Application of Bayesian methods to parameter estimation and model seletion in galaxy evolutionary studies

Stefano Andreon
INAF-Oss. Di Brera
stefano.andreon@brera.inaf.it

This talk is an open discussion.

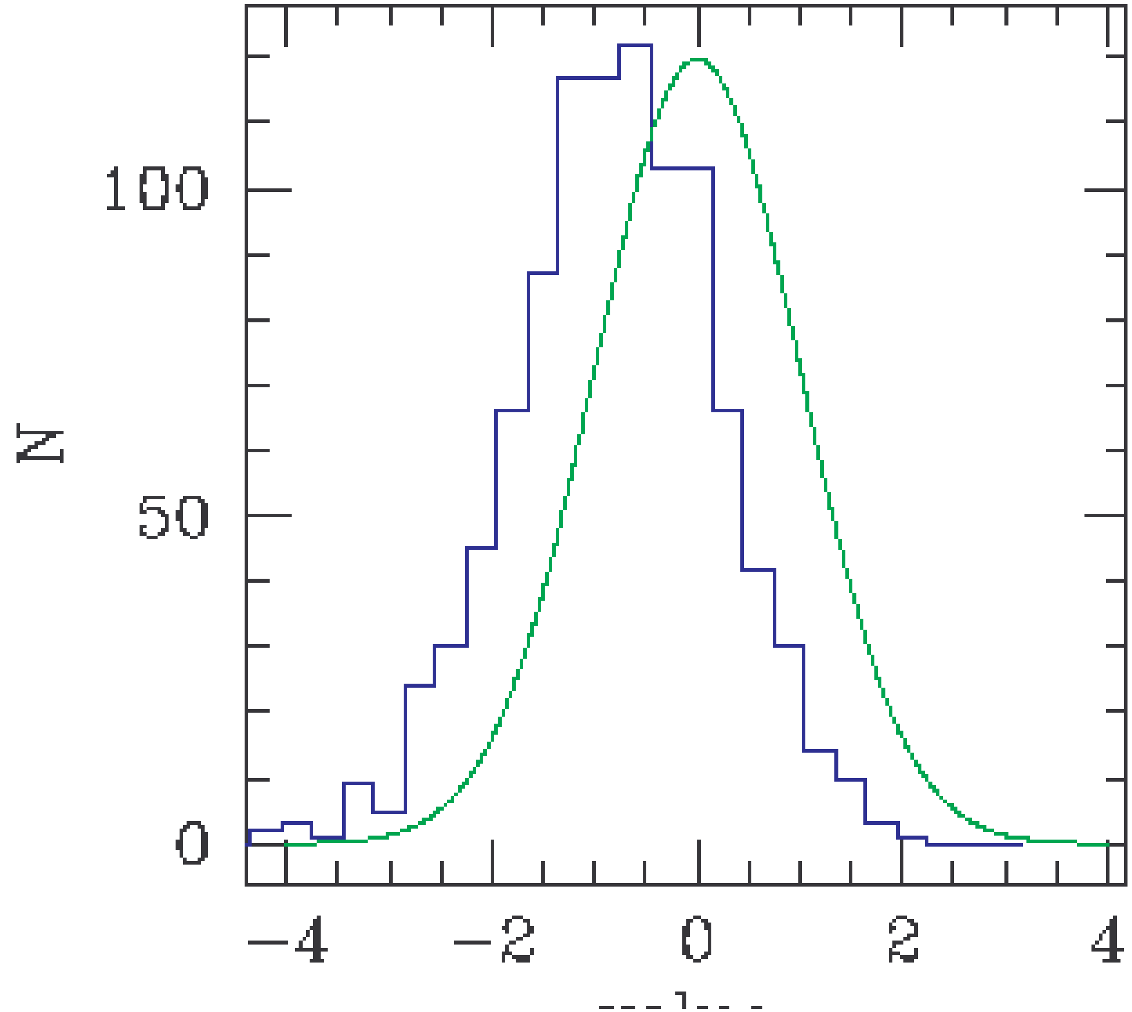
Don't esitate: ask me when I am unclear, when you want to make comments or to pose questions, to suggest a solution, or ... when you want

I'm provocative, but serius.

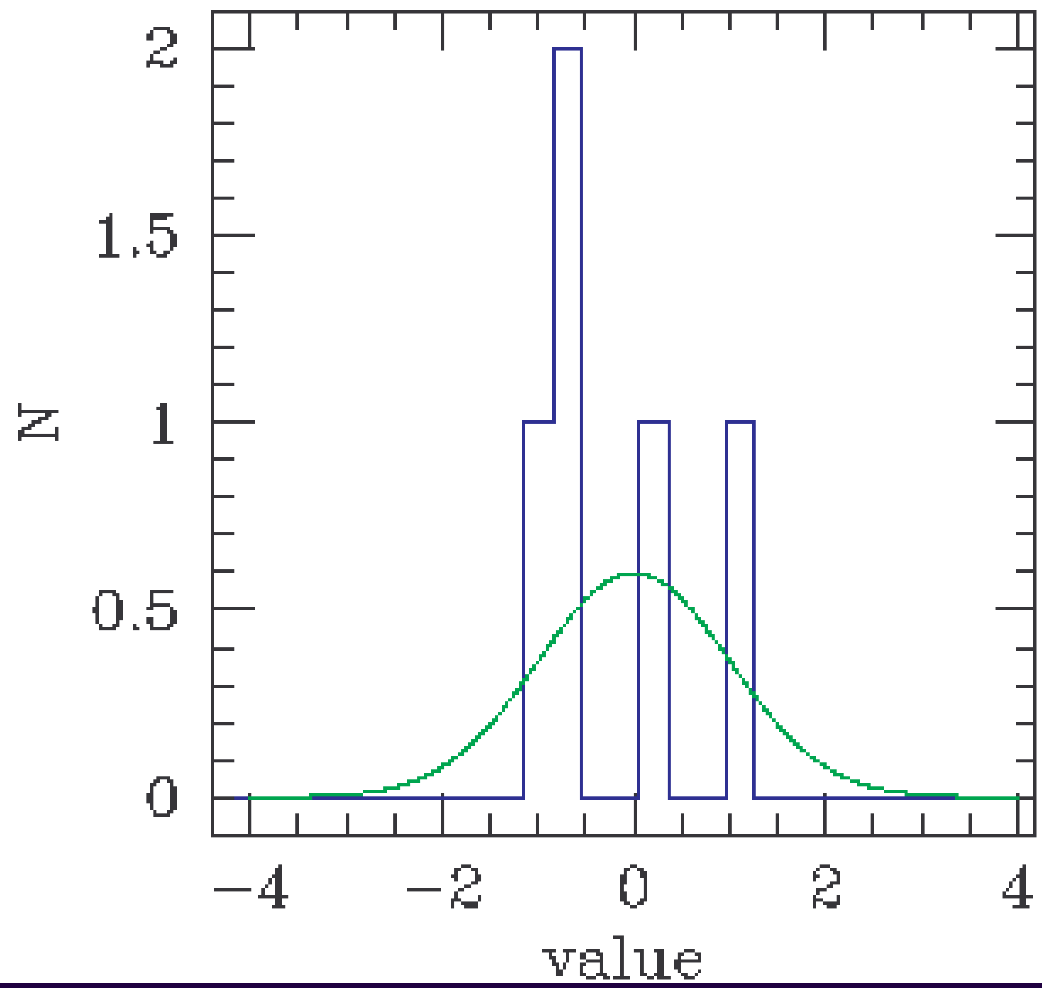
Caveat: talk for astronomers.

I am an astronomer, why be bothered with stats?

We don't need stats to say that these data and model differ, right?

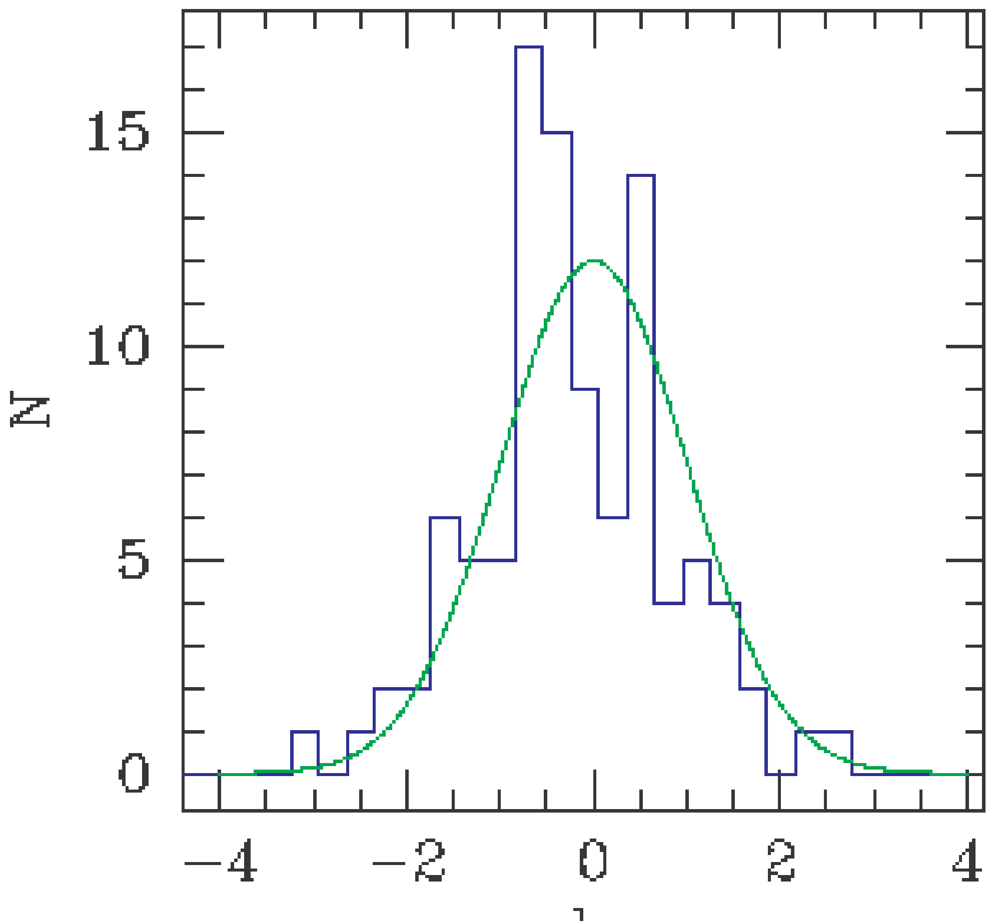


We don't need stats neither to say that the only thing to do is to go back to the telescope, right?



When we need stats, then? When the significance of the effect is not overwelming, like 3 sigma or so. In such cases, the uncertainty should be determined better than a factor 2 (unless you like 1.5 sigma results as you like 6 sigma results).

Therefore, the uncertainty should be carefully determined, to discriminate significant from insignificant results.

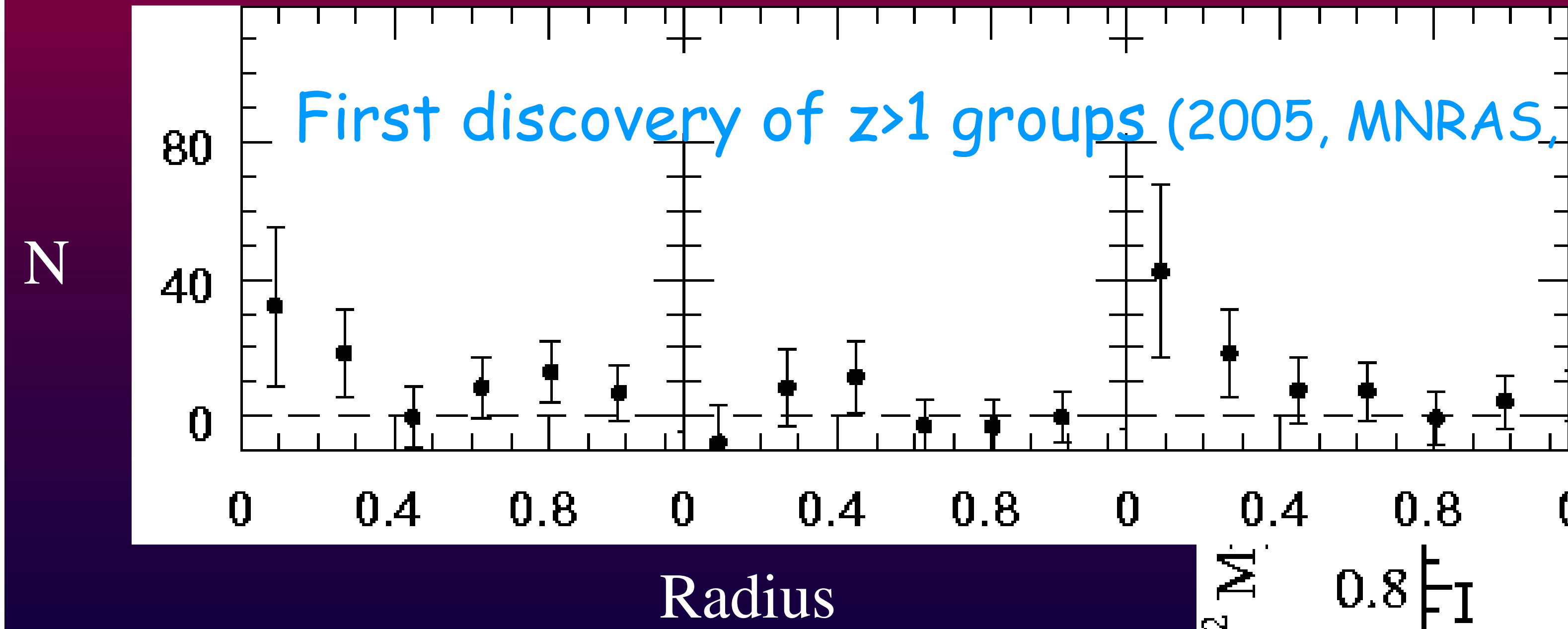


One more reason to pay attention to the statistical analysis: not to find impossible (unphysical) results i.e. finding something that cannot occur, like negative masses

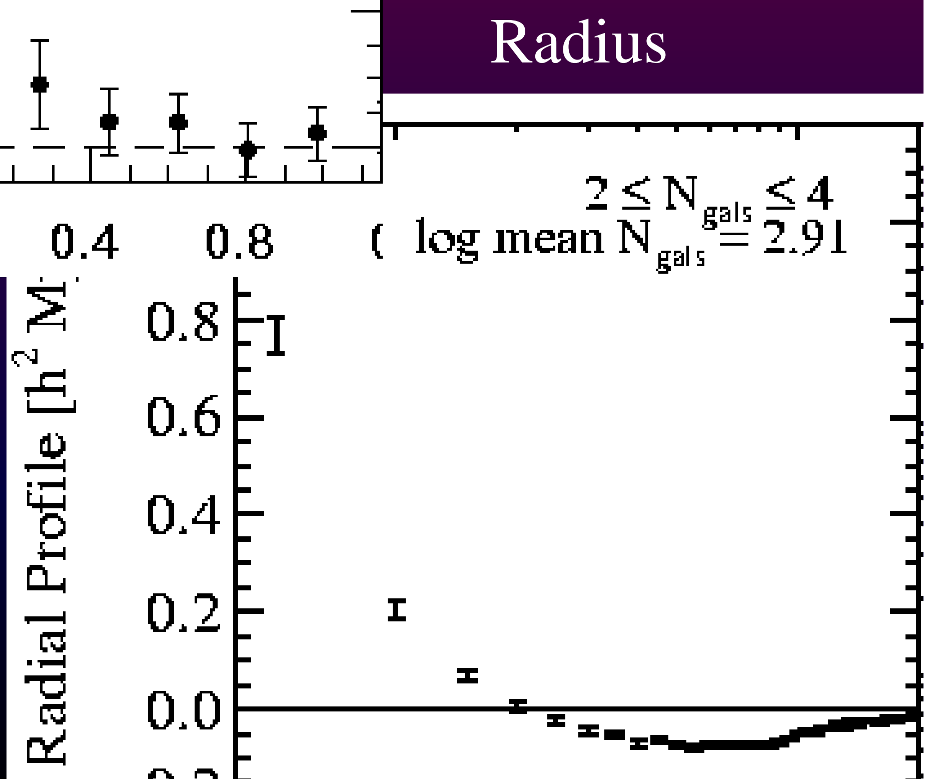
Here are some published (and sometime famous) examples

Galaxies come in positive units. Maybe.

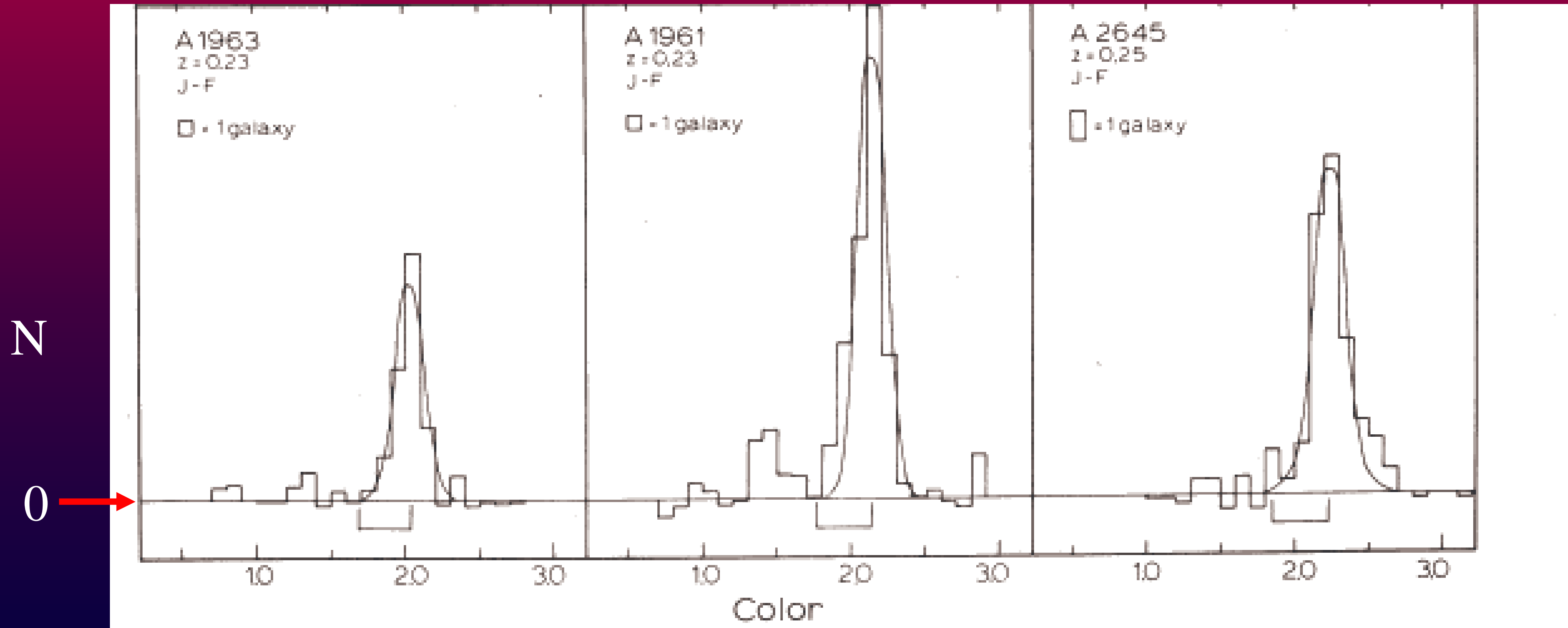
First discovery of z>1 groups (2005, MNRAS, 357, 1357)



Average of many local groups (2005, ApJ 633, 122)

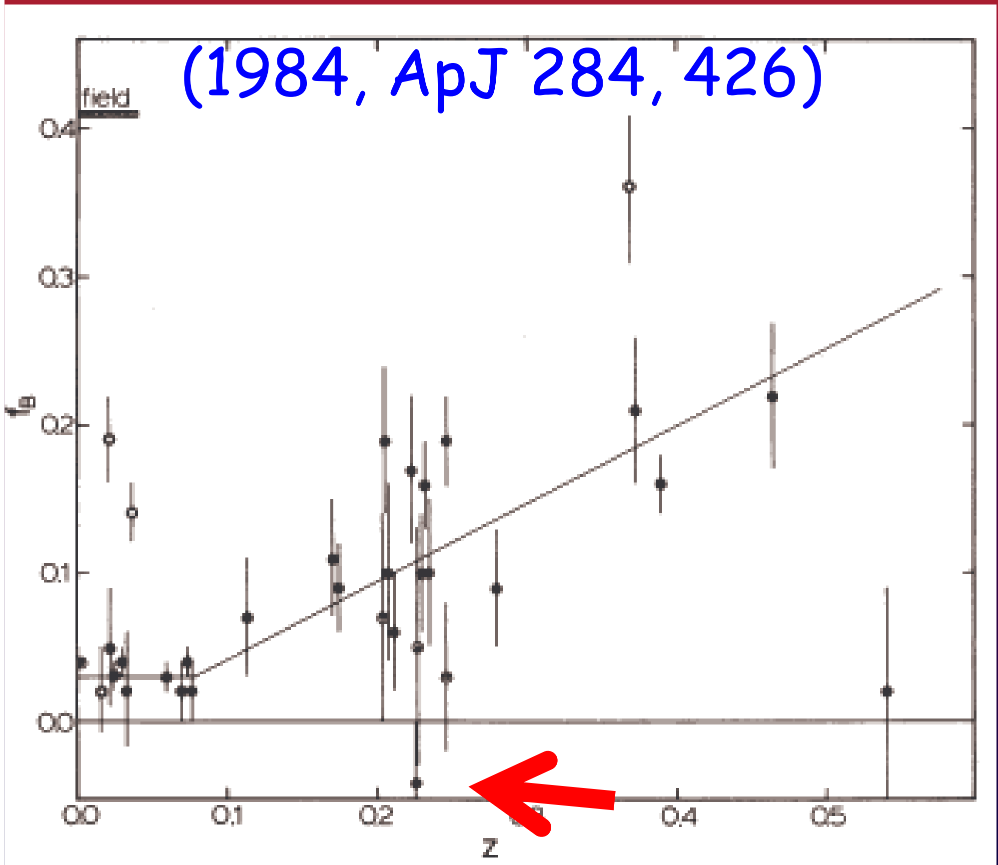


colour distribution



(1984, ApJ 284, 426)

… fractions smaller than 0



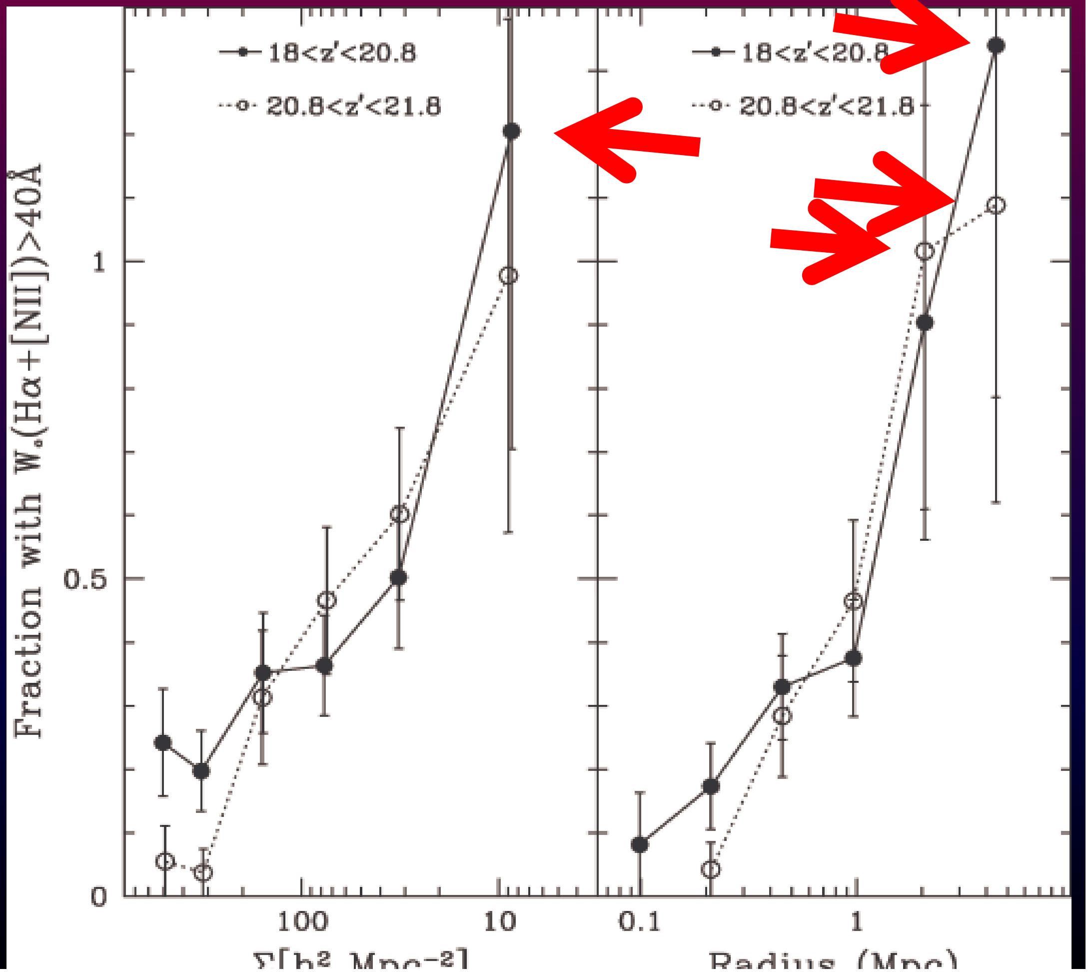
Maximum likelihood (best) estimates sometime fail to provide acceptable results.

Every physically acceptable value is better than the claimed best value!

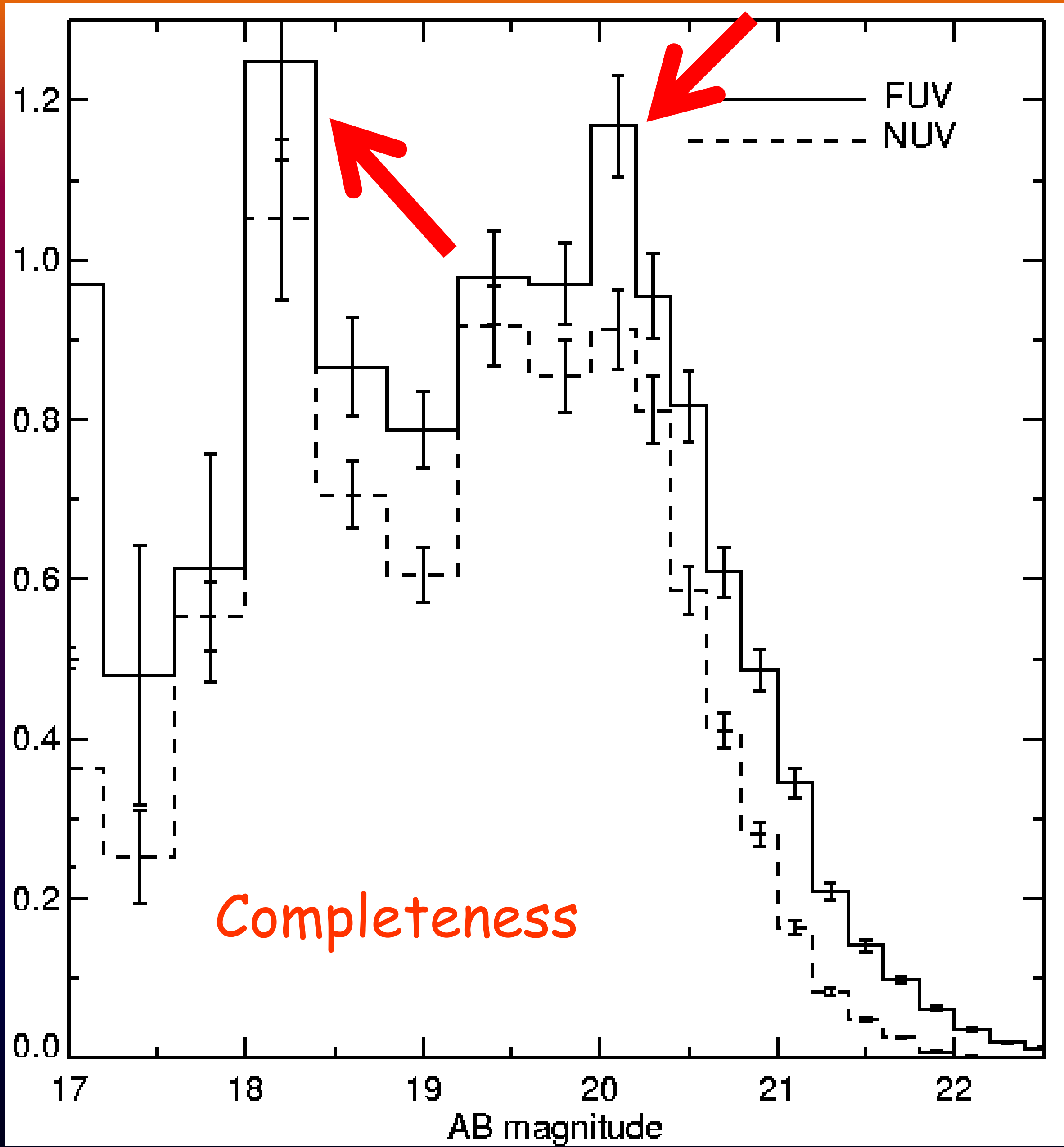
What "best" does it means?

… fractions larger than one

(2004, MNRAS 354, 1103)

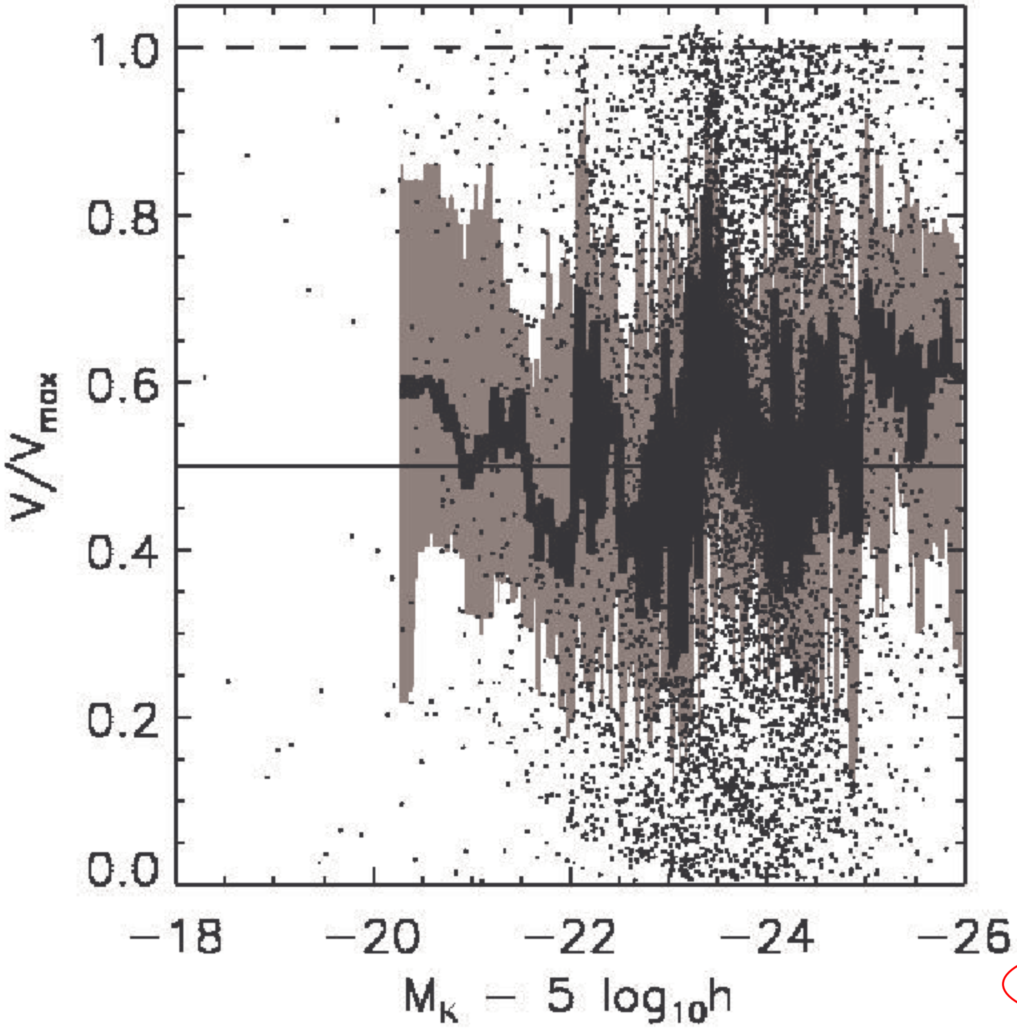


Completeness larger than 100 %



(2005, ApJ 619, L15)

V/Vax > 1



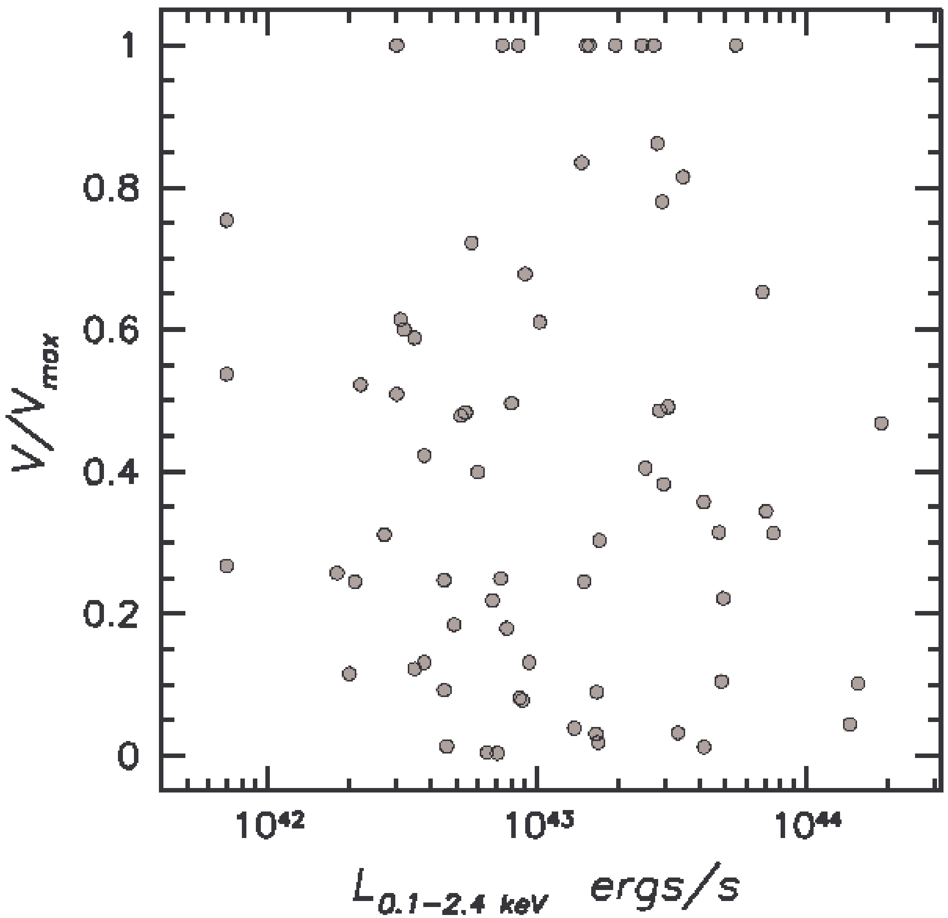
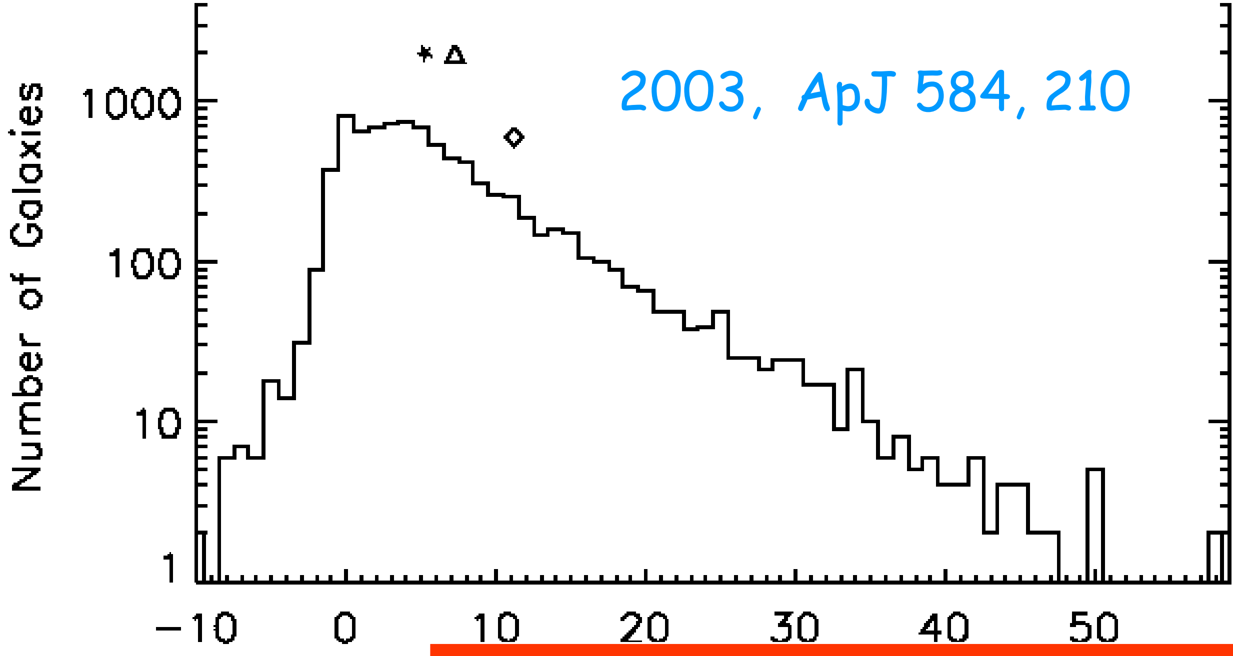


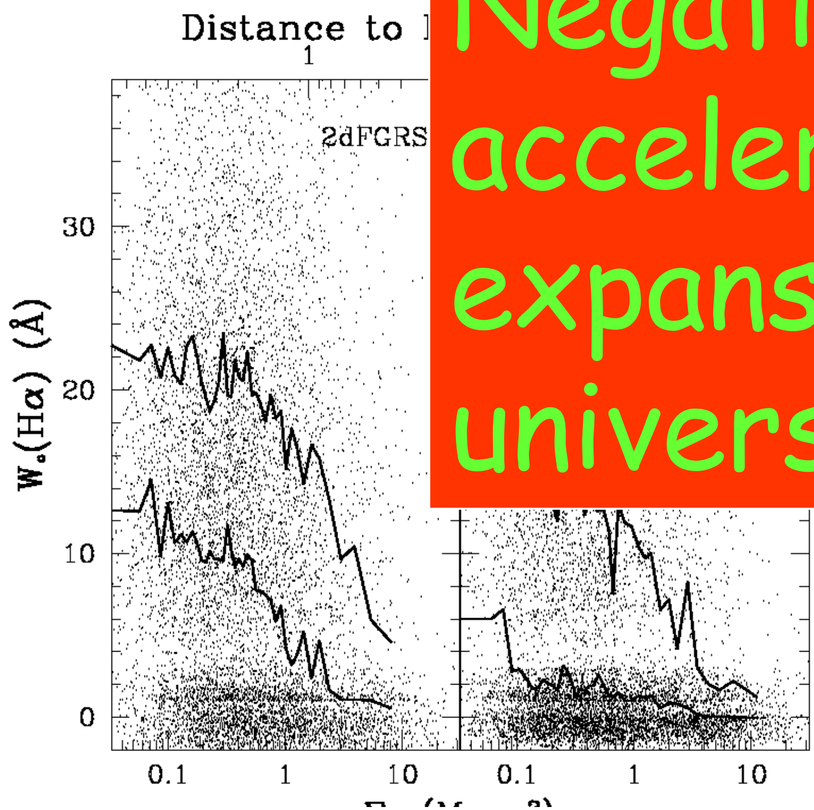
FIG. 8.— Test for the sample redshift completeness (V/V_{max}). The estimates exceeding 1 (due to the scatter in the flux-area relation) are substituted with 1. The number of clusters above and

ApJS, 149, 289

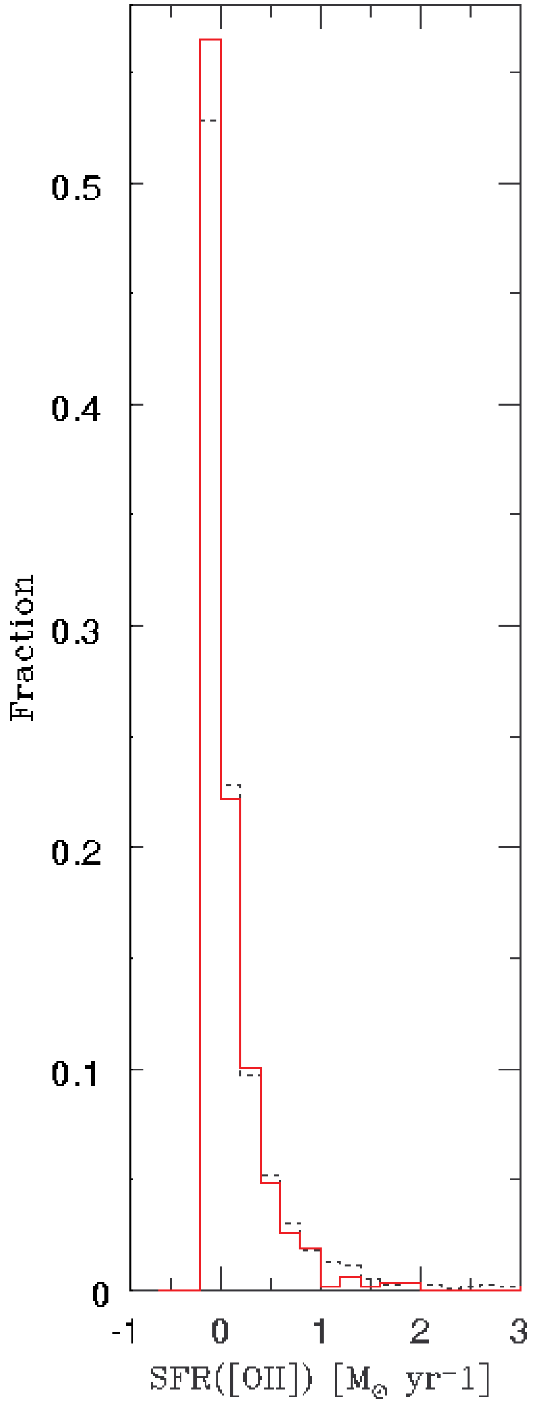
ApJ, in press (special issue)

2003, *ApJ* 584, 210



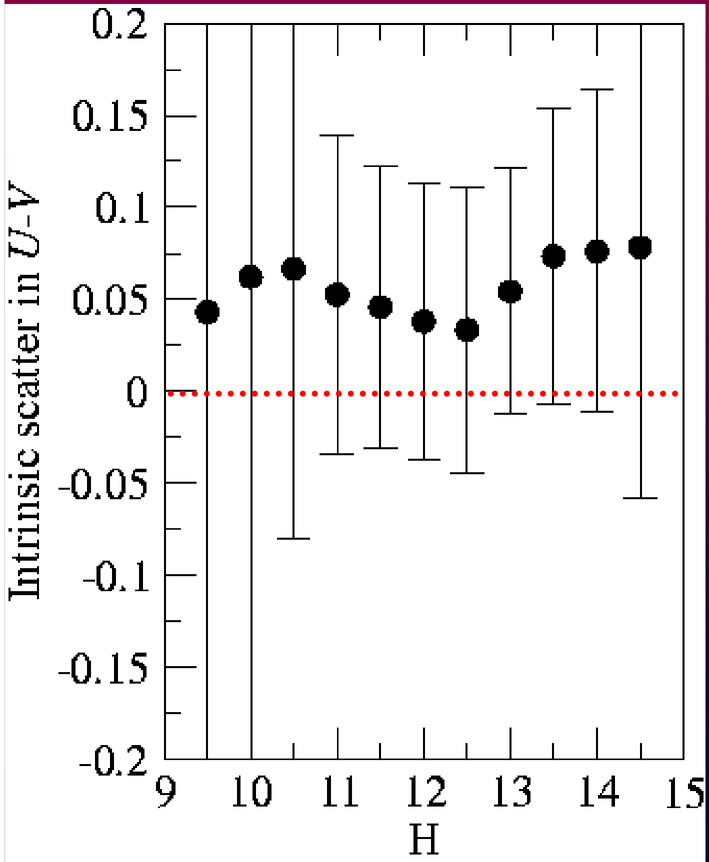


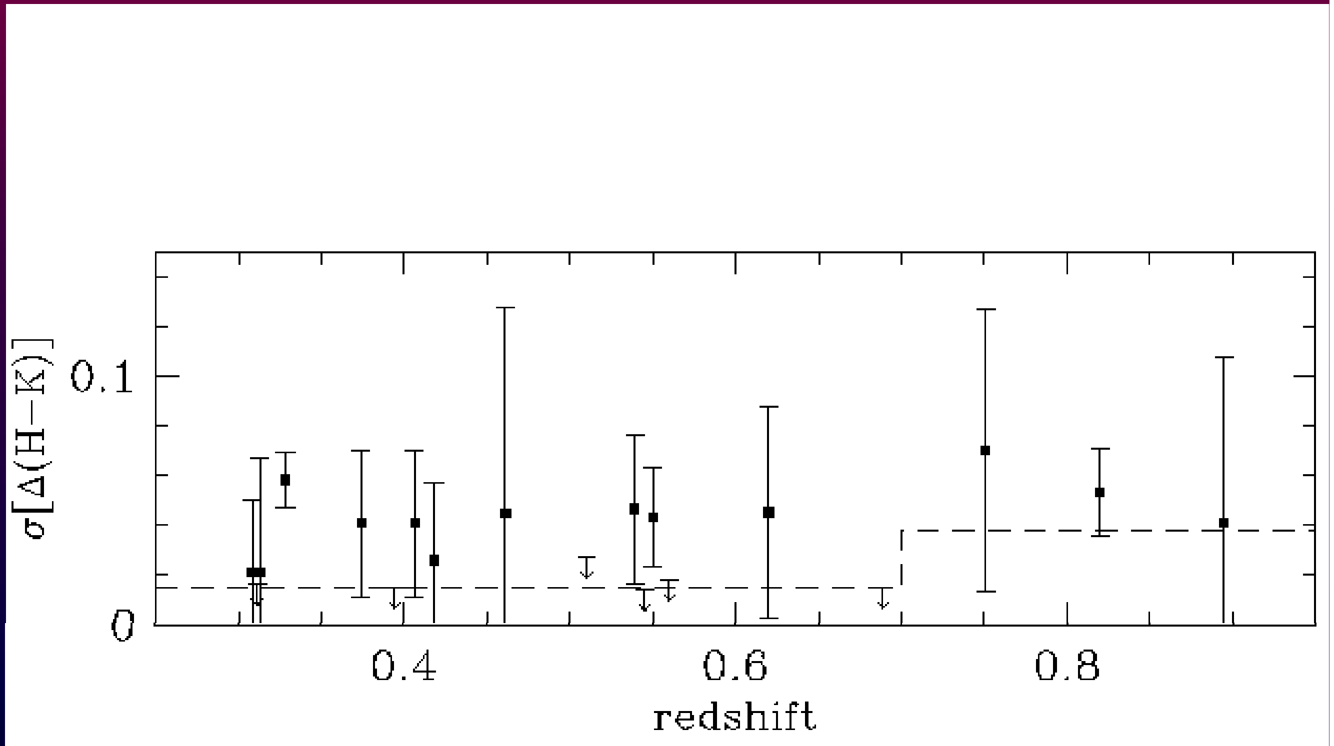
Negative masses accelerating the expansion of the universe? (joke)



2005, ApJ 624, 571

Scatter may plausibly be negative



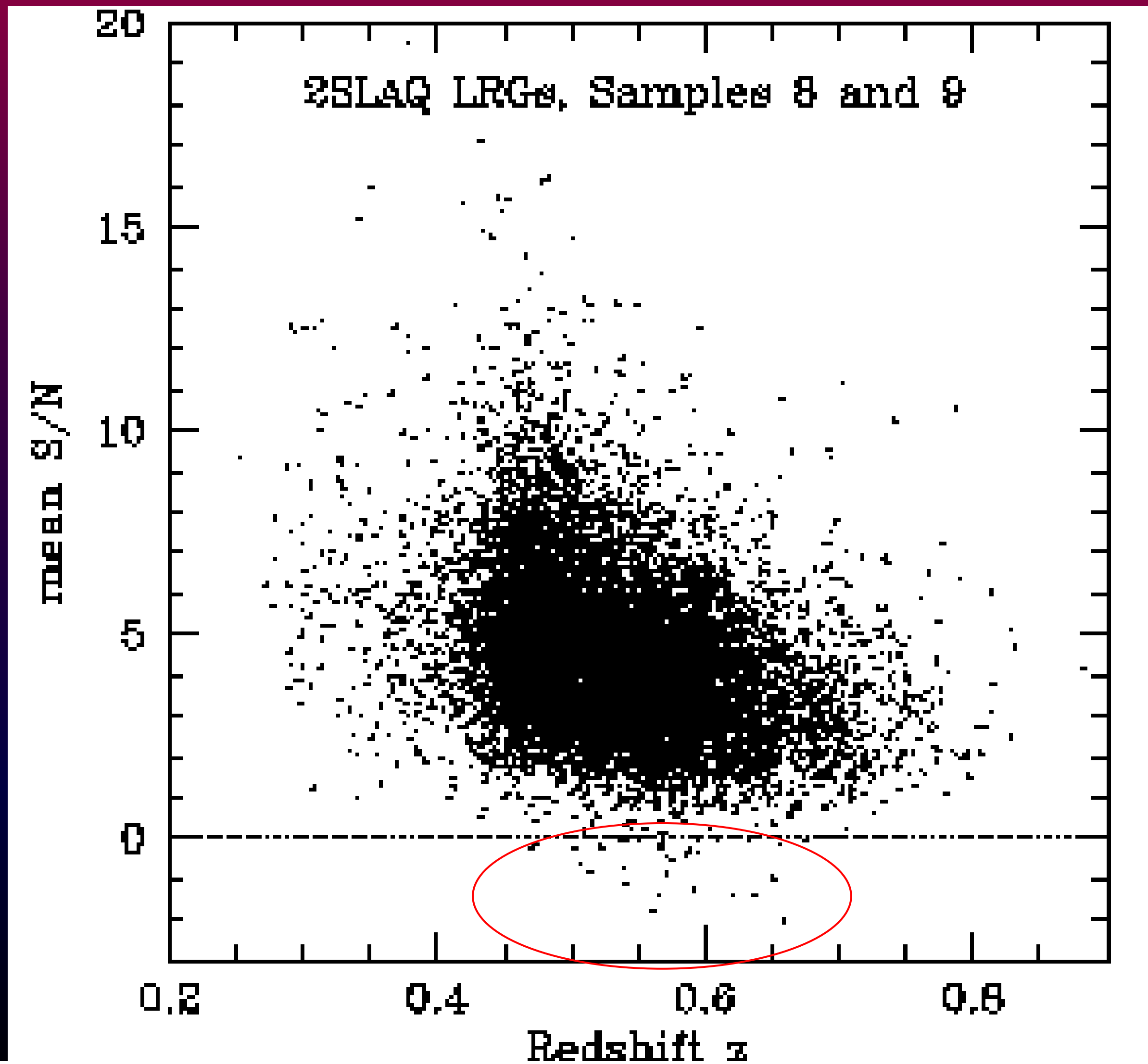


ApJ, in press (Coma cluster)

ApJ 492, 461

S/N can be negative

2006, MNRAS 372, 425



In the previous plot, the error bar is perhaps too large (including impossible values), but often it is too small, not correctly quantifying the degree of uncertainty.

One example:

Temperature errors of faint extended xray sources

ApJ, in press, special number

TABLE 1
CATALOG OF THE IDENTIFIED X-RAY CLUSTERS.

ID	R.A. Eq.2000	Decl. Eq.2000	r_{500} '	flux 10^{-14} ergs cm^{-2} s^{-1}	z	M_{500} 10^{13} $M_\odot$	$L_{0.1-2.4keV}$ 10^{42} ergs s^{-1}	kT keV	z source
3	150.80244	1.98985	1.0	0.70 ± 0.16	0.25	0.69 ± 0.14	2.2 ± 0.5	0.56 ± 0.06	1
9	150.75121	1.52793	1.1	1.97 ± 0.38	0.75	9.62 ± 1.65	75.4 ± 14.5	2.62 ± 0.23	1
11	150.73676	2.82680	0.7	0.40 ± 0.08	0.60	1.88 ± 0.35	10.2 ± 2.1	1.06 ± 0.10	1
15	150.67342	2.09190	0.9	0.50 ± 0.09	0.34	0.88 ± 0.13	3.2 ± 0.5	0.65 ± 0.05	0
20	150.64041	2.12791	0.8	0.43 ± 0.06	0.55	1.78 ± 0.23	9.0 ± 1.3	1.01 ± 0.07	1
24	150.58962	2.87187	0.7	0.36 ± 0.08	0.95	3.41 ± 0.66	30.6 ± 6.6	1.61 ± 0.16	1
25	150.58631	1.92693	1.1	1.04 ± 0.10	0.30	1.36 ± 0.12	4.9 ± 0.5	0.81 ± 0.04	1
29	150.53842	2.37393	1.3	1.28 ± 0.15	0.18	0.62 ± 0.07	1.8 ± 0.2	0.52 ± 0.03	0
32	150.50535	2.22395	1.2	3.45 ± 0.11	0.90	18.64 ± 0.52	189.7 ± 5.9	3.90 ± 0.06	1
34	150.49330	2.06795	0.9	0.61 ± 0.07	0.40	1.47 ± 0.14	6.0 ± 0.6	0.87 ± 0.04	1
36	150.49048	2.74592	0.6	0.22 ± 0.05	0.65	1.34 ± 0.30	7.4 ± 1.9	0.90 ± 0.10	1
38	150.44824	1.91197	0.6	0.26 ± 0.04	1.25	3.49 ± 0.50	45.1 ± 7.3	1.79 ± 0.13	1
39	150.44827	2.04996	1.3	2.54 ± 0.13	0.55	8.10 ± 0.38	48.4 ± 2.5	2.25 ± 0.05	1
40	150.44523	1.88197	0.6	0.26 ± 0.04	0.70	1.70 ± 0.23	10.4 ± 1.6	1.03 ± 0.07	1
41	150.44229	2.15796	0.7	0.30 ± 0.06	0.40	0.77 ± 0.13	3.0 ± 0.6	0.62 ± 0.05	1
42	150.41533	2.43096	2.8	11.45 ± 0.26	0.12	2.26 ± 0.05	7.1 ± 0.2	1.01 ± 0.01	0
44	150.42124	1.98397	0.8	0.42 ± 0.07	0.45	1.26 ± 0.19	5.4 ± 0.9	0.81 ± 0.06	1
45	150.42121	1.84898	0.7	0.49 ± 0.05	0.85	3.70 ± 0.35	29.5 ± 3.1	1.63 ± 0.08	1

1 % accuracy?

All the above oddities are related to:

1) Measurements near boundaries (fractions, completeness, hardness ratios, sources/features with few counts ...)

-> accounting for boundaries

2) nuisance parameters

I.e. I would like to measure an interesting parameter, without precise knowledge of another parameter that I known to influence the measurement:

-> parameter estimation in presence of a nuisance parameters

Bayesian methods solve mentioned odities

Lesson zero: all what you need to assume is:

a) product rule/axiom

$$p(x,y) = p(x|y)*p(y)=p(y|x)*p(x)$$

ex: in a bag I have 4 blue balls and 10 red balls. If I extract two of them without replacement, what is the probability that both are red?

The probability of getting red the first ball, $p(x)$, is =10/14

The probability of getting red the second ball, after having get a red ball in the first extraction, $p(y|x)$, is = 9/13

The probability of getting red both, $p(x,y)$, is the product

10/14* 9/13= $p(y|x)*p(x)$

b) the sum rule/axiom

$$p(x) = \Sigma_y\, p(x,y) = \int p(x,y)dy$$

Ex: in a bag I have 4 blue balls, 10 red balls, 5 blue dies and 3 green cards. What is the probability of extracting something blue?
The sum of the probability of extracting a blue ball (4/22) and the probability of extracting a blue dies (5/22), =9/22

usefull to measure uncertainty in presence of nuisance parameters, as explained later.

c) Bayes theorem

$$p(\theta|data) = c * p(data|\theta) * p(\theta)$$

Posterior = c* Likelihood* prior

Can be derived from the product rule, or, in alternative, assumed as axiom, and the product rule derived.

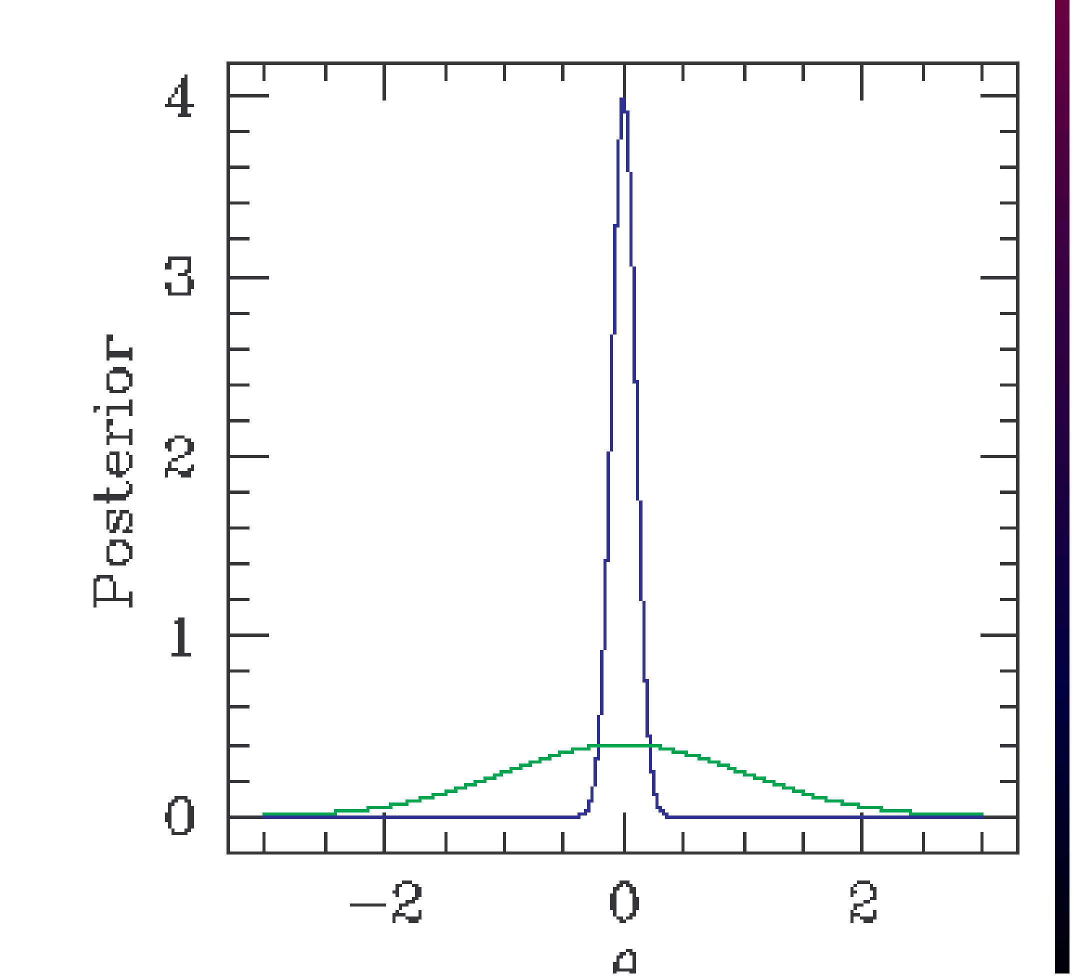
Central tool for parameter estimation

The posterior width quantifies the uncertainty

$$p(\theta|data) = c * p(data|\theta) * p(\theta)$$

if $p(\theta|data)$ is a narrow (almost delta) function, θ is very well determined
if $p(\theta|data)$ is a flat function, θ is badly determined

Do you want to known the uncertainty? Compute the posterior, and its width! This is the mantra of most applications: spell a prior, compute the likelihood, multiply them, and compute the width of the result.



Everything comes from the two axioms, no other ingredients used.

Lesson one: where unrealistically small errors comes from and how to avoid them

Bayesian solution to nuisance parameters

Empirically, claimed LF errors are underestimated by a factor 2 at least (Andreon, 2004, A&A 416, 865)

consequence of (having forgot) the sum rule of probabilites:

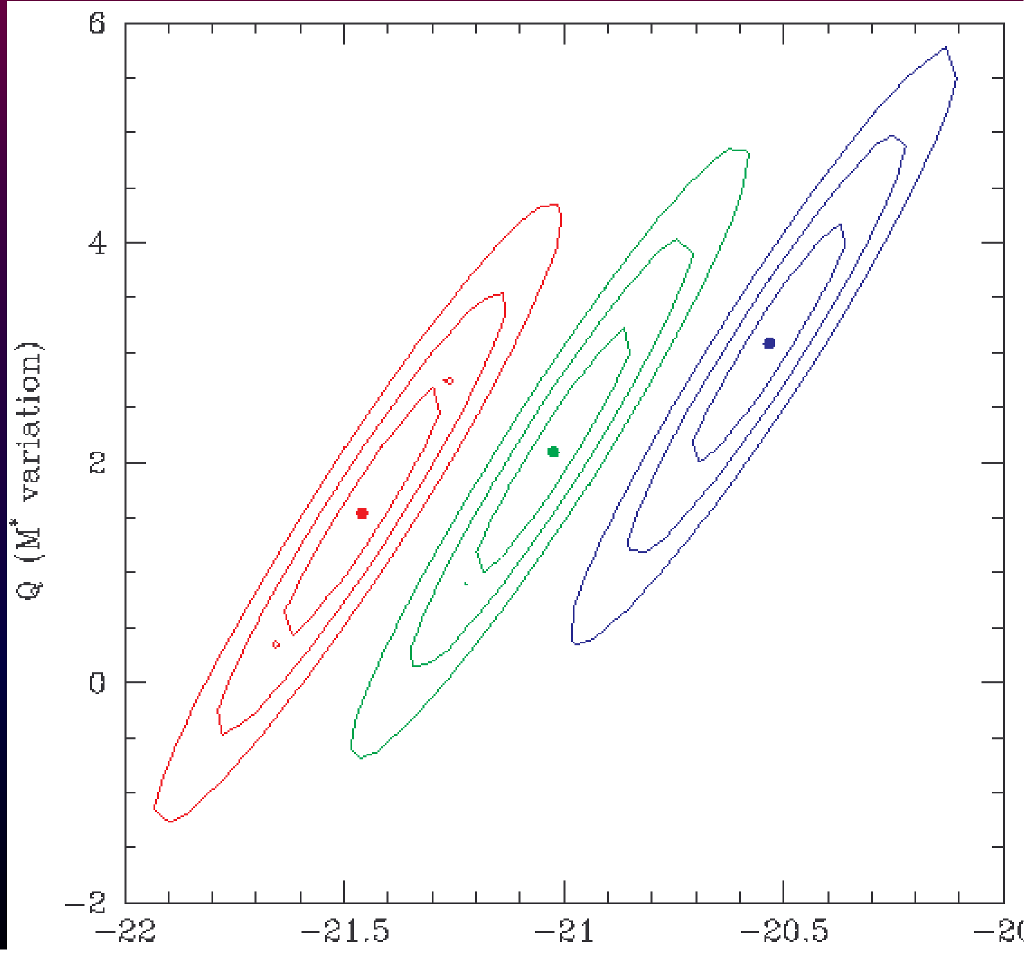
$$p(x) = \int p(x,y)dy$$

Std derivation assumes no evolution on M* in order to derive it, i.e.

$$Q = \frac{\partial M^*}{\partial z} = 0$$

the unknown nuisance parameter Q has been taken fixed instead of marginalizing over it.

Common mistake. Other example: keep alpha fixed when observations do not constraint it, keep T fixed in Lx dets



Lesson two: where unpossible results (often) come from and how avoid them

Poisson signal in presence of a bkg.

The astronomical recipe for background subtraction of Poisson signals, 'unbiased estimate' of sampling theory:

n_net = n_tot - n_bkg (Zwicky 1957, Oemler 1973, etc.)

But, if

n_tot = 3 (galaxies, photons, whatever)

n_bkg = 5 (idem) true (average) mean value perfectly known

what about n_net?

n_net = n_tot - n_bkg = -2 ?

what does it means to have observed a negative signal when it is defined to be positive?

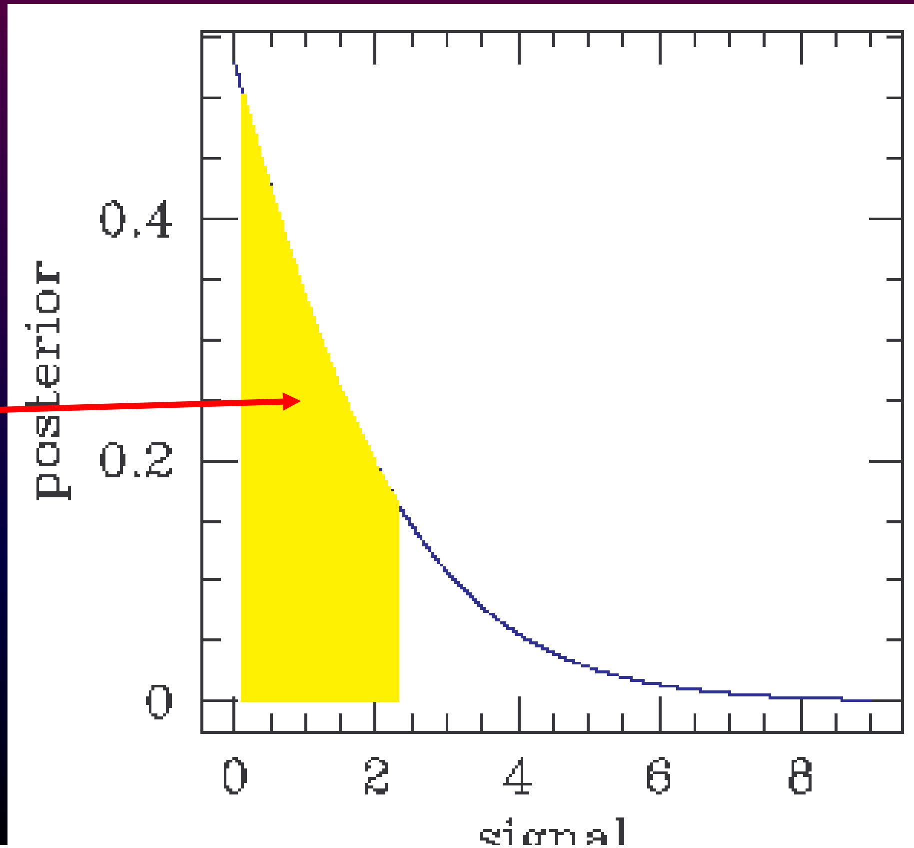
Bayesian solution to boundary problem: the Bayes theorem

p(signal|data)=c*p(data|signal)*p(signal)

= c* Poisson(x=3; λ=5+signal) * p(signal)

Because of the prior (p(signal)=0 if signal<0), posterior($\forall$signal<0)=0

uniform prior, mean: 1.63; shading: shortest 68 % confidence interval



Lesson three: Bayes theorem embodies the correction for Malmquist bias

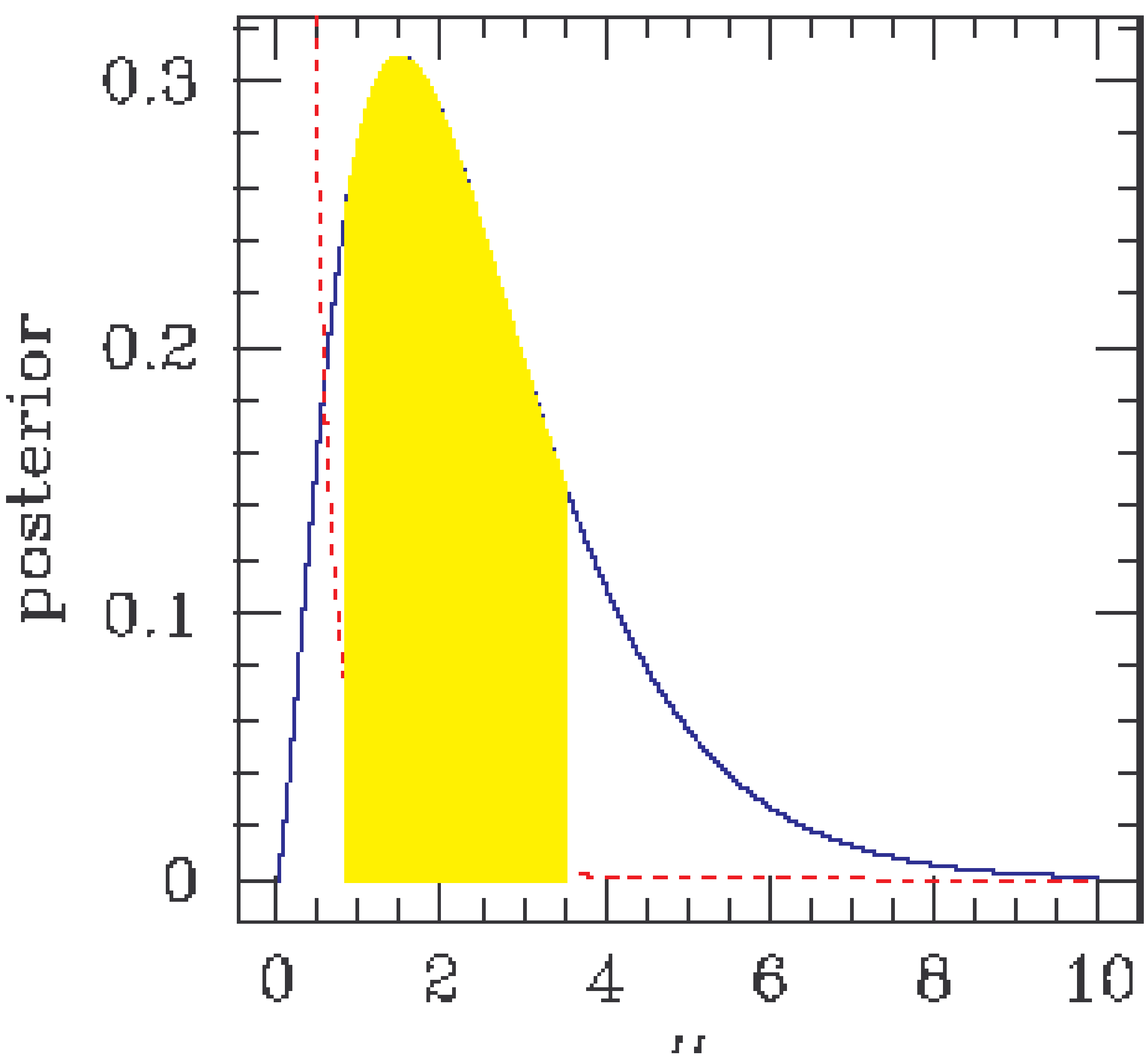
p(θ|data)= c * p(data|θ) * p(θ)

ex. modelled around Kenter et al 2005, ApJS 161, 9 (x-ray survey with sources as faint as 2-4 photons).

$p(\mu|4) = c^* \; p(4|\,\mu)\; p(\mu)$

at the studied fluxes, the prior $p(\mu)$ (=number counts for astronomers) is well known, $p(\mu)= \mu^{-\beta}$ with beta approx 2.5 (euclidian slope).

4 photons are observed but the maximum a posteriori (most probable) is about 1.5!

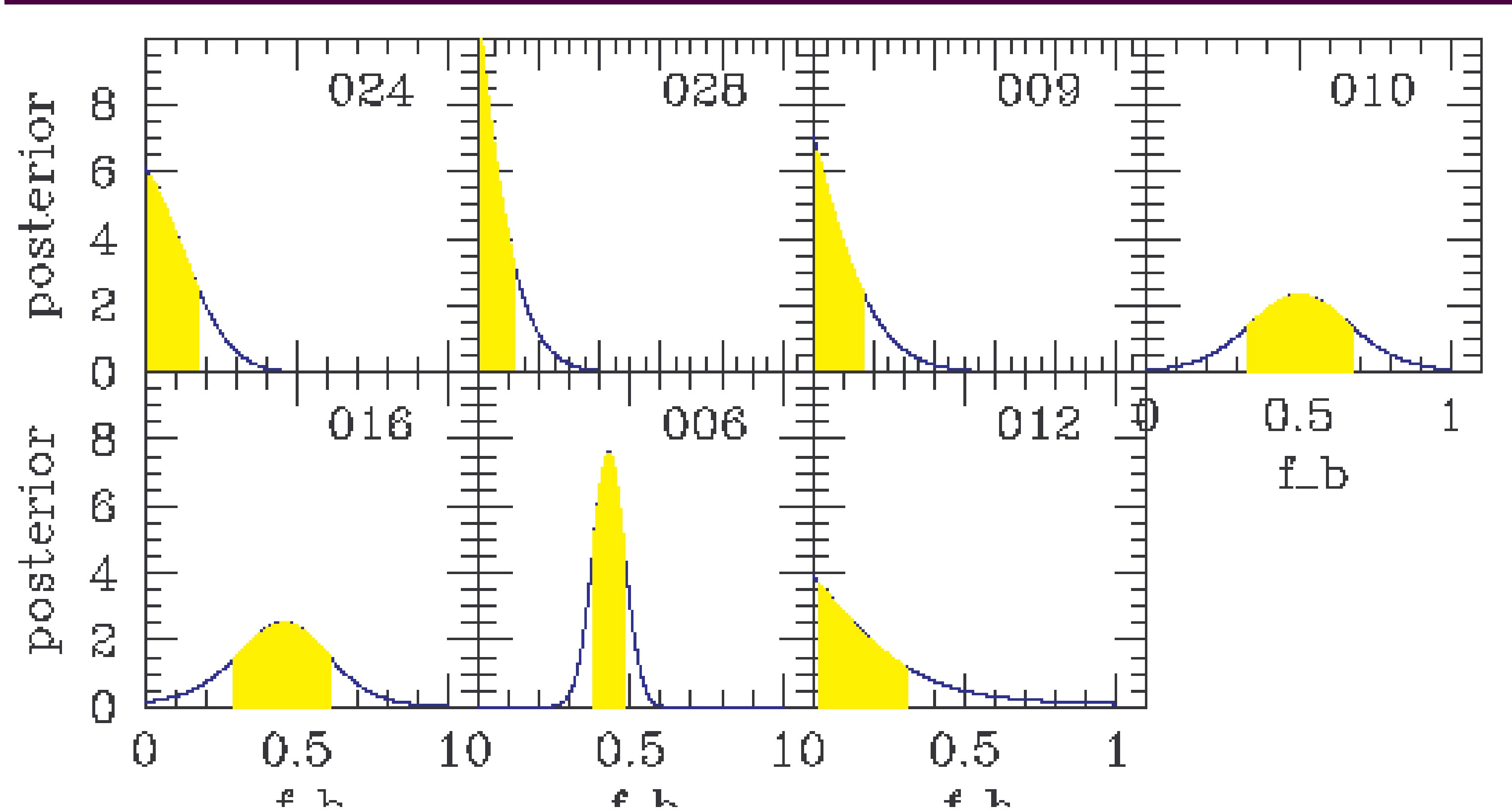


Lesson four: moving from textbook examples to research applictions

Life is hard: boundaries and nuisance parameters often got toghether.

Astro: The evolution of the blue fraction (Butcher-Oemler effect) measures the change with cosmic times of the star formation activity in clusters.

Stats: Fraction in presence of a background: D'Agostini (2004, physics/0412069)



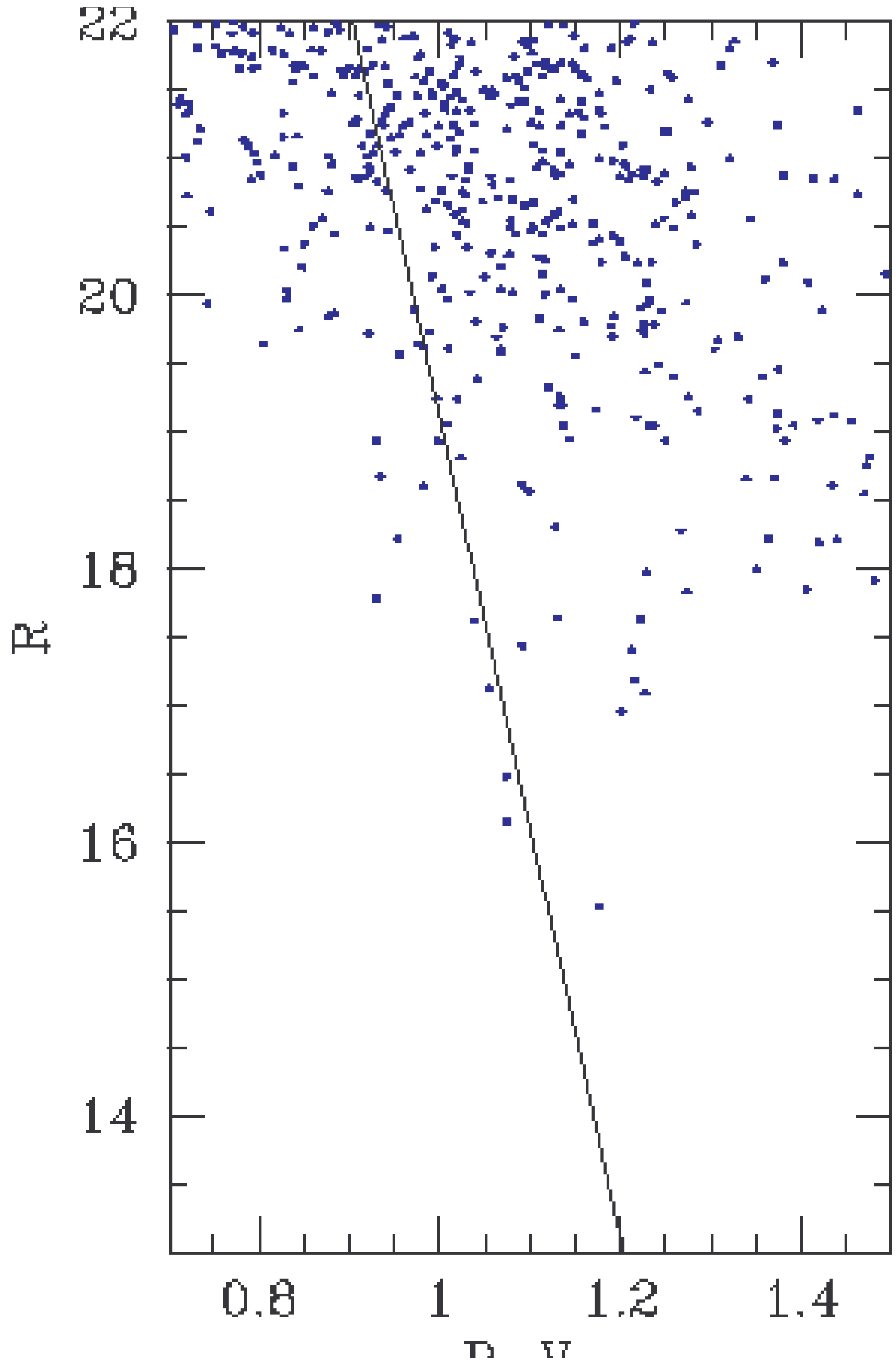
Andreon et al. (2006, MNRAS 365, 915).

Harder and harder.

Astro: the scatter around the colour-magnitude relation put a strong constraint on the age of stars in these objects and, indirectly, on the ages of the galaxies themselves.

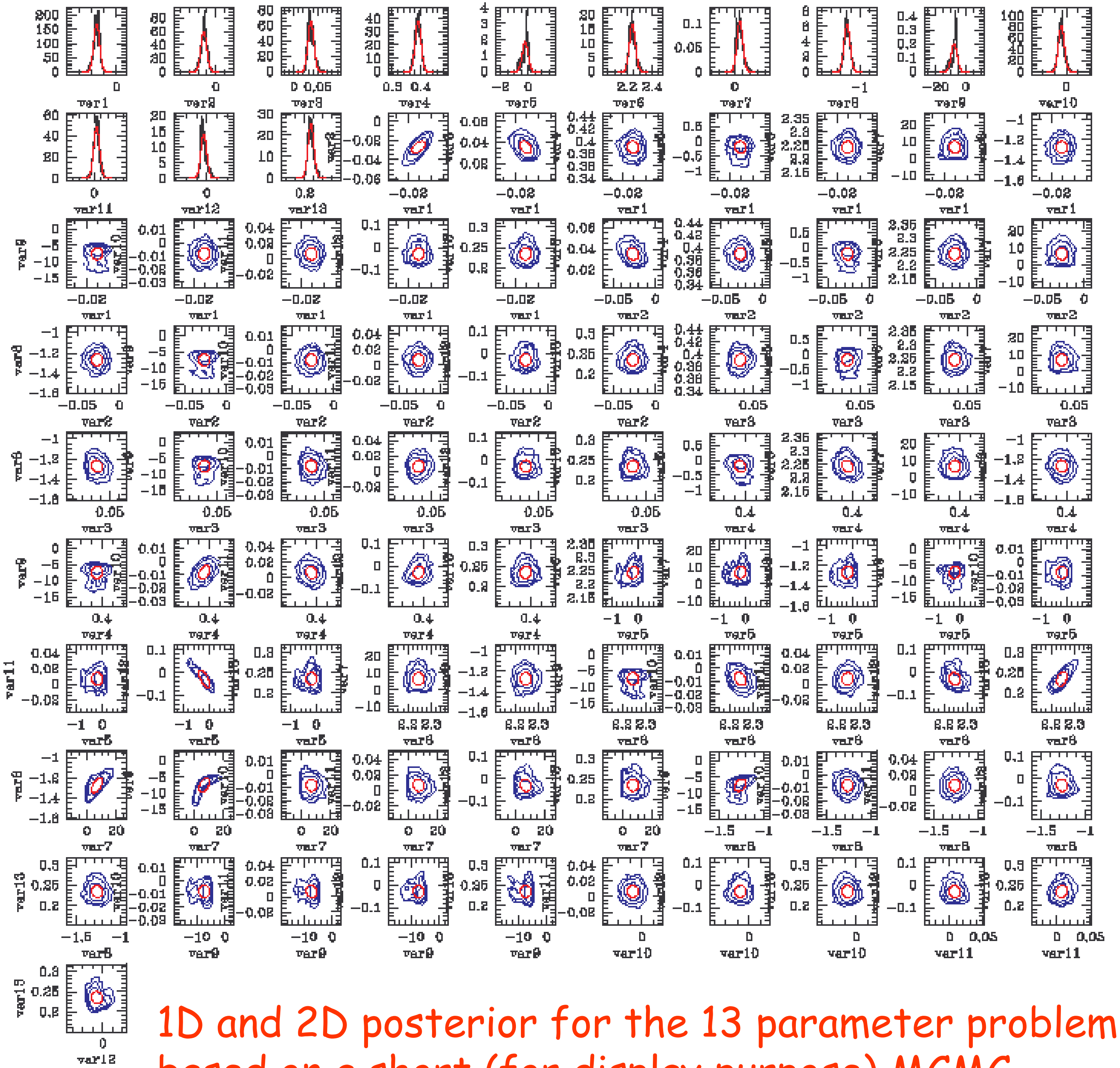
Problem: background galaxies (about 4 for every cluster galaxy).

Abell 1185 cluster (Andreon, Cuillandre et al. 2006, MNRAS 372, 60)



Stats: determination of the intrinsic scatter of a correlation in presence of heteroscedastics errors on both x and y (solution due to D'Agostini 2005, physics/0511182) without the precise knowledge of which galaxies are cluster members, i.e. in presence of a background possibly displaying another correlation (solution due to Andreon 2006, MNRAS, 372, 60).

Marginalization in a large dimensional space by using MCMC stochastical computations.



1D and 2D posterior for the 13 parameter problem, based on a short (for display purpose) MCMC

Lesson five: model selection

I'm often confronted with questions like: my models or hypothesis have to be rejected/refined, or a model is to be preferred to another one?

e.g.

i) do my data provide evidence for a luminosity evolution ?

ii) What fits better: a de Vaucouleurs or an exponential law?

$\ln(I(R))=R^{1/n}$ with N=4 (de Vaucouleurs) or N=1 (exponential)

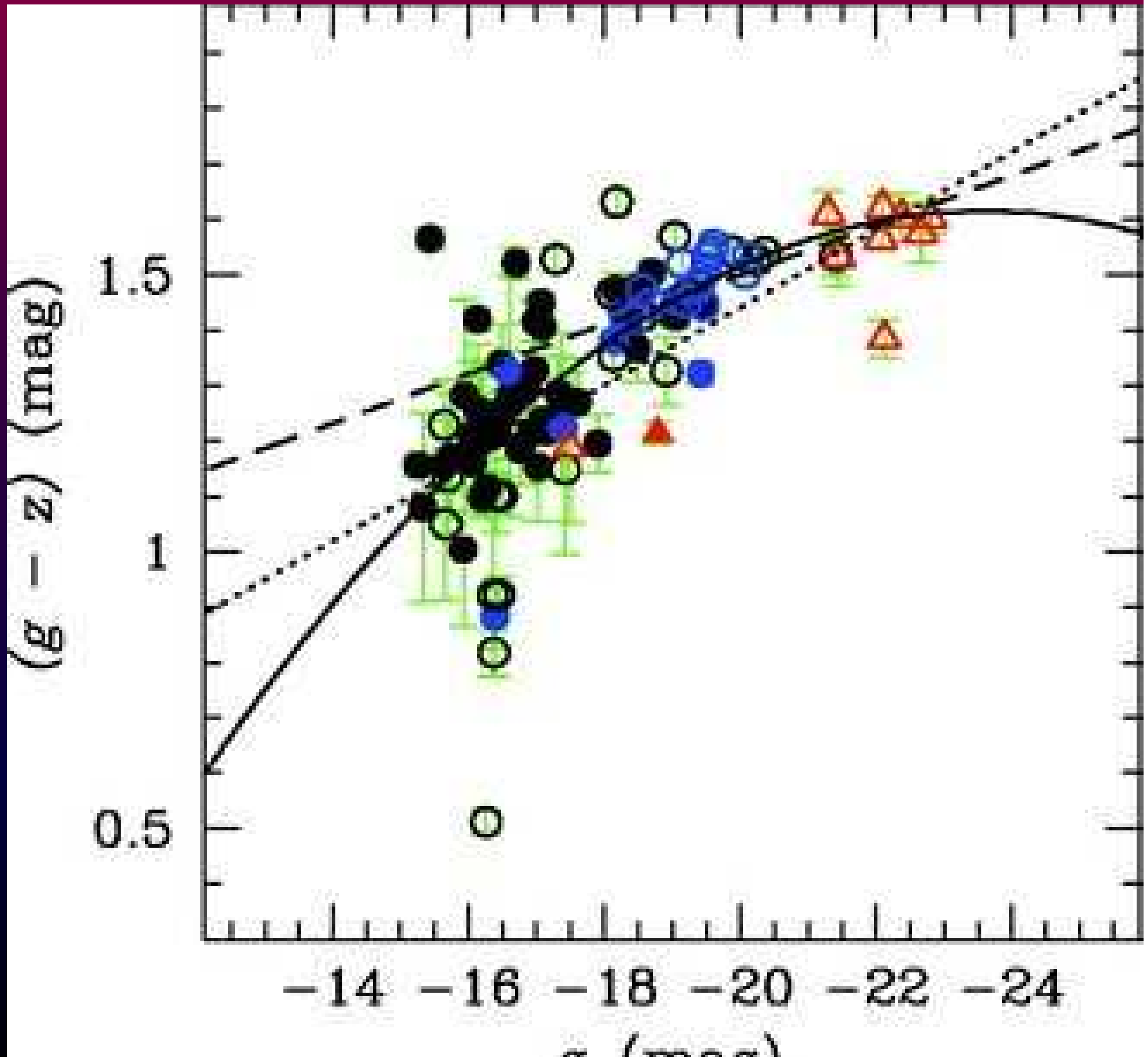
-> this task is named model selection

models are often not hierarchically nested.

Does the relation is linear or bended?

bendend, 2006, ApJS 164, 334

cannot said with these data, other data say “no”, Andreon et al. 2006, MNRAS 372, 60



How model selection works:

Directly from bayes theorem:

$$\frac{P(H_1 \mid D)}{P(H_2 \mid D)} = \frac{P(H_1)}{P(H_2)} \frac{P(D \mid H_1)}{P(D \mid H_2)}$$

evidence ratio = prior ratio * Bayes factor

includes the Occam razor, ie penalizes models with unnecessary complexity

allow to compare models, including not hierarchically nested models.

allow to accumulate evidence in favour of the simplest model.

Often hard to calculate, require integrations over large dimensional space.

Approximations are welcome, BIC is one of them.

BIC $= -2 \ln L_{best} + k \ln N$

A real (usual) case: boundaries, marginalization and model selection

Astro: How the assembly of galaxy masses proceed? Evolution of the 3.6 micron LF measures the growth history of galaxy masses.

Stats: determination of the 3.6 micron LF and model selection among various possible mass growth histories.

Data: 1000 member galaxies (plus a 4500 background galaxies, whose distribution is estimated from a larger background formed by 107000 galaxies) from one of Legacy Spitzer surveys (SWIRE)

Two derivations: standard and bayes, both published in Andreon (2006, A&A 448, 447), so I can blame without bless anyone …

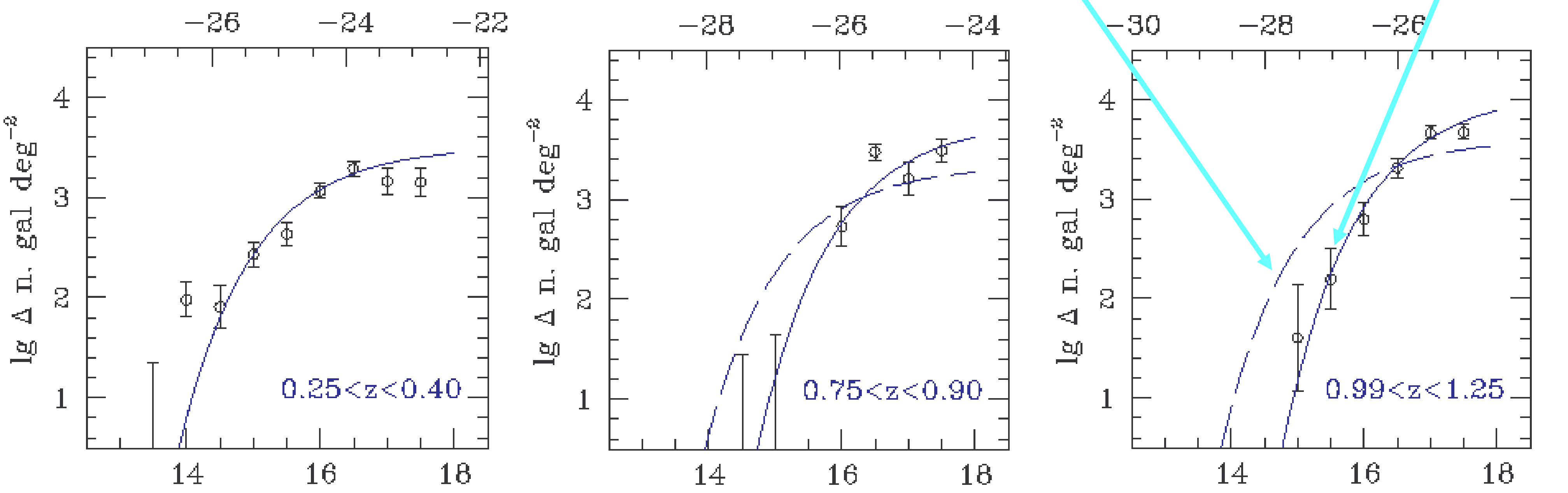
Simplistic (std) analysis.

Step 1: parameter estimation

- Bin in z and mag, don't care if bins are optimally chosen
- Don't care if n(L) is defined to be positive and found negative (positive background fluctuation)
- Assume no evolution, don't worry the risk of a circular raisoning
- take unconstrained parameters (alpha) fixed, i.e. neglect the role of nuisance parameters

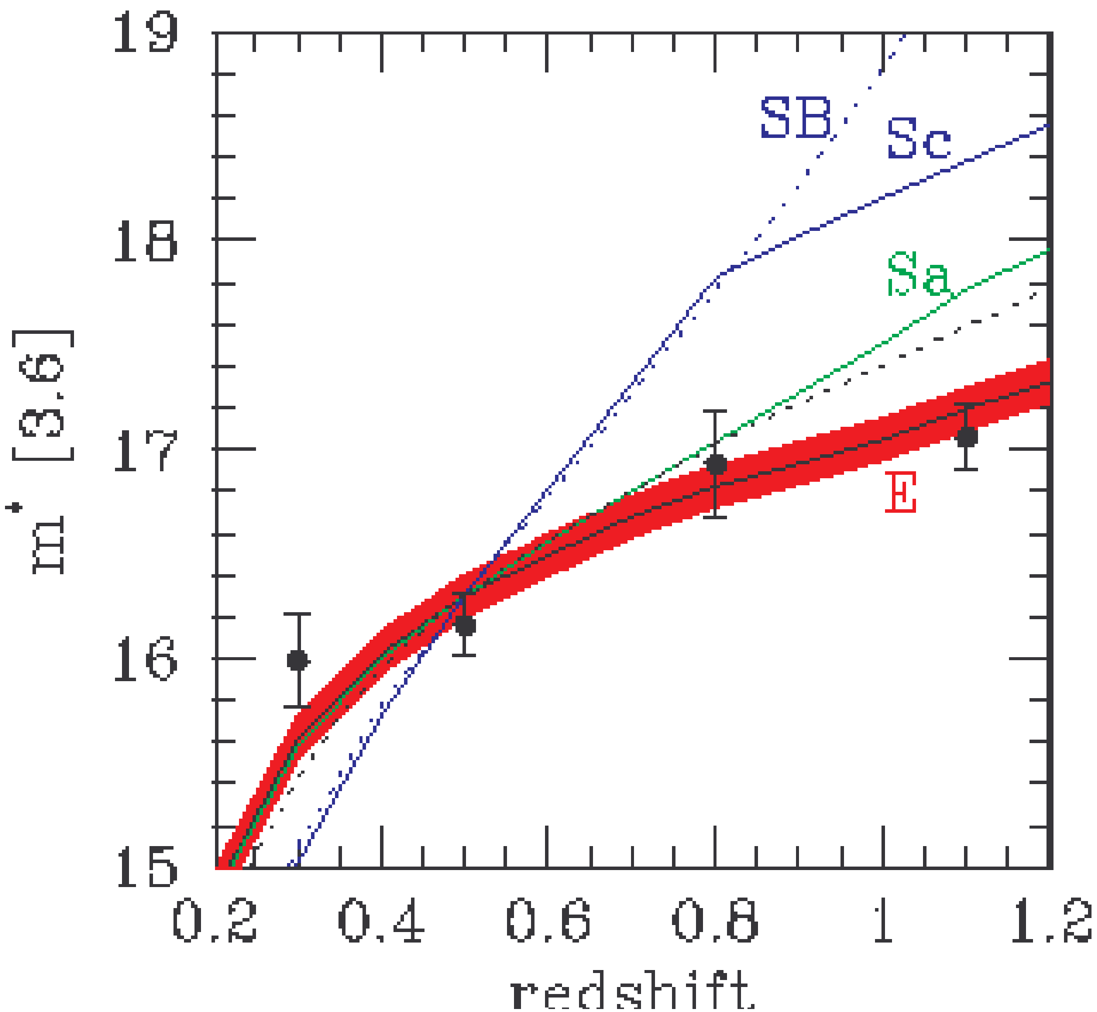
Redshift~0.3 fit

Best fit



Simplistic (std) analysis

Step 2: simplistic model selection. Compare data and models in order to select the best model.



E models are obviously best, but does other model are rejected?

Do the best model need to be refined? (model complexity issue)

At this point of the talk, we known that problems are there. Therefore I stop with the simplistic analysis, and I use bayesian methods.

Bayesian analysis

Don't bin in z and mag, bayesian methods don't require bins,

Account for boundaries: n(Mass)dMass is positively defined.

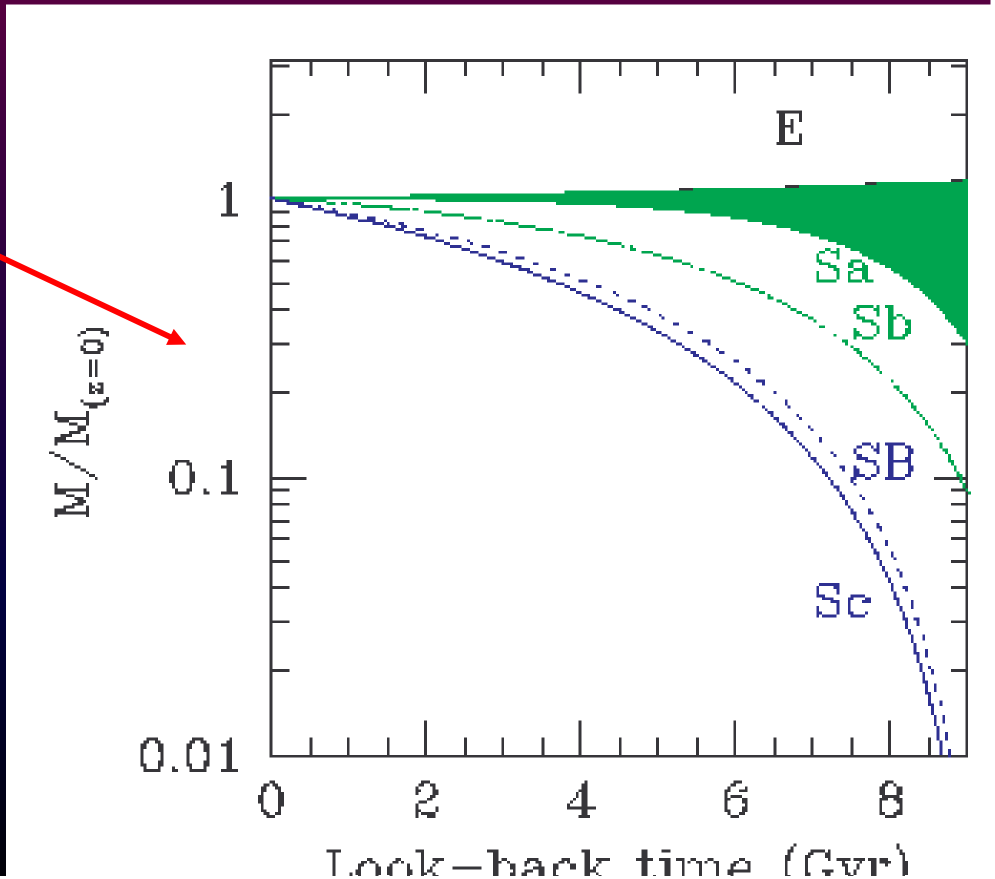
Model evolution   and, eventually

refine the model

Marginalize over nuisance parameters

(alpha).

Mass growth histories, converted in 3.6 micron luminosity evolution by using Grasil models



Bayesian analysis

Step 1: model selection.

Models are not hierarchically nested, likelihood ratio test cannot be used.

Questions to answer:

1) Which model best describes the data? Which models are rejected?
2) Do models need to be refined by a further evolutionary term (taken prop to z)? (model complexity)

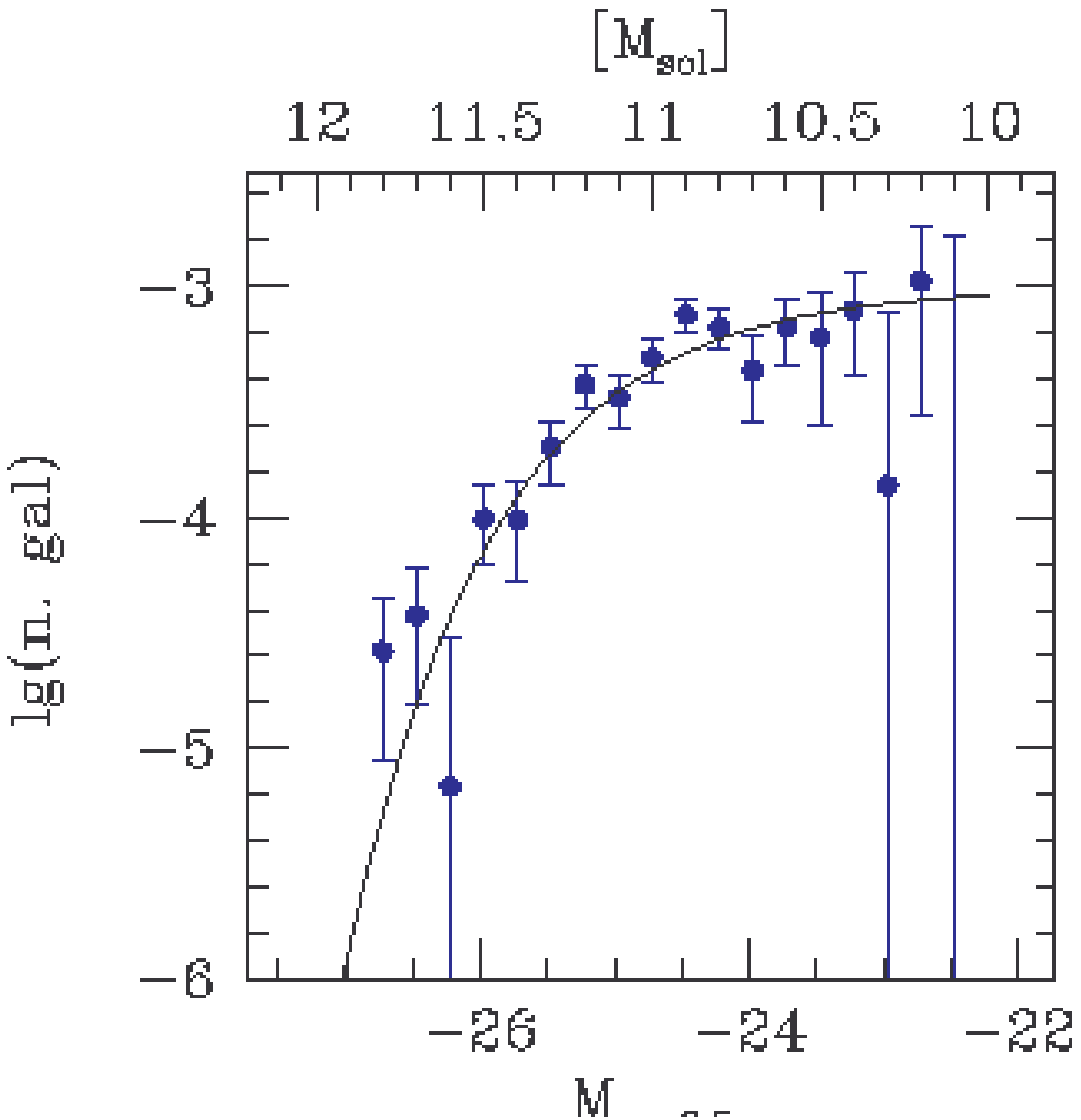
I used the $BIC = -2 \ln L_{best} + k \ln N$

R1: E (no mass growth) models are preferable to all the other ($\Delta BIC > 5$)
R2: no

Step 2: once I have selected the model, I can compute the mass function because …

Mass Function

… I known the mass evolution and I known at which mass I should put galaxies observed to have $mass_i$ at z_i .



Summary

In my field, Bayesian methods are almost unknow.

Imaginary values of velocity dispersions

Possible negative M/L ratios of clusters

Negative star formation rates

Fractions of blue galaxies outside [0-1]

Spectroscopic of photometric completeness larger than 1

Number density profiles of cluster systematically negative

Conclusions in contradictions with hypothesis

Fun results can be avoided by using Bayesian methods.

Thank you