ulm university universität ulm

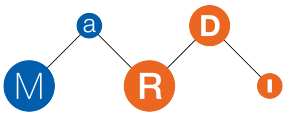
Statistical Computing 2024

Abstracts der 54. Arbeitstagung

HA Kestler, JM Kraus (eds)

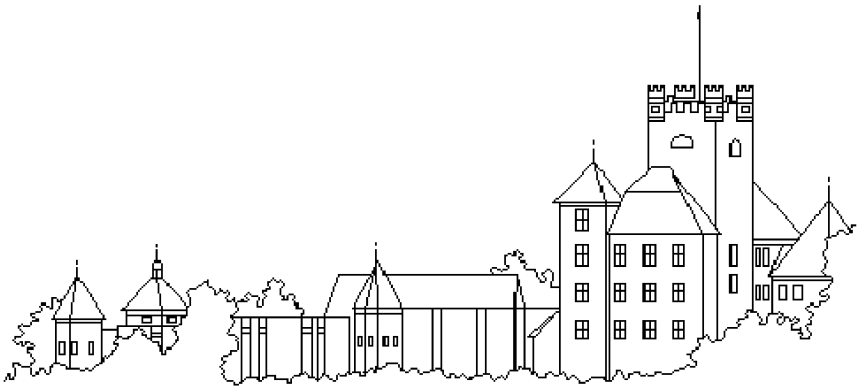
Berichte des Instituts für Medizinische Systembiologie

Nr. 2024-01
July 2024



Universität Ulm

Statistical Computing 2024



54. Arbeitstagung

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

28.07. - 31.07.2024, Schloss Reisensburg (Günzburg)

Workshop Program

Sunday, July 28, 2024

18:30 – 20:00		**Dinner**
		Introduction: H. A. Kestler
20:00 – 21:00	Achim Tresch (Köln)	Feature extraction for multivariate spatial data

Monday, July 29, 2024

08:20		**Opening of the workshop: H. A. Kestler**
08:30 – 10:35		**Chair: M. Schmid**
08:30 – 08:55	Alina Schenk (Bonn)	Modeling the restricted mean survival time as a function of time horizons with pseudo-value regression trees
08:55 – 09:20	Lukas Burk (Bremen)	A Large-Scale Neutral Comparison Study of Survival Models on Low-Dimensional Data
09:20 – 09:45	Gunther Schauberger (München)	Boosting for Conditional Logistic Regression
09:45 – 10:10	Colin Griesbach (Göttingen)	Additive Mixed Models for Location, Scale and Shape via Gradient Boosting Techniques
10:10 – 10:35	Alexandra Daub (Göttingen)	A Balanced Statistical Boosting Approach for GAMLSS via New Step Lengths
10:35 – 11:00		**Coffee Break**
		Introduction: H. A. Kestler
11:00 – 12:00	Eyke Hüllermeier (München)	Uncertainty Quantification in Machine Learning: From Aleatoric to Epistemic
12:00 – 13:30		**Lunch**
		Introduction: C. Griesbach
13:30 – 15:30	Bernd Bischl, Fiona Ewald (München)	Tutorial: Interpretable Machine Learning
15:30 – 16:00		**Coffee Break**
16:00 – 17:40		**Chair: A. Mayr**
16:00 – 16:25	Ludwig Bothmann (München)	Causal Fair Machine Learning
16:25 – 16:50	Max Westphal (Bremen)	Estimand-aligned data splitting and performance estimation in applied machine learning
16:50 – 17:15	Pegah Golchian (Bremen)	MissARF: Adversarial random forests for imputing missing values
17:15 – 17:40	Marietta Hamberger (Ulm)	bioXfuse: An Integrated Toolkit for In-Depth Exploration of Semantic and Contextual Landscapes in Biomedical Research
		Introduction: H. A. Kestler
17:40 – 18:30	Göran Kauermann (München)	Human in the Loop – Uncertainty of Labelling in Supervised Machine Learning
18:30 – 20:00		**Dinner**

Tuesday, July 30, 2024

08:30 – 10:35		**Chair: S. Hoffmann**
08:30 – 08:55	Tobias Weckop (Erlangen)	Robust estimation of distributional regression models using artificial neural networks
08:55 – 09:20	Moritz Hermann (München)	Dimensionality and Distance: Curse or Blessing? Geometrical Aspects of Nearest Neighbor Computation in High-Dimensional Data
09:20 – 09:45	Paul Kaftan (Ulm)	Registration of Dynamic 1H-MRI Series for Fourier-based Lung Functional Analysis
09:45 – 10:10	Metehan Oruc (Ulm)	Strategies for reducing heterogeneity in clinical trials through optimized group assignment
10:10 – 10:35	Jörn Lötsch (Frankfurt)	How to impute if you must: Selecting the Appropriate Missing Value Imputation Strategy for Cross-Sectional Biomedical Numerical Data
10:35 – 11:00		**Coffee Break**
		Introduction: M. Schmid
11:00 – 12:00	Sarah Friedrich (Augsburg)	Regularization methods in clinical biostatistics: State-of-the art and possibilities for improvement
12:00 – 13:30		**Lunch**
		Introduction: S. Haug
13:30 – 14:30	Sara El-Gebali (Schweden)	From Data to Discovery: DataCite's Role in FAIR Data Management
14:30 – 15:30	Roman Hornung (München)	Reproducibility at the Biometrical Journal: Simple tips for writing and publishing clear code to ensure reproducible results
15:30 – 16:00		**Coffee Break**
16:00 – 16:25		Poster session
16:25 – 18:05		**Chair: J. Kraus**
16:25 – 16:50	Anna von Plessen (Göttingen)	Extending Gradient Boosting Frameworks for High-Dimensional MEG Data Analysis in Neurophysiological Research
16:50 – 17:15	Marisa Lange (Göttingen)	Distributional Regression for Lungfunction of Cystic Fibrosis Patients with a Special Focus on Center Specific Effects
17:15 – 17:40	Denna Langhans (Ulm)	Determining the urgency of surgery of retinal detachment based on deep learning
17:40 – 18:05	Sabine Hoffmann (München)	Accounting for complex structures of aleatoric and epistemic uncertainty through problem-tailored MCMC algorithms
18:05 – 18:30		Working group meeting – **Statistical Computing 2025**
18:30 – 20:00		**Dinner**

Wednesday, July 31, 2024

08:30 – 10:35		**Chair: O. Zadorozhnyi**
08:30 – 08:55	Matthias Medl (Wien)	Comparison of Design of Experiments and Gaussian Process optimization for the optimization of a simulated bioprocess
08:55 – 09:20	Tobias Hepp (Erlangen)	Component-wise gradient boosting for mixtures of distributional regression models
09:20 – 09:45	Milena Wünsch (München)	On the handling of method failure in comparison studies
09:45 – 10:10	Lisa Schönenberger (Dornbirn)	Optimal Scaling of an Algorithmic Parameter in Restart Strategies
10:10 – 10:35	Andreas Mayr (Bonn)	A copula boosting approach for dependent censoring
10:35 – 11:00		**Coffee Break**
11:00 – 11:50		**Chair: C. Griesbach**
11:00 – 11:25	Hannah Marchi (Bielefeld)	Development of a recommender system for targeted antibiotic therapy in sepsis
11:25 – 11:50	Tobias Nietsch (Ulm)	Inductive Logic Programming for Single-cell Analysis
12:00 – 13:30		**Lunch**

Feature extraction for multivariate spatial data

Achim Tresch[1]

1 Institute for Medical Statistics and Computational Biology, Cologne

achim.tresch@uk-koeln.de

Modeling the restricted mean survival time as a function of time horizons with pseudo-value regression trees

Alina Schenk[†], Matthias Schmid[†]

In recent years, the restricted mean survival time (RMST) has become an increasingly important estimand in time-to-event studies. Defined as the restricted area under the survival function over a specified follow-up period $[0, \tau]$, the RMST serves as a comprehensive summary measure on the survival time scale, representing the average event-free survival time up to the time horizon τ. In practice, directly modeling the RMST conditional on a set of covariates X is particularly valuable for investigating the effects of treatments, exposures, and other variables of interest on the expected lifetime. Several methods are available for regression modeling of RMST, most of which rely on leave-one-out jackknife pseudo-values or employ an inverse-probability-of-censoring weights approach to account for censored survival times. However, most of these approaches model the RMST at one single fixed time horizon $\tau > 0$, which must be choosen before estimating covariate effects. Selecting an appropriate value for τ can be challenging and is extensively discussed in the literature. One way to avoid fixing the time horizon τ is to model the RMST as a function of τ, as demonstrated by Zhong and Schaubel (2022) [1]. This approach allows to estimate the RMST at various time horizons τ through one single model and enables the estimation of time-varying covariate effects. The authors employ an inverse-probability-of-censoring weights approach for modeling the RMST as a function of τ. However, this method requires pre-selection and predefinition of covariates and more complex interaction terms. Including flexible effect terms, such as interactions, by pre-specification is often infeasible, as it would require detailed knowledge on the data structure.

To enable data-driven variable selection and identification of interaction effects on the RMST, we propose modeling the RMST as a function of τ using pseudo-value regression trees (PRT) [2]. PRT are characterized by a multivariate regression tree built on a pseudo-value outcome and by successively fitting a set of regularized additive models to the data in the nodes of the tree using gradient boosting. Initially, PRT were developed as a direct modeling approach for survival probabilities on a grid of time points. We suggest to adapt this approach to flexibly model the RMST as a function of τ using continuous pseudo-values for the RMST on a grid of time horizons. Similar to the approach presented by Zhong and Schaubel (2022), our alternative modeling approach models RMST values at various time horizons τ simultaneously and incorporates time-varying covariate effects. We will present a simulation study and a real-world application to demonstrate the properties of the proposed method.

[†] Institute for Medical Biometry, Informatics and Epidemiology, Medical Faculty, University of Bonn, Venusberg-Campus 1, 53127 Bonn, Germany

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

References

1 Zhong Y, Schaubel DE. *Restricted mean survival time as a function of restriction time*. Biometrics, 78(1), 192-201 (2022)

2 Schenk, A., Berger, M. & Schmid, M. *Pseudo-value regression trees*. Lifetime Data Anal 30, 439–471 (2024)

A Large-Scale Neutral Comparison Study of Survival Models on Low-Dimensional Data

Lukas Burk[1,2,3,4], John Zobolas[5], Bernd Bischl[2,4], Andreas Bender[2,4], Marvin Wright[1,3,6], Raphael Sonabend[7,8]

This work presents the first large-scale neutral benchmark experiment focused on single-event, right-censored, low-dimensional survival data — the most common type of data present in clinical research. Benchmark experiments are essential in methodological research to scientifically compare new and existing model classes through proper empirical evaluation. Existing benchmarks in the survival literature are often narrow in scope, focusing, for example, on high-dimensional data. Additionally, they may lack appropriate tuning or evaluation procedures, or are qualitative reviews, rather than quantitative comparisons. This comprehensive study aims to fill the gap by neutrally evaluating a broad range of methods and providing generalizable conclusions. We benchmark 18 models, ranging from classical statistical approaches to many common machine learning methods, on 32 publicly available datasets. The benchmark tunes for both a discrimination measure and a proper scoring rule to assess performance in different settings. Evaluating on 8 survival metrics, we assess discrimination, calibration, and overall predictive performance of the tested models. Using discrimination measures, we find that no method significantly outperforms the Cox model. However, Accelerated Failure Time models were able to achieve significantly better results with respect to overall predictive performance as measured by the right-censored log-likelihood. Machine learning methods that performed comparably well include Oblique Random Survival Forests under discrimination, and Cox-based likelihood-boosting under overall predictive performance. We conclude that in the standard survival analysis setting of low-dimensional, right-censored data, the Cox Proportional Hazards model remains a simple and robust method, sufficient for practitioners.

[1] Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen, Germany
[2] Department of Statistics, LMU Munich, Munich, Germany
[3] Faculty of Mathematics and Computer Science, University of Bremen, Bremen, Germany
[4] Munich Center for Machine Learning (MCML), Munich, Germany
[5] Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital, Oslo, Norway
[6] Department of Public Health, University of Copenhagen, Copenhagen, Denmark
[7] OSPO Now, London, UK
[8] Imperial College, London, UK

burk@leibniz-bips.de

Boosting for Conditional Logistic Regression

Gunther Schauberger[1] and Stefanie J. Klug[1]

In matched case-control studies, each diseased person (*case*) is matched to one or more *controls*. The controls are persons free from the disease under consideration but are similar or equal to the respective case with respect to defined matching variables (e.g., age, sex, and area of residence). Matching is an important tool to account for multiple sources of confounding. Matching creates strata within the data, which have to be considered during the analysis of matched case-control studies. Conditional logistic regression is the standard method for the analysis of matched case-control data.

However, conditional logistic regression implicitly uses rather restrictive assumptions about the data-generating process. It assumes linearity and requires the user to incorporate interactions. Machine-learning techniques could be valuable alternatives as they typically allow for more flexibility beyond linear functions. Standard versions of popular machine-learning techniques are unable to deal with the special stratified nature of the data. Up to now, only a tree-based method [1] exists as a more flexible alternative to ordinary conditional logistic regression.

We propose to use boosting for the analysis of matched case-control studies based on the general principle of gradient-based boosting [2] as implemented in the add-on package *mboost* [3] in R. The main idea of boosting is to update the model step by step using so-called base-learners. Besides simple linear base-learners, also p-splines, decision trees, Markov-random fields, or random effects could, for example, be chosen. In each iteration, only one base-learner is selected for a model update. The main tuning parameter is the number of boosting iterations, resulting in "early stopping". Using boosting for the analysis of matched case-control studies allows to build a large variety of models simply by combining different types of base-learners. Typically, when analysing matched case-control studies, a dedicated exposure variable is of special interest. This exposure variable can be modelled using linear base-learners for easier interpretation, while other (confounding) variables can be modelled more flexibly. Variable selection can easily be performed by using the framework provided by *mboost*, either using stability selection or cross-validation.

The proposed method is illustrated using data from a matched case-control study on cervical cancer [4]. In an outlook, we will show how the method can also be applied outside of matched case-control studies, in particular in discrete choice modelling. An example from travel-mode choice is presented.

References

[1] Schauberger G, Tanaka LF, Berger M. A tree-based modeling approach for matched case-control studies. Statistics in Medicine. 2023; 42(5): 676–692. doi:10.1002/sim.9637

[2] Hofner, B., A. Mayr, N. Robinzonov, and M. Schmid (2014). Model-based boosting in R: A hands-on tutorial using the R package mboost. Computational Statistics 29, 3–35.

[3] Hothorn, T., P. Buehlmann, T. Kneib, M. Schmid, and B. Hofner (2010). Model-based boosting 2.0. Journal of Machine Learning Research 11, 2109–2113.

[4] Tanaka, L. F., D. Schriefer, K. Radde, G. Schauberger, and S. J. Klug (2021). Impact of opportunistic screening on squamous cell and adenocarcinoma of the cervix in Germany: A population-based case-control study. PLOS ONE 16 (7), 1–17.

---

[1] Technical University of Munich, Germany; TUM School of Medicine and Health, Chair of Epidemiology

gunther.schauberger@tum.de, stefanie.klug@tum.de

Additive Mixed Models for Location, Scale and Shape via Gradient Boosting Techniques

Colin Griesbach[1] *and Elisabeth Bergherr*[1]

In this work we adapt recent findings from statistical boosting [1, 4, 5] in order to construct an estimation approach for distributional regression [2] including random effects. The algorithm is applied to registry data provided by the German Cystic Fibrosis Registry [3] where the subject-specific evolution of each patients lung function and its corresponding distributional parameters are modelled.

References

[1] Griesbach, C., Säfken, B. and Waldmann, E.: Gradient Boosting for Linear Mixed Models. The International Journal of Biostatistics. 17(2), 317–329 (2021).

[2] Mayr, A., Fenske, N., Hofner, B. et al.: Generalized additive models for location, scale and shape for high dimensional data — a flexible approach based on boosting. Journal of the Royal Statistical Society, Series C. 61(3), 403–427 (2012).

[3] Nährlich, L., Burkhart, M. and Wosniok, J.: Deutsches Mukoviszidose-Register — Berichtsband 2022. Bonn: Mukoviszidose e.V. & Mukoviszidose Institut gGmbH (2022).

[4] Thomas, J., Mayr, A., Bischl, B. et al.: Gradient boosting for distributional regression: faster tuning and improved variable selection via noncyclical updates. Statistics and Computing, 28, 673–687 (2018).

[5] Waldmann, E., Taylor-Robinson, D., Klein, N., et al.: Boosting joint models for longitudinal and time-to-event data. Biometrical Journal, 59(6), 1104–1121 (2017).

[1] Georg-August-Universität Göttingen, Platz der Göttinger Sieben 3, 37073 Göttingen

colin.griesbach@uni-goettingen.de, elisabeth.bergherr@uni-goettingen.de

A Balanced Statistical Boosting Approach for GAMLSS via New Step Lengths

Alexandra Daub[1], Andreas Mayr[2], Boyao Zhang[1], Elisabeth Bergherr[1]

Component-wise gradient boosting algorithms are popular for their intrinsic variable selection and implicit regularization, which can be especially beneficial for very flexible model classes. When estimating generalized additive models for location, scale and shape (GAMLSS) by means of a component-wise gradient boosting algorithm, an important part of the estimation procedure is to determine the relative complexity of the different submodels. Existing methods (e.g., [1], [2]) either suffer from a computationally expensive tuning procedure or can be biased by structural differences in the negative gradients' sizes, which, if encountered, lead to imbalances between the different submodels.
Shrunk optimal step lengths have been suggested by Zhang et al. [3] to replace small fixed step lengths for a non-cyclical boosting algorithm limited to a Gaussian response variable in order to address this issue. We propose a new adaptive step length approach that accounts for the relative size of the fitted base-learners to ensure a natural balance between the different submodels. The balanced non-cyclical boosting algorithm was implemented for a Gaussian, a negative binomial as well as a Weibull distributed response variable. In a simulation studies as well as for real world data sets the competitive performance of the new approach is shown.

References

1 Mayr, A., Fenske, N., Hofner, B., Kneib, T., and Schmid, M. (2012). Generalized additive models for location, scale and shape for high dimensional data - a flexible approach based on boosting. *Journal of the Royal Statistical Society: Series C (Applied Statistics).* 61(3), 403-427.

2 Thomas, J., Mayr, A., Bischl, B., Schmid, M., Smith, A., and Hofner, B. (2018). Gradient boosting for distributional regression: faster tuning and improved variable selection via noncyclical updates. *Statistics and Computing.* 28(3), 673-687.

3 Zhang, B., Hepp, T., Greven, S., and Bergherr, E. (2022). Adaptive step-length selection in gradient boosting for Gaussian location and scale models. *Computational Statistics.* 37(5), 2295-2332.

[1] Chair of Spatial Data Science and Statistical Learning, University of Goettingen
[2] Department of Medical Biometrics, Informatics and Epidemiology, University Hospital Bonn

alexandra.daub@uni-goettingen.de

Uncertainty Quantification in Machine Learning: From Aleatoric to Epistemic

Eyke Hüllermeier[1]

[1] Institut für Informatik, Artificial Intelligence and Machine Learning, LMU München

eyke@ifi-lmu.de

Tutorial: Interpretable Machine Learning

Bernd Bischl[1], *Fiona Ewald*[1]

[1] Institut für Statistik, LMU München

bernd.bischl@stat.uni-muenchen.de

Causal Fair Machine Learning

Ludwig Bothmann[1,2], Kristina Peters[3], Susanne Dandl[1,2], Michael Schomaker[1], Bernd Bischl[1,2]

A growing body of literature in fairness-aware ML aspires to mitigate machine learning (ML)-related unfairness in automated decision-making (ADM) by defining metrics that measure the fairness of an ML model and by proposing methods that ensure that trained ML models achieve low values in those metrics (see, e.g., Verma & Rubin, 2018, Caton & Haas, 2023). However, the underlying concept of fairness, i.e., the question of what fairness is, is rarely discussed, leaving a considerable gap between centuries of philosophical discussion and the recent adoption of the concept in the ML community.

We bridge this gap by formalizing a consistent concept of fairness and translating the philosophical considerations into a formal framework for training and evaluating ML models in ADM systems (Bothmann et al., 2024). We argue why and how causal considerations are necessary when assessing fairness in the presence of protected attributes (PAs) by proposing a fictitious, normatively desired (FiND) world where the PAs have no (direct or indirect) causal effect on the target. In practice, this unknown FiND world must be approximated by a warped world, for which the causal effects of the PAs must be removed from the real-world data. We propose rank-preserving interventional distributions to define an estimand of this FiND world and a warping method for estimation (Bothmann et al., 2023). Evaluation criteria for both the method and the resulting ML model are presented. Experiments on simulated data show that our method effectively identifies the most discriminated individuals and mitigates unfairness. Experiments on real-world data showcase the practical application of our method.

References

1 Bothmann, L., Dandl, S. & Schomaker, M. (2023) Causal Fair Machine Learning via Rank-Preserving Interventional Distributions. *In: Proceedings of the 1st Workshop on Fairness and Bias in AI co-located with 26th European Conference on Artificial Intelligence (ECAI 2023)*. CEUR Workshop Proceedings. URL: `https://ceur-ws.org/Vol-3523/paper1.pdf`

2 Bothmann, L., Peters, K. & Bischl, B. (2024) What Is Fairness? On the Role of Protected Attributes and Fictitious Worlds. *arXiv.2205.09622*. DOI: 10.48550/arXiv.2205.09622

3 Caton, S. & Haas, C. (2023) Fairness in Machine Learning: A Survey. *ACM Comput. Surv. 3616865* DOI: 10.1145/3616865

4 Verma, S. & Rubin, J. (2018) Fairness Definitions Explained. *In: Proceedings of the International Workshop on Software Fairness*, 1–7 (ACM, Gothenburg, Sweden). DOI: 10.1145/3194770.3194776

[1] Department of Statistics, LMU Munich, Ludwigstr. 33, 80539 Munich, Germany
[2] Munich Center for Machine Learning (MCML)
[3] Faculty of Law, LMU Munich, Ludwigstr. 29, 80539 Munich, Germany

`ludwig.bothmann@lmu.de, kristina.peters@jura.uni-muenchen.de,`
`susanne.dandl@stat.uni-muenchen.de, michael.schomaker@stat.uni-muenchen.de,`
`bernd.bischl@stat.uni-muenchen.de`

Estimand-aligned data splitting and performance estimation in applied machine learning

Rieke Alpers[1] and Max Westphal[1]

Data splitting is an important step in applied machine learning (ML), primarily to avoid overfitting and over-optimism. A wide variety of methods exists to randomly partition the available data for training, validation and testing purposes (e.g., holdout, cross-validation, bootstrap and variations thereof) [1, 2]. However, concrete guidance for practitioners which method should be preferred, depending on the ML problem at hand, is scarce.

We are connecting the choice of the data splitting method to the definition of an adequate estimand for the model or algorithm evaluation study [3]. In particular, our framework focuses on a specification of constraints (inclusion- or exclusion criteria) on (a) test observations, (b) training data and (c) the relation between the two. These constraints describe what kind of generalization or transferability (out-of-distribution) performance is actually the inference target in the evaluation study.

In this talk, we will focus on two aspects. Firstly, we showcase the recently developed R package {mldesign} which allows to derive a concrete data split corresponding to an estimand specification in the above sense [4]. We illustrate how the computational demand of this derivation can be significantly reduced by considering appropriate equivalence classes of observations (implied by the estimand-defining variables) instead of the raw data. Secondly, we propose a Bayesian model for estimation of and uncertainty quantification for the estimand-aligned generalization (transferability) performance. For this purpose, we extend existing individual participant data (IPD) meta-analysis approaches by adding a non-linear (fixed) effect for the training sample size to the model [5, 6]. In effect, the estimand turns out to be an entire learning curve and not only the expected (scalar) performance for a fixed training sample size. Our findings are illustrated by the numerical results of different hypothetical ML studies performed on the International Stroke Trial dataset [7].

References

1. Raschka S. Model evaluation, model selection, and algorithm selection in machine learning. arXiv [cs.LG]. 2018. Available from: http://arxiv.org/abs/1811.12808
2. Borra S, Di Ciaccio A. Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods. Comput Stat Data Anal. 2010;54(12):2976–89. Available from: https://www.sciencedirect.com/science/article/pii/S0167947310001064
3. Alpers R, Westphal, M. An estimand framework to guide model and algorithm evaluation in predictive modelling. Manuscript in preparation. 2024.
4. Westphal M. mldesign: Meaningful Data Splitting in Applied Machine Learning. R package. Available from: https://github.com/maxwestphal/mldesign
5. Riley, Richard D., et al. "External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges." *bmj* 353 (2016).
6. Viering, Tom, and Marco Loog. "The shape of learning curves: a review." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.6 (2022): 7799-7819.
7. Sandercock, Peter AG, et al. "The international stroke trial database." Trials 12.1 (2011): 101.

[1] Fraunhofer MEVIS, Institute for Digital Medicine, Max-von-Laue-Str. 2, 28359 Bremen, Germany

Contact:
max.westphal@mevis.fraunhofer.de

MissARF: Adversarial random forests for imputing missing values

Pegah Golchian[1,2],Jan Kapar[1,2],Kristin Blesch[1,2], David S. Watson[3] and Marvin N. Wright[1,2,4]

Applying machine learning to real datasets can be challenging because we often have to face the problem of missing values. However, when making statements about a population, it is essential to properly address this issue. A common approach is to fill these values with so-called imputation methods, which are categorized into single and multiple imputation [1]. In a statistical setting, multiple imputation is recommended since it accounts for uncertainty. It reaches its limits when the number of variables is large and the sample size is moderate, leading to computational costs and overparameterization [2]. Moreover, it has problems in tackling complex interactions and nonlinearity of variables [2]. Here, machine learning could lead to improvements. Machine learning imputation methods, such as MissForest [3], are convenient for returning a single complete dataset on the one hand but do not account for uncertainty on the other. Another promising approach is to address the missing value problem from a generative modeling perspective, as seen in GAIN [4].

We propose a fast and easy-to-use imputation method based on generative machine learning that offers single and multiple imputation. It is based on the method adversarial random forest (ARF) [5] for density estimation and data synthesis. ARF uses a recursive variant of unsupervised random forests [6], inspired by the idea of generative adversarial networks (GANs) [7]. We extend ARF for imputing by using conditional sampling on the non-missing values and call the method Miss-ARF.

For comparison with other imputation methods, we use simulation studies and real data, where we simulate the three missing data mechanisms missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) for a range of missingness proportions. We compare the imputed dataset with the real dataset by normalized root mean squared error (NRMSE) and proportion of falsely classified entries (PFC). Since we consider a generative modeling setting, we also compare the performance of machine learning models with the imputed dataset and the actual dataset by using machine learning efficacy/utility [8], which is a suitable measure for synthetic tabular data.

[1] Leibniz Institute for Prevention Research and Epidemiology – BIPS. Achterstraße 30, 28359 Bremen, Germany

[2] University of Bremen, Faculty of Mathematics and Computer Science. Post office box 330 440, 28334 Bremen, Germany

[3] King's College London, Faculty of Natural, Mathematical & Engineering Sciences. (N)5.16, Bush House, Strand campus, 30 Aldwych, London, WC2B 4BG, United Kingdom

[4] University of Copenhagen, Department of Public Health. Øster Farimagsgade 5, 1353 København, Denmark

golchian@leibniz-bips.de, kapar@leibniz-bips.de, blesch@leibniz-bips.de, david.watson@kcl.ac.uk, wright@leibniz-bips.de

References

[1] Van Buuren S. Flexible imputation of missing data. 2nd ed. CRC press; 2018.

[2] Tang F, Ishwaran H. Random forest missing data algorithms. Statistical Analysis and Data Mining: The ASA Data Science Journal. 2017;10(6):363-77.

[3] Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012;28(1):112-8.

[4] Yoon J, Jordon J, van der Schaar M. GAIN: Missing data imputation using generative adversarial nets. In: Proceedings of the 35th International Conference on Machine Learning. vol. 80. PMLR; 2018. p. 5689-98.

[5] Watson DS, Blesch K, Kapar J, Wright MN. Adversarial random forests for density estimation and generative modeling. In: Proceedings of The 26th International Conference on Artificial Intelligence and Statistics. vol. 206. PMLR; 2023. p. 5357-75.

[6] Shi T, Horvath S. Unsupervised learning with random forest predictors. Journal of Computational and Graphical Statistics. 2006;15(1):118-38.

[7] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Advances in Neural Information Processing Systems. vol. 27; 2014. p. 2672-80.

[8] Choi E, Biswal S, Malin B, Duke J, Stewart WF, Sun J. Generating multi-label discrete patient records using generative adversarial networks. In: Proceedings of the 2nd Machine Learning for Healthcare Conference. vol. 68. PMLR; 2017. p. 286-305.

bioXfuse: An Integrated Toolkit for In-Depth Exploration of Semantic and Contextual Landscapes in Biomedical Research

Marietta Hamberger[1], Nensi Ikonomi[1], Julian D. Schwab[1] and Hans A. Kestler[1]

Achieving a holistic perspective of biological systems often relies on in-depth analysis and the integration of existing knowledge. Traditionally, this involves sequentially reviewing documents to identify key elements. This method not only requires significant manual effort but also makes it challenging for researchers to detect subtle deviations in co-occurrence or changes in text semantics, especially when processing large amounts of literature.

To address these challenges, we introduce bioXfuse, a natural language processing (NLP)-based toolkit that employs a multi-layered analysis approach to uncover potential links between documents, words, and their contexts. The graphical user interface (GUI) is designed for a flexible exploration of the semantic landscapes and contextual motifs within documents, enabling a richer, more holistic perspective of the data.

BioXfuse utilizes advanced NLP techniques and deep learning models to identify biomedical entities and central phrases in large datasets. By visualizing co-occurrence networks, the tool highlights how terms are interconnected and co-mentioned within the literature. It further enriches this comprehensive map by clustering documents and entities based on context and metadata, enabling researchers to uncover patterns and correlations that might otherwise remain hidden. Additionally, incorporating word embeddings introduces a powerful semantic layer, simplifying the recognition of contextual similarities and possible functional parallels. This can be crucial for discovering nuanced relationships or roles within the data.

The toolkit supports researchers by automatically detecting key phrases and their textual connections within diverse contexts. In gene regulatory network studies, for instance, the identification of central players and interaction hubs can be streamlined, supporting the semi-automated construction of complex network models. Not limited to specific research fields, the toolkit's versatile functions can be applied broadly, such as in disease pathway analysis, drug discovery, or biomarker identification, to uncover new insights and highlight potential research gaps.

Despite the automation, users always retain veto power, reducing the black box issue and ensuring transparency and control. This integrated perspective can enhance the research process and provide deeper, more comprehensive insights.

[1] Ulm University, Institute of Medical Systems Biology

marietta.hamberger@uni-ulm.de, nensi.ikonomi@uni-ulm.de, julian.schwab@uni-ulm.de, hans.kestler@uni-ulm.de

In conclusion, bioXfuse supports literature synthesis and exploration by combining advanced NLP techniques with flexible visualization and analytical tools. This fusion encourages multi-layered semantic exploration and contextual understanding, fostering reproducibility and keeping researchers in the loop at each stage. bioXfuse equips scientists with a powerful resource for uncovering hidden connections within existing knowledge.

Human in the Loop – Uncertainty of Labelling in Supervised Machine Learning

Göran Kauermann

Image classification based on supervised machine learning is often prone to ambiguities in the labelled training data. The same applies to sentiment analysis in NLP. To generate suitable training data, images or texts are labelled according to evaluations of human experts. This can result in ambiguities, which will affect subsequent machine learning models.
We present two examples, where the human labelling is influenced by intrinsic uncertainty. The examples come from remote sensing (classification of satellite images) and computer linguistics (classification of text). We construct a multinomial mixture model given the evaluations of multiple experts. This is based on the assumption that there is a ground truth, which however remains unknown to the annotators. The model parameters can be estimated by a stochastic EM algorithm. Analyzing the estimates gives insights into sources of label uncertainty and allows to estimate the confusion matrix. The results can next be employed to train machine learning models, which is sketched in the talk.

German Data Science Society

Göran Kauermann

Die German Data Science Society (GDS e.V.) wurde 2018 mit dem Ziel gegründet, die führende Vereinigung von akademisch ausgebildeten Data Scientists zu sein. Dabei geht es um die Vernetzung von Data Scientists in Unternehmen und deren Verzahnung mit Forschungsinstituten und Ausbildungsstätten im Bereich Data Science. Die GDS ist als gemeinnützig anerkannt und verfolgt Ihr Ziel durch unterschiedliche Aktivitäten wie die German Data Science Days und spezialisierte Workshops. Die GDS wird dabei durch zahlreiche namenhafte Unternehmen unterstützt.

Robust estimation of distributional regression models using artificial neural networks

Tobias Weckop[1] and Tobias Hepp[1,2]

Distributional regression models using the generalized additive models for location, scale and shape (GAMLSS) framework [1] enable researchers to flexibly model multiple parameters of a target distribution in relation to a set of covariates. However, the accuracy and reliability of the estimates depend considerably on the distributional assumptions and may be strongly affected by the presence of outliers. Recent approaches to solve this problem rely on the use of a robustified log-likelihood to give less weight to data points which seem unlikely to fit the data generating process of interest [2, 3]. While the initial problems of a maximum-likelihood based procedure [2] have already been addressed via the use of gradient boosting [3], current approaches seem to be computationally intensive and therefore slow.

In this work, we suggest an alternative method of fitting a GAMLSS by combining distributional neural networks [4] with the robustified log-likelihood concept. Using a simulated dataset containing nonlinear relationships between the covariates and the parameters of the distribution of interest, our neural network was compared to both established methods in terms of speed and quality of model parameter estimation when outliers are present. Initial simulations indicate that our proposed neural network is significantly faster than both previous methods, without any notable loss of accuracy with respect to the estimated mean and standard deviation of the target distribution.

Literatur

1 R. A. Rigby, D. M. Stasinopoulos, Generalized Additive Models for Location, Scale and Shape, Journal of the Royal Statistical Society Series C: Applied Statistics, Volume 54, Issue 3, June 2005, Pages 507–554

2 Aeberhard, W.H., Cantoni, E., Marra, G. et al. Robust fitting for generalized additive models for location, scale and shape. Stat Comput 31, 11 (2021).

3 Speller, J., Staerk, C., Gude, F. et al. Robust gradient boosting for generalized additive models for location, scale and shape. Adv Data Anal Classif (2023).

4 Marcjasz, G., Narajewski, M., Weron, R., Ziel, F. Distributional neural networks for electricity price forecasting. Energy Economics 125 (2023).

[1] Institut für Medizininformatik, Biometrie und Epidemiologie, Friedrich-Alexander-Universität Erlangen-Nürnberg

[2] Professur für Raumbezogene Datenanalyse und Statistische Lernverfahren, Georg-August-Universität Göttingen

tobias.weckop@fau.de

Dimensionality and Distance: Curse or Blessing? Geometrical Aspects of Nearest Neighbor Computation in High-Dimensional Data

Moritz Herrmann[1,2]

When it comes to computation, it is often said that high-dimensional data is particularly challenging, known as the *curse of dimensionality*. For example, in their seminal work, Beyer et al [1] study the impact of high-dimensional data on nearest neighbor computation. They show that in a wide range of settings, including IID data, the difference between the distance to the nearest neighbor and the distance to the most distant neighbor vanishes as the dimension increases. However, it is arguably often overlooked that they also point out that this result does not hold in certain situations, in particular when the intrinsic dimension of the data is low and/or when the data is distributed in well separable subsets. More generally, it is probably less well known that high dimensionality can make computation easier, to the extent that Kainen [2] even speaks of a blessing of dimensionality. Given these different aspects, a natural question to ask is: when is high dimensionality a curse and when is it not (or even a blessing)? In this talk we approach this question from a geometric point of view. Focusing on the aspect of nearest neighbor (and hence distance) computation, we show that high-dimensional data need not be more challenging than low-dimensional data in many practically relevant situations. In particular, using results from extensive experiments on synthetic and real data, we show that this can be the case for both outlier detection and cluster analysis, and for a range of different data types, including image and functional data [3, 4]. Moreover, based on concepts from manifold learning and topological data analysis, we show that these observations can be explained using a common conceptual foundation.

[1] Institute for Medical Information Processing Biometry and Epidemiology, Faculty of Medicine, LMU Munich, Marchioninistr. 15, 81377 Munich, Germany
[2] Munich Center for Machine Learning (MCML)

moritz.herrmann@ibe.med.uni-muenchen.de

References

1 Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U. (1999). When Is "Nearest Neighbor" Meaningful?. In: Beeri, C., Buneman, P. (eds) Database Theory — ICDT'99. ICDT 1999. Lecture Notes in Computer Science, vol 1540. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-49257-7_15

2 Kainen, P.C. (1997). Utilizing Geometric Anomalies of High Dimension: When Complexity Makes Computation Easier. In: Kárný, M., Warwick, K. (eds) Computer Intensive Methods in Control and Signal Processing. Birkhäuser, Boston, MA. https://doi.org/10.1007/978-1-4612-1996-5_18

3 Herrmann, M., Pfisterer, F., Scheipl, F. (2023). A geometric framework for outlier detection in high-dimensional data. WIREs Data Mining and Knowledge Discovery, 13(3), e1491. https://doi.org/10.1002/widm.1491

4 Herrmann, M., Kazempour, D., Scheipl, F., Kröger, P. Enhancing cluster analysis via topological manifold learning. Data Mining and Knowledge Discovery (2023). https://doi.org/10.1007/s10618-023-00980-2

Registration of Dynamic ^{1}H-MRI Series for Fourier-based Lung Functional Analysis

Paul Kaftan[1,2], Volker Rasche[2], and Hans Kestler[1]

Patients with chronic lung diseases such as interstitial lung disease (ILD) are regularly examined to monitor the disease progression. While *global* functional parameters like lung volume can be measured non-invasively through spirometry, *regional* ventilation and perfusion are usually examined using CT imaging [1]. Repeated CT scans for disease monitoring subject the patients to high doses of ionizing radiation [2]. Furthermore, the scans are performed in a breath-hold manner, which is very challenging for individuals with chronic lung conditions [3]. In this project, we investigate into an alternative method for regional lung functional analysis based on free-breathing, non contrast-enhanced ^{1}H-MRI sequences [4,5]. In a real-time MRI sequence, the signal intensity in the parenchyma changes during the respiratory and cardiac cycles, which can be extracted using Fourier analysis [5,6]. Before this is possible, registration over time is used to compensate for the respiratory and cardiac motion. Accurately matching the anatomically corresponding points in the image sequence is crucial for a meaningful quantification of lung function.

In this work we will assess the applicability of different paradigms to perform this registration task. The task is especially challenging due to low contrast in the lung parenchyma compared to surrounding anatomies. Registration algorithms include variational methods as well as deep learning-based approaches.

[1] Ulm University, Institute of Medical Systems Biology
[2] Ulm University, MoMAN Center for Translational Imaging

paul.kaftan@uni-ulm.de, volker.rasche@uni-ulm.de, hans.kestler@uni-ulm.de

References

1 Lilian Lonzetti, Matheus Zanon, Gabriel Sartori Pacini, Stephan Altmayer, Diogo Martins de Oliveira, Adalberto Sperb Rubin, Fernando Ferreira Gazzoni, Marcelo Cardoso Barros, and Bruno Hochhegger. Magnetic resonance imaging of interstitial lung diseases: A state-of-the-art review. *Respiratory Medicine*, 155:79–85, August 2019. ISSN 0954-6111. doi: 10.1016/j.rmed.2019.07.006.

2 John R. Mayo. CT Evaluation of Diffuse Infiltrative Lung Disease: Dose Considerations and Optimal Technique. *Journal of Thoracic Imaging*, 24(4):252, November 2009. ISSN 0883-5993. doi: 10.1097/RTI.0b013e3181c227b2.

3 Burton Marks, Donald G. Mitchell, and John P. Simelaro. Breath-holding in healthy and pulmonary-compromised populations: Effects of hyperventilation and oxygen inspiration. *Journal of Magnetic Resonance Imaging*, 7(3):595–597, 1997. ISSN 1522-2586. doi: 10.1002/jmri.1880070323.

4 Maren Zapke, Hans-Georg Topf, Martin Zenker, Rainer Kuth, Michael Deimling, Peter Kreisler, Manfred Rauh, Christophe Chefd'hotel, Bernhard Geiger, and Thomas Rupprecht. Magnetic resonance lung function – a breakthrough for lung imaging and functional assessment? A phantom study and clinical trial. *Respiratory Research*, 7(1):106, August 2006. ISSN 1465-993X. doi: 10.1186/1465-9921-7-106.

5 Grzegorz Bauman, Michael Puderbach, Michael Deimling, Vladimir Jellus, Christophe Chefd'hotel, Julien Dinkel, Christian Hintze, Hans-Ulrich Kauczor, and Lothar R. Schad. Non-contrast-enhanced perfusion and ventilation assessment of the human lung by means of fourier decomposition in proton MRI. *Magnetic Resonance in Medicine*, 62(3):656–664, 2009. ISSN 1522-2594. doi: 10.1002/mrm.22031.

6 Åsmund Kjørstad, Dominique M. R. Corteville, Thomas Henzler, Gerald Schmid-Bindert, Erlend Hodneland, Frank G. Zöllner, and Lothar R. Schad. Quantitative lung ventilation using Fourier decomposition MRI; comparison and initial study. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 27 (6):467–476, December 2014. ISSN 1352-8661. doi: 10.1007/s10334-014-0432-9

Strategies for Reducing Heterogeneity in Clinical Trials through optimized group assignment

M Oruc[1], JM Kraus[1], HA Kestler[1]

In existing clinical trials, heterogeneity in certain variables, such as age, is a challenge. Randomised controlled trials (RCTs) are the gold standard, but are often limited in practice for ethical and cost reasons [1]. The aim of this research in observational studies is to provide optimal group assignement, identifying a near-optimal subset of both groups with maximal samples in which certain variables do not differ significantly.

Brute force methods for determining the optimal groups are time and resource intensive due to their exponential complexity ($O(2^n)$), where 'n' represents the number of samples. This paper examines alternative strategies that do not require searching the entire population: a genetic algorithm (GA) that evolves optimal sample combinations through selection, and propensity score matching (PSM) with a greedy approach, which groups participants based on confounding variables, as shown in Figure 1.

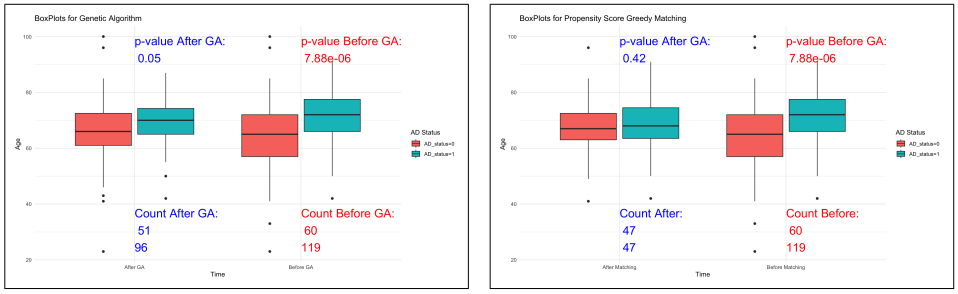


Figure 1: Comparative boxplots illustrating the impact of the Genetic Algorithm (GA) (left) and Propensity Score Matching (PSM) with the greedy approach (right) on age distribution and sample size homogeneity in clinical trial groups. Significant differences are observed before matching (p = 7.88e-06). P-values from the Wilcoxon signed-rank test) after matching indicate improved comparability: 0.05 after GA (unbalanced data) and 0.42 after PSM with the greedy approach (balanced data).

Our insights confirm the suitability of these methods for forming comparable study groups for clinical trials. Future research will extend these methods to consider multiple confounding variables.

[1] Ulm University, Insitute of Medical Systems Biology

metehan.oruc@uni-ulm.de, johann.kraus@uni-ulm.de, hans.kestler@uni-ulm.de

References

1 Zhao, Qin-Yu, Jing-Chao Luo, Ying Su, Yi-Jie Zhang, Guo-Wei Tu, and Zhe Luo. (2021). *Propensity score matching with R: conventional methods and new features. Annals of Translational Medicine*, Volume 9.

2 M. Watabe-Rudolph, Z. Song, L. Lausser, C. Schnack, Y. Begus-Nahrmann, M.-O. Scheithauer, G. Rettinger, M. Otto, H. Tumani, D.R. Thal, J. Attems, K.A. Jellinger, H.A. Kestler, C.A.F. von Arnim, K.L. Rudolph. (2012). *Chitinase enzyme activity in CSF is a powerful biomarker of Alzheimer disease. Neurology*, 78(8): 569-577.

3 Lori S. Parsons. (2001). *Reducing Bias in a Propensity Score Matched-Pair Sample Using Greedy Matching Techniques*, Proceedings of the Twenty-Sixth Annual SAS® Users Group International Conference, Long Beach, California, 214-26.

4 André Kratzer, Linda Karrer, Nikolas Dietzel, Franziska Wolff, Manuela Hess, Peter Kolominsky-Rabas, Elmar Gräßel. (2020). *Symptombelastung, Inanspruchnahme des Gesundheitssystems und Todesumstände von Menschen mit Demenz in der letzten Lebensphase: der Bayerische Demenz Survey (BayDem)*, *Das Gesundheitswesen*, 82(01): 30-39.

5 Kuss, O., Blettner, M., Boergermann, J. (2016). *Propensity Score - eine alternative Methode zur Analyse von Therapieeffekten. Dtsch Arztebl*, 597-603, 113(35-36): 0597.

6 Maekawa, M., Tanaka, A., Ogawa, M., Roehrl, M.H. (2024). *Propensity score matching as an effective strategy for biomarker cohort design and omics data analysis*. PLOS ONE, 19(5): e0302109. Editor: Raffaele Serra, University Magna Graecia of Catanzaro, ITALY.

How to impute if you must: Selecting the Appropriate Missing Value Imputation Strategy for Cross-Sectional Biomedical Numerical Data

Jörn Lötsch[1] and Alfred Ultsch[3]

j.loetsch@em.uni-frankfurt.de, ultsch@Mathematik.Uni-Marburg.de

Missing value imputation is a common data preprocessing task in biomedical research, but the choice of a suitable imputation method is often arbitrary. This report proposes a method for rational imputation model selection, focusing on cross-sectional numerical tabular data. Starting from the premise that an optimal imputation method should restore true values without bias and with minimal error, a combined metric was applied to evaluate various univariate and multivariate imputation models, including poisoned (biased) and calibrating (predefined imprecision) algorithms. Diagnostic missing values were used to rank imputation models based on deviation from true values, evaluated on four biomedical and three artificial datasets. When diagnostic missing values were used for model ranking, a set of "A" grade imputation methods appeared as the best strategy, generalizable to truly missing values. Instances were highlighted where multivariate methods offered no advantage or when datasets could not be imputed satisfactorily. An algorithm is proposed that provides a ranking of models, estimates imputation accuracy, evaluates multivariate imputation, and signals when available methods are unsatisfactory. The data and R code are available in the "opImputation" package (CRAN upload pending).

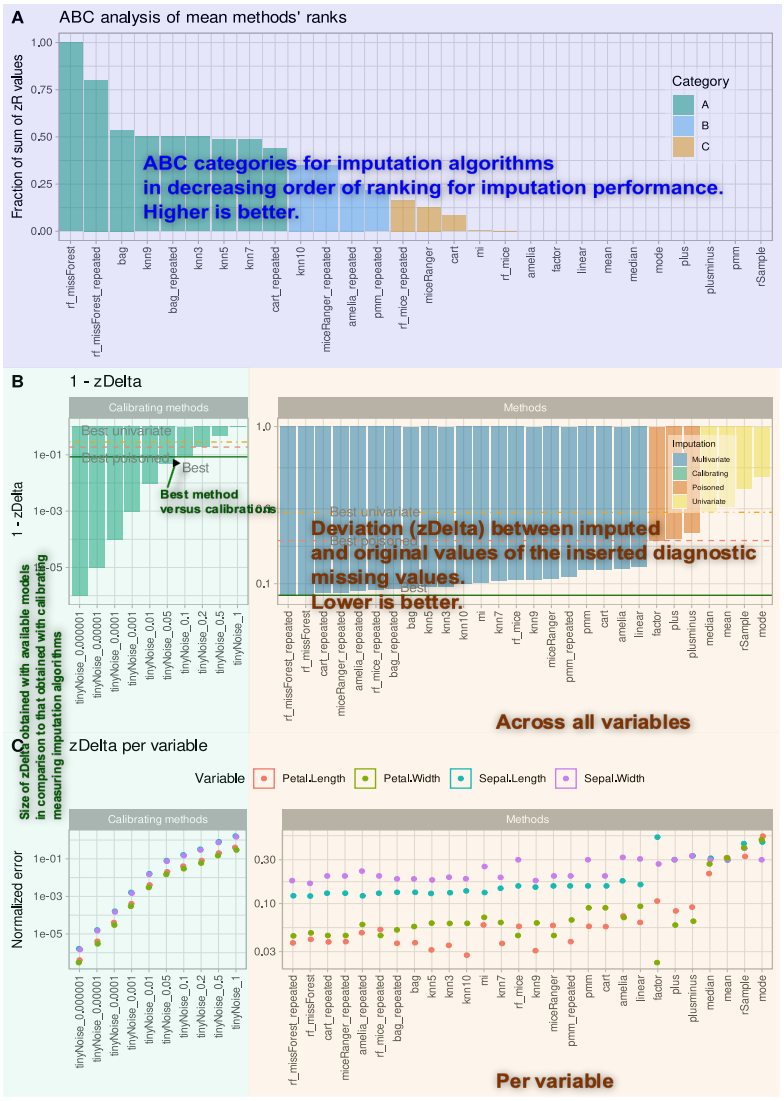


Figure 1: Annotated summary of results from comparative analyses of expected imputation results from different multivariate, univariate, and diagnostic algorithms

[1] Institute of Clinical Pharmacology, Goethe - University, Theodor Stern Kai 7, 60590 Frankfurt am Main, Germany

[2] DataBionics Research Group, University of Marburg, Hans - Meerwein - Straße, 35032 Marburg, Germany

Regularization methods in clinical biostatistics: State-of-the art and possibilities for improvement

Sarah Friedrich[1]

[1] Mathematical Statistics and Artificial Intelligence in Medicine, Augsburg

sarah.friedrich@math.uni-augsburg.de

From Data to Discovery: DataCite's Role in FAIR Data Management

Sara El-Gebali[1]

[1] DataCite, Schweden

sara.elgebali@gmail.com

Reproducibility at the Biometrical Journal: Simple tips for writing and publishing clear code to ensure reproducible results

Roman Hornung[1]

[1] Biometry in Molecular Medicine, LMU München

roman.hornung@ibe.med.uni-muenchen.de

Addressing Missing Data in Clinical Metadata involving Patients with Alzheimer's Disease

Charles Kaumbutha[1,2],Shubhi Ambast[1,2], Karola Mai[1,2],Marie Oestreich[1,2],Karoline Mauer[2,3] and Matthias Becker[1,2]

With the growing significance in studying complex diseases in biomedical research, data quality has been a key challenge owing to the complex data pipelines applied to fully understand the diseases. The large number of demographic and clinical variables collected in clinical research is plagued with inevitable missing values, and inadequate handling can lead to biased estimates and decreased statistical power. This has led to a hindrance in machine learning (ML) applications as high quality data is crucial in ensuring robust predictive performance and responsible usage in decision making. We therefore aim to examine extant evaluation metrics, to in turn find the optimal imputation approaches for the clinical data.

The incompleteness of clinical metadata particularly in Alzheimer's disease is mainly driven by the age of the participants and nature of the disease. These participants have a high risk of comorbidities and the demanding nature of the examinations poses a huge burden not only to the AD patients, but also the relatives accompanying them, who also undergo proxy evaluations[1]. This results in loss of motivation which in turn influences the refusal or withdrawal from the studies.

While multiple imputation (MI) is widely recognized by statisticians and ML experts as a valuable strategy for managing missing data, there remains uncertainty regarding which methods are most effective and how their performance should be quantified[2]. Inadequate imputation can have far-reaching consequences, as downstream ML applications, such as classification tasks, may suffer from poor feature importance allocation when trained on suboptimally imputed datasets. To mitigate these risks, a variety of evaluation metrics have been proposed to gauge the performance of imputation models. One promising approach is the sliced Wasserstein distance metric, which evaluates the authenticity of data reconstruction across the entire feature value distribution and has shown potential as a superior alternative to both sample-wise and feature-wise metrics. In this context, we conduct comparative analysis of these evaluation metrics on popular imputation techniques, and how well they capture the distributional discrepancy. Imputation is an imperative step in the data processing within the Prisyn project, where delivery of high quality data will ensure reliability of generators and classification models, ultimately enhancing the prediction of Alzheimer's disease.

[1] Modular High Performance Computing and Artificial Intelligence, German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany
[2] Systems Medicine, German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany
[3] PRECISE Platform for Genomics and Epigenomics at DZNE and University of Bonn, Bonn, Germany

CharlesMwangi.Kaumbutha@dzne.de

References

[1] Shadbahr, T., Roberts, M., Stanczuk, J. et al. The impact of imputation quality on machine learning classifiers for datasets with missing values. Commun Med 3, 139 (2023). https://doi.org/10.1038/s43856-023-00356-z

[2] van Buuren, S. Flexible Imputation of Missing Data, 2nd edn. (CRC Press, 2018)

TSCPDetector: A Comprehensive Approach to Change Point Detection in Time Series Models

Mehdi Lotfi[1], *Lars Kaderali*[1]

Change point detection refers to identifying specific times when features of a time series dataset change abruptly, often due to factors such as policy adjustments. For instance, during the COVID-19 pandemic, interventions like lockdowns and vaccinations can trigger sudden shifts in infection or hospitalization rates, leading to alterations in the COVID-19 dataset. In epidemiological modeling utilizing Ordinary Differential Equations (ODEs), detecting change points entails recognizing shifts in model parameters, indicating instances where the existing model fails to accurately simulate the dataset and necessitates re-estimation for the new circumstances.

Our method offers a unique advantage in handling scenarios where certain characteristics of a dataset change at specific points while others remain constant. For example, in COVID-19 simulations, infection and hospitalization rates may vary at certain time points, while recovery rates or hospital stay durations remain consistent throughout the dataset. Our approach introduces a novel perspective by incorporating a cost function to assess the goodness of fit between the time series dataset and the underlying model. This emphasizes the importance of evaluating model fit rather than relying solely on statistical distributions, which are commonly used in change point detection problems but may not guarantee accurate identification of the number and positions of change points. The method comprises two main modules. The first module employs binary segmentation to investigate potential change points, while the second module utilizes a modified genetic algorithm to determine the optimal combination of the model's parameters for each segment caused by change points using the predefined objective function.

It is important to note that our developed method is flexible and can be applied to any mathematical modeling. In this poster presentation, we introduce the developed method and apply it to simulated datasets and COVID data simulations

---

[1] Institute for Bioinformatics, University Medicine Greifswald, Greifswald, Germany

mehdi.lotfi@uni-greifswald.de, lars.kaderali@uni-greifswald.de

Extending Gradient Boosting Frameworks for High-Dimensional MEG Data Analysis in Neurophysiological Research

Anna von Plessen[1], Nadia Müller-Voggel[2], Elisabeth Bergherr[1]

Understanding brain function through Magnetoencephalographic (MEG) recordings involves managing high-dimensional data, as neural oscillatory activity is measured across multiple frequency bands and at numerous brain locations. Traditional methods often fall short when dealing with the complexity and dimensionality of such data, leading to overfitting or limited explanatory power. To address these challenges, we propose extending gradient boosting frameworks [1] to MEG data that records oscillatory activity during an on-off-experiment with the aim of identifying key brain regions involved in conscious auditory perception [2]. Gradient boosting is adept at variable selection and effect shrinkage, thereby mitigating overfitting while enhancing model interpretability. As a first step, we will develop a base-learner that integrates three-dimensional spatial effects [3] of MEG measurement locations, crucial for accurately modelling the intricate anatomy of the brain. To construct this base-learner, we will leverage methodology that employs non-parametric regression models with differential regularisation [4]. This method allows us to capture the complex spatial dependencies in the brain, providing a more precise and nuanced understanding of the neural mechanisms underlying conscious perception.

References

1 Bühlmann, P., & Hothorn, T. (2007). Boosting Algorithms: Regularization, Prediction and Model Fitting. Statistical Science, 22(4), 477-505.

2 Schmitt, A., Kim, C., Rampp, S., Buchfelder, M., & Müller-Voggel, N. (2023). Neurophysiological correlates of Somatosensory tinnitus modulation. bioRxiv, 2023-06. https://doi.org/10.1101/2023.06.28.546718

3 Kneib, T., Hothorn, T., & Tutz, G. (2009). Variable Selection and Model Choice in Geoadditive Regression Models. Biometrics, 65(2), 626-634.

4 Arnone, E., Negri, L., Panzica, F., & Sangalli, L. M. (2023). Analyzing data in complicated 3D domains: Smoothing, semiparametric regression, and functional principal component analysis. Biometrics, 79(4), 3510-3521.

[1] Chair of Spatial Data Science and Statistical Learning, Georg-August-University Göttingen
[2] Center for Biomagnetism, Department of Neurosurgery, University Hospital Erlangen

anna.plessen@uni-goettingen.de

Distributional Regression for Lungfunction of Cystic Fibrosis Patients with a Special Focus on Center Specific Effects

Marisa Lange [1], Colin Griesbach[1], Elisabeth Bergherr [1]

Rapid lung function decline is a severe problem for cystic fibrosis patients throughout their lives. We have access to data from the German Cystic Fibrosis Registry [1], which includes thousands of patients and hundreds of thousands of observations collected repeatedly each year, covering hundreds of variables such as sociodemographic information, biomarkers, and gene expression data.
We plan to develop a prediction model for lung volume measured by the %FEV1 value, a key indicator of healthy lung function [2]. Since both the expected volume and its variation are critically important for patients, we will use a Gaussian distributional regression model [3]. Given the vast number of potential explanatory variables, a robust selection algorithm is necessary, and gradient boosting [4] is particularly well-suited for this task. We will incorporate information about the treatment centers where individual patients are treated. Two approaches will be compared: first, by introducing the center variable as a random effect in the model, and second, by performing a spatial triangulation to include the center as a discrete spatial effect in the model.

References

1 Deutsches Mukoviszidose-Register. (2022). *Berichtsband 2022*. Mukoviszidose e.V., Bundesverband Cystische Fibrose (CF).

2 Taylor-Robinson, D., Whitehead, M., Diderichsen, F., Olesen, H. V., Pressler, T., Smyth, R. L., Diggle, P. (2012). Understanding the natural progression in % FEV1 decline in patients with cystic fibrosis: a longitudinal study. Thorax, 67(10), 860-866.

3 Stasinopoulos, D. M., Rigby, R. A. (2008). Generalized additive models for location scale and shape (GAMLSS) in R. Journal of Statistical Software, 23, 1-46.

4 Mayr, A., Hofner, B. (2018). Boosting for statistical modelling-A non-technical introduction. Statistical Modelling, 18(3-4), 365-384.

[1] Chair of Spatial Data Science and Statistical Learning, Georg-August-Universitat Gottingen

marisa.lange@uni-goettingen.de

Determining the urgency of surgery of retinal detachment based on deep learning

Denna S. Langhans[1], Efstathios Vounotrypidis[1], Hans-Jürgen Buchwald[1], Melih Parlak[1], Kevin Koray Bayhan, Armin Wolf[1], Carolin Elhardt[1*], Christian M. Wertheimer[1*]

[1]Department of Ophthalmology, University Hospital Ulm, Prittwitzstraße 43, 89075 Ulm

* Both last authors contributed equally to this work

Purpose

The retina is a thin layer of light-sensitive tissue at the back of the eye that acts like the image sensor in a camera, capturing light and sending visual information to the brain. Retinal detachment, a rare but serious eye condition, occurs when this layer separates from its supporting tissue. Treatment is surgical, and timing has been identified as a critical factor in postoperative outcome. (1) The longer the retina remains detached, the greater the likelihood of irreversible damage to the retinal cells. In practice, however, environmental factors and limited resources must also be taken into account. Therefore, the consensus is to classify retinal detachment according to whether the central region of the retina is affected, which has a significantly worse prognosis. The aim of this study was to develop a neural network capable of detecting the status of the central retina. Due to the difficulties of deep learning approaches with limited data in rare diseases, a special network design was used.

Methods:

A single-center, ethics committee-approved, retrospective study included 475 eyes before surgical treatment of retinal detachment. Eyes were carefully labeled by determining the central retinal status using preoperative optical coherence tomography imaging (cross-sectional images providing detailed insights into the layers of the retina and bordering structures), with 229 still attached and 246 already detached central regions. 90-degree retinal images and clinical risk factors were presented to customised neural networks, depending on the data type. The clinical factors included age, biological sex, side, lens status and best corrected visual acuity. The data were divided into training, validation and test sets. The performance of the network was measured using the hold-out test data set. The influence of each data input on the overall performance was evaluated using class activation maps, ablation studies, receiver operating characteristic and threshold analysis.

Results:

If images alone were used for training, an accuracy of 0.76, a sensitivity of 0.52, and a specificity of 0.95 were determined. The class activation maps demonstrated the ability to detect the spatial region or the edges of the retinal detachment. When combined with clinical data, the accuracy improved to 0.92. Clinical data alone reached an accuracy of 0.93 with a sensitivity of 0.94 and a specificity of 0.93. The ablation study demonstrated that visual acuity is the primary factor in class determination. A threshold for visual acuity alone, without the use of deep learning to determine, was found to be within a range of 0.55 to 0.85 logMAR visual acuity.

Conclusions:

It should be emphasized that only small convolutional networks such as GoogLeNet combined with a small multilayer perceptron and a careful factorial analysis of input data for dimension reduction could reliably recognize relationships in the given data. The neural network then reached a very high accuracy, sensitivity and specificity on a held-out dataset. Clinical applicability needs to be evaluated in future studies.

References

1 Angermann R, Bechrakis NE, Rauchegger T, Casazza M, Nowosielski Y, Zehetner C. Effect of Timing on Visual Outcomes in Fovea-Involving Retinal Detachments Verified by SD-OCT. J Ophthalmol. 2020;2020:2307935.

2 Frings A, Markau N, Katz T, Stemplewitz B, Skevas C, Druchkiv V, et al. Visual recovery after retinal detachment with macula-off: is surgery within the first 72 h better than after? The British journal of ophthalmology. 2016;100(11):1466-9.

3 Sothivannan A, Eshtiaghi A, Dhoot AS, Popovic MM, Garg SJ, Kertes PJ, et al. Impact of the Time to Surgery on Visual Outcomes for Rhegmatogenous Retinal Detachment Repair: A Meta-Analysis. American journal of ophthalmology. 2022;244:19-29.

Correspondence to:

Denna Sabrina Langhans
Department of Ophthalmology, Ulm University
Prittwitzstraße 43
89075, Ulm, Germany
Phone: +49 731-500-59088
E-Mail: denna.langhans@uniklinik-ulm.de

Accounting for complex structures of aleatoric and epistemic uncertainty through problem-tailored MCMC algorithms

Sabine Hoffmann[1]

When analyzing data in biomedical research, researchers are often faced with various sources of uncertainty ranging from measurement error and uncertain selection mechanisms to uncertain decisions in the definition of exposure and outcome variables and the treatment of outliers and missing values. Lacking suitable tools to account for these complex sources of uncertainty, it is common to make the problem fit the tools by ignoring uncertainty to subsequently discuss the resulting biases. The Bayesian hierarchical framework provides a coherent and flexible framework to account for complex sources of uncertainty by combining different sub-models through conditional independence assumptions. Owing to the modular nature of the Bayesian hierarchical framework, it is possible to build highly flexible models. However, in realistic settings, the considered uncertainty structures are often too complex to be able to conduct statistical inference using standard implementations. This talk will present problem-tailored Markov Chain Monte Carlo solutions developed to conduct statistical inference for Bayesian hierarchical models describing complex structures of uncertainty that would be very time-consuming or even impossible to fit with current implementations of Bayesian inference with applications in occupational epidemiology and infectious disease modeling. Moreover, it will discuss how these algorithms can be extended to address the problem of apples and oranges in evidence synthesis and to account for the analytical variability arising from the multiplicity of possible analysis stategies that is often observed in multianalyst studies were multiple researchers are asked to answer the same research question using the same data set.

---

[1] The Institute for Medical Information Processing, Biometry, and Epidemiology, Ludwig-Maximilians-Universität München

sabine.hoffmann@stat.uni-muenchen.de

Comparison of Design of Experiments and Gaussian Process optimization for the optimization of a simulated bioprocess

Matthias Medl[1], Theresa Scharl[1], Bernhard Spangl[1], Friedrich Leisch[1] and Johannes Buyel[2]

Operating biopharmaceutical manufacturing processes at optimal conditions is critical to maximize productivity and product quality, while minimizing manufacturing cost and environmental impact. Therefore, employing efficient and effective optimization strategies is essential. These span from simple ones like one-factor-at-a-time, which assess the influence of process variables in isolation, to more advanced ones such as Design of Experiments (DoE) or Gaussian Process-based (GP) optimization, which also account for variable interactions.

In practice, a lot of decision making is involved to parameterize the optimization strategies themself. For instance, when choosing GP optimization, besides selecting the bioprocess parameters and their respective ranges, one additionally has to decide on a) the kernel of the GP itself, b) the acquisition function, c) hyperparameters concerning the acquisition function, d) the approach for sampling initial bioprocess parameter settings, e) protocols for managing stalled optimizations, and f) criteria for terminating the optimization. Configuring these parameters correctly is a challenge and continues to be a focus of ongoing research.

For DoE optimization, decision making involves selecting a) the general design, b) the number of center points, c) additional hyperparameters such as the distance between center points and star points of central composite designs, and d) the strategy for selecting the final response surface model.

Furthermore, the determination of the best parameter settings and the most effective strategy is substantially influenced by often unknown bioprocess parameters. These include factors, such as process and measurement noise, the location of the optimum within the design space, and the geometry of the response surface.

The aim of this study is to compare the performance and robustness of multiple optimization strategies across a diverse range of parameter and process configurations of a simulated ion exchange chromatography process. The final outcome of this empirical study is to provide guidance navigating the large decision space encountered when performing bioprocess optimization.

[1] University of Natural Resources and Life Sciences Vienna, Institute of Statistics, Peter-Jordan-Straße 82/I, 1190 Vienna, Austria
[2] University of Natural Resources and Life Sciences Vienna, Institute of Bioprocess Science and Engineering, Muthgasse 18, 1190 Vienna, Austria

matthias.medl@boku.ac.at, theresa.scharl@boku.ac.at, bernhard.spangl@boku.ac.at, friedrich.leisch@boku.ac.at, johannes.buyel@boku.ac.at

Component-wise gradient boosting for mixtures of distributional regression models

Tobias Hepp[1,2], Hannah Klinkhammer[3], Andreas Mayr[3], Elisabeth Bergherr[2], Colin Griesbach[2]

Applying statistical models usually requires the assumption of a probability density function to describe the (conditional) distribution of the variable of interest. However, in the case of unobserved heterogeneity, e.g. if the data consists of two or more unlabeled subpopulations, a single density function may not be sufficient to describe the distribution of the outcome. Then, given the number of latent components, a weighted sum of probability density functions can be used to construct a finite mixture distribution to account for unobserved heterogeneity. Mixture regression models [1] extend the basic mixture model by allowing one or more of the distribution parameters to be functions of the observed covariate vector.

Component-wise gradient boosting algorithms [2,3] are iterative updating schemes in which the gradient of the loss-function in the current step is fitted separately to multiple base-leaners Selecting only the best-performing learner in each iteration allows for regularization and variable selection during model fitting.

In this work, we present a gradient boosting algorithm for the estimation of finite mixture regression models. The proposed algorithm is evaluated in a simulation study and demonstrated on a dataset for the prediction of LDL-cholesterol.

References

1 DeSarbo, W.S. and Cron, W.L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of classification*, **5**(2), 249 – 282.

2 Bühlmann, P. and Yu, B. (2003) Boosting with the L2 loss: regression and classification. *Journal of the American Statistical Association*, **98**, 324 – 339.

3 Mayr, A., Binder, H., Gefeller, O. and Schmid, M. (2014). The evolution of boosting algorithms—From machine learning to statistical modelling. *Methods of Information in Medicine*, **53**(6), 419 – 427.

[1] Institut für Medizininformatik, Biometrie und Epidemiologie, Friedrich-Alexander-Universität Erlangen-Nürnberg

[2] Professur für Raumbezogene Datenanalyse und Statistische Lernverfahren, Georg-August-Universität Göttingen

[3] Institut für Medizinische Biometrie, Informatik und Epidemiologie, Rheinische Friedrich-Wilhelms-Universität Bonn

tbs.hepp@fau.de

On the handling of method failure in comparison studies

Milena Wünsch[1,2], Elisa Noltenius[3], Mattia Mohr[3], and Anne-Laure Boulesteix[1,2]

Neutral comparison studies aim to compare methods in an evidence-based manner, offering guidance to data analysts in selecting a suitable method for their application. To be reliable, they must be carefully designed, implemented, and reported, especially given the high degree of flexibility. A common challenge in comparison studies is to handle the failure of one or more methods in producing a result for some (real or simulated) data sets, such that their performances cannot be measured in those instances. Despite an increasing emphasis on this topic in recent literature (focusing on non-convergence as a common source of failure), guidance on proper handling is still scarce and transparent reporting of the chosen approach is often neglected [1, 2]. A common approach is thus to (silently) discard the corresponding data sets, similarly to complete or available case analysis in the "regular" missing data context. It has been argued, however, that this might lead to biased results and misleading method recommendations [1]. Our work thus aims to provide concrete guidance for handling method failure. First, we review commonly applied (and reported) approaches in peer-reviewed comparison studies using simulation studies and real data-based benchmarking in different contexts, including regression modelling, statistical testing, and machine learning as well as low- and high-dimensional data. Then, we present two exemplary comparison studies, from machine learning and classical statistics, in which method failure occurs for different reasons. Based on these examples, we investigate how applying available and complete case analysis affects the results of the comparison studies. Finally, we demonstrate that proper handling of missing performance values often requires closely investigating the underlying modelling processes and implementations, especially when method failure is not caused by non-convergence, and that it ultimately depends on the study's goal.

References

1 Morris, T. P., White, I. R., and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102.

2 Pawel, S., Kook, L., and Reeve, K. (2024). Pitfalls and potentials in simulation studies: Questionable research practices in comparative simulation studies allow for spurious claims of superiority of any method. *Biometrical Journal*, 66(1):2200091.

[1] Institute for Medical Information Processing, Biometry, and Epidemiology, Faculty of Medicine, LMU Munich, Marchioninistr. 15., 81377 Munich
[2] Munich Center for Machine Learning, Munich
[3] Department of Statistics, LMU Munich, Ludwigstr. 33, 80539 Munich

milena.wuensch@ibe.med.uni-muenchen.de, elisa.noltenius@campus.lmu.de,
mattia.mohr@campus.lmu.de, boulesteix@ibe.med.uni-muenchen.de

Optimal Scaling of an Algorithmic Parameter in Restart Strategies

Lisa Schönenberger and Hans-Georg Beyer

Optimization algorithms are often confronted with various challenges, such as becoming trapped in a local optimum or very long runtimes. Restart strategies have been demonstrated to be an effective method for overcoming these obstacles. They can significantly enhance the performance and robustness of optimization algorithms.

In restart strategies, the current search process is regularly stopped, and the optimization algorithm is restarted. Often a different starting point is chosen, or an algorithmic parameter is changed. The specific implementation of restart strategies can vary considerably, ranging from simple random restarts to more sophisticated techniques that adapt to the particular algorithm and problem characteristics. In general, it is not clear what the optimal implementation of a restart strategy is. Obviously, this question cannot be answered in a general way for each type of optimization algorithm. Therefore, assumptions and constraints must be made. This optimal implementation of a restart strategy has already been studied for Las Vegas algorithms, those that are always successful, but whose running time is a random variable. In [1] it was possible to derive an optimal restart strategy under the assumption of certain knowledge about the algorithm.

The following investigations are limited to optimization algorithms whose success depends on a certain parameter λ. This parameter can be, for example, the population size in evolutionary strategies and particle swarm optimization, or a predefined runtime. It is assumed that this parameter is changed after each restart, so that the restart strategy can be defined as $\mathcal{R} = (\lambda_0, \lambda_1, \lambda_2, \ldots)$. Often used for such problems is $\mathcal{R} = \left(\lambda_0, \lambda_0\rho, \lambda_0\rho^2, \lambda_0\rho^3, \ldots\right)$ where $\rho > 1$. In [2] this was applied to evolution strategies where the population size was increased after each restart with the above scaling rule. It was mentioned that experiments indicate that the optimal value for the increasing constant ρ lies between 2 and 3.

The goal of this investigation was to find the optimal choice for the increasing constant ρ. For this purpose, the loss function was introduced, which indicates how much computation time was wasted compared to the optimal strategy. An upper and lower bound for the loss function was found and it was possible to minimize the upper bound of the relative loss and to determine the optimal increasing constant.

References

[1] M. Luby, A. Sinclair, and D. Zuckerman. "Optimal speedup of Las Vegas algorithms". In: *[1993] The 2nd Israel Symposium on Theory and Computing Systems*. 1993, pp. 128–133. DOI: 10.1109/ISTCS.1993.253477.

[2] A. Auger and N. Hansen. "A Restart CMA Evolution Strategy with Increasing Population Size". In: *2005 IEEE Congress on Evolutionary Computation*. Vol. 2. 2005, 1769–1776 Vol. 2. DOI: 10.1109/CEC.2005.1554902.

A copula boosting approach for dependent censoring

Annika Stroemer[1], Nadja Klein[2], Guillermo Briseño Sanchez[2] and Andreas Mayr[1]

In survival analysis, censoring is an inherent observation that is usually assumed to be unrelated to the event of interest. When this assumption is not fulfilled, traditional methods like the Cox model may yield skewed or biased results. For example, if a patient's health deteriorates and the patient chooses to withdraw from the trial due to a poor prognosis, the time of censoring depends on the patient's health status. To deal with dependent censoring, in this work we propose to utilize distributional copula regression via model-based boosting. This approach allows to model the joint distribution of survival and censoring times by linking appropriately marginal distributions for T and C through a parametric copula. Rather than assuming the marginals are known, all distribution parameters (including the copula parameter) are estimated simultaneously as functions of (potentially different) covariates.

A key merit of boosting is that estimation is even feasible for high-dimensional data with $p > n$, when classical estimation frameworks easily meet their limits. In addition, the boosting algorithm includes data-driven variable selection. To investigate the performance of our approach under controlled conditions, we first conduct a simulation study. Furthermore, we illustrate its practical application analysing the survival of colon cancer patients from an observational study.

[1] Department of Medical Biometrics, Informatics and Epidemiology, University of Bonn, Bonn, Germany
[2] Chair of Uncertainty Quantification and Statistical Learning, Research Center Trustworthy Data Science and Security (UA Ruhr) and Department of Statistics (Technische Universität Dortmund), Dortmund, Germany

stroemer@imbie.uni-bonn.de, nadja.klein@tu-dortmund.de, briseno@statistik.tu-dortmund.de, amayr@uni-bonn.de

Development of a recommender system for targeted antibiotic therapy in sepsis

Hannah Marchi[1,2], Sophie Schmiegel[1], Christiane Fuchs[1,2]

Antibiotic resistances represent a major challenge for society, health policy and the economy. Due to genetic changes in the bacteria, resistances might already arise three to five years after the introduction of a new antibiotic. The use of broad-spectrum antibiotics, which cover a wide range of pathogens, further promotes the spread of resistance.
At present, however, sepsis is often treated with broad-spectrum antibiotics as initial therapy. Starting the treatment within the first three hours after occurrence is crucial for the survival of sepsis patients. Due to a lack of information about the underlying pathogen and effective targeted antibiotics, a broad-spectrum antibiotic is a safe first choice to save patients' lives as its broad coverage is likely to have a combative effect on the sepsis causing pathogens.
However, the use of such broad-spectrum antibiotics exacerbates the spread and severity of antibiotic resistances, both locally and globally in the long term. An important measure for reducing antibiotic resistances is the use of more targeted antibiotics wherever possible.
In this project, we aim to individually identify narrow spectrum antibiotics for treating sepsis patients which are equally or even better suitable than initially prescribed broad-spectrum antibiotics. We base our considerations on data about sepsis patients who were admitted to the intensive care unit of the Evangelisches Klinikum Bethel (EvKB, Germany) between 2012 and 2023. The data includes clinical information such as age, gender and weight, time series of laboratory and vital signs, e.g. heart rate, body temperature, procalcitonin and lactate, as well as information on ventilation, dialysis requirements and the suspected focus of infection. Further it contains administered medication and results of microbiological analyses, including the resistance status of the pathogens present. Statistical method evaluation and validation, however, relies on tailored synthetic data.
We present our approach to find a targeted therapy at the time of the sepsis diagnosis and discuss upcoming challenges. We use a hybrid recommender system in which the patients represent the users, and the various therapies can be seen as items. In an upstream step, we investigate how to model the effectiveness of a prescribed antibiotic. We use the outcome of this investigation to build a therapy-patient matrix, which contains a value for the effectiveness per patient for each prescribed antibiotic. This matrix is rather sparse, as each patient has received only a few of all possible antibiotics. In the next steps, we assume both that a therapy has similar effectiveness values in similar patients and that similar therapies have comparable effectiveness values in a patient. Therefore, based on the success (effectiveness) of the prescribed therapies, we use both patient similarity (collaborative filtering) and therapy similarity (content-based filtering) to estimate

[1] Faculty of Business Administration and Economics, Bielefeld University, Germany
[2] Institute of Computational Biology, Helmholtz Munich, Germany

hannah.marchi@uni-bielefeld.de, sophie.schmiegel@uni-bielefeld.de,
christiane.fuchs@uni-bielefeld.de

the effectiveness of non-prescribed therapies and thus fill the empty cells of the matrix. Patient similarity considers clinical data, laboratory values, vital signs and the infection focus and is calculated using Gower's similarity coefficient, whereby the weighting of the covariates is determined according to medical relevance. For therapies, we assume that those which partially cover the same pathogens have similar effectiveness values. The similarity of two therapies is calculated by determining specific similarity values for matches and mismatches per pathogen and subsequently calculating the arithmetic mean over all pathogens.
Additionally, stopping criteria (e.g. allergies, guidelines) are applied to exclude certain therapies, and finally, a recommendation is made for the most promising therapies by ranking the effectiveness values of all therapies for each patient.
We present steps taken to answer the presented research question with the final goal to develop a clinical decision support system which recommends an effective and targeted initial antibiotic with minimal side effects. This system would provide everyday support for doctors and contribute to reducing the global problem of antibiotic resistances.

Inductive Logic Programming for Single-cell Analysis

Tobias Nietsch[1], Julian D. Schwab[1], Hans A. Kestler[1]

Inductive Logic Programming (ILP) describes an intersection between the areas of Machine Learning and Logic Programming [1]. The general goal of the approach is to find a hypothesis, consisting of a set of logic rules, that describes a given set of positive examples, while rejecting a given set of negative examples. The generated logic rules represent underlying relations that are consistent with the given observations (examples) and if provided, with additional background knowledge that can further guide the search [2]. Two substantial advantages of ILP in comparison to many other Machine Learning approaches such as Neural Networks (NNs) are that reliable rules can already be learned from a small amount of data (vs. large amounts of training data needed for NNs) as well as an increased explainability of the resulting logic programs (vs. "black boxes" employed in NNs) [3]. ILP was shown to be a suitable approach in bioinformatics applications given that biological structures, described as interaction networks, can be represented as relations [3]. To overcome the lack of dynamic descriptions in such interaction networks, Boolean Networks (BNs) are used for investigating dynamic behavior of underlying biological systems based on inferred state transitions of the respective systems over time. One application of increasing interest is the analysis of single cells, which yields insights into heterogeneities within cell populations and tissues as well as composition and regulatory mechanisms at the individual cell level [4]. Proposed algorithms following the ILP approach for learning logic programs from state transitions [5], in view of the mentioned advantages, raised our interest in integrating ILP in systems biology. To begin with, we compared a proposed extension of this learning approach for the analysis of single-cell RNA-sequencing (scRNA-seq) data [6] to a previously published method for analyzing structural and dynamic properties based on reconstructed BN ensembles [7]. Promising initial results support our aim to further investigate how single-cell analysis can benefit from using ILP by tackling limitations of current approaches.

[1] Ulm University, Institute of Medical Systems Biology

tobias.nietsch@uni-ulm.de, julian.schwab@uni-ulm.de, hans.kestler@uni-ulm.de

References

1 Muggleton, S., 1991. Inductive Logic Programming. New Generation Computing 8, no. 4: 295–318.

2 Muggleton, S., de Raedt, L., 1994. Inductive Logic Programming: Theory and methods. The Journal of Logic Programming, Special Issue: Ten Years of Logic Programming 19–20, 629–679.

3 Cropper, A., Dumančić, S., 2022. Inductive Logic Programming At 30: A New Introduction. Journal of Artificial Intelligence Research 74, 765–850.

4 Yuan, G.-C., Cai, L., Elowitz, M., Enver, T., Fan, G., Guo, G., Irizarry, R., Kharchenko, P., Kim, J., Orkin, S., Quackenbush, J., Saadatpour, A., Schroeder, T., Shivdasani, R., Tirosh, I., 2017. Challenges and emerging directions in single-cell analysis. Genome Biology 18, 84.

5 Inoue, K., Ribeiro, T., Sakama, C., 2014. Learning from interpretation transition. Machine Learning 94, 51–79.

6 Buchet, S., Carbone, F., Magnin, M., Ménager, M., Roux, O., 2021. Inference of Gene Networks from Single Cell Data through Quantified Inductive Logic Programming, in: The 12th International Conference on Computational Systems-Biology and Bioinformatics, CSBio2021, 48–63.

7 Schwab, J.D., Ikonomi, N., Werle, S.D., Weidner, F.M., Geiger, H., Kestler, H.A., 2021. Reconstructing Boolean network ensembles from single-cell data for unraveling dynamics in the aging of human hematopoietic stem cells. Computational and Structural Biotechnology Journal 19, 5321–5332.

Berichte des Instituts für Medizinische Systembiologie

Herausgeber:
Institut für Medizinische Systembiologie
Universität Ulm
89081 Ulm