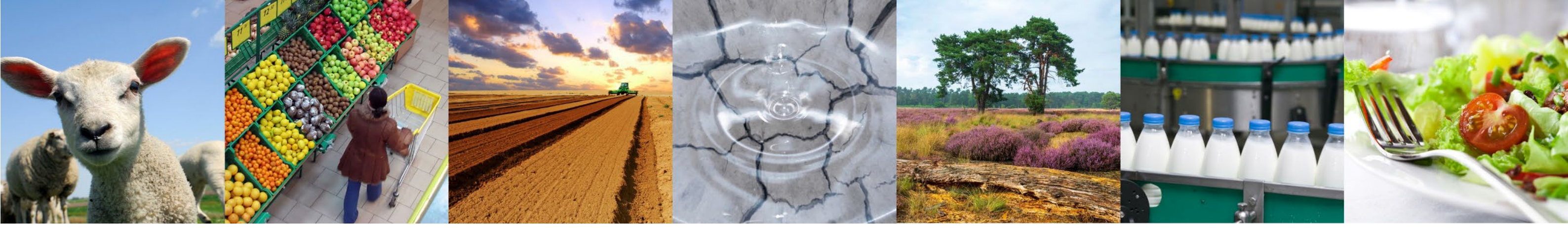
Channel Network Conference 2023

August 23 – 25, 2023

Wageningen

Book of Abstracts



Sponsors

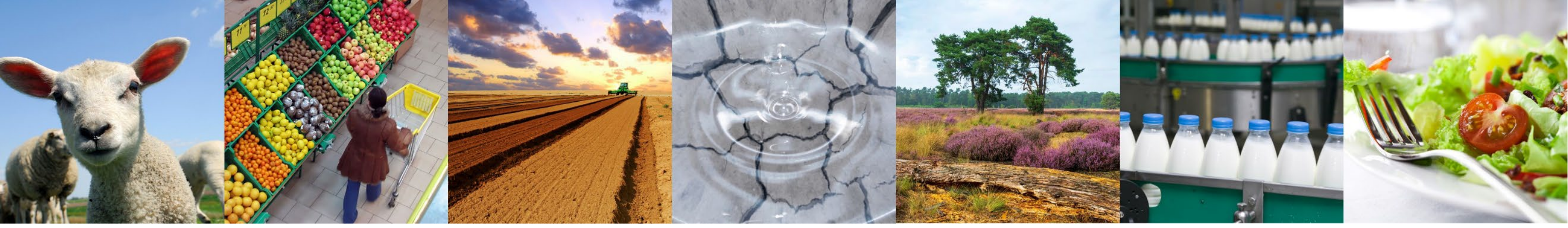
Biometris

Quantitative Methods brought to Life

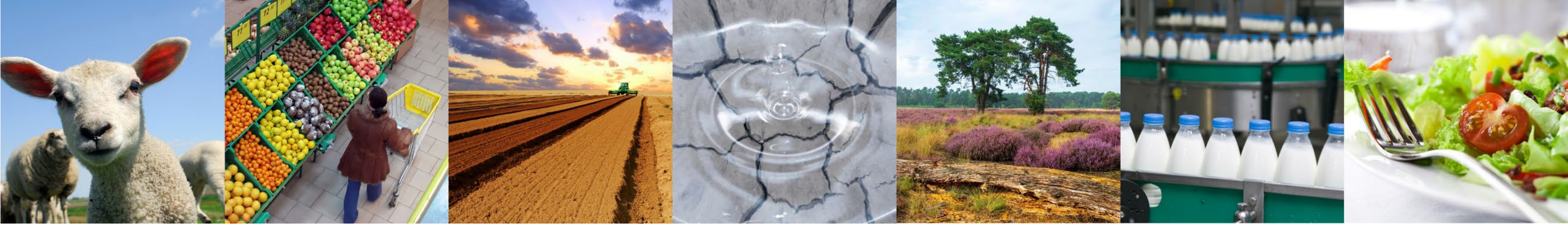
VSNi

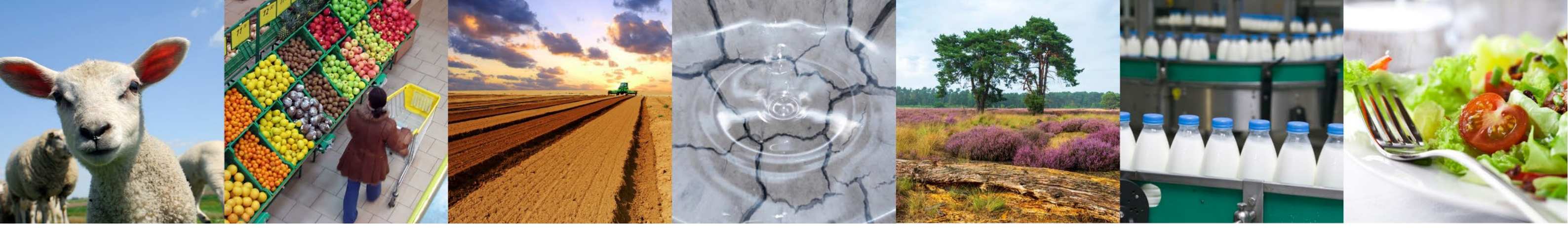
IBS

Afdeling Nederland



Contents

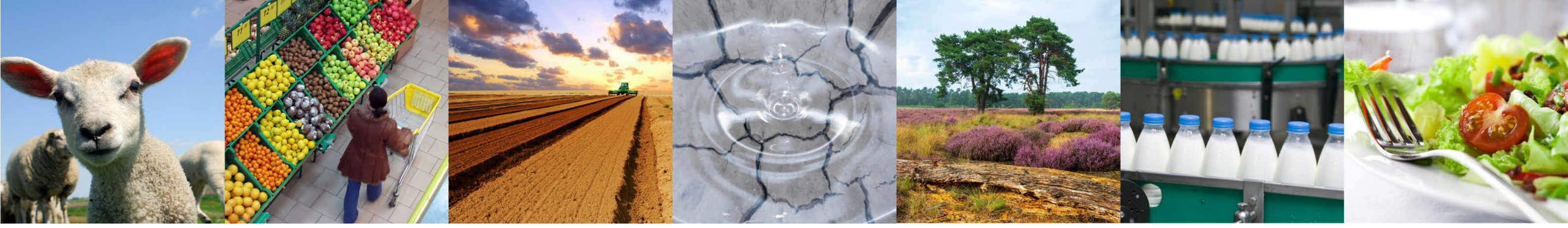


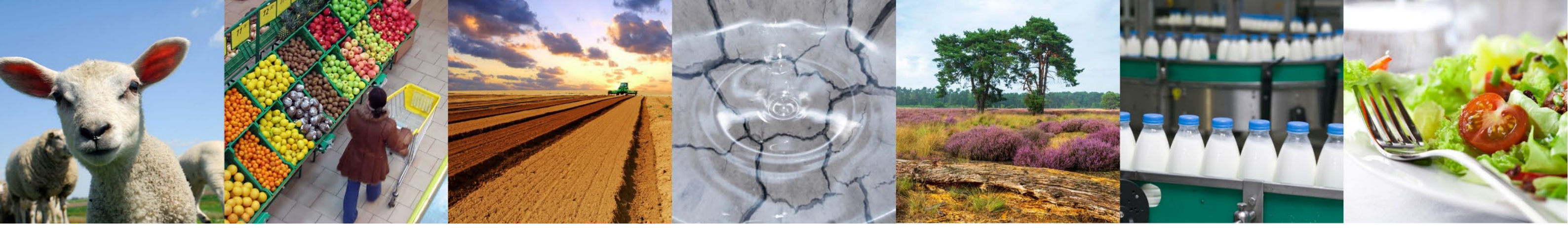


About CNC23

The Channel Network Conference is a biennial conference organized by the International Biometric Society (IBS) channel network. The channel network comprises the Belgian, French, British & Irish, and Dutch regions of the IBS. This conference gathers statisticians, mathematicians, and data scientists to discuss the latest methodology for the analysis of data in the biosciences, including agriculture, biomedical science, environmental science, and allied disciplines. It is a 3-day conference with short courses, invited and contributed sessions. In 2023, the Dutch region hosts the conference in lovely Wageningen, the Netherlands.

We are proud to announce that Aad van der Vaart (Delft Institute of Applied Mathematics) and Marc Chadeau-hyam (School of Public Health, Imperial College London) will respectively deliver the opening and closing keynote address. In addition, we have three invited sessions, three short courses, and many contributed sessions that reflect the wide range of methodological topics and application areas pursued by society members.





A word of welcome from the LOC

It is our pleasure, as co-chairs of the local organizing committee, to welcome you to Wageningen University & Research (WUR) for the 2023 Channel Network Conference (CNC23). During the 8th edition of the CNC we had to forego Paris due to COVID. We expressed the hope that we would be able to see you, the attendees, again without the aid of any screen. We are thus happy to return, for this 9th edition, to an in-person meeting.

The location of our meeting is Wageningen. While maybe not as appealing as Paris, it is a very interesting place. It is, among others, the birthplace of national agricultural education in The Netherlands with the (friendly) state takeover of a very modest community-run agricultural college. From these humble beginnings WUR has developed itself into an internationally recognized hub for technological research and education in the Life Sciences in the broadest sense. Despite its very modest size, Wageningen was transformed into a vibrant university town and one of the most international cities in the world.

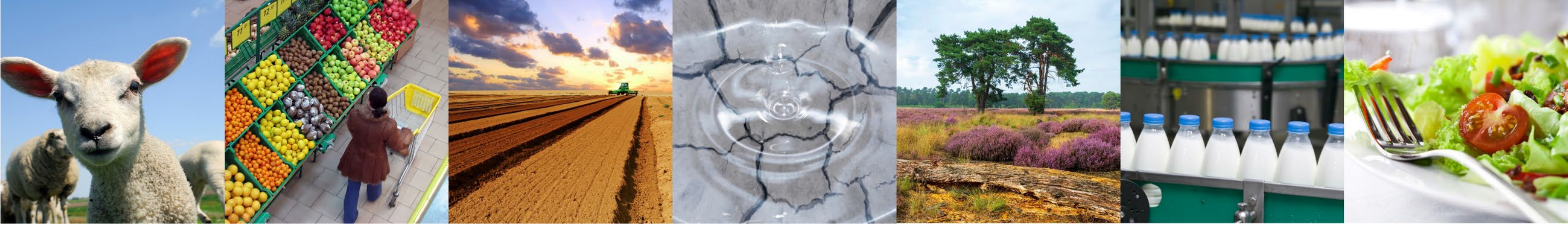
As such, we feel Wageningen embodies the heart of the CNC: togetherness through diversity. CNC traditionally attracts academics as well as industry professionals, theoreticians as well as practitioners, and the medically as well as the agriculturally inclined, in the noble art of data analysis. In this age of data-driven discovery, rife with buzzwords and in which information has become cheap in many ways, we understand, whatever our diverse inclinations, that modeling remains of pivotal importance. It is through modeling that we make sense of intricate patterns hidden within data, allowing us to help shape the future of healthcare, agriculture, and beyond.

Working in tandem with the Scientific Committee, we have provided an inspiring programme. Through interactive short courses, exciting invited speakers, and the breadth of the contributed sessions, CNC23 will explore the latest challenges and opportunities for the analysis of data in the biosciences. We are confident that this edition too will serve as a platform for fostering collaboration and knowledge exchange.

Thank you for coming to Wageningen. We hope you have a fruitful experience.

Carel F.W. Peeters
(co-chair LOC)

Jos Hageman
(co-chair LOC)



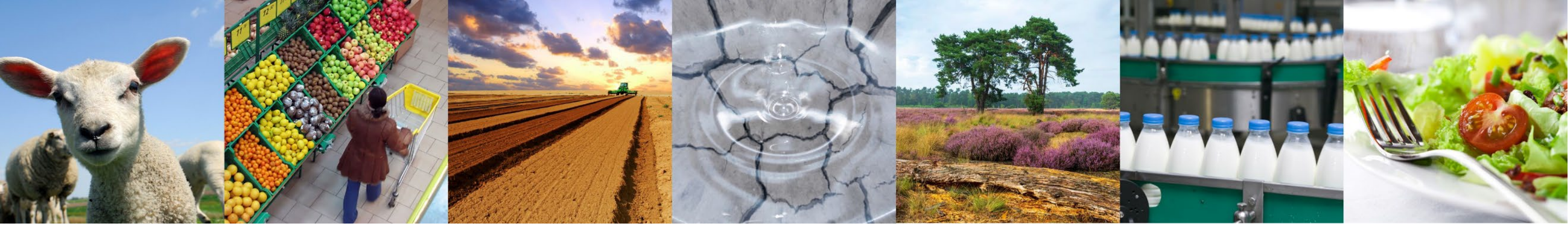
Committees

Scientific Committee (SC)

- Olivier Thas, Chair, Ghent University, Belgian region
- Pierre Lebrun, Co-chair, PharmaLex, Belgian region
- Jelle Goeman, Leiden University, Dutch region
- Carel F.W. Peeters, Wageningen University & Research, Dutch region
- Sophie Ancelet, IRSN, French region
- Boris Hejblum, INSERM, Bordeaux, French region
- Rafael de Andrade Moral, Maynooth University, British and Irish region
- Nicole Augustin, University of Edinburgh, British and Irish region

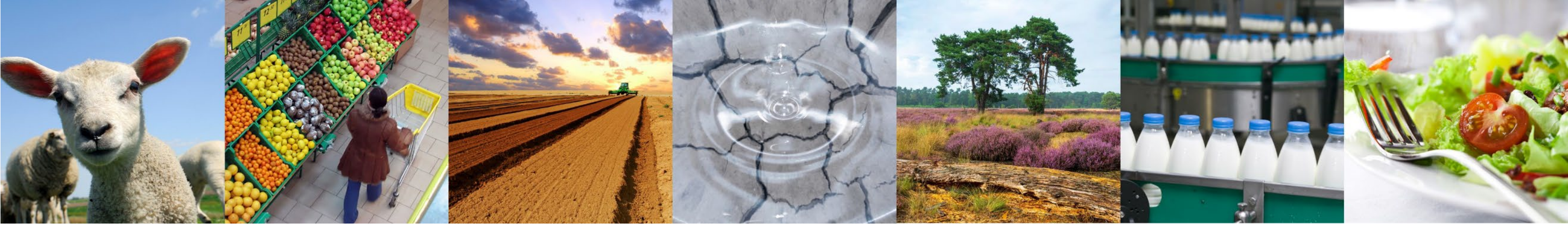
Local Organizing Committee (LOC)

- Carel F.W. Peeters, Chair, Wageningen University & Research, The Netherlands
- Jos Hageman, Chair, Wageningen University & Research, The Netherlands
- Dennis te Beest, Wageningen University & Research, The Netherlands
- Jonathan Kunst, Wageningen University & Research, The Netherlands
- Dinie Verbeek, Wageningen University & Research, The Netherlands

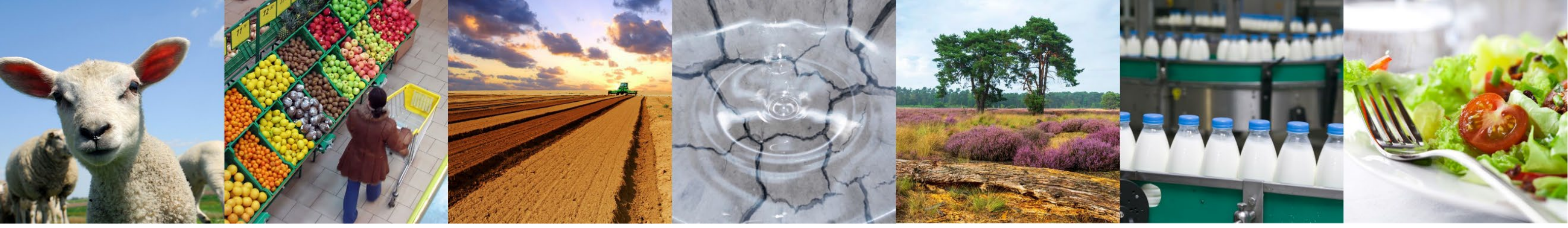


Programme

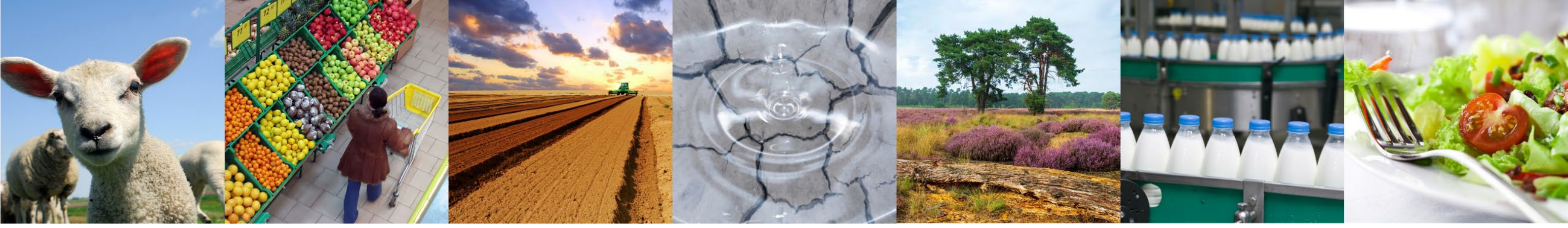
TIME	WEDNESDAY AUGUST 23		
09:00 – 12:30	**Short Course 1** **Integration of Multiple Omics Data Sets** Room: Quantum 1	**Short Course 2** **Methodology for Plant Breeding** Room: Quantum 2	**Short Course 3** **Detection of Structures in Ecological Networks** Room: Quantum 3
12:30 – 13:30	Lunch Break		
13:30 – 14:00	**Opening Ceremony** Room: Podium		
14:00 – 15:00	**Keynote Address: Aad van der Vaart** Chair: Carel F.W. Peeters Room: Podium		
15:00 – 15:30	Tea & Coffee Break		
15:30 – 17:00	**Contributed Session 1** **Statistics & Data Science in Agriculture** Chair: John Addy Room: Podium *Speakers:* Andrew Mead Ian Nevison K. Melsen & J. Kunst Martin Boer	**Contributed Session 2** **Advances in Survival Modeling** Chair: Marianne Jonker Room: Momentum 2/3 *Speakers:* Marije Sluiskes Ronald Geskus Mar Rodríguez-Girondo Mari Brathovde	**Contributed Session 3** **Signal and Image Processing** Chair: Xiaodong Chen Room: Momentum 1 *Speakers:* Carel Peeters Paul Eilers Ba Kalidou Tatsiana Khamiakova
17:00 – 17:30	**Poster Lightning Presentations** Chair: Carel F.W. Peeters Room: Podium *Lightning presentations:* Adrian Roberts Sandra Keizer Alfred Montero Fernández Gabriel Palma Fakhira Rifanti Maulana Ron Wehrens Kai Ruan Harimurti Buntaran Sabine Schnabel Iris van den Boomgaard Jos Hageman		
17:30 – 19:00	Poster Session & Drinks Reception		

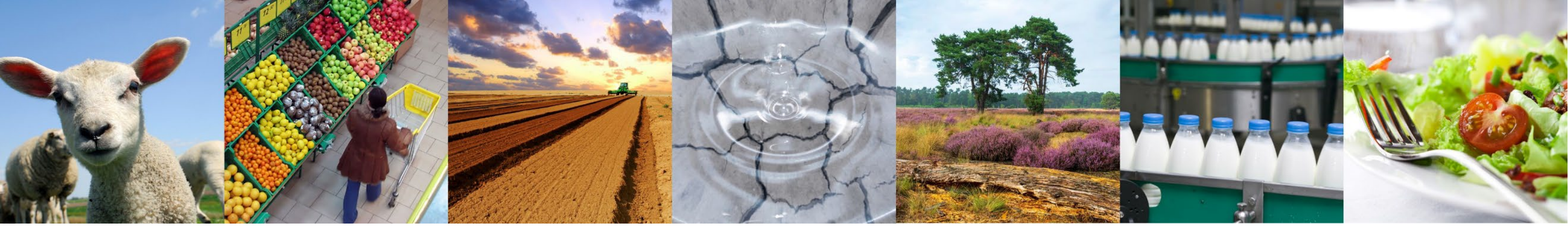


TIME	THURSDAY AUGUST 24		
09:00 – 10:30	**Invited Session I: Challenges in Genomic Prediction** Chair: Fred van Eeuwijk Room: Podium *Speakers:* Moritz Hermann Gregor Gorjanc Laura Zingaretti		
10:30 – 11:00	Tea & Coffee Break		
11:00 – 12:30	**Contributed Session 4** **Methods for High-Dimensional & Big Data** Chair: Jeanine Houwing-Duistermaat Room: Podium *Speakers:* Jürgen Claesen Kayané Robach Yuchen Guo He Li	**Contributed Session 5** **Advances in Survival Modeling II** Chair: Ronald Geskus Room: Momentum 2/3 *Speakers:* Daniel Gomon Chengyuan Lu Lars van der Burg	**Contributed Session 6** **Topics in Testing** Chair: Jelle Goeman Room: Momentum 1 *Speakers:* Erik van Zwet Stefan Böhringer Dominique-Laurent Couturier
12:30 – 13:30	Lunch Break		
13:30 – 15:00	**Contributed Session 7** **Joint and Multistate Models** Chair: Liesbeth de Wreede Room: Podium *Speakers:* Floor van Oudenhoven Roula Tsonaka Chiara Degan Michel Hof	**Contributed Session 8** **Multi-Environment Genomic Prediction** Chair: Gregor Gorjanc Room: Momentum 2/3 *Speakers:* Jip Ramakers Harimurti Buntaran Willem Kruijer	**Contributed Session 9** **Bayesian Methods** Chair: Mark van de Wiel Room: Momentum 1 *Speakers:* John Addy Richard Post Marianne Jonker Cécile Levrault
15:00 – 15:30	Tea & Coffee Break		
15:30 – 17:00	**Invited Session II: Advances in (Network) Meta-Analysis** Chair: Olivier Thas Room: Podium *Speakers:* Sylwia Bujkiewicz Leonhard Held		
17:00 – 18:00	**Contributed Session 10** **Clinical & Medical Statistics** Chair: Roula Tsonaka Room: Podium *Speakers:* Vera Arntzen Minh Hanh Nguyen Aiden O'Keeffe	**Contributed Session 11** **Statistical Genetics** Chair: Andrew Mead Room: Momentum 2/3 *Speakers:* Adrian Roberts James Adams Chaozhi Zheng	**Contributed Session 12** **Statistical Kaleidoscope** Chair: Jos Hageman Room: Momentum 1 *Speakers:* Stijn Hawinkel Máté Kormos Hideyasu Shimadzu
18:00 – 19:00	Pre-dinner drinks		
19:00 – 21:00	Conference Dinner		



TIME	FRIDAY AUGUST 25
09:00 – 10:30	**Invited Session III: Bayesian Machine Learning & Optimal Design** Chair: Pierre LeBrun Room: Podium *Speakers:* Estevão Prado Matthias Brückner James Willard
10:30 – 11:00	Tea & Coffee Break
11:00 – 12:00	**Keynote Address: Marc Chadeau-Hyam** Chair: Jelle Goeman Room: Podium
12:00 – 12:30	**Closing & Awards Ceremony** Room: Podium
12:30 – 13:30	Farewell Lunch





Keynote Address I: Aad van der Vaart

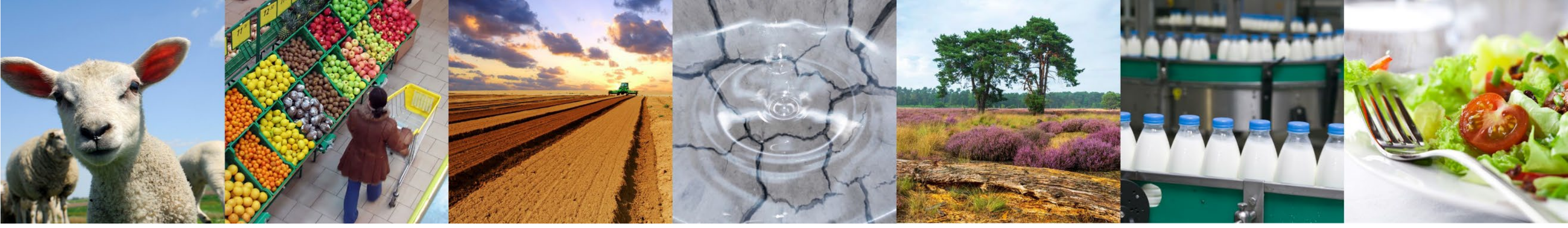
Chair: Carel F.W. Peeters
Room: Podium

Sensitivity analysis in causal analysis

Aad van der Vaart

Delft Institute of Applied Mathematics (DIAM), TU Delft

Abstract
Causal inference is based on the assumption of “conditional exchangeability”. This is not verifiable based on the data when using nonparametric modelling. A “sensitivity analysis” considers the effect of deviations from the assumption. In a Bayesian framework we could put a prior on the size and structure of the deviation and obtain an ordinary posterior. We review possible approaches and present some results on one approach, comparing different ways of nonparametric modelling. (Based on joint work with Bart Eggen, Stéphanie van der Pas and Chris van Vliet.)



Contributed Sessions 1-3

Contributed Session 1: Statistics & Data Science in Agriculture

Chair: John Addy
Room: Podium

Extending meta-analysis approaches to address routes to Ecological Intensification using agricultural Long-Term Experiments

Andrew Mead[1], Chloe MacLaren[2], Jon Storkey[3]

[1]Intelligent Data Ecosystems, Rothamsted Research, Harpenden, UK
[2]Swedish University of Agricultural Sciences, Uppsala, Sweden
[3]Protecting Crops and the Environment, Rothamsted Research, Harpenden, UK

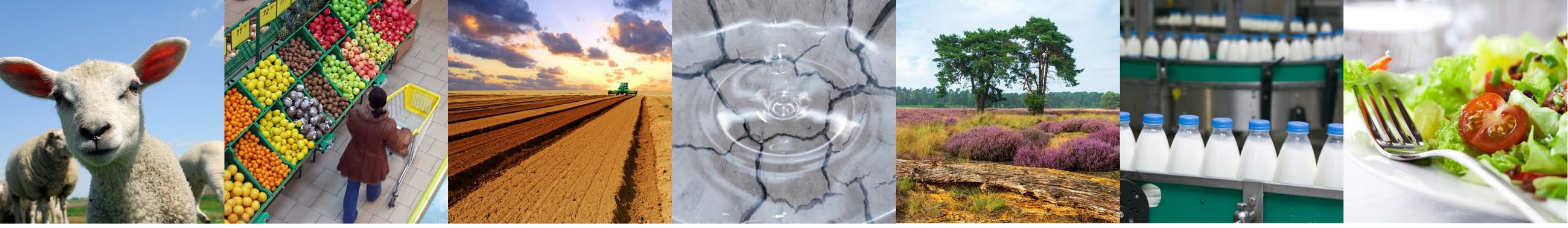
Abstract

Agricultural Long-Term Experiments (LTEs) provide a rich data source for exploring impacts of ecological intensification (EI) practices for transforming agricultural systems, against the backdrop of weather variation. We extend and apply a meta-analysis approach to data from LTEs collated within the Global Long-Term Experiment Network (GLTEN – www.glten.org), demonstrating the potential to address questions about EI. We first develop common indices to define key EI components across different trials. Unusually for a meta-analysis approach we then analyze the raw data from each trial to extract contrasts for different aspects of ecological intensification, using replication over time to assess the associated uncertainty. Separate analyses are applied to subsets of the generated data, focused on different EI practices, including mainly qualitative moderators to identify the impacts of different combinations of EI practices, with trial as a random model component. Model selection takes account of the factorial structure of the combinations of EI practice moderators, using AIC and QM/QE test statistics to select the best model. Results are presented for the main and interactions effects amongst the moderating context variables, identifying the relative importance of terms. The approach assesses the contributions from different trials to the consensus and identifies potential remaining sources of variation.

Key words
meta-analysis, agricultural systems, long-term experiments, model selection

References

[1] MacLaren, C., Mead, A., Van Balen, D., Claessens, L., Etana, A., De Haan, J., Haagsma, W., Jack, O., Keller, T., Labuschagne, J., Myrbeck, A., Necpalova, M., Nziguheba, G., Six, J., Strauss, J., Swanepoel, P. A., Thierfelder, C., Topp, C., Tshuma, F., Verstegen, H., Walker, R., Watson, C., Wesselink, M. and Storkey, J. 2022. Long-term evidence for ecological intensification as a pathway to sustainable agriculture. Nature Sustainability, 5, p770-779. https://doi.org/10.1038/s41893-022-00911-x



Quantifying and mitigating shading effects in winter barley trials

Ian Nevison[1], Tess Vernon[1], Adrian Roberts[1]

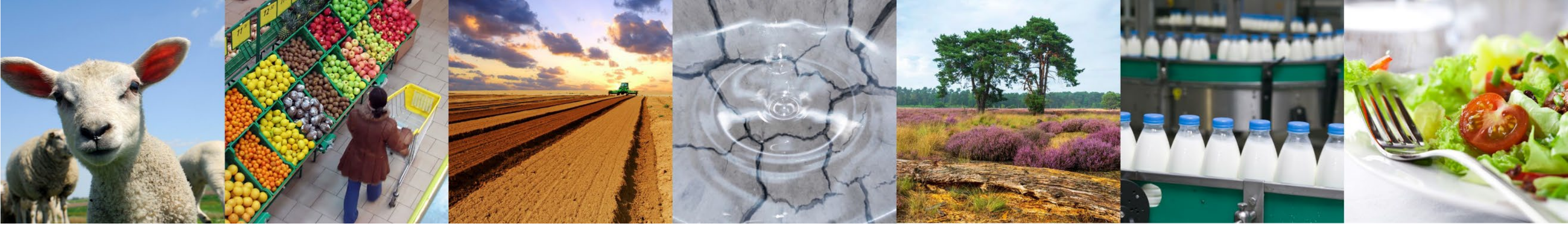
[1]Biomathematics and Statistics Scotland, Edinburgh, United Kingdom

Abstract

In the UK, candidate winter barley varieties must perform well in two years of regulatory testing before they can be marketed. For inclusion in the list of recommended varieties they need to successfully undergo a further year of trialing. Although disease and agronomic responses are taken into consideration, the primary criterion for decision-making is yield. Varieties are sown in series of replicated alpha-lattice design trials comprising small plots (approx. 10m x 2m) of each variety unlike standard agricultural practice. There are appreciable differences in height between varieties, leading to concerns that taller varieties might shade neighbouring shorter varieties and hence depress their yields. Such effects would result in biased comparisons between varieties and therefore unfairly disadvantage some varieties in the decision-making process.

We consider various alternative models for quantifying the magnitude of varietal shading effects, the extent to which estimates vary between models, and the potential impact of ignoring shading effects completely. We also discuss potential ways in which the impact of such shading effects could be mitigated.

Key words
Crop; Varieties; Shading; Mitigation; Comparisons



gfBLUP: Integrating high-dimensional phenotypic data in genomic prediction models using latent factor analysis

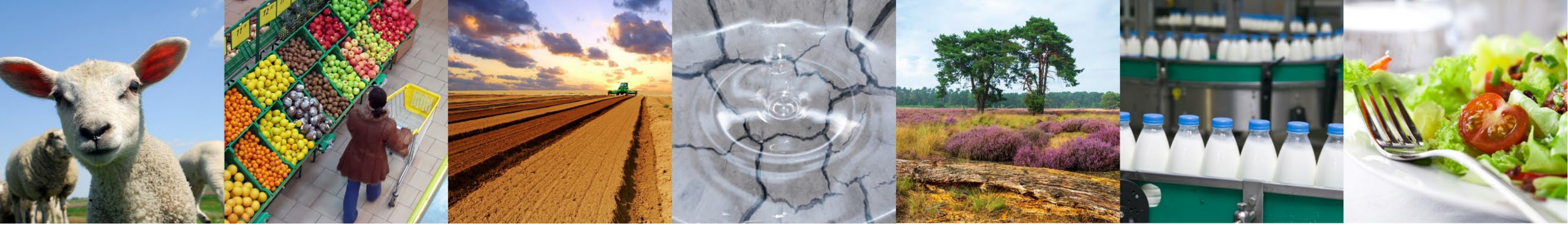
Killian Melsen[1], Jonathan Kunst[1], Fred van Eeuwijk[1], Willem Kruijer[1], Carel F.W. Peeters[1]

[1]Mathematical & Statistical Methods group (Biometris), Wageningen University & Research, Wageningen,
The Netherlands

Abstract
The study of plant breeding populations is an increasingly technological endeavour. In addition to genetic information many traits are routinely evaluated using high-throughput technologies.

This has led to the availability of high-dimensional data of many observed traits within breeding populations. Consequently, we are faced with the challenge of interpretably integrating these data to support data-driven breeding decisions. Here we present gfBLUP, an efficient method for genomic prediction utilising high-dimensional secondary trait information and genomic data to better predict complex traits such as yield. In short, high-dimensional traits are reduced to interpretable latent factors, which are subsequently integrated into multivariate genomic prediction models for a complex trait. We show the usefulness of this method in real breeding studies.

Key words
Factor analysis; Genomic prediction; Plant breeding; High-dimensional data



Fast multi-dimensional mixed model P-splines analysis of High Throughput Phenotype data in breeding trials

Martin Boer[1], Bart-Jan van Rossum[1], Daniela Bustos Korts[2], Jip Ramakers[1], Jesse Hemerik[1], Scott Chapman[3], Fred van Eeuwijk[1]

[1]Biometris, Wageningen University and Research Centre, Wageningen, the Netherlands
[2]Institute of Plant Production and Protection, Faculty of Agricultural Sciences, Campus Isla Teja,
Universidad Austral de Chile, Valdivia, Chile
[3]School of Agriculture and Food Sciences, The University of Queensland, St. Lucia, Queensland 4072, Australia.

Abstract

In modern breeding trials, many observations are made during the growing season; for example, spectral image data obtained from unmanned aerial vehicles (UAV). We analyse the spatial patterns using multi-dimensional penalized splines (P-splines; Eilers and Marx 1996), to separate the genetic effects from the spatial trend during the growing season. To analyse such type of data computational efficient methods are needed.

A new approach to represent multi-dimensional P-splines as a mixed model is presented (Boer, 2023). The precision matrices are sparse allowing the new approach can find the optimal values of the penalty parameters in a computationally efficient manner. An important feature ensuring that the entire computation is fast is a sparse implementation of the Automated Differentiation of the Cholesky algorithm (Smith 1995). The methodology has been implemented in the R-package LMMsolver available on CRAN. The new fast method will be illustrated with two examples in wheat from the INVITA (Innovations in Variety Testing in Australia) project. The first example will illustrate the use of mixed model P-splines for UAV data in plant breeding trials. The second example will use two-dimensional P-splines to predict yield as function of latitude and longitude.

Key words

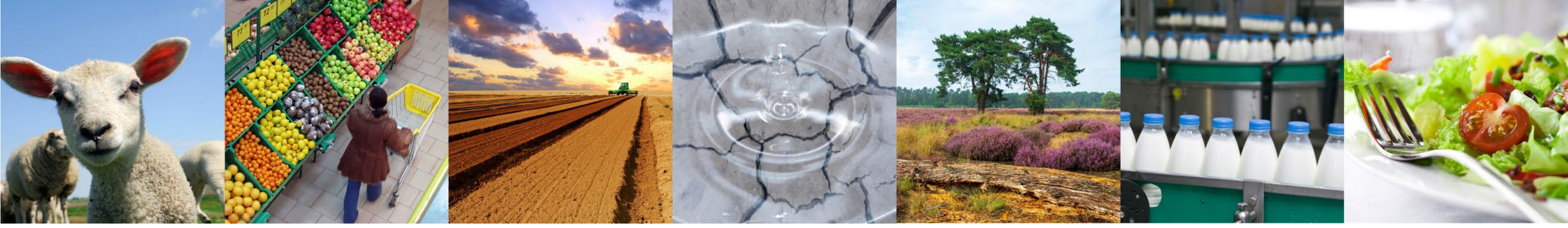
Sparse Linear Algebra; Cholesky; Automated Differentiation

References

[1] Boer MP (2023) Tensor product P-splines using a sparse mixed model formulation. Statistical Modelling (accepted).

[2] Eilers PHC and Marx BD (1996) Flexible smoothing with B-splines and penalties Statistical science p. 89-121

[3] Smith SP (1995) Differentiation of the Cholesky Algorithm.J. Comput. Graph. Stat. p. 134-147



Contributed Session 2: Advances in Survival Modeling

Chair: Marianne Jonker
Room: Momentum 2/3

A reduced rank proportional hazards model for age-related multimorbidity event data

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

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

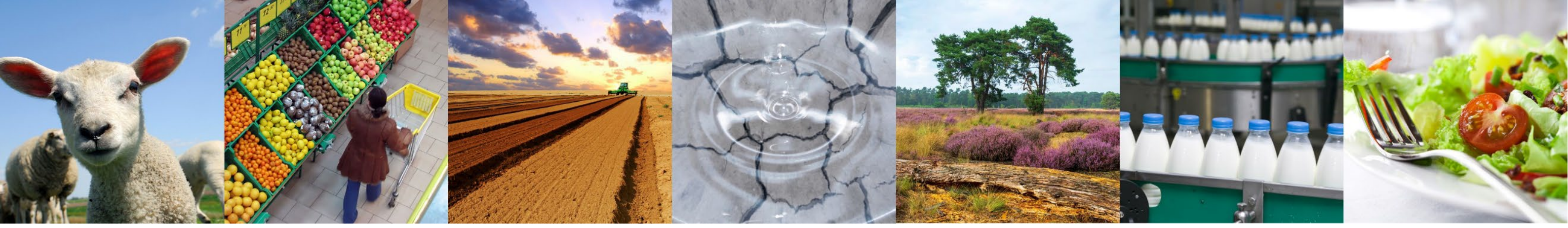
Abstract
The identification of biomarkers of aging is an important biomedical research theme. Most current statistical methods that aim to capture the aging process either use chronological age or time-to-mortality as the outcome of interest. There is however a shift in the field towards the study of health span and patterns of age-related multi-morbidity, as aging entails more than lifespan duration alone.

Several large epidemiological studies, such as the UK Biobank and the Leiden Longevity Study, have recently incorporated detailed age-at-disease-onset profiles, obtained from electronic health records. The availability of these data opens new analytical possibilities. Nevertheless, analyses conducted thus far oversimplify the complexity of multi-morbidity patterns, for instance by ignoring information on age-at-disease-onset or by failing to acknowledge that age-related diseases are likely driven by a shared set of underlying factors.

We propose a new methodological framework for the analysis of age-related multi-morbidity data, based on multiple-outcome survival modelling. Specifically, we propose to use a reduced rank proportional hazards model. This model can be fitted on the (possibly right-censored and left-truncated) age-at-disease-onset of several age-related diseases simultaneously. It assumes that there is a set of shared latent factors that drive all age-related diseases considered, thereby reducing the dimensionality of the problem and providing additional insight into different facets of the aging process. As there is a large interest in the use of high-dimensional omics data as potential biomarkers of aging, we also discuss some ideas to include penalization in the reduced rank proportional hazards framework.

The use and intuitive interpretation of the reduced rank proportional hazards model is illustrated by applying it to age-related multimorbidity and mortality data from the UK Biobank, using metabolomics data as predictor variables. Comparison of the reduced rank model to simpler alternative models is shown.

Key words: multiple-outcome survival modelling, electronic health records, omics



Requiring two positive tests as event definition: consequences for censoring strategy

Ronald Geskus[1,2]

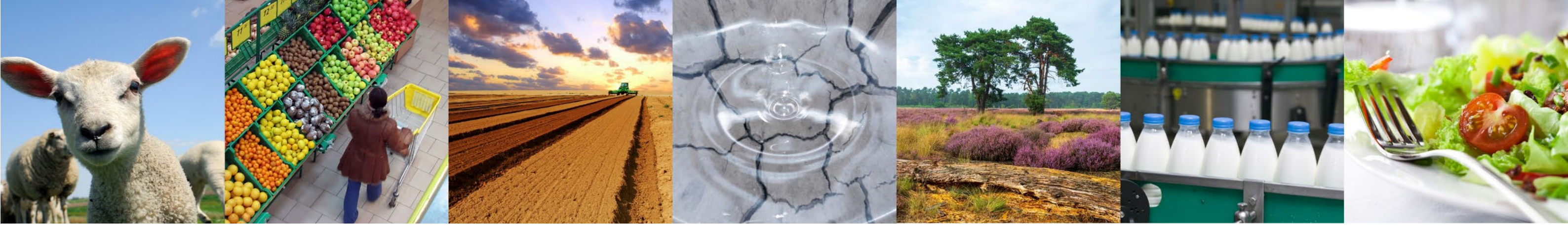
[1]Oxford University Clinical Research Unit [Ho Chi Minh City]
[2]Nuffield Department of Clinical Medicine [Oxford]

Abstract

In the typical time-to-event setting, a single positive observation defines occurrence of the event. Sometimes more than one positive observation is needed. For subclinical events with suboptimal sensitivity and/or specificity, the event may be defined by having two consecutive positive tests. To distinguish between temporary and chronic disease in studies with intermittent observations, the condition needs to be present at two or more visits. Fever clearance is often defined as being without fever for 48 hours.

For the proper definition of risk set, the basic principle is: only include individuals while the event can be observed. In the medical literature, the last negative test is commonly used as censoring time, violating this principle in the above settings. We investigate the amount of bias in two simulation studies. One mimics a cohort study with short follow-up on the incidence of infection with human papilloma virus. The other mimics a cohort study on chronic disability that has the additional complication that disability can go unobserved due to mortality. In both, the penultimate test has the last disease-free information. Bias between 10% and 50% was observed using the incorrect censoring strategies. Using our suggested strategy is essential to reduce bias.

Key words
censoring strategy, periodic observations, illness death model



Inverse probability weighted Cox regression to correct for ascertainment bias

Mar Rodríguez-Girondo

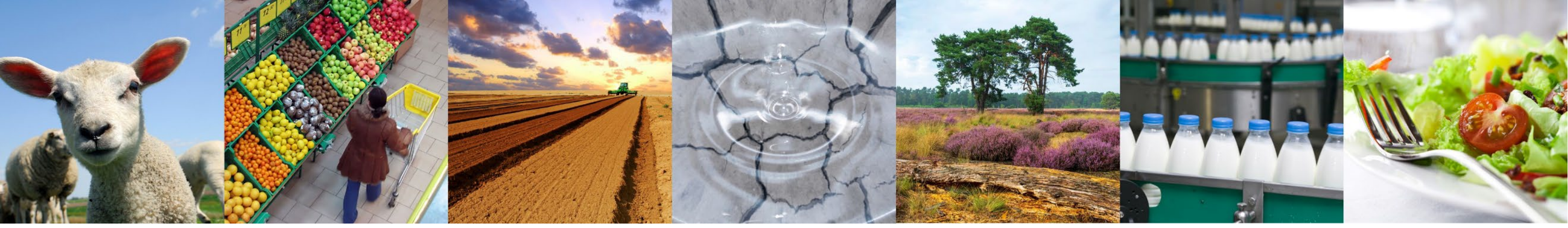
Leiden University Medical Center, Leiden, The Netherlands

Abstract

Motivated by the study of genetic effect modifiers of cancer, we examine weighting approaches to correct for ascertainment bias of covariate effects in the context of Cox proportional hazards regression. Family-based outcome-dependent sampling is common in genetic epidemiology leading to samples with an overrepresentation of young, affected subjects. A usual approach for correcting for ascertainment bias in this setting is to use an inverse probability-weighted Cox model, using weights based on external available population-based age-specific incidence rates of the type of cancer under investigation. However, the current approach relies on the assumption of oversampling of cases of all ages and the absence of unobserved heterogeneity which is not realistic in relevant practical settings. We propose a new, more general approach based on the same principle of weighting observations by their inverse probability of selection. We compare the methods in simulations and illustrate the advantage of our new method with several real datasets. In all the applications, the goal is to assess the association between common susceptibility loci identified in Genome-Wide Association Studies (GWAS) and cancer (colorectal, breast, and melanoma) using data collected through genetic testing in clinical genetics centers of the Netherlands.

Key words

survival analysis, outcome-dependent sampling, weighting, Cox regression, genetic epidemiology



A lean additive frailty model: with an application to clustering of melanoma in Norwegian families

Mari Brathovde[1,2], Tron A. Moger[3], Marit B. Veierød[2], Tom Grotmol[4], Odd O. Aalen[2], Morten Valberg[1]

[1]Oslo Centre for Biostatistics and Epidemiology, Oslo University Hospital, Oslo, Norway
[2]Oslo Centre for Biostatistics and Epidemiology, Dept. of Biostatistics, Institute of Basic Medical Sciences, University of Oslo, Oslo, Norway
[3]Department of Health Management and Health Economics, University of Oslo, Oslo, Norway
[4]Cancer Registry of Norway, Oslo, Norway

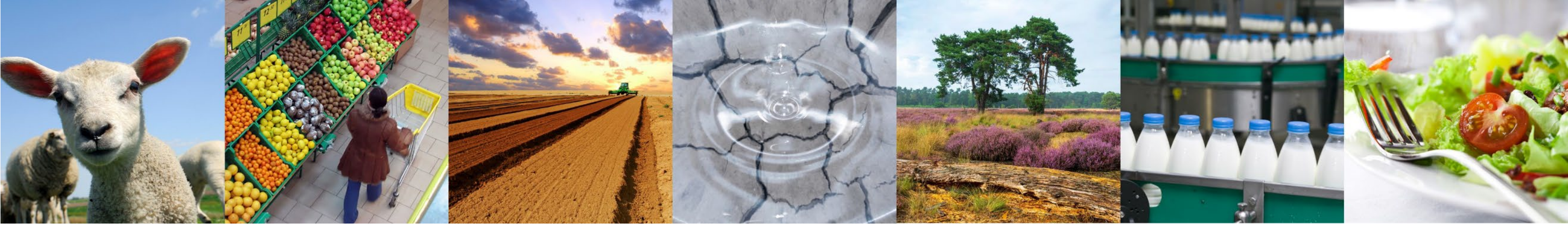
Abstract

Large-scale health registries enable detailed studies of clustering of cancers in families. Frailty models provide a framework for conducting such studies at a much higher level of detail than conventional studies. However, these models' complexity grows with family size and the number of cancer diagnoses in each family. Consequently, these models have primarily been used in settings where cluster sizes are small, e.g. for twin pairs, or by considering only a few first-born children in a family. This poses a challenge for fully utilizing the detailed data available in the registries. We present a modification of the additive genetic gamma frailty model, which alleviates some of these problems by using a leaner additive decomposition of the frailty. The modified model is then used to analyze population-wide data on clustering of melanoma in 2,391,125 Norwegian families.

Using a first-order approximation of the genetic structure in nuclear families of parents and children, we obtain a model that reduces the complexity wrt. family size. Although a large number of cases within the same family still poses a challenge, this allows us to analyze a far greater class of datasets. An additional major benefit of the lean model is a significant speed-up in model fitting. This enables fitting even more complex models and makes model fitting on a desktop computer feasible without needing a high-performance cluster. We demonstrate in a simulation study that our proposed lean model gives a good approximation to the original additive genetic gamma frailty model.

Using the lean model, we can analyze the complete population-wide data set on melanomas in all Norwegian families registered from 1960-2016. We find a substantial clustering of melanomas in Norwegian families and a large heterogeneity in melanoma risk across the population. We estimated that there is a large inequality in frailty in the population, where 46% of the frailty could be attributed to the 10% of the population at the highest unobserved risk.

In conclusion, additive frailty models can be used to study relatively large clusters. Furthermore, there is a substantial clustering of melanomas in Norwegian families and a large heterogeneity in melanoma risk across the population.



Contributed Session 3: Signal and Image Processing

Chair: Xiaodong Chen
Room: Momentum 1

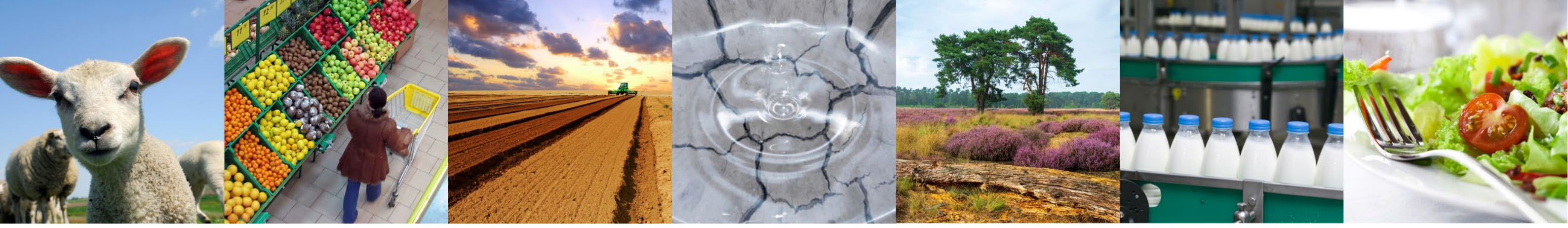
Representation Learning: Shallowed be Thy Name

Carel F.W. Peeters

Mathematical & Statistical Methods group (Biometris), Wageningen University & Research, Wageningen, The Netherlands

Abstract
We will focus on representation learning (RL), i.e., learning representations of data that make subsequent learning tasks easier. From a probabilistic perspective RL can be viewed as the recovery of a low-dimensional set of latent random variables that captures the information contained in the observed data. We will treat the case in which a one-layer generative neural network concurs with the classic factor analytic model. We will see that, for high-dimensional data, such shallow representations can be sufficient for data representation and downstream modeling support. We will exemplify this with both medical and agricultural imaging data.

Key words
Data representation; High-dimensional data; Latent variable modeling; Representation learning



Super-fast Image Deconvolution for Super-resolution

Paul H.C. Eilers[1], Cyril Ruckebusch[2]

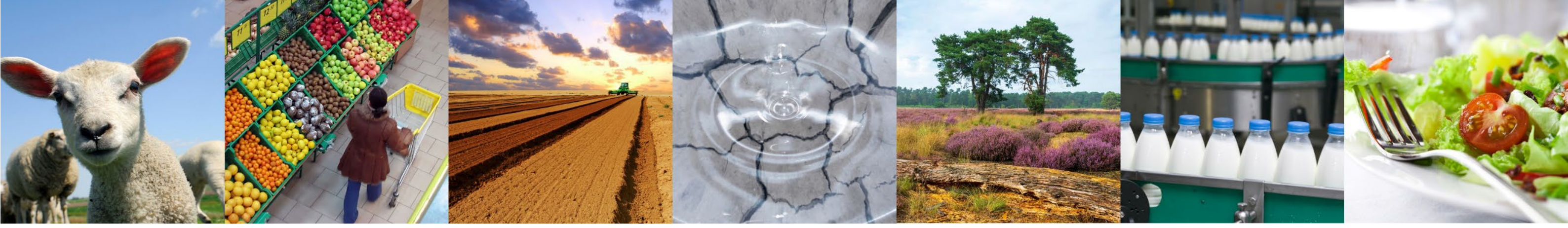
[1]Erasmus University Medical Center, Rottterdam, The Netherlands
[2]University of Lille, Lille, France

Abstract
The resolution of an optical microscope is limited by the wavelength of light. The observed image can be seen as the convolution of the true image and the spread function of the optical instrument. One way to improve the resolution of a digitized image is deconvolution by penalized regression. The principle is simple, but practical implementation is challenging. To increase the resolution of an image of 250 by 250 pixels with a factor 4, a linear system with one million unknowns has to be solved. A straightforward implementation would be very demanding in computation time and computer memory.

The conjugate gradients method for solving (large) linear systems of equations uses only products of matrices and vectors (Hestenes and Stiefel, 1952). The optical spread function can be written as the outer product of two vectors. The two-dimensional convolution can thus be written as a onedimensional convolution of the rows of the source image, followed by a convolution of the columns of the result. Combining these ideas leads to a very fast and compact algorithm. The algorithm needs less than ten iterations for a sharp result. Already after three or four iterations a useful image is obtained. On a PC with an I5-8400 processor and Matlab 2012, one iteration of our algorithm takes one second when increasing the resolution of a 250 by 250 pixels image by a factor 8. We will discuss and illustrate the theory and show applications to biological images. The software is
freely available.

Key words
Convolution; Conjugate gradients; Tensor products

References
[1] Hestenes MR and Stiefel E (1952). Methods of Conjugate Gradients for Solving Linear Systems. Journal of Research of the National Bureau of Standards, 49: 409–436.



When less is not more: the negative impact of incomplete signature reference matrices on cellular frequency deconvolution performance

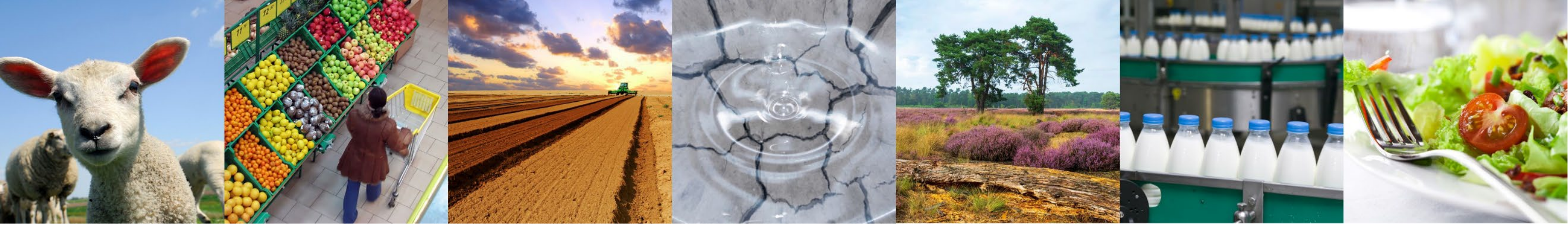
Kalidou BA[1,2], Rodolphe THIÉBAUT[1,2,3], Xavier HINAUT[4,5,6], Boris HEJBLUM[1,2]

[1]Univ. Bordeaux, INSERM, INRIA, SISTM team, BPH, U1219, F-33000 Bordeaux, France
[2]Vaccine Research Institute, F-94000 Créteil, France
[3]CHU Pellegrin, Groupe Hospitalier Pellegrin, Bordeaux F-33076, France
[4]INRIA Bordeaux Sud-Ouest, France
[5]LaBRI, Bordeaux INP, CNRS, UMR 5800, France
[6]Institut des Maladies Neurodégénératives, Université de Bordeaux, CNRS, UMR 5293, France

Abstract

Cellular deconvolution refers to the estimation of cellular population frequencies from gene expression measurements in a biological sample. While numerous supervised approaches have been proposed (such as CibersortX or scaden), their good performance critically depends on the reference signature matrix. This matrix encodes the gene expression profiles of the different cell types from external prior knowledge.However, addressing the common scenario of missing cellular populations from the reference matrix has received limited attention compared to the profusion of proposed deconvolution algorithms. We assess the lack of robustness of the state-of-the-art deconvolution methods in both simulations and benchmarking real data. Our simulations designs, based on either a Poisson or a Gaussian multivariate distribution, are validated against real data from the literature. Results from simulations and multiple real datasets, demonstrate that deconvolution performance remains relatively unaffected as long as the reference matrix includes most cellular populations present in the sample. Conversely, performance rapidly deteriorates for all deconvolution methods as the reference matrix becomes increasingly incomplete. Moreover, the impact of missing cell populations in the reference matrix depends on their actual frequency in the sample.

Key words
Cellular deconvolution; Bulk RNA-seq; Reference matrices; Cell abundance; Factor model.



Deconvolution of UPLC data to streamline multivariate quality control of oligonucleotide synthesis

Tatsiana Khamiakova[1] , Tiny Deschrijver[2] , Nicolas Sauwen[3]

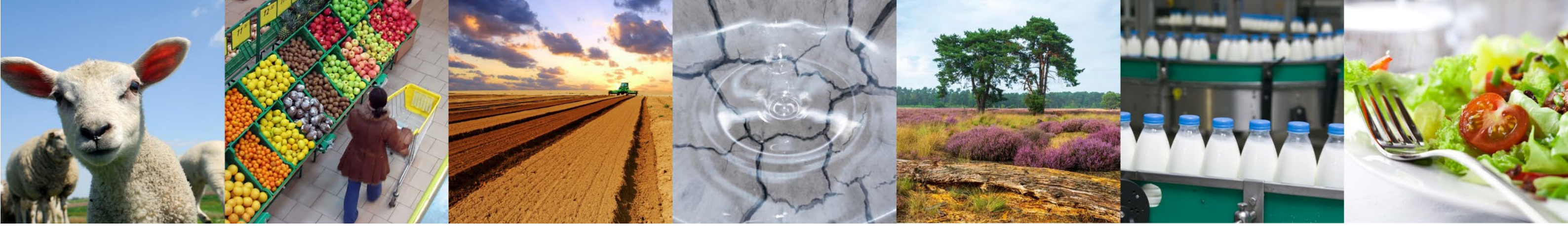
[1]Statistics and Decision Sciences, Janssen Pharmaceutica NV., Beerse, Belgium
[2]Chemical process research and development, Janssen Pharmaceutica NV., Beerse, Belgium
[3]Open Analytics NV, Antwerp, Belgium

Abstract

To assure the delivery of high quality product of synthesis for oligonucleotide-based drug substance, the result of synthetic reaction is checked by ultra-high pressure chromatography (UPLC). The UPLC data coming in a shape of chromatogram are used to monitor whether the signals from new samples are similar to the expected signal based on the historical data. However, the task of constructing a multivariate control chart with interpretable parameters is complicated due to the presence of multiple modifications of the main compound in a sample. A typical chromatogram contains several peaks which have (quasi) Gaussian shape with a lot of overlap which makes quantification of concentrations of each species present in a sample cumbersome.

In this work we use exponentially modified gaussian mixture modeling to extract the most relevant information related to the concentration of different oligonucleotide species to track individual components in both univariate and multivariate ways which provides a practitioner with a toolkit to identify outliers in the process. The advantage of this approach compared to the unsupervised multivariate methods is in its interpretability and ability to detect the components of a mixture with a small number of samples. In addition, we shall demonstrate the practical implementation of this workflow, including aspects of chromatographic alignment and normalization.

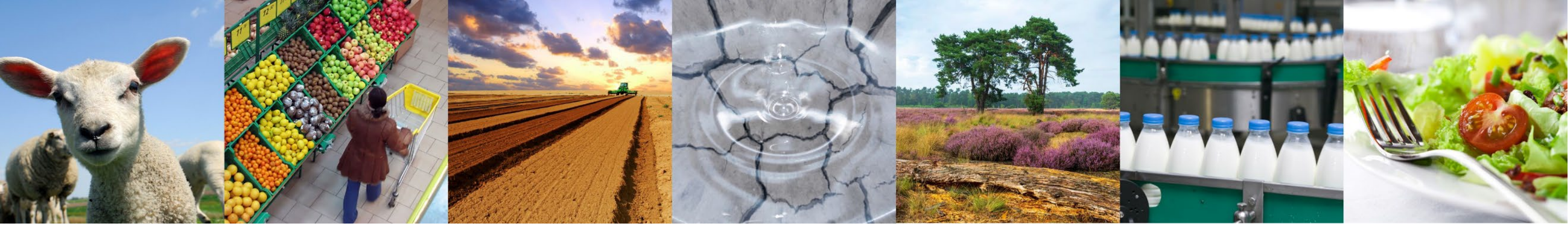
Key words
Signal processing, gaussian mixture modeling, multivariate analysis, statistical process control



Poster Lightning Presentations

Chair: Carel F.W. Peeters
Room: Podium
Poster area: Foyer

Poster	Presenter	Title
1	Adrian M.I. Roberts	Assessing genetic uniformity of crop varieties based on pooled samples
2	Sandra P. Keizer	Joint modeling with time-dependent treatment and heteroscedasticity: Bayesian analysis with application to the Framingham Heart Study
3	Alfred Montero Fernández	Enhancing Bioassay Validation: SmartSTATS/Enoval, An Efficient Statistical Tool for Compliance with ICH Q2 Guideline
4	Gabriel R. Palma	Forecasting the abundance of agricultural pests: a new machine learning framework
5	Fakhira Rifanti Maulana	System Identification of a Greenhouse Crop Model using the Physics-Informed SINDy-PI Machine Learning Algorithm
6	Ron Wehrens	Analysing postprandial amino-acid responses in crossover studies with the aaresponse package for R
7	Kai Ruan	Targeted Structural Difference Testing on Sparse Gaussian Graphical Model
8	Harimurti Buntaran	Projecting results of zoned multi-environment trials to new locations using environmental covariates with random coefficient models: accuracy and precision
9	Sabine K. Schnabel	Challenges in statistical consulting for Animal Science
10	Iris van den Boomgaard	A system-level representation of kwashiorkor pathophysiology with multi-omics models
11	Jos Hageman	Validating QSAR models: an anti-MRSA case study



Invited Session I: Challenges in Genomic Prediction

Chair: Fred van Eeuwijk
Room: Podium

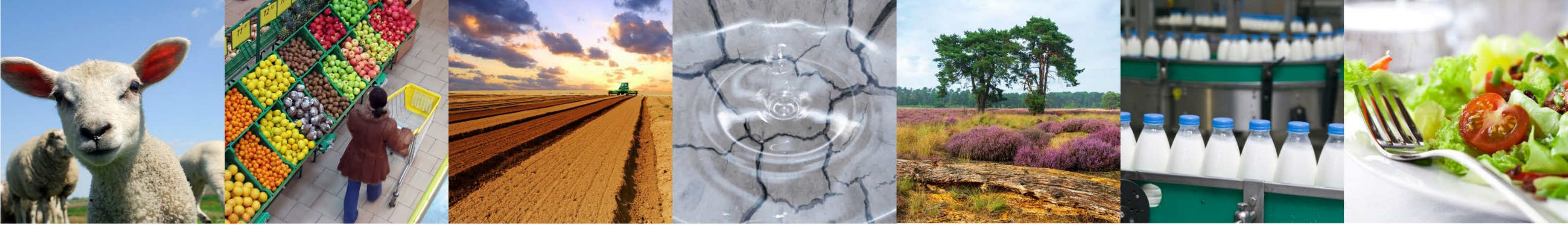
On the challenges of multi-omics data and method comparisons: Insights from a systematic benchmark study on survival prediction

Moritz Herrmann

Institute for Medical Information Processing, Biometry and Epidemiology, Ludwig-Maximilians-University

Abstract

Predicting disease outcomes, such as patients' survival times, using genomic data poses several challenges. This talk will focus on the challenges arising from the presence of multiple groups of molecular data types, so-called multi-omics data, on the one hand, and the challenges posed by the complexity of benchmark studies, on the other hand. With regard to the first aspect, the presence of multiple feature groups in multi-omics datasets requires effective handling strategies. For example, while high-dimensional sets of molecular features may provide rich information, there is a risk that important but low-dimensional clinical features may be lost in the prediction process. The effective incorporation of different types of high-dimensional molecular features alongside a small number of often highly informative clinical variables is critical to predictive performance. In contrast, the second aspect addresses the challenges of comparing prediction methods and selecting an appropriate method for the problem at hand. Identifying the most appropriate methods for a given dataset and outcome from the plethora of prediction methods available requires careful consideration of their strengths, limitations and ability to effectively use multi-omics data. Neutral benchmark studies, designed to systematically compare and evaluate different prediction methods on a large number of datasets, play a crucial role in this selection process. However, the variety of design and analysis options available can lead to biased interpretations and overly optimistic conclusions in favor of a particular method. Addressing these issues involves navigating the complexities of benchmark study design, performance measures, handling missing values in benchmarking results and aggregating results across datasets. By recognizing and addressing these challenges, we can improve the predictive power, harness multi-omics data while maintaining clinical relevance, and ensure rigorous and reliable evaluation and comparison of prediction methods. Ultimately, these efforts will help advance personalized medicine and improve patient outcomes.

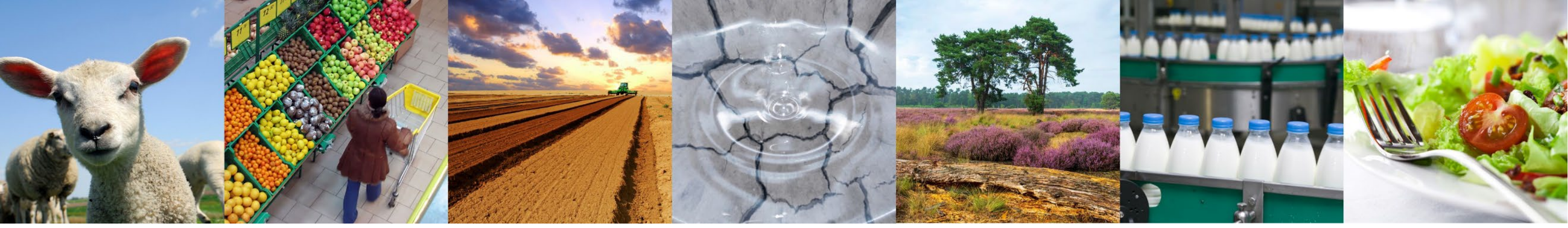


Leveraging ancestral recombination graphs to store genome-wide data and enable origin-aware genomic predictions

Gregor Gorjanc

The Roslin Institute and Royal (Dick) School of Veterinary Medicine, The University of Edinburgh, Edinburgh, UK

Abstract
We now have abundant genome-wide data that fuels quantitative genetic applications and biological discovery. The amount of this data is growing rapidly, challenging data storage and computation. However, the established statistical models don't seem to be able to leverage this data growth fully to increase the accuracy of genomic predictions. In this talk I will describe our work on leveraging ancestral recombination graphs, encoded with the tree sequence data format. We are working with ancestral recombination graphs for two reasons. First, to store mega-scale whole-genome sequence data. I will present two ongoing projects in our lab - encoding data from the globally diverse set of 1000 bull genomes and from the more narrowly diverse set 1 million pig genomes. Second, to build a richer statistical model that can leverage the full description of recombinations and mutations to enable origin-aware genomic predictions.



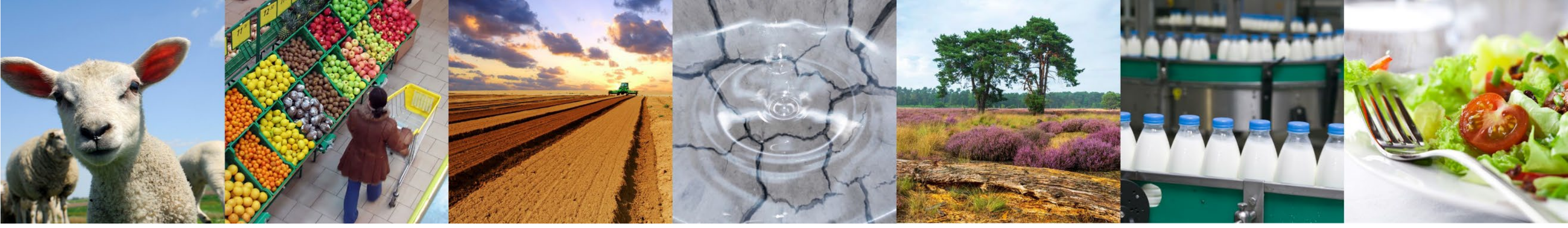
New strategies for old problems: should we integrate Deep Learning in plant breeding programs?

Laura Zingaretti

Nunhems (BASF), Nunhem, The Netherlands

The primary goal of any breeding program is to improve complex trait in a cost-effective way. Since Lush times, breeders have strongly relied on the breeding equation (BE) to guide plant breeding decisions. Basically, BE expresses genetic gain as a function of the selection intensity, the genetic variation, the selection accuracy, and the generation interval. Genomic Selection (GS) consist of using a dense set of markers to select superior genotypes without identifying the QTL individually. GS adds value by increasing the accuracy of breeding value predictions in comparison with pedigree-based models, helps to increase the intensity of selection and reduces the generation interval. Despite it is a well-established technology, its application in many plant breeding programs is still immature. GS is typically implemented by linear mixed models or Bayesian equivalent approaches. Although powerful, there are impractical to model non-additive genetic effects (such as dominance or epistasis) and assume a homogeneous variable input.

In recent years, the use of Deep Learning (DL) in GS has become a popular alternative. Although promising, the application of DL in GS in both plant and animals has not given clear signs of outperforming the standard methods. One of the reasons that have been argued is that the variable to be predicted (the genetic value) is not actually observed. Other reasons may be that input variables (i.e., SNPs) have a highly leptokurtic distribution. Some scenarios, though, may be more beneficial: in plants, the use of DL has been shown to be powerful in capturing complex (non-additive) patterns. Importantly, applications of DL in breeding can span a broader spectrum than just GS: generative neural networks (GANs) or variational autoencoders hold promise for conducting more realistic (data-driven) simulations, as they can learn complex distributions and can be used to produce synthetic DNA or images. DL are also powerful techniques to combine different data sources, e.g., Long- Short Term Memory (LSTM) can be used to model a temporal dimension with minimal preprocessing, allowing climate and environmental information to be easily added into GS or phenotypic models. Many of these applications are still in their infancy, so we expect rapid development of these technologies and breeders should take advantage of it.



Contributed Sessions 4-6

Contributed Session 4: Methods for High-Dimensional & Big Data

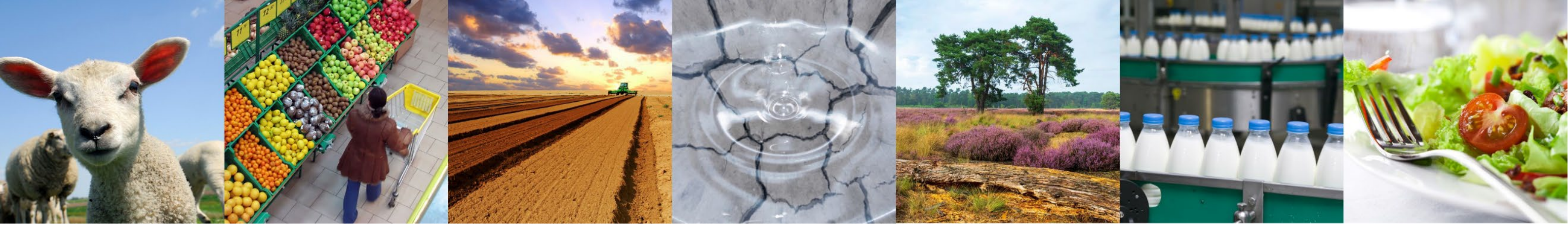
Chair: Jeanine Houwing-Duistermaat
Room: Podium

A nonlinear mixed-effects model for protein quantitation based on mass spectrometry data

Mateusz Staniak[1], Tomasz Burzykowski[2], Jürgen Claesen[3]

[1]University of Wroclaw
[2]Hasselt University
[3]Amsterdam University Medical Centre

Abstract
In bottom-up mass-spectrometry-based proteomics, proteins are digested into peptides prior to measuring them with mass spectrometry. Therefore, before determining the protein abundance, peptides have to be identified and assigned to the corresponding protein. The latter is a difficult task as multiple peptides cannot be assigned unambiguously to one protein. Commonly, these shared peptides are ignored during protein quantitation. Here, we propose a non-linear mixed effects model that can be used for protein quantitation that accounts for the presence of shared peptides. We illustrate that accounting for shared peptides leads to an improved precision of the estimated protein abundances with several real-life datasets.



A flexible Record Linkage model

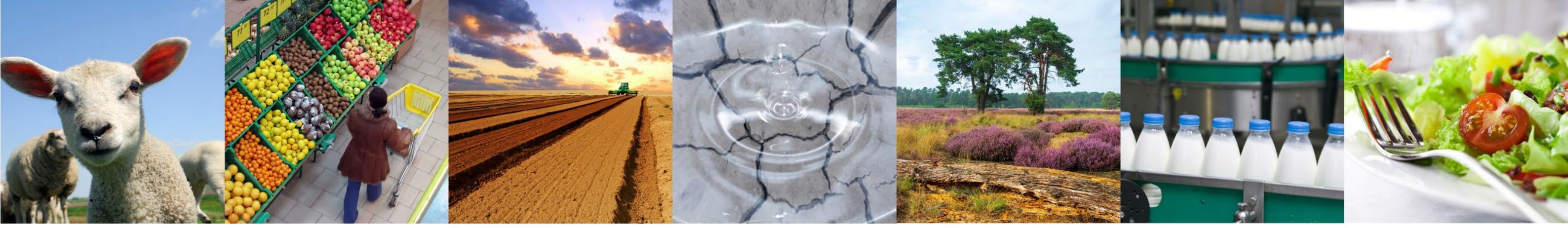
Kayané Robach[1], Stéphanie van der Pas[1], Mark van de Wiel[1], Michel Hof[2]

[1]Amsterdam Universitair Medische Centra, Vrije Universiteit, Epidemiology & Data Science, Amsterdam,
The Netherlands
[2]Amsterdam Universitair Medische Centra, Universiteit van Amsterdam, Epidemiology & Data Science,
Amsterdam, The Netherlands

Abstract

Combining different existing data sources empowers researchers to explore innovative questions, including those raised by tallying casualties and conducting healthcare monitoring studies. However, the lack of availability of a unique identifier often poses challenges. Record linkage procedures identify whether pairs of observations collected on different occasions belong to the same individual (referred to as matches) using partially identifying variables (e.g. initials, birthday, zipcode). Existing solutions attempt to simplify this task through condensing information but neglect dependencies among linkage decisions and disregard the one-to-one relationship required for building matches. The resulting reduction in computational burden comes at the price of inaccuracies and limited applicability. To avoid those issues, we propose to model the data generating process. We determine the set of matches (known as the linkage) incorporating complex correlation structures based on a latent variable model. We develop an MC-EM algorithm and estimate the linkage using maximum likelihood. Simulations demonstrate the robustness of our model to the linking variables quality and its ability to better connect observations. We illustrate its scalability using the Perinatal Registry of the Netherlands (approximately 500,000 observations per data source) to identify first and second born children belonging to the same mother.

Key words
Record linkage; Partially identifying variables; Latent variables; MC-EM



Generation of realistic virtual populations with a Copula approach

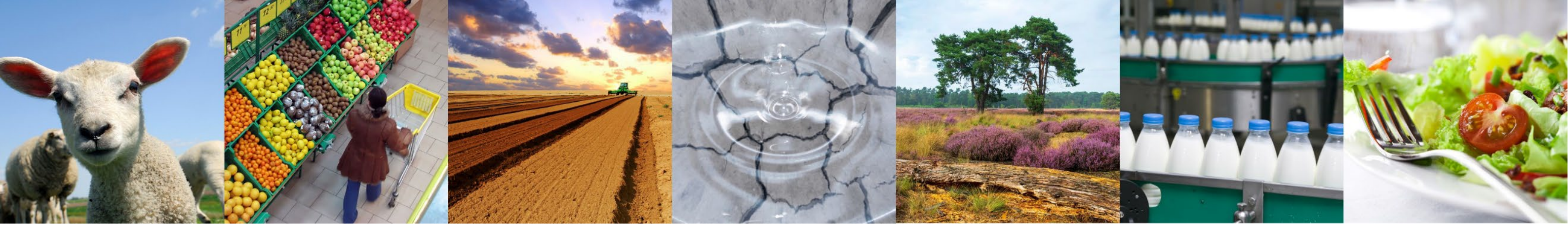
Yuchen Guo[1], Laura B. Zwep[1], Tingjie Guo[1], J. G. Coen van Hasselt[1]

[1]Division of Systems Pharmacology and Pharmacy, Leiden Academic Center for Drug Research, Leiden University, Leiden, The Netherlands

Abstract
Generation of virtual populations (VPs) which consist of sets of patient-associated covariates is commonly performed in model-based clinical trial simulation or the evaluation of optimized dosing strategies. To obtain realistic simulation results, the VPs generated as part of such model-based simulations should be representative of the real-world target population. In this context, a key hallmark of realistic VPs is that the simulated patient-associated covariates reflect not only the marginal distributions of the population of interest, but also their dependency structure. Recently, we have proposed the use of Copula models as a relevant strategy to support the simulation of virtual patient populations and demonstrated favorable performance of Copula in comparison to alternative simulation approaches [1]. Importantly, since Copula models are distribution-based, they enable sharing of patient-specific covariate data in the community without sharing original individual-level data. To encourage the application of copula in pharmacometrics and reduce the barriers of data sharing, a general workflow and a user-friendly platform are needed to guide and support the modelers and non-modelers to generate VPs using the Copula approach. In this study, we demonstrated a typical Copula model development workflow for virtual population simulation; based on the real-world population data available in the NHANES database [2], we systematically developed and evaluated a Copula model for simulation of commonly used covariates in adult individuals; moreover, a user-friendly web application is also developed to facilitate the use and share of Copula models.

Key words
multivariate prediction, Copula, virtual patient population, pharmacometrics, NHANES

References
[1] Zwep, Laura B., et al. PAGE 30 (2022) Abstr 10099 [www.page-meeting.org/?abstract=10099]
[2] Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention [https://wwwn.cdc.gov/nchs/nhanes/Default.aspx].



Statistical Modeling of Omics Data Using Two-Stage-PO2PLS

He Li[1], Zhujie Gu[2,3], Said el Bouhaddani[2], Jeanine Houwing-Duistermaat[1,4]

[1]Dept. of Mathematics, Radboud University, Nijmegen, The Netherlands
[2]Dept. of Data Science and Biostatistics, Julius Centre, UMC Utrecht, The Netherlands
[3]Medical Research Council Biostatistics Unit, University of Cambridge, UK
[4]Dept. of Statistics, University of Leeds, UK

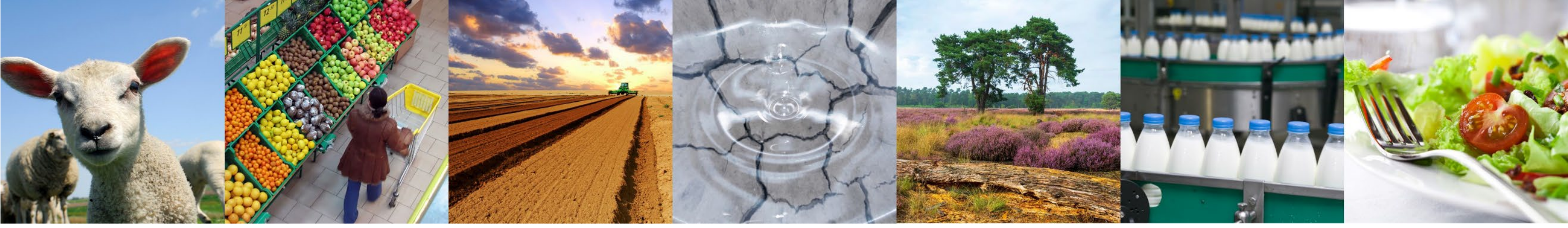
Abstract

Many studies are interested in the relationship between omic variables and outcomes. While omic variables change with age, genetic components of these variables are time invariant. Our work is motivated by the cohort ORCADES with information on single nucleotide polymorphisms (SNPs), glycomics and metabolomics, and the outcome variable body mass index (BMI). To estimate the genetic part of the omics data, polygenic risk scores for each omic variable (Omics-PRS), linear combinations of SNPs weighted by regression coefficients can be computed and included in a model for BMI. However, these methods ignore the genetic correlation between omic variables. An alternative would be the joint components of a latent variable model such as the PO2PLS model. A simulation study is performed to compare the performance of Omics-PRS and PO2PLS. We generate data with various dimensions using different models. For computing Omics-PRS, we use Lasso and Ridge to deal with the correlation between genetic markers. For modeling the outcome, we apply Ridge regression to deal with the large number of Omics-PRS variables. We evaluate the performance of methods using R square. We will show the results of simulation study and data analysis. Preliminary results show that PO2PLS outperforms Omics-PRS.

Key words
Latent Variable Models; Data Integration; Dimension Reduction; Polygenic Risk Score

References

[1] Choi SW, Mak TS and O'Reilly PF (2020). Tutorial: a guide to performing polygenic risk score analyses. Nature Protocols, 15(9), 2759–2772.
[2] Cook RD (2022). A slice of multivariate dimension reduction. Journal of Multivariate Analysis, 188, 104812.
[3] El Bouhaddani S, Uh HW and Houwing-Duistermaat J (2022). Statistical integration of heterogeneous omics data: probabilistic two-way partial least squares (PO2PLS). Journal of the Royal Statistical Society: Series C, 71(5), 1451–1470.
[4] Trygg J, Wold S (2002). Orthogonal projections to latent structures (O-PLS). Journal of Chemometrics, 16(3), 119–128.
[5] Trygg J, Wold S (2003). O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter. Journal of Chemometrics, 17(1), 53–64.



Contributed Session 5: Advances in Survival Modeling II

Chair: Ronald Geskus
Room: Momentum 2/3

Dynamic prediction of survival using multivariate Functional Principal Component Analysis: a strict landmarking approach

D. Gomon[1], M. Fiocco[1], H. Putter[1] and M. Signorelli[1]

[1]Leiden University; Leiden (the Netherlands)

Abstract
Dynamically updating the prediction of patient survival probabilities using longitudinal measurements has become of great importance with routine data collection becoming more common. Landmark analysis is a very popular approach for this problem due to its simplicity and computational feasibility. Recently, multi-step landmarking procedures have been developed where the longitudinal trajectories are first summarised using an appropriate method, such as Functional Principal Component Analysis (FPCA). The benefit of using FPCA is that no underlying structure needs to be specified for the longitudinal trajectories. Afterwards, these summaries are used in a survival model (e.g. Cox regression) to make predictions. Many of these approaches however fail to landmark the training data, an approach we call “relaxed” landmarking.

Longitudinal outcomes are often compared between subjects on a study time scale, even though the time of entry into the study might not be clinically relevant. Especially in an observational study this might not be the case, as participants will differ significantly in age at baseline. We would therefore like to eliminate the natural variation in the longitudinal trajectories caused by the age disparity between subjects before performing further analyses.

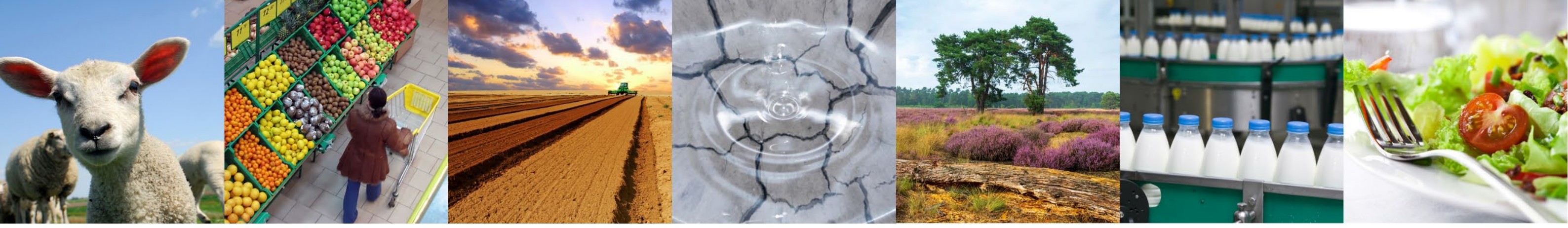
We develop an Age-based Centered multivariate Functional Principal Component Analysis (ABC mFPCA) technique to describe subjects using their age-at-observation and extend the multi-step landmarking approach proposed by Li and Luo (2019). We show in a simulation study that erroneously modelling covariates using time-on-study instead of using age-at-observation drastically reduces prediction accuracy. We formalise the difference between a “relaxed” landmarking approach where only validation data is landmarked and a “strict” landmarking approach where all data is landmarked. An application of our method to an observational study on Alzheimer’s disease (ADNI) shows that strict landmarking approaches significantly improve prediction accuracy.

Relaxed landmarking approaches introduce bias in the survival model, thereby failing to effectively use the information contained in the longitudinal outcomes and do not significantly improve prediction accuracy, whereas a strict landmarking approach can substantially improve prediction accuracy. Modelling longitudinal covariates on an appropriate and meaningful time scale is vital to extract relevant information for prediction purposes.

Key words
Dynamic prediction; Functional Principal Component Analysis; Landmark Analysis

References

[1] K. Li and S. Luo (2019): Dynamic prediction of Alzheimer’s disease progression using features of multiple longitudinal outcomes and time-to-event data. Statistics in Medicine, 38(24):4804-4818.



Smoothed likelihood loss for cross-validation in semi-parametric survival analysis

C. Lu[1], J. Goeman[1], H. Putter[1]

[1]Leiden University Medical Center; Leiden (The Netherlands)

Abstract

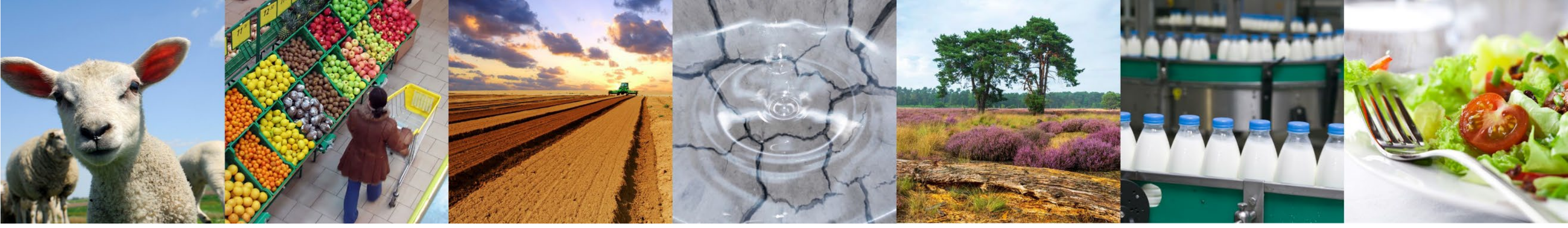
When building survival models for prediction, it is important to be able to assess predictive ability. It is common to use the train/test data set-ups and cross-validations with suitable loss functions. Many popular survival analysis methods, such as Cox models, additive hazards models, frailty models , etc, are semi-parametric or even non-parametric. So they return a step function as an estimation of the survival curve, with steps only at the locations of the event times in the training data set. Direct application of predictive likelihood as a loss function is therefore ineffective. The solution of Verweij [Verweij] (1993) propose replacing predictive likelihood by predictive partial likelihood. However, this solution is only suitable for Cox models. In this report, we propose solving the zero predictive likelihood problem by using a nearest neighbor kernel smoothing method. We compare our method with other existing methods, such as Verweij's method and the (integrated) Brier score, in simulated data in the Cox model, where Verweij's method can be used, and in a frailty models where it cannot. We apply the method to the problem of finding the optimal tuning parameter in a penalized non-parametric additive hazards model.

Key words

survival models; likelihood function; cross-validation

References

[1] Verweij (1993). Cross-validation in survival analysis. Statistics in Medicine: volume 12, 2305–2314.



Analyzing coarsened and missing data by imputation methods

Lars L.J. van der Burg[1], Stefan Böhringer[1], Jonathan W. Bartlett[2], Liesbeth C. de Wreede[1,3] and Hein Putter[1]

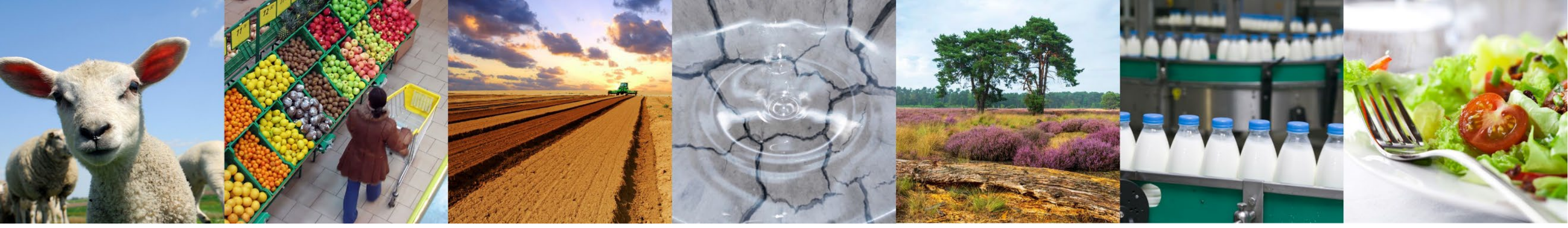
[1]Biomedical Data Sciences, LUMC, Leiden, The Netherlands
[2]Department of Medical Statistics, London School of Hygiene & Tropical medicine, London, UK
[3]DKMS, Dresden/Tübingen, Germany

Abstract
Many methods are available to handle missing data. For these methods, missing observations are assumed to be unrestricted with respect to the underlying distribution and are imputed with the help of correlated covariates and outcome. However, in many situations for some individuals it is known that their true value is restricted to a subset of the sample space. This phenomenon is known as coarsening. When imputing these coarsened observations using default imputation methods, their imputed values can violate restrictions, possibly leading to biased results. Here, we propose and investigate methods to correctly analyze such data for both linear regression and Cox proportional hazards models.

Imputations are performed by two popular methods: MICE and SMC-FCS, where various straightforward approaches of incorporating restrictions in the imputation step are examined. These approaches were compared to a complete case analysis and analyses that ignore restrictions. Additionally, an extension of the SMC-FCS algorithm is proposed, where restrictions are incorporated into the rejection sampling, thereby ensuring compatible imputation.

All methods were compared in a simulation study where a single main categorical predictor was coarsened. An auxiliary variable was created to improve performance of standard imputation methods. Varying choices for the correlation structure between predictor and auxiliary variable and proportion of coarsening in the categorical predictor were investigated in several scenarios.

Preliminary results show that considering coarsened data as complete missing data results in high misclassification. Incorporation of the restriction information in the current approaches generally leads to improvements. The new SMC-FCS extension leads to good results. Next, the approaches will be applied to a real cervix carcinoma dataset. Here, an important predictor is present, where for 20% of the individuals the observation is coarsened, such that only two of the three biomarker levels are compatible. These coarsened observations need to be modelled appropriately, in order to optimally estimate regression coefficients.



Contributed Session 6: Topics in Testing

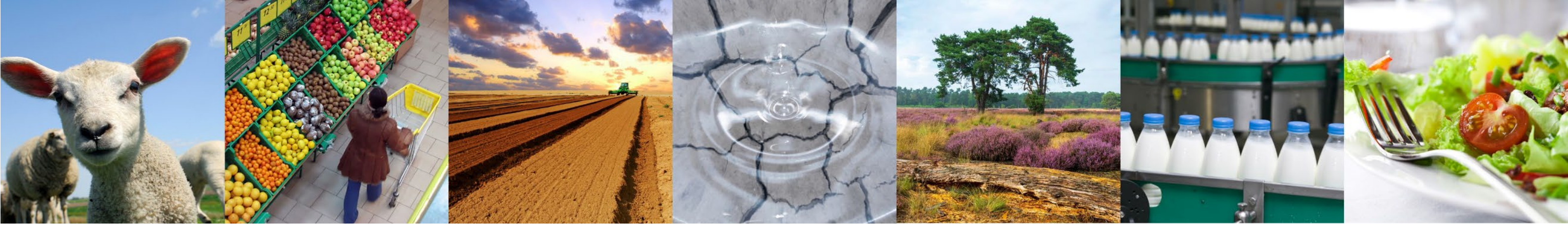
Chair: Jelle Goeman
Room: Momentum 1

Meta-meta-analysis

Erik van Zwet

Leiden University Medical Center, Leiden, The Netherlands

Abstract
There is much that can be learned about a particular research area from a collection of meta-analyses. The Cochrane Database of Systematic Reviews (CDSR) is a well-known example, but there are similar databases from psychology, ecology and economics. To do a "meta-meta-analysis", we summarize the result of each single study as a triple (beta,b,s) where beta is the "true" parameter, b is the unbiased, normally distributed estimator, and s is the standard error of b. We do not observe beta, but we do observe the pair (b,s). We define the z-value z=b/s and the signal-to-noise ratio SNR=beta/s. Despite the fact that the beta are not observed, it is possible estimate the joint distribution of the z-values and the SNRs across the database. This joint distribution is very useful because many important statistical concepts such as power, coverage, relative bias, probability of the correct sign, and probability of replication depend on (beta,b,s) only through z and SNR. We can also use the distribution as prior information to improve the inference in any study that may be considered to be "exchangeable" with the database. In the case of the CDSR, we can reduce the mean squared error by almost 50%. If time permits, we discuss how the extent of publication bias can be estimated, and how the joint distribution of the z-values and the SNRs can be adjusted for it.



Calculating exact, small P-values for high-dimensional data using plausibility models

Stefan Böhringer[1,2], Jesse J. Swen[1], Liesbeth de Wreede[2]

[1]Department of Clinical Pharmacy and Toxicology, Leiden University Medical Center, Leiden, The Netherlands
[2]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

Abstract

Instead of relying on asymptotic approximations, plausibility models rely on calculating the cumulative distribution function (CDF) of the test statistic exactly for a wide range of parametric models[1]. Additionally, exact finite sample P-values can be derived for model comparisons while allowing for nuisance covariates[2]. This applies to both small sample and high-dimensional settings. Exactness is achieved by optimizing the CDF over all possible nuisance parameter values. In most situations, closed form formulas do not exist and stochastic integration is necessary. As the CDF changes with changing nuisance parameters, optimization is hampered by variability induced by stochastic sampling.

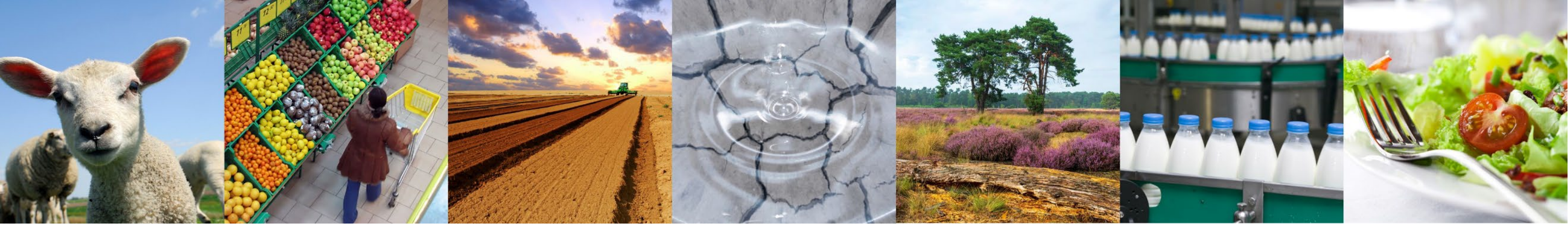
We develop a re-weighing scheme – borrowing from importance sampling – to avoid repetition of stochastic sampling throughout the optimization. Simultaneously, the sampling approach allows for the approximation of small P-values, where we aim for P-values as small as 10^{-10}. We investigate standard errors of P-values in simulations to prove reproducibility of computations. Validity of P-values is demonstrated through synthetic examples in the linear setting, where exact P-value computations under the alternative are possible.

We analyze genome wide SNP data with a binary toxicity outcome and correction for five covariates. We compare global plausibility P-values derived from SNP sets based on the local correlation structure. Plausibility uses elastic net regression for the plausibility weighting scheme, and P-values are also derived from SNP-by-SNP regressions. The global, plausibility-based P-value is 4.5×10^{-5} for the most strongly associated region, being genome wide significance (#regions=10^3) as opposed to single SNP findings (best P=5.8×10^{-8}, #snps=10^6). The standard error of the global P-value was ~10^{-6}, therefore being reproducible.

References

[1] Martin, Ryan. "Plausibility functions and exact frequentist inference." *Journal of the American Statistical Association* 110.512 (2015): 1552-1561.

[2] Böhringer, Stefan, and Dietmar Lohmann. "Exact model comparisons in the plausibility framework." Journal of Statistical Planning and Inference 217 (2022): 224-240.



Finite Sample Corrections for Average Equivalence Testing

Younes Boulaguiem[1], Julie Quartier[2,3], Maria Lapteva[2,3], Yogeshvar N Kalia[2,3], Maria-Pia Victoria-Feser[1], Stéphane Guerrier[1,2,3], Dominique-Laurent Couturier[4,5]

[1]Geneva School of Economics and Management, University of Geneva, Switzerland
[2]School of Pharmaceutical Sciences, University of Geneva, Switzerland
[3]Institute of Pharmaceutical Sciences of Western Switzerland, University of Geneva, Switzerland
[4]Medical Research Council Biostatistics Unit, University of Cambridge, England
[5]Cancer Research UK, Cambridge Institute, University of Cambridge, England

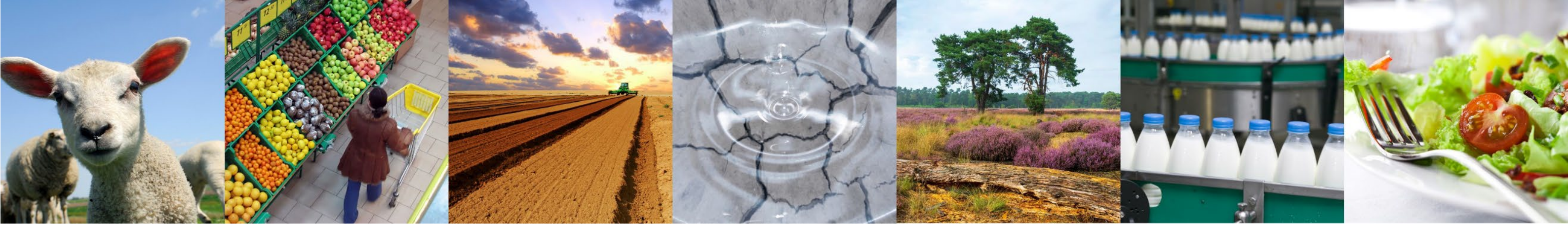
Abstract
Average (bio)equivalence tests are used to assess if a parameter, like the mean difference in treatment response between two conditions for example, lies within a given equivalence interval, hence allowing to conclude that the conditions have `equivalent' means. The two one-sided tests (TOST) procedure, consisting in testing whether the target parameter is respectively significantly greater and lower than some pre-defined lower and upper equivalence limits, is typically used in this context, usually by checking whether the confidence interval for the target parameter lies within the same limits. This intuitive and visual procedure is however known to be conservative, especially in the case of highly variable drugs, where it shows a rapid power loss, often reaching zero, hence making it impossible to conclude for equivalence when it is actually true.

Here, we propose a finite sample correction of the TOST procedure, the α-TOST, which consists in an adjustment of the level of the TOST allowing to guarantee a type-I error rate of α. This new procedure essentially corresponds to a finite sample and variability adjustment of the TOST procedure. We provide an algorithm to estimate the corrected α level and compare the type I error rate and power of our method to the ones of competing methods. A case study about econazole nitrate deposition in porcine skin is used to illustrate the benefits of the proposed method and its advantages compared to other available procedures.

We show that this procedure, also appealingly relying of the assessment of equivalence by means of confidence intervals, is uniformly more powerful than the TOST, easy to compute, and that its operating characteristics are better that the ones of its competitors.

Key words
Average bioequivalence; Finite sample correction

References
[1] Berger RL and Hsu JC (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. Statistical Science, 11, 283-319.
[2] Boulaguiem Y, Quartier J, Lapteva M, Kalia YN, Victoria-Feser MP, Guerrier S and Couturier DL (2023). Finite Sample Adjustments for Average Equivalence Testing. bioRxiv, 1-22.



Contributed Sessions 7-9

Contributed Session 7: Joint and Multistate Models

Chair: Liesbeth de Wreede
Room: Podium

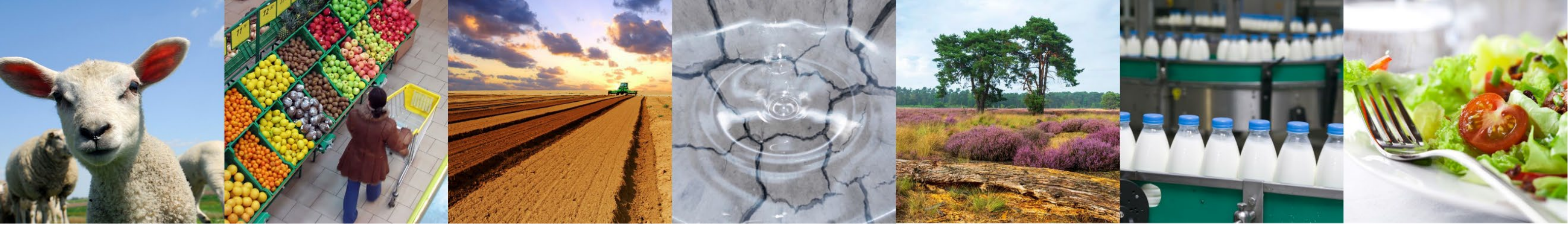
Modeling the underlying biological processes in Alzheimer's disease using a multivariate competing risk joint model

Floor van Oudenhoven[1,2,3], Sophie Swinkels[3], Tobias Hartmann[4,5], Dimitris Rizopoulos[1,2]

[1]Department of Biostatistics, Erasmus MC, Rotterdam, The Netherlands
[2]Department of Epidemiology, Erasmus MC, Rotterdam, The Netherlands
[3]Danone Nutricia Research, Utrecht, The Netherlands
[4]German Institute for Dementia Prevention (DIDP), Saarland University, Homburg, Germany
[5]Department of Experimental Neurology, Saarland University, Homburg, Germany

Abstract
Many clinical trials repeatedly measure several longitudinal outcomes on patients. Patient follow-up can discontinue due to an outcome-dependent event, such as clinical diagnosis or dropout. Joint modeling is a popular choice for the analysis of this type of data. Using data from a prodromal Alzheimer's disease trial, we propose a new type of multivariate joint model in which longitudinal brain imaging outcomes and memory impairment ratings are allowed to be associated with time to open-label medication and dropout, but also to directly affect one another. We model the dependence between the longitudinal outcomes so that a first longitudinal outcome affects a second one. Specifically, for each longitudinal outcome, we use a linear mixed-effects model to estimate its trajectory, where, for the second longitudinal outcome, we include the linear predictor of the first outcome as a time-varying covariate. This facilitates an easy and direct interpretation of the association between the longitudinal outcomes and provides a framework for latent mediation analysis to understand the underlying biological processes. The proposed joint models are fitted using a Bayesian framework via MCMC simulation.

Key words
Joint models; Competing risks; Mediation analysis; Clinical trials



Modelling dynamic associations in longitudinal studies with time-varying exposure

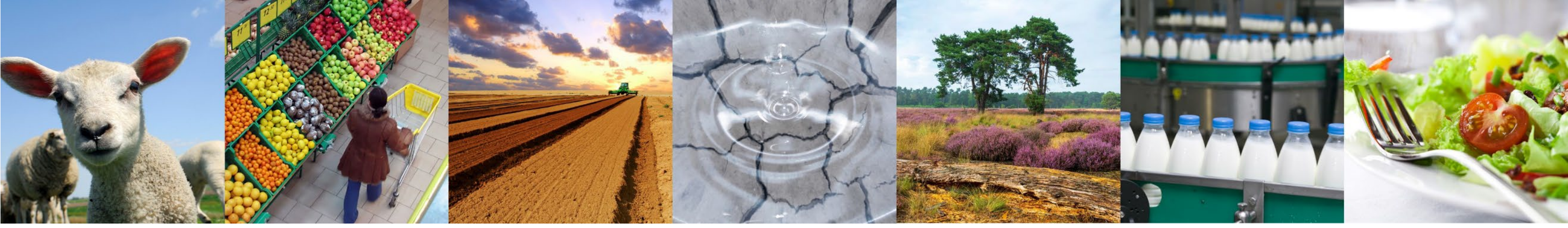
Roula Tsonaka[1], Georgy Gomon[2], Dimitris Rizopoulos[3], Bart Mertens[1]

[1]Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands
[2]Oncology Department, Leiden University Medical Center, Leiden, The Netherlands
[3]Department of Biostatistics, Erasmus University Medical Center, Rotterdam, The Netherlands

Abstract

In longitudinal studies measuring the association between the longitudinal response variable and time-varying covariates, such as the treatment dose or biomarkers, can be challenging for several reasons. First, the time-varying covariate may be endogenous, and thus standard multivariate models, e.g., mixed effects models for the longitudinal outcome, will be biased unless the covariate process is explicitly modeled. Second, the outcome and the covariate process may not be measured simultaneously, or missing data in any of the two processes can complicate their joint analysis. Finally, the functional form of the association is often not known beforehand, and thus several options need to be investigated. To address these challenges, we consider joint models that factorize the joint distribution of the two processes in terms of conditional densities. Our methods are exemplified using daily disease severity and biomarker data from the BEAT-COVID initiative at our home University Medical Center.

Key words
Joint Outcome Models; Random effects.



Multivariate longitudinal modeling with bounded outcomes and endogenous covariates

Chiara Degan[1], Pietro Spitali[1], Erik Niks[1], Bart Mertens[1], Roula Tsonaka[1], Jelle Goeman[1]

[1]Leiden University Medical Center, Leiden, The Netherlands

Abstract

Measuring the relationship between bounded longitudinal responses and endogenous time-varying covariates is not always trivial, and the use of simple Beta Mixed Models is no longer appropriate for mainly two reasons. First, it introduces bias when it fails to properly account for the dependence between the endogenous variable and the outcome history. Second, the longitudinal response and covariate can be measured at different points in time and may contain missing values. Multivariate models, on the other hand, could be utilized to analyze the relationship between the response and the endogenous variables. In this talk we consider two types of multivariate models, each assuming a different form of the association. One induces the association by jointly modeling the random effects, called Joint Mixed Models; the other quantifies the association using a scaling factor, called Joint Scaled Models. However, fitting these models is not straightforward, and their computational intensity, due to the high-dimensional integration over the random effects terms, limits their applicability. A flexible Bayesian estimation approach, known as INLA, will be used to fill this gap. Analytical work on these models will be presented along with the results of a clinical study conducted at Leiden University Medical Center.

Key words

Beta mixed models; Joint models; INLA; Endogenous covariates

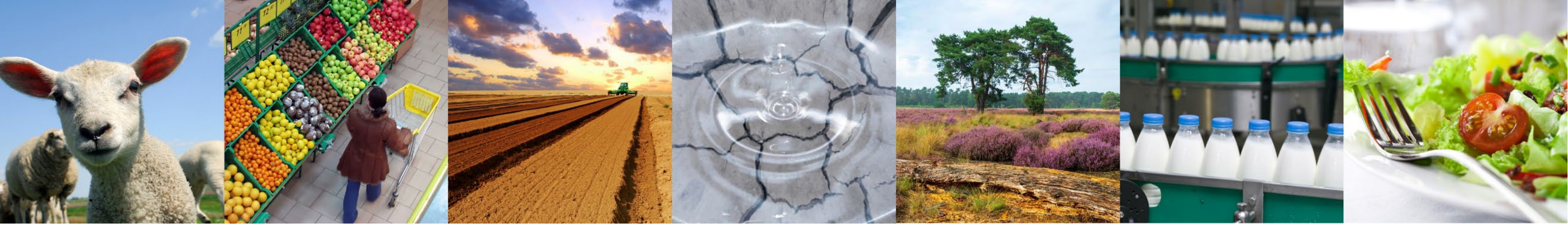
References

[1] Bonat WH, Ribeiro PJ, and Zeviani WM (2015). Likelihood analysis for a class of beta mixed models. Journal of Applied Statistics 42 (2): 252–66.

[2] Qian T, Klasnja P, and Murphy SA (2020). Linear Mixed Models with Endogenous Covariates: Modeling Sequential Treatment Effects with Application to a Mobile Health Study. Statistical science 35 (3): 375-390.

[3] Rue H, Martino S, and Chopin N (2009). Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71 (2): 319–92.

[4] Verbeke G, Fieuws S, Molenberghs G, and Davidian M (2014). The Analysis of Multivariate Longitudinal Data: A Review. Statistical Methods in Medical Research 23 (1): 42–59.



Estimating progressive multi-state disease models using imperfect screening data

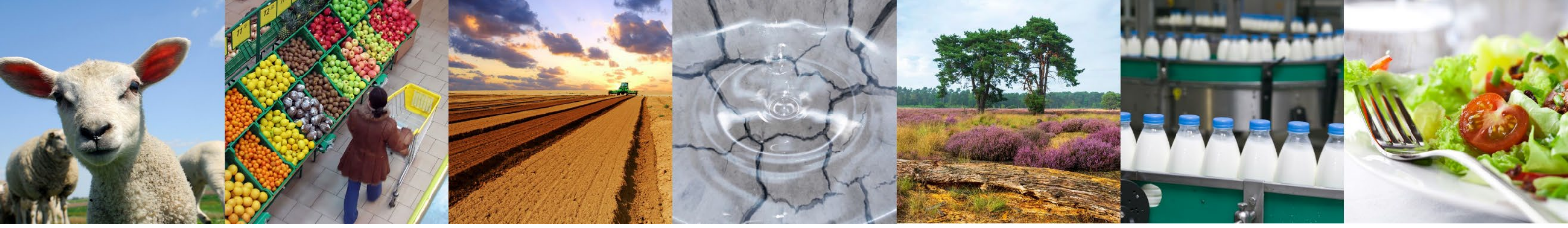
Michel Hof

Department of Epidemiology & Data Science, Amsterdam UMC, Amsterdam, the Netherlands

Abstract
Early detection of progressive diseases is vital for prompt treatment and effective management, which has led to the implementation of numerous population screening programs. We can use the screening data from these programs to estimate the individual-specific progression through disease states. In general, screening programs utilize multiple types of screening tests. These tests are often imperfect leading to miss-specified disease states. In addition, we only observe these imperfect states at irregular moments. Furthermore, clinical interventions might happen after a positive test and individuals are typically removed from the screening program afterwards. All this leads to complex interval-censored data.

To deal with these challenges, a novel model is proposed in which the unobserved disease process and screening routine are simultaneously estimated. The disease process is described by a progressive multi-state model. Given a particular disease progression, it is possible to model the screening data if the sensitivity and specificity of the screening tests are known. Parameters are estimated using the expectation-maximization algorithm by treating the disease process as missing data. The new method was applied to screening data from the POBASCAM study, containing longitudinal data on cervical cancer from approximately 40000 women.

Key words
Progressive disease model; Time-to-event data; Population screening; EM-algorithm; Imperfect tests



Contributed Session 8: Multi-Environment Genomic Prediction

Chair: Gregor Gorjanc
Room: Momentum 2/3

Predicting yield in multi-environment breeding trials using penalized regression and multiple-covariate random-regression models

Jip Ramakers[1], Jesse Hemerik[1], Martin Boer[1], Daniela Bustos-Korts[1], Javier Fernandez[2], Pengcheng Hu[2], Vivi Arief[2], Scott Chapman[2], Fred van Eeuwijk[1]

[1]Biometris, Wageningen University & Research, Wageningen, the Netherlands
[2]School of Agriculture and Food Sciences, The University of Queensland, Gatton, QLD, Australia

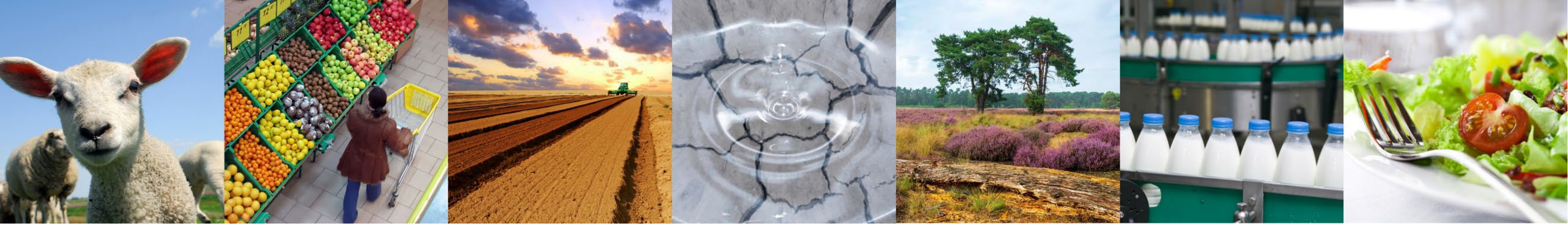
Abstract

The prediction of crop yield and other important agronomic traits from multi-environment breeding trials (METs) in new, untested environments (new locations in the next year) remains a major challenge in plant breeding (Malosetti *et al.* 2016). Traditionally, mixed-model approaches (anything between compound symmetry unstructured) are used to estimate genotype-by-environment interactions (G×E) in the MET and to extract predictions for new environments, calculated as genotypic averages across the MET. When G×E is substantial, however, these predictions will be inaccurate. With the advent of envirotyping and crop-growth modelling methods, however, continuous descriptors of the environment are becoming more widely available, providing us with the opportunity to make the modelled G×E predictable (e.g. Jarquín *et al.* 2014). At the same time, the sheer volume of environmental information brings challenges as to how to condense it into a manageable set of informative environmental predictors (EPs) for the estimation of G×E and prediction of yield.

We will describe a modelling framework used to describe G×E and predict yield, applied to a large multi-environment wheat (*Triticum aestivum*) dataset from Australian national variety trials, spanning 6 years, 150 locations and >100 EPs. We propose the use of genotype-specific penalized regression analysis across all EPs to identify a set of informative covariates across all genotypes. We then construct a set of random-regression models (RRMs) with orthogonal polynomials to predict yield in novel environments, ranging from single-EP RRMs, zone-specific RRMs that allow borrowing information from different agroecological zones, and multi-EP RRMs, and benchmark the models against standard models (e.g., compound symmetry and factor-analytic models). The first results suggest a quite substantial increase in predictive accuracy to be gained from the RRMs compared to conventional models.

References

[1] Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., Piraux, F., Guerreiro, L., Pérez, P., Calus, M., Burgueño, J. & de los Campos, G. (2014) A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and Applied Genetics,* **127,** 595-607.
[2] Malosetti, M., Bustos-Korts, D., Boer, M.P. & van Eeuwijk, F.A. (2016) Predicting responses in multiple environments: issues in relation to genotype× environment interactions. *Crop Science,* **56,** 2210-2222.



Assessing the response to genomic selection by simulation

Harimurti Buntaran[1], Angela Maria Bernal-Vasquez[2], Andres Gordillo[3], Morten Sahr[3], Valentin Wimmer[2], Hans-Peter Piepho[1]

[1]Biostatistics Unit, Institute of Crop Science, University of Hohenheim, Fruwirthstraße 23, 70599 Stuttgart, Germany
[2]KWS SAAT SE Co. KGaA, Grimsehlstaße. 31, 37574 Einbeck, Germany
[3]KWS-LOCHOW GmbH, Ferdinand-von-Lochow-Straße 5, 29303 Bergen, Germany

Abstract

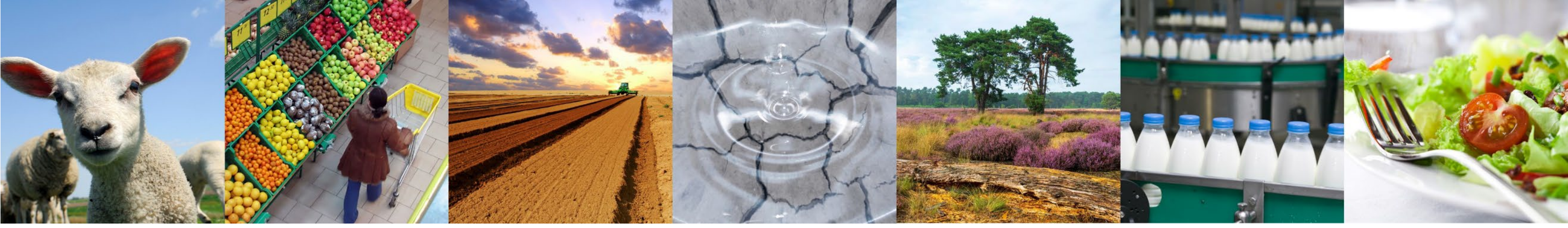
Any plant breeding program aims to maximize genetic gain for traits of interest. In classical quantitative genetics, the genetic gain can be obtained from what is known as “Breeder’s equation”. In the past, only phenotypic data were used to compute the genetic gain. The advent of genomic prediction (GP) has opened the door to the utilization of dense markers for estimating genomic breeding values or GBV. The salient feature of GP is the possibility to carry out the genomic selection with the assistance of the kinship matrix, hence improving the prediction accuracy and accelerating the breeding cycle. However, estimates of GBV as such do not provide the full information on the number of entries to be selected as in the classical response to selection. In this paper, we use simulation, based on a fitted mixed model for GP in a multi-environmental framework, to answer two typical questions of a plant breeder: (1) How many entries need to be selected to have a defined probability of selecting the truly best entry from the population; (2) what is the probability of obtaining the truly best entries when some top-ranked entries are selected.

Key words
genomic selection; multi-environment; response to selection; simulation

References

[1] Buntaran H, Bernal-Vasquez, AM, Gordillo, A, Sahr M, Wimmer, V, and Piepho H-P (2022). Assessing the response to genomic selection by simulation. Theor Appl Genet 135, 2891–2905.

[2] Piepho H-P, and Möhring J (2007). Computing heritability and selection response from unbalanced plant breeding trials. Genetics 177:1881–1888



Improving multi-environment genomic prediction with penalized factorial regression

Willem Kruijer[1], Vahe Avagyan[1], Martin Boer[1], Aalt D. J. van Dijk[2], Daniela Bustos Korts[3], Killian Melsen[1], Jip Ramakers[1], Fred van Eeuwijk[1]

[1]Biometris, Wageningen University and Research, Wageningen, The Netherlands
[2]Bioinformatics, Wageningen University and Research, Wageningen, The Netherlands
[3]Faculty of Agricultural Sciences, Universidad Austral de Chile, Valdivia, Chile

Abstract

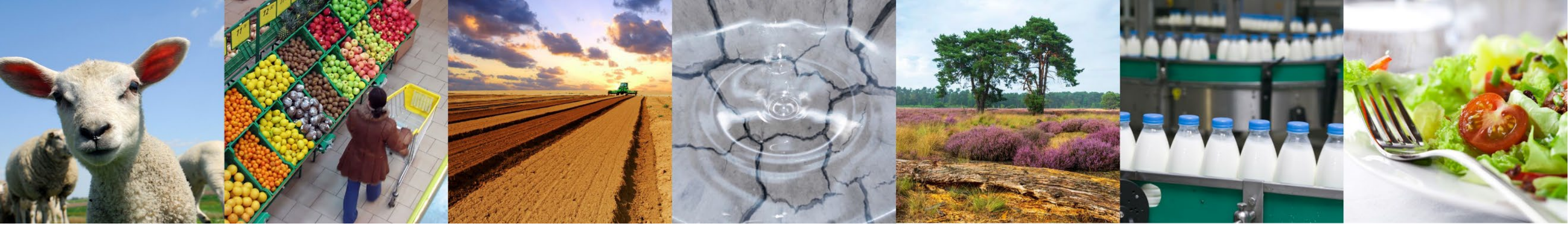
Genomic prediction has become an important tool in applications ranging from genomic selection to personalized medicine. While methodology for univariate genomic prediction is well-established, prediction for new environments remains a notorious challenge, which is however of great interest in plant 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 maximum daily temperature in a given stage of the growing season, which quantify both tested and untested environments. The most popular Bayesian approaches (Jarquin et al., 2014) are however computationally infeasible when the number of training environments gets beyond 30-50, while deep learning approaches (e.g., Khaki and Wang, 2019) only work well for very large numbers of environments.

Extending the work by Millet et al. (2019) we show that fast and accurate prediction is possible using penalized factorial regression (Avagyan et al., 2023). This can be implemented within a penalized regression framework (as in our R-package factReg) or alternatively in a mixed model framework (Boer, 2023). Using real and simulated data, our methodology achieves similar or better accuracies than existing approaches, with much lower computational cost. Finally we show how accuracies can be further improved if High-throughput phenotyping data is available. This is achieved by applying recent work on multi-trait genomic prediction to the sensitivities.

Key words
Multi-environment genomic prediction; penalized regression; deep learning

References

[1] Avagyan V., Boer M., van Dijk A. D. J., et al. (2023). Improving multi-environment genomic prediction with penalized factorial regression. Working paper.
[2] Boer M. (2023). Tensor product P-splines using a sparse mixed model formulation. Statistical Modelling (accepted for publication)
[3] 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.
[4] Khaki, S. and Wang, L. (2019). Crop yield prediction using deep neural networks. Frontiers in plant science, 10 , 621.
[5] Millet, E., Kruijer, W., Coupel-Ledru, A., et al. (2019). Genomic prediction of maize yield across European environmental conditions. Nature genetics, 51 (6), 952{956.



Contributed Session 9: Bayesian Methods

Chair: Mark van de Wiel
Room: Momentum 1

Identifying yield stability over time via time-varying GARCH processes and multivariate Horseshoe priors

John W. G. Addy[1], Chloe Maclaren[2], Kirsty Hassall[1]

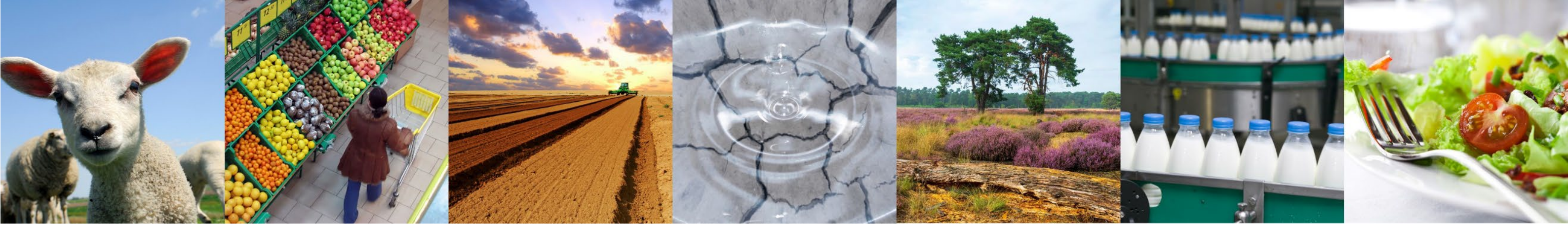
[1]Rothamsted Research, Hertfordshire, United Kingdom
[2]Swedish University of Agricultural Sciences, Uppsala County, Sweden

Abstract
Little is known about how grassland yield stability changes over time, where yield stability reflects the variance in yield, with higher-yielding grassland typically being less stable. Although Time-Varying Generalised Autoregressive Conditional Heterogeneity (TV-GARCH) processes are used heavily in economics and the medical sciences, presented is an application of these methods to better understand how conditional heterogeneous terms change over time for grassland ecosystems using data from the Park Grass Experiment, Hertfordshire.

Time varying processes can be modelled as a smooth non-parametric function, where Horseshoe priors (Carvalho et al. 2010) have been developed to prohibit overfitting of smooth functions through localized shrinkage of smooth coefficients through a series of zero-centred Normal distributions. Although Horseshoe priors provide good estimates for smoothed objects by localised shrinkage, we investigate the effects of covariate shrinkage through multivariate Horseshoe priors where the covariance matrix of a zero-centred multivariate Normal prior is decomposed into a vector of standard deviations and a matrix of correlations (Barnard et al. 2000). Covariate shrinkage allows for pairs of smooth coefficients to assume larger or smaller values depending on the correlation between coefficients and the shape of the overall smooth function. We develop a Bayesian TV-GARCH model using the variance-parameterised Gamma likelihood function (Addy et al. 2022). With this formalised modelling procedure we were able to identify time-varying changes in yield stability, providing novel insight into the underlying process and confirming previous findings of yield stability changing during the 1990s on Park Grass.

Key words
Bayesian Statistics; Horseshoe Priors; Co-Regulation; Heterogenity; Time-Series;

References
[1] Addy, JWG, Ellis, RH, MacLaren, C., Macdonald, AJ., Semenov, MA & Mead, A (2022). A heteroskedastic model of Park Grass spring hay yields in response to weather suggests continuing yield decline with climate change in future decades. Journal of the Royal Society Interface 19 (193).
[2] Barnard J, McCulloch, R & Meng, X.-L (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. Statistica Sinica pp. 1281–1311.
[3] Carvalho, CM, Polson, NG & Scott, J. G. (2010). The horseshoe estimator for sparse signals. Biometrika 97(2), 465–480.



On the distribution of individual causal effects of binary exposures using latent variable models

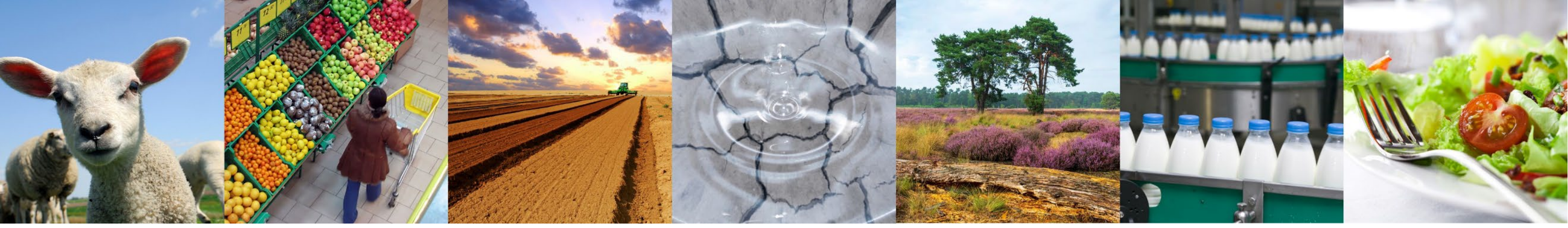
Richard Post[1], Edwin van den Heuvel[1]

[1]Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven,
Netherlands

Abstract
In recent years, the field of causal inference from observational data has emerged rapidly. This literature has focused on (conditional) average causal effect estimation. When (remaining) variability of individual causal effects (ICEs) is considerable, average effects may be less informative and possibly misleading for an individual. The fundamental problem of causal inference precludes estimating the joint distribution of potential outcomes without making assumptions, while this distribution is necessary to describe the heterogeneity of causal effects. In this chapter, we describe these assumptions and present a family of flexible latent variable models that can be used to study individual effect modification and estimate the ICE distribution from cross-sectional data. We will also discuss how the distribution is affected by misspecification of the error distribution or ignoring possible confounding effect heterogeneity. How latent variable models can be applied and validated in practice is illustrated in a case study on the effect of Hepatic Steatosis on a clinical precursor to heart failure. Assuming that there is (i) no unmeasured confounding and (ii) independence of the individual effect modifier and the potential outcome under no exposure, we conclude that the individual causal effect distribution deviates from Gaussian. We estimate that the 'treatment' benefit rate in the population is 23.7% (95% Bayesian credible interval: 2.6%, 53.7%) despite a harming average effect.

Key words
Causal inference; Heterogeneity of treatment effects; Latent variable modeling; Bayesian Methods

[1] Post RAJ, van den Heuvel ER (2022). On the distribution of individual causal effects of binary exposures using latent variable models. arXiv preprint, https://arxiv.org/abs/2210.16563.



Bayesian Federated Inference for Statistical Models

Marianne Jonker[1], Hassan Pazira[1], Emanuele Massa[2], Anthony Coolen[2,3]

[1]Department for Health Evidence, Section Biostatistics, Radboudumc, Nijmegen, The Netherlands
[2]Donders Institute, Faculty of Science, Radboud University, Nijmegen, The Netherlands
[3]Saddle Point Science Europe, Mercator Science Park, Nijmegen, The Netherlands

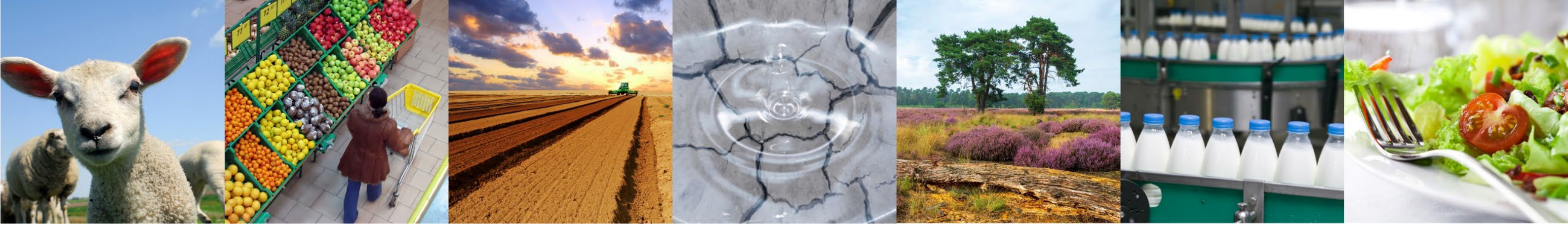
Abstract

Identifying predictive factors via multivariable statistical analysis is for rare diseases often impossible because the data sets available are too small. Combining data from different medical centers into a single (larger) database would alleviate this problem but is in practice challenging due to regulatory and logistic problems. Federated Learning (FL) is a machine learning approach that aims to construct from local inferences in separate data centers what would have been inferred had the data sets been merged. It seeks to harvest the statistical power of larger data sets without actually creating them. The FL strategy is not always feasible for small data sets. Therefore, we refine and implement an alternative Bayesian Federated Inference (BFI) framework for multi center data with the same aim as FL (Jonker et al, 2023). The BFI framework is designed to cope with small data sets by inferring locally not only the optimal parameter values, but also additional features of the posterior parameter distribution, capturing information beyond that is used in FL. BFI has the additional benefit that a single inference cycle across the centers is sufficient, whereas FL needs multiple cycles. We quantify the performance of the proposed methodology on simulated and real-life data.

Key words
Federated Learning; data integration; multi center data; MAP estimator; small data sets.

References

[1] Jonker MA, Pazira H, Coolen ACC (2023). Bayesian Federated Inference for Statistical Models. arXiv:2302.07677



Methane production rates of lactating dairy cows: A hierarchical Bayesian stochastic modelling approach

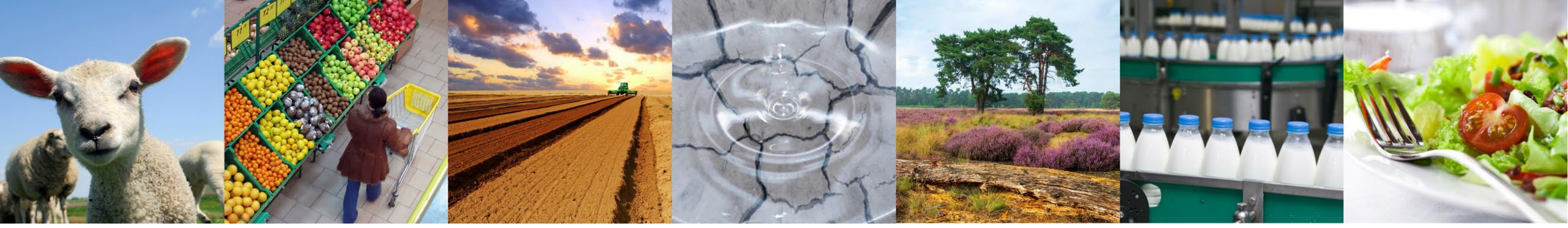
Cécile Levrault[1], Jan Dijkstra[2], Fred Van Eeuwijk[3], Peter Groot Koerkamp[1], Kelly Nichols[2], Nico Ogink[1,4], Carel F.W. Peeters[3]

[1]Farm Technology Group, Wageningen University & Research
[2]Animal Nutrition Group, Wageningen University and Research
[3]Mathematical and Statistical Methods Group (Biometris), Wageningen University & Research
[4]Wageningen Livestock Research, Wageningen University & Research

Abstract

Monitoring methane production from individual cows is crucial for the implementation of greenhouse gas reduction strategies. However, converting non-continuous measurements of methane concentration into daily methane production rates (MPR) remains challenging due to the non-linearity of the methane production curve. In this paper, we propose a Bayesian hierarchical stochastic kinetic equation approach to this challenge. Modelling was used to fit a non-linear curve on climate respiration chamber data before computing an area under the curve, therefore providing an estimate of methane production rate from individual cows. The shape parameters of this model were pooled across cows (population-level), while the scale parameter varied between individuals. This allowed for the characterization of variation in methane production rates within as well as between cows. Model fit was thoroughly investigated through posterior predictive checking, which showed that the model could reproduce the climate respiration chamber data of twenty-eight cows well. Comparison with a fully pooled model (all parameters constant across cows) was evaluated through cross-validation, where the Hierarchical Methane Rate (HMR) proved to perform better. Concordance between observed and predicted values was assessed with R2 (0.937), r (0.968), RMSE (9.95 g/d), and CCC (0.982, p = 3.81e-17). Overall, the predictions made by the HMR model appeared to reflect individual variation between cows better than the fully pooled model, and placed it as a promising approach for converting discrete measurements from practical monitoring methods into daily MPR.

Key words
dairy cows, diurnal production rate, enteric methane, Bayesian modelling.



Invited Session II: Advances in (Network) Meta-Analysis

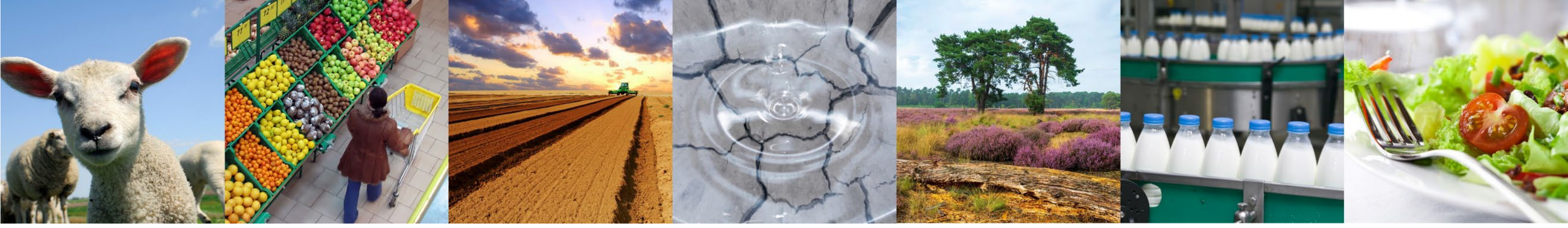
Chair: Olivier Thas
Room: Podium

Generalisability in surrogate endpoint evaluation and borrowing information across diverse sources when evaluating novel health technologies: meta-analytic approaches

Sylwia Bujkiewicz

Biostatistics Research Group, University of Leicester, UK

Abstract
Precision medicine research identifies subgroups of population, for example defined by genetic biomarkers in oncology, to which targeted therapies can be delivered successfully. As target populations become small, such novel therapies are increasingly evaluated in trials with short-term surrogate endpoints (such as progression free survival as a surrogate for overall survival) or single arm studies. As data on the final clinical outcome are either limited or not available at early stages of drug development, regulatory agencies grant accelerated licensing approval conditional on a surrogate marker. It is important that a surrogate endpoint is a reliable predictor of clinical benefit not only to ensure robust licensing approvals, but also to allow HTA agencies (such as Belgian KCE, Dutch ZIN, French HAS or NICE in the UK) to draw inferences from such limited evidence for reimbursement decisions - whether the new treatment is a good value for money to be recommended for use by health services. With small patient populations, synthesis of diverse sources of data and generalisability of surrogate relationships across treatments or even indications (disease areas) will play an increasingly important role in such decisions. We will discuss meta-analytic methods for borrowing of information from indirect sources of data; for example, across treatment classes or indications when evaluating surrogate endpoints and novel therapies.



The assessment of replicability with an application to meta-analysis

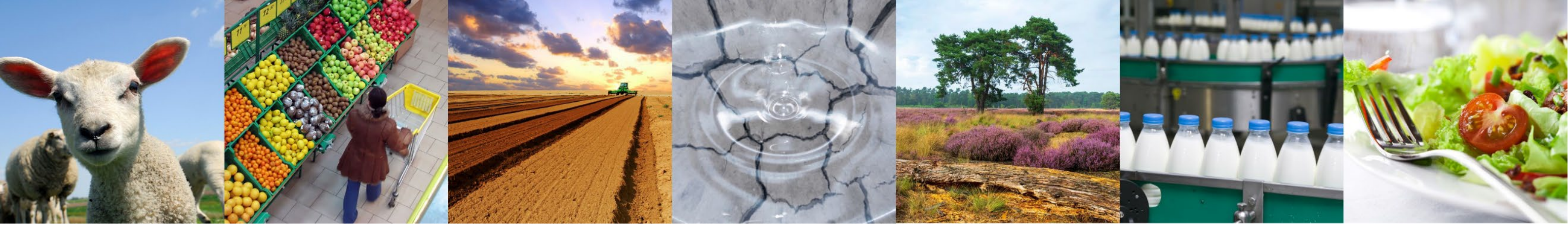
Leonhard Held

Epidemiology, Biostatistics and Prevention Institute, University of Zurich

Replicability of experiments is a cornerstone of the scientific method. A standard way to provide evidence of replicability is the two-trials rule by the US FDA which requests "at least two adequate and well controlled studies, each convincing on its own, to establish effectiveness" for drug approval. I will describe some alternatives based on p-value combination methods and compare the different methods with respect to relevant operating characteristics (Held, 2023). The corresponding p-value functions (Infanger and Schmidt-Trucksäss, 2019) can then be used to provide novel meta-analytic confidence intervals for the combined treatment effect with some interesting properties. A comparison with more standard approaches will be provided based on results from an extensive simulation study.

References

[1] L. Held (2023). Beyond the two-trials rule. http://arxiv.org/abs/2307.04548
[2] D. Infanger and A. Schmidt-Trucksäss (2019). P value functions: An underused method to present research results and to promote quantitative reasoning. Statistics in Medicine, 38(21):4189–4197. doi: 10.1002/sim.8293.



Contributed Sessions 10-12

Contributed Session 10: Clinical & Medical Statistics

Chair: Roula Tsonaka
Room: Podium

When exactly? Two overlooked biases in SARS-CoV-2 incubation time estimation related to information regarding exposure

V.H. Arntzen[1], M. Fiocco[1,2,3], R.B. Geskus[4,5]

[1]Mathematical Institute, Leiden University, Leiden (The Netherlands)
[2]Biomedical data, Leiden University Medical Center, Leiden (The Netherlands)
[3]Princess Maxima Center for child oncology, Utrecht (The Netherlands)
[4]Oxford University Clinical Research Unit, Ho Chi Minh City (Viet Nam)
[5]Centre for Tropical Medicine and Global Health, University of Oxford, Oxford (United Kingdom)

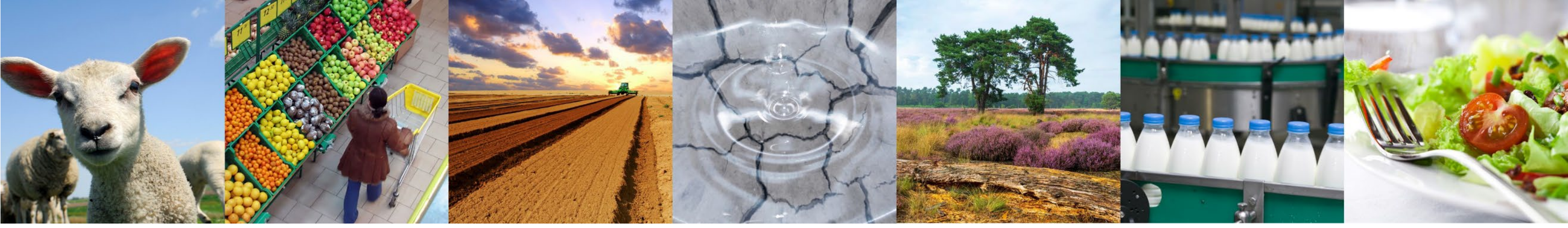
Abstract
Exposure information is essential for estimating the incubation time of an infectious disease (infection to symptom onset) and is typically interval censored. Interval censored time origins complicate analysis [1]. For SARS-CoV-2, data on exposure has been collected retrospectively, mostly by interviewing detected cases. However, our memory is imperfect, with recent exposures being more precisely recalled than older ones (differential recall). Interval-censored observations with less precise exposures are often excluded from analysis. This creates bias, because longer incubation times are more likely to be excluded.

Another bias in the initial estimates of SARS-CoV-2 incubation time arises from data collected from travellers from Wuhan. Only individuals who developed symptoms after departure were included, resulting in an underrepresentation of cases with short incubation times (left truncation).

A simulation study with incubation times generated from a Weibull distribution (median 5.4 days, 95th percentile 9.8 days) shows that in presence of differential recall, restricting the analysis to a subset of narrow exposure windows leads to underestimation up to 5 days in the median and even more in the 95th percentile, depending on the rate of differential recall. Neglecting left truncation leads to considerable overestimation, inflating the median and 95th percentile by multiple days.

Key words
differential recall; left truncation; incubation time; SARS-CoV-2; interval-censoring.

References
[1] Arntzen VH, Fiocco M, Leitzinger N, Geskus RB (2023). Towards robust and accurate estimates of the incubation time distribution, with focus on upper tail probabilities and SARSCoV-2 infection. Statistics in Medicine, 2023;1-20.



The impact of national and international travel on spatio-temporal transmission of SARS-CoV-2 in Belgium in 2021

Minh Hanh Nguyen[1], Thi Huyen Trang Nguyen[1], Geert Molenberghs[1,2], Steven Abrams[1,3], Niel Hens[1,3,4], Christel Faes[1,2]

[1]Data Science Institute, I-BioStat, Hasselt University, Hasselt, Belgium
[2]I-BioStat, Katholieke Universiteit Leuven, Leuven, Belgium
[3]Global Health Institute, University of Antwerp, Antwerpen, Belgium
[4]Vaccine and Infectious Disease Institute, University of Antwerp, Antwerpen, Belgium

Abstract

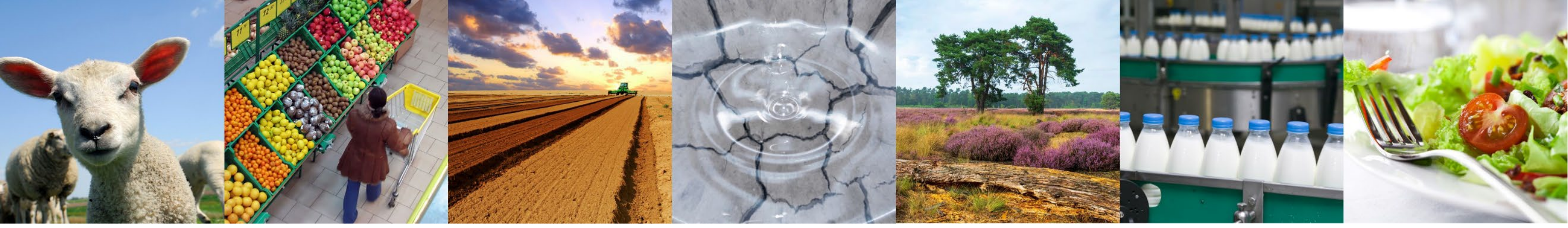
Background: Understanding the mechanism responsible for the spread of the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) and the impact of specific factors will help authorities to tailor interventions. We aim to analyze the spatio-temporal transmission of SARS-CoV-2 in Belgium at municipality level between January and December 2021 and explore the effect of different levels of human travel on disease incidence.

Methods: We applied the endemic-epidemic modelling framework. The spatial dependencies among areas are adjusted based on actual connectivity through mobile network data. We also considered other important factors such as international mobility, vaccination coverage, population size and the stringency of restriction measures.

Results: The results demonstrate the aggravating effect of international travel on the incidence, and simulated counterfactual scenarios further stress the alleviating impact of a reduction in national and international travel on epidemic growth. Local transmission contributed the most during 2021, and municipalities with a larger population tended to attract a higher number of cases from neighboring areas.

Conclusions: Although transmission between municipalities was observed, local transmission was dominant. We highlight the positive association between the mobility data and the infection spread. Our study provides insight to assist health authorities in decision-making, particularly when the disease is airborne.

Key words
Spatio-temporal model; COVID-19; human mobility; international travel.



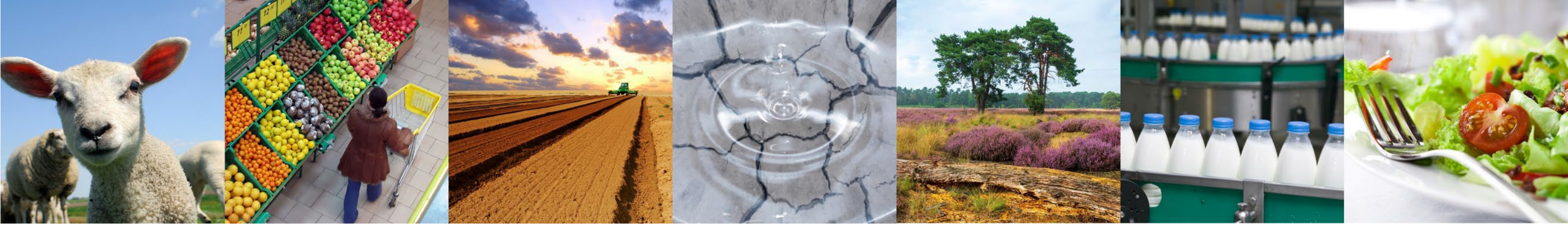
Risk ratio estimation in a regression discontinuity design: Application to the prescription of statins for cholesterol reduction in UK Primary Care

Aidan O'Keeffe[1], Mariam Adeleke[2], Gianluca Baio[2]

[1]University of Nottingham
[2]University College, London

Abstract
In recent years regression discontinuity (RD) designs have been used increasingly for the estimation of treatment effects in observational medical data where a rule-based decision to apply a treatment is taken using a continuous assignment variable. Most RD design applications have focused on effect estimation where the outcome of interest is continuous, with scenarios with binary outcomes receiving less attention, despite their ubiquity in medical studies. In this work we develop an approach to estimation of the risk ratio in a fuzzy RD design (where treatment is not always strictly applied according to the decision rule), derived using common RD design assumptions. This method compares favourably to other risk ratio estimation approaches: the established Wald estimator and a risk ratio estimate from a multiplicative structural mean model, with promising results from extensive simulation studies. A demonstration and further comparison is made using a real example to evaluate the effect of statins (where a statin prescription is made based on a patient's 10-year cardiovascular disease risk score) on LDL cholesterol reduction in UK Primary Care.

Key words
Regression discontinuity design ; Risk ratio ; Observational study ; Binary data.



Contributed Session 11: Statistical Genetics

Chair: Andrew Mead
Room: Momentum 2/3

A new approach for assessing distinctness of new crop varieties using genetic markers

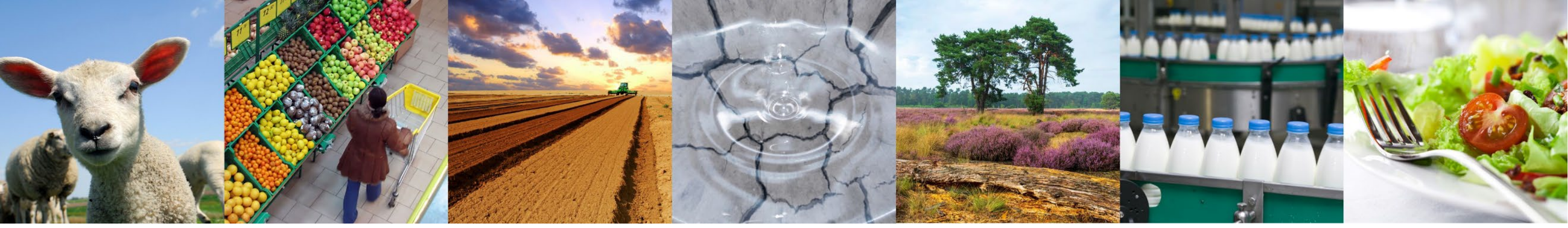
Adrian M.I. Roberts[1], Konrad Neugebauer[1]

[1]Biomathematics and Statistics Scotland, Edinburgh, United Kingdom

Abstract
To register a new crop variety for sale or to gain property rights (https://www.upov.int/overview/en/), it must be shown to be distinct from existing varieties. The assessment of distinctness is based on physical characteristics, such as morphology or phenology. Field trials to assess distinctness can be costly, due to the need to compare with all existing varieties. In some crops, existing varieties may be found distinct *a priori*, for example based on information supplied by applicants, thus reducing the trial size. However, this is difficult for outcrossing species. The use of genetic markers, such as SNPs, is being explored to identify which existing varieties are clearly different from candidates. A commonly-used method is to compare genetic distance with a phenotypic distance, with the latter computed by combining information over characteristics. In this case, the correlation between the two distances is often poor, reducing its usefulness.

We propose a more targeted approach, based on predicting distinctness decisions for each characteristic in turn using markers. Genomic prediction methods are adapted to predict distinctness decisions for each characteristic in turn. Information can then be combined to provide an overall probability of distinctness based on the markers.

Key words
Distinctness; Variety; Genomic Prediction; Decision-making



Applying IBD-based methods for QTL identification in hybrid potato breeding populations

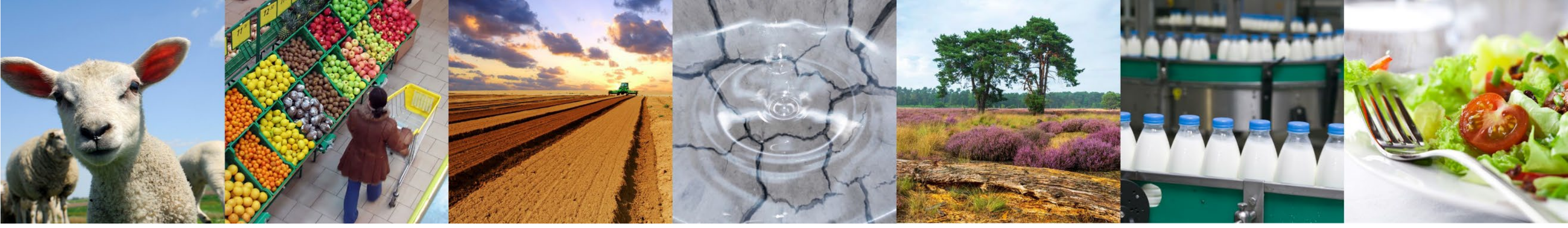
James Adams[1,2], Chaozhi Zheng[1], Fred van Eeuwijk[1]

[1]Biometris, Wageningen University, Wageningen, The Netherlands
[2]Solynta Hybrid Potato Breeding, Wageningen, The Netherlands

Abstract
Tracking the presence of favourable quantitative trait loci (QTL) within active breeding programmes has been a methodological problem for applied geneticists for decades. The challenge of identification of QTL in such populations is the inflexibility of classical analyses to deal with multi-parent origin and deviation from idealised allelic frequencies due to high selection pressure. Development of analyses for such multiparental populations (MPP's) have been an active area of research, but there are still very few practical examples of their use in populations with no formal design (Eeuwijk et al., 2010). We apply several identity by descent (IBD) based models using founder QTL effects on a panel of inbred diploid potato lines derived from 16 common founders. IBD probabilities were estimated using the reconstructing ancestry blocks bit by bit (RABBIT) software based on a Hidden Markov modelling methodology (Zheng, Boer, & Eeuwijk, 2015). Using the linear mixed modelling approach of (Li et al., 2021), we investigate the genetic control of several yield components in diploid hybrid potato based upon a common set of founder QTL effects. We hope that this demonstrates the potential of IBD-based models in the most challenging scenarios in genetic modelling.

Key words
Identity by Descent; Linear Mixed Modelling; Hybrid Potato; Multi-parent populations

References
[1] Li, Wenhao, Martin P. Boer, Chaozhi Zheng, Ronny V. L. Joosen, and Fred A. van Eeuwijk (2021). An IBD-Based Mixed Model Approach for QTL Mapping in Multiparental Populations. Theoretical and Applied Genetics: 134(11) 3643–60. https://doi.org/10.1007/S00122-021-03919-7.
[2] Zheng, Chaozhi, Martin P Boer, and Fred A Van Eeuwijk (2015). Reconstruction of Genome Ancestry Blocks in Multiparental Populations Genetics: 200, 1073-1087. https://doi.org/10.1534/genetics.115.177873.
[3] Eeuwijk, Fred A. van, Martin Boer, L. Radu Totir, Marco Bink, Deanne Wright,Christopher R. Winkler, (2010). Mixed Model Approaches for the Identification of QTLs within a Maize Hybrid Breeding Program. Theoretical and Applied Genetics: 120(2) 429–40. https://doi.org/10.1007/s00122-009-1205-0.



Consensus genetic map construction in connected multiparental populations

Chaozhi Zheng[1], Eligio Bossolini[2], Antje Rohde[2], Martin Boer[1], Fred van Eeuwijk[1]

[1]Biometris, Wageningen University, Wageningen, The Netherlands
[2]BASF Innovation Center Gent, Gent, Belgium

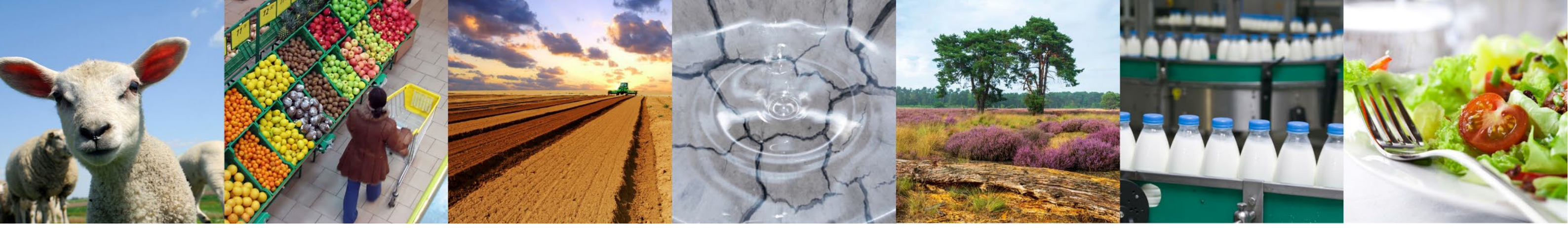
Abstract

To dissect the genetic architecture of complex traits, many multiparental populations have been produced in crops, since they have high genetic diversity in compare to traditional biparental populations. Additionally, many plant breeding programs rely on multiparental populations. While these population types can increase genetic diversity and the power to detect quantitative trait loci (QTL), they pose new challenges to data analysis. Previously, we have developed a hidden Markov framework MagicMap for map construction in a single multiparental population (Zheng et al., 2019). In this work, we extend MagicMap for multiple multiparental populations that may be connected by shared parents. The new algorithm increases computational efficiency by iteratively analyzing each subpopulation, as compared to the previous version that considered all founders simultaneously and thus its computational time increased quickly with the number of parents. In addition, the new algorithm is written in the high-performance Julia language, instead of Mathematica for the previous version. We demonstrate MagicMap by constructing a consensus map from three maize multiparental populations: the nested association mapping (NAM) (Bauer et al., 2013), the multi-parent advanced generation inter-cross (MAGIC) (Dell'Acqua et al., 2015), and the twelve recombination inbred line (RIL) populations (Pan et al., 2016).

Key words
QTL mapping, Connected multiparental populations, Consensus linkage map, Map construction.

References

[1] Bauer E, Falque M, Walter H, et al. (2013). Intraspecific variation of recombination rate in maize. Genome Biology, 14:R103.
[2] Dell'Acqua M, Gatti D, Pea G, et al. (2015). Genetic properties of the MAGIC maize population: a new platform for high definition QTL mapping in Zea mays. Genome Biology, 16, 167.
[3] Pan Q, Li L, Yang X, et al. (2016). Genome-wide recombination dynamics are associated with phenotypic variation in maize. New Phytol. 210, 1083–1094.
[4] Zheng C, Boer MP, and Eeuwijk FA (2019). Construction of genetic linkage maps in multiparental populations. Genetics, 212, 1031-1044.



Contributed Session 12: Statistical Kaleidoscope

Chair: Jos Hageman
Room: Momentum 1

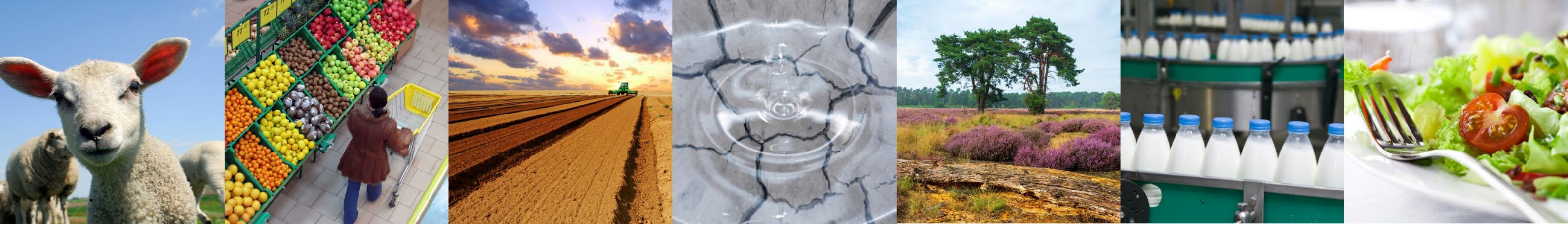
The out-of-sample R^2: estimation and inference

Stijn Hawinkel[1,2], Willem Waegeman[3], Steven Maere[1,2]

[1]Department of Plant Biotechnology and Bioinformatics, Ghent University, Belgium
[2]VIB Center for Plant Systems Biology, Belgium
[3]Department of Data Analysis and Mathematical Modelling, Ghent University, Belgium

Abstract
Out-of-sample prediction is the acid test of predictive models, yet an independent test dataset is often not available for assessment of prediction error. For this reason, out-of-sample performance is commonly estimated using data splitting algorithms such as cross-validation or the bootstrap. For quantitative outcomes, the ratio of variance explained to total variance can be summarized by the coefficient of determination or in-sample R^2, which is easy to interpret and to compare across different outcome variables. As opposed to the in-sample R^2, the out-of-sample R^2 has not been well defined and the variability on the out-of-sample $\hat{R}^2$ has been largely ignored. Usually only its point estimate is reported, hampering formal comparison of predictability of different outcome variables. Here we explicitly define the out-of-sample R^2 as a comparison of two predictive models, provide an unbiased estimator and exploit recent theoretical advances on uncertainty of data splitting estimates to provide a standard error. The performance of the estimators for the R^2 and its standard error are investigated in a simulation study. We demonstrate our new method by constructing confidence intervals for prediction of quantitative Brassica napus and Zea mays phenotypes based on gene expression data. Our method is available in the R-package oosse.

Key words
Prediction, Coefficient of Determination, Standard error, Cross-validation, Bootstrap



Asymptotics of Caliper Matching Estimators for Average Treatment Effects

Máté Kormos[1], Stéphanie van der Pas[2,3], Aad van der Vaart[1]

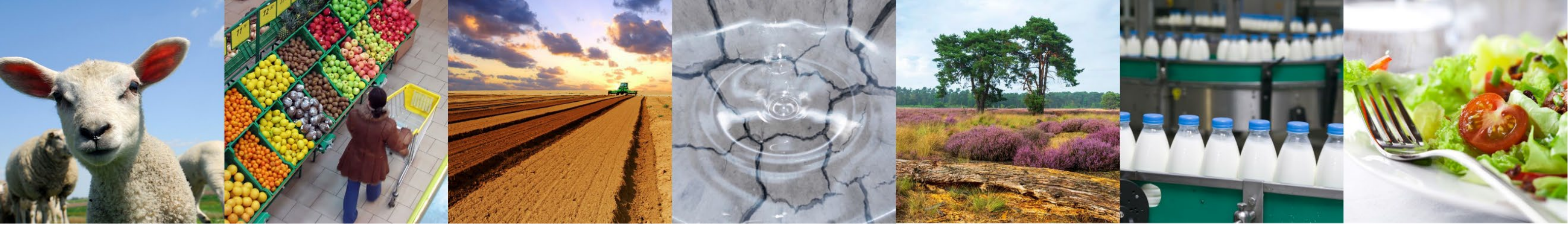
[1]Delft Institute of Applied Mathematics, Delft University of Technology, Delft, The Netherlands
[2]Amsterdam UMC location Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
[3]Amsterdam Public Health Methodology, Amsterdam, The Netherlands

Abstract

Caliper matching is used to estimate causal effects of a binary treatment from observational data by comparing matched treated and control units. Units are matched when their propensity scores, the conditional probability of receiving treatment given pretreatment covariates, are within a certain distance called caliper. So far, theoretical results on caliper matching are lacking, leaving practitioners with ad-hoc caliper choices and inference procedures. We bridge this gap by proposing a caliper that balances the quality and the number of matches. We prove that the resulting estimator of the average treatment effect, and average treatment effect on the treated, is asymptotically unbiased and normal at parametric rate. We describe the conditions under which semiparametric efficiency is obtainable, and show that when the parametric propensity score is estimated, the variance is increased for both estimands. Finally, we construct asymptotic confidence intervals for the two estimands.

Key words

caliper matching estimator; radius matching estimator; average treatment effect; causal inference; asymptotic statistics



Decoding Biodiversity Change

Hideyasu Shimadzu[1,2]

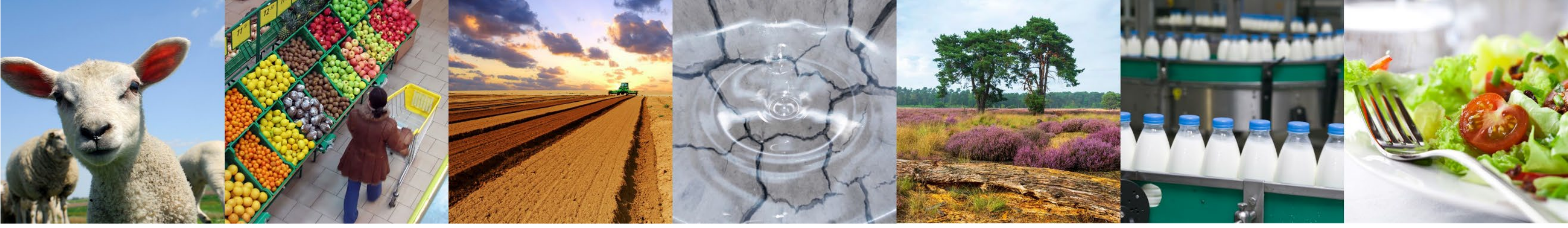
[1]Department of Mathematical Sciences, Loughborough University, Loughborough, UK
[2]Department of Data Science, Kitasato University, Kanagawa, Japan

Abstract

Increasing concerns underline the unprecedented pressures contributing to the ongoing biodiversity crisis on Earth. Understanding the intricacies of contemporary biodiversity change has thus been a pivotal pursuit within the fields of ecological and conservation sciences. Despite the fact that various biodiversity indices are proposed to quantify the state of biodiversity, these indices sometimes can yield conflicting outcomes. A missing yet crucial aspect lies in formulating a unified framework that accommodates a more cohesive understanding of these diverse measures of biodiversity.

The talk embarks on exploiting the classical ecological biodiversity concepts, namely alpha-, beta- and gamma-diversities. We then propose a little formal interpretation grounded in abundance distributions. Subsequently, we derive compelling evidence that shifts in species abundance distribution correspond to a specific type of biodiversity change. In other words, the widely employed biodiversity indices fundamentally operate as estimators of distributional deviation from one state to another. Through the exposition of our framework, we illustrate critical concepts in biodiversity study, providing some insights into community dynamics modelling.

Key words
Abundance distributions; Biodiversity; Divergence; Ecology; Species.



Invited Session III: Bayesian Machine Learning & Optimal Design

Chair: Pierre LeBrun
Room: Podium

Extensions to probabilistic tree-based machine learning algorithms

Estevão Prado

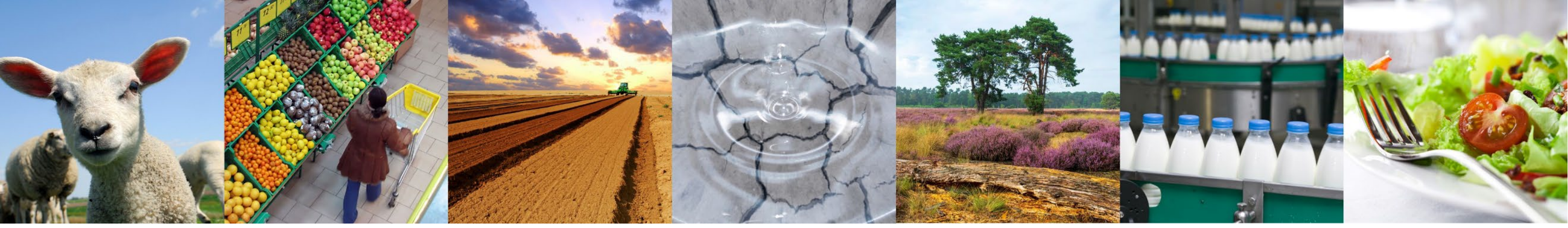
Department of Mathematics & Statistics, Lancaster University, UK

Abstract
Bayesian additive regression trees (BART) is a tree-based machine learning method that has been successfully applied to regression and classification problems. BART assumes regularisation priors on a set of trees that work as weak learners and is very flexible for predicting in the presence of non-linearities and low-order interactions. In this talk, we present two extensions to semi-parametric models based on BART. First, we propose a new class of models for the estimation of genotype-by-environment interactions in plant-based genetics. Our approach uses semi-parametric BART to accurately estimate marginal genotypic and environment effects along with their interaction in a cut Bayesian framework. We demonstrate that our approach is competitive or superior to similar models widely used in the literature via both simulation and a real-world dataset. Second, we extend semi-parametric BART models with a view to analysing data from an international education assessment, where certain predictors of students' achievements in mathematics are of particular interpretational interest. Through additional simulation studies and another application to a well-known benchmark dataset, we also show competitive performance when compared to regression models, alternative formulations of semi-parametric BART, and other tree-based methods.

This talk is based on

1 - Sarti, D.A.*, Prado, E.B.*, Inglis, A.N., dos Santos, A.A.L, Hurley, C.B., Moral, R.A, Parnell, A.C. Bayesian additive regression trees for genotype by environment interaction models. *Annals of Applied Statistics (to appear)*. * joint first authors.

2 - Prado, E.B., Parnell, A.C., Murphy, K., McJames, N., OShea, A. & A., Moral, R.A. Accounting for shared covariates in semi-parametric Bayesian additive regression trees (2022). https://arxiv.org/pdf/2108.07636.pdf

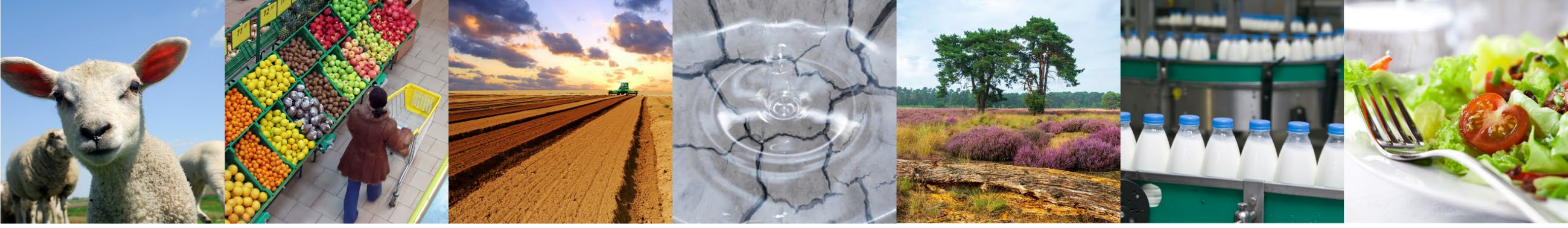


Optimal design of experiments using amortized Bayesian inference and conditional normalizing flows

Matthias Brückner

Janssen Pharmaceutica, Beerse, Belgium

Bayesian optimal design of experiments involves choosing a design that maximizes the expected utility, where the utility function depends on the posterior distribution of the model parameters. Analytically calculating the expected utility is infeasible in most cases. Estimates of the expected utility can be obtained using Monte-Carlo methods by averaging over a large number of datasets sampled from the prior predictive distribution. Evaluating the utility function in turn requires Monte-Carlo Markov Chain methods to obtain posterior samples except in simple models with conjugate priors. Repeating this for thousands of simulated datasets for every design option can be computationally prohibitive. Recent advances in amortized Bayesian inference offer a potential solution to this problem. Here the Bayesian inference process is divided into a costly training phase and a much cheaper inference phase. During the training phase a neural network learns an estimator for the probabilistic mapping from data to underlying model parameters. The trained neural network can then without additional training or optimization efficiently sample from the posterior distributions for arbitrary many datasets involving the same model family. We explore the computational advantages and possible applications in the design of non-clinical experiments.

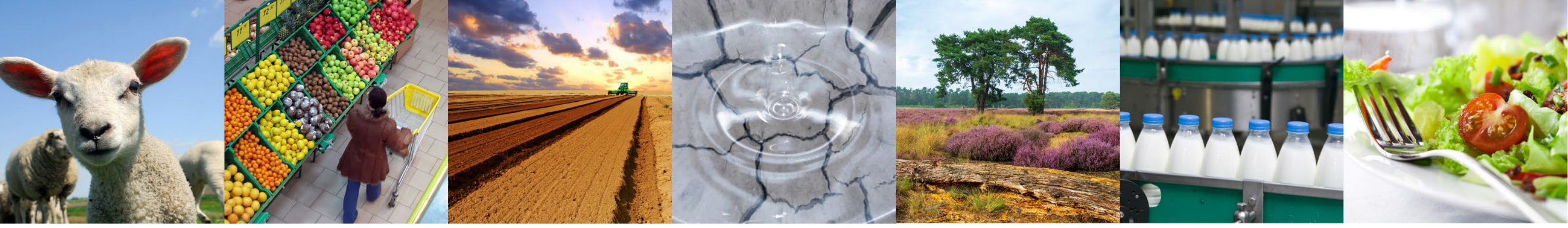


Bayesian Optimization approaches for optimal dose combination identification in early phase dose finding trials

James Willard

School of Population and Global Health, McGill University, Montreal, Canada

Identification of optimal dose combinations in early phase dose-finding trials is challenging due to the trade-off between precisely estimating a large number of parameters required to flexibly model the dose-response surfaces, and the small sample sizes in early phase trials. Existing methods often restrict the search to pre-defined dose combinations, which may fail to identify regions of optimality in the dose combination space. These difficulties are even more pertinent in the context of personalized dose-finding, where patient characteristics are used to identify tailored optimal dose combinations. To overcome these challenges, we propose the use of Bayesian optimization for finding optimal dose combinations in standard ("one size fits all") and personalized multi-agent dose-finding trials. Bayesian optimization is a method for estimating the global optima of expensive-to-evaluate objective functions. The objective function is approximated by a surrogate model, commonly a Gaussian process, paired with a sequential design strategy to select the next point via an acquisition function. This work is motivated by an industry-sponsored problem, where focus is on optimizing a dual-agent therapy in a setting featuring minimal toxicity. To illustrate and compare the performance of the standard and personalized methods under this setting, simulation studies are performed under a variety of scenarios.



Keynote Address II: Marc Chadeau-Hyam

Chair: Jelle Goeman
Room: Podium

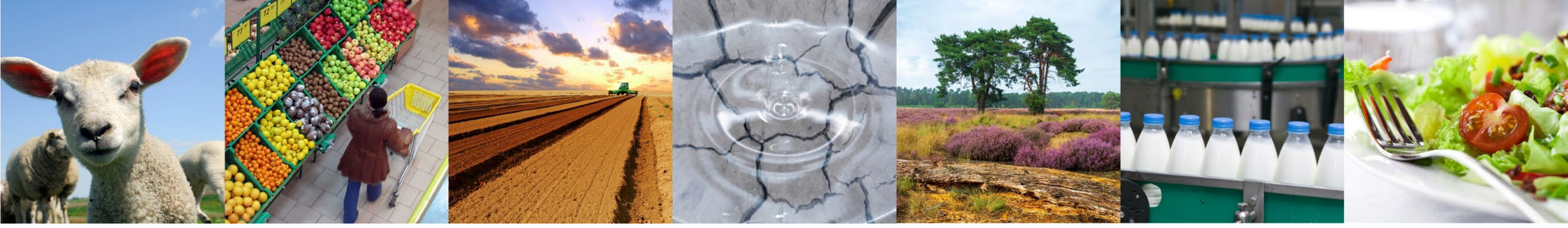
Exposome Analytics: methodological developments and needs

Marc Chadeau-Hyam

School of Public Health, Imperial College London

The Exposome concept has been developed as a necessary complement to the genome to better understand the determinants of health and of the risk of chronic diseases. The external exposome combines a large range of external stressors (i.e. non-genetic) factors potentially impacting human health from conception onwards. These external exposures (i) are heterogeneous in nature, scale, and variability, (ii) feature complex correlation patterns and (iii) may operate as mixtures. The internal exposome can be defined as the way these exposures are embodied and its exploration relies on the screening and integration of high-resolution molecular data. While methods for omics data analyses are established, their application in an exposome context is raising specific methodological challenges including the analysis of complex and correlated exposures. Furthermore, the isolated exploration of an omic profile offers the possibility to capture stressor-induced biological/biochemical alterations, potentially impacting individual risk profiles, but this may only yield a fractional picture of the complex molecular events involved, therefore limiting our understanding of the effective mechanisms mediating the effect of the exposome.

Taking examples from real-life exposome projects we will illustrate the use of statistical and machine learning techniques to accommodate co-occurring exposures contributing to population stratification, explore the links between these and cardiometabolic outcomes, and investigate the (multi)-omic response to these sets of exposures.



Notes