
8th Channel Network Conference
April 7-9, 2021
Organization: Conservatoire National des Arts et Métiers (CNAM)
FULLY ONLINE CONFERENCE
PARIS
INTERNATIONAL BIOMETRIC SOCIETY
SFdS

Contents

1. Welcome messages

By the scientific committee

Dear virtual attendee of the 8th IBS Channel Network Conference,

When preparations started in April 2020, we had faith that, come April 2021, we would surely be able to have a conference on-site in Paris. Alas, the reality of the global COVID-19 pandemic has proven unpredictable.

Notwithstanding the additional challenges, the local organizing committee as well as the members of the scientific committee have proven very flexible. Working in tandem they have provided a wonderful virtual infrastructure to support the inspiring programme. We extend our sincere thanks and admiration for all their efforts.

We are proud to have a broad and exciting offering of short courses, posters and contributed talks of excellent quality, invited speakers at the current frontiers of Biostatistics and BioData Science, and distinguished keynote lecturers. The quality and depth of these offerings will surely make up for yet some more screen time and the fact that we have to forego Paris.

We hope you enjoy the conference. We also express the hope to see you soon without the need of any screen.

Stay healthy,
Jelle Goeman & Carel F.W. Peeters

By the Local Organizing Committee

Dear attendee of the 8th IBS Channel Network Conference

We had hoped to welcome you to our wonderful city of Paris, always at its best in Springtime. Alas, that has not been possible. So, in addition to providing the virtual infrastructure for the conference, we have sought to bring at least a sense of Paris to you – through some images and sounds of our bustling, busy, enthralling, beautiful, and at times – it has to be said – maddening city. So please visit our Virtual Tour – and allow your imagination to do the rest. Paris, after all, is a state of mind: un état d'esprit.

There is much to be downcast about in this pandemic year, above all owing to the immense toll of human lives and livelihoods. But there are also positives, which this conference's virtual location in Paris encapsulates. We are a community, linked by ties of solidarity and friendship as well as profession. We are fully engaged in seeking the solutions to this pandemic and to addressing the inequalities it has laid bare. And we will, soon, retrieve the freedoms to meet in person as well as in spirit.

We hope you enjoy the conference and we look forward to welcoming you back to Paris in person.

Mounia N. Hocine & David Causeur

2. Foreword

The **Channel Network Conference** is a biennial conference organized by the International Biometric Conference (IBS) regions; Belgium, France, Great-Britain/Ireland and the Netherlands. This conference aims at gathering statisticians to discuss the newest statistical methodology for the analysis of biological and medical data. It is a 3-day conference with short courses, invited and contributed sessions. Usually, around 150 researchers from both academia and industry participate in the conference. In 2021, the French region will host the conference, supported by the French Statistical Association (SFdS).

Due to the global COVID-19 pandemic, the local organizing committee has decided to offer a fully virtual CNC21 conference. By taking this decision early, our wish is to offer a convenient and robust online video-conferencing sytem to ensure the best audio and video quality for talks and short courses and to favor stimulating discussions among participants and between particpants and speakers.

We are proud to announce that **Jeanine Houwing-Duistermaat** (University of Leeds, School of Mathematics, Statistics Institute) and **Mathias Drton** (Technical University of Munich, Department of Mathematics, see here for a short biography) will be the keynote speakers. In addition we will have three Invited Sessions:

- → Integrating and analyzing data from different sources (Data Integration)
- → Infectious diseases
- → Statistical modeling in movement ecology

Participants will additionally be able to choose one of three short courses:

- → Practical deep learning with R (S. Keydana, Rstudio)

- → Adaptive group-regularization in prediction (M. van Nee, M. Münch and M. van de Wiel, Amsterdam University medical centers, Department of Epidemiology and Data Science)
- → Non-parametric Bayesian methods for classification (B. Hejblum and A. Rouanet, University of Bordeaux)

As for all IBS conferences, contributed abstracts across the wide range of biological and biomedical application areas pursued by society members, for both oral and poster presentations demonstrate the diverse array of methodological topics required to address the challenges that these application areas bring.

Prizes will be awarded to best student oral presentations and also best poster presentations. These will be presented during the closing ceremony.

Registration provides an unlimited online access to all presentations and post-conference access to video recordings of the talks.

3. Committees

Scientific Committee

- Carel Peeters, Chair, Wageningen University and Research, The Netherlands
- Jelle Goeman, Co-chair, Leiden University, The Netherlands
- Nicole Augustin, University of Edinburgh, United Kingdom
- Anestis Touloumis, University of Brighton, United Kingdom
- Beatrijs Moerkerke, Ghent University, Belgium
- Olivier Thas, Ghent University, Belgium
- Cécile Proust-Lima, University of Bordeaux, INSERM BPH, Bordeaux, France
- David Causeur, Institut Agro, Rennes, France

Organization committee

- Mounia N. Hocine, Chair, Conservatoire National des Arts et Métiers (CNAM), Paris
- David Causeur, Co-chair, Institut Agro, Rennes, France
- Pascal Wild, French Research and Safety Institute (INRS), Nancy, France
- Cécile Proust-Lima, University of Bordeaux, INSERM BPH, Bordeaux, France
- Sophie Ancelet, Institut de Radioprotection et de Sûreté Nucléaire, Paris
- Pascale Tubert-Bitter, High-Dimensional Biostatistics for Drug Safety and Genomics, Inserm, CESP, Villejuif, France
- Robert Faivre, INRAe MIA, Toulouse, France
- Hélène Jacqmin-Gadda, University of Bordeaux, INSERM BPH, Bordeaux, France
- Boris Hejblum, University of Bordeaux, INSERM BPH, Bordeaux, France

4. Information for speaker

4.1 Oral Presentations

A full schedule of all contributed presentations is provided in the next Sections (also available through the website: https://cnc21.sciencesconf.org)

All contributed oral presentations are 15 mins + 3 mins for questions.

Students are eligible to apply to be considered for the **Best Student Oral Presentation**. To apply you need to have registered as a student delegate by the early-bird registration deadline date 7th March 2021 and also have submitted a final draft of your presentation (PowerPoint or PDF) by 31st March 2021 to cnc21@cnam.fr.

Speakers should ensure they upload their presentation before the start of their session.

4.2 Posters

The poster session will start on Wednesday 7th April 2021 with 3 minute flash talks. To aid with these flash talks, it is requested that a pdf of the poster is sent to cnc21@cnam.fr by 5th April. The flash talks will be followed by an open poster session.

All posters are eligible for the **Best Poster Presentation** which will be judged by the scientific programme committee.

All contributed poster presenters should register by 7th March 2021 "early bird" deadline.

5. Timetable

Wednesday, April, 7

Time			
09:00 - 12:30	**Course 1**: Practical deep learning in R	**Course 2**: Co-data learning in high dimensional prediction problems	**Course 3**: Non-parametric Bayesian methods for classification
12:30 - 13:30	Lunch break		
13:30 - 14:00	Opening ceremony		
14:00 - 15:00	**Keynote presentation**: Pr. Jeanine Houwing-Duistermaat		
15:00 - 15:20	Tea/coffee break		
15:20 - 16:50	**Contributed session 1**: Survival analysis	**Contributed session 2**: High-dimensional and functional data analysis	**Contributed session 3**: Multivariate data analysis
17:00 - 17:30	**Poster lightning presentations**		
17:30 - 18:30	**Poster session**		

Thursday, April, 8

Time			
09:00 - 10:00	**Invited session: Integrating and analyzing data from different sources (Data Integration)**		
10:00 - 10:30	Tea/coffee break		
10:30 - 12:00	**Contributed session 4**: GWAS and genomic prediction	**Contributed session 5**: Causal inference	**Contributed session 6**: Large scale hypothesis testing
12:00 - 13:30	Lunch break		
13:30 - 15:00	**Contributed session 7**: Spatial and environmental modelling	**Contributed session 8**: Penalized estimation methods	
15:00 - 15:20	Tea/coffee break		
15:20 - 16:20	**Invited session: Infectious diseases**		
16:30 - 18:00	**Contributed session 9**: Longitudinal data analysis	**Contributed session 10**: Bayesian methods	

Friday, April, 9

Time	
09:00 - 10:00	**Invited session: Statistical modelling in movement ecology**
10:00 - 10:30	Tea/coffee break
10:30 - 11:30	**Keynote presentation**: Pr. Mathias Drton
11:30 - 12:00	Closing ceremony - Best poster and best student oral presentation awards

6. Short courses

Wednesday, April 7, 9:00-12:30

6.1 Practical deep learning in R

Sigrid Keydana (Rstudio, Münich, Germany)

Course description

This short course will introduce you to torch for R, an R-native port of PyTorch that requires no Python installation. We start by a thorough introduction to the basics that make everything possible: tensors, automatic differentiation, and neural network modules. Equipped with that knowledge, participants will explore two applications: time series forecasting and numerical optimization. While the former showcases renowned deep learning architectures, the latter demonstrates the usefulness of torch as a high-performance tensor-computation library, going beyond the deep learning context. All modules will incorporate ample occasion for practice. This course does not presuppose familiarity with either deep learning concepts or frameworks; however, participants should have a basic knowledge of the R programming language, as well as of basic machine learning terminology.

6.2 Co-data learning in high dimensional prediction problems

Mirrelijn van Nee, **Magnus Münch** and **Mark van de Wiel** (Amsterdam University medical centers, Department of Epidemiology & Data Science, The Netherlands)

Course description

In many high dimensional prediction settings, extra information on the features, termed co-data, is available. This may benefit prediction if included in the analysis. Co-data comes in different forms: (i) group structures, (ii) hierarchical group structures, and (iii) continuous co-data. In genomics, for example, we may have type (i) co-data in the form of a classification of the genes into functional domains, type (ii) in the form of overlapping and hierarchically organised pathways, and type (iii) as p-values from a previous, related study.

In this course, we introduce several prediction methods that can include these co-data types to improve predictive performance. The penalty parameters are efficiently estimated using empirical Bayes techniques. The course covers technical aspects of co-data learning in ridge regression, elastic net regression, and the random forest. In addition, each of the methods is also investigated in a hands-on practical using the freely available R packages ecpc, gren, and CoRF.

The learning outcomes of this course are three-fold: (i) statistical theory, (ii) statistical application, and (iii) R computing skills. The learning balance between these three outcome may depend on the participants prior knowledge and skills. Some knowledge of statistics is assumed, which includes penalized regression, maximum likelihood estimation, and tree-based learning. Basic understanding of genetics, including the concepts of genome, DNA and phenotype is also useful, but not strictly necessary. Lastly, for the practical part, basic knowledge of R is required. The participants should be able to perform simple operations in R, such as installing packages, arithmetic, assigning and using variables, and applying functions.

6.3 Non-parametric Bayesian methods for classification

Boris Hejblum and **Anaïs Rouanet** (University of Bordeaux, France)

Course description

When performing clustering, a recurring issue is how to choose the appropriate number of clusters. The Bayesian non parametric framework allows to directly estimate the number of clusters within model-based clustering.

The first part of the course will be devoted to the Gaussian Dirichlet Process Mixture model and its Chinese Restaurant Process representation. We will cover theoretical concepts as well as hand-on R practicals illustrating those. The second part of the course will cover the case of supervised clustering where the clustering structure is guided by an outcome. We will illustrate this using the freely available R package PReMiuM on a real epidemiological data application.

Requirements: participants of this course should have a working knowledge of R. Previous exposure to the Bayesian framework analysis and MCMC algorithms would be helpful to understand the concepts covered in this course.

7. Opening keynote presentation

Wednesday, April 7, 14:00-15:00

Chairperson: Beatrijs Moerkerke (Ghent University)

Statistical sciences and interdisciplinary research

Jeanine Houwing-Duistermaat[1,2,3]
[1] Department of Statistics, University of Leeds, United Kingdom
[2] Department of Statistical Sciences, University of Bologna, Italy
[3] Department of Biostatistics and Research Support, University Medical Center Utrecht, The Netherlands.

Abstract: Interdisciplinary research engages multiple disciplines to contribute to science. Many statistical works are interdisciplinary. Without data and relevant questions from other domains, the research would not have taken place. Nowadays, due to advancements of science and technology many large datasets are available in various forms: omics, images, demographic, temporal. Joint analyses of these datasets may provide novel insights, but this often requires multiple disciplines to be involved. On the other hand, widely available open-source data make it easy to perform statistical research without input from other experts. Statistics should benefit from both developments.

For example, in omics research, a popular research topic is development of methods for integrated analysis of multiple omics datasets. These methods need to address high dimensionality, correlation within and between datasets and heterogeneity. Focussing on prediction, machine learning

methods address these challenges. However, they do not provide statistical inference and often do not provide information on which variables contribute most to associations. Further, the study design is often ignored.

I will share our work on statistical methods for omics data integration and suggest future research directions.

8. Contributed sessions 1-3

Wednesday, April 7, 15:20-16:50

8.1 Survival analysis

Chairperson: Niel Hens (University of Hasselt)

A flexible class of generalized joint frailty models for the analysis of survival endpoints

Chauvet Jocelyn[(1)], Rondeau Virginie[(1)]
[(1)] Bordeaux Population Health research center (Biostatistics team), University of Bordeaux, Bordeaux, France
E-mail for correspondence: jocelyn.chauvet@u-bordeaux.fr

Abstract: This communication addresses shared frailty models for correlated failure times, as well as joint frailty models for the simultaneous analysis of recurrent events (e.g., hospital readmissions)and a major terminal event (typically, death). As extensions of the Cox model, these joint models usually assume a frailty proportional hazards model for each of the recurrent and terminal event processes (Liu *et al.*, 2004; Rondeau *et al.*, 2007). To overcome this assumption, our proposal is to replace these proportional hazards models with generalized survival models (Liu *et al.*, 2017, 2018), for which the survival function is modeled as a linear predictor through a link function. Depending on the link function considered, these can be reduced to proportional hazards, proportional odds, additive hazards or probit models. We first consider a fully parametric framework for the time and covariate effects. For proportional and additive hazards models, our

approach also allows the use of smooth functions for baseline hazard functions and time-varying coefficients. The dependence between recurrent and terminal event processes is modeled by conditioning on a shared frailty acting differently on the two processes. Parameter estimates are provided using the maximum (penalized) likelihood method, implemented in the R package `frailtypack` (function `GenfrailtyPenal`). We perform simulation studies to assess the method, which is also illustrated on real datasets.
Key words: Joint Frailty Models; Generalized Survival Models; Recurrent Events; Terminal Event

Liu L., Wolfe R.-A. and Huang X. (2004) Shared frailty models for recurrent events and a terminal event. *Biometrics*, 60(3), 747–756.

Liu X.-R., Pawitan Y. and Clements M.-S. (2017) Generalized survival models for correlated time-to-event data. *Statistics in Medicine*, 36(29), 4743–4762.

Liu X.-R., Pawitan Y. and Clements M. (2018) Parametric and penalized generalized survival models. *Statistical Methods in Medical Research*, 27(5), 1531–1546.

Rondeau V., Mathoulin-Pélissier S., Jacqmin-Gadda H., Brouste V. and Soubeyran P. (2007). Joint frailty models for recurring events and death using maximum penalized likelihood estimation: application on cancer events. *Biostatistics*, 8(4), 708–721.

Estimating sample size for biomarker-strategy designs with survival endpoints

Dinart Derek[1,2], Bellera Carine[1,2], Rondeau Virginie[3]
[1] Bordeaux Population Health Center,INSERM U1219, Bordeaux, France
[2] Clinical Research and Clinical Epidemiology Unit, Institut Bergonie, Comprehensive Cancer Center, Bordeaux, France
[3] Biostatistic team, University of Bordeaux, Bordeaux, France
E-mail for correspondence: derek.dinart@gmail.com

Abstract

Background. In response to the rapid growth of precision medicine and the number of molecules entering the drug development pipeline, several study designs including the biomarker-strategy design (BSD) have been proposed. Contrary to traditional designs, the emphasis here is on the comparison of treatment strategies and not on the treatment molecules as such. Patients are assigned to either a biomarker-based strategy (BBS) arm where biomarker-positive patients receive an experimental treatment or a non-biomarker-based strategy (NBBS) arm where patients receive a treatment regardless of their biomarker status.

Methods. We examined several designs of BSD according to the biomarker assessment and the treatment received in NBBS arm and used frailty survival models to analyse them. Depending on the limits and specificity of each, we proposed statistical models that best described each design. We thus developed a partially clustered frailty model (PCFM) for the (standard) case where the biomarker status is only known in BBS arm. The PCFM allows us to account for the complex structure of BSD that may consider clustering only in one arm. In addition, we proposed an approach to calculate sample size for survival data relying on PCFM. We also proposed statistical tests to measure the overall strategy effect as well as the biomarker-by-strategy interaction effect.

Results. We conducted extensive simulations to assess, for each design, the robustness and performances of the different statistical models. We also performed power analysis and sample size estimation to compare the performance of PCFM to more traditional frailty models, and provided guidelines on the use of BSD in survival analysis.
Key words: Biomarker-strategy, Sample size, frailty model, heterogeneity, randomized

Use of semiparametric frailty model in analysis of survival data in twins

Muli Annah[(1)], Gusnanto Arief[(1)], Houwing-Duistermaat Jeanine[(1,2,3)]
[(1)] Department of Statistics, University of Leeds, United Kingdom.
[(2)] Alan Turing Institute, United Kingdom
[(3)] Department of Biostatistics and research support, Utrecht University Medical Center, The Netherlands.
E-mail for correspondence: staammu@leeds.ac.uk

Abstract: Family based studies help to investigate traits that segregate within families e.g., human longevity which is known to cluster within families. This is attributed to the fact that there is correlation between family members (e.g., twin pairs) as they share genetic and environmental factors. Analysis of such data is challenging due to censoring and correlation among the survival times. The shared frailty model is commonly used for analysis of such correlated survival data. A parametric frailty distribution is assumed, for example the gamma distribution which is computationally convenient. Via simulation we have shown that if the frailty distribution is not correctly specified the estimates of the regression coefficient and survival probabilities may be biased.

We consider a nonparametric specification of the baseline hazard by making use of splines. Simulations showed that replacing the parametric baseline hazard by a flexible baseline hazard can adjust for the incorrect frailty distribution and may improve the estimators of the population survival probability.

We therefore propose to use a semiparametric frailty model to estimate individual specific probabilities of fracture in the next time period given covariates using the TwinsUK data. The event of interest is time to fracture for twins aged 50 years and above. About 1500 people developed a fracture. Results of this analysis will be shown.

Key words: frailty; twins; fractures; survival; misspecification

Bayesian profile regression mixture models to estimate an instantaneous excess risk of cancer from highly correlated exposures and censored survival data

Ancelet Sophie[(1)], Belloni Marion[(1)], Laurent Olivier[(1)], Guihenneuc Chantal[(2)]
[(1)] Institut de Radioprotection et de Sûreté Nucléaire, PSE-SANTE/SESANE /LEPID, Paris, France
[(2)] Université de Paris, Unité de Recherche "Biostatistique, Traitement et Modélisation des données biologiques" BioSTM -UR 7537, Paris, France
E-mail for correspondence: sophie.ancelet@irsn.fr

Abstract: In this work, we focused on the problem of estimating a disease risk from a few highly correlated environmental exposures and a highly censored survival outcome. We extended Bayesian profile regression mixture (PRM) models to this context by assuming an instantaneous excess hazard ratio disease sub-model and conducted a simulation study to explore the performance of these models to deal with co-exposures. Our hierarchical model incorporates a truncated Dirichlet process mixture as an attribution sub-model. An adaptive Metropolis-Within-Gibbs algorithm, including label-switching moves, was implemented to infer the model. This allows simultaneously clustering individuals with similar risks and similar exposure characteristics and estimating the associated instantaneous excess risk for each group. Our Bayesian PRM model was applied to the estimation of the risk of death by lung cancer in a cohort of French uranium miners who were chronically and occupationally exposed to multiple and correlated sources of ionizing radiation. This case study shows that PRM models are promising tools for exposome research and opens new avenues for methodological research in this class of models. It also highlights the potential limit of using standard MCMC algorithms to fit these models, even if the updating schemes for the class labels incorporate label-switching moves.
Key words: Bayesian inference, ionizing radiation, lung cancer, multicollinearity, truncated Dirichlet process mixture

Wild C.-P. (2005) Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol Biomark Prev.* 14:1847–50. doi: 10.1158/1055-9965. EPI-05-0456

Rage E., Caër-Lorho S., Drubay D., Ancelet S., Laroche P., Laurier D. (2015) Mortality analyses in the updated French cohort of uranium miners (1946–2007). *Int Archiv Occupat Environ Health.* 88:717–30. doi: 10.1007/s00420-014-0998-6

Molitor J., Papathomas M., Jerrett M., Richardson S. (2010) Bayesian profile regression with an application to the National Survey of Children's Health. *Biostatistics.* 11:484–98. doi: 10.1093/biostatistics/kxq013

8.2 High-dimensional and functional data analysis

Chairperson: Renee Menezes (Netherlands Cancer Institute)

Improving model performance estimation in high-dimensional data settings by using learning curves

Goedhart Jeroen[(1)], Van De Wiel Mark[(1)], Klausch Thomas[(1)]
[(1)] Department of Epidemiology and Biostatistics, Amsterdam University Medical Centers, Amsterdam, the Netherlands
E-mail for correspondence: j.m.goedhart@amsterdamumc.nl

Abstract: In high-dimensional prediction settings, i.e. when $p > n$, it remains challenging to estimate the test performance (e.g. AUC). Conventional K-fold cross-validation and subsampling methods aim to balance between enough samples to reliably learn the model and estimate its performance. We show that combining estimates from a trajectory of subsample sizes, rendering a learning curve (1), leads to several benefits. Firstly, use of a smoothed curve can improve the performance estimate. Secondly, a still growing-or saturating learning curve indicates whether or not additional samples will boost the prediction accuracy. Thirdly, comparing the trajectories of different learners results in a more complete picture than doing so at one sample size only. Fourthly, the learning curve allows computation of a lower confidence bound for the performance. Standard cross-validation suffers from a limited amount of test samples, whereas the learning curve finds a better trade-off between training-and test sample sizes. This confidence bound is proven to be valid. We show coverage results from a simulation, and compare those to a state-of-the-art technique based on asymptotics (2). Finally, we demonstrate the benefits of our approach by applying it to several classifiers of tumor location from blood platelet RNAseq data.
Key words: High-Dimensional Data; Classification; Learning Curve; Confidence Interval; Omics

(1) Mukherjee *et al.* (2003) Estimating dataset size requirements for classifying DNA microarray data. *Journal of computational biology*, 10, 119-142

(2) LeDell, E. *et al.* (2015). Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. *Electronic journal of statistics*, 91, 1583-1607

Probabilistic PLS method for statistical integration of omics data (PO2PLS)

El Bouhaddani Said[(1)], Uh Hae-Won[(1)], Houwing-Duistermaat Jeanine[(2,3)]
[(1)] Dept. of Biostatistics and Research support, UMC Utrecht, The Netherlands
[(2)] Dept. of Statistics, University of Leeds, Leeds, The UK
[(3)] Dept. of Statistical Sciences, University of Bologna, Bologna, Italy
E-mail for correspondence: s.elbouhaddani@umcutrecht.nl

Abstract: Nowadays, data are collected on several biological levels, e.g., genomics, transcriptomics, epigenetics. We focus on the joint analysis or "integration" of these datasets. Challenges for data integration are high dimensionality, strong correlations, and heterogeneity across omics datasets (due to differing biological levels and measurement platforms). Several methods address parts of these challenges and are popular for data integration. However, they do not provide statistical evidence for a relation between the datasets.

We propose PO2PLS, a probabilistic latent variable framework for the relation between two datasets. The PO2PLS model includes joint and specific components that are linear combinations of the original variables. PO2PLS reduces dimensionality, captures correlations and addresses heterogeneity. For estimation, we develop a memory-efficient EM algorithm, and we show that the estimator is consistent and asymptotically normal, even for high dimensional data. We propose a "global" test for the relation and derive its asymptotic distribution.

We evaluate the estimation and testing performance of PO2PLS with simulations. We illustrate the PO2PLS inference framework with two motivating studies: a population cohort with genetics and glycomics data, and a case-control cohort on hypertrophic cardiomyopathy with epigenetics and transcriptomics data. This demonstrates the potential of PO2PLS as a statistical framework in data integration.

Key words: Joint principal components; Omics data; Inference; High dimensionality; PLS

el Bouhaddani, S., Houwing-Duistermaat, J., Salo, P., Perola, M., Jongbloed, G. & Uh, H.-W. (2016). Evaluation of PO2PLS in Omics data integration. *BMC Bioinformatics*, 17(S2), S11.

el Bouhaddani, S., Uh, H.-W., Hayward, C., Jongbloed, G. & Houwing-Duistermaat, J. (2018). Probabilistic partial least squares model: Identifiability, estimation and application. *Journal of Multivariate Analysis*, 167, 331–346

Sparse inverse time correlation model for signal identification in functional Near Infrared Spectroscopy data

David Causeur[(1)], Ching-Fan Sheu[(2)]
[(1)] Agrocampus Ouest, Irmar, UMR 6625 CNRS, 65 rue de St-Brieuc - CS 84215, 35042 Rennes Cedex, France
[(2)] National Cheng-Kung University, Institute of Education, 1 University Road, Tainan 701, Taiwan
E-mail for correspondence: david.causeur@agrocampus-ouest.fr

Abstract: Functional near infrared spectroscopy (fNIRS) uses the absorption of near infrared light by hemoglobin to record changes in blood oxygenation as signals of functional brain activity. For designs in which subjects are instructed to execute a specific mental task under different experimental conditions with pre-determined levels for covariates, fNIRS provides real-time cerebral hemodynamic responses for studying neural correlates of task-related experimental variables. Data obtained from such designs are discretized observations of the hemodynamic curves on a high-resolution time scale. Testing for overall group mean differences among curves or, more generally, relationships between curves and explanatory variables can be addressed by using functional Analysis of Variance (fANOVA) procedures in a general multivariate linear regression framework where additional assumptions are made to account for the regularity of mean curves and for the strong time-dependence across residuals.

Causeur *et al.* (2020) demonstrated that how way time dependence is modeled in such fANOVA testing procedures is crucial and should account for the interplay between the pattern of regression parameter curves and the distribution of the time correlations. To address the challenging issue of identifying time points for which the association signal is nonzero, we propose a doubly penalized estimation procedure assuming that both the association signal and the inverse time correlation matrix are sparse. We show how the tuning of penalty parameters enables a flexible handling of dependence and deduce optimal signal identification procedures.
Key words: Functional data; fNIRS data; Inverse correlation model; High-dimensional data; Penalized estimation.

Causeur, D., Sheu, C.-F., Perthame, E. and Rufini, F. (2020). A functional generalized F-test for signal detection with applications to event-related potentials significance analysis. *Biometrics*. 76(1), 246–256.

Data-Driven Tail-Greedy Unbalanced Haar-Fisz Method for Copy Number Alteration Data

Ummi Maharani Ahsani[(1)], Gusnanto Arief[(1)], Barber Stuart[(1)]
[(1)] Department of Statistics, University of Leeds, Leeds LS2 9JT, United Kingdom
E-mail for correspondence: mmmau@leeds.ac.uk

Abstract: Copy Number Alterations (CNA) are genomic aberrations, in which some regions of a genome exhibit more or less copy number than the normal two. They appear as 'gains' or 'losses' of copy number along the genome and play a key role in cancer diagnosis. In particular, the locations of the gains and losses are of major interest. However, estimation of CNA is a challenging process because CNA data contain inconsistent error variability. Several segmentation methods have been proposed to estimate CNA and many of them perform well for data whose error variability is relatively constant. In practice, real CNA data deviate from this assumption and indicate some dependencies of the variance on the mean value. To address this problem, we have developed a method called Data-Driven Tail-Greedy Unbalanced Haar-Fisz (DDTF) segmentation. The proposed method performs variance stabilization via a Fisz transform to bring the problem into a homoscedastic model before applying a denoising procedure. The use of the Tail-Greedy Unbalanced Haar wavelet also makes it possible to estimate CNA location more precisely compared to the traditional Haar wavelet. Furthermore, our simulation study shows the superiority of DDTF in estimating short segments, which are often difficult to detect by existing methods.

Key words: Copy number alteration; unbalanced Haar wavelet; data-driven denoising; heteroscedasticity; change-point detection

Fryzlewicz P. (2008) Data-driven wavelet-Fisz methodology for nonparametric function estimation. *Electronic Journal of Statistics* 2:863–896.

Fryzlewicz P. (2018) Tail-greedy bottom-up data decompositions and fast multiple change-point detection. *Annals of Statistics* 46.

8.3 Multivariate data analysis

Chairperson: Paul Eilers (Erasmus University Medical Centre)

Statistical modelling of in vitro pepsinolysis using peptidomic data
Suwareh Ousmane[(1)], Causeur David[(2)], Jardin Julien[(1)], Briard-Bion Valérie[(1)], Le Feunteun Steven[(1)], Pezennec Stéphane[(1)], Nau Françoise[(1)]
[(1)] STLO, INRAE, Institut Agro, 65 rue de Saint-Brieuc, 35042 Rennes, France
[(2)] IRMAR UMR6625, CNRS, Institut Agro, 65 rue de Saint-Brieuc, 35042 Rennes, France
E-mail for correspondence: ousmane.suwareh@agrocampus-ouest.fr

Abstract: The digestion process is a complex phenomenon not yet completely understood, despite a significant amount of studies on this topic. In the case of protein foods, peptidomic data generated using mass spectrometry can be used to identify the protein fragments released due to the action of digestive enzymes, and therefore to identify the peptide bonds cleaved. Using an in vitro model of digestion focused on the gastric phase, our goal is to propose a statistical framework for the probability that pepsin cleaves a sequence of aminoacid residues at a given peptide bond. The tested variables include the composition of the sequence itself around the peptide bond, and a large set of physicochemical features of its three-dimensional environment. The statistical framework introduces large-dimensional propensity scores, one for each aminoacid residue, at each position flanking the peptide bonds, in a logistic regression model.

In order to assess the specificity of pepsin action, namely that cleavage sites along protein sequences are not distributed randomly, significance tests were first implemented. However, the large dimension of the model questioned the accuracy of the standard likelihood-ratio tests. An alternative penalized estimation procedure was also proposed to select the features favouring cleavage by pepsin, assuming that the number of aminoacid residues influencing pepsin action is low. The presentation will focus on the comparison of these two approaches to analyse experimental data in a large design involving six proteins intentionally chosen to cover a large scope of physicochemical properties.
Key words: Peptidomic data; Large-dimensional propensity scores; Pepsin specificity

Improved Multivariate Extensions of McNemar's Test

Touloumis Anestis[(1)]
[(1)] University of Brighton, Brighton, UK
E-mail for correspondence: A.Touloumis@brighton.ac.uk

Abstract: McNemar's test is used on paired binary data to detect differences in the marginal probabilities (marginal homogeneity assumption). Klingenberg and Agresti (2006) considered extensions of McNemar's test to the global assumption of marginal homogeneity between paired vectors of binomial responses when paired multivariate binary data are collected. They suggested test statistics that are based on a generalized estimating equations (GEE) model (Liang and Zeger, 1986), bootstrap tests, and permutation tests. However, for sparse or imbalanced data, these approaches might either be intractable or computationally infeasible. To circumvent this, we propose test statistics that can be derived from a penalized GEE model. This ensures the finiteness of the regression parameters of the GEE model, and hence the existence of the proposed test statistics. We derive closed-form formulae for the proposed tests, investigate their theoretical properties, and assess their performance in finite samples via simulation. The proposed tests are applied to safety data for a drug, in which two doses are evaluated by comparing multiple responses by the same subjects to each one of them.

Key words: Correlated Binary Responses; Generalized Estimating Equations; Hypothesis Testing; McNemar Test; Paired Data

Klingenberg B., and Agresti A. (2006) Multivariate Extensions of McNemar's Test. *Biometrics*, 62(3), 921–928.

Liang K.-Y., and Zeger S.-L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13–22.

Categorizing variables after using regression calibration

Boshuizen Hendriek[1,2], Stratos-Initiative Topic Group 4 On Measurement Error And Misclassification Subgroup Categorized Variables[3]
[1] Biometris/Human Nutrition, Wageningen University and Research, Wageningen, Netherlands
[2] National Institute of Public Health and the Environment, Bilthoven, The Netherlands
[3] Stratos-Initiative Topic Group 4 on measurement error and misclassification, subgroup categorized variables
E-mail for correspondence: hendriek.boshuizen@wur.nl

Abstract: In nutritional epidemiology consumption of food intake is assessed using instruments (questionnaires, diaries, apps) subject to error. Such measurement error, even when non-differential (i.e. independent of disease status), will bias results of a regression of disease on intake. When intake is used as a continuous measurement, there are many methods available to deliver unbiased estimates, of which regression calibration (replacing the observed intake with a predicted value of true intake, the calibrated value) is one of the most popular.

However, in nutritional epidemiology the intake is often categorized, in order to study the shape of the relationship or to avoid assuming linearity. Measurement error in dietary intake will result in misclassification of subjects in these categories. As shown before (1), even when measurement error is non-differential, the misclassification is not.

Here we used a "worst case" simulation to study whether applying categorization on calibrated intake, although theoretically incorrect, would still approximately correct bias in practice, an approach close to that proposed by McMahon (2). Simulations showed that without confounding this approach works well. However, with strong confounding results are still biased. The likely cause is that a model using categorization misspecifies the relation between intake and therefore biases the estimated confounder-disease relation.

Key words: regression calibration; measurement error; categorization; variables-with-error; dietary assessment

1) Flegal, K.M., Keyul, P.M., Nieto, F.J. (1991) Differential Misclassification Arising from Nondifferential Errors in Exposure Measurement. *Am J Epidiol.* 134(10):1233-44

2) MacMahon, S., Peto, R., Cutler, J., Collins, R., Sorlie, P., Neaton, J., Abbott, R., Godwin, J., Dyer, A., and Stamler, J. (1990). Blood pressure, stroke, and coronary heart disease. Part 1, prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. *Lancet*, 335: 765–774.

3) Keogh, R.H., Strawbridge, A,D., White, I.R. (2012) Effects of Classical Exposure Measurement Error on the Shape of Exposure-Disease Associations, *Epidemiologic Methods*, 1(1):article 2. DOI: 10.1515/2161-962X.1007

Combined-information criterion for clusterwise elastic-net regression. Application to omic data

Bougeard Stéphanie[(1)], Bry Xavier[(2)], Verron Thomas[(3)], Niang Ndeye[(4)]
[(1)] ANSES (French agency for food, environmental and occupational health safety), Ploufragan, France
[(2)] University of Montpellier, France
[(3)] SEITA, Paris, France
[(4)] CEDRIC CNAM, Paris, France
E-mail for correspondence: stephanie.bougeard@anses.fr

Abstract: Many research questions pertain to a regression problem assuming that the population under study is not homogeneous with respect to the underlying model. In this setting, we propose an original method called Combined Information criterion CLUSterwise elastic-net regression (CICLUS). This method handles several methodological and application-related challenges. It is derived from both the information theory and the microeconomic utility theory and maximizes a well-defined criterion combining three weighted sub-criteria, each being related to a specific aim: getting a parsimonious partition, compact clusters for a better prediction of cluster-membership and a good within-cluster regression fit. The solving algorithm is monotonously convergent under mild assumptions. The CICLUS method provides an innovative solution to two key issues: the automatic optimization of the number of clusters and the issue of a prediction model. We applied it to elastic-net regression in order to be able to manage high-dimensional data involving redundant explanatory variables. CICLUS is illustrated through a real example in the field of omic data, showing how it improves the quality of the prediction and facilitates the interpretation. It should therefore prove useful whenever the data involve a population mixture as for example in biology, social sciences, economics or marketing.
Key words: Clusterwise regression; Typological regression; Elastic-net regularization

Charles C. (1977) Régression typologique et reconnaissance des formes. PhD, University of Paris IX, France.

Mortier F., Ouedraogo D.-Y., . . . and Picard N. (2015) Mixture of inhomogeneous matrix models for species-rich ecosystems. *Environmetrics*, 26(1), 39–51.

Suk H.-W. and Hwang H. (2010) Regularized fuzzy clusterwise ridge regression. *Advances in Data Analysis and Classification*, 4, 35–51.

9. Poster session

Wednesday, April 7, 17:00-18:30

Compromise clustering for quaternion time-series and its application for gait analysis in Multiple Sclerosis

Pierre Drouin[1,2], Stamm Aymeric[1], Chevreuil Laurent[2], Graillot Vincent[2], Laplaud David[3], Bellanger Lise[1]
[1] Laboratoire de Mathématiques Jean Leray, Faculté des Sciences et Techniques, Nantes, France
[2] Uman E-Health Solution, UmanIT, Nantes, France
[3] Centre d'investigation Clinique, équipe Neurologie, Centre Hospitalier Universitaire, Nantes, France
E-mail for correspondence: pdrouin@umanit.fr

Abstract: Recent approaches in gait analysis involve the use of quaternion time-series representing the body segments rotation and/or orientation in 3D space during the walk. Following the proposition of Motl (2017) in, we hereby propose a method to analyze walking data measured on patients with Multiple Sclerosis (MS). The subject matter is to find groups of multiple sclerosis patients with similar walking deficiencies. Expanded Disability Status Scale is currently used to assess overall disability in MS. This information may be taken into account when searching for the grouping structure to provide clinically relevant partition. There are two ways of incorporating external information: either through constraints or through a compromise. The former usually forces some observations to belong to the same cluster or some cluster to be structured in a particular fashion (Dinler(2016)) while the latter only uses the external information to guide the search of the grouping structure. A recent clustering approach known as `perioclust`(Bellanger(2020)) builds

on the principal of hierarchical agglomerative clustering and accounts for external information in the way similarity between observations is measured. We propose its generalisation for quaternion time-series and present results of its application on the walking data of patients with MS.
Key words: Time series; Compromise Clustering; Quaternion; Human Gait

Motl R.-W., Cohen J.-A., Benedict R., Phillips G., LaRocca N., Hudson L.-D., Rudick R. (2017). Validity of the timed 25-foot walk as an ambulatory performance outcome measure for multiple sclerosis. *Multiple Sclerosis Journal,* Vol. 23(5) 704–710.

Dinler D. and Tural M.-K. (2016). Analysis of Longitudinal Data. Unsupervised learning algorithms.

Bellanger L., Coulomb A., Husi P. (2020). Models for Discrete Longitudinal Data. Data Analysis,and Rationality in a Complex World.

Abundance and distribution of the blue shark in the Bay of Biscay

Lea Pautrel[(1)], Rindra Ranaivomanana[(1)], Emma Rouault[(1)], Mathieu Genu[(2)], Matthieu Authier[(2)], Marie-Pierre Etienne[(1)]
[(1)] Agrocampus Ouest, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France
[(2)] Observatoire PELAGIS, UMS 3462 CNRS-La Rochelle Université, 5 allée de l'Océan, La Rochelle, France
E-mail for correspondence: marie-pierre.etienne@agrocampus-ouest.fr

Abstract: The SPEE program has been developed to monitor over time the abundance of many marine megafauna species of conservation interest, such as marine mammals (harbour porpoise, common dolphin, ...), seabirds (Northern gannet, Black-headed gull , ...) or elasmobranchs (sharks, rays, ...)in the newly implemented Marine Protected Area (MPA) of the Gironde estuary and Pertuis sea, France. SPEE consists in seasonal surveys over the MPA throughout a whole year and allows for the observation of many large fish, in particular Blue sharks (Prionace glauca), a species widely distributed in the world's ocean. Using a distance sampling approach to estimate blue shark abundance and distribution within the MPA, this work provides the first abundance estimate in the area and shows a very strong seasonal pattern and highlights a strong spatial structuration of the blue shark population with few environmental explanation. The sharks population were more abundant and widespread in spring and tended to concentrate in summer. Since only 96 sharks have been spotted in 2019, there is a large uncertainty in the abundance estimate and future surveys will be useful to obtain more precise estimate of the detection function and consequently produce more robust abundance estimation.

Key words: Distance sampling; Abundance survey; Blue sharks

Understanding Variations in Mean Polyps Detection as a Key Performance Indicator (KPI) for Endoscopists in England: A Simulation Study

Kharati Ehsan[1,2], Wagnild Janelle[1,2], Catlow Jamie[3], Lu Liya[3], Sharp Linda[3], Matt Rutter[4], Kasim Adetayo[1,2]
[1] Durham University, Anthropology Department, Durham, United Kingdom of Great Britain and Northern Ireland
[2] Durham University, Durham Research Methods Centre, Durham, United Kingdom of Great Britain and Northern Ireland
[3] Newcastle University Centre for Cancer – Populations Health Sciences Institute, Newcastle Upon Tyne, United Kingdom of Great Britain and Northern Ireland
[4] North Tees and Hartlepool NHL Foundation Trust –Gastroenterology, Stockton on Tees, United Kingdom of Great Britain and Northern Ireland
E-mail for correspondence: a.s.kasim@durham.ac.uk

Abstract: Colorectal cancer (CRC), the fourth most common cancer in the UK, arises from benign polyps over a long period of time. Detection and removal of polyps through colonoscopy is therefore a critical step in the reduction of CRC incidence and mortality. There is substantial variation in polyp detection rate, and therefore colonoscopy quality, and the optimal statistical approach for quantifying this variation in performance is unknown. In this study we compare the performance of different statistical methods in estimating KPI for endoscopists using routinely collected big data in the National Endoscopy Database. When there was substantial variation between sites, generalised mixed-effects models (GLMM) showed higher accuracy in ranking performance of endoscopists (96% and 81% accuracy for top 25% and bottom 25%, respectively) compared to Poisson and Negative Binomial models (88% and 87%, respectively). When there was variation between endoscopists and not between sites, GLMM showed higher accuracy in identifying endoscopists in the top 25% compared to Poisson and Negative Binomial models (92% and 82%, respectively). However, Poisson and Negative Binomial model (81%) performed better than generalised linear mixed effects (75%) model in ranking endoscopists in the bottom 25%.

Key words: Generalised Mixed Effect Models; Polyp Detection; KPI, Poisson Regression and Negative Binomial Regression.

Barclay R.-L., *et al.* (2006). Colonoscopic Withdrawal Times and Adenoma Detection during Screening Colonoscopy. *N Engl J Med*:355(24), 2533–2541.

Chen, S.-C. and Rex, D.-K. (2007). Endoscopist can be more powerful than age and male gender in predicting adenoma detection at colonoscopy. *American Journal of Gastroenterology*:102(4), 856–861.

Kaminski M.-F., *et al.*(2010). Quality Indicators for Colonoscopy and the Risk of Interval Cancer. *N Engl J Med*: 362(19), 1795–1803.

Baxter N.-N., *et al.* (2011). Analysis of Administrative Data Finds Endoscopist Quality Measures Associated With Postcolonoscopy Colorectal Cancer. *Gastroenterology*: 140(1), 65–72.

Corley, D.-A. *et al.* (2014). Adenoma detection rate and risk of colorectal cancer and death. *New england journal of medicine*:370(14),1298–1306.

Bretagne, J.-F., *et al.* (2016). Interendoscopist variability in proximal colon polyp detection is twice higher for serrated polyps than adenomas. *World journal of gastroenterology*:22(38), 8549.

Catlow, J., *et al.* (2020). The National Endoscopy Database (NED) Automated Performance Reports to Improve Quality Outcomes Trial (APRIQOT) randomized controlled trial design. *Endoscopy International Open*: 8(11), p.E1545.

Comparison of statistical methods for estimating the effect of time-varying treatment on the risk of adverse event

Manitchoko Liliane[(1)], Bénichou Jacques[(1,2)], Thiébaut Anne[(1)]
[(1)] High-Dimensional Biostatistics for Drug Safety and Genomics, Université Paris-Saclay, UVSQ, Inserm, CESP, Villejuif, France
[(2)] Department of Biostatistics, Rouen University Hospital, Rouen, France
E-mail for correspondence: liliane.manitchoko@inserm.fr

Abstract:

Background: In pharmacoepidemiology, assessing the effect of drug exposure on an adverse event risk is challenging because exposure can vary over time and its effect can be complex. Cohort and nested case-control (NCC) designs are widely used in this context. However, evaluation of their relative performance is limited.

Methods: We simulated 1000 prospective cohorts of 5000 individuals for both fixed and time-varying exposure with a unique change (from "unexposed" to "exposed") during follow-up. We varied exposure prevalence, hazard ratios of event associated with exposure, proportions of subjects experiencing the event (cases) and matching control:case ratio in the NCC design.

Results: In all scenarios, the cohort design had small bias, was more precise and had greater power than the NCC design. For both types of exposure, bias in the NCC design tended to increase with lower exposure prevalence and higher proportions of events while it decreased with increasing matching ratio and higher hazard ratios.

Conclusion: Results with the NCC design should be interpreted with caution given its potential limitations. More complex exposures (e.g., time-varying with multiple changes or decreasing hazard ratios) are under evaluation, guided by the analysis of breast cancer risk associated with menopausal hormone therapy in the E3N cohort.

Key words: Pharmacoepidemiology; Cohort design; Nested case control design; Simulations.

Robustness of Supervised Clustering Methods to Different Types of Inactive Variables

Marion Rebecca[(1)], Lederer Johannes[(2)], Govaerts Bernadette[(1)], Von Sachs Rainer[(1)]
[(1)] ISBA/LIDAM, Université catholique de Louvain, Louvain-la-Neuve, Belgium
[(2)] Department of Mathematics, Ruhr-Universität Bochum, Bochum, Germany
E-mail for correspondence: rebecca.marion@uclouvain.be

Abstract: Model regularization methods that perform embedded variable selection during the model estimation process have great potential for increasing model interpretability. In cases where the predictor variables form clusters (such as in gene expression data, where genes belong to different regulatory pathways), the selection of important variables using model regularization is more challenging. This is especially true when the variable clusters are not known a priori. "Supervised clustering" methods in the literature, such as Pairwise Absolute Clustering and Sparsity (Sharma (2013)), Simultaneous Supervised Clustering and Feature Selection (Shen (2012)) and Cluster Elastic Net (Witten (2014)) are able to learn variable clusters from the data and select important clusters during model estimation. However, the correlation structure of inactive variables (i.e. variables that do not predict the response) can impact the variable selection, clustering and prediction quality of these methods. Inactive variables come in several types: they can be correlated or uncorrelated with each other, as well as correlated or uncorrelated with active variables. This poster presents a simulation study comparing state-of-the-art supervised clustering methods and demonstrating the impact of the correlation structure of inactive variables on prediction and clustering performance.

Key words: Variable clustering; Variable selection; Regression; Regularization.

Sharma D., Bondell H.-D. and Zhang H. (2013). Consistent group identification and variable selection in regression with correlated predictors. *Journal of Computational and Graphical Statistics*:22(2), 319–340.

Shen X., Huang H.-C. and Pan W. (2012). Simultaneous supervised clustering and feature selection over a graph. *Biometrika*: 99(4), 899– 914.

Witten D.-M., Shojaie A. and Zhang F. (2014). The cluster elastic net for high-dimensional regression with unknown variable grouping. *Technometrics*: 56(1), 112–122.

Predicting yield in new environments from a variety-testing trial network using random regression models

Ramakers Jip, Bustos-Korts Daniela, Boer Martin, Kruijer Willem, Van Eeuwijk Fred

Abstract: Predicting yield in new environments from VCU (Value for Cultivation and Use) crop variety trials remains a challenge in plant breeding. In VCU trials, new varieties are typically tested in field trials for a few consecutive years in multiple locations before being released and replaced by others. Genotype-by-trial (or genotype-by-year-by-location) analysis (GxE) of these trials focusses on the ranking of varieties in terms of mean yield across environments, using variance component linear mixed models (LMM). A major question in the VCU context is whether the current trial system can be improved (e.g. by increasing the number of test years per variety) to get better predictions for each genotype, and whether environmental covariates can be used to predict phenotypes in novel environments.

Using a fully balanced (simulated) dataset of wheat varieties across multiple years and locations, we tested the predictive performance of random regression models (RRM) within and outside observed environments, and compared it to conventional LMMs. We specifically looked at the number of training and test years needed for accurate prediction, as well as whether leveraging phenotypic information from a secondary location (or region) in the RRM would further increase accuracies. Our results show that RRMs performed admirably well compared to traditional LMMs and that prediction accuracies in unsampled environments were moderate to high, depending on the location, the number of years used for training and prediction, and the incorporation of phenotypes from a secondary location. RRMs are thus a useful tool for prediction in a VCU context, but the design of the VCU network will be an important factor that determines the performance of these models.

Non-parametric clustering of longitudinal functional data with application to NMR spectra of 18 kidney transplant patients

Xie Minzhen[(1)], Houwing-Duistermaat Jeanine[(1)], Liu Haiyan[(1)]
[(1)] Department of Statistics, University of Leeds, Leeds, UK
E-mail for correspondence: mmmxi@leeds.ac.uk

Abstract: The amount of data in the health domain is growing rapidly, which have different forms: omics, imaging and functional. A relevant question is whether these data can be used to cluster subjects to reveal their underlying health status. The aim of this research is to cluster the 18 kidney transplant patients based on their NMR spectra. An NMR spectrum is a function in mass, and the NMR spectra of each patient are recorded up to nine times (longitudinal design). It might loss relevant information, if we apply multivariate clustering method such as K-means on extracted scalar summaries of longitudinal measured NMR spectra.

To use all available information, we propose a non-parametric clustering method for multivariate functional data. Firstly, the distance of multivariate curves is defined and small ball probability is computed based on defined distance. Secondly, we compute the heterogeneity of original sample and partitioned samples by using mean, median and mode. Thirdly, we define a criterion to determine whether the obtained clustering provides homogeneous clusters or require further splits. We compare our method with other (non) functional clustering methods via simulation and apply the method to the kidney data.

Key words: Nonparametric clustering; Longitudinal functional data.

Philippe Vieu, Frédéric Ferraty (2006). *Nonparametric Functional Data Analysis*. New York: Springer-Verlag.

Ana-Maria Staicu, So Young Parka (2012). Longitudinal Functional Data Analysis. *The ISI's Journal for the Rapid 4*.

10. Invited session 1

Thursday, April 8, 09:00-10:00

Chairperson: Nicole Augustin (The University of Edinburgh)

Integrating and analyzing data from different sources (Data Integration)

Integrating and analysing -omics data from different sources in Alzheimer research

Cornelia van Duijn[(1)]
[(1)] Nuffield Department of Population Health, University of Oxford, UK.
E-mail for correspondence: Cornelia.vanDuijn@ndph.ox.ac.uk

Abstract: Large-scale genomics studies have plaid a key role in unravelling the causes and consequences of Alzheimer's disease (AD). Historically, genetic research has been a major driver of our understanding the aetiology of AD. The past decade has seen rapid changes in the availability of publicly accessible data sets involving -omics data beyond the genome: transcriptomics, proteomics, metabolomics and metagenome (microbiome). These data are not only generated in patient and controls but increasingly in cellular models. The integration of these data with genetic data have fuelled Alzheimer research. However, the wealth of data raises new questions on how to optimize the analysis of high-dimensional correlated data and how to deal with possible confounders. Examples of cross-omics studies integrating (epi)genetic, transcriptomic, proteomic, metabolomic and microbiome data in Alzheimer research will be discussed.

mixOmics: an R package for the integration of biological data sets

Sébastien Déjean[(1)]
[(1)] University of Toulouse, Toulouse Mathematics Institute
E-mail for correspondence: sebastien.dejean@math.univ-toulouse.fr

Abstract: It is generally admitted that single 'omics analysis does not provide enough information to give a deep understanding of a biological system, but we can obtain a more holistic view of a system by combining multiple 'omics analyses. In this context, this talk will present the mixOmics R packagethat proposes multivariate statistical methods to explore and integrate 'omics data sets with a specific focus on variable selection and visualisation.

Multi-omics data analysis using sets of variables: many needles in multiple haystacks

Renée Menezes[(1)]
[(1)] Netherlands Cancer Institute. Biostatistics Unit, Amsterdam.
E-mail for correspondence: r.menezes@nki.nl

Abstract: Understanding of molecular regulation mechanisms is important to better understand how processes in the human body occur, in particular in disease onset and development such as that of cancer. To help with this understanding, studies gather increasing amounts of molecular profiling data, such as DNA (copy number, methylation), mRNA and protein profiles. Such studies could help us better understand how often a gene dosage effect (change in copy number) occurs at the same time as a gene silencing one (hypermethylation). However, the analysis of all these data still represents a challenge. Often researchers still choose to study the association between pairs of features, for example the expression of one gene together with the copy number status at the beginning of the gene. Such pairwise analyses not only require stringent multiple testing correction (as the number of pairs of features is very large), but also cannot easily incorporate multiple omics datasets simultaneously.

We propose to use an approach that enables testing of the effect of a large number of variables on one response. By focusing on testing, no estimation of parameters is required, what makes this approach applicable to problems with $p \gg n$. We have applied this approach to studying the association between copy and gene expression in colon and breast cancer, uncovering interesting patterns that better characterize differences between these two cancer types. We have also extended the approach to explain gene expression by multiple omics data, for example by both copy number and methylation. In this way, we have uncovered genes which, when in a region with copy number gain, are silenced via methylation. We have also uncovered genes that achieve overexpression in different ways: in some samples by DNA copy gain, in others by hypomethylation. Another extension involved testing for spliceQTL, which required using multiple responses: here we tested for association between counts for all individual exons of a given gene, and the number of minor alleles of SNPs located between the start and end of the gene. By considering multiple relevant features at the same time, results yielded by this approach are more often replicated than those obtained by pairwise testing.

In conclusion, we have proposed efficient approaches to perform joint analyses of multi-omics datasets. By focusing on testing, rather than on estimation, our approaches can be used in $p \gg n$ problems without incurring very stringent multiple testing. In addition, the focus on testing helps further with separating signal from noise, a recurring problem in high-dimensional data analysis. Furthermore, by enabling the use of variable sets both as "responses" as well as "explanatory", we can study complex problems such as spliceQTL. Finally, by considering more data at once, results yielded are more often replicated than low-dimensional approaches.

11. Contributed sessions 4-6

Thursday, April 8, 10:30-12:00

11.1 GWAS and genomic prediction

Chairperson: Mark van de Wiel (Amsterdam University Medical Center)

Causal multi-trait analysis of Genome-Wide Association Studies data
Ahmed Azza[1,2], Kruijer Willem[3], Wit Ernst[4], Grzegorczyk Marco[1]
[1] Bernoulli Institute, University of Groningen, Groningen, the Netherlands
[2] Center for Bioinformatics and Systems Biology and Department of Electrical and Electronic Engi-neering, University of Khartoum, Khartoum, Sudan
[3] Biometris, Wageningen University and Research, Wageningen, the Netherlands
[4] Institute of Computational Science (ICS) and Social Network Analysis Research Center (SoNAR-C),Universita della Svizzera italiana, Lugano, Switzerland

Abstract: Univariate linear mixed models are commonly used in the analysis of genome-wide association studies (GWAS). However, with the rise of deeply phenotyped data in population-scale biobanks, there are now better opportunities for understanding relations between traits (in light of their genetic and non-genetic causes) by employing multivariate approaches. Causal inference (from observational data) of directed biological networks between traits and their causes lead to a better understanding of disease etiology, identification of risk factors, and ultimately, more efficient diagnosis

and treatment options. This talk gives an extension to a causal discovery framework, pcgen (Kruijer *et al.*, 2020), that employs the PC algorithm for building a network between traits and genetic causes. An overview of pcgen will be provided, along with extensions: 1) to make it applicable to human studies by accounting for genetic relatedness, 2) based on individual-level genotype data from a homogeneous cohort, on whom multiple phenotypes measurements are available, 3) at a reduced computational cost. A case study applying this approach to a human GWAS dataset will also be presented, where direct genetic and non-genetic effects between traits are delineated.
Key words: GWAS; Causality; pc algorithm, multivariate analysis

Kruijer W., Behrouzi, P. Bustos-Korts D., *et al* (2020) Reconstruction of networks with directand indirect genetic effects. *Genetics*: 214(4), 781–807

Fast and accurate multi-environment genomic prediction using penalized factorial regression

Kruijer Willem[(1)], Millet Emilie[(1,2)], Van Dijk Aalt-Jan[(1,3)], Bustos-Korts Daniela[(1)], Avagyan Vahe[(1)], Ramakers Jip[(1)], Boer Martin[(1)], Van Eeuwijk Fred[(1)]
[(1)] Biometris, Wageningen University and Research, Wageningen, The Netherlands
[(2)] INRA Montpellier, Montpellier, France
[(3)] Laboratory of Bioinformatics, Wageningen University and Research, Wageningen, The Netherlands
E-mail for correspondence: willem.kruijer@wur.nl

Abstract: Genomic prediction has become an important tool in applications ranging from genomic selection to personalized medicine. While methodology for univariate genomic prediction is well-established, prediction for new environments remains a notorious challenge, which is however of great interest in plant breeding, where new varieties need to be adapted to a range of increasingly extreme conditions. At least in theory, phenotypes in new environments can be predicted using environmental covariates (ECs) such as temperature, which quantify both tested and untested environments.

Millet *et al* (2019) recently showed that this is indeed possible using factorial regression models, in which Genotype-by-Environment (GxE) interactions are modelled using genotype-specific sensitivities to ECs. Their approach however relied on 3 predefined ECs, driven by biological knowledge. Such knowledge is however often unavailable, with potentially hundreds of ECs to choose from. Here we consider various ways to penalize factorial regression models, and simultaneously regularize SNP, EC and GxE effects, implicitly assuming different causal models. We show that for the most appropriate of these models, penalized factorial regression leads to accurate predictions for new genotypes in new environments. For simulated and real data with medium to high heritabilities, the within-environment accuracy is on average r = 0.68, outperforming a state-of-the-art deep-learning approach (Khaki and Wang 2019), as well as a popular Bayesian method (Jarquin et al 2014).
Key words: Multi-environment genomic prediction; penalized regression; deep learning.

Jarquin D., Crossa J., Lacaze X. *et al* (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and Applied Genetics*: 127(3): 595–607.

Khaki S., and Wang L. (2019) Crop Yield Prediction Using Deep Neural Networks. *Frontiers in Plant Science*:https://doi.org/10.3389/fpls.2019.00621

Millet, E.-J., Kruijer, W., Coupel-Ledru, A. *et al* (2019). Genomic prediction of maize yield across European environmental conditions. *Nature Genetics*: 51: 952-956.

Improving genomic prediction using secondary phenotypes

Arouisse Bader[(1)], , van Eeuwijk Fred[(1)], Kruijer Willem[(1)]
[(1)] Biometris, Wageningen University and Research, (Wageningen, The Netherlands)
E-mail for correspondence: willem.kruijer@wur.nl

Abstract: In the past decades, genomic prediction has had a large impact on plant and animal breeding. Given the current advances of high-throughput phenotyping, it is increasingly common to measure a large number of traits, in addition to the target trait of interest. This raises the important question whether these 'secondary' traits or 'co-data' can be used to improve genomic prediction for the target trait. With only a small number of secondary traits, this is known to be the case, given sufficiently high heritability and genetic correlations. Here we focus on the more challenging situation when many secondary traits are available. Secondary traits are then usually modelled using separate ridge penalties (Van De Wiel *et al.* (2016)), or, equivalently, through additional kernels (relatedness matrices). This approach is however infeasible when secondary traits are not measured on the test set, and cannot distinguish between genetic and residual correlations. Here we discuss three approaches that could potentially overcome one or both of these limitations. Our first approach relies on a dimension reduction, achieved using either random forests or penalized regression of the target on the secondary traits, while the second approach reduces the dimension using penalized selection indices (Lopez-Cruz *et al.* (2019)). In both cases, we use the bivariate GBLUP, with the fitted values as secondary trait. Finally, we extend existing multi-kernel approaches by replacing secondary traits by their genomic predictions, which we denote GM-BLUP. For most of our simulated datasets, prediction using selection indices was most accurate, while on real maize and arabidopsis data, the dimension reduction using random forests performed best. GM-BLUP performed only slightly better than existing multi-kernel methods, however being also applicable when secondary traits are not measured on the test set.

Key words: GBLUP, secondary traits, selection indices, penalized regression, random forests.

Marco Lopez-Cruz, Eric Olson, Gabriel Rovere, Jose Crossa, Susanne Dreisigacker, Suchismita Mondal, Ravi Singh and Gustavo de los Campos (2020). Regularized selection indices for breeding value prediction using hyperspectral image data. *Sci Rep* 10, 8195.

Mark A. van de Wiel, Tonje G. Lien, Wina Verlaat, Wessel N. van, Wieringen and Saskia M. Wilting (2015). Better prediction by use of co-data: adaptive group-regularized ridge regression. *Statistics in Medicine*.

11.2 Causal inference

Chairperson: Cécile Proust-Lima (Bordeaux Population Health, INSERM)

Causal assessment of surrogacy for time-to-event endpoints using meta-analytic data

Le Coënt Quentin[(1)], Legrand Catherine[(2)], Rondeau Virginie[(1)]
[(1)] Department of Biostatistics, Bordeaux Population Health Research Center (INSERM U1219), Université de Bordeaux, France
[(2)] ISBA/LIDAM, Université catholique de Louvain, Louvain-la-Neuve, Belgium
E-mail for correspondence: quentin.le-coent@u-bordeaux.fr

Abstract: Surrogate endpoints can be required to carry out trials that would be unfeasible if based on true endpoint but must have been statistically validated prior to their use (Burzykowski, 2006). In this work we propose a new approach for surrogate validation when both the surrogate and the true endpoint are time-to-event. This approach is based on the causal framework of mediation analysis and is developped for meta-analytic data. It uses a joint regression model for the hazard functions of both endpoints (Rondeau, 2007). The meta-analytic nature of the data is taken into account by using shared random effects at both the individual and trial levels. The mediation analysis enables one to study the decomposition of the total effect of the treatment on the true endpoint into a direct effect and an indirect effect through the surrogate (Tchetgen, 2011). The indirect effect of the treatment on the true endpoint through the surrogate is allowed as the composition of a direct effect of the treatment on the surrogate and a direct effect of the surrogate on the true endpoint. A measure of surrogacy is taken as the ratio of indirect effect over total effect. We applied this method for the assessment of the desease-free survival as a surrogate of the overall survival for adjuvant chemoterapy in the context of resectable gastric cancers.
Key words: Surrogacy; Mediation Analysis; Joint Modeling; Meta-analysis

Burzykowski T., Molenberghs G. and Buyse M. (Eds.) (2006). The evaluation of surrogate end-points. *Springer Science & Business Media*.
Rondeau V., *et al.* (2007). Joint frailty models for recurring events and death using maximum penalized likelihood estimation: application on cancer events. *Biostatistics*, 8(4), 708-721.

Tchetgen, E. J. T. (2011). On causal mediation analysis with a survival outcome. *The international journal of biostatistics*, 7(1).

Mediation analysis with a time-to-event outcome and time-varying mediator: an application to cystic fibrosis-related diabetes

Tanner Kamaryn T.[(1)], Keogh Ruth H.[(1)], Sharples Linda D.[(1)], Daniel Rhian M.[(2)]
[(1)] The London School of Hygiene and Tropical Medicine, Dept of Medical Statistics, London, UK
[(2)] Cardiff University, Division of Population Medicine, Cardiff, UK
E-mail for correspondence: kamaryn.tanner1@lshtm.ac.uk

Abstract: Cystic fibrosis-related diabetes (CFRD) is a common comorbidity of cystic fibrosis (CF). CFRD negatively affects survival but the mechanism is not well understood. We investigate different approaches to estimating how much of the impact of CFRD on mortality is mediated by lung function using the UK CF Registry. The time-dependent nature of the mediator and some covariates poses identification and analytical challenges. The time-to-event outcome provides an additional definitional challenge since the length of the mediator process differs under hypothetical assignment to different exposures. We compare two recently proposed methods. Aalen *et al.* (2018) described a method based on exposure splitting, combining a sequential linear model, an additive hazards model and a mediational g-formula. Vansteelandt *et al.* (2019) use a nested counterfactual framework that allows for time-dependent confounding and accommodates Cox regression or flexible parametric outcome models. We discuss the results of both methods when applied to the UK CF Registry, including the sensitivity of each to model misspecification and data availability. Finally, we outline areas for future methodological developments.
Key words: Cystic fibrosis; Mediation analysis; Longitudinal data; Survival outcome

Aalen, O.-O., Stensrud M.-J., Didelez V., Daniel R., Røysland K., Strohmaier S. (2019). Time-dependent mediators in survival analysis: modelling direct and indirect effects with the additive hazards model. *Biometrical Journal* 1-18.

Vansteelandt S., Linder M., Vandenberghe S., Steen J., Madsen J. (2019). Mediation analysis of time-to-event endpoints accounting for repeatedly measured mediators subject to time-varying confounding. *Statistics in Medicine* (38) 4828-4840.

G-computation and machine learning for causal inference

Chatton Arthur[1,2,*], , Le Borgne Florent[1,2,*], Foucher Yohann[1,3]
[1] INSERM UMR 1246 - SPHERE, Nantes University, Tours University, Nantes, France
[2] IDBC-A2COM, Pacé, France
[3] Nantes University Hospital, Nantes, France
[*] Co-first authors
E-mail for correspondence: arthur.chatton@univ-nantes.fr

Abstract: While machine learning approaches are increasingly used in prediction, their applications for causal inference are more recent. We propose an approach combining machine learning and G-computation (Robins, 1986) to estimate the causal effect of a binary exposure on a binary outcome. We evaluated and compared, through a simulation study, the performances of penalized logistic regressions, neural network, support vector machine, boosted classification, regression trees, and an ensemble method called super learner (van der Laan *et al.*, 2007). We proposed six different scenarios including various sample sizes and relationships between covariates, binary exposure, and binary outcome. We reported that, used in a G-computation approach to estimate the individual outcome probabilities, the super learner tended to outperform other approaches both in terms of bias and variance, especially for small sample sizes. Support vector machine also resulted in performant properties, albeit the mean bias was slightly higher compared to the super learner. In conclusion, the use of machine learning approaches can be pertinent to draw causal inference. Contrary to a preconception, this is true even for sample constituted by several hundred subjects, as in the majority of medical studies. The G-computation with the super learner is available in the R package RISCA.
Key words: Causal inference; G-computation; Model specification; Super learner; Simulation study.

Robins J.-M. (1986). A new approach to causal inference in mortality studies with a sustained exposureperiod—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9), 1393–1512.

van der Laan M.-J., Polley E.-C., and Hubbard A.-E. (2007). Super learner. *Stat. Appl. Genet. Mol. Biol.*, 6(1), Article 25.

11.3 Large scale hypothesis testing

Chairperson: Jelle Goeman (Leiden University)

Cluster extent inference revisited: quantification and localization of brain activity

Jelle Goeman[(1)], Weeda Wouter[(2)], Monajemi Ramin[(1)], Chen Xu[(1,2)], Górecki Paweł[(3)]
[(1)] Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands
[(2)] Methodology and Statistics, Leiden University, Leiden, The Netherlands
[(3)] Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland
E-mail for correspondence: j.j.goeman@lumc.nl

Abstract: Cluster inference based on spatial extent thresholding is the most popular analysis method for finding activated brain areas in neuroimaging. However, the method has several well-known issues. While powerful for finding brain regions with some activation, the method as currently defined does not allow any further quantification or localization of signal. In this paper we repair this gap. We show that cluster-extent inference can be used (1.) to infer the presence of signal in anatomical regions of interest and (2.) to quantify the percentage of active voxels in any cluster or region of interest. These additional inferences come for free, i.e. they do not require any further adjustment of the alpha-level of tests, while retaining full familywise error control. We achieve this extension of the possibilities of cluster inference by an embedding of the method into a closed testing procedure, and solving the graph-theoretic k-separator problem that results from this embedding. The new method can be used in combination with random field theory or permutations. We demonstrate the usefulness of the method in a large-scale application to neuroimaging data from the Neurovault database.
Key words: fMRI; Spatial specificity paradox; Closed testing; Familywise error rate

Distribution-free complex hypothesis testing for single-cell RNA-seq differential expression analysis

Gauthier Marine[1,2,3], , Agniel Denis[5,6], Thiébaut Rodolphe[1,2,3,4], Godot Véronique[3,7], Hejblum Boris[1,2,3]
[1] Univ. Bordeaux, Inserm, Bordeaux Population Health Research Center,UMR 1219, Bordeaux, France
[2] INRIA Bordeaux Sud Ouest, Talence, France
[3] Vaccine Research Institute, Créteil, France
[4] CHU, Bordeaux, France
[5] Rand Corporation, Santa Monica (CA), USA
[6] Harvard Medical School, Boston (MA), USA
[7] Inserm, U955, Team 16, Univ. Paris-Est, Créteil, France
E-mail for correspondence: marine.gauthier@u-bordeaux.fr

Abstract: Single-cell RNA-seq (scRNA-seq) quantifies gene expression at the cell resolution. State-of-the-art methods for scRNA-seq Differential Expression Analysis (DEA) often rely on strong distributional assumptions that are difficult to verify in practice. Furthermore, while the increasing complexity of clinical and biological single-cell studies calls for greater tool versatility, the majority of existing methods only tackles the comparison between two conditions. We propose a novel, distribution-free, and flexible approach to DEA for single-cell RNA-seq data. This new method, called `ccdf`, tests the association of each gene expression with one or many variables of interest (that can be either continuous or discrete), while potentially adjusting on additional covariates. To test such complex hypotheses, `ccdf` uses a conditional independence test relying on the conditional cumulative distribution function, estimated through multiple regressions. `ccdf` includes an asymptotic test as well as a permutation test (when the number of observed cell is not sufficiently large). `ccdf` exhibits good statistical performance in various simulation scenarios considering complex experimental designs (i.e.beyond the two condition comparison), while retaining competitive performance with the state-of-the-art in a two condition benchmark.
Key words: single-cell; conditional cumulative distribution function; conditional independence test; differential expression analysis; distribution-free test.

A semiparametric approach for differential abundance analysis in microbiome experiments

Kodalci Leyla[(1)], Thas Olivier[(1,2,3)]
[(1)] Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Data Science Institute, Hasselt University, Hasselt, Belgium
[(2)] Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Gent, Belgium
[(3)] National Institute for Applied Statistics Research Australia (NIASRA), University of Wollongong, Wollongong, Australia
E-mail for correspondence: leyla.kodalci@uhasselt.be

Abstract: Microbiome data obtained from high-throughput sequencing are considered as compositional data, which is characterised by a sum-constraint. Hence, only ratios of observations are informative. Furthermore, microbiome data are overdispersed and have many zero abundances. Many compositional data analysis methods make use of log ratios of the components of the observation vector. However, the many zero abundances cause problems when calculating ratios and logarithms.

In this work, we focus on the identification of taxa that are differentially abundant between two groups. We have developed a semiparametric method targeting the probability that the outcome of one taxon is smaller than the outcome of another taxon (a probabilistic index). The estimation of this probability only requires information about the pairwise ordering of the taxa, and hence zero observations cause no problems. Testing for differential abundance then reduces to testing that the probabilistic indexes are the same in the two treatment groups. We have constructed the semiparametric efficient estimator of the effect size parameter in the model, and a hypothesis test based on this estimator. Results from a simulation study indicate that our methods control the FDR at the nominal level and have good sensitivity compared to competitors.

Key words: differential abundance analysis; high-dimensional; large scale hypothesis testing, probabilistic index; rank method; semiparametric.

12. Contributed sessions 7-8

Thursday, April 8, 13:30-15:00

12.1 Spatial and environmental modelling

Chairperson: Richard Glennie (University of St Andrews)

Do's and don'ts of model comparison techniques
Vranckx Maren[(1)], Neyens Thomas[(1,2)], Faes Christel[(1)]
[(1)] UHasselt - Hasselt University, Data Science Institute (DSI), The Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), Diepenbeek, Belgium
[(2)] KU Leuven, The Interuniversity Institute for Biostatistics, and statistical Bioinformatics (I-Biostat),Leuven, Belgium
E-mail for correspondence: maren.vranckx@uhasselt.be

Abstract: Several model comparison techniques exist to select a best model from a set of candidate models. This study explores the performance of model comparison statistics among several Bayesian software packages that are often used for spatially discrete disease modelling: the deviance information criterion (DIC), the Watanabe-Akaike information criterion (WAIC) and the log marginal predictive likelihood (LMPL). We focus on the software packages CARBayes, OpenBUGS, NIMBLE and Stan, in which we fit Poisson models to disease incidence outcomes with intrinsic conditional autoregressive, convolution conditional autoregressive and log-normal error terms. From three data analyses, that differ in the number of areal units

and disease prevalence, we learn important disparities in model selection. Based on these conclusions, we provide recommendations on the optimal use of model comparison statistics for all kind of applications.
Key words: DIC; Disease mapping; LMPL; Software packages; WAIC

Vranckx M., Neyens T., and Faes C. The (in)stability of Bayesian model selection criteria in dis-ease mapping. *Manuscript submitted for publication.*

Scan statistics for multiple spatial clusters to investigate geographical disparities of air pollution data

Ahmed Mohamed-Salem[1,2], Genin Michael[1], Marbac Matthieu[3]
[1] Univ. Lille, CHU Lille, ULR 2694 - METRICS: Évaluation des technologies de santé et des pratiques médicales, F-59000 Lille, France.
[2] Socit d'Alicante, Seclin, France.
[3] Univ. Rennes, Ensai, CNRS, CREST - UMR 9194, F-35000 Rennes, France
E-mail for correspondence: matthieu.marbac-lourdelle@ensai.fr

Abstract: Motivated by the investigation of spatial disparities of air pollution data, we develop a new method for detecting multiple spatial clusters and testing their significance. In this context, a geographical cluster has a general parametric shape (allowing for elliptic and rectangular clusters) and is defined by a change of the conditional mean assessed by a regression model of the target variable (here the concentration of particles PM10) given the spatial coordinates and other covariates. We introduce a numerical approach to detect the potential clusters avoiding the use of suboptimal or exhaustive (but unfeasible for large samples) approaches. We present a new Monte-Carlo procedure used for assessing quantiles of the scan statistics under the null hypothesis. We adress the consistency and asymptotic efficiency of the procedure. Contrary to the standard approach, the method permits all the alternative hypothesis to be detected. Finally, the procedure provides a data-driven selection of the number of clusters. The proposed approach is used for analyzing air pollution data given by the European Environmental Agency that the standard spatial scan methods fail to analyze due to the data dimension and to their suboptimality for detecting some alternative hypothesis that leads to fewer but far away larger detected clusters.
Key words: Cluster detection; Quasi likelihood; Spatial scan statistics.

Statistical Downscaling for the Fusion of In-river, Drone and Satellite Water Quality Data in a River Network

Wilkie Craig[(1)], Ray Surajit[(1)], Scott Marian[(1)], Miller Claire[(1)], Sinha Rajiv[(2)], Bowes Mike[(3)]
[(1)] University of Glasgow, Glasgow, United Kingdom.
[(2)] Indian Institute of Technology Kanpur, Kanpur, India.
[(3)] UK Centre for Ecology & Hydrology, Wallingford, United Kingdom
E-mail for correspondence: craig.wilkie@glasgow.ac.uk

Abstract: Rivers are a vital part of the hydrosphere, but our understanding of spatial and temporal patterns in water quality is often limited due to a lack of available data. River health can change abruptly in space and time due to impacts of pollutants from industry, farming and human populations, and knowledge of these changes is needed to inform mitigation efforts. High resolution hyperspectral satellite and drone data are needed to provide this knowledge, but they must be combined with laboratory-analysed data from in-river samples to ensure validity. This talk presents a statistical downscaling method for the fusion of river data from multiple sources with different spatio temporal support, to provide a fully calibrated high resolution data product enabling predictions to be made at any spatial location along the river or at any timepoint. This work extends the method of Wilkie *et al.* (2019) to river data, with data at each location treated as observations of smooth functions over time, while spatially-varying coefficients regression (Gelfand *et al.*, 2003) accounts for smooth spatial changes in relationships between the data sources. An illustration of the method will be presented, namely an application to simulated water quality data for the Ramganga river in northern India.
Key words: Rivers; Downscaling; Fusion; Satellite; Spatiotemporal.

Gelfand A.-E., Kim H.-J., Sirmans C.F. and Banerjee S. (2003). Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98(462), 387–396.

Wilkie C.J., Miller C.A., Scott E.M., O'Donnell R.A., Hunter P.D., Spyrakos E., Tyler A.N. (2019). Nonparametric statistical downscaling for the fusion of data of different spatiotemporal support. *Environmetrics*, 30(3), e2549.

A flexible dynamic occupancy model to estimate non-linear effects in Odonata population dynamics across the UK

Belmont Jafet[(1)], Miller Claire[(1)], Scott Marian[(1)], Wilkie Craig[(1)], August Tom[(2)], Taylor Philip[(3)], Brooks Steve[(4)]
[(1)] School of Mathematics and Statistics, University of Glasgow, Glasgow, U.K.
[(2)] UK Centre for Ecology and Hydrology, Wallingford, U.K.
[(3)] UK Centre for Ecology and Hydrology, Edinburgh, U.K.
[(4)] Department of Life Sciences, Natural History Museum, London, UK
E-mail for correspondence: j.belmont-osuna.1@research.gla.ac.uk

Abstract: A major task in ecological studies is to account for the various sources of uncertainty that occur at different spatial and temporal scales to provide a more accurate description of how biodiversity responses are affected by environmental changes (Cressie *et al.*, 2009). However, this is not an easy task and very often ecological data are prone to an observational error induced by the species imperfect detection. Over the last decade, the increasing awareness of accounting for species imperfect detection in ecological studies has led to the development of different species distribution models (Elith and Leathwic, 2009; Devarajan *et al.*, 2020). Particularly, dynamic occupancy models have proven to be a powerful tool to estimate temporal changes in species occurrences by incorporating populations' extinction and colonization dynamics while accounting for false absences (Rushing *et al.*, 2019). Thus, in this work, we propose a multiple species flexible dynamic model that incorporates a non-linear effect on the colonization and survival dynamics to estimate Odonata occupancy patterns in waterbodies across the UK. Data were provided by Hydroscape (web:hydroscapeblog.wordpress.com), a project investigating how anthropogenic stressors and connectivity interact to influence biodiversity in UK freshwaters. We discuss issues regarding study designs and new approaches to modelling and collection of new data.
Key words: Detectability, Colonization, Flexible Model, Occupancy, Survival

Cressie, N., Calder, C. A., Clark, J. S., Hoef, J. M. V., and Wikle, C. K. (2009). Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecological Applications*, 19, 553-–570.

Devarajan, K., Morelli, T. L., and Tenan, S. (2020). Multispecies occupancy models: review, roadmap, and recommendations. *Ecography*, 43, 1612–1624.

Elith, J., and Leathwick, J. R. (2009). Species distribution models: ecological explanation and prediction across space and time. *Annual review of ecology, evolution, and systematics*, 40, 677—697.

Rushing, C. S., Royle, J. A., Ziolkowski, D. J., and Pardieck, K. L. (2019). Modeling spatially and temporally complex range dynamics when detection is imperfect. *Scientific reports*, 9, 1–9.

Wilkie C.J., Miller C.A., Scott E.M., O'Donnell R.A., Hunter P.D., Spyrakos E., Tyler A.N. (2019). Nonparametric statistical downscaling for the fusion of data of different spatiotemporal support. *Environmetrics*, 30(3), e2549.

12.2 Penalized estimation methods

Chairperson: Anestis Touloumis (The University of Brighton)

Network modeling with covariates for high-dimensional longitudinal data
Pazira Hassan[(1)], Ciocanea-Teodorescu Iuliana[(2)], Van Wieringen Wessel[(1,3)]
[(1)] Department of Epidemiology and Biostatistics, Amsterdam UMC, location VUmc, The Netherlands
[(2)] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
[(3)] Department of Mathematics, VU University Amsterdam, Amsterdam, The Netherlands
E-mail for correspondence: w.vanwieringen@amsterdamumc.nl

Abstract: The longitudinal data containing high-dimensional outcomes and covariates are prevalent in a wide range of scientific disciplines, including healthcare and medicine. Alongside the characterization of the individual's traits molecularly (as outcomes), clinical information of the individual (as covariates) could be available, which may be time-varying (e.g. pre-post treatment indicator) or constant (e.g. age/gender). In this work, we developed a network model including the covariates effects to detect changes in networks. Indeed, we are interested in potential effect modification by covariates in the relation among the individual's traits, e.g. the cohesion of the molecular entities. To investigate theses effect modifications, using the penalized mixed models as prior information from networks reconstructed from observation studies, the (fixed and random effects) parameters of the model were estimated by the generalized (fused) ridge penalties. The efficacy and performance of the proposed network model (against networks with no covariates effects) is evaluated in the simulation and application studies.
Key words: network; high-dimensionality; longitudinal data; penalized mixed model.

Fast marginal likelihood estimation of penalties for group-adaptive elastic net

Van Nee Mirrelijn[(1)], Van De Brug Tim[(1)], Van De Wiel Mark[(1,2)]
[(1)] Epidemiology and Data Science, Amsterdam University Medical Centers, Amsterdam, The Netherlands
[(2)] MRC Biostatistics Unit, University of Cambridge, Cambridge, UK
E-mail for correspondence: m.vannee@amsterdamumc.nl

Abstract: Nowadays, clinical research routinely uses omics, such as gene expression, for predicting clinical outcomes or selecting markers. Additionally, so-called co-data are often available, providing complementary information on the covariates, like groups of genes corresponding to pathways. Elastic net is widely used for prediction and covariate selection. Group-adaptive elastic net learns from co-data to improve prediction and selection, by penalising important groups of covariates less than other groups. Existing methods are, however, computationally expensive. Here we present a fast method for marginal likelihood estimation of group-adaptive elastic net penalties for generalised linear models. The method uses a low-dimensional representation of the Taylor approximation of the marginal likelihood and its first derivative for group-adaptive ridge penalties, to efficiently estimate these penalties. Then we show by using asymptotic normality of the linear predictors that the marginal likelihood for elastic net models may be approximated well by the marginal likelihood for ridge models. The ridge group penalties are then transformed to elastic net group penalties by using the variance function. The method allows for overlapping groups and unpenalised variables. We demonstrate the method in a cancer genomics application. The method substantially decreases computation time while outperforming or matching other methods by learning from co-data.
Key words: Clinical prediction; Empirical Bayes; Omics; Penalised generalised linear models; Prior information.

Are reconstructed molecular networks reproducible?

Van Wieringen Wessel[(1)] (Joint work with Chen Yao)
[(1)] Department of Epidemiology & Data Science, Amsterdam UMC, Amsterdam, The Netherlands
[(2)] Department of Mathematics, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
E-mail for correspondence: w.vanwieringen@amsterdamumc.nl

Abstract: Starting from the observation that the reconstruction of the cohesion among the variates of a multivariate random variable by means of Gaussian graphical model usually takes only sampling variation into account, we point out the consequences of this practice for the reconstruction of the underlying conditional independence graph. When replicates are included in the study, these consequences are overcome by the separation of sampling from other sources of variation. Hereto a simple 'signal+noise' model for the description of the multivariate data has been put forward. We present a penalized EM algorithm for the estimation of the model's parameters, alongside a discussion of cross-validation for choosing the penalty parameter(s). Through simulation we investigate how much is won by the inclusion of replicates, and compare the presented method to obvious alternatives. Finally, in an illustration using oncogenomics studies with replicates, we further investigate the effect of ignoring variation due to other sources than sampling variation and assess the reproducibility of the reconstruction of the conditional independence graph.
Key words: Conditional Independence Graph; Inverse covariance; Network; Reproducibility; Ridge penalty.

van Wieringen W.-N., Chen, Y. (2020). Penalized estimation of the Gaussian graphical model from data with replicates. *submitted*.

Penalized regression calibration: a method for the prediction of survival outcomes using complex longitudinal and high-dimensional data

Mirko Signorelli[(1)], Pietro Spitali[(2)], Roula Tsonaka[(1)]
[(1)] Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, NL
[(2)] Department of Human Genetics, Leiden University Medical Center, Leiden, NL
E-mail for correspondence: m.signorelli@lumc.nl

Abstract: Longitudinal and high-dimensional measurements have become increasingly common in biomedical research. However, methods to predict survival outcomes using covariates that are both longitudinal and high-dimensional are currently missing. We propose penalized regression calibration (PRC, Signorelli *et al.*, 2021), a method that can be employed to predict survival in such situations. The method is implemented in the R package pencal, available from CRAN.

PRC comprises three modelling steps: first, the trajectories described by the longitudinal predictors are flexibly modelled through the specification of multivariate latent process mixed models. Second, subject-specific summaries of the longitudinal trajectories are derived from the fitted mixed effects models. Third, the time to event outcome is predicted using the subject-specific summaries as covariates in a penalized Cox model.

To ensure a proper internal validation of the fitted PRC models, we furthermore develop a cluster bootstrap optimism correction procedure that allows to correct for the optimistic bias of apparent measures of predictiveness.

After studying the behaviour of PRC via simulations, we conclude by illustrating an application of PRC to data from an observational study that involved patients affected by Duchenne muscular dystrophy, where the goal is predict time to loss of ambulation using longitudinal blood biomarkers.

Key words: survival analysis; risk prediction modelling; longitudinal data analysis; high-dimensionality; optimism correction.

Signorelli M., Spitali P., Al-Khalili Szigyarto C., The MARK-MD Consortium, Tsonaka R. (2021). Penalized regression calibration: a method for the prediction of survival outcomes using complex longitudinal and high-dimensional data. arXiv preprint number: arXiv:2101.04426.

13. Invited session 2

Chairperson: Olivier Thas (Hasselt University)

Thursday, April 8, 15:20-16:20

Infectious diseases

Statistical Preparedness in a pandemic
Deborah Ashby[(1)]
[(1)] School of Public Health, Imperial College London

Abstract: Faced with a new disease such as COVID-19, we need to be able to diagnose it, describe its prevalence in the population, map its time course and spatial distribution, learn about what treatments work, develop and test a vaccine, and communicate with decision-makers and the population. Each of these present challenging statistical issues. We describe how the UK mounted responses to these, including large scale prevalence surveys, and platform adaptive trials that informed policy and clinical practice and show how a rapid response is possible where the statistical community is well-prepared.

Leveraging random effects to estimate the impact of non-pharmaceutical interventions on epidemic dynamics across French regions

Mélanie Prague, Annabelle Collin, Linda Wittkop, Dan Dutartre, Quentin Clairon, Philippe Moireau, Rodolphe Thiébaut, Boris Hejblum*

Abstract: We developed a multi-level model of the French COVID-19 epidemic at the regional level. We rely on a global extended Susceptible-Exposed-Infectious-Recovered (SEIR) mechanistic model as a simplified representation of the average epidemic process, with the addition of region specific random effects. Combining several French public datasets on the early dynamics of the epidemic, we estimate region-specific key parameters conditionally on this mechanistic model through Stochastic Approximation Expectation Maximization (SAEM) optimization using Monolix software. We thus estimate the basic reproductive numbers by region before lockdown, attack rates (i.e. percentages of infected people) over time per region, and the impact of nationwide lockdown on the infection rate. These results confirm the low population immunity, the strong effect of the lockdown on the dynamics of the epidemics. This methodology can also be applied to assess the impact of various other non pharmaceutical interventions such as school closing or curfews.

Mathematical and statistical epidemiology of COVID-19 in Belgium used to inform decision making

Niel Hens[(1)]
[(1)] Hasselt University, Center for Statistics - CenStat, Belgium.
E-mail for correspondence: niel.hens@uhasselt.be

Abstract: In this presentation, I will give an overview of the work that was done by the SIMID group (www.simid.be) to inform policy makers on how the COVID-19 pandemic spread(s) in Belgium. During my presentation, I will highlight the importance of collecting data and proper use of statistical and mathematical methodology, ie doing research in a crisis situation.

14. Contributed sessions 9-10

Thursday, April 8, 16:30-18:00

14.1 Longitudinal data analysis

Chairperson: Kamran Safi (Max Planck Institute for Ornithology)

Individual Reference Intervals for Personalized Interpretation of Clinical and Physiological Measurements

Pusparum Murih[1,2], Ertaylan Gökhan[2], Thas Olivier[1,3,4]
[1] Data Science Institute, Hasselt University, Hasselt 3500, Belgium
[2] Flemish Institute for Technological Research (VITO), Mol 2400, Belgium
[3] Department of Data Analysis and Mathematical Modelling, Ghent University, Ghent 9000, Belgium
[4] National Institute for Applied Statistics Research Australia (NIASRA), Wollongong 2500, NSW, Australia
E-mail for correspondence: murih.pusparum@uhasselt.be

Abstract: A reference interval (RI) of a clinical or physiological outcome refers to the range of outcomes that is expected in a healthy population. This interval is widely used in daily clinical practice for interpreting laboratory tests: when the outcome falls outside the RI, the physician will consider further examination. Whereas the conventional RI refers to a single population of healthy subjects, we argue that each individual may have different biological traits and therefore, may have its own Individual Reference Interval (IRI). We have developed methods for the estimation of IRIs when

time series data on multiple subjects are available. In a first approach we used linear quantile mixed models (LQMM) to separately estimate the lower and the upper bounds of the IRIs. We have extended this method for simultaneously estimating the two bounds by constructing a joint LQMM for the lower and upper quantile. Parameter estimation is based on the asymmetric Laplace distribution for a likelihood function construction, and a Monte-Carlo Expectation-Maximization (EM) algorithm. The methods' performance is evaluated in a simulation study. Finally, we demonstrate the validity of the proposed methods on real life data including several clinical and physiological measurements collected within the VITO IAM Frontier study.

Key words: Reference interval; Linear quantile mixed models; EM algorithm; Personalized health.

Combined shrinkage of fixed and random effects in linear mixed models using empirical Bayes

Amestoy Matteo[(1)], Van De Wiel Mark[(2)], Van Wieringen Wessel[(1,2)]
[(1)] Department of Epidemiology and Data Science, Amsterdam University Medical Center, Amsterdam, The Netherlands
[(2)] Department of Mathematics, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
E-mail for correspondence: m.amestoy@amsterdamumc.nl

Abstract: The ExposomeNL consortium aims to unravel the effect of individual and exposure-related variables on cardiovascular-related outcomes. Hereto longitudinal cohort data, combining individual-level medical records with exposure data, are available. Repeated measures over time and shared spatial information generate a complex correlation structure that is of substantive interest. While standard applications of linear mixed models (LMM) limit the number of random-effects variables, we use a high (medium) dimensional design matrix to fully capture this correlation structure. However, the high (medium) dimensionality of both the fixed- and random-effects compromise the stability of the associated estimators. We add a prior distribution to shrink both estimators. The prior's hyperparameters are estimated using an Empirical Bayes method, with an exact computation of the Hessian matrix in the Laplace approximation of the marginal likelihood. We compare the performance of our method to standard LMM algorithms using simulated data. We show that we capture complex correlation structures and improve the accuracy of the estimates. This methodology is then applied to Exposome data to analyse the combined effect of individual and exposure related variables, unravelling hitherto unknown patterns in the data that are overlooked by traditional methods.
Key words: Longitudinal data; High dimensionality; Covariance matrix; Regularization.

Individual dynamic prediction of clinical endpoint from large dimensional longitudinal biomarker history: a landmark approach

Devaux Anthony[(1)], Genuer Robin[(1,2)], Karine Pérès[(1)], Proust-Lima Cécile[(1)]
[(1)] Univ. Bordeaux, INSERM, BPH, U1219, F-33000 Bordeaux, France
[(2)] INRIA Bordeaux Sud-Ouest, Talence, France
E-mail for correspondence: anthony.devaux@u-bordeaux.fr

Abstract: The individual data collected throughout patient follow-up constitutes crucial information for assessing the risk of a health event. Some statistical methods were proposed to compute dynamic individual predictions from longitudinal information. However, they hardly handle a large number of markers, partly due to numerical issues. By relying on a landmark approach, we propose a methodology for computing individual dynamic predictions that may take into account a very large number of repeated markers over time while remaining computationally feasible. We model each marker trajectory up to the landmark time, and derive summary variables that best capture the individual trajectories. The summaries and additional covariates are then included in different prediction methods to predict the event from the landmark time. As we need to handle a possibly large dimensional history, we rely on machine learning methods adapted to survival data, namely regularized regressions and survival random forests, and show how they can be combined into a superlearner. We demonstrate in a simulation study the benefits of machine learning survival methods, especially in the case of multiple and/or nonlinear relationships between the predictors and the event. We also illustrate the methodology to predict death in the general elderly population at different ages.
Key words: Individual prediction; Landmark; Longitudinal data; Survival data; Machine learning methods.

Exploring disease progression using a latent class approach for multiple longitudinal markers and event history: example with Multi-System Atrophy

Proust-Lima Cécile[(1)], Saulnier Tiphaine[(1)], Pavy-Le Traonc Anne[(2,3)], Rascol Olivier[(2,4)], Meissner Wassilios G.[(5,6,7)], Philipps Viviane[(1)], Foubert-Samier Alexandra[(1,5,6)]
[(1)] Univ. Bordeaux, INSERM, BPH, U1219, F-33000 Bordeaux, France
[(2)] French Reference Centre for MSA, University Hospital Toulouse, Toulouse, France
[(3)] Institut des Maladies Métaboliques et Cardiovasculaires, Inserm U 1048, Toulouse University, Toulouse, France
[(4)] Inserm, Toulouse University and CHU Toulouse, Clinical Investigation Center CIC 1436 and Departments of Neurosciences and Clinical Pharmacology, Toulouse, France
[(5)] French Reference Centre for MSA, University Hospital Bordeaux, Bordeaux, France
[(6)] Institut des Maladies Neurodégénératives, CNRS, UMR 5293, Bordeaux University, Bordeaux, France
[(7)] Dept. Medicine, University of Otago, Christchurch, and New Zealand Brain Research Institute, Christchurch, New Zealand
E-mail for correspondence: cecile.proust-lima@inserm.fr

Abstract: Some diseases are characterized by numerous markers of progression. Although not specific to, this is particularly the case in neurodegenerative diseases where pathological brain changes may induce multiple clinical signs on which the progression of a patient is assessed. For instance, Multiple System Atrophy (MSA), a rare neurodegenerative synucleinopathy, is characterised by various combinations of progressive autonomic failure and motor dysfunction (parkinsonism and cerebellar ataxia), and by a very poor prognosis with a median survival of a few years after diagnosis. Describing the progression of such complex and multi-dimensional diseases is particularly difficult. One has to simultaneously account for the assessment of multivariate markers over time, the occurrence of clinical endpoints, and the highly suspected heterogeneity between patients which is partly due to the difficulty to formally diagnose the disease. Yet, such description is crucial for understanding the natural history of the disease, stage patients diagnosed with the disease, unravel subphenotypes, and predict the prognosis. Through the example of MSA progression, we show how a latent class approach can help describe complex disease progression measured by multiple repeated markers and clinical endpoints,and identify subphenotypes for exploring new pathological hypotheses.
Key words: Disease progression; Joint models; multivariate longitudinal data; Heterogeneity.

14.2 Bayesian methods

Chairperson: Boris Hejblum (Bordeaux Population Health, INSERM)

Fast approximate inference for multivariate longitudinal data
Hughes David[(1)], Garcia-Finana Marta[(1)], Wand Matt P[(2)]
[(1)] Department of Health Data Science, Universityof Liverpool, Liverpool, UK
[(2)] School of Mathematical and Physical Sciences, University of Technology Sydney, Australia
E-mail for correspondence: dmhughes@liverpool.ac.uk

Abstract: Collecting information on multiple longitudinal outcomes is increasingly common in many clinical settings. In many cases it is desirable to model these outcomes jointly. However, in large datasets, with many outcomes, computational burden often prevents the simultaneous modelling of multiple outcomes within a single model.

We develop a mean field variational Bayes algorithm, to jointly model multiple Gaussian, Poisson or binary longitudinal markers within a multivariate generalised linear mixed model.

Through simulation studies and clinical applications (in the fields of sight threatening diabetic retinopathy and primary biliary cirrhosis) we demonstrate substantial computational savings of our approximate approach when compared to a standard Markov Chain Monte Carlo, while maintaining good levels of accuracy of model parameters.

This talk will give a brief overview of variational Bayes approaches and discuss some of the algebraic tools needed to streamline model estimation. We will also detail some of the factors that can affect performance of variational Bayes approximations.
Key words: Variational Bayes; Longitudinal Data; mixed models; Bayesian modelling.

Two-dimensional Intrinsic Gaussian Markov Fields in blood pressure data

Spyropoulou Maria-Zafeiria[(1)], Bentham James[(1)]
[(1)] Department of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, UK
E-mail for correspondence: mzs2@kent.ac.uk

Abstract: Raised blood pressure is a key risk factor for non-communicable diseases, and is estimated to affect 1.13 billion people worldwide. A Bayesian hierarchical model for the variables of diastolic bloodpressure (DBP), systolic blood pressure (SBP) and especially the interaction (INT) of these two is proposed. We separate the globe into groups of countries whereas each country is a member of a region and super-region. This structure allows to borrow strength across regions and super-regions when no data exist. Within each country, data are correlated temporally and within each region and super-region data have temporal and between-countries correlation. A two dimensional second order Intrinsic Gaussian Markov Fields (IGMRF) will be used as a covariance matrix in a prior for DBP and SBP accounting for the interaction of these two as well. Hence, every possible combination between the years of DBP and SBP variables will be taken into account having the possibility to observe their INT. For the computational process, we use canonical parametrisation for the Block-Metropolis' updates and Cholesky factorisation for the Gibbs' sampler updates. Age, diet types, urbanization and studies'coverage are also included. Performance is demonstrated with simulation studies and real data.
Key words: Intrinsic Gaussian Markov Random Fields; B-splines; Block-Metropolis sampling; MCMC.

Danaei G., Finucane M.-M., Lin J.-K., *et al* (2011) National, regional, and global trends in systolic blood pressure since 1980: systematic analysis of health examination surveys and epidemiological studies with 786 country-years and 5·4 million participants. *Lancet*, 377(9765), 568-577.
Rue H. and Held L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Vol. 104. Monographs on Statistics and Applied Probability. London: Chapman & Hall.
Yue Yu and Speckman Paul L. (2010). Nonstationary spatial Gaussian Markov random fields. *Journal of Computational and Graphical Statistics*, 19(1), 96-116.

The Statistical Properties of RCTs

Van Zwet Erik[(1)], Schwab Simon[(2,3)], Senn Stephen[(4)]
[(1)] Department of Biomedical Data Sciences, Leiden University Medical Center, the Netherlands
[(2)] Center for Reproducible Science, University of Zürich, Switzerland
[(3)] Epidemiology, Biostatistics and Prevention Institute, University of Zürich, Switzerland
[(4)] Statistical Consultant, Edinburgh, United Kingdom
E-mail for correspondence: E.W.van_Zwet@lumc.nl

Abstract: We abstract a "study" as a triple (β, b, s) where β is the parameter of interest, b is an unbiased, normally distributed estimate of β, and s is the standard error of b. We do not observe β, but we do observe the pair (b, s). We define the z-value $z = b/s$ and the signal-to-noise ratio SNR=β/s. Note that the z-value is the sum of the SNR and independent standard normal "noise". This means that the distribution of the z-value is the convolution of the distribution of the SNR with the standard normal density.

We have collected a very large sample of pairs (b, s) from randomized controlled trials (RCTs) in the Cochrane Database of Systematic Reviews. We used these pairs to estimate the distribution of the z-values. Next, we obtained the distribution of the SNRs by deconvolution. Since we already know the conditional distribution of the z-value given the SNR, we now have the joint distribution of the pair $(z$,SNR$)$.

Many important statistical quantities depend on (β, b, s) only through the pair $(z$,SNR$)$. In particular, the exaggeration ratio $|b|/|\beta|$ and the indicator variables for the events: $\{|b|/s > 1.96\}$, $\{b - 1.96s < \beta < b + 1.96s\}$ and $\{\text{sign}(b) \neq \text{sign}(\beta)\}$. These quantities are closely related to the type M (magnitude) error, achieved power, coverage and type S (sign) error, respectively. We have computed their distribution across the Cochrane database both unconditionally and conditionally on the observed z-value. We find that the achieved power is often low and the exaggeration is typically large. However, conditionally on statistical significance, the probability of a type S (sign) error appears to be quite low.

Key words: Power; coverage; type M error; type S error; Cochrane database.

Schwab S. (2020). Re-estimating 400,000 treatment effects from intervention studies in the Cochrane Database of Systematic Reviews (Data set). https://doi.org/10.17605/OSF.IO/XJV9G.
van Zwet E., and Gelman A. (2020). A Proposal for Informative Default Priors Scaled by the Standard Error of Estimates. http://arxiv.org/abs/2011.15037.
van Zwet E., Schwab S., and Senn S. (2020). The Statistical Properties of RCTs and a Proposal for Shrinkage. http://arxiv.org/abs/2011.15004/.

A Bayesian model for heterogeneous treatment effects on the additive risk scale in meta-analysis

Thomassen Doranne[(1)], Steyerberg Ewout[(1)], Le Cessie Saskia[(1)]
[(1)] Biomedical Data Sciences, Leiden University Medical Center, Leiden, Netherlands
E-mail for correspondence: d.thomassen@lumc.nl

Abstract: Faced with a newly diagnosed patient, clinicians consider which of the available treatments will provide the largest absolute risk reduction. To answer this question requires statistical methods that quantify heterogeneous 'personalized' treatment effects on the clinically relevant scale.

We propose a Bayesian (meta-)regression model for binary outcomes on the additive risk scale. The model was applied in single trial analysis, meta-analysis and network meta-analysis of 20 hepatitis C trials. We compared our model to two other approaches: an alternative additive risk model (Warn *et al.* (2002)) and a logistic model that transforms predictions back to the natural scale after regression (Chalkou *et al.* (2020)).

Some trials had cure rates close to 100%, illuminating the main differences. Our model is very sensitive at the boundaries of the risk parameter support $[0, 1]$, whereas patients with predicted risks close to 0 or 1 contribute little to posterior precision in the model by Warn *et al.* In such cases, the logistic model sometimes produced extreme effect estimates, leading to instability in the network setting.

Their respective characteristics make the compared models suitable for different analysis settings. Our proposed model contributes to the statistical methods to model heterogeneous treatment effects on the additive risk scale.

Key words: treatment effect, heterogeneity, trial, risk difference, meta-analysis, Bayesian methods, personalized medicine.

D. E. Warn, S. G. Thompson, and D. J. Spiegelhalter (2002). Bayesian random effects meta-analysis of trials with binary outcomes: methods for the absolute risk difference and relative risk scales. *Statistics in Medicine*, 21(11), 1601–1623.

K. Chalkou, E. Steyerberg, M. Egger, A. Manca, F. Pellegrini, and G. Salanti (2020). A two-stage prediction model for heterogeneous effects of many treatment options: application to drugs for multiple sclerosis. preprint, 04 2020. URL:https://arxiv.org/abs/2004.13464.

15. Invited session 3

Friday, April 9, 09:00-10:30

Chairperson: David Causeur (Institut Agro, Agrocampus Ouest)

Statistical modeling in movement ecology

Identifying stationnary phases in animal movement
Marie-Pierre Etienne[(1)], Julien Chiquet[(2)], Sophie Donnet[(2)], Adeline Samson[(3)]
[(1)] UMR IRMAR, Université de Rennes, Agrocampus Ouest, France
[(2)] MIA-Paris, UMR 518 AgroParisTech, INRAE, Université Paris-Saclay, Paris, France
[(3)] Laboratoire Jean Kuntzmann, Univ. Grenoble Alpes, F-38000 Grenoble, France
E-mail for correspondence: marie-pierre.etienne@agrocampus-ouest.fr

Abstract: Movement of organisms is one of the main mechanisms who govern relations between species. Advances in biologging open promising perspectives in the study of animal movements at numerous scales. It is now possible to record time series of animal locations over extended areas and long durations with a high spatial and temporal resolution. Such time series are rarely stationary as the animal may alternate between different movement phases. Those phases are assumed to be linked to some internal states of the animals. Movement models which include a hidden state driven by a Markov process are classically used to address this non stationarity and to assign a cluster to each identified phase. However the Markov assumption induces constraints on the distribution of the length

of the different movement phases. We explore an alternative to Markov model based on a segmentation clustering method using change point detection approach which can be combined with different movement models.

Key words: Movement Ecology; Change point detection; Euler approximation.

Modelling latent animal movement in distance sampling and spatial capture-recapture

Richard Glennie[(1)]
[(1)] Centre for Research into Ecological and Environmental Modelling, University of St. Andrews, St.Andrews, UK
E-mail for correspondence: rg374@st-andrews.ac.uk

Abstract: Distance sampling and spatial capture-recapture are statistical methods to estimate the number of animals in a wild population based on encounters between these animals and scientific detectors. Both methods estimate the probability an animal is detected during a survey, but do not explicitly model animal movement and behaviour. The primary challenge is that animal movement in these surveys is unobserved; one must average over all possible histories of each individual. In this talk, a general statistical model, with distance sampling and spatial capture-recapture as special cases, is presented that explicitly incorporates animal movement. An efficient algorithm to integrate over all possible movement paths is given to overcome the computational obstacles. The model is then applied to both simulated and real data to estimate abundance, detection, and movement jointly.
Key words: Distance sampling; capture-recapture; animal movement; hidden Markov model.

Glennie R. *et al* (2020). Incorporating animal movement into distance sampling. *Journal of the American Statistical Association*, 1-9.
Glennie R. *et al*. (2019). Open population maximum likelihood spatial capture-recapture. *Biometrics*, 75(4), 1345-1355.

Deciphering the traces of animal-environment interactions by interpreting position through time: The challenges of making sense of movement data from an ecologist's perspective

Kamran Safi[(1)]
[(1)] Max Planck Institute for Ornithology, Münich, Germany
E-mail for correspondence: ksafi@ab.mpg.de

Abstract: Movement is a biological characteristic of all living organisms across all scales. From molecules in and between cells, to entire species communities across continents, all living is tightly associated and maintained by movement. With the miniaturization of electronic devices many organismal subdisciplines in Biology are catapulted not only into an era of getting the means to record movement, but increasingly so at an unprecedented level of detail. While the amount of data accumulating grows exponentially, the development of the theory and statistical methods to address the research questions lag behind. Very much like the revolutionary transitions in the "omics" decades ago, movement ecology is facing new challenges in data management, visualisation and analysis that provide a rich and rewarding area for descriptive as well as inferential statistical methods and models.

I will present some of the recent developments and approaches in the field of movement ecology from the perspective of an ecologist rather than a statistician. I will try to address the difficult question of relating animals to the environment and understanding movement as a process that is an expression of this relationship. Empirical random trajectories and continuous time movement models represent two ends of a continuum of approaches used to overcome the challenges in reaching for the holy grail of movement ecology: predictive models of animal movement. A goal that seems to be in our reach if we continue to build synergies between the disciplines.

Predicting Marine Birds Foraging Behaviour with Deep Learning

Amédée Roy[(1)]
[(1)] Institute of Research for Development, UMR MARBEC, Montpellier
E-mail for correspondence: amedee.roy@ird.fr

Abstract: Seabirds are often considered as suitable indicators for the study of marine ecosystems, since their foraging strategies give us a real-time response to the complex dynamics of the ecosystem. In particular, by deploying sufficiently light GPS sensors on seabirds, it is possible to obtain their trajectories, to identify behaviors at sea and foraging areas. The objective of this work was therefore to infer seabird diving behaviour from GPS data using Deep Learning methods. More precisely, from a database of about 250 foraging trajectories derived from GPS data deployed simultaneously with pressure sensors for the identification of dives the idea was to train a deep network in a supervised manner for the prediction of seabird dives from GPS data only. In this work, two network architectures were compared (Fully Connected Network vs U-Network), and different trajectory representation were used (Time-series vs Distance Matrix). These approaches were also applied to two tropical seabird species with distinct diving behaviour (Boobies vs Cormorants). Finally, the impact of these methods on the estimation of dives distribution was evaluated.

16. Closing keynote presentation

Friday, April 9, 11:00-12:00

Chairperson: Carel Peeters (Wageningen University and Research)

Causal discovery from observational data

Mathias Drton[(1)]
[(1)] Technical University of Münich, Department of Mathematics, Munich, Germany

Abstract: Graphical causal models represent the joint distribution of a collection of observations in a convenient and accessible form in terms of a directed graph. The models are causal in the sense that they also furnish a model for the joint distribution in settings in which the considered system is subject to external interventions, thus providing a means to predict the effects of such interventions. Much work has gone into statistical methods that infer the graph underlying a causal model from observational data a problem frequently termed causal discovery. Such methods either infer a class of graphs that are equivalent in the face of merely observational data, or base themselves on additional assumptions that render the causal graph identifiable. Focusing on the setting of linear causal models, this talk will review key results in causal discovery and highlight our recent work (joint with Y. Samuel Wang) on using non-Gaussianity to learn causal graphs in high-dimensional settings as well as in the presence of latent confounders.