Le conflit "Entropie vs Variance" pour des familles de lois bivariées

Rémy Landri

Le conflit « Entropie *vs* Variance » pour des familles de lois bivariées

Rémy Landri

STID Institut Universitaire et Technologique
Domaine universitaire d'Auriac
Avenue Dr Suzanne Noël
11000 Carcassonne

Dans la réalité, on est souvent amené à prédire, anticiper, contrôler un phénomène engendré par un processus partiellement ou totalement inconnu. Dans ce qui suit, on suppose que la sortie de ce processus est une variable que l'on note y^*, dont le comportement est affecté par un groupe de variables $\mathbf{x}^* = (x_1^*, x_2^*, ...)$ qui est inconnu. Pour étudier ce comportement, on considère un modèle produisant en sortie une variable à expliquer y ayant vocation à être proche de la sortie réelle y^*. Ce modèle utilise des variables explicatives regroupées dans un vecteur $\mathbf{x} \equiv (x_1, x_2, ..., x_m)$. On veut que $\mathbf{x}$ soit aussi proche que possible de la vraie valeur $\mathbf{x}^*$. Certains de ces x_i peuvent correspondre à des x_i^*, d'autres non. Ce vecteur est lié à sa valeur de sortie y par une équation du type $y = f(\mathbf{x})$ complexe.

Dans un contexte environnemental, certains des x_i, ou tous, sont incertains et incorporent donc de la variabilité, c'est-à-dire une incertitude irréductible sur la valeur exacte de l'entrée. Tout devient plus compliqué puisqu'en affectant à ces variables aléatoires (*v.a.*) X_1, X_2, etc., des distributions de probabilité L_1, L_2, etc., la sortie Y devient elle-même aléatoire et l'incertitude augmente. Cette incertitude augmente de façon complexe avec la non-linéarité, la présence d'interactions, ou de fonctions qui s'emboîtent successivement l'une dans l'autre.

On s'interroge ici sur la fiabilité du vecteur $\mathbf{X}$. Travailler sur ce type de problème, traité jusqu'à présent de façon périphérique par les statisticiens, présente un intérêt certain et implique de comprendre quelles sont les variables importantes et donc la sensibilité de chacune dans le modèle.

L'analyse de sensibilité cherche à identifier les incertitudes « importantes » et donc à accroître l'information sur le modèle en quantifiant l'impact des X_i. Elle étudie comment l'incertitude dans la sortie Y d'un modèle peut être attribuée aux différentes sources d'incertitude à l'entrée du modèle [9].

De façon générale, pour mener une analyse de sensibilité, on considère la variance $\mathbb{V}(Y)$ de la sortie Y. La popularité de la variance présente un grand attrait pour les modélisateurs. Elle provient pour l'essentiel de sa simplicité qui lui confère un rôle central en statistique, à la fois comme une mesure de référence de la dispersion et comme mesure de la qualité de l'ajustement d'un modèle. Cette popularité s'explique aussi par le fait que la statistique a longtemps été exclusivement au monde gaussien, caractérisé par ses deux premiers moments. L'espérance, linéaire, et la variance, quadratique, débouchent aussi naturellement sur une géométrie euclidienne.

Bien qu'étant pour l'instant assez peu répandue [8], [3] dans le domaine de l'analyse de sensibilité, l'entropie de Shannon, définie comme :

$$\mathbb{H}(\mathbf{X}) = \left\{ \begin{array}{ll} -\sum_{\mathbf{X}} p_{\mathbf{X}}(\mathbf{x}) \log p_{\mathbf{X}}(\mathbf{x}) = \mathbb{E}_X\left[\log p_{\mathbf{X}}(\mathbf{x})\right] & \text{dans le cas discret} \\ -\int_{\mathbb{R}^n} f_{\mathbf{X}}(\mathbf{x}) \log f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} = \mathbb{E}_X\left[\log f_{\mathbf{X}}(\mathbf{x})\right] & \text{dans le cas continu} \end{array} \right\},$$

représente une mesure de l'incertitude et de la variabilité qui peut être une alternative à la variance en analyse de la sensibilité. Cette réflexion est motivée par le fait que l'entropie donne des indications sur la dispersion des *v.a.*, sans toutefois en apporter autant que la variance, tout comme la variance n'est pas une mesure bien adaptée, et donc critiquable, de l'incertitude. On peut donc chercher à explorer les relations et les différences qui existent entre elles.

On propose une comparaison « entropie *vs* variance », à différents niveaux, en termes de variation des paramètres des lois. Pour une loi univariée, la comparaison entre $\mathbb{H}(X)$ et *vs* $\mathbb{V}(X)$ est une comparaison scalaire. Ce problème a été traité par plusieurs auteurs. Maurin [5] s'est intéressé à des comparaisons entre les deux mesures pour les lois gamma et béta. Mukherjee et Ratnaparkhi [6] ont approfondi graphiquement les relations fonctionnelles entre les deux mesures pour différentes lois de probabilité. Ebrahimi, Maasoumi et Soofi [2] s'intéressèrent en particulier au comportement, en termes de croissance ou de décroissance : ce qu'ils appelèrent les ordres de $\mathbb{H}$ et $\mathbb{V}$, en fonction de variations des paramètres d'échelle, de position ou de forme des distributions. Après avoir discuté du cas univarié, on se concentre sur une variété de familles de distributions bivariées, qui inclue les copules, les lois normales, exponentielles, de Pareto, Gamma, de Dirichlet, etc.

Ce cadre bivarié est plus intéressant pour l'analyse de la sensibilité. On mènera l'étude comparative à partir de certains éléments clefs comme $\mathbb{E}_Y\mathbb{V}(\mathbf{X}|\mathbf{Y} = \mathbf{y}^*)$ pour l'analyse de sensibilité fondée sur la variance et $\mathbb{H}(\mathbf{X}|\mathbf{Y})$ pour celle fondée sur l'entropie.

Finalement, si on considère une loi bivariée $f_{X,Y}(x, y)$, on va donc comparer les différents niveaux de mesure :

1. les entropies et les variances marginales :
   * $\mathbb{H}(\mathbf{X})$ *vs* $\mathbb{V}(\mathbf{X})$,
   * $\mathbb{H}(\mathbf{Y})$ *vs* $\mathbb{V}(\mathbf{Y})$,
2. les entropies des lois conditionnelles et les variances conditionnelles :
   * $\mathbb{H}(\mathbf{X}|\mathbf{Y} = \mathbf{y}^*)$ *vs* $\mathbb{V}(\mathbf{X}|\mathbf{Y} = \mathbf{y}^*)$,
   * $\mathbb{H}(\mathbf{Y}|\mathbf{X} = \mathbf{x}^*)$ *vs* $\mathbb{V}(\mathbf{Y}|\mathbf{X} = \mathbf{x}^*)$,
3. les entropies conditionnelles avec les espérances des variances conditionnelles :
   * $\mathbb{E}_Y\mathbb{H}(\mathbf{X}|\mathbf{Y} = \mathbf{y}^*) \equiv \mathbb{H}(\mathbf{X}|\mathbf{Y})$ *vs* $\mathbb{E}_Y\mathbb{V}(\mathbf{X}|\mathbf{Y} = \mathbf{y}^*)$,
   * $\mathbb{H}(\mathbf{Y}|\mathbf{X})$ *vs* $\mathbb{E}_X\mathbb{V}(\mathbf{Y}|\mathbf{X} = \mathbf{x}^*)$,

   Et enfin, on doit comparer :
4. l'entropie jointe $\mathbb{H}(\mathbf{X}, \mathbf{Y})$ *vs* la matrice de variance de $(\mathbf{X}, \mathbf{Y})$.

Malheureusement, le problème posé est de savoir sur quelle base comparer $\mathbb{H}(\mathbf{X}, \mathbf{Y})$, qui est un scalaire, avec une matrice de variance. En l'état, cette comparaison des deux éléments, l'un scalaire et l'autre une matrice de dimension 2 dans le cas bivariée, est impossible. En revanche, on propose que cette comparaison entre $\mathbb{H}$ et $\mathbb{V}$ ait lieu à partir de certains pivots essentiels de $\mathbf{\Sigma}$, c'est-à-dire la variance totale : la trace $tr(\Sigma)$, la variance généralisée : le déterminant $|\mathbf{\Sigma}|$ et enfin les valeurs propres $\lambda_1(\mathbf{\Sigma})$ et $\lambda_2(\mathbf{\Sigma})$. Ces éléments de $\mathbf{\Sigma}$ permettent cette comparaison avec l'entropie bivariée, qui reste unidimensionnelle.

Enfin, nous développons l'analyse de la sensibilité fondée sur l'entropie et l'appliquons à des cas-types. Nous comparons les résultats avec l'analyse de la sensibilité fondée sur la variance.

Références

[1] Darbellay, G.A. et Vajda, I. (2000). Entropy expressions for multivariate continuous distributions. *IEEE Transactions on Information Theory*, **46** : 2, 709–712.

[2] Ebrahimi, N., Maasoumi, E. et Soofi, E.S. (1999). Ordering univariate distributions by entropy and variance. *Journal of Econometrics*, **90**, 317–336.

[3] Krzykacz-Haussmann, B. (2001). Epistemic sensitivity analysis based on the concept of entropy. *Proceedings of SAMO 2001*, eds. P. Prado and R. Bolado, Madrid, CIEMAT, 31–35.

[4] Landri, R., Ducharme, G. et Bonnard, R. (2007). Entropy-based sensitivity analysis in health risk assessment. *Fifth International Conference on Sensitivity Analysis of Model Output*, Eötös University, Budapest, Hungary, 18–22 June.

[5] Maurin, M. (1999). Jeux de variance et d'entropie. *XXXI Journées de Statistique de la SFDS*, 17-21 mai, Grenoble, France, 4p.

[6] Mukherjee, D., Ratnaparkhi, M.V. (1986). On the functional relationship between entropy and variance with related applications. *Communications in Statistics-Theory and Methods*, **15**, 291–311.

[7] Nadarajah, S., Zografos, K. (2005). Expressions for Rényi and Shannon entropies for bivariate distributions. *Informations Sciences*, **170**, 173–189.

[8] Sacks, J., Welch, W.J., Mitchell, T.J. et Wynn, H.P. (1989). Design and Analysis of Computer Experiments. *Statistical Science*, **4** : 4, 409–435.

[9] Saltelli, A. (2002). Sensitivity analysis for importance assessment. *Risk Analysis*, **22** : 3, 579–590.

[10] Shannon, C.E. (1948). A mathematical theory of communication. *Bell system technical journal*, **27**, 379–423 (I-II) and 623–656 (III-IV).