Statistical Challenges in Astronomy
Program

Day 1:
9:00 Welcome: Chris Tinney (Associate Dean Research)
9:05 – 9:50 Keynote: Daniel Foreman-Mackey, Gaussian processes
10:00 – 10:30 Invited talk: Sanjib Sharma, Bayesian methods

10:30 – 11:00: Morning tea

11:00 – 1:00 Workshop session 1: introduce projects, discuss methods, start working

1:00- 2:00 Lunch

2:00 - 3:00 Contributed talks:

2:00 - 2:20 Sebatian Haan (Center for Translational Data Science, The University of Sydney)

Probabilistic Data Fusion for SAMI Cube Construction

A key challenge for astronomical data processing is to optimally reconstruct the sky image from a set of observations with incomplete sampling. In particular 3D surveys such as the SAMI Galaxy Survey observe thousand of spectra at different spatial locations across a galaxy that need to be combined in a single coherent cube. We present a probabilistic framework for astronomical data fusion that applies a Gaussian process prior and allows to transform a set of fiber-optic data into a data cube by taking into account the time and wavelength dependent PSF. Our results show robust sky reconstruction and uncertainty estimates while providing optimal spatial resolution - in some cases even reaching super-resolution. The proposed method offers several advantages over current ad-hoc solutions such as the Drizzle method which can cause aliasing artefacts due to differential atmospheric refraction and does not fully take into account measurements of the atmospheric seeing.

2:20 - 2:40 Manodeep Sinha (Center for Astrophysics and Supercomputing, Swinburn)

Accurate Modeling of Galaxy Clustering on Small Scales: Testing the Standard ΛCDM + Halo Model

The large-scale distribution of galaxies can be explained fairly simply by assuming i) all galaxies are hosted by halos and ii) a cosmological model. This simple framework, called the 'halo-model', has been remarkably successful at reproducing the large-scale clustering of galaxies observed in various galaxy redshift surveys. However, none of these studies have truly tested the 'halo-model' by carefully modeling the systematics. We present the results from a fully-numerical, accurate 'halo-model' framework and show that the theory can not simultaneously reproduce the galaxy projected correlation function and the group multiplicity function in the SDSS main samples. In particular, the bright galaxy sample shows significant tension with theory. We discuss the implications of our findings, as well as how to constrain different aspects of galaxy formation by simultaneously fitting multiple statistics.

2:40 – 3:00 Dan Taranu (University of Western Australia)

Bayesian Physical Galaxy Modelling Using MagRite

Galaxies consist of multiple distinct structural components, including thin, rotating stellar and gas disks, spheroidal stellar bulges, and extended dark matter halos. We have developed a method and code (Magrite; Taranu et al. 2017) to build stable dynamical models of multi-component galaxies, describing the full 6D

distribution function of each component i.e. the density of mass as a function of position and velocity. We integrate these models to simultaneously fit observational galaxy data, including images and maps of the stellar velocity moments derived from spectra, using maximum likelihood and Bayesian MCMC methods. MagRite provides complete estimates of fundamental galaxy parameters like mass, energy, angular momentum and size using well-justified assumptions. I will also discuss some of the trickier practical aspects of obtaining sensible parameter confidence intervals for the all-too-frequent case where the data are not well-described by a simple, smooth model.

3:00 - 3:30: Afternoon tea

3:30 - 5:00: John Ormerod: Tutorial on MCMC using R / STAN

STAN is a software package for performing state of the art MCMC using the Hamiltonian Monte Carlo with a No-U-Turn Sampler (Hoffman and Gelman, 2012, Journal of Machine Learning Research). In this session we will provide a gentle tutorial of performing a MCMC analysis in the R software environment using the package RStan, the R Interface to STAN. We will cover R basics, how to specify and fit a model, and how to summarize display and interpret results.

5:00 - 6:00 Discussion: methods for current astrophysical problems, adjustments to projects

(Dinner together somewhere in Kingsford)

Day 2:

9:05 – 9:55 Keynote: Inge Koch (Australian Mathematical Sciences Institute & School of Mathematical Sciences, The University of Adelaide)

SOPHE: second order polynomial histogram estimators for multivariate densities, clustering and estimation of modes

Density estimation provides information about structure in data. As the number of variables increases, emphasis shifts to the estimation of modes and data clusters. In this talk I describe second order polynomial histogram estimators (SOPHE), a method for estimating modes and clusters in large multivariate data, and show some of their theoretical properties. Unlike kernel-based density estimation which becomes computationally infeasible for four and more variables, SOPHE are calculated efficiently for many variables and more than one million observations. Applications of SOPHE to multivariate data from flow cytometry illustrate the performance of these estimators.

10:00 – 10:30 Invited talk: Minh Ngoc Tran (Business Analytics, University of Sydney Business School)

Enabling Bayesian inference for big data

Recent advances in technology have produced increasingly large volumes of data. Data are big in terms of both the number of observations (tall data) and the number of observed variables (high-dimensional data). This leads to many research opportunities as well as challenges in statistical inference, in particular simulation-based Bayesian inference. In this talk, I will summarize some recent advances in Bayesian computation that enables Bayesian inference for big data. In particular, I will talk about subsampling-based Markov chain Monte Carlo and Hamiltonian Monte Carlo for tall data, and Variational Bayes estimation methods for extremely high-

dimensional data. The talk is based on recent work of our research group: Robert Kohn, Matias Quiroz, Khue-Dung Dang (UNSW), David Nott (NUS), Mattias Villani (Linkoping), Nghia Nguyen (Usyd)

10:30 – 11:00 Morning tea

11:00 - 12:30 Tomasz Bernarz (UNSW Art & Design, CSIRO/Data61)
Data visualization tutorial
Focus is on more general visual analytics and modern webGL to do simulations and creative maths. Also demonstrate nbody simulation on the GPU.

12:30 - 2:00 lunch

2:00 – 3:00 Contributed talks:
2:00 – 2:20: Yuguang Wang (Latrobe & UNSW)
Statistical Analysis of Cosmic Microwave Background Data
Cosmic Microwave Background (CMB) Radiation provides physical evidence of the inflationary model of the Universe. Accumulated over more than a decade by COBE, WMAP and Planck, the CMB data are stored at around 50 million points located on a two-dimensional unit sphere. Analysing these big data by statistical tools is critical to estimate accurate cosmological parameters, such as the density of matter and dark matter in the Universe, the distribution of the seeds of stars and galaxies, gravitational lensing, CMB power spectra, non-Gaussianity, isotropy.

We model the CMB map which shows the sky temperature of the early Universe after the big bang as a realization of a random field on the sphere. I will show how we can effectively simulate the random field of CMB data using needlet (wavelet) representation, and the dynamical evolution of the random field using fractional stochastic partial differential equations. I will report results of detection of non-Gaussianity of the Universe by multifractal analysis and detection of fluctuations in the CMB map which provides the seeds for stars and galaxies to form out of gravitational instability. This is joint work with Vo Anh, Phil Broadbridge, Danny Fryer, Nikolai Leonenko, Ming Li and Andriy Olenko.

2:20 – 2:40: Mathew Varidel (University of Sydney, Sydney Institute for Astronomy)
Bayesian Inference for Gas Kinematics using a Transdimensional Hierarchical Gaussian Mixture Model
Integral Field Spectroscopy has been a revolution in optical astronomy as it records spatially resolved spectra, from which we can infer properties across a single galaxy. However, the seeing convolves the data by blending the emission lines at different locations with different velocity profiles, which affects naive gas kinematic measurements. Instead, we must infer the underlying gas kinematics from the convolved data cube. In this talk, I will present a new 3D galaxy modelling approach which decomposes the galaxy by using a transdimensional hierarchical gaussian mixture model. The model can be used to infer local galaxy kinematics via the multiple gaussian decomposition, while constraining local properties with respect to galaxy-wide hyperpriors.

2:40 – 3:00: Dan Taranu (University of Western Australia)
Astronomical Image Processing and Source Modelling with ProFit
ProFit (Robotham et al. 2017; github.com/asgr/ProFit) is a new R code for Bayesian modelling of astronomical images. ProFit is designed to model the light profiles of galaxies with multiple components like disks and bulges using a variety of analytic profiles. ProFit is built on a fast C++ image generation and convolution library

(libprofit) and has access to a variety of optimizers in R, including MCMC samplers from LaplacesDemon. While ProFit is intended to model galaxies, I will show worked examples of how it can be used as a general Bayesian image modelling framework, including fitting the point spread function that describes how sources are blurred by the atmosphere in images from ground-based telescopes, and simultaneous modelling of resolved and unresolved sources including galaxies and stars.

3:00 – 3:30 Afternoon tea

3:30 - 4:00 Workshop session 3: work on projects, present methods and progress