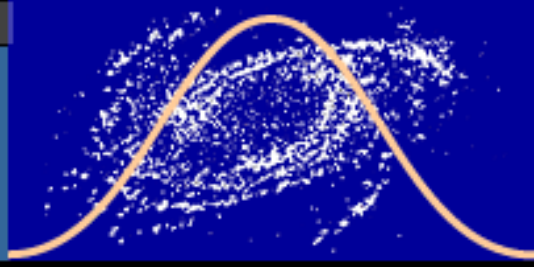


Analysis of astronomical datacubes

G. Jogesh Babu

http://astrostatistics.psu.edu

Penn State University

Most 20-th century astronomical data have been 2-dimensional images or 1-dimensional spectra/time series. But 3-dimensional hyperspectral images and videos are becoming increasingly prevalent:

1. Datacubes from radio interferometers
   (once restricted to 21-cm and small molecular line maps, nearly all data from the EVLA and ALMA will be 3-dim spectro-image datacubes)

2. Integral Field Units spectrographs
   (bundled-fiber-fed spectrographs give spectro-image cubes)

3. Multiepoch visible-light surveys
   (Palomar QUEST, Pan-STARRS, LSST, etc. produce huge datasets of time-image video-like cubes)

Extensive methodology developed in other fields for 3D analysis: digital video processing, remote sensing, animation, etc

Astronomical datacubes today

All major new telescopes produce datacubes as their primary data products: ALMA, ELVA, ASKAP, MeerKAT, LOFAR, SKA.

For many observations, ALMA will produce 1-100 GBy datacubes consisting of ~1-10 million spatial pixels with 1-10 thousand frequency channels. Over a decade, ALMA produces petabytes of datacubes.

Main interest here is in LSST-type video problems

- There are some differences from radio. E.g., night-to-night differences in PSF shape that affects the unresolved sources, whereas radio has RFI autocorrelated noise across some 2D images.

- The overall approach of computationally efficient, intermediate-level statistics (e.g. Mahalabonis distance) and computer vision techniques holds in both cases

Data analysis & science goals include:

a. Faint continuum source detection

   Source catalogs (logN-LogS), multiwavelength & transient studies

b. Faint spectral lines & transient detection

   Redshifted Lyα & HI proto/dwarf galaxies, SN Ia surveys, orphan GRBs, etc.

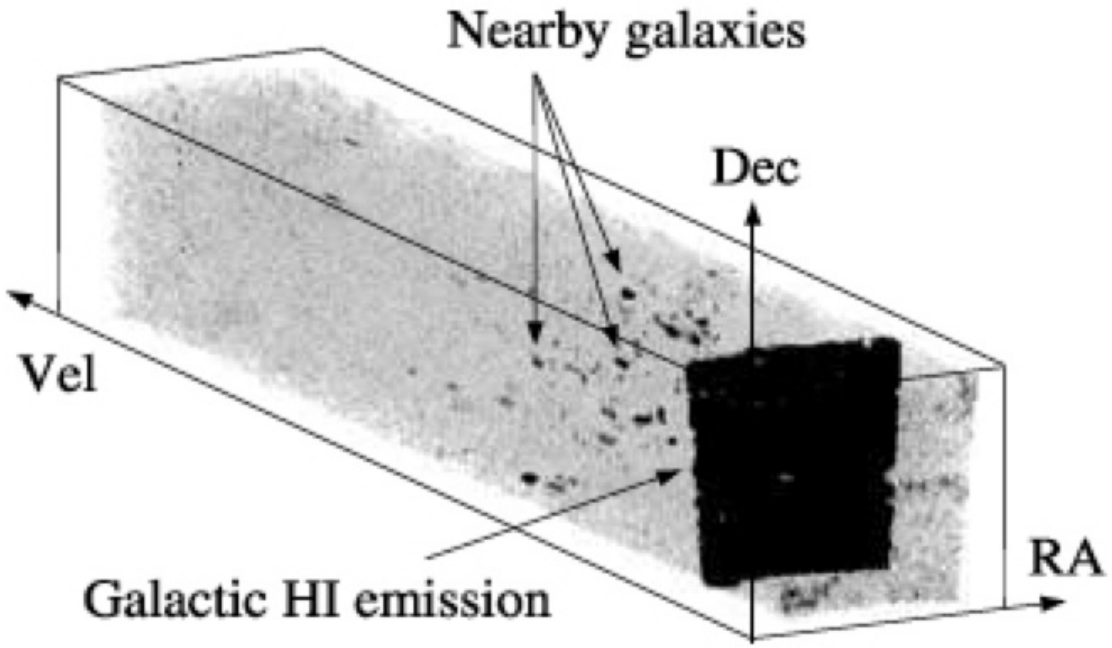
c. Characterization of bright features

   Photometry, galaxy morphology, radio jets/lobes, molecular clouds, etc.

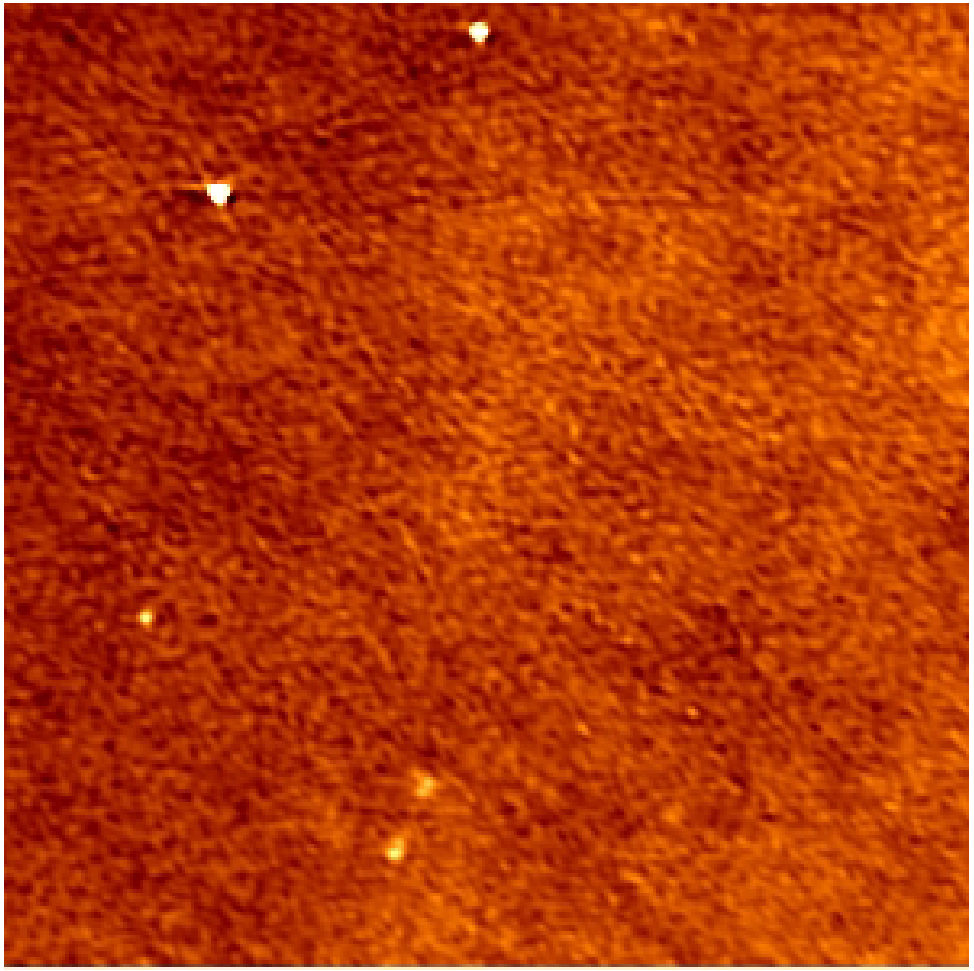
Methodologists have to think very broadly about all types of data cubes

Bump hunting in radio astronomical datacubes

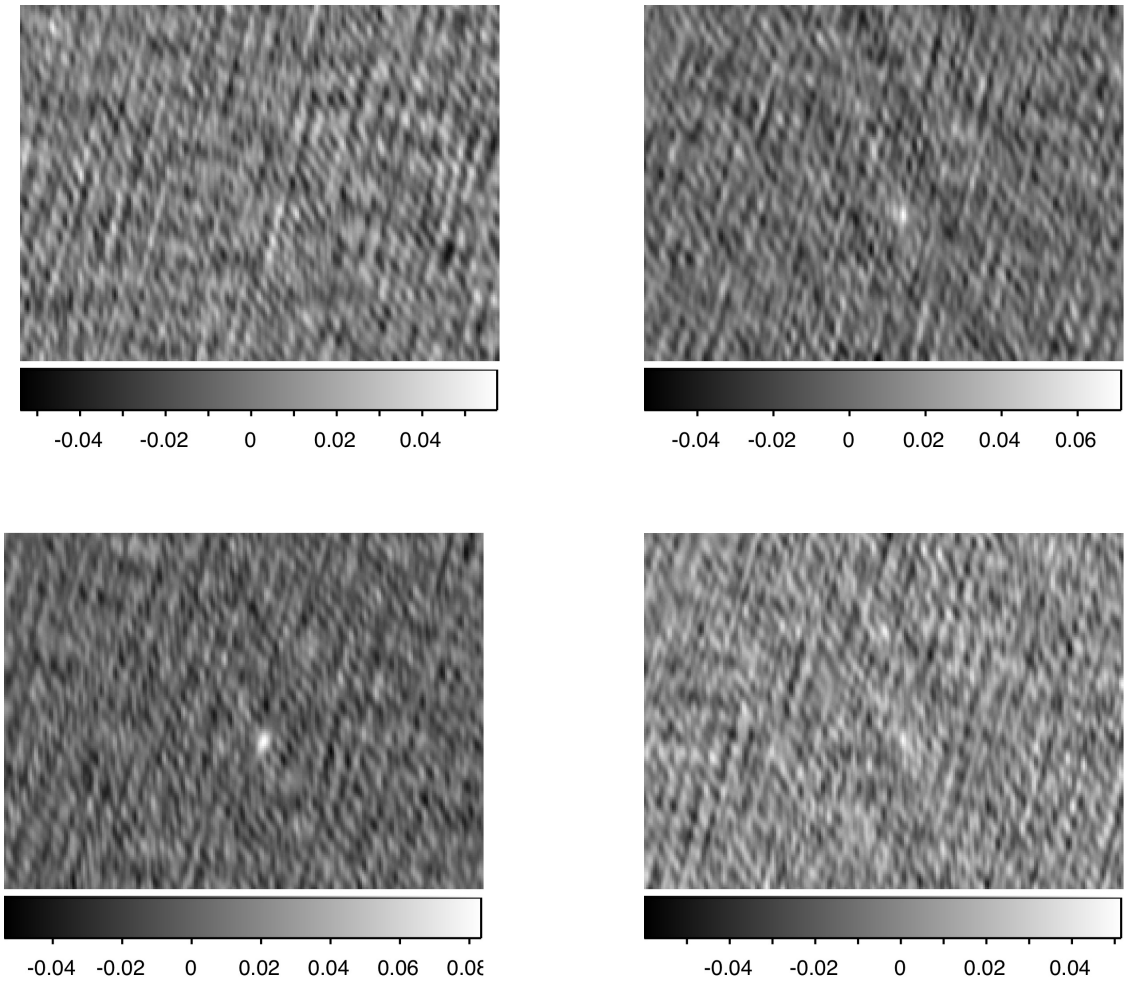
Single-dish data



HI Parkes All-Sky Survey (HIPASS) 21-cm data cube showing nearby galaxies (dark spots) and the Galactic Plane (dark sheet). (Meyer et al. 2004). Understanding the noise properties is particularly important for finding the faintest sources.

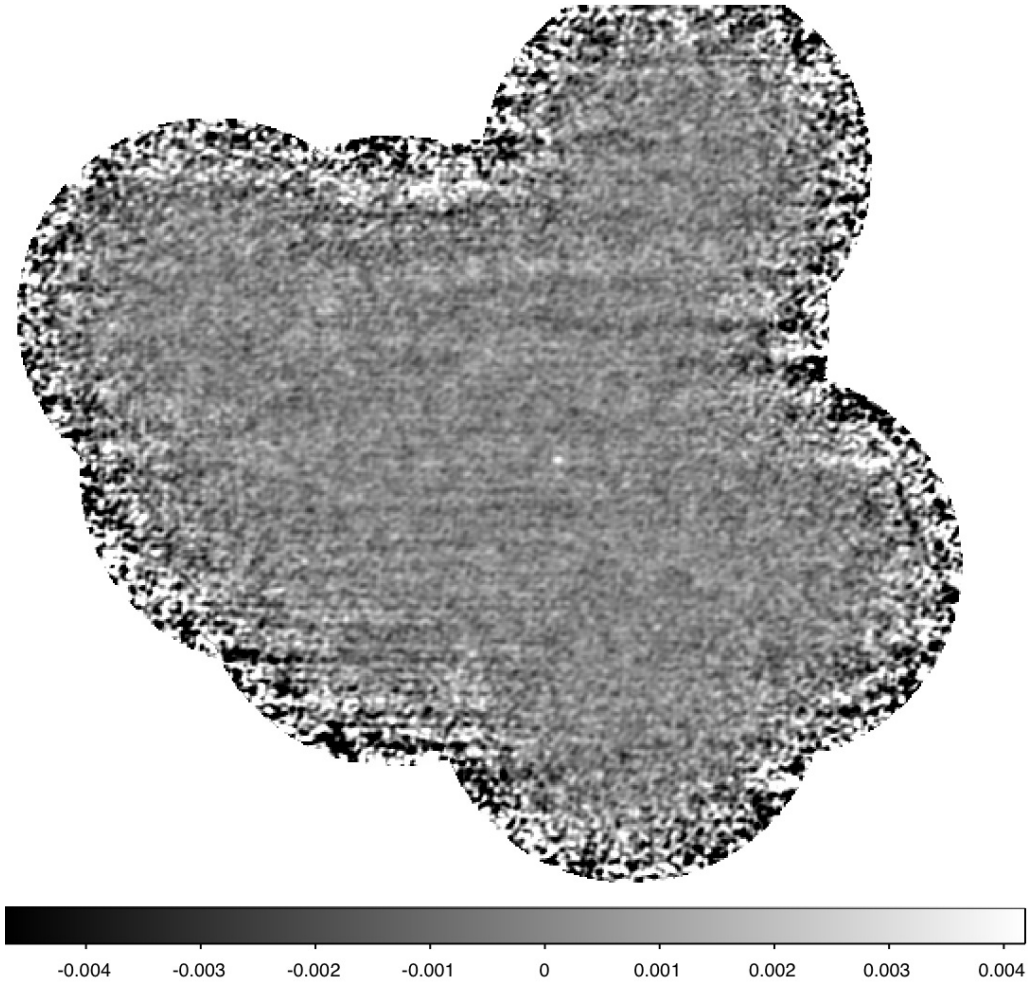


VLA 20 cm image in the Galactic Plane. This 10’ x10 ‘ is selected to illustrate the non-Gaussian backgrounds. (Becker et al. 2007)



Interferometric datacubes

VLA subcube showing four adjacent channels of a large datacube of molecular maser emission in a Galactic star forming region. This shows the spatially correlated, non-Gaussian noise and faint sources common in radio datacubes. (Courtesy NRAO)



A channel of a HI mosaic from the VLA illustrating spatially heteroscedastic noise from primary beam and ripples from mild radio frequency interference. (Courtesy J. van Gorkom)

Noise structure

- ***Spatial autocorrelation*** from Fourier transform of visibilities
- ***Non-Gaussian `tails'*** due to incomplete visibility coverage, un-CLEANed sidelobes of bright sources, un-flagged RFI
- ***Heteroscedasticity*** (i.e. varies across the 3-dimensional image) due to primary beam, sidelobes and RFI

If we are lucky, these problems will not be severe in ALMA data. But the methodology should be able to treat them.

Steps towards ALMA feature detection & characterization

Datacube construction: RFI removal from visibilities, Image formation from visibilities, CLEAN sidelobes and construct datacube

Feature identification

- **Local noise characterization –** Heteroscedasticity & non-Gaussianity
- **Signal detection -** Procedures to mark regions in the datacube with likely continuum and/or line emission. Multiscale for both unresolved and extended structures. Most sources are near the noise and detection thresholds should control for false positives. Important development: False Detection Rate (Benjamini & Hochberg 1995).

Feature characterization

- **Source consolidation** - Procedures to merge adjacent marked subregions into distinct continuum sources and line structures must be adopted. `Image segmentation' from computer vision technology.
- **Source characterization** - Source outline, line center and width, total flux, normal mixture model (Gaussian components), spectral indices, etc.

Other methods used in radio astronomy

Visual examination: Commonly used and praised for sensitivity. But not practical for Tby archives

ClumpFind, Picasso, MultiFind: User-specified global S/N threshold after image cleaning (e.g. polynomial fits to continuum; Cornwell et al. 1992, Minchin 1999). ClumpFind then applies: local median/noise threshold, reject peaks smaller than beam, reject peaks near edge (Williams et al. 1994; Di Francesco et al. 2008).

MultiFind uses S/N threshold with noise is the robust median absolute deviation measured aftern Hanning smoothing (Meyer et al. 2004).

TopHat: 3-step procedure that scans through frequency planes, removes continuum ripples with median filter, cross-correlates image with multiresolution tophat filter weighted by local noise, groups features in adjacent velocity planes (Meyer et al. 2004).

Multiresolution wavelet decomposition (Motte et al. 1998, Knudsen et al. 2006).

Matched filter algorithms: CLEAN-type algorithms plus peak-finding (Enoch et al. 2006; Young et al. 2006). Also developed for ALFALFA (Sointonge 2007) where it cross-correlates a Fourier transform of datacube with Hermite polynomials. Effective in locating low-surface brightness galaxies, robust to baseline fluctuations, performs all scales in a single calculation.

Methods from other fields

Wavelet transform: Effective for multiscale image decomposition and denoising on both small- & large-scale (low- & high-pass filtering). Strong mathematical foundation (Mallat 1999; Starck & Murtagh 2006).

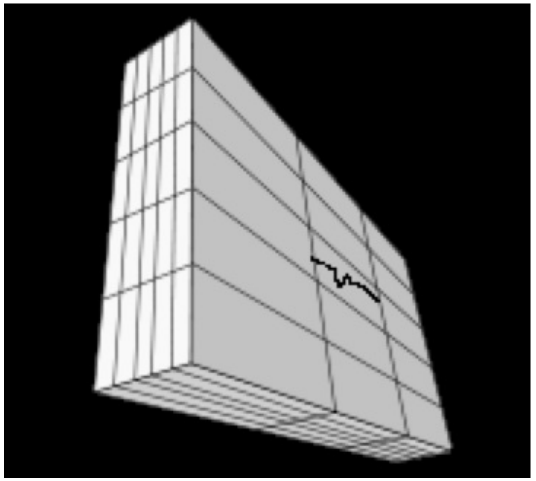
Local regression: Long-established techniques for smoothing nonlinear data (Cleveland 1988) and mapping geological strata (kriging, Journel & Huijbregts 1978). Procedures estimate value and noise (variogram) around each location with local windows. Can use robust estimation. Effective for spatial background gradients and heteroscedasticity.

Semiparametric regression: New techniques extending nonparametric density estimation (e.g. smoothing, spline fits) to give data-dependent confidence intervals around the estimator (Ruppert et al. 2003; Wasserman 2006). Particularly effective for spatial heteroscedasticity. Not computationally efficient.

Gamma Test: First Hanning smooth data with user-specified bandwidth. Second, pass moving window of the q-th nearest neighbors for specified range of q. Local background is then estimated for each location by linear regression of noise vs. q. Signal is then plotted over this background. Developed by Stefansson et al. (1997), Jones et al. (2006, 2007). Applied to 1-dim radio astronomical spectra by Boyce (2003).

Overview of the suggested procedure

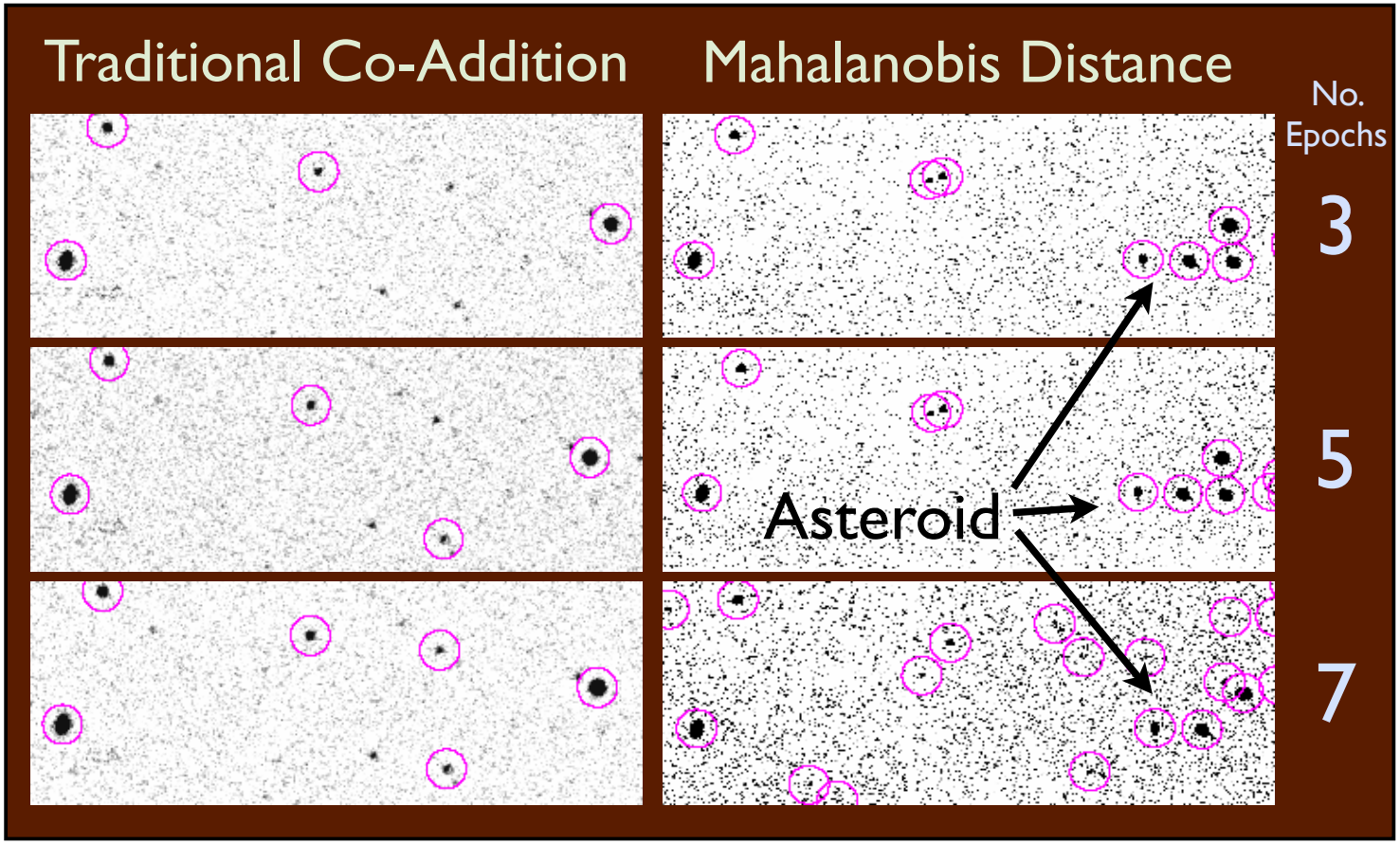
- Robust rejection of bad planes (unflagged RFI)
- Divide the data cube into small overlapping cubes
- Fully 3-dim signal detection in locally homoscedastic subcubes using non-parametric regression
- Pixel-based identification of continuum/ constant sources (rods in 3-dim) and line/transient sources (spots in 3-dim) using thresholds based on the local variance structure and autocorrelation between spectral/temporal planes
- False Discovery Rate control for false positives
- Image segmentation to unify adjacent hits
- Pyramid aggregation or `Active contour' of resulting 3-dim structures, including non-convex topologies



Local 3-dim subcube for source detection

Statistical approach focuses on:

- The noise distribution in 2Dimages or 3D subcubes, that exhibit locally homoscedastic noise.
- Local regression methods, if large-scale correlations are present.
- Pixel-level signal detection based on Mahalanobis-type distance [$D_M(\mathbf{x},\boldsymbol{\mu})=\sqrt{(\mathbf{x}-\boldsymbol{\mu})^T S^{-1}(\mathbf{x}-\boldsymbol{\mu})}$]
- To address spatial correlation from the PSF and from extended sources, the joint quantiles in adjacent channels within a subcube should be examined to set local thresholds.



Preliminary results on source detection using Mahalanobis distance.

B band images of a small portion of the sky from the Palomar-QUEST multi-epoch survey.

The three panels show co-adds from 3 (top), 5 (middle) and 7 (bottom) epochs. The left images are formed by a traditional pixel-by-pixel averaging method with 4σ sources circled. The images on the right are formed by the statistic based on Mahalanobis distance with significant sources circled. Mahalanobis distance techniques can provide a richer view near the detection threshold.

The emergence of new objects in the right-hand panels is likely due to the passage of a minor planetary body (an asteroid) through the field.

Analysis

- The analysis proceeds by calculating flux averages in each subcube to treat the heteroscedastic errors.

- The data vectors will be compared with averages over the entire image to set thresholds based on quantiles to identify subcubes with potential source contributions. Noise-only subcubes are discarded.

- Thresholds for multi-pixel triggers will be set based on a Mahalanobis-type statistics.

- Operationally, overlapping windows are passed over each homoscedastic subcube.

- The additive model

$$f^r_i = E(f^r_i) + (f^r_i - E(f^r_i)) = M^r_i + \varepsilon^r_i, \quad i=1, \ldots, s,$$

is considered for the data at location r and channel i, where

$$Var(f^r_i) = Var(\varepsilon^r_i)$$

may vary over the spatial domain. The parameter M^r_i may be viewed as the true signal of the source at location r and channel i.

As all subcubes have an autocorrelation structure due to the synthesized beam:

- We first consider the case where the noise distribution is heteroscedastic (varies with location), but roughly normal (Gaussian).

- In this case the distribution of $(f^r_1, \cdots, f^r_s)$ is multivariate normal and generalized χ^2-type statistics can be constructed to test hypotheses of signal at different locations.

- If an autocorrelation structure does not hold and/or Gaussianity fails to hold, **block bootstrap methods** to estimate covariance structure may help in the development of corresponding test statistics.

- In some cases, the multi-epoch correlations of the noise of a datacube may follow an AR(1) model

- That is, a signal at a location i is correlated with its neighbors one epoch apart according to

$$f^r_1 = \varepsilon^r_1, \quad f^r_i = \mu^r_i + \lambda f^r_{i-1} + \varepsilon^r_i,$$

where ε^r_i is zero mean noise.

- The hypothesis H_0: $(\mu^r_1, \cdots, \mu^r_s) = (0, \ldots, 0)$ may be tested using a quadratic form $(f^r_1, \cdots, f^r_s)\, S^{-1} (f^r_1, \cdots, f^r_s)^T$, where the entries of the matrix S are polynomials of estimate of λ.

- The autocorrelated structure at longer lags can capture much of the deviations from normality in the noise.

- Improved False Discovery Rates (FDR) procedures help to control false positive detections for autocorrelated data.

- Popular procedures used to control the FDR in large-scale multiple testing are stepwise procedures where the p-values are ordered and compared with specified cutoffs according to a stopping rule.

- Starting with the most significant p-value, the step-down procedure rejects each null hypothesis as long as the corresponding cutoff is not exceeded.

- The step-up procedure starts with the least significant p-value and proceeds in the opposite direction and accepts each null hypothesis, provided the p-value does not exceed its cutoff. These procedures have been shown to control the FDR under certain types of dependencies.

We recently started investigating a generalization of these stepwise procedures that allows it to continue rejecting as long as the fraction of p-values not exceeding their cutoffs is sufficiently large.

Preliminary studies including large-sample results indicate that, for appropriate choices of this fractional bound, increased statistical power may be obtained.

A modified FDR procedure that controls for $P(V/R > c) < \alpha$, where c and α are user-defined is more appropriate. We are investigating such a procedure.

CONCLUSION

Analysis of astronomical datacubes encounter difficulties due to:

- non-Gaussianity and heteroscedasticity of the noise
- evaluation of FDR in the presence of spatial autocorrelation
- need for computational efficiency.

A variety of approaches can be explored involving (semi-)parametric techniques and enhanced FDR calculations.

- The methods described here are designed for, and will be useful for faint source detection in datacubes from new EVLA and ALMA radio/millimeter-band telescopes.

- Techniques developed for the faint line sources in multi-channel radio band surveys may also be directly applicable for detection of faint transients in multi-epoch visible band surveys.