Final Report for Summer Program on Modern Statistical and Computational Methods for Analysis of Kepler Data, 10-28 June 2013

Summary. This three-week Summer Research Program can be viewed as a follow up of the Spring Program on Astrostatistics in 2006.

The Kepler program was characterized by a high level of excitement, enthusiasm and productivity, as well as a superb synergy between statistics and astronomy. Statisticians learnt about astronomy and the fine points of the Kepler data, while astronomers learnt about advanced statistical techniques. Then members from both groups started immediate collaborations to put statistics to good use in the search for small, rocky and potentially habitable planets.

The Kepler participants were a well-balanced mix mixture of astronomers and statisticians at different career stages. The organizers had made an extraordinary effort to include many graduate students and postdocs, as well as a surprisingly large number of women.

Three main working groups (WGs) formed and collaborated over three weeks. Rapid progress lead to several publications likely to be completed in the near future. Specific plans were also made for topics and logistics of long-term future collaborations.

Overview. We give a brief motivation for the program (Section 1), followed by list of participants (Section 2), the program schedule (Section 3), and a sketch of the working group research (Section 4).

In the text to follow, asterisks denote women or members of under-represented groups; and members of the Kepler Science Team are also explicitly identified.

Organizers: Eric Ford (Astro, U. Florida, Kepler Team), Paul Baines (Stat, UC Davis), Jim Berger (Stat, Duke), David Hogg (Physics, NYU)
SAMSI Directorate Liaison: Ilse Ipsen* (Math, NCSU)
Web page: `http://www.samsi.info/Kepler`

1 Purpose

For centuries, theories of planet formation were guided exclusively by our solar system. Discoveries during the past two decades, however, have revealed planets that orbit other stars. These so-called *exoplanets* illustrate that nature can produce planetary systems that are quite different from our own.

NASA's Kepler Mission is a search for habitable planets. The search is conducted with a specially designed 0.95-meter diameter telescope, situated aboard a space craft on a heliocentric orbit. Since 2009, the telescope has been almost continuously observing over 190,000 stars, once every 1 or 30 minutes, with a duty cycle of about 95 percent. While Kepler's high-precision photometry is revolutionizing several areas of astronomy, including exoplanets, astroseismology, and variable stars, it is also raising several new statistical challenges.

The SAMSI Kepler Program was designed to meet these challenges, by bringing together astronomers, astrophysicists, and statisticians, in the hope that the statisticians will help to improve the statistical tools and to develope new techniques geared towards the analysis of exoplanet data.

Why Now?

In October 2012, observations from the first three years of NASA's Kepler Mission were made public, and subsequent observations are becoming public immediately upon processing. Since these data are no longer proprietary, members of the Kepler Science Team, who are well versed with the Kepler data, are free to discuss all data with all participants. This makes it easier for the broader community of astrophysicists and statisticians to contribute to the analysis of this revolutionary data set. All data from the nominal mission, that is, the first 3.5 years of spacecraft operations, were available for study by all participants.

2 Participants

The organizers made a concerted effort to attract as many early-career participants and women as possible.

Graduate students

Ruth Angus* (Physics, U. Oxford, UK)
Mary Beth Broadbent* (Stat, Duke)
Shi-Han Chang (Math/Stat, Duke)
Rebekah Dawson* (Astro, Harvard)
Daniel Foreman-Mackey (Physics, NYU)
Kelly Hambleton* (Astro, Villanova)
Fengji Hou (Physics, NYU)
Michael Lindon (Stat, Duke)
Robert Morehead (Astro, U. Florida)
Ben Montet (Astro, Caltech)
Ben Nelson (Astro, U. Florida)
Megan Shabram* (Astro, U. Florida)
Angie Wolfgang* (Astro, UC Santa Cruz)

Postdocs and junior faculty

Joshua Carter (CfA[1], Kepler Team)
Jessi Cisewski* (Stat, CMU)
Gal Matijevic (Astro, Villanova)
Hannu Parviainen (Astro, Oxford, UK)
Matthew Payne (CfA[1], Kepler Team)
Andrej Prsa (Astro, Villanova)
Billy Quarles (NASA)
Darin Ragozzine (Astro, U. Florida, Kepler Team)
Leslie Rogers* (Astro, Caltech)
Meg Schwamb* (Astro, Yale)

Senior Statisticians

Jogesh Babu (Stat, Penn State)
Paul Baines (Stat, UC Davis)

Merlise Clyde* (Stat, Duke University)
Jim Berger (Stat, Duke University)
Robert Wolpert (Stat, Duke University)

Senior astronomers

Thomas Barclay (NASA Ames)
Jessie Christiansen* (NASA Ames)
Eric Ford (Astro, U. Florida, Kepler Team)
David Hogg (Physics, NYU)
Jon Jenkins (NASA Ames/SETI, Kepler Team)
Tom Loredo (Astro, Cornell)

3 Program Schedule

The Kepler Porgram consisted of roughly three parts: Invited talks and WG formation during the first two days (Section 3.1) subsequent research collaborations in WGs (Section 3.2) until the last day, when results were summarized and plans for future collaborations were made (Section 3.3).

3.1 Talks and WG formation (10 and 11 June 2013)

The invited talks were designed to help participants understand the nature of Kepler data, and to provide an introduction to relevant statistical methods. With a few exceptions, the talks on the first day were devoted to Kepler data and astronomy, while those on the second day focussed on statistical methods. All talks were scheduled for 30 minutes.

Talks on 10 June 2013

Eric Ford (Astro, U. Florida, Kepler Team) *Welcome and Overview*
Thomas Barclay (NASA Ames) *Overview of Kepler Mission*
Matthew Payne (CfA[1], Kepler Team) *Long Period Planets from the Kepler Data*
Thomas Barclay (NASA Ames) *Searching for Planet Candidates in Kepler Data*
Andrej Prsa (Astro, Villanova) *Behind the Scenes of Kepler Eclipsing Binary Science*
Tim Morton (Astro, Caltech) *Validation of Kepler Planet Candidates* (via webex)
Joshua Carter (CfA[1], Kepler Team) *Characterizing Planets via Transit and/or Eclipse Timing Variations*
Eric Ford (Astro, U. Florida, Kepler Team) *Characterizing the Distribution of Planetary Architectures with Kepler*
Darin Ragozzine (Astro, U. Florida, Kepler Team) *Understanding New Systems of Multiple Planets from the Kepler Space Telescope*
Tom Loredo (Astro, Cornell) *Graphical Models on Graphics Processors*
David Hogg (Physics, NYU) *Hierarchical Inference for Characterizing Populations*
Paul Baines (Stat, UC Davis) *Efficient Bayesian Computation*

[1]Harvard-Smithsonian Center for Astrophysics

Talks on 11 June 2013
Jim Berger (Stat, Duke) *Bayesian Adjustment for Multiplicity*
Merlise Clyde* (Stat, Duke) *Model Selection, Interim Priors and Hierarchical Models*
Jessie Christiansen* (NASA Ames) *Characterizing the Completeness of the Kepler Planet Candidate List*
Robert Wolpert and Mary Beth Broadbent* (Stat, Duke) *LARK: One Way to Quantify Uncertainty about Light Curves*
Jessi Cisewski* (Stat, CMU) *Approximate Bayesian Computing*
Meg Schwamb* (Astro, Yale) *Planet Hunters: Searching for Exoplanets with 500,000 Eyes*

The formal talks were followed by informal "introductions" where the remaining participants, mostly students and postdocs, gave 5 minute descriptions of their research interests.

In the afternoon, participants brain-stormed about research topics and then organized themselves into the three main working groups (WGs) below:

1. *Light curve detrending and planet detection*
2. *Exoplanet populations*
3. *Exoplanet characterization*

3.2 Research (12 - 27 June 2013)

Each week day, the WGs met at SAMSI roughly from 8 to 5, to conduct collaborative research. A specific time of day was allocated for each WG to hold "office hours" and present updates.

In the middle of the day, all participants got together to listen to talks of general interest (in astronomy or statistics) while eating lunch.

3.3 Last day (28 June 2013)

The WGs presented their findings, and discussed how to continue their collaboration beyond the SAMSI Kepler Program.

The following participants gave summaries of research results on behalf of their WGs.

1. *Planet detection and light curve detrending*
   Ruth Angus, Bekki Dawson, Hannu Parviainen, Bill Quarles
2. *Exoplanet populations*
   Robert Morehead, Matt Payne
3. *Exoplanet characterization*
   Jesse Cisewski, Eric Ford, Angie Wolfgang

The talks were intermingled with lively and constructive discussions about the details and structure of the resulting papers. Then the participants talked about the future, and in particular:

1. How to continue regular meetings and collaborations beyond the Kepler Program;
2. Which platforms to use for communication (phone, skype, Google Hangout, WebEx, etc);
3. Where to deposit software (GitHub); and

4. Whether to open up WG collaborations to outsiders (mostly no, but with a few exceptions).

The participants requested that the SAMSI Kepler WG pages be kept open for disseminating documents, write-ups, and discussions.

4 Working Group Research

A history of WG development (Section 4.1), and brief reports from individual participants (Section 4.2) are followed by a summary of the discussions on the last day (see Section 3.3) about short-term plans (Section 4.3), and long-term plans for collaborations beyond the Kepler Program (Section 4.4).

4.1 WG Development

Below is a short history of the WG development throughout the course of the Kepler Program. Three WGs were established at the outset of the program (see Section 3.1), three WGs were active throughout the program, and three WGs plan to continue (see Section 4.3). However the configurations were changing. In general, there was a lot of communication among the different WGs.

1. *Light curve detrending and planet detection*
   Informal subgroups: Gaussian Processes, Wavelets, Development of testbeds for different methods

   This WG was established at the outset, remained active, and plans to continue.

2. *Exoplanet populations*
   Informal subgroups: MCMC and Hierarchical Bayesian Models, Approximate Bayesian Computing and Hierarchical Bayesian Models, Long-period planets

   This WG was established at the outset, remained active, and plans to continue. However, the long-planets subgroup split off.

3. *Long-period planets*

   This WG started originally as a subgroup of the WG on exoplanet populations. It turned out, though, that its problems were somewhat different from those of the rest of the WG, and in some aspects more similar to those of the WG on light curve detrending and planet detection. Hence it made sense to spin off into a separate WG.

4. *Exoplanet characterization*

   This WG was established at the outset. Members met regularly for the first week, but the activity died down during the second week, partly because members started spending time with other WGs. This WG doesn't really plan to continue its operation.

4.2 Individual reports

Ruth Angus worked in the light curve detrending WG on searching for small transit signals. She collaborated with Gal Matijevic, Ben Montet and Dan Forman-Mackey to develop an optimised box-least-squares search algorithm with notched median filter detrending. This code was designed to iteratively search for and remove transit-like signals in Kepler light curves in order to find small planets. She plans to continue to develop this search method and will compare the performance of different detrending algorithms for a future publication.

Jogesh Babu had several discussions with Jon Jenkins that helped clarify the data structure. This helps in formulating appropriate statistical procedures. Participation in the light curves WG was useful in collaborating in the formulation of the likelihood function.

Paul Baines was part of the detrending WG, and worked with Bekki Dawson, David Hogg, Dan Foreman-Mackey and Ben Montet on a wavelet-based approach to simultaneously model the detrending process and estimation of transit parameters. By adopting more flexible models for the non-transit astrophysical and noise processes, the approach seeks to account for the wide variety of features in Kepler light curves that cause false positives in transit detection. The motivation behind the approach is to retain computational efficiency and scalability while introducing enough complexity to handle the many challenges in extracting the transit signal from Kepler light curves.

Jessi Cisewski worked with the populations modeling WG on inference of hierarchical Bayesian (HB) models in order to characterize the orbital eccentricity distribution of extrasolar planets. In particular, she worked with Eric Ford, Merlise Clyde, Robert Wolpert, Darin Ragozzine, and remotely with Chad Shafter (Stat, CMU) on building an approximate Bayesian computing (ABC) algorithm for the HB models.

Rebekah Dawson worked with detrending WG on testing and developing methods for detrending transit light curves, searching for planets, and developing a likelihood function for detection and fitting that is appropriate for the non-Gaussian, correlated noise in the Kepler light curves. She explored the properties of light curve noise in the time domain and wavelet domain and gained knowledge of Kepler instrumental noise from Tom Barclay, Jessie Christiansen, and Jon Jenkins.

She collaborated with Paul Baines, Mary Beth Broadbent, Josh Carter, David Hogg, Tom Loredo, and Robert Wolpert on developing a wavelet-based likelihood function and tested the assumption that the wavelet coefficients are Gaussian and uncorrelated on a given scale.

She worked with Ruth Angus, Hannu Parviainen, and Billy Quarles to test detrending light curves and searching for Earth-size planets on a set of planets injected into Kepler data by Dan Foreman-Mackey. She tested the performance wavelet likelihood in searching for and fitting planets. She searched for Earth-like planets in a subset of bright Kepler stars that fall on modules subject to relatively low instrumental noise, and are known to already host small planets on 10-200 day orbits.

She also started a collaboration with Mary Beth Broadbent to use reversible jump MCMC to distinguish between planet vs astrophysical false-positive light curve models.

Eric Ford worked with the populations modeling WG on the construction of hierarchical Bayesian (HB) models in order to characterize the orbital eccentricity distribution of extrasolar planets. He collaborated with Merlise Clyde, Tom Loredo and Robert Wolpert on eliciting appropriate priors, approximating the likelihood for accelerating MCMC-based HB.

He assisted Leslie Rodgers, Megan Shabram and Angie Wolfgang in developing HB models and implementing MCMC samplers to allow for rapid testing of hierarchcial models for the eccentricity distribution, mass-radius distribution and distribution of planet envelope fractions using various priors, generative models and likelihood functions, as well as exploring the consequences of truncation and censoring for such models.

Ford also worked with Jessi Cisewski, Darin Ragozzine, and remotely with Chad Shafer (Stat, CMU) on developing and testing distance metrics for applying Approximate Bayesian Computing to HB to characterize the eccentricity distribution.

David Hogg spent his time at SAMSI learning about Kepler systematics and probabilistic non-parametric models based on Gaussian Processes (GPs). The idea of the latter is that non-trivial time dependence of stellar variability signals can be captured by a Gaussian Process. Diagonalization of an arbitrary GP might be challenging, but since the standard forms for GPs are chosen heuristically and constitute relatively unjustified assumptions, there is no reason not to choose a sensible basis and choose the form of the GP that has a diagonal variance tensor in that basis. Work by the detrending WG suggested that a sensible wavelet transform would lead to light-curve statistics that are empirically consistent with independence of the right form.

By the end of his visit at SAMSI, Hogg and his student Dan Foreman-Mackey had been able to demonstrate that GPs can make a very precise probabilistic model of an exoplanet transit light-curve, and that exoplanet inference can be performed realistically in such a model. It remained an open question whether the GPs are fast enough to permit search of light-curves for potential periodic transit signals. The results of the search would be hypotheses to test with proper probabilistic inference.

On the Kepler systematics side, Hogg learned, again with the detrending WG, that the housekeeping data (meta data) recorded by spacecraft temperature sensors and focal-plane mapping systems are strongly correlated with spacecraft-induced variability in the calibrated light curves. Naive experiments to "detrend" the light-curves using the housekeeping data as "templates" were not successful: For one, the light-curves contain much more information than the housekeeping data, so the latter are not sufficient to "de-noise" the former. For another, the relationship between housekeeping quantities and light-curve deviations is not well explained by a stationary, linear, process, even within one relatively homogeneous quarter of observing time. These "failures" led to specification of a project in which light-curve data are used to improve the housekeeping data (if the housekeeping data can't de-noise the light-curves, then the light-curves can probably de-noise the housekeeping data), and a realization that a hierarchical (multi-level) model that generates the light-curves and housekeeping data simultaneously must perform better than models in which these data are treated separately.

In addition to this work with the detrending WG, Hogg participated in discussions of hierarchical inference with the population modeling WG, and parameter estimation with the modeling WG. Indeed, every project under discussion at the SAMSI workshop were of benefit to his work in some way. He also started a scientific paper on likelihood functions for Kepler data (based on the GPs) with the detrending WG and a note on marginalized likelihood and utility (related to a seminar given at SAMSI by Jim Berger).

Daniel Foreman-Mackey worked in several subgroups of the "systematics and de-trending" WG. Part of his time was spent working with Ruth Angus, Thomas Barclay, Gal Matijevic, and Ben Montet to develop algorithms for injecting synthetic transit signals into raw Kepler data to asess the performance of de-trending prescriptions and transit search methods.

Foreman-Mackey also collaborated with Paul Baines, Bekki Dawson, and David W. Hogg to develop flexible probabilistic models for simultaneously characterizing an exoplanet transit signal, stellar variability, and the poorly-understood systematic effects induced by the detector.

All collaborations resulted in open source software and the second collaboration led to drafts of two possible papers discussing the results of the experiments. In particular, a simple model for the likelihood function based on Gaussian Processes was shown to be a good model for the data and computationally feasible to compute.

Matthew Payne formed new collaborations with Robert Morehead and Meg Schwamb to investigate the reliability of transit detection methods for discovering long period planets, comparing human visual detection (Planet Hunters) with algorithmic detection. They developed methods to

1. Extract candidate long-period transits from multiple light-curve inspections by human volunteers;
2. Extract candidate long-period transits algorithmically;
3. Automate the precise fitting of all candidates from (1) and (2) in order to allow quantified comparison of the different methods.

He also collaborated with Mary Beth Broadbent on reversible-jump methods for distinguishing single transit events from multiple transit events, with particular thought on applying this to rigorously decide whether a given event is a planetary transit, or a planet plus a moon.

Billy Quarles: Quarles acquired several new insights into the detrending and transit search of Kepler data. This experience has allowed him to build an international collaboration with Ruth Angus, Hannu Parviainen, and Rebekah Dawson that will address the ability of existing techniques to find Earth-analogs within the Kepler data.

They also investigated statistical methods to improve the existing techniques using conversations with Paul Baines, Josh Carter, and David Hogg. Through this synergistic collaboration they intent to bring forth several publications that build upon the work performed here at SAMSI.

Angie Wolfgang: I have learned and accomplished quite a lot at SAMSI: not only have I applied Markov Chain Monte Carlo to a research problem for the very first time, but I learned about Hierarchical Bayesian Modeling and Approximate Bayesian Computing, neither of which I'd heard of, let alone used, before coming to SAMSI. Over the last three weeks I have developed a simple yet nontrivial hierarchical Bayesian model for the composition of Kepler's planets that are Neptune-sized and smaller, specifically looking at the mass fraction of gas in these planets; this project will culminate in a publication in an astronomical journal.

4.3 Short-term plans

The following will be responsible for setting up regular meetings and making sure that the WG collaborations continue beyond the Kepler Program.

1. *Light curve detrending and planet detection:* Bekki Dawson
2. *Exoplanet populations:* Darin Ragozzine, Matt Payne
3. *Long-period planets:* Meg Schwamb

The following publications are likely to result from the Kepler Program:

1. Eric Ford, Meg Shabram, *Hierarchical Bayesian analysis of exoplanet eccentricity distribution from occultations*
   (some progress, need to finish simulations and analyze results)
2. Tom Barclay, Eric Ford, *Hierarchical Bayesian analysis of exoplanet eccentricity distribution from transit light curves and astroseismology*
   (just started, need to improve MCMC mixing, need to analyze SC Kepler data)

3. Jessi Cisewski, Eric Ford, Meg Shabram, *Comparison of HB + MCMC versus HB + ABC for eccentricity distribution*
   (significant progress, need more code development)

4. Leslie Rogers, *Capabilities, limitations and biases in planet mass-radius relationship for planets with both transit and RV constraints*
   (significant progress, still developing models of increasing complexity)

5. Angie Wolfgang, *Planet volatile fraction from Kepler data and planet interior models*
   (just started, still developing tractable model)

6. Jesse Cisewski, Eric Ford, Darin Ragozzine, *An ABC approach to characterizing the architectures of planetary systems*
   (learned about basics)

7. David Hogg et al., *Likelihood functions for Kepler data based on GPs*

8. David Hogg, *A note on marginalized likelihood and utility*

4.4 Long-term plans & Goals

The following research topics and possible participants were suggested for collaborations in the long-term.

1. *Detrending at the pixel level, in the context of PSFD photometry or aperture photometry*
   Ruth Angus, Tom Barclay, Jessie Christiansen, Bekki Dawson, Jon Jenkins, Robert Wolpert

2. *Analysis of period-radius distributions*
   Angie Wolfgang

3. *Reversible Jump MCMC to minimize false positives in planet detection*
   Mary Beth Broadbent, Bekki Dawson

4. *Search for small planets: Picking stars, improving detrending, searching*

5. *Using the experience with Kepler data analyses to guide design choices for the TESS mission*