10th Channel Network Conference

May 19 – 21 2025
Liège Belgium

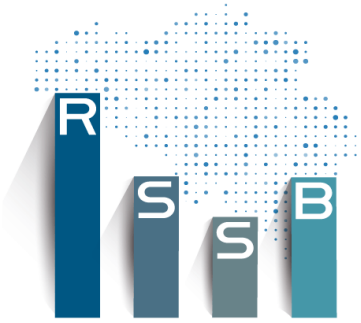
cnc25.sciencesconf.org

Book of Abstracts

Sponsors


fnrs
LA LIBERTÉ DE CHERCHER


LIÈGE
université


R
S
S
B


QUETELET SOCIETY

About CNC25

The Channel Network Conference is a biennial conference organized by the International Biometric Society (IBS) channel network. The channel network comprises the Belgian, French, British & Irish, and Dutch regions of the IBS. This conference gathers statisticians, mathematicians, and data scientists to discuss the latest methodology for the analysis of data in biosciences, including agriculture, biomedical science, environmental science, and allied disciplines. It is a 3-day conference with short courses, invited and contributed sessions. In 2025, the Belgian Region will host the conference in Liege city.

As for all IBS conferences, contributed abstracts for both oral and poster presentations have come across the wide range of methodological topics and application areas pursued by society members. We are proud to announce that Maria Xosé Rodriguez-Alvarez (University of Vigo) and Ruth Keogh (London School of Hygiene and Tropical Medicine) will deliver keynote sessions. In addition, we will have three Invited Sessions and three Short Courses.

A word of welcome from the local organizing committee

It is our pleasure, as co-chairs of the Local Organizing Committee, to welcome you to Liège for the 2025 Channel Network Conference (CNC25).

Located in the French-speaking part of Belgium, the University of Liège (ULiège) hosts nearly 27,000 students from 123 different nationalities in a dynamic and multicultural city. Thanks to its ideal situation - less than an hour away from Brussels and Cologne, two hours from Paris and three hours from London and Amsterdam - the University of Liège was a natural choice for hosting the 10th Channel Network Conference.

In times when science is under threat and "fake" information is spreading rapidly, it is our duty to continue delivering reliable, trustworthy, and explainable results. Making sense of the overwhelming amount of data across the various disciplines of the IBS network is both an exciting and challenging task.

Let the motto of the University of Liège, "*Scientia Optimum*", meaning "excellence through science" or "the best through knowledge" - serve as an inspiration throughout these three days. Together with the Scientific Committee, we have prepared a rich program of short courses, invited and contributed sessions, covering most aspects of the biostatistics and showcasing the latest and most advanced developments in biosciences data analysis.

Thank you for coming to Liège. We hope you will enjoy this 10th CNC conference.

Arnaud Monseur

(co-chair LOC)

Pierre-Yves Sacré

(co-chair LOC)

Committees

Scientific Committee (SC)

- Brian Tom, University of Cambridge, British & Irish Region
- Nicole Augustin, University of Edinburgh, British & Irish Region
- Boris Hejblum, INSERM, Bordeaux, French region
- Sophie Ancelet, IRSN, French region
- Philippe Lambert, University of Liège, Belgian Region
- Arnaud Monseur, Cencora-PharmaLex, Belgian Region
- Carel F.W. Peeters, Wageningen University & Research, Dutch Region

Local Organizing Committee (LOC)

- Arnaud Monseur, *Co-Chair*, Cencora-PharmaLex, Belgium
- Pierre-Yves Sacré, *Co-Chair*, University of Liège, Belgium
- Catherine Legrand, UCLouvain, Belgium
- Gentiane Haesbroeck, University of Liège, Belgium
- Anne-Françoise Donneau, University of Liège, Belgium
- Eric Ziemons, University of Liège, Belgium

TIME	MONDAY MAY 19		
09:00 - 12:30	**short course 1** **Time-dependent effects and time-dependent covariates in survival analysis** Teacher: Hein Putter Room: A1	**short course 2** **Introduction to modern Generalised Additive Models in R** Teacher: Simon Wood Room: A300	**short course 3** **Modern Factorial Design of Experiments** Teacher: Peter Goos Room: A2
12:30 - 13:30	*lunch break*		
13:30 - 14:00	**Opening Ceremony** Room: A300		
14:00 - 15:00	**Keynote Address: Maria Xosé Rodriguez-Alvarez** Chair: Philippe Lambert Room: A300		
15:00 - 15:30	*Tea & Coffee Break*		
15:30 - 17:00	**Contributed session 1: Experimental design, Tests & Variable selection** Chair: Stijn Jaspers Room: A500 Speakers: *Dominique-Laurent Couturier* *Peter Goos* *Stijn Jaspers* *Kayane Robach* *Ophelie Schaller*	**Contributed session 2: Survival Analysis** Chair: Hein Putter Room: A300 Speakers: *Liesbeth C. De Wreede* *Morine Delhelle* *Yilin Jiang* *Philippe Lambert* *Mar Rodríguez-Girondo*	**Contributed session 3: Joint and Latent Variable Modelling** Chair: Mark Brewer Room: A303 Speakers: *Auriane Gabaut* *Jonathan Kunst* *He Li* *Killian A.C. Melsen*

TIME	TUESDAY MAY 20		
09:00 - 10:30	**Invited session 1: Statistical Methods for Sustainable Management of Natural Resources** Chair: Nicole Augustin Room: A300 Speakers: *Alison Johnston* *Verena Trenkel* *Frederic Mortier*		
10:30 - 11:00	*Tea & Coffee Break*		
11:00 - 12:30	**Contributed session 4: Clinical and Medical Statistics** Chair: Peter Goos Room: A500 Speakers: *Said El Bouhaddani* *Steven Gilmour* *Yongxi Long* *Erik Van Zwet* *James Willard*	**Contributed session 5: Big Data and Machine Learning** Chair: Gentiane Haesbroeck Room: A300 Speakers: *Stefan Böhringer* *Michel Hof* *Connie Musisi* *Kai Ruan* *Giorgio Spadaccini*	**Contributed session 6: Functional Data Analysis** Chair: Maria Xosé Rodriguez-Alvarez Room: A303 Speakers: *Quentin Clairon* *Paul Eilers* *Marion Kerioui* *Tatsiana Khamiakova* *Corentin Segalas*
12:30 - 13:30	*lunch break*		
13:30 - 15:00	**Contributed session 7: Epidemiology & Causal Inference** Chair: Ruth Keogh Room: A500 Speakers: *Mailis Amico* *Yuwen Ding* *James Murray* *Chin Yang Shapland*	**Contributed session 8: Omics Data Analysis** Chair: Carel Peeters Room: A300 Speakers: *Mathilde Bruguet* *Jeroen Goedhart* *Madeline Vast* *Shizhe Xu*	**Contributed session 9: Bayesian Methods** Chair: Philippe Lambert Room: A303 Speakers: *Hortense Doms* *Leandro Garcia Barrado* *Thomas Klausch* *Clement Laloux*
15:00 - 15:30	*Tea & Coffee Break*		
15:30 - 17:00	**Invited session 2: Methods for high-dimensional feature selection** Chair: Carel Peeters Room: A300 Speakers: *Renee De Menezes* *Angel Reyero Lobo* *Michel Verleysen*		

17:00 - 17:30	**Poster Lightning Presentations** Chair: Room: A300 Speakers: *Guillaume Deside* *Lennart Hoheisel* *Christiana Kartsonaki* *Valeria Leiva-Yamaguchi* *Kato Michiels* *Areti Papadopoulou* *Henk Van Der Pol* *Celia Vidal*
17:30 - 18:00	**Poster session**
19:30	*Conference dinner*

TIME	WEDNESDAY MAY 21
09:00 - 10:30	**Invited session 3: Double/Debiased Machine Learning** Chair: Stefan Böhringer Room: A300 Speakers: *Oliver Dukes* *Mathieu Evens* *Shaun Seaman*
10:30 - 11:00	*Tea & Coffee Break*
11:00 - 12:00	**Keynote Address: Ruth Keogh** Chair: Arnaud Monseur Room: A300
12:00 - 12:30	**Closing and Awards ceremony** Room: A300
12:30 - 13:30	*Farewell Lunch*

KEYNOTE ADDRESS 1:

MARIA XOSE RODRIGUEZ-ALVAREZ

Computational Methods for Multidimensional P-Splines: Advances and Applications in Agriculture and Neuroscience

MARIA XOSE RODRIGUEZ-ALVAREZ

Departamento de Estatística e Investigación Operativa, Universidade de Vigo, 36310, Vigo, Spain

E-mail for correspondence: mxrodriguez@uvigo.es

Abstract: Multidimensional P-splines are a powerful tool for modelling complex, non-linear interactions. However, their application is often constrained by high computational costs, especially when complex smoothing structures, such as anisotropy or local adaptiveness, are required. This talk focuses on recent advancements in computational methods for multidimensional P-splines, offering efficient solutions to these challenges. Through two case studies, we demonstrate how these methods alleviate the computational burden, facilitating the practical application of multidimensional P-splines to complex, real-world data across diverse research fields. The first case study presents a spatio-temporal P-spline hierarchical model for analysing high-throughput phenotyping (HTP) data in agricultural experiments. Using this approach, we model the temporal evolution of genetic effects on a given phenotype, while accounting for spatio-temporal noise and experimental design/post-blocking factors. In the second case study, we introduce a novel anisotropic, locally adaptive P-spline model designed to capture sharp transitions in spatial and spatio-temporal data. The model is applied to the analysis of data from experiments studying neuronal activity in the visual cortex, where, unlike traditional/non-adaptive P-splines, it effectively identifies the spatio-temporal properties of receptive fields.

KEYNOTE ADDRESS 2:

RUTH KEOGH

Chances, choices and challenges: Trial Emulation using patient registry data in cystic fibrosis

RUTH KEOGH

Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London, UK

E-mail for correspondence: Ruth.Keogh@lshtm.ac.uk

Abstract: Randomised controlled trials (RCTs) are the gold standard for generating evidence on the effects of treatments, but they have limitations and are not always feasible. Observational data offer alternative chances to study the causal effects of treatments, and 'trial emulation' has emerged as a powerful framework for helping in designing such studies. This talk will discuss trial emulation in the field of cystic fibrosis (CF). There remain many unanswered questions about the effects of treatments in CF and national patient registries offer the possibility of assessing a range of questions that are not feasible to address in RCTs. The challenges and opportunities of trial emulation in this context will be discussed through several examples using the UK Cystic Fibrosis Registry, which contains detailed and carefully collected longitudinal data on >99% of the UK CF population. This will include a discussion of the process of conducting "benchmarking" studies in which we emulate previously conducted trials using patient registry data with the aim of exploring the reliability of the data for providing evidence about treatment effects. A further example will consider the challenges of estimating the long-term effects of disease-modifying treatments introduced for large segments of the CF population, including by using negative controls. Some of the choices that need to be made when conducting trial emulations will be discussed, which include how to use annually collected confounder and outcome data in combination with treatment prescriptions data.

INVITED SESSION:

DOUBLE/DEBIASED MACHINE LEARNING

An introduction to double/debiased machine learning

SHAUN SEAMAN

MRC Biostatistics Unit, University of Cambridge, UK

E-mail for correspondence: shaun.seaman@mrc-bsu.cam.ac.uk

Abstract: In problems involving missing data or causal inference, biometricians may wish to estimate one or a small number of quantities of interest but find that this requires estimating more complex 'nuisance' functions. For example, we might wish to estimate the average causal effect (ACE) of an exposure, defined as the expected difference between an individual's outcome if exposed and outcome if unexposed, from a sample where exposure is not randomly assigned. Commonly used techniques for this are inverse probability weighting (IPW) and regression imputation (RI). Both involve estimating a nuisance function: the conditional probability of exposure for IPW and the conditional expectation of the outcome for RI. Parametric models could be specified for these, but it is tempting to use flexible machine-learning techniques, to reduce the risk of model misspecification. There is, however, a problem associated with naive use of such methods for this purpose: machine-learning estimators of nuisance functions typically converge slowly, and this slow convergence may affect the convergence rate of the estimator of the quantity of ultimate interest, e.g. the ACE. Such slow convergence greatly complicates the construction of valid confidence intervals. Debiased machine learning is a group of techniques designed to address this problem. I shall provide an introduction to these techniques.

INVITED SESSION:

DOUBLE/DEBIASED MACHINE LEARNING

Rethinking the Win Ratio: A Causal Framework for Hierarchical Outcome Analysis

MATHIEU EVEN

INRIA-INSERM, University of Montpellier, France

E-mail for correspondence: mathieu.even@inria.fr

Abstract: Quantifying causal effects in the presence of complex and multivariate outcomes is a key challenge to evaluate treatment effects. For hierarchical multivariate outcomes, the FDA recommends the Win Ratio and Generalized Pairwise Comparisons approaches. However, as far as we know, these empirical methods lack causal or statistical foundations to justify their broader use in recent studies. To address this gap, we establish causal foundations for hierarchical comparison methods. We define related causal effect measures, and highlight that depending on the methodology used to compute Win Ratios or Net Benefits of treatments, the causal estimand targeted can be different, as proved by our consistency results. Quite dramatically, it appears that the causal estimand related to the historical estimation approach can yield reversed and incorrect treatment recommendations in heterogeneous populations, as we illustrate through striking examples. In order to compensate for this fallacy, we introduce a novel, individual-level yet identifiable causal effect measure that better approximates the ideal, non-identifiable individual-level estimand. We prove that computing Win Ratio or Net Benefits using a Nearest Neighbor pairing approach between treated and controlled patients, an approach that can be seen as an extreme form of stratification, leads to estimating this new causal estimand measure. We extend our methods to observational settings via propensity weighting, distributional regression to address the curse of dimensionality, and a doubly robust framework. We prove the consistency of our methods, and the double robustness of our augmented estimator. These methods are straightforward to implement, making them accessible to practitioners. Finally, we validate our approach using synthetic data, and observational oncology dataset.

INVITED SESSION:

DOUBLE/DEBIASED MACHINE LEARNING

Nonparametric tests of treatment effect heterogeneity for policy-makers

OLIVER DUKES

Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Gent, Belgium

E-mail for correspondence: Oliver.Dukes@UGent.be

Abstract: Recent work has focused on nonparametric estimation of conditional treatment effects, but inference has remained relatively unexplored. We propose a class of nonparametric tests for both quantitative and qualitative treatment effect heterogeneity. The tests can incorporate a variety of structured assumptions on the conditional average treatment effect, allow for both continuous and discrete covariates and does not require sample splitting. Furthermore, we show how the tests are tailored to detect alternatives where the population impact of adopting a personalised decision rule differs from using a rule that discards covariates. The proposal is thus relevant for guiding treatment policies. The utility of the proposal is borne out in simulation studies and a re-analysis of an AIDS clinical trial.

This is joint work with Mats Stensrud, Riccardo Brioschi and Aaron Hudson

INVITED SESSION:

STATISTICAL METHODS FOR SUSTAINABLE MANAGEMENT OF NATURAL RESOURCES

Modelling genetics data for fisheries management

VERENA TRENKEL

DECOD (Ecosystem Dynamics and Sustainability), Institut Agro, IFREMER, INRAe, Nantes, France

E-mail for correspondence: Verena.Trenkel@ifremer.fr

Abstract: Increased availability of genetics data has made it possible to attain sufficient samples sizes for population level applications in fisheries science. In this presentation, I will be talking about the statistical methods underlying the close-kin mark-recapture approach, which is increasingly used for abundance estimation in support of fisheries management or conservation. The approach consists of recapturing individuals via their relatives, identifying them using genetic information such as single-nucleotide polymorphism markers (SNP). Various simplifying assumptions can be made to overcome data limitations, such as uncertain or absent age information for sampled individuals. Two marine fish species with contrasting biology, thornback ray and meagre, will serve as case studies to evaluate the sensitivity of results to different modelling decisions and illustrate the additional biological insights that can be gained.

INVITED SESSION:

STATISTICAL METHODS FOR SUSTAINABLE MANAGEMENT OF NATURAL RESOURCES

Tree Pólya splitting models with zero inflation. Application to forecast tree species in Central African forests

FRÉDÉRIC MORTIER

CIRAD La Recherche Agronomique pour le développement, AMAP Unit, Montpellier, France

E-mail for correspondence: frederic.mortier@cirad.fr

Abstract: Understanding the impact of climate change on tropical rainforest ecosystems is crucial to promote efficient conservation strategies. The classical approach remains the use of species-specific distribution model. However, in species-rich ecosystems with many rare species, such an approach is doomed to failure. Moreover, univariate approaches ignore species dependencies. However, biodiversity is not merely the sum of species but the result of multiple interactions. Modelling multivariate count data allowing for flexible dependencies as well as zero inflation and overdispersion is challenging. In this presentation, we develop a new family of models called the zero-inflated binary tree Pólya-splitting models. This family allows the decomposition of a multivariate count data into a successive sub-model along a known binary partition tree. In the first part, I will present the general form of the zero-inflated binary tree Pólya-splitting model, studying the properties of the specified model in terms of marginal and conditional properties (distribution and moment). The second part presents the extension to the regression context. Finally, we finish presenting results on a real case study based on an impressive dataset consisting of the abundance of more than 180 tree taxa sampled on 1,571 plots covering more than 6 million hectares from the Congo Basin tropical rainforests.

INVITED SESSION:

STATISTICAL METHODS FOR SUSTAINABLE MANAGEMENT OF NATURAL RESOURCES

People and birds: Analytical methods for citizen science biodiversity data

ALISON JOHNSTON

Centre for Research into Ecological and Environmental Modelling, University of St Andrews, St Andrews, UK

School of Mathematics and Statistics, University of St Andrews, St Andrews, UK

E-mail for correspondence: alison.johnston@st-andrews.ac.uk

Abstract: Citizen science data have become a key source for understanding ecological systems and informing conservation, but working with these data brings many statistical challenges. Key challenges to address analytically include spatial and temporal biases, observer preferences, and observer experience. I'll outline how we've approached these challenges to analyse eBird data, to learn more about avian ecology. eBird is the largest biodiversity citizen science project in the world and contains over 1 billion bird observations contributed by over a million participants. We can use these data to estimate bird species distributions, migratory movements, demographics, and population trends. Each of these targets of ecological inference requires different consideration of the datasets and different analytical methods. I'll outline how we use a spatio-temporal ensemble of machine learning models to estimate bird distributions and migratory movements and use double machine learning methods to account for confounding factors in estimating bird population trends. I'll also outline how we can for the first time use large-scale observational data to estimate demographic parameters. All of these methods allow us to create new ecological knowledge from unstructured citizen science data.

INVITED SESSION:

METHODS FOR HIGH-DIMENSIONAL FEATURE SELECTION

When Knockoffs fail: diagnosing and fixing non-exchangeability of Knockoffs

ANGEL REYERO LOBO

Institut de Mathématiques de Toulouse ; UMR5219 Université de Toulouse, France

E-mail for correspondence: angel.reyero-lobo@inria.fr

Abstract: Knockoffs provide a powerful statistical framework for conditional variable selection in high-dimensional settings while ensuring statistical control, which is crucial for reliable inference. However, the validity of knockoff-based guarantees depends on an exchangeability assumption that is difficult to verify in practice, and there is limited discussion in the literature on how to address violations of this assumption. Exchangeability requires that knockoff variables be indistinguishable from the original data that is, any entry in the original input can be swapped with its corresponding knockoff while preserving the exact same distribution.
To maintain reliable inference, we introduce a diagnostic tool based on Classifier Two-Sample Tests, which effectively detects the shortcomings of classical knockoff generators in both simulated and real data. The results show that using non-exchangeable knockoffs leads to a severe inflation of false positives. As a solution, we present an alternative knockoff construction that builds a predictor for each variable based on all others. We prove that this approach restores error control. Since this approach is computationally expensive, we also present a more efficient variant which maintains error control in simulated data and semi-simulated experiments based on neuroimaging data.

This is joint work with Alexandre Blain, Bertrand Thirion, Julia Linhart and Pierre Neuvial.

Conflicting Feature Selection Requirements: How Mutual Information Can Show a Trade-off

MICHEL VERLEYSEN

UCLouvain - ICTEAM/ELEN, Belgium

E-mail for correspondence: michel.verleysen@uclouvain.be

Abstract: Feature selection is an essential task within a data modelling process: it helps to design better, simpler, more efficient, and more explainable models, and it helps users make the connection between complex, non-interpretable, or black-box models and their understanding of the data. Feature selection, coupled or not with prediction and classification models, has attracted much attention in the machine learning community for more than two decades.

An ideal feature selection method should combine conflicting properties related to nonlinearities, multivariate data, computational complexity, interpretability, and independence from the prediction/classification model, among others. Few methods excel at combining these properties. We will first show how the user's preference for some of these properties can influence the choice of feature selection method. Then, we will show that mutual information-based methods can provide a good compromise in many real-world situations.

INVITED SESSION:

METHODS FOR HIGH-DIMENSIONAL FEATURE SELECTION

Integrating multiple types of (omics) data: connecting many needles in multiple haystacks

RENÉE X. DE MENEZES

Biostatistics Centre, Department of Psychosocial and Epidemiological Research, Netherlands Cancer Institute

E-mail for correspondence: r.menezes@nki.nl

Abstract: Biomedical research currently collects large amounts of data from various sources, structured as well as unstructured. Making sense of these data can help us understand disease progression, such as cancer, as well as predict treatment response. There is great need for methods which allow us to combine multi-typed complex data for inference and prediction, in a way which also enables us to identify important data features. The challenge is to do this with methods powerful enough in high-dimensional spaces.

In this talk I will present a framework to test for associations between (multi-)omics data. This framework can be used with two or more omics datasets. Correlation between datasets, or between features of the same data, can be taken into account in the method. The framework can be used for a single response variable at a time, or a set of response variables, making it flexible as well as robust to effects found for only a few response variables, if desired. Reassuringly, tests yield replicable results across datasets. I will also briefly talk about an approach to analyse single-cell data, beyond using projections on lower-dimensional spaces, as well as current efforts to combine imaging data with other complex data in various studies.

Multivariate Adjustments for Average Equivalence Testing

Dominique-Laurent Couturier[1], Younes Boulaguiem[2], Luca Insolia[2], Maria-Pia Victoria-Feser[2], Stéphane Guerrier[3]

[1] Medical Research Council Biostatistics Unit, University of Cambridge, United Kingdom
[2] Faculty of Science, University of Geneva, Switzerland
[3] Department of Statistics, University of Bologna, Italy

E-mail for correspondence: dominique.couturier@mrc-bsu.cam.ac.uk

Abstract: Multivariate (average) equivalence testing is widely used to assess whether the means of two conditions of interest are 'equivalent' for different outcomes simultaneously. The multivariate Two One-Sided Tests (TOST) procedure is typically used in this context by checking if, outcome by outcome, the marginal $100(1-2\alpha)\%$ confidence intervals for the difference in means between the two conditions of interest lies within pre-defined lower and upper equivalence limits. This procedure leads to a rapid power loss when the number of outcomes increases, especially when one or more outcome variances are relatively large.

We propose a finite-sample adjustment for this procedure, the multivariate α-TOST, that consists in a correction of α, the significance level, taking the (arbitrary) dependence between the outcomes of interest into account and making it uniformly more powerful than the conventional multivariate TOST. We present an iterative algorithm allowing to efficiently define $\alpha^\star$, the corrected significance level, a task that proves challenging in the multivariate setting due to the inter-relationship between $\alpha^\star$ and the sets of values belonging to the null hypothesis space and defining the test size. We finally apply the α-TOST in a case study on ticlopidine hydrochloride when simultaneously assessing bioequivalence for multiple pharmacokinetic parameters.

Key words: Multivariate (Bio)Equivalence Testing; Finite-Sample Adjustments; Two One-Sided Tests

Optimal Experimental Designs for Process Robustness Studies

Peter Goos[1], Ying Chen[2], Bernard Francq[3]

[1] KU Leuven - Division MeBioS
[2] KU Leuven
[3] GSK

E-mail for correspondence: peter.goos@kuleuven.be

Abstract: In process robustness studies, experimenters are interested in comparing the responses at different locations within the normal operating ranges of the process parameters to the response at the target operating condition. Small differences in the responses imply that the manufacturing process is not affected by the expected fluctuations in the process parameters, indicating its robustness. In this presentation, we propose an optimal design criterion, named the generalized integrated variance for differences (GID) criterion, to set up experiments for robustness studies. GID-optimal designs have broad applications, particularly in pharmaceutical product development and manufacturing. We show that GID-optimal designs have better predictive performances than other commonly used designs for robustness studies, especially when the target operating condition is not located at the center of the experimental region. In some situations that we encountered, the alternative designs typically used are roughly only 50% as efficient as GID-optimal designs. We will demonstrate the advantages of tailor-made GID-optimal designs through an application to a manufacturing process robustness study of the Rotarix liquid vaccine.

Using the Probability of Improved Prediction for Model Selection in the Presence of Outliers

Stijn Jaspers[1], Olivier Thas[1,2,3]

[1] Data Science Institute, I-BioStat, Hasselt University, 3590 Diepenbeek, Belgium
[2] Department of Applied Mathematics, Computer Science and Statistics, Faculty of Sciences, Ghent University, Krijgslaan 281, 9000 Ghent, Belgium
[3] National Institute of Applied Statistics Research Australia (NIASRA), University of Wollongong, Northfield Ave, Wollongong, NSW 2522, Australia

E-mail for correspondence: stijn.jaspers@uhasselt.be

Abstract: Outlying observations are known to have a large influence on the model selection process. Rabbi et al. (2022) make a comparison of four robust linear model selection criteria that are all based on two key components that are derived from Müller and Welsh (2005), i.e. a good model has:

i. the capability to fit the sample data y and $\mathbf{X}$ reasonably well, and
ii. the ability to predict future observations with great accuracy.

All four methods accommodate for the presence of outliers by considering a bounded robust loss function. The methods differ in the way they quantify the requirements above, with the most advanced method employing an out-of-bag error estimate based on a stratified bootstrap procedure. The values of the respective criteria are computed for a set of candidate models and the optimal model is selected based on the minimum value.

As an alternative, we propose the Probability of Improved Prediction (PIP), which measures how more often a model gives better predictions than another model, where better is determined based on a user-defined loss function. In contrast to the four methods above, the PIP is less sensitive towards the choice of loss function as it directly compares between two models (m^0 and m^1) on an individual level by computing a score between zero and one that reflects the percentage of times a new observation y^* is predicted more accurately by model m^1 as compared to the prediction by model m^0.

A simulation study and data application show the performance of our new concept as compared to the existing methodology. Moreover, although the original four methods could probably be modified to also fit in the setting of more complicated machine learning models, we show that the PIP can directly be applied. In this perspective, some results with respect to gradient boosting machines are presented as well.

Keywords: Model Selection; Outliers; Probability of Improved Prediction

Müller S and Welsh A (2005). Outlier robust model selection in linear regression. Journal of the American Statistical Association 100(472), 1297-1310

Rabbi F, Khalil A, Khan I, Almuqrin MA, Khalil U and Andualem M. (2022). Robust model selection using the out-of-bag bootstrap in linear regression Scientific Reports, 12, 1-10.

False Discovery estimation in Record Linkage

Kayané Robach[1,2], Michel H. Hof[1,2], Mark A. van de Wiel[1,2]

[1] Department of Epidemiology and Data Science, Amsterdam UMC, Amsterdam, Netherlands
[2] Amsterdam Public Health, Methodology, Amsterdam, Netherlands

E-mail for correspondence: k.c.robach@amsterdamumc.nl

Abstract: This data era enables combining information to broaden research opportunities without costly new data collection. However, since data are not collected with specific future research questions in mind, and lack unique identifiers for privacy reasons, Record Linkage (RL) algorithms are used to assemble observations. The task poses challenges due to the sub-par reliability of partially identifying variables. Estimating the False Discovery Rate (FDR) associated with RL therefore holds importance for later inference. In particular in healthcare studies, estimating the Type I error of a set of linked records is crucial to determine the reliability of the inference drawn from the linked data. We introduce a new method to estimate the FDR and give guidelines for applying it on any sort of RL algorithm. Our recipe consists in linking records from real and synthesised data, estimating the FDR with the synthetic set. Our procedure enables identifying a threshold on the posterior linkage probabilities for which the RL process may be reliable. We investigate the performance of this methodology with well-known RL algorithms and data sets before applying it to the Netherlands Perinatal Registry to show the importance of the FDR in RL when studying children/mother dynamics in healthcare records.

Keywords: Record Linkage; False Discovery Rate

Absorbing List in Multiple Systems Estimation

Ophélie Schaller[1], Andrew Titman[1], Rachel McCrea[1]

[1] Lancaster University, Lancaster, United Kingdom

E-mail for correspondence: o.schaller@lancaster.ac.uk

Abstract: Multiple Systems Estimation is a family of statistical methods used to estimate wildlife and human population sizes when a total enumeration is unfeasible. These methods rely on combining information from partial lists that uniquely identify some individuals in the total population. In a social science setting, MSE methods most often suppose closeness of the population, however this assumption is not always justified. An absorbing list is a partial observation of a population happening at the moment when the individuals leave the population. In the presented work, we describe a modified model to the standard log-linear in order to include or detect potential absorbing lists. The new model is built by considering the order of the lists an individual has been observed by, taking into consideration the asymmetry of interactions between an absorbing list and other lists.

Keywords: Multiple Systems Estimation; Absorbing List; Hidden Population; Open Population

Bird S. M, and King R (2018). Multiple Systems Estimation (or Capture-Recapture Estimation) to Inform Public Policy. Annual Review of Statistics and Its Application.

Worthington H, McCrea R, King R and Vincent K (2021). How Ideas from Ecological Capture-Recapture Models May Inform Multiple Systems Estimation Analyses. Crime & Delinquency.

A multi-state model incorporating relative survival to estimate excess mortality during the Covid-19 pandemic in the Netherlands

Liesbeth C. de Wreede[1], Marije H. Sluiskes[1], Damjan Manevski[2], Hein Putter[1], Eva A. S. Koster[1]

[1] Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands
[2] Institute for Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia

E-mail for correspondence: l.c.de_wreede@lumc.nl

Abstract: The Covid-19 pandemic has led to excess mortality, especially in older persons. We investigated the impact of infection with and vaccination against Covid-19 on excess mortality during 2020-2021 in persons aged 65 years or older in the Netherlands.
We used relative survival methods outside their traditional context of estimating net survival of cancer patients, by incorporating them into a multi-state model considering vaccination, infection, and death based on the methodology by Manevski et al. This enabled splitting all mortality in background and excess mortality with and without intermediate events. The multi-state model is a time-inhomogeneous non-parametric Markov model. Transition probabilities were estimated by the Aalen-Johansen estimator. We applied the model on real-world, nationwide data from Statistics Netherlands.
We found that the probability of excess mortality increased with age, and was especially large for the oldest men. Almost all excess mortality took place after an infection, but its probability was much lower for vaccinated persons. Vaccinated persons who did not become infected experienced a negative excess hazard.
The current application of the novel model shows its value in the pandemic context. Further extensions incorporating regression modelling of background and (negative) excess hazards by means of additive hazards models are under development.

Keywords: Covid-19 Pandemic, Excess Mortality, Vaccination, Relative Survival, Multi-State Modelling

Manevski D, Putter H, ..., de Wreede LC (2022). Integrating relative survival in multi-state models-a non-parametric approach. Stat Methods Med Res., 31(6), 997-1012.

Copula based dependent censoring in cure models with covariates

Morine Delhelle[1], Anouar El Ghouch[1], Ingrid Van Keilegom[1,2]

[1] Institute of Statistics, Biostatistics and Actuarial Sciences, UCLouvain, Louvain-la-Neuve, Belgium
[2] Operations Research and Statistics Research Group, KU Leuven, Leuven, Belgium

E-mail for correspondence: morine.delhelle@uclouvain.be

Abstract: In survival data analysis, datasets that include both a cure fraction (i.e., individuals who will never experience the event of interest) and dependent censoring (loss to follow-up for a reason related to the event of interest before its occurrence) are not scarce. It is therefore essential to consider appropriate models and methods in order to avoid biased estimators of the survival function or incorrect medical conclusions. Delhelle and Van Keilegom (2025) proposed a fully parametric mixture cure model for the bivariate distribution of survival and censoring times (T, C), which deals with all these features. The model depends on a parametric copula and on parametric marginal distributions for T and C. A significant advantage of this approach in comparison to existing ones is that the copula which models the dependence between T and C is not assumed to be known, nor is the association parameter. Furthermore, the model allows for the identification and estimation of the cure fraction and the association between T and C, despite the fact that only the smallest of these variables is observable. This talk presents an improvement of this model. Administrative censoring is considered separately from dependent censoring, and covariates are included in the model.

Keywords: Dependent censoring; Cure models; Copulas; Covariates

Delhelle M, and Van Keilegom I (2025). Copula based dependent censoring in cure models. TEST (to appear - DOI 10.1007/s11749-024-00961-7)

A general approach to fitting multistate cure models based on an extended long data format

Yilin Jiang[1,2], Harm van Tinteren[2], Marta Fiocco[1,2,3]

1 Mathematical Institute, Leiden University, Leiden, the Netherlands
[2] Trial and Data Center, Princess Maxima Center, Utrecht, the Netherlands
[3] Department of Biomedical Data Science, Leiden University Medical Center, Leiden, the Netherlands

E-mail for correspondence: y.jiang@math.leidenuniv.nl

Abstract: A multistate cure model is a statistical framework used to analyse and represent the transitions that individuals undergo between different states over time, considering the possibility of being cured by initial treatment. This model is particularly useful in paediatric oncology where a proportion of the patient population achieves cure through treatment and therefore, they will never experience certain events. Our study provides a novel framework for defining such models through a set of non-cure states. We develop a generalized algorithm based on the extended long data format, an extension of the traditional long data format, where a transition can be split up to two rows each with a weight assigned reflecting the posterior probability of its cure status. The multistate cure model is fit on top of the current framework of multistate model and mixture cure model. The proposed algorithm makes use of the Expectation-Maximization (EM) algorithm and weighted likelihood representation such that it is easy to implement with standard package. Additionally, it facilitates dynamic prediction. The proposed algorithm is applied on data from the European Society for Blood and Marrow Transplantation (EBMT). Standard errors of the estimated parameters in the EM algorithm are obtained via a non-parametric bootstrap procedure.

Keywords: Multistate Cure Model; EM Algorithm; Weighted Likelihood; Extended Long Data Format; Dynamic Prediction

Beesley, L. J., & Taylor, J. M. (2019). EM algorithms for fitting multistate cure models. Biostatistics, 20(3), 416–432.

Blumen, I., Kogan, M., & McCarthy, P. J. (1955). The industrial mobility of labor as a probability process. Cornell University.

Breslow, N. E. (1972). Contribution to discussion of paper by D. R. Cox, Regression Models and Life Tables. Journal of the Royal Statistical Society, Series B, 34, 216–217.

Conlon, A., Taylor, J., & Sargent, D. J. (2014). Multi-state models for colon cancer recurrence and death with a cured fraction. Statistics in medicine, 33(10), 1750–1766.

de Wreede, L. C., Fiocco, M., & Putter, H. (2011). mstate: an R package for the analysis of competing risks and multi-state models. Journal of statistical software, 38, 1–30.

Fiocco, M., Putter, H., & van Houwelingen, H. C. (2008). Reduced-rank proportional hazards regression and simulation-based prediction for multi-state models. Statistics in Medicine, 27(21), 4340–4358.

Goodman, L. A. (1961). Statistical methods for the mover-stayer model. Journal of the American Statistical Association, 56(296), 841–868.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society Series B: Statistical Methodology, 44(2), 226–233.

Oakes, D. (1999). Direct calculation of the information matrix via the EM. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61(2), 479–482.

Putter, H., Fiocco, M., & Geskus, R. B. (2007). Tutorial in biostatistics: competing risks and multi-state models. Statistics in medicine, 26(11), 2389–2430.

Sommer, H., Wolkewitz, M., Schumacher, M., & Consortium, C.-N. (2017). The time-dependent cure-death model investigating two equally important endpoints simultaneously in trials treating high-risk patients with resistant pathogens. Pharmaceutical Statistics, 16(4), 267–279.

Sundin, P. T., Aralis, H., Glenn, B., Bastani, R., & Crespi, C. M. (2023). A semi-Markov multistate cure model for estimating intervention effects in stepped wedge design trials. Statistical Methods in Medical Research, 32(8), 1511–1526.

Sy, J. P., & Taylor, J. M. (2000). Estimation in a Cox proportional hazards cure model. Biometrics, 56(1), 227–236.

Van Houwelingen, H., & Putter, H. (2011). Dynamic prediction in clinical survival analysis. CRC Press.

Van Houwelingen, H. C., & Putter, H. (2008). Dynamic predicting by landmarking as an alternative for multi-state modeling: an application to acute lymphoid leukemia data. Lifetime data analysis, 14, 447–463.

Wei, G. C., & Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. Journal of the American statistical Association, 85(411), 699–704.

Fast Bayesian inference in additive cure survival models

Philippe Lambert[1,2]

[1] Institut de Mathématique, Université de Liège, Belgium
[2] Institut de Statistique, Biostatistique et Sciences Actuarielles (ISBA), UCLouvain, Belgium

E-mail for correspondence: p.lambert@uliege.be

Abstract: Markov chain Monte Carlo (MCMC) methods are widely used for Bayesian inference due to their efficiency and reliability in exploring the joint posterior distribution of model parameters. However, in complex additive models, slow mixing and potential convergence issues can significantly prolong the time required to obtain satisfactory results, which can be especially frustrating in interactive modelling or when working with large datasets. Often, it turns out that most conditional posteriors for both the regression and penalized parameters in the additive component can be well approximated by Gaussian distributions (Lambert & Gressani 2023).
Our focus will be on semi-parametric additive models for non-Gaussian censored data, where additive terms are represented as linear combinations of penalized B-splines (Marx and Eilers 1998). Our approach builds on the work of Rue et al. (2009) and the methodology underlying INLA. By employing Laplace approximations to the conditional posterior of the penalized parameters, we offer not only accurate and computationally efficient approximations to the multivariate posterior of regression and spline parameters but also facilitate simultaneous selection of smoothing parameters.
After a presentation of the general methodology, we will illustrate the use of Laplace P-spline models in flexible additive regression models for interval- and right-censored data (Lambert 2021), with a special focus on bounded hazard models with time-varying covariates (Lambert & Kreyenfeld 2025; Lambert & Bremhorst 2020). We will conclude the presentation with preliminary results on mixture cure models and their combination with flexible accelerated failure time models.

Keywords: Additive Model; Cure Survival Models; Laplace Approximation; Bayesian P-splines.

Gressani, O. and Lambert, P. (2021). Laplace approximations for fast Bayesian inference in generalized additive models based on P-splines. Comput. Stat. Data Anal., 154, 107088.

Lambert, P. (2021). Fast Bayesian inference using Laplace approximations in non- parametric double additive location-scale models with right- and interval-censored data. Comput. Stat. Data Anal., 161, 107250.

Lambert, P. and Bremhorst, V. (2020). Inclusion of time-varying covariates in cure survival models with an application in fertility studies. J. R. Stat. Soc. Ser. A, 183(1): 333-354.

Lambert, P. and Gressani, O. (2023). Penalty parameter selection and asymmetry corrections to Laplace approximations in Bayesian P-splines models. Statistical Modelling, 23(5-6): 409-423.

Lambert, P. and Kreyenfeld, M. (2025). Time-varying exogenous covariates with frequently changing values in double additive cure survival models: an application to fertility. J. R. Stat. Soc. Ser. A, qnaf035.

Marx, B.D., Eilers, P.H.C. (1998). Direct generalized additive modeling with penalized likelihood. Comput. Stat. Data Anal. 28 (2), 193-209.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J. R. Stat. Soc. Ser. B, 71(2), 319-392.

Wood, S. N. and Fasiolo, M. (2017). A generalized Fellner-Schall method for smoothing parameter optimization with application to Tweedie location, scale and shape models. Biometrics, 73, 1071-1081.

A penalized reduced rank regression model for multi-outcome survival data with applications to ageing

Mar Rodríguez-Girondo[1], Marije Sluiskes[1], Jelle Goeman[1], Hein Putter[1]

[1] Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

E-mail for correspondence: m.rodriguez_girondo@lumc.nl

Abstract: In recent years, large-scale epidemiological studies, such as the Leiden Longevity Study and the UK Biobank, have increasingly incorporated detailed age-at-disease-onset profiles derived from electronic health records. This integration offers valuable opportunities to explore the factors contributing to age-related multimorbidity and to develop risk scores that capture variations in the ageing process. Given that ageing is a complex phenomenon - encompassing both lifespan and healthspan, as well as the onset of age-related diseases - these data enable a more comprehensive understanding of its underlying mechanisms. However, the complexity of the data presents challenges due to the presence of multi-dimensional time-to-event outcomes and a high number of covariates.
We propose a novel methodological framework for analysing this type of data using a novel Lasso-penalized reduced-rank proportional hazards model. This model enables simultaneous fitting on the age-at-disease-onset of multiple age-related diseases, assuming the existence of shared latent factors underlying all considered age-related diseases. To deal with high-dimensional omics covariates, we propose incorporating a Lasso-type penalization. The performance of the new method is illustrated using UK Biobank data, utilizing metabolomics data as predictor variables.

Keywords: Reduced Rank Regression; Multiple-Outcome Data; High-Dimensional Survival Analysis; Multimorbidity; Metabolomics.

Regularized estimation of non-linear mixed effect models in the presence of high-dimensional biomarkers

Auriane Gabaut[1], Ariane Bercu[2], Mélanie Prague1*, Cécile Proust-Lima[2*]

[1] Université de Bordeaux, Inria, Inserm, Bordeaux Population Health Research Center; Vaccine Research Institute, Créteil, France
[2] Université de Bordeaux, Inserm, Bordeaux Population Health Research Center
* co-last authors

E-mail for correspondence: auriane.gabaut@inria.fr

Abstract: Mechanistic models, defined as nonlinear mixed-effects models based on differential equations, are relevant to describe temporal relationships between biological compartments. They are valuable for understanding dynamic interactions within systems, such as immune system mechanisms or cells interactions. However, model identifiability often requires observations across multiple compartments, which we believe could be inferred from high-dimensional transcriptomic data. Thus, we propose a regularization method to estimate nonlinear mixed-effects models involving unobserved compartments, measured through high-dimensional biomarkers. We aim to identify relevant biomarkers by regularizing the parameters linking them to the latent compartments while simultaneously estimating the population parameters from the structural mechanistic model.
To achieve this, we employ an iterative algorithm for Lasso-penalized maximum likelihood estimation. The estimation algorithm iterates between a regularization step and a mechanistic inference step. The regularization step updates coefficients linking latent compartments to biomarkers by solving penalized log-likelihood derivatives, approximated via second-order Taylor development. The mechanistic inference step focuses on estimating the mechanistic parameters using the Stochastic Approximation Expectation-Maximization (SAEM) algorithm, implemented through the Monolix software, considering the updated regularized coefficients from the first step.
We demonstrate our method's performance, in terms of biomarkers selection and parameters estimation, through simulations and a real-world application on COVID-19 mRNA vaccine data.

Keywords: Statistical Learning; Regularization; High Dimension; Model Selection

A latent factor approach to hyperspectral time series data for multivariate genomic prediction of grain yield in wheat

Jonathan Kunst[1], Killian Melsen[1], Chris Maliepaard[2], Willem Kruijer[1], Fred van Eeuwijk[1], Carel Peeters[1]

[1] Biometris, Wageningen University & Research, Wageningen, The Netherlands
[2] Plant Breeding, Wageningen University & Research, Wageningen, The Netherlands

E-mail for correspondence: jonathan.kunst@wur.nl

Abstract: High-dimensional time series data is becoming increasingly common within plant breeding programs and research using high-throughput phenotyping. However, analysing and integrating such data for genetic analysis and genomic prediction remains difficult. Here we use classic factor analysis with Procrustes rotation to approximate a dynamic factor-analytic process to extract relevant latent features from the genetic correlation matrix of hyperspectral secondary phenotype data. We use a subset of the CIMMYT wheat trials, consisting of 1,033 genotypes, which were measured across three irrigation treatments at several timepoints, using manned airplane flights with hyperspectral sensors capturing 62 bands in the spectrum of 385-850 nm. We perform multivariate genomic prediction using these latent variables to improve within-trial genomic prediction accuracy of wheat grain yield within three distinct watering treatments. By integrating latent variables of the hyperspectral data in a multivariate genomic prediction model, we improved prediction accuracy by 10 to 30 per cent, depending on treatment and data availability. Furthermore, the method elucidates which timepoints within a trial are important and how these relate to plant growth stages. This showcases how the combination of domain knowledge with data-driven approaches can increase accuracy and gain new insights from sensor data of high-throughput phenotyping platforms.

Keywords: Factor Analysis; Procrustes Rotation; Genomic Prediction; Time Series; Hyperspectral Data

Dynamic Principal Components Modelling of Longitudinal Omics Data

He Li[1], Said el Bouhaddani[2], Jeanine Houwing-Duistermaat[1,3]

[1] Dept. of Mathematics, Radboud University, Nijmegen, The Netherlands
[2] Dept. of Data Science and Biostatistics, UMC Utrecht, Utrecht, The Netherlands
[3] Dept. of Statistics, University of Leeds, Leeds, United Kingdom

E-mail for correspondence: he.li@ru.nl

Abstract: For a longitudinally measured omics dataset, the interest might be identifying a set of variables representing the dynamic structure. PCA approaches enable identifying sets of omics variables representing the cross-sectional structure. On the other hand, univariate analysis with fixed and random effects provides insights into whether a variable changes over time but ignores the joint distribution. We propose a novel multivariate dynamic probabilistic PCA-approach (DPPCA) which models the scores over time using a mixed model.
For estimation of the parameters, we maximise the log-likelihood using the EM algorithm. Via an extensive simulation study, we evaluate the performance of DPPCA for varying numbers of omics variables, dynamic components and time points. Finally, we apply DPPCA to a longitudinal metabolomics dataset from the TwinsUK study.
The first simulation results show that the parameter estimators for the time effect and random intercept variance related to the first component are unbiased (RMSE 0.10 to 0.83, 8.52 to 18.08, respectively). Concerning the data, a scree plot suggests to model five dynamic DPPCA components. Preliminary results show that the components change over time ($\hat{\beta}_1 = -0.21$), and are highly correlated over time ($\hat{\rho} = 0.99$). In the first component, most metabolites with high weights belong to lipoprotein subclasses.

Keywords: Latent Variable Models; Linear Mixed Models; Dimension Reduction; Metabolomics

Nyamundanda G, Gormley IC, and Brennan L (2014). A dynamic probabilistic principal components model for the analysis of longitudinal metabolomics data. Journal of the Royal Statistical Society: Series C, 63(5), 763-782.

Verdi S, Abbasian G, Bowyer RC, et al. (2019). TwinsUK: the UK adult twin registry update. Twin Research and Human Genetics, 22(6), 523-529.

El Bouhaddani S, Uh HW, Jongbloed G, et al. (2022). Statistical integration of heterogeneous omics data: Probabilistic two-way partial least squares (PO2PLS). Journal of the Royal Statistical Society: Series C, 71(5), 1451-1470.

Improving Genomic Prediction using High-dimensional Secondary Phenotypes: The Genetic Latent Factor Approach

Killian A.C. Melsen[1], Jonathan F. Kunst[1], José Crossa[2], Margaret R. Krause[3], Fred A. van Eeuwijk[1], Willem Kruijer[1], Carel F.W. Peeters[1]

[1] Mathematical and Statistical Methods group (Biometris), Wageningen University & Research, Wageningen, The Netherlands
[2] Global Wheat Program, International Maize and Wheat Improvement Centre (CIMMYT), Texcoco, Mexico
[3] College of Agricultural Sciences, Oregon State University, Corvallis, OR, USA

E-mail for correspondence: killian.melsen@wur.nl

Abstract: Decreasing costs and new technologies have led to an increase in the amount of data available to plant breeding programs. High-throughput phenotyping (HTP) platforms routinely generate high-dimensional datasets of secondary features that may be used to improve genomic prediction accuracy. However, integration of these data comes with challenges such as multicollinearity, parameter estimation in $p > n$ settings, and the computational complexity of many standard approaches. Several methods have emerged to analyse such data, but interpretation of model parameters often remains challenging. We propose genetic latent factor best linear unbiased prediction (glfBLUP), a prediction pipeline that reduces the dimensionality of the original secondary HTP data using generative factor analysis. In short, glfBLUP uses redundancy filtered and regularized genetic and residual correlation matrices to fit a maximum likelihood factor model and estimate genetic latent factor scores. These latent factors are subsequently used in multi-trait genomic prediction. Our approach performs better than alternatives in extensive simulations and a real-world application, while producing easily interpretable and biologically relevant parameters. We discuss several possible extensions and highlight glfBLUP as the basis for a flexible and modular multi-trait genomic prediction framework.

Keywords: Empirical Bayes; Factor Analysis; Genomic Prediction; High-Dimensional Data

A novel statistical inference approach to temporal clustering stability of longitudinal patient trajectories

Said el Bouhaddani[1], Emma Rademaker[1,2], Lennie P.G. Derde[2], Harm-Jan de Grooth[2], Olaf L. Cremer[2]

[1] Department of Data Science & Biostatistics, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands
[2] Department of Intensive Care Medicine, University Medical Center Utrecht, Utrecht, The Netherlands

E-mail for correspondence: s.elbouhaddani@umcutrecht.nl

Abstract:

Background: Identifying clinically meaningful subphenotypes to direct potential interventions in sepsis patients and provide personalized medicine, has been a major focus in intensive care research. To investigate sepsis heterogeneity over time and across patients, a retrospective study was conducted where immunobiomarkers were measured every 8 hours on 345 sepsis patients in Amsterdam and Utrecht UMC. Unsupervised clustering is a popular approach to assess sepsis immunobiomarker stability. Several metrics exist that visualize clustering stability, but a formal assessment of the significance of temporal cluster variability remains challenging.
Methods: We consider two approaches to formally assess the temporal stability of clustering structures in longitudinal patient trajectories. We propose a simple method using conditional bootstrap with Rand index and multinomial logistic regression to evaluate the cluster assignment consistency over time. Next, we derive an (approximate) asymptotic test for a time trend within the finite latent mixture modelling framework. A simulation study with realistic synthetic data is conducted to evaluate the type I error and power of the tests.
Application: We present the results of our simulation study and the application of our methodology to our sepsis dataset. Preliminary results show clusters of clinically meaningful immunobiomarker profiles, but with a high temporal crossover of patients between clusters.

Keywords: Finite Mixture Model; Clustering; Hypothesis Testing; Personalized Medicine; Sepsis

Boldea O, Magnus JR (2009). Maximum likelihood estimation of the multivariate normal mixture model. Journal of the American Statistical Association, 104(488), 1539–1549.
Marshall JC (2014). Why have clinical trials in sepsis failed? Trends in Molecular Medicine, 20(4), 195–203.

Test-Treat Clinical Trial Designs

Steven G. Gilmour[1], Rebecca E. A. Walwyn[2]

[1] King's College London, London, UK
[2] University of Leeds, Leeds, UK

E-mail for correspondence: steven.gilmour@kcl.ac.uk

Abstract: Clinical trials in various settings can have interventions which depend on the outcome of some tests. An old example is the FEVER trial, in which care home residents received an initial dose of flu vaccine. They were then randomised to have an antibody test or not. Those who were tested got a booster dose if their antibody response was low and no booster otherwise; those who were not tested got no booster. This is a special case of a more general test-treat design, in which the intervention patients receive depends on the outcome of some test. Identifying the appropriate design and analysis requires drawing a clear distinction between treatment strategies, to which patients are randomised, and treatment pathways, which are the sequences of interventions patients receive. A basic analysis, justified by the randomisation, suggests a comparison of treatment strategies often in the form of main effects and interactions of treatments with a factorial structure. Building on this, treatment pathways can be compared by identifying an appropriate set of orthogonal contrasts among treatment strategies. It will be shown that the almost-forgotten method of using orthogonal contrasts allows different clinical questions to be answered with simple changes to the design and analysis.

Keywords: Dynamic Treatment Regime; Factorial Design; Orthogonal Contrasts; Randomisation

e-values instead of p-values in clinical trials: what happens?

Yongxi Long[1], Erik van Zwet[1]

[1] Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

E-mail for correspondence: y.long@lumc.nl

Abstract: E-values have been proposed as a flexible alternative to p-values. In the context of clinical trials, they allow continuous inspection of the data ("peeking") and arbitrary stopping rules (Grünwald et al., 2024). However, this flexibility comes at the cost of reduced statistical power. Our goal is to make an empirical assessment of this trade-off.

We used the results of more than 20,000 randomized controlled trials from the Cochrane Database of Systematic Reviews (CDSR). We estimated the distribution of effect sizes, which was used to simulate trial trajectories up to and beyond their original sample sizes. Among various possible constructions of e-values, we focus on the Bayes factor of the estimated effect size distribution relative to the null model. We evaluated the type I error and the power of this e-value under continuous monitoring, stopping whenever a large enough e-value was observed.

To enable "anytime-valid testing", the studied e-value results in more conservative Type I error control than p-values. Moreover, it requires approximately 3.5 times the original sample size to match the power of the usual fixed-sample p-value.

The great flexibility of e-values comes at the price of a reduction of statistical power which translates to larger sample sizes. Although anytime-valid testing is very attractive, its practical implementation in clinical trials may be constrained by resource limitations.

Keywords: E-Value; Optional Stopping; Clinical Trials; Cochrane Database of Systematic Reviews

Grünwald, P., de Heide, R., & Koolen, W. (2024). Safe testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *86*(5), 1091–1128.

An empirical assessment of the cost of dichotomization of the outcome of clinical trials

Erik van Zwet[1], Frank Harrell[2], Stephen Senn[3]

[1] Leiden University Medical Center, Leiden, The Netherlands
[2] Vanderbilt University Medical Center, Nashville, TN, USA
[3] Statistical Consultant, Edinburgh, UK

E-mail for correspondence: E.W.van_Zwet@lumc.nl

Abstract: In the context of clinical trials, it is well known that binary endpoints are less informative than continuous (numerical) ones. This must be compensated by larger sample sizes to maintain sufficient power. We have used 21,435 unique randomized controlled trials (RCTs) from the Cochrane Database of Systematic Reviews (CDSR). Of these trials, 7,224 (34%) have a continuous (numerical) outcome and 14,211 (66%) have a binary outcome. We find that trials with a binary outcome have larger sample sizes on average, but also larger standard errors and fewer statistically significant results. We conclude that researchers do tend to increase the sample size to compensate for the low information content of binary outcomes, but do not do so sufficiently.
Binary endpoints are often the result of dichotomizing a continuous outcome which is sometimes called "responder analysis". In that case, the probit transformation of the responder probabilities is equal to Cohen's d. We use this equivalence to compare the required sample size with and without dichotomization. We hope that this will guide researchers during the planning phase. We also provide a method to calculate the loss of information after a responder analysis has been done. We hope that this will motivate researchers to abandon dichotomization in future trials.

Keywords: Dichotomization; Responder Analysis; Power; Sample Size

Senn SJ (2005). Dichotomania: an obsessive compulsive disorder that is badly affecting the quality of analysis of pharmaceutical trials. Proceedings of the International Statistical Institute, 55th Session, Sydney.

Toward dose optimisation in early phase oncology trials – escalation and expansion

James Willard[1], Thomas Jaki[1,2], Burak Kürsad Günhan[3], Christina Habermehl[3], Anja Victor[3], and Pavel Mozgunov[1]

[1] MRC Biostatistics Unit, University of Cambridge, Cambridge, United Kingdom
[2] Faculty of Informatics and Data Science, University of Regensburg, Regensburg, Germany
[3] Merck Healthcare KGaA, Darmstadt, Germany

E-mail for correspondence: james.willard@mrc-bsu.cam.ac.uk

Abstract: Early phase dose finding trials in oncology historically focused on identifying the correct doses of cytotoxic chemotherapies, where more benefit was expected from higher doses. Following this assumption, many trials have been designed to find a maximally tolerated dose (MTD), defined as the largest dose which satisfies specific toxicity constraints.
Recently, the FDA's Project Optimus highlighted the need to revisit this paradigm as modern therapies can provide benefit at doses lower than the MTD and so identifying these doses via dose optimisation has become a major objective of early phase trials.
Unfortunately, the small sample sizes and short observation periods of these trials makes dose optimisation challenging, since it is difficult to collect comprehensive information on the dose response curves under these settings. This may result in suboptimal doses being recommended for future development, adversely affecting later phase studies and ultimately patients. To help remedy this and collect more information before recommending doses for further study, the use of expansion cohorts has been proposed.
During dose expansion, cohorts of patients are assigned to a small number (usually 2-3) of the most promising doses determined at the end of dose escalation. In this work, we examine the relationship between the escalation and expansion components of dose optimisation. We compare the performance of a variety of adaptive stopping rules which determine when to terminate escalation and transition to expansion. Findings from an extensive simulation study will be discussed and recommendations for performing dose optimisation with expansion cohorts will be provided.

ForwardFlow: A deep-learning framework for robust frequentist and Bayesian inference

Stefan Böhringer[1]

[1] Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

E-mail for correspondence: s.boehringer@lumc.nl

Abstract: For Bayesian models, normalizing flow models have been introduced. Using simulated data, deep networks are trained to map standard normal random variables to the posterior distribution of given data. These models involve two networks, a mapping from data to parameters, and an invertible network mapping the posterior to a standard normal distribution.
Here, we introduce a forward model using a deep neural network that learns the full estimator of parametric models, mapping data to parameters. This allows to derive the confidence distribution using bootstrap methods and obviates the need for learning a normalizing flow. Bayesian models can still be dealt with using approximate Bayesian computations (ABCs). Forward models are flexible in that they can be used to perform inference on contaminated data such as missing values or outliers as long as the model has been trained on such data. Models trained on missing data represented by a separate indicator matrix allow for implicit imputation, i.e. efficient estimates in the presence of missing data.
In a simulation study we demonstrate robustness properties and also highlight finite sample exactness which is achieved by training on variable sample size.

Keywords: Deep Learning; Parametric Model; Bootstrap; ABC; Robust Inference

Radev ST et al. (2020). *BayesFlow*: Learning complex stochastic models with invertible neural networks. IEEE TNNLS, 32(4): 1452-66

Scalable Bayesian Record Linkage

Michel H. Hof[1]

[1] Department of Epidemiology & Data Science, Amsterdam UMC, Vrije Universiteit Medical Centre, Amsterdam, the Netherlands

E-mail for correspondence: m.h.hof@amsterdamumc.nl

Abstract: With the rise of digitization, more data is electronically stored, and combining datasets to answer new research questions has become common. In the absence of a unique identifier (e.g. citizen service number), record linkage relies on partially identifying variables (e.g. gender, place of residence, and initials) to classify pairs of observations as matches or non-matches.
In this work, data from two overlapping random samples of the same population are combined. Each entity has at most one observation per file, meaning each observation can match with at most one from the other file. Recent Bayesian approaches address this complex correlation structure but are computationally feasible only for small datasets (e.g. $\leq$ 5000 observations) or rely on ignoring aspects of the data-generating process.
To reduce computational burden, a novel blocking method is proposed, treating registered values as imperfect measurements of true values. Since most partially identifying variables remain stable over time, their true values must agree in matches. Additionally, we can use the blocking method to account for complex registration errors and changing true values (e.g. changes in residence). To demonstrate scalability, the model was applied to the Perinatal Registry of the Netherlands, linking firstborn (n=600000) and second-born (n=400000) children to the same mother.

Keywords: Gibbs Sampler; Blocking; Registries

An Adaptive Test for Differential Abundance in Microbiome Studies

Connie Musisi[1], Olivier Thas[1,2,3], Leyla Kodalci[1], Stijn Jaspers[1], Johniel Babiera[4]

[1] Data Science Institute, Hasselt University
[2] Department of Data Analysis and Mathematical Modelling, Ghent University
[3] National Institute for Applied Statistics Research Australia (NIASRA)
[4] Department of Mathematics, Mindanao State University, Philippines

E-mail for correspondence: connie.musisi@uhasselt.be

Abstract: Microbiome research is key to understanding human health and various ecosystems. Detecting taxa that are differentially abundant (DA) between conditions, such as healthy vs. diseased individuals, is critical in gaining insights into microbial dynamics. Many existing DA methods for microbiome data struggle with issues such as false discovery rate (FDR) control and compositionality. We present an innovative, data-driven approach that leverages order statistics to address these challenges.

Our proposed approach takes in three different steps. Firstly, we begin by constructing a pseudo-dataset from pairing taxa from one of the original groups to avoid bias when constructing test statistics. These are referred to as pseudo-taxa. The pseudo dataset mimics the structure of the original data. Using user-defined thresholds on the effect sizes, the pseudo-taxa are then subsetted into non-DA and DA, from which a training dataset is constructed. Secondly, we treat the DA problem as a prediction problem and build a linear prediction model with the order statistics of the relative abundances as predictors. Using linear (partial least squares) regression, the regression coefficients are estimated. Finally, the resulting linear model is considered to be the test statistic and can be applied to the original dataset. Permutation-based testing is then used to generate p-values, and these are adjusted using classical methods like Benjamini-Hochberg to control the FDR.

Extensive simulation studies demonstrate that our method achieves good FDR control and competitive sensitivity when compared with other differential abundance analysis methods, making it a promising, robust tool for microbiome research.

Signpost testing to navigate the parameter space of the Gaussian graphical model with high-dimensional data

Kai Ruan[1], Mark A. van de Wiel[1], Wessel N. van Wieringen[1,2]

[1] Department of Epidemiology and Data Science, Amsterdam Public Health Research Institute, Amsterdam UMC, location VUmc, Amsterdam, The Netherlands
[2] Department of Mathematics, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

E-mail for correspondence: k.ruan@amsterdamumc.nl

Abstract: We consider the learning of the precision matrix of a Gaussian graphical model for a pathway in the low prevalent oestrogen negative breast cancer subtype. Hereto we have, next to high-dimensional in-house gene expression data, external data of the more prevalent oestrogen positive subtype available. The latter data come, for privacy reasons, as a parameter estimate. We propose the signpost test to assess the external precision matrix estimate's relevance for the in-house estimation problem.
The signpost test considers a null value of the precision matrix that represents the absent/current knowledge of the parameter. The test takes the external precision matrix estimate as the alternative value of our parameter. We parameterise the line segment between the null and alternative value by a one-dimensional parameter. The test statistic optimises, within the class of ridge precision matrix estimators, the loss over this line segment. With the null distribution at hand, we evaluate the signpost test's power. It compares favourably to the likelihood ratio test.
We re-analyse breast cancer studies involving aforementioned subtypes. It shows that the signpost test statistic is an empirically valid metric for relevance of external information, reveals significance of the external information, and provides diagnostics of the test's results.

Keywords: Gaussian Graphical Model; Hypothesis Testing; Shrinkage

Van Wieringen, W. N., Peeters, C. F. (2016). Ridge estimation of inverse covariance matrices from high-dimensional data. Computational Statistics & Data Analysis, 103, 284-303.
Kai R., Van Wieringen, W. N. (2025). Signpost testing to navigate the parameter space of the Gaussian graphical model with high-dimensional data. submitted.

Posterior inference for Shapley values through Bayesian horseshoe estimation of tree-based prediction rule ensembles: Abstract

Giorgio Spadaccini[1,2], Marjolein Fokkema[2], Mark van de Wiel[1]

[1] Department of Epidemiology and Data Science, Amsterdam UMC
[2] Leiden University

E-mail for correspondence: g.spadaccini@amsterdamumc.nl

Abstract: In many prognostic settings, Machine Learning (ML) is gaining increasing popularity in hypothesis-free discovery of risk (or protective) factors and groups. ML is strong at discovering nonlinearities and interactions, but this power of ML is compromised by a lack of methods to reliably infer such effects on a local level. This is needed as the high complexity of both reality and ML models imply that the influence of risk factors strongly varies across subjects and their unique combination of features. While local measures of feature attributions can be combined with e.g. tree ensembles, uncertainty quantifications for these measures remain only partially available and oftentimes unsatisfactory. We propose RuleSHAP, a framework for using rule-based, hypothesis-free discovery that combines sparse Bayesian regression, tree ensembles and Shapley values to carry out a one-step procedure that both detects and tests complex patterns at the individual level. We compare our model with linear regression and with state-of-the-art tree ensemble models on simulated data. Moreover, we apply our machinery to data from an epidemiological cohort to detect and infer several effects for high cholesterol level and blood pressure. We illustrate Shapley values and their uncertainties for the most important features, allowing to infer what is relevant and for whom. From these experiments, we conclude that the combination of rule-based prediction, the horseshoe prior and our derived Shapley values provide a powerful and interpretable method that combines the flexibility of tree ensembles with statistical inference.

Curve registration for non-linear mixed effect ordinary differential equation models

Quentin Clairon[1], John Fricks[2], Mélanie Prague[1]

[1] SISTM team, Université de Bordeaux, Inria Bordeaux Sud-Ouest, Bordeaux, France
2 School of Mathematical and Statistical Sciences, Arizona State University, Tempe, USA

E-mail for correspondence: quentin.clairon@u-bordeaux.fr

Abstract: We tackle the curve registration problem of time-warping functions $\{hi\}_{i=1,..,n}$ learning from noisy observations of registered curves $\{Xi \circ hi\}_{i=1,..,n}$. Still, in our case a priori knowledge regarding the unregistered curve dynamics is available under the form of a parametric ordinary differential equations (ODE)s $\dot{X}_i = f(X_i, t)$. From this combination of descriptive nonparametric model and causal parametric one, we aim to locate as accurately and exhaustively as possible the effect of a given therapy on a treated population. From the causal representation, we quantify treatment effects on well identified mechanisms, specified as ODE parameter covariates. From the descriptive one, we infer global action of treatment due to other mechanisms missed by the ODE but accounted for by time-warping functions, leading to distorted dynamics for treated subjects compared to the control group. The joint estimation of $\{hi\}_{i=1,..,n}$ and ODE parameters is then cast as a non-linear regression problem in a mixed effect setting to account for inter-subject variability. We then confirm on simulated data the capacity of our method to estimate treatment effects on the general evolution of some variables of interests as well as on specific mechanisms acting on the patient dynamic. We conclude this work by analysing pre-clinical data trials testing HIV cures in non-human primates models.

Keywords: Curve Registration, Nonlinear Mixed Effect Models, Mechanistic Modelling, Clinical Trials.

Challenging Copulas with Smooth Density Decompositions

Paul Eilers[1]

[1] Erasmus University Medical Center, Rotterdam, The Netherlands

E-mail for correspondence: p.eilers@erasmusmc.nl

Abstract: Given observed two-dimensional data, we can estimate the logarithm of their density as a sum of three components: $u(x)$, $v(y)$ and $c(x,y)$, all three assumed to be smooth. Starting from a two-dimensional histogram with narrow bins, all components can be estimated with penalized Poisson regression. Modelling logarithms automatically guarantees positive density estimates that integrate to 1. To get unique results, $u(x)$ and $v(x)$ are estimated as the logarithms of the marginal densities, using P-splines. To estimate $c(x,y)$, tensor product P-splines are used, with $u(x)$ and $v(x)$ as offsets. Four penalty parameters are present: in addition to one for $u(x)$ and one for $v(y)$, two are needed for the x- and y-directions of $c(x,y)$. The values of all parameters are estimated using fast mixed model technology, as implemented in the R packages LMMsolver.
This model can be used as an alternative to copulas. Once $c(x,y)$ has been estimated, $exp(c(x,y))$ can be combined with new marginal densities for x and y to construct a two-dimensional density with the given interaction.
In contrast to copulas, no cumulative distributions are involved. That is an advantage for circular and directional data, for which they are not defined uniquely.
Interesting variations are possible. An example is modelling multiple densities with one shared interaction component $c(x)$ and individual $u(x)$ and $v(x)$. I will discuss technical details and present applications.

Keywords: P-Splines; Copula; Smoothing.

Bayesian functional principal component analysis for partitioning high-dimensional longitudinal data

Marion Kerioui[1], Daniel Temko[1], Shahin Tavakoli[2], Helene Ruffieux[1]

[1] MRC Biostatistics Unit, University of Cambridge, Cambridge, United Kingdom
[2] RISIS, GSEM, University of Geneva, Geneva, Switzerland

E-mail for correspondence: marion.kerioui@mrc-bsu.cam.ac.uk

Abstract: Longitudinal data are increasingly collected in many biomedical applications. When variables are frequently observed over time, one can use Multivariate Functional Principal Component Analysis (MFPCA) to disentangle different sources of variability while accounting for the variables covariance by sharing individual scores. However, when the number of variables becomes large, only subgroups of variables are expected to vary coordinately. Here, we combine the MFPCA framework with a mixture model whereby we simultaneously assign variables to different groups and estimate the MFPCA parameters within each group. Previous work showed good empirical properties of Bayesian variational algorithms to infer MFPCA model parameters within a reasonable timeframe. In simulations, we show good estimation accuracy of our proposed Partition FPCA (PFPCA) model compared to applying MFPCA separately to each group of variables, assuming that the true group structure is known. We apply the PFPCA to 36 biomarkers sampled over time in covid patients hospitalized in an academic French hospital between January and July 2020. Our approach allows us to uncover groups of variables driven by the same biological pathway.

Keywords: Longitudinal data, High-dimensional data, Bayesian inference, Mixture model, SARS-CoV-2

Happ C, and Greven S (2018) Multivariate Functional Principal Component Analysis for Data Observed on Different (Dimensional) Domains, *Journal of the American Statistical Association,*113(552), 649-659

Nolan T et al (2025) Efficient Bayesian functional principal component analysis of irregularly observed multivariate curves, *Computational Statistics and Data Analysis,* 203, 108094

Lavalley-Morelle A, et al (2023). Multivariate joint model under competing risks to predict death of hospitalized patients for SARS-CoV-2 infection, *Biometrical Journal,* 66(1), 2300049

Semi-nonnegative matrix factorization and variance modelling on near infrared spectroscopy PAT data to determine the blend uniformity endpoint

Tatsiana Khamiakova[1], Ana Tavares Da Silva[2], Adriaan Blommaert[3], Nicolas Sauwen[3]

[1] Statistics and Decision Sciences, Johnson and Johnson, Beerse, Belgium
[2] Therapeutics, Development and Supply, Johnson and Johnson, Beerse, Belgium
[3] Open Analytics, Antwerp, Belgium

E-mail for correspondence: tkhamiak@its.jnj.com

Abstract: Process analytical technology (PAT) is defined as a system for designing, analysing, and controlling manufacturing processes through timely measurements (i.e., during processing) of critical quality and performance attributes [1]. Regulatory agencies encourage the pharmaceutical industry to use PAT strategy for better manufacturing process understanding and control [2].
We focus on a PAT application using near-infrared (NIR) spectroscopy for blending monitoring and blending endpoint detection defined as the time at which the variation of the different blend components during the process reaches the steady state. To extract the information on the concentration of blend components from the first derivative NIR spectra, semi-nonnegative matrix factorization is applied. This is a multivariate blind source separation technique, where the components are initialized by known spectra of the blend components so that the extracted trends are directly related to the concentration of specific compounds in a blend. For the model-based determination of blending endpoint, two approaches are proposed: (1) the extracted blend trends are modelled by parametric additive model for variation to predict end variability of a trend of specific blend component and (2) generalized additive model for mean and variance. Final decision can be based either on a threshold or on the visual inspection of modelling outcome.

Keywords: Nonnegative Matrix Factorization; Endpoint Determination; Generalized Additive Models; Chemometrics

[1] Assessment of Recent Process Analytical Technology (PAT) Trends: A Multiauthor Review. (2015) Org. Process Res. Dev. 19, 3−62.
[2] Guidance for Industry PAT — A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance (2004). FDA guidance: Pharmaceutical CGMPs.

Nonparametric Dynamic Random Survival Forests for prediction of survival outcomes from time-varying predictors

Corentin Segalas[1], Robin Genuer[1], Cécile Proust-Lima[2]

[1] Univ. Bordeaux, INSERM, INRIA, BPH, U1219, Bordeaux, France
[2] Univ. Bordeaux, INSERM, BPH, U1219, Bordeaux, France

E-mail for correspondence: corentin.segalas@u-bordeaux.fr

Abstract: Predicting individual survival outcomes from medical history presents significant statistical challenges, due to the presence of time-dependent predictors measured irregularly and with error. Regression calibration models, landmark models and joint models are used but they respectively ignore informative truncation, fail to use all available data or becomes computationally intractable as the number of longitudinal predictors increases.
In this work, we combine functional principal component analysis (FPCA) and random survival forests (Iswharan et al., 2008) (RSF) to predict a survival outcome from longitudinal predictors. While RSF captures complex interactions between predictors, it does not accommodate longitudinal predictors. However, FPCA scores of their trajectories can be included as time-independent predictors. Implementations of FPCA for longitudinal data – sparse and irregular functional data – have been developed (Yao et al., 2005) and shown to be robust to missing at random data (Segalas et al., 2024). In the proposed model, inside each node, longitudinal predictors are summarized using their FPCA scores so that informative truncation of predictors by the event is taken into account as nodes become homogeneous. In a simulation study, the performance of the proposed model under various missing data scenarios is evaluated and compared to alternative models. The model was applied to pbc2 dataset to predict death using all longitudinal information.

Keywords: Dynamic Survival Prediction; Functional Principal Component Analysis; Longitudinal Data; Missing Data; Random Survival Forests.

Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS (2008) Random survival forests. Annals of Applied Statistics, 2(3), 841–860.

Segalas C, Helmer C, Genuer R, Proust-Lima C (2024). Functional Principal Component Analysis as an Alternative to Mixed-Effect Models for Describing Sparse Repeated Measures in Presence of Missing Data. Statistics in Medicine, 43, 4899–4912.

Yao F, Muller HG, Wang JL (2005). Functional Data Analysis for Sparse Longitudinal Data. Journal of the American Statistical Association, 100(470), 577–590.

Emulating a Target Clinical Trial in a Context of Multiple Outcomes and Missing Data: the example of the PROPENSLEEVE Study

Maïlis Amico[1,2], Bader Al Taweel[3], David Nocca[4], and Marie-Christine Picot[2]

[1] Desbrest Institute of Epidemiology and Public Health, Univ Montpellier, Inserm, Montpellier, France
[2] Clinical Research and Epidemiology Unit, CHU Montpellier, Univ Montpellier, Montpellier, France
[3] Digestive Surgery and Transplantation, CHU Montpellier, Univ Montpellier, Montpellier, France
[4] Digestive and Bariatric Surgery, CHU Montpellier, Univ Montpellier, Montpellier, France

E-mail for correspondence: mailis.amico@umontpellier.fr

Abstract: When it is not possible to perform a randomized clinical trial for ethical, feasible or timely reasons, target trial emulation (Hernán & Robins (2016)) provides a framework to draw conclusions using observational data. It consists in 1) defining the target trial, 2) selecting observational data that fit target trial characteristics, and 3) using causal inference methods to answer the question of interest. In the PROPENSLEEVE study, we aim to emulate a target trial to compare two surgeries for obese patients using cohort data. We plan to analyse two outcomes, excess weight loss and proportion of de novo gastroesophageal reflux, using two approaches, matching and inverse probability of treatment weighting based on propensity score (Rosenbaum & Rubin (1983)). We face, however, two issues: baseline characteristics used for propensity score computation have missing values, and we want to perform propensity score analysis for more than one outcome, which raises two questions : how can we handle multiple imputation with propensity score analysis, and do we need to use one or several propensity scores when more than one outcome are of interest? In this talk, we propose to present implementation of a target trial emulation and to answer these questions based on PROPENSLEEVE study example.

Keywords: Target Trial Emulation; Propensity Score; Multiple Imputation; Epidemiology; Causal Inference

Hernán MA, and Robins JM (2016). Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. American Journal of Epidemiology, 183(8), 758–764.

Rosenbaum PR, and Rubin DB (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika, 70, 41–55.

Parameter Estimation in Compartmental Epidemic Models with Heterogeneity in Susceptibility

Yuwen Ding[1], Jacco Wallinga[1,2], Hein Putter[1]

[1] Leiden University Medical Center, Leiden, the Netherlands
[2] National Institute for Public Health and the Environment, Bilthoven, the Netherlands

E-mail for correspondence: y.ding1@lumc.nl

Abstract: The susceptible-infectious-recovered (SIR) model is widely used in epidemic modelling but assumes a homogeneous population. To account for heterogeneity in susceptibility, we incorporate a frailty model by scaling the transmission parameter with individual random effects. In a completely observed epidemic scenario, we estimate the transmission parameter and frailty variance using three distributions from the power variance function family—gamma, inverse Gaussian, and compound Poisson with probability mass at zero. Epidemic outbreaks are simulated in R, and the observed data likelihood function is derived via the Laplace transform. We obtain the maximum likelihood estimates using the optim function, with a profile Expectation-Maximization (EM) algorithm as an alternative approach. Simulation results show consistently accurate estimates for both parameters across all frailty distributions, with low root mean squared error and confidence interval coverage close to the target 95% level. Estimates remain stable across various parameter combinations and sample sizes, each evaluated over 1,000 replications. The EM algorithm produces similar estimates but is computationally inefficient. Incorporating frailty distributions into the SIR model provides a more nuanced representation of epidemic dynamics. Our estimation approach demonstrates both efficiency and precision. Future work will extend the method to more realistic data mechanisms based on daily counts of new infections and to real-world epidemic scenarios.

Keywords: Epidemic Modelling; SIR Model; Frailty Models

Trial emulation to assess the effect of surgery on survival when there are competing risks, with application to patients with thoracic aortic aneurysms

James Murray[1], Caroline Chesang[1], Steve Large[2], Colin Bicknell[3], Carol Freeman[4], Ruth H. Keogh[1], Linda D. Sharples[1]

[1] Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, United Kingdom
[2] Department of Cardiac Surgery, Royal Papworth Hospital NHS Foundation Trust, Cambridge, United Kingdom
[3] Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, London, United Kingdom
[4] Papworth Trials Unit Collaboration, Royal Papworth Hospital NHS Foundation Trust, Cambridge, United Kingdom

E-mail for correspondence: james.murray@lshtm.ac.uk

Abstract: Randomised controlled trials (RCTs) are the gold standard for causal inference but are often infeasible for surgical interventions due to logistical and ethical complexities. Selection-for-treatment bias arises from surgeon and patient preferences, and immortal time bias occurs because patients must survive until surgery. Failure to account for these biases overestimates surgical benefits.
We use trial emulation and a cloning-censoring-weighting approach (CCW) to estimate the causal effect of aneurysm-repair surgery on (i) overall and (ii) aneurysm-related mortality in patients from the Effective Treatments for Thoracic Aortic Aneurysms (ETTAA) cohort. CCW artificially creates patient 'clones' labelled as surgery or no surgery within a 12-month grace period, and adjusts for informative censoring, thereby mitigating biases.
Estimated seven-year survival if all patients have surgery was 57.4% (95% CI: 47.3%, 67.4%) compared to 49.9% (44.0%, 55.0%) if no patients had surgery. The benefit was primarily due to reduced aneurysm-related deaths (risk difference -8.7%, 95% CI -14.0%, -3.9%) with no significant effect on other causes.
This study demonstrates trial emulation's feasibility for causal inference in the presence of competing risks. Findings suggest survival benefit from aneurysm-repair surgery to seven years post-enrolment to ETTAA. This approach targets causal effects of surgical interventions when RCTs are infeasible.

Keywords: Trial Emulation; Survival; Aortic Aneurysm; Surgery

Maringe C, Majano SB, Exarchakou A, *et al.* (2020). Reflection on modern methods: trial emulation in the presence of immortal-time bias. Assessing the benefit of major surgery for elderly lung cancer patients using observational data. International Journal of Epidemiology, 49, 1719–29.

Murray J, Chesang C, Large S, *et al.* (2025). Trial emulation to assess the effect of surgery on survival when there are competing risks, with application to patients with thoracic aortic aneurysms. Journal of Clinical Epidemiology, 181.

Forecast-Attribution reveals enhanced heat-mortality from climate change in British Columbia Heatwave

Chin Yang Shapland[1,2], Y. T. Eunice Lo[3,4], Nicholas J. Leach[5], Éric Lavigne[6,7], Kate Tilling[1,2] and Dann M. Mitchell[3,4,8]

[1] MRC Integrative Epidemiology Unit at the University of Bristol, U.K.
[2] Population Health Sciences at the University of Bristol, U.K.
[3] Cabot Institute for the Environment, University of Bristol, Bristol, UK.
[4] Elizabeth Blackwell Institute for Health Research, University of Bristol, Bristol, UK
[5] Atmospheric, Oceanic, and Planetary Physics, Department of Physics, University of Oxford, Oxford, UK·
[6] School of Epidemiology & Public Health, Faculty of Medicine, University of Ottawa, Ottawa, Canada.
[7] Environmental Health Science and Research Bureau, Health Canada, Ottawa, Canada.
[8] School of Geographical Sciences, University of Bristol, Bristol, UK.

E-mail for correspondence: chinyang.shapland@bristol.ac.uk

Abstract: In 2021, British Columbia (BC) experienced a record-breaking temperature of 49.6 °C and estimated that 619 heat-related deaths occurred in this event. Assessing the range of possible extreme heat and the resulting mortality in the current climate is critical to preparing our population and policy makers for unprecedented heat-health events. For the first time, we use a weather forecast model to attribute health impacts to climate change.

We fitted a distributed lag non-linear model to estimate the temperature-mortality relationship for the key health region in BC. We apply forecast-based climate simulations designed for extreme event attribution, which provide many realisations of heatwave events in current, future and pre-industrial climates.

We show that our method for climate health attribution can give different results compared to other methods, with implications for using such methods to examine events from different regions of the world. We show that under the "current" climate conditions, 30% more deaths were possible than were observed during the heatwave. We show that 15% of heat-related deaths during the observed heatwave are due to human induced carbon emission.

We argue that this novel method gives particularly reliable impact attribution results and is therefore strongly defensible in decision making and legal settings.

Keywords: Climate Change; Heat-Related Mortality; Attribution

Genome-wide association between gene expression and chromatin accessibility

Mathilde Bruguet[1], David Causeur[1], Nadia Ponts[2], Gaël Le Trionnaire[3]

[1] Department of Statistics and Computer Science UMR 6625, Institut Agro Rennes Angers/IRMAR CNRS, Rennes, France
[2] IGEPP, INRAE, Rennes, France
[3] MycSA, INRAE, Bordeaux, France

E-mail for correspondence: mathilde.bruguet@agrocampus-ouest.fr

Abstract: Plant pathogens must adapt to various environmental stresses to grow and survive. Epigenetic variations play a crucial role to shape short-term phenotypic responses to these stresses by modifying the gene regulatory network without altering the genomic DNA sequence. The filamentous fungus *Fusarium graminearum* is an example of a highly resistant plant pathogen, reproducing by cloning and capable of adapting rapidly to variations in environmental conditions, producing mycotoxins responsible for crop damage.
Among key epigenetic mechanisms regulating gene expression, variations in chromatin accessibility driven by changes in nucleosome positioning are particularly important. High-throughput sequencing technologies, such as MAINE-seq, are specifically designed to capture genome-wide epigenetic variations. However, comparative studies, employing a broad range of statistical learning methods including tree-based ensemble models and neural networks exhibit limited predictive performance, often only distinguishing between low and high gene expression levels differences. Our initial analyses confirm that both linear and non-linear whole-genome approaches, such as Random Forest and Neural Networks, fail to identify meaningful associations between chromatin accessibility signals and gene expression on *Fusarium g.* genome. To address this limitation, we propose a mixture model of penalized signal-to-scalar regressions, which uncovers specific chromatin accessibility patterns strongly linked to gene expression in large gene clusters. The extent to which this approach enhances whole-genome predictive performance will be discussed in the presentation.

Keywords: Deep-Learning; Mixture Modelling; Functional Data; Genomics; Large-Scale Prediction

Fused Estimation of Varying Omics Effects for Clinicogenomic Data

Jeroen M Goedhart[1], Mark A van de Wiel[1], Wessel N van Wieringen[1,2], Thomas Klausch[1]

[1] Department of Epidemiology and Data Science, Amsterdam University Medical Center, Amsterdam, the Netherlands
2 Department of Mathematics, Vrije Universiteit, Amsterdam, the Netherlands

E-mail for correspondence: j.m.goedhart@amsterdamumc.nl

Abstract: Cancer prognosis is often based on a set of omics covariates and a set of established clinical covariates such as age and tumor stage. Combining these two sets of covariates in a so called clinico-genomic model poses challenges. First, dimension: clinical covariates should be favored because they are low-dimensional and usually have stronger prognostic ability compared to high-dimensional omics covariates (De Bin et al., 2011). Second, interactions: genetic profiles and their prognostic effects may vary across patient subpopulations (Martinez-Jimenez et al., 2023). Last, redundancy: a (set of) gene(s) may encode similar prognostic information as a classical covariate (Ng et al., 2023). To address these challenges, we combine regression trees, employing clinical covariates only, with a unique fusion-like penalized regression framework in the leaf nodes for the omics covariates. The fusion penalty controls the modelled variability in genetic profiles across subpopulations defined by the tree. We prove that the shrinkage limit of our penalized framework equals a benchmark model: a ridge regression with penalized omics- and unpenalized clinical covariates. The proposed method also allows researchers to evaluate, for different patient subpopulations, whether the added overall omics effect enhances prognosis compared to only employing clinical covariates. We illustrate the strengths of the proposed method in simulations and in an application to colorectal cancer prognosis.

Keywords: Fused Estimation; Omics; High-Dimensional Data; Cancer Prognosis; Shrinkage

De Bin R., Sauerbrei W., and Boulesteix A. L. (2011). Investigating the prediction ability of survival models based on both clinical and omics data: two case studies. Statistics in Medicine, 33 (30), 5310-5329

Martınez-Jimenez F., Movasati A., Brunner S. R., Nguyen L., Priestley P., Cuppen E., and Van Hoeck A. (2023). Pan-cancer whole-genome comparison of primary and metastatic solid tumours. Nature, 618 (7964), 333–341

Ng H. M., Jiang B., and Wong K. Y. (2023). Penalized estimation of a class of single-index varying-coefficient models for integrative genomic analysis. Biometrical Journal, 65(1), 2100139

Review of multi-table integration methods for longitudinal multi-omics data

Madeline Vast[1], Laura Symul[1]

[1] Institute of Statistics, Biostatistics, and Actuarial Sciences, UCLouvain, Louvain-la-Neuve, Belgium

E-mail for correspondence: madeline.vast@uclouvain.be

Abstract: Data integration refers to the joint analysis of several multivariate tables. Unsupervised integration methods aim to identify latent structures that may be shared across tables or specific to a subset. While some methods have been proposed for multi-omics data which are often highly heterogeneous in type (count *vs.* continuous), distribution (heteroscedasticity, sparsity), and highdimensional (small n, large p), few methods are suitable for analysing longitudinal multi-omics datasets. Here, we review and compare DiSTATIS and MEFISTO for an application to longitudinal microbiome-related data. DiSTATIS is a method for the analysis of dissimilarity matrices, which is particularly interesting for multi-omics applications as dissimilarities between samples are often more directly comparable than the original distributions. MEFISTO proposes a functional version of a probabilistic factor analysis framework that takes into account known temporal or spatial dependencies between observed samples. Both methods are promising but have limitations. While DiSTATIS is fast and easy to interpret, it is not adapted to longitudinal data and cannot handle missing values. In contrast, MEFISTO is adapted to both, but is limited to a small number of distributions, has a long computation time, and the results are sensitive to imbalances in table size.

Keywords: Multi-Omics Integration; Longitudinal Integration; Unsupervised Integration; Microbiome

Abdi H, O'Toole AJ, Valentin D and Edelman B (2005). DISTATIS: The Analysis of Multiple Distance Matrices. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops, 3, 42–42.

Velten B, Braunger JM, Argelaguet R, Arnol D, Wirbel J, Bredikhin D, Zeller G and Stegle O (2022). Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO. Nature Methods, 19(2), Article 2.

Unraveling Breast Cancer Genetic Risk in Chinese Women: Integrating GWAS, Fine-Mapping, and Machine Learning in the China Kadoorie Biobank

Shizhe Xu[1,2], Christiana Kartsonaki[1,2], Kyriaki Michailidou[3], Kuang Lin[1]

[1] University of Oxford
[2] Nuffield Department of Population Health, University of Oxford
[3] Cyprus Institute of Neurology and Genetics

E-mail for correspondence: shizhe.xu@ndph.ox.ac.uk

Abstract:

Background / Introduction: Genome-wide association studies (GWAS) have identified approximately 200 genomic regions containing common genetic variants associated with breast cancer risk. However, their target genes remain uncertain mainly due to linkage disequilibrium (LD) and the prevalence of variants in non-coding regions. To address this, fine-mapping methods have been introduced to pinpoint the most likely causal variants from a set of credible candidate variants and identify target genes. Most previous GWAS and fine-mapping studies have primarily focused on European-ancestry individuals. Given differences in genetic architecture and environmental exposures between Asian and European populations, our study aims to conduct GWAS on Chinese women and perform fine-mapping with summary statistics to uncover additional association signals and candidate susceptibility genes for breast cancer. Furthermore, we integrate machine learning models into the fine-mapping process.
Methods: First, we performed a GWAS on 57,660 Chinese women from the China Kadoorie Biobank using two software packages, SAIGE and REGENIE. Second, to understand how these packages handle complex living regions in China, we analysed specific loci and explicitly compared their approaches to computing relatedness and performing association testing by a Firth logistic regression model or a linear mixed model. Third, to distinguish true causal variants from significant signals, we applied fine-mapping methods such as SuSiE-RSS and PolyFun to our generated summary statistics. Fourth, to investigate computational trade-offs, we conducted finemapping on individual-level data and summary statistics to evaluate loss in accuracy. Finally, we applied a sequence-based deep learning model to assign functional annotations to variants in noncoding regions and incorporated a supervised learning approach, such as random forest.
Results: Summary statistics, Manhattan plots, QQ plots and LD score regression files were generated. Among Chinese women, several genetic loci associated with breast cancer were identified. The signals detected by SAIGE were slightly more significant than those identified by REGENIE due to differences in their underlying algorithms and correction thresholds. A comparison was conducted between fine-mapping results derived from individual-level data and summary statistics. A systematic evaluation was conducted to assess whether functional annotations and supervised learning enhance fine-mapping accuracy.
Conclusion: This study provides new insights into breast cancer genetics based on data from Chinese women. The comparison between REGENIE and SAIGE enhances our understanding of their strengths and limitations. This study also visualises the differences between fine-mapping using individual-level data and summary statistics. Finally, a machine learning-based framework in fine-mapping paves the way for more explicit analysis.

Joint modelling of longitudinal HRQoL data accounting for the risk of competing dropouts

Hortense Doms[1], Philippe Lambert[1,2], Catherine Legrand[1]

[1] LIDAM/ISBA, UCLouvain, Louvain-la-Neuve, Belgium
[2] Institut de Mathématiques, Université de Liège, Belgium

E-mail for correspondence: hortense.doms@uclouvain.be

Abstract: In cancer clinical trials, health-related quality of life (HRQoL) is an important endpoint, providing information about patients' well-being and daily functioning. However, missing data due to premature dropout can lead to biased estimates, especially when dropouts are informative. We introduce the extJMIRT approach, a novel tool that efficiently analyses multiple longitudinal ordinal categorical data while addressing informative dropout. Within a joint modelling framework, this approach connects a latent variable, derived from HRQoL data to cause-specific hazards of dropout. Unlike traditional joint models, which treat longitudinal data as a covariate in the survival submodel, our approach prioritizes the longitudinal data and incorporates the log baseline dropout risks as covariates in the latent process. This leads to a more accurate analysis of longitudinal data, accounting for potential effects of dropout risks. Through extensive simulation studies, we demonstrate that extJMIRT provides robust and unbiased parameter estimates and highlight the importance of accounting for informative dropout. We also apply this methodology to HRQoL data from patients with progressive glioblastoma, showcasing its practical utility.

Keywords: Bayesian Joint Models; Informative Dropout; Item Response Theory; Quality of Life

Early-outcome-based Interim Decisions Regarding Treatment Effect on a Long-term Endpoint

Leandro Garcia Barrado[1], Tomasz Burzykowski[1,2]

[1] International Drug Development Institute, Louvain-la-Neuve, Belgium
[2] Hasselt University, Hasselt, Belgium

E-mail for correspondence: tomasz.burzykowski@uhasselt.be

Abstract: In clinical trials that use a long-term efficacy endpoint T, the follow-up time necessary to observe T may be substantial. In such trials, an attractive option is to consider an interim analysis based solely on an early outcome S that could be used to expedite the evaluation of treatment's efficacy. Garcia Barrado and Burzykowski (2024) developed a methodology that allows introducing such an early interim analysis for S and T of any type. It appears that such a design may offer substantial gains in terms of the expected trial duration and sample size. A prerequisite, though, is that the treatment effect on S has to be strongly correlated with the treatment effect on T, i.e., that S is a good trial-level surrogate for T. Garcia Barrado and Burzykowski (2024) assumed that the coefficients defining the (trial-level) model used to evaluate properties of S as a surrogate for T were known. In practice, only estimates of the coefficients would be available. In the current manuscript, we address this issue. The obtained results allow designing trials with an interim analysis based only on S, while properly adjusting for the estimation of the model capturing properties of S as a surrogate for T.

Keywords: Early Outcome; Long-Term Endpoint; Interim Analysis; Randomized Clinical Trial

Garcia Barrado L, and Burzykowski T (2024). Using an early outcome as a sole source of information of interim decisions regarding treatment effect on a long-term endpoint: the non-Gaussian case. Pharmaceutical Statistics, 23(6), 928–938.

A Bayesian prevalence-incidence mixture model for screening outcomes with misclassification

Thomas Klausch[1], Birgit Lissenberg-Witte[1], Veerle Coupé[1]

[1] Amsterdam University Medical Center, Department of Epidemiology and Data Science, Amsterdam, The Netherlands

E-mail for correspondence: t.klausch@amsterdamumc.nl

Abstract: We present BayesPIM (Klausch et al., 2024), a Bayesian prevalence-incidence mixture model for estimating time- and covariate-dependent disease incidence from screening and surveillance data with accompanying R package (Klausch, 2025). The method is particularly suited to settings where some individuals may have the disease at baseline, baseline tests may be missing or incomplete, and the screening test has imperfect test sensitivity. This setting is present in data from high-risk colorectal cancer (CRC) surveillance through colonoscopy, where adenomas, precursors of CRC, are already present at baseline and remain undetected due to imperfect test sensitivity. By including covariates, the model can quantify heterogeneity in disease risk, thereby informing personalized screening strategies. Internally, BayesPIM uses a Metropolis-within-Gibbs sampler with data augmentation and weakly informative priors on the incidence and prevalence model parameters. In simulations based on the real-world CRC surveillance data, we show that BayesPIM estimates model parameters without bias while handling latent prevalence and imperfect test sensitivity. However, informative priors on the test sensitivity are needed to stabilize estimation and mitigate non-convergence issues. We also show how conditioning incidence and prevalence estimates on covariates explains heterogeneity in adenoma risk and how model fit is assessed using information criteria and a non-parametric estimator.

Keywords: Bayesian; Survival; Screening; Misclassification; Latent-Class

Klausch T, Lissenberg-Witte BI, Coupé VMH (2024). A Bayesian prevalence-incidence mixture model for screening outcomes with misclassification. arXiv:2412.16065.

Klausch T (2025). BayesPIM: Bayesian prevalence-incidence mixture model. github.com/thomasklausch2/BayesPIM

Penalized Bayesian Methods for Product Ranking Using Both Positive and Negative References

Clément Laloux[1], Bruno Boulanger[1], Philippe Bastien[2], Bradley P. Carlin[3], Arnaud Monseur[1], Carole Guillou[2], Daiane Garcia Mercurio[2], and Hussein Jouni[2]

[1] Cencora-PharmaLex Belgium, 5 Rue Edouard Belin, 1435 Mont-Saint-Guibert, Belgium
[2] L'Oréal Research & Innovation, 1 Av. Eugène Schueller, 93600 Aulnay-sous-Bois, France
[3] PhaseV Trials, Inc., 200 Portland St., Boston, MA, 02114, USA

E-mail for correspondence: Clement.Laloux@pharmalex.com

Abstract: Product ranking according to pre-specified criteria is essential for developing new technologies, allowing identification of more preferable candidates for further development. Such ranking often builds on the results of a network meta-analysis, where the relative or absolute performances of the various products are synthesized across multiple clinical studies, each of which considered only a subset of the products. Ranking involving both a negative and a positive reference enables the scientist to directly compare tested products against known benchmarks (Cipriani et al., 2013). Here, more preferable candidates are those products that approach the positive reference while remaining distant from the negative reference. We provide a new metric to quantify this multivariate distance following Bayesian meta-analysis. Our method does not simply rely on point estimates to perform the comparisons, but also accounts for their uncertainties via their posterior distributions. For each product, posterior probabilities of being comparable to the positive reference are computed, and subsequently penalized by the posterior probability of performing worse than the negative reference. Each product is then compared to a hypothetical product about which we have no knowledge, as captured by a uniform distribution. The result is a prospective metric that is directly interpretable as the improvement of any product beyond this state of ignorance.

Keywords: Bayesian Statistics; External References; Meta-Analysis; Product Ranking

Cipriani A, Higgins JP, Geddes JR, and Salanti G (2013). Conceptual and technical challenges in network meta-analysis. Ann Intern Med, 159(2), 130-137.

Errors-in-Variables Bayesian Model of Glycemic Response to Lifestyle Factors

Guillaume Deside[1], Laura Symul[1]

[1] ISBA - Institut de statistique, biostatistique et sciences actuarielles, UCLouvain LIDAM/ISBA, UCLouvain, Louvain-la-Neuve, Belgium

E-mail for correspondence: guillaume.deside@uclouvain.be

Abstract: Modelling glycemic responses to nutriment intakes, lifestyle, and physiological factors using *in situ* data presents significant challenges due to noisy sensor measurements, irregular sampling frequencies, and incomplete or imprecise self-reported records (Salathe 2024). Our approach builds on the work of Zhang et al. (2021), which highlighted the importance of explicitly modelling measurement errors in input variables. We extend it beyond meal timing to incorporate a broad set of lifestyle factors such as physical activity and sleep. By jointly modelling the glucose temporal variations and the uncertainty in the measurement of lifestyle factors, our model aims to provide biologically and medically interpretable insights into the contribution of diverse behavioral and physiological inputs. We specifically address inter- and intra-person heterogeneity while incorporating prior information about glycemic responses. We apply our model to longitudinal data from studies such as the Food&You cohort (Héritier et al. 2024) to study inter-individual differences in glycemic regulation. Our method will facilitate robust hypothesis testing on glycemic regulation, ultimately advancing personalized medicine with a tool that can be deployed for clinical research and personal health management.

Keywords: Bayesian Errors-In-Variables; Continuous Blood Glucose Monitoring (CGM); Glycemic Response Modelling; Measurement Error Modelling; Longitudinal Data Analysis

Zhang G, Ashrafi RA, Juuti A, Pietiläinen K, and Marttinen P (2021). Errors-in-Variables Modeling of Personalized Treatment-Response Trajectories. IEEE Journal of Biomedical and Health Informatics, 25(1), 201–208.

Héritier H, et al. (2024). Food & You: A digital cohort on personalized nutrition. PLOS Digital Health, 2, e0000389.

Paranjape K, Schinkel M, and Nanayakkara P (2020). Short Keynote Paper: Mainstreaming Personalized Healthcare–Transforming Healthcare Through New Era of Artificial Intelligence. IEEE Journal of Biomedical and Health Informatics, 24, 1860–1863.

Salathé M, Toumi M, and Singh R (2024). Personalized glucose prediction using in situ data only. Preprint, April 2024.

Forest Structure Alters Biomass-Backscatter Relationship

Lennart Hoheisel[1], Nicole Augustin[1], Casey Ryan[2]

[1] University of Edinburgh, School of Mathematics, Department of Statistics
[2] University of Edinburgh, School of Geosciences

E-mail for correspondence: L.J.Hoheisel@sms.ed.ac.uk

Abstract: Above-ground biomass (AGB), an essential climate variable, has been identified as a critical input to the United Nations' Reducing Emissions from Deforestation and Forest Degradation-plus (REDD+) program (Duncanson et al. (2021)). Estimating and mapping AGB using satellite aperture radar (SAR) backscatter has gained popularity with increasing data availability. Although many of the currently available map products utilize a similar model structure, discrepancies can be observed at individual grid points (Ploton et al. (2020)). In addition to overlooking spatial autocorrelation (Ploton et al. (2020)), signal saturation (especially for L-band or shorter wavelength instruments) and confounding variables - such as soil, forest, and tree characteristics - could explain these differences (Woodhouse et al. (2012)). To address these issues, we define a Generalised Additive Model (GAM, Wood (2017)) that performs parameter estimation through neighbourhood cross-validation (NCV, Wood (2024)) and treats correlation structures through random effects. Using a large dataset of field measurements and ALOS-1 and ALOS-2 PALSAR data observations, we find that the differences in the shape of the backscatter-AGB relationship at the observation level could be explained through vegetation structure and composition and topographic characteristics. Additionally, we provide a model trained exclusively on geospatial product data that achieves results comparable to the aforementioned model, which could help improve global wall-to-wall prediction of biomass.

Duncanson, L., Armston, J., Disney, M., Avitabile, V., Barbier, N., Calders, K., Carter, S., Chave, J., Herold, M., MacBean, N., et al. (2021). Aboveground woody biomass product validation good practices protocol.

Ploton, P., Mortier, F., Réjou-M´echain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., et al. (2020). Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nature communications, 11(1):4540.

Wood, S. N. (2017). Generalized additive models: an introduction with R. CRC press.

Wood, S. N. (2024). On neighbourhood cross validation. arXiv preprint arXiv:2404.16490.

Woodhouse, I. H., Mitchard, E. T. A., Brolly, M., Maniatis, D., and Ryan, C. M. (2012). Radar backscatter is not a 'direct measure' of forest biomass. Nature climate change, 2(8):556–557.

Matched pairs with rare event outcomes

Christiana Kartsonaki[1]

[1] Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom

E-mail for correspondence: christiana.kartsonaki@dph.ox.ac.uk

Abstract: We consider studies in which the same broad context, say sets of hospitals, is used to compare the rate at which a rare event, such as diagnosis of a rare disease, occurs under different conditions, for example the availability of a particular diagnostic test, which are present during different time periods. The patient populations are broadly similar, but the individual patients studied are distinct in different time periods. We are interested in the effect of the different conditions on the rate of occurrence of the event of interest. We assume that the number of events in each group has a Poisson distribution. The random variables (Y, Z) associated with a specific hospital are assumed to have a common value of an underlying rate μ multiplied by a factor representing the availability of the test. The nuisance parameters μ are eliminated from the likelihood by conditioning on the total number of events to yield a log likelihood that can be maximised to estimate the parameters representing the effect of the condition on the outcome.

Acknowledgement: I would like to thank David R. Cox for his help with this work.

Keywords: Matched Pairs; Rare Events; Poisson Distribution

Multi-State models for Breast Cancer: A Bayesian Non-parametric Approach

Valeria Leiva-Yamaguchi[1], Inés M. Varas[2], Fernando Quintana[2], Oscar M. Rueda[1]

[1] MRC Biostatistics Unit, University of Cambridge, Cambridge, UK
[2] Department of Mathematics, Pontificia Universidad Católica de Chile, Chile

E-mail for correspondence: christiana.kartsonaki@dph.ox.ac.uk

Abstract: We propose a Dependent Dirichlet Process (DDP) model for multistate survival analysis, extending the DDP ANOVA framework to accommodate both continuous and categorical covariates while incorporating random effects and censored data. A key advantage of our approach is its flexibility as it does not impose a constant hazard function over time. To illustrate our model, we apply it to breast cancer data from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort. We determine the survival transition probabilities in different stages of the disease, such as locoregional recurrence, distant recurrence, death from breast cancer and death from other causes. Our analysis integrates key clinical variables measured at diagnosis, such as age, tumor grade, tumor size, number of tumor-positive lymph nodes, and estrogen receptor status (negative / positive), allowing the visualization of the risks of disease progression between subgroups of patients and the quantification of uncertainty in individual predictions. We also compare our results to the multi-state Cox model using calibration and discrimination metrics. This shows that our BNP-based approach is more adaptable to capture complex survival dynamics.

Keywords: Censoring; Dependent Dirichlet process; Markov Chain Monte Carlo.

Andersen, P. K., Hansen, L. S., and Keiding, N. (1991). Non-and semi-parametric estimation of transition probabilities from censored observation of a non-homogeneous Markov process. Scandinavian Journal of Statistics 18, 153-167.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. The annals of statistics 1, 209-230.

Blackwell, D. and J. B. MacQueen (1973). Ferguson distributions via Polya urn schemes. The Annals of Statistics 1(2), 353-355.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. Statistica Sinica 4, 639-650.

MacEachern, S. (1999). Dependent nonparametric processes. In ASA Proceedings of the Section on Bayesian Statistical Science, pp. 50-55.

De Iorio, M., P. Müller, G. Rosner, and S. MacEachern (2004). An ANOVA model for dependent random measures. Journal of American Statistical Association 99, 205-215.

Semiparametric Bayesian latent trajectory models. Technical report, ISDS Discussion paper 16, Duke University, NC, USA.

De la Cruz, R., A. A. Quintana, and P. Müller (2007). Semiparametric Bayesian classification with longitudinal markers. Applied Statistics 56(2), 119-137.

Rueda OM, Sammut SJ, Seoane JA, et al. (2019) Dynamics of breast-cancer relapse reveal late-recurring ER-positive genomic subgroups. Nature; 567(7748): 399-404.

Menopause and the vaginal microbiome: an Isala cohort study

Kato Michiels[1,2], Thies Gehrmann[2], Sarah Ahannach[2], Stijn Wittouck[2], Sandra Condori[2], Camille Allonsius[2], Olivier Thas[1,3], Sarah Lebeer[2]

[1] Data Science Institute, Hasselt University, Hasselt, Belgium
[2] Lab of Applied Microbiology and Biotechnology, Department of Bioscience Engineering, Antwerp University, Antwerp, Belgium
[3] Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium

E-mail for correspondence: kato.michiels@uhasselt.be

Abstract: All women experience menopause, typically in their fifties, and the vaginal microbiome may play a key role in this transition. The Isala citizen science project, involving over 3000 women, aims to better understand the vaginal microbiome during menopause and the postmenopausal period. A subset of 159 peri/postmenopausal women was analysed to explore how menopause affects the vaginal microbiome. Confounders such as age, parity, smoking status, and hormonal treatments were identified using a causal Directed Acyclic Graph (cDAG) and included in analyses.
The study investigates how menopause alters bacterial composition and interactions in the vaginal microbiome, identifying microbial biomarkers and predictors of menopausal status. Differential abundance analyses revealed differences between (post)menopausal, perimenopausal, and fertile women. The modulation of the microbiome community correlation structure was studied by defining modules for all menopausal participants and comparing these to the findings of the main Isala study. Machine learning and structural equation modelling were used to identify taxa as biomarkers for perimenopause prediction. Based on the proposed cDAG, we adapt a non-linear mediation method proposed by Wen Wei Loh et. al. (2020) for vaginal microbiome data, which handles high-dimensional data with unknown causal structure.
Preliminary results showed a decrease in Lactobacillus species and an increase in Peptococcus, Atopobium, and Aerococcus in menopausal women. Overall, the effect of menopause on the microbiome is shown to be either limited, or highly personal.

Keywords: Microbiome Data

Validation of PREdiction of DELIRium in ICu patients (PRE-DELIRIC) model for ICU delirium in general ICU and patients with liver disease: A retrospective cohort study

Areti Papadopoulou[1], Sarah L. Cowan[2], Jacobus Preller[2,3], Robert J.B. Goudie[1]

[1] MRC Biostatistics Unit, School of Clinical Medicine, University of Cambridge, Cambridge, CB2 0SR, UK.
[2] Addenbrooke's Hospital, Cambridge, CB2 0QQ, UK.
[3] Department of Medicine, University of Cambridge, Cambridge, CB2 0QQ, UK.

E-mail for correspondence: areti.papadopoulou@mrc-bsu.cam.ac.uk

Abstract: Delirium is a common complication among ICU patients. Several models have been developed aiming to identify patients at high risk of delirium. PRE-DELIRIC (PREdiction of DELIRium in ICu), has been externally validated in several studies, but its performance remains unclear, particularly in patients with liver disease.

This study evaluated the PRE-DELIRIC using detailed electronic health records from 3,312 ICU patients at Cambridge University Hospital. Delirium was defined as either a positive Confusion Assessment Method or any administration of antipsychotic medication. We also conducted subgroup analyses in patients with liver disease, sedated patients, and across varying opioid dosing.

Delirium occurred in 32.9% of patients. Overall, PRE-DELIRIC demonstrated moderate discriminative performance (AUROC 0.74; 95% C.I. 0.72–0.76); but the model was poorly calibrated. Among liver disease patients, discrimination was similar to the overall cohort. Discrimination was significantly poorer in both sedated patients and patients receiving high opioid dosing.

To date, this is the largest validation study of the PRE-DELIRIC model, showing moderate discriminative predictive performance both overall and in liver disease patients. However, calibration was only moderate overall, and significantly under-predicted risk in patients with liver disease. Recalibration of the model and further subgroup-specific adjustments may enhance its utility in clinical practice.

Keywords: Delirium; PRE-DELIRIC; Validation; Liver Disease; Opioids

American Psychiatric Association, D. and A.P. Association, Diagnostic and statistical manual of mental disorders: DSM-5. Vol. 5. 2013: American psychiatric association Washington, DC.

van den Boogaard, M., et al., (2014) Recalibration of the delirium prediction model for ICU patients (PRE-DELIRIC): a multinational observational study. Intensive Care Med, 40(3): p. 361-9.

Amerongen, H.V.N., et al., (2023) Comparison of Prognostic Accuracy of 3 Delirium Prediction Models. Am J Crit Care, 32(1): p. 43-50.

Prediction with Logistic Regression in Binary Class Imbalance: Comparison of Re-sampling Techniques with Proper Threshold Probability Assignment under Varying Predictive Power of Covariates

Henk van der Pol[1,2], Ragnhild Sorum Falk[3], Marta Fiocco[2,4,5], Arnoldo Frigessi[6], Euloge Clovis Kenne Pagui[3,6]

[1] Department of Medical Onocology, Leiden University Medical Centre, Leiden, the Netherlands
[2] Mathematical Institute, Leiden University, Leiden the Netherlands
[3] Oslo Centre for Biostatistics and Epidemiology, University Hospital Oslo, Oslo, Norway
[4] Department of Biomedical Data, Leiden University Medical Center, Leiden, the Netherlands
[5] Princess Maxima Center for Pediatric Oncology, Utrecht, the Netherlands
[6] Oslo Centre for Biostatistics and Epidemiology, University of Oslo, Oslo, Norway.

E-mail for correspondence: h.van_der_pol@lumc.nl

Abstract: Prediction of rare events in binary classification is a well-studied topic in biostatistics and has gained renewed interest in the context of Machine Learning. Class Imbalance is addressed through resampling methods such as random over and undersampling, and synthetic minority over-sampling technique. However, recent studies suggest that these may not be necessary and may even harm clinical prediction Carriero et al. (2025), Goorbergh et al. (2022), Piccininni et al. (2024). This study extends current research, whether proper threshold assignment may overcome problems related to lass imbalance. Importantly, we explore how the strength of covariates and correctly specifying the model impact the prediction performance. Building on previous research, we performed a Monte Carlo simulation study, based on a logistic regression model with various settings: prevalence of disease, resampling techniques, model specification, and strength of covariate. Performance is assessed by precision and recall measures. We compare resampling techniques with appropriate threshold selection. In all simulation scenarios, proper threshold selection has comparable or better performance, compared to resampling techniques. Specifically, the threshold that maximizes the area under the ROC curve returns the best overall performance. Lastly, we reinforce current findings in which we show that avoiding resampling the dataset returns better performance in imbalanced data.

Keywords: Class Imbalance; Resampling Techniques; Threshold Selection; Logistic Regression Model; Simulation

Carriero A, Luijken K, de Hond A, Moons K.G.M., Van Calster B (2025). The harms of class imbalance corrections for machine learning based prediction models: A simulation study. Statistics in Medicine, 44(3-4):e10320.

van der Goorbergh R, van Smeden M, Timmerman D, Van Calster B (2022). The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. Journal of the American Medical Informatics Association, 29(9):1522–1531.

Piccininni M, Wechsung M, Van Calster B, Rohmann J.L., Konigorski S, van Smeden M (2024). Understanding random resampling techniques for class imbalance correction and their consequences on calibration and discrimination of clinical risk prediction models. Journal of Biomedical Informatics, 104666.

Hierarchical Bayesian model to predict CPAP usage in sleep apnea syndrome patients

Celia Vidal[1,2], Nicolas Molinari[1,2,3], Christophe Abraham[1,4], Dany Jaffuel[2,5,6]

[1] IDESP, INSERM, Montpellier University, Montpellier, France
[2] Groupe Adène, Montpellier, France
[3] PreMedical INRIA, Montpellier University Hospital, Montpellier, France
[4] Institut Agro Montpellier — SUPAGRO, Department of Sciences for Agro-Bio Processes, Montpellier, France
[5] Department of Respiratory Diseases, Montpellier University Hospital, Arnaud de Villeneuve Hospital, Montpellier, France
[6] PhyMedExp (INSERM U 1046, CNRS UMR9214), Montpellier University, Montpellier, France

E-mail for correspondence: c.vidal@groupe-adene.com

Abstract: Although significant progress has been made in the development of alternative therapies, continuous positive airway pressure (CPAP) remains the main treatment for sleep apnea syndrome. Patients use CPAP with a mask (three different types) placed over the face. The initial choice of mask type is important because it can influence mask acceptance, adverse effects, and CPAP treatment adherence. The objective of this research is to predict the best type of mask for each new patient in order to increase CPAP treatment adherence (Genta PR et al. (2020)). For each patient, we have longitudinal data, CPAP usage (number of hours) and mask type used daily. To predict mean CPAP usage and mean CPAP usage by mask type for each patient, we use a two-level hierarchical Bayesian model incorporating clinical and demographic variables for patients with sleep apnea syndrome. At both levels, we describe the patient's mean and mask type CPAP usage according to a Gaussian distribution. Variances of Gaussian distributions have been generated by Inverse Gamma distributions. We will also use a Dirichlet process to cluster patients according to their mean CPAP usage (Dahl D.B. (2003)).

Keywords: Hierarchical Bayesian Model; Longitudinal Data; CPAP; Sleep Apnea Syndrome

Dahl D.B. (2003). An improved merge-split sampler for conjugate dirichlet process mixture models. Technical report, 1086.

Genta PR et al. (2020). The Importance of Mask Selection on Continuous Positive Airway Pressure Outcomes for Obstructive Sleep Apnea. An Official American Thoracic Society Workshop Report. Ann Am Thorac Soc, 17(10), 1177-1185.

Index

Index