Offre de stage en statistique

Université Rennes 2 / Laboratoire IRMAR

Présentation

Sujet. Propriétés des tests d'indépendance basés sur les mesures HSIC et les échantillons LHS.

Encadrant. Anouar Meynaoui (anouar.meynaoui@univ-rennes2.fr),
Institut de recherche mathématique de Rennes, Université Rennes 2.

Durée. 4 à 6 mois.

Description du Stage

Cadre. L'étude des tests d'indépendance entre deux variables aléatoires X et Y est un sujet fondamental en statistique. Plusieurs tests d'indépendance ont été proposés dans la littérature, et leurs propriétés théoriques ont été étudiées. En particulier, les tests basés sur les mesures HSIC (Hilbert-Schmidt Independence Criterion) (Gretton et al., 2005, 2007) ont progressivement gagné en popularité ces dernières années, et leurs propriétés théoriques ont été étudiées (Albert et al., 2022; Gretton et al., 2007). Ces tests présentent l'avantage de s'adapter à plusieurs types de données grâce à l'astuce du noyau et ont montré une bonne efficacité dans les applications pratiques. Tous ces travaux ont été développés sous l'hypothèse que l'échantillon observé $(X_i, Y_i), i = 1, \dots, n$ est indépendant et identiquement distribué. Cependant, dans certains domaines, cette hypothèse n'est pas toujours garantie, notamment dans le cas de l'analyse de sensibilité globale des codes numériques (Iooss and Marrel, 2017). Des plans d'expériences appelés *space-filling* sont généralement utilisés pour mieux explorer l'espace de variation des données et garantir de bonnes propriétés des estimateurs en grande dimension (Joseph, 2016). Cependant, l'utilisation de ces plans d'expérience ne garantit pas les propriétés théoriques habituelles des tests statistiques, notamment le contrôle de l'erreur de première espèce (El Amri and Marrel, 2022). Dans ce stage, on considère en particulier des plans d'expérience de type LHS (Latin Hypercube Sampling) (McKay et al., 2000).

Objectifs du stage. On s'intéresse au modèle suivant : $Y = F(X_1, \dots, X_p) = F(\mathbf{X})$ où $F : \mathbb{R}^p \to \mathbb{R}^q$ est une fonction déterministe, mais inconnue (représentant un simulateur numérique), et $X_1, \dots, X_p$ sont des variables aléatoires indépendantes et de même loi uniforme sur $[0, 1]$. On suppose qu'on dispose d'un échantillon $\mathcal{S}_n = (\mathbf{X}^{(i)}, Y^{(i)}), i = 1, \dots, n$, où les $\mathbf{X}^{(i)}$ sont générées suivant un plan LHS. On souhaite tester l'indépendance entre une variable X_j et la variable Y en utilisant des estimateurs des mesures HSIC. Les objectifs du stage sont les suivants :

1. S'approprier les outils statistiques nécessaires en étudiant la bibliographie.
2. Étudier les propriétés asymptotiques des estimateurs des HSIC sous plans LHS. Il est possible de s'inspirer des résultats de Loh (1996).
3. Construire un test de niveau asymptotique prescrit $\alpha \in (0, 1)$. En particulier, l'approximation Gamma de la loi asymptotique sous l'hypothèse nulle dans le cas i.i.d. (Gretton et al., 2007) est à étudier dans le cadre de l'échantillonnage LHS.

4. (Optionnel) Étudier l'extension du lemme de Romano and Wolf (Romano and Wolf, 2005) dans le cadre de l'échantillonnage LHS pour contrôler l'erreur de première espèce dans le régime non-asymptotique.

La bibliographie ci-dessous donne quelques éléments indicatifs.

Références

Albert, M., Laurent, B., Marrel, A., and Meynaoui, A. (2022). Adaptive test of independence based on hsic measures. *The Annals of Statistics*, 50(2) :858–879.

El Amri, M. R. and Marrel, A. (2022). Optimized hsic-based tests for sensitivity analysis : Application to thermalhydraulic simulation of accidental scenario on nuclear reactor. *Quality and Reliability Engineering International*, 38(3) :1386–1403.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer.

Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., and Smola, A. (2007). A kernel statistical test of independence. *Advances in neural information processing systems*, 20.

Iooss, B. and Marrel, A. (2017). An efficient methodology for the analysis and modeling of computer experiments with large number of inputs. *arXiv preprint arXiv :1704.07090*.

Joseph, V. R. (2016). Space-filling designs for computer experiments : A review. *Quality Engineering*, 28(1) :28–35.

Loh, W.-L. (1996). On latin hypercube sampling. *The annals of statistics*, 24(5) :2058–2080.

McKay, M. D., Beckman, R. J., and Conover, W. J. (2000). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 42(1) :55–61.

Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469) :94–108.