Playing in High Dimensions

Bob Nichol

ICG, Portsmouth

Thanks to all my colleagues in SDSS, GRIST & PiCA

Special thanks to Chris Miller, Alex Gray, Gordon Richards, Brent Bryan, Chris Genovese, Ryan Scranton, Larry Wasserman, Jeff Schneider

Outline

Cosmological data is exploding in both size and complexity. This drives us towards searching non-parametric techniques, and the need to model data in high dimensions. Two examples:

1. Detection of quasars in multicolor space
2. Detection of features in data

Hunting for quasars

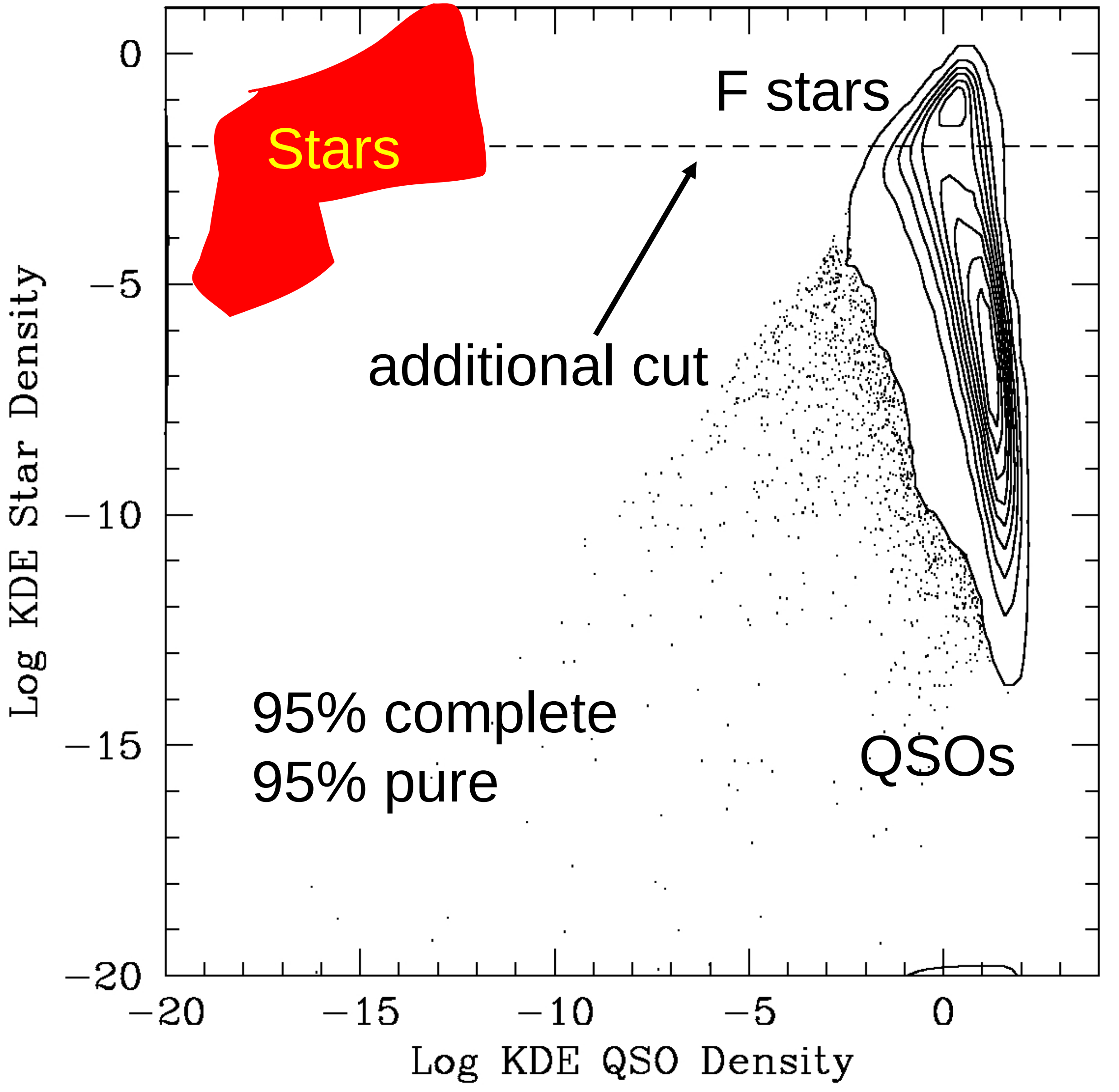
Quasi-stellar sources: by definition they look like stars!

Traditional approaches have used UVX approach to finding quasars, i.e., quasars are “very blue” so can be isolated in color-color space using simple hyper-planes (see Richards et al. 2002). However, there is significant contamination (~40%), thus demanding spectroscopic follow-up which is very time-consuming.

Use a probabilistic approach

- Use Kernel Density Estimation (KDE) to map color-color space occupied by known stars and quasars ("training sets")
- Use cross-validation to "optimal" smooth the 4-D SDSS color space and obtain PDFs
- Fast implementation via KD-trees (Gray & Moore)
- ~16,000 known quasars and ~500000 stars
- Using a non-parametric Bayes classifier (NBC)

$$P(C_1|x) = \frac{p(x|C_1)P(C_1)}{p(x|C_1)P(C_1) + p(x|C_2)P(C_2)}$$


Stars
F stars
additional cut
95% complete
95% pure
QSOs
Log KDE Star Density
Log KDE QSO Density
0
−5
−10
−15
−20
−20
−15
−10
−5
0

Advancements

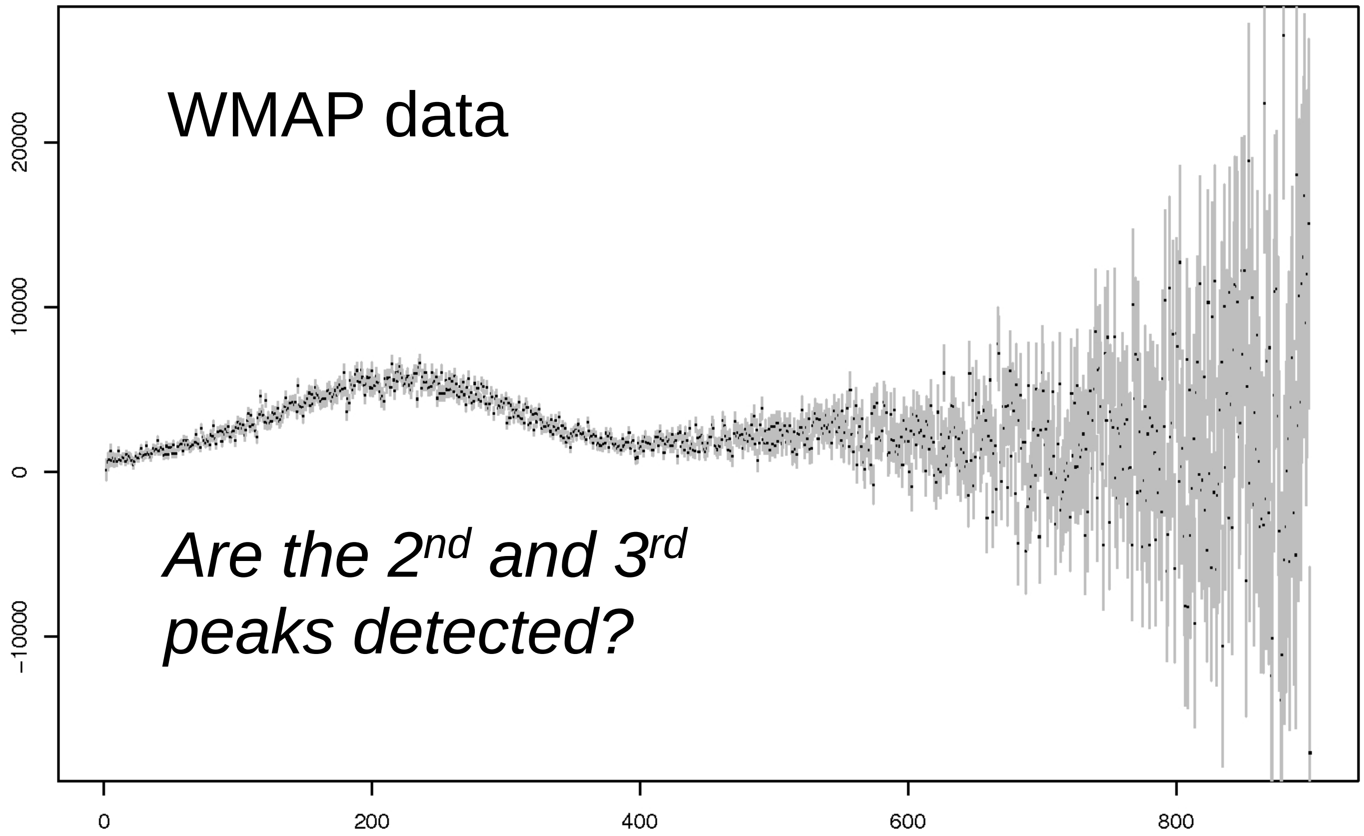
- Present prior for $P(C_1)$ is 0.88 - the ratio of stars to galaxies in our dataset
- Add magnitude and redshift information, either via increasing dimensionality or through priors

Non-parametric Techniques

- The wealth of data demands non-parametric techniques, ie., can one describe phenomena using the less amount of assumptions?

- The challenges here are computational as well as psychological

CMB Power Spectrum



In parametric models of the CMB power spectrum the answer is likely "yes" as all CMB models have multiple peaks. But that has not really answered our question!

Can we answer the question non-parametrically e.g.,

$$Y_i = f(X_i) + c_i$$

Where Y_i is the observed data, $f(X_i)$ is an orthogonal function ($\beta_i \cos(i\pi X_i)$), c_i is the covariance matrix. The challenge is to "shrink" $f(X_i)$, we use

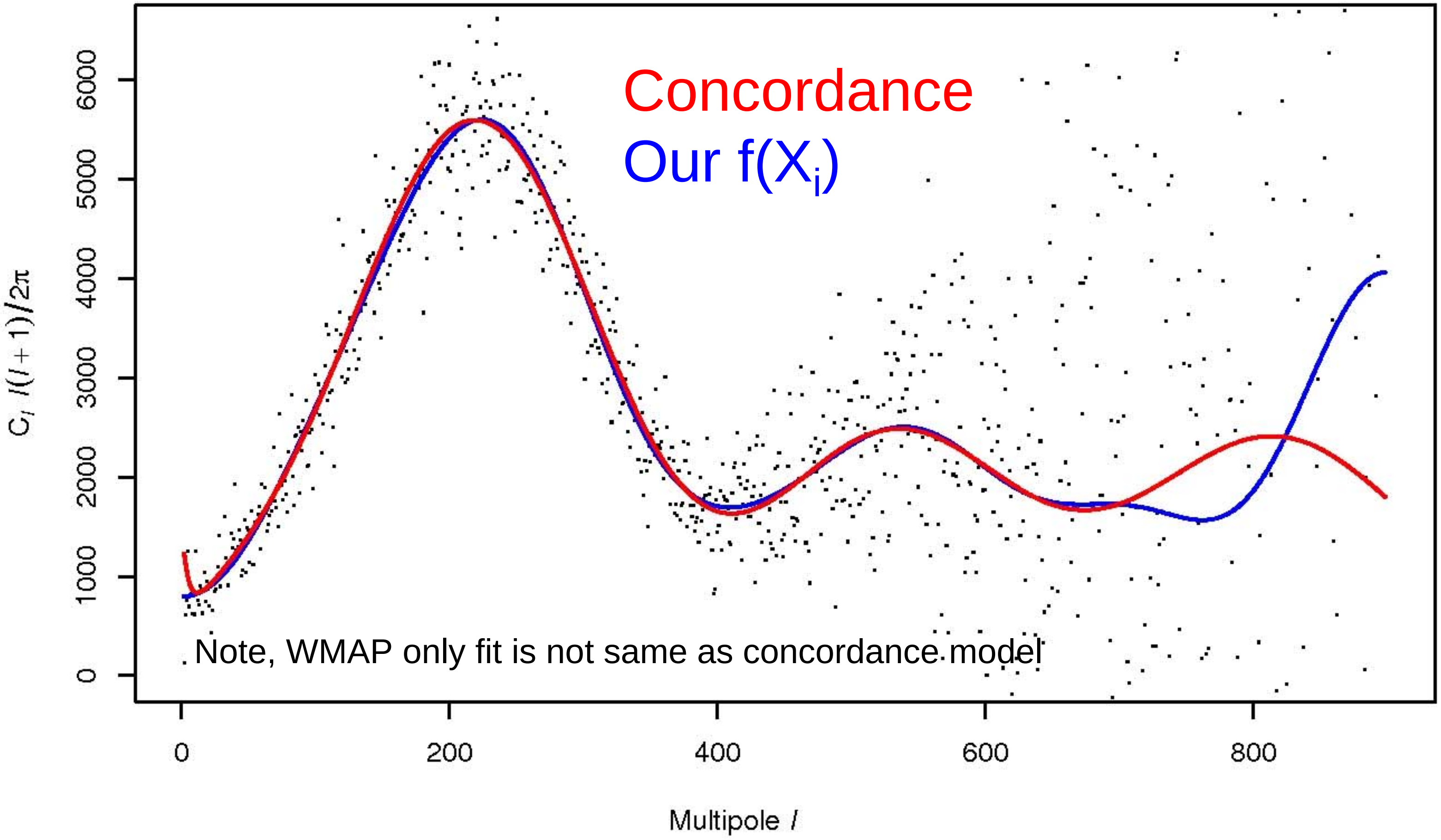
- Beran (2000) to strink $f(X_i)$ to N terms equal to the number of data points - optimal for all smooth functions and provides valid confidence intervals

- Monotonic shrinkage of β_i - specifically nested subset selection (NSS)

See Genovese et al. (2004) astro-ph/0410104

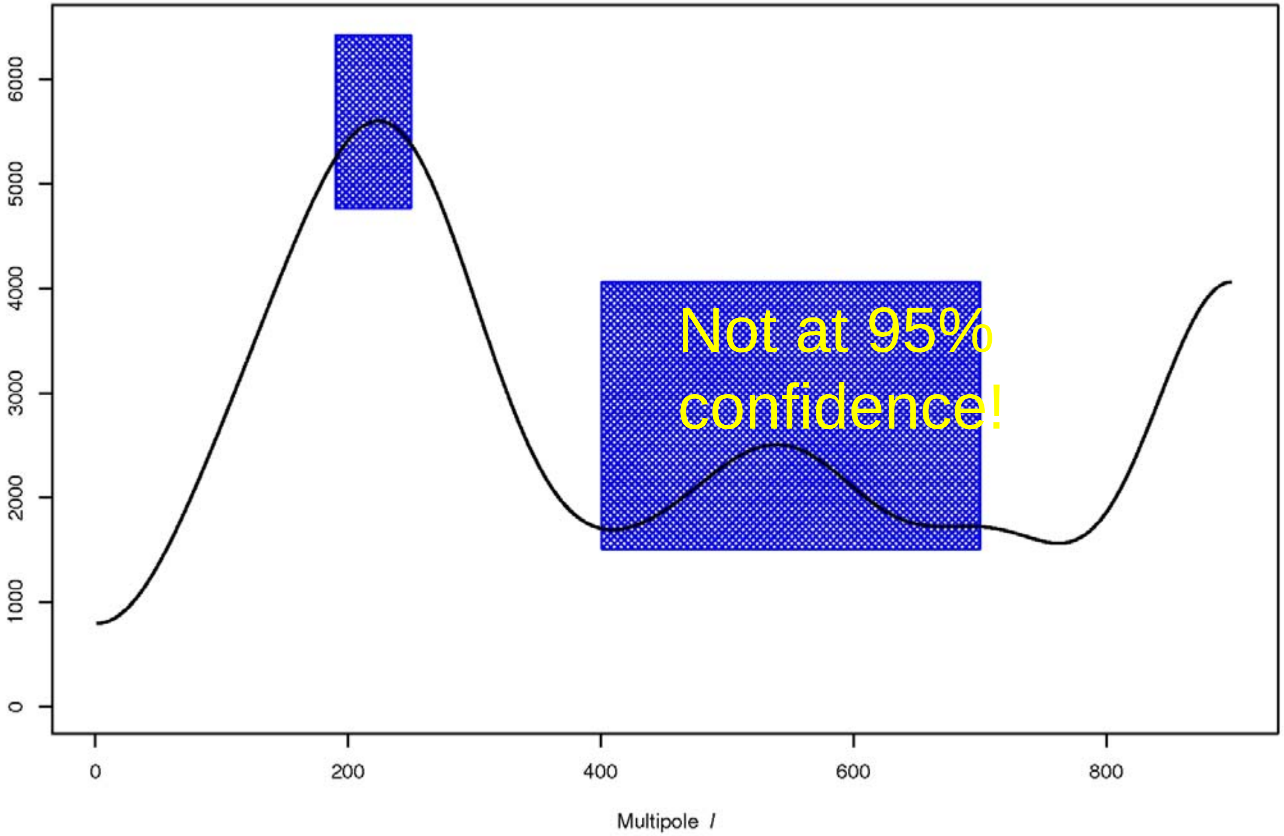
Results

(optimal smoothing through bias-variance trade-off)



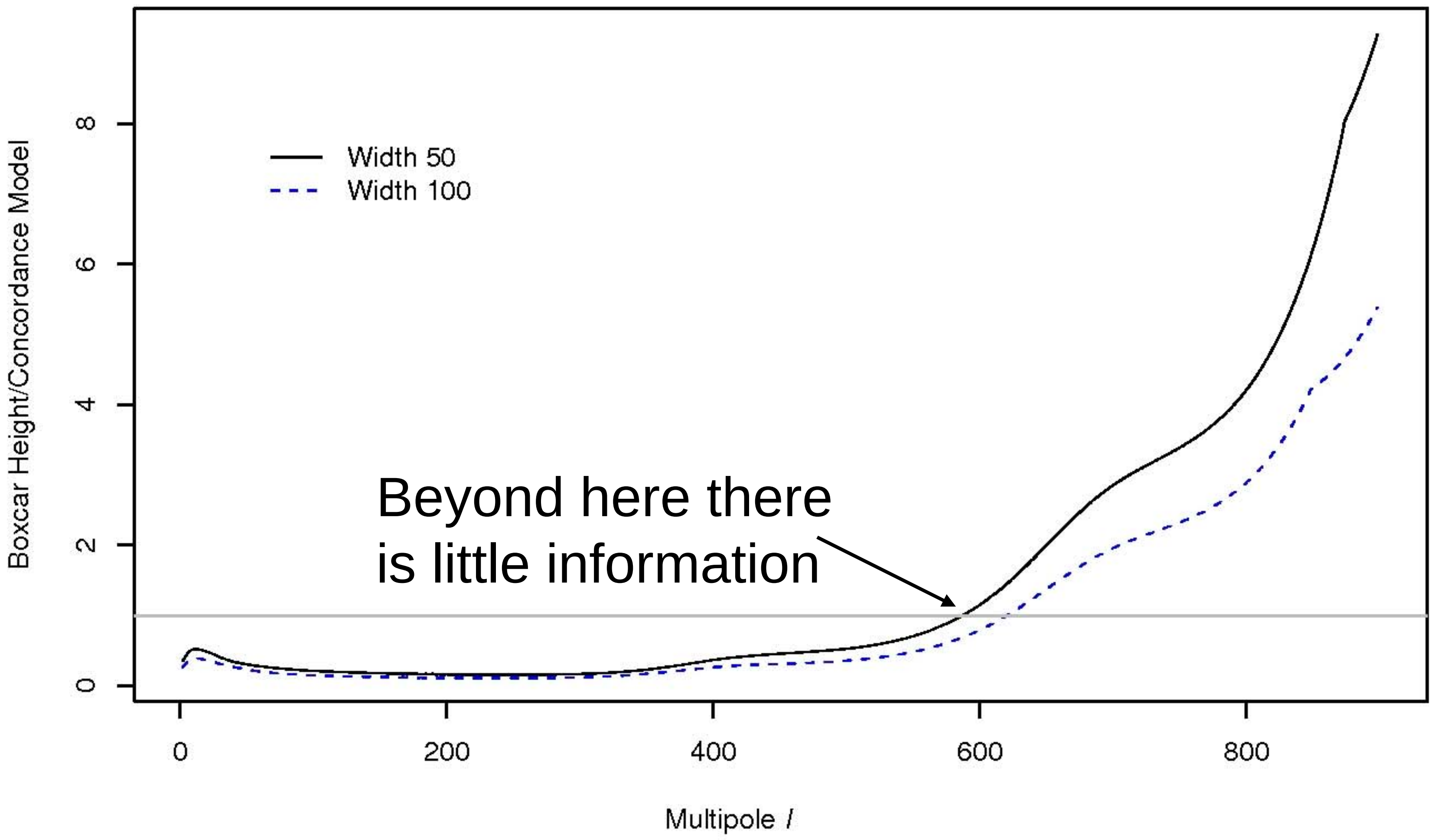
Testing models

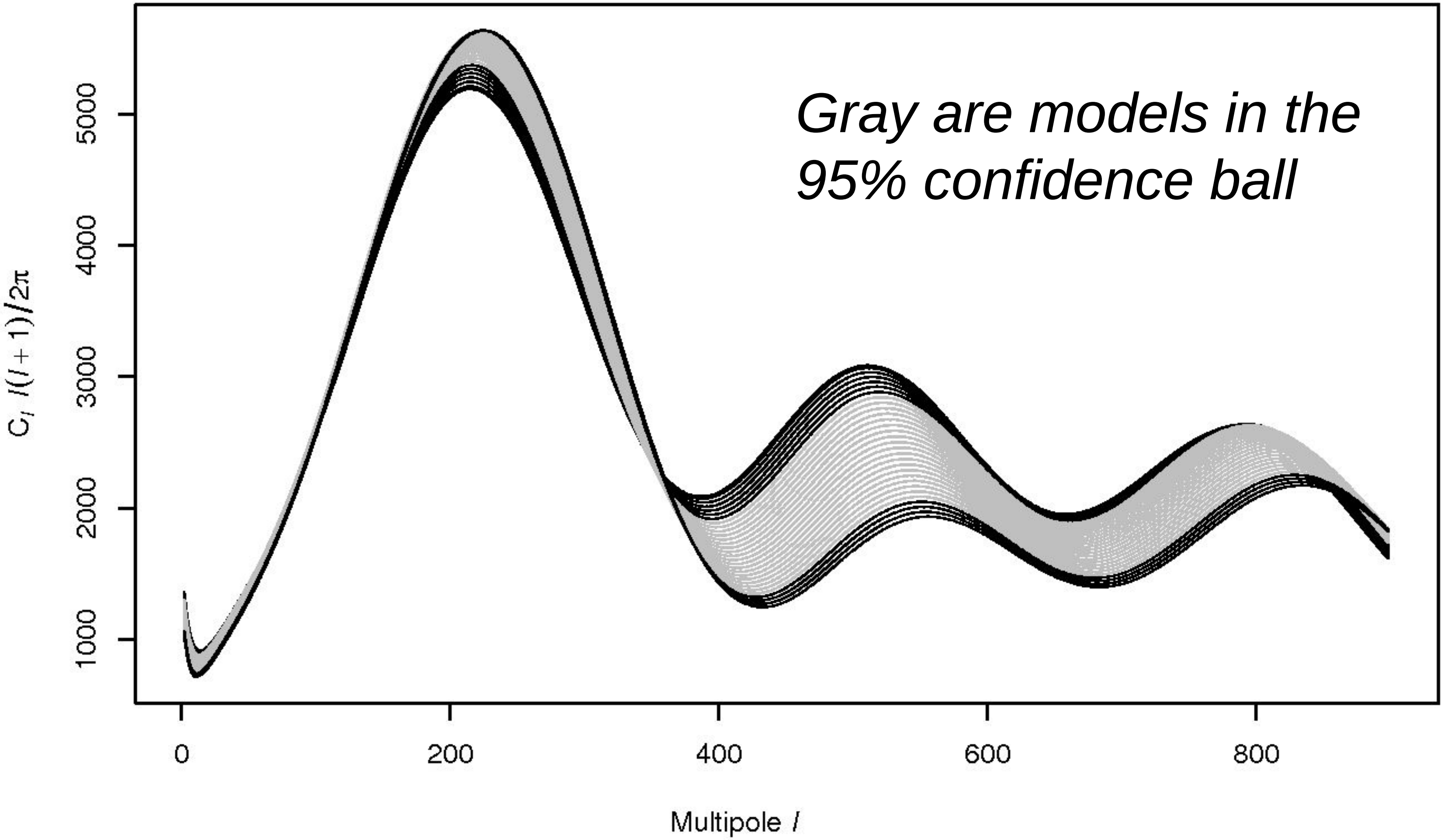
- The main advantage of this method is that we can construct a “confidence ball” (in N dimensions) around $f(X_i)$ and thus perform non-parametric interferences e.g. is the second peak detected?



Information Content

$f_h(X_i) = f(X_i) + b*h$

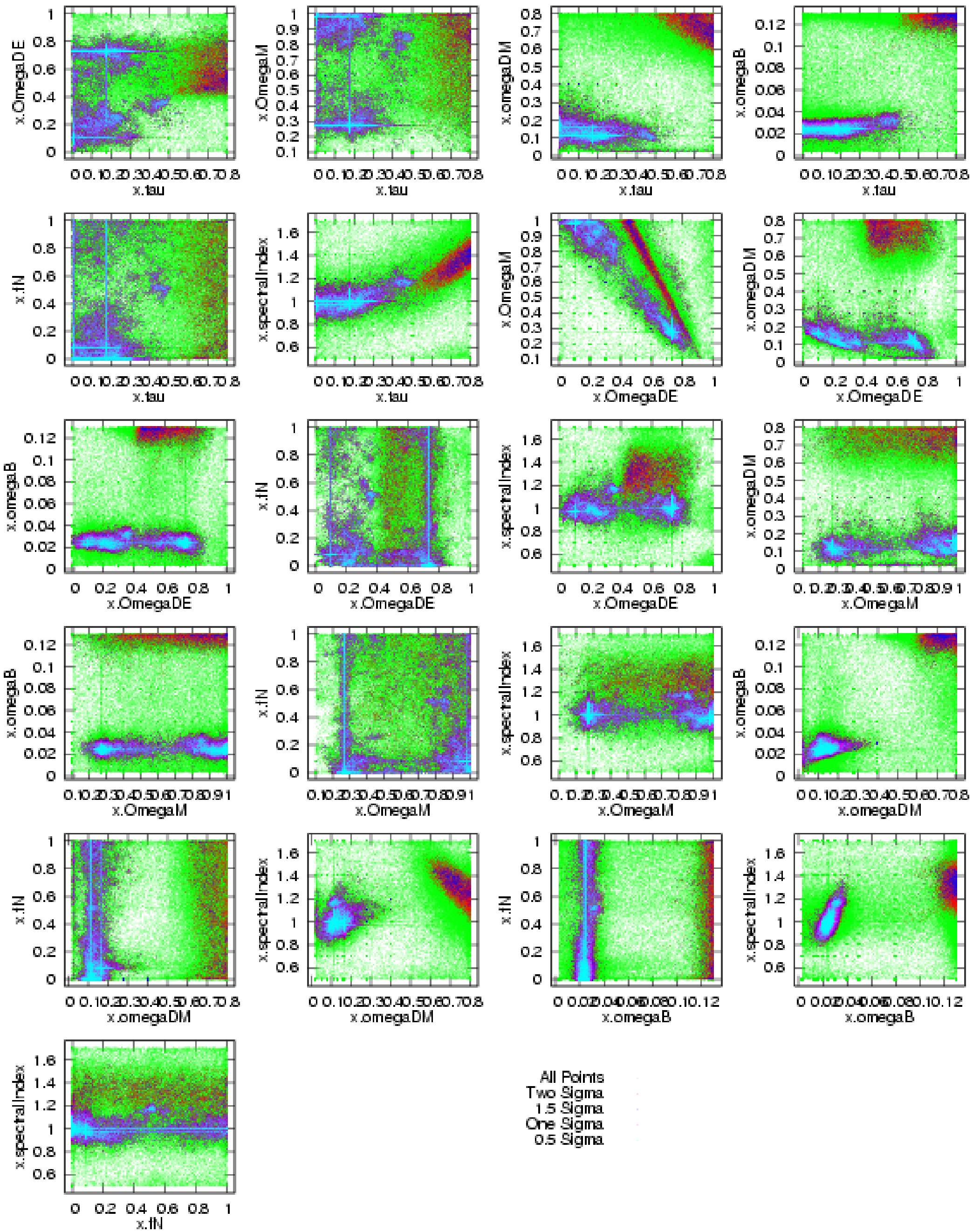




Using CMBfast we can make parametric models (11 parameters) and test if they are within the “confidence ball”. Varying Ω_b we get a range of 0.0169 to 0.0287

Testing in high D

- Now we can now jointly search all 11 parameters in the parametric models and determine which models fit in the confidence ball (at 95%).
- Traditionally this is done by marginalising over the other parameters to gain confidence intervals on each parameter separately. This is a problem in high-D where the likelihood function could be degenerate, ill-defined and under-identified
- This is computational intense as billions of models need to searched, each of which takes ~minute to run
- Use Kriging methods to predict where the surface of the confidence ball exists and test models there.



7D parameter space

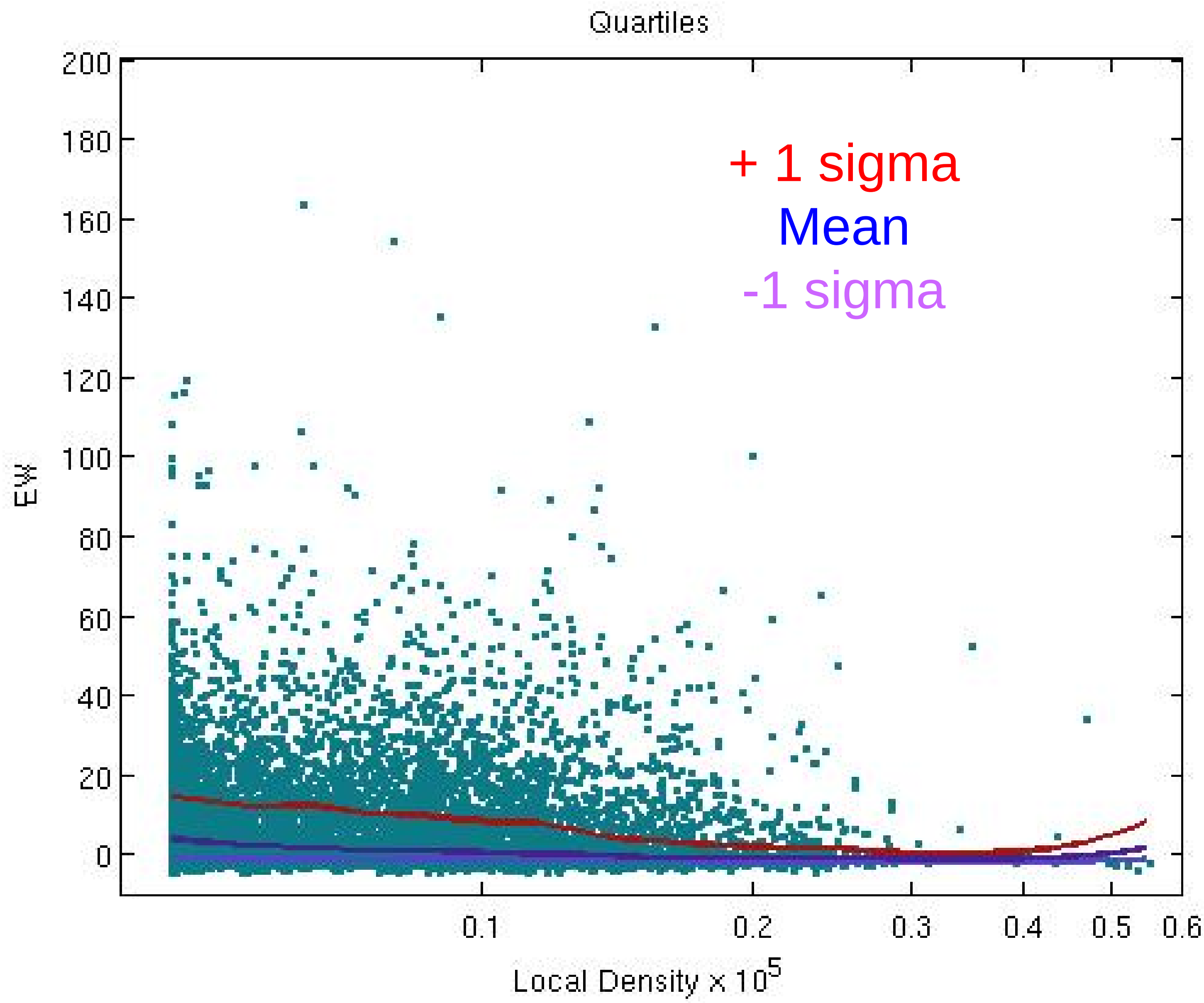
400000 samples

Cyan : 0.5σ
Purple: 1σ
Blue : 1.5σ
Red : 2σ
Green : >2σ

Bimodal dist. for several parameters

Other applications

- Physics of CMB is well-understood and people counter that parametric analyses are better (including Bayesian methods) *[however, concerns about CI]*
- Other areas of astrophysics have similar data problems, but the physics is less developed
  - Galaxy and quasar spectra (models are still rudimentary)
  - Galaxy clustering (non-linear gravitational effects are not confidently modeled)
  - Galaxy properties (e.g. star-formation rate)


Quartiles
+ 1 sigma
Mean
-1 sigma
EW
200
180
160
140
120
100
80
60
40
20
0
0.1
0.2
0.3
0.4
0.5
0.6
Local Density x 10^5