Loi jointe et concentration des p-valeurs conformelles

Ulysse Gazin [1] & Gilles Blanchard [2] & Étienne Roquain[3]

[1] *Université Paris Cité and Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation. Email: ugazin@lpsm.paris*
[2] *Université Paris Saclay, Institut Mathématique d'Orsay. Email: gilles.blanchard@universite-paris-saclay.fr*
[3] *Sorbonne Université and Université Paris Cité, CNRS, Laboratoire de Probabilités, Statistique et Modélisation. Email: etienne.roquain@upmc.fr*

Résumé. L'inférence conformelle permet de construire des ensembles de prédiction en régression ou en classification et de faire de la détection de nouveautés via des méthodes issues de l'apprentissage. Cette approche permet de quantifier l'incertitude des procédures d'apprentissage. Les p-valeurs conformelles sont les outils de base de l'inférence conformelle, d'où l'importance d'une étude théorique de ces objets. Prendre m décisions, ou bien construire m régions de prédiction ou de confiance, demande de comprendre la loi des p-valeurs conformelles sous l'hypothèse classique d'échangeabilité. Même si les lois marginales sont connues, la loi jointe reste inconnue dans la littérature et on propose ici son étude. On présente la loi jointe des p-valeurs sous différents aspects: séquentiel avec les urnes de Pólya, bayésien avec une variable latente de Dirichlet, ainsi qu'explicite via une formule combinatoire. In fine, on montre une inégalité de concentration de type DKW pour la fonction de répartition empirique des p-valeurs conformelles. Ce résumé reprend l'article "Transductive conformal inference with adaptive scores" de U. Gazin, G. Blanchard et E. Roquain en se focalisant sur les propriétés théoriques des p-valeurs conformelles.

Mots-clés. Inégalité de concentration, p-valeur, Urne de Pólya, Inférence conformelle.

Abstract. Conformal inference have recently emerged as one of the fundamental tools in various domains of statistics by allowing to construct prediction sets for regression or classification and for machine learning novelty detection. This approach allows us to quantify the uncertainty of any machine learning procedure. Conformal p-values are the main tool of conformal inference and so a theoretical study is fundamental. Providing m multiple decision (or simultaneous inference) in these contexts strongly relies on the joint distribution of the conformal p-values under an exchangeability assumption. However, while the marginal distribution is well known, only little is known about the joint distribution in the litterature and we propose to fill the gap. We present the joint distribution of conformal p-values under various aspects: a sequential one with a Pólya urn model, a Bayesian one with a latent Dirichlet random variable, and an explicit one with a combinatory formula. Finally, we propose a DKW-type concentration inequality for the empirical cumulative distribution function of m conformal p-values. This report is a summary of the article "Transductive conformal inference with adaptive scores" by U. Gazin, G. Blanchard and E. Roquain with a focus on the theoretical properties of conformal p-values.

Keywords. Concentration inequality, p-value, Pólya urn model, Conformal inference.

1 Introduction

Les inégalités de concentration sont fondamentales en théorie des tests. Que ce soit pour la création de tests non asymptotiques lorsque l'on ne connaît pas explicitement la loi sous l'hypothèse nulle ou ses quantiles (pensons à l'inégalité DKW (Massart, 1990) pour le test non-asymptotique de Kolmogorov-Smirnov) ou pour contrôler un taux d'erreur en test multiple (Meah et al., 2023). Ici, nous cherchons à obtenir une propriété de concentration pour des variables aléatoires particulières: les p-valeurs conformelles, d'abord introduites par Saunders et al. (1999) pour l'inférence conformelle (conformal inference). Nous nous intéressons donc à l'étude de la fonction de répartition empirique de m p-valeurs conformelles, que nous voyons ici comme un taux d'erreur, afin d'en déduire une inégalité de concentration de type DKW. Ce résumé est une relecture de l'article (Gazin et al., 2023) des mêmes auteurs; mais pris sous l'angle de l'étude théorique des p-valeurs conformelles. Nous renvoyons donc vers l'article en question pour voir plus en détails les preuves, les liens de ces p-valeurs avec les applications en prédiction conformelle et en détection de nouveautés ainsi que différents exemples d'applications en lien avec l'apprentissage, comme par exemple le transfer learning.

On pose, pour tout $(r < n) \in \mathbb{N}^2$, $[\![r, n]\!] := \{r, r+1, \cdots, n\}$, et $[\![n]\!] := [\![1, n]\!] = \{1, 2, \cdots, n\}$.

1.1 Les p-valeurs conformelles

On se donne deux familles de variables aléatoires réelles $\mathcal{D}_{\text{cal}} = \{S_1, \cdots, S_n\}$ et $\mathcal{D}_{\text{test}} = \{S_{n+1}, \cdots, S_{n+m}\}$ que nous appellerons scores et on définit comme suit nos p-valeurs conformelles.

Définition 1.1 (p-valeurs conformelles)**.** Soit deux familles de variables aléatoires réelles $\mathcal{D}_{\text{cal}} = \{S_1, \cdots, S_n\}$ et $\mathcal{D}_{\text{test}} = \{S_{n+1}, \cdots, S_{n+m}\}$. On définit, pour tout $i \in [\![m]\!]$, la i-ième p-valeur conformelle p_i:

$$p_i = \frac{1}{n+1}\left(1 + \sum_{k \in [\![n]\!]} \mathbb{1}\{S_k \geq S_{n+i}\}\right). \tag{1}$$

On définit également la fonction de répartition empirique de ces p-valeurs conformelles:

$$\widehat{F}_m : t \in \mathbb{R} \mapsto \frac{1}{m} \sum_{i \in [\![m]\!]} \mathbb{1}\{p_i \leq t\} \in [0, 1]. \tag{2}$$

Si les variables aléatoires $(S_i)_{i \in [\![n+m]\!]}$ sont iid de loi P_0 de fonction de survie caglad $\overline{F}_- = \mathbb{P}(S_1 \geq \cdot)$, alors p_i est proche de l'estimateur plug-in de $\overline{F}_-(S_{n+i})$ qui est une p-valeur (au sens de Lehmann and Romano (2005)) pour un test d'adéquation à la loi P_0 :

$$\begin{aligned} &(\mathrm{H}_0) : \text{" La loi de } S_{n+i} \text{ est } P_0 \text{ "}, \\ \text{contre } &(\mathrm{H}_1) : \text{" La loi de } S_{n+i} \text{ n'est pas } P_0 \text{ "}. \end{aligned}$$

En ce sens, les p-valeurs conformelles peuvent être vue comme des p-valeurs empiriques, et sont d'ailleurs utilisées par Bates et al. (2023) et Marandon et al. (2022) pour faire de la détection de nouveautés en appliquant une procédure de test multiples.

On introduit les deux hypothèses nécessaires pour l'étude de ces p-valeurs.

Hypothèse 1. *Les scores* $(S_1, \cdots, S_n, S_{n+1}, \cdots, S_{n+m})$ *sont échangeables.*

Cette hypothèse, un peu plus faible que i.i.d., provient de la littérature classique en inférence conformelle (on pourra lire Angelopoulos and Bates (2021) pour une introduction), mais se justifie par l'égalité suivante. En effet, $(n+1)p_i = \left|\{k \in [\![n]\!] \cup \{n+i\}, S_k > S_{n+i}\}\right|$ ce qui correspond, dans le cas où les $(S_j)_j$ sont p.s. deux à deux disjoints au rang de S_{n+i} dans l'ensemble $\{S_1, \cdots, S_n, S_{n+i}\} = \left\{S_{(1)} > S_{(2)} > \cdots > S_{(n)} > S_{(n+1)}\right\}$ des scores ordonnés dans l'ordre décroissant. La loi des statistiques de rang étant libre de la loi de $(S_j)_j$ et connue dès que les scores sont échangeables et distincts p.s., l'hypothèse que les scores soient indépendants est superflue. L'hypothèse suivante complète la première afin d'utiliser cette loi des statistiques de rang.

Hypothèse 2. *Les scores* $(S_1, \cdots, S_n, S_{n+1}, \cdots, S_{n+m})$ *sont deux à deux distincts presque sûrement.*

Cette hypothèse est faible dans le sens où si les scores ne la vérifient pas il est toujours possible de les perturber en rajoutant un "petit" bruit indépendant pour avoir de nouveaux scores proches des premiers; mais vérifiant désormais l'hypothèse 2.

1.2 Lois marginales

Tout d'abord, les p-valeurs conformelles sont bien des p-valeurs car elles sont sur-uniformes.

Théorème 1.2 (Romano and Wolf (2005))**.** *Supposons que l'hypothèse 1 soit vraie. Alors, pour tout* $i \in [\![m]\!]$*,* p_i *est sur-uniforme, i.e.*

$$\forall \alpha \in [0,1],\ \mathbb{P}(p_i \leq \alpha) \leq \alpha.$$

De plus, on connaît la loi marginale si les hypothèses 1 et 2 sont vérifiées.

Théorème 1.3. *Supposons que les hypothèses 1 et 2 soient vérifiées. Alors, pour tout* $i \in [\![m]\!]$*,* p_i *suit la loi uniforme sur* $[\![n+1]\!]/(n+1)$*.*

On peut remarquer, que du théorème 1.2, on peut construire un test non-asymptotique d'échangeabilité de $(S_1, \cdots, S_n, S_{n+1})$ avec la p-valeur p_1 en rejetant l'hypothèse nulle si $p_1 \leq \alpha$. En faisant cela pour tout $i \in [\![m]\!]$, $\widehat{F}_m(\alpha)$ correspond au nombre d'hypothèses nulles rejetées au niveau α divisé par le nombre m de tests. De ces garanties marginales, on peut déjà obtenir un premier résultat sur la fonction de répartition empirique. On introduit d'abord, pour tout $n \in \mathbb{N}$, I_n la fonction de répartition de la loi uniforme sur $[\![n+1]\!]/(n+1)$. On a, pour tout $t \in \mathbb{R}$, $I_n(t) = \mathbb{1}\{t \geq 1\} + \mathbb{1}\{t \in]0,1[\} \frac{\lfloor (n+1)t \rfloor}{n+1}$.

Corollaire 1.4. *Si les hypothèses 1 et 2 sont vérifiées, alors pour tout $t \in \mathbb{R}$, $\mathbb{E}\Big(\widehat{F}_m(t)\Big) = I_n(t)$.*

Cependant, on ne peut en dire plus sur la loi jointe et sur $\widehat{F}_m$ sans étudier plus particulièrement la loi jointe des $(p_i)_{i\in[\![m]\!]}$. En effet, en regardant la définition des p-valeurs conformelles, on remarque dans (1) que les scores de calibration $S_1, \cdots, S_n$ sont présents dans tous les $(p_i)_{i\in[\![m]\!]}$, et de là on peut supposer que les p-valeurs conformelles ne sont pas indépendantes. Cette dépendance est un problème puisque l'on ne peut désormais plus utiliser les théorèmes usuels sur les variables aléatoires i.i.d. (Boucheron et al., 2013; Shorack and Wellner, 1986).

2 Loi jointe des p-valeurs et une inégalité de concentration de type DKW

2.1 $P_{n,m}$: la loi jointe des p-valeurs conformelles

Comme expliqué ci dessus, les p-valeurs conformelles données par (1) ne sont pas indépendantes; mais on remarque qu'elle le sont conditionnellement à $\mathcal{D}_{\text{cal}}$. Ce résultat, classique dans la littérature (on le trouve par exemple dans la preuve du théorème 6 de Bates et al. (2023)), nous indique que nous manipulons des p-valeurs "quasiment" indépendantes, et qui le deviennent lorsque l'on conditionne par les bonnes variables aléatoires. On définit ici une loi $P_{n,m}$ particulière vérifiant de telles propriétés.

Définition 2.1 ($P_{n,m}$)**.** Soit $n \geq 1$ et $m \geq 1$ deux entiers. Soit $U = (U_i)_{i\in[\![n]\!]}$ un échantillon de variables aléatoires i.i.d. de loi uniforme sur $[0, 1]$. On définit la distribution discrète P^U sur l'ensemble $\left\{\frac{k}{n+1}, k \in [\![n+1]\!]\right\}$ comme suit:

$$\forall k \in [\![n+1]\!],\ P^U(\{k/(n+1)\}) = U_{(k)} - U_{(k-1)},$$

où $0 =: U_{(0)} < U_{(1)} < \cdots < U_{(n)} < U_{(n+1)} := 1$ sont les valeurs ordonnées de $U = (U_1, \cdots, U_n)$ dans l'ordre croissant. On définit désormais $(q_1, \cdots, q_m)$ m variables aléatoires qui conditionnellement à U sont i.i.d. de loi P^U. On note $P_{n,m}$ la loi inconditionnelle de $(q_1, \cdots, q_m)$. En résumé, $P_{n,m}$ est la loi sur $[0, 1]^m$ définie comme suit:

$$P_{n,m} = \mathcal{D}\big(q_i, i \in [\![m]\!]\big), \text{ où} \tag{3}$$

$$\begin{cases} \quad (q_1, \dots, q_m | U) & \overset{\text{i.i.d.}}{\sim} P^U; \\ \text{et } U = (U_1, \dots, U_n) & \overset{\text{i.i.d.}}{\sim} \mathcal{U}(0, 1). \end{cases} \tag{4}$$

La mesure aléatoire P^U est un processus de Dirichlet discret, dans le sens où le vecteur $\left(P^U\left(\frac{k}{n+1}\right)\right)_{k\in[\![n+1]\!]}$ suit une loi de Dirichlet de paramètre $(1, \cdots, 1) \in \mathbb{R}^{n+1}$.

Dans le cas où les scores sont i.i.d. et distincts p.s., il est montré dans la littérature que les p-valeurs conformelles ont pour loi $P_{n,m}$: les $(p_1, \cdots, p_m)$ sont i.i.d. conditionnellement à $\mathcal{D}_{\text{cal}}$ avec pour loi marginale $P^{U(\mathcal{D}_{\text{cal}})}$ où $U(\mathcal{D}_{\text{cal}}) = (1 - F(S_i))_{i \in [\![n]\!]}$ avec F la fonction de répartition de S_1. De là, en remarquant que les $(p_i)_{i \in [\![m]\!]}$ se définissent à partir des statistiques de rang seules, on peut en déduire la loi jointe des p-valeurs conformelles en utilisant uniquement l'échangeabilité des scores $(S_i)_{i \in [\![n+m]\!]}$ (et non plus l'indépendance).

Proposition 2.2 (Gazin et al. (2023))**.** *Si les hypothèses 1 et 2 sont vérifiées, alors les p-valeurs conformelles $(p_i)_{i \in [\![m]\!]}$ ont pour loi $P_{n,m}$ qui est donc libre de la loi des scores.*

L'universalité de $P_{n,m}$, dans le sens où la loi des p-valeurs conformelles est libre de la loi sous-jacente des scores n'est pas surprenante. En effet, en théorie des tests on cherche à construire des statistiques de test dont nous pouvons contrôler la loi sous (H_0), ce qui revient à trouver des statistiques de test libres de la loi sous (H_0). Par exemple, lors d'un test d'adéquation à une loi continue, il est connu que la loi de la statistique de test de Kolmogorov-Smirnov sous (H_0) est libre de la loi sous-jacente, et peut donc être simulée en utilisant des variables aléatoires uniformes. De la même façon, la proposition 2.2 montre que l'on peut simuler la loi jointe de $(p_i)_{i \in [\![m]\!]}$ à partir de variables aléatoires uniformes. La loi $P_{n,m}$ n'est pas une nouveauté en probabilités, et correspond à un phénomène bien connu: les $(p_1, \cdots, p_m)$ correspondent à m tirages dans une urne de Pólya contenant $n+1$ boules de couleurs différentes. On définit pour tout $\boldsymbol{j} = (j_1, \ldots, j_m) \in [\![n+1]\!]^m$, le vecteur $\boldsymbol{M}(\boldsymbol{j}) := (M_k(\boldsymbol{j}))_{k \in [\![n+1]\!]}$ où pour tout $k \in [\![n+1]\!]$, $M_k(\boldsymbol{j}) := |\{i \in [\![m]\!] \ : \ j_i = k\}|$ est le nombre de coordonnées dans $\boldsymbol{j}$ égales à k et dans ce cas on pose $\boldsymbol{M}(\boldsymbol{j})! := \prod_{k=1}^{n+1} (M_k(\boldsymbol{j})!)$.

Théorème 2.3. *La loi $P_{n,m}$ correspond à la distribution des couleurs de m tirages successifs dans une urne de Pólya avec $n+1$ couleurs numérotées $\left\{\frac{\ell}{n+1}, \ell \in [\![n+1]\!]\right\}$ et initialement une seule boule de chaque couleur. C'est-à-dire, si $\boldsymbol{p} \sim P_{n,m}$ définie en (3), on a les propriétés suivantes.*

(i) Description dynamique : pour tout $i \in [\![0, m-1]\!]$, la loi de p_{i+1} conditionnellement à $p_1, \ldots, p_i$ ne dépend pas de m et est donnée par

$$\mathcal{D}(p_{i+1} \,|\, p_1, \ldots, p_i) = \sum_{j=1}^{n+1} \frac{1 + \sum_{k=1}^{i} \mathbb{1}\{p_k = j/(n+1)\}}{n+1+i} \delta_{j/(n+1)}. \tag{5}$$

(ii) Loi jointe : pour tout $\boldsymbol{j} \in [\![n+1]\!]^m$,

$$\mathbb{P}\left(\boldsymbol{p} = \frac{\boldsymbol{j}}{n+1}\right) = \boldsymbol{M}(\boldsymbol{j})! \frac{n!}{(n+m)!}. \tag{6}$$

(iii) Distribution de l'histogramme: l'histogramme de $\boldsymbol{p}$ est distribué uniformément sur l'ensemble des histogrammes d'un m-échantillon en $n+1$ subdivisions, i.e. pour tout $\boldsymbol{m} = (m_1, \ldots, m_{n+1}) \in [\![0, m]\!]^{n+1}$ tel que $m_1 + \cdots + m_{n+1} = m$, on a

$$\mathbb{P}\big(\boldsymbol{M}\big((n+1)\boldsymbol{p}\big) = \boldsymbol{m}\big) = \binom{n+m}{m}^{-1}. \tag{7}$$

En particulier, conditionnellement à $\boldsymbol{M}\big((n+1)\boldsymbol{p}\big)$, la variable $\boldsymbol{p}$ est uniformément distribuée sur l'ensemble des trajectoires possibles, c'est à dire, pour tout vecteur $\boldsymbol{j} \in [\![n+1]\!]^m$,

$$\mathbb{P}\left(\boldsymbol{p} = \frac{\boldsymbol{j}}{n+1} \;\middle|\; \boldsymbol{M}\big((n+1)\boldsymbol{p}\big) = \boldsymbol{M}(\boldsymbol{j})\right) = \frac{\boldsymbol{M}(\boldsymbol{j})!}{m!}. \tag{8}$$

Un corollaire intéressant du point (iii) consiste à dire que $\widehat{F}_m$ suit une loi uniforme sur l'ensemble des fonctions de répartition des variables aléatoires à valeurs dans $[\![n+1]\!]/(n+1)$ faisant m sauts, i.e., telle que la probabilité de valoir $k/(n+1)$ soit de la forme ℓ/m avec $\ell \in [\![0, m]\!]$ un entier, et cela pour tout $k \in [\![n+1]\!]$.

2.2 Concentration de la fonction de répartition empirique

La quantité $\widehat{F}_m(t)$ est intéressante car elle représente la proportion d'erreurs faites au niveau t sur m points de tests comme énoncé dans Gazin et al. (2023). En utilisant la proposition 2.2 avec le théorème 2.3 il est possible de retrouver la distribution de $\widehat{F}_m(t)$ à $t \in [0,1]$ fixé et nous permet de retrouver les résultats de Marques F. (2023). Cependant dans le cadre d'un contrôle d'un taux d'erreur il faut connaître le processus $(\widehat{F}_m(t))_{t\in[0,1]}$. À cette fin la proposition 2.2 nous indique qu'il existe une certaine variable aléatoire $U = (U_1, \cdots, U_n)$ telle que conditionnellement à U, $\widehat{F}_m$ soit la fonction de répartition empirique d'un m-échantillon de loi P^U. Cela nous permet donc d'appliquer les théorèmes usuels, comme l'inégalité DKW par exemple, lorsque l'on conditionne, puis d'intégrer contre le conditionnement en sachant que le vecteur U est un n-échantillon de loi $\mathcal{U}(0,1)$. On peut ainsi obtenir l'inégalité de concentration de type DKW suivante.

Théorème 2.4 (Gazin et al. (2023))**.** *Supposons que les hypothèses 1 et 2 soient vérifiées. Alors on a les inégalités suivantes. Pour tout $\lambda \geq 0$,*

$$\begin{aligned}
\mathbb{P}\left(\sup_{t\in\mathbb{R}}\Big(\widehat{F}_m(t) - I_n(t)\Big) > \lambda\right) &\leq \mathbb{1}\{\lambda < 1\}\left[1 + \frac{2\sqrt{2\pi}\tau_{n,m}}{(n+m)^{1/2}}\lambda\right] e^{-2\tau_{n,m}\lambda^2};\\
\mathbb{P}\left(\sup_{t\in\mathbb{R}}\Big(\widehat{F}_m(t) - I_n(t)\Big) < -\lambda\right) &\leq \mathbb{1}\{\lambda \leq 1\}\left[1 + \frac{2\sqrt{2\pi}\tau_{n,m}}{(n+m)^{1/2}}\lambda\right] e^{-2\tau_{n,m}\lambda^2};\\
\mathbb{P}\left(\Big\|\widehat{F}_m - I_n\Big\|_\infty > \lambda\right) &\leq \mathbb{1}\{\lambda \leq 1\}\left[1 + \frac{2\sqrt{2\pi}\tau_{n,m}}{(n+m)^{1/2}}\lambda\right] 2e^{-2\tau_{n,m}\lambda^2};
\end{aligned} \tag{9}$$

avec $\tau_{n,m} = \frac{nm}{n+m} \in [n \wedge m/2, n \wedge m]$.

Cette inégalité de concentration nous donne une vitesse de concentration de l'ordre de $\sqrt{\tau_{n,m}}$, qui est également de l'ordre de l'écart type, puisque l'on peut obtenir, si les hypothèses 1 et 2 sont vérifiées, que pour tout $(t,s) \in \mathbb{R}^2$,

$$\operatorname{Cov}\Big(\widehat{F}_m(t) - I_n(t), \widehat{F}_m(s) - I_n(s)\Big) = \frac{m+n+1}{m(n+2)}(I_n(t) \wedge I_n(s) - I_n(t)I_n(s)),$$

avec $\frac{m+n+1}{m(n+2)} \sim \tau_{n,m}$ quand $n \wedge m \to +\infty$. On peut remarquer, qu'en prenant le bon λ, on peut construire un test de niveau α testant si les scores sont échangeables en utilisant comme statistique de test $\left\|\widehat{F}_m - I_n\right\|_\infty$.

Corollaire 2.5. *Soit $\delta \in]0,1[$. On définit, pour tout $r \in \mathbb{N}\backslash\{0\}$,*

$$\lambda^{\scriptscriptstyle DKW}_{\delta,n,m} = \Psi^{(r)}(1) \tag{10}$$

$$\text{où pour tout } x \in [0,1],\ \Psi(x) = 1 \wedge \left(\frac{\log\left(\frac{1}{\delta}\right) + \log\left(1 + \frac{2\sqrt{2\pi}\tau_{n,m}}{(n+m)^{1/2}}x\right)}{2\tau_{n,m}}\right)^{1/2},$$

où $\Psi^{(r)}$ est la fonction Ψ composée r fois. Alors, on a:

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}}\left(\widehat{F}_m(t) - I_n(t)\right) > \lambda^{\scriptscriptstyle DKW}_{\delta,n,m}\right) \leq \delta;$$
$$\mathbb{P}\left(\sup_{t\in\mathbb{R}}\left(\widehat{F}_m(t) - I_n(t)\right) < -\lambda^{\scriptscriptstyle DKW}_{\delta,n,m}\right) \leq \delta;$$
$$\mathbb{P}\left(\left\|\widehat{F}_m - I_n\right\|_\infty > \lambda^{\scriptscriptstyle DKW}_{\delta/2,n,m}\right) \leq \delta.$$

La formule (10) pour $r = 1$ permet de comprendre facilement la dépendance en les paramètres n et m issus des données et δ fixé à priori. Par contre, afin d'obtenir un intervalle de confiance plus petit (ou une zone de rejet plus grande dans le cadre d'un test) tout en restant au niveau $1 - \delta$, il convient de choisir un r grand. Dans le cadre des simulations effectuées dans Gazin et al. (2023), le paramètre r est choisi égal à 8.

L'uniformité des bornes (9) de ce théorème permet également de construire des bornes "post hoc" (Blanchard et al., 2020) dans le sens où l'on peut prendre n'importe quel niveau $\hat{\alpha} := \hat{\alpha}(S_1, \cdots, S_{n+m})$ dépendant des données tout en ayant la garantie

$$\mathbb{P}\left(\widehat{F}_m(\hat{\alpha}) > I_n(\hat{\alpha}) + \lambda^{\mathrm{DKW}}_{\delta,n,m}\right) < \delta.$$

3 Conclusion

Nous avons ici présenté certaines propriétés théoriques des p-valeurs conformelles, et explicité leur loi jointe ainsi qu'une propriété de concentration de leur fonction de répartition empirique. Ces résultats théoriques permettent notamment de quantifier l'incertitude des tâches de machine learning dans lesquelles $\widehat{F}_m$ représente un taux d'erreur. Nous renvoyons la lectrice intéressée ou le lecteur intéressé aux articles Marandon et al. (2022) et Gazin et al. (2023) pour plus de détails.

References

Angelopoulos, A. N. and Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

Bates, S., Candès, E., Lei, L., Romano, Y., and Sesia, M. (2023). Testing for outliers with conformal p-values. *Ann. Statist.*, 51(1):149–178.

Blanchard, G., Neuvial, P., and Roquain, E. (2020). Post hoc confidence bounds on false positives using reference families. *Ann. Statist.*, 48(3):1281–1303.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities*. Oxford University Press, Oxford. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

Gazin, U., Blanchard, G., and Roquain, E. (2023). Transductive conformal inference with adaptive scores. *arXiv preprint arXiv:2310.18108*.

Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, third edition.

Marandon, A., Lei, L., Mary, D., and Roquain, E. (2022). Adaptive novelty detection with false discovery rate guarantee. *To appear in Annals of Statistics*.

Marques F., P. C. (2023). On the universal distribution of the coverage in split conformal prediction. *arXiv preprint 2303.02770*.

Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.*, 18(3):1269–1283.

Meah, I., Blanchard, G., and Roquain, E. (2023). False discovery proportion envelopes with consistency.

Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *J. Amer. Statist. Assoc.*, 100(469):94–108.

Saunders, C., Gammerman, A., and Vovk, V. (1999). Transduction with confidence and credibility. In *16th International Joint Conference on Artificial Intelligence (IJCAI 1999)*, pages 722–726.

Shorack, G. R. and Wellner, J. A. (1986). *Empirical processes with applications to statistics*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York.