The European Conference on Data Analysis (ECDA)

Book of Abstracts

2024 Gdansk (Poland)

Table of contents

Keynote Speakers

Evrim Acar Ataman, Extracting Insights from Complex Data: (Coupled) Tensor Factorizations & Applications

There is an emerging need to jointly analyze data sets collected from different sources in order to extract insights about complex systems such as the human brain or human metabolome. For instance, joint analysis of omics data (e.g., metabolomics, microbiome, genomics) holds the promise to improve our understanding of the human metabolism and facilitate precision health. Such data sets are heterogeneous – they are a collection of static and dynamic data sets. Dynamic data can often be arranged as a higher-order tensor (e.g., subjects by metabolites by time) while static data can be represented as a matrix. Tensor factorizations have been successfully used to reveal the underlying patterns in higher-order tensors, and extended to joint analysis of multimodal data through coupled matrix/tensor factorizations (CMTF). However, joint analysis of heterogeneous data sets still has many challenges, especially when the goal is to capture the underlying patterns. In this talk, we discuss CMTF models for temporal and multimodal data mining. We focus on a flexible, accurate and computationally efficient modelling and algorithmic framework that facilitates the use of a variety of constraints, loss functions and couplings with linear transformations when fitting CMTF models. Through various applications, we discuss the advantages and limitations of available CMTF methods.

Bettina Grün, Clustering Data Using Bayesian Mixture Models

Cluster analysis aims a grouping objects and is a main task in exploratory data analysis, statistical data analysis and machine learning. Model-based approaches have the advantage that model specification and selection are performed within a principled statistical framework, facilitating interpretation, improving validation and including uncertainty quantification. Pursuing a Bayesian approach allows the specification of suitable priors which can include a-priori knowledge about the cluster structure to be detected as well as regularizing the likelihood.

We will give an overview on recent advances in Bayesian model-based clustering, including prior specifications as well as computational inference tools. Suitable prior specifications need to enable the selection of the number of clusters in the data set as well as appropriate cluster distribution approximation and potentially also variable selection.

Inference methods need to cover approximation methods for the posterior such as MCMC schemes, but also post-processing methods for model selection and model identification.

Paweł Lula, Topic identification in analysis of scientific productivity – models, methods, and tools

The identification, modelling and evaluation of research trends should be considered as a very important part of the analysis of scientific achievements. The results of the analysis of research topics are indispensable in: evaluating the course of scientific development, analyzing leading scientific centers, designing and monitoring the implementation of scientific development policies, predicting the most promising directions of scientific developments, managing research units, and assessing interdisciplinarity in research.

In identifying research trends, the most important source of data is textual information in the form of scientific publications or their abstracts. During the initial stage of analysis, it is necessary to choose the appropriate method of representing textual information (frequency matrix, word sequences, embeddings). Then choose the right approach to the model building process (supervised, unsupervised). The next step involves the process of building and evaluating the quality of the model. A positive evaluation of the model justifies its implementation.

The objectives of this presentation are: to present the essential methods of identifying research topics, to conduct an evaluation of discussed algorithms, and to present software tools to implement the topic identification process.

The presentation will include the results of research on the development of research trends in Poland in the area of social sciences.

Line Clemmensen, Machine learning in psychiatry with distribution shifts, fairness and explainability in mind

Machine learning (ML) finds many applications within psychiatry using multiple modalities like speech, video, biosensors, and medical health records. ML models developed on open source large datasets can be challenged by distribution shifts and a lack of explainability or fairness for minority groups. We will dive into some of the challenges and look at ways of addressing these concerns going from data representativity and fairness to explainability of models.

Speakers in alphabetical order

Aurore Archimbaud, Robust matrix completion for rating-scale data

We introduce a new Low-rank matrix completion (LRMC) algorithm designed specifically for discrete rating-scale data and robust to the presence of corrupted observations. Furthermore, we evaluate the performance of our proposed method and several competing approaches through simulation studies using discrete rating-scale data instead of continuous data, considering various attack scenarios. I am a speaker of the “Advances in robust methodologies” section.

Daniel Baier, Measuring Technology Acceptance over Time by Online Customer Reviews Based Transfer Learning

Online Customer Reviews (OCRs) are user-generated semi-formal evaluations of objects (brands, products). They typically consist of a time stamp, a star rating, and – in many cases – a detailing comment. Many methodological approaches have been developed and applied to analyze and aggregate OCRs as well as to improve products and services based on this knowledge. In this paper, we present a new transformers based approach for the same purpose.

Marco Berrettini, Mean-restricted Matrix-variate Normals with an application to clustering

This work introduces different novel approaches to model mean structures in matrix-variate normals, addressing the over-parameterization issue commonly encountered in model-based clustering. The methodology employs parsimonious parameterization to reveal the multivariate interdependence underlying the datau0027s mean structure. An Expectation-Maximization (EM) algorithm is developed for model estimation. Simulations and real-world examples show significant improvements over traditional methods.

Eva Boj, Aurea Grané, Agustín Mayo-Íscar, Robust distance-based generalized linear models: A new tool for classification

Distance-based generalized linear models are prediction tools which can be applied to any kind of data whenever a distance measure can be computed among units. In this work, robust ad-hoc metrics are proposed to be used in the predictors’ space of these models, incorporating more flexibility to this tool. Their performance is evaluated by means of a simulation study and several applications on real data are provided. Computations are made using the dbstats package for R.

Agnieszka Brelik, Katarzyna Cheba, Arkadiusz Malkowski, Michał Pietrzak, Magdalena Olczyk, Mapping Sustainable Transformation in Co-creation Perspective: Applying Multidimensional Comparative Methods in Systematic Literature Review

In recent years much attention has been paid to the different types of tools and methods used in bibliographic research. These methods are also used in the mapping of new technologies. The paper presents examples of the use of selected methods of multivariate statistical analysis to identify new areas of research. Publications in the field of sustainable transformation that incorporate a co-creation perspective will be selected for analysis.

Zino Brystowski, Multi-view stacking for theory development

Theories are essential in research for understanding phenomena. Using predictive accuracy as an alternative criterion to model fit can improve robustness in theory development. Here we introduce a framework for theory development in multidisciplinary research based on predictive modeling. Base models representing different theories are integrated into a meta-model that uses sparse penalized regression. We demonstrate this method with simulated data and discuss future directions.

Justyna Brzezińska, Measurement of financial literacy using IRT models

Financlial literacy is understood as a dombination of awareness, knowledge, skills, attitudes and behaviors necessary to make sound financial decisions and ultimately achieve individual financial prosperity.In this paper we present the theory, role and components of financial literacy, as well as results of own study on financial literacy in Poland using item response theory models. We also compare alpha`s reliability coefficient with McDonalds Omega coefficient and present results using graphs.

Katarzyna Cheba, Michał Pietrzak, Fuzzy methods and their impact on the results of socio-economic studies

Models of socio-economic behaviour are based on the relationships between different variables. As the complexity of the research problem increases, it is difficult to establish these relationships and make properly decisions. In this case the knowledge of experts and fuzzy methods can be helpful. In this paper, the authors present examples of the application of fuzzy methods in socio-economic research and carry out a sensitivity analysis of the results obtained depending on the method chosen.

Luca Coraggio, Quadratic discriminant score for selecting number of clusters, clustering models and algorithms.

We address the problem of selecting optimal clustering solutions by developing two cluster-quality criteria based on the quadratic discriminant score, consistent with clusters from a broad class of elliptic–symmetric distributions. We review this methodology, validated through extensive experimental analysis, and introduce companion software for efficient estimation of the proposed (bootstrap) quadratic scores. --- Invited speaker of the Advances in robust methodologies section.

Anna Denkowska, Deep Neural Networks in the Modeling of the Dependence Structure in Risk Aggregation

European Insurers calculate the Solvency Capital Requirement using the Standard Formula given by the Directive (2016) or internal models developed taking into account their business profiles. Our research indicates that it is this model, based on DNN used to estimate the multivariate distribution, that gives the Diversification Effect estimate at the right level.

Andrzej Dudek Symbolic Data Processing with Deep Neural Networks Autoencoders,

The presentation aims to bridge the gap between symbolic data analysis and deep learning, providing new insights into how complex data structures can be represented and learned by deep models, expanding the applicability of autoencoders in domains where symbolic data is prevalent.Symbolic data, representing complex and structured objects, such as intervals, distributions, and multi-valued categories, poses unique challenges for traditional machine learning models. Autoencoders, known for their ability to learn compressed and meaningful representations of high-dimensional data, offer a promising solution to these challenges. The presentation explores how deep neural network autoencoders can be adapted to efficiently handle symbolic data. It addresses various architectures and modifications required to process and extract features from symbolic data, enabling tasks such as clustering, classification, and anomaly detection.

Andreas Geyer-Schulz, Meta-Genetic Algorithms with the xega R-Package

In this contribution we give a tutorial presentation of implementing and running meta-genetic algorithms based on the xega R-package (https://CRAN.R-project.org/package=xega). With xega, computational experiments for finding hyperparameters of genetic and evolutionary algorithms can be represented by several types of genes (e.g. vectors of reals, derivation trees). Performance results of several experiments will be reported across platforms.

Aurea Grané, Silvia Salini, Gabriele Infante, A new distance for categorical data with moderate association

Categorical variables coming from surveys usually share high percentage of information. This redundancy may lead to misleading results in data visualization and clustering. In this work we propose a new distance for categorical data, able to take into account the association/correlation structure of the data. Its performance is evaluated and compared to Hamming distance via a simulation study. The methodology is illustrated on a novel dataset on co-creation antecedents of telemedicine.

Francesca Greselin ‘Operational Risk mitigation: utilizing AI and Social Media for Early Event Detection’

Operational risk (OpRisk) is a crucial non-financial issue for financial institutions, traditionally focused on data collection, capital requirements, and reporting. Recently, OpRisk functions have shifted to proactive strategies using AI for deeper data insights. This study advances the use of text analysis and topic modeling to identify root causes of OpRisk. It enhances data sources using tweets to identify early-stage risks, integrating these methods into a holistic management approach.

Ana Keney, Simultaneous Feature Selection and Outlier Detection Using Mixed-Integer Programming Under Varying Data Structures

Biomedical research is increasingly data rich, with studies comprising growing numbers of features. The larger a study, the higher likelihood that a substantial portion of features may be redundant and/or contain contamination (outlying values). Furthermore, these may be complex in nature with inherent dependence structure (e.g., longitudinal). We develop a flexible framework using mixed-integer programming for simultaneous feature selection & outlier detection accounting for data structure.

Bartosz Kocot, Paweł Krawczyński ‘Identification of factors influencing customer choice and attrition on streaming platforms’

A streaming platform is an online service that allows many users to watch movies, TV series, or TV shows online. This service involves real-time data transmission to different devices. In the context of many different providers, it is crucial to understand the factors influencing the choice of a particular platform, as well as the factors determining the risk of customer churn. The article uses symbolic decision trees to identify key factors in both cases.

Arkadiusz Kozłowski, Accuracy of Complex Estimation Based on Nonprobability Samples in a Social Survey ? Simulation Based on EU-SILC Microdata

The aim of the study is to assess the accuracy of a hypothetical sampling strategy, in which the basis for inference is a nonprobability sample, and the estimation is strengthened by additional information on auxiliary variables from an independent probability sample and the entire population, and the subject of the study is the parameters of household income distribution. The study is based on a computer simulation on a set of unit data from the EU-SILC survey carried out in Poland in 2022.

Zuzanna Krysiak, Demographic effects of population migration for selected countries

The research examined the impact of climate migration on the demographic structure of societies, analyzing real data from Germany, Canada and Japan. GIS techniques were used to analyze spatial migration patterns and their impact on demography by examining the labor market, age distribution of the population, education level and temperature changes in selected countries over the years. Studies of temperature changes enable the analysis of the phenomenon of climate migration and their scale.

Ludwig Maximilian Lausser, Vindicating Ordinal Relations

We discuss vindicating ordinal relations in experiments with ordinal classifier cascades. Those constraint classification models can only separate classes if their underlying feature representation fulfills predefined ordering constraints. Any violation of the constraints leads to severely decreased class-wise sensitivities. Uniformly high sensitivities provide evidence for the existence of an assumed ordinal relation. Our approach allows for exhaustive screens for ordinal relations of classes.

Karolina Lewandowska-Gwarda ‘Who is in pole position on the labor market in Poland? Evaluation of spatial diversification of men and women situation on the labor market in Poland (2019-2022)’

The main aim of the study was to answer the question, who has a better position on the labor market in Poland - men or women? Does the spatial dimension, the place where we live, matter in this case? The study use a taxonomic measure of development to evaluate and compare the situation of men and women on local labor markets in Poland, in 2019-2022. GIS and ESDA methods were used for visualization and statistical evaluation of the obtained results.

Karsten Lübke, Florian Seliger ‘Failures-to-Deliver on New York Stock Exchange: A Forecasting Approach utilizing Tree-Boosting Modeling’

Failures-to-Deliver (FTDs) occur every day in US markets. As equity FTDs leave behind phantom shares, regulatory authorities consider this as a systemic risk. However, only a portion of delivery failures is reported with a considerable delay. This paper aims to develop a forecasting model for FTDs by utilizing Gaussian process boosting, consisting of a tree-ensemble fixed-effects function and temporal random effects estimated by separate Gaussian processes.

Giancarlo Manzi, Qi Guo, Aurea Grané, Marco Zanotti, Transforming social media data to survey data through a chatbot-based approach: A case on the elderly well-being

Transforming social media data to survey data through a chatbot-based approach: A case on the elderly well-being Giancarlo Manzi, University of Milan, Italy Qi Guo, University Carlos III, Madrid, Spain Aurea Grané, University Carlos III, Madrid, Spain Marco Zanotti, University of Milan-Bicocca, Italy Abstract. We present an original approach to transform social media in survey data with the use of chatbot technologies. ChatGPT APIs are used to associate each X message we collected about elderly people condition in Spain to the most probable question which hypothetically could have been generated such message among a grid of few questions, possibly representing a survey questionnaire. We also asked ChatGPT to give, according to the meaning of each X message, a possible answer option, either a "yes/no" answer or a rating answer. In this way we form a survey-like dataset about the elderly people in Spain.

Małgorzata Markowska, Andrzej Sokołowski ‘Graphical Illustration of 2xk Contingency Table and Post-hoc Ad Hoc Inference’

Graphical illustration method for 2xk contingency table is proposed in the paper. It is suitable for table with just two rows and at least three columns. Rows are represented as two axes on the coordinate system. Each column is represented by the circle with the middle defined by row frequencies (or percentages), and the radius equal to the half of confidence interval $1{,}96\sqrt{((w(1-w))/n)}$. Disjoint cicles show columns with statistically different structure of row answers.

Stefan Mathes, Dynamic pricing model for hotel & tourism revenue management systems

Modern revenue management systems can optimize revenues but often rely on outdated assumptions of full occupancy and fixed room quotas. In contrast, we forecast actual occupancy and build the RMS based on these predictions. We are analyzing

forecasting models that account for profound seasonality in tourism. Contrary to our hypothesis, our initial results indicate that, when comparing less complex SARIMA with complex Random Forest models, Random Forest, after tuning, achieves superior accuracy.

Paweł Miłobędzki, Sabina Nowak, The components of Bitcoin?s bid-ask spread. Does the change in tick size matter?

We disentangle the bid-ask spread of Bitcoin traded at Bitstamp into the private information, buy-sell imbalances and price clustering components. We use the transaction data from Mar 2022 through Feb 2023 and nest the analysis within the GMM and quantile regression frameworks. We show the impact of the tick size update from USD 0.01 to USD 1.00 on those components, effected in Aug 2022, and reveal how their shares in the spread vary across the quantiles of Bitcoin’s price change distribution.

Krzysztof Najman, Kamila Migdał Najman, Angelika Kędzierska Szczepaniak, Krzysztof Szczepaniak, Assessing the progress of Agenda 2030 implementation in EU member countries using AI modeling and cluster analysis

By adopting the Agenda 2030 document, over 130 countries have committed to implementing 17 sustainable development goals. These goals can be described by a large no. of parameters. One of the groups of goals is Prosperity, the level of which in EU countries seems to be significantly different. The aim of the presented research is to assess the differentiation of Prosperity in EU countries in 2022/23. The research will use data clustering and classification methods, including selected AI models.

Benjamin Kwaku Nimako, Multi-criteria Decision Analysis (MCDA) based on performance indicators of energy scenarios

The transition to a 100% renewable energy system is vital for addressing climate change. This study uses advanced data science techniques, specifically Multi-Criteria Decision Analysis (MCDA), to evaluate and optimize energy scenarios for Bozen-Bolzano, Italy. Incorporating indicators such as economic, environmental, and technical factors, the research provides a robust decision-making framework. This approach ensures sustainable and resilient energy solutions.

Marcin Pełka, Transformation of symbolic variables for ensemble clustering

In the case of symbolic data analysis, two different approaches can be applied to the variable transformation problem. These are the principal component analysis and spectral clustering. In all cases, we start initially with a set of symbolic variables and

after transformation, we get a transformed dataset. The paper compares both approaches for artificial datasets with a known cluster structure. Results suggest that spectral clustering reaches better results for single and ensemble models.

Edoardo Redivo, Cinzia Viroli ,Efficient classification with integrated depth functions

The recently introduced integrated rank-weighted depth is based on computing the halfspace depth on projections of multivariate data. We show that this depth function is flexible, computationally efficient, and uniquely characterizes a probability distribution. Through simulations and real data applications, it is shown that this depth is a useful tool in supervised classification for both the DD-classifier and the maximum depth classifier, for which it enjoys an asymptotic optimality property.

Dorota Rozmus, Assessment of the impact of the COVID pandemic on the clustering of polish regions in terms of gross value added

Statistical Office in Katowice implemented a project: Statistical analysis of the impact of disturbing factors on selected macroeconomic indicators at the regional level. The data concerned output, intermediate consumption and gross value added by PKD sections and regions. By comparing actual values and forecast values, the aim will be to check, using clustering methods, whether the COVID-19 pandemic has influenced the clusters of Polish regions in terms of gross value added by kind of activity.

Adam Sagan, Anna Myrda, Emergent Causality in Family System

The paper explores the Hoel and Rosas concept of emergent causality within family systems and its consequences for the influence on consumer behavior. On the basis of effective information and coarse-graining proceure, we identify micro and macro networks of causal relationships within the sample of 75 family members and introduce a logic regression as a method for modeling macro dynamics within family networks. Key words: family network, effective information, logic regression

Zdenek Sulc, New approaches to hierarchical clustering of mixed-type data

The contribution compares selected hybrid similarity measures for mixed-type data, which are adjusted for hierarchical clustering, e.g., Huangu0027s distance from the k-prototypes method. It also proposes several modifications to the original Gower measure. All the examined measures are compared on generated datasets using

different data characteristics, e.g., ratio of categorical variables. The created clusters are evaluated regarding cluster recovery performance using the adjusted Rand index.

Gero Szepannek, Explanation Groves -- Analyzing the Trade off Between Appropriateness and Complexity of a Model Explanation

In the presentation, explanation groves are presented as a tool to extract a set of understandable rules in order to explain arbitrary machine learning models. The degree of complexity of the resulting explanation can defined be the user. This allows in addition to analyze the trade of between the complexity of a given explanation and how well it represents the original model. Explanation groves are available in the R package xgrove.

Mirosław Szreder ‘Representative sample – the need and the proposal for a definition’

A representative sample belongs to few categories present in statistical inference which do not have a precise and commonly recognized definition. The need for such definition does not seem to be high in natural sciences, where empirical experiments require rarely having a representative sample. Exploratory objectives of those sciences cause that the role of statistical inference is there different than in social sciences and economics. The latter ones focus mainly on descriptive and confirmatory goals and therefore they need samples which enable to make generalizations about the populations they represent accompanied by errors of estimates. The major aim of this paper is to discuss the role of the representative sample in scientific inference and statistical inference, and additionally to propose a formal definition of the representative sample.

Key words: representative sample, statistical inference, scientific inference, exploratory objectives, descriptive goals, confirmatory objectives

Marcin Szymkowiak, Maciej Beręsewicz, Quantile balancing inverse probability weighting for non-probability samples

In this paper we propose quantile balancing (QB) inverse probability weighting estimator for non-probability samples. Our method allows to include quantile information in the estimation process. Our simulation study has demonstrated that the estimators in question are more robust against model mis-specification and, as a result, help to reduce bias and improve estimation efficiency. We applied the proposed methods to estimate the share of vacancies aimed at Ukrainian workers in Poland

Michael Thrun ‘Deployment of an Explainable AI system for medical diagnosis as a second opinion’

The XAI provides transparent computational support for medical experts evaluating FCS data. Domain experts sign up via a website that connects to the XAI. Once new data is uploaded, they oversee the data analysis process online through an overview interface. The results, presented within a ticket system, include a classification, a self-competency estimation, and an explanation in the language of the domain expert. The enriched data, complete with explanations, can be downloaded by the experts.

Grażyna Trzpiot, Re-definitions of measures of demographic burden ratio

In this article we will present proposals for new measures of demographic burden ratio, based on quantiles of age distribution. Demographic dividends, which are the subject of my recent work, will appear in the context of the dynamics of changes in demographic structures. Longevity risk, which is at the root of research on changes in the description of age in the populations under study: chronological age, biological or prospective age are used in the construction of the proposed measures.

Max Welz, Robust estimation and inference with categorical data

I propose a general framework for robustly estimating models of possibly multivariate categorical data, such as questionnaire responses. The estimator generalizes maximum likelihood, is consistent, and asymptotically Gaussian. I verify the theoretical properties of the proposed methodology in simulation studies, and demonstrate its practical usefulness in an empirical application on structural equation modeling on a popular survey instrument, where I find evidence for inattentive responding.

Adalbert Wilhelm, Advancements in Semi-Supervised Clustering for Image Analysis: A Review

This presentation explores recent developments in semi-supervised clustering techniques for image analysis. It provides a comprehensive overview of the current state-of-the-art algorithms, their applications, and the challenges in leveraging limited labeled data for clustering large image datasets. We also discuss the potential impact of these advancements on various fields such as computer vision, medical imaging, and remote sensing.