ulm university universität ulm

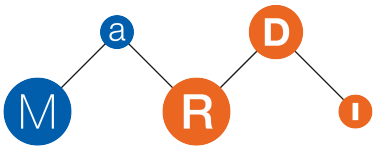
Statistical Computing 2023

Abstracts der 53. Arbeitstagung

HA Kestler, JM Kraus (eds)

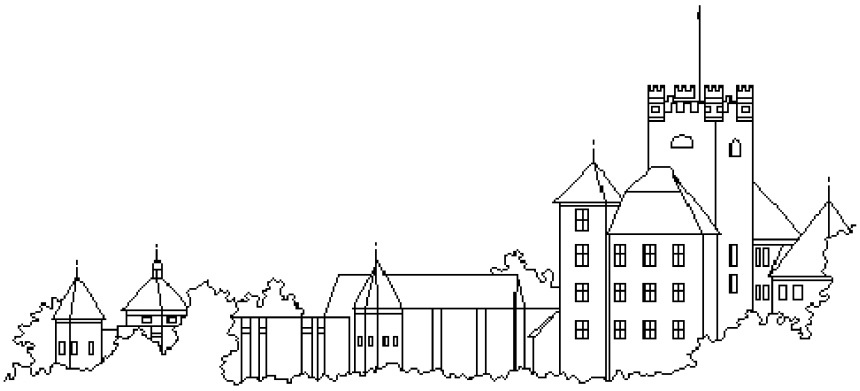
Berichte des Instituts für Medizinische Systembiologie

Nr. 2023-01
July 2023



Universität Ulm

Statistical Computing 2023



53. Arbeitstagung

der Arbeitsgruppen Statistical Computing (GMDS/IBS-DR), Klassifikation und Datenanalyse in den Biowissenschaften (GfKl).

30.07. - 02.08.2023, Schloss Reisensburg (Günzburg)

Workshop Program

Sunday, July 30, 2023

18:00 – 20:00		**Dinner**
		Introduction: H. A. Kestler
20:00 – 21:00	Anne-Laure Boulesteix (München)	Towards reliable empirical evidence in methodological computational research: Recent developments and remaining challenges

Monday, July 31, 2023

08:50		**Opening of the workshop: H. A. Kestler**
09:00 – 10:30		**Chair: J. Kraus**
09:00 – 09:30	Elham Shamsara (Tübingen)	Assessing the impact of vaccination and predicting the emergence of the Omicron-Variant: Modeling the dynamics of COVID-19 in the Omicron wave
09:30 – 10:00	Marcus Vollmer (Greifswald)	Paradoxes in sample size determination when using Fisher's exact test and accounting for drop out interference
10:00 – 10:30	Thomas Welchowski (Bonn)	Interaction difference hypothesis test for prediction models
10:30 – 11:00		**Coffee Break**
		Introduction: H. A. Kestler
11:00 – 12:00	Nico Pfeifer (Tübingen)	Efficient privacy-preserving machine learning for precision medicine
12:00 – 14:00		**Lunch**
14:00 – 16:00		**Chair: M. Drton / S. Haug**
14:00 – 14:30	Oleksandr Zadorozhnyi (München)	Statistics and ML within MaRDI project
14:30 – 15:00	Sebastian Fischer (München)	mlr3torch – integrating torch and mlr3
		Introduction: S. Haug
15:00 – 16:00	Sebastian Meyer (Erlangen)	Changes in R
16:00 – 16:30		**Coffee Break**
16:30 – 18:00		**Chair: A. Mayr**
16:30 – 17:00	Maren Hackenberg (Freiburg)	Combining componentwise boosting with neural networks for structuring latent representations of singlecell RNA-sequencing data
17:00 – 17:30	Marietta Hamberger (Ulm)	Accelerating biological network reconstruction with machine learning and natural language processing
17:30 – 18:00	Carlo Maj (Marburg)	A phenome-wide boosting analysis for polygenic prediction models in UK Biobank
18:00 – 20:00		**Dinner**

Tuesday, August 01, 2023

09:00 – 10:30		**Chair: T. Welchowski**
09:00 – 09:30	Colin Griesbach (Erlangen)	Confidence intervals for finite mixture regression based on re-sampling techniques
09:30 – 10:00	Nikolai Spuck (Bonn)	A tool to detect nonlinearity and interactions in generalized regression modeling
10:00 – 10:30	Ludger Sandig (Dortmund)	A Julia package for Bayesian optimal design of experiments
10:30 – 11:00		**Coffee Break**
		Introduction: M. Schmid
11:00 – 12:00	Tim Friede (Göttingen)	From trial simulation to in-silico trials
12:00 – 14:00		**Lunch**
14:00 – 16:00		**Chair: M. Schmid**
14:00 – 14:30	Sebastian Krey (Göttingen)	Using model based optimization for exploring the performance profile of large scale storage systems
14:30 – 15:00	Matthias Medl (Wien)	Application on CNN ensembles to model Fourier transform infrared spectra
15:00 – 15:30	Andreas Mayr (Bonn)	Using the R^2 measure on test data? The evaluation of polygenic prediction models across populations
15:30 – 16:00	Hannah Klinkhammer (Bonn)	Advanced statistical modelling of polygenic risk scores via boosting targeted objective functions
16:00 – 16:30		**Coffee Break**
16:30 – 18:00		Working group meeting – **Statistical Computing 2024**
18:00 – 20:00		**Dinner**
20:00 – 21:00		Poster session

Wednesday, August 02, 2023

09:00 – 10:30		**Chair: M. Vollmer**
09:00 – 09:30	Jörn Lötsch (Frankfurt)	A clinical sensory test recognized as random walk
09:30 – 10:00	David Köhler (Bonn)	Reduction by spatial components analysis improves pattern detection in multivariate spatial data
10:00 – 10:30	Felix Weidner (Ulm)	Constraint-based attractor search in quantum Boolean networks
10:30 – 11:00		**Coffee Break**
11:00 – 12:00		**Chair: S. Krey**
11:00 – 11:30	Ana Stolnicu (Ulm)	Establishing a trustworthy signalling entropy calculation for biological processes analysis
11:30 – 12:00	Alina Schenk (Bonn)	A random forest pseudo-value approach for modeling restricted mean survival times
12:00 – 14:00		**Lunch**
14:00		**Departure**

Contents

Towards reliable empirical evidence in methodological computational research: Recent developments and remaining challenges

Anne-Laure Boulesteix[1]

[1] Ludwig-Maximilians-Universität München, Institut für medizinische Informationsverarbeitung, Biometrie und Epidemiologie

boulesteix@ibe.med.uni-muenchen.de

Assessing the impact of vaccination and predicting the emergence of the Omicron variant: Modeling the dynamics of COVID-19 in the Omicron wave

Elham Shamsara[1], Florian König[1], Nico Pfeifer[1]

The COVID-19 pandemic has brought unprecedented challenges to healthcare systems globally. This highly contagious virus spreads through respiratory droplets produced when an infected person coughs, sneezes, or talks, leading to a rapid spread. In November 2021, a novel variant of SARS-CoV-2, Omicron (B.1.1.529), was first identified in Botswana and South Africa, which has now spread to several countries. The World Health Organization declared it a variant of concern (VOC) on November 26, 2021. Omicron has distinct characteristics and did not emerge from earlier known variants, evolving into five lineages: BA.1, BA.2, BA.3, BA.4, and BA.5 [1]. In this study, we employed a compartmental model to investigate the impact of the Omicron variant and its lineages BA.1, BA.2, and BA.5 on populations. The model consists of several compartments, including Susceptible (S), Infected (I), Hospitalized (H), ICU, Recovered (R), and Death (D). We incorporated a rate of vaccination efficacy, which removes individuals from the S compartment, with the remaining fraction becoming recovered [6]. However, upon examining the model, we found that considering recovered or vaccinated groups with a rate becoming susceptible improved the model's accuracy. Our model aimed to address principal concerns surrounding the Omicron variant, including its transmissibility and severity compared to other VOCs and its ability to circumvent vaccine protection. The study investigates the effect of vaccination and immunity rates concerning the Omicron variant across various countries [3]. We utilized the Disease Informed Neural Network (DINN) [4] to estimate the rate of the SIHICURD model in this study. To capture the time-dependent behavior of the parameters, we incorporated sliding windows of size 3 months, which we determined after testing various window sizes and interpolation techniques. By considering this interval, the neural network can learn that the solutions are not solely increasing or decreasing, which is essential for modeling the virus's different waves and fluctuations. This approach enables us to develop a more accurate and comprehensive understanding of Omicron's dynamics and how it evolved over time. We implemented our model to simulate the compartments for France, Italy, and Germany. Our results demonstrated that the model was capable of accurately simulating the dynamics of the different compartments in each country. Our study highlights the importance of utilizing mathematical models in comprehending and forecasting the dynamics of pandemics, including the emergence, and spread of new variants such as Omicron and its sub-lineages in the ongoing COVID-19 pandemic.

[1] University of Tübingen, Methods in Medical Informatics, Department of Computer Science

elham.shamsara@uni-tuebingen.de, florian.koenig@uni-tuebingen.de,
nico.pfeifer@uni-tuebingen.de

References

1 Tegally, Houriiyah, et al. "Emergence of SARS-CoV-2 omicron lineages BA. 4 and BA. 5 in South Africa." Nature medicine28.9 (2022): 1785-1790.

2 Dashtbali, Mohammadali, and Mehdi Mirzaie. "A compartmental model that predicts the effect of social distancing and vaccination on controlling COVID-19." Scientific Reports 11.1 (2021): 8191.

3 Karim, Salim S. Abdool, and Quarraisha Abdool Karim. "Omicron SARS-CoV-2 variant: a new chapter in the COVID-19 pandemic." The lancet 398.10317 (2021): 2126-2128.

4 Shaier, Sagi, Maziar Raissi, and Padmanabhan Seshaiyer. "Data-driven approaches for predicting spread of infectious diseases through DINNs: Disease Informed Neural Networks." arXiv preprint arXiv:2110.05445 (2021).

Paradoxes in sample size determination when using Fisher's exact test and accounting for drop out interference

Marcus Vollmer[1]

Objective: Sample size determination is crucial while planning clinical trials. The basis for the determination is the statistical method to be used to test the primary hypothesis. Typically, a statistical power of 80% holding an alpha level of 5% is used for the determination, and the sampling numbers were usually corrected for drop outs by dividing the expected drop out rate. The paradoxon of weaker statistical power despite increasing sample size that can occur when using Fisher's exact test is examined.

Methods: The exact statistical power is calculated for pairwise combinations in an example study when the number of events (e.g., development of cancer) is compared between two experimental groups to the control group. Strong effect sizes we set as assumptions (reduction from 20% to 1%) as both experimental groups should result in a significantly lower event rate. Then, drop out rates were included in the calculations to examine the impact on filtering out paradoxical effects.

Results: The statistical power in the example shown in Figure 1 surprisingly leads to a loss of power up to 8% when the sample size in the control group is increased by 1. The paradox is also evident in the optimal combination of sample sizes that results in a total sample size of 123 (n_C=53, n_E=35). Increasing either n_C or n_E would result in a power below 80%. Simulating the realized drop out acts as a filter on the determined power and lowers the risk of a steep paradoxical power decline.

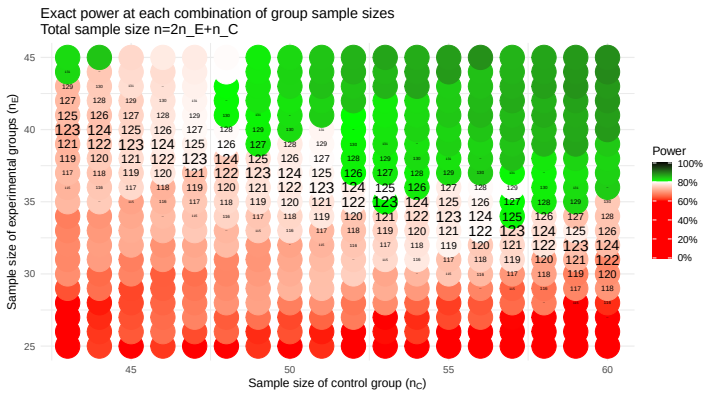


Figure 1: Statistical power obtained from Fisher's exact tests to illustrate the paradox. Some combinations with larger sample sizes result in reduced statistical power.

[1] University Medicine Greifswald, Institute of Bioinformatics

marcus.vollmer@uni-greifswald.de

Interaction difference hypothesis test for prediction models

Thomas Welchowski[1], *Matthias Schmid*[1]

Machine learning research focuses on the improvement of prediction performance. During the past decade, major advances were made in the field of deep learning with imaging, audio or video data and ensemble models like bagging or boosting with matrix type data. These so called black-box models flexibly adapt to the given data and involve fewer assumptions about the data generating process than standard methods like linear regression or single decision trees. However, due to their increased complexity, black-box models are more difficult to interpret. To address this issue, techniques for interpretable machine learning have been developed; yet there is still a lack of methods to reliably identify interaction effects between predictors under uncertainty. In this work we present a model-agnostic asymptotic hypothesis test for the identification of interaction effects in black-box machine learning models. The null hypothesis assumes that a given set of covariates does not contribute to interaction effects in the prediction model. The test statistic is based on the difference of variances of partial dependence functions with respect to the original black-box predictions and the (more restrictive) predictions under the null hypothesis. The proposed hypothesis test can be applied to any black-box prediction model, and the null hypothesis of the test can be flexibly specified/modified according to the research question of interest. Furthermore, the test is computationally fast to apply as the null distribution does not require resampling and/or re-fitting black-box prediction models.

[1] University of Bonn, Department of Medical Biometrics, Informatics and Epidemiology

welchow@imbie.uni-bonn.de, matthias.c.schmid@uni-bonn.de

Efficient privacy-preserving machine learning for precision medicine

Nico Pfeifer[1]

[1] University of Tübingen, Methods in Medical Informatics, Department of Computer Science

nico.pfeifer@uni-tuebingen.de

Statistics and ML within MaRDI project

Zadorozhnyi Oleksandr [1]

Machine learning and statistical methods have become essential tools for processing and analyzing large and complex datasets and are becoming part of day to day life. We at MaRDI (Mathematical Research Data Infrastructure)[1], a project which is funded by the German Research Foundation that aims to establish a sustainable, open, and decentralized infrastructure for research data in the field of mathematics, aim to tackle the challenges connected to the task of managing research data in statistics and machine learning. The goal of our task area is to develop and implement machine learning and statistical methods to improve the analysis and interpretation of mathematical data and make them FAIR via various platforms and by creating versatile toolboxes. In my talk I give an overview about consortia MaRDI, problems with which we are dealing and various toolboxes to solve them. I illustrate my talk presenting the resources[2] we use to achieve findability and accessibility of the research data. The work of task area "Statistics and Machine Learning" will help researchers to gain a deeper understanding of mathematical research data processes and to create infrastructure for typical ML tasks. By applying machine learning and statistical methods to research data our project within MaRDI aims to promote the discovery of new knowledge and to support evidence-based decision-making in the field of statistics.

References

1 https://www.mardi4nfdi.de/about/mission

2 https://zenodo.org/communities/mardigmci/

[1] TUM School of Computation Information and Technology

oleksandr.zadorozhnyi@tum.de

mlr3torch - Integrating torch and mlr3

Martin Binder[1], Bernd Bischl[1], Lukas Burk [1,2], Sebastian Fischer[1], Florian Pfisterer [1]

The mlr3 framework is a collection of R packages providing a unified interface to machine learning in the R language. With the release of torch in 2020, R has gained a native deep learning framework, enriching its capabilities in the domain of neural network training. The R package mlr3torch seamlessly integrates torch and mlr3, simplifies the training process, and offers a language to represent neural networks using the graph language defined in mlr3pipelines. In this presentation, we will showcase the capabilities of mlr3torch and demonstrate its benefits for researchers.

[1] Ludwig-Maximilians-Universität München
[2] Leibniz-Institut für Präventionsforschung und Epidemiologie - BIPS

mb706@gmail.com, bernd_bischl@gmx.net, lukas.burk@gmail.com, sebf.fischer@gmail.de, pfistererf@googlemail.com

Changes in R

Sebastian Meyer[1]

I will give an overview of recent changes in base R and of my experiences in testing and contributing such changes.

[1] Friedrich-Alexander-Universität Erlangen-Nürnberg, Institute of Medical Informatics, Biometry, and Epidemiology

seb.meyer@fau.de

Combining componentwise boosting with neural networks for structuring latent representations of single-cell RNA-sequencing data

Maren Hackenberg[1], *Niklas Brunn*[1], *Harald Binder*[1]

Dimension reduction is an important step in any exploratory analysis workflow of single-cell RNA-sequencing (scRNA-seq) data for identifying underlying patterns of cellular heterogeneity. To facilitate biological insight, specific properties of the corresponding low-dimensional representation are desirable. For example, it should be disentangled, i.e., different dimensions correspond to distinct underlying factors of variation, and interpretable, i.e., latent patterns can be attributed to a small set of characteristic variables such as genes driving a cellular differentiation process. For structuring the representation of scRNA-seq data accordingly, we propose to combine variable selection by componentwise likelihood-based boosting with neural networks for dimension reduction. Specifically, we implicitly regularize an autoencoder by componentwise boosting when minimizing the reconstruction loss. Thus, a small number of explanatory features is identified for each dimension of the latent representation, ensuring interpretability. As targets for the boosting approach, we use constrained negative gradients of the current reconstruction loss w.r.t. the individual latent dimensions, where the constraint can be tailored to different criteria which correspond to specific structural properties of the latent representation. Specifically, a constraint ensures that, for a given dimension, only features are selected that are complementary to the information already encoded in the other dimensions, thus arriving at a disentangled representation. Differentiable programming allows for jointly optimizing both components by differentiating through the boosting step, such that the latent space is adapted to and structured by the variable selection component. In an application on scRNA-seq data from cortical neurons of mice and a corresponding simulation design, we show that our approach captures distinct cell types in different latent dimensions while simultaneously identifying a set of characterizing genes for each dimension. We further illustrate how our approach can be extended to capture temporal development patterns in time-series scRNA-seq data by linking latent dimensions that represent the same developmental pattern across time points, thus capturing gene expression dynamics in cellular differentiation programs.

[1] University of Freiburg, Institute of Medical Biometry and Statistics
[2] University of Freiburg, Freiburg Center for Data Analysis and Modelling

maren.hackenberg@uniklinik-freiburg.de, niklas.brunn@uniklinik-freiburg.de, harald.binder@uniklinik-freiburg.de

Accelerating biological network reconstruction with machine learning and natural language processing

Marietta Hamberger[1], *Hans A. Kestler*[1]

The exploration of intricate gene regulatory networks in Systems Biology is often hindered by the time-consuming and biased manual reconstruction of molecular interactions, which is traditionally undertaken by experienced modelers. To overcome these challenges, our objective is to develop a streamlined interface that enables efficient and extensive investigations in this field. The interface aims to support the modelers at multiple stages, spanning from gene detection to regulatory network reconstruction.

In this study, we sought to assess the performance of our automated procedure by conducting a comparative analysis with a manual reconstruction study focused on understanding the crosstalk between IGF and Wnt signaling in aging satellite cells [1]. To achieve this, we extracted and processed textual content from the articles cited by the study, resulting in a data set of approximately 31,000 sentences. Leveraging BioBERT, a pre-trained language model specialized in biomedical text mining, we improved its classification capabilities by fine-tuning on diverse data sets, including the GENIA corpus [2, 3]. This optimization allowed us to identify various key biomedical entities within the text, such as genes, diseases, cell lines, and species.
Additionally, we devised a trigger word system to identify phrases associated with gene regulation and implemented a rule-based relation extraction model as an exploratory strategy. Notably, our model successfully detected the majority of the gene regulatory network which had been manually curated by human experts.

This research showcases the feasibility of using automated methods to extract gene regulatory statements from scientific literature, highlighting the potential for harnessing machine learning and natural language processing to assist human modelers and improve the overall efficiency of the reconstruction process.

References

1 Siegle, L., Schwab, J. D., Kühlwein, S. D., Lausser, L., Tümpel, S., Pfister, A. S., Kühl, M., and Kestler, H. A. (2018). A Boolean network of the crosstalk between IGF and Wnt signaling in aging satellite cells. PLoS One, 13(3), e0195126.

2 Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234-1240.

3 Kim, J. D., Ohta, T., Tateisi, Y., & Tsujii, J. I. (2003). GENIA corpus - a semantically annotated corpus for bio-textmining. Bioinformatics, 19(suppl_1), i180-i182.

[1] Ulm University, Insitute of Medical Systems Biology

marietta.hamberger@uni-ulm.de, hans.kestler@uni-ulm.de

A phenome-wide boosting analysis for polygenic prediction models in UK Biobank

Carlo Maj[1,2], Hannah Klinkhammer[2,3], Christian Staerk[3], Peter Krawit[2], Andreas Mayr[3]

The availability of large-scale biobank data enables the fitting of polygenic models directly on individual genotype data. The derived prediction models by incorporating the conditional joint effect across variants can lead to substantially improved predictions with respect to models based on univariate associations from genome-wide association studies (GWAS) [1]. In our work we used a statistical boosting algorithm (snpboost) we recently developed to derive sparse (i.e., based on relatively few variants) polygenic risk score (PRS) models from high-dimensionality genotype data for a phenome-wide analysis in UK Biobank [2]. We applied snpboost as well as other PRS methods on a wide range of qualitative and quantitative phenotypes from the UK Biobank with different heritabilities and polygenicities. Using snpboost we were able for several traits to derive sparse PRS models achieving better performances with respect to summary statistics-based approaches. Our findings suggest that the genetic architecture of the underlying traits plays a major role in model prediction performance with respect to model sparsity. In particular, both single-nucleotide-polymorphism-heritability (SNP-h^2) and polygenicity jointly contribute to the heterogeneity of prediction performances across different methods and considering different model sparsities. Our study demonstrated that, given the current dimensionality of large-scale biobank data, PRS models for quantitative and highly prevalent phenotypes can be derived with methods based on individual-level genotype data which can yield substantially improved prediction performances with respect to GWAS based approaches. Our analysis additionally shows that polygenicity (i.e., number of independently associated loci) and SNP-h^2 are also strongly related to the relationship between model sparsities and model predictions.

References

1 Maj, Carlo, Christian Staerk, Oleg Borisov, Hannah Klinkhammer, Ming Wai Yeung, Peter Krawitz, and Andreas Mayr. "Statistical learning for sparser fine-mapped polygenic models: The prediction of LDL-cholesterol." Genetic epidemiology 46, no. 8 (2022): 589-603.

2 Klinkhammer Hannah, Christian Staerk, Carlo Maj, Peter Michael Krawitz, and Andreas Mayr. "A statistical boosting framework for polygenic risk scores based on large-scale genotype data." Frontiers in Genetics 13 (2023): 1076440.

[1] University of Marburg, Center for Human Genetics
[2] University Hospital Bonn, Institute for Genomic Statistics and Bioinformatics
[3] University Hospital Bonn, Institute for Medical Biometry, Informatics

carlo.maj@uni-marburg.de, klinkhammer@imbie.uni-bonn.de, staerk@imbie.uni-bonn.de, pkrawitz@uni-bonn.de, amayr@uni-bonn.de

Confidence intervals for finite mixture regression based on resampling techniques

Colin Griesbach[1] *and Tobias Hepp*[1,2]

Mixture Regression Models [4] are widely used to quantify associations between outcomes and various covariates in scenarios with unobserved heterogeneity. However, meaningful uncertainty estimates are not immediately available as regular statistical inference neglects any variance regarding class assignments yielding biased results [3]. This issue has been addressed for ordinary mixture models by employing resampling techniques like various bootstrapping routines or the jackknife [1, 5] and in the case of mixture regression models, [2] already used bootstrapping to detect identifiability issues of fitted mixture regression models.
In this work, we propose a resampling approach for uncertainty estimates of regression parameters in finite mixture regression models. The method applies empirical bootstrapping and in addition uses a matching mechanism based on correlations of posterior class probabilities to aggregate estimates across all bootstrapping iterations and prevent label switching. Simulations and real world applications reveal that applying the proposed resampling approach results in slightly wider confidence intervals which are now capable of holding the type-I error threshold.

References

1 Basford, K., Greenway, D., McLachlan G. and Peel, D. (1997). Standard Errors of Fitted Component Means of Normal Mixtures. *Computational Statistics*, **12(1)**, 1 – 17.

2 Grün, B. and Leisch, F. (2004). Bootstrapping Finite Mixture Models. In: *COMPSTAT 2004 – Proceedings in Computational Statistics*. Heidelberg: Physica Verlag, 1115 – 1122.

3 Grün, B. and Leisch, F. (2008). FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters. *Journal of Statistical Software*, **28(4)**, 1 – 35.

4 McLachlan, G. and Peel, D. (2000). *Finite Mixture Models.* New York: John Wiley and Sons Inc.

5 O'Hagan, A., Murphy, T., Scrucca, L. and Gormley, I. (2019). Investigation of parameter uncertainty in clustering using a Gaussian mixture model via jackknife, bootstrap and weighted likelihood bootstrap. *Computational Statistics*, **34**, 1779 – 1813.

[1] Georg-August-Universität Göttingen
[2] Friedrich-Alexander-Universität Erlangen-Nürnberg

colin.griesbach@uni-goettingen.de, tbs.hepp@fau.de

A tool to detect nonlinearity and interactions in generalized regression modeling

Nikolai Spuck[1], Matthias Schmid[1], Moritz Berger[1]

Generalized linear models (GLMs) are a popular tool for regression analysis. They are based on the assumption that the relationship between the modeled outcome of interest and the covariates is linear. In addition, it is frequently assumed that the effect of a covariate is independent of the values of other covariates, neglecting possible interactions. These assumptions, however, may be too restrictive in many applications and lead to biased effect estimates. There are numerous alternative approaches for modeling continuous covariates like categorization, polynomial regression, generalized additive models (GAMs) and tree-based methods. However, while the application of variable selection methods in regression analysis has become increasingly common, methods that provide guidance regarding the choice of suitable functional forms for continuous covariates are still lacking. To address this issue, we propose an algorithm that examines various modeling alternatives and is able to detect nonlinearity and interactions between covariates if they are present. The algorithm utilizes tree-based splits which makes the resulting effects easily interpretable. More specifically, it indicates whether (i) linear effects are sufficient (indicating the use of a simple GLM), (ii) varying linear effects should be included in the model formula, (iii) one or several covariates exhibit non-linear effects (calling for the use of a GAM), or (iv) interaction effects occur in the data (hinting that the use of a tree-based method may be beneficial). We illustrated the algorithm by an application to data from patients who suffered from chronic kidney disease. The performance of the algorithm was assessed based on detection rates in a simulation study. Results of the simulation study indicate that the algorithm is able to proficiently detect nonlinearity and identify the correct functional form for a continuous covariate in settings with medium to high sample sizes and moderate noise. Some specific interactions structures were less likely to be identified correctly.

[1] University of Bonn, Institute of Medical Biometry, Informatics and Epidemiology

spuck@imbie.uni-bonn.de, matthias.c.schmid@uni-bonn.de, moritz.berger@imbie.uni-bonn.de

A Julia package for Bayesian optimal design of experiments

Ludger Sandig[1]

Suppose a toxicologist wants to study the effects of a drug. But which initial dose(s) should they administer, and at which point(s) in time should they take measurements? This question can be framed mathematically as a problem of optimal experimental design: selecting values for the covariates (design points) and corresponding samples sizes (weights) such that the measurements will be as informative as possible about the unknown parameters of an underlying nonlinear regression model. In a Bayesian context this means maximizing the expected difference in Shannon information between prior and posterior distribution.

Existing free and open-source software packages for finding optimal designs are typically implemented in R or MATLAB. Unfortunately, they often include only a highly domain-specific collection of models and design criteria. When user-defined functions can be supplied at all, they run much slower than internal ones, or fiddly interfacing with external C/C++ code is necessary. Moreover, the package code is often not well modularised, making it hard to comprehend and difficult to extend with new functionality. For these reasons it is not straightforward to adapt existing software for complex new experimental setups.

In this talk, we present a Julia package that tries to address these issues. Julia is a scientific high-level dynamic programming language with a performance comparable to statically compiled C. The optimization problems of experimental design can be elegantly implemented using Julia's user-extendable type system in conjunction with multiple dispatch. We demonstrate our package's flexibility on a range of examples from toxicology and pharmacometrics.

[1] Technische Universität Dortmund, Fakultät Statistik

sandig@statistik.tu-dortmund.de

From trial simulation to in-silico trials

Tim Friede[1]

Investigators make increasingly use of adaptive designs to increase the efficiency of their studies. This applies in particular to clinical trial, but also to experiments in other research areas. Trial simulations are an important tool in the planning of such complex designs (e.g. Friede et al, 2020; Jobjörnsson et al, 2022). Recent work in this area also investigates the use of efficient optimization methods to conduct the simulation studies in a timely and resource economic manner (Richter et al, 2022). Beyond the planning of experiments, simulations are used to assess statistical methods comparatively. Although this approach is very common in statistical science, in other areas such as computer science benchmarking based on established databases are more frequently used. Simulations and benchmarking will be contrasted and recommendations on their potentially combined used will be discussed (Friedrich and Friede, 2022) Musuamba et al (2021) define in-silico clinical trials as a "Class of trials for pharmacological therapies or medical devices based on modelling and simulation technologies. Such trials produce digital evidence that can serve in complement to or replacement of in vivo clinical trials for the development and regulatory evaluation of medical therapies." (Table 1, Musuamba et al, 2021). In the context of small populations (e.g. rare diseases, paediatrics) the use of modelling and simulation techniques to substitute clinical data will be discussed.

[1] Universitätsmedizin Göttingen, Institut für Medizinische Statistik

tim.friede@med.uni-goettingen.de

References

1 Benda N, Branson M, Maurer W, Friede T (2010) Aspects of modernizing drug development using scenario planning and evaluation. Drug Information Journal 44: 299-315.

2 Friede T, Nicholas R, Stallard N, Todd S, Parsons N, Valdés-Márquez E, Chataway J (2010) Refinement of the clinical scenario evaluation framework for assessment of competing development strategies with an application to multiple sclerosis. Drug Information Journal 44: 713-718.

3 Friede T, Stallard N, Parsons N (2020) Adaptive seamless clinical trials using early outcomes for treatment or subgroup selection: Methods, simulation model and their implementation in R. Biometrical Journal 62: 1264–1283.

4 Friedrich S, Friede T (2022) On the role of benchmarking data sets and simulations in method comparison studies. (in revision for Biometrical Journal) Preprint https://arxiv.org/abs/2208.01457

5 Jobjörnsson S, Schaak H, Mußhoff O, Friede T (2022) Improving the statistical power of economic experiments using adaptive designs. Experimental Economics (in press).

6 Musuamba et al (2021) Scientific and regulatory evaluation of mechanistic in silico drug and disease models in drug development: Building model credibility. CPT: Pharmacometrics & Systems Pharmacology 10: 804–825.

7 Richter J, Friede T, Rahnenführer J (2022) Improving Adaptive Seamless Designs through Bayesian optimization. Biometrical Journal 64: 948–963.

Using model based optimization for exploring the performance profile of large scale storage systems

Sebastian Krey[1], Nellie Marie Lackschewitz[1], Olaf Mersmann[2] and Julian Kunkel[1]

Storage systems for large scale compute clusters like they are used in High Performance Computing (HPC) clusters are becoming larger and more complex, with changing input and output (I/O) demands depending on usage patterns. Traditional simulation-based HPC workloads create large files that are sequentially written out at the end of the job in a short time period using many nodes in parallel. In contrast, data analytics and machine learning workloads process massive amounts of input data stored in small files and read in random batches over the whole lifetime of the job. This requires different types of storage and configurations that cater to the different workloads.

Optimizing the performance of storage systems is crucial for operating HPC systems efficiently, as the I/O system's performance gap with compute performance is increasing - it is much easier to add additional compute resources compared to scaling a storage system. Moreover, because storage is a shared resource, a single misbehaving job can negatively impact all other users of the storage system. This is in contrast to compute resources which are allocated exclusively to a single job.

Although there are general rules how to achieve optimal I/O performance for some usage patterns, they are often conflicting and heavily dependent on the storage system's technology. As a result, creating an accurate and comprehensive I/O profile for a storage system is time-consuming and site specific.

This work presents a methodology based on design of experiments and model-based optimization to generate such an I/O profile covering sequential, random, metadata-heavy, distributed and shared file, read and write patterns using different file and block sizes while minimizing the number and size of benchmark runs required.

[1] Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG)
[2] Technische Hochschule Köln

sebastian.krey@gwdg.de, nellie-marie.lackschewitz@gwdg.de, olaf.mersmann@th-koeln.de and julian.kunkel@gwdg.de

Application of CNN ensembles to model Fourier transform infrared spectra

Matthias Medl[1], Theresa Scharl[1], Friedrich Leisch[1]

The field of chemometrics concerns itself with the analysis of chemical data by means of mathematical and statistical models. The analysis of spectral data e.g., Fourier transform infrared (FTIR) spectra is one of the central challenges in chemometrics. Historically, partial least squares (PLS) based methods were most commonly used to model spectral data, however with the emergence of larger datasets and advancements in deep learning (DL), DL models have been reported to outperform PLS based methods in multiple studies [1–3]. In all of these studies single convolutional neural networks (CNNs) were used. According to literature, DL ensembles have been capable of outperforming single DL models in many instances [4]. One of the main criticisms of DL ensembles is that their training is computationally more costly compared to single models, which is especially problematic for huge datasets (e.g. ImageNet; $n > 14,000,000$) where training single models can take multiple weeks. This, however, is not a significant challenge when analyzing FTIR datasets as the number of observations ($n < 100,000$) and features ($p < 2,000$) is comparatively small leading to non-prohibitive ensemble training times. Therefore, DL ensembles can be considered an attractive choice to achieve higher predictive performance. Additionally, DL ensembles allow for the easy estimation of the epistemic model uncertainty by simply calculating the variance of predictions made by the individual ensemble members for any given datapoint [5]. When the ensemble members are trained to not only predict the mean response, but also the variation of the response, the aleatoric uncertainty can be quantified as well [6]. These concepts have not been explored in the context of FTIR analysis so far. Thus, we will investigate the potential of DL ensembles in context of FTIR modelling. At first, a benchmark study will be conducted to compare the performance of single CNNs and CNN ensembles resulting from multiple ensemble generation strategies: (i) simple averaging, (ii) weighted averaging, (iii) CNN weight averaging, (iv) Monte Carlo dropout, (v) model fusion, (vi) bootstrapping, (vii) boosting, (viii) stacking and (ix) negative correlation learning. Performance metrics of all methods will be compared for two public FTIR datasets; one dataset where the dry-matter content of mangos should be predicted from FTIR spectra [7] and another one where chemical compounds within soil samples should be quantified [8]. To the author's knowledge no methodological benchmark study has been performed with more than one dataset at a time. Furthermore, uncertainty estimates for the trained ensembles will be analyzed and discussed.

[1] University of Natural Resources and Life Sciences Vienna

matthias.medl@boku.ac.at , theresa.scharl@boku.ac.at , friedrich.leisch@boku.ac.at

References

1 Cui C, Fearn T (2018). Modern practical convolutional neural networks for multivariate regression: Applications to NIR calibration. Chemometrics and Intelligent Laboratory Systems, 182:9–20

2 Mishra P, Passos D (2021). A synergistic use of chemometrics and deep learning improved the predictive performance of near-infrared spectroscopy models for dry matter prediction in mango fruit. Chemometrics and Intelligent Laboratory Systems, 212:104287

3 Ng W, Minasny B, Mendes WdS, Demattê JAM (2020). The influence of training sample size on the accuracy of deep learning models for the prediction of soil properties with near-infrared spectroscopy data. Soil, 6:565–578

4 Ganaie MA, Hu M, Malik AK, Tanveer M, Suganthan PN (2022). Ensemble deep learning: A review. Engineering Applications of Artificial Intelligence, 115:105151

5 Lakshminarayanan B, Pritzel A, Blundell C (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems, 30

6 Scalia G, Grambow CA, Pernici B, Li Y-P, Green WH (2020). Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. Journal of chemical information and modeling, 60:2697–2717

7 Anderson N, Walsh K, Subedi P (2020). Mango DMC and spectra Anderson et al. 2020. https://data.mendeley.com/datasets/46htwnp833/1 References

8 D.E. Beaudette, J.L. Nemecek. Interactive NCSS Map. https://ncss-tech.github.io/AQP/soilDB/NCSS-interactive-map.html

Using the R^2 measure on test data? The evaluation of polygenic prediction models across populations

Andreas Mayr[1], Hannah Klinkhammer[1,2], Tobias Wistuba[1], Carlo Maj[3], Peter Krawitz[2], Christian Staerk[1]

Polygenic risk scores (PRS) quantify the genetic predisposition for different diseases or traits based on individual genotype data and should play an increasingly important role in personalized risk assessment and prevention. In the context of PRS models for continuous traits, the R^2 is a commonly used measure of prediction accuracy. While the R^2 as the coefficient of determination is a well-defined goodness of fit measure for linear models on training data, surprisingly there exist different conflicting definitions for its application to assess the prediction accuracy on test data, which complicates the correct interpretation and comparison of results. Based on large-scale genotype data from the UK Biobank, we compare three common definitions of the R^2 for evaluating the predictive performance of PRS models. Polygenic models for several phenotypes (including height, BMI and lipoprotein A) are derived based on European training data using state-of-the-art boosting methods and are evaluated on various test populations with different ancestries. Our analysis shows that the choice of the R^2 definition can lead to severely different results. In particular, while the definition of the R^2 as the squared correlation between predicted and observed phenotypes always yields values between 0 and 1, definitions of the R^2 incorporating the squared prediction error can also result in negative values, particularly in cases of miscalibrated prediction models for test populations different from the training population. We argue that the choice of the most appropriate definition of the R^2 depends on the aim of the PRS analysis, i.e. whether the PRS should be mainly used for risk stratification in a given cohort or also for the prediction of continuous traits for individual risk assessment. In the latter case, alternative and unambiguously defined measures of prediction accuracy may be preferred.

[1] University of Bonn, Institute for Medical Biometry, Informatics and Epidemiology
[2] University of Bonn, Institute for Genomic Statistics and Bioinformatics
[3] Philipps University of Marburg, Center for Human Genetics

amayr@uni-bonn.de

Advanced statistical modelling of polygenic risk scores via boosting targeted objective functions

Hannah Klinkhammer[1,2], Christian Staerk[1], Carlo Maj[2,3], Peter Krawitz[2], Andreas Mayr[1]

Polygenic risk scores (PRS) can be used to predict a trait or phenotype based on the genetic information of a patient and are based on common genetic variants with low to medium effect sizes. As genotype data are high-dimensional in nature, from a technical perspective it is crucial to develop algorithms that can be applied on large-scale data (large n and large p). A wide range of PRS methods focus on summary statistics from genome-wide association studies (GWAS) based on univariate effect estimates and combine them to a single score. More recently, methods have been developed that can be applied directly on individual-level genotype data to model the variants' effects simultaneously. In this context, we introduced snpboost, a framework that applies statistical boosting on individual-level genotype data to estimate PRS directly via multivariable regression models. By iteratively working on batches of variants, snpboost can deal with large-scale cohort data, e.g. from the UK Biobank ($n \approx 500.000$ and $p \approx 10.000.000$). As the technical obstacles due to the dimensionality of the data are therefore solved, the methodological scope can be now broadened – focusing on the objectives that are really key for the clinical application of PRS. Similar to many other methods, so far, via snpboost we also have focused solely on quantitative and binary traits based on common loss functions such as the squared error and logistic loss functions. Exploiting the modular structure of statistical boosting, we now incorporated more advanced alternatives. As the loss function defines the type of regression problem that is optimized, we effectively extended the snpboost framework to further data situations such as time-to-event and count data. Furthermore, alternative loss functions for continuous outcomes allow us to focus not only on the mean of the conditional distribution but also on other aspects that may more helpful in the risk stratification of individual patients, e.g. median or quantile regression.

References

1 Klinkhammer, H, Staerk, C, Maj, C, Krawitz, P, Mayr A. A statistical boosting framework for polygenic risk scores based on large-scale genotype data. Frontiers in Genetics. 2022(13). doi: 10.3389/fgene.2022.1076440.

[1] University of Bonn, Institute for Medical Biometry, Informatics and Epidemiology
[2] University of Bonn, Institute for Genomic Statistics and Bioinformatics
[3] Philipps University of Marburg, Center for Human Genetics

klinkhammer@imbie.uni-bonn.de, christian.staerk@imbie.uni-bonn.de,
carlo.maj@uni-marburg.de, pkrawitz@uni-bonn.de, amayr@uni-bonn.de

A clinical sensory test recognized as random walk

Jörn Lötsch[1], Alfred Ultsch[3]

Random walks describe stochastic processes that result from a sequence of indeterminate changes in a random variable that are not correlated with past changes. We describe an established sensory test of odor detection threshold as random walks in two sequential parts. Both parts can be described as a biased random walk with highly unequal probabilities of moving toward higher (11%) or lower scores (89%). The walk is complicated by the nesting of two components, the first consisting of the determination of the starting point for the next walk, which consists of the determination of the subsequent turning points for threshold calculation. The first component is unidirectional, i.e., the movement can only go in the direction of lower values. The second part of the test is an up and down movement. The first random walk determines the starting point for the next. Let $t = 0, \ldots, N$ (time) be the length of the Biased Random Walk (BRW) i.e., the number of sniff tests with a guessed result. The time $t = 0$ is the start time of the BRW, which starts at Tstart. For the staircase paradigm to determine the starting points (SPSP), the step width is $Sw = 2$ and the initial starting point is $Tstart = 16$ at $t = 0$. The evolution of mean expectation and the 99.7% range in this algorithm. For the staircase paradigm at threshold T, Tstart is the result of the start point finding (SPSP) and $Sw = 1$. The distribution of the threshold pdf(t) reached at t can generally be calculated as a Gaussian $pdf(t) = N(M, V) = N(M(t), V(t))$, with mean $M(t) = 2 * p * t - sw * t + Tstart$ and variance $V(t) = p * q * t$. The position from which the first test part begins, and the length of the random walk in the subsequent second part were critical factors in the probability of accidentally achieving high test scores (Figure 1).

Recognition of the probabilistic background of the sensory test led to the proposal of a modification that raised the established cut-off of test failure, indicating inability to smell, from representing the 87th quantile of random test scores to representing the 97th quantile. The results likely apply to other sensory tests using the staircase paradigm that can also be described as random walks.

[1] Goethe University, Institute of Clinical Pharmacology
[2] University of Marburg, DataBionics Research Group

j.loetsch@em.uni-frankfurt.de, ultsch@mathematik.uni-marburg.de

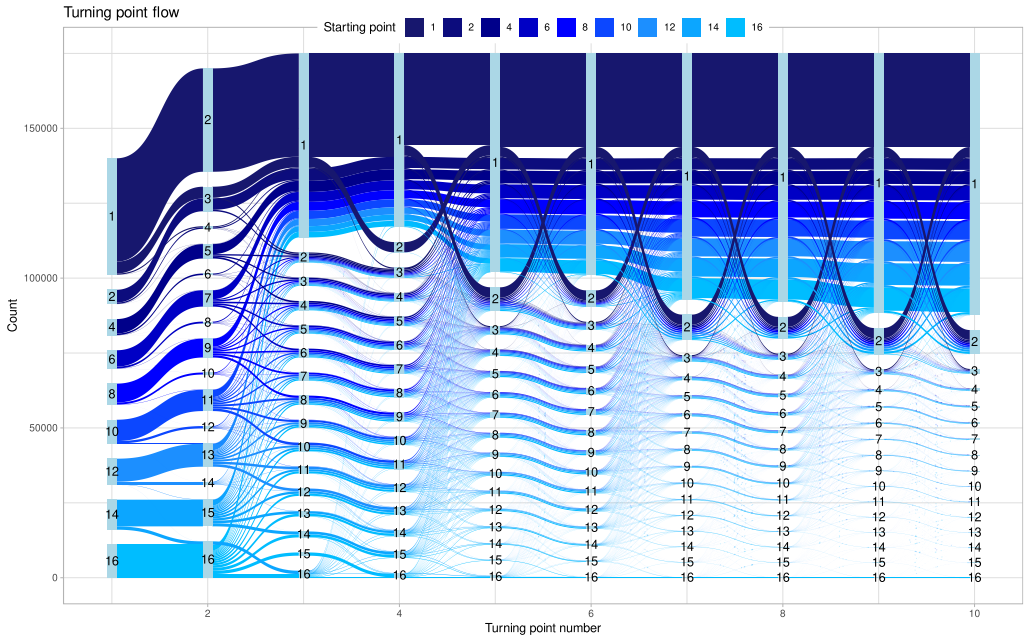


Figure 1: Dependency of the final test score when all responses were guessed from the starting point (25,000 simulations). The Sankey plot shows the flow of turning points through the test. The bands are colored according to the starting point. Starting from a high position in a two-dimensional space involves a non-negligible chance of staying in high positions, even if the ability that would drive an upward movement is lacking. According to the rules of the test, the probability of a downward movement is still quite high. On the other hand, if the starting position is low, it is very difficult to reach a high position by chance.

Dimension reduction by spatial components analysis improves pattern detection in multivariate spatial data

Niklas Kleinenkuhnen[1,2], David Köhler[3], Matthias Schmid[3], Peter Tessarz[2], Achim Tresch[1]

The field of spatial transcriptomics (ST) measures the transcription of genes with regard to the location of gene expression on a dense grid of spots, enabling insights into cellular processes[1]. It is a highly active research field with many applications in life sciences and pathology, leading to its selection as Method of the Year 2020 by Nature[2]. Despite its relevance, there is a lack of tools specialised for the analysis of ST data. Such a tool should be able to identify spatial patterns as well as genes following these patterns while being robust regarding the often noisy data. For this task, a novel algorithm, Spatial Components Analysis (SpaCo), combines the widely applied method of principal component analysis with known statistics for univariate spatial data such as Geary's C to multivariate data. It can be used for dimension reduction, inference on a gene's spatial variability and further downstream tasks such as clustering. The algorithm includes a novel statistic to evaluate spatial variability including spatial interaction between genes, enabling identification of spatially variable genes and regions that are not detected when only considering each gene's spatial activity separately. This way, SpaCo is in several aspects superior to commonly used methods in ST analysis such as principal component analysis as a preprocessing step, or gene-wise Geary's C for the identification of spatially variable genes. SpaCo can also be applied for spatial denoising of genes by using information of similarly expressed genes to mitigate issues in data generation. The performance of SpaCo was demonstrated on the example of mouse brain and liver ST data as well as simulated data. It is assessed by SpaCo's ability to detect genes with spatial variability as measured in the widely applied spatial statistic Geary's C. The simulation study confirmed these results while also demonstrating the robustness of the algorithm towards noise.

References

1 Lowe, Rohan et al. (2017). "Transcriptomics technologies". In: PLoS computational biology 13.5, e1005457

2 Marx, Vivien (2021). "Method of the Year: spatially resolved transcriptomics". In:Nature methods 18.1, pp. 9–14.

[1] University of Cologne, Institute of Medical Statistics and Computational Biology
[2] Max Planck Institute for Biology of Ageing, Cologne
[3] University of Bonn, Department of Medical Biometrics, Informatics and Epidemiology

nkleinen@smail.uni-koeln.de, koehler@imbie.uni-bonn.de, matthias.c.schmid@uni-bonn.de, achim.tresch@uni-koeln.de, Peter.Tessarz@age.mpg.de

Constraint-based attractor search in quantum Boolean networks

Felix M. Weidner[1], Mirko Rossini[2], Joachim Ankerhold[2], Hans A. Kestler[1]

Boolean networks (BNs) are simple dynamical models for gene interactions, in which each gene is either active or inactive at a given point in time. The network's state is updated by the application of logical rules connecting genes via the operators AND, OR and NOT, leading to biologically relevant stable states called attractors.
Our previous work has implemented a proof of principle for the simulation Boolean networks on quantum computers [1]. This offers various benefits over classical hardware. Notably, a linearly growing number of qubits can be used to perform computations in an exponentially growing high-dimensional state space, thus capturing the complex dynamics of BNs. Furthermore, dedicated quantum algorithms such as Grover's search algorithm [2] offer further improvements in complexity.
Constraint-based heuristics based on Boolean satisfiability (SAT) have proven useful for attractor search in BNs [3]. We present a modification of Grover's algorithm for this purpose, in which previously measured attractors are added as constraints to a quantum circuit. This leads to a suppression of parts of the state space leading to already known states [4], aiding also in the identification of rare phenotypes.
We further investigate the use of quantum annealers for such constraint-based search approaches, as these provide a significantly larger number of qubits [5].

References

1 Weidner, F. M., Schwab, J. D., Wölk, S., Rupprecht, F., Ikonomi, N., Werle, S. D., ... & Kestler, H. A. (2023). Leveraging quantum computing for dynamic analyses of logical networks in systems biology. Patterns, 4(3), 100705.

2 Grover, L. K. (1997). Quantum Mechanics Helps in Searching for a Needle in a Haystack. Physical Review Letters, 79(2), 325.

3 Dubrova, E., & Teslenko, M. (2011). A SAT-Based Algorithm for Finding Attractors in Synchronous Boolean Networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 8(5), 1393-1399.

4 Liu, Y., & Ouyang, X. (2013). A quantum algorithm that deletes marked states from an arbitrary database. Chinese Science Bulletin, 58(19), 2329-2333.

5 Su, J., Tu, T., & He, L. (2016, June). A Quantum Annealing Approach for Boolean Satisfiability Problem. In Proceedings of the 53rd Annual Design Automation Conference (pp. 1-6).

[1] Ulm University, Institute of Medical Systems Biology
[2] Ulm University, Institute for Complex Quantum Systems

felix.weidner@uni-ulm.de, mirko.rossini@uni-ulm.de, joachim.ankerhold@uni-ulm.de, hans.kestler@uni-ulm.de

Establishing a trustworthy signalling entropy calculation for biological processes analysis

Ana Stolnicu[1], Nensi Ikonomi[1*], Johann M. Kraus[1], Hans A. Kestler[1]*

Signalling entropy calculation could provide valuable insights in the mechanisms underlying complex biological systems, including cell differentiation, cancer initiation, and development [1,2]. For estimating biological changes within a system, this measure can be computed by combining the information acquired from pattern profiles with a protein interaction network (PIN). Since it is assumed that the available PINs contain a significant amount of false interactions [3], various correction strategies have been proposed to enhance accuracy by removing irrelevant information [4-7]. The primary objective of this study is to identify interaction networks that exhibit high reliability, by reducing the occurrence of false positive protein interactions, thus enhancing quality of the signalling entropy calculation. At first, we examined the impact of different percentages of randomly incorporated protein-protein interactions in the network on the signalling entropy of a synthetic gene expression dataset. Subsequently, we analysed the alterations among the signalling entropies after integrating distinct PINs with real datasets. In this context, we considered interactions from well-known databases like Pathway Commons (PC)[8], STRING[6,7] and the Biological General Repository for Interaction Datasets (BioGRID)[9], together with the union and the intersection of those, in conjunction with cell differentiation or cancer datasets. The study revealed a trend wherein the entropies exhibit a declining pattern with an increase in the quantity of randomly added interactions. However, this trend reverses after a certain threshold, approximately at 50% of the inserted edges, where the entropies start to increase again. Upon reaching a threshold of 50% to 90% of interactions, the network may exhibit properties similar to those of a random graph, wherein the nodes possess roughly equivalent degrees of connectivity. When analysing cell differentiation and tumour data, we examined the statistical significance of the entropy differences among the classes. It was observed that the entropies did not exhibit significant differences when the intersection of the PINs was combined with any of the datasets. Although certain cases showed improvement in the outcome after applying a correction, the intersection might be unsuitable due to the restricted overlap observed across the protein interaction databases. Meanwhile, the utilisation of any of the other PINs yielded encouraging results, and in certain instances, exhibited improvement subsequent to the application of a correction. The STRING database, more than BioGRID and PC, improved substantially after filtering for misleading protein links. An effective strategy could involve consolidating protein interactions from various databases, potentially mitigating the occurrence of false negative interactions and minimising the influence of false positive interactions on the final results.

[1] Ulm University, Institute of Medical Systems Biology
[*] Equal contribution

ana.stolnicu@uni-ulm.de, nensi.ikonomi@uni-ulm.de, johann.kraus@uni-ulm.de, hans.kestler@uni-ulm.de

References

1 Banerji, C. R. S. et al. (2013). Cellular network entropy as the energy potential in Waddington's differentiation landscape. Scientific Reports, 3(1):3039.

2 Cheng, F. Et al. (2016). Investigating cellular network heterogeneity and modularity in cancer: a network entropy and unbalanced motif approach. BMC Systems Biology, 10(S3):65.

3 Kuchaiev, O. et al. (2009). Geometric De-noising of Protein- Protein Interaction Networks. PLOS Computational Biology, Volume 5(8) :e1000454

4 Csárdi, G. and Nepusz, T. (2006). The igraph software package for complex network research. InterJournal Complex systems.

5 Yu, G. et al. (2010). Gosemsim: an r package for measuring semantic similarity among go terms and gene products. Bioinformatics.

6 Franceschini, A. et al. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Research, 41(D1):D808-D815.

7 Szklarczyk, D. et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Research, 43(D1):D447-D452.

8 Cerami, E. et al. (2011). Pathway commons, a web resource for biological pathway data. Nucleic Acids Research.

9 Stark, C., Breitkreutz, B.-J., Reguly, T. et al. (2006). Biogrid: a general repository for interaction datasets. Nucleic Acids Research.

A random forest pseudo-value approach for modeling restricted mean survival times

Alina Schenk[1], Vanessa Basten[1], Matthias Schmid[1]

Because of its simple interpretation, the restricted mean survival time RMST is often suggested as measure for the treatment effect in time-to-event analysis in randomized controlled trials. The RMST is defined as the area under the survival function S(t) up to $\tau > 0$ and can readily be interpreted as the life expectancy between t = 0 and a specific time horizon ($t = \tau$). Besides, the usage of RMST is advantageous in the sense of having minimal assumptions on the survival process (such as, e.g. proportional hazards of a treatment effect) [1]. In practice, the direct modeling of the RMST conditional on a set of covariates X is of particular interest when investigating covariate effects (e.g. treatment effects) on the expected life time. However, due to (right-censored) data leading to partly unobserved survival times, modeling of the RMST is not straightforward and requires special estimation and modeling techniques. Here, we use individual-specific leave-one-out jackknife pseudo-values providing, on average, a distribution-free estimate of the restricted mean survival time at τ for the entire sample [1]. These values can be treated in the same way as a continuous outcome in standard regression models for estimating the RMST [2]. Currently, the most popular strategy for modeling those pseudo-values is the estimation of main covariate effects on the RMST via a generalized estimation equation (GEE) approach. While the GEE method yields unbiased estimates in the case of correct model specification and independent censoring, the commonly applied model is restricted to the inclusion of main covariate effects only. The inclusion of more flexible effect terms (e.g. interactions between the covariates) by pre-specification is often infeasible, as it would require prior knowledge on the, usually hidden, interaction structure in the data. To extend standard pseudo-value models by higher-order interaction terms, we propose an alternative modeling approach with the aim of the estimation of individual-specific RMST, while at the same time selecting the most relevant covariates in a data-driven way. Conceptually, our modeling approach comprises the estimation of pseudo-values for RMST used for the specification of a flexible random forest regression algorithm. We will present a simulation study and an application investigating the ability of our method to filter out relevant covariates and to accurately estimate RMST as well as a comparison to established methods for RMST estimation and modeling.

[1] University of Bonn, Department of Medical Biometry, Informatics and Epidemiology

schenk@imbie.uni-bonn.de, basten@imbie.uni-bonn.de, matthias.c.schmid@uni-bonn.de

References

1 P. Royston and M. K. B. Parmar. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. Statistics in Medicine, 30:2409–2421, 2011.

2 P. K. Andersen and M. Pohar Perme. Pseudo-observations in survival analysis. Statistical Methods in Medical Research, 19:71–99, 2010.

Distributional analysis and data augmentation of (multi) organ 3D data

Michael Selle[1], Magdalena Kircher[1], Cornelia Schwennen[2], Christian Visscher[2], Klaus Jung[1]

In medicine, a reference population, commonly healthy individuals from a stratified sample group, form together with established reference intervals a popular tool for medical decision-making [1]. Both unsupervised and supervised learning can be used to identify 'outliers' that deviate from a reference population. In medical imaging, the geometry of an organ can be a useful indicator for diagnosis [2]. However, patient data is often limited due to small population sizes and privacy concerns [3,4]. Therefore we follow two approaches: (1) unsupervised learning to quickly identify abnormal shapes within medical imaging data and (2) supervised learning of medical imaging data with paired clinical data after data augmentation. For all analyses, publicly available datasets were used: CT-ORG [5] and AbdomenCT-1k [6] for the unsupervised learning as well as TCGA-LIHC [7] and HCC-TACE-Seg [8] for the supervised learning. Three datasets contained organ annotations from preceding segmentation while on the remaining dataset segmentation still had to be performed. To reduce noise and organ complexity, various 2D-morphological operations were employed. Then, for each organ, 3D mesh objects were created. The surface mesh objects were downsampled and smoothened prior to point-set registration. From each set of registered point clouds, shape features such as local curvatures and distances were computed. For unsupervised learning, the feature matrices from multiple organs were projected altogether into the same 2-dimensional space via multiple co-inertia analysis or separately via principal component analysis or t-SNE to determine the distributional location of a single organ or an individual in comparison to the sample population. For supervised learning, a statistical shape model was created that represents the mean shape and its variation within the population. New clinically plausible shapes were generated and used to augment paired tabular clinical patient data and vice versa. The unsupervised learning revealed that single organs or entire invidiuals can be separated by shape from the sample population. Since no patient records were available for this approach, it cannot be concluded if an 'outlier' is of technical or biological nature. As for supervised learning, the results are currently being collected.

[1] University of Veterinary Medicine Hannover, Department of Animal Breeding and Genetics
[2] University of Veterinary Medicine Hannover, Institute for Animal Nutrition

michael.selle@tiho-hannover.de, magdalena.kircher@tiho-hannover.de,
cornelia.schwennen@tiho-hannover.de, christian.visscher@tiho-hannover.de,
klaus.jung@tiho-hannover.de

References

1 Mave, V., Kulkarni, V., Bharadwaj, R., Khandekar, M., Gupta, A., & Gupte, N. (2012). Determination of a reference interval in a population. The National Medical Journal of India, 25(1), 33-34.

2 Razzak, M. I., Naz, S., & Zaib, A. (2018). Deep learning for medical image processing: Overview, challenges and the future. Classification in BioApps: Automation of Decision Making, 323-350.

3 Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of big data, 6(1), 1-48.

4 Garcea, F., Serra, A., Lamberti, F., & Morra, L. (2022). Data augmentation for medical imaging: A systematic literature review. Computers in Biology and Medicine, 106391.

5 Rister, B., Yi, D., Shivakumar, K., Nobashi, T., & Rubin, D. L. (2020). CT-ORG, a new dataset for multiple organ segmentation in computed tomography. Scientific Data, 7(1), 381.

6 Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., ... & Yang, X. (2021). Abdomenct-1k: Is abdominal organ segmentation a solved problem?. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10), 6695-6714.

7 Erickson, B., Kirk, S., Lee, Y., Bathe, O., Kearns, M., Gerdes, C., ... & Lemmerman, J. (2016). Radiology data from the cancer genome atlas liver hepatocellular carcinoma [TCGA-LIHC] collection. Cancer Imaging Arch, 10, K9.

8 Moawad, A. W., Fuentes, D., Morshid, A., Khalaf, A. M., Elmohr, M. M., Abusaif, A., ... & Elsayes, K. M. (2021). Multimodality annotated HCC cases with and without advanced imaging segmentation [data set]. The Cancer Imaging Archive.

A semiparametric thin plate spline spatial model using Bayesian computation

Joaquin Cavieres[1], Cole.C. Monnahan[2], Paula Moraga[3]

Modelling a large number of spatial observations is expensive in computational terms. Often, georeferenced data are modeled using a Gaussian random field as spatial random effect. However, this implies a large computational cost of estimation (O(n3)) due to the factorization of a dense n × n covariance matrix for a large number of observations. To deal with this, [1] proposed to use a Gaussian Markov random field as an approximation of the Gaussian random field. This methodology works well for a special class of models, the called "latent Gaussian models", but if the interest is in making full Bayesian inference (not approximated), the chains convergence is difficult to obtain given the correlation between the hyperparameters of the Gaussian random field. In this work, we propose to use a low rank approximation of a thin plate spline as spatial random effect in a semiparametric Bayesian spatial model, based mainly in the results obtained by [3]. The structure of the thin plate spline is known and determined completely by a kernel (conditionally positive definite) function, hence we do not need to specify a particular covariance function and neither be worried about the estimation of parameters which govern this structure. Since the Kernel matrix is dense as well, we use the proposal by [2], who uses the Lanczos iteration to get the truncated eigen- decomposition in O(kn2) operations, by iteratively building up a tri-diagonal matrix, the eigenvalues of which converge to those required, as iteration proceeds. The Bayesian inference is done by using the Hamiltonian Monte Carlo algorithm of the probabilistic software Stan [4]. This algorithm has, in some cases, better performance than Metropolis- Hasting algorithm since it uses the derivatives of the target distribution to explore the space's parameters in an efficient way. The preliminary results from the simulations showed that the computational time of estimation for the thin plate spline model is less than the model that uses a Gaussian Markov random field as spatial random effect. In a real application, the thin plate spline model has a better performance than the Gaussian Markov field model using the Leave-one-out cross-validation criterion, and the computational cost of estimation was considerably less as well. The main advantage to use a the thin plate spline as spatial random effect is the simple parameterization and the fast convergence of the chains considering the complexity of the model.

[1] Bayreuth University, Geography department
[2] National Marine Fisheries Service (NOAA), Seattle
[3] KAUST, Computer, Electrical and Mathematical Sciences and Engineering Division

joaquin.cavieres@uni-bayreuth.de

References

1 Lindgren,F.,Rue,H.,&Lindström,J.(2011).An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(4), 423-498.

2 Wood, S. N. (2003). Thin plate regression splines. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65(1), 95-114.

3 Cavieres, J., Ibacache-Pulgar, G., & Contreras-Reyes, J. E. (2023). Thin plate spline model under skew-normal random errors: estimation and diagnostic analysis for spatial data. Journal of Statistical Computation and Simulation, 93(1), 25-45.

4 Gelman, A., Lee, D., & Guo, J. (2015). Stan: A probabilistic programming language for Bayesian inference and optimization. Journal of Educational and Behavioral Statistics, 40(5), 530-543.

Berichte des Instituts für Medizinische Systembiologie

Herausgeber:
Institut für Medizinische Systembiologie
Universität Ulm
89081 Ulm