

EXPERIMENTAL DESIGN AND BIG DATA
MAY 08 2015, THE UNIVERSITY OF WARWICK
WORKSHOP PROGRAM

REGISTRATION (09:45 - 10:00). ROOM MS.03, ZEEMAN BUILDING

SESSION 1 (10:00 - 11:30). ROOM MS.03, ZEEMAN BUILDING

10:00 **David Rossell** (University of Warwick, UK)

Welcome address

10:10 **Matthias Seeger** (Amazon, Germany)

Large scale variational Bayesian inference and sequential experimental design for signal acquisition optimization

Abstract: I will give a brief introduction to sequential Bayesian experimental design, in the sense of greedy maximization of information gain. I will motivate the challenges this program places on approximate Bayesian inference, if it is to be used for high-dimensional signal acquisition optimization. I will outline a framework for variational Bayesian inference in large sparse linear models, with which BED can be implemented for such scenarios.

11:00 **Tristan Henderson** (University College of Saint Andrews, UK)

Reliable, reproducible and responsible data collection from online social networks

Abstract: The use of online social networks (OSNs) such as Facebook and Twitter for research has exploded in recent years, as researchers take advantage of access to the hundreds of millions of users of these sites to understand social dynamics, health, mobility, psychology and more. But there are myriad challenges in collecting the appropriate data from OSNs for an experiment.

In this talk we will discuss three of these challenges. First, we will look at differences in passive collection of OSN data (e.g., crawling Facebook) versus actively requesting information from OSN users. Secondly, we will examine the state-of-the-art in reproducible OSN research; that is, appropriate documentation of OSN experiments to enable replication and indeed understanding of an experiment. Finally, we will look at responsible data collection; in particular, collecting data in an ethical fashion that respects the desires of the OSN users themselves.

11:50 **Jason McEwen** (University College London, UK)

Optimising radio interferometric imaging with compressive sensing

Abstract: We are about to enter a new era of radio astronomy with new radio interferometric telescopes under design and construction, such as the Square Kilometre

Array (SKA). While such telescopes would provide many scientific opportunities, they will also present considerable modelling and data processing challenges. Novel modelling and imaging techniques will be required to overcome these challenges. The theory of compressive sensing is a recent, revolutionary development in the field of information theory, which goes beyond the standard Nyquist-Shannon sampling theorem by exploiting the sparsity of natural images. Compressive sensing suggests a powerful framework for solving linear inverse problems through sparse regularisation, such as recovering images from the incomplete Fourier measurements taken by radio interferometric telescopes. I will present recent developments in compressive sensing techniques for radio interferometric imaging, which have shown a great deal of promise. Furthermore, by appealing to the theoretical foundations of compressive sensing, I will discuss how telescope configurations can be optimised to further enhance imaging fidelity via the spread spectrum effect that arises in non-coplanar baseline and wide field-of-view settings.

12:40 Floor discussion

LUNCH BREAK (12:50 - 14:00)

SESSION 2 (14:00 - 15:50). ROOM MS.03, ZEEMAN BUILDING

14:00 **Yuan Ji** (University of Chicago, USA)

Subgroup-Based Adaptive (SUBA) Designs for Multi-Arm Biomarker Trials

Abstract: Targeted therapies based on biomarker profiling are becoming a mainstream direction of cancer research and treatment. Depending on the expression of specific prognostic biomarkers, targeted therapies assign different cancer drugs to subgroups of patients even if they are diagnosed with the same type of cancer by traditional means, such as tumor location. For example, Herceptin is only indicated for the subgroup of patients with HER2+ breast cancer, but not other types of breast cancer. However, subgroups like HER2+ breast cancer with effective targeted therapies are rare and most cancer drugs are still being applied to large patient populations that include many patients who might not respond or benefit. Also, the response to targeted agents in humans is usually unpredictable. To address these issues, we propose SUBA, subgroup-based adaptive designs that simultaneously search for prognostic subgroups and allocate patients adaptively to the best subgroup-specific treatments throughout the course of the trial. The main features of SUBA include the continuous reclassification of patient subgroups based on a random partition model and the adaptive allocation of patients to the best treatment arm based on posterior predictive probabilities. We compare the SUBA design with three alternative designs including equal randomization, outcome-adaptive randomization and a design based on a probit regression. In simulation studies we find that SUBA compares favorably against the alternatives.

14:50 **Camille Stephan-Otto Attolini** (IRB Barcelona, Spain)

A Bayesian framework for personalized design in alternative splicing RNA-seq studies

Abstract: I will present a very useful (and nice) application of Bayesian predictive simulation to the problem of sample size calculation in the context of expression estimation from RNA sequencing data. New technologies have made possible the scrutiny of

gene expression at unprecedented levels and the analysis of these data has generated a large number of models and tools. Despite this, little effort has been done to address the problem of sample size calculation in setups that involve thousands of euros in the simpler experiment. We use a Bayesian probabilistic model to simulate reads from pilot data in order to compute optimality measures for different combination of experimental parameters. We focus on coverage calculation minimising estimation error for the single sample problem, while we try to optimise the number of differentially expressed isoforms in multi-samples experiments. Our results show that optimal parameters depend on characteristics such as species, tissue and conditions under study, resulting in the necessity of personalized designs. We found that large savings can arise from a well planned experiment, and suggest sequential acquisition of data in order to optimise resources.

15:40 **Floor discussion**

15:50 **Workshop ends**

TRAVEL DIRECTIONS

If you are either arriving by train, bus or flying into Birmingham International Airport, the easiest option to reach campus is to take a taxi (about 20min from airport/train/bus station). Taxi drivers should be able to take you directly to the Zeeman building (also known as the Maths building), where the workshop takes place. Further directions to the University of Warwick: <http://www2.warwick.ac.uk/about/visiting/directions>.