SmallData Symposium 2024

Section I
Presentation Abstracts

Session 4: Navigating similarity & uncertainty - statistical approaches for robust predictions and inferences in the small data setting

Section II
Poster Abstracts

Poster Pitch Tour #1

N	Yang, H	Calibrating representations of expert knowledge with patient data in latent spaces for synthetic trajectories	37
O	Bratulić, J	Taxonomy-aware continual semantic segmentation in hyperbolic spaces for open-world perception	38
P	Walter, S	SPARQL knowledge graph question answering over Wikidata via constrained language modeling	39
Q	Ging, S	Image-text representation learning	40
R	Arnold, P	Comparing the performance of open and close sourced Large Language Models for automatic CAD-RADS 2.0 classification from cardiac computer tomography radiology reports	41
S	Fässler, D Huang, C	Methodologies to improve the scope and accuracy of whole-body models of human metabolism	42

Poster Pitch Tour #2

A	Scherer, N	Coupling of metabolomics and exome sequencing reveals graded effects of rare damaging heterozygous variants on gene function and human traits and diseases	43
B	Hoffman, L	Being certain of uncertainty	44
C	Mesuere, G	Locally stationary hidden Markov models	45
D	Müller, J	Efficacy of psychotherapy, pharmacotherapy, or their combination in chronic depression: a systematic review and network meta-analysis using aggregated and individual patient data	46
E	Neubrand, N	1000+ synthetic benchmark problems for parameter estimation in dynamic modelling	48
F	Kord, Y	Enhancing SNLS optimisation via deep reinforcement learning for adaptive tolerance setting	49
G	Hasan, M	Generating optimal small datasets for efficient offline reinforcement learning training	50
H	Zhang, B	Exploration cocktail: automating exploration in reinforcement learning	51
I	Zabërgja, G	Empirical assessment of paradigms in tabular classification	52
J	Purucker, L	Applying a foundation model to small tabular data	53
K	Kabus, F Hackenberg, M	An end-to-end modeling approach for capturing spatiotemporal patterns in two-photon imaging data	54
L	Döhler, S	Small data meets high dimensions: some approaches from multiple testing	55
M	O'Brien, T	Challenges of small data in biomedical and environmental research	56
N	Moringen, A	A meta unit for co-constructing a computational scaffold model to guide human motor learning	57
O	Wendland, P	OptAB - an optimal antibiotic selection framework for sepsis patients with artificial intelligence	58
P	Dümpelmann, M	Denoising of low dimensional EEG data with deep learning for improved seizure detection	59
Q	Krutsylo, A	Forward-forward optimization in small data	60

R	Brunn, N	Similarity-based refinement of single-cell interactions	61
S	Rollin, J	Prediction of cell lineage trajectories by integration of small single-cell RNA datasets into a large reference dataset	62

Poster Pitch Tour #3

A	Brombacher, E	Characterizing the omics landscape based on 10,000+ datasets	63
B	Mahendra, M	Convex space learning for tabular synthetic data generation	64
C	Umesh, C	Preserving logical and functional dependencies in synthetic tabular data	65
D	Archer, L	Uncertainty in clinical risk prediction: perspectives and approaches	66
E	Legha, A	Uncertainty-based sequential sample size calculations for developing clinical prediction models using regression or machine learning methods	67
F	Pierre Paul, D	Speeding up the clinical studies with biomarker-based enrichment	68
G	Schneider, J	Multimodal outcomes in N-of-1 trials: deep-learning based effect estimates in a small data study design	69
H	Papakonstantinou, E	Multidimensional investigation of response to treatment with inhaled corticosteroids in COPD patients: insights from the HISTORIC study	70
I	Lang, T	AI & statistics in preclinical research and development	71
J	Bonetti, M	Two small-sample problems in optimal and exact inference	72
K	Eggert, A	When only small data is available in livestock research	73
L	Farhadyar, K	Impact of different longitudinal data representations on transformer performance in small data applications	74
M	Bodden, D	Allocation bias in group sequential designs	75
N	Schoenen, S	Quantifying the impact of allocation bias in randomised clinical trials with multi-component endpoints	76
O	Bordoloi, R	Multivariate functional linear discriminant analysis of partially-observed time series	77

Section I

Presentation Abstracts

Integrating gene expression from whole blood into dynamical systems: illustration with a mechanistic model of the antibody response to COVID vaccination

Rodolphe Thiébaut[1]

[1]University of Bordeaux, Inserm, Inria, CHU, Bordeaux, France

Abstract

Dynamical systems applied to within host mechanistic modeling of virus, immune response to vaccines and other biological pathways have been successfully used and are a promising approach for personalized medicine. A corner stone is the availability of data to be fitted such as repeated measurements of viral or cell concentrations in blood or other tissues. However, such measurements are often sparse, leading to challenges for practical identifiability of model parameters. This issue is particularly acute for cell concentrations, as acquiring repeated venous samples in sufficient quantities for flow cytometry analysis is highly restrictive. Instead, we propose a methodology that capitalizes on dynamics that can be observed from high-dimensional gene expression measurements from whole blood. Our approach leverages cellular deconvolution algorithms to reduce the dimension of gene expression signal and obtain cell-type specific dynamics from whole blood gene expression. Subsequently, these cell-specific expression dynamics can be incorporated into the observation model of a dynamical system, overcoming the obstacles related to sparse observation and limited venous sampling. We study structural and practical identifiability of a system describing the establishment of humoral response after vaccination. We demonstrate that this methodological advancement promises to yield more accurate parameter estimates and deepen our comprehension of viral and immune dynamics. We demonstrate the practical application of this approach on COVID-19 vaccination data. New research projects including deep learning approaches are ongoing and will be mentioned.

A national research data infrastructure for data science and artificial intelligence

Sonja Schimmler[1,2]

[1]Technical University of Berlin, Berlin, Germany

[2]Fraunhofer Institute for Open Communication Systems, Berlin, Germany

Abstract

In my research, I focus on the digitalization and opening up of science with a special emphasis on research data infrastructures. My central hypothesis is that scientific progress will be furthered only if (1) digital artifacts become available at large scale, (2) these digital artifacts are linked and machine-interpretable, and (3) the specifities of the individual disciplines are taken into account.

In this talk, I will give an overview of several of my current research projects, especially NFDI4DataScience. The overarching objective is the development, establishment, and sustainment of a national research data infrastructure (NFDI) for the data science and artificial intelligence community in Germany. The key idea is to work towards increasing the transparency, reproducibility and fairness of data science and artificial intelligence projects, by making all digital artifacts available, interlinking them, and offering innovative tools and services.

Development of CRISPert, a novel deep learning-based tool enabling efficient and safe application of CRISPR-Cas (CRC A05)

Masako Monika Kaufmann[1], William Jobson Pargeter[2], Rolf Backofen[2], Toni Cathomen[1]

[1]Institute for Transfusion Medicine and Gene Therapy, Medical Center-University of Freiburg, Freiburg; Center for Chronic Immunodeficiency, Medical Center-University of Freiburg, Freiburg, Germany

[2]Bioinformatics Group, Department of Computer Science, University of Freiburg, Freiburg, Germany

Background

CRISPR-Cas9-based genome editing has revolutionized biomedical research by enabling the targeted modification of complex genomes in various cell types, including clinically relevant human cells. It consists of a Cas9 protein and a synthetic guide RNA (gRNA) with a 20-nucleotide long spacer matching the target site. However, it is known that the CRISPR-Cas9 system can also be active at so-called off-target sites which share high homology to the intended target site. Although several algorithms have been developed to predict CRISPR-Cas9 at on- and off-target activity, they are either not specific or not sensitive enough for use in clinical applications.

Methods

We are about to develop a deep learning-based method for cell type-specific prediction of CRISPR-Cas9 on- and off-target activity. As our approach integrates different types of information, including CRISPR-Cas9 binding and cleavage sites, the genetic sequence context of the cleavage sites, and cell type-specific chromatin and epigenetic features, we are about to generate CRISPR-Cas9 binding data by ChIP-seq, off-target data by CAST-seq, and on-target activity by high-throughput sequencing (HTS).

Results

We have nominated off-target sites of numerous CRISPR-Cas9 nucleases in hematopoietic stem cells and T cells in human, mouse and rhesus macaque by CAST-seq and validated the mutagenesis frequencies at on- and off-target sites by multiplexed HTS. Using publicly available cell type-specific ATAC-seq and RNA-seq data we further confirmed that off-target sites preferentially lie in sites with open chromatin and actively transcribed genes.

Conclusions

Unspecific DNA cleavage of CRISPR-Cas9 is a concern for clinical application because it can lead to genotoxicity, including gross chromosomal rearrangements or loss-of-function and gain-of-function mutations. In the future CRISPert will help scientists to design highly active and highly specific CRISPR-Cas9 nucleases by predicting cell type-specific on- and off-target activity.

From population studies to modeling of human metabolism - and back

Anna Köttgen[1]

[1]Institute of Genetic Epidemiology, Medical Faculty and Medical Center, University of Freiburg, Freiburg, Germany

Abstract

Rare variants in genes controlling central functions of human metabolism can cause altered levels of metabolites, small molecules that are absorbed, distributed, metabolized, or excreted by the respective gene products. This can lead to severe diseases such as autosomal-recessively inherited inborn errors of metabolism (IEMs). Studies of IEMs are limited because of the very low number of homozygous carriers of causative variants, constituting a small data challenge even in the setting of very large population studies.

This presentation will outline several approaches to tackle this challenge, including external replication testing, experimental confirmation, statistical approaches to leverage information from more commonly observed heterozygous variant carriers, and from *in silico* gene-knockouts in whole-body models of metabolism. The approaches will be exemplified by the study of several human conditions and diseases and emphasize their biomedical relevance. This presentation is connected to Small Data (CRC 1597) project B06.

Sparse clinical data: a call for population-level models

Jan Hasenauer[1]

[1]Hausdorff Center for Mathematics, Life & Medical Sciences Institute, University of Bonn, Germany

Background

The progression of diseases is a dynamic process, with clinical data providing insights into the trajectories of individual patients. However, individual patient data are often sparse, making it challenging to distinguish between natural variations and detrimental changes. Integrating data from different individuals is essential for robust analysis.

Methods

We explore the use of population-level models to integrate sparse datasets from individual patients. First, we employ nonlinear mixed-effects models to describe heterogeneous patient populations. To leverage large patient cohorts with limited information per individual, we introduce a novel inference scheme based on neural posterior approximation. Additionally, we utilize multi-state stochastic models to analyze patient trajectories, focusing on sparse longitudinal observations. Finally, we apply federated learning techniques to enhance data accessibility and integration across multiple clinical institutions.

Results

Our application of neural posterior approximation demonstrated effective integration and analysis of large, sparse datasets, providing insights into patient heterogeneity. The multi-state stochastic models facilitated the understanding of breast cancer metastasis development and the impact on different patient subgroups. Federated learning improved data accessibility, enabling more comprehensive analysis without compromising patient privacy.

Conclusions

Innovative methodologies, such as neural posterior approximation, multi-state stochastic models, and federated learning, significantly enhance the integration and analysis of sparse clinical data. These approaches provide robust insights into disease progression and patient outcomes, paving the way for improved clinical decision-making and personalized medicine.

No free lunch: why small data tasks require big data models

Thomas Brox[1]

[1]Department of Computer Science, Faculty of Engineering, University of Freiburg, Germany

Background

The abilities of present AI systems are based on a handful of fundamental principles, which I will go through in the talk. One of them is scaling. This principle is obviously in conflict with the idea of learning from small data. I will explain under which circumstances learning from small data makes sense, despite this conflict. I will argue that the major approach to learn from small data is the use of priors, and that deriving these priors requires big data at some point. An alternative approach can be targeted search for non-redundant data, but it also comes with a price tag. Finally, I will argue that world models are a perspective that serves both

What matters for in-context learning: a balancing act of look-up mechanism and in-weight learning (CRC B04)

Jelena Bratulić[1], Sudhanshu Mittal[1], Christian Rupprecht[2], Thomas Brox[1]

[1]Department of Computer Science, Faculty of Engineering, University of Freiburg, Germany

[2]University of Oxford, Oxford, United Kingdom

Background

In-context learning (ICL) is the ability in recent transformer-based models that enable them to learn new tasks from few-shot examples without updating the model weights. This has sparked large interest, since it makes models more ubiquitous and convenient to use and is closer to the way natural intelligence takes up a new task. ICL contrasts the standard in-weight learning (IWL) where the knowledge needed for inference tasks is embedded directly within the model weights. ICL has been theoretically and mechanistically studied to understand the model components and mechanism that establish the in-context learning ability.

Methods

A recent work, using GPT-2 on simple visual data, showed indication that ICL occurs due to certain data distributional properties like burstiness in the training sequences, large vocabulary and long-tail tokens. In this work, we show that the conditions for ICL are much less strict and come down to a balancing of in-context look-up and model parameter optimization during the training process, which can be achieved in multiple ways. We introduce a simple data augmentation method iCopy which is based on instance copying within the data sequences.

Results

Our proposed iCopy strategy achieves strong and stable ICL performance on various image classification datasets under both supervised and self-supervised pretraining tasks. We show how different approaches of regulating the IWL task difficulty influences the performance suggesting that a proper choice of the pretraining task is needed.

Conclusions

We identify that the relationship between the in-context look-up mechanism and the complexity of the model training objective influences the strength and stability of ICL. We propose a strong in-context look-up strategy with which we can maintain both ICL and IWL performance across multiple training objectives and datasets.

TabPFN v2, a foundation model for small tabular data

Frank Hutter[1]

[1]Department of Computer Science, Faculty of Engineering, University of Freiburg, Germany

Abstract

Tabular data, spreadsheets organized in rows and columns, is ubiquitous in many fields, prominently including medicine. The fundamental prediction task of filling in a label column based on the rest of the columns (the so-called features), is essential for predictive diagnostics, biomedical risk models, and drug discovery. Yet, in contrast to the deep learning revolution for text and images, the traditional machine learning methods used for such tabular data have made almost no progress for a decade. In this talk, I discuss TabPFN, the first foundation model for tabular data to dramatically improve predictive performance. TabPFN yields better performance in seconds than any previous method in hours, especially for small datasets, for which classical methods overfit. This makes it particularly promising for low-data applications in medicine.

Fusing deep learning and statistics towards understanding structured biomedical data

Sonja Greven[1], Marco Simnacher[1], Xiangnan Xu[1], Hani Park[2], Christoph Lippert[2]

[1]The Humboldt University of Berlin, Germany

[2]The Hasso Plattner Institute, University of Potsdam, Germany

Abstract

In biomedicine, structured data in the form of image stacks, genome sequences or time series are collected in increasing number and volume from high-throughput technologies. They are characterised by their inherent interdependencies between measurements, their often non-vector nature and the presence of confounding influences and sampling biases.

Deep learning (DL) excels in many applications on structured data and allows for accurate predictions due to its ability to capture complex dependencies within and between inputs and outputs. Despite recent advances, DL still has limitations in terms of uncertainty assessment, interpretability, and validation. However, these are essential components to go beyond prediction towards understanding the underlying biology. To this end, statistics has traditionally been used in the biomedical sciences to obtain interpretable model results and statistical inference, i.e., to quantify uncertainty, correct confounding, and test hypotheses with statistical error control. However, classic statistical methods are limited in their modeling flexibility for structured data and in their ability to capture complex nonlinearities in a data-driven manner. The DFG funded research unit KI-FOR 5363 Fusing Deep Learning and Statistics towards Understanding Structured Biomedical Data (DeSBi) brings together experts from machine learning and statistics with a track record in biomedical applications to develop methods that integrate DL and statistics. We aim to improve interpretability, uncertainty quantification and statistical inference for DL, and to improve modeling flexibility of statistical methods for structured data, in a feedback loop with biomedical applications. I will introduce the research unit and then talk in more depth about one of the projects developing **Deep Nonparametric Conditional Independence Tests for Images**, integrating statistical nonparametric Conditional Independence Tests with Deep Learning (embeddings) for images.

.

Statistical optimal transport meets life sciences

Axel Munk[1]

[1]Department of Mathematics and Computer Science, Göttingen University, Göttingen, Germany

Abstract

While optimal transport has been a long standing mathematical, physical and economic concept for more than two centuries, recent developments in statistics, optimization and machine learning suggests its use as a tool for modern data analysis. Variants, such as Gromov-Wasserstein transport respect the inner metric structure of data sets and have been proven to be useful for image registration and object matching. In this talk we introduce some basic concepts and aim to illustrate optimal transport data analysis with different examples from cell biology.

Improving machine learning for small genetic data using mathematical statistics (CRC C01-C05)

Carola Sophia Heinzel[1], Lennart Purucker[2], Frank Hutter[2,3], Peter Pfaffelhuber[1]

[1]Department of Mathematical Stochastics, University of Freiburg, Germany

[2]Machine Learning Lab, University of Freiburg, Germany

[3]ELLIS Institute Tübingen, Tübingen, Germany

Background

When using machine learning for tabular data, one must decide, among other things, (i) which model to select for use in practice and (ii) how to preprocess the data. Previous work on our application setting, i.e., classifying genotype data in forensic genetics, focuses exclusively on naive Bayes models with limited preprocessing. The performance of state-of-the-art machine learning models, which usually excel with larger datasets, remains unexplored.

Methods

We answer (i) for our application setting by benchmarking the performance of state-of-the-art machine learning algorithms for tabular data compared to naive Bayes classifiers. We used data from the SNIPPER reference database, consisting of only 3800 individuals from six populations with 100 markers, and the evaluation metric ROC AUC. With the best model from (i), we evaluate two approaches for (ii): using a statistical test based on the admixture model to remove unreliable individuals from the training data and generate more informative features.

We define unreliable individuals as those whose self-disclosure about their ancestry is wrong or who cannot be classified into one single population, e.g., if the parents come from different populations. We create new features by leveraging biological insights to fully utilize all available information. We then show how other machine learning models can exploit the new features and, by extension, mathematics.

Results

Our benchmark shows that more sophisticated, well-trained models outperform naive Bayes classifiers, even for small data. Moreover, we demonstrate the impact of data preprocessing on classifier performance.

Conclusions

Our work improves the state-of-the-art method for classifying individuals while providing evidence for the benefit of more sophisticated models in small data applications. In addition, we show how statistics and probability theory can aid general-purpose machine learning for forensics genetics applications. Moreover, our methods can be applicable to other tabular, small data classification problems.

Nonparametric maximum likelihood estimation of monotone binary regression models under weak feature impact

Angelika Rohde[1], **Dario Kieffer**[1]

[1]Department of Mathematical Stochastics, Faculty of Mathematics and Physics, University of Freiburg, Germany

Abstract

Nonparametric maximum likelihood estimation in monotone binary regression models is studied when the impact of the features on the labels is weak. We introduce a mathematical model that describes the weak feature impact. Consistency of the nonparametric maximum likelihood estimator (NPMLE) in Hellinger distance as well as its pointwise, L^1 and uniform consistency are proved in this model. Moreover, rates of consistency and limiting distribution of the NPMLE are derived. They are shown to exhibit a phase transition depending on the level of feature impact. Statistical properties of functionals in the weak feature impact scenario are also discussed.

Learning the differential equation of the tumour density of one breast cancer patient

Arnoldo Frigessi[1]

[1]Integreat, University of Oslo, Norway

Abstract

Our task is to predict the evolution in time of the density of cancer cells in a breast tumour under treatment. Typically, on has good data at treatment start, including histopathology from biopsies, clinical data, genomic data, MRIs, and possibly some repeated measurements after a first cycle of the treatment, which might last for a few months. However, there are not many observations of the tumour in time. One way to describe the time dynamic of the tumour density of the individual tumour is through the ordinary differential equation followed by the tumour density. Symbolic regression is an interpretable approach to learn the time dynamics of a system, as it estimates the differential equation. However, it requires a significant number of observations in time of the tumour density. Our approach is to combine inference on the differential equation with a stochastic simulation of the system under study, so to be able to generate synthetic data which are then useful for inference. This is joint work with Håkon Tasken, Alvaro Köhn-Luque and Benjamin Ricaud.

Analysis of CRISPR data and prediction for design of gene editing experiments

Jan Gorodkin[1,2]

[1]Center for non-coding RNA in Technology and Health, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark

[2]Computational Biology and Bioinformatics group, Department of Veterinary and Animal Sciences, University of Copenhagen, Denmark

Abstract

With the introduction of CRISPR editing DNA has been made extremely agile opening numerous applications over all kingdoms of life from enhancing microbial cell factories, crops and therapeutic possibilities. To obtain a desired change in the genome, e.g., changing a nucleotide with another in the DNA, there are often several possibilities, and to choose among these possible on-targets computational tools which can predict how efficient the edits will be are required. The computational tools are also critical to uncover unintended edits (off-targets) for each of the on-targets and can in the end influence the selection of location to edit. To construct the computational methods, from base line models to deep learning, data from various experimental screens needs to be processed. In the presentation I will present our work on human data, CRISPRon/off and the further ongoing methodological development.

Network meta-regression (CRC C02)

Nana-adjoa Kwarteng[1], Theodoros Evrenoglou[1], Adriani Nikolakopolou[1,2]

[1]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Centre, University of Freiburg, Freiburg, Germany

[2]Department of Hygiene, Social-Preventive Medicine and Medical Statistics, School of Medicine, Aristotle University of Thessaloniki, Thessaloniki, Greece

Background

Network meta-analysis (NMA) is a statistical technique that aims to provide relative effect estimates between an array of alternative treatments for a health outcome. As with regression within individual studies, it is often of additional interest to examine how certain characteristics affect NMA estimates. To this end, network meta-regression (NMR) enables the examination of treatment-by-covariate interactions within networks of treatments. When performing NMR, each treatment comparison becomes a unique subgroup within the network for interaction estimation. However, in networks where data availability is sparse, estimating the regression coefficients poses difficulties. Moreover, conceptual issues in NMR have hindered the wide adoption of NMR by evidence synthesis end-users.

Methods

We propose models for NMR under various combinations of NMR interaction and consistency assumptions and implement them within a frequentist framework. Using data from a published NMA of glucose-lowering agents and unpublished aggregate data on psychotherapies, we implement the proposed frequentist models and validate their output against existing Bayesian models. Detailed examination of the output highlights the importance of considering directionality in some models, and we introduce visualization tools for NMR results.

Results

Initial analyses reveal that the frequentist and Bayesian NMR methods provide comparable estimates under four combinations of NMR interaction and consistency assumptions. However, differences arise in scenarios with small subgroups where the observed covariate distribution typically deviates from the approximate normality assumption. Finally, in cases with extremely sparse covariate data, both frequentist and Bayesian NMR models face challenges in parameter estimation.

Conclusions

This work elucidates some assumptions of NMR, the importance of the network structure, and the pivotal role of small data in examining covariates in NMR. The proposed frequentist models can improve the adoption of NMR, facilitating the exploration of sources of heterogeneity and inconsistency in NMA.

The vision of SmallData

Harald Binder[1]

[1]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Centre, University of Freiburg, Freiburg im Breisgau, Germany

Abstract

Reflecting on the wide range of contributions to small data methodology and applications leads to the question of whether there can be a unified small data movement, and consequently what the contribution of individual researchers can be. In particular, it raises the question of how we might need to adapt our research practices to enable a small data community, perhaps even against the headwinds of our respective disciplines. Of course, this also leads to the question of short-term incentives. For example, doctoral researchers need to carefully plan their research agendas, and this may not fit well with community building, which may only pay off in the longer term, especially in the context of the AI movement at large, where a potential AI winter is looming. Using examples in the context of the small data concepts of similarity and uncertainty, I will suggest some possible answers. In particular, I will outline the immediate benefits of engaging with a small data community, including how to overcome some major hurdles. This includes the data engineering gap between methods research and applications in biomedical research, where I will discuss a potential role for statistics as a discipline that could guide developments.

Section II

Poster Abstracts

Identifying data similarity across subgroups and sites (CRC A01)

Max Behrens[1], Gabriele Bellerino[2], Nils Kober[2], Angelika Rohde[2], Daiana Stolz[3], Moritz Hess[1], Harald Binder[1]

[1]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Germany

[2]Department of Mathematical Stochastics, University of Freiburg, Germany

[3]Department of Pneumology, Faculty of Medicine and Medical Center, University of Freiburg, Germany

Background

Clinical prediction models often struggle with the heterogeneity across patient subgroups and multiple data sites. This challenge is even more pronounced in small data scenarios where accommodating this complexity is difficult. To address these issues, we propose a novel approach using localized regressions within an autoencoder network. We observe that different models are valid for different subgroups within a single site. Extending this to another data site, we find that while some subgroups align between sites, others differ significantly. Our goal is to identify these subgroups and determine when to enhance the site-specific model with external data and when to exclude it.

Methods

To identify subgroups, we fit a localized regression for each observation with weights reflecting proximity to other observations within the autoencoder's latent space. The autoencoder reduces data dimensionality, ensuring a dense representation for subgroup-specific model fitting. By differentiating through both the autoencoder and the localized regressions, we train a latent representation that reconstructs the data well and learns differences between subgroups from the two data sites.

Results

We demonstrate our approach on data from two sites with patients suffering from chronic obstructive pulmonary disease (COPD). When fitting our method to the data, we see patterns in the latent space and can identify subgroups with unique characteristics. Some subgroups can be combined between data sites as site-specific models are similar, while other subgroups are unique to each site. Our method serves as a diagnostic tool to identify subgroup-specific disease dynamics within the data.

Conclusions

Our approach effectively addresses the heterogeneity across subgroups and multiple data sites. We can identify subgroups and determine when to enhance them with external data and when to exclude such data. These subgroups can be a valuable basis for subsequent analysis with the potential to improve individual site-specific clinical healthcare.

ID 102

The following topics will be presented as a joint poster:

Similarity weights in the nonparametric maximum likelihood estimator (CRC A01)

Nils Kober[1], Angelika Rohde[1], Gabriele Bellerino[1], Max Behrens[2], Harald Binder[2]
[1]Department of Mathematical Stochastics, Mathematical Institute, University of Freiburg, Freiburg, Germany
[2]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Germany

Background
The nonparametric maximum likelihood estimator (NPMLE) is a highly versatile tool for estimating distribution functions without parametric assumptions, which we consider in the context of one-sided interval-censored data. We propose leveraging external data with similar underlying distributions to enhance estimation for small target datasets by introducing appropriate weights in the criterion function.

Methods
We analyze the weighted NPMLE and its convergence rates in asymptotic scenarios specifically tailored to small data problems. These allow for simultaneous convergence of sample sizes and the underlying distributions, aiming to reveal desirable properties of the weighted NPMLE that would remain obscured in conventional settings.

Results
Under suitable convergence assumptions for the external distribution, we have established the consistency of the NPMLE with uniform weights.

Conclusions
The weighted NPMLE appears to be a promising approach to leverage external data. The optimal choice of weights will be investigated as part of this project.

Limit theorems for Markov processes (CRC A01)

Gabriele Bellerino[1], Angelika Rohde[1], Nils Kober[1], Max Behrens[2], Harald Binder[2]
[1]Department of Mathematical Stochastics, Mathematical Institute, University of Freiburg, Freiburg, Germany
[2]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Germany

Background
Markov Processes find wide applicability in several fields as they often represent an acceptable approximation of real-world processes. Therefore, the limit behaviour of statistics when the number of observations increases is of crucial interest for many applications. If the Markov process is not geometrically ergodic, however, such limit statements do not reflect small sample effects. To address this aspect, we study the asymptotic distribution for additive functionals along a triangular array of Markovian observations which approaches the null-recurrent case. Interesting phenomena arise which cannot be observed in the classical asymptotics.

Methods

We concentrate firstly on providing results for subclasses of Markov Processes. Then we eventually enlarge the class of applicability of these results.

Results

A preliminary result for additive functionals based on polynomials is available.

Conclusions

The stepwise generalization of the results could be a promising approach for such statistics, due to the particular structure of dependence of Markov Processes.

ID 103

Bias-corrected maximum likelihood estimation of parametric competing risk models for small data (CRC A02)

Anika Schlosser[1], Maryam Farhadizadeh[2], Harald Binder[1], Zoe Lange[3], Holger Dette[3], Nadine Binder[2]

[1]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany

[2]Institute of General Practice/Family Medicine, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany

[3]Faculty of Mathematics, Ruhr University Bochum, Bochum, Germany

Background

The A02 project aims to assess the similarity of healthcare pathways to identify best practice treatment strategies. The availability of routine clinical data provides new opportunities to analyze these pathways, which describe the sequence of diagnostic or treatment events for specific diseases. However, diverse treatment options lead to heterogeneous healthcare pathways, resulting in decreasing numbers of observed transitions over time and creating a small data challenge. Multistate models form a powerful class of statistical models for describing healthcare pathways. Moving between treatments as states over time is modeled by transition intensities. Estimating the intensities parametrically, common methods like maximum likelihood estimation (MLE) result in significant bias for small sample sizes.

Methods

To reduce bias in MLE estimates, we adapted the Cox-Snell bias approximation approach for competing risk models, a special case of multistate models. We theoretically develop corrected estimators for both uncensored and censored data, assuming underlying exponential and Weibull distributions. To validate our proposed estimators, we conducted a simulation study motivated by routine clinical data from urology. When simulating censored data for small sample sizes, datasets with no observed events are likely to be generated. Excluding such datasets would introduce further bias in the estimates. Instead, we introduce a new approach that adds artificial observations with small weight to all generated data sets.

Results

Our simulation study shows promising results, demonstrating that our methods significantly reduce bias while not increasing variance. Considering sample sizes ranging from 5 to 50 and censoring administratively up to a censoring rate of 80%, we observed consistent bias reduction across all scenarios. The obtained estimates demonstrated expected properties in terms of dependence on censoring rate and sample size. Furthermore, the pseudo-observations effectively handle inestimable datasets in small sample scenarios, introducing less bias compared to the exclusion of datasets.

Conclusions

In conclusion, we propose corrected estimators for small data, reducing the bias of MLE for parametric multistate models, while also introducing a novel method for avoiding scenarios with no observed transitions in data generation.

Identifying best practice treatment strategies by incorporating information from similar healthcare pathways (CRC A02)

Holger Dette[1], Maryam Farhadizadeh[2], Nadine Binder[2], Zoe Lange[1]

[1]Faculty of Mathematics, Ruhr University Bochum, Bochum, Germany

[2]Institute of General Practice/Family Medicine, Faculty of Medicine and Medical Center, University of Freiburg, Germany

Background

The identification of similar patient pathways is an important step in improving the overall healthcare practice. Despite the increasing availability of routine clinical data, performing inference on healthcare pathways remains a problem of small sample sizes due to the strong heterogeneity of the data. Testing the similarity of different samples of pathways helps to increase the sample sizes by pooling similar data and therefore, allows the study of typical pathways.

Methods

We model a pathway as a multistate model, namely a process in continuous time with possible transitions between a finite number of states. In this research, the similarity of such processes is analysed based on their transition probabilities. For two multistate models X^1 and X^2, describing the pathways of two different groups of patients, we test the hypotheses

$H_0: d(P^1,P^2) \geq \varepsilon$ versus $H_1: d(P^1,P^2) < \varepsilon$.

Here, P^1and P^2 are the transition probability matrices of X^1 and X^2, respectively, while $\varepsilon > 0$ is a threshold and d is a distance on the space of matrices. To account for the small sample sizes, a constrained, parametric bootstrap test for these hypotheses is suggested.

Results

We give a proof of the validity of the new bootstrap test for different distances, defining different measures of similarity between multistate models. In this context, we also prove a conditional Lindeberg-Feller Central Limit Theorem that could be a useful tool in different bootstrap settings as well.

Conclusions

Testing the similarity of multistate models based on their transition probabilities is a new and promising approach. Further research will focus on testing the performance of the bootstrap for different sample sizes by means of a simulation study and on comparing it to the performance of approaches that use transition intensities for similarity testing.

Analysis of treatment effects despite switches in measurement instruments by combining variational autoencoders with mixed effects models (CRC A03)

Clemens Schächter[1], Thorsten Schmidt[2], Felix Ndonfack[2], Astrid Pechmann[3], Janbernd Kirschner[3], Harald Binder[1]

[1]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany

[2]Department of Mathematical Stochastics, Faculty of Mathematics and Physics, University of Freiburg, Freiburg, Germany

[3]Department of Neuropediatrics and Muscle Disorders, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany

Background

In a longitudinal clinical registry dedicated to rare diseases, patients could undergo a variety of treatments and may transition between different therapeutic options as new medications emerge. To mitigate the challenge of small patient populations with few follow-up visits, physicians may adopt a comprehensive approach to patient characterization. This includes detailed baseline assessments and characterizing individuals at the few observed time points utilizing a variety of tests or measurement tools. Integrating these high-dimensional and incomplete observations selecting the relevant baseline information is infeasible with conventional statistical methods.

Methods

For a better understanding of individual disease dynamics, for example, under a treatment regime, we model the patient progression within a dimension-reduced latent space. Our methodology utilizes variational autoencoders to map high-dimensional observations to the latent space. We use separate encoder and decoder networks for every test and align their latent representations. The latent trajectory incorporates treatment effects and switches through a mixed-effects model. Here, available baseline information is incorporated through fixed effects, ensuring that the model accurately captures their effect on each patient's health status.

Results

Our model is able to handle multiple modalities with missing observations, changes in measurement instruments, and irregularly timed visits. By imposing a mixed model on the latent space, we achieve a nuanced understanding of each patient's individual disease progress and the impact of various treatment options. We correct for potential overfitting of the encoder models to the latent dynamics to perform model selection more accurately. Our methodology is evaluated on the SMArtCARE registry, which contains patients with spinal muscular atrophy (SMA).

Conclusions

Our approach can model disease progression and treatment impact in rare diseases, as shown in the SMArtCARE registry for spinal muscular atrophy. By imposing mixed effects models on the dimension-reduced latent representation, we offer a detailed perspective on individual patient disease dynamics.

Advanced filtering theory and the Zakai equation for jump-diffusion stochastic processes (CRC A03)

Felix B. Tambe-Ndonfack[1], Clemens Schächter[2], Harald Binder[2], Thorsten Schmidt[1]

[1]Department of Mathematical Stochastics, Mathematical Institute, University of Freiburg, Freiburg, Germany

[2]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany

Background

In the theory of stochastic processes, a central challenge is the estimation of the state of a signal from noisy observations, a fundamental problem in filtering theory. Filtering has diverse applications across various fields, including signal processing and finance. Filtering problems can be classified into two types: linear and nonlinear filtering.The linear filtering problem was initially formulated and solved by Norbert Wiener in 1949 and Andrey Kolmogorov in 1941. Rudolf Emil Kalman later redefined the linear filtering problem by presenting it in the form of a stochastic system in state space, leading to the development of the Kalman filter.This study focuses on nonlinear filtering, specifically addressing systems of stochastic differential equations involving unobservable signals and observable processes. These systems are crucial in scenarios where direct observation of signals is impossible or impractical, making advanced filtering techniques essential for accurate state estimation.

Methods

We employ rigorous mathematical frameworks, including the theory of Poisson random measures (PRMs) to model the random occurrence of jumps. This approach ensures synchronisation in the jump processes affecting both signal and observation process. Due to the complexity to obtain an explicit form solution of the problem, we start by designing a specific approach: **Formulation of Assumptions:** Careful formulation to ensure the existence and uniqueness of solutions. **Predictable Projection:** Finding the predictable projection with respect to a filtration generated by observations. **Change of Probability Measure Method:** Modifying the probability measure to transform the observation process into a Brownian motion using Girsanov's theorem.

Results

The results include: **Formulation using PRMs:** Ensuring jumps occur according to a Poisson process aligned with the system's stochastic nature which after a certain binding will be a dependency of an aleatory measurement respecting the jump of the aleatory process. **Ongoing work**, such as the dual compensator and equations like the Kushner-Stratonovich and Zakai equations, which are integral to extending the study's findings. Parameter estimation using particle algorithms and EM algorithms is also studying. In addition, the extension of this study in the context of semimartingales (which seems a little more rigorous).

Conclusions

In conclusion, filtering theory offers a powerful tool for addressing small data problems in dynamic contexts like stochastic differential equations. The study contributes to filtering theory by providing a robust mathematical framework for systems involving stochastic differential equations with jumps. This work not only advances theoretical understanding but also offers practical insights into modeling and predicting complex stochastic systems.

Dissecting the molecular basis of monogenic neurodevelopmental disorders (CRC A04)

Esma Secen[1], Rolf Backofen[2], Miriam Schmidts[1]

[1]University Hospital Freiburg, Center for Pediatrics and Adolescent Medicine, Freiburg, Germany

[2]Department of Computer Science, University of Freiburg, Freiburg, Germany

Background

The prevalence of neurodevelopmental disorders accompanied by intellectual disability (ID) poses significant healthcare challenges, affecting 1 to 3% of the population. Hereby monogenic forms constitute a major etiological factor. Disease gene identification in ID is hampered by excessive genetic heterogeneity and up to 1000 rare gene variants putatively impacting the protein sequence identified per individual in exome data.

Methods

Our research aims to elucidate the genetic architecture underlying intellectual disability and to dissect underlying molecular disease mechanisms. This includes gene discovery approaches using exome sequencing and development of a neural network approach to facilitate causal gene priorisation in exome datasets.

Results

So far, analysing over 200 exomes from human ID cases, we have identified to novel human disease genes, *TTBK1* and *FKBP4*, causing a syndromal ID phenotype upon loss if function. TTBK1 has been previously associated with Alzheimer`s disease playing a role in Tau phosphorylation, however loss of function effects during neurodevelopment have not been studied to date. FKBP4 has been suggested to play a role in protein folding and mTor signalling and loss of function of *Fkbp4* in mice has been shown to cause a sexual development phenotype, likely through defective androgen receptor signalling.

Currently, we are generating in-vitro models to study TTBK1 and FKBP4 function and dissect possibly disease mechanisms.

Conclusions

We have identified two novel human disease genes, TTBK1and FKBP4. Future research will include confirmation of in-vitro findings also in vivo.

ID 109

CRISPert: A transformer-based model for CRISPR-Cas off-target prediction (CRC A05)

William Jobson Pargeter[1], Masako Monika Kaufmann[5], Van Dinh Tran[3,4], Toni Cathomen[5], Rolf Backofen[1,2]

[1]Bioinformatics Group, Department of Computer Science, University of Freiburg, Freiburg, Germany

[2]Signalling Research Centre CIBSS, University of Freiburg, Freiburg, Germany

[3]Department of Information and Computer Science, King Fahd University of Petroleum and Minerals, Saudi Arabia

[4]SDAIA-KFUPM Joint Research Center for Artificial Intelligence, Saudi Arabia

[5]Institute for Transfusion Medicine and Gene Therapy, Medical Center-University of Freiburg, Freiburg; Center for Chronic Immunodeficiency, Medical Center-University of Freiburg, Freiburg, Germany

Background

CRISPR-Cas9 has emerged as a popular gene-editing technique due to its flexibility, precision, and ease of use. It involves a complex that consists of a Cas9 protein and a designed, synthetic single guide-RNA(sgRNA) that guides the Cas9 protein to its intended genomic target site, where it induces editing of the DNA through cleavage. Despite its popularity, the potential side effects caused by unintended cleavage of CRISPR-Cas9 have been a critical issue that hinders its development and clinical applications. Many methods have been proposed for off-target site prediction. However, they only obtain moderate results. This is partly due to the high imbalance of data, the choice of network architecture, and the neglect of additional useful information.

Methods

Therefore we present CRISPert, a transformer-based model for predicting CRISPR-Cas9 off-target sites. The model integrates sequence data from the sgRNA and off-target site with additional contextual features such as Cas9 binding profiles. Pretraining allows the model to effectively learn from limited and imbalanced datasets.

Results

We show CRISPert outperforms state-of-the-art models on the main off-target prediction benchmark dataset. Additional tests show that the inclusion of Cas9 binding profiles leads to improved performance.

Conclusions

Predicting the potential off-target sites can help evaluate the safety of a designed CRISPR-Cas9 system. Empirical results from various experimental settings show that our proposed model outperforms all compared methods and confirms its potential for practical use.

CoordConformer: Decoding heterogeneous EEG datasets using transformers (CRC 801)

Samuel Boehm[1], Sharat Patil[2], Robin Tibor Schirrmeister[1], Frank Hutter[2], Tonio Ball[1]

[1]Department of Neurosurgery, Medical Center, University of Freiburg, Freiburg, Germany

[2]Department of Computer Science, Faculty of Engineering, University of Freiburg, Freiburg, Germany

Background

With modern machine learning capabilities, it is possible to interpret the electrical activity of the brain and classify, for example, between different pathological brain activities, sleeps states or emotions. A reliable method to measure these signals is electroencephalography (EEG), as it is applied to the scalp. However, most current machine learning models for EEG face a dilemma: they are very specific to the task for which they are trained. To make them more versatile, they need to be trained on many different datasets. This, in turn, is not trivial to accomplish: For many datasets, different brain areas have been recorded, with a varying number of electrodes and experimental setups, making it impossible to feed all the data to the same model.

Methods

With the introduction of our CoordConformer model, we provide a new model that utilizes self attention mechanisms to not only learn from the EEG signal, but also take the recorded brain area into account by paying attention to the coordinates where each electrode was placed. Furthermore, and probably more importantly, this method allows us to train with any arrangement of electrodes, meaning a varying number of electrodes in the datasets does not prevent the model from learning from the dataset. With this method we aim to learn the general underlying statistics of EEG.

Results

Our CoordConformer shows state of the art accuracy in common motor imagery classification tasks.

Conclusions

The here introduced architecture has the potential to form a backbone by training over many different datasets. We anticipate that the potential can be further realized by training the model on unsupervised tasks, as this has been shown to be very effective for this type of models in other domains.

Meta-learning population-based methods for reinforcement learning (CRC B01)

Johannes Hog[1], Noor Awad[1], André Biedenkapp[1], Raghu Rajan[1], Vu Nguyen[2], Frank Hutter[1]

[1]Department of Computer Science, Faculty of Engineering, University of Freiburg, Freiburg, Germany

[2]Amazon, Australia

Background

Reinforcement learning (RL) environments have become more complex and time-intensive to train. This limits the amount of data that can be used for hyperparameter optimization, which is crucial for well-performing RL agents. One prominent approach that handles this limitation through the use of parallel resources is Population-based Bandits (PB2). One inefficiency of PB2 is that it is prone to slow starts due to the lack of observations at the beginning of optimization.

Methods

We tackle this inefficiency and speed up the optimization with meta-learning, which uses the information contained in past training runs to inform the training in new environments.
We propose four different meta-learning extensions to PB2. Our first approach learns a hyperparameter portfolio to warmstart the optimization. Our other approaches meta-learn different parts of the BO component used by PB2, namely a prior to the surrogate model, the surrogate model, and the acquisition function. We evaluate our methods and compare them to standard baselines on two different environment families.

Results

Our results show that MultiTaskPB2, our method that meta-learns the surrogate model itself, achieves the best overall anytime and final performance compared to all baselines. Additionally, we quantify the overhead introduced by our methods and show that it is negligible for expensive RL environments.

Conclusions

We demonstrate that meta-learning can speed up and robustly improve the performance of hyperparameter optimization in expensive RL environments. We also motivate further research into suitable meta-features for RL.

Dynamic integration of process models and neural networks to improve predictive performance in ecology (CRC B02)

Hanne Raum[1], Hannah Habenicht[2], Joschka Boedecker[1], Carsten Dormann[2]

[1]Neurorobotics Lab, Department of Computer Science, Faculty of Engineering, University of Freiburg, Freiburg, Germany
[2]Institute of Biometry and Environmental Analysis, Faculty of Environment and Natural Resources, University of Freiburg, Freiburg, Germany

Background

When studying complex phenomena in nature, a simplified version of a system is often used to describe the underlying effects and processes. In forest science, this is done by process models (PM), which combine empirical measurements with theoretical understanding of the underlying processes. However, the PMs representation of complex phenomena may be incomplete or oversimplified.

Machine learning models, particularly neural networks (NNs), can outperform PMs with more flexible representations when large data sets are available. For small and sparse datasets, NNs can benefit from the integration of domain knowledge such as PMs. However, due to the simplification of the underlying processes, the accuracy of process forest models is limited. In such cases, the integration of PMs into NNs may worsen the results.

Methods

We propose a model framework, where we identify a dynamic training approach that identifies which temporal part of process model predictions are accurate enough to be integrated into NNs, thus improving our training process. Furthermore, we aim to fill these identified knowledge gaps by suggesting additional terms through scientific discovery methods.

Results

Initial experiments demonstrated that NNs can successfully predict errors in a forest PM and improve its predictions. Additionally, the PM varied in its prediction accuracy over time.

Conclusions

Identifying parts of PM predictions that are accurate enough to support NNs can not only lead to better prediction accuracy, but also highlight knowledge gaps. Furthermore, addressing these knowledge gaps through methods such as scientific discovery and identifying additional terms for process models can improve our understanding and suggest ways to improve process models.

Similarity evaluation of training vs test data and the potential of process knowledge (CRC B02)

Hannah Habenicht[1], Hanne Raum[2], Joschka Boedecker[2], Carsten Dormann[1]

[1]Institute of Biometry and Environmental Analysis, Faculty of Environment and Natural Resources, University of Freiburg, Freiburg, Germany

[2]Neurorobotics Lab, Department of Computer Science, Faculty of Engineering, University of Freiburg, Freiburg, Germany

Background

We explore “informed machine-learning (ML)”, which integrates both scientific insights and data into the development and training of ML models, moving beyond traditional data-only or theory-only approaches. Scientific insights can take the form of mechanistic process models (PMs), which, in the environmental sciences, provide a simplified representation of realistic environmental processes. Although these hybrid-ML models approach significantly improve over PM predictions, challenges in capturing certain data patterns persist, particularly in sparse-data settings. To move beyond mere black-box predictions, we enhance explainability of hybrid models by analysing boundaries and decay of prediction quality in our models. Data constraints often lead to misleading extrapolations to unknown environments (even for well-fitting models), typically caused by the limited environmental range covered by the training data and the specific assumptions of each algorithm, when extrapolating beyond that range.

Methods

To gain a better understanding when both the process and hybrid-ML models extrapolate their predictions we apply data analysis techniques, used to understand the structure, distribution, and similarity of multi-dimensional training vs test data. In essence, we examine the informative content available in our data, especially for rare or extreme values. We expect that with sparse data availability, the ML would rely more heavily on the information provided in the PM, thereby preventing a drop in extrapolation performance.

Results

Initial experiments demonstrate a considerable improvement in predictions by both our ML and hybrid-ML models over the standalone PM, in both spatial and temporal extrapolation. As similarity of test data with training data decreases, PMs suffer less than ML, with hybrid-ML moderately affected. Further analyses of these results are to be conducted.

Conclusions

Hybrid-MLs show great promise not only in improving the prediction capabilities of process models, but particularly in the context of sparse data availability. Specifically, within sparse data, knowing the informational similarity available in data may provide useful indications of how well a model will perform.

A systematic review and metanalysis for inflammation parameters in dystrophic epidermolysis bullosa (CRC B03)

Meropi Karakioulaki[1], Nana-Adjoa Kwarteng[2], Adriani Nikolakopoulou[2], Hanning Yang[2], Moritz Hess[2], Kilian Eyerich[1], Cristina Has[1]

[1]Department of Dermatology and Venereology, Faculty of Medicine and Medical Center, University Hospital Freiburg, Freiburg, Germany

[2]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany

Background

Dystrophic epidermolysis bullosa (DEB) is a rare inherited skin blistering disorder caused by mutations in the collagen type VII gene, leading to mucocutaneous blistering. Subsequent inflammation contributes to chronic wounds, scarring, and systemic complications. There is controversy over if and how inflammation should be therapeutically targeted. This systematic review and meta-analysis aim to analyze patterns of tissue and systemic inflammation in DEB and identify research gaps to improve patient management.

Methods

Studies examining the association between DEB and tissue and systemic inflammation were eligible for inclusion. A comprehensive search of MEDLINE via PubMed for studies published in English from 14.01.2024 to 18.03.2024 was conducted, using specific terms for epidermolysis bullosa and inflammation. Out of 663 studies found, 37 met the inclusion criteria and were included in the systematic review. Data for the meta-analysis were collected from studies investigating cytokine levels in DEB patients. The dataset included threenodes: DEB patients, healthy controls, patients with other forms of EB. After excluding studies without measures of central tendency, one node and nodes with one participant, the final dataset comprised 32 markers within 11 studies.

Results

IL-6 was consistently identified as a key factor in tissue and systemic inflammation in DEB across most studies. Other elevated inflammation parameters in DEB included CRP, IL-10, IL-31, IL-1, IL-2, TSLP, SAA, TNF-β, and IFN-γ. Autoantibodies such as anti-BP180 and anti-BP230 were positively correlated with IL-6, the IL-6/IL-10 ratio, disease severity, the BEBS score, and IFN-γ in various studies.

Conclusions

Existing studies have limitations, such as heterogeneous patient groups, reliance on observational and retrospective descriptive studies, and a lack of large-scale prospective and interventional studies. Advanced AI modeling tools can help study complex and rare diseases like DEB. New well-designed clinical trials and prospective studies are necessary to address the unmet needs of patients suffering from this devastating genetic disease.

ID 116

Calibrating representations of expert knowledge with patient data in latent spaces for synthetic trajectories (CRC 803)

Hanning Yang[1], Meropi Karakioulaki[2], Cristina Has[2], Harald Binder[1], Moritz Hess[1]

[1]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany

[2]Department of Dermatology, Medical Faculty and Medical Center, University of Freiburg, Freiburg, Germany

Background

In longitudinal clinical studies on rare diseases like Epidermolysis Bullosa (EB), various interrelated measures such as blisters, wounds, inflammation, anemia, and physical growth need to be considered. However, modeling patient trajectories is challenging due to limited observations and complex missing data patterns in individual patient data (IPD). Representing expert knowledge from the scientific literature via quantitative models can help augment IPD with synthetic data. Nonetheless, calibration with real data is often lacking. We construct ordinary differential equations (ODEs) to simulate the evolution of EB biomarkers and use deep neural networks for ODE calibration.

Methods

Drawing upon expert knowledge , we developed a system of five ordinary differential equations (ODEs) to model key biomarkers like C-Reactive Protein and hemoglobin, representing the EB progression. Initial conditions are sampled from Gaussian distributions. To calibrate the ODE system with real data, we use an autoencoder, i.e., a neural network based approach, for dimension reduction. We calibrate initial condition distributions using a centering approach in the latent space. Realistic data scenarios with various noise and missing data are simulated. Missing variables are handled by a separately trained imputation layer. Performance is measured using mean squared error (MSE) for judging imputation accuracy and Euclidean distance between true and calibrated ODE parameters in simulations.

Results

Our method calibrates the ODE system in a low-dimensional latent space, ensuring reliable results despite various noise and complex missing data patterns, which is justified by the stability of the gradients. Furthermore, our centering approach maximizes the utilization of available information.

Conclusions

We demonstrate calibrating a complex expert-informed synthetic data model by comparing real and generated observations in a low-dimensional latent space. Combining ODEs with an autoencoder allows effective dimension reduction and flexibility in handling data complexities, enabling realistic synthetic IPD while having control over the generative processes.

ID 118

Taxonomy-aware continual semantic segmentation in hyperbolic spaces for open-world perception (CRC B04)

Julia Hindel[1], Daniele Cattaneo[1], Abhinav Valada[1]

[1]Department of Computer Science, University of Freiburg, Freiburg, Germany

Background

Semantic segmentation models are typically trained on a fixed set of classes, limiting their applicability in open-world scenarios. Class-incremental semantic segmentation aims to update models with emerging new classes while preventing catastrophic forgetting of previously learned ones. However, existing methods impose strict rigidity on old classes, reducing their effectiveness in learning new incremental classes.

Methods

We propose Taxonomy-Oriented Poincaré-regularized Incremental-Class Segmentation (TOPICS) that learns feature embeddings in hyperbolic space following explicit taxonomy-tree structures. This supervision provides plasticity for old classes, updating ancestors based on new classes while integrating new classes at fitting positions. Additionally, we maintain implicit class relational constraints on the geometric basis of the Poincaré ball. This ensures that the latent space can continuously adapt to new constraints while maintaining a robust structure to combat catastrophic forgetting.

Results

We establish nine realistic incremental learning protocols for autonomous driving scenarios, where novel classes can originate from known classes or the background. Extensive evaluations of TOPICS on the Mapillary Vistas 2.0 and Cityscapes benchmarks demonstrate that it achieves state-of-the-art performance.

Conclusions

Our method is one of the early works that uniformly address the bifurcation of previously observed classes and incremental classes from the background. Consequently, we motivate future research to target these realistic open-world scenarios. Further, we emphasize the benefit of hierarchical modeling in hyperbolic space and want to motivate future work to explore its potential for various visual challenges.

ID 119

SPARQL knowledge graph question answering over Wikidata via constrained language modeling (CRC B05)

Sebastian Walter[1], **Hannah Bast**[1]

[1]Department of Computer Science, University of Freiburg, Freiburg, Germany

Background

SPARQL-based knowledge graph question answering (KGQA) is the task of answering natural language questions by generating corresponding SPARQL queries that can be executed over a knowledge graph. KGQA is challenging because it is a combination of several different subtasks, like language understanding, semantic parsing or entity and relation recognition and disambiguation. Even seemingly simple questions can lead to very complex SPARQL queries, see e.g. https://qlever.cs.uni-freiburg.de/wikidata/FMZ8Sr for the query returning the "highest peak per country".

Methods

Our approach is based on large language models that are trained to translate natural language questions into natural language representations of SPARQL queries. We use a combination of self-supervised and supervised learning to train our models on a large dataset of SPARQL queries and question-query pairs, currently containing around 1.4 million examples in total. In addition, we guide and constrain the models during inference in various ways to improve the quality of the generated queries, including:

1. Natural language index for generating valid entities and relations
2. Subgraph constraining
3. SPARQL grammar constraining
4. Similarity-based in-context examples

And more.

Results

At this point, we evaluated our approach on a KGQA benchmark dataset containing only simple questions, called SimpleQuestionsWikidata. We achieve state-of-the-art performance (0.88 F1-Score), outperforming previous approaches by a large margin (see https://ad-blog.cs.uni-freiburg.de/post/deep-knowledge-graph-question-answering/ for more details). Initial experiments on more challenging datasets like LC-Quad v2 show promising results, but are still ongoing. We also make all of our code already available during development, including a demo at https://ad-sparql-kgqa.cs.uni-freiburg.de.

Conclusions

Our current focus lies on the Wikidata knowledge graph, the largest and most widely used one. We are working towards a generalist KGQA system capable of answering questions over any knowledge graph with only a few examples.

Image-text representation learning (CRC 805)

Simon Ging[1], Sebastian Walter[1], Max Argus[1], Hannah Bast[1], Thomas Brox[1]
[1]Department of Computer Science, Faculty of Engineering, University of Freiburg, Freiburg, Germany

Background

Contrastive Language-Image Pretraining (CLIP) has emerged as a strong way to learn high-quality visual features by training on pairs of images and their corresponding text descriptions collected from the web. These features can either be used directly, e.g. for Zeroshot Classification or Text-to-Image retrieval, or as input to a Large Language Model (LLM) to enable dialogs with visual input.

However, training such a model is prohibitively expensive, since it requires large amounts of data and compute (around 12 days using 256 GPUs with 32GB for training a large Vision Transformer).

Methods

In this project, we investigate training such a model with significantly less compute. To simplify the task, we train a specialized model for understanding animals and plants.

We build a dataset of around 5 million image-text pairs to enable understanding unseen recombinations of the training input. To achieve this, we define our entities of interest using knowledge graphs (WordNet and WikiData) and define visual attributes using LLMs.

We use an image search to mine image-text pairs concerning these specific entities and entity-attribute combinations. To help resolving ambiguous synonyms, we cluster the images and discard clusters with information that is unrelated to the search query. Finally we rate and clean the descriptions using LLMs.

Results

Initial results on pretraining a CLIP model using our dataset show that we can achieve similar performance to a model trained on noisy web-data, while using an order of a magnitude less data.

Additionally, we finetune an existing pretrained CLIP model and significantly improve its specialized knowledge for animals and plants, while retaining its pretrained knowledge in other domains.

Conclusions

In conclusion, we achieve an important step in training a VLM on small data by showing strong preliminary results on our newly created image-text dataset.

Comparing the performance of open and close sourced Large Language Models for automatic CAD-RADS 2.0 classification from cardiac computer tomography radiology reports (CRC B05)

Philipp Arnold[1], **Maximilian Russe**[1], **Muhammad Taha Hagar**[1], **Elmar Kotter**[1]

[1]Department of Radiology, Faculty of Medicine and Medical Centre, University of Freiburg, Freiburg, Germany

Background

Coronary Artery Disease-Reporting and Data System (CAD-RADS) 2.0 provides standardized reporting for coronary artery disease in cardiac computer tomography (CT) reports. Accurate and consistent scoring is essential for clinical decision-making. This study explores the potential of large language models (LLMs) in generating CAD-RADS 2.0 scores from synthetically created cardiac CT reports.

Methods

A dataset of 200 synthetically created cardiac CT reports was generated and used to evaluate the performance of several state-of-the-art LLMs in generating CAD-RADS 2.0 scores from in context learning. The models tested included GPT-3.5, GPT-4o, Mistral 7b, Mixtral 8x7b, and LLama3 8b, Llama3 8b with 64k context length and Llama3 70b. The accuracy of each model was assessed by comparing the generated scores to the ground truth.

Results

The performance of the LLMs varied, with GPT-4o achieving the highest accuracy at 97.5%, followed by LLama3 70b at 97%, LLama3 8b at 96%, Mixtral 8x7b and GPT 3.5 at 95%, Mistral 7b at 93.5%, and LLama3 8b with 64k context at 86.5%. Overall, the models demonstrated high levels of accuracy in generating CAD-RADS 2.0 scores, indicating their potential utility in clinical settings.

Conclusions

GPT-4o exhibited superior performance in generating CAD-RADS 2.0 scores from synthetic cardiac CT reports, followed closely by the open source models LLama3 70b and Mixtra 8x7b. The results highlight the efficacy of advanced LLMs in automating the scoring process, which could enhance the efficiency and consistency of cardiac CT report evaluations. As open-source models that can be hosted locally and thus avoid data privacy concerns demonstrate comparable performance to state-of-the-art models from OpenAI, further research is possible to validate these findings using real-world clinical data using open-source models.

Methodologies to improve the scope and accuracy of whole-body models of human metabolism (CRC B06)

Daniel Fässler[1], Chenglong Huang[1], Jordi Roma[2], Nora Scherer[3,4], Anna Köttgen[3,5], Johannes Hertel[1]

[1]Department of Psychiatry and Psychotherapy, University Medicine Greifswald, Greifswald, Germany

[2]University of Lorraine, Vandoeuvre-les-Nancy, France

[3]Institute of Genetic Epidemiology, Faculty of Medicine and Medical Center - University of Freiburg, Freiburg, Germany

[4]Spemann Graduate School of Biology and Medicine (SGBM), University of Freiburg, Freiburg, Germany

[5]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA

Background

Whole-body models (WBMs) of human metabolism are comprehensive, organ- and sex-specific representations of human metabolism built on the formalism of constraint-based reconstruction and analysis (COBRA). By translating accumulated knowledge of gene functions into mathematical form, WBMs enable the investigation of human metabolism in a condition-specific and holistic manner. Nevertheless, these models are limited by our current understanding of gene functions, predetermined constraints, and the chosen optimization strategy to investigate the large variety of possible states. Furthermore, they focus on metabolism, overlooking other crucial biological processes, such as signal transduction and regulation.

Methods

We integrated microbiome community models into WBMs and computed unique flux distributions for each microbiome-personalized wild-type and *KYNU* gene-knockout model. We then calculated pseudo-effect sizes of in silico gene knockouts on urinary secretion fluxes across the population of personalized whole-body models. Secondly, we incorporated regulatory network constraints to develop a comprehensive multi-scale modeling framework. Additionally, we enhance the kidney model in the current WBM by integrating single-cell omics data to capture the unique metabolism of different kidney cell types and improve model resolution.

Results

Quantitative metabolite changes resulting from the loss of *KYNU* gene function, as observed in gene-based tests for rare, deleterious variants in the German Chronic Kidney Disease cohort (GCKD), align with effect sizes from in silico *KYNU*gene-knockouts in microbiome-personalized WBMs. In-silico simulations of more than 50 tissues, integrated with regulatory networks, showed more accurate tissue groupings and clearer differentiation than non-regulatory simulations.

Conclusions

Using in vivo metabolome and whole exome sequencing data from the GCKD cohort and patients with inborn errors of metabolism, the in silico gene-knockout approach, utilizing personalized microbiome models, can validate metabolite-gene associations identified by gene-based tests of rare variants. Integrating regulatory network constraints, and cell-specific metabolism into WBMs may enhance the precision and applicability of WBMs.

ID 123

Coupling of metabolomics and exome sequencing reveals graded effects of rare damaging heterozygous variants on gene function and human traits and diseases (CRC 806)

Nora Scherer[1,2], Daniel Fässler[3], Oleg Borisov[1], Kai-Uwe Eckardt[4,5], Matthias Wuttke[1], Johannes Hertel[3,6], Anna Köttgen[1,7]

[1]Institute of Genetic Epidemiology, Faculty of Medicine and Medical Center - University of Freiburg, Freiburg, Germany

[2]Spemann Graduate School of Biology and Medicine (SGBM), University of Freiburg, Freiburg, Germany

[3]Department of Psychiatry and Psychotherapy, University Medicine Greifswald, Greifswald, Germany

[4]Department of Nephrology and Hypertension, University Hospital Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

[5]Department of Nephrology and Medical Intensive Care, Charité - Universitätsmedizin Berlin, Berlin, Germany

[6]School of Medicine, University of Galway, Galway, Ireland

[7]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA

Background

Rare variants in genes encoding for enzymes or transporters can result in changes of levels of their metabolic substrates/products and potentially cause severe inborn errors of metabolism (IEMs). Studies of IEMs are limited due to the very low number of homozygous carriers of causative variants. We hypothesized that exome-wide genetic studies of paired plasma and urine metabolomes on a population level can reveal metabolite-associated genes and variants, and establish connections to related diseases.

Methods

We performed gene-based burden tests focusing on the aggregated effect of rare, putatively damaging variants identified by whole-exome sequencing on the levels of >1,294 plasma and urine metabolites from 4,737 GCKD study participants. Rare qualifying variants (QVs) were selected informed by their predicted consequences and individual QV contributions were prioritized via forward selection. Identified gene-metabolite associations were integrated with health outcomes and compared to *in silico* gene-knockouts in whole-body models (WBMs) of human metabolism.

Results

We identified 192 significant (P<5.04e-9) gene-metabolite associations involving 73 genes. Metabolite-associated QVs were almost exclusively present in the heterozygous state, but nevertheless informative about the gene's function and corresponding metabolic effects. First, 38% of identified genes harbor known causative variants for recessive IEMs. Second, measurements in homozygous IEM patients (*KYNU*, *PAH*) reflected metabolic changes observed in heterozygous carriers. Third, male carriers of QVs in X-chromosomal *TMLHE* showed significantly more extreme levels of the encoded enzyme's substrate and product than female carriers, reflecting hemi- versus heterozygosity. Fourth, allelic series of functional QVs in plasma sulfate-associated *SLC13A1* and *SLC26A1* provided links to related musculoskeletal phenotypes. Last, *in silico* gene-knockouts in WBMs were predictive for observed metabolic changes based on heterozygous carriers.

Conclusions

This genetic study identified known and novel gene-metabolite relationships and demonstrated that aggregated metabolite-associated rare heterozygous damaging variants inform about a gene's function and capture metabolic effects.

Being certain of uncertainty (CRC C01)

Lukas Hoffmann[1,2], Ferdinand Suchanek[1], Carola Heinzel[3], Maria de la Puente[4], Christopher Philipps[4,5], Peter Pfaffelhuber[3], Sabine Lutz-Bonengel[1]

[1]Institute of Forensic Medicine, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany

[2]Faculty of Biology, University of Freiburg, Freiburg, Germany

[3]Institute of Mathematics, Department of Mathematical Stochastics, University of Freiburg, Freiburg, Germany

[4]University of Santiago de Compostela, Instituto de Ciencias Forenses "Luis Concheiro", Santiago de Compostela, Spain

[5]King's Forensics, Department of Analytical, Environmental and Forensic Sciences, King's College, London, UK

Background

While estimating the phenotype of a potential suspect is legalised in the field of so-called 'extended forensic DNA analysis' in Germany (§81e STPO) since 2019, the estimation of the biogeographic ancestry is still not permitted. To address prospective questions about the uncertainty in the estimation of the biogeographic ancestry several, already existing panels (VISAGE Basic Tool, Verogen ForenSeq DNA Signature Kit, The Force Panel AIMs subset, and TFS Precision ID Ancestry Panel) were compared to evaluate their performance in distinguishing continental and sub-continental populations. Migrants from the Middle East show the second highest proportion of migrants in Germany. Therefore, particular attention was centred on the differentiation between Europe and the Middle East.

Methods

Divergence, Informativeness I_n, and Delta Δ are key traits of genetic markers in population genetics. Divergence reflects genetic accumulation of two populations over time, while I_n and Δ assess a genetic marker's ability to differentiate two populations. With the Naïve-Bayes classifier SNIPPER, the divergence of genetic markers was estimated which enabled the calculation of I_n and Δ to evaluate marker performance. By solving an optimization problem, we evaluated an in-house in silico panel, called FAME (Freiburg Panel for admixed individuals) to enhance distinction of admixed individuals. This panel, along with established ancestry panels, was examined using STRUCTURE with admixed individuals to compare their similarity.

Results

By comparing these panels, we showed that the performance in distinguishing continental and sub-continental regions differs and that there is no panel able to cover all regions equally well. We present our results and outline the marker assessment algorithm with which we compiled FAME.

Conclusions

We evaluated the performance of several panels for estimating the biogeographic ancestry. With the obtained data we could evaluate an own in silico panel called FAME (Freiburg panel for admixed populations). We present our results and the marker assessment algorithm with which we compiled FAME.

Locally stationary hidden Markov models (CRC C01)

Angelika Rohde[1], Gerard Mesuere[1]

[1]Department of Mathematical Stochastics, Faculty of Mathematics and Physics, University of Freiburg, Freiburg, Germany

Background

Hidden markov models are frequently used to analyse data and in particular genetic data. An extensive mathematical theory is available for the case where the underlying Markov chain is homogeneous. In the inhomoegeneous case however, this mathematical theory is currently less extensive. Our overall aim is to develop a theory for locally stationary hidden markov models. In such a setting the non-stationary process of interest can be locally approximated by a stationary one. Therefore, only a limited number of observations is available for the maximum likelihood estimate (MLE) at each point in time, rendering the corresponding task a small data problem.

Methods

As to our methods, apart from the literature on the maximum likelihood estimation for hidden markov models we rely on the literature where a theory for locally stationary processes is developed by Rainer Dahlhaus.

Results

So far it has become evident that some preliminary fundamental results from the homogeneous case such as the recursive representation of the conditional distribution also hold in the inhomogeneous setting. Furthermore, the SISR algorithm appears to be well behaved under conditions of inhomogeneity.

Conclusions

Therefore, we believe it could be possible to develop algorithms which can be used to compute the MLE approximately, once the mathematical theory has been established. The latter will hopefully not only include a result on the consistency of the MLE but also a limiting distribution and a good rate of convergence.

ID 128

Efficacy of psychotherapy, pharmacotherapy, or their combination in chronic depression: a systematic review and network meta-analysis using aggregated and individual patient data (CRC C02)

Julia Müller[1], Moritz Elsaesser[1], Nana-Adjoa Kwarteng[2], Theodoros Evrenoglou[2], Pim Cuijpers[3,4], Orestis Efthimiou[5], Daniel N. Klein[6], Martin B. Keller[7], Toshi A. Furukawa[8], Adriani Nikolakopoulou[2], Elisabeth Schramm[1]

[1]Department of Psychiatry and Psychotherapy, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany
[2]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany
[3]Department of Clinical, Neuro and Developmental Psychology, WHO Collaborating Centre for Research and Dissemination of Psychological Interventions, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
[4]Babeș-Bolyai University, International Institute for Psychotherapy, Cluj-Napoca, Romania
[5]Institute of Primary Health Care (BIHAM), University of Bern, Bern, Switzerland
[6]Department of Psychology, Stony Brook University, Stony Brook, New York, USA
[7]Department of Psychiatry and Human Behavior, The Warren Alpert Medical School of Brown University, Providence, USA
[8]Department of Health Promotion and Human Behavior, Kyoto University Graduate School of Medicine / School of Public Health, Kyoto, Japan

Background

Chronic depression represents a common, highly disabling disorder. Several randomized controlled trials investigated the effectiveness of psychological, pharmacological, and combined treatments for chronic depression. This is the first overarching systematic review and network meta-analysis based on aggregated and individual patient data (IPD-NMA) comparing the efficacy and acceptability of various treatment options for all subtypes of chronic depression.

Methods

We searched Cochrane Library, MEDLINE via Ovid, PsycINFO, Web of Science, and metapsy databases to identify randomized controlled trials (RCTs) that investigated treatment effects in adults with a primary diagnosis of chronic depression. The main outcome is observer-rated depression severity at six months post treatment (range 3-12 months). Two reviewers independently screened and selected eligible studies based on the pre-defined inclusion and exclusion criteria. Risk of bias will be assessed using Version 2 of the Cochrane risk-of-bias tool for randomized trials (Rob 2.0). Individual patient data (IPD) will be requested and used for the network meta-analysis (NMA) when provided. If not available, aggregated data (AD) will be extracted and incorporated in the network. An NMA comparing psychotherapies and a network meta-regression (NMR) estimating individualized treatment effects of psychotherapy will be implemented assuming a Bayesian framework. All models will be fitted in R with calls to JAGS. Empirical informative prior distributions will be used for model parameters where available. If not available, non-informative priors will be used.

Results

We screened 3,490 studies for eligibility, selecting 192 for full-text review. To date, we have reviewed 47 studies, with 25 deemed eligible for inclusion in our analyses.

Conclusions

This systematic review and meta-analysis aims to clarify and quantify the impact of various treatment options and the role of predictive and modifying factors in chronic depression, providing insights that can inform personalized treatment strategies for patients with chronic depression.

1000+ synthetic benchmark problems for parameter estimation in dynamic modeling (CRC C03)

Niklas Neubrand[1,3], **Yaser Kord**[2,3], **Jens Timmer**[2,3,4], **Clemens Kreutz**[1,3,4]

[1]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Centre, University of Freiburg, Germany

[2]Faculty of Mathematics and Physics, Institute of Physics, University of Freiburg, Germany

[3]Freiburg Center for Data Analysis and Modeling, University of Freiburg, Germany

[4]Centre for Integrative Biological Signalling Studies (CIBSS), University of Freiburg, Germany

Background

Systems Biology seeks to find a mechanistic understanding of intracellular processes by modeling the dynamics of bio-chemical species through Ordinary Differential Equations (ODEs). The best-fit parameters of the models, e.g. rate constants or initial concentrations, are estimated from experimental data. Even with the large set of existing tools, this optimization can be challenging numerically due to limitations of the ODE solvers, non-linearity of the models and limited observations. Our project aims to learn an optimal optimization strategy by a reinforcement learning approach. This requires a comprehensive set of parameter estimation problems for training and testing which can also be useful for benchmarking existing strategies.

Methods

To generate novel parameter estimation problems, we use a set of 22 published ODE models and replace the original experimental data with realistic synthetic data. Here, we build on a method by Egert and Kreutz for drawing realistic observables and creating time-course data based on the simulated model dynamics. We extend this method to distribute the drawn observables realistically to the experimental conditions of the original problems. This way, multiple different views on the underlying system are included in the data, which resembles real experimental design more closely.

Results

For each of the 22 pre-existing problems, we generated 50 synthetic data realizations, resulting in 1100 synthetic parameter estimation problems. We present the distributions of relevant characteristics, e.g. the number of data points and parameters. Here, we observe that the synthetic problems substantially extend not only the size but also the variability of our problem set.

Conclusions

The creation of 1100 synthetic parameter estimation problems provides a substantial resource for improving and benchmarking optimization methods in dynamic modeling. One possible approach to this is our research in project C03. These efforts will ultimately remove computational limitations and advance the field of Systems Biology.

ID 130

Enhancing SNLS optimisation via deep reinforcement learning for adaptive tolerance setting (CRC C03)

Yaser Kord[1,3], **Niklas Neubrand**[2,3], **Baohe Zhang**[5], **Clemens Kreutz**[2,3,4], **Jens Timmer**[1,3,4]

[1]Faculty of Mathematics and Physics, Institute of Physics, University of Freiburg, Germany

[2]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Centre, University of Freiburg, Germany

[3]Freiburg Center for Data Analysis and Modeling, University of Freiburg, Germany

[4]Centre for Integrative Biological Signalling Studies (CIBSS), University of Freiburg

[5]Department of Computer Science, Faculty of Engineering, University of Freiburg

Background

The Sparse Nonlinear Least Squares (SNLS) optimiser is widely used for parameter estimation in Ordinary Differential Equations (ODEs), particularly within biomedical applications. Setting appropriate tolerance values during the SNLS optimisation process is crucial for convergence speed and accuracy. Traditional approaches using fixed tolerance values can lead to suboptimal performance. Systems Biology seeks to achieve a mechanistic understanding of intracellular processes by modeling the dynamics of biochemical species through ODEs. This project explores the use of Deep Reinforcement Learning (DRL) to dynamically adjust these tolerance settings, aiming to enhance SNLS optimiser performance.

Methods

We integrated a DRL agent, implemented in Python, with the SNLS optimiser coded in MATLAB as part of the Data2Dynamics software suite. The DRL agent functions by adjusting the tolerance settings at each iteration of the SNLS optimisation process. Specifically, we employed the Proximal Policy Optimisation (PPO) algorithm with Long Short-Term Memory (LSTM) networks to handle the delayed feedback inherent in the SNLS optimisation process. The agent was trained in an episodic manner, where each episode involved running the SNLS optimiser, providing the agent with feedback on its tolerance adjustments. The state representation for the DRL agent included metrics such as current error and error reduction rate.

Results

Our experimental results demonstrate that the DRL-enhanced SNLS optimiser achieves improved performance compared to traditional fixed-tolerance methods. Specifically, the DRL-driven approach resulted in faster convergence and higher accuracy in parameter estimation tasks. The adaptive tolerance settings recommended by the DRL agent enabled the SNLS optimiser to balance convergence speed and solution precision effectively.

Conclusions

The integration of DRL for adaptive tolerance setting in the SNLS optimisation process presents a notable improvement in parameter estimation for ODEs. This approach not only accelerates convergence but also enhances the accuracy of the solutions obtained, making it a valuable tool for complex biomedical modeling.

Generating optimal small datasets for efficient offline reinforcement learning training (CRC C04)

Noor Awad[1], M Asif Hasan[1]

[1]Department of Computer Science, Faculty of Engineering, University of Freiburg, Freiburg, Germany

Background

Offline reinforcement learning (RL) is crucial in scenarios where real-world interactive learning is impractical or expensive. Unlike online RL, where agents learn through continuous interaction with the environment, offline RL relies on pre-collected datasets. This method mitigates risks and reduces costs. However, these are typically large datasets, necessary for effective RL agent training, pose computational and cost challenges.

To address these challenges, constructing the smallest possible dataset that ensures high trajectory quality (TQ) and state-action coverage (SACo) is essential. Smaller datasets reduce computational burden, costs, and enable faster training cycles. Generating small, high-quality datasets helps maintain the benefits of offline RL while improving its practicality and efficiency, making it more feasible for real-world applications.

Methods

Our approach involves generating an offline RL dataset through multiple runs. Initially, an agent is trained with an RL algorithm, and its performance is evaluated to ensure it meets satisfactory criteria. After achieving satisfactory performance, datasets of various sizes are collected at different time points and assessed based on trajectory quality (TQ) and state-action coverage (SACo) scores. The high-performing datasets are then merged and iteratively refined to improve these scores.

The refined datasets are evaluated using an offline RL algorithm to assess the performance of policies learned from these offline data without further environmental interaction. This process aims to identify the smallest dataset containing optimal state-action pairs, ensuring efficient and effective training.

Results

Preliminary results indicate that our process successfully created minimal-sized datasets containing a maximum number of unique state-action pairs. These datasets enable faster, more cost-effective offline RL training, achieving performance comparable to or better than larger datasets.

Conclusions

Constructing minimal, high-quality datasets for offline RL training is feasible and can significantly enhance training efficiency while reducing computational costs. We will use these methods in developing the benchmark for HPO for our task.

Exploration cocktail: Automating exploration in reinforcement learning (CRC C04)

Baohe Zhang[1], Shengchao Yan[1], Raghu Rajan[1], André Biedenkapp[1], Joschka Bödecker[1]
[1]Department of Computer Science, Faculty of Engineering, University of Freiburg, Freiburg, Germany

Background
From Atari game, Go to autonomous driving, Reinforcement Learning (RL) has shown its potential to solve complicated tasks. However, exploration still plays the essential role when applying RL to new tasks. Gaussian noise is usually added to RL policy, but this may not work across different tasks when multi-modal actions are preferred.

Methods
We propose to use population-based training (PBT) to automate the exploration noise selection. A group of different exploration strategies are initialized differently and then combined together. Each member in the PBT will be assigned a different exploration policy to discover different data regimes. We then aggregate all collected data for policy improvement. With mutation and exploitation, PBT allows to learn automatically the best exploration strategy for a given task. We also provided a vectorized PBT version that allows training PBT in a single machine.

Results
Our results show that exploration cocktail can improve the data efficiency of the existing RL algorithms, both for on-policy and off-policy methods. We further show that, exploration cocktail can significantly improve the performance in tasks that contains multiple local minimum and are hard to explore.

Conclusions
We created a universal exploration strategy that allows RL to solve more difficult tasks in a handful trials. Our method also shows the insight of different exploration strategies are necessary in order to solve complicated tasks.

Empirical assessment of paradigms in tabular classification (CRC C05)

Guri Zabërgja[1], Arlind Kadra[1], Josif Grabocka[2]

[1]Department of Computer Science, University of Freiburg, Freiburg, Germany

[2]Department of Computer Science, University of Technology Nuremberg, Germany

Background

Deep Learning has revolutionized the field of machine learning, becoming the go-to method for various tasks. Despite the rapid advancements in deep learning methods, which have shown unmatched performance in various subfields of machine learning, the realm of tabular data remains unconquered, with the most widely used methods not being deep learning-based. In this work, we empirically assess the performance of existing paradigms in tabular classification tasks, including recent foundation models for tabular data.

Methods

In our study, we empirically compare various methods designed for tabular data. We evaluate these methods on the OpenMLCC18 benchmark, which is widely used in the community. Our comparison includes gradient-boosted decision trees such as CatBoost and XGBoost, neural networks like ResNet, and attention-based models such as FT-Transformer, SAINT, and TabNet. We also assess the performance of recently proposed foundation models, including TabPFN, XTab, TPBerta, and CARTE. Additionally, we include AutoGluon, recognized as one of the best AutoML systems for tabular data.

Results

The results of our study, highlight that AutoGluon is the top-performing method across all datasets, followed by XGBoost and CatBoost, indicating their robustness for diverse tabular data tasks. For smaller datasets (≤ 1000 rows), TabPFN demonstrates better performance, showing its effectiveness in small data scenarios.

Conclusions

In this study, we empirically compared various methods for tabular data classification. AutoGluon emerged as the top performer across all datasets, while TabPFN excelled in smaller datasets. This evaluation provides valuable insights for practitioners aiming to optimize their approach to tabular data tasks.

Applying a foundation model to small tabular data (CRC C05)

Lennart Purucker[1], **Frank Hutter**[1,2]
[1]Machine Learning Lab, University of Freiburg, Freiburg, Germany
[2]ELLIS Institute Tübingen, Tübingen, Germany

Background
Small tabular data, spreadsheets organized in rows and columns, is omnipresent across scientific fields. A widespread application for tabular data is using informative columns to learn to predict unobserved values for a target column. For example, medical professionals can use tabular data in the form of patient characteristics in computer-aided disease diagnosis. Yet, when faced with a small number of observations (e.g., patients), traditional learning algorithms struggle to yield meaningful predictions.

Methods
Therefore, we propose to apply TabPFN, a foundation model, to small tabular data. TabPFN is a deep learning model pre-trained on 100 million synthetic datasets. When applied to a real-world dataset, TabPFN demonstrates its unique capability to effectively learn from a small number of observations by transferring knowledge from its extensive pre-training. Moreover, TabPFN offers additional foundation model abilities, such as data generation, data embeddings, and support for fine-tuning to a specific task.

Results
We show that TabPFN outperforms all other state-of-the-art algorithms for a wide range of small tabular datasets. In addition, we created an automated interface for using TabPFN for small tabular data.

Conclusions
For small tabular data, using the foundation model TabPFN not only yields better predictions but also offers more abilities for scientific discovery than traditional algorithms. This superiority of TabPFN is a testament to its effectiveness and potential for small tabular data.

An end-to-end modeling approach for capturing spatiotemporal patterns in two-photon imaging data (CRC F)

Fabian Kabus[1], Maren Hackenberg[1], Thibault Cholvin[2], Antje Kilias[2], Harald Binder[1], Marlene Bartos[2]

[1]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany

[2]Institute for Physiology I, University of Freiburg, Medical Faculty, Freiburg, Germany

Background

In vivo two-photon calcium imaging (2PCI) is a widely used method to measure physiological activity in neurons. Data processing pipelines for 2PCI recordings typically rely on steps, registration, region of interest (ROI) segmentation, extraction of neuronal traces, followed by modeling of auto- and cross- correlation. Consequently, the performance of each step depends on the outcome of the previous steps. As a result manual tuning and repeated execution of each step are required, and information might get lost. Therefore an end-to-end approach that does not require segmentation and extraction might be attractive.

Methods

Instead of splitting the analysis into multiple independent steps, we develop an artificial neural network approach that operates at the level of the original temporal sequence of images. Specifically, our proposed neural network architecture comprises an outer convolutional autoencoder that learns a compressed representation of the input video. We impose a statistical model on the resulting inner dimension-reduced representation. To obtain parameters that reflect auto- and cross-correlation, we fit a vector autoregressive model (VAR) to the latent time series. We regularize the VAR parameters during training to encourage the emergence of structure in the latent space.

Results

We evaluate our approach by visualizing the VAR parameters. Our proposed approach is seen to have pattern-finding capabilities on a 2PCI datasets. The model captures spatiotemporal features, showcasing its adaptability. Importantly, the integrated architecture minimizes the need for ROI curation, streamlining the analysis process.

Conclusions

We showcase an integrated method to characterize 2PCI recordings. By infusing knowledge-based constraints into an autoencoder with a latent VAR model, we overcome the limitations of traditional approaches. The demonstrated adaptability to the dataset underscores the generalizability of our approach. As we move towards a more automated methodology, our holistic approach stands out as a valuable tool for advancing *in vivo* two-photon calcium imaging studies.

Small data meets high dimensions: Some approaches from multiple testing

Sebastian Döhler[1], **Etienne Roquain**[2]

[1]Faculty of Mathematics and Sciences, Darmstadt University of Applied Sciences, Darmstadt, Germany

[2]Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Universite, Paris, France

Background

Multiple testing issues frequently arise in the analysis of high-dimensional data e.g. in astrophysics, biology or economics where thousands or millions of statistical tests are performed simultaneously. A large body of research deals with such problems by developing methods that control error rates like the false discovery rate (FDR). When the sample size is small, exact statistical tests like Fisher's exact test are often used for analysing such data. This typically leads to p-value distributions that are discrete and conservative (under the null hypothesis). Classical methods like the Benjamini-Hochberg procedure, which are based on uniformly distributed p-values, can still be applied to such small but high-dimensional data, however the conservatism inherent in the (single) tests, may make these methods inefficient.

Methods

We describe some mathematical approaches that make classical multiple testing methods more powerful for discrete data. The main idea is to incorporate the distributional information on the discrete p-values in order to obtain stochastically tighter bounds on error rates like the FDR.

Results

We give an overview over several methods that were recently developed for controlling various error rates. We also present some real data results that illustrate the benefit of taking discreteness into account.

Conclusions

The work presented here is part of an ongoing project that aims to provide statisticians and data scientists with improved mathematical methods and software tools for analysing data that is both small (in the sense of discreteness) and high-dimensional.

Challenges of small data in biomedical and environmental research

Timothy E. O'Brien[1]

[1]Department of Mathematics and Statistics, School of Environmental Sustainability, Loyola University Chicago, USA

Background

Biomedical, toxicological and environmental/ecological researchers often base their preliminary go/no-go decisions on very limited sample sizes due to cost and other imposed constraints. Additionally, these researchers often find that nonlinear models better fit their system and data than linear ones, and this small data-nonlinear combination can often yield inaccurate and/or inconclusive estimation and assessment decisions.

Methods

Our proposed hybrid strategy is one of targeting key research parameter(s) and optimal experimental strategies before engaging in experimentation and data acquisition. This methodology has been tested on and confirmed with several real-data situations; when prior experience is available, it is easily extended to include a Bayesian approach. Importantly, when model uncertainty is present, this uncertainty can easily be built in to the data acquisition process.

Results

Although at the initial stage, our results demonstrate increased power over conventional methods in the 20%-40% range. Further, software programs to implement our methods are made available in both R and SAS, thereby providing the end-user with the means to implement these methods in their practical work.

Conclusions

Our preliminary results suggest that integrating robust optimal design methods (via use of strategic hyperparameters) can alleviate some of the small-sample challenges in nonlinear modeling. Ongoing and future research will continue to focus on improving our methods and exploring the use of AI in enhancing our methods.

A meta unit for co-constructing a computational scaffold model to guide human motor learning

Alexandra Moringen[1]
[1]Institute for Data Science, University of Greifswald, Greifswald, Germany

Background

Learning a dexterous motor skill, such as necessary in medical training, sports or playing a musical instrument is time-costly and needs supervision from a teacher, e.g. to avoid injury. Learning a motor skill requires active practice, adjusting one's own individual embodiment to the target movement. The questions that we are interested in exploring are: How to balance self-guided exploratory practice with teacher-guided practice? How to improve beyond the performance of the teacher? How to adjust the teacher strategy that has been optimized by teacher to suite their embodiment to one's own embodiment through exploration?

Methods

We have used Gaussian Process (GP) to model the teacher policy. The model was trained to offer suitable practice units to a learner, given their state and some controllable parameters of the target task. Now we are interested to explore a further optimization of the above policy within an epsilon greedy Q-learning framework. Here the human learner, the Q-learner and the teacher represented by a GP co-construct a policy that optimizes human learner's individual long-term progress. The idea of this poster is to discuss different alternatives for an implementation of such a meta unit that coordinates the above mentioned interactive co-construction by the three actors.

Results

In the current setup tested only in simulation, the meta unit samples from a distribution that is being parameterized by the average reward achieved by the human learner during practice. A low reward makes sampling from the teacher policy more likely and results in a teacher-guided practice, while a high reward results in a higher probability of a Q-learner-driven practice.

Conclusions

We are interested in exploring different options for the co-construction coordinated by the meta unit, as well as different experimental settings for guided learning.

OptAB - an optimal antibiotic selection framework for sepsis patients with artificial intelligence

Philipp Wendland[1], Christof Schenkel-Häger[2], Ingobert Wenningmann[3], Maik Kschischo[1,4]

[1]University of Applied Sciences Koblenz, Department of Mathematics and Technology, Remagen, Germany

[2]University of Applied Sciences Koblenz, Department of Economics and Social Studies, Remagen, Germany

[3]University Hospital Bonn, Department of Anesthesieology and Operative Intensive Care Medicine, Bonn, Germany

[4]University of Koblenz, Department of Computer Science, Koblenz, Germany

Background

Current methods for forecasting disease progression or optimizing treatments typically often rely solely on initial measurements taken at disease or treatment onset. In a hospital setting, patients are monitored over time. Consequently, treatment optimization models should assimilate these data to update their predictions (online-updateable data assimilation). A representative example is the optimal selection of antibiotics for sepsis,a leading cause of avoidable deaths worldwide. The antibiotic treatment of Sepsis is challening, because the initial treatment decision has to be made quickly, but results for microbiological cultures are usually only available with a delay of two or three and in more than 30% of all patients no pathogen is detected at all.

Methods

We propose OptAB, the first completely data-driven online-updateable optimal antibiotic selection model for real-world Sepsis patients. OptAB is based on Treatment-Effect Controlled Differential Equations and aims to minimize the Sepsis-related organ failure score (SOFA-Score) as a measure for treatment success while accounting for antibiotic-specific side-effects. OptAB can handle the special characteristics of patient data including irregular measurements, a large amount of missing values and time-dependent confounding.

Results

OptAB learns realistic treatment influences for (combinations of) the antibiotics Vancomycin, Ceftriaxone and Piperacillin/Tazobactam on the SOFA-Score and the side-effect indicating laboratory values creatinine, bilirubin total and alanine-transaminase. OptAB's selected optimal antibiotics exhibit faster efficacy than the administered antibiotics. We provide evidence, that OptAB captures the toxic side effects and contraindications, which are important for treatment decisions.

Conclusions

The results suggest that dynamic models like OptAB can be used to optimize the antibiotic selection for individual patients, based on their specific characteristics. Such dynamic treatment regimes can potentially lead to better Sepsis-outcomes, shorter stays on ICU and less side effects. Further research will focus on optimizing dosing regimes and refining the interpretability of OptAB.

Denoising of low dimensional EEG data with deep learning for improved seizure detection

Matthias Dümpelmann[1,2], Romina Roshani[2], Sotirios Kalousios[1], Andreas Schulze-Bonhage[1]

[1]Department of Neurosurgery, Epilepsy Center, Medical Center - University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany

[2]Department of Microsystems Engineering (IMTEK), University of Freiburg, Freiburg, Germany

Background

New implantable subcutaneous EEG devices allow the recording of EEG activity over months using minimal invasive surgical approaches. These devices are restricted to record EEG with a few channels and exhibit physiologic artifacts, such as interferences from muscle activities (electromyographic (EMG) noise) and eye blinks (electrooculographic (EOG) noise), comparable to scalp EEG. Such artifacts hamper both automatic detection and human review of epileptic seizures.

Methods

Recordings of 182 sleep-related subclinical seizures from 50 patients were annotated by expert epileptologists and formed the basis for training the denoising algorithms. Two derived data sets were created by adding muscle activity and activity from eye blinks from a public database. Three different autoencoder architectures, an autoencoder with GRU unit, a CNN and a LSTM model were trained to clean these data. The models were applied to a separate dataset including 358 seizures representing a four channel electrode layout of a neurostimulation device.

Results

Among all the models, LSTM and CNN show the best performance in removing EOG contaminations in terms of signal-to-noise ratio, peak signal-to-noise ratio, root mean square error, and Pearson correlation. Autoencoder-based models exhibit the best performance when employed for removing EMG noise. Signal denoising with deep learning approaches outperformed the application of traditionally used linear filters. Improvement of the seizure detection results was moderate, but especially the amount of false detections could be reduced.

Conclusions

Two expert annotated EEG datasets and a public dataset containing typical artifacts in EEG signals were the basis of the study and the creation of models to clean EEG data. Application of the models achieved a moderate improvement in seizure detection for a new class of implants for patients with epilepsy. Access to high quality and expert annotated EEG datasets was crucial for the implementation of the study.

Forward-forward optimization in small data

Andrii Krutsylo[1]
[1]Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

Background
The Forward-Forward (FF) optimization algorithm, proposed by Geoffrey Hinton, offers an alternative to traditional optimization methods like Stochastic Gradient Descent (SGD), particularly in small data settings. This method utilizes a greedy multi-layer learning procedure inspired by Boltzmann machines and Noise Contrastive Estimation, replacing the backward pass of backpropagation with two forward passes, which can lead to improved performance with limited data.

Methods
The FF algorithm was tested against SGD on small subsets of the MNIST, FashionMNIST, and CIFAR-10 datasets. These subsets were either sampled randomly or selected using submodular optimization. The FF algorithm's unique approach involves separate forward passes for positive (real) and negative (generated) data, each adjusting weights to increase or decrease goodness, respectively. This method was compared to standard SGD in terms of accuracy and training efficiency.

Results
Experiments demonstrate that while the FF algorithm is slower than SGD, it significantly improves model accuracy, particularly with smaller datasets. The FF algorithm achieved up to a 6% increase in accuracy over SGD, with the difference being more pronounced as the number of training samples decreased. This suggests that FF can effectively leverage the limited data to improve model performance.

Conclusions
The FF algorithm offers a valuable alternative to traditional optimization methods like SGD, especially in small data settings. Its ability to outperform SGD in accuracy, despite its slower convergence, highlights its potential for applications where data is scarce. Future research will focus on optimizing the FF algorithm to reduce training time while maintaining or improving its accuracy benefits.

Similarity-based refinement of single-cell interactions

Niklas Brunn[1,2], **Manolo Blaufuß**[3], **Harald Binder**[1,2,4]

[1]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany

[2]Freiburg Center for Data Analysis and Modeling, University of Freiburg, Freiburg, Germany

[3]Mathematical institute, Faculty of Mathematics and Physics, University of Freiburg, Freiburg, Germany

[4]Centre for Integrative Biological Signaling Studies (CIBSS), University of Freiburg, Freiburg, Germany

Background

Emerging transcriptomics technologies have enhanced the study of cellular interactions by providing gene expressions of cells with spatial locations. At the same time, several computational tools have been developed to reconstruct communication events using prior knowledge of known signaling mechanisms. However, there is a lack of ground truth, which complicates the validation of predicted interactions by the existing approaches, especially for small cell groups. Therefore, inferred interactions between single cells or groups of cells may consist of many false or missing connections.

Methods

To this end, we develop an iterative approach that is able to refine a priori inferred cell interactions from transcriptomics data by coupling neural network-based dimensionality reduction with sparse additive models. Specifically, we assume that the expression of genes involved in active signaling processes can be predicted across individuals based on the expression of signaling genes in interaction partners, provided that the initial matching of cells is reasonably accurate. Thus, in each iteration, we fit a componentwise boosting model that, given the low-dimensional representation of gene expressions in sender cells, predicts the representation of expressions in currently matched receiver cells. The matching of interaction partners is then refined based on the similarities between the predicted and the original gene expression of the cells.

Results

As a proof of concept, we evaluate the approach on simulated gene expression data with varying signal-to-noise ratios, demonstrating its functionality and robustness. In real-world applications with pre-computed interactions, the proposed refinement algorithm was able to converge to a state, where re-matching has no further effect on the expression patterns across matched cells.

Conclusions

Single-cell interactions can be substantially refined based on present patterns in gene expression profiles of previously matched interaction partners. Furthermore, our refinement algorithm can help to reveal key genes involved in upstream and downstream signaling mechanisms.

Prediction of cell lineage trajectories by integration of small single-cell RNA datasets into a large reference dataset

Tanja Vogel[1], Johan Rollin[1], Maren Hackenberg[2]

[1]Institute for Anatomy and Cell Biology, Faculty of Medicine, University of Freiburg, Freiburg, Germany

[2]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany

Background

Single-cell sequencing technologies help increase knowledge about cell origins and interaction. In embryologic development, it helps to understand spatial and temporal relationships between cell types. Unfortunately, data scarcity limits the prediction ability of precise cell subtypes, especially with several cell maturation paths. The hippocampus subfield CA3 provides a good use case for studying this single-cell limitation.

Methods

We sequenced scRNA data from embryonic (E16.5) mouse hippocampus for which CA3 lineage could not be determined. Integration of that dataset into a larger reference dataset covering several brain structures (helping discover migration path) and several embryonic times should be performed. This should compensate for the data scarcity and help determine the lineage(s) leading to CA3 formation. Ideally, if an analysis with that level of precision within the diversity of the reference dataset is successful, the integration method can be reused to allow the automatic identification of yet unknown cell transient states (or subtypes) that are key for embryologic development in future datasets.

Results

The initial analysis demonstrates that CA3 formation cannot be resolved alone, and local integration does not group the CA3 cells together. Which may indicate a multiple origin for those cells. A larger scale integration may help us understand the context (spatial and temporal) for each possible origin resolving CA3 origin question.

Cell annotation usually uses gene markers linked to specific cell types. Those annotations are not efficient in all cell types, especially for embryologic studies. Our integration should allow projection of cells of interest in a reference dataset and then, transfer of reference cell annotations to the projected one, overcoming current annotation limitations.

Conclusions

Global integration of single-cell data should provide more cell-specific information than traditional analysis. Further work needs to be done on refining integration process, especially on cell-to-cell comparison and confidence in annotation transfer.

Characterizing the omics landscape based on 10,000+ datasets

Eva Brombacher[1,2,3,4] , **Oliver Schilling**[5,6,7] , **Clemens Kreutz**[1,2]

[1]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center-University of Freiburg, Freiburg, Germany

[2]Centre for Integrative Biological Signaling Studies (CIBSS), University of Freiburg, Freiburg, Germany

[3]Spemann Graduate School of Biology and Medicine (SGBM), University of Freiburg, Freiburg, Germany

[4]Faculty of Biology, University of Freiburg, Freiburg, Germany

[5]Institute for Surgical Pathology, Faculty of Medicine, Medical Center – University of Freiburg, Freiburg, Germany

[6]German Cancer Consortium (DKTK) and German Cancer Research Center (DKFZ), Heidelberg, Germany

[7]BIOSS Centre for Biological Signaling Studies, University of Freiburg, Freiburg, Germany

Background

The characteristics of data generated by omics technologies are crucial as they critically impact the effectiveness and feasibility of computational approaches used in downstream analyses, such as data harmonization and differential abundance analyses. Variability in these data characteristics across different datasets can lead to varying results in benchmarking studies, which are essential for selecting the appropriate analysis methods across all omics disciplines. Additionally, downstream analysis tools are developed and employed in specific omics communities due to presumed differences in data characteristics linked to each omics technology.

Methods

We analyzed over ten thousand datasets to examine how data characteristics differ among proteomics, metabolomics, transcriptomics, and microbiome data.

Results

Our study identified distinctive patterns in data characteristics for each omics type and introduced a tool that helps researchers evaluate how representative a given omics dataset is for its respective omics field. Moreover, we illustrate how data characteristics can impact analysis at the example of normalization in the presence of sample-dependent proportions of missing values.

Conclusions

Given the variability of these omics data characteristics, we recommend inspecting them in the context of benchmark studies and downstream analyses to avoid suboptimal method selection and unintended biases. Finally, in a small data setting with limited available datasets, knowledge of the characteristics of large omics datasets can be leveraged for transfer learning, where the similarity to known datasets is used for weighting.

Convex space learning for tabular synthetic data generation

Manjunath Mahendra[1], Chaithra Umesh[1], Saptarshi Bej[1,4], Kristian Schultz[1], Olaf Wolkenhauer[1,2,3]

[1]Department of Systemsbiology and Bioinformatics, Institute of Computer Science, University of Rostock, Rostock, Germany

[2]Leibniz-Institute for Food Systems Biology, Technical University of Munich, Freising, Germany

[3]Stellenbosch Institute for Advanced Study, South Africa

[4]Indian Institute of Science Education and Research, Thiruvananthapuram, India

Background

Generating synthetic samples from the convex space of the minority class is a popular oversampling approach for imbalanced classification problems. Recently, deep-learning approaches have been successfully applied to modeling the convex space of minority samples. Beyond oversampling, learning the convex space of neighborhoods in training data has not been used to generate entire tabular datasets.

Methods

We introduced a deep learning architecture (NextConvGeN) with a generator and discriminator component that can generate synthetic samples by learning to model the convex space of tabular data. The generator takes uniquely accessed data neighborhoods as input and creates synthetic samples within the convex space of that neighborhood. After that, the discriminator tries to classify these synthetic samples against a randomly sampled batch of data from the rest of the data space.

Results

We compared our proposed model with five state-of-the-art tabular generative models across ten publicly available datasets from the biomedical domain, out of which eight are smaller datasets with sample sizes ranging from 200 to 1000. Our analysis reveals that synthetic samples generated by NextConvGeN can better preserve classification and clustering performance across real and synthetic data than other synthetic data generation models. Synthetic data generation by deep learning of the convex space produces high scores for popular utility measures. We further compared how diverse synthetic data generation strategies perform in the privacy-utility spectrum and produced critical arguments on the necessity of high-utility models.

Conclusions

In conclusion, the performance of the NextConvGeN model is comparable to diffusion-based tabular generative models and surpasses that of generative adversarial networks and variational autoencoders in terms of utility measures. Our research on deep learning of the convex space of tabular data opens up new opportunities in clinical research, machine learning model development, decision support systems, and clinical data sharing.

Preserving logical and functional dependencies in synthetic tabular data

Chaithra Umesh[1], Kristian Schultz[1], Manjunath Mahendra[1], Saptarshi Bej[2], Olaf Wolkenhauer[1,3]

[1]Department of Systems Biology and Bioinformatics, Faculty of Computer Science and Electrical Engineering, University of Rostock, Germany

[2]Indian Institute of Science Education and Research, Thiruvananthapuram, India

[3]Stellenbosch Institute for Advanced Study, South Africa

Background

Dependencies among attributes are a common aspect of tabular data. However, whether existing synthetic tabular data generation algorithms preserve these dependencies while generating synthetic data is yet to be explored. Moreover, no well-established measures can quantify logical dependencies among attributes.

Methods

We provide a measure to quantify logical dependencies among attributes in tabular data. Utilizing this measure, we compare several state-of-the-art synthetic data generation algorithms and test their capability to preserve logical and functional dependencies on several publicly available datasets.

Results

We demonstrate that currently available synthetic tabular data generation algorithms do not fully preserve functional dependencies when they generate synthetic datasets. In addition, we also showed that inter-attribute logical dependencies can be preserved by some tabular synthetic data generation models.

Conclusions

Our review and comparison of the state-of-the-art reveal research needs and opportunities to develop task-specific synthetic tabular data generation models. Maintaining functional and logical dependencies in synthetic data drives medical progress and enhances clinical decision-making.

Uncertainty in clinical risk prediction: perspectives and approaches

Richard D. Riley[1,2], Gary S Collins[3], Kym Snell[1,2], Joie Ensor[1,2], Rebecca Whittle[1,2], Paula Dhiman[3], Lucinda Archer[1,2]

[1]Institute of Applied Health Research, College of Medical and Dental Sciences, University of Birmingham, Birmingham, United Kingdom

[2]National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre, United Kingdom

[3]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, United Kingdom

Background

Each year, thousands of prediction models are published in the medical literature aiming to inform diagnosis or prognosis in a target population. These models enable an individual's risk of a health-related outcome to be estimated, though most provide only a point-estimate of this risk and do not present any information on the corresponding uncertainty in their estimate. Before a prediction model can be considered for use in clinical practice, it should be critically appraised and rigorously evaluated. Unfortunately, many published models are not fit for purpose due to poor methodological standards leading to high levels of uncertainty in predictions. There is inconsistency, however, in the presentation of uncertainty around risk estimates, and much debate surrounding its usefulness. In general, uncertainty in predicted risk is ignored.

Methods

We outline perspectives on the presentation of uncertainty in risk estimates from a clinical prediction model and propose methods to quantify this uncertainty in practice.

Results

Deriving uncertainty intervals fully conditional on key patient attributes is the most appropriate option for individual-level uncertainty estimates, though is computationally complex. We demonstrate intervals derivation using the variance-covariance matrix of parameter estimates (following a frequentist approach) or by sampling from the posterior distribution for an individual's risk (when using a Bayesian framework). More generally, bootstrapping can be used to gain uncertainty estimates conditional on the individual's predicted risk, regardless of model development approach, though such estimates may not suitably account for patient-level characteristics.

Conclusions

Presenting uncertainty in predicted risk allows end-users and stakeholders to evaluate and critically appraise a prediction model, and can direct further research for model development and updating. Although point-estimates of risk are often sufficient for individual decision making, acknowledging uncertainty may enhance the patient-doctor consultation, though more research is also needed on how best to communicate uncertainty information to patients.

Uncertainty-based sequential sample size calculations for developing clinical prediction models using regression or machine learning methods

Amardeep Legha[1,2], Joie Ensor[1,2], Lucinda Archer[1,2], Rebecca Whittle[1,2], Kym I.E. Snell[1,2], Paula Dhiman[3], Ben Van Calster[4,5], Gary S. Collins[3], Richard D. Riley[1,2]

[1]Institute of Applied Health Research, College of Medical and Dental Sciences, University of Birmingham, Birmingham, United Kingdom

[2]National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre, United Kingdom

[3]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, United Kingdom

[4]Leuven Unit for Health Technology Assessment Research (LUHTAR), KU Leuven, Leuven, Belgium

[5]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands

Background

Clinical prediction models estimate an individual's risk of particular outcomes to inform decision-making. Small sample sizes may lead to unreliable predictions, but what constitutes 'small' is difficult to gauge in advance of analysis. To address this, here we propose adaptive sample size calculations to be used during data collection, to sequentially examine when additional data are (not) required to improve model robustness.

Methods

Our method extends previous work, to focus on the uncertainty of individual-level predictions and misclassification probabilities based on risk thresholds for decision making. Starting with small sample sizes, we derive and sequentially update prediction and classification instability plots as new participants are recruited. A stopping rule is based on the perceived value of additional information; crucially this is clinical context specific. Our approach is illustrated using real examples, including regression and machine learning approaches.

Results

Our findings show that what constitutes an adequate sample size is strongly dependent on the risk thresholds of interest for decision making, and the level of prediction and classification instability deemed acceptable by stakeholders. In particular, the sequential approach may identify that small sample sizes may still be acceptable if wide individual-level uncertainty intervals fall mainly in regions less relevant to clinical decision-making, and thus mis-classification probabilities are low. We also demonstrate the impact of stopping rules based on targeting suitable precision in groups of individuals defined by protected characteristics (e.g., ethnicity), to help improve model fairness.

Conclusions

An uncertainty-based sequential sample size approach allows users to dynamically monitor and identify when enough participants have been recruited to reliably develop their prediction model. This new approach for studies carrying out prospective data collection helps ensure more reliable models that optimise clinical decisions for individuals, and may help identify when smaller sample sizes are sufficient for decision making.

Speeding up the clinical studies with biomarker-based enrichment

Djuly Pierre Paul[1,2], Hong Sun[1], Irina Irincheeva[1]

[1]Bristol Myers Squibb, Boudry, Switzerland

[2]Nantes University, Nantes, France

Background

Identifying patients groups based on biomarkers is crucial in developing immunotherapies and targeted therapies in oncology. However, validating a biomarker as a stratification criterion in clinical trials can take several years. Choosing the threshold for continuous biomarkers is particularly challenging, often relying on limited number of values evaluated with simplistic statistical approaches. Early dichotomization ignores the actual distribution of values and the potentially informative *grey zone*.

Methods

In this work we adapt a biomarker enrichment design to identify the optimal threshold to determine patients who will benefits the most from the experimental treatment. We simulate Simon & Simon design for binomial endpoint and survival endpoint [1]. Various scenarios of chosen thresholds are studied through simulations inspired by existing studies. The impacts of measurement variability due to differences between laboratories on study power are investigated. ROC curve-based approach to determine the threshold, as well as a decision tree to improve the biomarker enrichment strategy are explored.

Results

Initial results suggest that this design can control the type I error but provides medium statistical power. Using maximum likelihood to determine the threshold in the initial design fits well in the data with binomial endpoint, but does not seem optimal for data with survival endpoint.

Conclusions

In conclusion, this study highlights the importance of a more nuanced approach in selecting biomarker threshold for oncology clinical trials. It's essential for more robust and adaptive methods that could accelerate the development of personalized therapies while optimizing the efficiency of clinical trials. These advanced methodology have the potential to significantly improve the precision of personalized medicine in oncology.

References

[1] Simon N, Simon R. Adaptive enrichment designs for clinical trials. Biostatistics. 2013 Sep;14(4):613-25. doi: 10.1093/biostatistics/kxt010. Epub 2013 Mar 21. PMID: 23525452; PMCID: PMC3769998.

Multimodal outcomes in N-of-1 trials: deep-learning based effect estimates in a small data study design

Juliana Schneider[1], Thomas Gärtner[1], Stefan Konigorski[1,2]

[1]Digital Health Center, Hasso Plattner Institute for Digital Engineering, University of Potsdam, Potsdam, Germany

[2]Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai, New York, USA

Background

N-of-1 trials are randomized multi-crossover trials in single participants with the purpose of investigating the possible effects of one or more treatments. These effects can be modeled individually or be aggregated to estimate population effects. Here, a participant alternates (randomly) between periods of treatment and non- or alternative treatment. This trial design is especially useful for rare diseases, chronic diseases and personalized analyses.

N-of-1 trials research has primarily focused on scalar outcomes. We propose to adapt this design to multimodal outcomes that later could easily be collected by trial participants on their personal mobile devices. In a first multimodal N-of-1 trial, Fu et al. [1] recently investigated the effect of creams on acne severity by tracking the outcome with images of the affected areas and applying supervised deep learning models in a series of 5 individual N-of-1 trials.

Methods

We present an approach for analyzing multimodal N-of-1 trials by combining an unsupervised deep learning model with statistical inference for omitting the need for outcome labels [2]. First, we trained an Autoencoder on the skin images to create lower dimensional embeddings. Afterwards, these were reduced to a single dimension by extracting the first principal component. Finally, statistical hypothesis tests were computed on an individual level to provide personalized estimates of effectiveness. However, the model training and dimensionality reduction leveraged the data of all participants and data augmentation to avoid overfitting.

Results

Our findings indicate that our unsupervised model approach captures relevant features for an effect estimate in the small data that N-of-1 trials provide. We show the suitability of deep learning approaches for N-of-1 trials.

Conclusions

In future work, we will experiment with incorporating transfer learning with pre-trained models and investigate more realistic simulations of N-of-1 trial image sequences in order to further mitigate the challenge of small sample sizes.

References

[1] Jingjing Fu, Shuheng Liu, Siqi Du, Siqiao Ruan, Xuliang Guo, Weiwei Pan, Abhishek Sharma, and Stefan Konigorski. Multimodal n-of-1 trials: A novel personalized healthcare design, 2023.

[2] Juliana Schneider, Thomas Gärtner, and Stefan Konigorski. Multimodal outcomes in n-of-1 trials: Combining unsupervised learning and statistical inference

Multidimensional investigation of response to treatment with inhaled corticosteroids in COPD patients: insights from the HISTORIC study

Eleni Papakonstantinou[1], Moritz Hess[2], Leticia Grize[1], Harald Binder[2], Daiana Stolz[1]

[1]Clinic of Respiratory Medicine, University of Freiburg, Germany; Faculty of Medicine, University of Freiburg, Freiburg, Germany

[2]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany

Background

Deep neural networks, specifically variational autoencoders (VAEs), can learn low-dimensional representations from multiple variables collected over time in clinical trials, increasing the power to detect effects without explicit variable selection. This study uses VAEs to assess the treatment-response to inhaled corticosteroids (ICS) on a dimension-reduced representation of the data from the HISTORIC study, a double-blind, randomized, placebo-controlled trial. The study tested the hypothesis that COPD patients with high area of airway smooth muscle cells (HASMC) in their endobronchial biopsies respond better to ICS than patients with low area (LASMC).

Methods

Longitudinal data from the HISTORIC study over 12 months, containing lung function and physical activity variables, was variance-filtered to 100 observed variables, with three removed manually. The remaining 97 variables were converted to z-scores. A VAE was trained on data from 177 patients with up to seven observations each, totaling 1073 observations, to learn a five-dimensional representation. We used a naive VAE approach and another with time structure enforced through polynomial regression. The five latent representations were analyzed for aggregated differences over time using Welch's t-test. To determine the clinical variable contributions to the latent variables, we dichotomized the latent variables at the median and used t-statistics.

Results

Of the five learned latent variables, one displayed lower values in the LASMC and HASMC ICS groups compared to placebo (LASMC: -0.38 vs. 0.160, p=0.086; HASMC: 0.032 vs. 0.638, p=0.112). This latent variable showed a strong negative association with Expiratory reserve volume and vital capacity, related variables, indicating higher values for these variables in the ICS-treated groups.

Conclusions

Using neural networks for dimension reduction of clinical trial data allows simultaneous assessment of treatment effects in longitudinal clinical trial data. This approach simplifies analysis by enabling simultaneous investigation of all potential outcomes. The implicit variable weighting simplifies model building and result interpretation.

AI & statistics in preclinical research and development

Tina Lang[1], Bernd-Wolfgang Igl[2], René Kubiak[3], Jörg Rahnenführer[4]

[1]Data Science & Artificial Intelligence, Bayer AG, Wuppertal, Germany

[2]Global Biostatistics & Data Sciences, Boehringer Ingelheim, Biberach, Germany

[3]Early Development & Non-Clinical Biostatistics Statistics, Sanofi, Frankfurt, Germany

[4]Department of Statistics, TU Dortmund University, Dortmund, Germany

Background

Pharmaceutical research and development in the early process phases, preceding Human Trials, involves critical stages such as target identification, lead discovery, lead optimization, and preclinical evaluation. Unlike the later Clinical Phases, these early-phase studies are usually small and exploratory. Most data will be contained within a company and are not publicly available. This setting brings special statistical challenges, particularly in the context of small sample sizes and limited prior knowledge.

Ideation

In this poster, we ideate on the potential integration of Artificial Intelligence (AI) into the early-phase processes. To this stage, we will focus on AI supported knowledge generation and only hint on possibilities for AI in data analysis. One possible role of AI is verifying exploratory results based on small data, single animal models, and specific strains by assessing publications. Additionally, the application of AI in data base screenings for target identification will be discussed. If AI-supported literature research on an innovative approach results into lists of potentially relevant papers, statistical background knowledge can help sort the publications into useful and unwanted and thus help generating new research hypothesis. If it was possible to get hold of enough small data sets for one common research question, the use of Machine Learning methods could be an option.

Conclusions

Overall, this poster aims to shed light on the unique statistical challenges and opportunities within early-phase pharmaceutical research and development, with a specific focus on the integration of traditional statistical methods and emerging AI technologies.

Two small-sample problems in optimal and exact inference

Marco Bonetti[1], Laura Bondi[2], Marcello Pagano[3], Pasquale Cirillo[4], Anton Ogay[5]

[1]Dondena Research Center, Bocconi University, Milan, Italy

[2]Health Data Science Centre, Human Technopole, Milan, Italy

[3]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA

[4]ZHAW School of Management and Law, Winterthur, Switzerland

[5]TU Delft, Delft, The Netherlands

Background

In the era of big data, we discuss two examples of optimal, exact inference from small samples.

Methods

The first example involves the optimal unbiased estimation of the sparsity index (the reciprocal of the intensity) when data are obtained from size-biased Poisson sampling. In the second example, we revive some algorithms - originally proposed in Rappeport (1968) - for the computation of the exact distribution of some functions of the multinomial counts under the hypothesis of equiprobability. The examination of, say, the largest observed count can represent an alternative approach to testing such null hypothesis to the standard chi-square test, or to the multiple testing of individual cell-specific probabilities.

Results

We propose two exact algorithms for the Poisson problem, which are computationally burdensome even for small sample sizes. As an alternative, we describe a third, approximate algorithm based on the inverse fast Fourier transform. An exact confidence interval based on the optimal estimator is also proposed. The performance of the estimation procedure is compared to classical maximum likelihood inference in terms of mean squared error and average coverage and width of the corresponding confidence intervals. For the multinomial problem, we revisit and expand on the original exact algorithms to compute the distribution of the maximum, the minimum, the range and the sum of the J largest order statistics of a multinomial random vector. We explore the improvement obtained when using the exact vs. the approximate distributions.

Conclusions

Exact inference is relevant for small sample problems. In our two examples, the improvement over large-sample methods is illustrated.

When only small data is available in livestock research

Anja Eggert[1], Anne-Marie Galow[1]
[1]Research Institute for Farm Animal Biology (FBN), Dummerstorf, Germany

Background
In today's "Big Data" era, Smart Farming is transforming agriculture. However, in livestock research, we often encounter challenges associated with "Small Data". Adhering to the 3Rs principle (Replace, Reduce, Refine) is essential, and optimized experimental design, including sample size calculation, is required to minimize the use of animals in scientific studies. Additionally, animal data are often noisy, incomplete, or expensive and labor-intensive to obtain.

Methods
This scarcity of data hinders the development of robust prediction models, particularly in the application of machine learning (ML). Training ML models on small datasets can lead to overfitting, where models fail to generalize patterns and lack statistical power. Unsupervised models may struggle to identify patterns, while supervised models may exhibit low prediction accuracy due to insufficient data.

Results
Despite these challenges, innovative strategies such as data augmentation and oversampling have been developed to mitigate the effects of small data. From the ML perspective, algorithms like support vector machine or random forest and strategies including active learning and transfer learning have shown to be effective methods of handling small datasets.

Conclusions
Attending this symposium, we aim to connect with peers in the domain of machine learning methods for small data. We will present a use case focused on developing a biomarker as an indicator for stress resilience in pigs. Our objective is to gain feedback on this ongoing project and discuss the application of innovative machine learning methods to provide a clear evaluation of the new biomarker system. Engaging with experts at the symposium will help us refine our approach and explore collaborative opportunities to enhance the project's impact.

Impact of different longitudinal data representations on transformer performance in small data applications

Kiana Farhadyar[1,2], Lukas Königs[1], Harald Binder[1,2]

[1]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Centre, University of Freiburg, Freiburg, Germany

[2]Freiburg Center for Data Analysis and Modeling, University of Freiburg, Freiburg, Germany

Background

In biomedical research, longitudinal cohort data are crucial for understanding the developmental paths of individuals. However, the analysis of these data can be complicated by complex temporal patterns and often limited sample sizes, requiring special care for using advanced methodologies. This study, in particular, aims to explore the use of attention mechanisms and different representations of longitudinal data within and outside the transformer architecture.

Methods

To address the challenge of analyzing cohort data in a small data application, we propose converting longitudinal datasets into a discrete event format that is more in line with the application transformers have originally been developed for, i.e., natural language processing. We accordingly use different representations of data, inside and outside of the transformer architecture, e.g., regression, to evaluate the capabilities of each in prediction tasks. As the first method, we use a pragmatic one-hot-encoding approach. Then, we investigate the potential of large language models (LLMs), e.g., pre-trained embedding models, to extract semantic representations of events. For this, we use different approaches, e.g., averaging or concatenating single events, to aggregate sequence events that happen at one time point. In the end, we compare different embedding approaches using visualization techniques and the prediction performance of such approaches.

Results

We illustrate the advantages of such event codings and embeddings with data that comprises self-reported stressors collected in a longitudinal resilience assessment study. We show which word embeddings at the event level, i.e., reported stressors, produce more meaningful attention scores that align with expert knowledge on stressors. Additionally, our work demonstrates the impact of semantic representation combined with attention mechanisms in simple statistical methods, e.g., regression, and more complex models, e.g., transformers.

Conclusions

This provides a promising foundation for maximizing the potential of various approaches when analyzing longitudinal cohort data.

Allocation bias in group sequential designs

Daniel Bodden[1], Nicole Heussen[1,2], Ralf Dieter Hilgers[1]

[1]Department of Medical Statistics, RWTH Aachen University, Aachen, Germany

[2]Center for Biostatistics and Epidemiology, Medical School, Sigmund Freud Private University, Vienna, Austria

Background

Randomized controlled clinical trials (RCTs) are the gold standard in clinical research, but they are not immune to biases. Both the FDA and EMA stress the need to address these biases, even in RCTs. While allocation bias effects have been studied in various trial designs, its impact on group sequential designs remains underexplored. This is especially relevant for rare disease trials, where group sequential designs can reduce expected sample size. However, allocation bias poses a significant concern in this context, as researchers may find it easier to track and predict future allocations, potentially leading to allocation bias.

Methods

This study investigates the influence of allocation bias on the testing decisions in group sequential clinical trials across different randomization procedures. We introduce a statistical model based on a biasing policy to calculate the Type I Error probability for different randomization sequences. Monte Carlo simulations are employed to assess Type I Error probabilities under various randomization procedures in the presence of allocation bias.

Results

For Lan & DeMets alpha spending functions, our findings indicate that among the randomization procedures evaluated, permuted block randomization with a block length of 4 resulted in the highest Type I Error probability in the presence of allocation bias. This was followed by Chen's design with p=0.67 and an imbalance factor of 3, as well as Efron's biased coin design with p=0.67. Conversely, complete randomization and the random allocation rule led to only a slight increase in Type I Error probability.

Conclusions

The choice of the randomization procedure in group sequential trials is crucial to mitigate biases. It is recommended to use complete randomization or random allocation rule, when alllocation bias is a concern. When additional biases are expected, evaluating the pros and cons of each randomization procedure becomes essential.

Quantifying the impact of allocation bias in randomised clinical trials with multi-component endpoints

Stefanie Schoenen[1], Nicole Heussen[2], Ralf-Dieter Hilgers[1]

[1]Institute of Medical Statistics, RWTH Aachen University, Aachen, Germany

[2]Center for Biostatistics and Epidemiology, Medical School, Sigmund Freud Private University, Vienna, Austria

Background

Randomized clinical trials (RCTs) are the gold standard for reducing bias. However, in rare diseases, RCTs are often unblinded, which can result in allocation bias. This bias arises when researchers, aware of or predicting the next patient's allocation, influence the allocation process by selecting a patient who best fits the anticipated group assignment. The EMA recommends considering the potential impact of bias to ensure the validity of trial results. While allocation bias has been studied in various trial designs, its impact on trials with multi-component endpoints—beneficial in rare diseases because their ability to reduce sample size—has not been analyzed.

Methods

To assess allocation bias in multi-component endpoint trials, we developed a biasing policy based on the assumptions that researchers know prior allocations and that well-responding patients score higher on endpoint components. Following Blackwell and Hodges' convergence strategy, the next patient is more likely to get assigned to the less frequent group. If assigned to the experimental group, a good responder is chosen; otherwise, a worse responder. A neutral responder is allocated only when group sizes are balanced.

Results

Simulations show that allocation bias inflates the type I error rates when using the Wei-Lachin test. Even small bias effects cause exceeding the 5% significance level. The amount of inflation depends on the chosen randomisation procedure and the number of components combined in the multi-component endpoint. More components lead to increased error inflation.

Conclusions

Analysing bias effects during the planning phase of a clinical trial and choosing a bias-mitigating randomisation procedure enhances the validity of the trial. Thus, the developed methodology can be applied to base the selection of a randomisation procedure on scientific arguments, to facilitate the design of more robust and valid rare disease clinical trials, and to perform bias corrected statistical tests in the trial's analysis phase.

Multivariate functional linear discriminant analysis of partially-observed time series

Rahul Bordoloi[1], **Clémence Réda**[1], **Orell Trautmann**[1], **Saptarshi Bej**[2], **Olaf Wolkenhauer**[1,3,4]

[1]Institute of Computer Science, University of Rostock, Germany

[2]School of Data Science, IISER Thiruvananthapuram, India

[3]Leibniz-Institute of Food Systems Biology, Technical University of Munich, Germany

[4]Stellenbosch Institute of Advanced Studies (STIAS), South Africa

Background

Functional Linear Discriminant Analysis (FLDA), proposed by James and Hastie in 2001, extends multi-class classification and dimensionality reduction with Linear Discriminant Analysis to fragmented observations of an ultrashort univariate time series with missing data. However, for ultrashort multivariate time series, a remaining major challenge is disentangling statistical dependencies between the different features of the multivariate function. A further challenge is provided by heterogeneous sampling times and gaps with missing feature values. Such cases can, for example, arise in clinical and psychological studies, where patients are surveyed for several variables (features) sparsely over weeks or months.

Methods

We developed Multivariate fUnctional linear DiscRiminant Analysis (MUDRA) as an extension of FLDA to the ultrashort multivariate time series setting. MUDRA leverages novel tensor decomposition and matrix equation-solving approaches in an efficient expectation-maximization algorithm, which already improve the computational cost in the uni-variate case compared to FLDA. The structure embedded in the tensor decomposition elegantly incorporates the statistical dependencies across different features of the multivariate function. We also theoretically show that MUDRA converges to the maximum likelihood estimate.

Results

Our experimental study applied to the “Articulary Words” data set shows that MUDRA outperforms state-of-the-art approaches such as the RandOm Convolutional Kernel Transform (ROCKET), especially when a significant portion of data is missing or when sparse representations (less than d=40 dimensions) are required. For instance, for d=40, a ridge regression-based classifier trained on MUDRA (respectively, ROCKET)-reduced feature vectors achieve an F1-score of 0.88 (resp., 0.83). Furthermore, even in scenarios where 55% of feature values are missing, MUDRA achieves significantly better F1-scores than ROCKET, whatever the value of d. Moreover, unlike ROCKET, MUDRA does not require any additional mechanism to deal with missing values.

Conclusions

Our approach allows us to infer sparse, meaningful embeddings from partially observed ultrashort multivariate time series data, significantly enhancing classification accuracy in challenging observational data. Our preprint is available at https://arxiv.org/abs/2402.13103 and the open-source Python package MUDRA at https://github.com/rbordoloi/MUDRA.