Draft version May 21, 2018
Preprint typeset using LaTeX style emulateapj v. 12/16/11

NOISE MAXIMUM SNR

Riccardo Pavesi

1. THEORY

All pixels in a noise cube have a Gaussian probability distribution function centered at 0 and let us take the variance to be normalized (SNR cubes). First, note that the distribution of all pixels, which is a realization of many short-scale correlated random variables X_n, is not exactly the same as the PDF for each pixel, but is very close because correlation takes place on short scales (beam size). So note that the pixel distribution is very close to Gaussian, with normalization equal to the total number of pixels. This would, at first sight, imply that the high tail end (i.e. the SNR of the highest pixel) is going to depend on the number of pixels in a beam, which cannot be true (in fact it is not, but the reason is subtle, the pixel distribution tail is where the difference from correlations shows up, but it is only an effect of $\sim 5\%$ or less in the maximum SNR, for large cubes).

So we look at the distribution of the maximum value for a correlated Gaussian field in a finite region, following Colombi et al. (2011). We are going to work in the regime of high SNR thresholds (significantly >1), and the regime where peaks above these thresholds are far enough that they do not feel any residual correlation from the presence of a finite beam.

The cumulative probability distribution function for the maximal value of the field in a region (i.e., the probability that the maximum of the field in the region is below a certain value) is the same as the probability of having no peaks above that same threshold in the given region. Under the previous simplifying conditions, the problem simplifies then, to a Poisson point process where there is a surface (volume) density of peaks (conditional to be above a given SNR) and the probability of finding N peaks (above a specified threshold) in the region is given by a Poisson distribution with expectation equal to the area (volume) of the region multiplied by the number density of peaks. Then, the cumulative function of interest is then the Poisson probability of finding 0 peaks above the given threshold (equivalent to saying that the maximum is below this threshold), i.e., $\mathrm{Poisson}(0|nV) = e^{-nV}$.

The peak density, conditional to lie above a fixed threshold, was calculated by Bardeen et al. (1986) for the 3D case and Bond et al. (1987) for the 2D case, and they are summarized by Colombi et al. (2011). Following the definitions of Colombi et al. (2011), we use $(L/l)^D$ to represent the "number of independent elements" here defined. The setup is such that the field region under study is a D-dim ball of radius L, and l represents the standard deviation of the isotropic Gaussian "beam" (the size of the Gaussian used to smooth the "original" white noise). Then the formulae for the cumulative distribution function for the field maximum are:

$$P(\nu_{max} < \nu) \sim \exp[-0.10\,(L/l)^2\,\nu\exp(-\nu^2/2)] \tag{1}$$

in 2D, and in 3D:

$$P(\nu_{max} < \nu) \sim \exp[-0.0375\,(L/l)^3\,(\nu^2-1)\exp(-\nu^2/2)] \tag{2}$$

Where ν is the SNR field threshold. The terms inside the exponential are the region volume multiplied by the peak number density.

2. RESULTS

There are two ways to look at the problem for interferometric line searches in data cubes. The first way is to just look at the starting data (SNR) cube. This has independent channels, and spatial noise correlation on the scale of the beam. To calculate the probability distribution of the highest noise contaminant (one realization also provided by looking at the negatives and minima, since noise is symmetric around 0) we then use the 2D formula, and consider the equivalent area which corresponds to the total area of putting all the channel maps together (since noise is uncorrelated, so we just add the region area). Hence set l to the standard deviation of the beam, and $\pi\,L^2 = A_{chan}\,N_{chans}$, where A_{chan} is the area in each channel map. $N_{chans} \sim 2000$ in our cubes, and A_{chan} =112267 pixels for COSMOS and 687260 pixels for GN (pixel size=0.5″). The beam FWHM can be taken to be approximately $\sim 2.5''$ (this is an important source of uncertainty in this estimate) implying std. of 2.1 pixels. The "effective area", total area divided by beam area (circular beam of radius equal to std), $(L/l)^2$, in independent elements is 1.6×10^7 for COSMOS and 10^8 for GOODS-N. For COSMOS then the 16th, 50th and 84th percentiles are (5.55, 5.73, 5.97) and for GOODS-N (5.88, 6.05, 6.28). The measured values are -5.91 for COSMOS and -6.07 for GOODS-N in the N-mosaic, exactly as expected (and lower in the S-moisac, as expected due to additional smoothing).

The second way to use these formulae is the 3D view, and considering a cube that was convolved (MF3D-style) both spatially and/or in frequency to consider noise peaks which contaminate the search for signal line candidates. The volume of the correlation element (3D version of the beam, i.e., simply template width along the spectral dimension) is measured in the various dimensions using the appropriate unit of the correlation length in that dimension. For example, if we consider the cube with no spatial convolving (point source search) and a frequency FWHM of 4 channels (~ 140 km/s; std of 1.7 channels), then the total volume in "independent elements", denoted $(L/l)^3$ in the formula above, is the ellipsoid volume in units of "effective beam" standard deviations (i.e., product of the standard deviations along the 3 axes, divided by $4/3\,\pi$): 9.5×10^6 for COSMOS and 5.8×10^7 for GOODS-N. Resulting in percentiles on the max SNR of (5.58, 5.76, 6.01) for COSMOS and (5.92, 6.09, 6.32) for GOODS-N. We measure -5.72 for COSMOS and -5.81

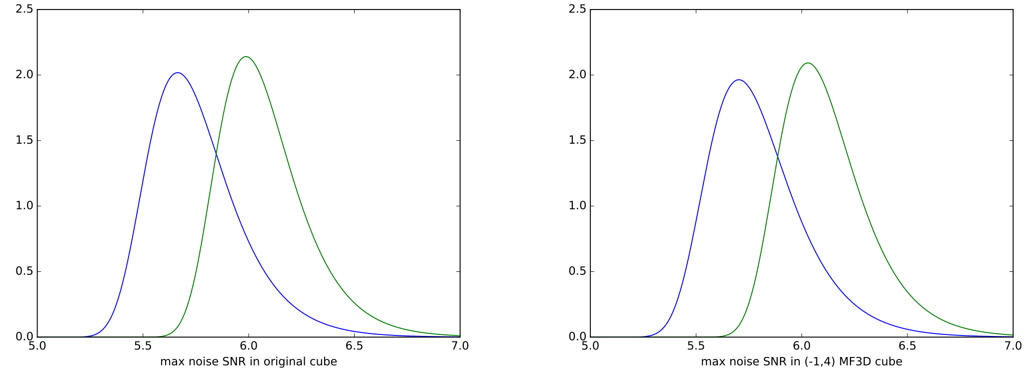

Figure 1. Probability distribution functions for the max SNR due to noise in the 2D and 3D cases. Blue curve for COSMOS, green for GOODS-N.

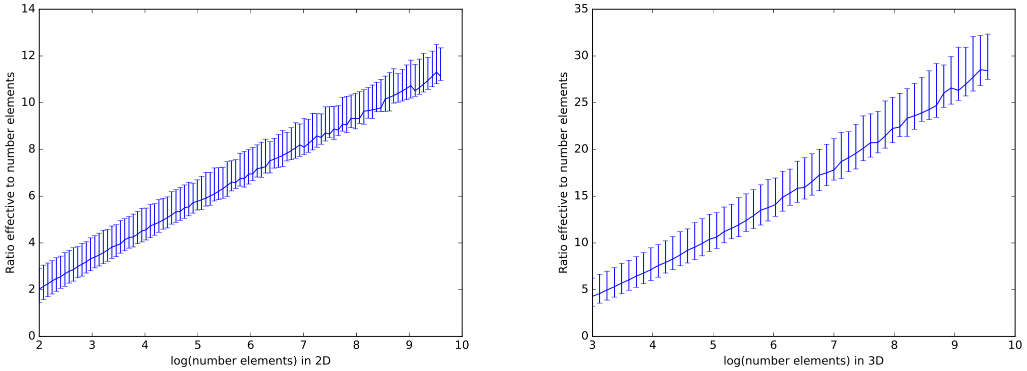

Figure 2. How many more "effective" independent elements determine the extent of the noise tail, relative to the naive number of elements, as a function of the naive number of elements.

for GOODS-N in the (-1,4) MF cubes. The GOODS-N peak is not as high as we might expect, and may suggest that the beam counting may not be quite correct, but close (also consider the effect of having ruled out, with additional data, the most significant negative feature as a formaldehyde absorption candidate). The highest negative line of the full MF3D is just the max over all the templates, and is in the same range (broader frequency templates have somewhat lower peak, as might be expected as they reduce the number of channels more significantly). Although it might appear that the difference between 5 and 6σ in a Gaussian tail might be large, the value of the maximum in a large cube is actually not well predicted, i.e., the tails are intrinsically subject to large fluctuations. We also note that the difference between SNR=5.7 and 5.8 is only $< 2\%$ and that at this small level of difference the discrete pixel sampling of the noise field may also potentially alter values slightly (it is not quite the peak of a continuous distribution that we will ever see, but some pixel samples), but we do not expect systematic shifts from this.

3. DISCUSSION

Let us consider the case of no pixel correlation (IID) first. The cumulative distribution for the maximum is $P_{CDF}^{N_{pix}}$ where P_{CDF} is the individual pixel CDF. Take $N_{pix} = 10^7$, then the usual 16th, 50th, 84th percentiles (i.e., probabilities that the maximum will be less than these SNR thresholds) are (5.07, 5.27, 5.51), and N_{pix}*(1-$P_{CDF}([5.07, 5.27, 5.51])$)=[1.86,0.69,0.18]. these are the expected number of of pixels with SNR> threshold. Poisson expectations of μ =[1.86,0.69,0.18] exactly imply that [16%,50%, 84%] of the cases will return 0 and the remaining will return $>$ 0 pixels with SNR>threshold, as expected.

Therefore, by comparing the calculated percentiles for the max SNR in the correlated case we can compute the "effective number of independent elements", which is the number of independent variables you would need, in order to reproduce such noise tails.

In the 2D case, we get for COSMOS 1.3–1.5$\times 10^8$ and for GOODS-N 9.1–11$\times 10^8$, which are 8–9× and 9–11× the previously defined "number of independent elements". In the 3D case, we get for COSMOS 1.5–1.9$\times 10^8$ and for GOODS-N 1.16–1.4$\times 10^9$, which are 16–20× and 20–24× the previously defined "number of elements", showing that **the naive estimate would greatly underestimate the extent of the noise tail. The more numerous the independent elements, the larger the *effective* number of independent elements (to the tail) becomes, relative to the naive counting.**

REFERENCES

Bardeen, J. M., Bond, J. R., Kaiser, N. et al., 1986, ApJ, 304, 15
Bond J. R. Efstathiou G., 1987, MNRAS, 226, 655
Colombi, S., Davis, O., Devriendt, J. et al., 2011, MNRAS, 414, 2436