PARACELSUS MEDIZINISCHE UNIVERSITÄT · salzburgresearch · FH Salzburg · IDA LAB · PARIS LODRON UNIVERSITÄT SALZBURG

Data Science Doctoral Retreat 2024

03. Juli 2024, 10:00-16:30

PLUS, FB Informatik/AIHI, 5020 Salzburg, Jakob Haringer Straße 2, HS T01 (JAK2EG0.01)

Programm:

10:00 - 10:10	**Begrüßung**		
10:10 - 11:10	**Session 1**		
10:10 - 10:30	G. Schäfer	FHS	*Comparison of model predictive control and proximal policy optimization for a 1-dof helicopter system*
10:30 - 10:50	H. Waclawek	FHS	*Utilization of Polynomial Models in Reinforcement Learning*
10:50 - 11:10	S. Baron	PLUS	*Sleep staging and apnea detection using low-cost wearables and deep learning*
11:10 - 11:40	**Kaffeepause**		
11:40 - 12:40	**Session 2**		
11:40 - 12:00	R.H. Kutil B. Strasser-Kirchweger	PMU PLUS	*From narrative to logic: Structuring diagnostic criteria for accurate disorder differentiation*
12:00 - 12:20	A.Götz L. Maislinger	PLUS PLUS	*Fungus/algae associations in lichens along climatic gradients in Antarctica*
12:20 - 12:40	D. Radovanovic	FHS	*The dark side of smart metering: Addressing emerging privacy Attacks*
12:40 - 13:40	**Mittagspause**		
13:40 - 15:00	**Session 3**		
13:40 - 14:00	N. Dietrich	PLUS	*Estimating scale-invariant directed dependence of bivariate Distributions*
14:00 - 14:20	M. Kiefel	PLUS	*Nonparametric analysis of multivariate data in factorial designs with nondetects: A case study with microbiome data*
14:20 - 14:40	J. Beck	PLUS	*Nonparametric statistical inference for niche overlap*
14:40 - 15:00	Y. Wang	PLUS	*Hierarchical variable clustering based on the predictive strength between random vectors*
15:00 - 15:30	**Kaffeepause**		
15:30 - 16:30	**Session 4**		
15:30 - 15:50	F. Köhnke	PLUS	*Portfolio optimization using variable-size genetic network programming*
15:50 - 16:10	M. Uray	FHS	*Topological data analysis for smart manufacturing in industry 4.0*
16:10 - 16:30	P. Langthaler	PLUS	*quantifying and estimating dependence via sensitivity of conditional distributions*
17:00	**Gasthof Imlauer**		

Comparison of Model Predictive Control and Proximal Policy Optimization for a 1-DOF Helicopter System

Georg Schäfer
FH Salzburg - Josef Ressel Centre for Intelligent and Secure Industrial Automation

This study conducts a comparative analysis of Model Predictive Control (MPC) and Proximal Policy Optimization (PPO), a Deep Reinforcement Learning (DRL) algorithm, applied to a 1-Degree of Freedom (DOF) Quanser Aero 2 system. Classical control techniques such as MPC and Linear Quadratic Regulator (LQR) are widely used due to their theoretical foundation and practical effectiveness. However, with advancements in computational techniques and machine learning, DRL approaches like PPO have gained traction in solving optimal control problems through environment interaction. This paper systematically evaluates the dynamic response characteristics of PPO and MPC, comparing their performance, computational resource consumption, and implementation complexity. Experimental results show that while LQR achieves the best steady-state accuracy, PPO excels in rise-time and adaptability, making it a promising approach for applications requiring rapid response and adaptability. Additionally, we have established a baseline for future RL-related research on this specific testbed. We also discuss the strengths and limitations of each control strategy, providing recommendations for selecting appropriate controllers for real-world scenarios.

Utilization of Polynomial Models in Reinforcement Learning

Hannes Waclawek

Josef Ressel Centre for Intelligent and Secure Industrial Automation
Salzburg University of Applied Sciences, Austria
hannes.waclawek@fh-salzburg.ac.at

Due to its close ties to control theory, Reinforcement Learning (RL) is an interesting candidate for utility in Mechatronics, however, approximate RL solution methods mostly rely on Neural Networks (NNs) as approximators, which lack explainability. When training and utilizing NNs, we arguably pass control onto a black box and receive results without knowing the exact inner workings (exact model) of the network. In Mechatronics, however, non-transparency to this extent is not always a desired approach and often the resulting model description is of interest. There simply is a certain necessity for explainability, mainly rooted in the fact that movement of motors and connected kinematic chains like robotic arms affects its physical environment with the potential of causing harm, therefore calling for predictable movement. This, in turn, calls for a "predictable model", that allows to reliably derive physical properties like velocity, acceleration or jerk. This is why polynomial models are commonly utilized as approximators in this engineering discipline.

In the context of a doctoral research project, we investigate how optimizers of modern Machine Learning (ML) frameworks can be utilized directly, outside of the scope of NNs, in order to optimize polynomial models. This allows for an optimization using state of the art methods, while at the same time working with an explainable model. In our current research, we want to build on the foundation that our approach is compatible to other iterative ML algorithms and utilize it for approximate solution RL methods. Generalizing our model to multiple variables enables the representation of features, value functions as well as policies in an attempt of fostering utility of RL in Mechatronics. Our goal is to improve explainability of approximators as well as sample efficiency of RL methods.

In this talk, we provide a glimpse into the early stages of the first research direction we want to pursue in this regard: Where the approximate function is linear in the weights. In this special case of approximate RL solution methods, convergence guarantees exist for the prediction as well as control task, providing further incentive for appliance in Mechatronics contrary to approximate solution methods utilizing NNs. We show how our intended approach aligns with the state of the art in this regard and briefly introduce our previous work on ML-optimized polynomials. We then discuss the basic idea of how this approach can be generalized to feature construction in RL and how we intend to evaluate results in an experimental setup. We discuss the potential of this approach and welcome your insights to refine our direction.

Keywords: Machine Learning · Reinforcement Learning · Polynomials · Gradient Descent Optimizers · Mechatronics

Sleep staging and apnea detection
using low-cost wearables and deep learning

Sebastian Baron,

Department for Artificial Intelligence & Human Interfaces, University of Salzburg,
Jakob-Haringer-Strasse 2, 5020 Salzburg, Austria

Sebastian.Baron@plus.ac.at,

Sleep staging based on polysomnography (PSG) performed by human experts is the de facto "gold standard" for the objective measurement of sleep. PSG and manual sleep staging is, however, personnel-intensive and time-consuming and is thus impractical for monitoring a person's sleep architecture over extended periods of time. The same holds true for sleep apnea detection, which is also based on PSG-analysis and performed by human experts. Both, sleep architecture as well as sleep apnea play a crucial role for individual health and well-being, presenting the need for a less resource- and time-intensive method to monitor a person's sleep on a regular basis. Over the course of my doctoral research, I investigate in possibilities to perform these tasks with the help of low-cost wearable sensors and deep-learning algorithms.

The key measurement for both tasks are inter-beat-intervals (IBIs), the duration between two consecutive heartbeats. Commonly, in automated sleep-classification, a variety of so-called heart-rate-variability (HRV-) features is calculated based on this IBI-time-series, which are subsequently fed into a classification model. However, some of these features are quite resource-demanding to calculate and a variety of features are needed for accurate classification. A consumer-friendly application, however, is expected to perform this classification task in a short amount of time. In my work, both the features-engineering as well as the classification-part are therefore implemented using suitable neural-network architectures that ensure a sufficiently quick inference time.

My doctoral work is based on co-operations with the startup-company sleep^2 as well as the DGH (Das Gesundheitshaus), an established company that focuses on specialized mattresses. The sleep-staging algorithm I developed is featured in a mobile app created by sleep^2 and we are planning to introduce a similar feature for apnea detection within this app. The DGH developed an IMU-Senor that can be placed inside a mattress and we are are currently working on adapting the sleep-staging-algorithm to the recordings of this sensor, which involves the more difficult task of working with Ballistocardiography data instead of IBIs directly.

From Narrative to Logic: Structuring Diagnostic Criteria for Accurate Disorder Differentiation

Kutil Raoul
raoulhugo.kutil@plus.ac.at
PMU, PLUS-AIHI, IDA-Lab

Strasser-Kirchweger Barbara
barbara.strasser-kirchweger@plus.ac.at
PLUS-Psychologie, CCNS

Accurate diagnosis is crucial for the timely and effective treatment of disorders. The complexity of disorder definitions and the need to consider all possible manifestations during diagnosis add to this challenge. For instance, the first diagnostic criterion of Major Depressive Disorder (MDD) alone can have over 7,000,000 possible manifestations. This complexity is further compounded by the overlap between multiple disorder groups, making discrimination between disorders essential. Over years, established disorders have seen iterative refinement of diagnostic criteria, improving differentiation. However, there is no systematic representation explicitly stating successful discrimination between these disorders. The emergence of new clinical conditions, such as Long Covid, which presents with symptoms similar to depressive disorders and chronic fatigue syndrome, underscores the need for quick and reliable diagnostic pathways. In such cases, where iterative refinement is not yet available, it is particularly important to systematically compare the similarities and overlaps with established disorders.

Diagnostic manuals like the Diagnostic and Statistical Manual of Mental Disorders (DSM) encompass a vast amount of knowledge that could benefit from AI support to assist healthcare professionals. However, due to the necessity for 100 percent accuracy in medical decisions, current Large Language Models (LLMs) are insufficient for providing reliable diagnostic knowledge, given the complex combinations of rules described in narrative form. Therefore, a systematic, machine-actionable representation that can quantify the similarities and overlaps of established disorders is needed. This would facilitate accurate and reliable diagnosis, especially in the context of emerging disorders.

Currently, diagnostic criteria are predominantly available in narrative form, which limits their utility for AI applications. The aim here is to present a systematic representation of diagnostic criteria using logical formalisms and to quantify the similarities and overlaps of disorders. By transforming narrative diagnostic criteria into a structured, logical format, we can leverage computational methods to enhance diagnostic accuracy. This approach not only aids in the discrimination of established disorders but also provides a robust framework for the integration of new clinical conditions into existing diagnostic paradigms. Ultimately, this systematic representation will support healthcare professionals in making more accurate diagnoses and improve patient outcomes through timely and appropriate treatment interventions.

FUNGUS/ALGAE ASSOCIATIONS IN LICHENS ALONG CLIMATIC GRADIENTS IN ANTARCTICA

Götz Anna[1a], Maislinger Lea[2a], Wolfgang Trutschnig[2b], Ulrike Ruprecht[1b]

[1]Dept. of Environment and Biodiversity; PLUS
[1a] anna.goetz@plus.ac.at; [1b] ulrike.ruprecht@plus.ac.at

[2]Dept. of Artificial Intelligence and Human Interfaces; PLUS
[2a] lea.maislinger@plus.ac.at; [2b] wolfgang.trutschnig@plus.ac.at

Polar cold deserts are climatically extreme and hostile ecosystems. Due to global warming, specialised, cold adapted species are shifting to the poles with increasing risk of extinction. Lecideoid lichens as dominant vegetation-forming organisms in the climatically harsh areas of the Antarctic continent show clear preferences in relation to environmental conditions caused by the macroclimate. 434 lichen samples collected between 62°S and 86°S latitude, covering all relevant ice-free areas with regard to macro-climate, were processed in this study.

We aim to investigate to which extent the composition as well as the associations of the two dominant symbiotic partners (fungus/algae) were set in context with environmental conditions along an environmental gradient to calculate the potential niche of the different species/operational taxonomical units (OTUs) of both symbionts over the continent.

Our results revealed the specificity of fungal species towards their algae decreased under more severe climate conditions. For instance, the generalist species *Lecidea cancriformis* in Antarctica is associated with all available algal OTUs specialized to different climatic conditions and can therefore survive under a large range of environmental conditions, having a broad environmental niche. In contrast, the species *Rhizoplaca macleanii* is only associated with one algal-OTU and is therefore restricted to areas with an intermediate climate, which means it has a small environmental niche.

To summarize, macroclimate is considered to be the main driver of species distribution, making certain species useful bioindicators of climate conditions and thus for assessing the impact of climate change.

The Dark Side of Smart Metering: Addressing Emerging Privacy Attacks

Dejan Radovanovic

dejan.radovanovic@fh-salzburg.ac.at

Center for Secure Energy Informatics at the Salzburg University of Applied Sciences, Salzburg, Austria

As smart meters revolutionize energy management with their promise of efficiency and real-time insights, they also unveil significant privacy concerns. This dissertation talk delves into the critical intersection of technological advancement and personal privacy, exposing the hidden risks within load data analysis. It explores how sophisticated machine learning and deep learning techniques can inadvertently strip away the anonymity of consumers, revealing detailed personal habits and socio-demographic characteristics. For instance, Figure 1 displays the energy consumption fingerprint of a household from May to September, illustrating distinct patterns over daily periods. By investigating novel privacy attacks and evaluating existing mitigation strategies, this research aims to strike a delicate balance between utility and confidentiality. Through the application of advanced anonymization methods and existing privacy-preserving techniques, this research aims to preserve consumer data while retaining the essential benefits of smart metering. It investigates strategies to address the complexities of privacy in the digital age, ensuring that smart technology enhances functionality without compromising privacy.

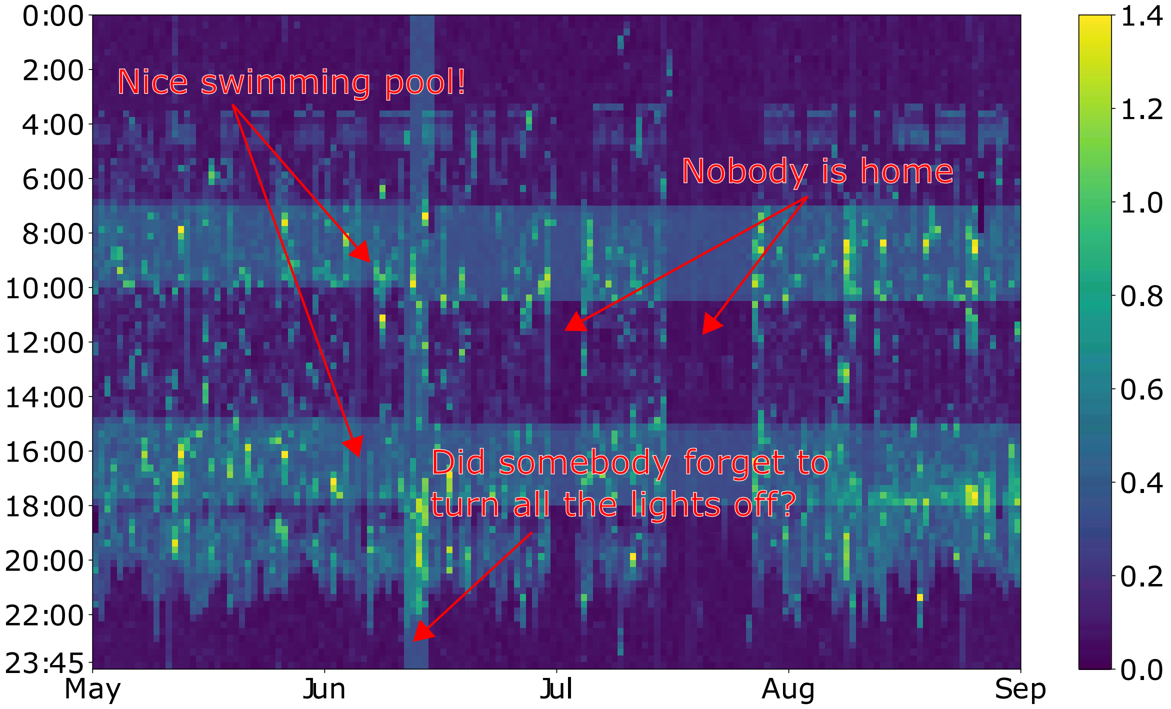

Figure 1: Load heat-map of one household depicting energy consumption patterns from May to September across daily hours (measurements are in kWh), uncovering potential personal habits and socio-demographic characteristics.

Estimating scale-invariant directed dependence of bivariate distributions

Nicolas Pascal Dietrich[a,*], Robert R. Junker[b], Florian Griessenberger[a], Wolfgang Trutschnig[a]

[a] *University of Salzburg, Department for Artificial Intelligence and Human Interfaces, Hellbrunnerstrasse 34, Salzburg, 5020, Salzburg, Austria*

[b] *Philipps-University Marburg, Department for Evolutionary Ecology of Plants, Karl-von-Frisch-Strasse 8, Marburg, 35043, Hessen, Germany*

Abstract

Asymmetry of dependence is an inherent property of bivariate probability distributions. Being symmetric, commonly used dependence measures such as Pearsons r or Spearmans ρ mask asymmetry and implicitly assume that a random variable Y is equally dependent on a random variable X as vice versa. A copula-based, hence scale-invariant dependence measure called ζ_1 overcoming the just mentioned problem was introduced in 2011. ζ_1 attains values in $[0, 1]$, it is 0 if, and only if X and Y are independent, and 1 if, and only if Y is a measurable function of X. Working with so-called empirical checkerboard copulas allows to construct an estimator ζ_1^n for ζ_1 which is strongly consistent in full generality, i.e., without any smoothness assumptions on the underlying copula. The *R*-package qad (short for quantification of asymmetric dependence) containing the estimator ζ_1^n is used both, to perform a simulation study illustrating the small sample performance of the estimator as well as to estimate the directed dependence between some global climate variables as well as between world development indicators.

Keywords: asymmetry, copula, correlation, dependence, direction, invariance

[*] Presenting author

Email addresses: `nicolaspascal.dietrich@plus.ac.at` (Nicolas Pascal Dietrich), `robert.junker@uni-marburg.de` (Robert R. Junker), `florian.griessenberger@plus.ac.at` (Florian Griessenberger), `wolfgang.trutschnig@plus.ac.at` (Wolfgang Trutschnig)

Nonparametric Analysis of Multivariate Data in Factorial Designs with Nondetects: a Case Study with Microbiome Data

Maximilian Kiefel[1] and Johanna Freidl[2]

ABSTRACT

The term *nondetects* describes observations that are not fully observed because the true value is below a detection threshold, and can therefore not be precisely detected. One may also consider them a special case of left-censored data.

Nondetects occur frequently, for instance, in life sciences research in medicine or microbiology.

We examine the use of nonparametric inference methods for multivariate data in factorial designs in situations where nondetects are present. In addition to a simulation study, we apply them to a real example from a recent study in the field of microbiology about farmed alps in Austria and their effects on human health.

The statistical focus is on testing hypotheses regarding factor effects on the microbiomial composition due to the locations where the samples were taken.

The nonparametric centerpiece of the methodology is assuming the nonparametric relative effect (probabilistic index) and its generalizations as the functional on which inference is built, along with the respective invariance properties of the resulting tests.

On this basis, we apply and evaluate recently proposed nonparametric analogs to some types of multivariate test statistics and its implementations in R, namely the
(1) Wald-type statistic (WTS), (2) ANOVA-type statistic (ATS), (3) Lawley-Hotelling trace, (4) Wilks Lambda (Likelihood ratio), (5) Bartlett-Nanda-Pillai trace.

Except for the WTS, all the mentioned methods are available through the R-package **nparMD**.

We demonstrate that the proposed methods can handle commonly occurring rates of nondetects without substantial impairment of specificity and sensitivity.

[1] Paris Lodron Universität Salzburg
maximilian.kiefel@plus.ac.at

[2] Paracelcus Medizinische Privatuniversität (PMU)
johanna.freidl@pmu.ac.at

Nonparametric Statistical Inference for Niche Overlap

Jonas Beck[1]

[1] *Paris Lodron University of Salzburg, Department of Artificial Intelligence and Human Interfaces, Austria*

Abstract

The understanding of species interactions and ecosystem dynamics hinges upon the study of ecological niches. Quantifying the overlap of Hutchinsonian-Niches has garnered significant attention, with many recent publications addressing the issue. Prior work on estimating niche overlap often often did not provide confidence intervals or assumed multivariate normality, seriously limiting applications in ecology and biodiversity research.

In this talk, we develop a fully nonparametric approach to statistical inference for the Overlap of Niches. Additionally we will discuss certain extensions, especially the challenges in the multiple species setup. The novel methodology is then applied to a study comparing the ecological niches of the Eurasian eagle owl, common buzzard, and red kite. These species share a habitat in Central Europe but exhibit distinct population trends. The analysis explores their breeding habitat preferences, considering the intricate competition dynamics and utilizing the nonparametric approach to niche overlap estimation. Our proposed method provides a valuable inferential tool for the quantitative evaluation of differences and overlap between niches.

Hierarchical variable clustering based on the predictive strength between random vectors

Sebastian Fuchs, Yuping Wang

Abstract

A rank-invariant clustering of variables is introduced that is based on the predictive strength between groups of variables, i.e., two groups are assigned a high similarity if the variables in the first group contain high predictive information about the behaviour of the variables in the other group and/or vice versa. The method presented here is model-free, dependence-based and does not require any distributional assumptions. Various general invariance and continuity properties are investigated, with special attention to those that are beneficial for the agglomerative hierarchical clustering procedure. A fully non-parametric estimator is considered whose excellent performance is demonstrated in several simulation studies and by means of real-data examples.

Portfolio Optimization using Variable-Size Genetic Network Programming

Fabian Köhnke[*]
Department of Artificial Intelligence
University of Salzburg
5020 Salzburg, Austria
fabian.koehnke@plus.ac.at

Christian Borgelt[†]
Department of Artificial Intelligence
University of Salzburg
5020 Salzburg, Austria
christian.borgelt@plus.ac.at

Abstract

We present an extension of a graph-based evolutionary algorithm called Genetic Network Programming (GNP) by a novel mutation operator, which allows for a variable number of nodes and edges per individual. With this operator, the search space is significantly extended, but without risk of incurring the bloat problem. Due to higher flexibility, it is now possible for GNP to automatically adapt to the complexity of a given task and to find suitable features, especially for high dimensional data sets. We applied our mutation operator successfully in a GNP for a financial data set where it improved over standard GNP and showed better performance concerning portfolio optimization in six out of seven simulations.

[*] https://fabiankoehnke.github.io/gnp/hp/

[†] https://borgelt.net/

Preprint. Under review.

Topological Data Analysis for Smart Manufacturing in Industry 4.0

Martin Uray
martin.uray@fh-salzburg.ac.at

Josef Ressel Centre for Intelligent and Secure Industrial Automation
Salzburg University of Applied Sciences, Salzburg, Austria

In this talk, we outline a doctoral research project focused on applying methods from Topological Data Analysis (TDA) to the domain of *Industry 4.0*.

Industry 4.0 represents a fundamental shift in how global production and supply chains are organized, aiming for more flexible, efficient, autonomous, and decentralized processes. To fully leverage its benefits, several challenges must be addressed. These challenges include integrating existing, partially mature hardware and machines, ensuring production process security, and managing high-dimensional data.

Various approaches in the literature employ different methods from Artificial Intelligence to address these challenges. However, these methods ignore the underlying topological and geometrical structure of the data. To leverage this information, methods from TDA are particularly compelling in this context. By applying techniques from algebraic topology, we can extract structures such as connected components, loops, and cavities from the data, providing a deeper description of its underlying patterns.

Specifically, we will explore the application of TDA to analyze and model time series production data. Our goals include the detection of abnormal behaviour within the temporal observation of the production process, which can support applications such as predictive maintenance or the detection of malicious behavior in OT security. Additionally, we aim to develop a framework capable of modeling the production process to generate synthetic data for various tasks.

Quantifying and estimating dependence via sensitivity of conditional distributions

Jonathan Ansari[1,a] , Patrick B. Langthaler[1,b] , Sebastian Fuchs[1,c], and Wolfgang Trutschnig[1,d]

[1] Department of Artificial Intelligence and Human Interfaces, University of Salzburg, Austria
[a] jonathan.ansari@plus.ac.at
[b] patrick.langthaler@plus.ac.at
[c] sebastian.fuchs@plus.ac.at
[d] wolfgang@trutschnig.net

Abstract
Recently established, directed dependence measures for pairs (X, Y) of random variables build upon the natural idea of comparing the conditional distributions of Y given X = x with the marginal distribution of Y . They assign pairs (X, Y) values in [0, 1], the value is 0 if and only if X, Y are independent, and it is 1 exclusively for Y being a function of X. We show that comparing randomly drawn conditional distributions with each other instead or, equivalently, analyzing how sensitive the conditional distribution of Y given X = x is on x, opens the door to constructing novel families of dependence measures Λ_φ induced by general convex functions $\varphi : R \to R$, containing, e.g., Chatterjee's coefficient of correlation as special case. After establishing additional useful properties of Λ_φ we focus on continuous (X, Y), translate Λ_φ to the copula setting, consider the L^p -version and establish an estimator which is strongly consistent in full generality. A real data example and a simulation study illustrate the chosen approach and the performance of the estimator.

Keywords: dependence measure, sensitivity, conditional distribution, Chatterjee's correlation coefficient, explainability, copula