University of California, Irvine
Statistics Seminar

Hybrid Statistical / Learning Method for Principled Astrostatistical Inference

David van Dyk
Professor of Statistics
Imperial College London

4 p.m., Tuesday, Dec. 3, 2024
6011 Donald Bren Hall

In recent years, technological advances have dramatically increased the quality and quantity of data available to astronomers. While this has enabled scientists to make impressive strides in our understanding of the physical Universe, it is also generating massive data-analytic challenges. At the same time, rapid progress in computational techniques under the guise of machine learning is providing powerful new tools, particularly for massive data sets. These tools, however, are not always well-suited to reliable scientific inference. In this talk, I illustrate how multi-level statistical models can combine direct science-based modelling with fast flexible learning methods for nuisance processes (e.g., data collection, non-representative sampling, instruments), thereby preserving scientific interpretability of the overall models and their parameters. The probabilistic nature of principled statistical frameworks, within which we embed learning algorithms, preserves the transparency of modelling assumptions and results in well-calibrated uncertainty assessments for estimates and predictions. I will discuss two examples. First, *StratLearn* is a statistically principled and theoretically justified method to improve supervised learning when the training set is not representative. *StratLearn* uses well-established techniques from causal inference to provide probabilistic classifications that can be carried forward to category-specific analyses. The second example develops *Jolideco* (short for joint likelihood deconvolution) for scientific analysis of high-energy astrophysical images. *Jolideco* is a fully Bayesian method and combines a Poisson likelihood with a patch-based prior. In this way, it favors correlation structures among the reconstructed pixel intensities that are characteristic of those observed in high-resolution training images. This Bayesian formulation allows us to meaningfully quantify uncertainty in features observed in fitted images.