New MCMC-Enabled Bayesian Statistical Methods for Complex Data and Computer Models in Astronomy

PI: Professor David A van Dyk, Imperial College London

Project Objectives. This project is a continuation of Professor van Dyk's highly productive work in tackling outstanding statistical problems in astrophysics. The overarching goals are to *establish unified frameworks for statistical analyses of complex data using state-of-the-art statistical, astronomical, and computer models* and to *develop sound statistical methods to solve data analytic problems in astronomy and to extend them for general statistical use*.

In recent years technological advances have dramatically increased the quality and quantity of data available to astronomers. New instrumentation provides massive surveys resulting in terabytes of data, high resolution spectrography and imaging across the electromagnetic spectrum, and incredibly detailed movies of dynamic and explosive processes in the solar atmosphere. These studies aim to improve our understanding of the evolution of the Universe and of our own origins. While many new instruments are making impressive strides in answering the central questions of astrophysics, they are also generating massive data-analytic and data-mining challenges. Subtle physical processes generate very weak signals relative to background; parameters of complex computer models designed to approximate physical processes have large uncertainties; and data truncation and measurement errors are unavoidable in studies of faint stars, super novae, galaxies, and solar features.

Analysis of the faint signals provided by new instrumentation requires sophisticated astronomical and statistical models that aim to link the underlying physics of the source with its observed spectrum, image, and/or light curve. This often requires a mixture of mathematical and computer models. Computer models are, for example, used to describe complex physical processes such as the evolution of stars and galaxies, the formation of the Universe, and the workings and calibration of sophisticated space-based telescopes. These models are increasingly popular throughout the physical sciences and are the subject of much interest in the statistical literature. Incorporating them into a principled statistical analysis, however, leads to significant modeling, inferential, and computational challenges. This project's primary goal is to develop principled statistical and computational techniques for complex physical systems, components of which may require computer models. In astronomy computer models are typically used in isolation—inputs are tweaked to improve agreement between predictions and data. The methods developed in this project, on the other hand, embed them into a principled statistical analysis that enables us to leverage statistical theory and methods for parameter estimation, uncertainty calculation, and model checking.

Principled statistical analyses allow proper assessment of deviations between predictions and observations. In this way, we are able to empirically compare models (e.g., based on different physical assumptions) with each other or with external predictions. Computation is challenging with complex models, especially those involving computer models and the development of new techniques for statistical computation is central to this project. Our methodological and computational development draws on a wide range of examples from astrophysics and related fields including calibration of x-ray detectors, stellar evolution, cosmology, spectral analysis, image analysis, gravitational lensing, solar physics, and particle physics. We both develop statistical methods specifically tailored to these settings and use them as a testing ground in the development of general statistical methods. Specific research objectives include:

Objective 1: Development of the first ever principled fully-Bayesian methods to account for calibration uncertainty in high-energy x-ray analysis, and extending these methods to other calibration-dependent analyses in astronomy.

Objective 2: Development sophisticated multilevel models and inference techniques for the complex astronomical and cosmological systems, thereby dramatically improving the estimates of their physical parameters.

Objective 3: The proposal of careful model diagnostics that allow us to detect inconsistencies between predictions and (external) data streams, to compare competing models, and to improve the underlying astronomical models.

Work to Date. Substantial progress has been made on all three project objectives.

Objective 1: Calibration of data-collecting instruments is critical to any signal processing task. Without reliable calibration, the data cannot be transformed into physically meaningful quantities that can be mapped to questions of scientific interest. Our approach uses sophisticated computer models to quantify calibration errors and their effects on analyses of astronomical observations. Work has focused on the the development of a method that embeds this computer model into a multilevel statistical model that allows for principled assessments of the effects of calibration

uncertainty on scientific findings. Specific examples have been developed in detail, including accounting for uncertainty (1) in calibration products used for high-energy spectral analysis and (2) in the atomic data base (AtomDB) used to map images of the Sun and/or stellar spectra to their underlying physical processes, temperatures, and compositions. The extension to AtomDB was funded in part by the International Space Science Institute (ISSI, Switzerland). A related project developed Bayesian hierarchical models to combine observations of multiple objects with multiple telescopes to estimate both calibration adjustments and their errors.

Objective 2: Analysis of the faint signals provided by new instrumentation requires sophisticated astronomical and statistical models. These models aim to link parameters that describe the underlying physics of the source with its observed spectrum, image, and/or light curve. This often requires a multilevel model composed of sophisticated physical, statistical, and computer models. This objective aims to design, deploy, and validate such multilevel models for a variety of astrophysical problems. Work has focussed on three classes of problems, those stemming from the study of stellar evolution, of cosmology, and of the solar cycle. In particular we have developed and/or extended models that allow us to reconstruct the physical characteristics of a single white dwarf star or a cluster of stars, to study the age of the galactic halo, to estimate the carbon fraction of white dwarf stars, to separate multiple stellar populations in a globular cluster, to leverage observations of super novae explosions and gravitationally lensed images to estimate cosmological parameters, and to more accurately predict the solar cycle.

Fitting such models requires sophisticated computational methods. We have extended the so-called Partially Collapsed Gibbs sampler to enable it to fit a broader class of models and developed the new Repulsive-Attractive Metropolis (RAM) algorithm. Applications have been derived for a number of astronomical models and for the multinomial probit model which is very popular in the social sciences.

Objective 3: Quantifying statistical evidence to choose between two of more physical models is a ubiquitous problem in the physical sciences. We are studying a set of powerful methods to make such choices for a class of problems that are common in high-energy astrophysics, photon science, and particle physics. This class includes the search for new physical particles, such as the Higgs boson. Other classes of problems under investigation include testing for unexpected structure in images of astronomical sources, apparent abrupt changes in the brightness and spectra of sources, and for the existence of x-ray dark sources in a population.

In the physical science, we are often study populations of objects and aim determine the best physical model for each. From a statistical perspective this can be formalized as classification of the objects into groups associated with each of several possible models. We have developed methods to automatically classify solar active regions, galaxies, and supernovae into their constituent types, and to segment solar images into regions with similar thermal activity.

Final Results. The advanced new statistical techniques developed under this project will allow astronomers to combine sophisticated statistical models and computer models of complex physical processes with state-of-the-art data steams in order to study the physical environment and structure of sources, the processes and laws which govern the birth and death of planets, stars, and galaxies, and ultimately the structure and evolution of the universe. The scientific impacts of this project fall into two categories. First is the impact of more reliable statistical methods on scientific findings in astronomy and the impact of the new statistical inference and computation methods in a wide range of scientific fields. Not only have new methods been developed and software freely distributed, but efforts have been made to educate the astronomical community as to the benefit of the careful use of sophisticated statistical methods. It is expected that a fundamental impact will be a more general acceptance and use of appropriate statistical methods among astronomers. Second, rather than using off-the-shelf models and methods, we develop application specific techniques that account for the particular complexities of a problem at hand. In this way we have developed numerous inferential and computational methods for handling multilevel models including complex computer models.

This project has helped fund 29 research papers (24 accepted or in print and 5 under review) with two more in preparation, 6 PhD theses with 4 more in progress, 59 professional presentations including 6 key-note addresses, 6 scheduled latter in 2017, and 13 educational or outreach presentations, the organization of 17 conferences and workshops including two near Imperial, and 7 publicly available software packages. Funding has enabled the PI to engage in new collaborations with researchers in Sweden, Spain, France, Belgium, Italy, UK, Greece, and the US.

Find out More! A more complete description of the project results can be found at
`http://wwwf.imperial.ac.uk/~dvandyk/`.