M2 CSMI : Incertitudes

Etienne Birmelé

Automne 2022

Table des matières

Chapitre 1

Introduction

On s'intéresse dans ce cours à un système représenté par l'équation

$$Y = Q(X_1, \dots, X_p)$$

Ce système peut être issu de la résolution d'un système d'équations différentielles, Y étant alors une mesure réelle dépendant de la solution. Il peut également représenter le risque d'un système aléatoire, Y étant la probabilité qu'une quantité d'intérêt dépasse un certain seuil critique. La variable Y sera désignée comme la **quantité d'intérêt**.

Les **variables d'entrée** X_i désignent les entrées du système, pas exemple les conditions initiales et les valeurs des coefficients constants dans un système d'EDO.

Le but de ce cours est de s'intéresser à l'incertitude liée aux valeurs des X_i et à la façon dont elle se propage en une incertitude sur la valeur de Y. Nous ne nous intéresserons pas ici à l'étude du système sous-jacent. Il est simplement fait l'hypothèse qu'il existe un moyen (plus ou moins coûteux) d'évaluer Y pour un certains nombre de valeurs des X_i.

Ce cours a été construit en particulier à partir des références (Duprez, 2022), (Garnier, 2017) et (**?**).

Chapitre 2

Variabilité, estimation et simulation de lois connues

Ce chapitre est essentiellement un chapitre de rappels, et ne rentre pas dans les détails pour les notions de base de probabilités et de statistiques, considérées comme connues.

2.1 Lois unidimensionnelles

2.1.1 Loi d'une variable aléatoire unidimensionnelle

Soit X une variable aléatoire réelle. Elle est définie (de façon équivalente) par

— sa loi $f_X : \mathbb{R} \to \mathbb{R}^+$ vérifiant $\int_{\mathbb{R}} f(x)dx = 1$ et telle que

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

— sa fonction de répartition $F_X : \mathbb{R} \to [0, 1]$ définie par

$$F_x(t) = P(X \leq t) = \int_{-\infty}^t f(x)dx$$

Remarque : Dans le cas de variables discrètes, les intégrales correspondent à des sommes.

2.1.2 Centre et dispersion

2.1.2.1 Espérance/variance/écart-type

L'**espérance** de X est définie par

$$\mathbb{E}(X) = \int_{\mathbb{R}} xf(x)dx$$

Elle s'interprète comme le centre de la distribution au sens où la loi des grands nombres certifie qu'elle correspond à la moyenne d'une infinité de tirages indépendants.

Théorème 1. *Loi des grands nombres Soit $(X_i)_{i\geq 1}$ des tirages i.i.d. suivant une loi d'espérance $\mathbb{E}X$. Alors, presque sûrement,*

$$\lim_{n\to+\infty} \frac{X_1 + ... + X_n}{n} = \mathbb{E}X$$

La **variance** et l'**écart-type** sont des indicateurs de la dispersion de la variable autour de son espérance définis par

$$var(X) = \mathbb{E}[(X - \mathbb{E}X)^2] \text{ et } \sigma(X) = \sqrt{var(X)}$$

— La variance est plus facile à manipuler mathématiquement
— L'écart-type est interprétable (même unité que X)

Théorème 2. *Théorème centrale limite Soit $(X_i)_{i\geq 1}$ des tirages i.i.d. suivant une loi d'espérance $\mathbb{E}X$ et d'écart-type $\sigma(X)$. Soit ϕ la fonction de répartition de la loi normale centrée réduite. Alors, pour tout z,*

$$\lim_{n\to\infty} \mathbb{P}(\frac{\frac{X_1+...+X_n}{n} - \mu}{\sigma/\sqrt{n}} \leq z) = \phi(z)$$

2.1.2.2 Centre et dispersion : quantiles

Soit $p \in]0, 1[$. Le **quantile** d'ordre p est défini comme l'inverse généralisé de la fonction de répartition F

$$q(p) = F^{\leftarrow}(p) = \inf\{x \| F(x) \geq p\}$$

— Dans le cas d'une distribution continue, c'est le réel q tel que

$$F(q) = \mathbb{P}(X \leq q) = p$$

— Une manière alternative d'aborder le centre et la dispersion d'une distribution est d'utiliser la **médiane** (quantile d'ordre $\frac{1}{2}$) et les **quartiles** (quantiles d'ordres $\frac{1}{4}$ et $\frac{3}{4}$).
— Dans le cadre de ce cours, les quantiles peuvent être la quantité d'intérêt. Trouver la valeur de Y qui est dépassée dans 1% des cas par exemple revient à déterminer le quantile d'ordre 0.99.

2.1.3 Estimation

En pratique, la difficulté est de choisir la densité qu'on utilise pour modéliser une variable aléatoire dans un problème réel : problème de l'**estimation statistique**.

Différentes questions se posent :

- Que faire en l'absence d'observations ?
- Que faire en toute généralité en présence d'observations (estimation non paramétrique) ?
- Avantages et inconvénients à se restreindre à une famille de lois (estimation paramétrique) ?
- Peut-on évaluer la pertinence du choix final ?

2.1.3.1 Choix d'une distribution en l'absence d'observations : l'entropie statistique

Considérons pour commencer le cas où aucune observation est disponible. Introduite par Shannon (1948), l'entropie d'une distribution f est définie par

$$H(X) = -\int f(x)\log(f(x))dx$$

L'entropie est une mesure (inverse) de l'information portée sur X par sa loi de probabilité. Elle est minimale (nulle) quand toute la masse est en un point et qu'il n'y a donc pas pas d'aléa.

Supposons qu'un expert donne des informations sur la loi du type $\int g_j(x)f(x)dx = c_j$, $j = 1..N$. La loi la moins informative (et donc la plus générale) les respectant est celle qui maximise l'entropie sous ces contraintes. Choisir une loi de X en l'absence d'observations, avec N contraintes suivant le principe du maximum d'entropie revient alors à résoudre

$$\underset{f\geq 0}{argmax} H(X) \quad \text{sous les contraintes}$$
$$\int f(x)dx = 1$$
$$\int g_j(x)f(x)dx = c_j, j = 1..N$$

Exemples :

- Sous la contrainte $a \leq X \leq b$: $X \sim \mathcal{U}(a,b)$
- Sous les contraintes $X \geq 0$ et $\mathbb{E}X = \mu$: $X \sim \mathcal{E}(1/\mu)$
- Sous les contraintes $\mathbb{E}X = \mu$ et $varX = \sigma^2$: $X \sim \mathcal{N}(\mu, \sigma^2)$

2.1.3.2 Estimation non-paramétrique d'une distribution

On se place maintenant dans le cas où un échantillon d'observation est disponible mais où on souhaite ne pas faire d'hypothèse sur la forme de la loi. Ceci n'est raisonnablement valable qu'en présence de beaucoup de données.

Sans lissage :

La fonction de répartition réelle F peut être approchée par

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{x_i \leq x}$$

La densité est alors approchée par un histogramme (problème du choix de la largeur)

Avec lissage :

Afin de considérer une distribution continue plutôt que discrète et ainsi atténuer les effets de bords liés à l'histogramme, on peut lisser l'estimation à l'aide d'un noyau. Un **noyau** est une fonction positive continue K, telle que $\int K(x)dx = 1$. On approche alors la fonction de répartition F par

$$\hat{f}_{n,h}(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - x_i}{h})$$

- Le noyau est en général une densité symétrique, par exemple la densité d'un loi normale $\mathcal{N}(0, 1)$
- Plus h est grand, plus la distribution est lisse
- Une difficulté est à nouveau le choix du paramètre h (bandwidth). Une large littérature existe à ce propos. Par example, dans un cadre d'apprentissage, supervisé, une des solutions fournie dans *sklearn* est de le régler par validation croisée. Dans le cadre général, des estimateurs sont proposés en fonction du contexte pour minimiser l'écart quadratique moyen de l'estimation finale.

2.1.3.3 Estimation paramétrique

On fait cette fois une hypothèse sur la forme de la loi, qu'on choisit dans une famille paramétrée (lois normales, exponentielle, binômiales…). Estimer la loi revient alors à estimer ses paramètres.

L'avantage de cette approche est de réduire la dimension du problème : elle est donc plus simple et plus précise si la vraie loi est bien dans la famille choisie. Son inconvénuent est qu'en réalité, la vraie loi n'est jamais réellement de la bonne forme. Cependant, si elle en est suffisamment proche, il peut être plus précis d'approcher la loi paramétrée la plus proche (espace de dimension 2 pour la famille des lois normales par exemple) que de chercher à approcher la loi en toute généralité (espace de dimension infinie).

On se ramène alors à des problèmes d'estimation ponctuelle des paramètres de la loi, regroupé dans un vecteur θ de dimension p. En fonction du problème, il existe de nombreuses méthodes, les deux principales étant la méthode des moments et le maximum de vraisemblance.

Méthode des moments :

On détermine la valeur théorique de p moments en fonction des paramètres (en général $\mathbb{E}X, ..., \mathbb{E}X^p$). On estime les moments par $\mathbb{E}\widehat{X}^k = \frac{x_1^k+...+x_n^k}{n}$ On obtient ainsi un système à p équations à p inconnues qu'on résout pour obtenir des estimations des p paramètres.

Exemple : Considérons par exemple que X suit une loi de type $Beta(a, b)$. Alors $\mathbb{E}X = \frac{a}{a+b}$ et $varX = \frac{ab}{(a+b)^2(a+b+1)}$ (estimer $varX$ ou $\mathbb{E}X^2$ revient au même). En notant $\bar{x}$ et $\hat{\sigma}^2$ les estimations de l'espérance et la variance via la moyenne et la variance de l'échantillon, on obtient $\hat{a} = \frac{\bar{x}(\bar{x}-\bar{x}^2-\hat{\sigma}^2)}{\hat{\sigma}^2}$ et $\widehat{b} = \frac{\bar{x}-2\bar{x}^2+\bar{x}^3-\hat{\sigma}^2+\bar{x}\hat{\sigma}^2}{\hat{\sigma}^2}$.

Méthode du maximum de vraisemblance :

La vraisemblance d'une observation x peut s'écrire $L_x(\theta) = f_\theta(x)$, où f_θ est la densité de la loi de paramètres θ. Les observations étant indépendantes, la vraisemblance et la log-vraisemblance (définie par $\ell_x(\theta) = \log L_x(\theta)$) d'un échantillon valent

$$L_{\mathbf{X}} = L_{(X_1,...,X_n)}(\theta) = \prod_{i=1}^{n} L_{X_i}(\theta)$$

$$l_{\mathbf{X}} = l_{(X_1,...,X_n)}(\theta) = \sum_{i=1}^{n} l_{X_i}(\theta)$$

L'estimateur du maximum de vraisemblance est défini par

$$\hat{\theta}_{EMV} = \underset{\theta\in\Theta}{\operatorname{argmax}} \quad l_{\mathbf{X}}(\theta)$$

Exemples : (à refaire en exercice)

- Supposons que $X \sim \mathcal{P}(\lambda)$. Alors $\lambda_{EMV} = \bar{x}$
- Supposons que $X \sim \mathcal{N}(\mu, \sigma)$. Alors $\mu_{EMV} = \bar{x}$ et $\sigma^2_{EMV} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Supposons que X suivent un mélange de gaussiennes à K classes : chaque observation tombe dans la classe i avec probabilité α_i, et alors $X \sim \mathcal{N}(\mu_i, \sigma_i)$. En notant f les distributions normales,

$$\mathcal{L}(\mathbf{X}|\Theta) = \prod_{i=1}^{n}\sum_{k=1}^{K} \alpha_k f_{\mu_k,\sigma_k}(X_i)$$

$$\log \mathcal{L}(\mathbf{X}|\Theta) = \sum_{i=1}^{n} \log\big(\sum_{k=1}^{K} \alpha_k f_{\mu_k,\sigma_k}(X_i)$$

Cette fois, la résolution théorique n'est pas possible (l'algorithme EM permet une résolution approchée).

2.1.3.4 Qualité de l'estimateur

Plusieurs estimateurs peuvent être construits pour la même quantité. Ils peuvent être comparés sur plusieurs bases :

- Le **biais** (estimerait-t-on la bonne valeur en moyenne en répétant les estimations ?)
$$b(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$$
- La **variance** $var(\hat{\theta})$ (à quel point des estimations répétées seraient-elles différentes)
- L'**erreur quadratique moyenne**
$$\mathbb{E}\left[(\hat{\theta} - \theta)^2\right] = var(\hat{\theta}) + b(\hat{\theta})^2$$
L'erreur quadratique moyenne est d'autant plus petite que l'estimation est peu biaisée et peu variable.

Exemple : On vérifie aisément que l'estimateur de la moyenne pour l'espérance est sans biais et de variance $\frac{Var(X)}{n}$.

Des critères asymptotiques sont également utilisés pour décrire les propriétés d'un estimateur

- La **consistance** traduit le fait que $\hat{\theta}$ converge vers θ quand le nombre d'observations tend vers 0
- Supposons qu'il existe une constante $C(\theta)$ ne dépendant pas de n, un réel positif α et une loi L connue, ne dépendant ni de n ni de θ, tels que la loi de la variable aléatoire $Cn^{\alpha}(\hat{\theta} - \theta)$ tend vers la loi de L. On dit alors que la l'estimateur converge à la **vitesse** $\frac{1}{n^{\alpha}}$. Un estimateur est d'autant meilleur que α est grand.

Exemple : L'estimateur de la moyenne pour l'espérance est consistant (loi des grands nombres) et de vitesse $\frac{1}{\sqrt{n}}$.

2.1.4 Intervalle de confiance

La consistance ou la vitesse d'un estimateur sont des propriétés asymptotiques. Cependant, le nombre d'observations est en pratique limité. Il est donc important de pouvoir, à taille d'échantillon fixé, estimer l'incertitude lié à une estimation. Pour cela, on construit des intervalles de confiance.

Soit $\alpha \in]0, 1[$. On appelle **intervalle de confiance du paramètre θ de niveau de confiance** $1 - \alpha$ (ou de risque α) un intervalle I_{α} tel que

$$\mathbb{P}(\theta \in I_{\theta,\alpha}) = 1 - \alpha$$

.

Remarques + L'intervalle est aléatoire dans le sens où des observations différentes donnent des intervalles différents.

- Plus le niveau de confiance $1-\alpha$ est grand, plus l'intervalle de confiance est de grande amplitude.
- Plus la taille de l'échantillon augmente, plus l'observation contient de l'information, plus l'amplitude de l'intervalle de confiance est faible.

Il y a plusieurs méthodes pour cela

2.1.4.1 Etude théorique de la loi de l'estimateur

Soit $\hat{\theta}$ une estimation sans biais d'un paramètre θ d'interêt. Il faut alors trouver une fonction de $\hat{\theta}$ et θ, appelée **pivot**, dont on sait déterminer la loi et donc les quantiles $\frac{\alpha}{2}$ et $1-\frac{\alpha}{2}$, ce qui permet de construire un intervalle de confiance.

Exemples : + Supposons que $X \sim N(\mu, \sigma^2)$ avec σ^2 connu et qu'on cherche à estimer μ. Alors $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$. En notant z les quantiles de la loi normale, un intervalle de confiance de niveau $1-\alpha$ pour μ est $\left[\overline{X} - \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}} \ ; \ \overline{X} + \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}\right]$

- Supposons que $X \sim N(\mu, \sigma^2)$ et qu'on cherche à estimer σ^2. Soit S^2 la variance de l'échantillon. Alors $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$. On peut en déduire qu'un intervalle de confiance de niveau $1-\alpha$ pour σ^2 est $[\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}]$

2.1.4.2 Approche bayésienne

Une approche très différente de l'estimation ponctuelle consiste à sortir du cadre dans lequel il existe une vraie valeur θ^* à trouver.

Dans l'approche dite **bayésienne**, on considère le paramètre θ comme étant lui-même une variable aléatoire. Dans ce cas, le rôle de l'échantillon est de *mettre à jour* la loi de cette variable aléatoire.

Dans le cadre bayésien :

- on considère une *loi à priori* pour le paramètre θ, qui est sa loi avant d'avoir considéré les données. On supposera que cette loi est de densité $p(\theta)$
- on considère un modèle paramétré de la loi de X sachant θ, définissant une densité $f_\theta(X) = p(X|\theta)$. Cette partie est identique à la partie modélisation de l'approche ponctuelle.
- la grandeur estimée est alors la loi de $\theta|\mathbf{x}$, appelée la **loi à postériori** de θ. Disposer d'une loi pour θ permet d'obtenir simplement un intervalle de confiance en considérant les quantiles d'ordre $\frac{\alpha}{2}$ et $1-\frac{\alpha}{2}$ de cette loi.

Pour réaliser cette estimation, on utilise la formule de Bayes, qui stipule que (p désignant ici les densités des lois dans le cas continu, des probabilités dans le cas discret)

$$p(\mathbf{x}, \theta) = p(\theta|\mathbf{x})p(\mathbf{x}) = p(\mathbf{x}|\theta)p(\theta)$$

donc

$$p(\mathbf{x}|\theta) = \frac{f_\theta(\mathbf{x})p(\theta)}{p(\mathbf{x})}$$

où $p(\mathbf{x})$ désigne la probabilité de l'observation sous la loi à priori, à savoir

$$p(\mathbf{x}) = \int p(\mathbf{x}|\theta)p(\theta)d\theta$$

Remarques :

— $p(\mathbf{x})$ est en général impossible à déterminer. Cependant, comme la loi à postériori est une loi en θ, on utilise

$$p(\mathbf{x}|\theta) \propto f_\theta(\mathbf{x})p(\theta)$$

puis le fait que c'est une densité pour déterminer la constante multiplicative manquante (ce qui est parfois impossible et justifie le développement de nombreuses algorithmes bayésiens).

— Il est important de noter que la partie modélisation (loi de $X|\theta$) est la même que précédemment, avec le même type d'hypothèses (X qui une loi normale, de Poisson, ...). La partie optimisation de la vraisemblance et l'hypothèse de normalité asymptotique nécessaire pour obtenir un intervalle de confiance ne sont plus nécessaires. Par contre, elles sont remplacées par le besoin de choisir une loi à-priori, qui influe sur le résultat.

Exemple : Supposons que le paramètre d'intérêt soit l'espérance d'une variable X que l'on modélise par une loi de Poisson $\mathcal{P}(\lambda)$: on cherche alors à estimer $\theta = \lambda$ dans le cadre d'une vraisemblance

$$p(\mathbf{x}|\lambda) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!}$$

Choisissons une loi à priori appartenant à la famille des lois Gamma dont la distribution, dépendant de deux paramètres α et β, est définie sur $\mathbb{R}^{+*}$ par

$$p(\lambda; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

α et β sont ici les **hyperparamètres** du problème (il faut en choisir des valeurs dans le cadre d'une application).

Etudions alors la distribution de la loi à postériori de λ. La densité cherchée étant une fonction de λ, toute constante multiplicative non dépendante de λ peut être mise dans le signe $\propto$ (*proportionnel à*)

$$\begin{aligned} p(\lambda|\mathbf{x}) &\propto p(\mathbf{x}|\lambda)\eta(\lambda) \\ &\propto (\prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!})\lambda^{\alpha-1}e^{-\beta\lambda} \\ &\propto \lambda^{\sum_{i=1}^{n} x_i}e^{-n\lambda}\lambda^{\alpha-1}e^{-\beta\lambda} \\ &\propto \lambda^{\alpha+\sum_{i=1}^{n} x_i-1}e^{-(n+\beta)\lambda} \end{aligned}$$

On constate alors que (comme fonctions de λ), la densité de la loi à postériori est proportionnelle à la loi *Gamma* de paramètres $\alpha' = \alpha + \sum_{i=1}^{n} x_i$ et $\beta' = n + \beta$ (puisque cette loi est également proportionnelle au second membre de la dernière ligne).

Or, deux lois de probabilité ne peuvent être proportionnelles que si elles sont égales. La loi à postériori est donc la loi *Gamma* de paramètres α' et β'.

Remarque : Dans cet exemple, on a choisi une distribution *conjuguée* à la loi choisie pour X, c'est-à dire que l'application de la formule de Bayes donne une distribution à postériori de la même famille que la distribution à priori. Seules les paramètres de ces distributions, appelés **hyperparamètres**, sont alors modifiés.

Un tel choix n'est pas toujours possible ou pertinent. Dans le cas général, il est courant qu'on ne sache pas étudier de façon théorique la loi à posteriori. On fait alors appel à des algorithmes de simulation suivant la loi à posteriori, le plus souvant des **Algorithmes MCMC** ou des **Gibbs samplers**.

2.1.4.3 Approche bootstrap

L'approche bootstrap est une approche basée sur la simulation, qui va permettre de s'affranchir des hypothèses de forme sur la loi de X. La procédure classique, introduite par Efron, correspond au tirage d'un grand nombre B d'échantillons suivant la loi empirique de l'échantillon $\mathcal{X} = (X_1, \dots, X_n)$. Or, tirer suivant cette loi revient à tirer uniformément un élément de l'échantillon. Le principe du bootstrap peut donc s'écrire simplement :

Pour i de 1 à B :

Tirer n fois avec remise dans $\mathcal{X}$ pour obtenir un échantillon bootstrap $X^{*i} = (X_1^{*i}, \dots, X_n^{*i})$

Obtenir une estimation bootstrap $\widehat{\theta^{*i}}$ du paramètre pour l'échantillon modifié.

Faire cela permet d'approcher l'estimation, et la loi de l'erreur, qui sont inatteignables dans le monde réel, par des approximations calculables à partir des échantillons bootstrap.

Monde réel

- loi P inconnue
- échantillon $\mathbf{X} = (X_1, ..., X_n)$
- estimateur $\hat{\theta} = T(\mathbf{X})$
- loi de $\hat{\theta} - \theta$ inconnue en l'absence d'hypothèses

Monde du bootstrap

- loi P_n connue : **première approximation**
- échantillons $\mathcal{X}_b^* = (X_{b,1}^*, ..., X_{b,n}^*)$
- estimateurs $\hat{\theta}_b^* = T(\mathcal{X}_b^*)$
- loi de $\hat{\theta}^* - \hat{\theta}$ approximable car $\hat{\theta}$ est connu et la loi de $\hat{\theta}^*$ peut être approchée par la loi empirique de $(\hat{\theta}_1^*, ..., \hat{\theta}_B^*)$: **deuxième approximation**.

L'utilisation sans doute la plus fréquente du bootstrap est la production d'intervalles de confiance. Pour cela, on note

$$\hat{\theta}_{(1)}^* \leq ... \leq \hat{\theta}_{(B)}^*$$

le vecteur des B estimations bootstrap ordonnées par ordre croissant.

Une possibilité pour cela est de considérer la loi de l'erreur $\hat{\theta} - \theta$, dont les réalisations bootstrap sont les $\hat{\theta}_b^* - \hat{\theta}$. Un intervalle de confiance de l'erreur commise est alors

$$[\hat{\theta}_{\lceil B\alpha/2 \rceil}^* - \hat{\theta}, \hat{\theta}_{\lceil B(1-\alpha/2) \rceil}^* - \hat{\theta}]$$

ce qui donne comme intervalle de confiance pour le paramètre

$$IC^*(1-\alpha) = [2\hat{\theta} - \hat{\theta}_{\lceil B(1-\alpha/2) \rceil}^*, 2\hat{\theta} - \hat{\theta}_{\lceil B\alpha/2 \rceil}^*]$$

Cette approche a l'avantage d'être toujours utilisable mais on cumule deux approximations en cours de route

- En simulant suffisamment (B grand), on peut faire tendre l'erreur de deuxième approximation vers 0
- L'erreur de première approximation par contre dépend uniquement de n, taille de l'échantillon observé. Le bootstrap n'est pas une technique miracle, un manque d'observations mènera à une mauvaise estimation.

2.1.5 Critère BIC et tests d'adéquation

Une fois une estimation réalisée, se pose la question de sa pertinence. Parmi les méthodes possibles, les tests d'adequation permettent de répondre à la question *Est-il crédible de penser que les données ont été générées suivant une loi*

donnée ? et les critères AIC/BIC permettent de comparer plusieurs estimations concurrentes, par exemple réalisées avec des familles de lois différentes.

2.1.5.1 Test de Kolmogorov-Smirnov

Il existe des tests statistiques dits tests d'adéquation qui permettent de trancher entre

H_0 : il est crédible que l'échantillon ait été tiré suivant la loi (ou famille de lois) de référence
H_1 : l'échantillon s'écarte significativement de cette loi (ou famille de lois)

Il existe trois grandes familles de tests d'adéquation :

- le test du χ^2 d'adéquation, qui est le moins puissant (il faudra donc un signal plus fort pour qu'il repère que la réalité est H_1). Il est cependant toujours utilisable. Il repose sur une découpage de l'univers en un nombre fini d'ensembles (typiquement des intervalles) et teste si les fréquences de chaque ensemble dans les données correspondent aux fréquences théoriques sous la loi de référence.
- le test de Kolmogorov-Smirnov compare la fonction de répartition empirique de l'échantillon avec la fonction théorique à l'aide de la distance L_1 (ou d'autres distances pour d'autres tests de la même famille). Le résultat est un test plus puissant mais valable uniquement pour les lois continues.
- des tests d'adéquation particuliers ont été élaborés pour des lois très usitées, notamment les test de Levène ou de Shapiro-Wilk pour la loi normale.

Il est à noter que les statistiques utilisées pour ces tests sont différentes suivant qu'on se compare à une loi, ou à une famille de lois.

Dans la plupart des cas pratiques, on choisit une famille de lois et on procède à l'estimation des paramètres. Il faut alors utiliser le test en version *famille de lois* et non pas le test pour une loi simple, le risque étant d'avoir une trop grande probabilité de choisir H_0 à tort. Cette possibilité existe par exemple dans les arguments d'OpenTurns, il faut simplement y faire attention.

2.1.5.2 Critères AIC et BIC

Supposons qu'on ait des estimations concurrentes pour une même loi, par exemple faites avec des hypothèses de loi Poisson et de loi Beta. On ne petu pas se contenter de comparer les vraisemblances, car les modèles ayant le plus de paramètres ont de ce fait un avantage, et on risque de toujours privilégier les modèles les plus compliqués et aboutir à de la sur-paramétrisation (ce qui donne du sur-apprentissage sur les prédictions).

Un manière de compenser cela est de considérer les critères AIC (Akaike Information Criterion) ou BIC (Bayesian Information Criterion) qui sont des critères

où la log-vraisemblance est pénalisée par la nombre de paramètres du modèle. L'idée est qu'un modèle avec plus de paramètres consitue une amélioration seulement si le gain en log-vraisemblance est suffisant pour compenser une pénalité fonction du nombre de paramètres ajoutés.

En notant p le nombre de paramètres du modèle, n la taille de l'échantillon et ℓ la log-vraisemblance du modèle appris, ces critères sont définis par

$$AIC = -2\log(\ell) + 2p$$
$$BIC = -2\log(\ell) + p\log n$$

Ces critères permettent de comparer plusieurs estimations candidates, en retenant celui ayant la plus petite valeur.

2.2 Lois multidimensionnelles

Dans la plupart des cas réels, les variables d'entrée sont multiples. Les traiter une à une permet d'obtenir F_{X_i} pour chacun des X_i, c'est-à-dire les **lois marginales** de la vraie loi $F_{X_1,\dots,X_d}$.

Dans le cas où les entrées sont **indépendantes**

$$f_{X_1,\dots,X_d}(x_1,\dots,x_d) = \prod_{i=1}^{d} f_{X_i}(x_i)$$

Dans le cas général cependant, les lois marginales ne permettent pas de remonter à la loi jointe. Des lois jointes différentes peuvent en effet avoir les mêmes lois marginales.

2.2.1 Corrélations

En plus de son centre et de la dispersion de chacune de ses marginales, un élément central de la description d'une variable multidimensionnelle est la manière dont ses coordonnées varient les unes en fonction des autres.

- La **covariance** $\mathrm{cov}(X_1, X_2) = \mathbb{E}((X_1 - \mathbb{E}X_1)(X_2 - \mathbb{E}X_2))$ qui peut être estimée par $\widehat{\mathrm{cov}}(X_1, X_2) = \sum_{k=1}^{n}(X_1^{(k)} - \bar{X}_1)(X_2^{(k)} - \bar{X}_2)$
- La **corrélation de Pearson** $\mathrm{cor}(X_1, X_2) = \frac{cov(X_1,X_2)}{\sigma(X_1)\sigma(X_2)}$ qui peut être estimée en utilisant les estimateurs de la covariance et des écarts-types. Par rapport à la covariance, la corrélation de Pearson a l'avantage d'être bornée entre -1 et 1 et sans unité.
- La **corrélation de Spearman** définie par la corrélation de Pearson des fonctions de répartition. Elle peut être estimée par la corrélation entre les vecteurs de rang des échantillons de X_1 et X_2. Sa différence avec la

corrélation de Pearson est d'être non-paramétrique : ne dépendre que des rangs fait perdre des informations, mais rend le résultat moins sensible aux erreurs de mesures ou valeurs extrêmes.
— Une autre valeur non-paramétrique est le **tau de Kendall**. Une paire d'observation est concordante si l'observation la plus petite pour X_1 l'est aussi pour X_2, discordante sinon. Alors $\tau = \frac{\#concordances - \#discordances}{n(n-1)/2}$

Ces indices permettent d'indiquer en quoi les variations de X_1 sont informatives des variations de X_2, **ce qui n'est pas indicatif d'un lien de causalité**. Cette notion sera néanmoins importante en termes de sensibilité de Y aux entrées X_i puisque des entrées fortement corrélées porteront le même type d'information.

Il est à noter également qu'on observe parfois des variables qui semblent plutôt décorrélées en général mais ne le sont plus lorsqu'on s'intéresse à leurs extrêmes. On parle alors de dépendance en queue de distribution et celle-ci se mesure aux deux extrémités par

$$\lambda_l(X_1, X_2) = \lim_{q \to 0} \mathbb{P}(X_2 \leq F_{X_2}^{-1}(q) | X_1 \leq F_{X_1}^{-1}(q))$$

et

$$\lambda_u(X_1, X_2) = \lim_{q \to 1} \mathbb{P}(X_2 \geq F_{X_2}^{-1}(q) | X_1 \geq F_{X_1}^{-1}(q))$$

2.2.2 Un exemple de loi multi-dimensionnelle : la loi normale multivariée

Un vecteur $X = \begin{pmatrix} X_1 \\ \cdots \\ X_d \end{pmatrix}$ suit une loi normale multivariée $\mathcal{N}(\mu, \Sigma)$ si sa distribution dans $\mathbb{R}^d$ est

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}(det\Sigma)^{1/2}} \exp(\frac{1}{2}(\mathbf{x} - \mu)'\Sigma^{-1}(\mathbf{x} - \mu))$$

La matrice Σ est la **matrice de variance-covariance** de X, car elle contient les variances des X_i sur la diagonale, les coveriances des couples sur les coefficients extra-diagonaux.

La loi marginale de X_i est la loi normale $\mathcal{N}(\mu_i, \Sigma_{ii})$.

L'estimation par maximum de vraisamblance à partir de n vecteurs observés $\mathbf{x}^{(k)}$, $1 \leq k \leq n$ donne

— $\hat{\mu} = \bar{\mathbf{x}}$. L'espérance de chaque X_i est estimé par sa moyenne.
— $\hat{\Sigma} = \frac{1}{n} \sum_{k=1}^{n} (\mathbf{x^{(k)}} - \ )'(\mathbf{x^{(k)}} - \ )$. Chaque variance ou convariance est estimée par la version biaisée de la variance ou de la covariance (en divisant par n au lieu de $n-1$).

2.2.3 Copules

2.2.3.1 Définition

Soit $(X_1, \dots, X_d)$ un vecteur aléatoire de marginales continues et $(U_1, \dots, U_d) = (F_1(X_1), \dots, F_d(X_d))$ le vecteur des fonctions de répartitions des lois marginales.

La **copule** de $(X_1, \dots, X_p)$ est la fonction définie sur $[0,1]^d$ par

$$\begin{aligned} C(u_1, \dots, u_d) &= \mathbb{P}(U_1 \leq u_1, \dots, U_d \leq u_d) \\ &= \mathbb{P}(X_1 \leq F_1^{-1}(u_1), \dots, X_d \leq F_d^{-1}(u_d)) \end{aligned}$$

Si la copule admet une densité continue c sur $[0,1]^d$, avec $c(x_1, \dots, x_d) = \frac{\partial}{\partial x_1 \dots \partial x_d} C(x_1, \dots, x_d)$, la distribution h de la loi jointe s'écrit

$$h(x_1, \dots, x_d) = c(F_1(x_1), \dots, F_d(x_d)) \quad f_1(x_1) \dots f_d(x_d)$$

Ainsi, pour toute fonction g,

$$\begin{aligned} \mathbb{E}[g(X_1, \dots, X_d)] &= \int_{\mathbb{R}^d} g(x_1, \dots, x_d) h(x_1, \dots, x_d) dx_1 \dots dx_d \\ &= \int_{\mathbb{R}^d} g(x_1, \dots, x_d) c(F_1(x_1), \dots, F_d(x_d)) f_1(x_1) \dots f_d(x_d) dx_1 \dots dx_d \end{aligned}$$

Connaître la densité de la copule et les densités marginales permet donc de décrire la loi jointe.

2.2.3.2 Exemples de copules

Il existe plusieurs familles de copules qui peuvent être utilisées pour modéliser ou simuler des lois multidimensionnelles. On peut citer par exemple les copules normales, les t-copules, les couples archimédiennes, de Franck ou de Clayton. Elles permettent de créer, à partir de lois marginales fixées, des lois de couple (ou en dimension supérieure) avec des corrélations spécifiées, avec des corrélations fortes ou faibles dans les queues de distribution. Le chapitre dédié dans (McClarren et al., 2018) présente des illustrations des différentes possibilités.

2.2.3.3 Estimation d'une copule

En pratique cependant, la démarche inverse est nécessaire : on observe n valeurs de la loi jointe et il faut estimer la copule sous-jacente. On peut alors utiliser la méthode suivante, qui est bien sûr d'autant meilleure que le nombre d'observations est grand :

— soit $\hat{F}_i$ la fonction de répartition empirique de la variable X_i à travers les n observations

— pour toute observation k, soit $(\tilde{U}_1^{(k)}, \dots, \tilde{U}_d^{(k)}) = (\hat{F}_1(X_1^{(k)}), \dots, \hat{F}_d(X_d^{(k)}))$
— On définit la copule $\hat{C}$ sur $[0,1]^d$ par

$$\hat{C}(u_1, \dots, u_d) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}(\tilde{U}_1^{(k)} \leq u_1, \dots, \tilde{U}_d^{(k)} \leq u_k)$$

Cette façon de faire est l'équivalent de l'utilisation de la distribution empirique d'un échantillon pour estimer une distribution. Des techniques similaires à ce qui a été développé dans le cadre des distributions, notamment l'estimation paramétrique ou l'estimation non-paramétrique à noyaux, est possible également pour des copules mais n'est pas développé ici.

2.2.4 ACP

Une autre manière de gérer le fait d'avoir plusieurs variables est de reparamétrer le problème en de nouvelles variables qui sont décorrélées. L'une de manière de faire cela est d'utiliser l'*Analyse en composante principales* (ACP ou PCA en anglais).

Cette méthode n'est pas développée en détails ici, mais revient, via une diagonalisation de la matrice de variance-covariance des données, à réaliser un changement de base tel que :

— les variables X_i sont remplacées par des variables Z_i qui sont des combinaisons linéaires des variables initiales
— les Z_i sont de corrélation nulles entre elles
— Z_1 est de variance maximale, puis Z_2 est de variance maximale parmi les variables décorrélées de Z_1, etc…

L'avantage de ce changement de base est que l'approximation de la loi jointe par le produit des lois marginales devient meilleure puisque les lois sont décorrélées. De plus, certains métamodèles, notamment le modèle linéaire, sont plus interprétables si les variables étudiées sont décorrélées. Enfin, quitte à accepter de perdre un peu de variance totale (= d'information), on peut ne garder que les k premières variables Z_i, réduisant ainsi la dimension du problème.

L'inconvénient principal de l'ACP est la perte en terme d'interprétation, les variables étudiées ne correspondant plus aux variables observées mais à des combinaisons linéaires de celles-ci après centrage/réduction.

Chapitre 3

Estimation dans le cadre de variables de loi inconnue

3.1 Méthode de Monte-Carlo

Considérons Y unidimensionnel et une quantité d'intérêt qui peut s'écrire

$$I = \mathbb{E}(\psi(Y))$$

pour une certaine fonction ψ. C'est le cas par exemple pour l'espérance, le moment d'ordre 2 (et donc la variance) ou la probabilité de dépassement d'un seuil.

Supposon que soit la fonction Q, soit la fonction ψ sont trop complexes pour pouvoir déterminer I théoriquement, mais qu'on dispose d'un échantillon $\left\{Y^{(i)}\right\}_{i=1,\dots,n}$ de valeurs de Y.

3.1.1 Principe

Le principe de la méthode de Monte-Carlo est de considérer l'estimateur de I défini par

$$\hat{I_n} = \frac{1}{n}\sum_{i=1}^{n} \psi(Y^{(i)})$$

L'estimateur de Monte-Carlo est un estimateur sans biais. En effet, par linéarité de l'espérance, $\mathbb{E}(\hat{I_n}) = \frac{1}{n} n \mathbb{E}(\psi(Y)) = I$.

Pa la loi des grands nombres, il est convergent et avec probabilité 1, $\hat{I_n} \to_{n\to+\infty} I$.

De plus, **si les valeurs de l'échantillon sont indépendantes**, son erreur quadratique moyenne est

$$\mathbb{E}((I-\hat{I_n})^2) = Var(\hat{I_n}) = \frac{1}{n}Var(\psi(Y))$$

On obtient ainsi une erreur relative variant en $\frac{1}{\sqrt{n}}$:

$$\frac{\mathbb{E}((I-\hat{I_n})^2)^{1/2}}{I} = \frac{1}{\sqrt{n}}\frac{\sigma(\psi(Y))}{\mathbb{E}(\psi(Y))}$$

Enfin, soit $\hat{\sigma_n}^2 = \frac{1}{n}\sum_{i=1}^n \psi(Y^{(i)})^2 - \hat{I_n}^2$ la variance de l'échantillon des $\psi(Y^{(i)})$. En notant q_α le quantile d'ordre α de la loi normale centrée réduite, des intervalles de confiance approchés au niveau $1-\alpha$ peuvent être obtenus en utilisant que

$$\lim_{n\to+\infty} \mathbb{P}(I \in [\hat{I_n} - q_{1-\alpha/2}\frac{\hat{\sigma_n}}{\sqrt{n}}, \hat{I_n} + q_{1-\alpha/2}\frac{\hat{\sigma_n}}{\sqrt{n}}]) = 1-\alpha$$

Remarques :

- pas régularité requise sur ψ, ni d'ailleurs sur Q dans le cas $Y = Q(\mathbf{X})$
- vitesse de convergence indépendante de la dimension de $\mathbf{X}$, mais lente
- il est nécessaire de savoir simuler suivant la loi de Y

Le chapitre 7 suivant étudie les façons de simuler afin de maintenir les propriétés de la méthode de Monte-Carlo tout en essayant de réduire au maximum la variance du numérateur.

3.1.2 Application à l'estimation de $\mathbb{E}(Y)$

Dans un cadre où on sait simuler sous la loi de Y sans savoir déterminer $\mathbb{E}(Y)$, par exemple si $Y = Q(X)$ avec une fonction Q trop complexe pour l'intégrer, on peut utiliser l'estimateur de Monte-Carlo pour $\psi = Id$ et obtenir l'estimateur habituel de la moyenne

$$\widehat{\mathbb{E}(Y)} = \frac{1}{n}\sum_{i=1}^n Y^{(i)}$$

L'estimateur est sans biais, consistant, permet de déterminer des intervalles de confiance asymptotiques. Son erreur relative

$$\frac{\mathbb{E}((\widehat{\mathbb{E}(Y)} - \mathbb{E}(Y))^2)^{1/2}}{\mathbb{E}(Y)} = \frac{1}{\sqrt{n}}\frac{\sigma(Y)}{\mathbb{E}(Y)}$$

peut elle-même être estimée à l'aide de la moyenne et de la variance de l'échantillon.

3.1.3 Application à une probabilité de dépassement $\mathbb{P}(Y > y_s)$

Considérons, dans le même cadre, un problème d'estimation du risque p_s de dépassement d'un seuil y_s. Alors $\psi(x) = \mathbb{1}_{x>y_s}$ et

$$\widehat{p_s} = \frac{1}{n} Card\{i : Y^{(i)} > y_s\}$$

L'estimateur est sans biais, consistant, permet de déterminer des intervalles de confiance asymptotiques. Son erreur relative

$$\frac{\mathbb{E}((\widehat{p_S} - p_s)^2)^{1/2}}{p_s} = \frac{1}{\sqrt{n}} \frac{\sqrt{p_s(1-p_s)}}{p_s} = \frac{1}{\sqrt{np_s}}$$

souligne le fait qu'il faut faire un nombre de simulations d'ordre supérieur à celui de $\frac{1}{p_s}$ pour obtenir une estimation convenable. Les techniques de simulation permettant de réduire la variance sont donc particulièrement pertinentes dans ce cas.

3.2 Méthodes de quasi Monte-Carlo

Les méthodes précédentes reposent sur un grand nombre de simulations. Or, si on se place dans le cadre de l'étude de $Y = Q(X)$, estimer $\mathbb{E}(\psi(Y)) = \mathbb{E}(\psi(Q(X)))$, chaque donnée simulée nécessite un appel à la fonction Q.

Cela peut se révéler très couteux, et choisir les valeurs de X pour lesquelles Y est estimée de façon non aléatoire peut s'avérer préférable en termes de vitesse de convergence de l'estimation. Quitte à appliquer la fonction quantile suivant chaque marginale, on se contente par la suite de générer des suites dans $[0,1]^d$.

Definition 3.1. La **discrépance** d'une famille de points $x_1, \dots, x_n$ de $[0,1]^d$ est définie par

$$D(x_1, \dots, x_n) = \sup_{B \in J} \left(\frac{Card(i : x_i \in B)}{n} - \lambda_d(B) \right)$$

où J est l'ensemble de tous les pavés de la forme $\prod_{i=1}^d [a_i, b_i]$ et λ_d est la mesure de Lebesgue sur $[0,1]^d$.

Definition 3.2. La **discrépance étoilée** d'une famille de points $x_1, \dots, x_n$ de $[0,1]^d$ est définie par

$$D^*(x_1, \dots, x_n) = \sup_{B \in J} \left(\frac{Card(i : x_i \in B)}{n} - \lambda_d(B) \right)$$

où J est l'ensemble de tous les pavés de la forme $\prod_{i=1}^d [0, b_i]$ et λ_d est la mesure de Lebesgue sur $[0,1]^d$.

Definition 3.3. Soit $f : [0,1] \to \mathbb{R}$. La variation de Hardy-Krause de f est donnée par :

$$V^{HK}(f) = \int_0^1 |f'(x)|dx$$

Soit $f : [0,1]^d \to \mathbb{R}$. La variation de Hardy-Krause de f est donnée par :

$$V^{HK}(f) = \int_0^1 |\frac{\partial^d f}{\partial x_1 ... \partial x_d}(x)|dx + \sum_{i=1}^{d} V^{HK}(f^{(i)})$$

où $f^{(i)} : [0,1]^{d-1} \to \mathbb{R}$ est la restriction de f sur l'hyperplan $x_i = 1$.

Théorème 3 (Koskma-Hlakwa)**.** *Soit f une fonction de variation de Hardy-Krause finie. Alors, pour tout ensemble de points $(\mathbf{x_1}, ..., \mathbf{x_n})$ de $\mathbb{R}^d$,*

$$|\frac{1}{n}\sum_{i=1}^{n} f(x_i) - \int_{[0,1]^d} f(x)dx\Big| \leq V^{HK}(f)D^*(x_1, ..., x_n)$$

De plus, pour tout ensemble de points et tout $\epsilon > 0$, il existe une fonction f telle que

$$|\frac{1}{n}\sum_{i=1}^{n} f(x_i) - \int_{[0,1]^d} f(x)dx\Big| \geq V^{HK}(f)(D^*(x_1, ..., x_n) - \epsilon)$$

En d'autre termes, ce théorème dit que la convergence la plus rapide des sommes de Monte-Carlo vers les espérances cibles se fait à l'aide de suite de discrépances étoilées minimales.

Les **méthodes de quasi-Monte Carlo** proposent ainsi d'accélérer la convergence des estimations de Monte-Carlo en utilisant des suites de discrépance plus faibles que des suites aléatoires, appelées **suites quasi-aléatoires**. Elles vérifient

$$D(x_1, ..., x_n) \leq c_d \frac{(\log n)^d}{n}$$

La plus simple est la suite de Van Corput, en dimension 1. Elle consiste à écrire $n = \sum_{j \leq 1} a_j 2^{j-1}$ en base 2, puis à définir $x_n = \sum_{j \leq 1} a_j 2^{-j}$.

Les suites de Halton en dimension plus grande généralisent ce principe suivant chaque dimension, en choisissant des décompositions suivant des nombres premiers différents suivant chaque dimension.

D'autres suites plus complexes comme les suites de Sobol existent également. Le désavantage principal de telles suites quand la dimension s'agrandit est que la propriété de faible discrépance n'est pas forcément vraie dans le cas de projections (si on se restreint à certaines coordonnées par exemple).

3.3 Plans d'expérience

Une procédure alternative, qui peut être moins bonne d'un point de vue de la discrépance mais est bien répartie suivant toutes les marginales est le **plan d'expérience en carré latin**. Il consiste :

1. A découper l'intervalle $[0,1]$ en n intervalles de longueur égale, suivant chaque variables. On obtient ainsi une partition de $[0,1]^d$ en n^d cubes de côté $\frac{1}{n}$.
2. Choisir n de telle façon à ce que suivant chaque dimension, il y ait exactement un cube sélectionné suivant chaque intervalle de la marginale. Cela revient à sélectionner d permutations $\sigma^{(j)}$ de $1,\dots,n$, $1 \leq j \leq n$. Le i^e cube choisi est alors celui qui correspond à l'intervalle $\sigma_i^{(j)}$ suivant la dimension j.
3. Tirer uniformément le point x_i dans le i^e cube.

Théorème 4. *Si la fonction $f : \mathbb{R}^d \to \mathbb{R}$ est monotone en tous ses arguments, si $x_1,\dots,x_n$ sont les points d'un hypercube latin et si $y_1,\dots,y_n$ sont les points d'un n-échantillon de Monte Carlo, alors*

$$Var(\frac{1}{n}\sum_{i=1}^{n} f(x_i)) \leq Var(\frac{1}{n}\sum_{i=1}^{n} f(y_i))$$

La qualité du remplissage de l'espace dépend du fait que les permutations suivant les différentes dimensions sont suffisamment différentes. Le plus simple est de déterminer plusieurs suites en hypercube latin et de garder celle qui maximise le critère de la distance minimale entre points de la séquence

$$D(\mathbf{x}) = \min_{i \neq j} \|x_i - x_j\|$$

3.4 Estimation de quantiles

Dans de nombreuses applications, le problème posé est le problème inverse de l'estimation du risque, à savoir l'estimation, pour une probabilité p donnée, du quantile y_p associé.

En supposant Y unidimensionnel et en notant F la fonction de répartition de Y,

$$y_p = \inf\{y : F(y) \leq p\}$$

3.4.1 Estimateur naturel

L'estimateur intuitif du quantile à partir d'un échantillon $(Y_j)_{1\leq j\leq n}$ est de considérer l'échantillon ordonné

$$Y_{(1)} \leq ... \leq Y_{(n)}$$

et de définir

$$\widehat{y_p} = Y_{(\lceil pn \rceil)}$$

On notera que $Y_{(\lceil pn \rceil)}$ est le quantile d'ordre p de la loi empirique de l'échantillon.

D'un point de vue de la qualité de l'estimateur, on peut démontrer qu'il est sans asymptotiquement biais, consistant et vérifie un théorème de type TCL. Plus exactement,

$$\begin{aligned}
\mathbb{E}(\widehat{y_p}) &= y_p + \mathcal{O}(\frac{1}{n}) \\
Var(\widehat{y_p}) &= \frac{p(1-p)}{n y_p^2} + \mathcal{O}(\frac{1}{n^2}) \\
\sqrt{n}(\widehat{y_p} - y_p &\to_{n\to+\infty} \mathcal{N}(0, \frac{p(1-p)}{y_p^2})
\end{aligned}$$

D'un point de vue pratique, cet estimateur a cependant deux limitations :

1. La probabilité d'avoir $\widehat{y_p} \leq y_p$ est d'environ $\frac{1}{2}$. Pour des raisons de risque, on peut souhaiter maîtriser la probabilité de cet évènement en la maintenant à un niveau bas. En effet, si $\widehat{y_p} < y_p$, la probabilité pour Y d'être supérieur au quantile estimé est plus grande que $1-p$.
2. Cette méthode se révèle impossible ou de mauvaise qualité pour les quantiles extrêmes, c'est-à-dire si $\lceil pn \rceil \approx n$.

3.4.2 Estimateur de Wilks

L'estimateur de Wilks répond à la première des deux questions en décalant le rang choisi dans l'échantillon ordonné

$$\widehat{y_p} = Y_{(\lceil pn \rceil + r)}$$

Cela permet de borner le risque $\mathbb{P}(\widehat{y_p} \leq y_p)$. En effet,

$$\begin{aligned}\mathbb{P}(Y_{(\lceil pn\rceil+r)}) \leq y_p) &= \mathbb{P}(\text{au moins } \lceil pn\rceil + r \text{ tirages sont inférieurs à} y_p) \\ &= \sum_{k=\lceil pn\rceil+r}^{n} \mathbb{P}(\text{exactement } k \text{ tirages sont inférieurs à} y_p) \\ &= \sum_{k=\lceil pn\rceil+r}^{n} \binom{n}{k} p^k (1-p)^{n-k}\end{aligned}$$

On remarquera qu'on peut également l'écrire

$$\begin{aligned}\mathbb{P}(Y_{(\lceil pn\rceil+r)}) \leq y_p) &= \mathbb{P}(\text{au plus } n - \lceil pn\rceil - r - 1 \text{ tirages sont supérieurs à} y_p) \\ &= \sum_{k=0}^{n-\lceil pn\rceil-r-1} \mathbb{P}(\text{exactement } k \text{ tirages sont supérieurs à} y_p) \\ &= \sum_{k=0}^{n-\lceil pn\rceil-r-1} \binom{n}{k} (1-p)^k p^{n-k}\end{aligned}$$

Quand r grandit, cette probabilité tend vers 0 et on peut donc choisir r comme étant le plus petit entier tel que cette probabilité est inférieure à $\beta > 0$, obtenant ainsi un estimateur tel que $\mathbb{P}(\widehat{y_p} \leq y_p) < \beta$

Remarques : - si l'on souhaite un quantile pour p proche de 1 avec un β faible, un échantillon trop petit ne va pas permettre d'utiliser cette méthode. Une borne inférieure simple peut être obtenue en constatant que pour pouvoir obtenir $\mathbb{P}(\widehat{y_p} \leq y_p) \leq \beta$, il faut avoir $\mathbb{P}(\widehat{y_p} > y_p) > 1 - \beta$ et donc en particulier $\mathbb{P}(\widehat{y_n} > y_p) > 1 - \beta$. Cela amène à $1 - p^n > 1 - \beta$, soit $n > \frac{\log \beta}{\log p}$.

- on peut montrer [Garnier] que à p et β fixé, un équivalent asymptotique de r est $r \sim \Phi^{-1}(1-\beta)\sqrt{p(1-p)n}$, où Φ désigne la fonction de répartition de la loi normale centrée réduite.
- le gain de la maîtrise du risque se traduit par contre par un estimateur biaisé : comme on se décale dans l'échantillon ordonné par rapport à l'estimateur naturel qui est non-biaisé, on a $\mathbb{E}(y_p) > y_p$.

3.4.3 Quantiles extrêmes

Considérons le problème de l'estimation de y_p lorsque $\alpha = 1 - p$ est très faible par rapport à la taille de l'échantillon. On parle de **quantile extrême**.

D'un point de vue théorique, on parle de quantiles extrêmes si $\lim_{n\to+infty} n\alpha_n = 0$ pour une suite de quantiles portant sur une suite d'échantillons de plus en

plus grands. En pratique, cette théorie s'applique notamment si α est inférieur (et le quantile est donc supérieur au maximum de l'échantillon) ou de l'ordre de $\frac{1}{n}$ (et le quantile est essentiellement le maximum de l'échantillon). Les méthodes des paragraphes précédents ne s'appliquent alors pas.

Exemple : La hauteur d'une crue centennale quand les données portent sur moins de plusieurs siècles est un problème d'estimation de quantile extrême.

Definition 3.4. Soit F et G deux fonctions répartition. F est dans le **domaine d'attraction** de G si il existe des suites (a_n) et b_n telles que

$$\lim_{n\to+\infty}[F(a_n y + b_n)]^n = G(y)$$

Ceci est équivalent, si on note Y_n un échantillon indépendant de taille n tiré suivant la loi de F, et L_G la loi correspondant à G, à dire que

$$\frac{max(Y_n) - b_n}{a_n} \xrightarrow{\mathcal{L}} L_G$$

On définit

$$H_\gamma(x) = \begin{cases} \exp[-(1+\gamma x)^{-1/\gamma}] & \text{si } \gamma \neq 0 \\ \exp[-e^{-x}] & \text{si } \gamma = 0 \end{cases}$$

Le théorème central de la théorie des valeurs extrêmes dit alors que

Théorème 5. *Soit F une fonction de répartition. L'une des quatre propositions suivantes est vraie :*

1. F fait partie du domaine d'attraction de H_γ pour un $\gamma > 0$. On dit alors que F a un domaine d'attraction de type Fréchet et que la loi est à queue lourde.

2. F fait partie du domaine d'attraction de H_γ pour $\gamma = 0$. On dit alors que F a un domaine d'attraction de type Gumbel et que la loi est à queue légère.

3. F fait partie du domaine d'attraction de H_γ pour un $\gamma < 0$. On dit alors que F a un domaine d'attraction de type Weibull et que la loi a un point terminal fini (la densité devient nulle au-delà d'un certain point).

4. F ne fait partie d'aucun domaine d'attraction.

On notera que le point 4. implique que les domaines d'attraction des fonctions de répartition de type H_γ sont les seuls possibles.

Fréchet	**Gumbel**	**Weibull**	**Aucun domaine**
Pareto	Normale	Uniforme	Poisson
Log-Gamma	Log-Normale	Beta	
Student	Exponentielle		
	Gamma		

Pour une distribution de Y appartenant à un domaine d'attraction,

$$\lim_{n\to+\infty}[F(a_n y + b_n)]^n = H_\gamma(y)$$

entraine, pour n assez grand,

$$n \log F(a_n y + b_n) \approx \log H_\gamma(y)$$

Comme on s'intéresse à la région extrême où $1 - F$ est proche de 0, en utilisant le DL de $\log(1-u)$ à l'ordre 1 pour $u = 1 - F(a_n y + b_n)$,

$$n(1 - F(a_n y + b_n)) \approx -\log H_\gamma(y) = (1 + \gamma y)^{-1/\gamma}$$

En choissant y tel que $a_n y + b_n = y_p$,

$$n(1 - F(y_p)) \approx (1 + \gamma \frac{y_p - b_n}{a_n})^{-1/\gamma}$$

En inversant cette relation,

$$y_p \approx b_n + \frac{a_n}{\gamma}([n(1-p)]^{-\gamma} - 1)$$

Estimer des quantiles extrêmes revient donc à estimer a_n, b_n et γ. Les façons de faire cela dépendent du bassin d'attraction considéré et dépassent le cadre de ce cours. On pourra se référer au poly de cours de Laurent Gardes (Gardes, 2020).

Chapitre 4

Analyse de sensibilité locale

L'analyse de sensibilité consiste à étudier l'effet des variables d'entrée X_i sur la quantité d'intérêt $Y = Q(X)$. Le but est de déterminer les variations de Y en fonction de celles des X, dans le but notamment de déterminer quelles sont les variables d'entrée ayant le plus d'influence sur l'incertitude de la sortie.

Le but de ce chapitre est de réaliser cette étude dans le cadre de petites variations autour des valeurs moyennes des entrées, ce qui est un cadre raisonnable dans beaucoup d'applications où ces valeurs correspondent à la valeur nominale des entrées.

Pour ce faire, on se place dans un cadre où on suppose connue la loi du vecteur $\mathbf{X}$ et donc en particulier son espérance $\mu \in \mathbb{R}^p$ et sa matrice de variance-covariance $\Sigma \in \mathcal{M}_{(p,p)}(\mathbb{R})$. On note $\sigma_i = \sqrt{\Sigma_i i}$ l'écart-type de X_i.

On suppose de plus qu'on se restreint à l'étude de faibles variations des entrées $\mathbf{X}$ autour de μ. Le développement de Taylor à l'ordre 2 de la fonction Q (ou de chacune de ses composantes si elle est multidimensionnelle) au point μ s'écrit :

$$\begin{aligned} Y = Q(\mu) &+ \sum_{i=1}^{p} \frac{\partial Q}{\partial X_i}(\mu)(X_i - \mu_i) \\ &+ \frac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{\partial^2 Q}{\partial X_i \partial X_j}(\mu)(X_i - \mu_i)(X_j - \mu_j) + \mathcal{O}(\|X - \mu\|_2^3) \end{aligned}$$

4.1 Développement d'ordre 1

Supposons que le voisinage considéré est suffisamment petit pour que l'approximation à l'ordre 1 soit valable, c'est-à-dire que Y est linéaire en les X_i.

4.1.1 Hiérarchisation des entrées : indices descriptifs

La formule de Taylor que la variation de X_i autour de μ_i est multiplié par le coefficient $\frac{\partial Q}{\partial x_i}$, qui semble par conséquent pertinent pour décrire l'influence sur Y d'une variation unitaire de X_i. Ce coefficient a cependant l'inconvénient de ne pas être sans unité, si bien qu'il sera affecté par un changement d'unité de mesure. On considère par conséquent plutôt le **coefficient de sensibilité**

$$\mu_i \frac{\partial Q}{\partial x_i}(\mu)$$

qui est sans échelle.

Remarque : Ce coefficient peut être calculé en d'autres points que μ.

L'inconvénient du coefficient de sensibilité est qu'il hiérarchise les entrées en fonction de la variation de Y pour une variation unitaire de l'entrée, mais qu'il ne prend pas en compte le réel niveau de variation des entrées, certaines pouvant être très peu variables en pratique alors que d'autres le sont beaucoup.

Une alternative est sans unité prenant en compte la varitable variabilité des entrées est le **l'indice de sensibilité**

$$\sigma_i \frac{\partial Q}{\partial x_i}(\mu)$$

Ces deux indices sont de bonnes manières de décrire les données dans un premier temps, afin de voir quelles sont les entrées plus ou moins importantes en cas de petite variation autour des moyennes. Elles ne disent rien cependant de l'incertitude sur Q. De plus, il est important de garder en mémoire qu'elles correspondent à une approximation d'ordre 1 et ne sont donc pas pertinente pour l'interprétation concernant de grandes variations.

D'un point de vue computationnel, il est à noter que les dérivées partielles peuvent être approchées par les taux d'accroissement , ce qui nécessite $p + 1$ évaluations de Q : en μ et pour chaque i en $\mu + \mathbf{e}_i$ où $\mathbf{e}_i$ est le vecteur valant 1 sur la coordonnée i et 0 ailleurs.

4.1.2 Espérance et variance pour Y unidimensionnel

Sous l'hypothèse

$$Y = Q(\mu) + \sum_{i=1}^{p} \frac{\partial Q}{\partial X_i}(\mu)(X_i - \mu_i)$$

on obtient

$$E(Y) = Q(\mu).$$

On peut également à la variance puisque

$$\mathbf{var}(Y) = \sum_{i,j=1}^{p} \frac{\partial Q}{\partial X_i}(\mu)\frac{\partial Q}{\partial X_j}(\mu)Cov(X_i - \mu_i, X_j - \mu_j) \quad = \sum_{i,j=1}^{p} \frac{\partial Q}{\partial X_i}(\mu)\frac{\partial Q}{\partial X_j}(\mu)\Sigma_{i,j}$$

Cette égalité peut également s'écrire de façon matricielle :

$$\mathbf{var}(Y) = \nabla Q(\mu)'\Sigma\nabla Q(\mu)$$

On remarque que **dans le cas indépendant**, cette formule devient

$$var(Y) = \sum_{i=1}^{p} \left(\frac{\partial Q}{\partial X_i}(\mu)\right)^2 \sigma_i^2.$$

En d'autres termes, la variance totale s'écrit comme la somme des carrés des indices de sensibilité des différentes entrées.

On peut alors isoler la contribution de la variance de chaque variable X_i sur la variance de la sortie Y par le *facteur d'importance*

$$\eta_i = \frac{\left(\frac{\partial Q}{\partial X_i}(\mu)\right)^2 \sigma_i^2}{var(Y)}$$

On a alors $\sum_i \eta_i = 1$ et l'importance relative d'une entrée en termes de variabilité est d'autant plus grande que η_i, ou le carré de l'indice de sensibilité, est grand.

4.1.3 Espérance et variances/covariances pour Y multidimensionnel

Pour toute composante j de Y,

$$Y_j = Q_j(\mu) + \sum_{i=1}^{p} \frac{\partial Q_j}{\partial X_i}(\mu)(X_i - \mu_i) + O(\|\mathbf{X} - \mu\|_2^2)$$

ce qui, en notant $J_Q(\mu)$ la jacobienne de Q évaluée en μ donne

$$\mathbf{Y} = Q(\mu) + J_Q(\mu)'(\mathbf{X} - \mu) + \mathcal{O}(\|\mathbf{X} - \mu\|_2^2)$$

Dans le cas de variations de $\mathcal{X}$ dans un domaine suffisamment petit autour de μ pour pouvoir se contenter de l'approximation à l'ordre 1, on obtient alors que

$$\mathbb{E}(\mathbf{Y}) = \mu$$

et que

$$\begin{aligned} cov(Y_j, Y_k) &= \mathbb{E}((Y_j - Q_j(\mu))(Y_k - Q_k(\mu))) \\ &= \sum_{s=1}^{p}\sum_{t=1}^{p} \frac{\partial Q_j}{\partial X_s}(\mu)\frac{\partial Q_j}{\partial X_t}(\mu)\mathbb{E}((X_s - \mu_s)(X_t - \mu_t)) \\ &= \sum_{s=1}^{p}\sum_{t=1}^{p} \frac{\partial Q_j}{\partial X_s}(\mu)\Sigma_{st}\frac{\partial Q_j}{\partial X_t}(\mu) \end{aligned}$$

d'où, en notant $Cov(\mathbf{Y})$ la matrice de variance-covariance des coordonnées de $\mathbf{Y}$,

$$Cov(\mathbf{Y}) = J_Q(\mu)'\Sigma J_Q(\mu)$$

4.1.4 Développement d'ordre 2

Supposons que l'espace de variation de $\mathbf{X}$ est un peu plus grand mais suffisamment faible pour que l'approximation à l'ordre deux soit raisonnable.

Alors

$$\begin{aligned} Y = Q(\mu) &+ \sum_{i=1}^{p} \frac{\partial Q}{\partial X_i}(\mu)(X_i - \mu_i) \\ &+ \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p} \frac{\partial^2 Q}{\partial X_i \partial X_j}(\mu)(X_i - \mu_i)(X_j - \mu_j) \end{aligned}$$

Or $\mathbb{E}(X_i) = \mu_i$. De plus, en notant σ_i l'écart-type de X_i et $\rho_{i,j}$ la corrélation entre X_i et X_j, $\mathbf{E}((X_i - \mu_i)(X_j - \mu_j)) = cov(X_i, X_j) = \sigma_i\sigma_j\Sigma_{i,j}$. D'où, par linéarité de l'espérance,

$$\mathbf{E}(Y) = Q(\mu) + \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p} \frac{\partial^2 Q}{\partial X_i \partial X_j}(\mu)\Sigma_{i,j} \quad = Q(\mu) + \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p} \frac{\partial^2 Q}{\partial X_i \partial X_j}(\mu)\sigma_i\sigma_j\rho_{i,j}$$

On constate que, dès qu'on sort du cadre linéaire de l'approximation d'ordre 1, la valeur centrale de Y dépend de la variabilité de $\mathbf{X}$. En d'autres termes, **le centre de la sortie dépend du centre mais également de la variabilité des entrées**.

Dans le cas où les entrées sont considérées comme indépendantes, cette formule se réduit à

$$\mathbf{E}(Y) = Q(\mu) + \frac{1}{2}\sum_{i=1}^{p} \frac{\partial^2 Q}{\partial X_i^2}(\mu)\sigma_i^2.$$

D'un point de vue computationnel, les dérivées partielles peuvent à nouveau être évaluées par des taux d'accroissement, ce qui nécessite $2p^2 + 1$ évaluations de Q :

- $p + 1$ pour les dérivées premières
- p pour $\frac{\partial^2 Q}{\partial X_i^2}(\mu) \approx \frac{Q(\mu+h_i\mathbf{e}_i)-2Q(\mu)+Q(\mu-h_i\mathbf{e}_i)}{h_i^2}$, les évaluations en $\mu + h_i\mathbf{e}_i$ ayant déjà été faites pour le calcul des dérivées premières.
- $2p(p-1)$ pour $\frac{\partial^2 Q}{\partial X_i \partial X_j}(\mu) \approx \frac{Q(\mu+h_i\mathbf{e}_i+h_j\mathbf{e}_j)-Q(\mu+h_i\mathbf{e}_i-h_j\mathbf{e}_j)-Q(\mu-h_i\mathbf{e}_i+h_j\mathbf{e}_j)+Q(\mu-h_i\mathbf{e}_i-h_j\mathbf{e}_j)}{4h_ih_j}$

4.2 Approche par modèle linéaire et sélection de variables

Les méthodes du premier et deuxième ordre ci-dessus nécessitent soit de pouvoir expliciter et dériver la fonction Q, soit de pouvoir l'évaluer en des points précis permettant l'estimation des dérivées partielles. Le but de ce paragraphe est d'introduire une alternative pour l'approximation au premier ordre utilisant la régression linéaire.

En reprenant la formule de Taylor au premier ordre, on peut écrire en première approximation que

$$Y = Q(\mu) + \sum_{i=1}^{p} \frac{\partial Q}{\partial X_i}(\mu)(X_i - \mu_i)$$

ce qui revient à un système

$$\tilde{Y} = \tilde{X}\beta$$

où

- $\tilde{Y}$ est un vecteur colonne de n observations avec $\tilde{Y}_j = Y_j - Q(\mu)$
- $\tilde{X} \in \mathcal{M}_{(d,n)}(\mathbb{R})$ est une matrice où $\tilde{X}_{i,j} = \frac{X_i^{(j)}-\mu_i}{\mu_i}$, $X_i^{(j)}$ étant la j^{eme} observation de la variable X_i
- β est un vecteur colonne de dimension d avec $\beta_i = \mu_i \frac{\partial Q}{\partial X_i}(\mu)$.

Estimer le vecteur β peut alors se faire par régression en résolvant le problème des moindres carrés suivant

$$\hat{\beta} = \underset{\beta\in\mathbb{R}^{p+1}}{argmin} \|\tilde{Y} - \tilde{X}\beta\|_2^2$$

Remarques :

- Cette approche s'applique également au voisinage de tout autre point que μ.
- La division par μ_i dans la définition de $\tilde{X}$ et la multiplication par μ_i dans celle de β ne sont pas obligatoires, mais permettent d'obtenir un vecteur β sans unité, ce qui est préférable en termes d'interprétation et de comparaison des différentes coefficients de β.

4.2.1 Cas $\tilde{X}'\tilde{X}$ inversible

Dans le cas où $\tilde{X}'\tilde{X}$ est inversible ($\tilde{X}'$ dénotant la transposée de $\tilde{X}$), la solution à ce problème est unique. En effet, $\tilde{Y} = \tilde{X}\beta$ entraîne $\tilde{X}'\tilde{Y} = \tilde{X}'\tilde{X}\beta$ puis $\beta = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y}$.

On obtient ainsi simplement l'ensemble des valeurs des dérivées partielles. Cependant, pour que la condition soit vérifiée, il faut que $\tilde{X}'\tilde{X}$ soit de rang plein, ce qui entraîne que $n \geq d$. En effet, le rang de X et donc de $\tilde{X}'\tilde{X}$ est borné par n et $\tilde{X}'\tilde{X}$ est carrée de dimension d.

Cette formule ne peut donc être utilisée telle quelle si on souhaite (ou ne peut) utiliser moins d'appels à la fonction Q que de variables d'entrée.

4.2.2 Régression pénalisée

Dans le cas où $\tilde{X}'\tilde{X}$ n'est pas inversible, il est possible de modifier le problème afin de le rendre résoluble. Pour cela, on minimise le problème des moindres carrés non pas dans l'espace entier des β, mais dans un espace borné.

4.2.2.1 Pénalisation L2 : pénalisation Ridge

La première possibilité pour effectuer une régression pénalisée est d'utiliser une pénalité de type Ridge. L'idée est de forcer le vecteur de coefficients β à être borné en norme L_2.

Considérons, pour $c > 0$ fixé, le problème

$$\text{Trouver} \quad \beta^{Ridge} = \underset{\beta}{argmin} \|\tilde{Y} - \tilde{X}\beta\|_2^2 \text{ sous la contrainte } \|\beta\|_2^2 \leq c$$

La méthode des multiplicateurs de Lagrange dit qu'il existe $\lambda \geq 0$ tel que la solution du problème satisfait les conditions de Karush-Kuhn-Tucker à savoir que

1. Le gradient de $\|\tilde{Y} - \tilde{X}\beta\|_2^2 + \lambda(\|\beta\|^2 - c)$ est nul
2. $\lambda(\|\beta\|^2 - c) = 0$

Une manière d'approcher le problème ci-dessous est donc de considérer, pour un coefficient $\lambda > 0$ donnée, le problème suivant :

$$\text{Trouver} \quad \beta^{Ridge}(\lambda) = \underset{\beta}{argmin} \|\tilde{Y} - \tilde{X}\beta\|_2^2 + \lambda\|\beta\|_2^2$$

On notera que pour λ tendant vers 0, on retrouve la régression classique, alors que pour λ infini, la solution est $\beta = 0$. En faisant varier λ, on obtient ainsi une famille de modèles de plus en plus pénalisés. "

Proposition 4.1. *La solution du problème est donnée par*

$$\beta^{Ridge} = (\tilde{X}'\tilde{X} + \lambda \mathbf{I}_p)^{-1}\tilde{X}'\tilde{Y}$$

Démonstration. Soit $f(\beta) = \|\tilde{Y} - \tilde{X}\beta\|_2^2 + \lambda\|\beta\|_2^2$.

$$\frac{\partial f}{\partial \beta} = -2\tilde{X}'(\tilde{Y} - \tilde{X}\beta) + 2\lambda\beta$$

En annulant cette dérivée, $(\tilde{X}'\tilde{X} + \lambda \mathbf{I}_d)^{-1}\tilde{X}\tilde{Y}$ est un point stationnaire. De plus,

$$\frac{\partial^2 f}{\partial \beta^2} = 2(\tilde{X}'\tilde{X} + \lambda \mathbf{I}_d)$$

qui est définie positive, si bien qu'il s'agit d'un minimum local.

De plus, la fonction $f(\beta)$ est strictement convexe comme somme d'une fonction convexe et d'une fonction strictement convexe. Le minimum local est donc unique et global. □

Remarques :

1. La matrice $\tilde{X}'\tilde{X} + \lambda \mathbf{I}_d$ est inversible pour tout $\lambda > 0$, le problème a donc bien une solution unique, même si $n < d$.
2. On constate que la pénalité rend le problème strictement convexe, et donc résoluble par une descente de gradient quelque soit la dimension.

Une question importante laissé est celle du choix de la valeur de λ à appliquer. Un choix automatique de λ peut être fait par validation croisée en minimisant le critère des moindres carrés pour la prédiction.

4.2.2.2 Sélection de variables : pénalisation Lasso

Quel que soit le nombre de variables ayant réellement une influence non nulle sur $\tilde{Y}$, tous les coefficients de β en utilisant une pénalité Ridge sont non-nuls. Or dans certains cas, la sélection des variables ayant localement le plus d'influence peut être le but de l'étude, auquel cas on cherche à forcer à 0 les coefficients des autres variables.

Un autre type de pénalisation consiste à favoriser les solutions ayant un grand nombre de coordonnées nulles, en considérant la norme 1 plutôt que la norme 2 du vecteur β. Cette norme aura en effet pour conséquence de mettre de nombreux coefficients exactement à 0. On parle alors de régression parcimonieuse.

$$\text{Trouver} \quad \beta^{Lasso} = \underset{\beta}{argmin} \|\tilde{Y} - \tilde{X}\beta\|_2^2 \text{sous la contrainte } \|\beta\|_1 \leq c$$

D'un point de vue géométrique, la forme des boules de la norme 1 va faire que de les solutions auront de nombreux coefficients nuls en grande dimension.

A nouveau, le lagrangien de la fonction à optimiser permet de définir un nouveau problème d'optimisation, à savoir, pour un coefficient $\lambda > 0$:

$$\text{Trouver} \quad \beta^{Lasso} = \underset{\beta}{argmin} \|\tilde{Y} - \tilde{X}\beta\|_2^2 + \lambda\|\beta\|_1$$

Il n'est plus possible de déterminer la solution à l'aide d'une formule close. Cependant, la fonction est convexe et admet donc un unique minimum qu'il est possible de trouver de façon algorithmique (algorithme LARS par exemple).

Le choix de λ se fait toujours par validation croisée, ou par **Stability selection**. Cette procédure consiste à effectuer un grand nombre d'apprentissage sur des sous-échantillonnages (80% des données par exemple) et à garder les variables sélectionnées le plus souvant par Lasso. On peut ensuite réaliser un apprentissage en petite dimension sur les variables sélectionnées.

Remarque : Il est possible de mélanger pénalités Ridge et Lasso en considérant une pénalité **Elastic-Net** qui consiste à résoudre, pour un paramètre $\alpha \in [0, 1]$ à choisir

$$\text{Trouver} \quad \beta^{Lasso} = \underset{\beta}{argmin} \|\tilde{Y} - \tilde{X}\beta\|_2^2 + \lambda(\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2)$$

Chapitre 5

Analyse de sensibilité globale - Indices de Sobol

On se place toujours dans le cas d'une fonction

$$Y = Q(X_1, ..., X_p)$$

et on s'intéresse à la part de la variabilité de Y qui est liée à la variabilité d'une ou plusieurs de ses entrées.

5.1 Influence d'une variable

Commençons par nous intéresser au cas d'une unique variable X_i. Quelle est la part de variabilité de Y liée à l'incertitude liée à X_i ?

Une façon de répondre à la question serait de fixer la valeur de X_i à une valeur x_i et de comparer la variance avec X_i libre, $Var(Y)$, à la variance avec X_i fixé, $Var(Y|X_i = x_i)$. Cette quantité serait cependant une fonction de x_i. Une manière de faire est alors d'intégrer l'influence de x_i et d'évaluer

$$\mathbb{E}_{X_i}(Var(Y|X_i))$$

Plus cette quantité est grande, moins l'incertitude sur X_i est importante dans l'incertitude de Y (Y varie beaucoup même si on connaît X_i). Cette quantité est cependant non normalisée : si l'échelle de mesure pour Y change (par exemple de mm à m pour une distance), sa valeur change également, ce qui rend son interprétation difficile.

On utilise la formule de la variance totale, à la fois pour normaliser et pour obtenir un indice d'autant plus grand que l'influence de X_i est grande. Cette formule est la suivante :

Proposition 5.1.

$$Var(Y) = \mathbb{E}(Var(Y|X_i)) + Var(\mathbb{E}(Y|X_i))$$

Démonstration.

$$\begin{aligned} Var(Y) &= \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 \\ &= \mathbb{E}(\mathbb{E}_{X_i}(Y^2|X_i)) - \mathbb{E}(\mathbb{E}_{X_i}(Y|X_i)^2) + \mathbb{E}(\mathbb{E}_{X_i}(Y|X_i)^2) - \mathbb{E}(\mathbb{E}_{X_i}(Y|X_i))^2 \\ &= \mathbb{E}(Var(Y|X_i)) + Var(\mathbb{E}(Y|X_i)) \end{aligned}$$

□

On obtient ainsi un indice normalisé, insensible au changement d'échelle et toujours compris entre 0 et 1, en considérant

$$S_i = \frac{Var(\mathbb{E}(Y|X_i))}{Var(Y)} = 1 - \frac{\mathbb{E}(Var(Y|X_i))}{Var(Y)}$$

S_i est appelé **l'indice de Sobol d'ordre 1** lié à la variable i.

Remarques :

- dans le cadre d'un modèle linéaire $Y = \beta_0 + \sum_i \beta_i X_i + \epsilon$ à variables d'entrées indépendantes, cette quantité correspond à $\frac{\beta_i^2 Var(X_i)}{Var(Y)}$. On retrouve η_i décrit dans le cadre de la méthode du cumul quadratique sous hypothèse d'un modèle linéaire (cf Chapitre Propagation).
- un indice proche de 0 signifie que les variations de X_i prennent une faible part dans les variations de Y, un indice proche de 1 qu'elle en prennent une part importante.

5.2 Cas général

5.2.1 Décomposition de Hoeffding

Theorem 5.1 (Décomposition de Hoeffding)**.** *Supposons que les X_i sont indépendants et Q est de carré intégrable sur son domaine de définition. Alors elle admet une unique décomposition*

$$Q(\mathbf{X}) = Q_0 + \sum_{1 \le i \le p} Q_i(X_i) + \sum_{1 \le i < j \le p} Q_{i,j}(X_i, X_j) + ... + Q_{1,...,p}(X_1, ..., X_p)$$

sous les contraintes

- Q_0 *est une constante*

- $\mathbb{E}(Q_I(X_I)|X_J) = 0$ *pour tout* $J \subsetneq I$

La deuxième contrainte implique que, pour toute fonction h de carré intégrable et tout J tel que $J \cap I \subsetneq I$,

$$\begin{aligned}\mathbb{E}(Q_I(X_I)h(X_J)) &= \mathbb{E}[\mathbb{E}(Q_I(X_I)h(X_J)|X_J)] \\ &= \mathbb{E}[h(X_J)\mathbb{E}(Q_I(X_I)|X_{J\cap I})] \\ &= 0\end{aligned}$$

Elle peut donc être vue comme une condition d'orthogonalité de $Q_I(X_I)$ par rapport à $L^2(X_J)$.

En particulier, pour $h = Q_J$, on obtient que

$$cov(Q_I(X_I), Q_J(X_J)) = 0, \quad \forall I \neq J$$

ce qui implique la décomposition de la variance suivante.

Proposition 5.2. $Var(Y) = \sum_{I\subset\{1,\dots,p\}} Var(Q_I(X_I))$

Il est de plus possible de déterminer explicitement l'écriture des différents termes par récurrence. La démonstration de l'existence de la décomposition et de cette écriture peuvent être retrouvées dans (Garnier, 2017)

Proposition 5.3. *Les termes de la décomposition de Hoeffding peuvent s'écrire*

$$\begin{aligned}Q_0 &= \mathbb{E}(Q(\mathbf{X})) \\ Q_i(X_i) &= \mathbb{E}(Q(\mathbf{X})|X_i) - Q_0 \\ Q_I(X_I) &= \mathbb{E}(Q(\mathbf{X})|X_I) - \sum_{J\subsetneq I} Q_I(X_I) \\ &= \sum_{J\subsetneq I} (-1)^{|I|-|J|}\mathbb{E}(Q(\mathbf{X})|X_J)\end{aligned}$$

5.2.2 Indice de Sobol

Definition 5.1. Pour tout ensemble $I \subset \{1, \dots, p\}$, on définit l'indice de Sobol

$$S_I = \frac{Var(Q_I(X_I))}{Var(Y)}$$

Ces indices sont normalisés, c'est-à-dire que

$$\sum_{I\subset\{1,\dots,p\}} S_I = 1$$

.

Remarque : Si $Y = f_1(X_1) + \dots + f_p(X_p)$, en particulier dans le cas linéaire, les seuls indices non nuls sont les indices d'ordre 1.

5.2.3 Indice total d'une variable

Une variable X_i influe sur la variabilité de Y à travers tous les indices de Sobol indexés par un ensemble contenant i. On définit

$$S_i^{tot} = \sum_{I \subset \{1,\dots,p\}, i \in I} S_I$$

Cet indice quantifie entièrement l'influence de X_i sur Y, que ce soit directement ou par interaction avec d'autres variables. En effet,

Proposition 5.4. *Soit $X_{-i} = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_p)$. Alors*

$$S_i^{tot} = 1 - \frac{Var(\mathbb{E}(Y|X_{-i}))}{Var(Y)} = \frac{\mathbb{E}(Var(Y|X_{-i}))}{Var(Y)}$$

La première égalité est démontrée dans (Garnier, 2017), la seconde résulte de la formule de la variance totale.

Exemple : Considérons $Y = X_1X_2 + X_3$, où les X_i sont indépendants, de loi $\mathcal{U}(0,1)$. Alors (calculs à faire en exercice) :

- $Var(Y) = \frac{19}{144}$
- $\mathbb{E}(Y|X_1) = \frac{1}{2}X_1 + \frac{1}{2}$, d'où $Var(\mathbb{E}(Y|X_1)) = \frac{3}{144}$ puis $S_1 = \frac{3}{19}$.
- Par symétrie, $S_1 = \frac{3}{19}$
- $\mathbb{E}(Y|X_3) = \frac{1}{4} + X_3$, d'où $Var(\mathbb{E}(Y|X_3)) = \frac{1}{12}$ puis $S_1 = \frac{12}{19}$.
- $Q_{1,2}(X_1, X_2) = X_1X_2 + \frac{1}{2} - \frac{1}{2}X_1 - \frac{1}{2} - \frac{1}{2}X_2 - \frac{1}{2} + \frac{3}{4} = X_1X_2 - \frac{1}{2}X_1 - \frac{1}{2}X_2 + \frac{1}{4}$. Alors, $Var(Q_{1,2}(X_1, X_2)) = \frac{1}{144}$ puis $S_{1,2} = \frac{1}{19}$.
- Les autres termes sont nuls.
- L'influence totale de X_1 est alors de $S_1^{tot} = \frac{4}{19}$. On peut également retrouver ce chiffre en calculant la variance de $\mathbb{E}(Y|X_{-1}) = \frac{1}{2}X_2 + X_3$ divisée par $Var(Y)$.

5.3 Estimation des indices de Sobol

5.3.1 Par méthode de Monte-Carlo

Proposition 5.5. *Soit i un indice entre 1 et p. Soit X_i' une copie indépendante de X_i, tirée suivant la même loi. On considère $Y = Q(X_i, X_{-i})$, $Y^i = Q(X_i, X'_{-i})$ et $Y^{-i} = Q(X_i', X_{-i})$. Alors*

$$S_i = \frac{Cov(Y, Y^i)}{Var(Y)}$$

et

$$S_i^{tot} = \frac{\mathbb{E}[(Y - Y^{-i})^2]}{2Var(Y)}$$

Remarque : La première égalité est encore valable si on remplace l'indice i par un ensemble d'indices I.

Démonstration. La première équation découle de

$$\begin{aligned} Cov(Y, Y^i) &= \mathbb{E}(YY^i) - \mathbb{E}(Y)\mathbb{E}(Y^i) \\ &= \mathbb{E}[\mathbb{E}(YY^i|X_i)] - \mathbb{E}(Y)^2 \\ &= \mathbb{E}[\mathbb{E}(Y|X_i)\mathbb{E}(Y^i|X_i)] - \mathbb{E}(Y)^2 \\ &= \mathbb{E}[\mathbb{E}(Y|X_i)^2 - \mathbb{E}[\mathbb{E}(Y|X_i)^2] \\ &= Var(\mathbb{E}(Y|X_i)) \end{aligned}$$

Pour la seconde, constatons d'abord que

$$\mathbb{E}[YY^{-i}] = \mathbb{E}(\mathbb{E}[YY^{-i}|X_{-i}]) = \mathbb{E}(\mathbb{E}[Y|X_{-i}]\mathbb{E}[Y^{-i}|X_{-i}]) = \mathbb{E}(\mathbb{E}[Y|X_{-i}]^2)$$

Alors

$$\begin{aligned} \frac{1}{2}\mathbb{E}[(Y - Y^{-i})^2] &= \frac{1}{2}(\mathbb{E}[Y^2] - 2\mathbb{E}[YY^{-i}] + \mathbb{E}[(Y^{-i})^2]) \\ &= \mathbb{E}[Y^2] - \mathbb{E}(\mathbb{E}[Y|X_{-i}]^2) \\ &= \mathbb{E}(\mathbb{E}[Y^2|X_{-i}]) - \mathbb{E}(\mathbb{E}[Y|X_{-i}]^2) \\ &= \mathbb{E}(\mathbb{E}[Y^2|X_{-i}] - \mathbb{E}[Y|X_{-i}]^2) \\ &= \mathbb{E}(Var(Y|X_{-i})) \end{aligned}$$

□

Cette reécriture permet d'aborder l'estimation des indices de Sobol à partir l'approche de Monte-Carlo.

Pour cela, on considère deux échantillons $\mathbf{X^{(k)}} = (X_1^{(k)}, \dots, X_p^{(k)}), 1 \leq k \leq n$ et $\mathbf{X'^{(k)}} = (X_1'^{(k)}, \dots, X_p'^{(k)}), 1 \leq k \leq n$ et les estimateurs suivants

$$\widehat{Q_0} = \frac{1}{n}\sum_{k=1}^{n} Q(\mathbf{X^{(k)}})$$

$$\widehat{Var(Y)} = \frac{1}{n}\sum_{k=1}^{n} Q(\mathbf{X^{(k)}})^2 - \widehat{Q_0}^2$$

$$\widehat{Cov(Y, Y^i)} = \frac{1}{n}\sum_{k=1}^{n} Q(X_1^{\prime(k)}, \dots, X_{i-1}^{\prime(k)}, X_i^{(k)}, X_{i+1}^{\prime(k)}, \dots, X^{\prime(k)_p})$$

$$\hat{S}_i = \frac{\widehat{Cov(Y, Y^i)}}{\widehat{Var(Y)}}$$

$$\mathbb{E}[\widehat{(Y - Y^{-i})^2}] = \frac{1}{n}\sum_{k=1}^{n} \left(Q(\mathbf{X^{(k)}}) - Q(X_1^{(k)}, \dots, X_{i-1}^{(k)}, X_i^{\prime(k)}, X_{i+1}^{(k)}, \dots, X^{(k)_p})\right)^2$$

$$\hat{S_i^{tot}} = \frac{\mathbb{E}[\widehat{(Y - Y^{-i})^2}]}{2\widehat{Var(Y)}}$$

Ces estimateurs sont biais, consistants et asymptotiquement normaux (Prieur, 2022). Il est possible de réduire le nombre de simulations en utilisant des méthodes de Monte-Carlo ou des hypercubes latins comme décrit au chapitre ??.

Chapitre 6

Métamodèles

Toutes les méthodes vues précédemment nécessitent un nombre assez important d'appels à la fonction Q pour obtenir des estimations suffisamment précis des quantités d'intérêt.

Dans le cas d'une fonction Q trop coûteuse à évaluer un grand nombre de fois, il peut être intéressant de la remplacer par une fonction moins coûteuse à évaluer et en réalisant une bonne approximation. On parle alors de métamodèle. Ce chapitre en présente certains d'entre eux.

6.1 Modèle linéaire

6.1.1 Modèle linéaire gaussien multiple et écriture matricielle

On se place dans le cas d'une variable à expliquer Y unidimensionnelle. Le modèle linéaire gaussien consiste à considérer que la relation entre Y et les variables explicatives X_i s'écrit

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

les erreurs ϵ étant indépendants entre les mesures.

Pour un ensemble de n mesures consistant, on utilise la notation matricielle

$$Y = X\beta + \epsilon$$

où

— $Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ est le vecteur des observations de la variable à expliquer

— $X = \begin{pmatrix} 1 & x_{11} & ... & x_{1p} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & ... & x_{np} \end{pmatrix}$ est le vecteur des observations des variables explicatives, x_{ij} désignant l'observation de la variable j pour l'individu i.

— $\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$ est le vecteur des coefficients de la relation linéaire.

— $\epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$ est le vecteur des erreurs.

Les hypothèses suivantes sont faites sur ce modèle

$$\begin{array}{ll} (H1) & rg(X) = p \\ (H2) & \epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n) \end{array}$$

On notera que l'hypothèse $(H1)$ nécessite pour être valable que $n \geq p$ (il faut qu'il y ait plus d'observations que de variables explicatives). De plus, elle entraîne que $X'X$ est inversible (si il existait un vecteur v tel que $X'Xv = 0$, on aurait $v'X'Xv = 0$ puis $Xv = 0$, ce qui contredirait $(H1)$).

Remarques :

— Dans le cas d'une variables Y multidimensionnelle, le modèle se généralise avec β qui devient une matrice dont chaque colonne définit les coefficients pour une dimension de Y, et le bruit modélisé par une normale multivariée de matrice de variance-covariance Σ.

— On remarque que le modèle ne prédit pas une valeur pour Y sachant X mais une loi. Il est cependant courant de considérer comme prédiction la valeur $Y^{pred} = \mathbb{E}(Y|X) = X\beta$, la loi normale permettant de prendre en compte une incertitude à X fixé.

6.1.2 Estimation

L'estimation des coefficients (hors σ^2) peut se faire en résolvant le problème des moindres carrés

$$\hat{\beta}_{mc} = \underset{\beta \in \mathbb{R}^{p+1}}{argmin} \|Y - X\beta\|_2^2$$

Elle peut également se faire (en incluant σ^2) par maximum de vraisemblance en résolvant

$$\hat{\beta}_{emv}, \hat{\sigma}^2_{emv} = \underset{\beta \in \mathbb{R}^{p+1}, \sigma^2 \in \mathbb{R}^+}{argmin} -\frac{n}{2} \log \sigma^2 + \sum_{i=1}^{n} \frac{(y_i - X_{i\bullet\beta})^2}{2\sigma^2}$$

Théorème 6.
— $\hat{\beta}_{mc} = \hat{\beta}_{emv} = (X'X)^{-1}X'Y$
— $\hat{\sigma}^2_{emv} = \frac{\|Y - \hat{Y}\|^2}{n}$

Les démonstrations et des énoncés des propriétés asymptotiques de ces estimateurs sont disponibles par exemple dans (Guyader, 2012).

6.1.3 Interprétation géométrique et coefficient R^2

On peut décomposer Y en

$$\begin{aligned} Y &= X\hat{\beta} + Y - X\hat{\beta} \\ &= X(X'X)^{-1}X'Y + (I_n - X(X'X)^{-1}X')Y \\ &= P_X Y + (I_n - P_X)Y \end{aligned}$$

avec $P_X = X(X'X)^{-1}X'$. Or, $P_X^2 = P_X$ et $P_X' = P_X$. P_X est donc un projecteur orthogonal. On peut ainsi voir $\hat{Y} = X\beta$ comme la projection orthogonale dans $\mathbb{R}^n$ de Y (l'échantillon à expliquer) sur l'espace engendré par les colonnes de X (les échantillons explicatifs).

Cette projection est difficile à se représenter, même pour une seule variables explicative, puisque le dimension n est celle de la taille de l'échantillon.

Cependant, elle est important puisqu'elle implique l'orthogonalité des deux termes de la décomposition $Y = \hat{Y} + (Y - \hat{Y})$.

Notons $\bar{Y} = \bar{y}1$ le vecteur dont toutes les coordonnées valent la moyenne des y_i. On constate, en reécrivant l'équation $\nabla C(\beta') = 0$ en $(\hat{Y} - Y)'X = 0$ et en regardant uniquement la première colonne de X, que $\hat{Y} - Y$ est également orthogonal à 1 donc à $\bar{Y}$.

Cela entraîne que les termes de la décomposition

$$Y - \bar{Y} = (\hat{Y} - \bar{Y}) + (Y - \hat{Y})$$

sont orthogonaux. Le théorème se Pythagore implique alors

$$\|Y - \bar{Y}\|^2 = \|\hat{Y} - \bar{Y}\|^2 + \|Y - \hat{Y}\|^2$$

soit, comme dans le cas de la régression simple,

$$SCT = SCE + SCR$$

On peut donc utiliser $R^2 = \frac{SCE}{SCT}$ comme indicateur de la variabilité de Y expliquée par le modèle.

6.1.4 Estimation de l'erreur de prédiction

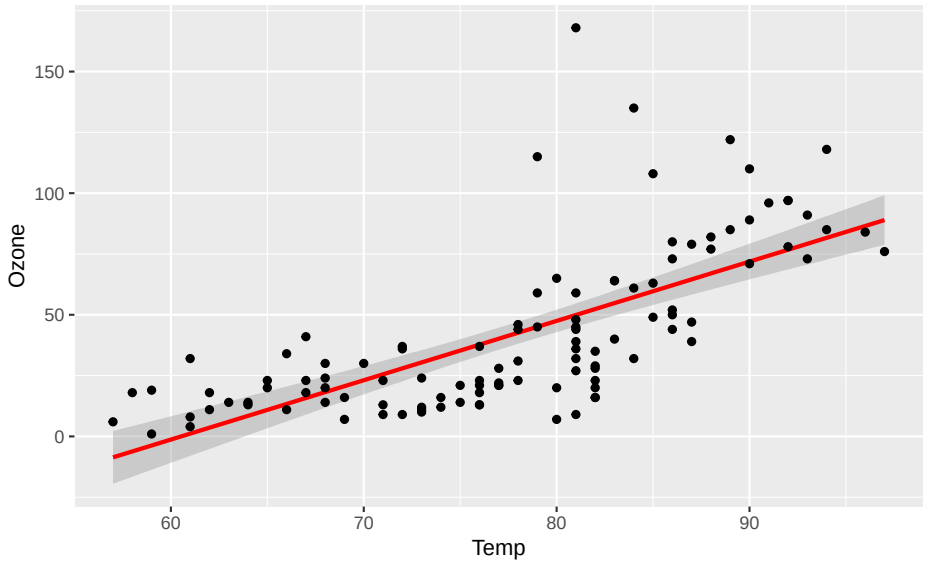
Il est intéressant de noter que dans ce modèle, il est possible d'obtenir une idée de l'incertitude qui porte sur une prédiction. Seul le résultat est donné ici, la démonstration pouvant être trouvé à nouveau dans (Guyader, 2012).

Théorème 7. *Soit $x_{new} \in \mathbb{R}^{p+1}$ un vecteur formé des observations des p variables pour un nouvel individu, précédées d'un 1. On note $\hat{y}_{new} = x'_{new}\beta$ la prédiction obtenue et $\hat{\epsilon}_{new} = y_{new} - \hat{y}_{new}$ l'erreur de prédiction.*

Alors,

$$\mathbb{E}(\hat{\epsilon}_{new}) = 0$$
$$Var(\hat{\epsilon}_{new}) = \sigma^2(1 + x'_{new}(X'X)^{-1}x_{new})$$

En remplaçant σ^2 par $\hat{\sigma}^2_{emv}$, on peut approcher des intervalles de prédiction dans lesquels la véritable prédiction se trouve avec probabilité $1 - \alpha$.



6.1.5 Indices de Sobol pour un modèle linéaire gaussien

Considérons un modèle linéaire gaussien avec des variables d'entrées **indépendantes**.

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Alors

$$\mathbb{E}(Y|X_i) = \beta_0 + \sum_{j \neq i} \beta_j \mathbb{E}(X_j) + \beta_i X_i$$

puis $Var(\mathbb{E}(Y|X_i)) = \beta_i^2 Var(X_i)$

On obtient ainsi

$$S_i = \frac{\beta_i^2}{\sum_i \beta_i^2 + \sigma^2} \quad S_\epsilon = \frac{\sigma^2}{\sum_i \beta_i^2 + \sigma^2}$$

Dans le cas de non-indépedance des variables d'entrée, il n'y a pas de formule simple mais le coût des estimations à l'aide de la méthode de Monte-Carlo est très faible.

6.1.6 Extensions

Le modèle linéaire gaussien a plusieurs limites pour lesquelles des extensions sont envisageables.

6.1.6.1 Termes non linéaires

Le modèle ne prend en compte que des effets linaéaires des entrées sur la sortie. Il est possible d'introduire des interactions entre variables, par exemple un terme $\beta_{ij}X_iX_j$, ou des termes non linéaires comme $\gamma_i X_i^2$. Il faut cependant une bonne connaissance de l'application dans la mesure où ces termes doivent être choisis en amont (ils ne sont pas appris par le modèle)

6.1.6.2 Régressions pénalisées

Dans le cas de variables explicatives corrélées, l'interprétation de leurs coefficients est difficile. En effet, une variable peut 'attirer' l'influence d'une autre variable qui lui est fortement corrélée.

La première possibilité pour effectuer une régression pénalisée est d'utiliser une pénalité de type Ridge, déjà rencontrée au chapitre 4. L'idée est de forcer le vecteur de coefficients β à être borné en norme L_2. Intuitivement, les problèmes de colinéarité sont alors réglés par le fait qu'une variable ne peut 'attirer' les coefficients des variables qui lui sont corrélées que dans une certaine mesure.

L'estimateur ridge est un estimateur biaisé, contrairement à l'estimateur des moindes carrés. Par contre, il est de moindre variance.

A contrario, la pénalité Lasso permet de faire de la sélection de variables, notamment dans le cas de la grande dimension où le nombre d'appels à Q et donc d'observations est trop faible.

Dans ce cas, les variables corrélées vont au contraire avoir tendance à se regrouper dans le sens où tout le poids va être accaparée par une seule d'entre elles. Ceci se traduit par un phénomène d'instabilité du modèle (des données légèrement différentes vont donner des sélections différentes).

Une manière d'essayer de tirer le meilleur des deux pénalités est de les mélanger via une pénalité Elastic-Net.

Remarques :

1. Les pénalités Ridge, Lasso ou Elastic-Net rendent le problème strictement convexe, et donc résoluble par une descente de gradient quelque soit la dimension.

2. Le choix des paramètres peut se faire par validation croisée ou, dans le cadre de la sélection de variables, par une étape de *Stability Selection* suivie d'une estimation de modèle non pénalisé
3. En cas de variables corrélées, la répartition entre les variables est meilleure : l'estimateur des moindres carrés risque de mettre tout le poids sur une variable, ce qui n'est pas le cas de l'estimateur ridge.
4. La corrélation des variables peut être gérée par l'application d'une ACP afin de n'avoir que des variables non corrélées dans le modèle. Le prix à payer est par contre une plus grande difficulté d'interprétations, les variablees explicatives étant du coup des combinaisons linéaires des véritables variables.

6.1.6.3 Variables non gaussiennes : modèle linéaire généralisé

Notons que le modèle linéaire gaussien peut s'écrire de façon équivalente

$$\begin{aligned} y_i|X_{i\bullet} &\sim \mathcal{N}(\mu_i, \sigma^2) \\ \mu_i &= X'_{i\bullet}\beta \end{aligned}$$

ce qui revient à dire à choisir une forme de distribution paramétrique (la loi normale) pour $Y|X$ et à définir une relation entre un paramètre de cette distribution et une combinaison linéaire des entrées.

Cette manière de faire peut se généraliser à d'autre lois que la loi normale, quand $Y|X$ ne peut pas être décrite par une loi normale. Cela est par exemple le cas si Y est une variable binaire (loi de Bernoulli), une variable de comptage (loi de Poisson) ou un temps d'attente (loi exponentielle). On parla alors de **modèle linéaire généralisé** et il possible de l'utiliser pour toute forme de loi faisant partie de la famille exponentielle.

A titre d'exemple, il est possible de traiter le cas de sorties binaires à l'aide du modèle suivant, appelé aussi régression logistique

$$\begin{aligned} y_i|X_{i\bullet} &\sim \mathcal{B}(p_i) \\ \log(\tfrac{p_i}{1-p_i}) &= X'_{i\bullet}\beta \end{aligned}$$

Ces modèles peuvent être estimés en maximisant la vraisemblance à l'aide d'algorithmes de descente de gradient.

6.2 Modélisation par processus gaussien - krigeage

La modélisation par processus gaussien est une généralisation du modèle linéaire qui permet d'introduire une forme non-linéaire pour $\mathbb{E}(Y|X)$

6.2.1 Processus gaussien

Un processus gaussien $Z(\mathbf{x})$ est un processus qui en tout point de $\mathbb{R}^p$ correspond à une loi définie par :

- une fonction $m(\mathbf{x})$ qui donne l'espérance de $Z(\mathbf{x})$
- le fait que pour tout ensemble de points $(\mathbf{x_1}, \dots, \mathbf{x_d})$ de $\mathbb{R}^p$, la loi jointe de $(Z(\mathbf{x_1}), \dots, Z(\mathbf{x_d}))$ est une loi gaussienne multidimmensionnelle Σ

En d'autres termes, définir un processus gaussien revient à définir la fonction moyenne $m(\mathbf{x})$ et une fonction de covariance $k(\mathbf{x}, \mathbf{x}') = cov(Z(\mathbf{x}), Z(\mathbf{x}'))$.

Le modèle linéaire est un cas particulier où $m(\mathbf{x}) = \mathbf{x^T}$ et $k(\mathbf{x}, \mathbf{x}') = \sigma^2 \mathbb{1}_{\mathbf{x},\mathbf{x}'}$.

L'approche la plus couramment utilisée est de considérer une fonction de covariance qui est décroissante en la distance entre $\mathbf{x}$ et $\mathbf{x}'$ et stationnaire, c'est-à-dire ne dépendant que de la différence $\mathbf{x} - \mathbf{x}'$. Un modèle couramment utilisé est par exemple le modèle exponentiel anisotrope

$$k(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\big(-\frac{1}{2} \frac{\|\mathbf{x} - \mathbf{x}'\|^2}{\theta} \big)$$

Il est à noter que l'utilisation d'une loi normale entraîne que le conditionnement de la loi normale jointe de $(Z(\mathbf{x}), Z(\mathbf{x_1}), \dots, Z(\mathbf{x_d}))$ par des valeurs observées $Z(\mathbf{x_1}) = z_1, \dots, Z(\mathbf{x_d}) = z_d$ donne toujours une loi normale pour $Z(\mathbf{x})|Z(\mathbf{x_1}) = z_1, \dots, Z(\mathbf{x_d}) = z_d$.

L'intérêt d'une approche par processus gaussien est de pouvoir introduire des formes non linéaires pour la moyenne, ce qui permet une flexibilité plus grande en termes de modélaisation. L'introduction d'une covariance non nulle entre points distincts permet également d'avoir une incertitude autour de la prediction moyenne qui dépend de la proximité des points auxquels ont été faits des mesures.

6.2.2 Régression par processus gaussien ou Krigeage

Pour définir la moyenne du processus gaussien, on fait le choix d'une base de fonctions

$$f = (f_1, ..., f_K)$$

et d'un paramètre $\in \mathbb{R}^K$. On définit la moyenne modèle linéaire :

$$m(\mathbf{x}) = \mathbf{f}(\mathbf{x})^T$$

Remarque : Les fonctions de base peuvent être choisis non linéaires (par exemple les polynômes d'un certain degré), ce qui permet d'introduire de la non-linéarité.

Deux étapes peuvent être distinguées dans le krigeage :

1. l'estimation des paramètres , σ et ;
2. le conditionnement du processus par les données $\mathbf{y}$.

6.2.2.1 Étape (1) : estimation des (hyper)paramètres

Une approche rencontrée dans la littérature, appelée *bayesian kriging* ou *full bayesian kriging*, consiste à traiter les variables , σ ou θ comme des paramètres incertains que l'on associe à une distribution de probabilité à priori.

Elles peuvent aussi être considérées comme des paramètres à estimer à partir de données censées provenir d'une même réalisation (de la même trajectoire) d'un processus gaussien. Cela est réalisé suivant le principe du maximum de vraisemblance (cf TP avec OpenTurns)

6.2.2.2 Étape (2) : conditionnement

Supposons les paramètres précédents coonus et considérons le problème de la prédiction en un point $\mathbf{x} \in \mathbb{R}^p$.

Faisons l'hypothèse que l'on connaît les valeurs de la fonction sur un plan d'expériences $\mathbf{x}^{(i)}$ pour $i = 1, ..., n$ où n est le nombre de simulations. Pour chacune de ces entrées, on suppose que l'on connaît la valeur de la sortie scalaire $y^{(i)}$ pour $i = 1, ..., n$. On note $\mathbf{y} = (y_1, ..., y_n)^T$ le vecteur des sorties observées.

Notons F la matrice de conception associée aux fonctions de base :

$$F = \left[f_j \left(\mathbf{x}^{(i)} \right), \quad i = 1, ..., n, \quad j = 1, ..., p \right].$$

Notons R la matrice de covariance associée au noyau de covariance :

$$R = \left[k \left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}, \right), \quad i, j = 1, ..., n \right]$$

Considérons un nouveau point $\mathbf{x} \in \mathbb{R}^d$ correspondant à une sortie y inconnue. Notons $\mathbf{k}(\mathbf{x})$ le vecteur des covariances entre le point $\mathbf{x}$ et les points du plan d'expériences :

$$\mathbf{k}(\mathbf{x}) = \left[k \left(\mathbf{x}, \mathbf{x}^{(i)}, \right), \quad i = 1, ..., n \right]^T$$

On peut démontrer que le vecteur aléatoire associé aux observations $\mathbf{Y}$ et la variable aléatoire $Y(\mathbf{x})$ sont liés par une loi normale :

$$\begin{pmatrix} \mathbf{Y} \\ Y(\mathbf{x}) \end{pmatrix} = \mathcal{N} \left(\begin{pmatrix} F \\ \mathbf{f}(\mathbf{x})^T \end{pmatrix}, \begin{pmatrix} R & \mathbf{k}(\mathbf{x}) \\ \mathbf{k}(\mathbf{x})^T & 1 \end{pmatrix} \right)$$

Le modèle prédictif est donné par loi de $Z(\mathbf{x})$ conditionnée par les observations connues du code :

$$\tilde{Y}(\mathbf{x}) = \left[\mathbf{Y}(\mathbf{x}) | \mathbf{Y} = \mathbf{y} \right].$$

On peut démontrer que $\tilde{\mathbf{Y}}(\mathbf{x})$ est également une variable aléatoire gaussienne :

$$\tilde{Y}(\mathbf{x}) \sim \mathcal{N}\left(\mu_{\tilde{Y}}(\mathbf{x}), \sigma_{\tilde{Y}}(\mathbf{x})^2\right)$$

où la moyenne $\mu_{\tilde{Y}}(\mathbf{x})$ et la variance $\sigma_{\tilde{Y}}(\mathbf{x})^2$ s'écrivent de manière explicite.

Les calculs liés au conditionnement n'impliquent que la résolution de systèmes d'équations linéaires. Néanmoins, si n est grand (par exemple $n = 10000$), alors la matrice de covariance R est de taille $n \times n$, ce qui peut poser des difficultés de performance, voire de mémoire. Pour résoudre ce problème, une alternative consiste à utiliser des techniques de compression de matrices, comme par exemple la technique des H-mat utilisée par OpenTURNS.

Remarque : Il est possible de modifier la définition de la fonction de covariance conditionnée pour tenir compte de l'incertitude sur l'estimation des paramètres du modèle, ou d'ajouter une incertitude sur la mesure des données observées.

6.2.3 Indices de Sobol

L'approche par processus gaussiens ne permet pas de déterminer les indices de Sobol de façon théorique. Elle permet cependant de bénéficier d'un métamodèle permettant de générer des échantillons pour les approches de type Monte-Carlo de façon peu coûteuse.

6.3 Polynôme de chaos

Le modèle linéaire a pour avantage la simplicité de son écriture et donc de son étude, notamment en termes de propagation de l'incertitude. Il est cependant limité d'un point de vue applicatif dans la mesure où il ne prend pas en compte les interactions entre variables à moins de les introduire explicitement à priori, et qu'il se limite à des effets linéaires des entrées sur les sorties.

Une manière alternative d'écrire Y en fonction des X_i est de considérer qu'elle s'écrit comme une série de fonctions, suivant une base de fonctions bien choisie, qu'on nommera polynômes de chaos.

6.3.1 Dimension 1

Definition 6.1. Soit μ une mesure positive sur $\mathbb{R}$. On peut définir un produit scalaire entre fonctions de carré intégrable dans $L^2(\mu)$ par

$$< f, g >_{L^2(\mu)} = \int_{-\infty}^{+\infty} f(x)g(x)\mu(x)dx$$

Une famille de polynômes $(P_n)_{n\geq 0}$ est othogonale si P_n est de degré n et si, pour tout $n \neq m$,

$$< P_n, P_m >_{L^2(\mu)} = 0$$

Exemples :

1. Les polynômes de Legendre sont définis par

$$L_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}[(x^2-1)^n]$$

Ils vérifient $\int_{-1}^{1} L_n(x)L_m(x)\frac{1}{2}dx = 0$ et $\int_{-1}^{1} L_n(x)^2\frac{1}{2}dx = \frac{1}{2n+1}$ Ce sont donc des polynômes orthogonaux pour la mesure uniforme sur $[-1,1]$.

2. Les polynômes d'Hermite sont définis par

$$H_n(x) = (-1)^n e^{x^2/2}\frac{d^n}{dx^n}[e^{-x^2/2}]$$

Ils vérifient $\int_{-\infty}^{+\infty} H_n(x)H_m(x)\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx = 0$ et $\int_{-\infty}^{+\infty} H_n(x)^2\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx = n!$ Ce sont donc des polynômes orthogonaux pour la mesure correspondant à la densité de la loi normale centrée réduite.

Théorème 8. *Soit P_n une suite de polynômes othogonaux pour la mesure μ et f une fonction continue telle que $\int f^2 d\mu < +\infty$.*

On définit, pour tout n,

$$a_n = \frac{< f, P_n >_{L^2(\mu)}}{< P_n, P_n >_{L^2(\mu)}}$$

et

$$f_n(x) = \sum_{k=0}^{n} a_k P_k(x)$$

Alors, la suite (f_n) converge uniformément vers f sur tout intervalle $[a,b]$.

Remarque : Cette convergence est d'autant plus rapide que la fonction cible est régulière (Garnier, 2017).

6.3.2 Dimension quelconque

Les notions et résultats du paragraphe précédents se généralise en dimension p, en remplaçant les monômes x^k par des monômes $\mathbf{x} = x_1^{\alpha_1}x_2^{\alpha_2}...x_p^{\alpha_p}$, le degré d'un monôme étant $\sum_i \alpha_i$.

Dans ce cas, pour une mesure sur $\mathbb{R}^p$, une famille de polynômes othogonaux est une famille indexée par $\mathbb{N}^p$ telle que

— Le monôme P a α comme monôme de plus grand degré.
— Pour tout $\neq$, $< P, P >_{L^2(\,)} = 0$

En particulier, si est une mesure produit

$$d\,(x) = d\mu_1(x_1)\,...\,d\mu_p(x_p)$$

et qu'on dipose d'une famille de polynômes orthogonaux $(P_k^{(i)})_{k\geq 0}$ suivant μ_i pour chaque $1 \leq i \leq p$, on peut construire une famille de polynômes orthogonaux pour en considérant

$$P_\alpha = P_{\alpha_1}^{(1)}\,...\,P_{\alpha_p}^{(p)}$$

A nouveau, il est possible d'approcher toute fonction f par sa projection sur l'ensemble des polynômes de la famille ayant un degré inférieur à n en considérant

$$f_n(\mathbf{x}) = \sum_{\alpha,|\alpha|\leq n} a_\alpha P_\alpha(\mathbf{x})$$

avec

$$a_\alpha = \frac{< f, P_\alpha >_{L^2(\mu)}}{< P_\alpha, P_\alpha >_{L^2(\mu)}}$$

6.3.3 Estimation des coefficients

Une manière de faire est d'écrire que, pour une variable X suivant une loi de distribution μ,

$$a_\alpha = \frac{\mathbb{E}(f(X)P_\alpha(X))}{\mathbb{E}(P_\alpha(X)^2)}$$

On peut alors utiliser une estimation de type Monte-Carlo.

Il est cependant plus efficace d'utiliser une approche par régression linéaire.

En effet, supposons que

$$f(X) = \sum_{\alpha\in\mathbb{N}^p} y_\alpha P_\alpha(X)$$

Alors, par orthogonalité des P_α,

$$< f_n - f, f_n - f >_{L^2(\mu)} = \sum_{\alpha,|\alpha|\leq n} (y_\alpha - a_\alpha)^2 < P_\alpha, P_\alpha >_{L^2(\mu)} + \sum_{\alpha,|\alpha|>n} y_\alpha^2 < P_\alpha, P_\alpha >_{L^2(\mu)}$$

On peut donc voir le vecteur des a_α comme la solution à l'équation

$$\underset{b_\alpha}{\operatorname{argmin}}(\|f(X) - \sum_{\alpha,|\alpha|\leq n} b_\alpha P_\alpha(X)\|^2_{L^2(\mu)})$$

c'est-à-dire comme les coefficients de la régression linéaire de f par les P_α.

Ainsi, en partant d'un échantillon de valeurs $(X^{(k)})_{1\leq k\leq K}$, on considère

$$\widehat{a_\alpha} = \underset{y_\alpha}{\operatorname{argmin}}(\sum_{k=1}^{K}(f(X^{(k)}) - \sum_{\alpha,|\alpha|\leq n} y_\alpha P_\alpha(X^{(k)})^2)$$

que l'on peut obtenir par

$$\widehat{a_\beta} = (P'P)^{-1}P'F$$

où P est la matrice des $P_\alpha(X^{(k)})$ dont les lignes sont indexées par les k et les colonnes par les α, et F est le vecteur des $f(X^{(k)}$.

6.3.4 Indices de Sobol pour une métamodélisation par polynôme de chaos

L'un des avantages de la décomposition en polynôme de chaos est la simplicité de l'écriture des indices de Sobol. En effet, considérons une décomposition en polynômes de chaos de $Y = Q(X_1, \dots, X_p)$ sous la forme

$$Y = \sum_{\alpha\in\mathbb{N}^p} y_\alpha P_\alpha(X)$$

Pour tout ensemble d'indices $I \subseteq \{1, \dots, p\}$, on définit $\mathcal{A}_I$ comme l'ensemble des indices α non nuls exactement sur I :

$$\mathcal{A}_I = \{\alpha | \alpha_i > 0 \text{ ssi } i \in I\}$$

Alors, la décomposition de Hoeffding s'écrit

$$Y = y_{\mathbf{0}} + \sum_{I\subseteq\{1,\dots,p\}, I\neq\emptyset} Q_I(X)$$

avec

$$Q_I(X) = \sum_{\alpha\in\mathcal{A}_I} y_\alpha P_\alpha(X)$$

Démonstration. - $\mathbb{E}(Y) = y_0$

- Soit $J \subset I$, $J \neq I$. Quitte à réordonner les indices, soit $I = (i_1, \dots, i_l)$ et $J = (i_1, \dots, i_k)$, $k < l$. Alors

$$\begin{aligned}\mathbb{E}(Q_I(X_I)|X_J) &= \sum_{\alpha\in\mathcal{A}_I} y_\alpha \mathbb{E}(P_\alpha(X)|X_J) \\ &= \sum_{\alpha\in\mathcal{A}_I} y_\alpha \mathbb{E}(P^{i_1}_{\alpha_1}(x_1)\dots P^{i_l}_{\alpha_1}(x_l)|X_J) \\ &= \sum_{\alpha\in\mathcal{A}_I} y_\alpha P^{i_1}_{\alpha_1}(x_1)\dots P^{i_k}_{\alpha_1}(x_k)\mathbb{E}(P^{i_{k+1}}_{\alpha_{k+1}}(x_{k+1})\dots P^{i_l}_{\alpha_1}(x_l)|X_J) \\ &= 0\end{aligned}$$

car $\mathbb{E}(P^{i_{k+1}}_{\alpha_{k+1}}(x_{k+1})\dots P^{i_l}_{\alpha_1}(x_l)|X_J) = 0$. □

Il découle alors de l'orthogonalité des P_α que

$$Var(Y) = \sum_{\alpha\in\mathbb{N}^p} y_\alpha^2$$

et que, pour toute variable X_i, on peut écrire l'indice de Sobol associé

$$S_i = \frac{\sum_{\alpha\in\mathcal{A}_i} y_\alpha^2}{\sum_{\alpha\in\mathbb{N}^p} y_\alpha^2}$$

ainsi que l'indice total

$$S_i^{tot} = \frac{\sum_{\alpha\in\mathcal{A}_I, i\in I} y_\alpha^2}{\sum_{\alpha\in\mathbb{N}^p} y_\alpha^2}$$

6.4 Réseaux de neurones

Il est également possible d'utiliser des réseaux de neurones comme métamodèles pour Y. Ils ne sont pas développés ici, mais ils représentent clairement une façon de modéliser Y en fonction des X_i, en permettant une approche non-linéaire et possiblement plus proche de la réalité que les méthodes précédentes. De plus, la propagation forward sur un réseau déjà appris permet d'obtenir facilement et de façon peu coûteuse de nombreuses réalisations du système, permettant de développer sans difficultés les approches de type Monte-Carlo ou quasi-Monte-Carlo.

La difficulté réside dans ce cas dans l'apprentissage du réseau, qui nécessite un grand nombre de données disponibles au préalable, et ne rend pas toujours cette approche possible.

Chapitre 7

Indices de Shapley

L'approche par Indices de Sobol nécessite que les variables d'entrée puissent être supposées indépendantes, afin que la décomposition de Hoeffding existe et de façon unique. En pratique, cette hypothèse peut être discutable.

Les indices de Shapley sont une manière d'aborder la même question en ne faisant pas cette hypothèse d'indépendance, inspiré de la théorie des jeux.

7.1 Définition

Soit I un ensemble $I \subset \{1, \dots, p\}$, où p est le nombre de variables d'entrée. On définit le gain d'explication lié à I comme l'indice de Sobol lié à I :

$$c(I) = \frac{Var(\mathbb{E}(Y|X_I))}{Var(Y)}$$

La **valeur de Shapley** de la variable j est alors définie comme

$$Sh_j = \frac{1}{p} \sum_{I, j \notin I} \binom{p-1}{|I|}^{-1} (c(I \cup j) - c(I))$$

Remarques : - il est possible de montrer qu'il est équivaletn de remplacer c par $\bar{c}(I) = \frac{Var(\mathbb{E}(Y|X_{-I}))}{Var(Y)}$ - Sh_j est un gain moyen en termes d'indices de Sobol quand on ajoute la variable j à un ensemble de variables explicatives. Cette moyenne est pondérée avec un poids $\binom{p-1}{|I|}^{-1}$ qui permet de donner le même poids à l'ensemble des passages de taille i à $i+1$, pour tout i. Sans les poids, une importance trop grande serait donnée aux ensembles de taille proche de $\frac{p}{2}$ qui sont beuacoup plus nombreux. - L'indice de Shapley est compris entre les

indices de Sobol de premier ordre et total de la variable j

$$S_j \leq \eta_j \leq S_j^{tot}$$

- L'indice de Shapley ne permet pas de différencier la partie dûe à la variable seule et celle due à ses interactions avec d'autres variables.

Le grand avantage des indices de Shapley par rapport à ceux de Sobol est qu'ils sont définis sans hypothèse d'indépendance entre entrées. Par contre, des variables très fortement corrélées seront indiscernables par Shapley.

Exemple : Exemple si $X_1 = X_2$, les modèles $Y = 1 \times X_1 + 0 \times X_2$, $Y = 0 \times X_1 + 1 \times X_2$ ou $Y = .5 \times X_1 + .5 \times X_2$ donneront tous $Sh_1 = Sh_2 = .5$.

7.2 Estimation

Il est parfois possible de faire des calculs explicites dans certains cas, comme celui du modèle linéaire où $S_j = Sh_j = S_j^{tot}$.

Pour les cas où ce n'est pas possible, le calcul exact pose un problème combinatoire : il faudrait déterminer les gains de tous les sous-ensembles de variables, soit 2^p calculs à faire.

Il est possible de montrer qu'on peut reécrire les indices sous la forme :

$$Sh_j = \frac{1}{p!} \sum_{\sigma \in \mathcal{S}_p} \left(c([\sigma]_j \cup j) - c([\sigma]_j) \right)$$

où $\mathcal{S}_p$ est l'ensemble des permutations de $\{1, \dots, p\}$ et $[\sigma_j]$ est l'ensemble des indices précédant j dans la permutation σ.

Quand le nombre de variables est trop important pour énumérer toutes les permutations, cette somme peut être approximée par

$$\frac{1}{m} \sum_{i=1}^{m} \left(c([\sigma^i]_j \cup j) - c([\sigma^i]_j) \right)$$

où $\sigma^1, \dots, \sigma^m$ sont des permutations tirées au hasard.

En y incorporant l'estimation $\hat{c}$ des indices de Sobol par Mante-Carlo, on obtient un estimateur de l'indice de Shapley sous la forme

$$\widehat{Sh}_j = \frac{1}{m} \sum_{i=1}^{m} \left(\hat{c}([\sigma^i]_j \cup j) - \hat{c}([\sigma^i]_j) \right)$$

Chapitre 8

Simulation de lois aléatoires et plans d'expériences

Les approches Monte-Carlo nécessaires pour des estimations liées à Y nécessitent de pouvoir simuler un grand nombre de valeurs de Y.

Pour cela, il faut pouvoir simuler des valeurs aléatoires de X puis de déterminer $Y = M(X)$. Ce chapitre présente les principales méthodes de simulation.

8.1 Méthodes de génération d'échantillons indépendants

8.1.1 Simuler des valeurs sous $\mathcal{U}(0,1)$

Il est possible de simuler des suites de nombres pseudo-aléatoires qui sont de bonnes approximation d'échantillons aléatoires indépendants suivant lois uniformes sur $[0,1]$.

Une manière courante de faire cela est de considérer une suite définie par

$$x_{n+1} = ax_n + b \text{ modulo } m$$

avec m très grand (2^{31} ou plus) et a et b bien choisis, notamment pour que la période soit m. La suite renvoyée est alors celle des $(\frac{x_n}{m})$.

8.1.2 Simuler suivant une loi de fonction quantile connue

Soit F la fonction de répartition d'une loi sous laquelle on souhaite simuler, et telle qu'on connaît la fonction quantile

$$F^{\leftarrow}(u) = inf\{x : F(x) \geq u\}$$

Si U suit une loi uniforme sur $[0,1]$, $F^{\leftarrow}(U)$ suit la loi correspondant à F. En effet,

$$\begin{aligned}\mathbb{P}(F^{\leftarrow}(U) \leq x) &= \mathbb{P}(F(F^{-}(U)) \leq F(x)) \\ &= \mathbb{P}(U \leq F(x)) \\ &= F(x)\end{aligned}$$

On peut ainsi simuler suivant la loi correspondant à F : il suffit de simuler suivant $\mathcal{U}(0,1)$ et d'appliquer $F^{\leftarrow}$.

Exemple : Si $X \sim \mathcal{E}(1)$, on a $F(x) = 1 - e^{-x}$. Son inverse est $F^{\leftarrow}(u) = -log(1-u)$. Par conséquent, si $U \sim \mathcal{U}(0,1)$, $-\log U \sim \mathcal{E}(1)$

8.1.3 Méthode d'acceptation-rejet

Les méthodes précédentes ne peuvent pas s'adapter à toutes les lois d'intérêt.

Soit f la densité sous laquelle on cherche à simuler, appelée *densité cible*. On considère une autre densité g, appelé *densité instrumentale*, telle que :

- il est aisé de simuler suivant g
- $supp(f) \subset supp(g)$
- il existe une constante C telle que $f(x) \leq Cg(x)$ pour tout x.

On génère alors un échantillon suivant l'algorithme suivant :

1. Générer X suivant la loi g.
2. Générer U suivant une loi $\mathcal{U}[0,1]$
3. Accepter (c'est-à-dire ajouter à l'échantillon) la valeur X si $U < \frac{f(X)}{Cg(X)}$

L'échantillon suit alors la loi de X.

En effet,

$$\begin{aligned}\mathbb{P}(X \leq x | U < \frac{f(X)}{Cg(X)}) &= \frac{\mathbb{P}(X \leq x, U < \frac{f(X)}{Cg(X)})}{\mathbb{P}(U < \frac{f(X)}{Cg(X)})} \\ &= \frac{\int_{-\infty}^{x} \int_{0}^{\frac{f(y)}{Cg(y)}} dug(y)dy}{\int_{-\infty}^{+\infty} \int_{0}^{\frac{f(y)}{Cg(y)}} dug(y)dy} \\ &= \frac{\int_{-\infty}^{x} \frac{f(y)}{Cg(y)} g(y)dy}{\int_{-\infty}^{+\infty} \frac{f(y)}{Cg(y)} g(y)dy} \\ &= \frac{\int_{-\infty}^{x} f(y)dy}{\int_{-\infty}^{+\infty} f(y)dy} \\ &= \mathbb{P}(X \leq x)\end{aligned}$$

Cette méthode permet ainsi de simuler sous f même sans être capable de déterminer la fonction quantile associée. Par contre, trouver une distribution g telle que C n'est pas trop grande peut être compliqué. Or, plus C est grande, plus l'algorithme va passer beaucoup de temps à générer des valeurs qui seront rejetées.

8.1.4 Echantillonnage préférentiel

On considère une estimation de Monte-Carlo consistant à estimer $\mathbb{E}_f(h(X))$ par $\frac{1}{n}\sum_{i=1}^{n} h(x_i)$, les x_i étant tirés indépendemment suivant f.

Soit g une densité définie sur Ω telle que $supp(h \times f) \subset supp(g)$, c'est-à-dire telle que $g(x) \neq 0$ si $f(x)h(x) \neq 0$. L'espérance à estimer peut se reécrire

$$\mathbb{E}_f(h(X)) = \int_\Omega \frac{h(t)f(t)}{g(t)} g(t)dt = \mathbb{E}_g(\frac{h(X)f(X)}{g(X)})$$

La méthode de Monte-Carlo peut alors être appliquée en échantillonnant les x_i suivant g plutôt que suivant f et en approchant l'intégrale par

$$\frac{1}{n}\sum_{i=1}^{n} \frac{h(x_i)f(x_i)}{g(x_i)}$$

La convergence vers $\mathbb{E}_f(h(X))$ quand n tend vers l'infini reste vraie, la différence étant que la variance de l'estimateur est alors

$$\frac{1}{n}\int_\Omega (\frac{h(t)f(t)}{g(t)} - \mathbb{E}_f(h(X)))^2 g(t)dt$$

Un choix judicieux de g peut réduire cette variance et donc l'amplitude des intervalles de confiance (un choix moins judicieux peut évidemment la faire exploser, voire rendre l'intégrale non convergente).

Exemple :(Robert et al., 2010)

On cherche à déterminer la p-valeur $\mathbb{P}(Z > 4)$ quand Z suit une loi normale centrée réduite. Soit f la densité d'une loi normale centrée réduite et $h(t) = \mathbb{1}_{t>4}$.

La méthode de Monte-Carlo appliquée à h et f va dans ce cas se révéler très lente puisque l'énorme majorité des valeurs échantillonnées suivant $\mathcal{N}(0,1)$ vont être inférieures à 4 (la vraie valeur recherchée étant de 3.2 10^{-5}, à peu près 1 valeur sur 30000 sera non nulle).

Une manière d'accélerer la convergence est alors de considérer la densité g d'une loi exponentielle de paramètre $\frac{1}{4}$. La proportion de valeurs échantillonées non nulle passe alors à plus d'un tiers, accélerant la convergence de l'algorithme.

La question du choix de la meilleure fonction g possible n'est pas abordée ici, mais l'idée est en général de prendre une fonction qui échantillonnera préférentiellement dans les régions dans lesquelles le produit fh est élevé et telle que $\int_\Omega \frac{f^2(t)h^2(t)}{g(t)}dt$ converge (afin que la variance de l'estimateur existe).

8.2 Méthodes MCMC

On considère toujours une distribution cible f, suivant laquelle on cherche à simuler. L'algorithme de Métropolis-Hastings repose sur le théorème central de la théorie des chaînes de Markov.

8.2.1 Chaînes de Markov continue

Une suite de variables aléatoires $(X_i)_{i\geq 0}$ définies sur un ensemble $\mathcal{X}$ est une **chaîne de Markov** si $X_{i+1}|X_0, \dots, X_i$ suit la même loi que $X_{i+1}|X_i$. La fonction K telle que

$$X_{i+1}|X_0, \dots, X_i \sim K(X_i, X_{i+1})$$

est appelé **noyau markovien**. Si f_i désigne la densité de X_i, on a alors

$$f_{i+1}(y) = \int_{\mathcal{X}} K(x,y) f_i(x) dx$$

La chaîne est **irréductible** si pour tout choix de la valeur initiale et tout ensemble A de mesure non nulle, la probabilité que la chaîne atteigne A est non nulle, ce qui est par exemple le cas si $K(x,y) > 0$, $\forall (x,y)$.

Théorème 9. *Si la chaîne est irréductible, il existe une unique loi stationnaire f, c'est-à-dire telle que*

$$f(y) = \int_{\mathcal{X}} K(x,y) f(x) dx$$

Cette loi est presque sûrement la loi limite de la chaîne de Markov.

En d'autres termes, en générant une chaîne de Markov très longue, on peut supposer qu'au bout d'un nombre conséquent de pas, quel que soit le point de départ, les x_i correspondent à des tirages suivant la loi f.

Ces tirages ne sont cependant pas indépendants, puisque x_{n+1} dépend clairement de x_n. Cependant, une autre propriété fondamentale des chaînes de Markov, appelée **ergodicité** permet d'utiliser le même estimateur que dans la méthode de Monte-Carlo

Théorème 10. *On considère une chaîne de Markov (X_i) de distribution limite f. Pout toute fonction intégrable h,*

$$\lim_{n\to+\infty} \frac{1}{n} \sum_{i=0}^{n-1} h(X_i) = \int_{\mathcal{X}} h(x) f(x) dx = \mathbb{E}_f(h(X))$$

8.2.2 Algorithme de Metropolis-Hastings

L'idée des algorithmes MCMC est de construire une chaîne de Markov dont la distribution cible f est la distribution limite. L'une des principale familles de tels algorithmes est celle des algorithmes de Metropolis-Hastings.

L'idée est similaire à celle de l'algorithme d'acceptation/rejet, à savoir partir d'une loi q dite de proposition et d'accepter ou non la valeur tirée suivant q. La différence est que là où les tirages étaient i.i.d. dans l'algorithme d'acceptation/rejet, le tirage et la probabilité d'acceptation vont dépendre ici de la valeur précédente, créant ainsi une chaîne de Markov.

Soit $q(y|x)$ une densité conditionnelle telle que :

1. $\frac{f(y)}{q(y|x)} \leq C$, $C > 0$
2. $q(.|x)$ a une dispersion suffisamment forte pour que la chaîne de Markov parcoure tout l'espace (c'est par exemple le cas si $supp(f) \subset supp(q(.|x), \forall x$.

On considère alors l'algorithme suivant :

Algorithme de Metropolis Hastings
Etant donné x_n,

1. Générer $y_n \sim q(y|x_n)$
2. Choisir

$$x_{n+1} = \begin{cases} y_n & \text{avec probabilité} \quad \rho(x_n, y_n) \\ x_n & \text{avec probabilité} \quad 1 - \rho(x_n, y_n) \end{cases}$$

où

$$\rho(x, y) = \min\{\frac{f(y)}{f(x)}\frac{q(x|y)}{q(y|x)}, 1\}$$

Théorème 11. *Les (x_n) forment une chaîne de Markov. Si q est tel que cette chaîne est irréductible, sa distribution limite est f.*

Démonstration. La construction de x_{n+1} dépend clairement uniquement de la valeur de x_n. Il s'agit donc bien d'une chaîne de Markov. Soit K le noyau de cette chaîne.

$$\begin{aligned} K(x, y) &= \mathbb{P}(X_{n+1} = y | X_n = x) \\ &= \mathbb{P}(X_{n+1} = y \cap \text{saut accepté} | X_n = x) + \mathbb{P}(X_{n+1} = y \cap \text{ saut refusé} | X_n = x) \\ &= \mathbb{P}(X_{n+1} = y | X_n = x \cup \text{saut accepté}) \mathbb{P}(\text{saut accepté} | X_n = x)) + \mathbb{P}(X_{n+1} = X_n | X_n = x) \delta_x(y) \\ &= \rho(x, y) q(y|x) + (\int (1 - \rho(x, z)) q(z|x) dz) \delta_x(y) \end{aligned}$$

Soit $x \neq y$. On peut supposer, quitte à échanger x et y, que $f(y)q(x|y) \leq f(x)q(y|x)$. Ceci implique que $\rho(x,y) = \frac{f(y)}{f(x)} \frac{q(x|y)}{q(y|x)}$ et $\rho(y,x) = 1$.

Alors

$$f(x)K(x,y) = f(x)\rho(x,y)q(y|x) = f(y)q(y|x) = f(y)\rho(y|x)q(y|x) = f(y)K(y,x)$$

Pour $y = x$, cette égalité est toujours vraie de façon évidente.

Par conséquent, on a toujours $f(x)K(x,y) = f(y)K(y,x)$. En intégrant des deux côtés en y, on en déduit que $f(x) = \int_y f(y)K(x,y)dy$. En d'autres termes, f est une distribtion invariante. En cas d'irréductibilité de la chaîne, on a donc convergence vers f de sa distribution. □

En pratique on élimine la partie initiale de la chaîne, pendant laquelle elle n'a pas encore convergé vers la distribution stationnaire (phase de *burn-in*). L'avantage de cette méthode est qu'elle est très flexible et s'applique même si la ditribution cible n'est connue qu'à une constante multiplicative près.

L'inconvénient principal est la lenteur de la convergence quand l'espace d'exploration est grand, et l'absence de critère certifiant la convergence. Il existe certains critères de qualité (cf (Robert et al., 2010)), mais on ne sait jamais s'il ne reste pas une partie de l'espace non exploré.

Bibliographie

Duprez, M. (2022). Incertitudes. communication personnelle.

Gardes, L. (2020). Théorie des valeurs extrêmes. https://irma.math.unistra.fr/~gardes/Poly_extreme.pdf.

Garnier, J. (2017). Gestion des incertitudes et analyse de risque. https://josselin-garnier.org/wp-content/uploads/2018/01/polyMAP568.pdf.

Guyader, A. (2012). Régression linéaire. https://perso.lpsm.paris/~aguyader/files/teaching/Regression.pdf.

McClarren, R. G., McClarren, P., and Penrose, R. (2018). *Uncertainty quantification and predictive computational science*. Springer.

Prieur, C. (2022). Global sensitivity analysis and dimension reduction. https://membres-ljk.imag.fr/Clementine.Prieur/teaching/CIMI/Prieur_GSAPoincare.pdf.

Robert, C. P., Casella, G., and Casella, G. (2010). *Introducing monte carlo methods with r*, volume 18. Springer.