Rapport de stage L3

Barbier Noé

Mai-juin 2025

Résumé

Dans le cadre d'un stage de six semaines au Laboratoire Jean Kuntzmann, encadré par Julien Chevallier. Cette introduction aux prédictions conformes s'appuie sur l'article "Distribution-Free Predictive Inference for Regression" [1], publié en 2017. L'objectif est de mieux comprendre ce domaine et, en particulier, de découvrir différents algorithmes afin de pouvoir les comparer. Le stage s'est articulé autour de deux grandes étapes : une première partie théorique, consacrée à l'appropriation des concepts fondamentaux et à l'étude détaillée des théorèmes et preuves de l'article ; puis une seconde partie appliquée, dédiée à l'implémentation en Python d'une partie des algorithmes étudiés.

1. https ://www.arxiv.org/pdf/1604.04173

Table des matières

1 Introduction **3**

2 Présentation des algorithmes **4**
2.1 Le naïf 5
2.2 Conformal prediction 6
2.3 Split conformal prediction 7
2.4 Jackknife 8

3 Présentation des hypothèses **8**
3.1 A0 (i.i.d data) 8
3.2 A1 (Independent and symetric noise) 8
3.3 A2 (Sampling stability) 8
3.4 A3 (Perturb-one sensitivity) 9
3.5 A4 (Consistency of base estimator) 9

4 Les oracles **9**
4.1 Le super oracle 10
4.2 Regular oracle 10
4.3 Résultats liés aux oracles 11
4.3.1 Les deux oracles 11
4.3.2 Conformal et regular oracle 11
4.3.3 Conformal et super oracle 12

5 Les configurations expérimentales **13**
5.1 Setting A (linear, classical) 13
5.2 Setting B (nonlinear, heavy-tailed) 13
5.3 Setting C (linear, heteroscedastic, heavy-tailed, correlated features) 14
5.4 Setting P5 14

6 Aide à la compréhension des preuves **16**
6.1 Théorème 2.4 16

7 Du côté du code **18**

1 Introduction

Motivation

Supposons que l'on cherche à modéliser la consommation électrique d'un bâtiment au cours du temps. On effectue l'hypothèse que cette quantité peut être approximée avec la connaissance de plusieurs facteurs : la température extérieure, le jour de la semaine, l'heure de la journée et le taux d'occupation du bâtiment.

On souhaite répondre à la question suivante : « Si demain à 10h, on prévoit une température de 15°C, en semaine, avec un taux d'occupation de 80 %, à quel point peut-on anticiper la consommation électrique, et avec quelle incertitude ? »

Pour cela, on construit un modèle en recueillant des observations répétées : à différents moments, on mesure ces variables explicatives ainsi que la consommation électrique correspondante. Ici, on a alors X_i est un vecteur de dimension 4 représentant les conditions de la i ème mesure, et Y_i est la consommation électrique de la i ème mesure.
L'objectif alors est de trouver un intervalle de prédiction $C(X)$ pour prédire Y dans les conditions X.

Formalisation mathématique

On considère $Z_1, \ldots, Z_n \sim P$ des variables aléatoires indépendantes et identiquement distribuées (i.i.d.), telles que pour tout i dans le segment $[\![1; n]\!]$, $Z_i := (X_i, Y_i)$ soit une variable aléatoire de $\mathbb{R}^d \times \mathbb{R}$.
Les X_i, appelées variables explicatives, sont les prédicteurs des Y_i, appelées variables expliquées. On définit alors la fonction de régression μ par $\mu(x) = \mathbb{E}[Y|X = x]$ pour tout $x \in \mathbb{R}^d$. On peut ensuite définir l'erreur, comme une variable aléatoire réelle ϵ par $\epsilon = Y - \mu(X)$. Dans la pratique, on peut l'interpréter comme un bruit : des valeurs différentes de la variable expliquée peuvent provenir d'observations identiques des variables explicatives.

L'objectif est alors de prédire Y_{n+1} quand X_{n+1} est donné, sans faire d'hypothèses sur μ et P. On choisit $\alpha \in]0; 1[$ la "marge d'erreur" et on cherche $C \subseteq \mathbb{R}^d \times \mathbb{R}$ qui dépend de $Z_1, \ldots, Z_n$ tel que

$$\mathbb{P}[Y_{n+1} \in C(X_{n+1})] \geq 1 - \alpha \tag{1}$$

Où $C(x) := \{y \in \mathbb{R} : (x, y) \in C\}$

En général, on a des résultats asymptotiques, mais ici, on veut trouver une construction de C valable lorsqu'on a un nombre fini de mesures. C'est

important en pratique d'avoir des résultats valides à n fini.

2 Présentation des algorithmes

On appelle *échantillon* de taille $n \in \mathbb{N}$ d'une variable aléatoire X, les variables aléatoires $(X_1, \ldots, X_n)$ telles que, pour tout $i \in [\![1; n]\!]$, X_i admet la même loi que X et que les X_i soit indépendantes. Si X modélise une grandeur physique, alors un échantillon correspond à une série de mesures de cette grandeur obtenues dans différentes conditions expérimentales ou à différents instants.

Définition (Échangeabilité)**.** *Soit $(U_1, .., U_n)$ des variables aléatoires à valeur dans un ensemble mesurable $(E, \mathcal{E})$, on dit que $(U_1, \ldots, U_n)$ sont* échangeables *si :*

$$\forall A \in \mathcal{E},\ \forall \sigma \in \mathfrak{S}_n,\ \mathbb{P}[(U_1, \ldots, U_n) \in A] = \mathbb{P}[(U_{\sigma(1)}, \ldots, U_{\sigma(n)}) \in A].$$

Dans le cadre des observations, elles sont échangeables si leur ordre n'a pas d'importance : les résultats auraient pu apparaître dans n'importe quel ordre sans changer la nature statistique du phénomène observé. Dans la pratique, c'est une hypothèse très faible.
On note aussi que i.i.d implique échangeable.

Définition (Statistique d'ordre)**.** *Soient $X_1, \ldots, X_n$ des variables aléatoires réelles. On appelle* statistiques d'ordre *les variables aléatoires $X_{(1)}, \ldots, X_{(n)}$ obtenues en réarrangeant les X_i par ordre croissant :*

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$$

Ainsi, pour tout $k \in [\![1; n]\!]$, $X_{(k)}$ représente la k-ième plus petite valeur parmi les X_i.

Proposition 1. *Soient $U_1, \ldots, U_{n+1}$ des variables aléatoires réelles échangeables. On note F leur fonction de répartition commune qu'on suppose continue (pas de cas d'égalité) et f leur densité. On pose $U_{(1)}, \ldots, U_{(n+1)}$ leur statistique d'ordre. Le rang de U_{n+1} parmi $U_1, \ldots, U_{n+1}$ suit la loi uniforme sur $[\![1; n+1]\!]$*

Démonstration. (uniquement pour le cas indépendant) : Soit R le rang de U_{n+1}, soit $k \in [\![1; n+1]\!]$. On peut poser

$$N = \sum_{i=1}^{n} \mathbb{1}_{U_i < U_{n+1}} \quad \text{tel que} \quad R = N + 1.$$

Par indépendance, on a $\mathbb{P}[U_i < x | U_{n+1} = x] = F(x)$. Donc, sous la condition $U_{n+1} = x$, on a $\mathbb{1}_{U_i<U_{n+1}} \sim \text{Ber}(F(x))$ et donc $N \sim \text{Bin}(n, F(x))$. Or, $\mathbb{P}[R = k | U_{n+1} = x] = \mathbb{P}[N = k-1 | U_{n+1} = x]$. Ainsi,

$$P[R = k | U_{n+1} = x] = \binom{n}{k-1} F(x)^{k-1} (1 - F(x))^{n-k+1}$$

On peut alors intégrer sur x :

$$\begin{aligned} \mathbb{P}[R = k] &= \int_{\mathbb{R}} \mathbb{P}[R = k | U_{n+1} = x] f(x)\, dx \\ &= \int_{\mathbb{R}} \binom{n}{k-1} F(x)^{k-1} (1 - F(x))^{n-k+1} f(x)\, dx. \end{aligned}$$

En effectuant le changement de variable $t = F(x)$ ($dt = f(x)\, dx$), on obtient $\mathbb{P}[R = k] = \int_0^1 \binom{n}{k-1} t^{k-1} (1-t)^{n-k+1}\, dt$. On reconnaît la fonction Bêta,

$$\begin{aligned} \mathbb{P}[R = k] &= \binom{n}{k-1} B(k, n-k+2) \\ &= \frac{n!}{(k-1)!(n-k+1)!} \frac{(k-1)!(n-k+1)!}{(n+1)!} = \frac{1}{n+1}. \end{aligned}$$

□

2.1 Le naïf

Proposition 2. *Soient $U_1, \ldots, U_{n+1}$ un échantillon i.i.d. (ou au moins échangeable) d'une v.a. réelle. On pose $U_{(1)}, \ldots, U_{(n)}$ la statistique d'ordre de $U_1, \ldots, U_n$.*

On définit le quantile $\hat{q}_{1-\alpha}$ basé sur les $(U_1, \ldots, U_n)$ comme

$$\hat{q}_{1-\alpha} = \begin{cases} U_{(\lceil (n+1)(1-\alpha) \rceil)} & \text{si } \lceil (n+1)(1-\alpha) \rceil \leq n \\ +\infty \quad \text{sinon} \end{cases}$$

Il vient alors :

$$\mathbb{P}[U_{n+1} \leq \hat{q}_{1-\alpha}] \geq 1 - \alpha \tag{2}$$

Démonstration. On procède par disjonction de cas :
Si $\lceil (n+1)(1-\alpha) \rceil \leq n$ alors $\hat{q}_{1-\alpha} = +\infty$ et $\mathbb{P}[U_{n+1} \leq +\infty] = 1 \geq 1-\alpha$
Sinon, par échangeabilité des $U_1, \ldots, U_{n+1}$, d'après la proposition 1, le rang de U_{n+1} parmi $U_1, \ldots, U_{n+1}$ suit une loi uniforme, on a alors :

$$\mathbb{P}[U_{n+1} \leq \hat{q}_{1-\alpha}] = \mathbb{P}[U_{n+1} \leq U_{(\lceil (n+1)(1-\alpha) \rceil)}] = \frac{\lceil (n+1)(1-\alpha) \rceil}{n+1} \geq 1-\alpha.$$

□

Ce résultat joue un rôle central dans le cadre des prédictions conformes, car il permet de majorer de U_{n+1} avec une probabilité fixée librement et en connaissant seulement $U_1, \ldots, U_n$.

Dans notre situation, on introduit $\hat{\mu}$ un estimateur de la fonction de régression. Ensuite, on peut définir les résidus quadratiques ajustés par $R_i := |Y_i - \hat{\mu}(X_i)|$. Enfin, on pose $\hat{F}_n$ la fonction de répartition des R_i. Elle est définie pour tout $x \in \mathbb{R}$ par $\hat{F}_n(x) = \mathbb{P}[R_i \leq x]$. $\hat{F}_n^{-1}(1-\alpha)$ représente alors le $(1-\alpha)$-quantile de $\hat{F}_n$. On a $\hat{q} \simeq \hat{\mu}(X_{n+1}) + \hat{F}_n^{-1}(1-\alpha)$

On peut alors construire un C_{naif} par

$$C_{\text{naif}}(X_{n+1}) = [\hat{\mu}(X_{n+1}) - \hat{F}_n^{-1}(1-\alpha),\ \hat{\mu}(X_{n+1}) + \hat{F}_n^{-1}(1-\alpha)] \qquad (3)$$

Le problème de C_{naf} est que pour avoir une couverture valide, il faut garantir l'échangeabilité sur les R_i. En effet, la construction de $\hat{\mu}$ dépend de $X_1, \ldots, X_n$ mais pas de $X_{(n+1)}$ donc les R_i sont échangeables pour $i \in [\![1; n]\!]$ mais en général, rien ne garantie l'echangeabilité avec $R_{(n+1)}$. L'objectif des algorithmes suivant est de contourner ce problème.

2.2 Conformal prediction

Une première manière de contourner le problème de l'échangeabilité est de considérer un esemble de test $\mathcal{Y}_{trial}$ et d'entrainer $\hat{\mu}_y$ sur l'ensemble $Z_1, \ldots, Z_n, (X_1, y)$, où $y \in \mathcal{Y}_{trial}$.

On peut alors calculer $\pi(y)$ la proportion des points dont l'erreur ajustée est plus petite que $|y - \hat{\mu}(X_{n+1})|$.

Dans leur article, les auteurs montrent que

$$\mathbb{P}[(n+1)\pi(Y_{n+1}) \leq \lceil (1-\alpha)(n+1) \rceil] \geq 1-\alpha$$

Donc, il est légitime de poser

$$C_{conf}(X_{n+1}) = \{y \in \mathbb{R} : (n+1)\pi(y) \leq \lceil (1-\alpha)(n+1) \rceil\} \qquad (4)$$

2.3 Split conformal prediction

L'idée est de découper les données observées en deux sous-ensembles de même taille, $\mathcal{I}_1, \mathcal{I}_2$. On entraîne $\hat{\mu}$ sur $\{Z_i : i \in \mathcal{I}_1\}$

On calcule les résidus ajustés R_i sur $\{Z_i : i \in \mathcal{I}_1\}$. Et on considère $R_{(q)}$, où $q = \lceil (n/2+1)(1-\alpha) \rceil$

On pose

$$C_{split}(X_{n+1}) = [\hat{\mu} - R_{(q)}, \hat{\mu} + R_{(q)}] \tag{5}$$

Et, les auteurs de l'article montre que :

$$\mathbb{P}[X_{n+1} \in C_{split}(X_{n+1})] \geq 1 - \alpha$$

La pondération des résidus absolus ajustés

On note qu'avec l'algorithme *split conformal*, la bande de prédiction possède une largeur constante. Or en pratique, la variance des données peut être non constante (comme avec Setting P5), et dans ce cas une bande à largeur variable serait plus pertinente. Une manière de remédier à ce problème est de pondérer les résidus absolus ajustés R_i. Pour ce faire, on entraine une fonction de régression $\hat{\rho} : \mathbb{R}^{\mathsf{d}} \to \mathbb{R}$ sur les $(X_i, R_i)_{i \in \mathcal{I}_2}$. Ensuite, on considère les

$$\tilde{R}_i = \frac{|Y_i - \hat{\mu}(X_i)|}{\hat{\rho}(X_i)}, \ i \in \mathcal{I}_2$$

Enfin, on pose $q = \lceil (n/2+1)(1-\alpha) \rceil$ et on renvoie l'intervalle ajusté :

$$C_{split}(x) = [\hat{\mu}(x) - \hat{\rho}(x)\tilde{R}_{(q)}, \hat{\mu} + \hat{\rho}(x)\tilde{R}_{(q)}]$$

En fait, les résidus absolus ajustés donnent l'information sur la variance des données. Pour s'en convaincre, on peut voir que si on remplace $\hat{\mu}$ par μ alors $R_i = |\mu(X) - Y| = \epsilon_i$. Or c'est justement le bruit ϵ entraine cette variance.

Le multi split conformal

Pour encore améliorer les performances de calcul, une idée serait de partager les données en plusieurs sous-ensembles, puis de prendre l'intersection des bandes trouvées.

On aurait $\mathcal{I}_1, \ldots, \mathcal{I}_m \subseteq [\![1; N]\!]$ puis, par une méthode analogue à celle précédemment décrite, en entrainant $\hat{\mu}_i$ sur $\mathcal{I}_i$ et en calculant les R_i sur $\mathcal{I}_i^c$,

on obtient $C_{split,i}$. On prend alors

$$C_{N,split}(X_{n+1}) = \bigcap_{i=1}^{N} C_{split,i}(X_{n+1})$$

Mais sous les hypothèses A0, A1, A2, on peut monter que l'intervalle ainsi obtenu sera asymptotiquement plus large que celui précédemment décrit.

2.4 Jackknife

L'idée est de laisser seulement une observation de côté et d'entraîner la fonction de régression sur le reste. Puis de répéter le processus avec chacune des données. On prend alors le quantile sur les résidus des données misent de côté. Le problème est que, mis à part sous des conditions très strictes, la bande ainsi obtenue n'a pas de bonnes propriétés de prédiction.

3 Présentation des hypothèses

3.1 A0 (i.i.d data)

Les données observées $Z_1, \ldots, Z_d$ sont i.i.d à valeur dans $\mathbb{R}^d \times \mathbb{R}$ issues d'une distribution P avec une fonction moyenne $\mu : x \mapsto \mathbb{E}[Y|X = x]$.

3.2 A1 (Independent and symetric noise)

Le bruit est symétrique et indépendant. C'est-à-dire que pour, $(X, Y) \sim P$ on a $\epsilon = Y - \mu(X)$ indépendant de X. De plus, la fonction densité de ϵ est symétrique et décroissante sur $[0, \infty[$.

3.3 A2 (Sampling stability)

Soit $\hat{\mu}_n$ la fonction de régression approchée entraînée sur n observations. Il existe (η_n) et (ρ_n) deux suites réelles convergeant vers 0 et une fonction $\tilde{\mu} : \mathbb{R}^d \to \mathbb{R}$ telles qu'à partir d'un certain rang n :

$$\mathbb{P}[\|\hat{\mu}_n - \tilde{\mu}\|_\infty \geq \eta_n] \leq \rho_n$$

Autrement dit à partir d'un certain rang, $\hat{\mu}_n$ est proche d'une fonction avec forte probabilité.

Il est à noter que sous cette hypothèse, la fonction de régression asymptotique $\tilde{\mu}$ n'est pas nécessairement proche de la fonction de régression effective.

3.4 A3 (Perturb-one sensitivity)

Pour $\mathcal{Y}$ un segment, en notant pour $y \in \mathcal{Y}$, $\hat{\mu}_{n,(X,y)}$ la fonction de régression entraînée sur n observation agrémentée du point (X, y). Il existe (η_n) et (ρ_n) deux suites réelles convergeant vers 0 telles que

$$\mathbb{P}[\sup_{y \in \mathcal{Y}} \|\hat{\mu}_n - \hat{\mu}_{n,(X,y)}\|_\infty \geq \eta_n] \leq \rho_n$$

Autrement dit, μ_n est peu sensible à une perturbation induite par l'ajout d'une nouvelle donnée d'entrainement à partir d'un certain rang.

3.5 A4 (Consistency of base estimator)

Il existe (η_n) et (ρ_n) deux suites réelles convergentes vers 0 telles qu'à partir d'un certain rang

$$\mathbb{P}[\mathbb{E}_X[(\hat{\mu}_n(X) - \mu(X))^2 | \hat{\mu}_n] \geq \eta_n] \leq \rho_n$$

Ici, la notation est un peu confusante car $\hat{\mu}_n$ n'est pas explicitement une variable aléatoire. En revanche comme elle dépend directement des données $X_1, \ldots, X_n$, $\mathbb{E}[\,\cdot\,|\hat{\mu}_n]$ revient à prendre $\mathbb{E}[\,\cdot\,|(X_1, \ldots, X_n)]$. Mais alors, il y a redondance, car $\mathbb{E}_X[\,\cdot\,]$ indique déjà qu'on considère l'espérance seulement sur la v.a. X

L'idée est d'assurer qu'avec forte probabilité, si la fonction de régression empirique $\hat{\mu}_n$ est fixée, alors sur une nouvelle donnée, elle n'est pas trop loin de la fonction de régression réelle μ.

4 Les oracles

L'intérêt des oracles est de comparer la bande déterminée par les algorithmes avec une bande "idéale".

4.1 Le super oracle

Le super oracle connaît la fonction de régression μ ainsi que la loi de ϵ Sous les hypothèses A0, A1, la bande de prédiction de cet oracle est : Avec q_α le quantile supérieur d'ordre α de la loi de $|\epsilon|$

$$C_s^*(x) = [\mu(x) - q_\alpha, \mu(x) + q_\alpha] \tag{6}$$

Je ne l'ai pas fait mais il est possible de montrer que cette bande vérifie ces trois propriétés.

i) $\mathbb{P}[Y \in C_s^*(x)|X = x] \geq 1 - \alpha$, la couverture conditionnelle est vérifiée

ii) C'est la plus courte bande qui vérifie la couverture conditionnelle

iii) En moyenne, c'est la plus courte bande qui vérifie la couverture marginale

Remarque. *Il ne faut pas confondre la couverture marginale et la couverture conditionnelle (qui est plus forte).*

La couverture marginale d'une bande de confiance $C(x)$ au niveau $1 - \alpha$ est définie par la propriété

$$\mathbb{P}\big(Y \in C(X)\big) \geq 1 - \alpha$$

où la probabilité est prise conjointement sur la distribution de (X, Y). Autrement dit, la bande couvre la vraie valeur Y avec au moins la probabilité $1 - \alpha$ en moyenne sur la loi de X.

En revanche, la couverture conditionnelle exige que pour chaque point x,

$$\mathbb{P}\big(Y \in C(x) \mid X = x\big) \geq 1 - \alpha,$$

c'est-à-dire que la bande couvre Y avec la probabilité requise, quelle que soit la valeur spécifique de x.

4.2 Regular oracle

L'oracle ordinaire connait la loi de $Y - \hat{\mu}_n(X)$ où $(X, Y) \sim P$ sont indépendantes des (X_i, Y_i) - les données utilisées pour entrainer $\hat{\mu}_n$.

On pose, pour $q_{n,\alpha}$ le quantile supérieur d'ordre α de la loi de $|Y - \hat{\mu}_n(X)|$

$$C_O^*(x) = [\hat{\mu}_n(x) - q_{n,\alpha}, \hat{\mu}_n(x) + q_{n,\alpha}]$$

La couverture de cet oracle n'est que marginale.

4.3 Résultats liés aux oracles

4.3.1 Les deux oracles

Théorème 1. Les notations :

- $\Delta_n : x \mapsto \hat{\mu}_n(x) - \mu(x)$ *l'écart relatif entre la fonction de régression empirique et celle théorique*
- $F : \mathbb{R} \to \mathbb{R}$ *la fonction de répartition de* $|\epsilon|$
- $f : \mathbb{R} \to \mathbb{R}$ *la densité de* $|\epsilon|$
- $F_n : \mathbb{R} \to \mathbb{R}$ *la fonction de répartition de* $|Y - \hat{\mu}_n(X)|$
- $f_n : \mathbb{R} \to \mathbb{R}$ *la densité de* $|Y - \hat{\mu}_n(X)|$

Les hypothèses :

- *A0, A1*
- f *admet une dérivée continue bornée par* $M > 0$

Les résultats :

$$\sup_{t>0} |F_n(t) - F(t)| \leq (M/2)\mathbb{E}[\Delta_n^2(X)] \tag{7}$$

Si on suppose de plus qu'il existe $r > 0$ *et* $\eta > (M/2r)\mathbb{E}[\Delta_n^2(X)]$ *tels que* f *soit minorée par* r *sur* $]q_\alpha - \eta, q_\alpha + \eta[$ *alors*

$$|q_{n,\alpha} - q_\alpha| \leq (M/2)\mathbb{E}[\Delta_n^2(X)] \tag{8}$$

Étant donné que la largeur des bandes est $2q_\alpha$ pour le *super oracle* et $2q_{n,\alpha}$ pour le *regular oracle*, on en déduit que sous ces hypothèses les deux bandes ont des tailles similaires.

4.3.2 Conformal et regular oracle

Définition. *Soient* $(u_n)_{n\geq 1}$ *et* $(v_n)_{n\geq 1}$ *deux suites de variables aléatoires définies sur un même espace probabilisable, avec* $v_n > 0$. *On dit que* $u_n = O_{\mathbb{P}}(v_n)$ *si :*

$$\forall \varepsilon > 0,\ \exists M > 0 \text{ tel que } \limsup_{n\to\infty} \mathbb{P}\left(\left|\frac{u_n}{v_n}\right| > M\right) < \varepsilon.$$

Cas split conformal

Théorème 2. Les notations :

- $\alpha \in]0, 1[$
- $C_{n,split}$ *l'intervalle de l'algorithme* split conformal prediction
- $\nu_{n,split}$ *la longueur de* $C_{n,split}$

— $\tilde{\mu}$, η_n *et* ρ_n *de A2*
— $\tilde{f}$ *la densité de* $\tilde{\mu}(X)$

Les hypothèses :

— *A0, A1 et A2*
— $\tilde{f}$ *est strictement minorée par* 0 *dans un voisinage ouvert de son quantile supérieur d'ordre* α

Le résultat :

$$\nu_{n,split} - 2q_{n,\alpha} = O_{\mathbb{P}}(\rho_n + \eta_n + n^{-1/2}) \tag{9}$$

Cas full conformal

Théorème 3. Les notations :

— $\alpha \in]0,1[$
— $C_{n,conf}$ *l'intervalle de l'algorithme* conformal prediction
— $\nu_{n,conf}$ *la longueur de* $C_{n,conf}$
— $\mathcal{Y}$, $\tilde{\mu}$, η_n *et* ρ_n *de A3*

Les hypothèses

— *Les mêmes que celles du cas* split conformal
— Y *est à support dans* $\mathcal{Y}$ *tel qu'on ait A3*

Le résultat :

$$\nu_{n,conf} - 2q_{n,\alpha} = O_{\mathbb{P}}(\rho_n + \eta_n + n^{-1/2}) \tag{10}$$

4.3.3 Conformal et super oracle

En considérants les résultats précédents de cette partie, sous les bonnes hypothèses, il vient

$$\mathbb{E}[\Delta_n^2(X)] = o(1) \implies \nu_{n,split} - 2q_\alpha = o_{\mathbb{P}}(1) \text{ et } \nu_{n,conf} - 2q_\alpha = o_{\mathbb{P}}(1)$$

On introduit le concept de *Couverture conditionnelle asymptotique* : Pour C_n une suite de bandes de prédiction (possiblement aléatoires), on dit que C_n admet une couverture conditionnelle asymptotique au niveau $(1-\alpha)$ s'il existe une suite d'ensembles $\Lambda_n \subseteq \mathbb{R}^d$ (possiblement aléatoires) tels que : $\mathbb{P}[X \in \Lambda_n | \Lambda_n] = 1 - o_{\mathbb{P}}(1)$ et

$$\left| \inf_{x \in \Lambda_n} \Big(\mathbb{P}[Y \in C_n(x) | X = x] \Big) - (1-\alpha) \right| = o_{\mathbb{P}}(1)$$

Théorème 4. Les notations :

— L *désigne la mesure de Lebesgue sur* $\mathbb{R}$

— Δ *désigne la différence symétrique*

Les hypothèses :

— *A0*, *A1*, *A4*

— $|Y - \mu(X)|$ *admet une densité strictement minorée par* 0 *dans un voisinage ouvert de son quantile supérieur d'ordre* α

Le résultat :

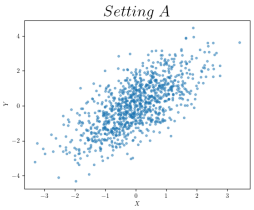
$$L(C_{n,split}(X)\Delta C^*_s(X)) = o_{\mathbb{P}}(1) \tag{11}$$

Donc $C_{n,conf}$ *a une couverture conditionnelle asymptotique au niveau* $(1-\alpha)$

Sous les mêmes hypothèses que le résultat précédent, en supposant de plus A3, $C_{n,conf}$ *admet aussi une couverture conditionnelle asymptotique au niveau* $(1-\alpha)$

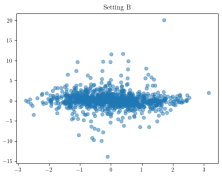
5 Les configurations expérimentales

5.1 Setting A (linear, classical)



Le modèle A est linéaire, c'est-à-dire que μ est linéaire. De plus, les variables explicatives $X_i(1), \ldots, X_i(d)$ sont i.i.d., suivant la loi $\mathcal{N}(0,1)$. L'erreur ϵ_i suit aussi la loi $\mathcal{N}(0,1)$ et est indépendante des variables explicatives.

5.2 Setting B (nonlinear, heavy-tailed)



Pour cette deuxième configuration, il va falloir introduire quelque concepts.

Premièrement, un *B-spline* est une fonction, polynomiale par morceaux qui peut être arbitrairement régulière (en fonction du degré des polynômes choisis). Ils ont notamment utilisé pour lisser une fonction.

Ensuite, la loi de Student à deux degrés de liberté, notée $t(2)$, est une loi symétrique, centrée. Elle a une queue qui est plus épaisse que celle de la loi normale. Elle admet une espérance, mais pas de variance. Dans un contexte pratique, l'avantage de la queue lourde est qu'elle prend en compte l'éventualité de valeurs extrêmes.

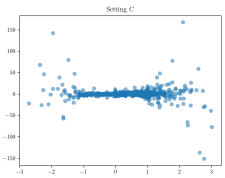
La fonction du modèle B est additive, c'est-à-dire qu'on peut écrire :

$$\mu(x) = f_1(x_1) + \ldots + f_d(x_d)$$

Et ici, chaque f_i est un B-spline.
De plus, le bruit ϵ_i suit une loi de Student $t(2)$ est indépendant des X_i

5.3 Setting C (linear, heteroscedastic, heavy-tailed, correlated features)

On introduit la loi normale asymétrique, notée $SN(a, \sigma^2)$. Il s'agit d'une loi d'espérance a et de variance σ^2, elle est proche de la loi normale $N(a, \sigma^2)$ mais est asymétrique.



Définition. *Dans un modèle de régression, on parle d'hétéroscédasticité lorsque la variance conditionnelle de l'erreur ϵ dépend de la variable explicative X, c'est-à-dire*

$$\mathrm{Var}(\epsilon \mid X = x) = \sigma^2(x) \quad \textit{avec } \sigma^2(x) \textit{ non constante en } x.$$

C'est-à-dire que la dispersion des erreurs varie selon les valeurs de X.

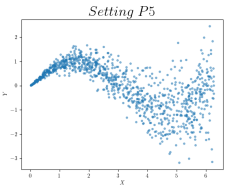
Ici, μ est linéaire en x.

Pour déterminer les X_i, on tire, dans un premier temps, de manière équiprobable et indépendante, pour chaque $X_i(k)$ ($k \in [\![1; d]\!]$ une distribution parmi la loi normale $N(0, 1)$, la loi normale asymétrique $SN(0, 1)$ et la loi de Brenoulli $B(0.5)$. Ensuite, on introduit de la corrélation en définissant séquentiellement les $X_i(k)$ comme combinaison convexe des (au plus) va $X_i(k), X_i(k-1), X_i(k-2), X_i(k-3)$. C'est-à-dire une combinaison linéaire où la somme des coefficients vaut 1.

Pour le bruit, on pose ϵ_i suivant une loi $t(2)$ puis on le multiplie par un facteur $1 + 2|\mu(X_i)|^3/\mathbb{E}[|\mu(X_i)|^3]$.
Cela permet le rendre dépendant des X_i et d'avoir de l'hétéroscédasticité (même si la loi de Student n'a pas de variance finie).

5.4 Setting P5

Dans la partie 5.2 de l'article, les auteurs utilisent une configuration où $\forall i \in [\![1; n]\!]$ X_i suit une loi uniforme sur $[0, 2\pi]$, le bruit ϵ_i suit la loi $\mathcal{N}(0, 1)$ et $Y_i = \sin(X_i) + \frac{\pi |X_i|}{20} \epsilon_i$.



L'intérêt de cette configuration est d'avoir une variance non constante. Cela permet d'illustrer le fait que l'algorithme de prédiction conforme donne des bandes de largeur constante

6 Aide à la compréhension des preuves

6.1 Théorème 2.4

La minoration de la longueur de $C^{(N)}_{split}(X)$

Proposition. *Soit $(I_i)_{i\in[\![1;n]\!]}$ une famille de segments de $\mathbb{R}$, où pour tout $i \in [\![1;n]\!]$, $I_i = [a_i, b_i]$. On définit : $I = \bigcap_{i=1}^{n} I_i$*
Si I est non vide, alors :

$$\lambda(I) \geq \min_{i\in[\![1;n]\!]}(\lambda_i) - 2\max_{i\in[\![1;n]\!]}|c - c_i|$$

où :
- *$\lambda(I)$ désigne la longueur du segment I*
- *λ_i désigne la longueur du segment I_i*
- *$c_i = \frac{a_i+b_i}{2}$ désigne le centre du segment I_i*
- *$c = \frac{\min a_i + \max b_i}{2}$ désigne le centre du segment I*

Démonstration. On note pour simplifier :

$$m = \min_{i\in[\![1;n]\!]} a_i \quad M = \max_{i\in[\![1;n]\!]} b_i \quad m_0 = \min_{i\in[\![1;n]\!]} \lambda_i, \quad \delta = \max_{i\in[\![1;n]\!]} |c - c_i|$$

et donc par définition $c = \frac{m+M}{2}$
On suppose que $I = \bigcap_{i=1}^{n} I_i$ est non vide donc, on peut écrire

$$I = \bigcap_{i=1}^{n}[a_i, b_i] := [A, B] \quad \text{où} \quad A = \max a_i \text{ et } B = \min b_i$$

avec $A \leq B$
On commence par majorer A :
Pour chaque i, on a $a_i = c_i - \frac{\lambda_i}{2}$ et, comme $|c_i - c| \leq \delta$, on en déduit

$$c_i \leq c + |c_i - c| \leq c + \delta$$

D'où pour tout i :

$$a_i = c_i - \frac{\lambda_i}{2} \leq (c + \delta) - \frac{\lambda_i}{2} \leq (c + \delta) - \frac{m_0}{2}$$

En passant au maximum sur i, on obtient :

$$A = \max_{1\leq i\leq n} a_i \leq (c + \delta) - \frac{m_0}{2}$$

Ensuite on minore B, par un argument similaire

Pour chaque i, on a $b_i = c_i + \frac{\lambda_i}{2}$ et, comme $|c_i - c| \leq \delta$, on en déduit

$$c_i \geq c - |c_i - c| \geq c - \delta$$

Donc pour tout i :

$$b_i = c_i + \frac{\lambda_i}{2} \geq (c - \delta) + \frac{\lambda_i}{2} \geq (c - \delta) + \frac{m_0}{2}$$

En passant au minimum sur i, on obtient :

$$B = \min_{1 \leq i \leq n} b_i \geq (c - \delta) + \frac{m_0}{2}$$

Comme $\lambda(I) = B - A$, on conclut que $\lambda(I) \geq m_0 - 2\delta$

□

La majoration de la longueur de $C_{split,1}(X)$

Si on suppose que $\|\hat{\mu}_1 - \tilde{\mu}\|_\infty < \eta_n$, les résidus construits autour de $\hat{\mu}_1$ peuvent être approximés par ceux construits autour de $\tilde{\mu}$, à une erreur près de l'ordre de η_n.

Or, par définition, la largeur de $C_{\text{split},1}(X)$ est $2\hat{F}^{-1}_{n,1}(1-\alpha)$

De plus, par hypothèse, on a pour tous $i \in [\![1; n]\!]$, on a

$$|Y_i - \hat{\mu}_1(X_i)| \leq |Y_i - \tilde{\mu}(X_i)| + \eta_n,$$

ce qui implique, par croissance de la fonction de répartition empirique, que :

$$\hat{F}^{-1}_{n,1}(1-\alpha) \leq \tilde{F}^{-1}_{n,1}(1-\alpha) + \eta_n.$$

On en déduit que la largeur de $C_{\text{split},1}(X)$ est majorée par :

$$2\hat{F}^{-1}_{n,1}(1-\alpha) \leq 2\tilde{F}^{-1}_{n,1}(1-\alpha) + 2\eta_n.$$

L'inégalité de Dvoretzky-Kiefer-Wolfowitz (DKW)

Théorème. *Soit $n \in \mathbb{N}^*$, $X_1, \ldots, X_n$ des var i.i.d..*
On définit la fonction empirique sur $\mathbb{R}$ par

$$F_n : t \mapsto \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq t\}}$$

Il vient, $\forall \varepsilon > 0$

$$\mathbb{P}\left(\|F_n(t) - F(t)\|_\infty > \varepsilon\right) \leq 2e^{-2n\varepsilon^2}$$

7 Du côté du code

Pour illustrer empiriquement les résultats décrits dans l'article, les auteurs effectuent une mise en pratique. Mon programme s'articule comme suit :

ExpSetups.py

- *Setting_A* : comme décrit dans la partie 5
- *Setting_B* : idem
- *Setting_C* : idem
- *Setting_P5* : idem
- *Setting_Trivial* : pour effectuer des tests, c'est un modèle très simple où X est déterminé avec *np.linspace*, ϵ suit la loi $\mathcal{N}(0,1)$ et $Y = 3 * X + \epsilon$

RegAlg.py

- *regalg_lasso* : une méthode de régression non linéaire
- *regalg_elasticnet* : idem
- *regalg_spline* : une méthode de régression non linéaire

AlgoPrediction.py

- *full_conformal_prediction* : l'algorithme décrit dans la partie 2. Par défaut la fonction prend pour $\mathcal{Y}_{trial}$, 25 point répartis uniformément entre le plus petit Y_i et le plus grand Y_i, et test pour 25 point X_{new} répartis uniformément entre le plus petit X_i et le plus grand X_i. La fonction renvoie tous les points qui vérifient la propriété demandée.
- *conf_band_full* : transforme la sortie de l'algorithme full_conformal_prediction pour obtenir seulement les extrémités du segment.
- *split_conformal_prediction* : l'algorithme décrit dans la partie 2, il renvoie une fonction C_split, qui à x associe l'intervalle de prédiction conforme.
- *conf_band_split* : transforme la sortie de l'algorithme split_conformal_prediction pour obtenir seulement les extrémités du segment.
- *split_conformal_prediction_LW* : comme pour split_conformal_prediction mais avec la pondération décrite dans la partie 2
- *conf_band_split_LW* : comme pour split_conformal_prediction mais pour split_conformal_prediction_LW

Control.py

- *coverage_test* : calcule la couverture marginale d'une fonction de prédiction. Ne fonctionne qu'avec les algorithmes *split_conformal_prediction* et *split_conformal_prediction_LW*

Enfin, **Main.py** et **Affichage.py** n'ont pas d'intérêt théorique.

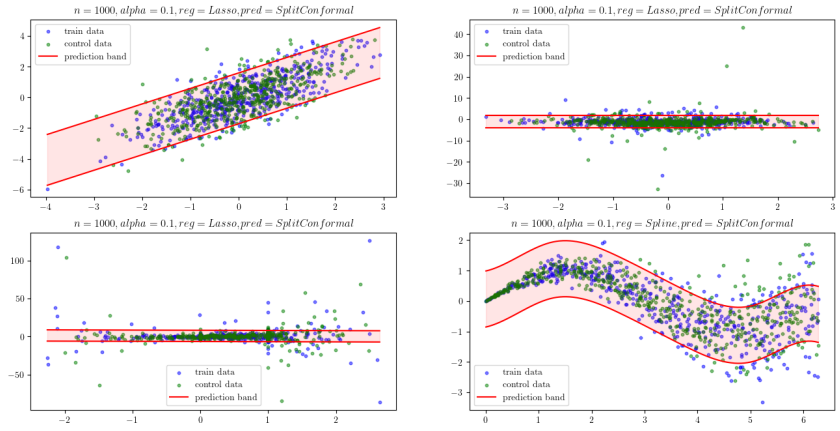


FIGURE 1 – Split Conformal Prediction

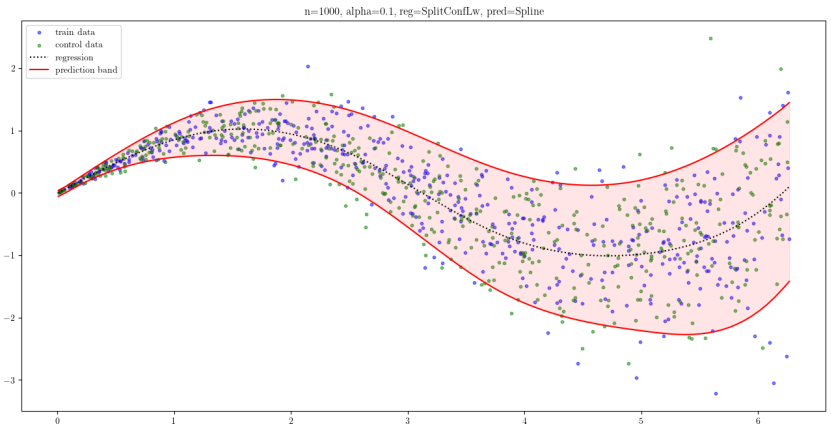


FIGURE 2 – Split Conformal Prediction Locally-Weighted

Pour obtenir les résultats suivants, j'ai fait tourner ce programme avec X de dimension 1 (ça facilite l'affichage de graphiques) :

- **Figure 1** : Elle concerne l'algorithme *Split Conformal*. J'ai généré les données suivant respectivement les quatre paramètres : *Setting A*, *Setting B*, *Setting C* et *Setting P5*. Pour les trois premiers, j'ai utilisé une régression lasso (linéaire) et pour la dernière la régression spline (non linéaire). Ensuite, j'ai affiché les bandes de confiances en rouge. Les données d'entraînement sont en bleu et les données de contrôle en vert.
- **Figure 2** : Comme pour le quatrième graphique de la **Figure 1**, mais avec l'algorithme *Split Conformal Prediction Locally-Weighted* décrit dans la partie 2. On voit bien que Contrairement à la **Figure 1**, ici, la largeur de la bande est non constante.

Enfin, en effectuant 100 répétitions de l'algorithme *split conformal*, avec la régression *spline*, j'ai pu obtenir ces résultats de couvertures. Ils sont tous du bon ordre de grandeur. Je n'ai pas pu faire de même avec le *full conformal* car il est trop long à process.

	Setting A	Setting B	Setting C	Setting Partie 5
Coverage	89.89	89.90	90.02	89.93