
Figure 2.18: The matched filtering algorithm, operating on a simulated narrow-band signal plus Gaussian noise, identifies the delineated region as the best signal candidate. The signal intensity is uniformly 10, in units of $\sigma$, and it occupies 2 $\times$ 200 samples. The filter does not capture the entire signal because only filter lenghts of $2^n$ are sampled. In this case, the $2 \times 128$ filter is the best match to the signal.

values decline with group size once the filter area exceeds the area of the signal ($2 \times 200$). Note that the range of filter sizes is discrete, not continuous. This is a result of how the filter sizes are stepped up, as described above.

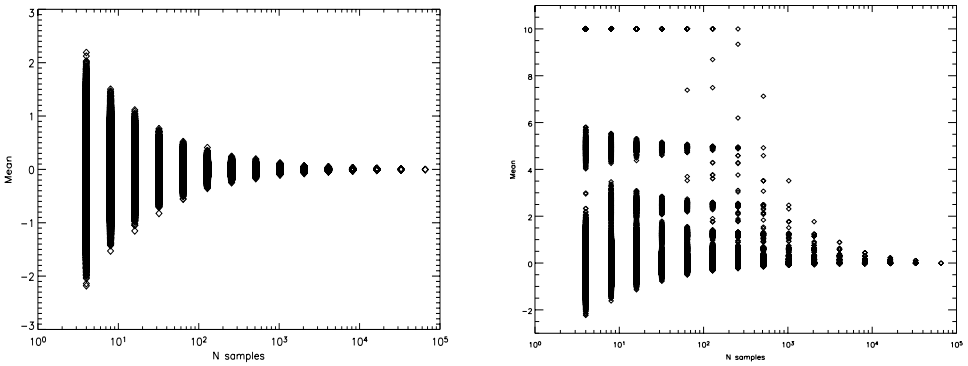


Figure 2.19: Region mean by filter size for pure noise, left, and noise plus signal, right. The mean for each region is shown in a scatter plot with the region size (number of samples contained in the filter region) on the x-axis.

The histograms in Figure 2.20 show the number of regions given fill factor values for pure noise (left) and noise plus signal (right), where the fill factor is defined as the fraction of samples within a region exceeding three times the rms of the entire data set. Regions with a fill factor of zero are not included in the histogram. In both simulations, $\sim 96\%$ of groups have zero fill factor. The discrete divisions between populated bins at high fill factor (which, as in Figure 2.21, correspond to small filter dimensions) are due to the exponentially increasing dimensions of the filter. In the pure noise case, no groups have a fill factor exceeding 0.3. However, the noise plus signal data set yields a significant number of groups with fill factors up to 0.6, and has a small spike at 1 (that is, 100%).

In Figure 2.21, we show the relationship between fill factor and region size for pure Gaussian noise (left) and noise plus signal (right). In the pure noise case,

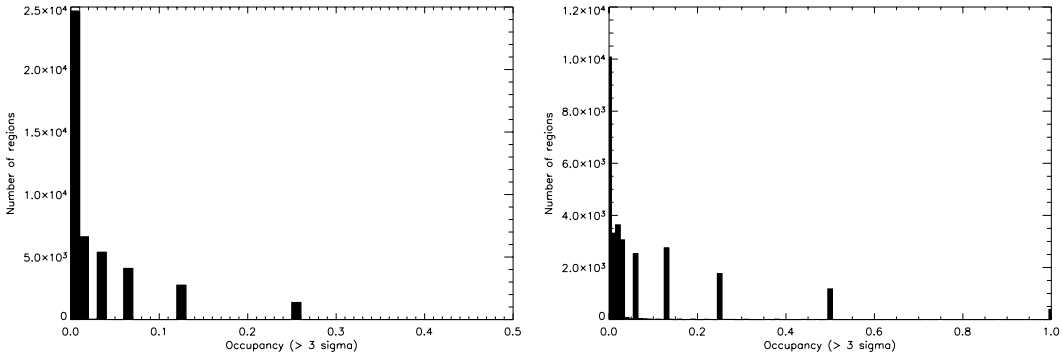


Figure 2.20: Region fill factor histogram for pure noise, left, and noise plus signal, right. Fill factor values are given on the x-axis and the number of regions with each fill factor is shown on the y-axis. Regions with zero fill factor are not included.

the highest fill factors are found in the smallest groups, as expected for normally distributed data. The signal plus noise case, however, displays a very different structure, with consistently high fill factors for group sizes smaller than the signal area. Once the number of samples exceeds the signal area, the falloff in fill factor takes a similar form to the pure noise case.

### 2.5.1 Discussion

The FOF and matched-filtering algorithms are both "smarter" than the threshold test in that they identify regions of interest, not just unrelated points. These smart results require additional computation time, though, which is a major challenge for large data sets such as those expected from an ALFA sky survey. In addition, they both require that the user make some assumptions about the strength, duration and bandwidth of the signal to be detected. For the FOF algorithm, these assumptions are minimal; for matched-filtering, they decrease as the number of test filters increases. One can also imagine using a genetic algorithm to create a

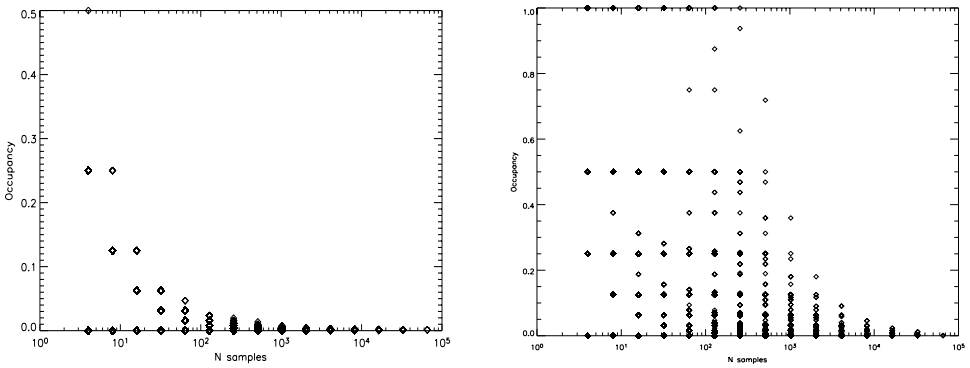


Figure 2.21: Scatter plot relating fill factor and filter size (number of samples contained in the filter) for normally distributed data, left, and the same normally distributed data plus a simulated narrowband signal, right. As expected for signal-less, normally distributed noise, only small filter sizes yield high fill factor values. For the noise plus signal case, high fill factors are observed for a wide range of filter sizes. Fill factor declines once the filter size exceeds the signal area.

filter that adapts to the data.

The FOF algorithm offers other advantages over the threshold test. The event list produced by the FOF algorithm is consirably shorter than the analogous list produced by a simple threshold test, meaning that the data can be probed at lower flux levels and still produce a managably-sized list of events. It is possible to probe lower flux levels in the threshold test by smoothing the data beforehand (Cordes & McLaughlin, 2003); however, this impedes our ability to identify very fast or narrowband signals.

Similarly, the proximity requirements of the FOF routine naturally produce fewer false positives than the basic threshold test. The cost is in computation time and in the loss of extremely fast, narrowband events, which fall short of meeting the minimum group size requirement ($p$).

Future work on this topic should include a quantitative study of the computa-

tion time required for each algorithm, and how that time depends on the size of the data set and the user-defined parameters input to the algorithm. This should go hand-in-hand with work to optimize each algorithm's running time.

## REFERENCES

Cordes, J. M. & McLaughlin, M. A. 2003, ApJ, 596, 1142

Huchra, J. P. & Geller, M. J. 1982, ApJ, 257, 423

Lundgren, S. C., Cordes, J. M., Ulmer, M., Matz, S. M., Lomatch, S., Foster, R. S., & Hankins, T. 1995, ApJ, 453, 433