School of Mathematics, Statistics and Applied Mathematics, NUIG

and

Biostatistics Unit, HRB Clinical Research Facility Galway

CASI 2018

38th Conference on Applied Statistics in Ireland

Wednesday 16th May to Friday 18th May 2018

Galway Bay Hotel, Salthill, Co. Galway

www.istat.ie
Contents

Welcome to Galway 1

Invited Speakers 3

Paul Rosenbaum 3
Carl Scarrott 4
Garrett Grolemund 5
Cóilín Minto 6
Trevor Bailey 7
Maria Durbán Reguera 8

Lunch: Wednesday 12.40 - 14.00

Session 1: Wednesday 14.00 - 15.40
Chair: John Ferguson 10

P.R. Rosenbaum: Addressing bias from unmeasured dispositions in observational studies 10
J. Madden, L. Browne, X. Li, P. Kearney and T. Fitzgerald: Morning surge in blood pressure using a random-effects multiple-component cosinor model 11
J. Leahy and C. Walsh: MAIC-ing the most of trials?: A Bayesian exploration Of matching adjusted indirect comparison 12
A. Simpkin, M. Durbán, D.A. Lawlor, C. MacDonald-Wallis, M.T. May, C. Metcalfe and K. Tilling: Derivative estimation for longitudinal data analysis: Examining features of blood pressure measured repeatedly during pregnancy 14

Break: Wednesday 15.40 - 16.00
Session 2: Wednesday 16.00 - 17.00
Chair: Simon Wilson


G. Kelly: Goodness-of-fit statistics and the birthday effect

Poster Session: Wednesday 18.00 - 19:30

A. Al-Ghamdi and B. Houlding: An integrated framework for estimating the extent of biodiversity

A. Alvarez-Iglesias, M. McGillycuddy and J. Ferguson: A risk prediction model for a matched case-control study

A. Boland: An R case study: From research to production. Putting a statistical algorithm in a production environment

C.M. Carbery, R. Woods and A.H. Marshall: Understanding complex interactions within a large-scale manufacturing system through applications of Bayesian networks

M. Cauchi: Conquering gut disorders using data science


B.M. Davis, M. Salinas-Navarro, M.F. Cordeiro, L. Moons and L. De Groef: Ripley’s K-statistics to resolve microglia territories and spatial dynamics

C. Donnelly, L.M. McFetridge and A.H. Marshall: A joint model for the analysis of longitudinal and survival data utilising the Coxian phase-type distribution

L. Fallah and J. Hinde: Discrete survival models including heterogeneity

J. Hinde, I. Lara, R. Moral, T. Oliveira, C. Demétrio, S. Zocchi and W. Godoy: Ireland – Brazil partnerships


J. Madden, L. Zgaga, F. Leacy and K. Bennett: A comparison of methodological approaches on the association between De Novo vitamin D supplement use post-diagnosis and breast cancer mortality

M.I. Mashinchi, D. Roshan, F.J. Sullivan and D. Rebholz-Schuhmann: Business intelligence dashboard solutions for improving decision making process: A focus on prostate cancer
M. McCague: Propensity score matching methods for observational studies  
M.G. McCurdy, A.H. Marshall and K.J. Cairns: Investigating the computational trade-off for computing matrix exponentials using \texttt{kexpmv}  
H.J. Mitchell, B.M. Davis, L. Guo, M.F. Cordeiro: Spatial point pattern analysis of rodent retinal ganglion cells  
S. Moghaddam, K. Murphy, L. Murphy, L. O’Gorman, T. Lynch, R. Power, K.J. O’Malley, T.B. Murphy and R.W. Watson: Predicting the stage of the prostate cancer for personalized treatment in an Irish cohort  
S. Unnikrishnan, J. Donovan, R. Macpherson and D. Tormey: A comparative statistical study of micrograph classification using repeatability and reproducibility analysis

Welcome Dinner: Wednesday 20.00 @ Oslo, Salthill

Session 3: Thursday 9.00 - 10.40  
Chair: Gabrielle Kelly  
C.J. Scarrott: Statistical modelling of air pollution and the challenges of communication  
J-L. Chapman: Changepoints to improve forecasts  
J. McDonnell, T. McKenna, K. Yurkonis, D. Hennessy and C. Brophy: Assessing spatial interactions among species in a grassland experiment  

Break: Thursday 10.40 - 11.00

Inside Insight Session: Thursday 11.00 - 12.40  
Chair: Nial Friel  
G. Grolemund: Teaching data science  
M. Fop, P-A. Mattei, T.B. Murphy and C. Bouveyron: Unobserved classes and extra variables in high-dimensional discriminant analysis  
J. Abedin and J. Newell: Text mining aided literature review: A case study in sports injuries  
K. Murphy and T.B. Murphy: Gaussian parsimonious clustering models with covariates

Lunch: Thursday 12.40 - 14.00
Session 5: Thursday 14.00 - 15.20
Chair: Cathal Walsh

R. Moral, J. Hinde, W. Bonat and C. Demétrio: Modelling bounded data in plant ecology


K. Hayes, T. Barr, E. Crotty, J. Holmes and D. Harrison: Statistical algorithms and issues arising during the development of a hand force-plate sensor system for false start detection in competitive sprinting

D. Peng, K. Burke and G. MacKenzie: Interval censored MPR modelling

Break: Thursday 15.20 - 15.40

Session 6: Thursday 15.40 - 17.00
Chair: Norma Bargary

C. Minto: Comparing catches from multiple fishing gears

T. Bailey: Dicing with Dengue

ISA AGM: Thursday 17.00 - 18.30

Conference Dinner: Thursday 20.00

CÚRAM Connections Session: Friday 9.00 - 11.00
Chair: John Hinde

M. Durbán Reguera: Smooth mixed models: Combining the best of both worlds

J. Chauvet, X. Bry and C. Trottier: Regularisation of generalised linear mixed models with autoregressive random effects

L. Boyle, L. McFetridge and Ö. Asar: The impact of time-varying outliers on mixed effects models: A simulation study motivated by renal data


O. Lyashevska, C. Minto and D. Brophy: Dynamic factor analysis to investigate coherency in growth patterns across various fish species in the Celtic Sea

Break: Friday 11.00 - 11.40
Session 8: Friday 11.40 - 13.00
Chair: Kathleen O’Sullivan

A. McCourt and B. Houlding: Modelling uncertainty and vagueness within recommender systems via nonparametric predictive inference

F-M. Jaouimaa, E. Park, I.D. Ha and K. Burke: Penalised multi-parameter regression survival modelling


J. Rizk, K. Burke and C. Walsh: Identifiability in Coxian phase type distributions - An alternative formulation

Conference Close and Lunch: Friday 13.00 - 14.00
Welcome to Galway!

The Statistics Group in the School of Mathematics, Statistics and Applied Mathematics, NUI Galway, in conjunction with the Biostatistics Unit in the HRB Clinical Research Facility Galway, are delighted to welcome you all to the 38th Conference on Applied Statistics in Ireland (CASI 2018).

This proceedings volume contains the papers and posters presented at CASI 2018 held in Galway, 16th to 18th May 2018. The scientific program is comprised of invited lectures, contributed talks and a pre-workshop short course. Special emphasis is given to student contributions with awards for the best poster presentation.

The conference aims to bring together statisticians, data scientists, researchers and all who are interested in the development and application of statistics. It provides a forum for the exchange of ideas over many different areas of applied statistics, and more recently data science, to stimulate new directions and collaborative work.

We would like to welcome our invited speakers Maria L. Durbán Reguera (University Carlos III University de Madrid), Paul Rosenbaum (University of Pennsylvania, USA), Trevor Bailey (University of Exeter, UK), Garrett Grolemund (Data Scientist at RStudio), Cóilín Minto (Galway-Mayo Institute of Technology) and Carl Scarrott (University of Canterbury, New Zealand) and thank them for accepting our invitation and preparing such interesting talks. We would also like to thank all of the authors who contributed to this volume for their care in preparing their manuscripts. We are very grateful for the generous financial support provided by Science Foundation Ireland, Insight Centre for Data Analytics NUI Galway, CÚRAM NUI Galway, CRC Press, Statistical Solutions, Zurich, The Central Bank of Ireland and Axis Consulting.

The spirit of CASI has always concentrated on statistical research that is motivated by real-life data and applications. It also has a long tradition of social interaction and networking, which is an important catalyst in preserving the coherence of the Irish statistical community. We hope that you will find much to interest and stimulate you at CASI 2018 and that you have an enjoyable, productive and fruitful stay in Galway.

John Newell    Neil O’Leary    Maeve McGillycuddy
John Ferguson  Carl Scarrott  Lida Fallah
Alberto Alvarez-Iglesias Jaynal Abedin  Kishor Das
Andrew Simpkin  Cara Dooley  Davood Roshan
John Hinde
Invited Speakers
Paul R. Rosenbaum is the Robert G. Putzel Professor of Statistics at the Wharton School and a Senior Fellow of the Leonard Davis Institute of Health Economics, University of Pennsylvania.

His research interests include design and analysis of experiments, design and analysis of observational studies and health outcomes research.

Paul’s latest book released in 2017, ‘Observation and Experiment: An Introduction to Causal Inference’, is described as ‘an introduction to causal inference from one of the field’s leading scholars. Using minimal mathematics and statistics, Paul Rosenbaum explains key concepts and methods through scientific examples that make complex ideas concrete and abstract principles accessible.’
Carlo Scarrott has a B.Sc. (Hons) in Mathematics and Ph.D. in Statistics from Lancaster University in the UK.

He joined the School of Mathematics and Statistics at the University of Canterbury in New Zealand in 2004 and became Associate Professor in Statistics in 2016.

His research focuses on statistical modelling and visualisation of complex data, with two main themes of extreme value modelling and biostatistics. All the better when the two come together. Application areas include environmental and industrial problems, along with clinical and non-clinical studies in health.

He is also passionate about teaching and statistical computing, with the latter leading to the development of the `evmix` package for R which includes functionality for extreme value threshold estimation, mixture modelling and boundary corrected kernel density estimation.
Garrett Grolemund

Garrett is the author of “Hands-On Programming with R” and co-author of “R for Data Science” from O’Reilly Media.

He is a Data Scientist at RStudio and holds a Ph.D. in Statistics, but specializes in teaching. He’s taught people how to use R at over 50 government agencies, small businesses, and multi-billion dollar global companies; and he’s designed RStudio’s training materials for R, Shiny, dplyr and is a frequent contributor to the RStudio blog. He wrote the popular lubridate package for R.
Cóilín Minto

Cóilín is a research fellow in quantitative ecology at the Marine and Freshwater Research Centre at the Galway-Mayo Institute of Technology, Galway.

His research focuses on the development and application of tailored statistical methodologies to understand dynamics at individual, population and community levels. Statistical research interests include: hierarchical analysis, longitudinal analysis, mixture modelling and dynamic time series analysis, particularly state space modelling. Current research includes: mixture modelling of life histories, discontinuous time series filters, design and analysis of fishing gear experimental trials, and management strategy evaluation for wild population monitoring.

Cóilín contributes to European stock assessments and is a founding developer of the RAM Legacy Stock Assessment Database.
Professor Bailey joined the University of Exeter in January 1986 from the University of New South Wales, Australia. He retired from Exeter in August 2017.

His research interests focused on applied statistical modelling (particularly spatial-temporal epidemiology, environmental impacts on public health, and multivariate spatial methods). Much of his research was collaborative involving academics and professionals from a variety of fields such as medicine, public health, geography, computer science, behavioural science and education. He also worked with commercial organisations and government agencies.

He received regular invitations to speak at specialist meetings both in the UK and overseas and frequently gave short courses or seminar series on various statistical topics at universities in Europe, South America and Australasia.
Maria Durbán Reguera

Mary Durbán Reguera has a B.Sc. in Mathematics at University of Granada, M.Sc. Statistics in Mathematics at Cambridge University and Ph.D. in Mathematics at Heriot-Watt University. She is a member Carlos III University since 2000, and she became Associate Professor at the Department of Statistics in 2005.

Her research is focused on methods for multidimensional regression smoothing, efficient algorithms for the analysis of complex data, and their application in areas such as Medicine, Environmental Sciences, Insurance and Demography. She is author of two books, publishes in international scientific journal, regularly teaches specialized courses at national and international institutions.

Since 2005, she combines her research with other management activities: she has been Deputy Director for Promotion at the Engineering School (2005-2008), Deputy Director of the Department of Statistics (2013-2015); since 2007 she is Director of the “Fernando Martorell Abrill” Hall of Residence, and currently she is Vice-President Deputy for Promotion.
Session 1: Wednesday 14.00 - 15.40
Chair: John Ferguson
Addressing bias from unmeasured dispositions in observational studies

Paul R. Rosenbaum

1Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, USA

There are two treatments, each of which may be applied or withheld, yielding a $2 \times 2$ factorial arrangement with three degrees of freedom between groups. The differential effect of the two treatments is the effect of applying one treatment in lieu of the other. In randomized experiments, the differential effect is of no more or less interest than other treatment contrasts. Differential effects play a special role in certain observational studies in which treatments are not assigned to subjects at random, where differing outcomes may reflect biased assignments rather than effects caused by the treatments.

Differential effects are immune to certain types of unobserved bias, called generic biases, which are associated with both treatments in a similar way. This is exemplified using three familiar models, a Rasch model, a symmetric multivariate logit model and a preference tree model. Differential effects are not immune to differential biases, whose possible consequences are examined by sensitivity analysis. Under certain conditions, the differential comparison of two treatments balances other treatments, including unmeasured treatments, that are governed by the same unmeasured disposition. Three scientific examples are presented.
Morning surge in blood pressure using a random-effects multiple-component cosinor model

Jamie Madden1,2, Leonard Browne2, Xia Li2, Patricia Kearney2 and Tony Fitzgerald1,3

1Division of Population & Health Sciences, RCSI, Dublin, Ireland
2Department of Epidemiology & Public Health, University College Cork, Ireland
3Department of Statistics, University College Cork, Ireland

Blood pressure (BP) fluctuates throughout the day and the pattern it follows represents one of the most important circadian rhythms in the human body yet there has been limited research into methods to model this pattern and to subsequently extract clinically relevant parameters. This study attempts to develop on the most common method to model 24h BP, the single cosinor approach, and use it to extract an estimate of morning BP.

Introduction
It is well known that BP does not remain stationary but fluctuates throughout the day and follows a circadian rhythm. Independent of mean BP, features of this circadian pattern such as dipping status or morning surge have been shown to be risk factors of cardiovascular events. Morning surge refers to the phenomenon that occurs in individuals during the first few hours after waking up in the morning when there is an exaggerated spike or surge in BP. Modelling the circadian trajectory of BP and extracting clinical features remains a challenge (Edwards and Simpson, 2014). Here, we aim (i) to develop on a model that can capture the circadian pattern and (ii) use the model to extract a measure of morning BP.

Methods
We demonstrate how the most commonly used method to model 24h BP, the single-cosinor approach, can be extended to a multiple-component cosinor random-effects model. We outline how this model can be used to obtain a measure of morning BP surge by obtaining derivatives of the model fit. The model is compared to a functional principal component analysis (FPCA) which determines the main components of variability in the data. The parameters from the cosinor model are compared to the functional principal component scores. The purpose of this was to determine if model fits from a model with fewer parameters which are more interpretable (cosinor) compare favourably with a more flexible data-driven approach (FPCA). Data from the Mitchelstown Study, a population based study of Irish adults (n = 2,047) was utilized where a subsample (n = 886) underwent 24h ABPM.

Results
The single component model offered a very simplistic curve that struggled to capture the shape of the data. We demonstrated that our two-component model provided a significant improvement in fit compared to a single model and a similar fit to a more complex model captured by B-splines using FPCA. The estimate of the average maximum slope was 2.857 mmHg/30min (bootstrap estimates; 95% CI: 2.855-2.858 mmHg/30min). Simulation results allowed us to quantify the between-individual SD in maximum slopes which was 1.02 mmHg/30min.

Conclusion
We have demonstrated that extending the traditional single cosinor to a multiple-component cosinor in a random-effects model can be achieved while offering a substantial improvement in fit to ABPM data. This simple application of derivatives allows us to quantify a measure of morning BP that specifically represents a surge parameter. This may be beneficial in future studies exploring the prognostic significance of morning BP and chronotherapy effects of antihypertensive medication. This is the first study to demonstrate the use of a multiple component cosinor random-effects model to obtain a measure of morning BP surge.

References
MAIC-ing the most of trials?: A Bayesian exploration of matching adjusted indirect comparison

Joy Leahy\textsuperscript{1} and Cathal Walsh\textsuperscript{2}
\textsuperscript{1}School of Computer Science and Statistics, Trinity College Dublin, Ireland
\textsuperscript{2}Professor of Statistics, University of Limerick, Ireland

Incorporating Individual Patient Data (IPD) into a Network Meta Analysis (NMA) is the gold standard of analysis, as it allows a more in depth analysis of the data. However, often times a researcher may have IPD for trials concerning a particular treatment (for example from a sponsor), but none for other trials. In this case we can check that results obtained in the sponsor’s trial hold for other populations by re-weighting the IPD so that the covariate characteristics in the IPD trials match that of the Aggregate Data (AgD) trials. This method is called Matching Adjusted Indirect Comparison (MAIC). We investigate the benefits of using this technique in a Bayesian setting, applied to a time to event outcome, through a simulation study. We apply this to an NMA for multiple myeloma in newly diagnosed patients (ndMM) post-autologous stem cell transplant.

Introduction
MAIC has been growing in popularity in recent years. As it is a relatively new method there is a need for comprehensive simulation studies to explore the properties (Phillippo, et al., 2016). In this work we perform a simulation study in order to:

1. quantify the effect of MAIC in a Bayesian NMA using multiple trials and
2. investigate two options of weighting covariates: both within each IPD trial (MAIC Separate Trials) or across all IPD trials (MAIC pooled trials).

Methods
We run a simulation study with 3 IPD trials comparing treatments A and B (AB-IPD trials), and another AgD trial comparing treatments B and C (BC-AgD trial). As the relative efficacy of treatments changes depending on the covariate value, we examine our results in both the full NMA population and the (target) BC-AgD trial.

![Figure 1: Average proportion with covariate: AB-IPD trials=0.45, AgD-BC trials=0.90](image)

Figure 1: Average proportion with covariate: AB-IPD trials=0.45, AgD-BC trials=0.90
Results
In Figure 1, the overall difference in covariates between the AB-IPD (average over three trials) and BC-AgD trials remains constant. However, as the $x$-axis increases, one IPD trial becomes more similar to the BC-AgD trial (and will have a higher weight in the MAIC Pooled Trials Model), while another trial becomes more different. As the difference between the AB-IPD trials increases there is less of a difference between reweighting by MAIC or not. Overall, MAIC will produce a less accurate estimate of the full patient population. However, if the goal is to check the relative efficacy in a particular patient population then MAIC can be quite beneficial.

We compare the results of the ndMM network as an example. Using a standard NMA based on published medians, we find Lenidomide (Len) to be superior to Thalidomide (Thal) for progression free survival: HR=0.67 (Credible Interval (CrI)=0.50-0.89). Using an MAIC adjustment for the covariates "International Scoring Stage" and "response prior to maintenance", we obtain a HR of 0.71 (CrI=0.54-0.95). Hence, the MAIC produces a HR slightly closer to one, but the MAIC gives us confidence in the result, as it shows that the superiority of Len vs Thal holds across populations. For further results on clinical efficacy in this network please see Schmitz, et al. (2017).

Conclusion
MAIC is beneficial as a sensitivity analysis to confirm results across patient populations. However, results are less accurate with regards to the full population than a standard NMA. Given the increasing use of MAIC, it is important that researchers think carefully about the population of interest before conducting an MAIC.

References

Derivative estimation for longitudinal data analysis: Examining features of blood pressure measured repeatedly during pregnancy

Andrew J. Simpkin1, Maria Durbán Reguera2, Debbie A. Lawlor3, Corrie MacDonald-Wallis4, Margaret T. May5, Chris Metcalfe5 and Kate Tilling3
1Insight Centre for Data Analytics, National University of Ireland Galway, Ireland
2Department of Statistics, Universidad Carlos III de Madrid, Spain
3DMRC Integrative Epidemiology Unit, University of Bristol, UK
4Population Health Sciences, Bristol Medical School, University of Bristol, UK
5Centre for Exercise, Nutrition & Health Sciences, School for Policy Studies, University of Bristol, UK

Estimating velocity and acceleration trajectories allow novel inferences in the field of longitudinal data analysis, such as estimating change regions rather than change points, and testing group effects on nonlinear change in an outcome (i.e. a nonlinear interaction). Here we present derivative estimation for two standard approaches using polynomial mixed models and spline mixed models. We compare their performance with an established method - Principal components Analysis through Conditional Expectation (PACE) through a simulation study. We then apply the methods to repeated blood pressure (BP) measurements in a UK cohort of pregnant women, where the goals of analysis are to

(i) identify and estimate regions of BP change for each individual and
(ii) investigate the association between parity and BP change at the population level.

The penalised spline mixed model had the lowest bias in our simulation study and we identified evidence for BP change regions in over 75% of pregnant women. Using mean velocity difference revealed differences in BP change between women in their first pregnancy compared with those who had at least one previous pregnancy. We recommend the use of penalised spline mixed models for derivative estimation in longitudinal data analysis.
Session 2: Wednesday 16.00 - 17.00
Chair: Simon Wilson
Using probabilistic sensitivity analysis in budget impact models for reimbursement recommendations in Ireland

Felicity Lamrock\(^1,2\), Lesley Tilson\(^1,2\) and Laura McCullagh\(^1,2\)

\(^1\)National Centre for Pharmacoeconomics, St James’s Hospital, Dublin, Ireland
\(^2\)Department of Pharmacology & Therapeutics, Trinity College Dublin, Ireland

National guidelines recommend the use of probabilistic sensitivity analysis in budget impact models submitted to the National Centre for Pharmacoeconomics (NCPE) for new drug assessment. A retrospective analysis of submitted models to the NCPE was performed to address the feasibility and use of probabilistic sensitivity analyses to better predict the five year budget impact if a new drug were reimbursed. Fourteen models were recreated and adapted to include a probabilistic sensitivity analysis. Even with probabilistic sensitivity analysis performed on the budget impact models, predictions of budget impact over the next 5 years need to be improved.

Introduction

The National Centre for Pharmacoeconomics (NCPE) assesses the cost-effectiveness and budget impact of new drugs for which reimbursement by the healthcare payer is sought, the Health Service Executive (HSE) in Ireland (McCullagh, et al., 2016). The NCPE assesses the budget impact of new drugs in accordance to national health technology assessment (HTA) guidelines (HIQA, 2014). The national guidelines specify that a probabilistic sensitivity analysis is to be performed on the budget impact model as part of the assessment to address the uncertainty surrounding the predicted impact on the budget in Ireland. In reality the budget impact models submitted to the NCPE to date, have not encompassed a PSA, but may prove useful for the decision maker. This work investigates the feasibility of using a probabilistic sensitivity analysis within the budget impact models to potentially aid the reimbursement decision maker.

Methods

A retrospective analysis of budget impact models that had been submitted to the NCPE was performed. All drugs that underwent a HTA between January 2010 and December 2017 inclusive were considered. Drugs were excluded if they were not reimbursed, or had not been reimbursed for a full year. Models were also excluded if their functionality or inputs did not permit a probabilistic sensitivity analysis to be performed. The final price at which the drug was reimbursed was input into the budget impact models. Each individual budget impact model was recreated and a macro programmed in Excel Visual Basic for Applications written and inserted into the model to perform a probabilistic sensitivity analysis. The total number of input parameters in each of the budget impact models was recorded and fell under one of the following headings: population, market share, dosing and costs. All analyses were conducted in Microsoft Excel 2010.

Results

A total of 14 HTA’s were included in the analysis, 9 of which fell under the High Tech Drugs Scheme (HTDS), and 5 fell under the General Medical Services (GMS), Long Term Illness (LTI) scheme, and Drugs Payment Scheme (DPS). Most of the drugs under predicted the resources used in Ireland. The parameters that effected the budget impact model predictions the most were the population estimates. The probabilistic sensitivity analysis results showed that there was a high range for the estimated budget predicted, but that the models varied in their ability to predict the number of resources used within a 95% confidence interval.

Conclusion

This research has demonstrated that the estimation of the budget impact of a new drug for which reimbursement is sought is difficult but a probabilistic sensitivity analysis should be considered as part of the formal HTA process by the NCPE to aid the HSE in their decision. Further research lies in exploring the practical considerations of applicants to include this information, and for research regarding an algorithm to better assess what factors effect whether a drug is over or under the predicted amount submitted by the company.

References


Trends in modern contraceptive prevalence, unmet need for and demand satisfied with modern methods in the focus countries of the Family Planning 2020 Initiative

Niamh Cahill1, Michelle Weinberger2, Emily Sonneveldt3, John Stover3, Chuchu Wei4, Win Brown5 and Leontine Alkema6

1School of Mathematics and Statistics, University College Dublin, Ireland
2Avenir Health, Washington D.C., USA
3Avenir Health, Glastonbury, USA
4Department of Biostatistics and Epidemiology, University of Massachusetts-Amherst, Amherst, USA
5The Bill & Melinda Gates Foundation, Seattle, USA

The London Summit on Family Planning in 2012 inspired FP2020 and the goal of ‘120X20’, that is, having an additional 120 million women and adolescent girls become users of modern contraceptives in the world’s poorest countries by 2020. We use a recently updated version of the Family Planning Estimation Tool (FPET) to construct estimates and projections of modern contraceptive prevalence, unmet need for and demand satisfied with modern methods of contraception among women of reproductive age who are married or in a union in the FP2020 countries. We assessed current levels of these indicators as well as changes between 2012 and 2017 and we used a counterfactual analysis to assess if recent levels of mCPR exceeded pre-FP2020 expectations. We estimate that between 2012 and 2017 the number of MWRA who use modern methods increased by 28.8 million (5.80- 52.5). Based on a counter-factual analysis, we find that 63% of the countries that have made a commitment to FP2020 exceeded pre-FP2020 expectations for modern contraceptive use. Country success stories include rapid increases in Kenya and Mozambique.

Introduction

The Family Planning 2020 (FP2020) initiative is a global movement that supports the rights of women and girls to decide freely and for themselves whether, when and how many children they want to have. The London Summit on Family Planning in 2012 inspired FP2020 initiative and the goal of ‘120X20’, that is, having an additional 120 million women and adolescent girls become users of modern contraceptives in 69 of the world’s poorest countries by the year 2020 [Brown et al., 2014]. Working towards achieving ‘120X20’ is critical for ultimately obtaining the Sustainable Development Goals (SDGs) of universal access and satisfying demand for reproductive health. Thus, a performance assessment is required to determine countries’ progress.

Data and Methods

Contraceptive prevalence is measured as the percentage of women who report themselves or their partners as currently using at least one contraceptive method of any type (modern or traditional). Unmet need for family planning is defined as the percentage of women who want to stop or delay childbearing but who are not currently using any method of contraception to prevent pregnancy. With the exception of India, the FP2020 country data uses the United Nations Population Division survey database for contraceptive prevalence and unmet need for family planning as a base (United Nations, 2017). In India, state level data for family planning indicators were used (New et al., 2017). In some countries, survey data for modern contraceptive use is supplemented with family planning service data (Weinberger, et al., 2017).

For every country, the family planning estimation tool (FPET; New and Alkema, 2015) models contraceptive prevalence with an expected trend that assumes contraceptive prevalence will begin with a gradual increase, it will subsequently become more rapid and then it will begin to slow down when high levels of prevalence are reached. The parameters that control the trend are estimated hierarchically, such that estimates are based on the data available in the country of interest, and also the sub-regional, regional, and global experience. Distortions are added to capture how rates of change in the observed data (i.e., faster/slower rates of change in contraceptive prevalence) deviate from the rates of change indicated by the expected trend. Projections are informed by recent changes that have occurred in contraceptive prevalence (i.e., the difference between the two most recent surveys) as well as past experience. Estimates of unmet need are obtained by modeling the relationship between contraceptive prevalence and unmet need. Similar to the model for contraceptive
prevalence, a hierarchical approach is used to estimate parameters. However for unmet need, time dependent distortions are added to capture country-specific changes in the level of the indicator (Alkema et al., 2013).

FPET was used to construct estimates and projections of the modern contraceptive prevalence rate (mCPR), unmet need for and demand satisfied with modern methods of contraception among women of reproductive age who are married or in a union (MWRA) in the focus countries of the FP2020 initiative (see Appendix). We assessed current levels of family planning indicators and changes between 2012 and 2017. Then a counterfactual analysis was used to assess if recent levels of mCPR exceeded pre-FP2020 expectations.

Results
In 2017, mCPR among MWRA in FP2020 focus countries is 45.7% (95% uncertainty interval 42.4-49.1), unmet need for modern methods is 21.6% (19.7-23.9), and the demand satisfied with modern methods is 67.9% (64.4-71.0). Between 2012 and 2017 the number of MWRA who use modern methods increased by 28.8 million (5.74-52.4). At the regional level, Asia has seen mCPR among MWRA grow from 51.0% (48.5-53.4) to 51.8% (47.3-56.5) between 2012 and 2017, which is slow growth, particularly when compared to a change from 23.9% (22.9-25.0) to 28.5% (26.8-30.2) across Africa.

At the country level, based on a counter-factual analysis, we find that 57% of FP2020 countries obtained levels of mCPR greater than or equal to pre-FP2020 expectations including 20 (63%) of the countries that have made a commitment to the initiative (Figure 1). Among the countries that we consider to have made the most progress relative to what was expected are: Chad; Sierra Leone; Mozambique and Kenya, with the latter three being commitment countries. These four countries attained levels of mCPR post-2012 that had a less than 10% chance of being observed pre-FP2020. In Kenya, the estimate for mCPR in 2015, the most recent observation year was 14.3 (0.0-28.7) percentage points higher than the pre-FP2020 expectations for this year. This equates to an additional 0.92 million (0.00-1.85) married or in-union women using modern methods of contraception in 2015 compared to what was expected. Conversely, Burundi, which is a commitment country, and Gambia, a non-commitment country, observed lower levels of mCPR relative to expectations with a greater than 90% chance that the mCPR would be higher than current estimates.

![Figure 1: Attainment probabilities for modern contraceptive prevalence (mCPR) estimates in the most recent observation year for countries that made a commitment or not to FP2020. Results are for countries that had data available after 2012.](image-url)
Conclusion

While the estimate of additional users up to 2017 for women who are married or in a union would suggest that the ‘120X20’ goal for all women is overly ambitious, the aggregate outcomes mask the diversity in progress at the country level. We identified countries with accelerated progress, that provide inspiration and guidance on how to increase the use of family planning and inform future efforts, especially in countries where progress has been poor.

References


Appendix: FP2020 Focus Countries

Eastern and Southern Africa
- Burundi
- Comoros
- Djibouti
- Eritrea
- Ethiopia
- Kenya
- Lesotho
- Madagascar
- Malawi
- Mozambique

Central Africa
- Cameroon
- Central African Republic
- Chad
- Congo
- DR Congo
- Sao Tome and Principe

South Asia
- Afghanistan
- Bangladesh
- Bhutan
- India
- Nepal
- Pakistan
- Sri Lanka

Western Africa
- Benin
- Burkina Faso
- Côte d’Ivoire
- Gambia
- Ghana
- Guinea
- Guinea-Bissau
- Liberia
- Mali
- Mauritania
- Niger
- Nigeria
- Senegal
- Sierra Leone
- Togo

Middle East and Northern Africa
- Egypt
- Iraq
- South Sudan
- State of Palestine
- Sudan
- Western Sahara \(^{a}\)
- Yemen

\(^{a}\)Western Sahara which has no data, has been left out of the analysis, as it is not currently included in the reporting for the initiative.
Eastern and Central Asia
- Kyrgyzstan
- Mongolia
- DPR Korea
- Tajikistan
- Uzbekistan

Latin America and Caribbean
- Bolivia
- Haiti
- Honduras
- Nicaragua

South East Asia and Oceania
- Cambodia
- Indonesia
- Lao PDR
- Myanmar
- Papua New Guinea
- Philippines
- Solomon Islands
- Timor-Leste
- Vietnam
Goodness-of-fit statistics and the birthday effect

Gabrielle Kelly
School of Mathematics and Statistics, University College Dublin, Ireland

The Kolmogorov-Smirnov (K-S), Cramér-Von Mises (CVM) and Anderson-Darling (AD) are popular tests for testing whether a sample of \( n \) points come from some theoretical distribution \( F \). Zhang (2002) provided more powerful versions of these tests. Kuiper (1960) and Watson (1961) extended these statistics so that for distributions on a circle, they are independent of the origin of the polar coordinates. Here the test statistics of Zhang are extended in a like manner. The statistics are used to investigate if there is an association between birthday and deathday in famous people, also known as the Phillips’ phenomenon. Null distributions are found by Monte Carlo simulation.

Introduction

Let \( X_1, X_2, \ldots, X_n \) be a random sample with distribution function \( F(x) \). We wish to test the null hypothesis \( H : F(x) = F_0(x) \) for all \( x \in (-\infty, \infty) \) against the general alternative \( H : F(x) \neq F_0(x) \) for some \( x \in (-\infty, \infty) \). The analogues of the K-S statistic and CVM statistics for distributions on a circle are

\[
V_n = \sup (F_n(x) - F_0(x)) - \inf (F_0(x) - F_n(x))
\]

i.e. the Kuiper statistic (1960) and from Watson (1961)

\[
U_n = n \int_{-\infty}^{\infty} [F_n(x) - F_0(x) - \int_{-\infty}^{\infty} (F_n(x) - F_0(y)) dF_0(y)]^2 dF_0(x)
\]

where \( F_n(x) \) is the sample distribution function, with continuity correction. Alternative versions of these statistics are derived and applied to the deathday-birthday effect (Phillips, 1972). Some studies have found an individual’s likelihood of death ‘dips’ before the birthday consistent with postponement of death until after the birthday while others have found a death rise consistent with a ‘birthday blues’ hypothesis. Questions have been raised regarding their statistical methodologies and therefore we have examined 8 groups of famous people in whom the phenomenon might be more pronounced: British Prime Ministers, U.S. presidents, Wimbledon men’s and ladies’ singles winners, Academy award best actor, best actress, best director and winners of the Nobel prize.

Methods

To test \( F(x) = F_0(x) \) two statistics are Pearson’s \( \chi^2(x) = \frac{n(F_n(x) - F_0(x))^2}{F_0(x)(1 - F_0(x))} \) and the likelihood ratio statistic

\[
G(x) = 2n \left[ F_n(x) \log \left( \frac{F_n(x)}{F_0(x)} \right) + (1 - F_n(x)) \log \left( \frac{1 - F_n(x)}{1 - F_0(x)} \right) \right].
\]

The first of these when used in the form \( \sup_{x \in (-\infty, \infty)} \{ \chi^2(x)w(x) \} \) for a suitable weight function \( w(x) \) yields the K-S statistic and when used in the form \( \int_{-\infty}^{\infty} \chi^2(x) dw(x) \) yields the CVM and A-D statistics (Zhang, 2002). Replacing the \( \chi^2(x) \) statistic with the LR statistic and adapting for distributions on a circle yields the statistics

\[
LR_1 = \max_{1 \leq i \leq n} G_{X(i)}^2 - \min_{1 \leq i \leq n} G_{X(i)}^2 \quad \text{and}
\]

\[
LR_2 = \sum_{i=1}^{n} [\log(F_0(X(i))^{-1} - 1) - b_{i-1} + b_i - \sum \{\log(F_0(X(i))^{-1} - 1) - b_{i-1} + b_i\}]^2
\]

where \( b_i = i \log(i/n) + (n - i) \log(1 - i/n) \) and \( X(i) \) are the order statistics.

In the birthday data, for each of the eight groups the difference in days between the deathday and birthday were calculated and these differences mapped to a circle. The empirical distribution function was then compared
with the uniform using the $V_n$, $U_n$, $LR_1$ and $LR_2$ statistics above. The Nobel prize winners were divided into their categories and the hypothesis examined for each category. $p$-values were found by simulating samples from $F_0$ – a uniform

**Results**

U.S. presidents, Wimbledon men and Nobel physicists show a ‘dip’ in deaths before and after the birthday. For Academy award winning directors, actors and actresses and Wimbledon ladies there was rise in deaths before and after birthday.

**Discussion**

The power of the statistics presented here is the subject of ongoing research.

**References**


Poster Session: Wednesday 18.00 - 19:30
An integrated framework for estimating the extent of biodiversity

Asmaa Al-Ghamdi\(^1\) and Brett Houlding\(^1\)

\(^1\)Department of Statistics, Trinity College Dublin, Ireland

Estimating the extent of biodiversity is one of the most important human endeavours. Human motivations for maintaining biodiversity vary from high moral levels to those of basic human needs. These include ethical responsibility, satisfying our curiosity, ensuring global safety through species adaptation and ecosystem balance, and maintaining the valuable variety of resources for nourishment and raw substances, Dash (2007).

In this purpose, the most practical and commonly used indicator of revealing the extent of biodiversity on Earth is by speculating species richness and the uncertainty around it. Many serious attempts have been done to estimate species richness but within a specific taxa or within a specific location. However, there does not seem to be integration or consistency in conclusions. Collectively, estimates have ranged from 2 to over 30 million species on Earth, Costello, \textit{et al.} (2012). This significant gap provokes doubt in the adequacy and applicability of previous approaches. A common drawback of all these attempts is that they have relied on one of two methods, either through sampling efforts, Bunge and Fitzpatrick (1993), or the discovery curve, Wilson and Costello (2005), neither of which take into account associated factors that may be linked with species discovery.

**Proposed Approach**

Here we formulate a model that directly describes the process of species sampling and discovery in parallel through a latent taxonomic effort process. By linking the two processes the model does not only involve sampling effort and discovery rates, but also a proxy for latent taxonomic efforts through the number of authors, Figure 1. This model is designed to provide a system of inter-relationships influencing the total number of species \(S\). The aim is to formulate a posterior distribution on \(S\) given the actual information about the number of distinct author surnames and discovery times, along with the assumptions imposed by the model. From this posterior distribution a reasonable sample size will be generated to allow estimation of \(S\) and its uncertainty via a credible interval. To achieve this aim all the components of the model should be updated in an efficient sequence using several simulation techniques.

![Bayesian Network](image)

\(S\)
---
\(\theta\)
---
\(N_k\)
---
\(k = 1:S\)
---
\(e_j\)
---
\(a_j\)
---
\(j = 1:T\)
---
\(t_i\)
---
\(m_i\)
---
\(i = 1:R\)

Figure 1: A Bayesian Network visually displaying the whole structure of the proposed model and the probabilistic causal relationships among its components.
Figure 1 reflects a joint probability distribution for the unknown components, in circles, given the knowledge of $t_i$ and $a_j$, where $t_i$ is the discovery dates that reflect the rate of discovery, $a_j$ is the number of authors which is a proxy for the taxonomic efforts $c_j$. Species abundance is $N_k$, $m_i$ and $n_j$ are the number of drawn individuals until meeting a new discovery and the sample size, respectively, that represent the sampling efforts. The total number of species is $S$, while the known number of them is $R$. $\theta$, $\alpha$ and $\sigma$ are the parameters of the marginal distributions. Note that this model is designed under the constraints:

$$\max \left\{ M_{i-1}, \sum_{j=1}^{t_i-1} n_j \right\} \leq M_i \leq \sum_{j=1}^{t_i} n_j \hspace{1em} m_i = M_i - M_{i-1}.$$

**Discussion**

The current study introduces a novel approach of estimating the number of unknown species that provides an integrated framework based on Bayesian uncertainty and which satisfies: (1) Involves the discovery variable and other related covariates. It does not only comprise the sampling efforts and the discovery rate factors but also the latent taxonomic efforts and the species abundances. (2) Allows measuring the uncertainty to be associated with that estimate, giving a better reflection of the accuracy. (3) Avoids the kind of limitations forced by alternative and unreasonable assumptions concerning correlations in biodiversity along taxonomic size or geographic habitat etc. (4) Employs a variety of simulation techniques such as MCMC and Integrated Nested Laplace Approximations.

**References**


A risk prediction model for a matched case-control study

Alberto Alvarez-Iglesias\(^1\), Maeve McGillycuddy\(^1\) and John Ferguson\(^1\)
\(^1\)Biostatistics Unit, HRB Clinical Research Facility, National University of Ireland Galway, Ireland

In matched case control studies, case individuals that have a binary outcome (often a disease) are matched to one or more control individuals without the outcome. Matching is generally performed on a set of selected variables that are known to have an effect on the probability of the outcome variable. In this poster, binary logistic regression models are used to model the incidence of disease in a large observational matched case-control study. Details on the use of these models in this setting will be presented, along with an interactive risk calculator visualization tool. A discussion on the role of the intercept and a method to obtain adequate predicted incidence rates will also be provided.

Introduction

Prospective cohort studies, in their simplest form, make use of the information collected about a set of covariates (or risk factors) in a random sample of individuals. After a fixed time period, a binary outcome, such as disease/non-disease, is measured, and the probability of disease, conditioned on the risk factors, can be modeled using a logistic model.

It is well known that binary logistic models can also be used when conditioning on the outcome measurement, like in case-control studies, by simply applying Bayes formula (Hosmer, and Lemeshow, 2000). Odds ratios can be estimated via maximum likelihood as if they were obtained from a prospective cohort study. However, to obtain appropriate estimates of disease prevalence, a correction has to be made for the intercept of the model. In particular, a version of the following equation, a weighted average of the stratum-specific prevalence of the disease, should hold in the population. Its sample analogue (below) can be solved to obtain the corrected intercept \(\beta_0\):

\[
\pi \left( \frac{1}{n_{\text{case}}} \sum_{i \in \text{case}} \frac{\exp(\beta_0 + \ldots)}{1 + \exp(\beta_0 + \ldots)} \right) + (1 - \pi) \left( \frac{1}{n_{\text{control}}} \sum_{i \in \text{control}} \frac{\exp(\beta_0 + \ldots)}{1 + \exp(\beta_0 + \ldots)} \right) = \pi
\]

where \(\pi\) is the prevalence (or incidence) and \(n_{\text{case}}\) and \(n_{\text{control}}\) are the number of sample cases and controls.

Dataset and Methods

A large international case-control study (O’Donnell, 2010) was used to develop a risk prediction model. Cases were subjects with acute first stroke and controls were participants with no history of stroke. Matching was done in a 1:1 ratio for age, sex and region. A logistic model for the occurrence of stroke was fitted using 10 potentially modifiable risk factors. Incidence rates for age, gender and country were obtained from the global burden of disease ‘Epi Visualization’ tool (IHME, 2017). Corrections were performed to obtain adequate conditional incidence predictions. Figure 1 shows an snapshot of an interactive Shiny risk calculator.

Figure 1: Snapshot of an interactive risk calculator for stroke.
Discussion
Prediction accuracy was assessed using cross-validated AUC values and the model assumes constant incidence rates over time. Other settings could be explored where incidence rates are modeled over time and extrapolated beyond the available data.

References


An R case study: From research to production.  
Putting a statistical algorithm in a production environment

Aidan Boland  
1Clavis Insight, Dublin, Ireland

In industry, bringing an algorithm from it’s research phase into a usable product can be an awkward hurdle to cross. During research the algorithm may be coded in one language, but in practice it may need to be used within a different language environment. It’s often not worthwhile re-writing the function. This case study discusses the methods used to bring a statistical algorithm written in R from it’s research phase, into daily use within a production environment. The algorithm in question is a supervised classification method and is used to automatically categorise e-commerce products from online stores.

R’s ecosystem makes it very simple to provide access to code without the end user needing any knowledge of R. These same methods can be used in academia to open up quick and easy access to new statistical methods.

Introduction
R is one of the most popular languages in statistics. However, in industry R is not often used within a production environment. Bringing research from locally run R scripts to production code can be a long and arduous process, especially if the code must be ported into a different language to match the companies standard.

Alternatively, if the R code is already robust, it can be made accessible through Graphical User Interfaces (GUI) and Application Programming Interfaces (API). The ecosystem of R libraries makes its extremely simple and straightforward to create these accessible methods.

Case Study
Clavis Insight are a leading firm in e-commerce analytics. Large amounts of data is processed from online stores across the globe every day; this allows clients to monitor their performance in the online marketplace. One issue faced by the company is to categorise products into groups which reflect a clients individual specification.

An algorithm based on multinomial logistic regression was created to automatically classify new products into their relevant categories. The research was completed and coded using R. Bringing the algorithm from research into production was done in 2 steps.

First, during a testing period a GUI was created which allowed users to load and classify products. The user could then download the categorised data.

After testing, an API was created which allowed the algorithm to be run directly from within the production code. This allowed the classification of new data without any manual intervention.

Methods

GUI (shiny)
A GUI gives users simple point and click access to algorithms. The shiny library in R makes it very easy to create a GUI. Users can easily change parameters and results can be displayed graphically on screen.

API (plumber)
An API is an endpoint which can be called from almost any computer language. Data and parameters can be passed to the API, and results can be returned. Creating an API for code is the most versatile way of allowing access to methods. The plumber library in R is a simple way to create API’s for R functions.
**Discussion**
There are many easy ways in R to provide users with access to algorithms without the user needing any knowledge of R. These methods provide a quick solution for bringing an algorithm from research into production use.

A related problem in academia is providing simple access to new and novel algorithms. While publishing raw code is one method of sharing work, this assumes that the end user will be able to run the code on their own machine. Making code accessible through a GUI allows users simple access to the methods. While creating an API allows users to build the methods into their own code without needing to fully understand the original code.

**References**
Trestle Technology, LLC. (2017). *Plumber: An API Generator for R.*

Understanding complex interactions within a large-scale manufacturing system through applications of Bayesian networks

Caoimhe M. Carbery\textsuperscript{1,2}, Roger Woods\textsuperscript{2} and Adele H. Marshall\textsuperscript{1}
\textsuperscript{1}Mathematical Sciences Research Centre (MSRC), Queen’s University Belfast, Northern Ireland
\textsuperscript{2}Electronic Computer Engineering (ECE), Queen’s University Belfast, Northern Ireland

Recent emphasis has been placed on the benefit of early prediction and detection of faults for manufacturing companies. Advanced knowledge of their product yield can impact future planning and avoid potential system faults. Big Data also plays a role within this application domain and brings forth further challenges with the capabilities of statistical analysis and machine learning, Cai, et al. (2017). Uncertainty is present in these systems which results in standard approaches often failing to appropriately represent the systems. To account for the uncertainties, the probabilistic model of Bayesian networks are chosen as the model for this research.

Introduction

Manufacturing companies are striving towards Industry 4.0. The means by which they can improve their performance is through advanced models to gain insight into their systems. Manufacturing lines produce key processing data from multiple stages of their build resulting in highly complex datasets. The question is whether one can utilise the abundance of data alongside machine learning and data analytics to construct complex models for a system that involves many interrelated components. The target of the analysis may be to utilise the models to improve systems, help with decision making, and identify crucial issues at an earlier stage.

Methods

Bayesian networks (BNs) are a type of probabilistic graphical model that can be used to capture complex relationships between variables in a system, Cai, et al. (2017). BNs are becoming increasingly popular in manufacturing analysis as they can handle uncertainties within real systems, and provide a mechanism to translate complex problems into a simpler viewpoint through using conditional probabilities.

Case study: Bosch Production Line

Bosch provided a large dataset containing instances from one of their production lines to encourage the development of methods that can handle highly complex data for predictive models, Kaggle (2017). The dataset consists of fully anonymised and scaled measurements of production jobs moving through different lines and stations in a factory, highlighting the complexity of the system.

The goal was to construct a classification/predictive model to identify which components will fail the final testing stage at the close of the production. The dataset has 1,183,747 observations and 986 variables at its raw state.
Results and Conclusions

BNs were utilised to generate an interpretable network for diagnosing faults in the Bosch production line through uncovering the interconnections amongst stations, Carbery, et al. (2018). The focus on this approach was to create a mechanism for learning the BN structure of a large complex system without prior knowledge which is often reserved for smaller data samples. Methods used to reduce the dimensionality of the large data through feature selection will be presented alongside the full BN learnt from the data. The relationships found amongst the variables highlight the complexity of the system and the model validation is also shown.

References


Conquering gut disorders using data science

Michael Cauchi
1Department of Mathematics and Statistics, University of Limerick, Ireland

The development of a non-invasive diagnostic tool for treating gut disorders caused by gastrointestinal diseases, particularly Crohn’s disease (CD), ulcerative colitis (UC) and irritable bowel syndrome (IBS) would be highly beneficial. Current “gold standard” diagnostic methods for these diseases are invasive, expensive and inconvenient. CD and UC tend to have similar symptoms making them hard to distinguish from one another, whilst IBS tends to be diagnosed following a process of elimination after a number of tests. It will be reported how the application of data science can overcome these limitations.

Introduction
Gut disorders caused by gastrointestinal diseases are a major global public health burden, leading to significantly increasing incidences of morbidity and mortality annually in developed countries. Of great concern are Crohn’s disease (CD), ulcerative colitis (UC) and irritable bowel syndrome (IBS). The gold standard diagnostic methods for CD and UC are colonoscopy and sigmoidoscopy. However, there have been many cases where individuals diagnosed originally with CD are then re-diagnosed with UC, and vice versa (Cauchi, et al., 2014). The diagnosis of IBS tends to be after the elimination of other gastrointestinal diseases such as CD and UC.

Recent studies have shown that volatile organic compounds (VOCs) present in the headspace of sample matrices such as urine could be used as diagnostic biomarkers, and thus lead to a non-invasive technique (Cauchi, et al., 2014 and 2015). Gas chromatography mass spectrometry (GC-MS), high performance liquid chromatography (HPLC-MS) and selected ion flow tube mass spectrometry (SIFT-MS) are three laboratory techniques that are able to separate a mixture of compounds and subsequently identify each compound thus generating copious amounts of data. These data are processed via multivariate statistical and machine learning techniques.

The outcome is presented here.

Methods

- 91 patient candidates were selected: 24 had CD, 19 had UC, 28 had IBS and the remainder deemed to be “healthy”. Each donated breath, blood, urine and faecal samples, which were stored/frozen immediately.
- All samples measured on GC-MS, HPLC-MS and SIFT-MS laboratory instruments.
- Pre-processing of data involved:
  - GC-MS and HPLC-MS: Normalisation against an internal standard, exploratory data analysis via principal components analysis (PCA), and alignment via correlation optimised warping (COW) (Cauchi, et al., 2014).
  - SIFT-MS: Normalisation against the H₃O⁺ precursor ion, and removal of known adducts (Cauchi, et al., 2015).
- Cross-model validation optimisation was performed via partial least squares discriminant analysis (PLS-DA), support vector machines (SVMs) and random forests (RFs).

Results and Discussion
For each instrument the better target case throughout was CD in both urine and faecal samples. The accuracy, specificity and sensitivity produced by the PLS-DA algorithm for GC-MS on faecal samples (CD v Control) were 85%, 78% and 93% respectively. All results were shown to be statistically significant when subjected to permutation testing.
Conclusion
A non-invasive procedure for the diagnosis of gastrointestinal diseases is possible, in particular CD, which compared favourably with the “gold standard” of colonoscopy, and was successfully distinguished from UC.

References

Sports injuries have a direct impact on player physiology, psychology and performance. In addition to that it has financial implication for the professional sport leagues such as Major League Baseball, NFL, NBA, and English Premier League. To alleviate this issue, studies relating to injury prediction in sports tend to follow four different analytical approaches: risk factor identification, prediction modelling, prevention using intervention, and identification of a single injury risk factor. However, most of the studies lack sufficient statistical power, causing unreliable prediction. Building registry to ensure enough number of events or devising proxy continuous variable could solve the issue. If prior information can be elicited from a relevant panel of experts Bayesian Networks could give more insight about the factors responsible for injuries and the causal pathway to injury occurrence.

Introduction
Sports injuries have a direct impact on player physiology, psychology and performance (Parry and Drust, 2006). The financial implications, in terms of salaries spent on injured players, equates to an estimated €281,000,000 in the English Premier League, $700,000,000 in Major League Baseball, $450,000,000 in the NFL, $350,000,000 in the NBA. In particular, injury rates in professional football are extremely high compared to high risk industrial occupation with an estimated risk of 1,000 times higher (Hawkins and Fuller, 1999). A squad of twenty-five players typically experiences 50 time-loss injuries per season (Ekstrand, et al., 2010). This rate of injury is increasing (Ekstrand, et al., 2010). The ability to predict injuries is clearly attractive. The purpose of this literature review is to explore current practices in the area of sport injury predictions.

To build a statistical model for valid and reliable injury prediction or prevention, high quality data are needed with sufficient statistical power. However most of the research published in this area tends to fail to meet or fail to report this crucial information regarding their study.

Published Models
Studies relating to injury prediction in sports tend to follow four different analytical approaches: risk factor identification, prediction modelling, prevention using intervention and identification of a single injury risk factor. A few studies used statistical tests such as chi-squared and t-test to identify (marginal) factors responsible for injuries (Hawkins and Fuller, 1999; Myklebust, et al., 2003; Dupont, et al., 2010; McDonough and Funk, 2014). Studies that involved modelling the probability of injury occurrence typically uses logistic regression (Brink, et al., 2010; Wilkerson, et al., 2012; Wilkerson and Colston, 2015; Hides and Stanton, 2017), tree based classification models (López-Valenciano, et al., 2017; Rossi, et al., 2017; Rossi, et al., 2017), ensemble methods, naive Bayes, support vector machine and neural networks (Ruddy, et al., 2017). In some cases a continuous proxy response for injury occurrence (e.g. knee abduction moment) linear regression models have been used (Myer, et al., 2010). Given the time to event nature of injury occurrence some studies (Venturelli, et al., 2011; Gabbett, et al., 2012; Bittencourt, et al., 2016) have used Cox regression, incorporating a frailty random effect. Given the large number of covariates typically available for inclusion, dimension reduction technique such as principal component analysis, singular value decomposition, and partial least squares correlation analysis have been used (Weaving, et al., 2017)

Number of Events
To ensure sufficient statistical power when modelling a binary outcome it is estimated that 10 events per variable are required (Steyerberg, 2008). However this requirement is typically not met (Altman, 2007) in general and in injury prediction problems in particular. For example, Table 1 lists a selection of papers (in chronological order) and includes the number of events and events per variable. Among 18 papers we reviewed, only 5 (28%) of them had adequate number of events per variable.
Table 1: Event per variable (EPV) and statistical methods used in the sample of papers reviewed.

<table>
<thead>
<tr>
<th>Author</th>
<th>Number of events</th>
<th>Number of variables</th>
<th>Events per variable</th>
<th>Statistical methods listed</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Brink, et al., 2010)</td>
<td>320</td>
<td>19</td>
<td>16.8</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>(Wilkerson, et al., 2012)</td>
<td>46</td>
<td>7</td>
<td>6.6</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>(Wadey, et al., 2012)</td>
<td>104</td>
<td>9</td>
<td>11.6</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>(Pontillo, et al., 2014)</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>(Myer, et al., 2014)</td>
<td>23</td>
<td>14</td>
<td>1.6</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>(Hides, et al., 2014)</td>
<td>191</td>
<td>3</td>
<td>63.7</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>(Grant, et al., 2015)</td>
<td>132</td>
<td>8</td>
<td>16.5</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>(Wilkerson and Colston, 2015)</td>
<td>132</td>
<td>5</td>
<td>26.4</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>(Ehrmann, et al., 2016)</td>
<td>16</td>
<td>5</td>
<td>3.2</td>
<td>Not mentioned</td>
</tr>
<tr>
<td>(Hides, et al., 2016)</td>
<td>16</td>
<td>4</td>
<td>4</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>(Hammes, et al., 2016)</td>
<td>67</td>
<td>10</td>
<td>6.7</td>
<td>Not mentioned</td>
</tr>
<tr>
<td>(Gribble, et al., 2016)</td>
<td>54</td>
<td>6</td>
<td>9</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>(Rossi, et al., 2017)</td>
<td>21</td>
<td>18</td>
<td>1.2</td>
<td>Decision tree, random forest</td>
</tr>
<tr>
<td>(Rossi, et al., 2017)</td>
<td>23</td>
<td>18</td>
<td>1.3</td>
<td>Decision tree</td>
</tr>
<tr>
<td>(López-Valenciano, et al., 2017)</td>
<td>32</td>
<td>13</td>
<td>2.5</td>
<td>Decision tree</td>
</tr>
<tr>
<td>(Hides and Stanton, 2017)</td>
<td>70</td>
<td>13</td>
<td>5.4</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>(Ruddy, et al., 2017)</td>
<td>27</td>
<td>4</td>
<td>6.8</td>
<td>Naive Bayes, logistic regression,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>random forest, support vector machine,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>neural network.</td>
</tr>
<tr>
<td>(Wilkerson, et al., 2018)</td>
<td>16</td>
<td>4</td>
<td>4</td>
<td>Logistic regression</td>
</tr>
</tbody>
</table>

Increasing the number of events is unlikely to be met with much approval. One alternative is to build a registry combining injury information from different teams across different seasons to increase the number of injuries to model. Building such a registry has its own challenges due to the competitive nature of elite sport and may need to involve the inclusion of non-competing teams (e.g. different leagues).

Another approach involves the use of a proxy continuous variable for injury occurrence. No such metric exists in elite football and is an avenue worth exploring.

If prior information can be elicited from a relevant panel of experts (e.g. sports scientists, sports medics, trainers or coaches) about the role different variables play in the likelihood of getting injured, Bayesian hierarchical models or Bayesian network models could be used.

A Bayesian network is a probabilistic directed acyclic graphical model that represents the set of random variables collected and their conditional dependencies and predictive ability of an unknown domain via a directed acyclic graph (DAG). A DAG contains a set of nodes to represent variables with edges, drawn as arrows between the nodes, used to indicate the direction of causality, representing hypotheses leveraged from a panel of domain knowledge experts. An example of a DAG for the modelling the probability of an injury occurring involving variables such as player load, history, wellness and mood state and their (possible) conditional dependencies on soft tissue injury is displayed in Figure 1.
Model Performance and Reporting
Regardless of the method used to model injuries, an evaluation of the performance of the proposed model and a validation of its predictive ability for future observations is needed. The usefulness of a statistical model for prediction purpose can only be verified when all its aspects are clearly reported. The Transparent Reporting of a multivariate prediction model for Individual Prognosis Or Diagnosis (TRIPOD) initiative recommends a minimum set of important aspect for reporting studies which are either developing, validating, or updating a prediction models in biomedical field (Collins, et al., 2015). The use of these guidelines for creating transparency in the value of proposed injury prediction models is crucial in sport science. To date very few sports science journals have adopted this policy.

Conclusion
Prediction models for injuries in elite sports are appealing. However, due to the small number of injuries occurring models proposed in the literature have not made it into practice. Given the sparsity of events available to model the question remains whether the focus should be on injury prediction or prevention? Bayesian networks could give more insight about the factors responsible for injuries and the causal pathway to injury occurrence.

References


Ripley’s $K$-statistics to resolve microglia territories and spatial dynamics

Benjamin Davis$^1$, Manual Salinas-Navarro$^2$, Francesca Cordeiro$^1$, Lieve Moons$^2$ and Lies De Groef$^2$

$^1$Institute of Ophthalmology, University College London, UK
$^2$Neural Circuit Development and Regeneration Group, KU Leuven, Belgium

Microglia are a type of cell in the central nervous system that play important roles in disorders of the retina and brain. The retina is part of the central nervous system that can be used to model neurodegenerative changes in response to injury. Using $k$-means clustering, rodent retinal microglial populations were first segmented into “low activity” and “high activity” subpopulations based on established cell morphology parameters before unmatched and matched Ripley’s $K$-functions and Dixon’s chi-squared test were used to evaluate deviations from spatial homogeneity in response to injury and potential biological implications.

Introduction

As part of the central nervous system, the retina is an attractive model for the study of neurodegenerative processes that play important roles in disorders of the retina and brain. This is because the retina contains many of the same cell types as found in the brain, but far fewer in number with a much better understood spatial organization. Microglia are a type of cell found throughout the central nervous system that are increasingly thought to play an important role in neurodegenerative disorders. There, however, remains significant uncertainty as to whether these cells have a predominantly neuroprotective or degenerative role due to a lack of techniques to adequately assess their behaviour in response to injury. This study describes a novel approach seeking to maximize information extraction from whole-retinal histological flat-mounts derived from rodent models of retinal injury. Here, Allograft inflammatory factor 1 (Iba-1$^+$) microglial cells were automatically segmented from histological retinal images while retaining information regarding the spatial and morphological distribution.

Methods

All studies were conducted in compliance with the ARVO Statement for the Use of Animals in Ophthalmic and Vision Research and the European Communities Council Directive of 22 September 2010 (2010/63/EU) and the Belgian legislation (KB of 29 May 2013), and were approved by the KU Leuven institutional ethical committee (P025-2013).

Wild type mice (C57Bl6/J; Charles River Laboratories), aged 2 to 4 months, were kept under a 12/12h light-dark cycle and had ad libitum access to food and water. Intraorbital optic nerve crush (ONC) was performed as previously described (De Groef, et al., 2016) in 14 animals under general anaesthesia. Naive eyes from a separate cohort of untreated animals served as controls ($N = 14$). 28 retinal whole mounts were dissected from three groups (ONC eyes, contralateral ONC eyes and naive control eyes) before immunohistochemistry for Iba-1 and imaged using a laser confocal scanning microscope (FV1000, Olympus), controlled with FluoViewer 4.2 software (Olympus) and microglial cells automatically segmented from resulting images.

Microglia centroids from each retina were used in conjunction with ‘low’ or ‘high’ activity designations to create spatial point processes bounded by retinal area. To determine the extent of microglia distribution deviation from spatial homogeneity, Ripley’s $K$-functions (Ripley, 1988) were applied to compare microglia spatial point patterns and assess deviation from compete spatial randomness. Initially, $K$-functions for unmarked spatial point processes were drawn as previously described (Kiskowski, et al., 2009). Compete spatial randomness was estimated using a random Poisson distribution $K(r) = \pi r^2$ where $r > 0$. Deviations from this estimation are indicative of incidence of clustering or dispersion. To facilitate estimation of microglia territory size, the variance stabilized Ripley’s $K$-function $L(r)$ and corresponding $H(r)$ was determined such that the expected value of $r$ under compete spatial randomness is zero. (Besag, 1977 and Ehrlich, et al., 2004).

Conclusion

Retinal microglia were found to demonstrate a dispersed distribution maintaining distinct territories of measurable size. In a naive retina, microglia typically have small somata and large territories, corresponding to a ‘low activity’ status. They are regularly interspersed with a minority of ‘high activity’ cells. Four days post optic nerve injury, microglia density increased and the majority of cells adopted a ‘high activity’ phenotype,
characterized by a reduction in territory size, enlargement of their cell body and a more irregular cell shape. Furthermore, although ‘low activity’ and ‘high activity’ microglia are regularly distributed, ‘high activity’ microglia had a tendency to self-associate.

References


A joint model for the analysis of longitudinal and survival data utilising the Coxian phase-type distribution

Conor Donnelly¹, Lisa M. McFetridge¹ and Adele H. Marshall¹
¹Mathematical Sciences Research Centre, Queen’s University Belfast, Northern Ireland

Joint modelling techniques are a relatively recent statistical development, capable of simultaneously modelling a longitudinal response and survival outcome so as to overcome the bias which occurs when there exists an association between the two processes.

This research explores the use of the Coxian phase-type distribution to represent the survival process within the joint likelihood, in place of the standard Cox proportional hazards model, allowing additional latent information to be ascertained about how the survival process behaves before the event occurs.

Introduction

In today’s digital society, the collection of longitudinal and survival data often happens concurrently, especially within the medical field, where it is typically observed that there exists an association between both processes. When such an association exists, previous research has shown that independent analysis of one of the processes, ignoring the effect of the other, can lead to biased inferences, Henderson, et al. (2000). In order to overcome this, the parameters of the longitudinal and survival sub-models can be estimated simultaneously through a single joint likelihood, where commonly a linear mixed effects (LME) model is used to represent the longitudinal process and a Cox proportional hazards (PH) model to represent the survival process.

More recently, it has been shown that the unspecified baseline of the Cox PH can cause the standard errors within the joint model to be underestimated, Hsieh, et al. (2006), making fully parametric models a potentially more attractive representation of the survival process. However, it has been noted within the literature that these standard models, which most often assume an exponential, Weibull or gamma distribution, are restricted in terms of the distributional shapes which they can represent, thus limiting their scope and motivating the development of alternative survival representations which assume piecewise constant baseline hazards or make use of $B$-splines to estimate the baseline hazard.

Methods and Discussion

Within this research, the survival process is instead represented by the Coxian phase-type distribution, which can be utilised to represent any positive distribution to an arbitrary degree of accuracy, therefore overcoming the limitations encountered with parametric distributions that have been used in previous literature. Further, as phase-type distributions can be considered to represent the time to absorption within a finite state Markov process with a single absorbing state, they offer the additional benefit of allowing inferences to be made regarding the rates of flow between the states of the underlying Markov process, Neuts (1989). As a result, more information can be gained relating to how individuals behave before experiencing their event of interest.

A joint likelihood for the longitudinal and survival processes is generated, as given below, and the unknown parameters are estimated using maximum likelihood:

$$L(\theta; y_i, \tau_i, \delta_i) = \prod_{i=1}^{m} \int f(y_i | b_i; \theta_y) f(\tau_i, \delta_i | b_i; \theta_\tau) f(b_i; \theta_b) \, db_i \quad (1)$$

where $f(y_i | b_i; \theta_y)$ represents the density of the longitudinal process conditional on the random effects, $b_i$, represented by a LME model, $f(\tau_i, \delta_i | b_i; \theta_\tau)$ represents the density of the survival process conditional upon the random effects, represented by a Coxian phase-type regression model, and $f(b_i; \theta_b)$ is the density of the random effects.
Results from the application of this methodology to data collected on individuals suffering from chronic kidney disease shall be presented, illustrating the Coxian phase-type distribution’s ability to uncover latent stages of the survival process and allow inferences to be made regarding the rates of flow through them.

References


Discrete survival models including heterogeneity

Lida Fallah\textsuperscript{1} and John Hinde\textsuperscript{1}  
\textsuperscript{1}School of Mathematics, Statistics, and Applied Mathematics, National University of Ireland Galway, Ireland

We applied generalized mixed effect models to a real dataset collected with the purpose of studying the pathogenicity and virulence of a specific type of fungus. The main aim is to model the mortality of insects which are exposed to a kind of toxicity.

Introduction

Time to event, or survival, data is common in the agricultural, biological and medical sciences. Although it is typically recorded as a continuous measurement, in some situations it can be discrete. For example, toxicology experiments may be designed such that treated insects are monitored with their mortality recorded only every couple of days.

This type of experiment results in datasets where the survival time responses are grouped over time intervals and we refer to these as discrete/clustered survival data. Notice that such experiments can also be considered as longitudinal studies in which individuals are repeatedly measured over adjacent time intervals.

To reduce the amount of pest damages to crops in Brazil, the Insect Pathology Laboratory of ESALQ-USP, Brazil, has collected a dataset focusing on the use of the fungus \textit{Beauveria bassiana} as a microbial control for \textit{Heterotermes tenuis} insects.

Solutions of the isolates were applied to groups (clusters) of \( n = 30 \) termites kept in plastic Petri-dishes, with five replicate dishes for each isolate. The mortality in the groups was measured by counting the dead termites daily for a period of eight days after application of the fungus.

In this work, we consider the above clustered survival dataset as sequences of repeated binomial observations/longitudinal multinomial observations with censoring after 8 days. We model the mortality rate of the insects with various random effect models and use these to cluster isolates with similar behavior.

Methods

Suppose in an experimental study the individuals are monitored over time intervals \( T_i = [\tau_{i-1}, \tau_i), i = 1, 2, \ldots, m \), where \( \tau_0 = 0 \) and \( \tau_m \) is assumed to be any time after the study termination time. The interval hazard is defined by Tibshirani and Ciampi (1983) as follows:

\[
p_i(z) = Pr(\tau_{i-1} \leq T \leq \tau_i | T \geq \tau_{i-1}),
\]

where \( p_i \) depends on the covariate \( z \) but does not vary over time within the interval, i.e., we assume that for a given \( z \), \( p_i(z) \) is the piecewise constant hazard within the time interval \( T_i \).

Modelling the hazard function for this kind of clustered longitudinal survival data is typically done using the Generalized Linear Model (GLM) and Generalized Linear Mixed Model (GLMM), with appropriate link functions. These both allow for modelling non-normal responses (dependent variables), while with GLMM, correlated and uncorrelated random effects can be incorporated in the model to account for additional variation, often referred to as heterogeneity in outcome values.
Results and Conclusion
A GLMM is fitted to the data considering random intercepts for isolates and replicates within isolates. Figure 1 (left) pictures the cumulative mortality rate (hazard function) over eight days and the right hand side plot is to demonstrate the model fit. To conclude this, as the observed mortality rate curves are quite well recovered by the fitted ones, the model’s variance components have captured the sources of variability and the model is well fitted.

References

This poster presents a summary of seven collaborative projects involving Brazilian postgrads and postdocs who have visited NUI Galway in recent years. This work arises from long term collaborations between John Hinde and Clarice Demétrio.

**Introduction**

International collaborations are a vital component of scientific research and beneficial for both the scientific developments and long-lasting links that can be established. There is special value for the developing/emerging world and the opportunities that can be created for young researchers to broaden their horizons and contacts.

**Selected Projects:**

**Modelling Competition Between Overlapping Niche Predators**

A biological control related experiment on the predatory behaviour of the stinkbug and the earwig when given the choice between parasitised and non-parasitised prey. Statistical modelling of the time-until-attack data used different accelerated failure rate models (exponential, Weibull, and exponentiated-Weibull) with simultaneous models for the location and the scale parameters.

**Half-Normal Plots and Overdispersed Models in R: The hnp Package**

This work developed an R package, hnp, that may be used to generate the half-normal plot with a simulated envelope for residuals from different types of models. The function hnp() can be used together with a range of different model fitting packages in R that extend the basic generalized linear model fitting in glm() and is written so that it is relatively easy to extend to new model classes and different diagnostics. The methodology has been used on a wide range of examples, including continuous and discrete responses, and can be used to inform model selection and diagnose overdispersion.

**Bivariate Residual Plots with Simulation Polygons**

This is ongoing work aimed at extending the half-normal plot with a simulated envelope to bivariate models. The resulting plot provides checks of both marginal and joint behaviour.

**Conditional and Marginal Models for Analysing Light Interception Data**

See the abstract by Rafael Moral for “Modelling bounded data in plant ecology” on page 77.

**Models for Jointly Estimating Abundances of Two Unmarked Site-Associated Species Subject to Imperfect Detection**

This extends the N-mixture modelling framework for species abundance with imperfect detection to model two site-associated species abundances jointly and proposes to measure the influence of one species’ abundance on the populations of the other, and study how this changes over time and space. By including an additional parameter in the abundance distribution of one of the species, linking it to abundance of the other, the proposed model treats extra-variability as an effect induced by an associated species’ abundance and allows the study of how environmental covariates may affect this. The approach is illustrated using data from bald eagles and mallards obtained in the 2015 survey of the North American Breeding Bird Survey. By using the joint model we were able to separate overdispersion from mallard-induced variability and hence what would be accounted for with a dispersion parameter in the univariate framework for the eagles was explained by covariates related to mallard abundance in the joint model.
Transition Models Applied to Pig Behaviour Data
This work looked at the use of Markov transition models for analysing longitudinal ordinal data. Methodological aspects included model selection and the assessment of stationarity. The application considered a study of the animal behaviour of pigs and the impact of environmental enrichment where the housing pens were equipped at different times with suspended chains, a suspended 5 litre plastic container, and a loose 50 litre container.

Longitudinal Concordance Correlation Function Based on Variance Components: An Application in Fruit Color Analysis
The maturity stages of papaya fruit based on peel color is frequently characterized from a sample of four points on the equatorial region measured by a colorimeter. However, this procedure may not be suitable for assessing the papaya’s overall mean color and an alternative proposal is to use image acquisition of the whole fruit’s peel. Questions of interest are whether a sample on the equatorial region can reproduce a sample over the whole peel region and if the colorimeter can compete with a scanner, or digital camera, in measuring the mean hue over time. The reproducibility can be verified by using the concordance correlation for responses measured on a continuous scale. In this work we propose a longitudinal concordance correlation (LCC), based on a mixed-effects regression model, to estimate agreement over time among pairs of observations obtained from different combinations between measurement method and sampled peel region. The results show that the papaya’s equatorial region is not representative of the whole peel region, suggesting the use of image analysis rather than a colorimeter to measure the mean hue. Moreover, in longitudinal studies the LCC can suggest over which period the two methods are likely to be in agreement and where the simpler colorimeter method could be used.

Conclusion
These collaborations, visits, and exchanges are an exciting and satisfying way of being involved in many different projects. Benefits accrue to both the visiting students and the host group – NUIG has been happy to welcome our Brazilian visitors and the two countries have much in common, just not the weather!

Acknowledgements
We are grateful for financial support from SFI (Ireland), FAPESP and CNPq (Brazil) and of our colleagues in NUI Galway and ESALQ/USP.

References

Papers:


**Software:**


Prostate cancer (PCa) represents a significant healthcare problem due to the dilemmas associated with its detection and treatment especially with the projected increase in its incidence in Ireland and internationally. An Irish PCa risk calculator created from a national collection of patients can allow for individualised risk stratification and can be used to improve clinical decision making. Statistical approaches applied to clinical parameters from an Irish dataset were used to build a risk calculator to inform clinicians and patients as to the need for a biopsy to diagnose PCa. The use of this risk calculator will impact on the patients’ outcome and quality of life but also alleviate the pressures on our already overburdened healthcare sector by reducing biopsies. The risk calculator will be presented, and its performance will be discussed using a collection of numerical and graphical performance outcome summaries. An interactive Shiny application will also be presented which was designed to aid our understanding of the result.

Introduction
Patients and clinicians are faced with dilemmas associated with the detection and treatment of Prostate cancer (PCa). One such dilemma is in the early stages of diagnosis when men are referred by their GP for suspicion of PCa due to their elevated Prostate-specific antigen (PSA) value or suspicious Digital rectal examination (DRE), but it is not clear if they need a biopsy. Accurate risk stratification of patients before biopsy would help to reduce overdiagnosis and lead to better clinical decision making. The ERSPC and the PCPT are two well-known international risk calculators available for diagnosis of PCa. They have been tested in an Irish population and proved to be beneficial; however, we hypothesis that the predictive accuracy will be significantly improved when built with an Irish population.

Method
A national dataset including the routinely used clinical information of 4808 patients from the eight Irish tertiary referral rapid access clinical centres were collected. A risk calculator for the diagnosis of PCa (and high-grade PCa) was created using a logistic regression model including components such as age, DRE, family history of PCa, prior negative biopsy and PSA level.

<table>
<thead>
<tr>
<th>Models</th>
<th>AUC</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSA</td>
<td>0.5948</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>PCPT</td>
<td>0.6308</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>IPRC</td>
<td>0.6741</td>
<td>-</td>
</tr>
</tbody>
</table>

The discriminate ability of the model was compared with the current biomarker indicator (PSA) and PCPT risk calculator using various graphical (e.g Figure 1) and numerical performance outcome summaries (e.g. Table 1).
Discussion
An Irish PCa risk calculator can allow for individualised risk stratification and can be used to improve clinical decision making in Irish men under investigation for PCa. The proposed model calculates the risk of PCa as a probability; however, an optimal threshold needed to be chosen to make the best clinical decision. The selection of this threshold is challenging as it depends on a trade-off between a more sensitive test or a more specific test. For this reason, an interactive Shiny application is created to combine the graphical and numerical summaries to convey the result of the final model in the most translated way to clinicians and decision makers.

References

A comparison of methodological approaches on the association between De Novo vitamin D supplement use post-diagnosis and breast cancer mortality

Jamie Madden¹, Lina Zgaga², Finbarr Leacy³ and Kathleen Bennett¹

¹Division of Population and Health Sciences, RCSI, Dublin, Ireland
²Department of Public Health and Primary Care, Trinity College Dublin, Ireland
³Data Science Centre, RCSI, Dublin, Ireland

In this study we compare different approaches to remove biases such as confounding by indication from non-randomized studies. The traditional conditioning approach (adjustment), propensity score analysis and marginal structural models will be examined. Data from the National Cancer Registry of Ireland (NCRI) will be used to examine the association between vitamin D supplement use and breast cancer mortality.

Introduction
Pharmacy claims data offers a unique opportunity for non-randomized comparative effectiveness research but studies can be inherently biased due to confounding by indication (Hernan and Robins, 2016). This study investigates different methodological approaches to overcome inherent bias while examining the association between vitamin D supplements initiated after breast cancer incident diagnosis on survival.

Method
Women aged 50-80 years with a record of invasive breast cancer were identified on the NCRI database (n = 5417). Initiation of De Novo vitamin D post-diagnosis was identified from linked prescription data (n = 2581, 49%). We first implemented a standard multivariate Cox proportional hazards (PH) model to estimate adjusted HRs (95% CIs) for breast cancer-specific mortality. We compared these findings to a propensity score analysis approach. We subsequently sought to account for the time-varying nature of vitamin D use and time-varying confounding by bisphosphonate use using inverse probability of treatment weighted marginal structural models (MSMs), exploring the impact various weight trimming approaches have on the effect estimates.

Results
Using the standard Cox PH models we found a 20% reduction in breast cancer-specific mortality in de novo vitamin D users compared to non-users (HR, 0.80; 95% CI, 0.64-0.99). A similar point estimate was observed, but larger CIs, for the association between vitamin D use and breast cancer-specific mortality (HR: 0.80; 95% CI: 0.60-1.06) when correcting for covariate imbalance between treatment groups at baseline using propensity score analysis approach. We subsequently sought to account for the time-varying nature of vitamin D use and time-varying confounding by bisphosphonate use using inverse probability of treatment weighted marginal structural models (MSMs), exploring the impact various weight trimming approaches have on the effect estimates.

Discussion
Our results will be discussed with an emphasis on the problem of large weights formed for marginal structural models when using administrative data that is collected daily over a long period of time.

References
Business intelligence dashboard solutions for improving decision making process: A focus on prostate cancer

Mona Isazad Mashinchi¹, Davood Roshan², Francis J. Sullivan³ and Dietrich Rebholz-Schuhmann¹
¹Insight Centre for Data Analytics, National University of Ireland Galway, Ireland
²School of Mathematics, Statistics and Applied Mathematics, National University of Ireland Galway, Ireland
³Prostate Cancer Institute, National University of Ireland Galway & Galway Clinic, Ireland

Decision making processes are nowadays driven by data, data analytics and Business Intelligence (BI). BI as a software platform can provide a wide variety of capabilities such as organisation memory, information integration, insight creation and presentation capabilities. Visualizing data through dashboards is one of the BI solutions (for a variety of areas) which helps managers in the decision making processes to expose the most informative information at a glance. In the healthcare domain to date, dashboard presentations are more frequently used to track performance related metrics and less frequently used to monitor those quality parameters which relate directly to patient outcomes. Providing effective and timely care for patients and improving the health outcome are highly dependent on presenting and visualizing data and information.

Introduction
In this research, the focus is on the presentation capabilities of BI to design a dashboard for prostate cancer (PC) data that allows better decision making for the patients, the hospital and the healthcare system related to a cancer dataset. The aim of this research is to customize a retrospective PC dataset in a dashboard interface to give a better understanding of data in the categories (risk factors, treatment approaches, disease control and side effects) which matter most to patients as well as other stakeholders. By presenting the outcome in the dashboard we address one of the major targets of a value-based healthcare (VBHC) delivery model which is measuring the value and presenting the outcome to different actors in HC industry (such as patients and doctors) for a better decision making.

Method
For visualizing the stored data to users, three interactive dashboards based on the PC dataset have been developed (using the Tableau Software) to provide better views to the risk factors, treatment approaches and side effects.

Results
Many benefits derived from interactive graphs and tables in dashboards which helped to easily visualize and see the patients at risk, better understanding the relationship between patient’s status after treatment and their initial status before treatment, or to choose better decision about treatments with fewer side effects regarding patient status and etc.

Conclusion
Building a well-designed and informative dashboard is related to three important factors including: the users, goals and the data types. Dashboard’s hierarchies, drilling and graphical features can guide doctors to better navigate through information. The features of the interactive PC dashboard not only let doctors ask specific questions and filter the results based on the key performance indicators (KPI) such as: Gleason Grade, Patient’s Age and Status, but may also help patients to better understand different treatment outcomes, such as side effects during the time, and have an active role in their treatment decisions. Currently, we are extending the results to the real-time interactive dashboard that users (either patients and doctors) can easily explore the data by choosing preferred attribute and data to make better near real-time decisions.

Acknowledgments
The project is supported by the Irish Research Council (IRC) and Galway Clinic.

References


Propensity score matching methods for observational studies

Michael McCague¹
¹Biostatistics Unit, HRB Clinical Research Facility, National University of Ireland Galway, Ireland

Propensity score matching is an increasingly popular statistical concept used for causal inference in observational studies. The goal is to match subjects in the treatment and control groups as closely as possible on their observed characteristics so as to approximate the conditions of a randomized controlled trial (which is the gold standard of evidence in estimating treatment effects). There are many different matching algorithms and settings advocated in current literature. In this poster, using a sample of data arising from an observational study of the effect of Clostridium Difficile Infection (CDI) on patient Length of Stay (LOS) in hospital, I present a comparative study of some of these methods and their merits.
Investigating the computational trade-off for computing matrix exponentials using \textit{kexpmv}

Meabh G. McCurdy\textsuperscript{1}, Adele H. Marshall\textsuperscript{1} and Karen J. Cairns\textsuperscript{1}

\textsuperscript{1}Mathematical Science Research Centre, School of Mathematics & Physics, Queen’s University Belfast, Belfast, Northern Ireland

The exponential of a matrix is a widely researched and utilised matrix operator. However, one of the main drawbacks associated with this matrix operator is its computational complexity. The standard methods in \texttt{R} and \texttt{Matlab}, such as the scaling and squaring method, are efficient for small dense matrices but can become problematic when the dimensions or sparsity of the matrix increases. Krylov subspace methods aim to minimise this computational issue and are particularly efficient when dealing with large sparse matrices. This research investigates the computational trade-off between matrix exponential routines deemed efficient for small dense and large sparse matrices by utilising functions within the \texttt{R} package \textit{kexpmv}.

Introduction

The matrix exponential is a well defined researched area, which is reflected in the vast number of current methods available to calculate this matrix operator, Moler and Van Loan (2003). The standard methods, such as scaling and squaring are seen as efficient when dealing with small dense matrices, however becomes problematic when the dimensions or sparsity of the matrix increase. The computational complexity associated with the matrix exponential is a major drawback and the severity of this issue can grow with the dimensions of the matrix, which can limit its use within different application areas. Krylov subspace methods have been seen to address this issue and are particularly efficient for large sparse matrices. This research investigates the computational trade-off between the different routines utilised for both small dense and large sparse matrices. The aim is to identify what properties and qualities a matrix must display in order to implement the most efficient routine resulting in the fastest computational time.

Method

The routines deemed efficient for large sparse matrices are those that utilise Krylov subspace methods. These methods have made significant advances to overcome the computational issue with matrix exponentials by replacing matrix-matrix operations with matrix-vector operations. The \texttt{Expokit} (Sidje, 1998) is an extensive software package that incorporates these Krylov subspace methods to efficiently calculate matrix exponentials. In doing so, the software also includes routines for both large sparse matrices along with small dense matrices. A selection of these routines are included and implemented in the \texttt{R} package \textit{kexpmv}. This package is aimed at improving the computational efficiency for calculating matrix exponentials in \texttt{R}. The routines utilised within this \texttt{R} package for small dense matrices incorporate the scaling and squaring algorithm combined with a Padé approximation, whereas the routines utilised for large sparse matrices incorporate Krylov subspace methods.

Results

Both real and simulated data was used in this study to investigate the computational trade-off for matrix exponentials. The dimension of the matrix proved to be crucial when determining the most efficient routine. The study concluded and agreed with the literature that for matrices classed as small ($n \leq 1200$) the small dense routine was deemed most efficient. However when the dimension of the matrix is large ($n > 1200$) the large sparse routines that incorporate Krylov subspace methods were not always the most efficient. It was evident that for large matrices the norm of the matrix played an important role when determining the most efficient routine. These conclusions are validated through comparing the computational times for matrix exponentials across the different routines within the \texttt{kexpmv} package.
Conclusion
The computational issue associated with computing matrix exponentials is crucial to the development and utilisation of this matrix operator. This issue can limit the use of this matrix operator in many different application areas. Therefore it is important to investigate and define which dimensions are classed as ‘large’ and ‘small’. Examining these matrix properties should ensure the most efficient routine is implemented hence resulting in faster and more efficient computational times.

References

Spatial point pattern analysis of rodent retinal ganglion cells

Hannah J. Mitchell\textsuperscript{1}, Benjamin M. Davis\textsuperscript{2}, Li Guo\textsuperscript{2}, M. Francesca Cordeiro\textsuperscript{2}
\textsuperscript{1}Mathematical Sciences Research Centre, Queen’s University Belfast, Northern Ireland
\textsuperscript{2}Glaucoma and Retinal Neurodegenerative Disease Research Group, Institute of Ophthalmology, University College London, UK

This research applies spatial data analysis techniques to replicated point patterns derived from the distribution of retinal ganglion cells (RGCs) in two well-established rodent glaucoma models, partial optic nerve transection (pONT) or ocular hypertension (OHT). In particular, the $K$ function is estimated for each of the different experimental groups and between-group comparisons are made.

Introduction

With the emergence of techniques capable of recording human retinal images approaching the single cell resolution, coupled with many diseases of the retina and brain not presenting symptoms until late in the disease process (after much irreversible damage has already occurred), there is an unmet-need for a better understanding of early neurodegenerative disease processes at the cellular resolution. This work seeks to apply spatial statistical approaches to investigate disease associated changes in the retina at the cellular level with a view to establishing models of disease processes that could improve our understanding mechanisms underlying neurodegenerative conditions of the eye and brain.

Retinal Ganglion Cells (RGCs) cells are thought to play a role in diseases such as glaucoma where they are progressively lost. The present study utilizes a dataset derived from two established rodent glaucoma models (partial optic nerve transection (pONT) and ocular hypertension (OHT)) where snapshots of the RGC population was assessed at multiple timepoints using whole-retinal histological approaches to identify early any disease specific patterns of RGC loss which could be used to inform the development of novel diagnostic and therapeutic approaches for this condition.

Within this research, statistical analysis of spatial point patterns is utilised to uncover differences in the retinal ganglion cells locations and how they change over time. The $K$ function is used to investigate different forms of departure from complete spatial randomness for replicated spatial point patterns.

Method

Ripley’s $K$ function (Ripley, 1976) describes characteristics of the point processes at many distance scales. An unbiased estimate of the $K$ function is,

$$
\hat{K}_{ij}(t) = n_{ij}^{-1} \lambda_{ij}^{-1} \sum_{g \neq h} e_{ij}^{-1}(g, h) I[d_{ij}(g, h) \leq t]
$$

where $d_{ij}(g, h)$ is the distance between the $g$th and $h$th cells for the $j$th subject in the $i$th group. $I[.]$ is an indicator function for cells at a distance less than or equal to $t$ from an arbitrarily chosen cell of the process and $e_{ij}(g, h)$ is a correction for edge effects. $\lambda_{ij}$ is the intensity and $n_{ij}$ is the number of data points for the $j$th subject in the $i$th group.

Diggle, et al. (2013) showed that a weighted average of the individual estimates $\bar{K}_i(t)$ can be calculated using,

$$
\bar{K}_i(t) = \sum_{j=1}^{n_i} w_{ij} \hat{K}_{ij}(t) \quad \text{for } i = 1, ..., g
$$

with an overall average mean function,

$$
\bar{K}(t) = \frac{1}{n} \sum_{i=1}^{g} n_i \bar{K}_i(t)
$$
where \( w_{ij} = \frac{n_{ij}}{n_i} \), \( n_i = \sum_{j=1}^{m_i} n_{ij} \) and \( n = \sum_{i=1}^{g} n_i \), (\( m_i \) is the number of subjects in the \( i \)th group). Diggle (2013) also showed that these \( K \) functions can be used to obtain a statistic to measure differences between groups.

**Conclusion**
The research presented analyses replicated spatial point patterns of rodent retinal ganglion cells. It investigates different forms of departure from complete spatial randomness as well as detecting differences between patterns across the experimental groups if they exist.

**References**

Predicting the stage of the prostate cancer for personalized treatment in an Irish cohort

Shirin Moghaddam¹, Keefe Murphy²,³, Lisa Murphy¹, Laura O’Gorman¹, Thomas Lynch¹, Richard Power⁵, Kieran J. O’Malley⁶, Thomas Brendan Murphy²,³ and William Watson¹

¹Conway Institute of Biomolecular and Biomedical Research, School of Medicine, University College Dublin, Ireland
²School of Mathematics and Statistics, University College Dublin, Ireland
³Insight Centre for Data Analytics, University College Dublin, Ireland
⁴Department of Urology, St. James Hospital, Ireland
⁵Department of Urology, Beaumont Hospital, Ireland
⁶Department of Urology, Mater Misericordiae University Hospital, Ireland

Prostate cancer is the most common malignancy among men in developed countries. The single biomarker test for prostate-specific antigen (PSA) has decreased the number of deaths from prostate cancer. However, it is controversial due to low specificity and inability to identify aggressive forms of cancer which has led to overdiagnosis and treatment. The main dilemma faced by the patient and clinician once prostate cancer has been detected is how best to treat it. Here we are using optimised logistic regression and multiple new biomarker based diagnostics to enable accurate staging of prostate cancer and guidance for appropriate choices of therapy.

Predicting the stage in prostate cancer

Predictive tools based on standard clinicopathologic variables have been developed for prostate cancer including look-up tables and nomograms. One such highly regarded tool is the Partin tables which was developed using multivariate logistic regression. While the Partin tables are well validated and used by clinicians, studies in Ireland were not able to validate the findings with the discriminate ability of less than 70% for the organ-confined and non-organ confined disease (Boyce, et al., 2013).

The main goal of this study is to test whether the detection of a panel of serum protein biomarkers can be used to accurately detect and establish the stage of prostate cancers. A panel of nine serum protein biomarkers were assessed by the MSD platform based from previous discovery studies (Oon, et al., 2012). In a cohort of 150 Irish patients undergoing a radical prostatectomy for localized prostate cancer as part of the Prostate Cancer Research Consortium bioresource. Discrimination between pathological stages was investigated using a logistic regression model.

Model evaluation was carried out by examining calibration, discrimination and decision curve analysis.

The discriminate ability of the models were compared by the area under the curve (AUC) values. Decision curve analysis is also used for evaluating and comparing prediction models based on new biomarkers compare to the Partin table. By using the stepwise logistic regression model, the variables with the significant effects are PSA, biopsy GS, CD14, IGFBP3, GcGlobulin, ZAG, IGF1.

Figure 1 shows the discriminate ability measured using ROC curves, AUC values and decision curves for the prediction based on the Partin table and also new significant biomarkers.
patients undergoing a radical prostatectomy for localized prostate cancer as part
of the Prostate Cancer Research Consortium bioresource. Discrimination between
pathological stages was investigated using a logistic regression model.

Model evaluation was carried out by examining calibration, discrimination and de-
cision curve analysis.

The discriminate ability of the models were compared by the area under the curve
(AUC) values. Decision curve analysis is also used for evaluating and comparing
prediction models based on new biomarkers compare to the Partin table. By using
the stepwise logistic regression model, the variables with the significant effects are
PSA, biopsy GS, CD14, IGFBP3, GcGlobulin, ZAG, IGF1. Figure 1 shows the
discriminate ability measured using ROC curves, AUC values and decision curves
for the prediction based on the Partin table and also new significant biomarkers.

![ROC curves and decision curves for predicting the stage of prostate cancer.]

Figure 1: ROC curves (left) and decision curves (right) for predicting the stage of the prostate cancer. Here
“None” means assuming no one has non-organ confined prostate cancer and “all” means assuming everyone
has non-organ confined prostate cancer.

It is shown that the predictive model proposed where new biomarkers included has the highest AUC and net
benefit, which means it is a superior approach in predicting the stage of the prostate cancer.

References
Boyce, S. et al. (2013). Evaluation of prediction models for the staging of prostate cancer, *BMC Medical
Informatics and Decision Making*, 13, pp. 126.

Oon, S.F. et al. (2012). The identification and internal validation of a preoperative serum biomarker panel to

Tosoian, J.J. et al. (2017). Prediction of pathological stage based on clinical stage, serum prostate-specific
antigen, and biopsy Gleason score: Partin Tables in the contemporary era. *BJU International*, 119, pp. 676 –
683.
Data linkage in medical settings using the resource description framework: The AVERT model

Brian P. Reddy1,2, Mark Little1,3, Lucy Hederman2,4, Alan Meehan2, Declan O’Sullivan2,4 and Brett Houlding4
1Trinity Health Kidney Centre, Tallaght Hospital, Dublin, Ireland
2ADAPT Centre for Digital Content, Trinity College Dublin, Ireland
3Irish Centre For Vascular Biology, Trinity College Dublin, Ireland
4School of Computer Science and Statistics, Trinity College Dublin, Ireland

There is an ongoing challenge as to how best manage and understand ‘big data’ in precision medicine settings. We describe the potential for a Linked Data approach using a Resource Description Framework (RDF) model, to combine multiple datasets with temporal and spatial elements of varying dimensionality. This “A VERT model” provides a framework for converting multiple standalone files of various formats, from both clinical and environmental settings, into a single data source. This data source can thereafter be queried effectively, shared with outside parties, more easily understood by multiple stakeholders using standardized vocabularies, incorporating provenance metadata and supporting temporo-spatial reasoning. The approach has further advantages in terms of data sharing, security and subsequent analysis. We use a case study relating to anti-Glomerular Basement Membrane (GBM) disease, a rare autoimmune condition, to illustrate its utility.

Introduction

The availability of relevant data in healthcare and other settings has been growing exponentially in recent years. This poses challenges in terms of data storage, management and security. We used a series of steps to transform heterogeneous data, from a variety of different formats, into a single queryable data source using the Resource Description Framework model (RDF) and semantic web technologies. This data source facilitates further insights through data enrichment, eases the application of machine learning approaches to be used and supports scientific data management best practice according to the FAIR principles (Wilkinson, et al., 2016). The aim of this paper is to describe a model which has the potential to be applicable for use in many health informatics settings. It was initially developed as part of a project investigating the environmental risk factors for anti-GBM disease, requiring the use of temporo-spatial reasoning.

Method

8 unrelated datasets were identified for inclusion in the analyses, but they differed in their descriptions of temporal-level and geospatial-level data, in what format they were available, and ranged in size from 14KB to 8.75GB. We developed a sequence of steps, shown in Figure 1, to transform the data into RDF format (a process known as “uplift”), which could be combined into a single data source (a “triplestore”). This could then be queried to extract the relevant data for subsequent analyses. Throughout the process we adhered to existing W3C (2018) standards, used standardized ontologies to facilitate both sharing and machine learning, and used existing languages and packages at each stage to transform or query the data.

![Figure 1](https://example.com/figure1.png)

Figure 1: The “AVERT model” for transforming siloed tabular datasets into RDF, and back into enriched file for analyses. Further datasets were also used in practice.
Results
The model is a pragmatic, standards-based solution to integrating temporo-spatial environmental data with patient-level clinical information, to address an epidemiological research question. It successfully transformed multiple standalone tabular files into a single, queryable and sharable data source. We are not aware of the existing processes being used together in sequence in this way before.

Discussion and Conclusion
The AVERT model can be used to uplift tabular data into a common RDF format. From this it can (1) be converted back into a tabular format (“downlifting”), enriched by incorporation of external data sources and reasoning algorithms, and (2) be managed in a codified format that follows well understood ontologies, facilitating sharing and understanding by both external groups and machine learning scenarios. A clear advantage of the AVERT model when compared to siloed tabular files is that the integration of data in RDF, alongside the use of SPARQL (a W3C standard language that facilitates geospatial RDF queries), allows quicker and more intuitive searching of big data. Merged datasets should help to ensure that data is managed effectively and reduces the risk of human error. Once data is linked, it may lead to new opportunities for understanding causal mechanisms under analysis. Some of these are relatively simple, such as allowing visualisations; others, such as allowing the use of machine learning approaches, are more complex and promising.

References

A comparative statistical study of micrograph classification using repeatability and reproducibility analysis

Saritha Unnikrishnan\textsuperscript{1,2}, John Donova\textsuperscript{1,2}, Russell Macpherson\textsuperscript{3} and David Tormey\textsuperscript{1,2}

\textsuperscript{1}Department of Mechanical and Electronic Engineering, Institute of Technology Sligo, Ireland
\textsuperscript{2}Centre for Precision Engineering, Materials and Manufacturing Research, Institute of Technology Sligo, Ireland
\textsuperscript{3}GlaxoSmithKline, Sligo, Ireland

The quality evaluation of colloidal dispersions has been typically based on human assessment of microscopic images (micrographs). Due to the subjective nature of this interpretation, a multivariate classification model was developed to classify micrographs of emulsion samples obtained at various processing intervals. A set of micrographs were classified by human examiners as well as the classification model. A repeatability and reproducibility (R&R) test was performed to compare the accuracy and precision of the evaluations. The results confirm a much higher accuracy, repeatability and speed for the computer classification model.

\textbf{Introduction}

Microscopic evaluation of emulsion samples by humans has been widely used in pharmaceutical industries for product quality assessment, which can lead to subjectivity of interpretation. Assessment of emulsion samples using droplet data obtained through image processing of micrographs has shown potential in the characterisation of emulsions (Hosseini, \textit{et al.}, 2015). We investigate the accuracy and precision of a classification algorithm, developed using droplet characteristics, in comparison to manual assessment in classifying in-process micrographs of an oil-in-water (O/W) emulsion.

\textbf{Method}

Micrographs were acquired at six different mixing intervals during the emulsification process of an O/W emulsion. Droplet characteristics such as size, intensity and count were obtained for all the micrographs through image processing performed using Fiji. Statistical analysis of the droplet data was conducted using R and the micrographs were further categorised into four classes based on the variation in the statistical distribution of droplet characteristics. The classification algorithm was based on a multivariate approach using the droplet characteristics. Principal Component Analysis (PCA) of the droplet features was performed to reduce dimensionality and retain maximum variability. A linear discriminant analysis (LDA) classification model was developed using the significant principal components as the predictors. The accuracy and precision of human assessment was studied by conducting an R&R test with a set of 20 micrographs. Two examiners were evaluated and each examiner randomly assessed each micrograph twice. The task of the examiners was to classify the micrographs into four predefined classes based on their individual assessment. The same set of micrographs were employed for the LDA model classification. An independent set of 12 micrographs were used for training both the examiners and the model. The R&R results were compared using attribute agreement analysis.

\textbf{Results}

The two individual examiners presented a classification accuracy of 50\% (95\% CI: 27.20\%, 72.80\%) and 55\% (95\% CI: 31.53\%, 76.94\%) respectively, while the computer assessment accuracy was 90\% (95\% CI: 68.30\%, 98.77\%). The precision results for repeatability within appraisers were 65\% and 75\% for the first and second appraisers and 50\% reproducibility between appraisers. The repeatability of computer assessment was 100\%. The computer based classification was executed in less than 5 seconds, whereas the human evaluation took around 15 to 20 minutes.

\textbf{Conclusion}

Human assessment of micrograph classification has been shown to have an accuracy less than 55\% with a repeatability of approximately 70\%. This compares rather poorly with a statistical classification algorithm such as LDA. The time taken for an individual to determine the classification is also much longer.

\textbf{References}

Session 3: Thursday 9.00 - 10.40
Chair: Gabrielle Kelly
Statistical modelling of air pollution and and the challenges of communication

Carl Scarrott

School of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand

In New Zealand, air quality standards aiming at a minimum level of health protection are prescribed by the National Environmental Standards, which are enacted by the regional councils and unitary authorities. This talk will consider the progress towards meeting the standard for particulate matter below 10µm, referred to as PM$_{10}$, which is linked to various respiratory and cardiac problems, including premature deaths.

The standard specifies an upper threshold for the 24 hour average PM$_{10}$ of 50µg/m$^3$, which should not be exceeded more than once a year by 2020. Planning decisions and the issue of emission permits are to be controlled to ensure this standard is met. Commonly in practice, the trends in second largest concentration per year are modelled along with the impact of possible future policy scenarios to ensure this order statistic meets the threshold of 50µg/m$^3$ by 2020. This method ignores the uncertainty due to the large inter-annual meteorological variation and will underestimate the reductions required to be sufficiently confident that the standard will be met.

A generalised additive mixed model is developed to describe the meteorological impacts and historical trends in the PM$_{10}$ concentrations. A simulation approach using randomised block bootstrapped meteorological conditions is implemented, along with potential policy scenarios, to determine the percentage reductions needed to achieve these future targets with a certain likelihood. The results are used to inform air quality management policy and planning. Some of the challenges of effectively communicating uncertainty and the study findings to a non-statistical audience, including environmental scientists and policymakers, will be discussed.
Changepoints to improve forecasts

Jamie-Leigh Chapman¹, Rebecca Killick¹ and Idris Eckley¹
¹Department of Mathematics and Statistics, Lancaster University, UK

Forecasting time series accurately is challenging, particularly so when the structure of the data changes through time. In recent years, research has focused on forecasting for non-stationary time series, particularly in a local evolutionary setting. However many time series encountered in practice may have a piecewise stationary form. In this talk we introduce a novel approach to improve forecasting using changepoints and apply it to Gross Domestic Product data.

Introduction

Forecasting models try to capture the behaviour of a time series through various components, including outliers, level, trend, volatility, seasonality, dependence and explanatory variables. In practice, these components are often assumed to be stationary, but the reality is that these properties may vary over time in a piecewise way. Consider for example the UK’s Gross Domestic Product (GDP) data in Figure 1, the volatility is higher prior to 1981 and in general the behaviour is not constant over time.

Methods

Suppose, by way of introduction, we wish to obtain forecasts for the UK’s GDP. As a precursor to formal forecasting, one might perform changepoint analysis to identify regions of piecewise stationary structure. More formally, suppose our time series \((y_1, \ldots, y_n)\) contains \(m\) changepoints with positions \(\tau = (\tau_1, \ldots, \tau_m)\). Each changepoint position, \(\tau_i\), is an integer between 1 and \(n - 1\) and we define: \(\tau_0 = 0\) and \(\tau_{m+1} = n\). The changepoints are ordered such that: \(\tau_i < \tau_j \iff i < j\). Consequently the \(m\) changepoints split the time series data into \(m + 1\) segments with the \(i^{th}\) segment containing \(y(\tau_{i-1}+1):\tau_i\). Each segment \(i\) is summarized by a set of parameters \(\{\theta_i, \phi_i\}\). Within this set, \(\phi_i\) is a set of nuisance parameters and \(\theta_i\) is a set of parameters that may contain changes (Eckley, et al. 2011).

One way to detect multiple changepoints is to minimize: \(\sum_{i=1}^{m+1} [C(y(\tau_{i-1}+1):\tau_i)] + \beta f(m)\), where \(C\) is some cost function for a segment and \(\beta f(m)\) is a penalty based on the number of changepoints \(m\) which prevents over-fitting (Killick, et al. 2012). For forecasting GDP, we can use the negative log-likelihood for an autoregressive integrated moving average (ARIMA) model as the cost function. In particular, we can use a dynamic cost function which searches over different models, fits the best one, and returns the log likelihood. This allows for both changes in coefficients and model order. If we locate a significant change in the ARIMA model, then we should forecast using only the data after that change. Figure 1 shows the detected changes in the ARIMA model for the UK’s GDP and forecasts using all of the historical data, and only using the data after 2013.

Figure 1: The United Kingdom’s Gross Domestic Product (GDP) quarter on quarter growth. The vertical (blue) solid line is a change in variance and the (orange) dashed lines are changes in ARIMA structure. The enclosed box shows forecasts using all of the data and using only data after the most recent change.
Results and Conclusions
Changepoint analysis is not only useful for the retrospective understanding of data, but it can also be used to improve future forecasts. It’s integration into existing time series forecasting models can be straightforward and we often achieve better out-of-sample forecasts based on less historical information.

References

Assessing spatial interactions among species in a grassland experiment

Jack McDonnell¹,², Thomas McKenna³, Kathryn Yurkonis³, Deirdre Hennessy² and Caroline Brophy¹
¹Department of Mathematics and Statistics, Maynooth University, Ireland
²Teagasc Animal and Grassland Research and Innovation Centre, Ireland
³Department of Biology, University of North Dakota, Grand Forks, ND, USA

Diversity-Interaction (DI) models have been used to model biological experiments that manipulate proportions of species, but have not previously accounted for spatial interactions between species. The weed biomass from a three year grassland plot experiment in North Dakota with sixteen species which included a spatial sowing pattern treatment was analysed by extending DI modelling methods to incorporate spatial effects. Species diversity suppressed weed biomass over time. Although spatial pattern did not appear to have an overall effect on weed biomass, it affected the performance of individual species in multispecies plots.

Introduction
Mixing species in grasslands has been shown to suppress weed invasion. This study investigates whether the sowing pattern makes a difference. DI models have been used to model grassland plot experiments in the past, but none with spatial pattern treatments.

Data and Methods
A grassland plot experiment was undertaken at the University of North Dakota from 2012 to 2014, inclusive. There were 170 plots planted with 1, 2, 4 or 8 species from a pool of sixteen grass species (four each from four functional groups with similar traits) in each plot. A spatial treatment was manipulated across plots with more than one species (mixture plots): individuals of the same species in each plot were either dispersed or aggregated. Weed biomass was collected, dried and weighed monthly.

DI models contain an identity effect and a diversity effect (Kirwan et al. 2009; Brophy et al. 2017). The identity effects are interpreted as the yields of each species in monoculture, while the diversity effects are the effects of mixing species additional to the weighted identity effects. Spatial treatment was allowed to interact with the identity and diversity effects in the model, and was tested as a fixed effect in the model by itself. Likelihood ratio tests and AIC values were used for model comparisons. The variance-covariance structures of the DI models were manipulated to incorporate a repeated measures structure and to test whether the variance was constant in all plots. The pairwise interactions between species were also included as random effects in an effort to explain extra variability not explained by the fixed effects.

Results
An initial analysis showed there was no difference between the means of the dispersed and aggregated mixture plots. However, DI modelling methods showed that the spatial pattern had an effect at the species level. This suggested that the weed invasion in mixtures of certain species would differ for dispersed and aggregated plots, which was supported by predictions from the model: a mixture of species 1 and 14 (a warm season grass and a legume) had predicted weed biomass in 2014 of 156.8 g DM and 33.0 g DM in dispersed and aggregated plots respectively. Each species contributed a unique fixed effect to an interaction with another species regardless of the species it was interacting with. The variability of weed biomass in monocultures (plots with one species) differed depending on the functional traits of species.

Discussion
DI modelling showed that the effects of spatial sowing pattern were species specific. This information can be used in management decisions to minimise weed invasion.

References

Variability of variance estimates: The parametric bootstrap approach

Anthony Kinsella¹, Kevin Hayes² and Kevin O’Brien²
¹Molecular and Cellular Therapeutics, Royal College of Surgeons in Ireland, Ireland
²Department of Mathematics and Statistics, University of Limerick, Ireland

The following linear mixed effects model underpins method comparison studies such as the calibration of two measurement instruments or quantitative analytical techniques where there is no within unit or sample replication:

\[ y_{ij} = \mu + \beta_i + \tau_j + \epsilon_{ij} \text{ for } i = 1, 2, \ldots, n \text{ and } j = 1, 2. \]

Here \( \beta_i \) denotes a random between unit or sample component assumed \( N(0, \sigma^2_\beta) \), \( \tau_j \) denotes the systematic difference between the instruments or techniques and \( \epsilon_{ij} \) denotes the random variation of the \( i \)th reading from the \( j \)th instrument or technique with \( \epsilon_{i1} \) assumed \( N(0, \sigma^2_1) \) and \( \epsilon_{i2} \) assumed \( N(0, \sigma^2_2) \).

Hayes, et al. (2017) show how the existence of any systematic difference between the instruments or techniques and equality of the within instrument variances, \( \sigma^2_1 \) and \( \sigma^2_2 \), can be tested.

Grubbs (1948) proposed point estimates of \( \sigma^2_\beta \), \( \sigma^2_1 \) and \( \sigma^2_2 \) which, assuming normality and the correct dimensionality, are unbiased and are simple functions of sufficient statistics. In practice these estimators, which are functions of the data variances and the covariance, can produce negative estimates despite the corresponding parameters being, by definition, positive. Thompson (1963), using a restricted maximum likelihood approach, developed estimators which avoids the negative estimate problem.

The parametric bootstrap is proposed as a method for quantifying the variability of the original data set point estimates where the individual data values are simulated as follows:

\[ \tilde{y}_{i1} = \tilde{\beta}_i + \tilde{\epsilon}_{i1} \]

and

\[ \tilde{y}_{i2} = \tilde{\beta}_i + \tilde{\epsilon}_{i2} \]

with \( \tilde{\beta}_i \) assumed \( N(0, \tilde{\sigma}^2_\beta) \), \( \tilde{\epsilon}_{i1} \) assumed \( N(0, \tilde{\sigma}^2_1) \) and \( \tilde{\epsilon}_{i2} \) assumed \( N(0, \tilde{\sigma}^2_2) \). Here \( \tilde{\sigma}^2_\beta \), \( \tilde{\sigma}^2_1 \) and \( \tilde{\sigma}^2_2 \) denote the original dataset estimates of these variances. Selected percentiles of the parametric bootstrap distribution of the Thompson non-negative estimates are computed using the simulated data values. This approach allows examination of the effect of changes in the magnitudes of the computed point estimates when \( \tilde{\epsilon}_{i1} \) is assumed \( N(0, k_1\tilde{\sigma}^2_1) \) and \( \tilde{\epsilon}_{i2} \) assumed \( N(0, k_2\tilde{\sigma}^2_2) \) where \( k_1 \) and \( k_2 \) are multiplicative factors.

The technique is illustrated using the results of an instrument calibration study.

References


Inside Insight Session:
Thursday 11.00 - 12.40

Chair: Nial Friel
Teaching data science

Garrett Grolemund

1RStudio Inc

Data Science. How should we teach it? Should we teach it? I make some recommendations based on my experiences teaching adults how to do data science with R. Like it or not, data science is here because many parts of classical statistics are no longer necessary for making discoveries with data. And yet, the message of our profession is more important than ever. If we can reform our curriculum to capture and serve future data scientists, we can ensure that data science remains an actual method of science. If we fail, I predict that data science will become just another body of knowledge pertaining to a topic, like computer science, library science, or managerial science. So then, let’s talk about what we should do.
Unobserved classes and extra variables in high-dimensional discriminant analysis

Michael Fop\(^1\), Pierre-Alexandre Mattei\(^2\), Thomas Brendan Murphy\(^1\) and Charles Bouveyron\(^3\)
\(^1\)School of Mathematics and Statistics, Insight Research Centre, University College Dublin, Ireland
\(^2\)Department of Computer Science, IT University of Copenhagen, Denmark
\(^3\)Laboratoire J.A. Dieudonné, UMR CNRS 7351 and Equipe Epione, INRIA Sophia-Antipolis, Université Côte d’Azur, France

In supervised classification problems, the test set may contain data points belonging to classes not observed in the learning phase. Moreover, the same units in the test data may be measured on a set of additional variables recorded at a subsequent stage with respect to when the learning sample was collected. In this situation, the classifier built in the learning phase needs to adapt to handle potential unknown classes and the extra dimensions. We introduce a model-based discriminant approach that can detect unobserved classes and adapt to the increasing dimensionality. The method is embedded in a more general framework for adaptive variable selection and classification particularly suitable for high-dimensional data. An application to spectrometry data for food authenticity studies is presented.

Introduction

Standard supervised classification assumes that all existing classes in the data have been observed during the learning phase. However, in some cases the test set could present units that belong to classes not previously observed (Bouveyron, 2014). Moreover, the new observations to be classified may be recorded on a collection of additional variables, other than those already observed in the learning data. In such setting, a classifier will fail to detect the novel classes and assign the observations only to the classes it is aware of; it will also need to adapt to the increasing dimensionality and parameter space.

Methods

We introduce Streaming-Adaptive Mixture Discriminant Analysis for model-based adaptive supervised classification. The model is a generalization of Gaussian mixture discriminant analysis and is designed to classify new observations \(y_i\) measured on additional variables and containing information about unobserved classes.

Let \(x_s\) be an observation of the training data and \(\ell_s\) the known class it belongs to, \(K\) the number of classes observed in the training data and \(H\) the number of extra “hidden” classes potentially present in the test data; then the total number of classes is \(C = K + H\). The model is depicted as follows:

\[
\text{Training data: } x_s | \ell_s \sim \mathcal{N}(\mu_k; \Sigma_k) \quad \ell_s \sim \prod_{k=1}^{K} \tau_k^1(\ell_s = k)
\]

\[
\text{New data: } y_i \sim \sum_{k=1}^{K} \tau_k \mathcal{N}(\mu_k^*; \Sigma_k^*) + \sum_{h=K+1}^{C} \tau_h \mathcal{N}(\mu_h^*; \Sigma_h^*),
\]

where the * symbol denotes the increased dimensionality. Model estimation is carried out via an EM algorithm and consists of two phases: learning phase, where the classifier is estimated using the training data, and adaptive-discovery phase, where the model adapts to the increasing dimensions and detects the new classes. The model is then embedded in a more general framework for adaptive classification and variable selection (Maugis, et al., 2011).

Conclusions

We proposed a general framework for classification and variable selection when the new data contain unobserved classes and extra variables. The approach is based on a fast and computationally efficient estimation procedure, suitable for high-dimensional data.
The method is shown in application to high-dimensional spectroscopy data concerning food authenticity in the situation where the new data has been collected at a subsequent stage and at a finer resolution (with an increased number of wavelengths/variables). The purpose is to detect potential unobserved contaminants and select a small subset of the spectrum containing as much information for authentication as the whole spectrum does. The method performs a parsimonious variable selection and attains good classification performance.

References

Text mining aided literature review: A case study in sports injuries

Jaynal Abedin\textsuperscript{1} and John Newell\textsuperscript{2}
\textsuperscript{1}Insight Centre for Data Analytics, National University of Ireland Galway, Ireland
\textsuperscript{2}School of Mathematics, Statistics and Applied Mathematics, National University of Ireland Galway, Ireland

A systematic literature review is one approach to summarise existing evidence relating to a certain question of interest. Considering the growing rate of publication it is time consuming to do a systematic review manually. In this work we apply Topic Modeling to a set of relevant abstracts to identify research themes and how they change over time. An application is given in the domain of sports injuries.

Introduction

Systematic literature reviews are a widely used approach to accumulate and summarise existing research findings and are a recognised way of informing policy makers about evidence relating to the key question of interest. With ongoing advancement in technology, publication rates continue to increase dramatically creating challenges for completing efficient systematic reviews. Due to the large number of published papers in a particular domain it is almost impossible to manually review all published papers available on a certain area. To address this, we applied exploratory Topic Modeling as a potentially useful tool to streamline the literature review process.

Methods (Literature search & selection: A case study in sport injuries)

Identifying modifiable risk factors to prevent injuries in elite sports is an active research area spanning many academic domains (e.g. medicine, physiology, sports science & data science). In order to undertake a literature review, the PubMed database was used where we restricted the search to within the last 6 years in order to focus on more recent journal and review articles. The following query was executed:

\[(\text{sports AND injury})\]

Figure 1: Literature review plot of the contribution of words to a theme over time
Results
The search query resulted in 17,663 hits between 2013 to 2018. On average there were 3,500 publications per year. Using Topic Modeling five underlying research themes were identified namely, head injury, knee injury, exercise and performance, recovery from injury (Figure 1). The majority of the research papers were focused on soccer compared to other sports.

Conclusions
Topic modeling is a useful literature review tool to consider when the number of papers to review is large.

References

Gaussian parsimonious clustering models with covariates

Keefe Murphy\(^1\)\(^,\)\(^2\) and Thomas Brendan Murphy\(^1\)\(^,\)\(^2\)

\(^1\)School of Mathematics and Statistics, University College Dublin, Ireland
\(^2\)Insight Centre for Data Analytics, University College Dublin, Ireland

We consider model-based clustering methods that account for external information available in the presence of covariates. The proposed MoEClust family of models address equivalent aims: incorporating covariates into \textit{mclust} models (Scrucca, \textit{et al.}, 2016) and incorporating parsimonious covariance structures into Mixtures of Experts. MoEClust models show significant improvement from both perspectives. We also introduce the accompanying \texttt{R} package \texttt{MoEClust}, the syntax of which will be familiar to users of \textit{mclust}.

**Introduction**

Mixtures of Experts (Jacobs, \textit{et al.}, 1991) extend finite mixture model to allow covariates guide the clustering process, and not only interpretation of the clusters, through the mixing proportions &/or component densities. They can provide richer insight into the type of observation that characterises each cluster. The proposed MoEClust models avail of the parsimony afforded by the various \textit{mclust} covariance parameterisations that have so far been lacking in the Mixture of Experts context.

**Methods**

MoEClust models continuous outcome variables \(y_i\) via finite Gaussian mixtures and models parameters as functions of related covariates \(x_i\): mixing proportions \(\tau_g(x_i)\) (‘gating’ networks) by multinomial logistic regression and component densities (‘expert’ networks) by multivariate linear regression. Thus, covariates can effect either, neither, or both the probability of cluster membership and the distribution of \(y_i\).

\[
f(y_i|x_i) = \sum_{g=1}^{G} \tau_g(x_i) \text{MVN}_p \left( y_i \mid \theta_g(x_i) = \left\{ \gamma_g^\top x_i, \Sigma_g \right\} \right)
\]

Parsimonious parameterisations of \(\Sigma_g\) are obtained via an eigen-decomposition of the form \(\Sigma_g = \lambda_g D_g A_g D_g^\top\). \(\lambda_g\), a scalar controlling the ellipsoid’s volume, \(A_g\) a diagonal matrix specifying the shape of the density contours, and \(D_g\) an orthogonal matrix governing the orientation, can all be (un)strained across clusters, yielding 14 possible models. Thus, for instance, an EVE model is one where clusters have equal volume and orientation, with varying shape. Existing software either offers VVV and VVI models only, or doesn’t allow dependence on covariates in any way.

We focus on maximum likelihood estimation using the EM algorithm. The complete data log-likelihood separates into portions due to the gating and expert networks, which can be maximised separately. Fitting \(G\) separate multivariate regressions (weighted by \(\hat{z}_{ig}\)) yields \(G\) sets of \(n \times p\) residuals \(\hat{r}_{ig} = y_i - \gamma_g^\top x_i\) which crucially satisfy \(\sum_{i=1}^{n} \hat{z}_{ig} \hat{r}_{ig} = 0\). Thus, the M-step amounts to minimising this expression:

\[
\sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}^{(t+1)} \log|\Sigma_g| + \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}^{(t+1)} \hat{r}_{ig}^\top \Sigma_g^{-1} \hat{r}_{ig}
\]

This has the same form as \textit{mclust}’s M-step criterion, with component means equal to 0, using augmented data \(\hat{R}\) which stacks the residuals into a \((n \times G) \times p\) matrix.

**Results**

Performance is assessed on well-known Australian Institute of Sport (AIS) data (Cook and Weisberg, 1994). Interest lies in clustering hematological measurements from 102 male and 100 female athletes with reference
to their ‘BMI’ and ‘sex’. The BIC is employed to choose the appropriate model structure and number of clusters, and guide the inclusion of covariates. Table 1 clearly shows incorporation of covariates improving the fit compared to \textit{mclust} models and the parsimonious covariance structures improving the fit compared to standard Mixtures of Experts.

Table 1: BIC values for a selection of MoEClust models fitted to the AIS data. Rows 1-3 give the optimal models corresponding to those available in previously extant software. Remaining rows give the optimal MoEClust models with covariates in the gating or expert networks only, or both.

<table>
<thead>
<tr>
<th>Gating</th>
<th>Expert</th>
<th>G</th>
<th>Model Type</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMI</td>
<td>sex</td>
<td>2</td>
<td>VVV</td>
<td>-4054.86</td>
</tr>
<tr>
<td>BMI +</td>
<td>sex</td>
<td>4</td>
<td>VVI</td>
<td>-4302.82</td>
</tr>
<tr>
<td></td>
<td>sex</td>
<td>2</td>
<td>EVE</td>
<td>-4146.16</td>
</tr>
<tr>
<td>sex</td>
<td></td>
<td>2</td>
<td>EVE</td>
<td>-4015.35</td>
</tr>
<tr>
<td></td>
<td>sex</td>
<td>3</td>
<td>EVE</td>
<td>-4037.32</td>
</tr>
<tr>
<td>BMI</td>
<td>sex</td>
<td>2</td>
<td>EVE</td>
<td>-4013.40</td>
</tr>
</tbody>
</table>

References


Session 5: Thursday 14.00 - 15.20
Chair: Cathal Walsh
Modelling bounded data in plant ecology

Rafael Moral\textsuperscript{1}, John Hinde\textsuperscript{2}, Wagner Bonat\textsuperscript{3} and Clarice Demétrio\textsuperscript{4}
\textsuperscript{1}Department of Mathematics and Statistics, Maynooth University, Ireland
\textsuperscript{2}School of Mathematics, Statistics, and Applied Mathematics, National University of Ireland Galway, Ireland
\textsuperscript{3}Laboratório de Estatística e Geoinformação, UFRP, Brazil
\textsuperscript{4}Departamento de Ciências Exatas, ESALQ-USP, Brazil

Here, we analyse data from an experiment in Biodiversity and Ecosystem Functioning theory. The response variable is the percentage of intercepted light by the canopy in forest patches that were previously restored with three different numbers of species (treatments). Because the experimental design includes multilevel and longitudinal sampling, we propose conditionally specified beta mixed models, where we include random effects to incorporate the correlation between observations made within the same plot, as well as between the longitudinal observations made within each subplot. We discuss different computational strategies for fitting these models and pinpoint advantages and potential drawbacks of the approach.

Introduction

An important variable measured by forest restoration practitioners is light interception, which is an indicator of self-sustainability of a forest. According to the Biodiversity and Ecosystem Functioning theory, with a higher species diversity there is an increase in the number and intensity of ecosystem functions.

To assess whether higher tree diversity promotes (i) higher light interception and (ii) a more even distribution of light, both horizontally and vertically in a forest, an experiment was conducted in patches of restored Atlantic Forest in Brazil. There were three different levels of species richness, 20, 60, and 117 species, with four replicates in a completely randomized design. In each plot, twelve subplots were sampled at 0, 1, 2, 3, and 4 metres high, hence giving a form of longitudinal (height) study. The observed variable was the percentage of light interception by the canopy.

Methods

Let $y_{ijkl}$ be the percentage of light intercepted by the canopy at the $j$-th replicate, $j = 1, \ldots, 4$, of the $i$-th treatment, $i = 1, 2, 3$, measured at the $k$-th subplot, $k = 1, \ldots, 12$, at the $l$-th height level, $l = 1, \ldots, 5$. The Beta distribution is a reasonable assumption for modelling the response variable, since it is highly flexible and bounded on the $(0, 1)$ interval (Cribari-Neto and Zeileis, 2010).

![Figure 1: Light interception data: observed responses (grey points), means (blue points), fitted curves and confidence intervals (shaded) for each treatment](image)

Clearly longitudinal observations taken on the same subplot may be correlated, as may all observations made on the same plot. To accommodate these correlations we include normally distributed random intercepts, $b_{1ij} \sim N(0, \sigma_1^2)$, for observations in the same plot, and random intercepts, $b_{2ijk} \sim N(0, \sigma_2^2)$, and slopes,
$b_{3ijk} \sim \mathcal{N}(0, \sigma_3^2)$, for observations on the same subplot, with $\text{Cov}(b_{2ijk}, b_{3ijk}) = \sigma_{23}$. Then, we take the conditional distribution of $Y_{ijkl} | b_{1ij}, b_{2ijk}, b_{3ijk}$ as $\text{Beta}(\mu_{ijkl}, \phi_{ijkl})$ and implement code in R to fit this model using maximum likelihood.

**Results and Discussion**

We note that the proportion of light intercepted by the canopy gets smaller as the height increases (see Figure 1). Patches with a higher number of species intercept more light, giving evidence that higher tree diversity promotes more light interception and therefore makes it more likely for systems to be self-sustainable. Additionally, the variability increases over height and is smaller for patches with higher diversity, giving evidence of a niche complementarity effect.

Further work also includes optimization of the code written for fitting the models, and further exploration of the behaviour of the model with simulation studies.

**References**

Accurately estimating parameters for a bovine tuberculosis epidemic model in Northern Ireland

Emma Brown¹, Adele H. Marshall¹, Andrew Byrne²,³ and Hannah Mitchell¹
¹Mathematical Sciences Research Centre, Queen’s University Belfast, Northern Ireland
²Agri-Food and Biosciences Institute, Belfast, Northern Ireland
³School of Biological Sciences, Queen’s University Belfast, Northern Ireland

Bovine Tuberculosis (bTB), caused by Mycobacterium bovis, is endemic throughout Northern Ireland, and has had a substantial economic burden on the region (Abernethy, et al., 2006). Modelling the disease accurately could reduce uncertainty surrounding within-herd transmission. Properly estimating the transmission (β) and incubation (σ) parameters is necessary to accurately predict disease levels. Approximate Bayesian Computation (ABC) methods have been used to estimate model parameters (Conlan, et al., 2012; Ciaravino, et al., 2018).

Least squares estimation and ABC methods were compared to produce accurate parameter estimates.

Introduction

This work focused on estimating parameters for a Susceptible-Exposed-Infectious (SEI) epidemic model, shown in order in the following equations, by comparing ordinary least squares and ABC estimation:

\[
\begin{align*}
\frac{dS}{dt} &= bN - \beta SI - \mu S \\
\frac{dE}{dt} &= \beta SI - \sigma I - \mu E \\
\frac{dI}{dt} &= \sigma E - \mu I
\end{align*}
\]

where \(N = S + E + I\) (total population), \(b\) is the birth rate, and \(\mu\) is the death rate.

Methods

The parameters \(\beta\) and \(\sigma\) were initially estimated using ordinary least squares (OLS), which were then used to initialise the ABC rejection algorithm. ABC-MCMC (Markov Chain Monte Carlo) was implemented as it estimates values when the prior distribution is uninformative or different to the posterior (Ciaravino, et al., 2018). A normal distribution, with the respective OLS estimates for \(\beta\) and \(\sigma\) as the mean and 0.005 as the standard deviation, was used as the proposal distribution. ABC-MCMC, however, can break down if the parameters are highly correlated. Due to the definition of \(\beta\) and \(\sigma\) this was expected.

To tackle this problem, ABC with Sequential Monte Carlo (ABC-SMC) is compared to ABC-MCMC, as it is hypothesised that the SMC algorithm will improve estimates as intermediary distributions are created between the prior and posterior. Data from ante-mortem skin tests was provided by DAERA to estimate the parameters.

Results

The \(\beta\) and \(\sigma\) estimated using OLS and ABC-MCMC are found in Tables 1 and 2 respectively.

Table 1: OLS estimate of \(\beta\) and \(\sigma\) to 3 decimal places

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Least Squares</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\beta)</td>
<td>0.012</td>
</tr>
<tr>
<td>(\sigma)</td>
<td>0.026</td>
</tr>
</tbody>
</table>

Table 2: ABC-MCMC estimate of \(\beta\) and \(\sigma\) to 3 decimal places

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean Time-Series standard error</th>
<th>95% C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\beta)</td>
<td>0.097 0.022</td>
<td>[0.005, 0.240]</td>
</tr>
<tr>
<td>(\sigma)</td>
<td>0.029 0.001</td>
<td>[0.021, 0.051]</td>
</tr>
</tbody>
</table>
Conclusions
While ABC-MCMC can possibly be influenced by correlated parameters, the OLS estimates were within the ABC-MCMCs 95% C.I. The ABC-MCMC mean estimate of $\sigma$ was also extremely close to the OLS value while the ABC-MCMC mean for $\beta$ was not. Interesting insights may be gained when the estimates are compared to those produced by ABC-SMC which should perform better given its robustness against correlated parameters.

References


Statistical algorithms and issues arising during the development of a hand force-plate sensor system for false start detection in competitive sprinting

Kevin Hayes¹, Thomas Barr², Evan Crotty², Jonathan Holmes² and Drew Harrison²
¹Department of Mathematics and Statistics, University of Limerick, Ireland
²Department of Physical Education and Sport Sciences, University of Limerick, Ireland

A series of experimental studies carried out by the Human Performance Laboratory at the University of Limerick have established that the first response all athletes make when executing a sprint start is to push the ground with their hands. This presentation reports on the statistical issues that arose during the development and testing of a new false start detection system using a prototype hand force-plate system.

Introduction

A new hand plate sensor system developed at the University of Limerick represents a major shift in design of sprint start detection technology because it fundamentally changes the way in which a false start is detected. This new system examines changes in the ground reaction force under the hands using a custom built hand plate sensor and signal processing algorithm. Our research has revealed that in all cases, the change in force at the hand ground interface precedes the response in the blocks by between 40 to 100 ms. Consequently, starting-block based systems are not capable of detecting an athlete’s first response to the start signal and this is a fundamental drawback in the technology currently in use.

Methods

Experimental: Elite sprinters (n = 20) executed 3 to 4 sprint starts and their reaction times were simultaneously determined by (i) an approved IAAF start system; (ii) accelerometers placed on the starting blocks; and (iii) a prototype hand force-plate detection system. An IAAF certified official starter administered the starts.

Figure 1: Accelerometer readings on blocks (top); hand force-plate readings (below).
Statistical: An algorithm based on CUSUM scores was developed to process the accelerometer readings from the blocks in place of the less sensitive threshold method used by the IAAF system. A signed score algorithm was developed to determine reaction times based on hand force plate detector. Graphical comparison of these determinations of reaction time were made using Bland-Altman plots showing linked replicates, (R package MethComp; Carstensen, et al., 2015).

Results
The accelerometer recordings at the blocks and measured force on the hand plate for subject 1 trial A are shown in Figure 1. This example is typical of the records collected in the study. The firing of the gun and the official IAAF reaction time are marked as solid vertical lines in each plot. Determinations of the reaction time by the hand force-plate is indicated by the vertical dotted line in the bottom panel. In all trials, force on the hand plate system was detected in advance of movement at the blocks. The statistical reliability of these detection systems is discussed.

References
Interval censored MPR modelling

Defen Peng\textsuperscript{1}, Kevin Burke\textsuperscript{2} and Gilbert MacKenzie\textsuperscript{3}
\textsuperscript{1}Centre for Improved Cardiovascular Health, University of British Columbia, Canada
\textsuperscript{2}Department of Mathematics and Statistics, University of Limerick, Ireland
\textsuperscript{3}CREST, ENSAI, Rennes, France

We extend the non-PH Weibull, multi-parameter regression (MPR) survival model with frailty (Burke & MacKenzie, 2016) to the interval censoring setting and evaluate its performance by means of simulation and real data analysis. In particular a re-analysis of the tooth data presented by Gomez (2009) is revealing. The standard Weibull (SPR) model suggests that two covariates are important: gender and dmf. However, only dmf is important in the MPR model and remarkably this model can also explain the apparent gender difference. This finding raises important questions about the relevance of SPR models and their widespread use in risk factor evaluation.

Introduction

Multi-parameter regression (MPR) survival models extend the usual single parameter survival models by modelling the scale and shape parameters by means of two separate linear predictors. In this paper we generalise the Weibull MPR model with gamma frailty to the interval censored setting. Interval censored data arise in a wide variety of studies. They are particularly common in longitudinal randomised controlled trials (MacKenzie & Peng, 2013) when the outcome is time to some event and the response is measured repeatedly at scheduled follow-up visits.

Methods

The Weibull multi-parameter survival regression model with gamma frailty is defined by

\[ \lambda(t_i; \alpha, \beta, \xi) = \xi \lambda t_i^{\gamma_i - 1}, \]

then,

\[ S(t_i; \alpha, \beta, \xi) = \exp[-\xi \lambda t_i^{\gamma_i}], \]

where, \( \lambda = \exp(x^T \beta), \gamma = \exp(z^T \alpha) \), and \( \xi \) is an unobserved frailty term.

Then the likelihood of Weibull model for interval censored data is

\[ L(\theta|t_i, \delta_i, x_i, z_i) = \prod_{i=1}^{n} \left\{ e^{-\lambda t_{i,k-1}^{\gamma_i}} \left[ 1 - e^{-\lambda t_{i,k}^{\gamma_i}} \right] \right\} \delta_i \left\{ e^{-\lambda t_{i,c_i}^{\gamma_i}} \right\}^{1-\delta_i}, \]

where, \( \theta = (\beta^T, \alpha^T)^T, d_i^*(t_k) = t_{i,k}^{\gamma_i} - t_{i,k-1}^{\gamma_i} \) and typically \( n_k \) patients fail between fixed scheduled examination times \( t_{k-1}^* \) and \( t_k^* \) for \( k = 1, \ldots, m \) and \( n_c \) patients censored or withdrawn at specific times such that \( n_c + \sum_{k=1}^{m} n_k = n \). Here \( \delta_i = 1 \) denotes an event and \( \delta_i = 0 \) denotes a censored observation.

Data

We re-analysed the data from the Signal Tandmobiel study. This is a longitudinal prospective oral health study conducted in Flanders (Belgium) from 1996 to 2001. A cohort of 4430 randomly sampled schoolchildren (2297 boys and 2133 girls) who attended the first year of the primary school at the beginning of the study were examined annually (6 times). The response was time (yrs.) to the emergence of the permanent upper left first premolars. Two covariates were analysed gender: 0 = boy (52%), 1 = girl (48%) and dmf which stands for the status of the primary predecessor of this tooth: 0 (57%) if the primary predecessor was sound, 1 (43%) if it was decayed, missing due to caries or filled.

Results

The standard Weibull (SPR) model suggests that two covariates are important: gender and dmf. However, only dmf is important in the MPR models with and without frailty. Of the 3 models the MPR model with frailty fits best based on the AIC. And remarkably this model can also explain the apparent (?) gender difference. This finding raises interesting questions about the application of Occam’s Razor. For example is the SPR model in gender and dmf simpler than the MPR frailty model in dmf? One view is that the latter model is simpler since it...
explains the effect of gender. If so, the SPR model may be held to have produced a spurious result implicating gender and potentially misdirecting the search for causative factors. One need only consider the widespread use of SPR models in practice to realise the potential for misinformation.

References

Session 6: Thursday 15.40 - 17.00
Chair: Norma Bargary
Comparing catches from multiple fishing gears

Cóilín Minto

Marine and Freshwater Research Centre, Galway-May Institute of Technology, Ireland

Changes to the European Union Common Fisheries Policy aim to eliminate the practice of discarding unwanted or over-quota catches. Part of the solution to not discarding includes changing species and size-selectivity of fishing gears. Experimental fishing gear trials are designed to compare catches of candidate modified gears. Modifications include changes to mesh size and orientation, and the inclusion of separation or escape panels. Traditional catch comparison analysis focusses on comparing two gears. More recent development of quad-rig gears allows for the comparison of more than two gears simultaneously. Multiple gear trials require the development of suitable modelling frameworks.

Here, two trials conducted by BIM are analysed: 1) a comparison of four mesh sizes; and 2) a comparison of a species separation panel. For the mesh size comparison, a multinomial mixed effects model was developed incorporating: case-specific and choice-specific covariates, haul-level variability and sub-sampling offsets. For the separator panel trial data, a conditional model with penalised spline length effects and haul-level deviations was developed.

Maximum likelihood with Laplace approximation was used to estimate the parameters of both models, facilitated by automatic differentiation of the likelihood. Overall, the models fit the data relatively well capturing considerable between-haul variability and allowing inference on the population-level size and separation effects. Some issues of bias relative to the empirical raised proportions remain.
Dicing with Dengue

Trevor Bailey¹
¹Mathematics and Physical Sciences, College of Engineering, University of Exeter, UK

The transmission of many infectious diseases can be affected by weather and climate variability, particularly those spread by arthropod vectors such as malaria and dengue. Previous epidemiological studies have demonstrated statistically significant associations between the incidence of such infectious diseases and climate variability, and have highlighted the potential for developing climate-based early warning systems for associated epidemics. Dengue fever is now one of the most important emerging climate sensitive tropical diseases worldwide and Brazil experiences a higher morbidity from this disease than any other country.

This talk describes results from a Brazil-UK collaborative project that modelled the spatio-temporal variation in dengue fever risk in Brazil using climate and non-climate information, with a view to developing a national early warning system for dengue epidemics.

The Bayesian spatiotemporal hierarchical model developed allows derivation of posterior predictive distributions for disease risk at each spatial location for a given month or season. This allows probabilistic forecasts to be issued and forecast uncertainty to be quantified.
CÚRAM Connections Session:
Friday 9.00 - 11.00
Chair: John Hinde
Smooth mixed models: Combining the best of both worlds

Maria Durbán Reguera
1
1Department of Statistics, Universidad Carlos III de Madrid Avda, Madrid, Spain

Smoothing and mixed models are two of the most widely used techniques in applied statistics: smoothing relaxes the linearity constraints imposed by standard generalized linear regression models, and mixed or multi-level models deal with different sources of random variation in the data. The connection between nonparametric regression and mixed models was first established over 30 years, but it was not until the late 1990s before it became a “hot” research topic, partly due to the developments in mixed model software. Since then, more than a thousand articles have been published on this topic.

In this talk I shall try to show the audience how powerful this alliance can be. I will focus on two recent advances in smooth mixed models: the use adaptive multidimensional smoothing in neuroscience, and how to combine information collected at different spatial resolution in disease mapping.
Regularisation of generalised linear mixed models with autoregressive random effects

Jocelyn Chauvet\textsuperscript{1}, Xavier Bry\textsuperscript{1} and Catherine Trottier\textsuperscript{1,2}
\textsuperscript{1}Institut Montpelliérain Alexander Grothendieck, CNRS, Université de Montpellier, France
\textsuperscript{2}Université de Paul Valéry Montpellier III, Montpellier, France

We address two regularised versions of the EM algorithm for Generalised Linear Mixed Models (GLMM) for panel data. A random response \( y \) is modelled by a GLMM, using a set \( X \) of explanatory variables and two random effects. The first random effect models the dependence within individuals on which data is repeatedly collected while the second one embodies the serially correlated time–specific effect shared by all the individuals. Variables in \( X \) are many and redundant, so that regression demands regularisation. In this context, we first propose a ridge–penalised EM algorithm and then a supervised component–based regularised EM algorithm as an alternative.

Introduction

In the context of GLMM’s having a large number of redundant covariates, penalty–based approaches such as ridge (Eliot, \textit{et al.}, 2011) or lasso (Groll and Tutz, 2014) on the one hand and component–based approaches on the other (Chauvet, \textit{et al.}, (2016)) have already been highlighted. Focussing on situations where variable selection is inappropriate, we propose both ridge and Supervised Component (SC) estimation techniques for fitting a GLMM in the context of panel data with an autoregressive time–specific random effect.

General Principles

We first consider a Gaussian balanced panel data with \( q_1 \) individuals, each of them observed at the same \( q_2 \) time–points. With \( n = q_1 q_2 \) and \( q = q_1 + q_2 \), the model is

\[
y = X\beta + U\xi + \varepsilon,
\]

where \( y \in \mathbb{R}^n \) is the response vector, \( \beta \in \mathbb{R}^p \) and \( \xi \in \mathbb{R}^q \) respectively the fixed and random effects vectors (\( X \) and \( U \) being their associated design matrices) and \( \varepsilon \sim N_n(0, \sigma^2 I_n) \) the residuals.

We further assume that \( \xi = (\xi_1, \xi_2)^T \), where \( \xi_1 \sim N_{q_1}(0, \sigma^2_1 I_{q_1}) \) is the individual–specific random effect and \( \xi_2 \sim N_{q_2}(0, \sigma^2_2 A_2(\rho)) \), with \( A_2(\rho) = \left( \frac{\rho^{j-i}}{1 - \rho^2} \right)_{1 \leq i, j \leq q_2} \), the order–1 autoregressive time–specific random effect. For the rank–1 component, we set \( \beta = u\gamma \), with \( \|u\| = 1 \) and \( \gamma \in \mathbb{R} \) the regression parameter associated with component \( Xu \). Instead of subtracting a penalty term to the likelihood, we add a bonus term favouring the alignment of the component on the most interpretable directions in the explanatory subspace. For that, we take into account the structural relevance of component \( Xu \), defined as \( \phi(u) = \left( \sum_j (u^T N_j u)^{\ell} \right)^{\frac{1}{2}} \), where the \( N_j \)'s are s.d.p matrices encoding the type of structures of interest in \( X \) and \( \ell \geq 1 \) is a parameter tuning the locality of bundles to be considered. \( L \) denoting the complete log–likelihood and \( \theta = (\beta, \sigma^2_0, \sigma^2_1, \sigma^2_2, \rho) \), we present the current iteration of both ridge– and single SC–regularised EMs.

\textbf{ridge E–step:} Define \( Q_{\text{rid}}(\theta | \theta^{[t]}) = E_{\xi|y} \left[ L(\theta; y, \xi) - \lambda \|\beta\|_2^2 | \theta^{[t]} \right] \)

\textbf{SC E–step:} Define \( Q_{\text{SC}}(\theta | \theta^{[t]}) = E_{\xi|y} \left[ (1 - s) L(\theta; y, \xi) + s \log \phi(u) | \theta^{[t]} \right] \)

\textbf{M–step:} Set \( \theta^{[t+1]} = \arg \max_{\theta} Q_{\text{rid}}(\theta | \theta^{[t]}) \) or \( \theta^{[t+1]} = \arg \max_{\theta} Q_{\text{SC}}(\theta | \theta^{[t]}) \)
The trade–off parameter $s \in [0, 1]$ and parameter $\ell$ are tuned by cross–validation (as with the shrinkage parameter $\lambda \geq 0$ for ridge) and higher rank components are computed like the rank–1, subject to extra orthogonality constraints.

The extension to GLMMs is inspired by the Schall’s iterative scheme alternating linearisation of the model and parameters’ estimation. The idea is to keep the same linearisation step, but replace the usual estimation step with a “local” regularised EM.

**Conclusion**
Both methods were tested on simulated Gaussian and Poisson data and perform well in terms of estimation and prediction. But unlike ridge, SC gives access to interesting graphical diagnoses that reveal multidimensional predictive structures and greatly facilitate the interpretation of the model.

**References**


The impact of time-varying outliers on mixed effects models: A simulation study motivated by renal data

Laura Boyle¹, Lisa McFetridge¹ and Özgür Asar²
¹Mathematical Sciences Research Centre, Queen’s University Belfast, UK
²Department of Biostatistics and Medical Informatics, Acıbadem Mehmet Ali Aydınlar University, Istanbul, Turkey

This research explores the impact of longitudinal outliers on standard and robust linear mixed effects (LME) models, which respectively have normality and t-distributional assumptions for the random terms. Both renal data and simulated data are used to demonstrate that robust LME models obtain more accurate parameter estimates than standard LME models in the presence of both outlying individuals and individuals with outlying observations. The performance of robust LME models are investigated under a range of outlier patterns, including time-varying outliers, a concept not considered in the current literature.

Introduction

Medical studies increasingly involve the collection of longitudinal data, where observations are made repeatedly on a set of variables for each individual over time. Standard mixed effects (LME) models are frequently used to analyse such data, accounting for the influence of both fixed and random effects on a longitudinal response variable, Laird and Ware (1982). Previous research has shown that the LME normality assumptions are violated in the presence of longitudinal outliers, which can cause inaccurate and inefficient parameter estimation, Pinheiro et al. (2001). Robust LME models alternatively down-weigh the influence of outliers through t-distributional assumptions. However, current methods have the restrictive assumption that the impact of outliers is constant, unchanging over time, a possible reason why these methods are not more readily used in current literature, Pinheiro et al. (2001). This study aims to address this current gap in the robust LME literature, by investigating the impact of various stationary and time-varying outlier patterns on the estimates obtained from both standard and robust LME models.

Data

This research is motivated by a renal dataset collected over a period of ten years on haemodialysis patients in Northern Ireland. The dataset consists of 27,113 repeated measurements collected from 1320 patients, with haemoglobin (Hb) level measured as the response variable of interest. The renal dataset contains outliers in both the random effects (b-outliers) and the random error terms (e-outliers). Graphical analysis of the dataset suggests that the number of outliers, and the extent to which they are outlying, vary over time.

Methods

The impact of longitudinal outliers is investigated using both renal data and simulated data. Three outlier scenarios are considered in the simulation study. In each case, the accuracy and efficiency of the parameter estimates are reported.

- Many of the existing studies on robust LME models assume a user-specified value for the degrees of freedom (dof). This study aims to investigate the impact of misspecifying the dof for robust LME parameter estimation.
- A prevalent assumption in the literature is that the random effects and random error distributions have the same dof. The detrimental impact of this assumption is investigated.
- Currently, robust LME models assume that the impact of outliers does not vary over time. The effect of time-varying outlier patterns is investigated.

Discussion

Results show that the robust LME improves the efficiency of parameter estimates with application to both renal data and simulated data. Results will be presented from a simulation study investigating the performance of robust LME models under a variety of outlier settings.

References

Dynamic reference ranges for blood biomarkers: An application in prostate cancer

Davood Roshan1, John Ferguson2, Francis Sullivan3 and John Newell1
1School of Mathematics, Statistics and Applied Mathematics, National University of Ireland Galway, Ireland
2HRB Clinical Research Facility, National University of Ireland Galway, Ireland
3Prostate Cancer Institute, National University of Ireland Galway and Galway Clinic, Ireland

Biomarkers are characteristics that play an important role in understanding the normal or abnormal biological process of an individual. For example, Prostate Specific Antigen (PSA), which is a glycoprotein produced by the prostate gland, plays a key role in diagnosis of Prostate Cancer (PC) as it is usually increased in men with PC. However, other factors like age or urinary tract infection can cause an increase in the PSA levels. Therefore, age-related PSA reference ranges have been derived for detecting patients in the risk of PC. Prostate biopsy is the only diagnostic test to prove the existence of prostate cancer for a man whose age-related PSA value is beyond the defined range. However, such age-related reference ranges sometimes can lead to an unnecessary biopsy which can result in complications for patients and an increase in the cost of health care delivery. To overcome this drawback, we propose the use of personalized dynamic ranges for PSA monitoring using Bayesian methods and streaming EM algorithms. Such age-related reference ranges weight all subjects equally in a certain age group of the population, while dynamic ranges are tailored to the observations measured on one individual, which is more accurate in detecting meaningful changes in PSA trajectory.

Introduction

Normal ranges, derived from healthy individuals, are typically used when interpreting a set of biomarker test results. Although these static normal ranges provide valuable guidelines for assessing an individual’s health status when there is only one observation recorded for the individual’s biomarker, they have limited applicability, even when stratified by sex or gender, when there are at least two or three serial measurements on the individual’s biomarker. In this situation, the subject has started to generate his own reference values and the interest now is on the significant changes in these values. Therefore, dynamic reference ranges which take into account both between and within subject variabilities are needed for effective diagnosis. This is particularly useful when the within subject variability (WSV) is less than between subject variability (BSV).

Bayesian approaches are a natural choice to combine population information with individual measurements in order to construct personalised ranges. For example let \(y_{i1}, \ldots, y_{in}\) represent a sample of biomarker measurements for an individual over \(n\) follow up times. A new measurement \(y_{i,new}\) is being considered as abnormal if it falls outside of the \(\frac{\alpha}{2} \times 100\% \) and \((1 - \frac{\alpha}{2}) \times 100\%\) quantiles of \(P(Y_{i,new} \mid data)\); where ‘data’ includes both population and subject information excluding \(y_{i,new}\). Calculating the proposed dynamic reference range using a Bayesian approach is computationally intensive especially for large datasets. To overcome this challenge, a time-efficient approximate EM algorithm is adapted to generate the proposed dynamic ranges from the distribution of \(Y_{ij} \mid X_{i1}, \ldots, Y_{i(j-1)}\). This method only incorporates a single data point \(Y_{ij}\), the previous estimates of the model parameters and some summary statistics on the individual level, to update the model parameters.

Method

A simulation study was run to assess the performance of the three different approaches (Normal ranges, Bayesian and EM dynamic ranges). They were compared by their ability in detecting atypical values by measuring their area under the ROC curve (AUC). Different scenarios were considered by varying the population size (N), sample size (Ni), and the ratio of WSV to BSV \(r_1 = \frac{\text{var}(\sigma_i^2)}{\tau^2}, r_2 = \frac{E(\sigma_i^2)}{\tau^2}\); where \(\sigma_i^2\) and \(\tau^2\) representing the WSV and BSV, respectively.

Results

The results suggest that the Bayesian approach generally preformed better than the other two. However, when \(r_2\) is increasing, all three approaches resulted in a similar performance suggesting the normal range is as good as the other two approaches if the WSV is larger than BSV (Figure 1(a)). Moreover, when both population
and sample size increases (Figure 1(b)), the Bayesian approach and the EM algorithm have relatively the same performance where both outperform static normal ranges. Additionally, the proposed models were implemented on a series of PSA values for an individual diagnosed with PC. Figure 2 shows that both the Bayesian method and EM algorithm are detecting the cancer at the sixth follow up time while normal reference range was unable to detect the patient with cancer as all of the PSA test results are within the normal range.

![Figure 1: The AUC distribution of the three approaches (Reference: normal ranges; GIBBS: Bayesian approach; EM: streaming EM algorithm) for detection of outliers for different combination of (a) \( r_1 \& r_2 \), and (b) population size \( (N) \& sample size \( (N_i) \).](image)

![Figure 2: Personalized dynamic ranges for a patient diagnosed with PC based on the three different methods](image)

**Acknowledgments**

This research is supported by the Prostate Cancer Institute, NUIG and the Irish Research Council. Furthermore, I would like to thank the Galway Clinic for sharing their anonymized data.

**References**


Dynamic factor analysis to investigate coherency in growth patterns across various fish species in the Celtic Sea

Olga Lyashevskaya, Cóilín Minto and Deirdre Brophy

Marine and Freshwater Research Centre, Galway-Mayo Institute of Technology, Galway, Ireland

Problem statement: Do fish species display synchronous changes in growth and are these changes indicative of a common response to broad scale environmental drivers? We address this question through a dynamic factor analysis with a varying number of trends. The best fitting model indicated two underlying trends in weight-at-age time series of various fish species.

Introduction

Fish growth rates fluctuate in response to environmental conditions (primarily food abundance and temperature) and are influenced by intra and interspecific competition and by size-selective effects of fishing. Modelling temporal growth patterns across multiple fish species can help to disentangle the various drivers of change and to detect synchronous responses to broad scale ecosystem change.

Methods

We use dynamic factor analysis to determine the extent to which weight-at-age time series display synchronous changes across multiple pelagic (herring, mackerel, blue-whiting) and demersal (cod, haddock, plaice, sole) fish species in the Northeast Atlantic (ISEC, 2017). Generalised additive mixed models are fit to the data for each species with age modelled as a spline and year-of-birth and year-of-capture as random effects. Extracted random effects and their variance-covariance matrices ($R$) for each year are then modelled using dynamic factor analysis to identify common trends (Harvey, 1989).

\[
\begin{align*}
x_t &= x_{t+1} + w_t & \text{where } w_t &\sim MVN(0, Q) \\
y_t &= Zx_t + a + v_t & \text{where } v_t &\sim MVN(0, R)
\end{align*}
\]

Where observations $y$ are modelled as a linear combination of hidden trends $x$ and factor loadings $Z$ plus some offsets $a$.

Results

The AICc values for year-of-birth and year-of-capture models were lowest with 2 trends (Table 1).

<table>
<thead>
<tr>
<th>random effect</th>
<th>2 trends</th>
<th>3 trends</th>
<th>4 trends</th>
</tr>
</thead>
<tbody>
<tr>
<td>year-of-birth</td>
<td>-995.78</td>
<td>-977.55</td>
<td>-960.79</td>
</tr>
<tr>
<td>year-of-capture</td>
<td>-932.05</td>
<td>-917.07</td>
<td>-900.35</td>
</tr>
</tbody>
</table>

The loadings and trends were calculated using varimax rotation (Figure 1). The first common trend shows an increase followed by a sharp decrease in the late 1970s. The second common term seem to alternate between increase and decrease: after a decrease it increases again in the late 1970s, followed by a decrease after 2000. The factor loading indicate that the first common trend is mostly related to herring, whereas the second is related to plaice (areas f and g).
Discussion and Conclusions

Relationships between the resultant trends and various potential drivers will be examined. Drivers of change include: estimates of recruitment from the annual stock assessments to account for density-dependent changes in size at the species level. The influence of density-independent environmental effects (e.g. temperature, zooplankton abundance, wind patterns and climatic indices) are also to be examined. The possible impacts of the observed trends on productivity and ecosystem resilience will be discussed.

References


Session 8: Friday 11.40 - 13.00
Chair: Kathleen O’Sullivan
Modelling uncertainty and vagueness within recommender systems via nonparametric predictive inference

Angela McCourt¹ and Brett Houlding¹
¹School of Computer Science and Statistics, Trinity College Dublin, Ireland

Learning preferences for decision-making processes has been an area of substantial research in recent years, especially given the introduction of Recommender systems (RSs). RSs help in decision-making processes by recommending items of interest and filtering out undesired items, they need to learn preferences by extracting information about both the user and the item. Such systems often give point-wise recommendations to its users. In this paper, we propose incorporating uncertainty and vagueness into RSs via Nonparametric Predictive Intervals (NPIs).

Introduction
Nonparametric methods are widely used in statistics, they allow more flexibility than their parametric counterparts as they require fewer assumptions to be met. Nonparametric Predictive intervals are based on the assumption \( A_{(n)} \) which was proposed by Hill (1968). This assumption allows for prediction in the case of extremely vague prior knowledge. Therefore, the assumption \( A_{(n)} \) is a distribution free, post-data assumption related to finite exchangeability (DeFinetti, 1937).

Uncertainty Interval
In order to develop what we term an Uncertainty Interval (UI), it was proposed to incorporate information that would not normally be utilised. It was noted that pairwise correlations generated missing values (MV) given very sparse matrices. Therefore it was decided to utilise the correlations that could be calculated in order to estimate a correlation coefficient of a MV.

If a MV is returned for a correlation calculation between item \( a \) and item \( b \), the correlations of item \( a \) and item \( i \) and the correlations between item \( b \) and item \( i \), where item \( i \) represents all other items in the matrix, would provide information about how item \( a \) and item \( b \) are related, see Figure 1. Hence, the UI approach is also a pairwise comparison method. The matrix to be considered is one which is exchangeable and positively correlated.

![Figure 1: Representation of pairwise comparisons](image)

Raw variables are simulated, \( R_i \sim MVN(\mu, \Sigma) \), such that the vector \( \mu \) contains only zeros and the covariance matrix \( \Sigma \) contain values between .95 and .99 to ensure high correlation. The raw variables generated formed a matrix which has no missing variables and is positive definite, it contains \( 1000 \times 1000 \) entries and is denoted \( C_{\text{complete}} \). Four UI approaches (\( UI_{\text{Untrans}}, UI_{\text{Abs}}, UI_{\text{Sq}} \) and \( UI_{\text{Rt}} \)) are explored and evaluated on a sparse matrix \( C_{5\%} \), which has 95% MVs. A pairwise correlation is calculated and the results are shown in Figure 2.
The proposed method is heuristic in nature and allows for the population of a correlation matrix that contains MVs by utilising information that is not ordinarily included, which is a common approach by RSs dealing with a cold-start problem.

Figure 2: Application of $UI_{\text{Untrans}}$, $UI_{\text{Abs}}$, $UI_{\text{Sq}}$, and $UI_{\text{Rt}}$ to estimate $\rho_{\text{complete}}(1, 2)$

References

Penalised multi-parameter regression survival modelling

Fatima-Zahra Jaouimaa1, Eunyoung Park2, Il Do Ha2 and Kevin Burke1
1Department of Mathematics and Statistics, University of Limerick, Ireland
2Department of Statistics, Pukyong National University, Busan, Korea

Introduction
Multi-parameter regression (MPR) modelling refers to the approach whereby covariates are allowed to enter the model through multiple distributional parameters simultaneously - in contrast to the standard approaches where covariates enter through a single parameter (e.g., a location parameter). Burke and MacKenzie (2017) explored the use of MPR models in a survival setting, and demonstrated the flexibility afforded by jointly modelling the scale and shape parameters of a Weibull distribution on covariates.

While Burke and MacKenzie (2017) proposed an information criterion based stagewise variable selection procedure adapted for MPR models, it is well known that such selection procedures can be unstable and computationally intensive; more modern variable selection procedures are based on penalised likelihood. Methods such as the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), and adaptive lasso (ALASSO) (Zou, 2006) are used to simultaneously select variables and estimate their regression coefficients. Therefore, in this work, we develop penalised multi-parameter regression methods and investigate their associated performance; as an example, we consider the Weibull model.

Methods
The Weibull distribution hazard function is given by \( \lambda(t) = \lambda \gamma t^{\gamma-1} \), where \( \lambda > 0 \) is the scale parameter and \( \gamma > 0 \) is the shape parameter. The Weibull MPR model is obtained by letting both distributional parameters depend on covariates as follows: \( \log(\lambda) = x^T \beta \) and \( \log(\gamma) = z^T \alpha \), where \( x = (1, x_1, \ldots, x_p)^T \) and \( z = (1, z_1, \ldots, z_q)^T \) are scale and shape covariate vectors which may or may not have covariates in common, and \( \beta = (\beta_0, \beta_1, \ldots, \beta_p)^T \) and \( \alpha = (\alpha_0, \alpha_1, \ldots, \alpha_q)^T \) are the corresponding regression coefficients.

<table>
<thead>
<tr>
<th>Table 1: Treatment model coefficients and standard errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>Surgery</td>
</tr>
<tr>
<td>Chemo</td>
</tr>
<tr>
<td>Radio</td>
</tr>
<tr>
<td>C+R</td>
</tr>
</tbody>
</table>

Shape

<table>
<thead>
<tr>
<th>No Penalty</th>
<th>LASSO</th>
<th>SCAD</th>
<th>A-LASSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.19 (0.04)</td>
<td>-0.19 (0.04)</td>
<td>-0.19 (0.04)</td>
</tr>
<tr>
<td>Surgery</td>
<td>0.59 (0.21)</td>
<td>0.26 (0.22)</td>
<td>0.59 (0.20)</td>
</tr>
<tr>
<td>Chemo</td>
<td>0.07 (0.14)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
</tr>
<tr>
<td>Radio</td>
<td>0.34 (0.07)</td>
<td>0.31 (0.07)</td>
<td>0.33 (0.07)</td>
</tr>
<tr>
<td>C+R</td>
<td>0.97 (0.16)</td>
<td>0.70 (0.18)</td>
<td>0.97 (0.16)</td>
</tr>
</tbody>
</table>

Parameter estimation within the unpenalised MPR model can be carried out in a standard fashion by maximising the associated log-likelihood function, \( \ell(\theta) \), where \( \theta = (\beta^T, \alpha^T)^T \). On the other hand, penalised MPR estimation can be developed on the basis of maximising a penalised log-likelihood, \( \ell_p(\theta) = \ell(\theta) - n \sum_j p_{\lambda_j}(|\theta_j|) \) where \( p_{\lambda_j}(\cdot) \) is a penalty function. The presence of two separate regression vectors in this MPR setting, \( \beta \) and \( \alpha \) (in contrast to standard settings where there is only a \( \beta \) vector), necessitates the investigation of various choices that can be made, i.e., the functional form of \( p_{\lambda_j}(\cdot) \), the number of and relative magnitude of the associated tuning parameters, \( \lambda_j \), and the selection of such tuning parameters.
Results and Discussion
Early results of simulation studies have been very promising, and provide important insights into the requirements for penalised MPR methods. Some of the methods explored to date are displayed in Table 1, applied to a lung cancer dataset which appeared in Burke and MacKenzie (2017). More detailed discussion of these results will appear in the presentation of our work.

References


Prognostic feature selection from complementary FDG-PET quantitation analyses in lung cancer using survival data

Eric Wolsztynski\(^1\), Aidan Marnane\(^3\), Janet O’Sullivan\(^1\), Nicola Hughes\(^2\), Tian Mou\(^1\), Peter Murphy\(^3\), Finbarr O’Sullivan\(^1\) and Kevin O’Regan\(^4\)

\(^1\)School of Mathematical Sciences, University College Cork, Ireland
\(^2\)Royal College of Surgeons in Ireland, Dublin, Ireland
\(^3\)PET/CT Unit (Alliance Medical), Cork University Hospital, Ireland
\(^4\)Department of Radiology, Cork University Hospital, Ireland

Positron Emission Tomography (PET) imaging is used routinely for multivariate characterization of tumour and patient risk, in lung and other cancers. We consider two methodologies for PET data analysis. One summarises the structure of volumetric uptake via spatial modelling, the other evaluates various aspects of image texture. These approaches yield a large number of potential predictors.

We compare feature selections with respect to patient censoring status and survival time information for a number of learning techniques. Significant multivariate prognostic models are obtained that combine both quantization methodologies, suggesting that spatial modelling merits consideration for modern PET-based assessment.

Introduction

PET imaging analysis assesses the uptake of a radioactive tracer (predominantly 18F-fluorodeoxyglucose, or FDG) to provide multivariate quantitative summaries of the metabolic profile of a tumour for diagnosis, prognosis and therapy planning.

An emerging methodology called radiomics combines crude morphologic descriptors and image processing approaches to texture analysis, motivated by the importance of intratumoral heterogeneity in clinical assessment for a number of cancers (Vallieres, et al., 2017). Among the questions raised by this high-throughput approach is that of feature selection.

Our group has recently examined the complementarity of texture analysis and an alternative, model-based methodology for the characterization of the volumetric uptake distribution in sarcoma and lung cancers (Wolsztynski, et al., 2017). Here we examine how these additional descriptors compete or cohabit with current radiomics variables, in terms of a number of statistical learning techniques applied to feature selection.

Methods

A first feature frame consists of routine variables (patient age, gender, tumour volume and stage). A second feature frame comprises of structural measures derived from spatial modelling. These include goodness-of-fit of the uptake distribution from a reference ellipsoidal pattern and the assessment of metabolic gradients (Wolsztynski, et al., 2017), which provides local rates of metabolic change within the volume. The third feature frame contains morphologic (including ellipsoidal axis lengths, flatness and elongation), image intensity and image texture features computed conventionally (Vallieres, et al., 2017). Feature selection is performed via stepwise selection, LASSO, random forests and survival random forests (Friedman, et al., 2009), and prognostic validation via Cox proportional hazards analysis.

Results

Feature selections vary with the selection strategy, however with significant overlap, indicating that some features are predominant (including model-derived heterogeneity) and capture essential and distinct regions of the information space. Cross-validated feature selections combine structural and textural variables in over 90% of experiments, suggesting complementarity of the structural and textural frames. Prognostic significance of the selected feature sets is verified for this lung cohort.
Conclusion
Feature selection based on patient survival data is viable for PET-derived prognostic modelling and provides insight on the nature of feature spaces that best capture lung tumour characteristics. Our preliminary results suggest that spatial modelling of FDG-PET uptake distribution can complement conventional radiomics analyses.

References


Identifiability in Coxian phase type distributions - An alternative formulation

Jean Rizk, Kevin Burke and Cathal Walsh

Mathematics Applications Consortium for Science and Industry, Department of Mathematics and Statistics, University of Limerick, Ireland

Coxian phase-type (CPH) distributions have been used in a wide range of stochastic modelling applications such as queueing modelling and survival analysis. However, fitting a CPH distribution can be problematic. Due to the non-unique representation of phase-type distributions and the dependence on initial values when estimating the parameters, the problem that often emerges is the inability to find unique parameter estimates for the data. In this paper we wish to re-visit the issue of non-uniqueness in the Coxian representation and we refer to it as a “Label Switching” problem.

Introduction

In recent decades, the use of Coxian phase-type (CPH) distributions has become increasingly more popular for modelling patient length of stay in healthcare.

An \( n \)-phase CPH distribution describes duration until absorption in terms of a continuous time Markov process consisting of a sequence of \( n \) latent phases. The process starts in the first phase with a probability of exiting from any phase. Despite convergence, when multiple fits to data are performed they output different permutations of parameters that correspond to the same maximum likelihood (Marshall and Zenga, 2012). This is due to the non-uniqueness of representation of phase-type distributions. This problem had been examined in the past using matrix-analytic methods (O’Cinneide, 1989).

In this work, we study the non-uniqueness problem using a parametric family of finite mixture densities. We present the CPH distribution not in its matrix form but as a mixture of Hypoexponential distributions. In mixture models, this non-uniqueness problem is known as “Label Switching” (Walsh, 2006). This is caused by symmetry in the likelihood of the model parameters. We discuss two approaches to overcome the label switching problem. The first approach is to impose an identifiability constraint on the parameter space. The second approach is to specify a loss function that quantifies the loss when choosing a specific permutation of parameters (Stephen, 1983).

Methods

We write the CPH distribution as a mixture of Hypoexponential distributions. The log-likelihood function is defined as follows:

\[
    l(\Theta | x) = \sum_i \log \left( \sum_j p_j f_j(x_i | \Lambda_j) \right),
\]

where \( f_j(x | \Lambda_j) \) is a hypoexponential distribution with parameters \( \Lambda_j \). \( \Theta = (p_i) \cup (\Lambda_i) \) and \( p_i \) are the mixture components with \( \sum_i p_i = 1 \). An extensive simulation study has been carried out to investigate the label switching issue and tested on real data. The model is fit using the Optimization Toolbox in MATLAB.

Conclusion

We identify the non-uniqueness of representation of the CPH distribution by modelling it as a mixture of densities. A frequent response to the label switching problem is to remove the symmetry by using identifiability constraints. Alternatively, we can deal with the problem by specifying a loss function. This may identify the best permutation of parameters that truly reflect the underlying model which gave rise to the data.

References


Index

Abedin, Jaynal, 72
Al-Ghamdi, Asmaa, 25
Alkema, Leontime, 17
Alvarez-Iglesias, Alberto, 27
Asar, Özgür, 92

Bailey, Trevor, 7, 87
Barr, Thomas, 81
Bennett, Kathleen, 49
Boland, Aidan, 29
Bonat, Wagner, 77
Boyle, Laura, 92
Brophy, Caroline, 66
Brophy, Deirdre, 96
Brown, Emma, 79
Brown, Win, 17
Browne, Leonard, 11
Bry, Xavier, 90
Burke, Kevin, 83, 101, 105
Byrne, Andrew, 79

Cahill, Niamh, 17
Cairns, Karen, 53
Carbery, Caoimhe, 31
Cauchi, Michael, 33
Chapman, Jamie-Leigh, 64
Charles, Bouveyron, 70
Chauvet, Jocelyn, 90
Cordeiro, Francesca, 38, 55
Crotty, Evan, 81

Das, Kishor, 35
Davis, Benjamin, 38, 55
De Groef, Lies, 38
Demétrio, Clarice, 44, 77
Donnelly, Conor, 40
Donova, John, 61
Durban, Maria Reguera, 8, 14, 89
Durkan, Garrett, 47

Eckley, Idris, 64
Fallah, Lida, 42
Ferguson, John, 27, 93
Fitzgerald, Tony, 11
Foley, Robert, 47
Fop, Michael, 70

Galvin, David, 47
Godoy, Wesley, 44
Grolemund, Garrett, 5, 69
Guo, Li, 55

Ha, Il Do, 101
Harrison, Drew, 81
Hayes, Kevin, 67, 81
Hederman, Lucy, 59
Hennessy, Deirdre, 66
Hinde, John, 42, 44, 77
Holmes, Jonathan, 81
Houlding, Brett, 25, 59, 99
Hughes, Nicola, 103

Jalali, Amirhossein, 47
Jaouimaa, Angela, 101

Kearney, Patricia, 11
Kelly, Gabrielle, 22
Killick, Rebecca, 64
Kinsella, Anthony, 67

Lamrock, Felicity, 16
Lara, Idemauro, 44
Lawlor, Debbie, 14
Leacy, Finbarr, 49
Leahy, Joy, 12
Li, Xia, 11
Little, Mark, 59
Lundon, Dara, 47
Lyashevska, Olga, 96
Lynch, Thomas, 47, 57

MacDonald-Wallis, Corrie, 14
MacKenzie, Gilbert, 83
Macpherson, Russell, 61
Madden, Jamie, 11, 49
Marnane, Aidan, 103
Marshall, Adele, 31, 40, 53, 79
Mashinchi, Mona Isazad, 50
Mattei, Pierre-Alexandre, 70
Maweni, Robert, 47
May, Margaret, 14
McCague, Michael, 52
McCourt, Angela, 99
McCurdy, Meabh, 53
McDonnell, Jack, 66
McFetridge, Lisa, 40, 92
McGillycuddy, Maevé, 27
McKenna, Thomas, 66
McCullagh, Laura, 16
Meehan, Alan, 59
Metcalfe, Chris, 14
Minto, Cóllín, 6, 86, 96
Mitchell, Hannah, 55, 79
Moghaddam, Shirin, 57
Moons, Lieve, 38
Moral, Rafael, 44, 77
Mou, Tian, 103
Murphy, Keefe, 47, 57, 74
Murphy, Lisa, 57
Murphy, Peter, 103
Murphy, Thomas Brendan, 47, 57, 70, 74

Newell, John, 35, 72, 93

O’Brien, Frank, 47
O’Brien, Kevin, 67
O’Gorman, Laura, 57
O’Malley, Kieran, 47, 57
O’Regan, Kevin, 103
O’Sullivan, Declan, 59
O’Sullivan, Finbarr, 103
O’Sullivan, Janet, 103
Oliveira, Thiago, 44

Park, Eunyoung, 101
Peng, Defen, 83
Power, Richard, 47, 57

Rebholz-Schuhmann, Dietrich, 50
Reddy, Brian, 59
Rizk, Jean, 105
Rosenbaum, Paul, 3, 10
Roshan, Daawood, 50, 93

Salinas-Navarroa, Manual, 38
Scarrott, Carl, 4, 63
Simpkin, Andrew, 14
Sonneveldt, Emily, 17
Stover, John, 17
Sullivan, Francis, 50, 93

Tilling, Kate, 14
Tilson, Lesley, 16
Tormey, David, 61
Trottier, Catherine, 90

Unnikrishnan, Saritha, 61

Walsh, Cathal, 12, 105
Watson, William, 47, 57

Wei, Chuchu, 17
Weinberger, Michelle, 17
Wolsztynski, Eric, 103
Woods, Roger, 31
Yurkonis, Kathryn, 66
Zgaga, Lina, 49
Zocchi, Silvio, 44
<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Speaker</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00-9:40</td>
<td>Carl Scarrott</td>
<td>(Keynote)</td>
<td>Statistical modelling of air pollution and the challenges of communication</td>
</tr>
<tr>
<td>9:00-9:40</td>
<td>Maria Durbán</td>
<td>(Keynote)</td>
<td>Smooth mixed models: Combining the best of both worlds</td>
</tr>
<tr>
<td>9:40-10:00</td>
<td>Jamie Leigh</td>
<td></td>
<td>Changepoints to improve forecasts</td>
</tr>
<tr>
<td>9:40-10:00</td>
<td>Jocelyn Chauvet</td>
<td></td>
<td>Regularisation of generalised linear mixed models with autoregressive random effects</td>
</tr>
<tr>
<td>10.00-10.20</td>
<td>Jack McDonnell</td>
<td></td>
<td>Assessing spatial interactions among species in a grassland experiment</td>
</tr>
<tr>
<td>10.00-10.20</td>
<td>Laura Boyle</td>
<td></td>
<td>The impact of time-varying outliers on mixed effects models: A simulation study motivated by renal data</td>
</tr>
<tr>
<td>10.20-10.40</td>
<td>Anthony Kinsella</td>
<td></td>
<td>Variability of variance estimates: The parametric bootstrap approach</td>
</tr>
<tr>
<td>10.20-10.40</td>
<td>Davood Roshansangachin</td>
<td></td>
<td>Dynamic reference ranges for blood biomarkers: An application in prostate cancer</td>
</tr>
<tr>
<td>10:40-11:00</td>
<td>Olga Lyashevska</td>
<td></td>
<td>Dynamic factor analysis to investigate coherency in growth patterns across various fish species in the Celtic Sea</td>
</tr>
<tr>
<td>11:00-11:40</td>
<td>Garrett Grolemund</td>
<td>(Keynote)</td>
<td>Teaching data science</td>
</tr>
<tr>
<td>11:40-12:00</td>
<td>Michael Fop</td>
<td></td>
<td>Unobserved classes and extra variables in high-dimensional discriminant analysis</td>
</tr>
<tr>
<td>11.40-12:00</td>
<td>Angela McCourt</td>
<td></td>
<td>Modelling uncertainty and vagueness within recommender systems via nonparametric predictive inference</td>
</tr>
<tr>
<td>12.00-12:20</td>
<td>Jaynal Abedin</td>
<td></td>
<td>Text mining aided literature review: A case study in sports injuries</td>
</tr>
<tr>
<td>12.00-12:20</td>
<td>Fatima-Zahra Jaouimaa</td>
<td></td>
<td>Penalised multi-parameter regression survival modelling</td>
</tr>
<tr>
<td>12.20-12:40</td>
<td>Keefe Murphy</td>
<td></td>
<td>Gaussian parsimonious clustering models with covariates</td>
</tr>
<tr>
<td>12.20-12:40</td>
<td>Eric Wolsztynski</td>
<td></td>
<td>Prognostic feature selection from complementary FDG-PET quantitation analyses in lung cancer using survival data</td>
</tr>
<tr>
<td>5:00-5:40</td>
<td>Jean Rizk</td>
<td></td>
<td>Identifiability in Coxian phase type distributions – An alternative formulation</td>
</tr>
<tr>
<td>5:00-5:40</td>
<td>Paul Rosenbaum</td>
<td>(Keynote)</td>
<td>Addressing bias from unmeasured dispositions in observational studies</td>
</tr>
<tr>
<td>5:40-6:00</td>
<td>Jamie Madden</td>
<td></td>
<td>Morning surge in blood pressure using a random-effects multiple-component cosinor model</td>
</tr>
<tr>
<td>5:40-6:00</td>
<td>Emma Brown</td>
<td></td>
<td>Accurately estimating parameters for a bovine tuberculosis epidemic model in Northern Ireland</td>
</tr>
<tr>
<td>6:00-6:20</td>
<td>Joy Leahy</td>
<td></td>
<td>MAIC-ing the most of trials?: A Bayesian exploration of matching adjusted indirect comparison</td>
</tr>
<tr>
<td>6:00-6:20</td>
<td>Kevin Hayes</td>
<td></td>
<td>Statistical algorithms and issues arising during the development of a hand force-plate sensor system for false start detection in competitive sprinting</td>
</tr>
<tr>
<td>6:40-7:00</td>
<td>Andrew Simpkin</td>
<td></td>
<td>Derivative estimation for longitudinal data analysis: Examining features of blood pressure measured repeatedly during pregnancy</td>
</tr>
<tr>
<td>6:40-7:00</td>
<td>Gilbert MacKenzie</td>
<td></td>
<td>Interval censored MPR modelling</td>
</tr>
</tbody>
</table>

**Tuesday**

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Speaker</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:40-11:00</td>
<td>Felicity Lamrock</td>
<td></td>
<td>Using probabilistic sensitivity analysis in budget impact models for reimbursement recommendations in Ireland</td>
</tr>
<tr>
<td>11:00-11:40</td>
<td>Coilin Minto</td>
<td>(Keynote)</td>
<td>Comparing catches from multiple fishing gears</td>
</tr>
<tr>
<td>11:40-12:00</td>
<td>Niamh Cahill</td>
<td></td>
<td>Trends in modern contraceptive prevalence, unmet need for and demand satisfied with modern methods in the focus countries of the Family Planning 2020 Initiative</td>
</tr>
<tr>
<td>12:00-12:40</td>
<td>Trevor Bailey</td>
<td>(Keynote)</td>
<td>Dicing with Dengue</td>
</tr>
<tr>
<td>12:40-13:00</td>
<td>Gabrielle Kelly</td>
<td></td>
<td>Goodness-of-fit statistics and the birthday effect</td>
</tr>
</tbody>
</table>