale :

$$f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x, y) dx = \int_{\mathbb{R}} f_{Y|X=x}(y) f_X(x) dx$$

On a dans ce cadre la *formule de Bayes pour les densités* : pour tout x tel que $f_X(x) > 0$ et tout y tel que $f_Y(y)) > 0$:

$$f_{X|Y=y}(x) = \frac{f(x, y)}{f_Y(y)} = \frac{f_{Y|X=x}(y) f_X(x)}{f_Y(y)}.$$

Exemple

Poursuivons l'exemple vu ci-dessus. On rappelle qu'on a déjà identifié la loi marginale de X ainsi que la loi conditionnelle de Y sachant $X = n$ pour $n \in \mathbb{N}$ que l'on rappelle ci-dessous :

$\mathbb{P}(X = n) = (1-\alpha)\alpha^n$, $n \in \mathbb{N}$ et pour tout $B \in \mathcal{B}(\mathbb{R})$, $\mathbb{P}_{Y|X=x}(B) = \int_{B \cap \mathbb{R}_+^\star} e^{-t} \frac{t^n}{n!} dt$

On peut en déduire la loi marginale de Y en utilisant la proposition ci-dessus et le théorème de convergence monotone :

$$\begin{aligned} \mathbb{P}_Y(B) &= \sum_{n \in \mathbb{N}} (1 - \alpha)\alpha^n \int_{B \cap \mathbb{R}_+^\star} e^{-t} \frac{t^n}{n!} dt \\ &= (1 - \alpha) \int_{B \cap \mathbb{R}_+^\star} e^{-t} \sum_{n \in \mathbb{N}} \frac{(\alpha t)^n}{n!} dt \\ &= \int_B 1_{\mathbb{R}_+}(t)(1 - \alpha) e^{-(1-\alpha)t} dt, \end{aligned}$$

de sorte que Y suit une loi exponentielle de paramètre $(1-\alpha)$.

En inversant les rôles, on va pouvoir identifier la loi de X sachant $Y \in B$ en notant que

$$\begin{aligned}\mathbb{P}_{X,Y}(\{n\} \times B) &= \mathbb{P}_X(\{n\})\mathbb{P}_{Y|X=n}(B) \\ &= \int_B \frac{(\alpha t)^n}{n!} e^{-\alpha t} dt \\ &= \int_B \mathbb{P}_{X=n|Y=t}(\{n\})\mathbb{P}_Y(dt)\end{aligned}$$

où l'on reconnaît que $\mathbb{P}_{X=n|Y=t}(\{n\}) = \frac{(\alpha t)^n}{n!} e^{-\alpha t}$, c'est-à-dire que X sachant $Y = t$ suit une loi de Poisson de paramètre αt pour $\mathbb{P}_Y$-presque tout t.

En utilisant, le théorème précédent, on obtient une nouvelle caractérisation de l'indépendance de deux variables aléatoires faisant intervenir les lois conditionnelles.

Proposition (critère d'indépendance)

1. X et Y sont indépendantes si et seulement si, pour $\mathbb{P}_x$-presque tout x, $\mathbb{P}_{Y|X=x}$ ne dépend pas de x et dans ce cas, on a $\mathbb{P}_{Y|X=x} = \mathbb{P}_Y$, c'est-à-dire que la loi conditionnelle est identique à la loi marginale.
2. Dans le cas où (X, Y) admet une densité, X et Y sont indépendantes si et seulement si la densité conditionnelle de Y sachant $\{X = x\}$ ne dépend pas de x.

Démonstration

1. Si X et Y sont indépendantes, pour tous B_1, B_2 boréliens de $\mathbb{R}$, $\mathbb{P}_{X,Y}(B_1 \times B_2) = \mathbb{P}_X(B_1)\mathbb{P}_Y(B_2) = \int_{B_1} \mathbb{P}_Y(B_2)\mathbb{P}_X(dx) = \int_{B_2} \mathbb{P}_X(B_1)\mathbb{P}_Y(dy)$. Le résultat d'unicité du théorème ci-dessus (à une égalité $\mathbb{P}_X$-presque sûre près), nous indique alors que $\mathbb{P}_{Y|X=x}(B_2) = \mathbb{P}_Y(B_2)$.
   Inversement, si $\mathbb{P}_{Y|X=x} = \mathbb{P}_Y$, alors $\mathbb{P}_{X,Y}(B_1 \times B_2) = \int_{B_1} \mathbb{P}_{Y|X=x}(B_2)\mathbb{P}_X(dx) = \int_{B_1} \mathbb{P}_Y(B_2)\mathbb{P}_X(dx) = \mathbb{P}_X(B_1)\mathbb{P}_Y(B_2)$.
2. Si X et Y sont indépendantes, $f_{X,Y}(x, y) = f_X(x)f_Y(y)$, d'où $f_{Y|X=x}(y) = f_Y(y)$.
   Inversement, si $f_{Y|X=x}(y) = f_Y(y)$ alors $f_{X,Y}(x, y) = f_{Y|X=x}(y)f_X(x) = f_Y(y)f_X(x)$ et X et Y sont indépendantes.

■

Espérance conditionnelle

Puisque $\mathbb{P}_{Y|X=x}$ est la loi d'une variable aléatoire, on peut définir l'espérance qui lui est associée et introduire la notion d'espérance conditionnelle dans le cas où Y est intégrable.

Définition

Soit Y une variable aléatoire intégrable.

1. L'*espérance conditionnelle de Y sachant* $\{X = x\}$ est définie par

$$\mathbb{E}(Y|X = x) = \int_{\mathbb{R}} y\mathbb{P}_{Y|X=x}(dy).$$

2. L'*espérance conditionnelle de Y sachant X* est la **variable aléatoire** définie par :

$$\mathbb{E}(Y|X) = \psi(X), \text{ avec } \psi(x) = \mathbb{E}(Y|X = x).$$

Remarques

1. $\psi(x)$ n'est définie que pour $x \notin B$, avec $\mathbb{P}(X \in B) = 0$. Par conséquent, la définition définit bien l'espérance conditionnelle $\psi(X) = \mathbb{E}(Y|X)$ $\mathbb{P}_X$-presque sûrement, autrement dit avec probabilité 1.
2. $\mathbb{E}(\mathbb{E}(|Y||X)) = \int_{\mathbb{R}} \left(\int_R |y| \frac{f(x,y)}{f_X(x)} dy \right) f_X(x)dx = \mathbb{E}(|Y|)$ où nous avons utilisé le théorème de Fubini. L'espérance conditionnelle de Y sachant X est bien définie dès que Y est intégrable.
3. Lorsque (X, Y) admet une densité, l'espérance conditionnelle de Y sachant $\{X = x\}$ s'écrit

$$\mathbb{E}(Y|X = x) = \int_{\mathbb{R}} y f_{Y|X=x}(y)dy.$$

On peut étendre cette définition à toute variable de la forme $h(X, Y)$.

Définition

Soit Y une variable aléatoire et h une fonction mesurable positive ou $\mathbb{P}_{X,Y}$-intégrable sur $\mathbb{R}^2$.

1. L'*espérance conditionnelle de $h(X, Y)$ sachant* $\{X = x\}$ est définie par

$$\mathbb{E}(h(X, Y)|X = x) = \int_{\mathbb{R}} h(x, y)f_{Y|X=x}(y)dy.$$

2. L'*espérance conditionnelle de $h(X, Y)$ sachant X* est la **variable aléatoire** définie par :

$$\mathbb{E}(h(X, Y)|X) = \psi(X), \text{ avec } \psi(x) = \mathbb{E}(h(X, Y)|X = x).$$

Théorème

Si Y est intégrable, alors $\psi(X) = \mathbb{E}(Y|X)$ est intégrable, et

$$\mathbb{E}(\psi(X)) = E(Y).$$

Démonstration On a

$$\begin{aligned}
\mathbb{E}(\psi(X)) &= \int_{\mathbb{R}} \psi(x) f_X(x) dx \\
&= \int_{\mathbb{R}} \left(\int_R y f_{Y|X=x} dy \right) f_X(x) dx \\
&= \int_{\mathbb{R}} \left(\int_R |y| \frac{f(x,y)}{f_X(x)} dy \right) f_X(x) dx \\
&= \mathbb{E}(Y).
\end{aligned}$$

où avons utilisé ici le théorème de Fubini dont l'application est justifiée par la remarque ci-dessus. ∎

Ce résultat permet de calculer $\mathbb{E}(Y)$ en conditionnant par une variable auxiliaire X :

$$\mathbb{E}(Y) = \int_{\mathbb{R}} \mathbb{E}(Y|X=x) \mathbb{P}_X(dx) dx$$

Il généralise la formule des probabilités totales, qui correspond ici à $Y = 1_A$, et $B_x = \{X = x\}$ où les B_x forment cette fois une partition non dénombrable de $\mathbb{R}$. On l'écrit souvent sous forme

$$\mathbb{E}\left(\mathbb{E}(Y|X)\right) = \mathbb{E}(Y)$$

et on l'appelle la *formule de l'espérance totale.*

L'espérance conditionnelle étant définie comme l'espérance de la loi conditionnelle, elle hérite des propriétés usuelles de l'espérance :

1. si Y et Z sont intégrables, $\mathbb{E}(aY + bZ|X) = a\mathbb{E}(Y|X) + b\mathbb{E}(Z|X)$,
2. $\mathbb{E}(Y|X) \geq 0$ si $Y \geq 0$,
3. $\mathbb{E}(1|X) = 1$. De plus, si g est mesurable positive ou $\mathbb{P}_X$-intégrable,

$$\mathbb{E}(Yg(X)|X) = g(X)\mathbb{E}(Y|X)$$

est une généralisation de l'égalité 1. ci-dessus, au cas où $a = g(X)$, qui doit être considéré "comme une constante" dans le calcul de l'espérance conditionnelle sachant X (X est fixée comme une donnée connue a priori). En effet, on a alors $\mathbb{E}(g(x)Y|X = x) = g(x)\psi(x)$.

Vecteurs Gaussiens à Densité

Dans le cas des vecteurs gaussiens, le calcul des lois conditionnelles de certaines composantes par rapport aux autres est particulièrement aisé. On verra en particulier que les lois conditionelles ont le bon goût d'être elles-mêmes gaussiennes, ce qui explique (en partie) le succès de ces modèles dans les applications.

Régression et espérance conditionnelle des variables de carré intégrable

La régression est un ensemble de méthodes d'apprentissage statistique très utilisées pour analyser la relation d'une variable par rapport à une ou plusieurs autres. Ces méthodes visent notamment à décrire les liens de dépendance entre variable mais aussi de prédire au mieux la valeur d'une quantité non observée en fonction d'une ou plusieurs autres variables. On va en décrire ici le principe du point de vue probabiliste dans le cas particulier des variables de carré intégrable. On verra dans ce cadre, que l'on rencontre très fréquemment en pratique, une interprétation géométrique très éclairante de l'espérance conditionnelle.

Régression linéaire

On considère deux variables aléatoires de carré intégrable dont on suppose connues les variances et la covariance. Nous souhaitons trouver la meilleure approximation de Y par une fonction affine de X de la forme $aX + b$, au sens des moindres carrés, c'est à dire qui minimise la quantité $\mathbb{E}((Y - (aX + b))^2)$. Il s'agit de déterminer les constantes a et b telles que $\mathbb{E}((Y - (aX + b))^2)$ soit minimale. Or, par linéarité,

$$\mathbb{E}((Y - (aX + b))^2) = \mathbb{E}(Y^2) - 2a\mathbb{E}(XY) + a^2\mathbb{E}(X^2) + 2ab\mathbb{E}(X) + b^2$$

L'annulation des dérivées partielles par rapport à a et b entraîne que les solutions sont

$$\begin{aligned} a &= \frac{\mathrm{Cov}(X, Y)}{\mathbb{V}(X)} = \rho(X, Y)\frac{\sigma_Y}{\sigma_X} \\ b &= \mathbb{E}(Y) - a\mathbb{E}(X) \end{aligned}$$

On vérifie aisément que ces valeurs donnent bien un minimum pour $\mathbb{E}((Y - (aX + b))^2)$, et déterminent ainsi la meilleure approximation linéaire de Y basée sur X au sens de l'erreur quadratique moyenne.

Cette approximation linéaire vaut

$$\mathbb{E}(Y) + \rho(X, Y)\frac{\sigma_Y}{\sigma_X}(X - \mathbb{E}(X))$$

et l'erreur quadratique moyenne vaut alors

$$\begin{aligned}\mathbb{E}\left(\left(Y-\mathbb{E}(Y)-\rho(X,Y)\frac{\sigma_Y}{\sigma_X}(X-\mathbb{E}(X))\right)^2\right)&=\sigma_Y^2+\rho^2(X,Y)\sigma_Y^2-2\rho^2(X,Y)\sigma_Y^2\\&=\sigma_Y^2(1-\rho^2(X,Y)).\end{aligned}$$

On voit ainsi que cette erreur est proche de 0 lorsque $|\rho(X,Y)|\approx 1$ tandis qu'elle est proche de $\mathbb{V}(Y)=\sigma_Y^2$ lorsque $\rho(X,Y)\approx 0$.

Dans ce paragraphe, on s'est intéressé à caractériser la relation linéaire entre deux variables aléatoires de carré intégrable. On va montrer dans ce paragraphe que la meilleure approximation, au sens des l'erreur quadratique moyenne, de Y par une fonction de X est précisément donnée par $\mathbb{E}(Y|X)$.

Espace de Hilbert des variables aléatoires de carré intégrable

Exercices

Un exercice tout bête

Soient X et Y de densité jointe $f_{X,Y}(x,y)=\frac{1}{x}1_T(x,y)$ où T est le triangle $T=\{0<y<x<1\}$.

1. Calculer la densité marginale de X
2. Calculer la densité conditionnelle de Y sachant $X=x$.
3. En déduire l'espérance conditionnelle de Y sachant X.

(?)

Mélanges de loi

mélanges de gaussienne, cf silvere-schmidt

Lois conjuguées

exemples utiles en bayésien : gauss-gauss, gauss-gamma,...

Randomisation

exemple de somme aléatoire de v.a.r.

Solutions

Un exercice tout bête

La densité marginale de X est donnée par $f_X(x) = \int f_{X,Y}(x,y)dy = 1_{]0,1[}(x)$ et pour $x \in]0,1[$,

$$f_{Y|X=x}(y) = \frac{1}{x} 1_{]0,x[}(y)$$

Ainsi X est uniformément distribué sur $]0,1[$, et la loi de Y sachant $X = x$ est uniforme sur $]0,x[$ pour $(0 < x < 1)$. Pour un tel x, l'espérance conditionnelle $\mathbb{E}(Y|X = x)$ vaut ainsi $x/2$ et nous obtenons $\mathbb{E}(Y|X) = \frac{X}{2}$.