Imperial College London

Data Science Institute, Imperial College London

Advanced Statistical and Computational Methods for Emerging Challenges in Astronomy and Solar Physics

Professor David A van Dyk, Statistics Section (Mathematics), CHASC International Center for Astrostatistics

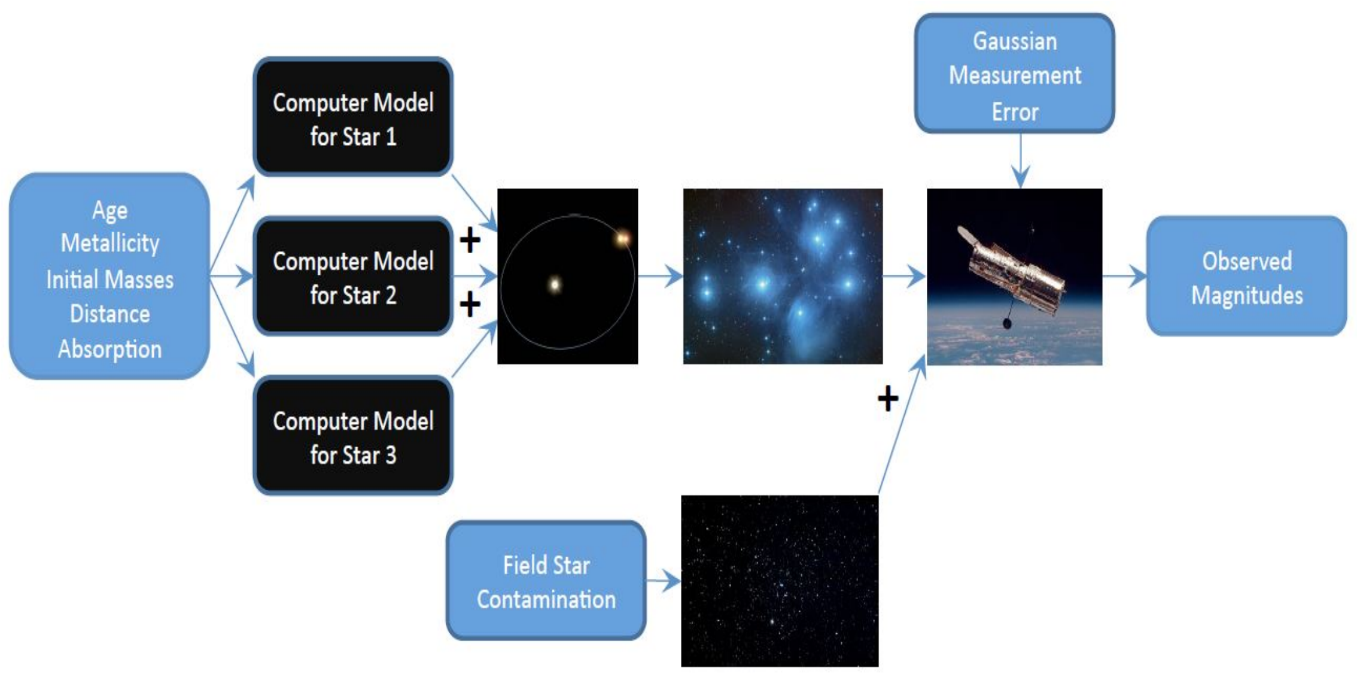
Big Data Challenges in Astrostatistics

In recent years, technological advances have dramatically increased the quality and quantity of data available to astronomers. Newly launched or soon-to-be launched space-based telescopes are tailored to data-collection challenges associated with specific scientific goals. These instruments provide massive new surveys resulting in new catalogs containing terabytes of data, high resolution spectrograph and imaging across the electromagnetic spectrum, and incredibly detailed movies of dynamic and explosive processes in the solar atmosphere. These new data streams are helping scientists make impressive strides in our understanding of the physical universe, but at the same time are generating massive data-analytic and data-mining challenges for scientists who study them.

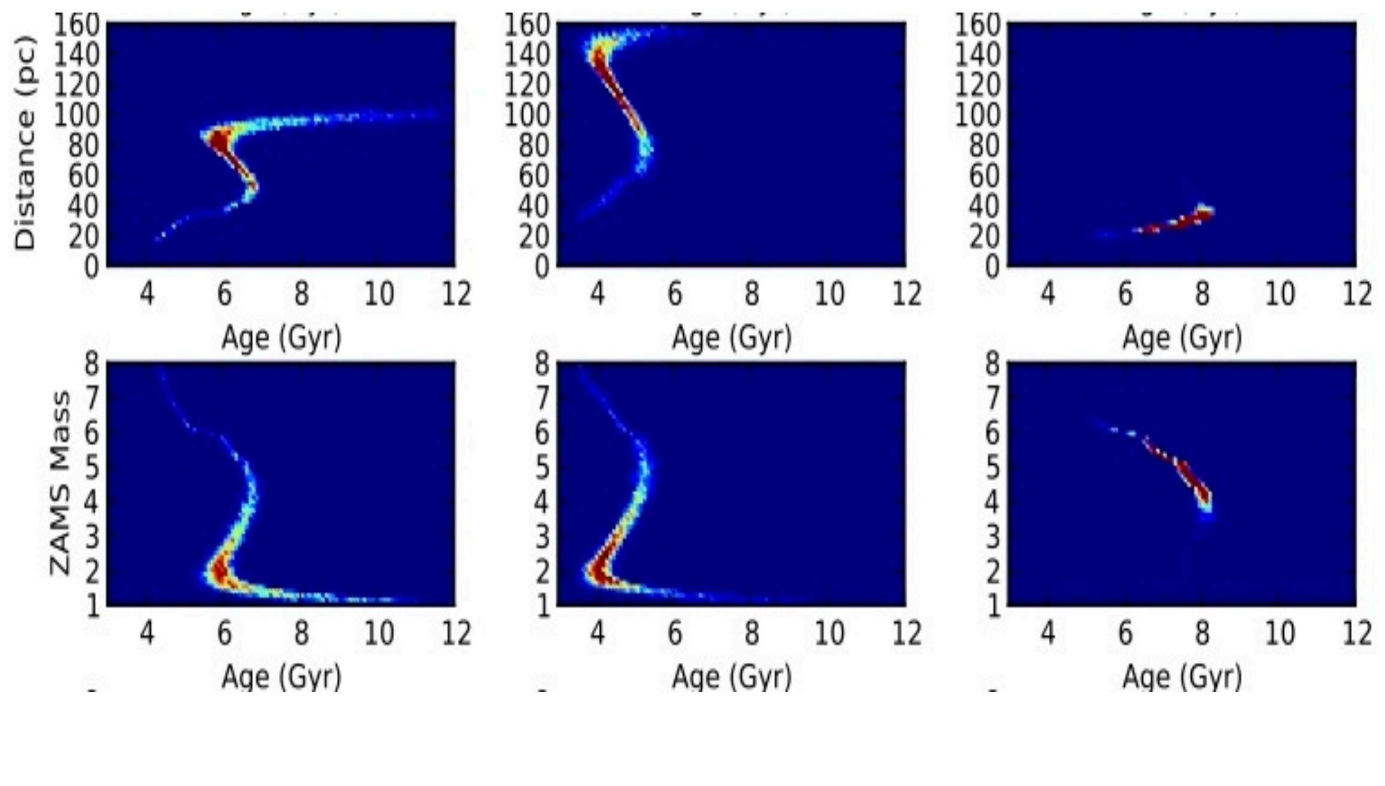
The complexity of the instruments, the complexity of the astronomical sources, and the complexity of the scientific questions lead to many subtle inference problems that require sophisticated statistical tools. For example, data are typically subject to non-uniform stochastic censoring, heteroscedastic errors in measurement, and background contamination. Scientists wish to draw conclusions as to the physical environment and structure of the source, the processes and laws which govern the birth and death of planets, stars, and galaxies, and ultimately the structure and evolution of the universe. Sophisticated astrophysics-based computer-models are used along with complex parameterized and/or flexible multi-scale models to predict the data observed from astronomical sources and populations of sources. The CHASC International Center for Astrostatistics tackles outstanding statistical problems generated in astro and solar physics by establishing frameworks for the analysis of complex data using state-of-the-art statistical, astronomical, and computer models. In doing so the researchers in the Center not only develop new methods for astronomy, but also use these problems as spring boards in the development of new general methods, especially in signal processing, multilevel modelling, computer modelling, and computational statistics.

Here we outline a number of our current research activities.

The Statistical Analysis of Stellar Evolution


Age
Metallicity
Initial Masses
Distance
Absorption
Computer Model for Star 1
Computer Model for Star 2
Computer Model for Star 3
+
+
+
Gaussian Measurement Error
Observed Magnitudes
Field Star Contamination

The physical processes that govern the evolution of sun-like stars first into red giants and then white dwarf stars can be described with mathematical models and explored using sophisticated computer models. These models can predict observed stellar brightness (magnitude) as a function of parameters of scientific interest, such as stellar age, mass, and metallicity. We embed these computer models into multilevel statistical models (see diagram) that are fitted using Bayesian analysis. This requires sophisticated computing, corrects for data contamination by field stars, accounts for complications caused by unresolved binary-star systems, and allows us to compare competing physics-based computer models for stellar evolution. Parameters of scientific interest can exhibit complex non-linear correlations (see figure below) that cannot be uncovered or summarized using standard methods. Principled statistical models and adaptive computational techniques are specially designed to fully explore such relationships.



Embedding the Big Bang Cosmological Model into a Bayesian Hierarchical Model

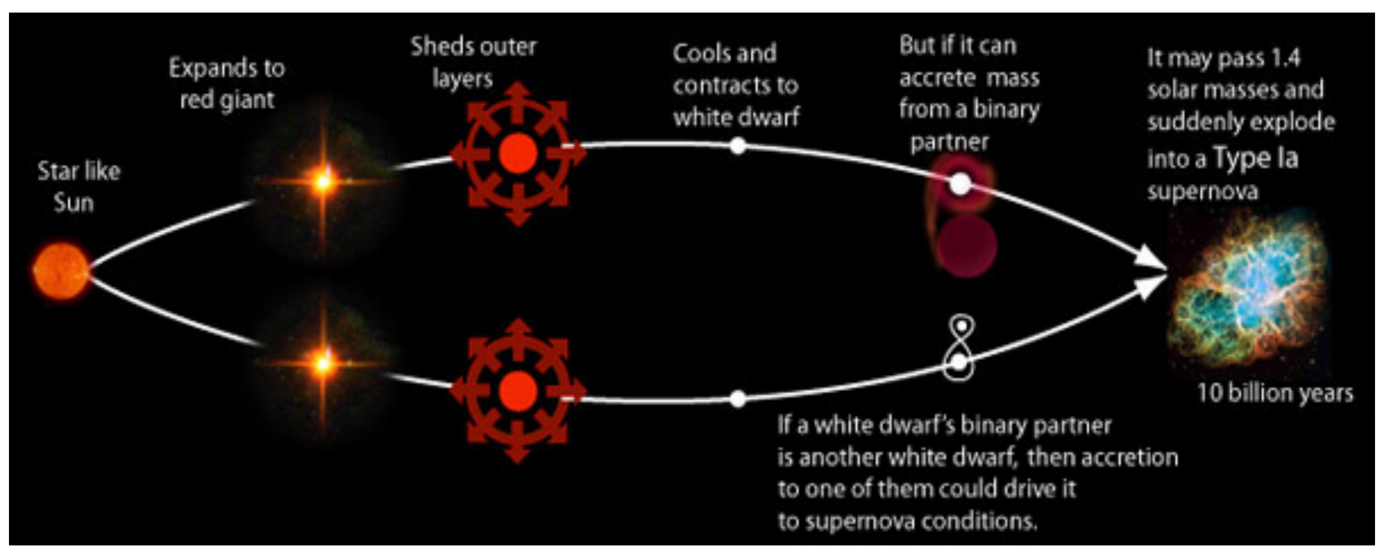

Star like Sun
Expands to red giant
Sheds outer layers
Cools and contracts to white dwarf
But if it can accrete mass from a binary partner
It may pass 1.4 solar masses and suddenly explode into a Type Ia supernova
10 billion years
If a white dwarf's binary partner is another white dwarf, then accretion to one of them could drive it to supernova conditions.

Image Credit: http://hyperphysics.phy-astr.gsu.edu/hbase/astro/snovcn.html

The 2011 Nobel Prize in Physics was awarded for the discovery that the expansion of the Universe is accelerating. We have developed a Bayesian model that relates the difference between the apparent and intrinsic brightnesses of objects to their distance which in turn depends on parameters that describe this expansion. Type Ia Supernova, for example, occur only in a particular physical scenario. This allows us to estimate their intrinsic brightness and thus study the expansion history of the Universe. Sophisticated Markov chain Monte Carlo methods are used for model fitting and a secondary Bayesian analysis is conducted for residual analysis and model checking.

Identifying Unspecified Structure in Low-Count X-ray Images

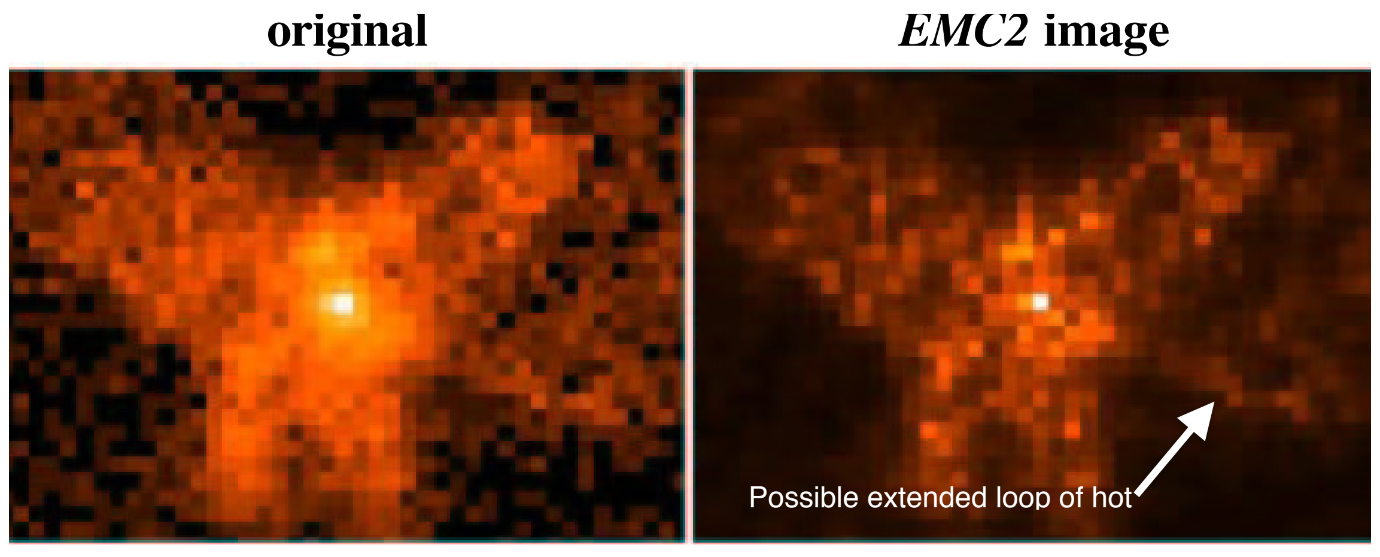
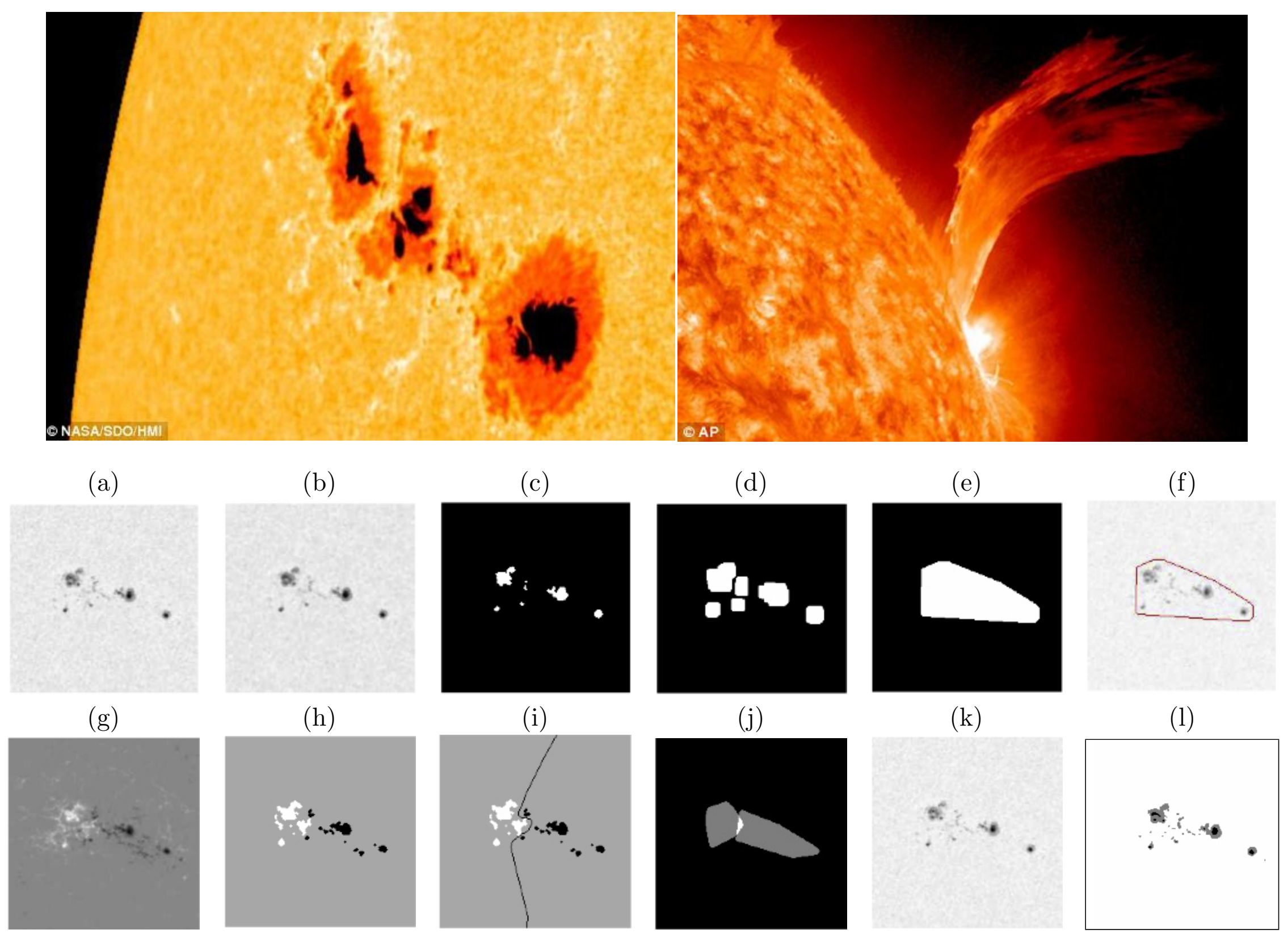


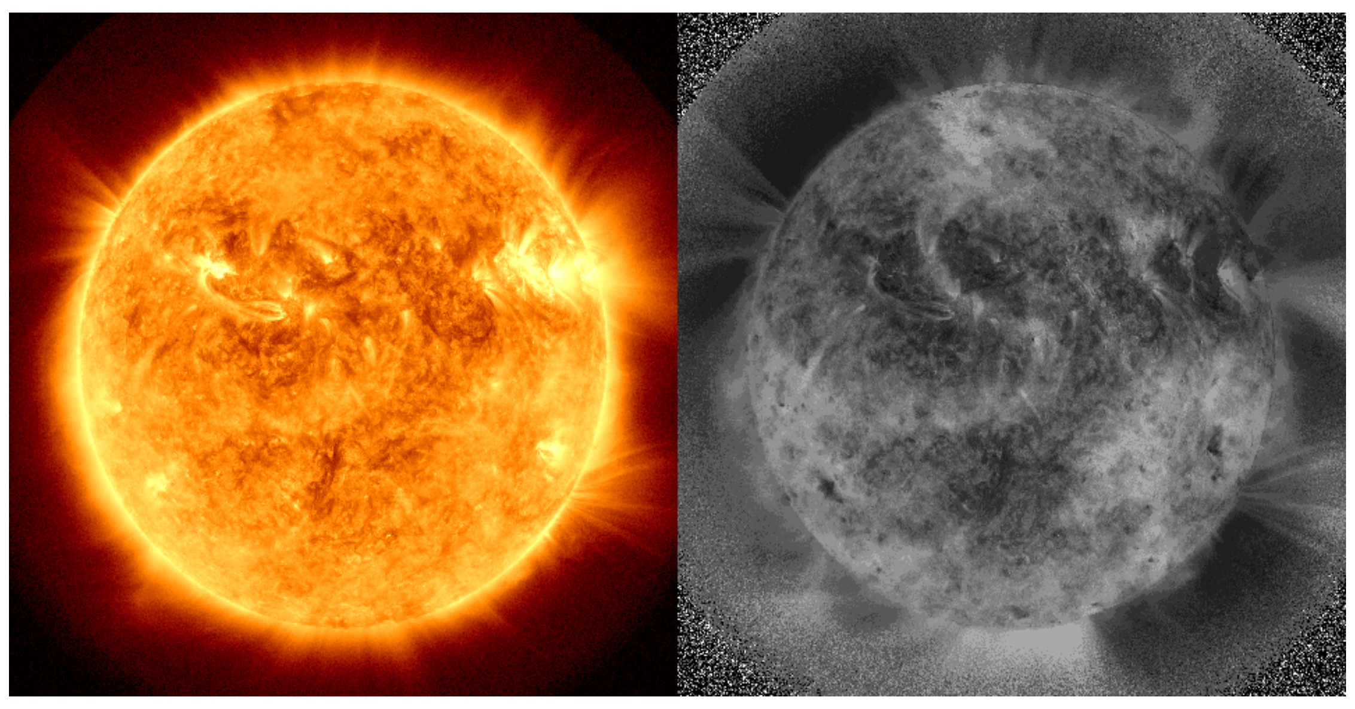
Image restoration including deconvolution techniques offers a powerful tool to improve resolution in images and to extract information on the multiscale structure stored in astronomical observations. Using a Bayesian model-based framework allows us both to quantify the uncertainty in the reconstructed images and to conduct formal statistical tests for unexpected structure in the image. NGC 6240, for example, is a nearby ultraluminous infrared galaxy that is the remnant of a merger between two smaller galaxies. The restored (EMC2) X-ray image of NGC 6240 shows a faint extended loop of hot gas. We are developing computationally efficient Monte Carlo techniques for quantifying the evidence for such structure.

Classification, Tracking and Prediction of Solar Features



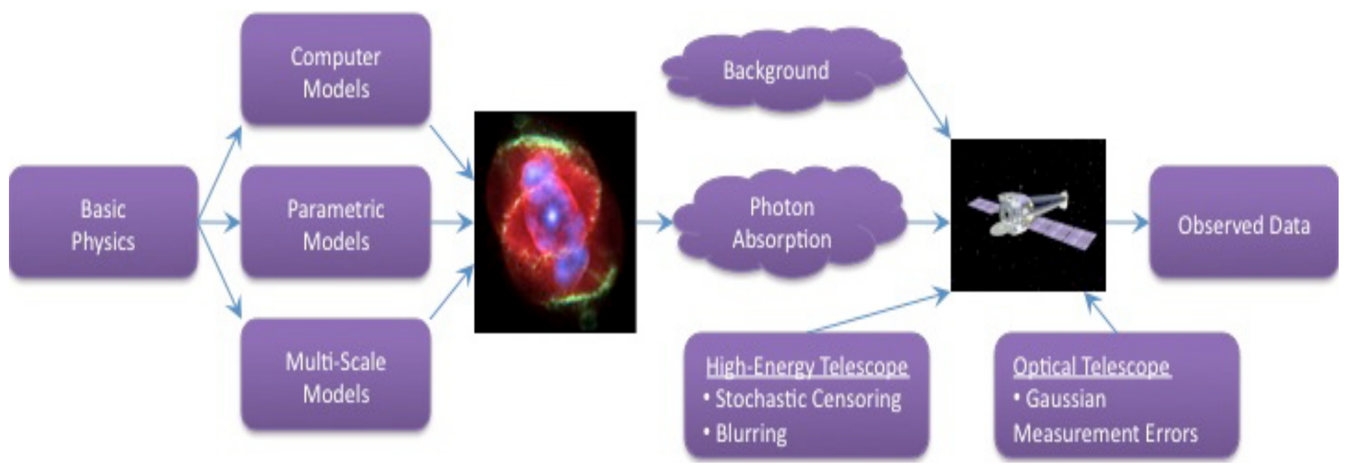
In order to take full advantage of the high-resolution and high-cadence Solar images that are now available, we must develop methods to automatically process and analyze large batches of such images. This involves reducing complex images to simple representations such as binary sketches or numerical summaries that capture embedded scientific information. The morphology of sunspot groups, for example, is predictive not only of their future evolution but also of explosive events associated with sunspots such as solar flares (top right) and coronal mass ejections. Using techniques involving mathematical morphology, we demonstrate how to reduce solar images into simple 'sketch' representations and numerical summaries that can be used as features for an automated classification and tracking.

Identifying Solar Thermal Features Using H-Means Image Segmentation

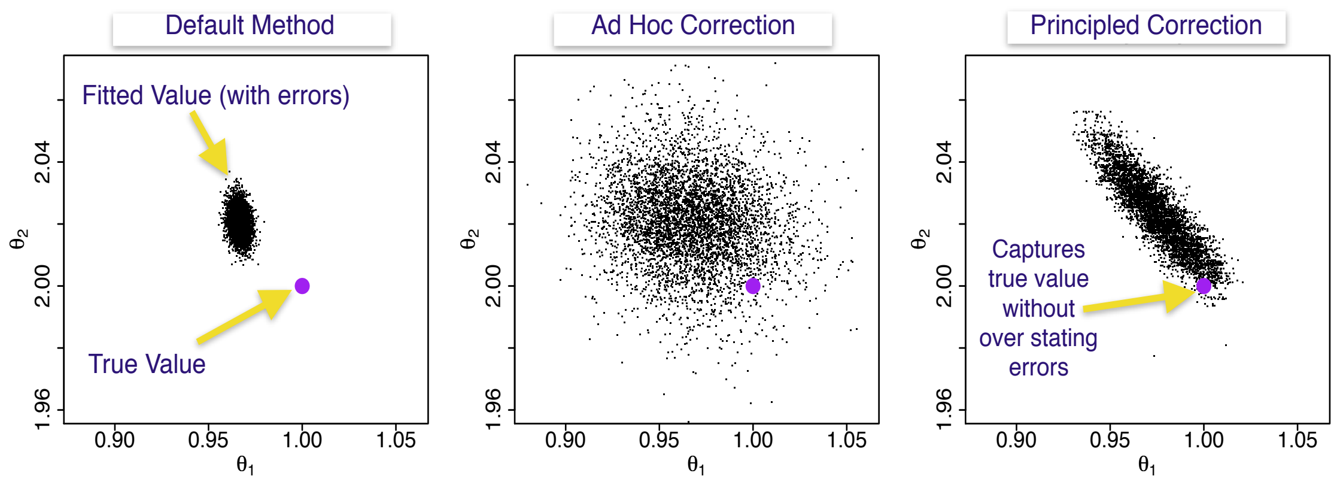


Properly segmenting multi-band images of the Sun by their thermal properties helps to determine the thermal structure of the solar corona. Off-the-shelf segmentation algorithms, however, are typically inappropriate because temperature information is captured by the relative intensities in different pass-bands, while the absolute levels are not relevant. Input features are therefore pixel-wise proportions of photons observed in each band. To segment solar images based on these proportions, we use a modification of k-means clustering that we call the H-means algorithm because it uses the Hellinger distance to compare probability vectors. H-means has a closed-form expression for cluster centroids, so computation is as fast as k-means. Application of our method reveals never before seen structure in the solar corona—see the large S-shaped feature in the right-hand panel of the figure.

Low-Count Spectral Analysis in High-Energy Astrophysics



Spectra describe the energy distribution of photon emitted from an astronomical source and carry information as to the composition and physical processes at work in the source. The space-based observatories that study high-energy (X-ray and γ-ray) spectra are subjected to stochastic data distortion processes (blurring, heterogeneous censoring, and background contamination). We build sophisticated multi-level statistical models that account for both physical-processes in the sources themselves and for stochastic data distortion (see diagram). Uncertainty in the calibration of the instruments is a particular challenge and is typically ignored in practice. The figure below illustrates the fitting of two spectral parameters ignoring calibration uncertainty (left panel) and with an ad hoc (middle panel) and a principled (right panel) correction. The statistically principled method captures the true parameter values without overstating uncertainty in the fit.



Acknowledgements

We gratefully acknowledge funding for this project partially provided by the Royal Society (Wolfson Merit Award), the US National Science Foundation (DMS-12-08791), the European Commission (Marie-Curie Career Integration Grant), and the UK Science and Technology Facilities Council.

Collaborators:

Elizabeth Jeffrey (James Madison), William Jeffreys (Texas and Vermont), Xiyun Jiao (Imperial), Vinay Kashyap (Harvard Smithsonian Center for Astrophysics), Thomas Lee (UC Davis), Xiao Li Meng (Harvard), Erin O'Malley (Dartmouth), Aneta Siegminowska (Harvard Smithsonian Center for Astrophysics), Shijing Si (Imperial), Nathan Stein (U Penn), David Stenning (UC Irvine), Roberto Trotta (Imperial), Ted von Hippel (Embry-Riddle), Jin Xu (UC Irvine), and Yaming Yu (UC Irvine).