Statistics Seminars 2017

Mathematical Sciences, School of Science, Loughborough University

Thursday 19 October 15:00-16:00

Room: N112

Speaker: Dr Theodore Kypraios (University of Nottingham)

Title: Recent Developments in Identifying Transmission Routes of Healthcare Associated Infections using Whole Genome Sequence Data.

Abstract:

Healthcare-associated infections (HCAIs) remain a problem worldwide and can cause severe illness and death. It is estimated that 5-10% of acute-care patients are affected by nosocomial infections in developed countries, with higher levels in developing countries.

Statistical modelling has played a significant role in increasing understanding of HCAI transmission dynamics. For instance, many studies have investigated the dynamics of MRSA transmission in hospitals, estimating transmission rates and the effectiveness of various infection control measures. However, uncertainty about the true routes of transmission remains and that is reflected on the uncertainty of parameters governing transmission.

Until recently, the collection of whole genome sequence (WGS) data for bacterial organisms has been prohibitively complex and expensive. However, technological advances and falling costs mean that DNA sequencing is becoming feasible on a larger scale.

In this talk we first describe how to construct statistical models which incorporate WGS data with regular HCAIs surveillance data (admission/discharge dates etc) to describe the pathogen's transmission dynamics in a hospital ward. Then, we show how one can fit such models to data within a Bayesian framework accounting for unobserved colonisation times and imperfect screening sensitivity using efficient Markov Chain Monte Carlo algorithms. Finally, we illustrate the proposed methodology using MRSA surveillance data collected from a hospital in North-East Thailand.

Thursday 02 November 15:00-16:00

Room: SMB002

Speaker: Dr Natsuhiko Kumasaka (Sanger Institute)

Title: Chromatin configuration QTL mapping using ATAC-seq

Abstract:

Gene regulatory elements frequently occur in regions of open or accessible chromatin. Mapping quantitative trait loci (QTLs) that associate with differences in chromatin openness (chromatin accessibility QTLs, caQTLs) is a powerful method for identifying likely causal regulatory variants. Here, we introduce a novel approach for mapping and interpreting caQTLs that uses a Hidden Markov Model and apply this method to a data set of 100 ATAC-seq samples derived from lymphoblastoid cell lines. Our model incorporates a number of novel features, including the ability to infer ensembles of nucleosomes, linkers and transcription factors (TFs) that are perturbed by specific

genetic variants. By explicitly incorporating information about TF binding affinities, our method can help refining predictions of the likely causal variants that drive differences in chromatin landscapes. We also show that our approach enables detection of a novel class of caQTLs that alter the shape of chromatin accessibility peaks, rather than peak height, which we term chromatin configuration QTLs. We also illustrate that there is abundant additional information on chromatin state that can be learned from utilizing information from paired end reads.

Thursday 23 November 15:00-16:00

Room: SCH013

Speaker: Dr Sara Wade (University of Warwick)

Title: Adaptive truncation of a Bayesian nonparametric multivariate regression model for a study of fertility and partnership patterns of Colombian women

Abstract:

We propose a flexible Bayesian nonparametric multivariate regression model, which can capture nonlinear regression functions and the presence of non-normal errors, such as heavy tails or multi-modality. The infinite mixture model has interpretable covariate-dependent weights constructed through normalization, allowing for combinations of both discrete and continuous covariates, and extends the model developed in Antoniano-Villalobos et al. 2014 for a multivariate and non-continuous response. The infinite number of components and intractable normalizing constant pose computational difficulties, which are overcome through an adaptive truncation algorithm (Griffin, 2014). The algorithm combines adaptive Metropolis-Hastings with sequential Monte Carlo to create a sequence of truncated posteriors and automatically determines the level of truncation. The model and algorithm are applied to a lifestyle study on Colombian women, which aims to understand the relationship between some focal life events (e.g. age at first sexual intercourse, relationship, child, presence in the labour market) and various baseline factors, such as year of birth, region of birth, and indicators of well-being in the family of origin. Regression function and conditional density estimates are presented, along with an analysis of the implied covariate-dependent clustering.

Thursday 07 December 15:00-16:00

Room: SCH013

Speaker: Richard Farley (Loughborough University)

Title: Evaluation of statistical methods and sample size requirements for untargeted metabolomics studies

Abstract:

Abstract: Over the past years there has been substantial interest in using quantitative metabolomics data for untargeted biomarker discovery. From a statistical perspective, this is a difficult problem because quantitative metabolomics data is high dimensional and prone to errors. There are many statistical techniques currently used for biomarker discovery in such data, including OPLS, penalised regression methods and multiple t-tests. There is no consensus about which methods perform best

in which situations, or about sample sizes needed for such untargeted studies to reduce the probability of identifying spurious markers.

The purpose of this simulation study is to investigate the performance of the most popular statistical methods for untargeted biomarker discovery using data that mimics compact mass spectrometry data from volatile organic compounds from bacterial cultures. Starting with data collected from an actual study, we build synthetic datasets of 1000 "control" samples and 1000 "treatment" samples, in which a set of channels has been augmented with a "biomarker signal". The size of this signal and the size of the noise are varied between datasets and the dimensional reduction methods are compared under these varying conditions and for sample sizes ranging from 10 to 1000.

We evaluated the above statistical methods under different conditions for their performance in terms of predictive power, ability to identify biomarkers channels and ability to avoid selecting non-biomarker channels. We present results about the strengths and weaknesses of the three methods alone and in combination and make suggestions on their suitability and appropriate sample sizes in different situations.

Seminar name: Statistics Seminars

Year: 2017/18

Semester 2

Date	Time/place	Speaker	Title
22nd March 2018	15:00-16:00/N.1.12 (Haslegrave Building)	Safa Elsheikh (University of Brighton)	Modelling Diffusion Directions of the Corpus Callosum using Von Mises-Fisher Distribution
19th April 2018	15:00-16:00/SCH.0.13 (Schofield Building)	Dr Ralf Weber (University of Birmingham)	Towards more complete annotation of metabolomes: analytical and computational approaches
10th May 2018	15:00-16:00/SCH.0.13 (Schofield Building)	Dr Dov Stekel (University of Nottingham)	Modelling antimicrobial resistance
24th May 2018	15:00-16:00/N.1.12 (Haslegrave Building)	Professor Kirsten Barrett (University of Leicester)	Remote sensing of disturbance-recovery cycles using time series data
31st May 2018	15:00-16:00/N.1.12 (Haslegrave Building)	Lei Ye (Loughborough University)	DTI Fiber Tracking