A model-free and dependence-based variable selection method for multi-outcome data

Sebastian Fuchs
Department for Artificial Intelligence and Human Interfaces
University of Salzburg, Austria
sebastian.fuchs@plus.ac.at

In regression analysis the main objective is to estimate the functional relationship between a set of $q \geq 1$ endogenous variables $\mathbf{Y} = (Y_1, \dots, Y_q)$ and a set of $p \geq 1$ exogenous variables $\mathbf{X} = (X_1, \dots, X_p)$. In view of constructing a good model, the question naturally arises to what extent $\mathbf{Y}$ can be predicted from the information provided by the multivariate exogenous variable $\mathbf{X}$, and which of the exogenous variables $X_1, \dots, X_p$ are relevant for the model at all.

We propose a direct and natural extension of Azadkia & Chatterjee's rank correlation T to a set of $q \geq 1$ endogenous variables. The novel measure T^q then quantifies the scale-invariant extent of functional dependence of the endogenous vector $\mathbf{Y}$ on the exogenous vector $\mathbf{X}$, characterizes independence of $\mathbf{X}$ and $\mathbf{Y}$ as well as perfect dependence of $\mathbf{Y}$ on $\mathbf{X}$ and hence fulfils all the desired characteristics of a measure of predictability. Aiming at maximum interpretability, we provide various general invariance and continuity conditions for T^q as well as novel ordering results for conditional distributions, revealing new insights into the nature of T.

Building upon the graph-based estimator for T, we present a non-parametric estimator for T^q that is strongly consistent in full generality, i.e., without any distributional assumptions. Based on this estimator we develop a model-free and dependence-based feature ranking and forward feature selection, called MFOCI, of data with multiple response variables, thus facilitating the selection of the most relevant explanatory variables. Several simulations as well as real-data examples for multi-response data illustrate T^q's broad applicability and the superior performance of MFOCI in comparison to existing procedures.

References

[1] Ansari, J. and S. Fuchs (2023). A simple extension of Azadkia & Chatterjee's rank correlation to a vector of endogenous variables. Available at arxiv.org/abs/2212.01621.

[2] Azadkia, M. and S. Chatterjee (2021). A simple measure of conditional dependence. *Ann. Stat. 49*(6), 3070–3102.