Sujet de thèse
Estimation non paramétrique de structure sparse de dépendance en grande dimension

23 février 2023

Sujet proposé par EDF - R&D - Département PERICLES

Encadrant EDF : Anne Dutfoy (anne.dutfoy@edf.fr)

Encadrants universitaires : J.F. Delmas (CERMICS, delmas@cermics.enpc.fr) et C. Butucea (ENSAE, cristina.butucea@ensae.fr)

1 Problématique : Estimation de copules

On considèrera des estimateurs non paramétriques de fonctions de répartition (ou densité) de copules. Les besoins décrits ci-dessous concernent un très grand nombre de projets au sein d'EDF. Notons que ces travaux font écho aux problématiques étudiées dans le cadre du partenariat OpenTURNS (www.openturns.org), qui propose notamment dans sa feuille de route, le développement d'indices de sensibilité différents de ceux reposant sur la décomposition de la variance (indices de Sobol). Les indices de Csiszar répondent à cet objectif.

De plus, le partenaire Airbus d'EDF au sien du partenariat OpenTURNS développe depuis plusieurs années en collaboration avec le LIP6 (Université Paris Sorbonne) une technique d'apprentissage de lois en grande dimension sur la base de graphes de réseaux bayésiens. L'établissement des tests d'indépendance conditionnelle simplifiera cet apprentissage en permettant l'exploitation de structures particulières de dépendance.

Par exemple, on pourra considérer en introduction un estimateur à partir de la copule de Bernstein qui est paramétrée par un nombre de points. Toutefois l'estimation optimale de ses paramètres est difficile et dépend de l'usage que l'on fait de l'etimateur, voir par exemple [1].

2 Besoins de techniques d'estimation de copules

L'estimation d'une copule permet de répondre aux besoins suivants.

2.1 Besoin principal : Calcul de divergences de Csiszar

Si P et Q sont deux mesures de probabilité absoluement continues par rapport à la mesure de Lebesgue, de densité p et q respectivement, et si f est une fonction convexe positive définie sur $\mathbb{R}^+$ telle que $f(1) = 0$,

alors on peut considérer la divergence suivante :

$$D_f(P||Q) = \int_\Omega f\left(\frac{p(x)}{q(x)}\right) q(x)dx \quad \in [0, +\infty]$$

Les auteurs de [3] et [4] ont défini un indice de sensibilité basé sur la divergence de Csiszar :

$$S = \mathbb{E}_X[D_f(p_Y \| p_{Y|X})] = D_f\left(p_{Y\otimes X} \| p_{(Y,X)}\right).$$

Il est facilde de voir que, avec Π la loi uniforme sur $[0,1]^2$:

$$S = \mathbb{E}_X[D_f(p_Y \| p_{Y|X})] = D_f\left(\Pi \| c_{(Y,X)}\right) \tag{1}$$

Cet indice de sensibilité basé sur la divergence de Csiszar ne repose donc pas sur la décomposition de la variance (part de la variance d'une variable due à la variance d'une autre variable). Il s'intéresse à l'influence de toute la loi d'une variable sur la loi d'une autre.

Si la fonction f est la fonction log, alors $D_f(P||Q)$ est l'entropie relative entre P et Q. En particulier, S dans (1) est l'entropie de la densité de copule c. Il existe des résultats sur l'estimation d'une entropie d'une densité de probabilité multivariée, e.g. [12] pour les lois discrètes et [11] pour les lois continues.

Une question est donc de généraliser ces approches à des fonctions f plus générales en vue d'améliorer les vitesses d'estimation des copules en présence de certaines structures de dépendance (par exemple quand des coordonées sont indépendantes conditionnellement aux autres).

Usage 1 : Sensibilité d'une variable scalaire Y à une variable scalaire X - Test d'indépendance - Hiérarchisation

On suppose que $Y = g(X)$, avec Y scalaire et X vectoriel. Le calcul des indices de Csiszar servent à plusieurs objectifs :

- Calculer la sensibilité de Y par rapport à X_i : définition d'un estimateur $\hat{S}_i$ de l'indice de Csiszar S_i et détermination de sa loi pour associer à chaque valeur un quantile.
- Réduire la dimension de X sur la base de tests d'indépendance entre Y et X_i : estimation de la loi de $\hat{S}_i$ pour voir si S_i est significativement différent de 0.
- Hiérarchiser les influences pour prioriser les actions de design : on suppose que $S_1 > S_2$. On dispose de n points et des estimations de $\hat{S}_1$ et $\hat{S}_2$. Quel est l'écart $S_1 - S_2$ minimum qui garantisse que $\mathbb{P}(\hat{S}_1 > \hat{S}_2) \geq 1 - \alpha$ pour un niveau de confiance $(1 - \alpha)$. Pour cela, il est nécessaire de disposer de la loi jointe de $(\hat{S}_1, \hat{S}_2)$.

Il est naturel de généraliser cette approche à plusieurs variables et de calculer la probabilité (test, niveau de confiance) pour que la hiérarchisation basée sur les estimateurs soit la vraie hiérarchisation.

Usage 2 : Tests d'indépendance conditionnelle (grande dimension et sparsité)

Dans notre contexte, l'indice de Csiszar permet de comparer les lois $\mathcal{L}_{(X,Y,Z)}$ et $\mathcal{L}_{(X,Z)\otimes(Y,Z)}$ où $\dim(X) = \dim(Y) = 1$ et $\dim(Z) = d$ (avec d grand). Remarquons que pour la loi $\mathcal{L}_{(X,Z)\otimes(Y,Z)}$, les variables aléatoires X et Y sont indépendant conditionellement à X.

L'objectif est de découvrir une structure de dépendance creuse de grande dimension (d grand) à partir d'un échantillon multivarié. Ce problème a été considéré par [13], voir aussi les références là-dedans, avec la séparation mesurée par la variation totale.

2.2 Autres besoins

Besoin 2 : Echantillonnage d'une loi multivariée. L'estimation de la fonction de répartition d'une copule de loi multivariée permet d'échantillonner cette loi et d'effectuer des transformations iso-probabilistes pour le calcul de probabilité de dépassement de seuil.

Besoin 3 : Détection d'outliers. L'objectif est de détecter dans un nuage de points multivariés, ceux qui sont associés à une vraisemblance très petite du modèle. Il est alors nécessaire d'estimer au mieux la densité de la copule.

Références

[1] *Multivariate Nonparametric Estimation of the Pickands Dependence Function using Bernstein Polynomials*, G. Marcon, S. A. Padoan, P. Naveau, P. Muliere, J. Segers, arXiv :1405.5228, 2014

[2] *Non parametric inference of dependence structures in high dimension with graphical models*, PH Wuillemin, Journée Utilisateurs OpenTURNS #11, www.openturns.org

[3] *Sensitivity analysis : A review of recent advances*, Borgonovo, Emanuele and Plischke, Elmar, European, vol 248 (3), pp Journal of Operational Research, 248 pp 869-887

[4] *Global Sensitivity Analysis with Dependence Measures*, Da Veiga, Sébastien, https ://hal.archives-ouvertes.fr/hal-00903283, 2013

[5] *Data-driven kernel representations for sampling with an unknown block dependence structure under correlation constraints*, G. Perrin, C Soize, N Ouhbi, Computational Statistics & Data Analysis, 119, 139-154

[6] *Efficient evaluation of reliability-oriented sensitivity indices*, G Defaux, G Perrin, Journal of Scientific Computing, 4, 2018

[7] *Data-driven kernel representations for sampling with an unknown block dependence*, G Perrin, C Soize, GDR MascotNum 2018

[8] *Moment-independent sensitivity analysis using copula*, Wei P, Lu Z, Song J., Risk Anal. 2014 Feb ;34(2) :210-22. doi : 10.1111/risa.12110. Epub 2013 Sep 11. PMID : 24024936.

[9] *Copula-based methods for global sensitivity analysis with correlated random variables and stochastic processes under incomplete probability information*, Shufang Song a, Zhiwei Bai a, Hongkui Wei b, Yingying Xiao b, Aerospace Science and Technology, Volume 129, October 2022, 107811

[10] *Estimation of Copulas via Maximum Mean Discrepancy*, P. Alquier, A.Derumigny, J.-D. Fermanian, Journal of the American Statistical Association, 2020.

[11] *Optimal rates of entropy estimation over Lipschitz balls*, Y. Han, J. Jiao, T. Weissman and Y. Wu, Ann. Statist. 48 (6) 3228 - 3250, 2020

[12] *Minimax rates of entropy estimation on large alphabets via best polynomial approximation*, Y. Wu and P. Yang, IEEE Transactions on Information Theory, 62(6) : 3702–3720, 2016.

[13] *Minimax optimal conditional independence testing*, M. Neykov, S. Balakrishnan and L. Wasserman, arXiv :2001.03039