Une méthode d'approximation des effets de Shapley en grande dimension

Bertrand Iooss, Laura Clouvel

Une méthode d'approximation des effets de Shapley en grande dimension

Bertrand Iooss [1] & Laura Clouvel [2]

[1] *EDF R&D, Département PRISME, Chatou, France et bertrand.iooss@edf.fr*
[2] *EDF R&D, Département PERICLES, Saclay, France et laura.clouvel@edf.fr*

Résumé. En apprentissage statistique supervisé et en analyse d'incertitudes de codes de calcul, les mesures d'importance (ou indices de sensibilité) ont pour but de quantifier l'influence des entrées (ou covariables) du modèle d'apprentissage ou du code de calcul sur sa sortie. Du fait de leur simplicité d'interprétation même en présence de fortes dépendances entre les entrées, les effets de Shapley ont récemment suscité un grand intérêt parmi les utilisateurs avides d'interprétabilité de modèles "boîte noires". Malheureusement, leur coût calculatoire pour les estimer limite leur utilisation à des cas de faible dimension (de l'ordre d'une dizaine d'entrées). Dans cette communication, nous étudions une méthode d'approximation des effets de Shapley, nommée l'allocation de poids relatifs ("Relative Weight Allocation"), développée dans le contexte de la régression linéaire. Une extension est proposée et testée pour des modèles non linéaires.

Mots-clés. Analyse de sensibilité, Apprentissage statistique, Interprétabilité, Indices de Johnson, Mesure d'importance, Régression linéaire, Shapley, Sobol.

Abstract. In supervised statistical learning and in uncertainty quantification of computer codes, the importance measures (or sensitivity indices) aim to quantify the influence of the inputs (or covariates) of the learning model or of the computer code on its output. Due to their ease of interpretation even in the presence of strong dependencies between inputs, Shapley effects have recently aroused great interest among users eager for the interpretability of "black box" models. Unfortunately, their estimation computational cost limits their use to low-dimensional cases (of the order of ten inputs). In this communication, we study a method of approximating Shapley's effects, called "Relative Weight Allocation", developed in the context of linear regression. An extension is proposed and tested for nonlinear models.

Keywords. Sensitivity analysis, interpretability, Importance measure, Johnson indices, Linear regression, Machine learning, Shapley, Sobol.

1 Introduction

En analyse de sensibilité de codes de simulation numérique (Da Veiga *et al.*, 2021) ou en interprétabilité de modèles d'apprentissage statistique (Lepore *et al.*, 2022), nous étudions la fonction suivante:

$$Y = f(\boldsymbol{X}),$$

où $f(\cdot)$ est le modèle numérique à analyser (souvent appelé "boîte noire"), $Y \in \mathbb{R}$ est la sortie et $\boldsymbol{X} = (X_1, \ldots, X_d) \in \mathbb{R}^d$ est le vecteur d'entrées. Dans ce travail, nous nous concentrons sur les mesures d'importance (MI) décomposant la variance de la sortie (Johnson and LeBreton, 2004, Iooss *et al.*, 2022).

En analyse du modèle de régression linéaire, pour contrer le problème de l'impact de dépendances statistiques entre entrées sur la compréhension de l'influence des entrées, certaines MI consistent à décomposer exactement le R^2 (pourcentage de variance expliquée par le modèle) en parts positives dues à chaque entrée (Johnson and LeBreton, 2004, Grömping, 2007). L'indice LMG (Lindeman *et al.*, 1980, Grömping, 2007), particulièrement intéressant, mesure la moyenne de la plus-value en R^2 qu'une entrée apporte dans tous les sous-modèles linéaires possible (i.e. ceux composés de toutes les combinaisons possibles des autres entrées). Soit u un sous-ensemble d'indices dans l'ensemble de tous les sous-ensembles de $\{1, \ldots, d\}$ et $\boldsymbol{X}_u = (X_j : j \in u)$ un groupe d'entrées, l'indice LMG s'écrit :

$$\mathrm{LMG}_j = \sum_{u \subseteq -\{j\}} \frac{(d-1-|u|)!\,|u|!}{d!} \left[R^2_{Y(\boldsymbol{X}_{v \cup \{j\}})} - R^2_{Y(\boldsymbol{X}_v)} \right].$$

où $R^2_{Y(\boldsymbol{X}_v)}$ désigne le R^2 du modèle linéaire entre Y et $\boldsymbol{X}_v$. La principale difficulté des indices LMG réside dans le coût computationnel pour les estimer, ce qui limite leur utilisation à des modèles à une dizaine d'entrées.

Les indices LMG ne sont autres que les valeurs dites de Shapley issues de la théorie des jeux coopératifs (Shapley, 1953), avec comme fonction coût le R^2 du modèle de régression linéaire. Dans le cadre général de modèles $f(\cdot)$ non linéaires, les effets de Shapley ont été introduits en analyse de sensibilité pour résoudre le problème de la dépendance entre entrées (Owen, 2014), et s'écrivent:

$$Sh_j = \sum_{u \subseteq -\{j\}} \frac{(d-1-|u|)!\,|u|!}{d!} \left[c(u \cup \{j\} - c(u) \right].$$

où $c(u) = \mathrm{Var}\left(\mathbb{E}\left[Y|\boldsymbol{X}_u\right]\right) / \mathrm{Var}(Y)$ correspond à la MI nommée indice de Sobol' fermé (Sobol', 1993, Da Veiga *et al.*, 2021). Si le modèle est linéaire, les effets de Shapley sont les indices LMG.

En pratique, de la même manière que les indices LMG, les effets de Shapley ne peuvent être estimés pour des modèles au-delà d'une dizaine d'entrées. Il existe des algorithmes permettant de les approximer (par exemple par la méthode des k plus proches voisins, cf. Broto *et al.*, 2020), mais des biais importants subistent, et ne peuvent être éliminés qu'en augmentant de manière considérable la taille des échantillons (ce qui n'est pas toujours possible), et donc aussi le coût computationnel de l'algorithme.

La section suivante présente la méthode d'allocation des poids relatifs pour le modèle de régression linéaire, permettant d'approximer les LMG de manière rapide. La section 3 tente une généralisation de cette méthode pour approximer les effets de Shapley pour des modèles non linéaires. Une dernière section contient quelques tests numériques préliminaires en faible dimension.

2 La méthode d'allocation des poids relatifs en régression linéaire

La méthode d'allocation des poids relatifs (Johnson, 1966, Johnson, 2000) utilise un échantillon de taille n d'entrées (noté $\mathbf{X}^n$) et l'échantillon de sorties correspondantes (noté $\mathbf{y}^n$). Elle consiste à effectuer d'abord une décomposition en valeurs singulières (SVD) de la matrice des entrées $\mathbf{X}^n$, afin de transformer les entrées corrélées $\boldsymbol{X}$ en variables non corrélées $\boldsymbol{Z}$, ce qui produit l'échantillon $\mathbf{Z}^n$ (cf. Fig. 1). Un processus de repondération adéquat est ensuite mise en œuvre. Celui-ci utilise les coefficients $(\alpha_i)_{i=1\ldots d}$ de la régression linéaire de $\mathbf{y}^n$ sur $\mathbf{Z}^n$, ainsi que les d combinaisons linéaires entre $\mathbf{X}^n$ et $\mathbf{Z}^n$ produisant la matrice des poids W estimé par $\hat{\mathbf{W}} = (\mathbf{Z}^n)^\top \mathbf{X}^n = (\widehat{w}_{ij})_{1\leq i,j\leq d}$.

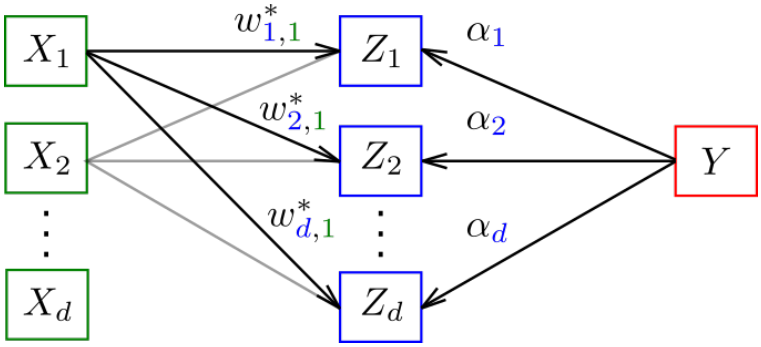

Figure 1: Représentation de la méthode d'allocation des poids relatifs pour l'entrée X_1.

L'allocation de poids relatifs (nommé RWA pour "Relative Weight Allocation") est réalisée en utilisant la matrice standardisée $\mathbf{W}^*$ (i.e. la matrice de corrélation entre $\boldsymbol{Z}$ et $\boldsymbol{X}$). La contribution proportionnelle de X_j sur Y est alors estimée en multipliant la proportion α_i^2 de la variance de Y expliquée par Z_i par la proportion w_{ij}^{*2} de chaque Z_i prise en compte par X_j. Pour une entrée X_j, la MI J_j, appelée indice de Johnson, s'obtient par

$$J_j = \frac{1}{\mathrm{Var}Y} \sum_{i=1}^{d} \alpha_i^2 w_{ij}^{*2}.$$

En partant d'entrées $\boldsymbol{X}$ standardisées, la somme des indices de Johnson est bien égale au R^2 (Johnson, 2000). Dans ce cas là, on parle d'indices de Johnson standardisés (Clouvel *et al.*, 2023), qui peuvent être aussi obtenus sans standardiser les entrées mais les $(\alpha_i)_{i=1\ldots d}$.

Les indices de Johnson sont utilisables sur des modèles à grands nombres d'entrée (pluieurs dizaines), car ils ne nécessitent qu'une SVD et une régression linéaire. Sur de nombreux cas tests, il a été montré qu'ils fournissent des résultats similaires à ceux obtenus via les indices LMG à un coût computationnel très réduit (Johnson and LeBreton, 2004, Nimon and Oswald, 2013, Clouvel *et al.*, 2019, Iooss *et al.*, 2022). Par ailleurs, il a été récemment démontré que les LMG et les indices de Johnson standardisés sont équivalents dans le cas particulier d'un modèle linéaire à deux entrées (Clouvel *et al.*, 2023).

3 Une extension pour les modèles non linéaires

L'idée est de ne plus utiliser un modèle de régression linéaire entre Y et $\boldsymbol{Z}$, mais de remplacer les coefficients de régression $(\alpha_j)_{j=1...d}$ par des MI plus générales (cf. Fig 1). Nous cherchons des MI que l'on peut estimer avec un grand nombre d'entrées d qui sont, cette fois-ci, non corrélées entre elles. La MI de chaque entrée Z_i doit porter les informations de ses effets non-linéaires sur Y, mais aussi les effets de ses interactions avec les autres entrées. Dans le cas d'entrées indépendantes, les effets de Shapley $(Sh_j)_{j=1...d}$ contiennent bien ces informations, ec qui est montré par leur expression en fonction des indices de Sobol' (Owen, 2014) :

$$Sh_j = \sum_{u \subseteq \{1...d\}, j \in u} \frac{S_u}{|u|}, \tag{1}$$

où $S_j = \mathrm{Var}(\mathbb{E}[Y|X_j]/\mathrm{Var}(Y)$ est l'indice de Sobol' du premier ordre (effet principal de X_j), $S_{jk} = \mathrm{Var}(\mathbb{E}[Y|X_jX_k]/\mathrm{Var}(Y) - S_j - S_k$ est l'indice de Sobol' du deuxième ordre (effet d'interaction entre X_j et X_k), etc. L'équation (1) montre clairement la répartition égalitaire des effets d'une interaction (entre les différentes entrées qui la composent) induite par les effets de Shapley.

Un algorithme récent et particulièrement efficace rend possible l'estimation de tous les indices de Sobol du premier ordre et deuxième ordre avec un coût indépendant de la dimension d (Gilquin *et al.*, 2019). Contrairement aux k plus proches voisins, cet algorithme n'est pas "given-data", i.e. qu'il doit générer lui-même, à partir d'un échantillon d'entrées, un nouvel échantillon particulier d'entrées (nommé plan d'expériences), qui doit être évalué avec le modèle $f(\cdot)$ afin d'en récupérer les sorties Y. Avec cet algorithme, les effets de Shapley sont donc approximés par :

$$Sh_j \simeq S_j + \sum_{k \neq j} \frac{S_{jk}}{2}, \tag{2}$$

Par ailleurs, cette relation est valide dans le cas d'entrées indépendantes, alors que les variables $\boldsymbol{Z}$ que l'on considère ne sont qu'orthogonales, ce qui induit une deuxième source d'erreur dans le cas d'entrées non gaussiennes.

L'algorithme que l'on propose, nommé RWA-Shapley, est le suivant :

- On part d'une matrice d'entrées $\mathbf{X}^n$;
- Une SVD de $\mathbf{X}^n$ produit $\mathbf{Z}^n$;
- Le plan d'expériences pour appliquer l'algorithme de Gilquin *et al.* (2019) est construit sur des lois uniformes entre 0 et 1 ;
- Chaque colonne de ce plan d'expériences est transformé en appliquant la fonction quantile empirique de chaque colonne de $\mathbf{Z}^n$;
- Cette nouvelle matrice est multiplié à $\mathbf{W}^*$ pour remonter à un échantillon de $\boldsymbol{X}$;
- On applique $f(\cdot)$ sur cet échantillon pour obtenir un vecteur de sorties de Y ;

- Les indices de Sobol du premier ordre et deuxième ordre (et donc les effets de Shapley par l'équation (2)) de $\boldsymbol{Z}$ sur Y peuvent alors êtres estimés :
- On estime les indices, que l'on nomme indices de Johnson-Shapley, par :

$$JSh_j = \frac{1}{\mathrm{Var}Y} \sum_{i=1}^{d} Sh_i w_{ij}^{*2}.$$

4 Tests numériques préliminaires

Nous illustrons les méthodes explicitées dans les sections précédentes sur deux modèles jouets à trois entrées gaussiennes ($X_1 \sim \mathcal{N}(1, 0.25)$, $X_i \sim \mathcal{N}(0, 1)$ pour $i = 1, 2$). Leur faible dimension permet de réaliser des estimations d'effets de Shapley fiables (par la méthode des k plus proches voisins) qui seront nos valeurs de référence. Pour les différentes estimations d'indiecs réalisées, de grandes tailles d'échantillon sont prises (10^5 pour les indices de Johnson/Johnson-Shapley et 10^4 pour les indices LMG/Shapley. Les incertitudes sur les indices de type Johnson ne sont pas évaluées (la méthode reste à développer) ; celles sur les indices LMG/Shapley sont inférieures à 1% (valeur correspondant à l'écart type de l'estimation évalué par bootstrap).

Le premier modèle étudié est le suivant :

$$f(\boldsymbol{X}) = X_1(1 + X_1(\cos(X_2 + X_3)^2)), \tag{3}$$

Ce modèle contient un effet linéaire prépondérant de X_1 et une interaction forte entre X_2 et X_3. En effet, dans le cas d'entrées indépendantes, les indices de Sobol' (estimés par l'algorithme de Gilquin *et al.* (2019)) valent $S_1 \simeq 60\%$, $S_2 \simeq S_3 \simeq 0\%$, $S_{23} \simeq 30\%$ alors que la régression linéaire donne $R^2 \simeq 59\%$ (seul X_1 a un effet dans le modèle linéaire).

Le Tableau 1 donne les résultats des MI sur le modèle (3) pour des entrées indépendantes. Les indices de Johnson approximent parfaitement les LMG (ils sont théoriquement égaux dans ce cas où les entrées sont indépendantes) et les indices de Johnson-Shapley approchent mais sous-estiment légèrement les effets de Shapley. Cette sous-estimation vient en partie du fait que les indices de Johnson-Shapley ne prennent pas en compte une interaction du troisième ordre.

	LMG	Johnson	Shapley	Johnson-Shapley
X_1	59	59	62	60
X_2	0	0	19	16
X_3	0	0	19	16
Sum	59	59	100	92

Table 1: Mesures d'importance (en %) sur le modèle (3) pour des entrées indépendantes.

Le Tableau 2 donne les résultats des MI sur le modèle (3) pour des entrées dépendantes : une corrélation de 0.9 est fixée entre X_1 et X_2. L'effet de la corrélation se voit immédiatement

sur les LMG et les effets de Shapley : par rapport au cas indépendant, X_1 transfère une part de son influence à X_2. Comme dans le cas indépendant les indices de Johnson approximent parfaitement les LMG et les indices de Johnson-Shapley approchent mais sous-estiment légèrement les effets de Shapley.

	LMG	Johnson	Shapley	Johnson-Shapley
X_1	35	35	38	35
X_2	23	23	38	37
X_3	0	0	24	23
Sum	58	58	100	96

Table 2: Mesures d'importance (en %) sur le modèle (3) pour des entrées dépendantes.

Le second modèle étudié est le suivant :

$$f(\boldsymbol{X}) = \sin(\frac{\pi}{2}X_1)(1 + X_1(\cos(X_2 + X_3)^2)), \tag{4}$$

Ce modèle contient un effet non linéaire de X_1 et une interaction très forte entre X_2 et X_3. En effet, dans le cas d'entrées indépendantes, les indices de Sobol' valent $S_1 \simeq 21\%$, $S_2 \simeq S_3 \simeq 1\%$, $S_{23} \simeq 71\%$ alors que la régression linéaire donne $R^2 \simeq 7\%$.

Le Tableau 3 donne les résultats des MI sur le modèle (4) pour des entrées indépendantes. Les indices de Johnson approximent parfaitement les LMG (ils sont théoriquement égaux dans ce cas où les entrées sont indépendantes) mais le R^2 de la régression linéaire, qui est expliqué par les LMG, est très faible. Les indices de Johnson-Shapley approximent également relativement correctement les effets de Shapley.

	LMG	Johnson	Shapley	Johnson-Shapley
X_1	7	7	25	22
X_2	0	0	37	37
X_3	0	0	37	36
Sum	7	7	99	96

Table 3: Mesures d'importance (en %) sur le modèle (4) pour des entrées indépendantes.

Le Tableau 4 donne les résultats des MI sur le modèle (4) pour des entrées dépendantes : une corrélation de 0.9 est fixée entre X_1 et X_2. Les indices de Johnson-Shapley approximent bien les effets de Shapley.

	LMG	Johnson	Shapley	Johnson-Shapley
X_1	4	4	22	19
X_2	2	2	34	36
X_3	0	0	44	43
Sum	6	6	100	99

Table 4: Mesures d'importance (en %) sur le modèle (4) pour des entrées dépendantes.

5 Travaux en cours et perspectives

Des tests numériques sur des modèles à plus grande dimension (jusqu'à 15) ont été réalisés. Ils montrent que les indices de Johnson et de Johnson-Shapley peuvent être estimés en quelques secondes (sur un PC standard), alors que les LMG et effets de Shapley ne peuvent pas être estimés en un temps de calcul raisonnable. Des tests numériques sur des modèles d'apprentissage statistique (par exemple le modèle des forêts aléatoires) appris sur des jeux de données publics sont également en cours de réalisation.

L'une des limites majeures de la méthode RWA-Shapley est l'orthogonalisation des entrées via la SVD, qui ne créent des variables Z indépendantes que dans certains cas précis (e.g., quand les entrées sont gaussiennes). De premiers tests numériques avec des lois d'entrées différentes montrent en effet que les indices Johnson-Shapley n'approximent plus les effets de Shapley correctement dans ces cas. Une transformation différente de la SVD est nécessaire pour obtenir un vecteur de variables indépendantes. Une autre perspective de recherche est l'utilisation d'un autre algorithme rapide pour estimer les effets de Shapley dans le cas de variables indépendantes. En effet, celui que l'on utilise néglige les effets des interactions entre entrées d'ordre supérieur ou égal à trois.

Bibliographie

Broto, B., Bachoc, F., and Depecker, M. (2020). Variance reduction for estimation of Shapley effects and adaptation to unknown input distribution, *SIAM/ASA Journal on Uncertainty Quantification*, 8:693–716.

Clouvel, L., Chabridon, V., Idrissi, M. I., Iooss, B., and Robin, F. (2023). Variance-based importance measures in the linear regression context: Review, new insights and applications. *Preprint*.

Clouvel, L., Mosca, P., Martinez, J., and Delipei, G. (2019). Shapley and Johnson values for sensitivity analysis of PWR power distribution in fast flux calculation. In *M&C 2019*, Portland, USA.

Da Veiga, S., Gamboa, F., Iooss, B., and Prieur, C. (2021). *Basics and Trends in Sensitivity Analysis. Theory and Practice in R.* SIAM.

Gilquin, L., Arnaud, E., Prieur, C., and Janon, A. (2019). Making the best use of permutations to compute sensitivity indices with replicated orthogonal arrays. *Reliability Engineering & System Safety*, 187:28–39.

Grömping, U. (2007). Estimators of relative importance in linear regression based on variance decomposition. *The American Statistician*, 61(2).

Iooss, B., Chabridon, V., and Thouvenot, V. (2022). Variance-based importance measures for machine learning model interpretability. In *Actes du 23ème Congrès de Maîtrise des Risques et de Sûreté de Fonctionnement ($\lambda\mu$23)*, Saclay, France.

Johnson, J. (2000). A heuristic method for estimating the relative weight of predictor

variables in multiple regression. *Multivariate Behavioral Research*, 35:1–19.

Johnson, J. and LeBreton, J. (2004). History and use of relative importance indices in organizational research. *Organizational Research Methods*, 7:238–257.

Johnson, R. (1966). The minimal transformation to orthonormality. *Psychometrika*, 31:61–66.

Lepore, A., Palumbo, B., and Poggi, J.-M., editors (2022). *Interpretability for Industry 4.0: Statistical and Machine Learning Approaches*. Springer.

Lindeman, R. H., Merenda, P. F., and Gold, R. Z. (1980). *Introduction to bivariate and multivariate analysis*. Scott Foresman and Company, Glenview, IL.

Nimon, K. and Oswald, F. (2013). Understanding the results of multiple linear regression: Beyond standardized regression coefficients. *Organizational Research Methods*, 16:650–674.

Owen, A. (2014). Sobol' indices and Shapley value. *SIAM/ASA Journal on Uncertainty Quantification*, 2:245–251.

Shapley, L. (1953). A value for n-persons game. In Kuhn, H. and Tucker, A., editors, *Contributions to the theory of games II, Annals of mathematic studies*. Princeton University Press, Princeton, NJ.

Sobol', I. (1993). Sensitivity estimates for non linear mathematical models. *Mathematical Modelling and Computational Experiments*, 1:407–414.