PHSX815_Project1: Distinguishing between instrumental noise and astronomical signals

Ryan Low

February 2021

1 Introduction

Modern astronomy relies on Charged Coupled Devices (CCDs) and other such imaging sensors for recording astronomical data. All of these technologies rely on photons exciting the electrons in some semiconducting material. Counting those electrons becomes a proxy for the number of photons detected. Because of this, recording astronomical data is a counting problem, and thus we can expect the number of photons recorded on a CCD to be distributed as a Poisson distribution. As with all electronic measurements, we must also be aware of sources of noise. Since the noise appears in our counts, we can also expect it to be distributed as a Poisson distribution. We will consider an idealized system of a single pixel and investigate whether we can distinguish between a noise source with rate parameter λ_{noise} and some astronomical source with rate parameter λ_{star}.

2 Problem Statement

Suppose we can successfully characterize the noise level for a CCD. In practice, we can do this by taking two types of calibration frames: bias frames and dark frames. Bias frames are zero second exposures that characterize the readout noise and the random bias signal present on the CCD. Dark frames are similar. Ideally, we expose the CCD to zero light for a fixed length of time. This allows us to measure the dark current on the CCD. Suppose that by taking these calibration frames, we find that $\lambda_{noise} = 10$ counts per second. Given some faint astronomical source with rate parameter λ_{star}, our task is to see whether we can distinguish between the noise or the source. To do so, we will perform simulations of measurements for a various number of observations. We will then choose some significance level, α, to perform hypothesis testing at. From this, we will be able to find the false negative rate β. Let our null hypothesis be that we are observing noise. We want to reject the null hypothesis as much as possible when we observe a source. Therefore, for a fixed number of observations, exposures per observation, and significance level, we want to find the λ_{star} such that our false negative rate is less than 1%. In practice, it is time inexpensive to take calibration frames since you can take them before it is dark outside. However, it is time expensive to take repeated exposures of the same source, since night time is limited. Therefore, we want to minimize the amount of observation time if possible. Hereafter, I will refer to the individual exposures as measurements and each observation as an experiment. Let's suppose we can perform 1000 experiments with 10 measurements each. We will arbitrarily set the significance level to $\alpha = 0.01$.

3 Algorithm Analysis

For this analysis, we generate simulated data using Python code. Since photon counting is a Poisson process, our simulation must be able to generate Poisson-distributed random numbers. We produce Poisson deviates using the implementation from section 7.3.12 in [1]. This method produces Poisson deviates from uniform deviates using the ratio-of-uniforms method, the product-of-uniforms method, and the rejection method [1].

For our test statistic, we use the Log Likelihood Ratio (LLR). The LLR is calculated using Equation 1.

$$\text{LLR} = \sum_{i}^{N} \log\left(\frac{P\left(x_i|H_0\right)}{P\left(x_i|H_1\right)}\right) \tag{1}$$

Where the probabilities are the Poisson probabilities of obtaining the measurement x_i for each hypothesis. The Poisson probabilities are given by Equation 2.

$$P\left(x|\lambda\right) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{2}$$

While easy to write down, this is incredibly difficult to calculate on a computer due to how large $x!$ becomes. In order to calculate the probabilities for large x, we take the logarithm

$$\begin{gathered}
\ln P = \ln\left(\lambda^x\right) + \ln\left(e^{-\lambda}\right) - \ln\left(x!\right) \\
\ln P = x\ln\lambda - \lambda - \ln x! \\
\ln P = x\ln\lambda - \lambda - \ln\Gamma\left(x+1\right) \qquad (3)
\end{gathered}$$

The Python package scipy includes a function for numerically calculating the log-gamma function. Therefore, we can easily implement Equation 3 in code while keeping the numbers reasonable, then obtain the probability by taking the exponential.

Once we have calculated the LLR for each dataset, it is convenient for the hypothesis analysis to sort the data. In general, if we know how the test statistic,

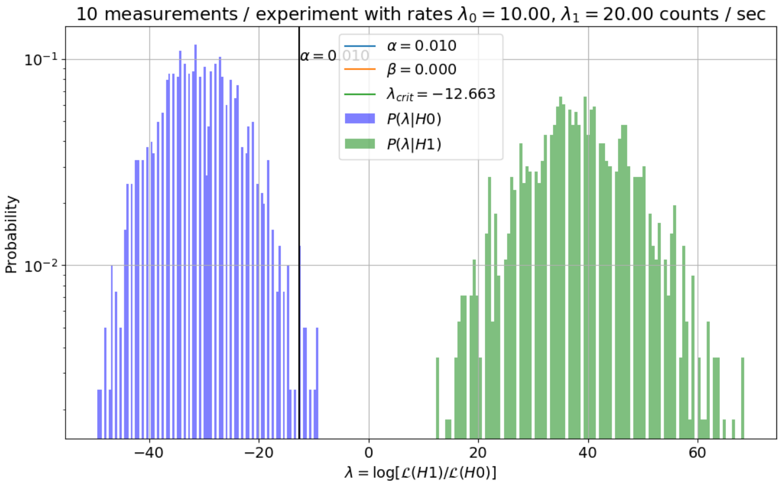

Figure 1: Histograms of the LLR comparing $\lambda_{noise} = 10$ and $\lambda_{star} = 20$.

λ, is distributed given that one of the hypotheses is true, the significance level, α, is obtained when

$$\int_{\lambda_\alpha}^{\infty} d\lambda \, P\left(\lambda|H_0\right) = \alpha$$

This gives us the critical value of λ, λ_α. From this, false negative rate is then

$$\beta = \int_{-\infty}^{\lambda_\alpha} d\lambda \, P\left(\lambda|H_1\right)$$

If our data is sorted in two arrays, one for H_0 and one for H_1, we can find λ_α and β without having to do numerical integration. λ_α is the λ at $(1 - \alpha)$ percent of the way through the H_0 array or at its end. Knowing λ_α, we can find all values in the H_1 array that are less than λ_α. β is the percentage through the H_1 array that is below λ_α. For our purposes, Python's default sort implementation is sufficient.

4 Analysis of Results

We ran simulations at different rate parameters ranging from 10 to 20 counts per second. As stated in Section 2, we will use $\lambda_{noise} = 10$ counts per second as the noise dataset, with the rest of the simulations representing the stellar dataset. The hypothesis testing plots for $\lambda_{star} = 20$ counts per second is presented in Figure 1. We see that even at $\lambda_{star} = 20$ counts per second, the two distributions are perfectly distinguishable. Adjusting λ_{star} by integer steps, the distributions do not begin to cross until $\lambda_{star} = 16$ counts per second. The criterion that

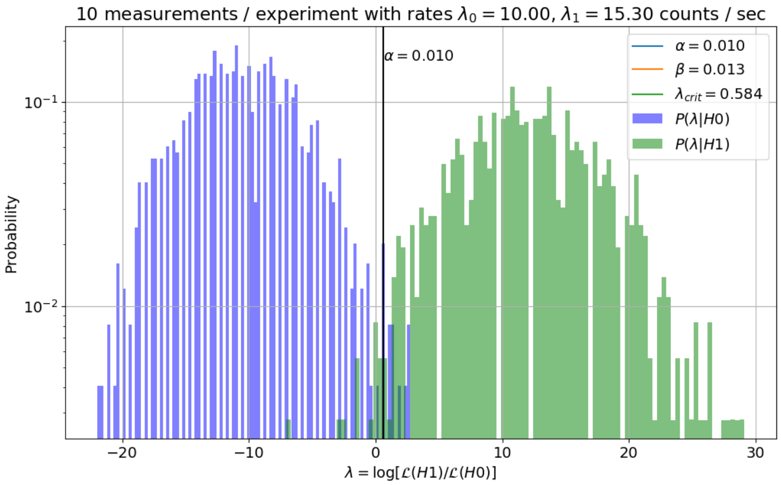

Figure 2: Histograms of the LLR comparing $\lambda_{noise} = 10$ and $\lambda_{star} = 15.3$.

$\beta < 1\%$ is not violated until $\lambda_{star} = 15.3$ counts per second (Figure 2), where $\beta = 1.3\%$. Below this rate, the β begins to increase substantially. When $\lambda_{star} = 11$ counts per second, $\beta = 92\%$. Therefore, we must be careful when observing sources that are very close to the noise level. However, away from the noise level, we should expect to be able to distinguish between signals and noise very well. This is a very important consideration when searching for low magnitude objects on the sky, where discovering new objects depends on observing near the magnitude limit of an instrument. Such surveys need to carefully characterize the noise level of their instruments and run a similar analysis to determine at what brightness they can successfully detect astronomical sources.

5 Conclusions

By modeling the measured signal and noise of a CCD as a Poisson process, we have shown that by taking many observations, we can successfully distinguish between signals and noise with $\beta < 1\%$ with λ_{star} close to λ_{noise}. However, as λ_{star} approaches λ_{noise}, the LLR distributions rapidly overlap, making distinguishing between the signal and noise difficult. This kind of modeling is useful in informing survey design, where the discovery of new objects relies on making observations close to the noise level of the instruments.

References

[1] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, USA, 3 edition, 2007.