


5-31-2013

# Discovering Exoplanets through Hidden Markov Model Analysis

Jon Drobny

*Rose-Hulman Institute of Technology*

Follow this and additional works at: [http://scholar.rose-hulman.edu/undergrad\\_research\\_pubs](http://scholar.rose-hulman.edu/undergrad_research_pubs)

 Part of the [Applied Mathematics Commons](#), [Longitudinal Data Analysis and Time Series Commons](#), and the [The Sun and the Solar System Commons](#)

---

## Recommended Citation

Drobny, Jon, "Discovering Exoplanets through Hidden Markov Model Analysis" (2013). *Rose-Hulman Undergraduate Research Publications*. 4.

[http://scholar.rose-hulman.edu/undergrad\\_research\\_pubs/4](http://scholar.rose-hulman.edu/undergrad_research_pubs/4)

This Article is brought to you for free and open access by Rose-Hulman Scholar. It has been accepted for inclusion in Rose-Hulman Undergraduate Research Publications by an authorized administrator of Rose-Hulman Scholar. For more information, please contact [weir1@rose-hulman.edu](mailto:weir1@rose-hulman.edu).

Jon Drobny  
Project Advisor: Dr. Shibberu

## **Research Project Report: Hidden Markov Models and the Search for Extrasolar Planets**

Inspiration for a research project involving the search for extrasolar planets and Hidden Markov Modeling came while I was attending a course in Pattern Recognition with Hidden Markov Modeling. At this time, the Kepler Space telescope, launched in 2009, began discovering planets at a breakneck pace. A website called Planet Hunters was created to allow and encourage crowd-sourced visual analysis of data from the Kepler Space Telescope. With a background in Hidden Markov Modeling from the course and a strong interest in planetary astronomy and a suggestion from Dr. Shibberu, my professor of the Hidden Markov Modeling course, to apply for a Weaver grant, my course was set. My goal for the project was to develop a Hidden Markov Model for the detection and characterization of extrasolar planets through the analysis of light curves.

Before work on the project could begin, I had to gain a deep and thorough understanding of the geometry of the transiting planet problem. A transiting planet passes at some point in its orbit between its host star and the observer. As it passes in front of its host star, the observed flux from the star decreases in a characteristic way. From first contact, when the stellar and planetary discs begin to intersect, to second contact, when the planetary disc appears entirely within the stellar disc, the flux decreases as the planet blocks more of the star's light. Since the light emitted by a star is not constant across the stellar disc, the flux changes as the planet crosses the face of the star. Once the star reaches the opposite side of the stellar disc, where third and fourth contacts occur, the flux increases symmetrically. In general, a sequence of flux measurements from a star is called a light curve. For the transit method, the flux is ideally measured over a

period long enough to observe several transits. With some assumptions, a transiting extrasolar planet light curve can be simply modeled.

The first of these assumptions is that the planet that we are searching for is Earthlike. For a planet to be Earthlike, it must be small in comparison to its host star, have a nearly circular orbit, and have a relatively long period. With these assumptions, the light curve for a transiting extrasolar planet can be modeled as a two state system. The transiting state occurs when the planet is in front of its host star, and the non-transiting state occurs when the planet is not. Assuming that the flux change due to limb darkening across the face of the star is small, the transiting state and the non transiting state will each have a characteristic flux. Since the planet is small, the time between first and second and between third and fourth contacts will be small compared to the overall period (e.g., the Earth's transit time is on the order of hours, and its period is one year), so the system can change instantaneously from one state to the next. This two-state system can be modeled using Hidden Markov Modeling techniques.

Hidden Markov Modeling is a method of statistical analysis with applications in speech and handwriting recognition, bioinformatics, and economics. More generally, a Markov Model is a statistical system that transitions from one state to another, obeying the Markov Property. The Markov Property states that a system must be memoryless - that is, the next state of the system can depend only on the current state and not on the previous states. This property is useful in simplifying the algorithms used in the analysis of a Markov Model. In real-world applications however, the state of a system is not directly observable. To model a system with such hidden states, a Hidden Markov Model must instead be used. A Hidden Markov Model contains all of the properties of a Markov Model, with the addition of observable symbols emitted from each state according to a probability distribution.

Hidden Markov Model analysis consists of three major problems: scoring, or how to rate a model's accuracy, training, or determining the parameters of an HMM from data, and prediction. To score HMMs, a log-likelihood measure is used. The likelihood of a sequence of data given a certain model is defined as the probability of the model given the sequence. The log of the likelihood is used because likelihoods can differ by many orders of magnitude. The log likelihood cannot alone be used as a measure of accuracy for a model, but must instead be compared to the log likelihood of an alternative model. Prediction for HMMs comes in two flavors. First, given an HMM and a sequence of states, what is the most likely state at a particular point in the sequence? The other type of prediction is, given an HMM and a sequence of emitted symbols, predicting the most likely sequence of hidden states that could have produced the sequence of emitted symbols. An algorithm called the Viterbi Algorithm is used to calculate the most likely series of hidden states given a model and a sequence of emitted symbols. There are two types of training in HMM analysis, supervised and unsupervised training. Supervised training of an HMM can be done only when both the sequence of emitted symbols and the underlying sequence of emitted states are known. Unsupervised training, however, requires only knowledge of the sequence of hidden states, but is a significantly harder problem. Unsupervised training is done with either the Baum-Welch algorithm or maximum likelihood estimation.

The first model I applied to the problem of transiting extrasolar planets was a simple two-state model with observable symbols with a continuous, Gaussian probability distribution. The data from the Kepler telescope exhibited Gaussian noise. This model had the advantage of being exceptionally computationally simple to perform data analysis with, but suffered from two fatal weaknesses. Due to the assumption of an Earthlike planet, the model must remain in the non-

transit state for long periods of time. However, the probability distribution for the state duration of a HMM is geometric, and long state durations become increasingly unlikely, reducing the ability of the model to positively identify transiting extrasolar planet light curves. The second weakness of the two-state model is that it does not take advantage of the periodicity of the light curve of a transiting planet. There have been many attempts at modeling state duration with Hidden Markov Models that I researched during the course of this project. These include Semi Hidden Markov Models, Explicit Duration HMMs, and Variable Transition HMMs. Additionally, an approximation of the EDHMM exists that is known as the Expanded State HMM that approximates state durations by allowing the model to transition through a number of sub-states in each state, but because its state duration probability distribution is only a sum of many geometric distributions, it has only a limited ability to change the state duration probabilities.

The alternative that I turned to after investigating these models is known as a left right model. In a left right model, state duration is modeled explicitly with many states, each of which represent one state of the system, that transition only from left to right. For the transiting exoplanet model I designed, the non transit state was modeled by a left right model, but with the modification of an added small probability of remaining in the last non-transit state to account for variations in the length of the non-transit state. The transiting state is short enough compared to the total period that it can be well modeled by a single state. The final model has the following parameters orbital period, planetary radius, and stellar radius. Due to Kepler's laws, the orbital period defines the length of the non-transit state of the model. The planetary radius, in conjunction with the stellar radius, which can be found from direct or indirect measurements of

the host star, can be used to determine the depth of the transit.. The mass of the planet can then be calculated from radial velocity measurements.

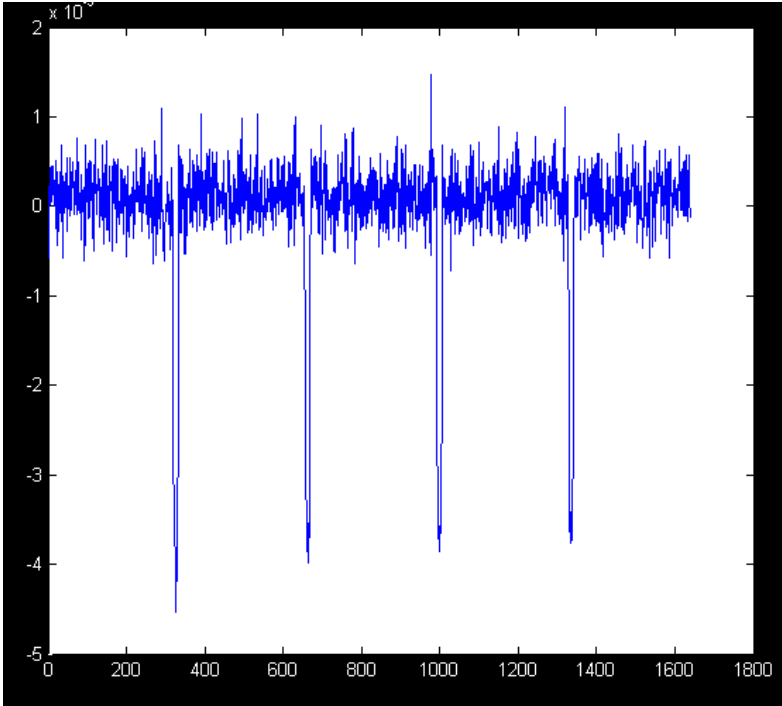
Before the model can be used to determine these parameters, it must be determined whether or not a light curve in fact contains planetary transits. This is done using log-likelihoods as described above. To pick out planetary transits, the following method was used to analyze a light curve. First a program was written to fit the model to a set of data given a certain period. A separate program performed this fit and returned the maximized log-likelihood for a range of possible periods. In this way a set of log-likelihoods over a range of periods can be computed. If the log-likelihoods are plotted, a clear spike in the log-likelihood strongly indicates a high likelihood of the existence of a transit with that period in the light curve. Interestingly, the log likelihood tends to increase as the period increases, whether or not there is strong planetary signal. Also of note are the harmonics that accompany the strong spike that indicates a planetary signal at integer multiples of the most likely period.

An inherent problem of the transit method is the prevalence of false positives. As mentioned before, false positives are a major problem with using the transit method. Many other astronomical phenomena can produce periodic variations in brightness. Stars often have an intrinsic variability which must be removed before attempting to search for transits, and this intrinsic variability is not easily modeled or filtered out. I attempted several methods to filter out intrinsic stellar variability, but because in general stellar variability is neither linear nor periodic, there is no general solution to this problem. Transit-like features can also be caused by eclipsing binary systems. Eclipsing binaries, interesting in their own right, have characteristic light curves with two alternating transits of different depths. The deeper of the two transits represents the larger of the two stars passing in front of the smaller, blocking out all of its light. The shallow

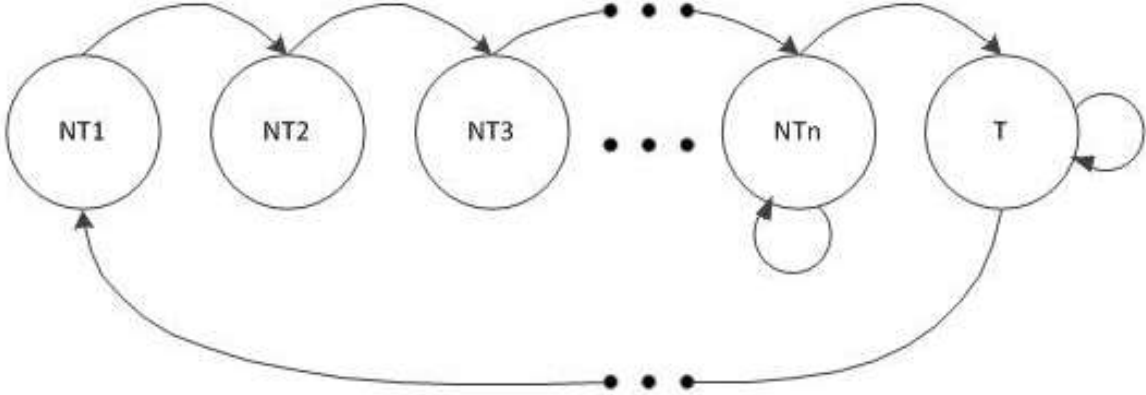
transit occurs when the smaller star passes in front of the larger one, blocking out some of the light of the larger star but contributing its own brightness to the light curve. A second problem that this model has is that it comes at a high computational cost. During this project, the analysis was done using the Dayhoff computer in the Rose-Hulman Theodrome running MATLAB. Even so, analysis of one set of data took as long as 10 hours to complete. Because the left-right model has as many states as there are time-steps in the orbital period, this computational time is very long, even when sparse matrices were used. In the future, I would like to investigate again the Explicit Duration HMMs. They were not used for this project due to the complexities involved in modifying the analysis algorithms, but since the left-right model is so computationally expensive, this initial investment could well prove its worth.

The model as constructed and used was able to identify the signal of a short period, transiting extrasolar planet that is not visible to the human eye. This power to pick up a small signal hidden in noise makes this method superior to crowd-sourced human-based efforts such as Planet-Hunters. Unfortunately, the massive computational expense that this method requires makes it less usable than I had hoped. The speed of the simple two-state HMM that inspired this effort did not scale with the addition of state duration to the HMM. Though the model does not live up to the optimistic initial predictions of speed, it is very powerful at sifting out noise and identifying a planetary signal. The characterization of a planet can be done with the orbital characteristics calculated from the fit model with the highest local log-likelihood score. Due to the extreme computational cost for this model, I have only been able to test it on short-period planets that are not Earthlike. As stated above, the other methods of including state duration in Hidden Markov Modeling may prove to be more applicable to this problem. A discussion of my project and its results was presented at the Rose-Hulman 2013 Undergraduate Math Conference.

Figures and Graphics

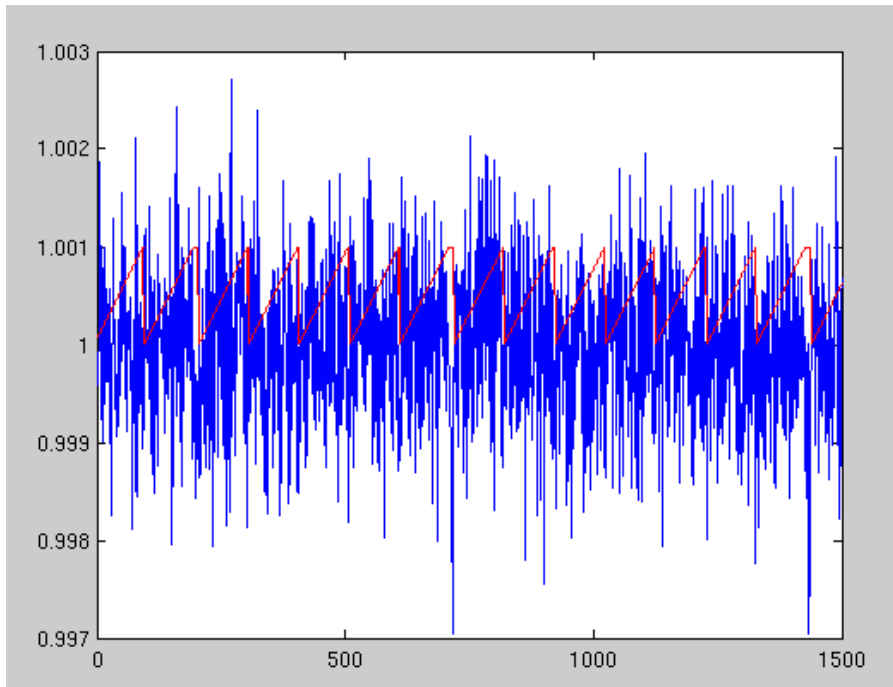


This graph represents a short-period light curve from the Kepler Space Telescope.

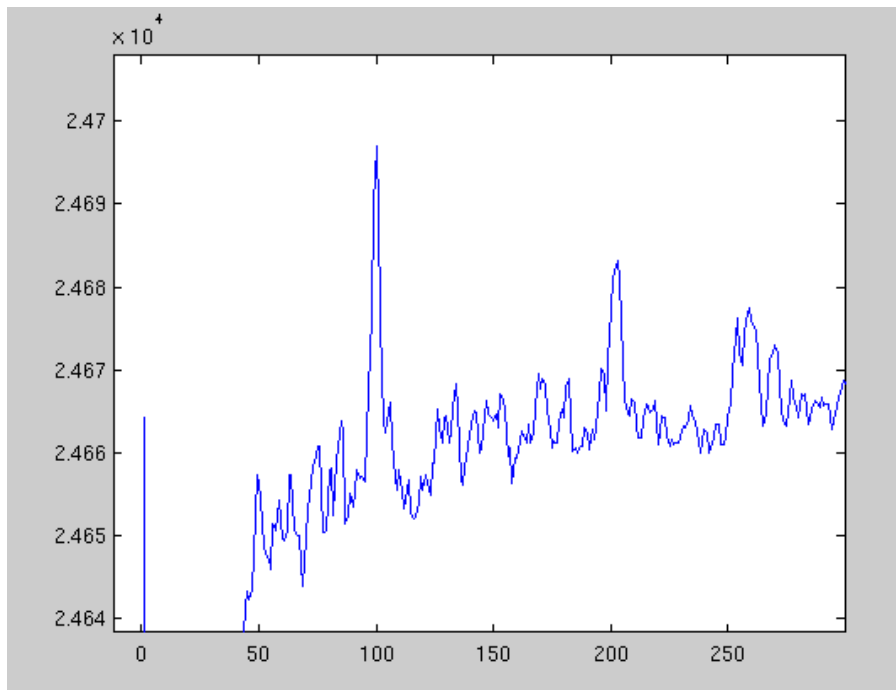


This figure represents the final, continuous, semi-left-right Hidden Markov Model developed and used in the project.





This graph shows a noisy light curve characterized by the final model shown above. The transits are not visible, but the state of the model showing identified periodic transits is shown in red.



This is the log likelihood plot of the above light curve, showing a positive identification of transit-like features with a period of approximately 100 time units, or approximately 2 days.