QUANTIFYING CORRELATIONS BETWEEN FREQUENCY BINS

EDGES MEMO

Steven G. Murray (on behalf of the EDGES collaboration)
School of Earth and Space Exploration
Arizona State University
Tempe, AZ
steven.g.murray@asu.edu

January 10, 2022

ABSTRACT

This memo quantifies the amount of correlation between adjacent frequency bins in EDGES spectra, both for pxspec and fastspec (used from ∼2020 onwards), and has recommendations for frequency-averaging to enable simpler independent noise models.

1 Introduction

While incoming radiation measured by our antenna is in principle normally distributed and independent between differing frequency channels, the spectrometer introduces correlations between the frequency channels by performing finite-range FFTs over the incoming time-series of voltages (thereby introducing a window function).

The exact form of this correlation is slightly dependent on the inherent spectrum of incident radiation, but is mostly dependent on the specifics of the spectrometer – the number of channels being combined, the window function it employs, etc. Since these settings are relatively constant between different measurements, it should be possible to do a single estimation of this correlation and use it for future measurements. One caveat to this is that in 2020 we upgraded our spectrometer from a basic windowed-FFT to a polyphase-filter-bank (PFB) approach. That is, we upgraded from `pxspec` to `fastspec`.

In this memo, I will outline a specific case for why knowing this correlation is important, and give rule-of-thumb procedures for avoiding its downfalls for both of the spectrometers mentioned above. It should be noted that any significant modification to the settings of the spectrometer (or a new spectrometer algorithm altogether) in the future would require a re-analysis in line with what we do below.

2 The effect of frequency correlations on a noise model

We are commonly concerned with estimating models that are functions of frequency, say $m(\nu|\theta)$, given our spectrum data, $d(\vec{\nu})$. Since our data is Gaussian, this leads to a likelihood:

$$\mathcal{L}(\theta|\vec{d}) \propto |\mathbf{\Sigma}| \exp\left(\vec{r}^T \mathbf{\Sigma}^{-1} \vec{r}\right), \tag{1}$$

where $\vec{r} = \vec{d} - m(\vec{\nu}, \theta)$ is the model residual. Here, $\mathbf{\Sigma}$ is the $N_\nu \times N_\nu$ model covariance matrix contains the correlations between different frequency channels. It is here that these correlations must be known, or else the likelihood is incorrect.

Nevertheless, in practice, our spectra only contain correlations between frequency channels that are close to each other, and therefore everything away from the diagonal of $\mathbf{\Sigma}$ is zero. Performing a full $\sim 8000 \times 8000$ matrix inversion is costly, especially in comparison to a simple diagonal matrix inversion. One wonders whether we could simply ignore the off-diagonal terms, since they should be reasonably small, and most of them zero. This would result in a very simple and efficient likelihood:

$$\mathcal{L}(\theta|\vec{d}) \propto \exp\left(\sum_i r_i^2 / \mathbf{\Sigma}_{ii}\right). \tag{2}$$

It is not clear that this approximation is reasonable, however. One way to assess whether this is a reasonable approximation is to average data over adjacent channels, i.e. compute

$$\bar{d} = \frac{1}{N}\sum_{i}^{N} \vec{d}_i, \tag{3}$$

where N is the number of channels being averaged together (in the same spectrum measurement). We can then compare the variance of $\bar{d}$ with theoretical expectation under the assumption of diagonal covariance. That is, the variance of $\bar{d}$ is

$$\mathrm{Var}(\bar{d}) = \frac{1}{N^2}\sum_{ij}^{N} \mathbf{\Sigma}_{ij}. \tag{4}$$

If frequencies were independent, we'd have

$$\mathrm{Var}_{\mathrm{ind}}(\bar{d}) = \frac{1}{N^2}\sum_{i} \mathbf{\Sigma}_{ii}, \tag{5}$$

$$= \sigma_d^2/N, \tag{6}$$

where in the second equality we've also assumed that frequency channels within the bin have essentially the same expectation and variance[1]

The ratio of the true variance to the independent variance gives a measure of how independent the channels are, i.e.

$$\xi = \frac{\frac{1}{N^2}\sum_{ij}^{N} \mathbf{\Sigma}_{ij}}{\sigma_d^2/N} - 1 \tag{7}$$

$$= \frac{\sum_{i\neq j} \Sigma_{ij}}{N\sigma_d^2}. \tag{8}$$

We expect $\xi = 1$ for uncorrelated samples. Let's imagine that the off-diagonals of the covariance are positive and non-zero for *one* channel (i.e. just the diagonal next to the main diagonal), and that the value along this off-diagonal is constant, σ^2_{off}. Then we'd have

$$\xi = 2\frac{\sigma^2_{\mathrm{off}}}{\sigma_d^2}, \tag{9}$$

i.e. the ratio converges to a positive quantity even for large N. This is telling us that the noise estimated even from a large bin, under the assumption of independence, is incorrect with respect to the true variance of that bin.

3 Reducing Correlations By Binning

Eq. 9 does not, however, tell us whether such a bin would be significantly correlated with the next adjacent bin. Intuitively, bigger frequency bins mean that 'most' of the data in one bin is not correlated with 'most' of the data in the next bin.

The proper measure of whether a given (averaged) bin is correlated with the next bin is to take the ratio ξ where the data is already binned. That is, let $\bar{d}_{i,N}$ be the i^{th} bin of data, where each bin is formed from N averaged channels. Then form the ratio

$$\xi_2(N) = \frac{\mathrm{Var}\left(\bar{d}_{i,N} + \bar{d}_{i+1,N}\right)}{\frac{1}{2}\mathrm{Var}(\bar{d}_N)} - 1. \tag{10}$$

Here again we assume that the variance of the adjacent bins is essentially the same [2].

For intuition, again assuming that only the first off-diagonal is populated with a constant positive value, we get

$$\frac{N\sigma_d^2 + (2N-1)\sigma^2_{\mathrm{off}}}{N\sigma_d^2 + 2(N-1)\sigma^2_{\mathrm{off}}} - 1 \tag{11}$$

which tends to zero as $N \to \infty$. That is, wider bins will tend to be uncorrelated if the channel-to-channel correlations are localized.

[1] This of course is not true in general – an inherent spectrum that is like a power-law will have different expectations every bin, but these differences should be very small for the fine channels we consider.

[2] This requires that N be not so large that the bins have appreciable bandwidth

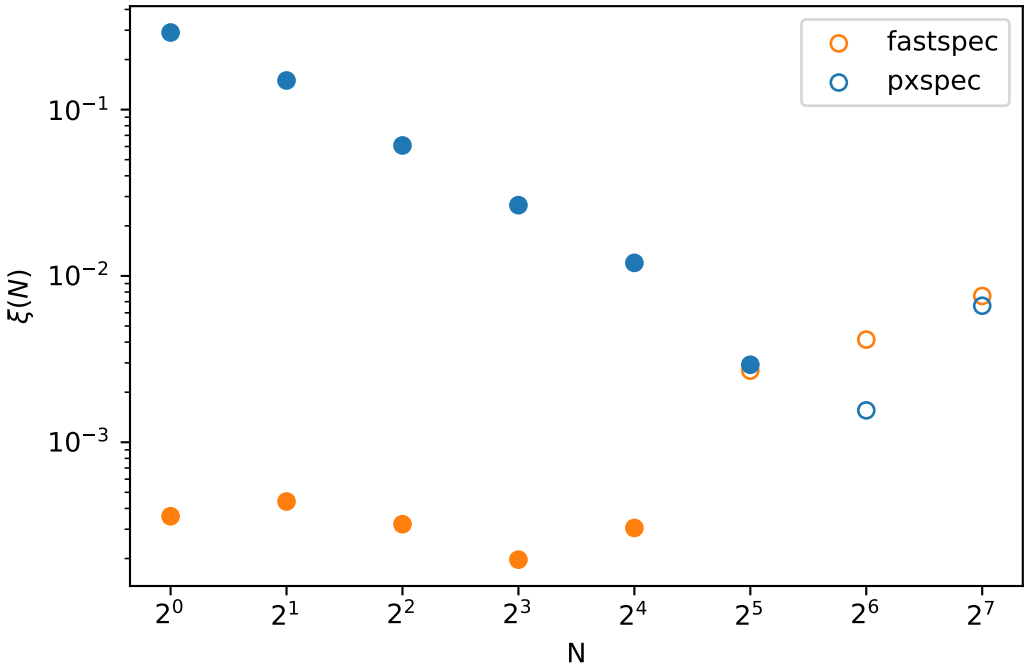

Figure 1: The ratio $\xi(N)$ for `pxspec` and `fastspec`. Open circles are negative values.

4 How Big Do The Bins Have to Be

Now, let us calculate $\xi_2(N)$ for real data. Since the correlation is expected to be *roughly* the same for different spectra, as long as the spectrometer is the same, we content ourselves to do the analysis on two measured spectra: one for `pxspec` and another for `fastspec`. For both, we use calibration observations, and just the ambient load spectra.

We discard the first 7% of spectra taken to allow time for warmup, and for both only use frequencies from 50-100 MHz. For `pxspec`, we use observations from `Receiver01_25C_2015_09_02`, while for `fastspec` we use `Receiver01_25C_2020_08_25`. To get the best estimate of the actual incident spectrum, we first calibrate the ambient load in each case (using the respective best-fit calibration from each observation).

To estimate the variance of a channel or bin, we simply use the sample variance along the time axis. This assumes stationarity of the signal over time, but this is well-supported by other tests. The numerator of ξ_2 is simply the variance of a bin two times the width of the denominator, where the binning is simple and unweighted.

We then plot the *mean* of $\xi(N, \nu)$ over all the frequency bins in the spectrum (remembering we just use from 50-100 MHz). The result is in Fig. 1.

The basic result is that for `pxspec` we should be using bins of $2^4 = 16$ channels to incur a $\sim 1\%$ error in the noise model (if assuming independence). For `fastspec`, we can ostensibly use the full resolution.

Note that for both spectrometers, the magnitude of the correlation actually goes up for still larger bins (though it is negative). This may be a consequence of having bins large enough to trace different parts of the spectrum that aren't homogeneous. It also may be telling us that there are negative correlations at some scale length in the spectrometer.

5 Takeaways

- Whenever fitting a model to spectral data, potential correlations between frequency channels must be accounted for to achieve reliable fits.
- In order to assume independence of adjacent bins, one can average channels within bins of some size.

- Fastspec seems to mitigate this issue with an error in the estimated variance of a binned spectrum of the order of 4×10^{-4} even for a single channel.
- `pxspec` requires binning to 16 channels per-bin in order to achieve $\sim 1\%$ level accuracy in the noise model.
- Note that getting this right is more important for Bayesian models where the covariance matrix plays a part in its own term (via its determinant) and is affected by the model parameters. For simple maximum-likelihood fits in which the data covariance is used simply to 'weight' the samples, the effect of having the overall wrong noise-level should be minimal.
- Following this analysis, `edges_cal.LoadSpectrum` now has the option to specify the level of binning in the spectrum. Typically one only uses the mean and variance (over time) of the spectrum read from the datafiles by this class. When specifying the binning (via `freq_bin_size=`), this mean and variance are taken *after* binning so that correlations are properly accounted for. The default of this parameter is unity (i.e. no binning) but should be manually set to 16 for pxspec.