From Z-Machines to ALMA: (Sub)millimeter Spectroscopy of Galaxies
ASP Conference Series, Vol. TBD, 2006
A. J. Baker, J. Glenn, A. I. Harris, J. G. Mangum, and M. S. Yun, eds.

A Statistical Method for Identifying Lines in Wideband Spectra

Stephanie G. Zonak[1], Andrew J. Baker[2,1], Andrew I. Harris[1]

[1]*Department of Astronomy, University of Maryland, College Park, MD 20742-2421*

[2]*Jansky Fellow, National Radio Astronomy Observatory*

Abstract.
We present a method and model results for determining the probability that a spectral line is present within a set of data. In wideband spectroscopy the noise across a spectrum varies significantly with frequency, complicating the usual methods for calculating noise. As an alternative to determining the noise across frequency bins in an averaged spectrum, it is possible to analyze the time sequence of signals to estimate the mean and uncertainty in the mean for each bin. A detection is then defined as a channel or set of channels whose mean is statistically significantly higher than those of its neighbors, making it possible to construct a plot of the confidence that a line is detected versus frequency. Calculating amplitude uncertainties for each channel also establishes errors on line intensities regardless of gain and noise variation with frequency. This method of spectral analysis is useful for wideband spectrometers such as the *Zpectrometer* (Harris, Baker & Jewell 2004), which will be installed on the Green Bank Telescope next year.

1. The Problem

Noise changes appreciably across the band for wideband spectra due to changes of receiver temperature and atmospheric opacity. As an example, Figure 1 plots the Ka-band receiver temperature as a function of frequency for the 100-m diameter Robert C. Byrd Green Bank Telescope and a synthetic spectrum with noise that reflects its change across the band.

The change in noise across the band complicates the process of estimating the significance of a line detection and setting error limits on the line's strength. Here we explore a method that uses the time series of data from spectral subscans to measure the noise and amplitude on a channel-by-channel basis as a function of frequency. Using this information, we rely on sub-regions of the wideband spectrum to evaluate the detection of a line. This approach will be applied to the data taken from an ultrawideband spectrometer, the *Zpectrometer*, being constructed at the University of Maryland for use at the GBT.

2. The Method

Our method entails two general steps:

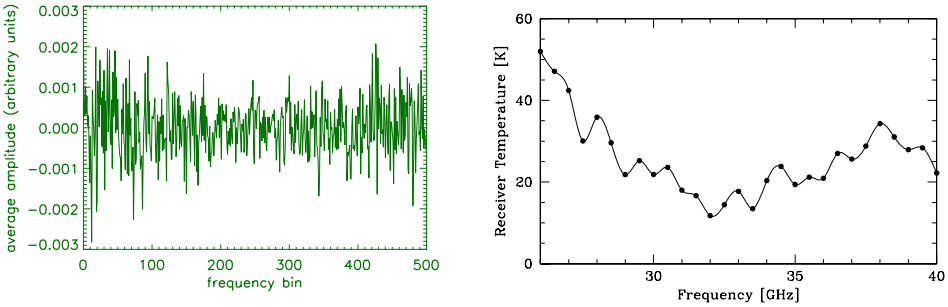

Figure 1. The time-averaged synthetic spectrum (left) was created using the measured receiver temperature as a function of frequency for the GBT (right).

1. We compare the mean amplitude of each bin with the average of the mean amplitudes of its neighbors. We declare a detection of an emission line if the bin's amplitude exceeds that of its neighbors by a statistically significant amount. The Student's t statistical test we employ depends on the estimates of the mean value and the variance in each bin which we calculate from the time sequence of subscans.

2. Instead of having the observer estimate the likelihood of a detection by looking at the signal to noise ratio in a spectrum of amplitude versus frequency, we directly plot line detection confidence versus frequency.

We begin by estimating the signal to noise ratio, defining the signal as the difference in amplitude between a bin and the mean of its neighbors, and the noise as a weighted sum of individual channel variances. This combination is insensitive to slowly-varying changes in amplitude, such as baseline structure, and accounts for changes in noise across the band.

Because both the mean and variance are estimated from the data, we use the Student's t distribution to describe the error in the calculated mean. This distribution is appropriate where the noise is normally distributed with the same variance for all samples (Ross 2004). As Figure 2 below shows, Student's t distribution has broader wings than a normal distribution, an important point when assessing the probability that a fluctuation has produced an extreme value at random.

We begin with a null hypothesis, meaning that we assume there is no line present, and mark a feature as a detected line if the test fails badly. The null hypothesis is that the true (rather than measured) mean of a channel μ_X is equal to the true local mean $\frac{1}{M}\sum_{i=1}^{M}\mu_{Y_i}$ (the mean of mean values μ_{Y_i} of the subscan-averaged channels surrounding channel X):

$$\Delta = \mu_X - \frac{1}{M}\sum_{i=1}^{M}\mu_{Y_i} = 0 \tag{1}$$

where M is the number of neighboring channels averaged over.

Although ideally the signal to noise would increase as $\sqrt{M}$, the number of neighbors to average over cannot be increased indefinitely. Because the noise

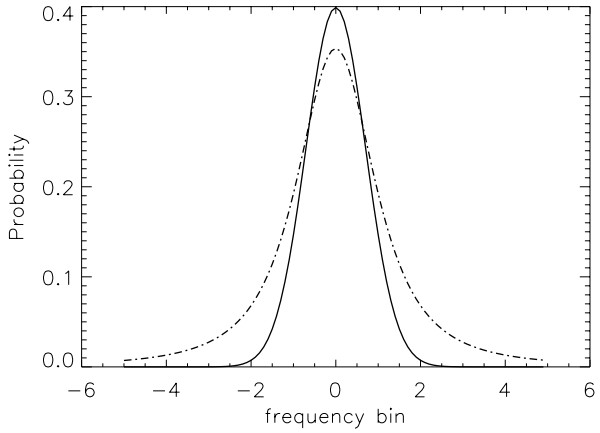

Figure 2. Shapes of Student's t distribution (dot-dash line) and a normal distribution (solid line). The broader wings of Student's t distribution arise from the extra parameter (the variance) that is estimated from the data.

structure changes across the band the estimated noise of the local mean (the mean of the neighboring channels) will no longer be representative of the noise in bin X if M is too large. This limit is determined experimentally by changing M until the greatest signal to noise is achieved.

Our test statistic d is a generalized version of the common two-sample Student's t test (Harris 2005) for determining whether the measured mean $\bar{X}$ averaged over N subscans and local mean averaged over M neighbors have statistically different means.

$$d = \frac{\bar{X} - \frac{1}{M}\sum_{i=1}^{M} \bar{Y}_i}{\sqrt{\frac{1}{NM}(S_X^2 + \sum_{i=1}^{M} S_{Yi}^2)}} \tag{2}$$

The numerator is the difference between test channel mean and the local mean. The denominator describes the weighted noise associated with these frequency bins: S_X is the standard deviation estimate from N subscans for the test channel, and S_{Y_i} are the corresponding standard deviation estimates for the noise in the neighboring bins. With this formulation the probability of a deviation d due to noise alone is given by a Student's t distribution with $(M+1)(N-1)$ degrees of freedom, or $T_{(M+1)(N-1)}$.

Fluctuations alone are unlikely to produce large values of d. If the chance, compared with the cumulative distribution of $T_{(M+1)(N-1)}$ of finding the measured quantity d is much smaller than one over the number of channels in the spectrum, then the null hypothesis fails and the likelihood of a line detection is high.

Figure 3 shows results from a synthetic data set. A comb of test lines with equal intensity tests the method across the band. The synthetic spectrum is from 60 subscans each and 500 channels, with noise shape matching that of the GBT's Ka-band receiver. Two artificial broad and time-varying baseline components were included. The top plot shows the average spectrum, the bottom a confidence plot for detection. For a spectrum with 500 channels, on average there should be one statistically significant noise spike per spectrum. The prob-

ability that one of the 500 bins will contain this noise spike is $1/500 = 0.002$ or $\log(0.002) = 2.7$ on the confidence plot. This is illustrated in Figure 3, where the bottom plot is renormalized so that a true value of 2.7 corresponds to 0. Therefore, on average, one point per spectrum should fall close to the horizontal line in the confidence plot by chance.

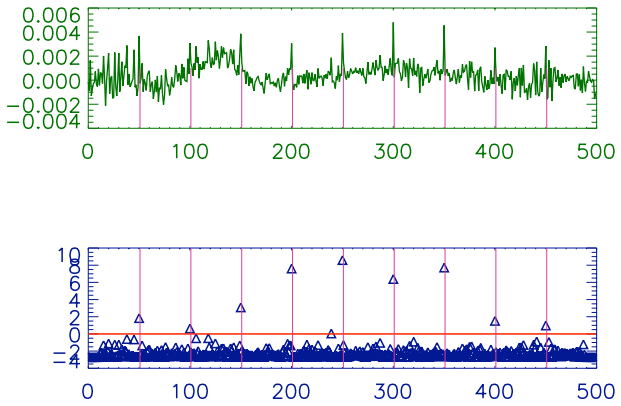

Figure 3. Top plot: average spectrum from 60 subscans. Each subscan has random noise scaled to match that of the GBT Ka-band receiver noise, two time-variable baseline structure components, and a regular array of identical spectral lines. The lower plot shows the neagaitve logarithm of the probability by which each channel is inconsistent with the null hypothesis (no line). Each spectrum contains 500 channels, with the solid line at zero indicating the level at which one channel in 500 will have a chance fluctuation. For noise alone, values below the line should occur often, while values above should occur rarely; a 0.002 percent chance of a fluctuation from noise corresponds to 0.0 on this scale.

Because of the variation of the receiver noise across the spectrum the spectral lines we introduce are easily detected in some parts of the band and are marginally detected toward the edges. We ran a grid of cases where the strengths and locations of the spectral lines were the same and the shape of the noise due to the receiver temperature was the same, but the the time-varying noise from other sources is different. We found that the success of our method is dependent upon the noise shape, but that in most cases our test consistently picked out the artifical lines even when they were not obvious by eye.

Acknowledgments. This work is supported by NSF grant AST-0503946

References

Ross, S.M., 2004, *Introduction to Probability and Statistics for Engineers and Scientists*, 3rd ed. Elsevier Academic Press, Burlington, MA

Harris, A. I., Baker, A.J, &Jewell, P.R., 2004. "*Zpectrometer*: An Ultra-wideband Spectrometer for the 100-meter Green Bank Telescope", NSF ATI program proposal

Harris, A. I., 2005, "Notes on spectral line detection: statistics for the amplitude difference between one spectral channel and the weighted sum of others", unpublished