Karlsruher Institut für Technologie

Institut für Programmstrukturen und Datenorganisation (IPD)
Lehrstuhl für Systeme der Informationsverwaltung, Prof. Böhm

Masterarbeit

Canonical Monte Carlo Dependency Estimation

Masterarbeit

Estimating the dependencies between variables is a fundamental task in data analysis. It allows to understand data and to identify the variables required to answer specific questions. In recent work, we introduced a framework to estimate multivariate dependencies in large data sets, called Monte Carlo Dependency Estimation (MCDE) [1, 2]. MCDE quantifies the dependency within a set of variables $X = \{X_1, \dots X_n\}$ as the average discrepancy between their marginal and conditional distributions.

However, while MCDE also works for more than two variables, it does not address the case of Canonical Correlation Analysis (CCA). CCA is a fundamental problem in statistics [3], in which one is interested in quantifying the correlation between two vectors $X = \{X_1, \dots, X_n\}$ and $Y = \{Y_1, \dots, Y_m\}$ of random variables [4]. CCA has been applied in numerous fields, such as representation learning [5], causal discovery [6], or neuroscience [7]. However, it still challenging, and new approaches are being published every year. Notable examples are [8, 9, 10, 11, 12].

Extending MCDE to CCA raises several interesting questions. For example, MCDE bases on an adaptive grid scheme to estimate conditional distributions — how to adapt this for comparing two multivariate components (X and Y)? Similarly, how to quantify the discrepancies within a multivariate conditional distribution? At the same time, would such an adaptation inherit the original properties of MCDE?

The focus of this thesis is to extend the ideas behind MCDE to CCA. This results in the following tasks:

- Literature review of CCA and related methods.
- Understanding of the mechanisms involved behind MCDE [1, 2].
- Proposal and implementation of a new variant of MCDE for CCA.
- Evaluation of the developed approach(es) against existing methods[a].

To successfully conduct this thesis project, the student must possess:

- Knowledge of Python or Scala programming. Interest in statistics and correlation analysis.
- The ability to plan and work independently. A working knowledge of English.
- A high level of motivation, enthusiasm and curiosity.

Throughout this work, the student will acquire knowledge and practical experience in the domain of Correlation Analysis, and a good understanding the state of research in Data Science.

[1] E. Fouché and K. Böhm. “Monte Carlo Dependency Estimation”. In: *SSDBM*. Best Paper Award. ACM, 2019, pp. 13–24.
[2] E. Fouché et al. “A Framework for Dependency Estimation in Heterogeneous Data Streams”. In: *DAPD* (2020).
[3] H. Hotelling. “Relations Between Two Sets of Variates”. In: *Biometrika* 28.3/4 (1936), pp. 321–377.
[4] “Canonical Correlation Analysis”. In: *Applied Multivariate Statistical Analysis*. Berlin, Heidelberg, 2007, pp. 321–330.
[5] D. R. Hardoon et al. “Canonical Correlation Analysis: An Overview with Application to Learning Methods”. In: *Neural Comput.* 16.12 (2004), pp. 2639–2664.
[6] K. Zhang et al. “Kernel-based Conditional Independence Test and Application in Causal Discovery”. In: *UAI*. 2011.
[7] X. Zhuang et al. “A technical review of canonical correlation analysis for neuroscience applications”. In: *Human Brain Mapping* 41.13 (2020), pp. 3807–3833.
[8] F. R. Bach and M. I. Jordan. “Kernel Independent Component Analysis”. In: *J. Mach. Learn. Res.* 3 (2002), pp. 1–48.
[9] G. J. Székely and M. L. Rizzo. “Brownian Distance Covariance”. In: *Ann. Appl. Stat.* 3.4 (2009), pp. 1236–1265.
[10] A. Gretton et al. “A Kernel Two-Sample Test”. In: *J. Mach. Learn. Res.* 13 (2012), pp. 723–773.
[11] D. López-Paz et al. “The Randomized Dependence Coefficient”. In: *NIPS*. 2013, pp. 1–9.
[12] S. Romano et al. “The randomized information coefficient: assessing dependencies in noisy data”. In: *Mach. Learn.* 107.3 (2018), pp. 509–549.

[a] Examples of existing implementations: `https://pypi.org/project/dcor/`, `https://github.com/amber0309/HSIC`, `https://cran.r-project.org/web/packages/dHSIC/dHSIC.pdf`, `https://en.wikipedia.org/wiki/File:Correlation_examples2.svg`, `https://www.statsmodels.org/stable/stats.html#multivariate`

Ansprechpartner

Dr.-Ing. Edouard Fouché | edouard.fouche@kit.edu | Raum: 342
Am Fasanengarten 5 | 76131 Karlsruhe | Gebäude: 50.34