

Introduction to Bayesian Statistics - 8

Edoardo Milotti

Università di Trieste and INFN-Sezione di Trieste

Image likelihood: 2. the noise model (degradation model)

Gaussian noise model

$$P(\mathbf{g}|\mathbf{f}) \propto \exp\left[-\frac{(\mathbf{g} - \mathbf{Hf})^2}{\sigma^2}\right]$$

Poisson noise model

$$P(\mathbf{g}|\mathbf{f}) \propto \prod_n \frac{(\mathbf{Hf})_n^{g_n}}{g_n!} \exp[-(\mathbf{Hf})_n]$$

(Poisson noise mostly from detection process, Gaussian noise mostly from electronics or from approximation of Poisson noise)

sometimes we can use the Gaussian approximation of Poisson noise

$$\begin{aligned} P(\mathbf{g}|\mathbf{f}) &\propto \prod_n \frac{(\mathbf{Hf})_n^{g_n}}{g_n!} \exp[-(\mathbf{Hf})_n] \\ &\approx \prod_n \exp\left[-\frac{(g_n - (\mathbf{Hf})_n)^2}{2(\mathbf{Hf})_n}\right] \\ &= \exp\left[-\sum_n \frac{(g_n - (\mathbf{Hf})_n)^2}{2(\mathbf{Hf})_n}\right] \end{aligned}$$

Gaussian noise only:

maximize linear combination of entropy and chi-square

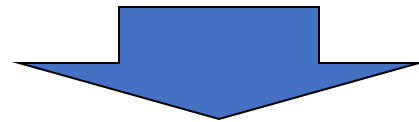
$$\begin{aligned}\ln P(\mathbf{f}|\mathbf{g}) &\approx \alpha S(\mathbf{f}) - \frac{(\mathbf{g} - \mathbf{H}\mathbf{f})^2}{2\sigma^2} \\ &= \alpha S(\mathbf{f}) - \sum_n \frac{(g_n - (\mathbf{H}\mathbf{f})_n)^2}{2\sigma^2} \\ &= \alpha S(\mathbf{f}) - 2\chi^2(\mathbf{f})\end{aligned}$$

Combined noise model

detector noise: Poisson noise

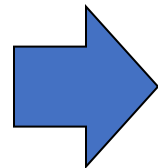
electronic noise: Gaussian noise

$$P(\mathbf{g}|\mathbf{f}) = \prod_n \sum_k \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(g_n - k)^2}{\sigma^2}\right] \frac{(\mathbf{H}\mathbf{f})_n^k}{k!} \exp[-(\mathbf{H}\mathbf{f})_n]$$



maximize

$$\log P(\mathbf{f}|\mathbf{g}) = \alpha S(\mathbf{f}) + \sum_n \log \left\{ \sum_k \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(g_n - k)^2}{\sigma^2}\right] \frac{(\mathbf{H}\mathbf{f})_n^k}{k!} \exp[-(\mathbf{H}\mathbf{f})_n] \right\}$$



numerical maximization procedure

Many related methods: e.g. the Richardson-Lucy (RL) algorithm

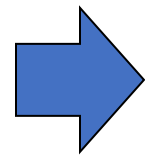
noise model: Poisson noise

prior: flat prior

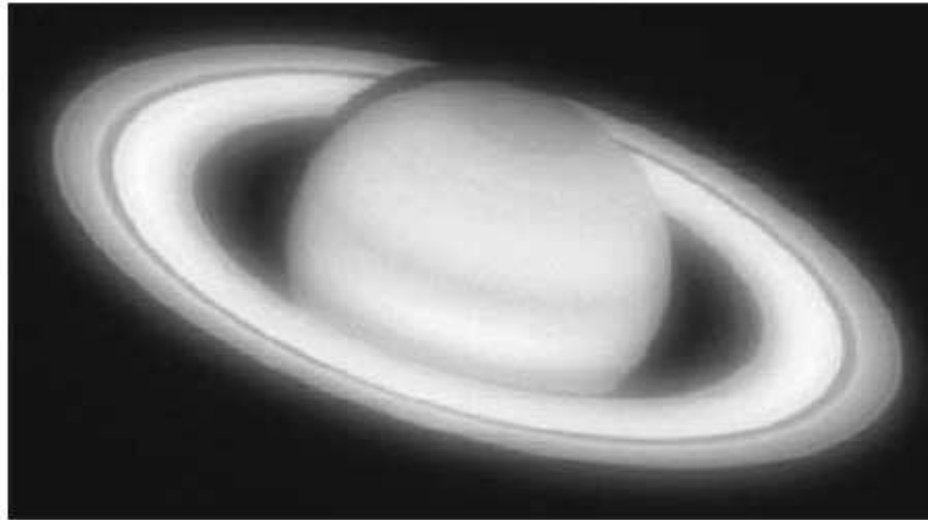
$$P(\mathbf{f}|\mathbf{g}) \propto \prod_n \frac{(\mathbf{H}\mathbf{f})_n^{g_n}}{g_n!} \exp[-(\mathbf{H}\mathbf{f})_n] P(\mathbf{f})$$

$$\log P(\mathbf{f}|\mathbf{g}) \approx \sum_n \left[-(\mathbf{H}\mathbf{f})_n + g_n \log(\mathbf{H}\mathbf{f})_n \right] + \text{const.}$$

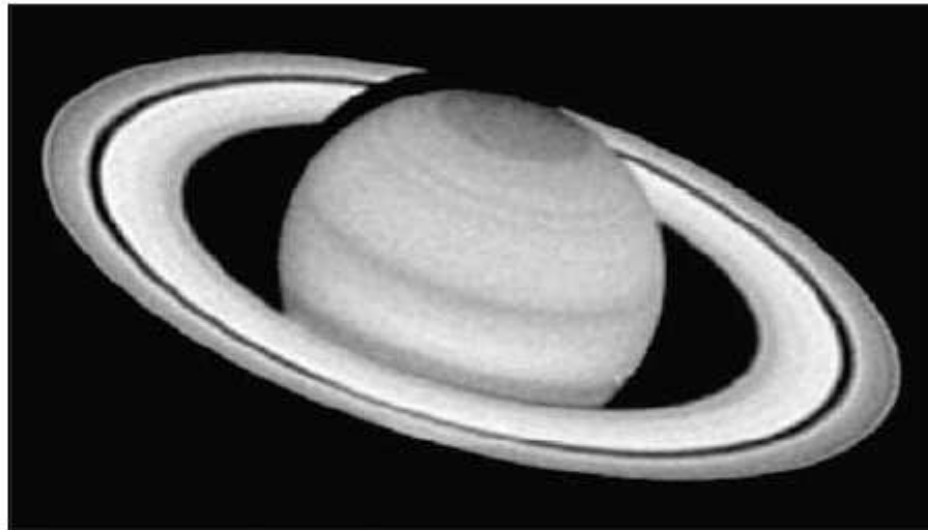
maximize this
posterior distribution



$$\hat{\mathbf{f}} = \arg \max_{\mathbf{f}} \sum_n \left[-(\mathbf{H}\mathbf{f})_n + g_n \log(\mathbf{H}\mathbf{f})_n \right]$$



▲ 8. *Raw image of planet Saturn obtained with the WF/PC camera of the HST.*

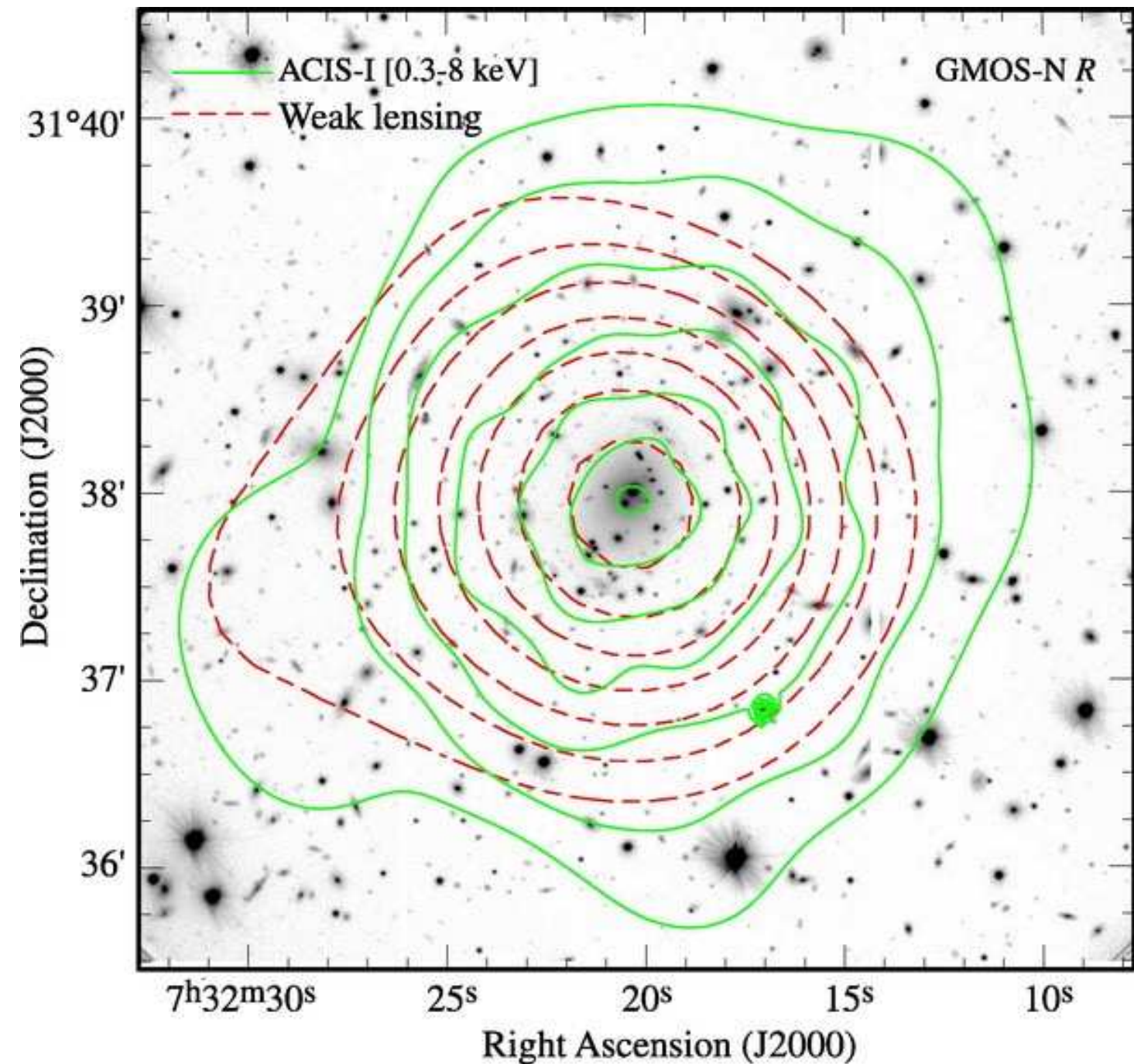


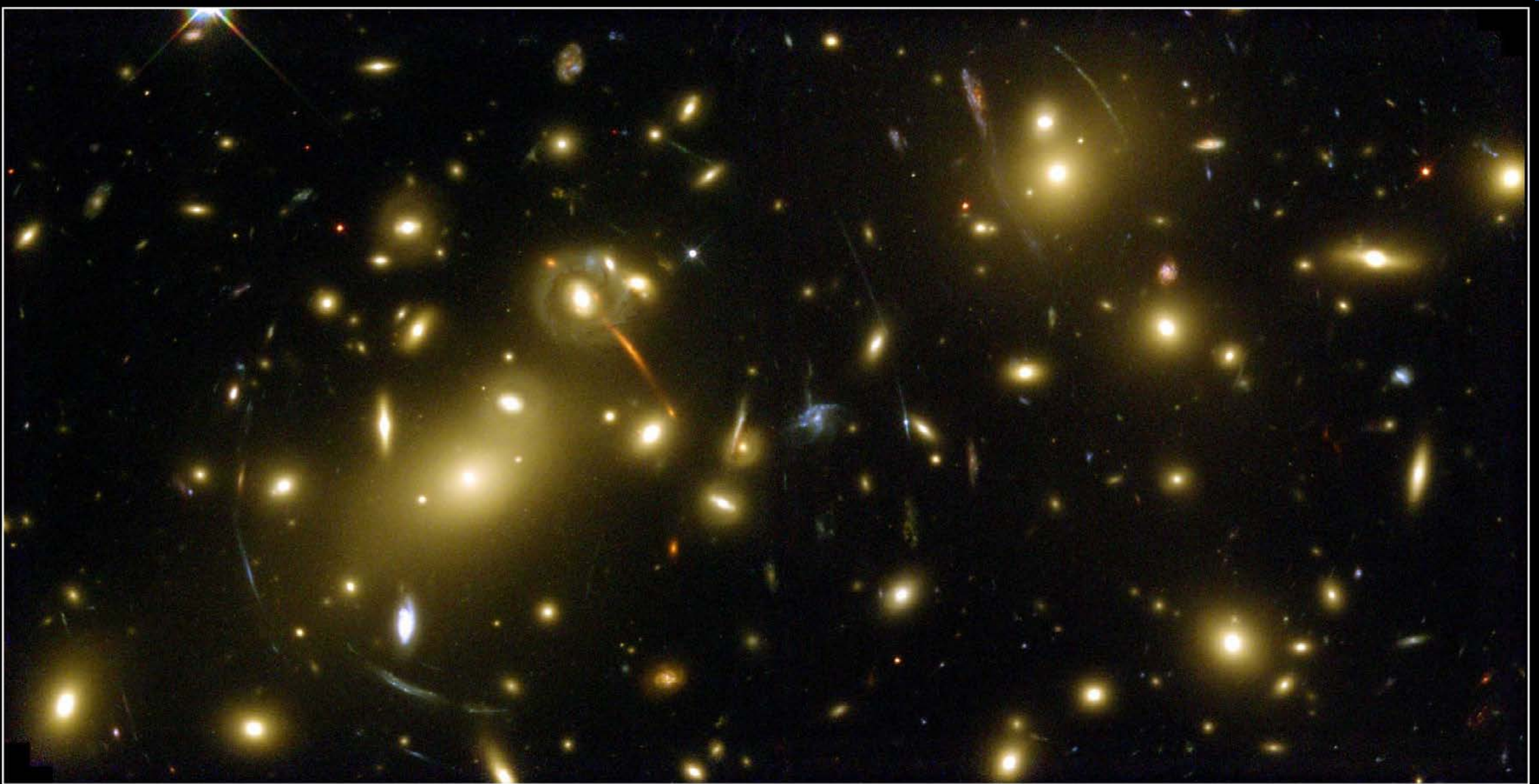
▲ 9. *Reconstruction of the image of Saturn using the R-L algorithm.*

NGC 604 in Spiral Galaxy M33



Example: reconstruction
of the mass distribution
of a gravitational lens

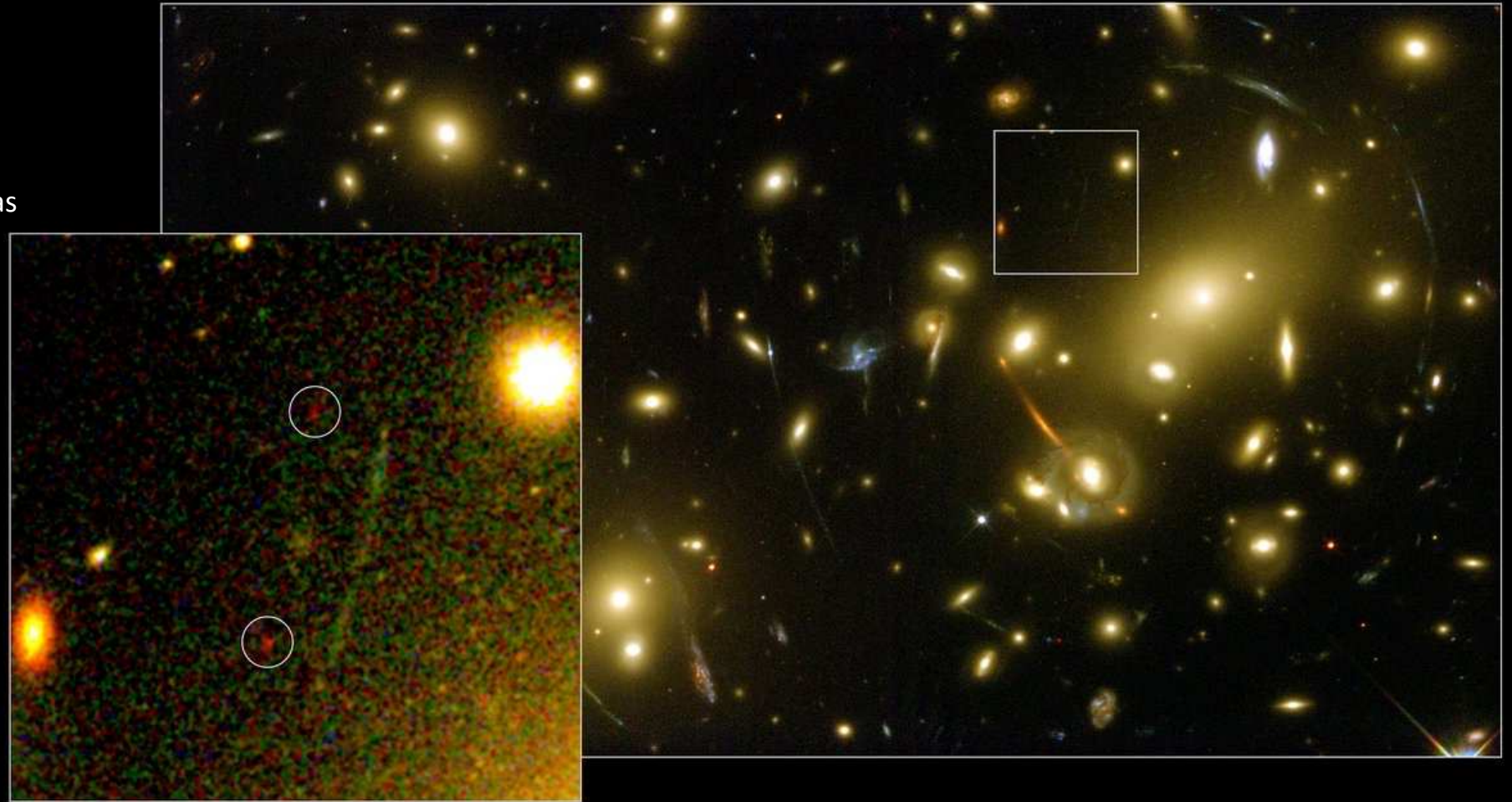




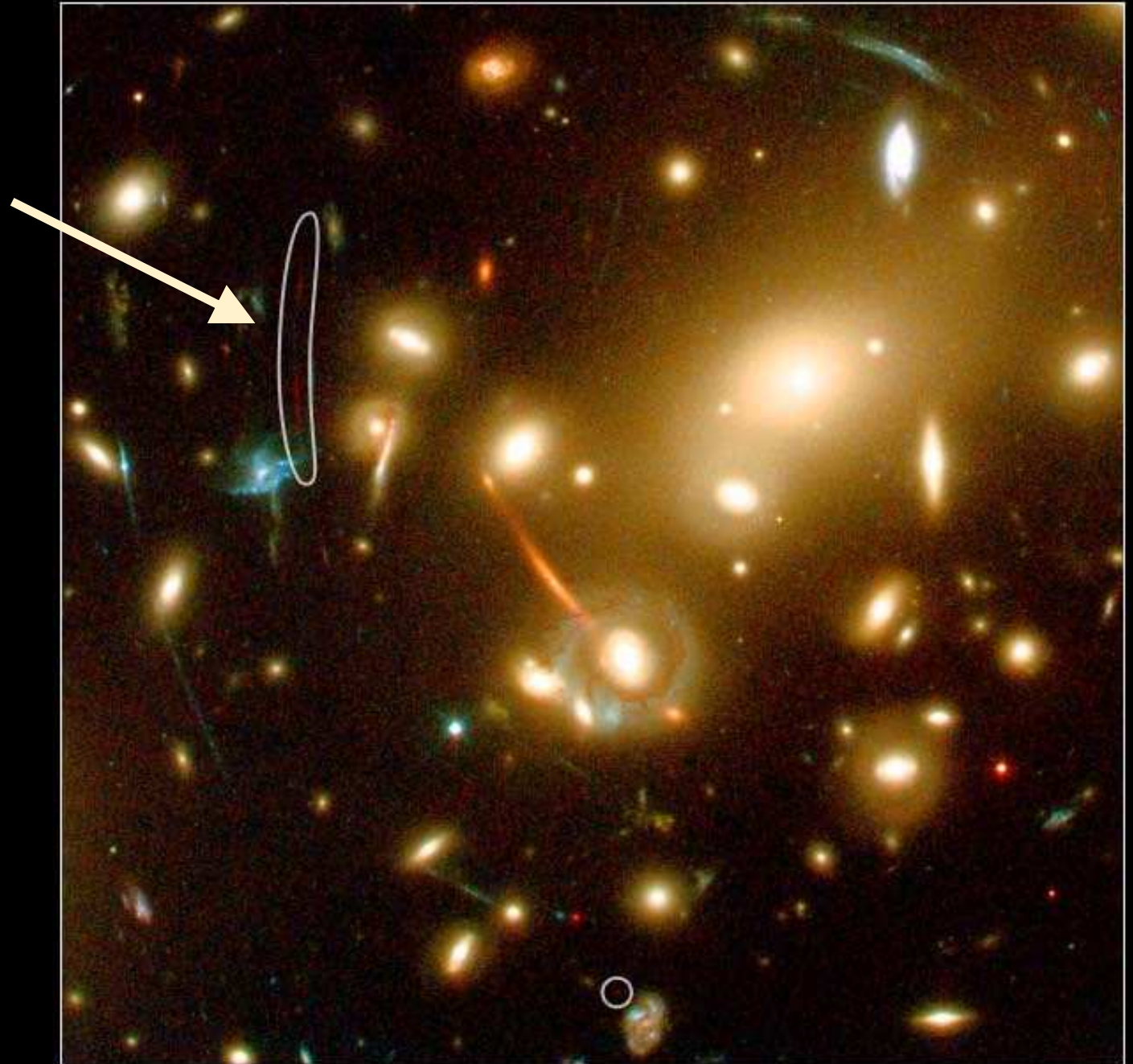
Galaxy Cluster Abell 2218
Hubble Space Telescope • WFPC2

A very small, faint galaxy - possibly one of the long sought 'building blocks' of present-day galaxies - has been discovered by a collaboration between the Hubble Space Telescope and the Keck Telescopes at a tremendous distance of 13.4 billion light-years (based on the estimate of 14 billion years as the age of the Universe). The discovery was made possible by examining small areas of sky viewed through massive intervening clusters of galaxies. These act as a powerful gravitational lens, magnifying distant objects and allowing scientists to probe how galaxies assemble at very early times. This has profound implications for our understanding of how and when the first stars and galaxies formed in the Universe.

(from:
<https://esahubble.org/news/heic0113/>)



The highlighted orange arc in this image taken by the Hubble Space Telescope's Advanced Camera for Surveys represent the stretched image of a galaxy some 13 billion light-years away from Earth. Analysis of spectra taken by Hubble and Keck determined that the galaxy has a redshift of about 7, meaning that the star system is seen as it was only 750 million years after the Big Bang. Light from the galaxy has been distorted by the gravitational-lensing effects of the intervening galaxy cluster Abell 2218.

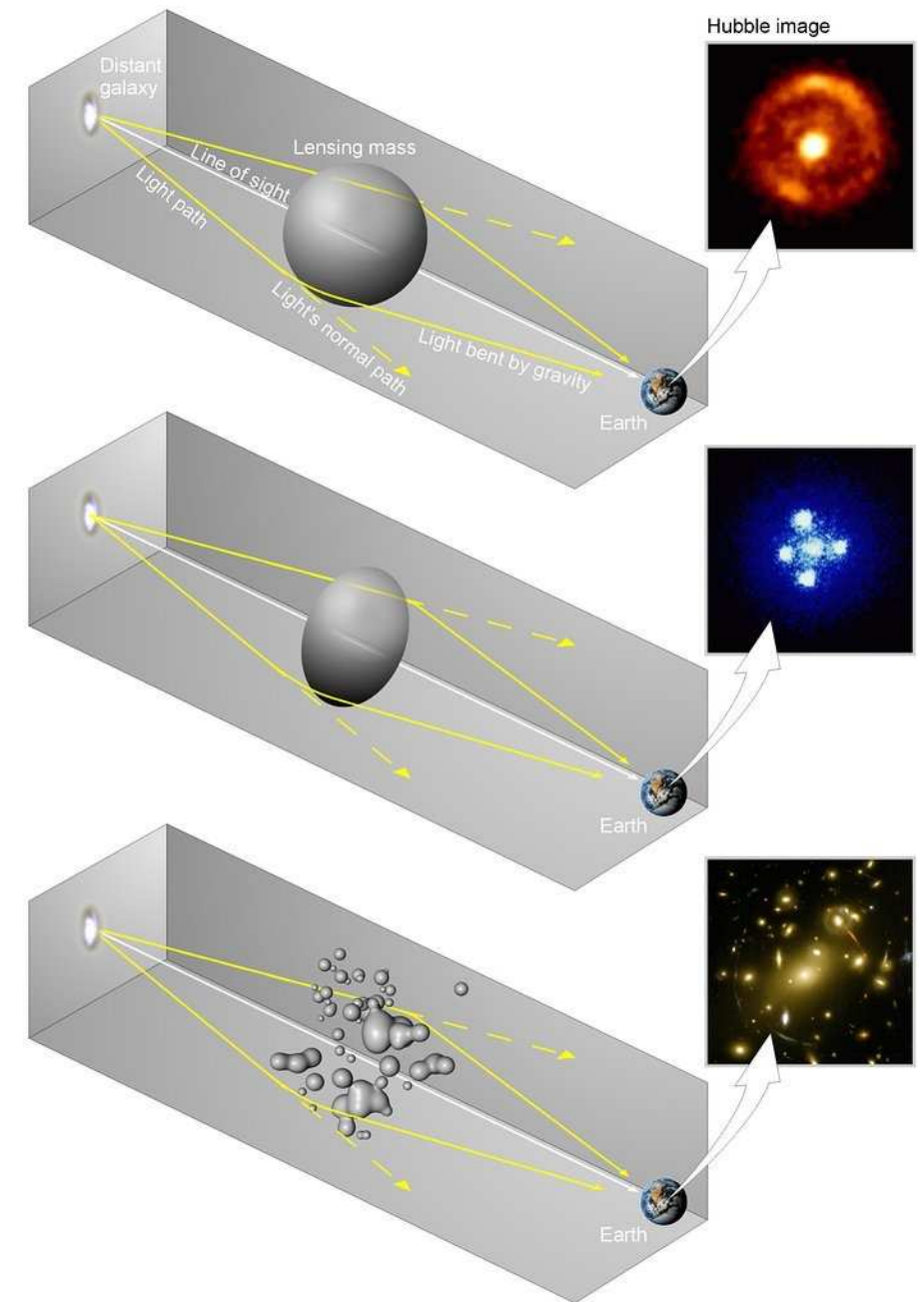


Gravitational lenses:

- Deflection angle due to a point-like mass

$$\alpha = \frac{4GM}{c^2 r}$$

- A weak gravitational lens is, in its essentials, equivalent to a flat space deflector with refractive index $1 - 2\phi/c^2$, where ϕ is the 3D Newtonian potential with respect to infinity. (Blandford & Narayan, *Cosmological Applications of Gravitational Lensing*, Annu. Rev. Astron. Astrophys. 1992, 30:311–358)



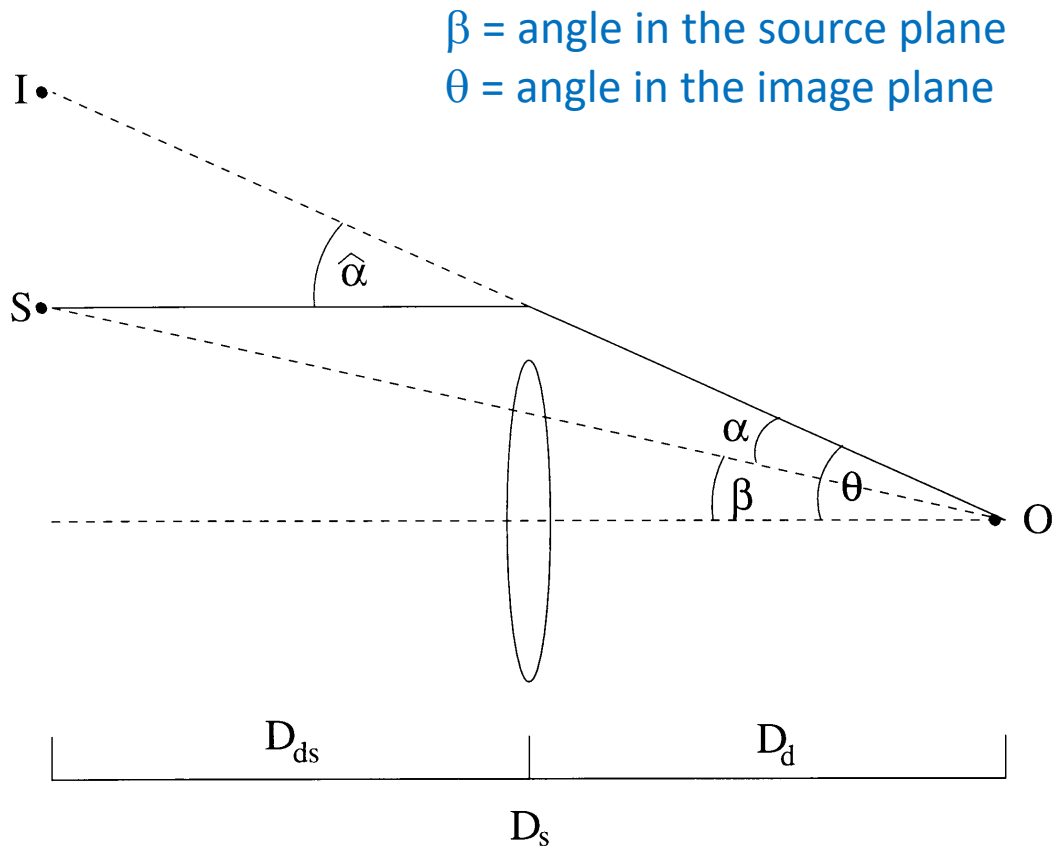


Figure 1. The gravitational lens geometry. The light ray propagates from the source S to the observer O and is deflected through an angle $\hat{\alpha}$ so that the image appears at I. The angular separations of the source and image from the optic axis are denoted by β and θ respectively. D_d , D_s and D_{ds} are respectively the angular diameter distances from the observer to the lens, the observer to the source and the source to the lens.

Vectors of angular distances in the source and image transform as follows:

$$d\boldsymbol{\beta} = \mathbf{A}d\boldsymbol{\theta}$$

where the tensor \mathbf{A} is the inverse of the magnification tensor and depends on the angle in the image plane

$$\mathbf{A}(\boldsymbol{\theta}) = \begin{pmatrix} 1 - \kappa - \gamma_1 & -\gamma_2 \\ -\gamma_2 & 1 - \kappa + \gamma_1 \end{pmatrix}$$

This tensor depends on two θ -dependent variables: κ which is called the *convergence*, and γ_1, γ_2 which are called the *shears*.

The convergence is directly related to the mass distribution in the source plane

$$\kappa(\boldsymbol{\theta}) = \frac{\Sigma(\boldsymbol{\theta})}{\Sigma_{\text{crit}}}; \quad \Sigma_{\text{crit}} = \frac{c^2}{4\pi G} \frac{D_s}{D_d D_{ds}}$$

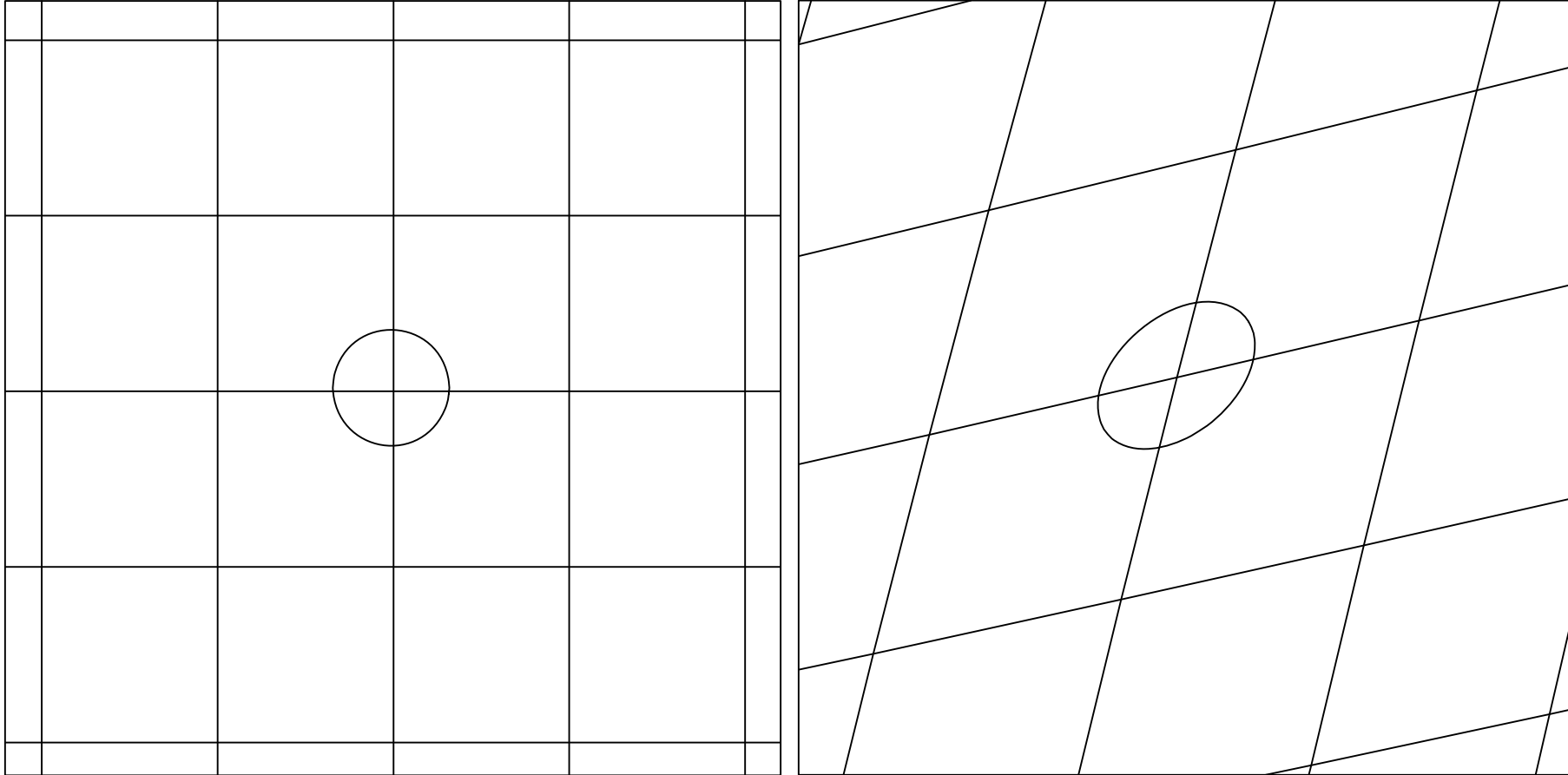
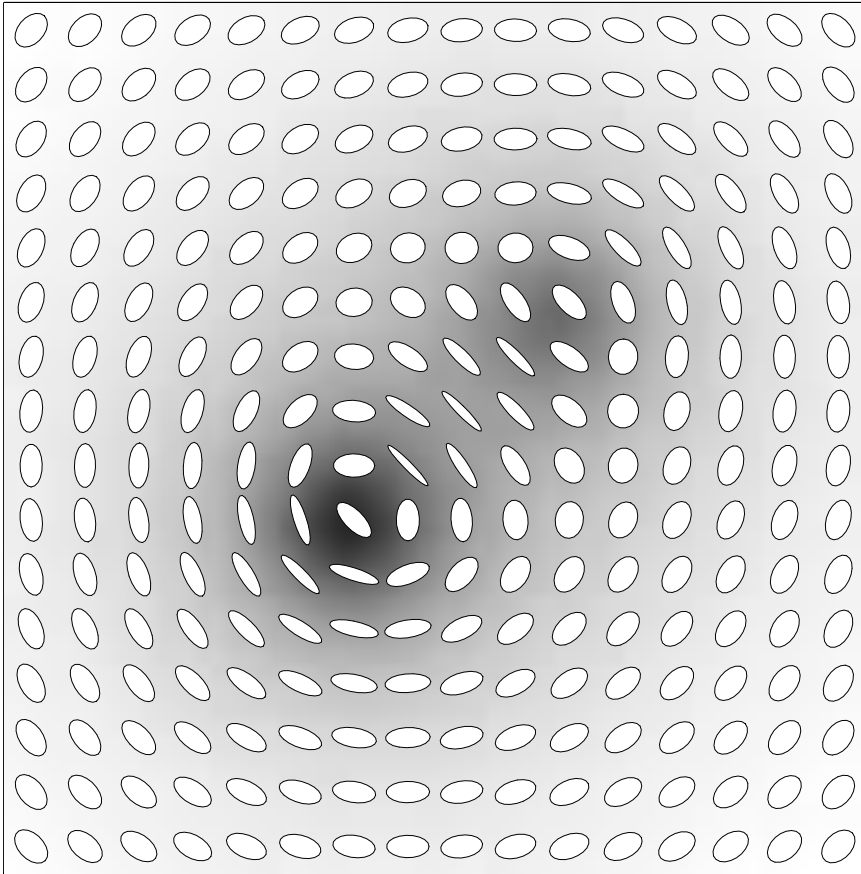


Fig. 6 Illustration of the effect of lensing: local deformation of a regular grid and a circle (*left*: source map) by a lens with constant value of the convergence κ and the shear γ over the region (*right*: image map)

Distortion field overlay



- Starting from an image, it is possible to recover magnification and (complex) ellipticity values (averaged over groups of pixels)
- These experimental values are used to reconstruct the convergence by means of a MEM approach
- The convergence distribution is simply proportional to the mass distribution

$$\kappa(\boldsymbol{\theta}) = \frac{\Sigma(\boldsymbol{\theta})}{\Sigma_{\text{crit}}}; \quad \Sigma_{\text{crit}} = \frac{c^2}{4\pi G} \frac{D_s}{D_d D_{ds}}$$

GEMINI AND *CHANDRA* OBSERVATIONS OF ABELL 586, A RELAXED STRONG-LENSING CLUSTER

E. S. CYPRIANO,^{1,2} G. B. LIMA NETO,³ L. SODRÉ, JR.,³ J.-P. KNEIB,^{4,5} AND L. E. CAMPUSANO⁶

Received 2004 August 16; accepted 2005 April 1

ABSTRACT

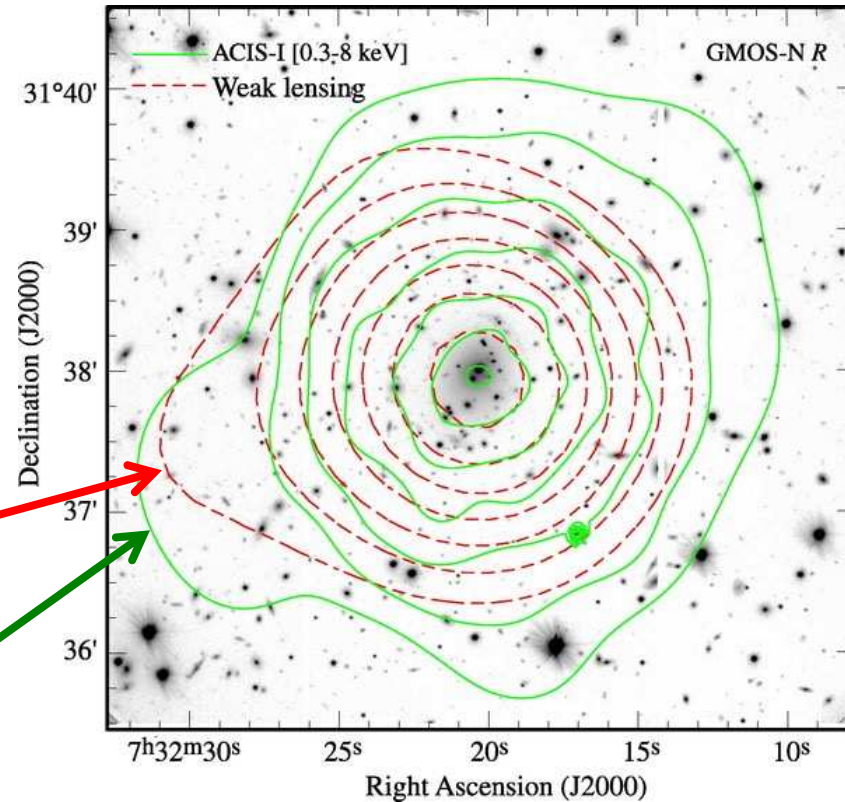
We analyze the mass content of the massive strong-lensing cluster Abell 586 ($z = 0.17$). We use optical data (imaging and spectroscopy) obtained with the Gemini Multi-Object Spectrograph (GMOS) mounted on the 8 m Gemini North telescope, together with publicly available X-ray data taken with the *Chandra* space telescope. Employing different techniques—velocity distribution of galaxies, weak gravitational lensing, and spatially resolved X-ray spectroscopy—we derive mass and velocity dispersion estimates from each of them. All estimates agree well with each other, within a 68% confidence level, indicating a velocity dispersion of $1000\text{--}1250\text{ km s}^{-1}$. The projected mass distributions obtained through weak lensing and X-ray emission are strikingly similar, having nearly circular geometry. We suggest that Abell 586 is probably a truly relaxed cluster whose last major merger occurred more than ~ 4 Gyr ago.

Example of LensEnt usage
(Bridle et al, 1998)

Reconstruction of mass
density from lensing data,
using Max Ent

reconstructed mass
density

X-ray emission data



Gemini Multi-Object Spectrograph image of the central region of Abell 586 with **logarithmically spaced X-ray isophotes (solid lines) and weak-lensing reconstructed mass density (dashed lines) superposed**. The X-ray point source near the southwest corner is the Seyfert 1 galaxy C171_3650.
(from Cypriano et al., ApJ, **630** (2005) 38-49)

Bayesian classification

data X , classes C

$$P(C|X) = \frac{P(X|C)}{P(X)} P(C)$$

Diagram illustrating the Bayesian classification formula:

- The term $P(X|C)$ is labeled "this likelihood is defined by training data" with a red arrow pointing to it.
- The term $P(C)$ is labeled "the prior is also defined by training data" with a red arrow pointing to it.

we can use the prior learning to assign a class to new data

$$C_k = \arg \max_{C_k} \frac{P(X|C_k)}{P(X)} P(C_k) = \arg \max_{C_k} P(X|C_k) P(C_k)$$

Consider a vector of N attributes given as Boolean variables $\mathbf{x} = \{x_i\}$ and classify the data vectors with a single Boolean variable.

The learning procedure must yield:

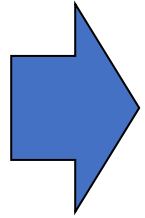
$P(y)$

it is easy to obtain it as an empirical distribution from an histogram of training class data: y is Boolean, the histogram has just two bins, and a hundred examples suffice to determine the empirical distribution to better than 10%.

$P(\mathbf{x}|y)$

there is a bigger problem here: the arguments have 2^{N+1} different values, and we must estimate $2(2^N-1)$ parameters ... for instance, with $N = 30$ there are more than 2 billion parameters!

How can we reduce the huge complexity of learning?



we assume the conditional independence of the x_n 's:
naive Bayesian learning

for instance, with just two attributes

$$P(x_1, x_2 | y) = P(x_1 | x_2, y) P(x_2 | y) = P(x_1 | y) P(x_2 | y)$$



conditional independence assumption

with more than 2 attributes

$$P(\mathbf{x} | y) \approx \prod_{k=1}^N P(x_k | y)$$

Therefore:

$$\begin{aligned} P(y_k|\mathbf{x}) &= \frac{P(\mathbf{x}|y_k)}{P(\mathbf{x})} P(y_k) = \frac{P(\mathbf{x}|y_k)}{\sum_j P(\mathbf{x}|y_j) P(y_j)} P(y_k) \\ &\approx \frac{\prod_{n=1}^N P(x_n|y_k)}{\sum_j P(y_j) \prod_{n=1}^N P(x_n|y_j)} P(y_k) \end{aligned}$$

and we assign the class according to the rule (MAP)

$$y = \arg \max_{y_k} \frac{\prod_{n=1}^N P(x_n|y_k)}{\sum_j P(y_j) \prod_{n=1}^N P(x_n|y_j)} P(y_k)$$

More general discrete inputs

If any of the N variables has J different values, and if there are K classes, then we must estimate in all $NK(J-1)$ free parameters with the Naive Bayes Classifier (this includes normalization) (compare this with the $K(J^N-1)$ parameters needed by a complete classifier)

Continuous inputs and discrete classes – the Gaussian case

$$P(x_n | y_k) = \frac{1}{\sqrt{2\pi\sigma_{nk}^2}} \exp\left[-\frac{(x_n - \mu_{nk})^2}{2\sigma_{nk}^2}\right]$$

here we must estimate $2NK$ parameters + the shape of the distribution $P(y)$ (this adds up to another $K-1$ parameters)

Gaussian special case with class-independent variance and Boolean classification (two classes only):

$$P(y = 0 | \mathbf{x}) = \frac{P(\mathbf{x} | y = 0)P(y = 0)}{P(\mathbf{x} | y = 0)P(y = 0) + P(\mathbf{x} | y = 1)P(y = 1)}$$

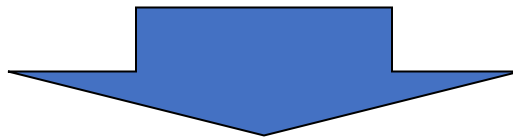
$$P(x_n | y = 0) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left[-\frac{(x_n - \mu_{n0})^2}{2\sigma_n^2}\right]$$

$$P(x_n | y = 1) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left[-\frac{(x_n - \mu_{n1})^2}{2\sigma_n^2}\right]$$

$$\begin{aligned}
P(y=0|\mathbf{x}) &= \frac{P(\mathbf{x}|y=0)P(y=0)}{P(\mathbf{x}|y=0)P(y=0) + P(\mathbf{x}|y=1)P(y=1)} \\
&= \frac{1}{1 + \frac{P(\mathbf{x}|y=1)P(y=1)}{P(\mathbf{x}|y=0)P(y=0)}} \\
&= \frac{1}{1 + \frac{P(y=1)}{P(y=0)} \prod_{n=1}^N \exp \left[-\frac{(x_n - \mu_{n1})^2}{2\sigma_n^2} + \frac{(x_n - \mu_{n0})^2}{2\sigma_n^2} \right]} \\
&= \frac{1}{1 + \exp \left\{ \ln \left(\frac{P(y=1)}{P(y=0)} \right) + \sum_{n=1}^N \left[\frac{(\mu_{n1} - \mu_{n0})x_n}{\sigma_n^2} + \frac{\mu_{n0}^2 - \mu_{n1}^2}{2\sigma_n^2} \right] \right\}}
\end{aligned}$$

$$w_0 = \ln \left(\frac{P(y=1)}{P(y=0)} \right) + \sum_{n=1}^N \left[\frac{\mu_{n0}^2 - \mu_{n1}^2}{2\sigma_n^2} \right]$$

$$w_n = \frac{(\mu_{n1} - \mu_{n0})}{\sigma_n^2}$$



logistic shape

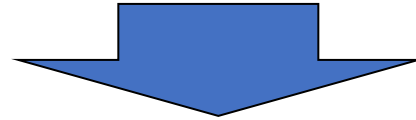
$$P(y=0|\mathbf{x}) = \frac{1}{1 + \exp \left(w_0 + \sum_{n=1}^N w_n x_n \right)}$$



$$P(y=1|\mathbf{x}) = 1 - P(y=0|\mathbf{x}) = \frac{\exp \left(w_0 + \sum_{n=1}^N w_n x_n \right)}{1 + \exp \left(w_0 + \sum_{n=1}^N w_n x_n \right)}$$

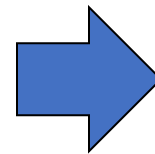
Finally an input vector belongs to class $y = 0$ if

$$\frac{P(y = 0|\mathbf{x})}{P(y = 1|\mathbf{x})} > 1$$

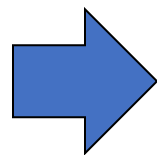


$$P(y = 0|\mathbf{x}) = \frac{1}{1 + \exp\left(w_0 + \sum_{n=1}^N w_n x_n\right)}$$

$$P(y = 1|\mathbf{x}) = \frac{\exp\left(w_0 + \sum_{n=1}^N w_n x_n\right)}{1 + \exp\left(w_0 + \sum_{n=1}^N w_n x_n\right)}$$



$$\exp\left(w_0 + \sum_{n=1}^N w_n x_n\right) < 1$$



$$w_0 + \sum_{n=1}^N w_n x_n < 0$$

Logistic regression (logit regression)

The odds ratio

$$\frac{P(y = 1|\mathbf{x})}{P(y = 0|\mathbf{x})} = \frac{p}{1 - p}$$

with the exponential expansion that we just found (the logistic expression for p) gives

$$\ln \frac{p}{1 - p} = w_0 + \sum_{n=1}^N w_n x_n$$

For a given set of K class determinations where the fraction of assignments to class 1 is $p^{(k)}$ for a parameter vector of $\{x_n^{(k)}\}_{k=1,K}$ this log odds ratio becomes

$$\ln \frac{p^{(k)}}{1 - p^{(k)}} = w_0 + \sum_{n=1}^N w_n x_n^{(k)}$$

The expression

$$\ln \frac{p^{(k)}}{1 - p^{(k)}} = w_0 + \sum_{n=1}^N w_n x_n^{(k)}$$

is the basis for a generalized linear regression, to determine the w parameters.

This can be done with the least squares method, where one minimizes

$$S = \sum_{k=1}^K \left[\ln \frac{p^{(k)}}{1 - p^{(k)}} - \left(w_0 + \sum_{n=1}^N w_n x_n^{(k)} \right) \right]^2$$

This logit regression is often used in classification problems.

The Gibbs sampler

(from Casella and George, *Explaining the Gibbs sampler* Am.Stat. 46 (1992) 167)

Let's start with an example, and consider the following joint distribution:

$$f(x, y) \propto \binom{n}{x} y^{x+\alpha-1} (1-y)^{n-x+\beta-1}, \quad x = 0, \dots, n \quad 0 \leq y \leq 1$$

We see that

$$f(x|y) \sim \text{Binomial}(n, y)$$

$$f(y|x) \sim \text{Beta}(x + \alpha, n - x + \beta)$$

Next we set up a simple Markov chain procedure ...

We generate a "Gibbs sequence" of random variables

$$Y'_0, X'_0, Y'_1, X'_1, Y'_2, X'_2, \dots, Y'_k, X'_k$$

where the initial value is specified and the others are computed with the rule

$$X'_j \sim f(x \mid Y'_j = y'_j)$$

$$Y'_{j+1} \sim f(y \mid X'_j = x'_j)$$

(Gibbs sampling).

We observe that for large enough k , the final X values have a fixed distribution that corresponds to the marginal pdf of the x variate.

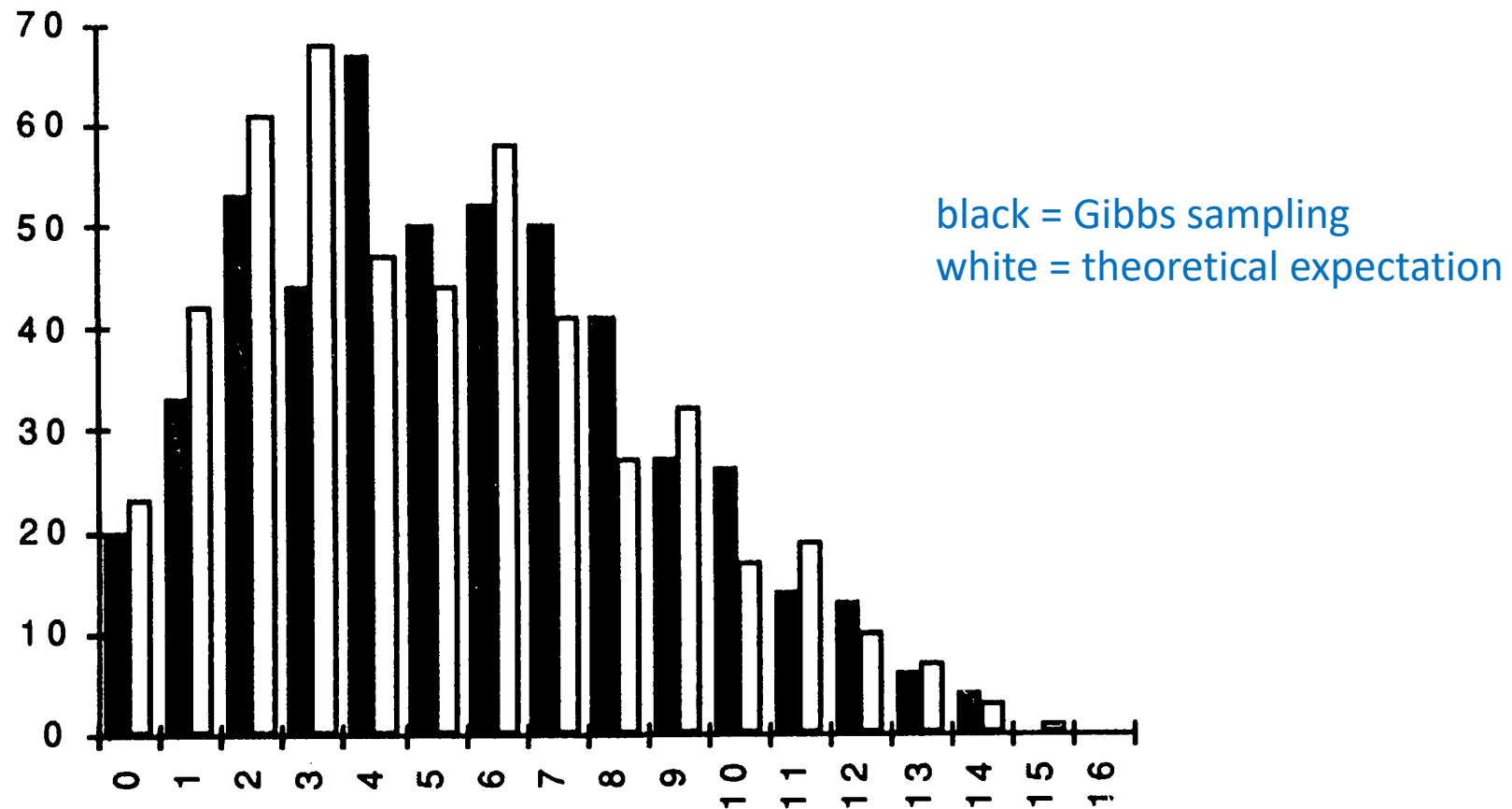


Figure 1. Comparison of Two Histograms of Samples of Size $m = 500$ From the Beta-Binomial Distribution With $n = 16$, $\alpha = 2$, and $\beta = 4$. The black histogram sample was obtained using Gibbs sampling with $k = 10$. The white histogram sample was generated directly from the beta-binomial distribution.

Should we expect this result?

Consider the following expectation value

$$E_y[f(x|y)] = \int_Y f(x|y)f(y)dy = \int_Y f(x,y)dy = f(x)$$

therefore we can estimate $f(x)$ with the sum

$$\hat{f}(x) = \frac{1}{m} \sum_{i=1}^m f(x | y_i)$$

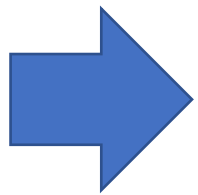
and finally the [Gibbs sampling provides representative samples that correspond to this marginal distribution](#). (for a better proof, check the paper by Casella&George)

Does Gibbs sampling converge?

We consider the following case: two discrete random variables with marginally Bernoulli distributions and with a joint probability distribution described by this matrix

		X	
		0	1
Y	0	p_1	p_2
	1	p_3	p_4

$p_i \geq 0, p_1 + p_2 + p_3 + p_4 = 1$



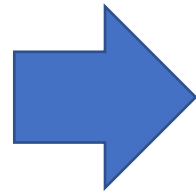
$$\begin{bmatrix} f_{x,y}(0,0) & f_{x,y}(1,0) \\ f_{x,y}(0,1) & f_{x,y}(1,1) \end{bmatrix} = \begin{bmatrix} p_1 & p_2 \\ p_3 & p_4 \end{bmatrix}$$

$$\begin{bmatrix} f_{x,y}(0,0) & f_{x,y}(1,0) \\ f_{x,y}(0,1) & f_{x,y}(1,1) \end{bmatrix} = \begin{bmatrix} p_1 & p_2 \\ p_3 & p_4 \end{bmatrix}$$



$$f_x = [f_x(0) \quad f_x(1)] = [p_1 + p_3 \quad p_2 + p_4]$$

marginal
distribution



$$A_{y|x} = \begin{bmatrix} \frac{p_1}{p_1 + p_3} & \frac{p_2}{p_1 + p_2} \\ \frac{p_3}{p_3 + p_4} & \frac{p_4}{p_3 + p_4} \end{bmatrix}$$

$$A_{x|y} = \begin{bmatrix} \frac{p_1}{p_1 + p_2} & \frac{p_2}{p_1 + p_2} \\ \frac{p_3}{p_3 + p_4} & \frac{p_4}{p_3 + p_4} \end{bmatrix}$$

transition
probabilities

Since we are only interested in the X sequence

$$P(X'_1 = x_1 \mid X'_0 = x_0) = \sum_y P(X'_1 = x_1 \mid Y'_1 = y) \\ \times P(Y'_1 = y \mid X'_0 = x_0).$$



the transition matrix for the X sequence is

$$A_{x|x} = A_{y|x} A_{x|y}$$

From the theory of Markov chains, we know that iterating this produces a fixed probability distribution, i.e., our marginal distribution for X.

So, what's the use of all this?

Consider the case where we want to compute the marginal pdf

$$f(x) = \int \dots \int f(x, y_1, \dots, y_p) dy_1 \dots dy_p$$

in a situation where the multidimensional integral can be hard to compute.

The Gibbs sampler completely bypasses the calculation of the multidimensional integral and affords an easy path to marginalization.

The procedure can be easily extended to multidimensional distributions, for example with two nuisance variables we produce the sequence

$$Y'_0, Z'_0, X'_0, Y'_1, Z'_1, X'_1, Y'_2, Z'_2, X'_2, \dots$$

by means of the conditional PDFs

Model selection

The generic purpose of a model selection statistic is to set up a tension between the predictiveness of a model (for instance indicated by the number of free parameters) and its ability to fit observational data. Oversimplistic models offering a poor fit should of course be thrown out, but so should more complex models that offer poor predictive power.

There are two main types of model selection statistic that have been used in the literature so far. Information criteria look at the best-fitting parameter values and attach a penalty for the number of parameters; they are essentially a technical formulation of “chi-squared per degrees of freedom” arguments. By contrast, the Bayesian evidence applies the same type of likelihood analysis familiar from parameter estimation, but at the level of models rather than parameters. It depends on goodness of fit across the entire model parameter space.

(Liddle & al., 2006)

Akaike Information Criterion (AIC).

This was derived by Hirotugu Akaike in 1974, and takes the form

$$\text{AIC} = -2 \ln \mathcal{L}_{\max} + 2k$$

*where k is the number of parameters in the model. The subscript “max” indicates that one should find the parameter values yielding the highest possible likelihood within the model. **This second term acts as a kind of “Occam factor”; initially, as parameters are added, the fit to data improves rapidly until a reasonable fit is achieved, but further parameters then add little and the penalty term $2k$ takes over.** The generic shape of the AIC as a function of number of parameters is a rapid fall, a minimum, and then a rise. The preferred model sits at the minimum.*

The AIC was derived from information-theoretic considerations, specifically an approximate minimization of the Kullback–Leibler information entropy which measures the distance between two probability distributions.

(Liddle & al., 2006)

Bayesian Information Criterion (BIC).

This was derived by Gideon Schwarz in 1978, and strongly resembles the AIC. It is given by

$$\text{BIC} = -2 \ln \mathcal{L}_{\max} + k \ln N$$

where N is the number of datapoints. Since a typical dataset will have $\ln N > 2$, the BIC imposes a stricter penalty against extra parameters than the AIC.

It was derived as an approximation to the Bayesian evidence, to be discussed next, but the assumptions required are very restrictive and unlikely to hold in practice, rendering the approximation quite crude.

(Liddle & al., 2006)

Bayesian evidence

Model selection aims to determine which theoretical models are most plausible given some data, without necessarily considering preferred values of model parameters.

Ideally, we would like to estimate posterior probabilities on the set of all competing models using Bayes' theorem:

$$P(M_i|D, I) = \frac{P(D|M_i, I)P(M_i|I)}{\sum_k P(D|M_k, I)P(M_k|I)}$$

and select the best model using the odds ratio

$$\mathcal{O}_{i,j} = \frac{P(M_i|D, I)}{P(M_j|D, I)} = \frac{P(D|M_i, I)P(M_i|I)}{P(D|M_j, I)P(M_j|I)}$$

or the Bayes factor, if we assume equal prior probabilities for the different models:

$$B_{i,j} = \frac{P(D|M_i, I)}{P(D|M_j, I)}$$

Thus we see that the Bayes factor is a ratio of evidences

$$B_{i,j} = \frac{P(D|M_i, I)}{P(D|M_j, I)}$$

As usual, each evidence is obtained by marginalizing the likelihood with respect to the (potentially different) parameters:

$$P(D|M_i, I) = \int_{\Theta_i} P(D|\boldsymbol{\theta}_i, M_i, I) p(\boldsymbol{\theta}_i|M_i, I) d\boldsymbol{\theta}_i$$

The evidence of a model is thus the average likelihood of the model in the prior.

Unlike the AIC and BIC, it does not focus on the best-fitting parameters of the model, but asks “of all the parameter values you thought were viable before the data came along, how well on average did they fit the data?”. Literally, it is the likelihood of the model given the data.

The evidence rewards predictability of models, provided they give a good fit to the data, and hence gives an axiomatic realization of Occam's razor.

A model with little parameter freedom is likely to fit data over much of its parameter space, whereas a model that could match pretty much any data that might have cropped up will give a better fit to the actual data but only in a small region of its larger parameter space, pulling the average likelihood down.

(Liddle & al., 2006)

Which statistics?

Of these statistics, we would advocate using – wherever possible – the Bayesian evidence, which is a full implementation of Bayesian inference and can be directly interpreted in terms of model probabilities. It is computationally challenging to compute, being a highly peaked multidimensional integral, but recent algorithm development has made it feasible in cosmological contexts.

If the Bayesian evidence cannot be computed, the BIC can be deployed as a substitute. It is much simpler to compute as one need only find the point of maximum likelihood for each model. However, interpreting it can be difficult. Its main usefulness is as an approximation to the evidence, but this holds only for gaussian likelihoods and provided the datapoints are independent and identically distributed. The latter condition holds poorly for the current global cosmological dataset, though it can potentially be improved by binning of the data, hence decreasing the N in the penalty term.

*The AIC has been widely used outside astrophysics, but is of debatable utility. **It has been shown to be “dimensionally inconsistent”, meaning that it is not guaranteed to give the right result even in the limit of infinite unbiased data.*** It may be useful for checking the robustness of conclusions drawn using the BIC. **The evidence and BIC are dimensionally consistent.**

(Liddle & al., 2006)

The EM algorithm

(Dempster, Laird & Rubin, 1977)

Recall the max. likelihood principle:

$$\begin{aligned} P(\boldsymbol{\theta} \mid \mathbf{d}, I) &= \frac{P(\mathbf{d} \mid \boldsymbol{\theta}, I)}{P(\mathbf{d} \mid I)} \cdot P(\boldsymbol{\theta} \mid I) \\ &= \frac{\mathcal{L}(\mathbf{d}, \boldsymbol{\theta})}{P(\mathbf{d} \mid I)} \cdot P(\boldsymbol{\theta} \mid I) \propto \mathcal{L}(\mathbf{d}, \boldsymbol{\theta}) \end{aligned}$$

uniform distribution
(usually an improper prior)

evidence

likelihood

in this (approximate) setting, the MAP estimate coincides with the ML estimate.

when data are independent and identically distributed (i.i.d.) we find the following likelihood function

$$\mathcal{L}(\mathbf{d}, \boldsymbol{\theta}) = \prod_i p(d_i | \boldsymbol{\theta})$$

and we estimate the parameters by maximizing the likelihood function

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{d}, \boldsymbol{\theta})$$

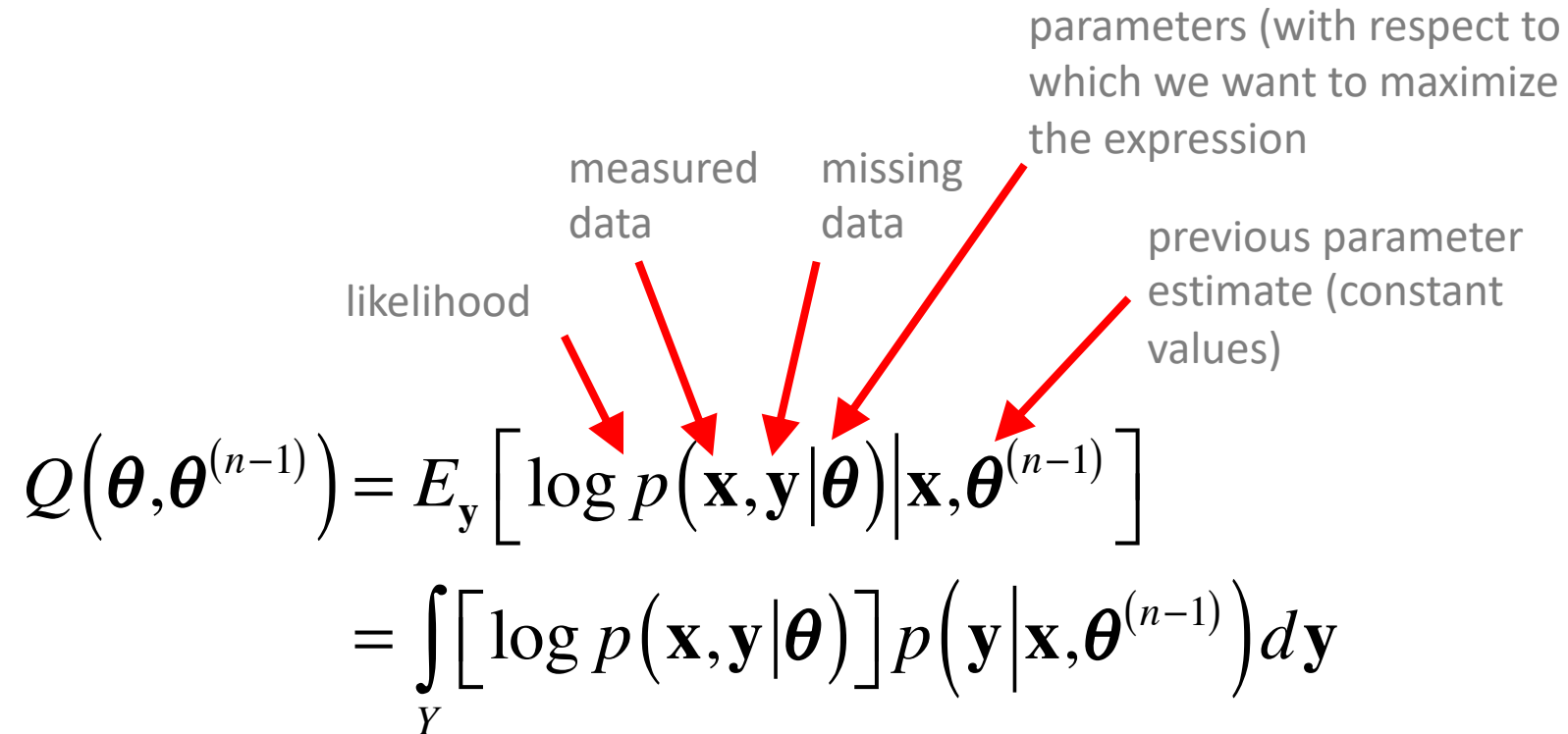
or, equivalently, its logarithm

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} [\log \mathcal{L}(\mathbf{d}, \boldsymbol{\theta})]$$

(in real life, this procedure is often complex and almost invariably it requires a numerical solution)

The EM algorithm is used to maximize likelihood with incomplete information, and it has two main steps that are iterated until convergence:

E. expectation of the log-likelihood, averaged with respect to missing data:


$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(n-1)}) = E_{\mathbf{y}} \left[\log p(\mathbf{x}, \mathbf{y} | \boldsymbol{\theta}) | \mathbf{x}, \boldsymbol{\theta}^{(n-1)} \right]$$
$$= \int_Y \left[\log p(\mathbf{x}, \mathbf{y} | \boldsymbol{\theta}) \right] p(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta}^{(n-1)}) d\mathbf{y}$$

M. maximization of the averaged log-likelihood with respect to parameters:

$$\boldsymbol{\theta}^{(n)} = \arg \max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(n-1)})$$

Example: an experiment with an exponential model (Flury and Zoppè)


Light bulbs fail following an exponential distribution with mean failure time θ

To estimate the mean two experiments are performed

1. n light bulbs are tested, all failure times u_i are recorded
2. m light bulbs are tested, only the total number r of bulbs failed at time t are recorded

$$1. \quad \mathcal{L} = \prod_{i=1}^n \frac{1}{\theta} \exp\left(-\frac{u_i}{\theta}\right) = \frac{1}{\theta^n} \exp\left(-\frac{\sum_i u_i}{\theta}\right) = \frac{1}{\theta^n} \exp\left(-\frac{n\langle u \rangle}{\theta}\right)$$

$$2. \quad \mathcal{L} = \prod_{i=1}^m \frac{1}{\theta} \exp\left(-\frac{v_i}{\theta}\right)$$

 missing data!

combined likelihood

$$\frac{1}{\theta^n} \exp\left(-\frac{n\langle u \rangle}{\theta}\right) \cdot \prod_{i=1}^m \frac{1}{\theta} \exp\left(-\frac{v_i}{\theta}\right)$$

log-likelihood

$$-n \ln \theta - \frac{n\langle u \rangle}{\theta} - \sum_{i=1}^m \left(\ln \theta + \frac{v_i}{\theta} \right)$$

expected failure time for a bulb that
is still burning at time t

$$t + \theta$$

expected failure time for a bulb that
is not burning at time t

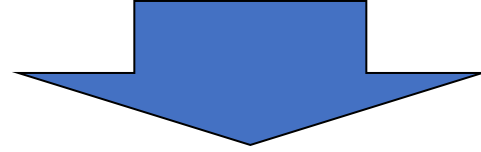
$$\theta - \frac{t \exp(-t/\theta)}{1 - \exp(-t/\theta)}$$

Note on mean failure time for a bulb that is not burning at time t

$$p(t') \propto \frac{1}{\theta} e^{-t'/\theta} \quad 0 \leq t' \leq t$$

$$\text{normalization} = \int_0^t p(t') dt' = \int_0^t \frac{dt'}{\theta} e^{-t'/\theta} = 1 - e^{-t/\theta}$$

$$\begin{aligned} \text{mean failure time} &= \int_0^t t' p(t') dt' = \frac{1}{1 - e^{-t/\theta}} \int_0^t t' e^{-t'/\theta} \frac{dt'}{\theta} \\ &= \frac{\theta}{1 - e^{-t/\theta}} \left[1 - e^{-t/\theta} - (t/\theta) e^{-t/\theta} \right] \\ &= \theta - \frac{te^{-t/\theta}}{1 - e^{-t/\theta}} \end{aligned}$$



average log-likelihood

$$\begin{aligned} Q &= E \left[-n \ln \theta - \frac{n \langle u \rangle}{\theta} + \sum_{i=1}^m \left(-\ln \theta - \frac{v_i}{\theta} \right) \right] \\ &= -(n+m) \ln \theta - \frac{n \langle u \rangle}{\theta} - \frac{r}{\theta} \left(\theta - \frac{t \exp(-t/\theta)}{1 - \exp(-t/\theta)} \right) - \frac{(m-r)}{\theta} (\theta + t) \end{aligned}$$

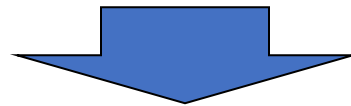
this ends the expectation step

the max of the mean likelihood

$$Q = -(n+m)\ln\theta - \frac{1}{\theta} \left[n\langle u \rangle + r \left(\theta - \frac{t \exp(-t/\theta)}{1 - \exp(-t/\theta)} \right) + (m-r)(\theta + t) \right]$$

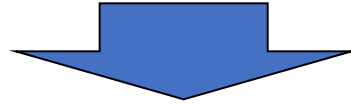
can be found by maximizing the approximate expression

$$Q \approx -(n+m)\ln\theta - \frac{1}{\theta} \left[n\langle u \rangle + r \left(\theta^{(k)} - \frac{t \exp(-t/\theta^{(k)})}{1 - \exp(-t/\theta^{(k)})} \right) + (m-r)(\theta^{(k)} + t) \right]$$



$$\frac{dQ}{d\theta} \approx -(n+m)\frac{1}{\theta} + \frac{1}{\theta^2} \left[n\langle u \rangle + r \left(\theta^{(k)} - \frac{t \exp(-t/\theta^{(k)})}{1 - \exp(-t/\theta^{(k)})} \right) + (m-r)(\theta^{(k)} + t) \right] = 0$$

$$\frac{dQ}{d\theta} \approx -(n+m)\frac{1}{\theta} + \frac{1}{\theta^2} \left[n\langle u \rangle + r \left(\theta^{(k)} - \frac{t \exp(-t/\theta^{(k)})}{1 - \exp(-t/\theta^{(k)})} \right) + (m-r)(\theta^{(k)} + t) \right] = 0$$



$$\theta^{(k+1)} = \frac{1}{n+m} \left[n\langle u \rangle + r \left(\theta^{(k)} - \frac{t \exp(-t/\theta^{(k)})}{1 - \exp(-t/\theta^{(k)})} \right) + (m-r)(\theta^{(k)} + t) \right]$$

this formula summarizes expectation
and maximization: therefore, the recipe
is to iterate this until convergence ...

Example with mean failure time = 2 (a.u.), and randomly generated data ($n = 100$; $m = 100$). In this example $r = 36$.

