Sensibilité des indices de qualité d'un classifieur probabiliste

Ndèye Awa Dieye[1,2], Ndèye Niang [1,2] & Giorgio Russolillo [1,2]

[1] *Laboratoire Cédric-Cnam, Paris*
[2] {*ndeye-awa.dieye, ndeye.niangkeita, giorgio.russolillo*}*@lecnam.net*

Résumé. Dans le domaine de la classification supervisée, évaluer les classifieurs probabilistes implique de regarder la discrimination, qui est la capacité à distinguer les classes, et la calibration, qui est la fiabilité de l'estimation des probabilités qui génèrent la variable de réponse. Dans ce papier, nous étudions la sensibilité de plusieurs mesures de qualité d'un classifieur probabiliste binaire via des simulations. Le but est de comprendre le comportement de ces mesures face à diverses formes de distribution des probabilités qui génèrent la variable de réponse et aux caractéristiques des écarts entre ces probabilités et leurs estimations.

Mots-clés. Classification supervisée, discrimination, calibration.

Abstract. In the field of supervised classification filed, assessing probabilistic classifiers involves examining discrimination, which is the ability to distinguish between classes, and calibration, which is the reliability of estimating the probabilities that generate the response variable. In this paper, we investigate the sensitivity of various quality measures for a binary probabilistic classifier through simulations. The objective is to comprehend the behavior of these measures in the presence of different shapes of the probability distribution generating the response variable, as well as the characteristics of deviations between these probabilities and their estimates.

Keywords. Supervised classification, discrimination, calibration.

1 Introduction

Les classifieurs probabilistes, largement utilisés dans divers domaines tels que la médecine, la finance et la reconnaissance d'images, estiment la probabilité π d'un évènement pour chaque observation $i \in \{1, \ldots, n\}$. Dans le contexte de la classification binaire, les observations sont décrites par un ensemble de variables explicatives et une variable de réponse binaire $y \in \{0, 1\}$, qui indique l'occurrence de l'évènement. Ainsi, π_i représente la vraie probabilité (inconnue en pratique) qui a généré la réponse y_i pour la $i$ème observation. Celle-ci est à distinguer de la probabilité p_i *estimée* par un classifieur binaire. Pour évaluer la qualité du classifieur, il faut mesurer la précision de l'estimation des probabilités π_i et la qualité des prédictions des valeurs y_i de la variable cible.

Plusieurs indices sont utilisés pour évaluer la qualité des classifieurs binaires. Ce travail présente une étude de la sensibilité de certains de ces indices. Nous définissons la sensibilité d'un indice comme son comportement face à des erreurs d'intensité différente dans

l'estimation des vraies probabilités. Via des simulations, on étudie cette sensibilité en fonction de 2 facteurs : la distribution des vraies probabilités π_i et les caractéristiques des écarts (monotonie, symétrie, etc.) entre les probabilités estimées et les vraies probabilités.

Dans les sections suivantes, nous détaillons les mesures utilisées pour évaluer la performance des classifieurs binaires, nous présentons notre méthodologie expérimentale, puis nous discutons des résultats obtenus et de leurs implications en pratique.

2 Mesures de la performance d'un classifieur probabiliste

L'évaluation d'un classifieur se fait en examinant son pouvoir discriminant, sa calibration ou les deux à la fois. Le pouvoir discriminant est la capacité à bien classer les individus. La calibration mesure la fiabilité de l'estimation de la probabilité d'appartenance aux classes.

Dans la pratique, la capacité discriminante est communément évaluée par le taux d'erreur ainsi que de nombreux indices (taux de vrais positifs, taux de vrais négatifs, f1-score, etc.). Toutefois, ces indices dépendent du seuil de probabilité choisi pour faire le classement.

Cependant, d'autres mesures existent prenant en compte tous les seuils possibles comme l'aire sous la courbe ROC (AUC). La courbe ROC représente le taux de vrais positifs et le taux de vrais négatifs pour différents seuils de classement et l'AUC exprime la probabilité qu'un exemple positif sélectionné au hasard x_j obtienne une probabilité plus élevée qu'un exemple négatif sélectionné au hasard x_i. Elle est donc une mesure de concordance. Soit y^0 l'ensemble des n_0 individus de classe 0 et y^1 l'ensemble des n_1 individus de classe 1, l'AUC est estimée comme (Calders & al. (2007)) :

$$AUC := \frac{1}{n_0 \cdot n_1} \left(\sum_{x_i \in y^0} \sum_{x_j \in y^1} \mathbb{1}[p_{x_i} < p_{x_j}] + \frac{1}{2} \sum_{x_i \in y^0} \sum_{x_j \in y^1} \mathbb{1}[p_{x_i} = p_{x_j}] \right)$$

où $\mathbb{1}[p_{x_i} < p_{x_j}]$ nous donne 1 si $p_{x_i} < p_{x_j}$ et 0 sinon. L'AUC varie dans $[0; 1]$.

Pour ce qui concerne l'évaluation de la calibration, on distingue la calibration *in the large* (calibration en moyenne par Van Calster & al. (2016)) et la calibration *in the small*. La première est issue de la comparaison du taux observé $\overline{y}$ de l'évènement et la moyenne $\overline{p}$ des probabilités estimées. Dans ce travail, nous choisissons de la calculer comme le rapport entre ces deux quantités :

$$Mean\_calib = \overline{y}/\overline{p}$$

Cet indicateur est égal à 1 lorsque le modèle est parfaitement calibré en moyenne. Lorsqu'il est supérieur à 1, il y a sous-estimation et surestimation dans le cas contraire.

Afin de mesurer la calibration *in the small*, on ordonne les individus selon les probabilités *estimées*, on les partitionne en G groupes de taille n_g et on compare le taux observé

d'évènement $\overline{y}_g$ et la moyenne des probabilités estimées $\overline{p}_g$ dans chaque groupe $g \in \{1, \ldots, G\}$. Dans ce travail, les groupes sont définis par les déciles de la distribution des probabilités estimées.

La calibration *in the small* peut être graphiquement évaluée à l'aide du diagramme de fiabilité (Figure 1). Il représente les taux d'évènement $(\overline{y}_g)$ en fonction des moyennes des probabilités prédites $(\overline{p}_g)$ de chaque groupe. Si le modèle est parfaitement calibré, alors les points seront tous sur la première bissectrice. Toute déviation, pour un groupe, de cette situation parfaite traduit une mauvaise calibration pour un certain niveau de probabilité (Pepe & al. (2013)).

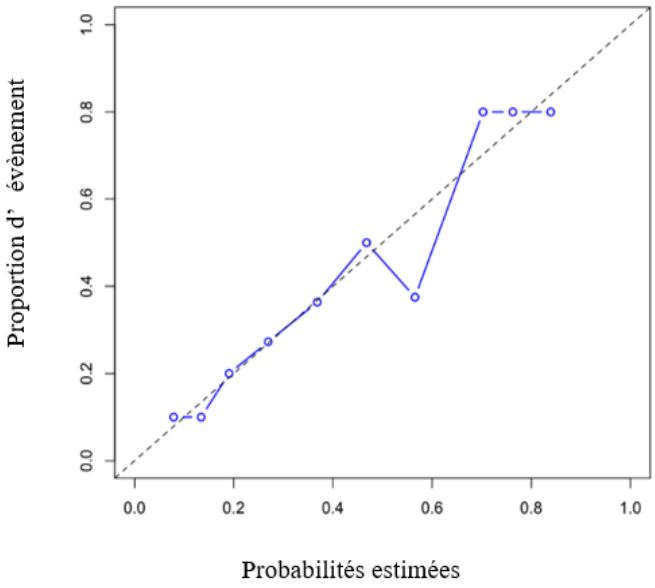

Figure 1: Exemple de diagramme de fiabilité. Un point se trouvant au-dessous de la bissectrice représente une probabilité surestimée et un point se trouvant au-dessus de la bissectrice représente une probabilité sous-estimée.

L'Expected Calibration Error (ECE- Naeini & al. (2015)) permet d'évaluer globalement la calibration *in the small* par une valeur numérique. Il est calculé comme la moyenne pondérée des valeurs absolues des écarts des taux observés à la moyenne des probabilités estimées dans chaque groupe :

$$ECE = \sum_{g=1}^{G} \frac{n_g}{n} |\overline{y}_g - \overline{p}_g|$$

Une bonne calibration correspond à un ECE faible, soit proche de 0.

Une mesure qui permet d'évaluer à la fois la discrimination et la calibration est le score de Brier. C'est la moyenne des carrés des différences entre les probabilités estimées et les

valeurs de y pour chaque individu (Brier, G.W. (1950)) :

$$B = \frac{1}{n}\sum_{i=1}^{n}(y_i - p_i)^2$$

Le score de Brier varie entre $[0; 1]$. Un score plus faible représente une meilleure performance.

Plusieurs décompositions du score de Brier existent en pratique dont l'une des premières en 2 éléments est donnée par Sanders, F. (1963) :

$$B = \frac{1}{n}\sum_{g=1}^{G} n_g \overline{y}_g(1 - \overline{y}_g) + \frac{1}{n}\sum_{g=1}^{G} n_g(p_g - \overline{y})^2 \tag{1}$$

Le premier terme de l'équation (1) est appelé *résolution* et le second terme correspond à la *calibration*. Vu que la résolution est liée à la discrimination (Huang, C. & al. (2021)), le score de Brier est généralement considéré comme une mesure de la qualité globale d'un classifieur.

Ainsi, nous utilisons l'AUC pour évaluer la discrimination ; la calibration en moyenne, le diagramme de fiabilité et l'ECE pour évaluer la calibration et le score de Brier pour les deux à la fois.

3 Méthodologie de simulation

Notre simulation vise à évaluer la sensibilité des indices cités dans la section précédente dans des conditions où la calibration est de plus en plus mauvaise. L'objectif est de comprendre comment ces indices réagissent aux déviations des probabilités estimées par rapport aux vraies probabilités, ce qui peut être crucial dans des applications réelles où la calibration des prédictions est nécessaire.

Pour ce faire, nous avons généré quatre vecteurs x_j ($j \in \{u, asym, unif, cloche\}$) de taille $n = 2000$ à partir de différentes lois (beta, normale et logistique). Chacun de ces vecteurs x_j est ensuite transformé en un vecteur ${\pi_j}^1$ de probabilités en utilisant le modèle logistique $\pi_j = \frac{e^{\beta x_j}}{e^{\beta x_j}+1}$ avec $\beta = 1$. Ces 4 vecteurs de probabilités présentent des distributions avec des formes différentes, comme illustrées dans la figure 2 :

- Forme en u : symétrique avec 2 modes aux extrêmes.
- Forme *asymétrique* : étalée à gauche (asymétrique positive).
- Forme *uniforme* : toutes les valeurs ont la même probabilité d'occurrence.
- Forme en *cloche* : symétrique avec mode à 0.5.

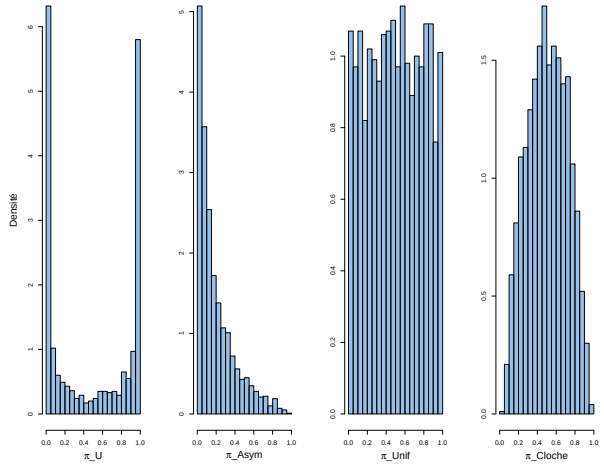

Figure 2: Distributions des probabilités π. De droite à gauche : *u*, *asymétrique* positive, *uniforme* et *cloche*

Pour la suite, 1000 valeurs binaires y_{ijk} ($k \in \{1, \ldots, 1000\}$), issues de lois de Bernoulli de paramètre π_{ij}, sont générées de façon aléatoire.

Afin d'obtenir des vecteurs de probabilités estimées p_j que nous considérons issus de classifieurs probabilistes, nous dévions les 4 vecteurs de vraies probabilités π_j en utilisant 3 méthodes de déviation qui présentent des caractéristiques différentes (figure 3). Pour chaque méthode, on varie l'intensité de l'erreur :

- **Méthode 1** : la valeur du coefficient β est remplacée par les valeurs 0.75, 0.5 et 0.25. Cette méthode donne des déviations monotones, symétriques, plus accentuées sur les valeurs extrêmes.
- **Méthode 2** : des bruits gaussiens de moyenne nulle et d'écart-type 0.2, 0.5, 1 et 2 sont ajoutés aux vecteurs π_j. Cette méthode donne des déviations non monotones, symétriques, plus accentuées sur les valeurs centrales.
- **Méthode 3** : le premier tercile de la distribution des π_j est multiplié par 1.075, 1.15, 1.3, 1.4, 1.5 et le dernier tercile par 0.925, 0.85, 0.7, 0.6, 0.5 (inspiré de Huang, Y. & al. (2020)). Cette méthode donne des déviations non monotones, asymétriques, plus accentuées sur les valeurs proches de 1.

Enfin, nous calculons les indices de performance, à savoir l'AUC, le score de Brier, l'ECE et la calibration en moyenne. Pour ce faire, nous comparons, pour chaque répétition k, les valeurs y_j de la réponse à une situation de référence (valeur optimale) où le modèle estime parfaitement les vraies probabilités (comparaison des valeurs y_j de la réponse aux vraies probabilités π_j).

[1] L'élément générique i du vecteur π_j est considéré comme la vraie probabilité associée à l'$i$ème observation.

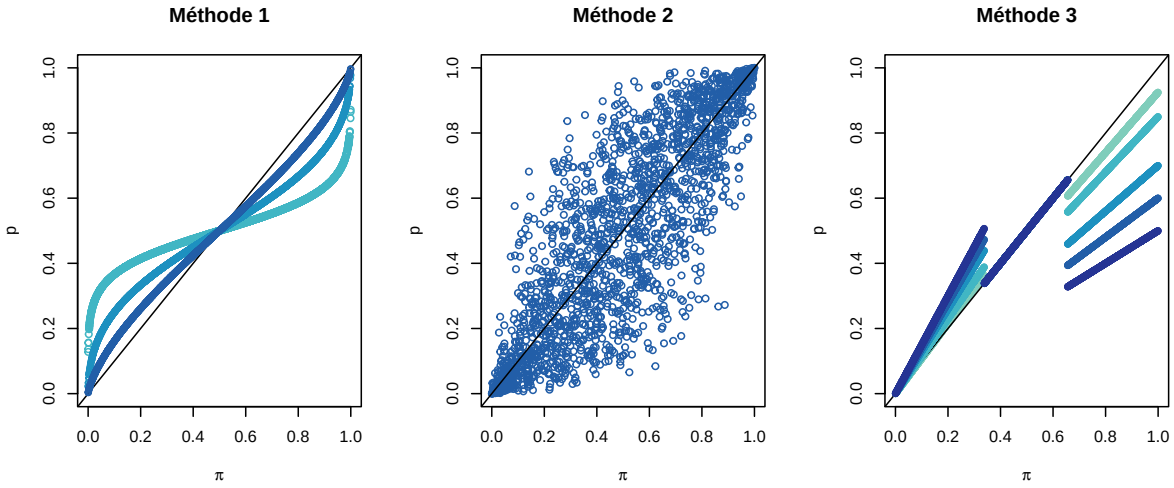

Figure 3: Caractéristiques des écarts des probabilités estimées p aux vraies probabilités π. Pour la méthode 2, $sd = 1$

Les valeurs y_j, pour chaque répétition k, sont ensuite comparées aux probabilités estimées p_j issues du croisement des 4 distributions des vraies probabilités et des déviations de tout type et intensité.

4 Résultats et discussion

La figure 4 nous donne une visualisation de l'influence de la distribution des probabilités sur les valeurs optimales de l'AUC et du score de Brier, en l'absence de toute déviation. Ces valeurs sont étroitement liées à la distribution des probabilités. Lorsque la distribution présente plus de valeurs extrêmes (forme en u), la valeur optimale tend à être meilleure. En revanche, lorsque la distribution est plus concentrée autour des valeurs centrales (forme en *cloche*), la valeur optimale tend à être moins bonne.

Les figures 5 et 6 illustrent respectivement la sensibilité de la calibration en moyenne et de l'ECE à différents niveaux d'écarts entre les probabilités estimées et les vraies probabilités, en fonction des distributions des probabilités et des méthodes de déviation.

La mesure de la calibration *in the large* est plus sensible lorsque la distribution de probabilités est *asymétrique* ou l'ampleur des écarts est asymétrique (figure 5). Plus spécifiquement, pour la méthode 3 présentant des déviations asymétriques, la calibration en moyenne est sensible pour toutes les formes de distribution, avec une sensibilité plus importante pour la distribution *asymétrique*. Ce qui n'a pas été noté avec les méthodes 1 et 2 (qui génèrent des écarts symétriques) où l'indice est sensible seulement lorsque la forme de distribution est *asymétrique*.

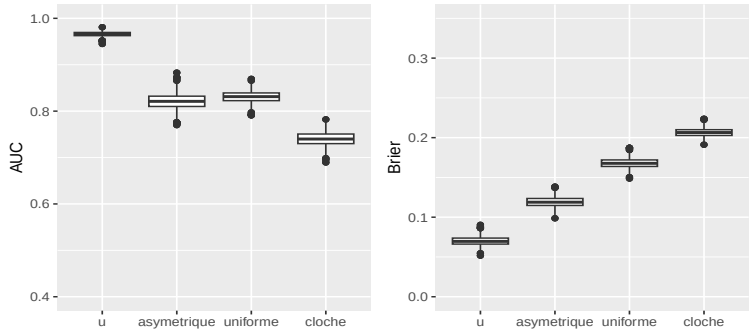

Figure 4: Boîtes à moustaches des 1000 valeurs optimales de l'AUC et du score de Brier pour chaque forme de distribution

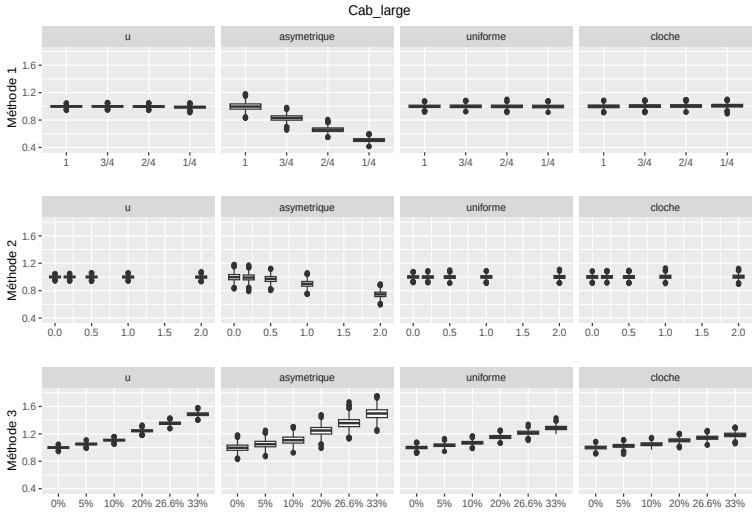

Figure 5: Boîtes à moustaches des 1000 valeurs obtenues de la calibration en moyenne par méthode et forme de distribution

L'ECE se révèle plus sensible lorsque les écarts [2] les plus importants se situent au niveau des probabilités présentant une densité plus élevée (figure 6). Pour la méthode 1, les plus larges écarts se trouvent au niveau des valeurs extrêmes (figure 3), ce qui rend l'ECE plus sensible pour les distributions en *u* et *asymétrique*. Pour la méthode 2, les plus grandes déviations se trouvent au niveau des valeurs centrales, rendant l'ECE plus sensible pour les distributions des vraies probabilités en *cloche* et *uniforme*. Enfin, avec la méthode 3, les plus larges écarts sont au niveau des probabilités proches de 1, ce qui fait qu'on observe une plus grande sensibilité pour la distribution des vraies probabilités en *u* par rapport au reste. Cela est d'ailleurs plus marquant en regardant la distribution des vraies probabilités *asymétrique* présentant une très faible sensibilité étant donné qu'elle est concentrée au niveau

[2] écarts entre les probabilités estimées et les vraies probabilités

des probabilités proches de 0.

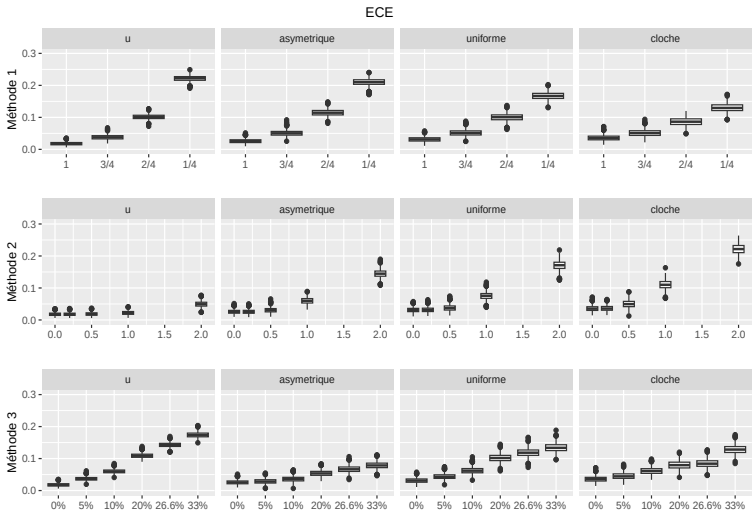

Figure 6: Boîtes à moustaches des 1000 valeurs obtenues de l'ECE par méthode et forme de distribution

Pour vérifier si une étude plus approfondie de la calibration pouvait nous aider à mieux détecter les estimations des probabilités de différente intensité par rapport à un indice global comme l'ECE, nous avons construit les diagrammes de fiabilité correspondants (voir figure 7).

Les diagrammes de fiabilité confirment que l'interprétation de la qualité de classifieurs peut changer selon la forme de la distribution des vraies probabilités lorsque la qualité des estimations reste la même.

Par exemple, à parité de type (méthode 2) et d'intensité ($sd = 1$: assez importante) de la déviation, le diagramme peut générer des interprétations très différentes de la qualité de la calibration selon la forme de la distribution des vraies probabilités. Si la forme est en u, l'estimation des probabilité semble presque parfaite, tandis que pour les autres formes la calibration ne semble pas satisfaisante.

5 Conclusion et perspectives

Cette étude sur la sensibilité des indices de qualité d'un classifieur probabiliste a permis de réaliser que ces indices dépendent non seulement de l'intensité des écarts des probabilités estimées aux vraies probabilités, mais aussi des caractéristiques de ces écarts et des vraies probabilités. Même si en théorie l'AUC et le score de Brier sont compris respectivement dans les intervalles $[0.5; 1]$ et $[0; 1]$, dans la pratique les valeurs optimales de ces indices dépendent de la distribution des vraies probabilités π, qui est inconnue en pratique. De plus, la vraie qualité de la calibration peut ne pas être détectée avec les indices considérés dans ce papier vu qu'ils dépendent de la distribution de π.

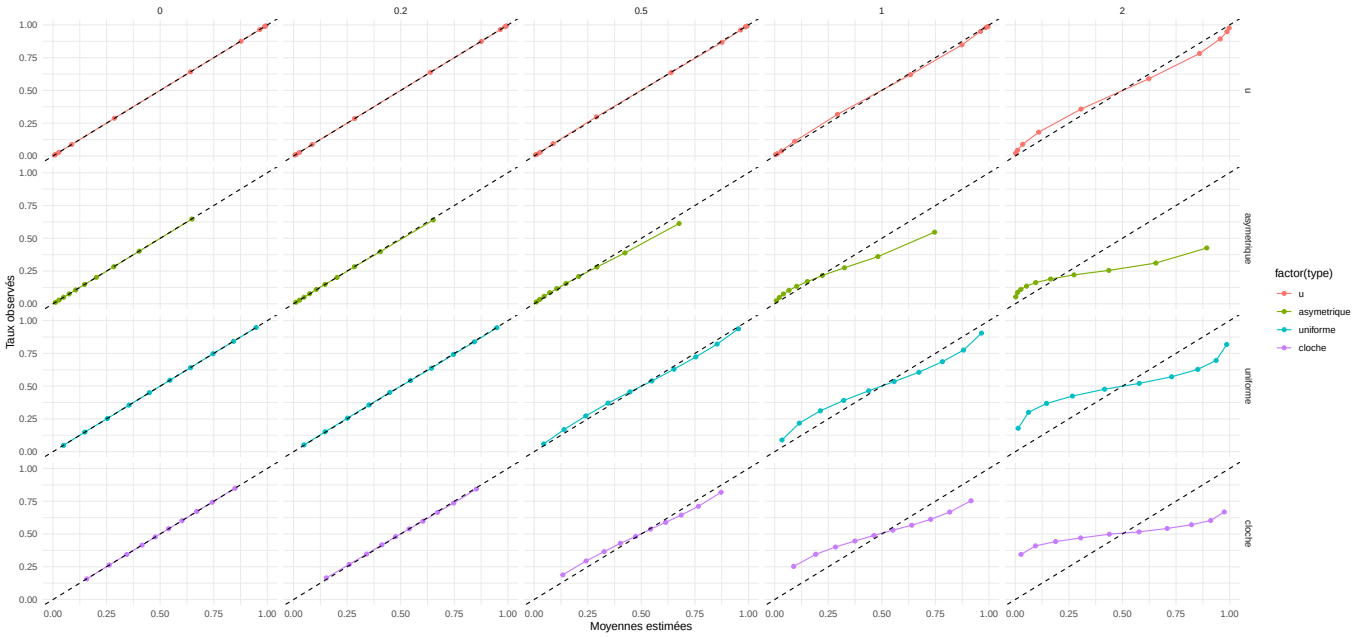

Figure 7: Diagramme de fiabilité pour chaque forme de la distribution des vraies probabilités et chaque niveau des déviations issues de la méthode 2

Il convient donc d'être prudent dans l'interprétation de la qualité des classifieurs probabilistes avec ces indices afin d'éviter des interprétations erronées ou biaisées. Des recherches futures sont nécessaires pour développer des méthodologies plus robustes permettant d'évaluer de manière fiable les performances des classifieurs probabilistes.

Bibliographie

Brier, G.W. (1950), *Verification of forecasts expressed in terms of probability*, Monthly Weather Review, 78 (1): 1-3.

Calders, T. and Jaroszewicz, S. (2007), *Efficient AUC Optimization for Classification*, Knowledge Discovery in Databases: PKDD 2007, Lecture Notes in Computer Science, vol 4702, Springer, Berlin, Heidelberg

Huang, Y. and Li, W. and Macheret, F. and Gabriel, R. A and Ohno-Machado, L. (2020), *A tutorial on calibration measurements and calibration models for clinical prediction models*, JAMIA, 27(4), pp. 621-633

Huang, C. and Li, S.X. and Caraballo, C. and Masoudi, F.A. and Rumsfeld, J.S. and Spertus, J.A. and Normand, S.T., Mortazavi, B.J. and Krumholz, H.M. (2021), *Performance Metrics for the Comparative Analysis of Clinical Risk Prediction Models Employing Machine Learning*, Circ Cardiovasc Qual Outcomes, 14(10):e007526

Pepe, M. and Janes, H. (2013), *Methods for Evaluating Prediction Performane of Biomarkers*

and Tests, In: Lee, ML., Gail, M., Pfeiffer, R., Satten, G., Cai, T., Gandy, A. (eds) Risk Assessment and Evaluation of Predictions. Lecture Notes in Statistics, vol 215. Springer, New York, NY

Naeini, P.M. and Cooper, G. and Hauskrecht, M. (2015), *Obtaining Well-Calibrated Probabilities Using Bayesian Binning*. Proceedings of the AAAI Conference on Artificial Intelligence, 29(1)

Sanders, F. (1963), *On Subjective Probability Forecasting*, J. Appl. Meteor. Climatol., 2

Van Calster, B. and Daan, N. and Yvonne, V. and Bavo, C. and Michael, J.P. and Ewout, W. S. (2016), *A calibration hierarchy for risk models was defined: from utopia to empirical data*, Journal of clinical epidemiology