

# Future of High-Dimensional Data-Driven Exoplanet Science

**Eric B. Ford**

Center for Astrostatistics, Institute for CyberScience, Center for Exoplanets and Habitable Worlds, Department of Astronomy & Astrophysics, The Pennsylvania State University, 525 Davey Laboratory, University Park, PA, 16802, USA

eford@psu.edu

**Abstract.** The detection and characterization of exoplanets has come a long way since the 1990's. For example, instruments specifically designed for Doppler planet surveys feature environmental controls to minimize instrumental effects and advanced calibration systems. Combining these instruments with powerful telescopes, astronomers have detected thousands of exoplanets. The application of Bayesian algorithms has improved the quality and reliability with which astronomers characterize the mass and orbits of exoplanets. Thanks to continued improvements in instrumentation, now the detection of extrasolar low-mass planets is limited primarily by stellar activity, rather than observational uncertainties. This presents a new set of challenges which will require cross-disciplinary research to combine improved statistical algorithms with an astrophysical understanding of stellar activity and the details of astronomical instrumentation. I describe these challenges and outline the roles of parameter estimation over high-dimensional parameter spaces, marginalizing over uncertainties in stellar astrophysics and machine learning for the next generation of Doppler planet searches.

## 1. Introduction

Historically, the field of exoplanets has shied away from the challenges of high-dimensional data analysis. Typically, astronomers rely on physical intuition to guide data analysis, reducing large and complex astronomical data into a few observable parameters for further analysis. Only recently has it become practical for astronomers to analyze their hard-won data in its full glory. This presents exciting opportunities for statisticians and computer scientists to contribute to advancing humankind's knowledge of planets beyond our solar system. However, doing so will require careful attention to many "real-world" effects that complicate the measurement process.

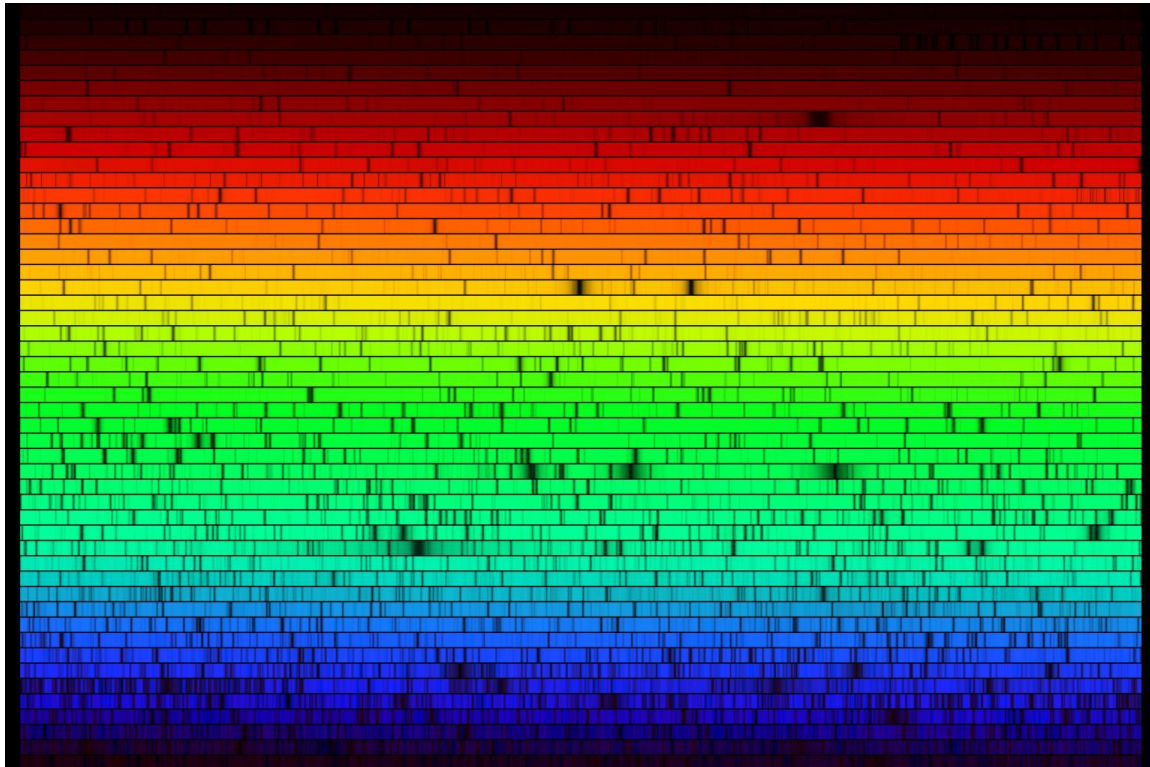
## 2. Doppler Planet Surveys

### 2.1. Doppler Observations

Doppler planet surveys aim to discover planets by repeatedly measuring the "radial velocity" of a star (i.e., stellar velocity projected onto the line-of-sight). For an isolated star, the star's velocity would be constant (relative to the solar system barycenter). For a star hosting one or more planets, the reflex velocity of the star about the planetary system's center-of-mass and its variation with time provides a mechanism for detecting and characterizing its planets. Each observation yields a



spectrum of the target star, including many absorption lines which provide information necessary to measure the star's velocity. The raw data from each Echelle spectrograph consists of one or more images measured with CCD chips, each containing several "orders" (rows) of the star's spectrum (see Figure 1). Astronomers take dozens or hundreds of spectra of a given target star, spread over days to decades. Discovering exoplanets requires performing differential radial velocity measurements, i.e., measuring very small shifts of the spectra lines, that correspond to just a few silicon atoms. Developing practical techniques to achieve the required Doppler precision and stability was a major accomplishment that enabled the discovery of exoplanets [1].



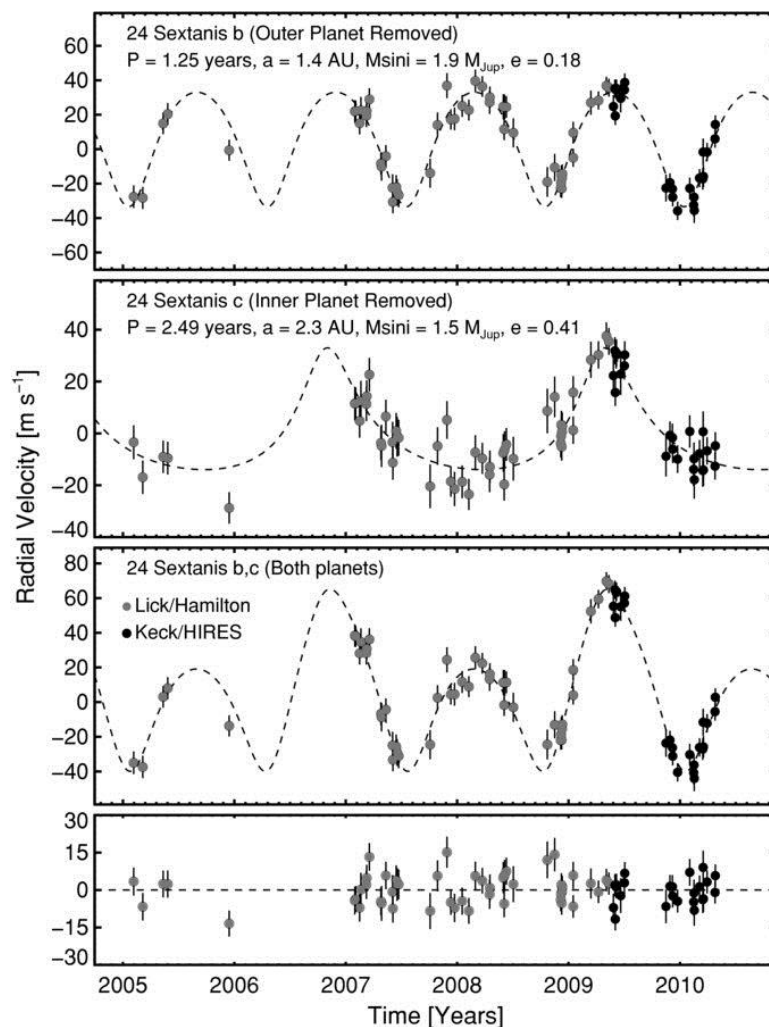
**Figure 1** High resolution spectrum of the Sun at visible wavelength. In each row, the wavelength varies by 6 nm, starting from 700 nm at the top and extending to 400nm at the bottom. Dark regions are due to stellar absorption lines, as the effects of Earth's atmosphere have been removed from this simulated image [2]. Real astronomical observations are further complicated by curvature of each row, non-linear wavelength scale, and non-uniform illumination as a function of wavelength. (Image Credit: N.A.Sharp, NOAO/NSO/Kitt Peak FTS/AURA/NSF.)

## 2.2. Bayesian Analysis of Doppler Observations

Existing Doppler survey rely heavily on astronomical expertise to reduce each spectrum into a few numbers, e.g., the time of observation, a differential velocity measurement of the star's radial velocity, and an estimate of the uncertainty in the velocity measurement. A time-series of such measurements can be analyzed to discover planets and characterize their masses and orbital properties (See Figure 2). The likelihood is derived by combining a physical model (e.g., a single planet traveling on a Keplerian orbit or a system of multiple planets interacting under Newton's laws of motion and gravity) and a simple statistical model (e.g., each observation includes an error modeled as uncorrelated, Gaussian noise). In a Bayesian context, the ideal priors would be based on the intrinsic distribution of exoplanet. Of course, those were completely unknown only a few decades ago. Astronomers typically adopt relatively weak priors, such as those developed during a 2006 program at the Statistical and Applied Mathematical Sciences Institute [3].

Fortunately, the choice of priors for orbital parameters typically has only a weak effect on posterior distributions.

A complete description of a planetary system would require seven parameters per body (mass, three position coordinates and three velocity coordinates at a reference epoch) plus any additional parameters to describe the observational setup. (E.g., if a star was observed by multiple instruments, then there may be offsets between radial velocity scales for each instrument.) However, several parameters are not important due to degeneracies (e.g., translations of the center-of-mass position of the planetary system, rotation of the orbital plane about the line of sight) or near degeneracies (e.g., the mutual inclination of orbital planes only affects the observed radial velocities if the planets' mutual gravitational interactions are detectable). Therefore, the observed radial velocity signature of most planetary systems with one planet or a few well-separated planets can be accurately modeled with approximately five model parameters per planet. For such systems, the star's motion can be modeled as a linear superposition of Keplerian orbits, allowing for very rapid model evaluation. In these cases, even simple Markov chain Monte Carlo algorithms provide a practical method for inferring posterior distributions for planet masses and orbits [4].



**Figure 2** Time series of measured radial velocities of the star 24 Sextanis [5]. The third panel from the top shows the original measurements, along with a two-planet model (dashed curve). The two panels above show the same measurements, after subtracting a model for either the outer (top) or inner (second from top) planet. The bottom panel shows the residuals to the two planet model. Reproduced by permission of the AAS.

When multiple planets orbital at similar distances from their host star, their mutual gravitational interactions can cause changes in their orbits. Such effects are particularly important when two or more planets have nearly commensurate orbital periods, such as the near 1:2 mean-motion resonance for the planets orbiting 24 Sextanis (see Figure 2). For such systems, accurately modeling the star's motion requires performing n-body integrations that are much more computationally demanding. In these cases, more efficient algorithms for Markov chain Monte Carlo are particularly valuable. Two ensemble samplers, differential evolution MCMC (DEMCMC) and affine invariant sampler, have found utility among astronomers [6,7]. They offer a reduced number of model evaluations and increased suitability for parallelization. With these algorithms, astronomers have been able to analyze even the most challenging of such extant data sets. For example, Nelson et al. (2016) combined DEMCMC with the Swarm-NG package for integrating an ensemble of planetary systems on a graphical processing unit (GPU) [7]. They applied the resulting code to analyze several of the most challenging planetary systems and Doppler data sets. For example, 1418 Doppler measurements collected from four observatories over more than two decades of observations (plus transit and inteferometric constraints) were included in an analysis of five planets orbiting 55 Cancri, including a super-Earth, Jupiter-analog and two planets near a 1:3 mean motion resonance [8]. In another example, they characterized the three dimensional architecture of a system of four planets orbiting GJ 876, including a super-Earth-mass planet and three planets near a 1:2:4 mean-motion resonance [9]. These analyzes required weeks of computation on GPUs to perform the posterior sampling over parameter spaces with ~30 to ~40 dimensions. In the case of GJ 876, the authors quantified the evidence for the fourth planet via Bayesian model comparison. After computing posterior samples conditioned on the number of planet, they constructed an importance sampling density for each model. When combined with the ratio estimator [3], this yielded robust estimates of Bayes factors and evidence for the four-planet model.

### 2.3. *The Challenge of Stellar Activity*

So far, we have assumed that astronomical observations directly measure the velocity of the host star. In practice, a variety of astrophysical effects (e.g., starspots, plage, pulsations and convection) can affect the star's spectrum, perhaps causing perturbations to the star's apparent radial velocity. We refer to these effects collectively as "stellar activity". To test whether stellar activity might be masquerading as a planet, astronomers often measure one or more astronomically-motivated activity indicators from each spectrum. Once one finds evidence for a statistically significant periodic variation of the measured radial velocity, it is wise to check whether there is a similar periodicity in any of the observed activity indicators. If so, then a cautious astronomer would refrain from announcing the discovery of a planet, out of concern that stellar activity might be responsible for the apparent radial velocity signal. Of course, sometimes indications of stellar activity are identified as part of follow-up analyses after the initial claims of planet detection [10,11,12]. In other cases, further analysis can provide evidence that a periodicity previously assumed due to activity is more likely a planet [11]. In some high-profile cases, different analyses have led to different interpretations about the most likely cause for an apparent Doppler signal [10,13,14,15]. These differing interpretations underscore the important of increased statistical rigor as the field of exoplanets pushes to identify increasingly Earth-like planets.

In the early days of Doppler planet searches, the measurement uncertainty for a star's radial velocity was "large" enough that stellar activity was a relatively minor concern. Once stellar activity could no longer be completely ignored, astronomers began treating stellar noise as "jitter", an additional source of uncorrelated Gaussian noise [4]. While this approach has served astronomers extremely well, the quest for Earth-mass planets, and particularly Earth-mass planets in the "habitable zone" of their host star, has motivated the development of increasingly precise

instruments. Several current spectrographs are so precise that stellar activity can be the dominant source of “noise”.

#### *2.4. Improving Doppler Surveys through Improved Modeling of Stellar Activity*

In order to realize the potential of current and next-generation spectrographs, astronomers must change their paradigm from treating stellar activity as noise to treating it as a signal that is included in their model [e.g., 15]. Unfortunately, modeling how astrophysics at the surface of a star affects the apparent radial velocity is extremely computationally intensive. While first-principles simulations can provide valuable physical insights, they will not be a practical tool for analyzing radial velocity observations for decades to come. Therefore, it is necessary to develop physically-motivated, semi-empirical models for stellar activity. For example, one could consider the starlight to be a linear superposition of light from many patches on a star’s surface, with each patch emitting the same spectrum, but shifted in wavelength (due to stellar rotation) and modulated in intensity (e.g., due to spots, plage, convection). While this represents a dramatic simplification, it still leaves a high-dimensional parameter space, since it requires specifying the intensity (or a proxy such as effective temperature) over the entire stellar surface and how it varies with time. Since the stellar rotation causes velocities order of magnitude larger than the gravitational perturbation of planets, even small star spots can cause an appreciable effect. Given the high contrast and sharp edges of star spots, they are not well-described by the obvious smooth functions (e.g., low-order spherical harmonics). Instead, one typically uses a high-resolution grid over the stellar surface (often approximated as a rotating sphere). Since distant main-sequence stars are not spatially resolved, solving the inverse problem - inferring the spot pattern from astronomical data - is impractical. Therefore, we must resort to marginalizing over uncertainty in the specific details stellar activity. The problem is further complicated by the temporal evolution of stellar activity. Star spots and plage can persist for timescales comparable to a stellar rotation period, but change in intensity and location. Thus, we must go beyond models consisting of a map on a rotating sphere. Developing models which are both astronomically accurate and sufficiently computationally efficient to enable marginalization over such a high-dimensional parameter space represents one major challenge in maximizing the scientific potential of next-generation Doppler planet surveys. Astronomers are making significant progress in developing empirically-motivated forward models [e.g., 16]. By performing such simulations for large ensembles of astrophysically plausible realizations of stellar activity, it may be possible to approximately marginalize over uncertainties in stellar activity. It will soon be possible to test such models using observations of the sun [17]. Since elements of the models were developed using data for the Sun, it will also be important to test such models on other stars. Verifying the accuracy of such models for stars other than the sun will be extremely demanding, astronomers must collect an unusually large number of extremely precise measurements.

#### *2.5. Improving Doppler Surveys through Improved Analysis of Spectroscopic Observations*

If the Doppler method is to discover Earth-mass planets in the habitable zone of sun-like stars, then astronomers must improve our ability to separate stellar activity from the Doppler signature of low-mass planets. So far, astronomers have used only a few indicators of stellar activity, each based on a very narrow range of wavelengths in the stellar spectrum. Further, these activity indicators were often developed for ease of measurement when computational resources were much more limited. Fortunately, there is reason for optimism. Remember, that each original observation contains megapixels of data. Rather than immediately reducing this down to measurements of the radial velocity and a few activity indicators, astronomers could analyze the full spectrum to extract additional information about the magnitude and nature of stellar activity. Just as combining measurements of the shift of many spectral lines allows for a much more precise measurement of the Doppler shift than is possible from any single spectral line, analyzing the full spectrum can provide information about changes in the shape of spectral lines that contain

information about stellar activity. Characterizing this relationship will be challenging. Machine learning algorithms can be applied to the large number of spectroscopic observations obtained as part of planet search program, as well as to simulated datasets based on semi-empirical models [17]. Some common statistical algorithms (e.g., principal components analysis, independent components analysis) have been applied to stellar spectra [e.g., 18] and suggest that future applications to planet search data may allow for a dramatic dimensional reduction from megapixels of data to a relatively small number of components that encode information about the extent, location and nature of stellar activity. Generalizing these machine learning methods to take advantage of physical intuition may provide additional benefits for separating stellar activity and the Doppler signature of low-mass planets.

### 3. Additional Roles for High-Dimensional Data Analysis in Exoplanet Science

#### 3.1. *Characterizing Exoplanets via Transit Timing Variations*

Recently, the method of transit timing variations has proven extremely powerful for measuring masses and densities of low mass planets. While the astronomical data differs significantly, the analysis methods developed for analysing Doppler observations of strongly-interacting planetary systems have proved invaluable for interpreting such transit timing observations [19]. The computational demands are considerable. The parameter spaces typically have a few dozen dimensions and each model evaluation is computationally more expensive than present analyses of Doppler observations. The resulting posterior distributions often have very strong and sometimes non-linear correlations, further complicating the sampling and interpretation of results. While current methodology is effective for exploring a posterior mode, further improvement in sampling algorithms will be important for accurately characterizing the wings of the posterior distributions for such systems (e.g., 99.9% credible regions) and for characterizing transit timing variations in systems where the perturbing planet is not observed to transit the host star.

#### 3.2. *Characterizing the Exoplanet Population*

So far, we have focused on the characterization of an individual planetary system. As the number of known exoplanets has grown, interest has shifted from individual planets to questions about the population of planets in general. A variety of detection limits and observational biases significantly complicate the interpretation of the overall population. During a 2013 SAMSI workshop on Modern Statistical and Computational Methods for Analysis of Kepler Data, several astronomers began applying hierarchical Bayesian modeling to characterize the exoplanet population [20]. Even when the number of parameters describing the properties of the population is relatively small, performing inference over the population often requires dozens, hundreds or thousands of model parameters. Thus, these analyses can be computationally demanding. It can be challenging to develop models with a minimal number of parameters, but sufficient flexibility to model observations accurately. Visualizing the results to assess convergence, robustness and draw conclusions is also challenging. While we have made great progress, we have a long way to go. I look forward to learning from statisticians and practitioners in other fields and to sharing insights with the exoplanet community.

### 4. Acknowledgements

E.B.F. thanks the organizers of the 2015 meeting on High-Dimensional Data-Driven Science in Kyoto, Japan for inviting this manuscript and for enabling his participation in the meeting. E.B.F. acknowledges discussions and correspondence with Jim Berger, Jessi Cisewski, Merlise Clyde, Allen Davis, Xavier Dumusque, Debra Fischer, Daniel Jontof-Hutter, Tom Lored, Paul Robertson, Sophia Araceli Sanchez-Maes, Robert Morehead, Benjamin Nelson, Darin Ragozzine, Megan Shabram, Sharon Wang, Angie Wolfgang, Robert Wolpert, Jason Wright and an anonymous referee. E.B.F. acknowledges support from the Pennsylvania State University's

Department of Astronomy & Astrophysics, Eberly College of Science, Institute for CyberScience, Center for Astrostatistics and Center for Exoplanets and Habitable Worlds. The Center for Exoplanets and Habitable Worlds is supported by the Pennsylvania State University, the Eberly College of Science, and the Pennsylvania Space Grant Consortium. This material was based upon work partially supported by the National Science Foundation under Grant DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute (SAMSI). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The results reported herein benefitted from collaborations and/or information exchange within NASA's Nexus for Exoplanet System Science (NExSS) research coordination network sponsored by NASA's Science Mission Directorate.

## References

- [1] Butler R P *et al* 1996 *Proc. Astronomical Society of the Pacific* **108** 500
- [2] Kurucz R L *et al* 1984 *National Solar Observatory Atlas* No. 1 (Sunspot, NM: NOAO)
- [3] Ford E B and Gregory P C 2007 Statistical Challenges in Modern Astronomy IV *Astronomy Society of the Pacific Conference Series* **371** 189
- [4] Ford E B 2005 *Astronomical J.* **129** 1706  
Gregory P C 2005 *Astrophysical J.* **631** 1198  
Ford E B 2006 *Astrophysical J.* **642** 505
- [5] Johnson J A *et al* 2010 *Astronomical J.* **141** 16
- [6] Ter Braak C J F 2006 *Stat. Comput.* **16** 239  
Foreman-Mackey D *et al* 2013 *Proc. Astronomical Society of the Pacific* **125** 306
- [7] Nelson B *et al* 2014 *Astrophysical J. Sup.* **210** 11
- [8] Nelson B E *et al* 2014 *Monthly Notices of the Royal Astronomical Soc.* **441** 442
- [9] Nelson B E *et al* 2016 *Monthly Notices of the Royal Astronomical Soc.* **455** 2484
- [10] Robertson P *et al* 2014 *Science* **345** 440
- [11] Robertson P and Mahadevan S 2014 *Astrophysical J. Letters* **793** L24
- [12] Robertson P *et al* 2015 *Astrophysical J.* **801** 79  
Robertson P *et al* 2015 *Astrophysical J. Letters* **805** L22
- [13] Vogt S S *et al* 2010 *Astrophysical J.* **723** 954  
Gregory P C 2011 *Monthly Notices of the Royal Astronomical Soc.* **415** 2523  
Baluev R V 2013 *Monthly Notices of the Royal Astronomical Soc.* **429** 2052  
Hatzes A P 2013 *Astronomische Nachrichten* **334** 616  
Anglada-Escudé G and Tuomi M 2015 *Science* **347** 1080  
Robertson P *et al* 2015 *Science* **345** 440
- [14] Dumusque X *et al* 2012 *Nature* **491** 207  
Hatzes A P 2013 *Astrophysical J.* **770** 133
- [15] Rajpaul V *et al* 2015 *Monthly Notices of the Royal Astronomical Soc.* **452** 2269  
Rajpaul V *et al* 2015 *Monthly Notices of the Royal Astronomical Soc. Letters* **456** L6
- [16] Dumusque X *et al* 2014 *Astrophysical J.* **796** 132
- [17] Dumusque X *et al* 2015 *Astrophysical J. Letters* **814** L21
- [18] Martínez González M J *et al* 2008 *Astronomy & Astrophysics* **486** 637
- [19] Carter J A *et al* 2012 *Science* **337** 556  
Jontof-Hutter D *et al* 2015 *Nature* **522** 321  
Jontof-Hutter D *et al* 2015 *Preprint* arXiv:1512.02003
- [20] Foreman-Mackey D *et al* 2014 *Astrophysical J.* **795** 64  
Rogers L A 2015 *Astrophysical J.* **801** 41  
Wolfgang A *et al* 2015 *Preprint* arXiv:1504.07557  
Shabram M *et al* 2015 *Astrophysical J.* in press (*Preprint* arXiv:1511.02861)