Image Restoration

The fields of image manipulation:

- **image reconstruction**
  to reconstruct a 2D or 3D image from some projections (TAC and PET algorithms)
- **image restoration**
  to find a true image μ starting from a degraded image n resulting from (a) the effect of the detector (a PSF R Point Spread Function) and (b) statistical noise r;

$$n_{ik} = \sum_{mn} R_{i-m,k-n}\, \mu_{mn} + r_{ik}$$

- **image enhancement**
  to manipulate an image to produce results more pleasing to an observer.

Often one use the chain

reconstruction + restoration + enhancement

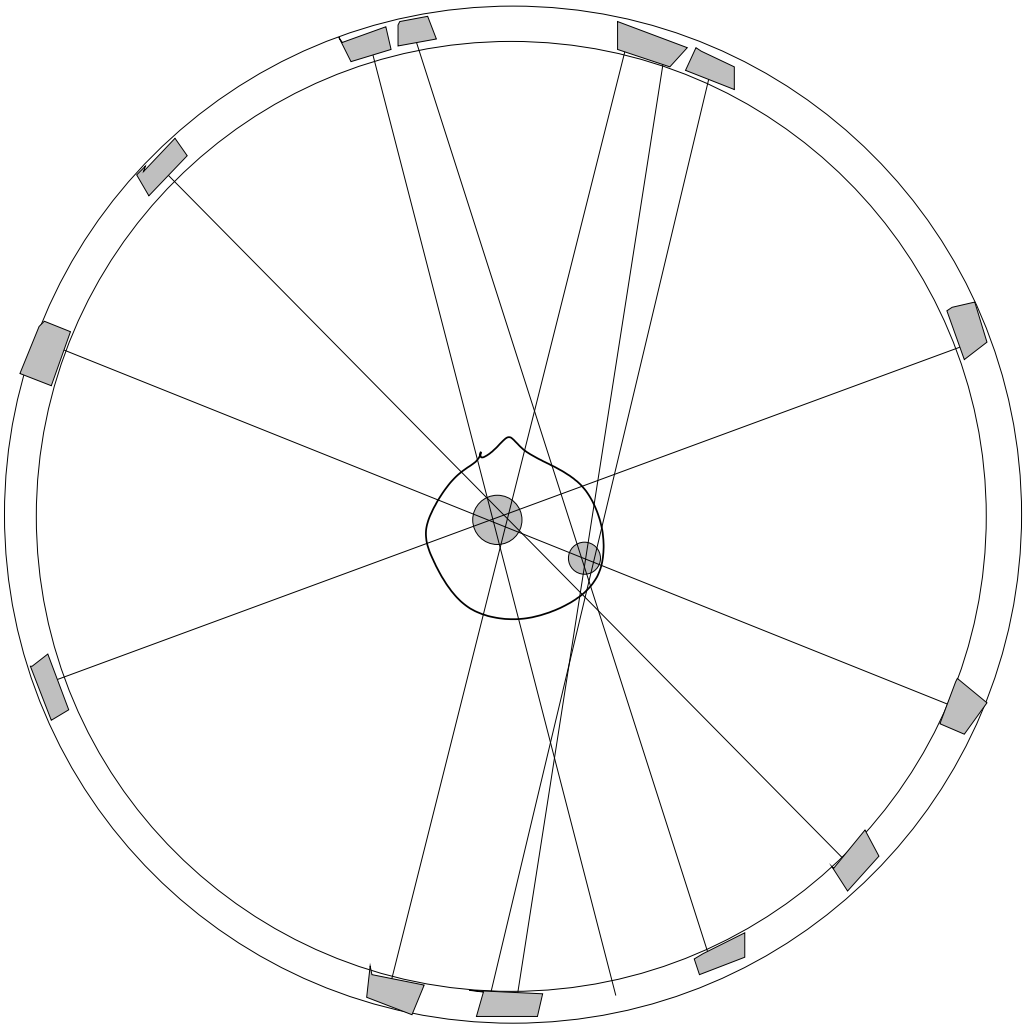
These lectures deal with some image restoration techniques

The fields involved are

- **statistics and estimation theory;**
- **ill posed inverse problems**
  where a small fluctuations in the data generates big differences in the solution;
- **linear algebra;**
- **numerical analysis;**

Positron Emission Tomography (PET)

An example of
reconstruction + restoration + enhancement



The 1D problem

In the reconstruction of an an histogram,

- **the true histogram (image) where the bin contents are the expected vales**

$$\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_N) \ , \quad \mu_j = \mu_{\rm tot} p_j = \mu_{\rm tot} \int_{\rm bin\,j} f_t(y)\,{\rm d}y$$

- **the expected histogram after the smearing effect of the detector**

$$\nu_i = \mu_{\rm tot} \int_{\rm bin\,i} {\rm d}x \int {\rm d}y\, s(x|y)\varepsilon(y) f_t(y)$$

$$\nu_i = \sum_{=1}^{M} \frac{\int_{\rm bin\,i} {\rm d}x \int_{\rm bin\,j} {\rm d}y\, s(x|y)\varepsilon(y) f_t(y)}{\mu_j/\mu_{\rm tot}} \mu_j = \sum_j R_{ij}\mu_j$$

where ε is the efficiency and s is the Point Spread Function: probability to be seen in the bin x when one falls in bin y.
$s(x|y)\,\varepsilon(y)$ **is called the response function.**

$$\begin{aligned} R_{ij} &= P(\text{observed in bin i AND value in bin j})/P(\text{value in bin j}) \\ &= P(\text{observed in bin i}|\text{true value in bin j}) \end{aligned}$$

- **background** $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_N)$

$$\nu_i = \sum_j R_{ij}\mu_j + \beta_i$$

- **The number of observed events in the case of random processes of detection:**

$$n_i = \frac{\nu_i^{n_i}}{n_i!}\,{\rm e}^{-\nu_i}$$

The problem in 2-D

A picture in a $x-y$ plane is the result of a double dimensional folding, where the true vertex points are smeared out by detector effects.

$$N = \sum_{ij=1}^{n_c} N_{ij}(\text{exp}) \ , \tag{1}$$

N is the total number of events and $N_{ij}(\text{exp})$ is the recorded number of event in the pixel placed at the ith-row and jth-column.

The observed $N_{ij}(\text{exp})$ events have to be compared with the expected values $N_{ij}(\text{th})$ predicted by a model.

$$N_{ij}(\text{th}) = N P_{ij}(\text{obs}) = N \sum_{i'j'} P_{i'j'}(\text{true})\, P_v(\text{obs}_{ij}|\text{true}_{i'j'}) \ , \tag{2}$$

that is, the number of vertices observed in the ijth-cell is due to the annihilations into the $i'j'$th-cell, times the probability P_v that the vertex algorithm shifts the point from the $i'j'$ to the ij-cell. One has to sum on all the cells near the $ij-$one.

In eq. (2) the normalization is understood. In practice, from (1, 2), one has

$$P(\text{true}_{ij}|\text{obs}_{i'j'}) = \frac{P_{i'j'}(\text{true})\, P_v(\text{obs}_{ij}|\text{true}_{i'j'})}{\sum_{ij}\left[\sum_{i',j'} P_{i'j'}(\text{true})\, P_v(\text{obs}_{ij}|\text{true}_{i'j'})\right]} \tag{3}$$

In the case of a two dimensional Gaussian point spread function PSF:

$$P_v(\text{obs}_{ij}|\text{true}_{i'j'}) = \frac{1}{2\pi\,\sigma_x\sigma_y}\ \exp\left[-\frac{(x_{ij}-x_{i'j'})^2}{2\,\sigma_x^2} - \frac{(y_{ij}-y_{i'j'})^2}{2\,\sigma_y^2}\right] , \tag{4}$$

PSF functions

The various form of the Blur or Point Spread function:

- **Gaussian Blur**

$$R(i,j) \propto \exp\left(-\frac{i^2+j^2}{2\sigma^2}\right)$$

  common in aerial imaging affected by atmospheric turbulence. Very common in astrophysics.

- **Uniform Out-of-Focus Blur**

$$R(i,j) \propto \begin{cases} \dfrac{1}{\pi R^2}\ , & \text{if } \sqrt{i^2+j^2} \leq R \\ \\ 0\ , & \text{otherwise} \end{cases}$$

  This is a defocussing of many imaging systems, sometimes parametrized as

$$R(i,j) \propto \begin{cases} \dfrac{1}{L^2}\ , & \text{if } \quad -L/2 \leq i,j \leq L/2 \\ \\ 0\ , & \text{otherwise} \end{cases}$$

These PSF functions acts as low-pass filter smoothing the original image and attenuating the edge information which is very important for the human visual perception.

The problem with fluctuations

From $\nu \to \mu$ deterministic methods can be used

$$\boldsymbol{\nu} = R\,\boldsymbol{\mu} \tag{5}$$

From $n \to \nu \to \mu$ statistical methods must be used

$$\boldsymbol{n} = \boldsymbol{\nu} + \boldsymbol{\rho} = R\,\boldsymbol{\mu} + \boldsymbol{\rho}$$

In the poissonian or binomial case we have to minimize:

$$-\ln L(\boldsymbol{\mu}) = -\sum_i \ln P(n_i, \nu_i)$$

In the gaussian case we must minimize

$$\chi^2(\boldsymbol{\mu}) = \sum_{ij} (\nu_i - n_i)(V^{-1})_{ij}(\nu_j - n_j)$$

These estimators are unbiased:

$$E[\hat{\mu}_j] = \sum_i (R^{-1})_{ji}\, E[n_i - \rho_i] = \sum_i (R^{-1})_{ji}\,(\nu_i - \rho_i) = \mu_j$$

In 2-D, when $N_{ij}(\mathrm{exp})$ contains fluctuations, we have to minimize:

$$\chi^2 = \sum_{ij} \frac{[N_{ij}(\mathrm{exp}) - NP_{ij}(\mathrm{obs})]^2}{NP_{ij}(\mathrm{obs})} \tag{6}$$

where

$$P(\mathrm{obs}) = P(\nu|\mu)\,P(\mu)$$

Fourier techniques

$$f(x) = \int F(t)\ \mathrm{e}^{2\pi i\, xt}\ \mathrm{d}t$$

Convolution:

$$f(x) = \int g(y)\delta(x-y)\,\mathrm{d}y$$

$$\int F(t)\,\mathrm{e}^{2\pi i x t}\ \mathrm{d}t = \int G(t)\ \mathrm{e}^{2\pi i\, yt}\ \Delta(t)\,\mathrm{e}^{2\pi i\,(x-y)t}\ \mathrm{d}t$$

$$\int F(t)\,\mathrm{e}^{2\pi i\, xt}\ \mathrm{d}t = \int G(t)\ \Delta(t)\,\mathrm{e}^{2\pi i\, xt}\ \mathrm{d}t \rightarrow F(t) = G(t)\,\Delta(t)$$

Correlation

$$\mathrm{Corr}(g,\delta) \equiv \int g(x+y)\,\delta(y)\,\mathrm{d}y \rightarrow G(t)\Delta^*(t)$$

if the functions are real

$$G(t) = G(-t)^*\ , \quad \mathrm{Corr}(g,\delta) \rightarrow G(t)\Delta(-t)$$

Autocorrelation (Wiener theorem)

$$\mathrm{Corr}(g,g) \rightarrow |G(t)|^2$$

Total Power:

$$P(f) \equiv \int |f(x)|^2\,\mathrm{d}x = \int |F(t)|^2\,\mathrm{d}t$$

Power Spectral Density (in the Fourier space):

$$PSD(f) \equiv |F(t)|^2 + |F(-t)|^2 \quad \overset{\text{f(x) real}}{\rightarrow} \quad 2|F(t)|^2 \quad 0 \le t \le \infty$$

Image Deconvolution

$$D(\boldsymbol{x}) = \int \mathrm{d}\boldsymbol{y} I(\boldsymbol{y})\, \delta(|\boldsymbol{x} - \boldsymbol{y}|)$$

In the absence of noise

$$I = F^{-1}\left[\frac{F(D)}{F(\delta)}\right]$$

where F is the Fourier transform.

For a real image $I(n_1, n_2)$ the Fourier transform is:

$$F(k_1, k_2) = \sum_{n_2=0}^{N_2-1} \sum_{n_1=0}^{N_1-1} \mathrm{e}^{2\pi i k_2 n_2/N_2}\, \mathrm{e}^{2\pi i k_1 n_1/N_1}\ I(n_1, n_2)$$

$$F(k_1, k_2) = FFT_2[FFT_1[I(n_1, n_2)]]$$

For the routine see for example *Numerical Recipes*

The use of pure inversion methods in the presence of statistical fluctuations gives often meaningless results

$$D(\boldsymbol{x}) = \int \mathrm{d}\boldsymbol{y} I(\boldsymbol{y})\, \delta(|\boldsymbol{x} - \boldsymbol{y}|) + \boldsymbol{\rho}$$

The problem with fluctuations

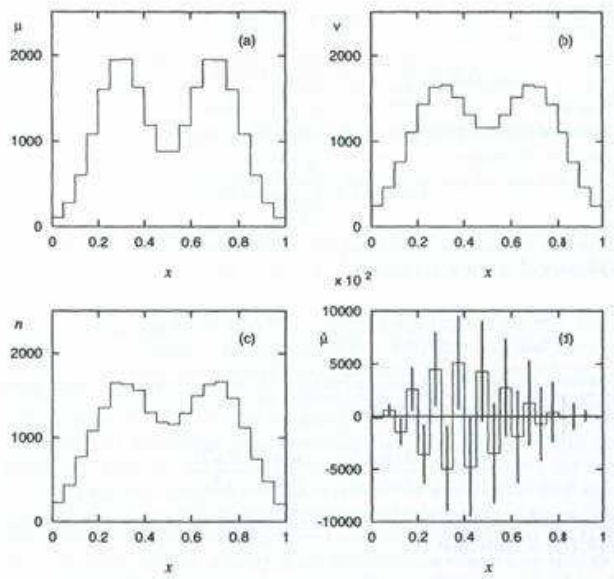


Fig. 11.1 (a) A hypothetical true histogram μ, (b) the histogram of expectation values $\nu = R\mu$, (c) the histogram of observed data n, and (d) the estimators $\hat{\mu}$ obtained from inversion of the response matrix.

Figure 1: **Lena restored by FFT: The original image (top left) is sampled with Poisson statistics (top right) and smeared with a 2D 10-bins Gaussian PSF (bottom left): the Fourier restored image (bottom right) is similar to the Poisson sampled image. In this case the noise term N is neglected.**

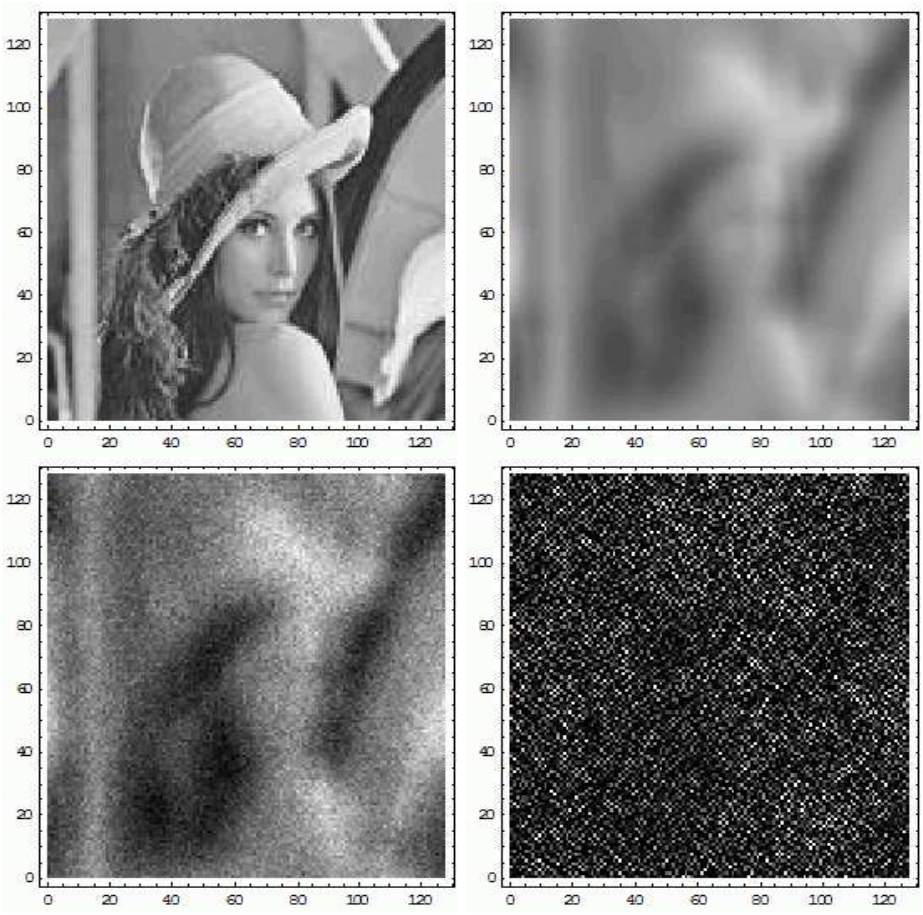


Figure 2: **Lena not restored by FFT: In this case the noise term N is not ignored: the original image (top left) is smeared with a 2D 10-bins Gaussian PSF (top right) and the result is sampled with Poisson statistics (bottom left): the Fourier restored image (bottom right) cannot recover the information lost in the noise. Another approach, statistical in nature, is required.**

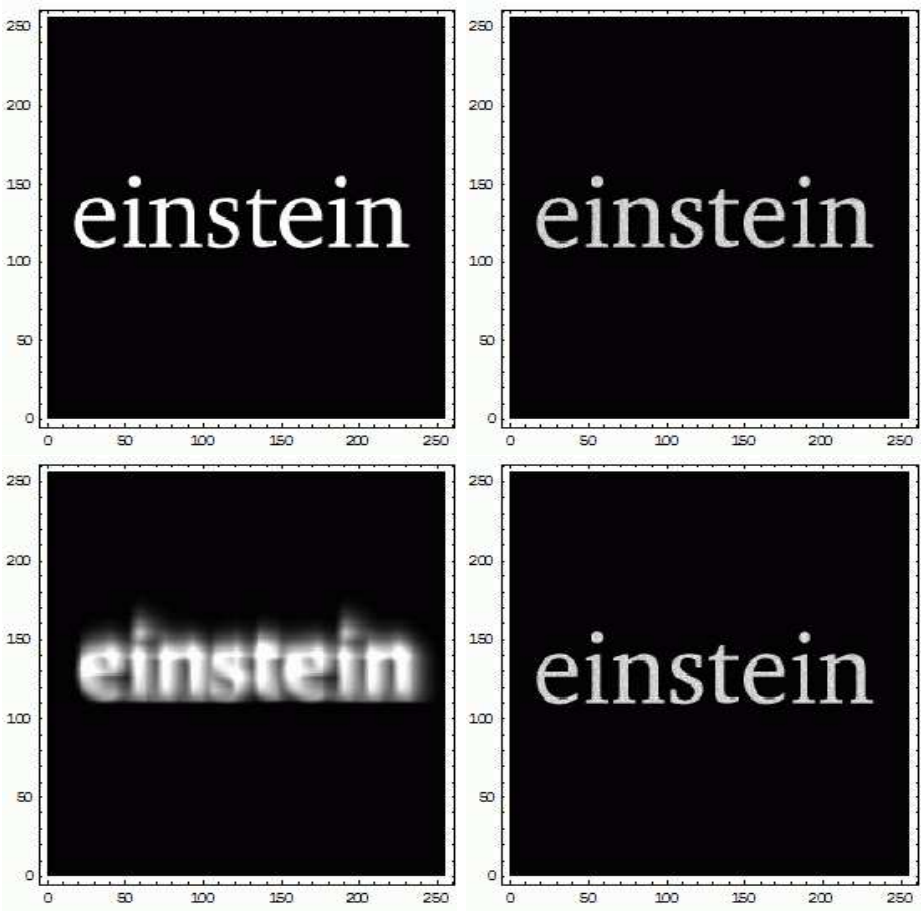


Figure 3: **Einstein restored by FFT: explanation as in Figure 1.**

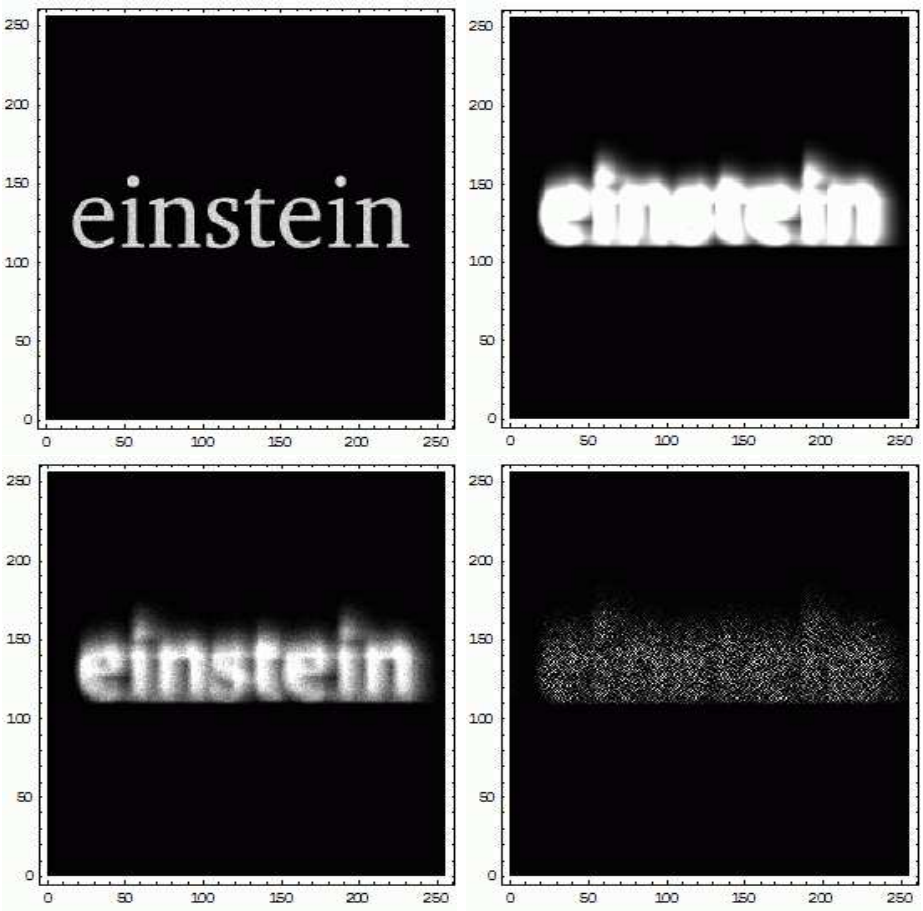
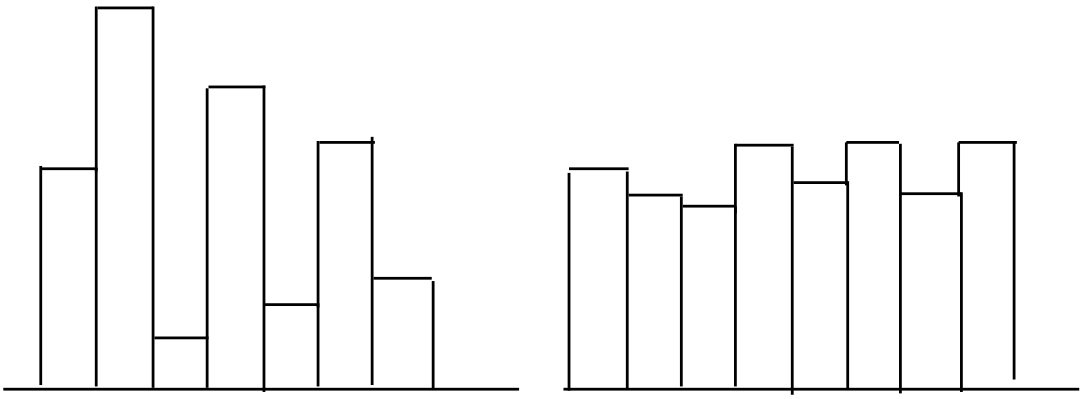


Figure 4: **Einstein not restored by FFT: explanation as in Figure 2.**

The problem with fluctuations



spike solution
HIGLY PROBABLE

smooth solution
UNLIKE

many solutions give a good χ^2

the spike ones are more probable!

Cure: to add to χ^2 an empirical regularization term $C[p]$.

$$\chi^2 \to \alpha\,\chi^2 + C[P(\text{true})]$$

or

$$\chi^2 \to \chi^2 + \alpha C[P(\text{true})]$$

The problem with fluctuations
General solution

Remember

$$\mu_i \overset{PSF}{\rightarrow} \nu_i \overset{random}{\rightarrow} n_i$$

and consider (3) as a form of the Bayes theorem

$$P(\text{true}_{ij}|\text{obs}) \propto \sum_{i',j'} P_{i'j'}(\text{true})\, P_v(\text{obs}_{ij}|\text{true}_{i'j'}) = L(\boldsymbol{n}|\boldsymbol{\mu})P(\boldsymbol{\mu})$$

Bayesians say: posterior = likelihood $\times$ prior One maximizes $P(\text{true}_{ij}|\text{obs}) \equiv F(\boldsymbol{\mu})$ (or minimize $-F(\boldsymbol{\mu})$):

$$F(\boldsymbol{\mu}) = \ln L(\boldsymbol{n}|\boldsymbol{\mu}) + \ln P(\boldsymbol{\mu}) \tag{7}$$

following the Maximum Likelihood (ML) principle.
The practical (no Bayesian) experimentalist introduces an empirical regularization parameter α and considers the prior $P(\boldsymbol{\mu})$ as a regularization function $C(\boldsymbol{\mu})$:

$$F(\boldsymbol{\mu}) = \alpha\, \ln L(\boldsymbol{n}|\boldsymbol{\mu}) + C(\boldsymbol{\mu}) \tag{8}$$

By keeping fixed the normalization:

$$\nu_T = \sum_i \sum_j R_{ij}\hat{\mu}_j + \rho_i = n_T$$

the objective function is

$$F(\boldsymbol{\mu}) = \alpha\, \ln L(\boldsymbol{n}|\boldsymbol{\mu}) + C(\boldsymbol{\mu}) + \lambda(n_T - \sum_i \nu_i) \tag{9}$$

where λ is a Lagrange multiplier

$$\frac{\partial F}{\partial \lambda} = 0 \rightarrow \sum_i n_i = n_T$$

Computational Aspects

If the number of μ_i is equal to the number of the pixels we have an ill-posed problem with zero degrees of freedom (DOF).
A first cure is the regularization term
A second cure is the normalization constraint given by the Lagrange Multiplier λ
In this case one minimizes

$$\chi^2 = -2 \ln L(\boldsymbol{n}|\boldsymbol{\mu}) \ , \quad p \text{ parameters}$$

with M equations of constraint

$$\Phi_k(\boldsymbol{n}, \boldsymbol{\mu}, \boldsymbol{\varphi}) = 0 \qquad k = 1, 2, \ldots, M \ , \tag{10}$$

where there are the p parameters and sometimes also q non-measured parameters $\boldsymbol{\varphi}$. We must have $M < p$
By minimizing χ^2 and from (10)

$$\mathrm{d}(\chi^2 + \Phi_k) = \sum_{j=1}^{p} \frac{\partial}{\partial \mu_j}[\chi^2 + \Phi_k] \, \mathrm{d}\mu_j + \sum_{n=1}^{q} \frac{\partial \Phi_k}{\partial \varphi_n} \, \mathrm{d}\varphi_n = 0 \ , \quad k = 1, 2 \ldots, M \ . \tag{11}$$

However, the differentials $\mathrm{d}\mu_j$ are not independent due to constraint equations. To restore the independence one introduces the M Lagrange Multipliers

$$\sum_{j=p-M}^{p} \frac{\partial}{\partial \mu_j} [\chi^2 + \lambda_k \Phi_k] = 0 \ . \tag{12}$$

which make the μ_j independent.

Computational Aspects II

We have:

$$\mathrm{d}(\chi^2 + \lambda_k \Phi_k) = \sum_{j=1}^{p-M} \frac{\partial}{\partial \mu_j}[\chi^2 + \lambda_k \Phi_k]\ \mathrm{d}\mu_j + \sum_{n=1}^{q} \lambda_k \frac{\partial \Phi_k}{\partial \varphi_n}\ \mathrm{d}\varphi_n = 0\ ,$$
$$k = 1, 2 \ldots, M\ .$$

By adding (12) finally we obtain:

$$\sum_{j=1}^{p} \frac{\partial}{\partial \mu_j}[\chi^2 + \lambda_k \Phi_k]\ \mathrm{d}\mu_j + \sum_{n=1}^{q} \lambda_k \frac{\partial \Phi_k}{\partial \varphi_n}\ \mathrm{d}\varphi_n = 0\ , \qquad (13)$$

where the parameter independence implies all null derivatives!
The constraint equations are equivalent to the derivatives w.r.t. λ: hence the objective function to be minimized w.r.t. $\mu_j, \varphi_n, \lambda_k$, with
$j = 1, 2, \ldots, p,\ \ n = 1, 2, \ldots, q,\ \ k = 1, 2, \ldots, M$ **is**

$$S(\boldsymbol{\theta}, \boldsymbol{\varphi}, \boldsymbol{\lambda}) = \left[\chi^2(\boldsymbol{\theta}) + 2 \sum_{k=1}^{M} \lambda_k\ \Phi_k(\boldsymbol{\theta}, \boldsymbol{\varphi})\right] \qquad (14)$$

The degrees of freedom are:

$$\nu = n - p + M - q$$

number of points - free parameters + constraints - non-measured parameters

Computational Aspects III

If the number of μ_i is equal to the number of the pixels we have an ill-posed problem with zero degrees of freedom (DOF).

First cure is the regularization term

Second cure is the normalization constraint given by the Lagrange Multiplier λ

Third cure: increase DOF by selecting the simplest model which yields an image that is statistically consistent with the data, where simplicity is measured by the lowest number of parameters required to describe the image. This technique reduces the number of free parameters μ_j by searching the number and size of the important homogeneous zones of the true image.
These zones are sometime called Pixons
Important groups of homogeneous pixels= Pixons
This technique is successful in Astrophysics (see literature)

Statistically, does an expression like

$$F(\boldsymbol{\mu}) = \alpha \ln L(\boldsymbol{n}|\boldsymbol{\mu}) + C(\boldsymbol{\mu}) + \lambda(n_T - \sum_i n_i)$$

make any sense?

Yes it is a typical form of Bayesian analysis...

A Bayesian Primer

- ***a problem of probability calculus*: if $p = 1/2$ for having head in tossing a coin, what is the probability to have in 1000 coin tosses less than 450 heads?**
- ***the same problem in statistics*: if in 1000 coin tosses 450 heads have been obtained, what is the estimate of the true head probability?**

Statistical error: $s \approx \sigma$

$$\begin{aligned} \mu \pm \sigma &= 500.0 \pm 15.8 \simeq 500 \pm 16 = [484, 516] \\ x \pm s &= 450.0 \pm 15.7 \simeq 450 \pm 16 = [434, 466] \end{aligned}$$

PROBABILITY CALCULUS	STATISTICS
probability of spectrum values	**estimate of parameters** $(\hat{p})$
true probability: $p = 0.5$	**frequency:** $f = x/n = 0.45$
expected value: $\langle X \rangle = 500$	**measurement:** $x = 450$
standard deviation: $\sigma[X] = \sqrt{np(1-p)} = 15.8$	**statistical error (incertitude):** $s = \sqrt{nf(1-f)} = 15.7$

Statistics

We have to inferences

- **parameter estimation: to estimate p from 1000 coin tosses**
- **hypothesis testing: in in two experiments of 1000 coin tosses 450 and 600 tosses have been obtained, how much is probable that the two experiments use two consistent coins?**

Parametric Statistics: the probability depends on θ:

$$\mathcal{E}(\theta) \equiv (S, \mathcal{F}, P_\theta)$$

corresponding to a density

$$P\{X \in A\} = \int_A p(x;\theta)\,\mathrm{d}x$$

Physics and Statistics

- **Higgs mass (PDG 2000):**

$$m > 95.3\ GeV, CL = 95\%$$

- W **mass:**

$$m_W = 80.419 \pm 0.056\ GeV$$

These are
confidence intervals

WHAT IS THEIR MEANING?

The hystorical path

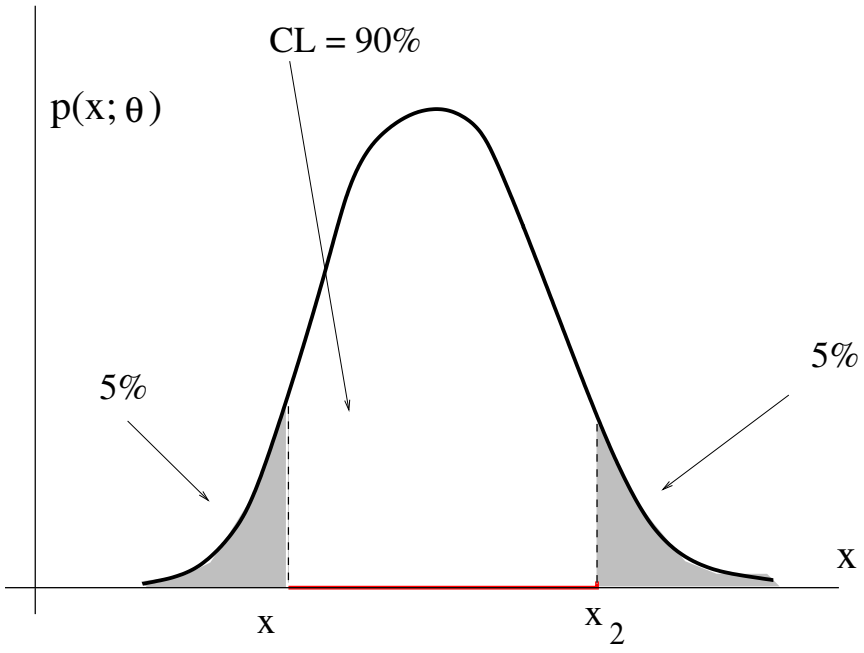
	FREQUENTISTS	BAYESIANS
1763		Thomas Bayes writes a fundamental paper. Bayesian age
1900	Karl Pearson proposes the χ^2 test	
1910	Robert Fisher invents Maximum Likelihood	
1937	The J. Neyman frequentist interval estimate	
1940	The Hypothesis testing of Pearson. Frequentist age The Popper scheme Frequentist teaching	
1990		rediscovering of the bayesian works of Jeffreys, De Finetti and Jaynes
now	the debate is open: see on Confidence Limits	the CERN Workshop (Geneva 2000) neo-Bayesian age?

Frequentist Confidence Intervals

One (Neyman, 1937) starts from probability calculus

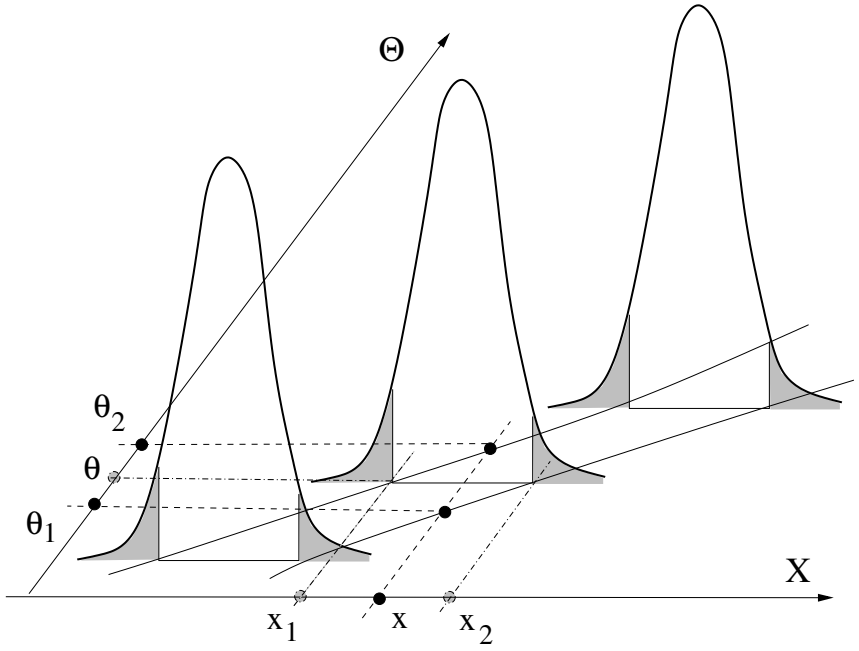
$$\int_{x_1}^{x_2} p(x;\theta)\,\mathrm{d}x = CL$$

and the procedure is repeated for all the pos-



sible θ values

Frequentist Confidence Intervals



It is possible to shown that

$$X \in [x_1, x_2] \text{ iff } \Theta \in [\theta_1, \theta_2]$$

Since

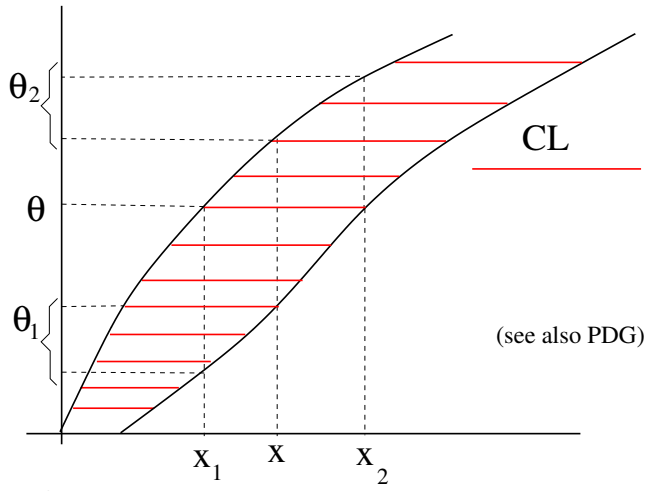
$$P\{X \in [x_1, x_2]\} = CL$$

then

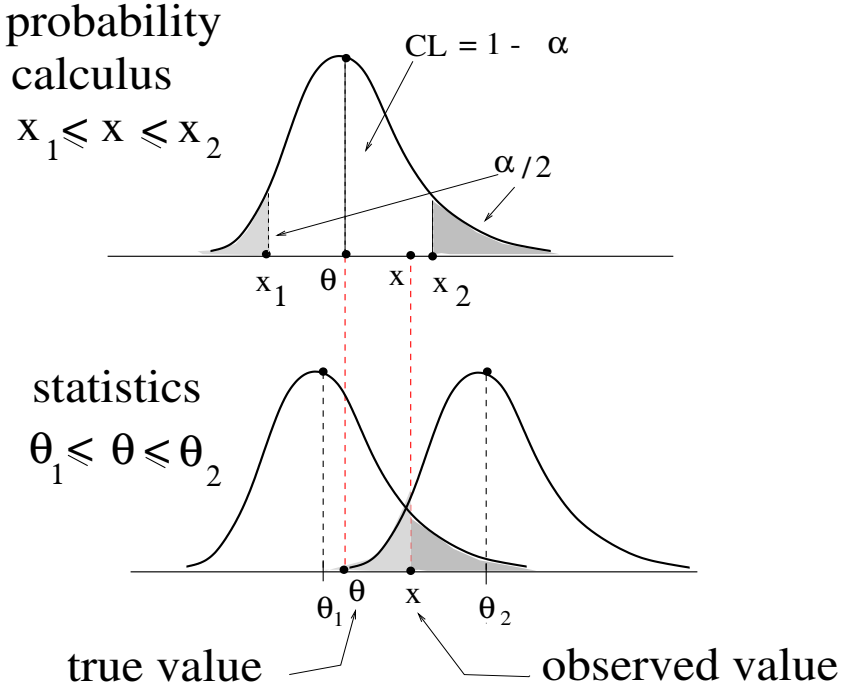
$$P\{\Theta \in [\theta_1, \theta_2]\} = CL$$

Fundamental Neyman result (1937)

Cut and top views of the Neyman construction:



$$P\{\theta_1 < \theta < \theta_2\} = P\{x_1 < x < x_2\} = CL$$



Frequentist Confidence Intervals mathematical definitions

If T_1 and T_2 are two statistics, the interval

$$I = [T_1, T_2]$$

is a confidence interval for θ, with $0 < CL < 1$ confidence level, if, for all $\theta \in \Theta$, the probability that I contains θ (*coverage*) is CL:

$$P\{T_1 \leq \theta \leq T_2\} = CL \ .$$

If T_1 e T_2 are discrete variables, the confidence interval satisfies the *minimum overcoverage*

$$P\{T_1 \leq \theta \leq T_2\} \geq CL \ .$$

Note: $[T_1, T_2]$ are random variables, the θ parameter is fixed

Frequentist C.I.
right and wrong definitions

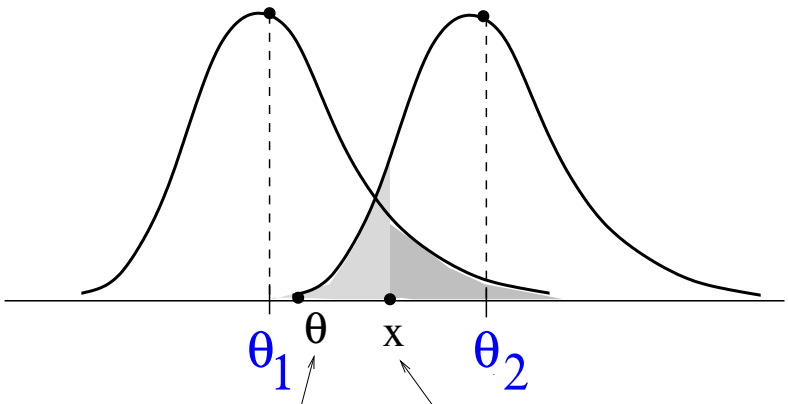
RIGHT quotations:

- **CL is the probability that the random interval $[T_1, T_2]$ covers the true value θ;**
- **in an infinite set of repeated identical experiments, a fraction equal to CL will succeed in assigning $\theta \in [\theta_1, \theta_2]$;**
- **if $\theta \notin [\theta_1, \theta_2]$, one can obtain $\{I = [\theta_1, \theta_2]\}$ in a fraction of experiments $\leq 1 - CL$**
- **if $H_0 : \theta \notin [\theta_1, \theta_2]$ the probability to reject a true H_0 is $1 - CL$ (falsification). see upper and lower limits estimates.**

WRONG quotations

- **CL is the degree of belief that the true value is in $[\theta_1, \theta_2]$**
- **$P\{\theta \in [\theta_1, \theta_2]\} = CL$ (θ is not a random variable!)**

Frequentist C.I. determination



true value measured value

$$\int_x^{\infty} p(x;\theta_1)\,\mathrm{d}x = c_1 \qquad \int_{-\infty}^{x} p(x;\theta_2)\,\mathrm{d}x = c_2$$

where

$$\theta \in [\theta_1, \theta_2] \ , \quad 1-(c_1+c_2) = CL$$

MC techniques can be used: grid over θ to find the values θ_1 and θ_2 satisfying these integrals

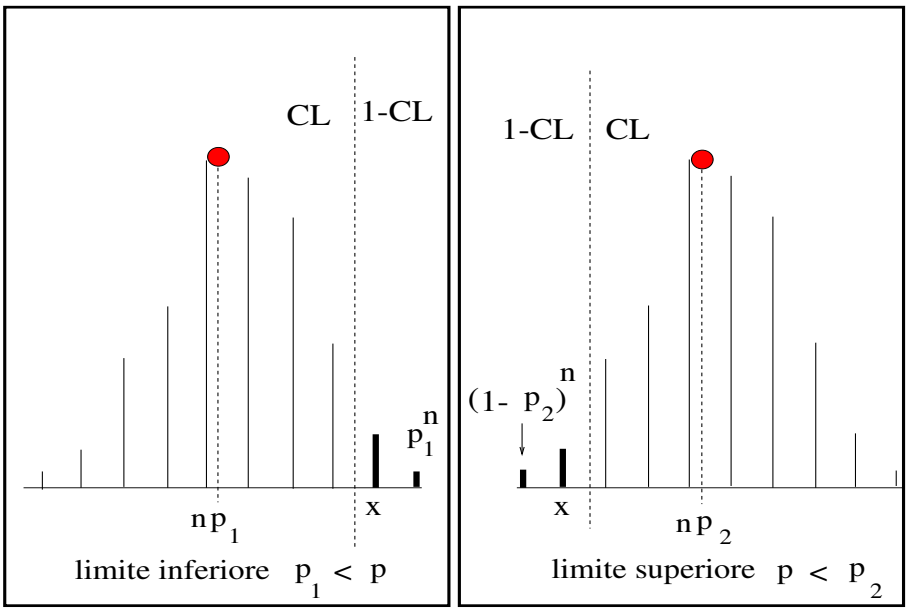
Important:

$$\int_{\theta_1}^{\theta_2} p(\theta;x)\,\mathrm{d}\theta = CL$$

WRONG!!!!

Upper and lower limits

x events are observed:



- **lower limit** $p \in [p_1, 1]$: **if $p < p_1$, one can observe at least x events, but in a fraction of experiments $< 1 - CL$.**
  If $x = n$, $p_1^n = 1 - CL$;
- **upper limit** $p \in [0, p_2]$: **if $p > p_2$, one can observe up to x events, but in a fraction of experiments $< 1 - CL$.**
  if $x = 0$, $(1 - p_2)^n = 1 - CL$.

Poisson Limits

$$\sum_{k=x}^{\infty} \frac{\mu_1^k}{k!} \exp(-\mu_1) = c_1 \ , \quad \sum_{k=0}^{x} \frac{\mu_2^k}{k!} \exp(-\mu_2) = c_2 \ ,$$

symmetric case: $c_1 = c_2 = (1 - CL)/2$.

Upper Limits to the mean number of events having obtained x events:

$$\sum_{k=0}^{x} \frac{\mu_2^k}{k!} \exp(-\mu_2) = 1 - CL \ .$$

For $x = 0, 1, 2$, where $\mu_2 \equiv \mu$

$$\begin{aligned} e^{-\mu} &= 1 - CL \ , \\ e^{-\mu} + \mu e^{-\mu} &= 1 - CL \ , \\ e^{-\mu} + \mu e^{-\mu} + \frac{\mu^2}{2} e^{-\mu} &= 1 - CL \end{aligned}$$

x	90%	95%	x	90%	95%
0	**2.30**	**3.00**	**6**	**10.53**	**11.84**
1	**3.89**	**4.74**	**7**	**11.77**	**13.15**
2	**5.32**	**6.30**	**8**	**13.00**	**14.44**
3	**6.68**	**7.75**	**9**	**14.21**	**15.71**
4	**7.99**	**9.15**	**10**	**15.41**	**16.96**
5	**9.27**	**10.51**	**11**	**16.61**	**18.21**

When $\mu > 2.3$, *one con observe no events* but in a number of experiments $< 10\%$.

The Bayes formula

$$P(B_k|A)P(A) = P(A|B_k)P(B_k)$$

if B_k are disjoint and cover the set S,

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

then $P(B_k|A)$ can be written as:

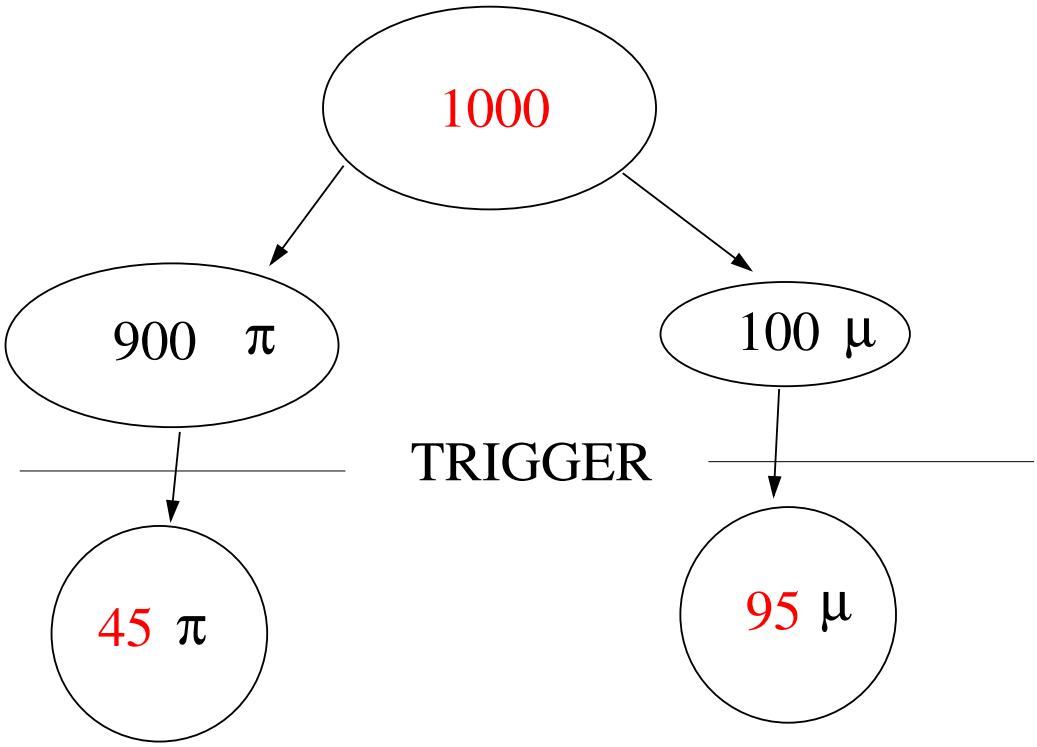
$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)} \quad , \qquad P(A) > 0 \ .$$

Example trigger problem
A μ-π trigger has $\varepsilon(\pi) = 0.05$ e $\varepsilon(\mu) = 0.95$. If the beam is 90% π and 10% μ find efficiency and enrichment factor

$$P(\mu|T) = \frac{P(T|\mu)P(\mu)}{P(T|\mu)P(\mu) + P(T|\pi)P(\pi)}$$
$$= \frac{0.95 \cdot 0.10}{0.95 \cdot 0.10 + 0.05 \cdot 0.90} = \frac{0.095}{0.14} = 0.678$$

efficiency $= 14\%$, enrichment ≈ 6.8

Bayes formula.
Physicist's point of view



$$P(\text{trigger}) = \frac{95 + 45}{1000} = 0.14$$

$$P(\mu|T) = \frac{95}{95 + 45} = 0.678$$

Bayesian use of Bayes formula

An attempt to solve the trigger problem without knowing
the beam percentages ...!!

The Bayes formula is employed starting from subjective probabilities

$$P(H_k|\text{data}) = \frac{P(\text{data}|H_k)P(H_k)}{\sum_{i=1}^{n} P(\text{data}|H_i)P(H_i)} \ .$$

an important step,

$$P(H_k|\text{data}) \rightarrow P_{n-1}(H_k)$$

iteration:

$$P_n(H_k|E) = \frac{P(E_n|H_k)P_{n-1}(H_k)}{\sum_{i=1}^{n} P(E_n|H_i)P_{n-1}(H_i)} \ ,$$

Bayesian use of Bayes formula

Example: from a box containing black and white balls, in unknown percentage, 1, 5 or 10 black balls are extracted consecutively. What is the black/white percentage??
Assumption: equal priors!!!

$$P_0(H_k|E) \equiv P(H_k) = 1/6$$

A posteriori probabilities $P(H|n = \ldots \bullet)$:

$$P(E|H_1) = 0 \ , P(E|H_2) = 1/5$$

$$P(E|H_3) = 2/5 \ldots P(E|H_6) = 1$$

$$P_n(H_k|E) = \frac{P(E|H_k)P_{n-1}(H_k|E)}{\sum_{i=1}^{n} P(E|H_i)P_{n-1}(H_i|E)} \ ,$$

hypothesis / **box content**	H_1 ○○○○○	H_2 ○○○○●	H_3 ○○○●●	H_4 ○○●●●	H_5 ○●●●●	H_6 ●●●●●
priors $P(H_i)$	**1/6**	**1/6**	**1/6**	**1/6**	**1/6**	**1/6**
$P(H_i\|n = 1\bullet)$	**0**	**0.07**	**0.13**	**0.20**	**0.27**	**0.33**
$P(H_i\|n = 5\bullet)$	**0.**	**0.00**	**0.01**	**0.05**	**0.23**	**0.71**
$P(H_i\|n = 10\bullet)$	**0.**	**0.00**	**0.00**	**0.00**	**0.11**	**0.89**

The two approaches

- **frequentist approach:**

  subjective probabilities
  for Hypotheses
  never
  are assumed.

  $P(\mathrm{H}|\mathrm{data})$ NO!!!!

- **Bayesian approach:**
  one must determine probabilities as

  $$P(\mathrm{H}|\mathrm{data}) \propto P(\mathrm{data}|\mathrm{H})P(\mathrm{H})$$
  $$\propto \text{likelihood x prior}$$

  which depend both on data and on the subjective initial hypotheses (priors)

The frequentist box problem

Example: 10 black balls are extracted consecutively. What is the black/white percentage??

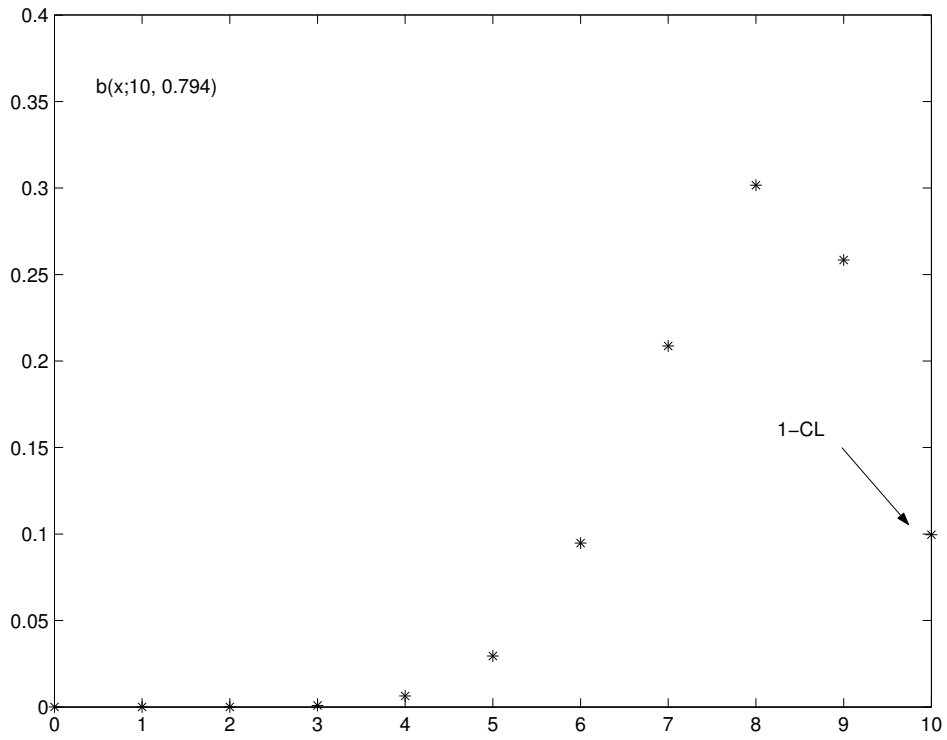
$$x = n = 10 \qquad CL = 90\%$$

$$p^n = 1 - CL \rightarrow p = (1 - CL)^{1/10} \approx 0.794$$

$$5 \text{ balls} \times 0.794 = 3.97 \approx 4$$

If the box contained less than 4 black balls, one could obtain such a result, but in less than 10% of the experiments ... or: the box contains more than 3 black balls. The probability to be wrong is $< 10\%$

No a priori probability has been assigned to the hypothesis on the box (state) preparation The frequentist result is independent of the box preparation


b(x;10, 0.794)
1−CL
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
1
2
3
4
5
6
7
8
9
10

The gambler problem
Bayesian approach

$$P(\text{Win}|C) = 1 \qquad P(\text{Win}|H) = 0.5$$

Problem: to find the probability that the gambler is cheat, as a function of the number of consecutive wins $\{W_n\}$

$$P(H) \equiv P(H|W_0)\,, \quad P(C) \equiv P(C|W_0) \quad P(H) = 1 - P(C)$$

Iteration:

$$P(C|W_n) = \frac{P(W_n|C)\,P(C|W_{n-1})}{P(W_n|C)\,P(C|W_{n-1}) + P(W_n|H)\,[1 - P(C|W_{n-1})]}$$

that is

$$P(C|W_n) = \frac{P(C|W_{n-1})}{P(C|W_{n-1}) + 0.5\,[1 - P(C|W_{n-1})]}$$

Bayes:

$P(C)/n$	**5**	**10**	**15**	**20**
1%	**24**	**91**	**99.7**	**99.99**
5%	**63**	**98**	**99.94**	**99.998**
50%	**97**	**99.9**	**99.997**	**99.999**

The gambler problem
Frequentist approach

Let us suppose 15 cosecutive wins

Hypothesis testing:
The null hypothesis H_0 (honest player) gives a significance level (p-value in this case)

$$0.5^{15} = 3.05\,10^{-5}$$

The probability to be wrong discarding the hypothesis is less then 0.003 %.
The player is cheat.

Cheat probability estimation:
with $n = 15$ and $CL = 90\%$ the probability is

$$p = (0.1)^{1/15} \approx 0.86 \ .$$

With a "cheat probability" $p < 0.86$ it is possible to win for 15/15 times, but in a percentage of plays $< 10\%$

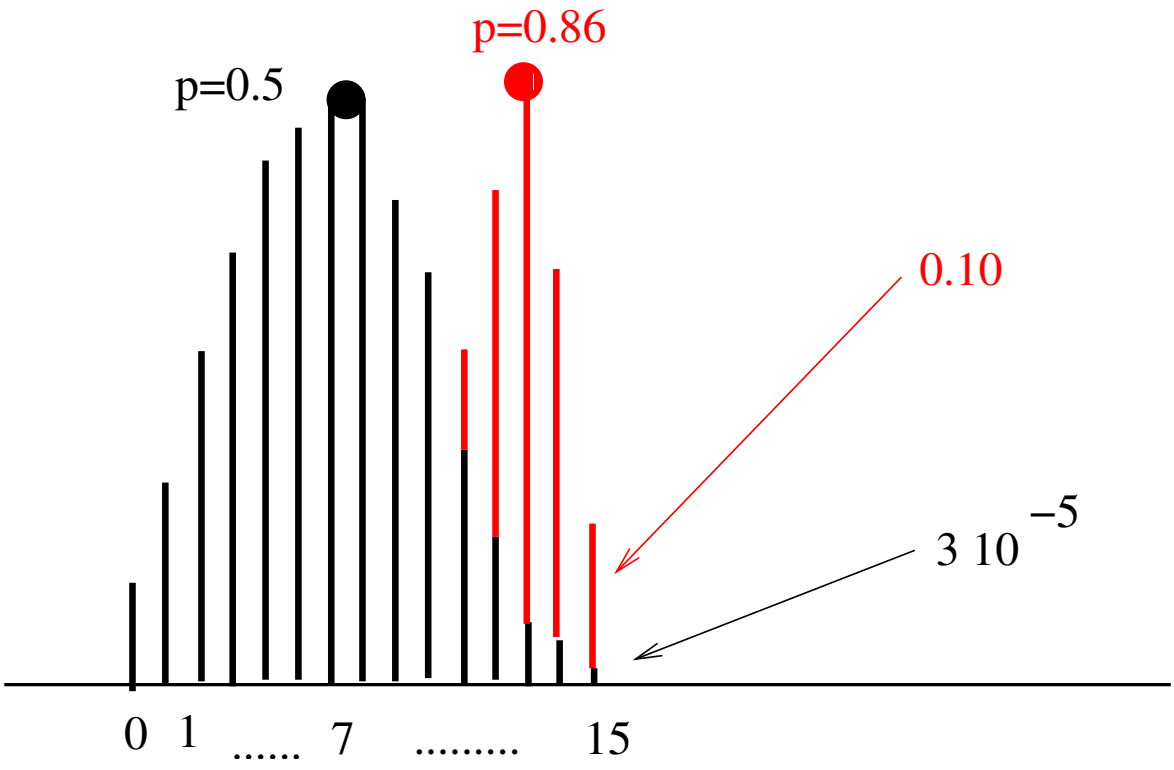
$$0.86 < p < 1 \qquad CL = 90\%$$

The gambler problem
Frequentist approach

Black: hypothesis testing

Red: probability estimation

These conclusions are independent of any a priori hypothesis!



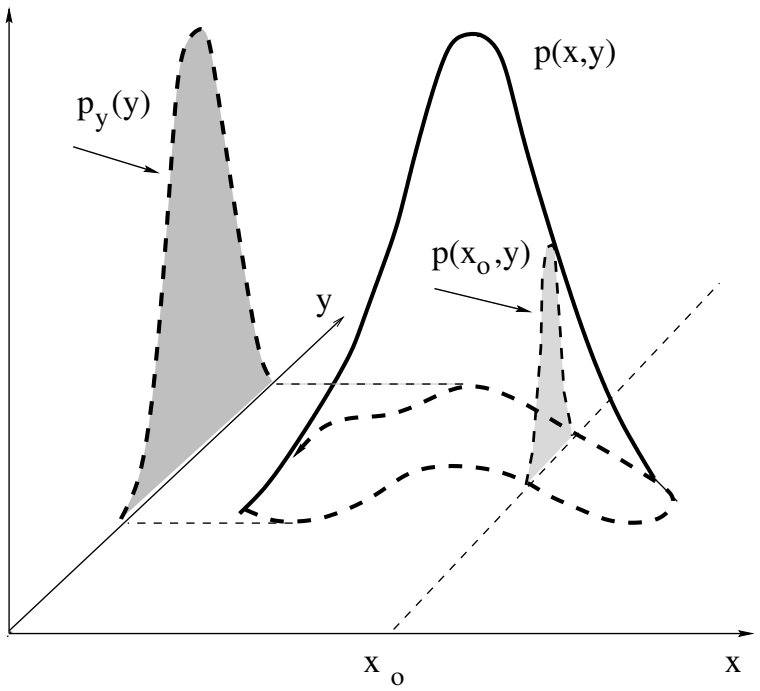
Marginal and conditional densities

probability product:

$$p(x,y) = p_Y(y)\, p(x|y) = p_X(x)\, p(y|x)$$

for independent variables:

$$p(x|y) = p_X(x) \ , \quad p(y|x) = p_Y(y) \ ,$$



Bayes for the continuum

$$p(x,y) = p_Y(y)\, p(x|y) = p_X(x)\, p(y|x)$$

hence

$$p(x|y) = \frac{p(y|x)\, p_X(x)}{p_Y(y)}$$

that is

$$p(x|y) = \frac{p(y|x)\, p_X(x)}{\displaystyle\int p(y|x)\, p_X(x)\, \mathrm{d}x}$$

Bayesian step:

$$p(\mu;x) = \frac{p(x;\mu)\, p_\mu(\mu)}{\displaystyle\int p(x;\mu)\, p_\mu(\mu)\, \mathrm{d}x}$$

that is

$$p(\mu;x) = \frac{\text{likelihood} \times \text{prior}}{\text{normalization}}$$

The prior

$$p_\mu(\mu)$$

that is the subjective probability assigned to μ, is NEVER used by frequentists

Bayesian coin tossing

$$p(p;n,x) = \frac{p^x(1-p)^{n-x}\, p_p(p)}{\displaystyle\int p^x(1-p)^{n-x}\, p_p(p)\,\mathrm{d}p}$$

With uniform prior,

$$p_p(p) = \text{const} \quad 0 < p < 1$$

Recalling the β function:

$$\int_0^1 p^x(1-p)^{n-x}\,\mathrm{d}p = \frac{x!(n-x)!}{(n+1)!}$$

one obtains the degree of belief of p

$$\begin{aligned} p(p;n,x) &= \frac{(n+1)!}{x!(n-x)!}\, p^x(1-p)^{n-x} \\ \langle p\rangle &= \frac{x+1}{n+1} \\ \mathrm{Var}[p] &= \frac{(x+1)(n-x+1)}{(n+3)(n+2)^2} \end{aligned}$$

Bayesian Interval estimate

$$p \in [p_1, p_2]$$

with degree of belief

$$\int_{p_1}^{p_2} \frac{(n+1)!}{x!(n-x)!}\, p^x (1-p)^{n-x}\, \mathrm{d}p$$

$x = n$**:**

$$p(p; x = n) = (n+1)p^n \ , \ F(p) = \int_0^p (n+1)p^n \, \mathrm{d}p = p^{n+1}$$

The 90% bayesian lower bound

$$0.10 = p^{n+1} \rightarrow p = (0.10)^{1/(n+1)}$$

$x = 0$**:**

$$\begin{aligned} p(p; x = 0, n) &= (n+1)(1-p)^n \ , \\ F(p) &= \int_0^p (n+1)(1-p)^n \, \mathrm{d}p = 1 - (1-p)^{n+1} \end{aligned}$$

The 90% bayesian upper bound

$$0.10 = p^{n+1} \rightarrow p = 1 - (0.10)^{1/(n+1)}$$

Frequentist $\rightarrow$ Bayesian

$n \rightarrow n+1$

but with a different meaning !!!

Bayesian Interval estimate

Degree of belief on μ for a measured x:

$$p(\mu; x) = \frac{L(x, \mu)\, p_\mu(\mu)}{\int L(x, \mu)\, p_\mu(\mu)\, \mathrm{d}p}$$

Estimate:

$$\mu \in [\mu_1, \mu_2]$$

with degree of belief

$$\int_{\mu_1}^{\mu_2} p(\mu; x)\, \mathrm{d}\mu = \text{degree of belief}$$

- **one integrates over μ considered as a random variable**
- **this coincides with the frequentist result if the prior $p_\mu(\mu)$ is uniform and the property**

  $$1 - F(\mu; x) = F(x; \mu)$$

  holds
- **but the interpretation is different!**

The neutrino mass
...here Bayes helps!

An experiment with a Gaussian resolution of

$$\sigma = 3.3 \text{ eV}/c^2$$

measures the ν_e mass as:

$$m = -5.41 \text{ eV}/c^2$$

make the Bayesian estimate of m_ν.

Bayes formula

$$p(m_\nu; m, \sigma) = \frac{p(m; m_\nu, \sigma)\, p_\nu(m_\nu)}{\int p(m; m_\nu, \sigma)\, p_\nu(m_\nu)\, \mathrm{d}m_\nu}$$

Choosing the prior:

- **define** $0 \leq m_\nu \leq 20 - 30 \text{ eV}/c^2$;
- **define** $\sigma_\nu = 10 \text{ eV}/c^2$
- **test three functional forms:**

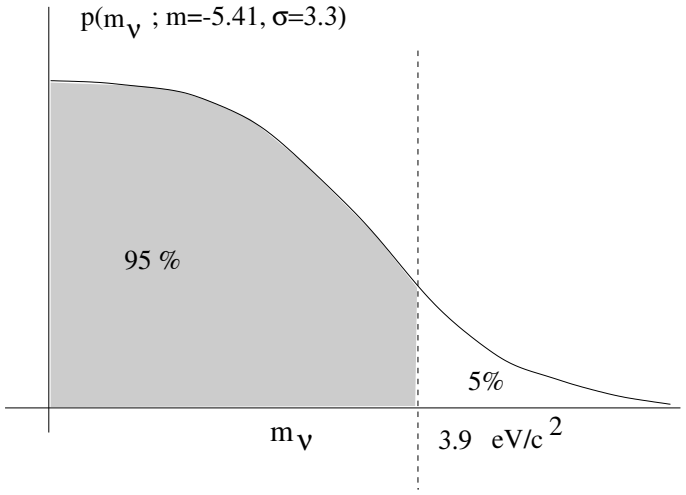
  1. **uniform:** $p_\nu = p_u(m_\nu) = 1/30$, $0 \leq m_\nu \leq 30$
  2. **Gaussian:**
  $$p_\nu = p_g(m_\nu) = \frac{2}{2\pi\sigma_\nu} \exp[-m_\nu^2/(2\sigma_\nu^2)]$$
  3. **triangular:** $p_\nu = p_t(m_\nu) = \frac{1}{450}(30 - m_\nu)$, $0 \leq m_\nu \leq 30 \text{ eV}/c^2$

The neutrino mass II

For example, using the uniform $p_u(m_\nu)$
and $\sigma = 3.3$, $m = -5.41$ **ev**$/c^2$**:**

$$p(m_\nu; m, \sigma) = \frac{\exp\left[-\dfrac{(m-m_\nu)^2}{2\sigma^2}\right]\dfrac{1}{30}}{\displaystyle\int_0^{30} \exp\left[-\frac{(m-m_\nu)^2}{2\sigma^2}\right]\frac{1}{30}\,\mathrm{d}m_\nu}$$

one obtains, at 95% **probability:**



- **uniform:** $0 \leq m_\nu \leq 3.9$ **eV**$/c^2$;
- **Gaussian:** $0 \leq m_\nu \leq 3.7$ **eV**$/c^2$;
- **triangular:** $0 \leq m_\nu \leq 3.7$ **eV**$/c^2$.

result "independent" of the prior!
Here the prior represent the knowledge, not the ignorance!!!

Bayesian/Frequentists Image Processing

Remember

$$\mu_i \overset{PSF}{\rightarrow} \nu_i \overset{random}{\rightarrow} n_i$$

and consider (3) as a form of the Bayes theorem

$$P(\text{true}_{ij}|\text{obs}) \propto \sum_{i',j'} P_{i'j'}(\text{true})\, P_v(\text{obs}_{ij}|\text{true}_{i'j'}) = L(\boldsymbol{n}|\boldsymbol{\mu})P(\boldsymbol{\mu})$$

Bayesians say: posterior = likelihood $\times$ prior One maximizes $P(\text{true}_{ij}|\text{obs}) \equiv F(\boldsymbol{\mu})$ (or minimize $-F(\boldsymbol{\mu})$):

$$F(\boldsymbol{\mu}) = \ln L(\boldsymbol{n}|\boldsymbol{\mu}) + \ln P(\boldsymbol{\mu}) \tag{15}$$

following the Maximum Likelihood (ML) principle.
The practical (no Bayesian) experimentalist introduces an empirical regularization parameter α and considers the prior $P(\boldsymbol{\mu})$ as a regularization function $R(\boldsymbol{\mu})$:

$$F(\boldsymbol{\mu}) = \alpha\, \ln L(\boldsymbol{n}|\boldsymbol{\mu}) + C(\boldsymbol{\mu}) \tag{16}$$

The present status of the Bayesian-Frequentist dispute:

- **The The Bayesian: always choose o noninformative prior; reject the concept of the ensemble of identical experiments and give a probability to all the hypotheses.**
- **The (frequentists) physicists: usually I avoid to give any probability to my hypotheses: they are rejected by the experience (falsification). I use only informative priors into the analysis when certain *a priori information* is available.**

Maximum Likelihood

Likelihood function:

$$L(\boldsymbol{\theta};\, \underline{\boldsymbol{x}}) \;=\; p(x_{11}, x_{21}, .., x_{m1}; \boldsymbol{\theta})\, p(x_{12}, x_{22}, .., x_{m2}; \boldsymbol{\theta}) \ldots$$
$$\times\, p(x_{1n}, x_{2n}, .., x_{mn}; \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\boldsymbol{x}_i; \boldsymbol{\theta}) \;\;,$$

the product covers
all the n values of the m variables $\boldsymbol{X}$.
Log-likelihood:

$$\mathcal{L} = -\ln\left(L(\boldsymbol{\theta};\, \underline{\boldsymbol{x}})\right) = -\sum_{i=1}^{n} \ln\left(p(\boldsymbol{x}_i; \boldsymbol{\theta})\right) \;\;,$$

Max L corresponds to Min $\mathcal{L}$.

For a given set of

$$\underline{\boldsymbol{x}} = \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$$

observed values, from a

$$\underline{\boldsymbol{X}} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n)$$

sample with density $p(\boldsymbol{x}; \boldsymbol{\theta})$, the ML estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is the maximum (if any) of the function

$$\max_{\Theta}\left[L(\boldsymbol{\theta};\, \underline{\boldsymbol{x}})\right] = \max_{\Theta}\left[\prod_{i=1}^{n} p(\boldsymbol{x}_i; \boldsymbol{\theta})\right] = L(\,\hat{\boldsymbol{\theta}};\, \underline{\boldsymbol{x}})$$

Maximum likelihood

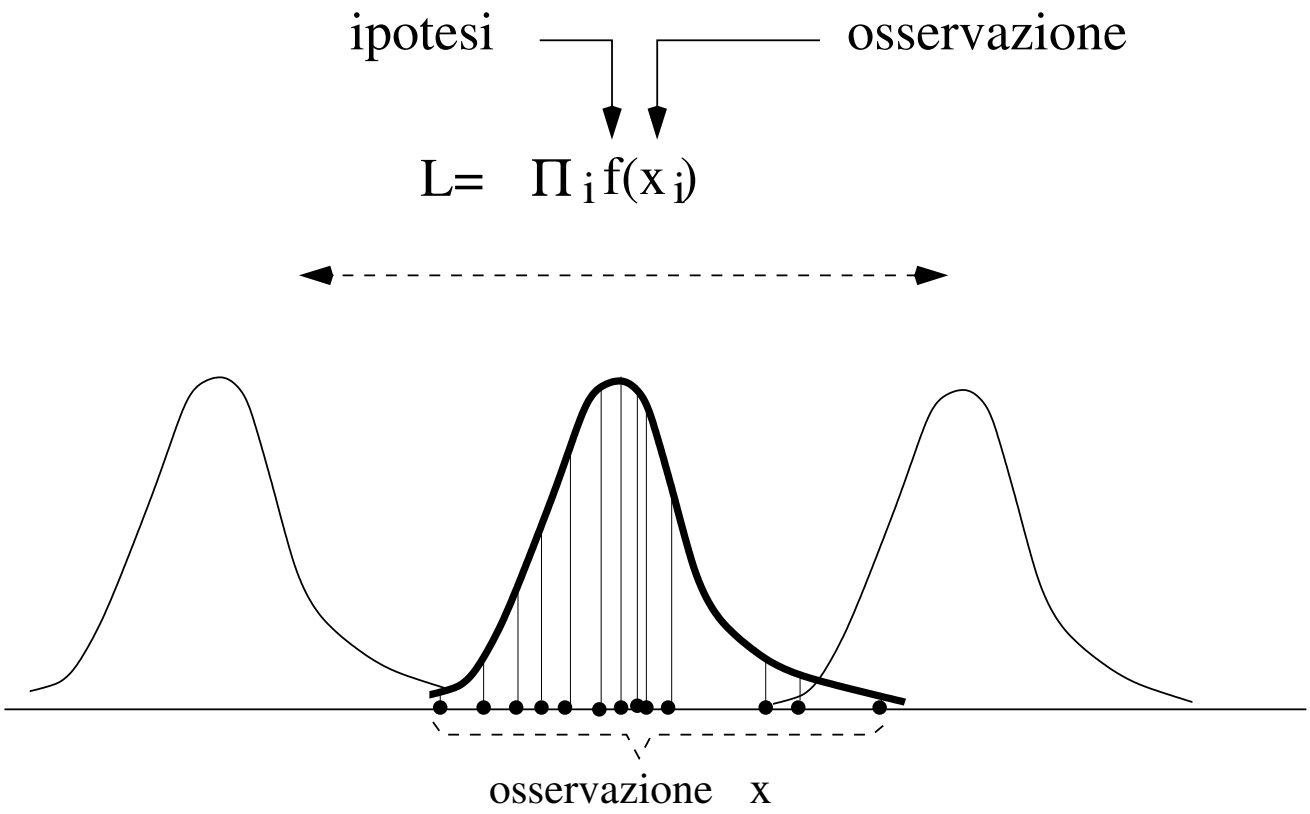
$$\frac{\partial L}{\partial \theta_k} = \frac{\partial \left[\prod_{i=1}^{n} p(\boldsymbol{x}_i; \boldsymbol{\theta}) \right]}{\partial \theta_k} = 0$$

or

$$\frac{\partial \mathcal{L}}{\partial \theta_k} = \sum_{i=1}^{n} \left[\frac{1}{p(\boldsymbol{x}_i; \boldsymbol{\theta})} \frac{\partial p(\boldsymbol{x}_i; \boldsymbol{\theta})}{\partial \theta_k} \right] = 0 \ , \quad (k = 1, 2, \ldots, p) \ .$$

- ***before the trial***, the likelihood function $L(\boldsymbol{\theta}; \underline{\boldsymbol{x}})$ is $\propto$ to the pdf of $(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots \boldsymbol{X}_n)$;
- ***before the trial***, the likelihood function $L(\boldsymbol{\theta}; \underline{\boldsymbol{X}})$ is a random function of X;
- ***frequentist view after the trial***: the likelihood function $L(\boldsymbol{\theta}; \underline{\boldsymbol{x}})$ depends on the values $\boldsymbol{\theta}$ only, representing the parameters (not the random variables) w.r.t. to maximize (or minimize)
- Bayesian view after the trial: $L(\boldsymbol{\theta}; \underline{\boldsymbol{x}})$ is the density function of the random variable $\boldsymbol{\Theta}$

ML estimators



The $p(x;\theta)$ form
is fitted to data
by maximizing
the ordinates of the observed data

Example

An urn with three marbles

$p = 1/3$ $\qquad$ $p = 2/3$

An experiment with 4 drawings:

$$p(x; n = 4, p) = \frac{4!}{x!(4-x)!} p^x (1-p)^{4-x}$$

	x=0	x=1	x=2	x=3	x=4
$p(x; 4, p = 1/3)$	16/81	32/81	24/81	8/81	1/81
$p(x; 4, p = 2/3)$	1/81	8/81	24/81	32/81	16/81

The likelihood estimate:

$\hat{p} = 1/3$ **if** $0 \leq x \leq 1$

$\hat{p} = 2/3$ **if** $3 \leq x \leq 4$

no maximum if $x = 2$

Example

In n trial x successes have been obtained. Make the ML estimate of p.

Binomial density

$$\mathcal{L} = -x \ln(p) - (n - x) \ln(1 - p) \ .$$

Minimum w.r.t. p:

$$\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}p} = -\frac{x}{p} + \frac{n - x}{1 - p} = 0 \implies \hat{p} = \frac{x}{n} = f$$

Make the ML estimate of p when x_1 successes on n_1 trials and x_2 successes on n_2 trials have been obtained.

Two binomials with the same p:

$$L = p^{x_1} \, p^{x_2} \, (1 - p)^{n_1 - x_1} \, (1 - p)^{n_2 - x_2} \ .$$

With logarithms:

$$\mathcal{L} = -(x_1 + x_2) \ln(p) - (n_1 - x_1 + n_2 - x_2) \ln(1 - p) \ ,$$

$$\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}p} = -\frac{x_1 + x_2}{p} + \frac{(n_1 + n_2) - x_1 - x_2}{1 - p} = 0$$

$$\implies \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Example

From the n values x_i of a Gaussian variable, find the ML estimate of mean and variance

Likelihood function:

$$L(\mu,\sigma) = \frac{1}{(\sqrt{2\pi}\,\sigma)^n}\, e^{-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2} \quad .$$

The log-likelihood:

$$\mathcal{L}(\mu,\sigma) = +\frac{n}{2}\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \ ,$$

Put the derivative =0:

$$\frac{\partial \mathcal{L}}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0$$

$$\Longrightarrow \quad \hat{\mu} = \sum_{i=1}^{n}\frac{x_i}{n} \equiv m$$

$$\frac{\partial \mathcal{L}}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

$$\Longrightarrow \quad \hat{\sigma^2} = \sum_{i=1}^{n}\frac{(x_i-m)^2}{n}$$

Estimators

- **Estimator of θ**
  If $\underline{\boldsymbol{X}}$ is a data sample with dimension n of a m-dimensional random variable $\boldsymbol{X}$ having $p(\boldsymbol{X};\theta)$ as a pdf, an estimator is a statistics

$$T_n(\underline{\boldsymbol{X}}) \equiv t_n(\underline{\boldsymbol{X}})$$

  for which $T : S \to \theta$.

- **Consistent estimator of θ**

$$\lim_{n\to\infty} P\left\{|T_n - \theta| < \epsilon\right\} = 1, \quad \forall\ \epsilon > 0\ .$$

- **Correct or unbiased estimator**

$$\langle T_n \rangle = \theta, \quad \forall\ n$$

- **The most efficient estimator**
  T_n is more efficient than Q_n if

$$\mathrm{Var}[T_n] < \mathrm{Var}[Q_n] \quad \forall\ \theta \in \Theta\ .$$

Theorems on $L(\theta; X)$

The mean value of the Score Function is zero:

$$\left\langle \frac{\partial}{\partial\theta} \ln p(\boldsymbol{X};\theta) \right\rangle = 0 \ .$$

The variance of the Score Function is the Fisher information:

$$\begin{aligned} \mathrm{Var}\left[\frac{\partial}{\partial\theta} \ln p(\boldsymbol{X};\theta)\right] &= \left\langle \left(\frac{\partial}{\partial\theta} \ln p(\boldsymbol{X};\theta) - \left\langle \frac{\partial}{\partial\theta} \ln p(\boldsymbol{X};\theta) \right\rangle \right)^2 \right\rangle \\ &= \left\langle \left(\frac{\partial}{\partial\theta} \ln p(\boldsymbol{X};\theta) \right)^2 \right\rangle \equiv I(\theta) \end{aligned}$$

These remarkable relations hold:

$$I(\theta) = \left\langle \left(\frac{\partial}{\partial\theta} \ln p(\boldsymbol{X};\theta) \right)^2 \right\rangle = -\left\langle \frac{\partial^2}{\partial\theta^2} \ln p(\boldsymbol{X};\theta) \right\rangle \ .$$

$$\left\langle \left(\frac{\partial}{\partial\theta} \ln L \right)^2 \right\rangle = \left\langle \left(\frac{\partial}{\partial\theta} \sum_i \ln p(\boldsymbol{X}_i;\theta) \right)^2 \right\rangle = n \left\langle \left(\frac{\partial}{\partial\theta} \ln p \right)^2 \right\rangle = nI(\theta) \ ,$$

The Cramér Rao theorem:
If T_n is an unbiased estimator

$$\mathrm{Var}[T_n] \geq \frac{1}{n \left\langle \left(\frac{\partial}{\partial\theta} \ln p(\boldsymbol{X};\theta)\right)^2 \right\rangle} = \frac{1}{nI(\theta)}$$

Binomial, Poisson, Gauss

$$\ln b(X;p) = \ln n! - \ln(n-X)! - \ln X! + X\ln p + (n-X)\ln(1-p)$$
$$\ln p(X;\mu) = X\ln\mu - \ln X! - \mu$$
$$\ln g(X;\mu,\sigma) = \ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2$$

These are random functions.

$$\begin{aligned}
\frac{\partial}{\partial p}\ln b(X;p) &= \frac{X}{p} - \frac{n-X}{1-p} = \frac{X-np}{p(1-p)} \\
\frac{\partial}{\partial \mu}\ln p(X;\mu) &= \frac{X}{\mu} - 1 = \frac{X-\mu}{\mu}\ , \\
\frac{\partial}{\partial \mu}\ln g(X;\mu,\sigma) &= -\frac{X-\mu}{\sigma}\left(-\frac{1}{\sigma}\right) = \frac{X-\mu}{\sigma^2}
\end{aligned}$$

according to $\left\langle \frac{\partial}{\partial\theta}\ln p(\boldsymbol{X};\theta)\right\rangle = 0$
Information:

$$\begin{aligned}
I(p) &= \frac{1}{p^2(1-p)^2}\left\langle (X-np)^2\right\rangle = \frac{np(1-p)}{p^2(1-p)^2} = \frac{n}{p(1-p)}\ , \\
I(\mu) &= \frac{1}{\mu^2}\left\langle (X-\mu)^2\right\rangle = \frac{\sigma^2}{\mu^2} = \frac{1}{\mu} = \frac{1}{\sigma^2}\ , \\
I(\mu) &= \frac{1}{\sigma^4}\left\langle (X-\mu)^2\right\rangle = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}\ ,
\end{aligned}$$

according to
$\mathrm{Var}[F] = \frac{1}{I(p)} = \frac{p(1-p)}{n}$, $\mathrm{Var}[M] = \frac{1}{nI(\mu)} = \frac{\sigma^2}{n}$

Golden results

1. If T_n is the best estimator of $\tau(\theta)$, it coincides with the ML estimator
   (if any)
   $$T_n = \tau(\hat{\theta}) \ .$$

2. the ML estimator is consistent

3. under broad conditions, the ML estimators are asymptotically normal.
   That is $(\theta - \hat{\boldsymbol{\theta}})$ is asymptotically normal with variance
   $$\frac{1}{nI(\theta)}$$

4. the score function $\partial \ln L / \partial \theta$ has zero mean, $nI(\theta)$ variance and is asymptotically normal

5. the variable
   $$2[\ln L(\hat{\boldsymbol{\theta}}) - \ln L(\boldsymbol{\theta})]$$
   tends asymptotically to $\chi^2(p)$, where p is the dimension of θ

6. In the case of gaussian variables the ML principle coincides with the Least Square method (χ^2 minimization).

Image Likelihood

The objective function to be minimized is

$$-F(\boldsymbol{\mu}) = -2\ln L(\boldsymbol{n}|\boldsymbol{\mu}) - \alpha\, C(\boldsymbol{\mu}) + \lambda(n_T - \sum_i n_i) \quad (17)$$

$$L(\boldsymbol{n}|\boldsymbol{\mu}) = \prod_{ij} \frac{\nu_{ij}^{n_{ij}}}{n_{ij}!}\, \mathrm{e}^{-\nu_{ij}}$$

By dropping the constant terms:

$$\ln L(\boldsymbol{n}|\boldsymbol{\mu}) = \sum_{ij} n_{ij} \ln \nu_{ij} - n u_{ij}$$

$$\nu_{ij} = n_T \sum_{i',j'} P_v(\nu_{ij}|\mu_{i'j'})\, P_\mu(\mu_{i'j'}) \quad (18)$$

In the case of a two dimensional gaussian point spread function PSF:

$$P_v(\nu_{ij}|\mu_{i'j'}) = \frac{1}{2\pi\,\sigma_x\sigma_y} \exp\left[-\frac{(x_{ij}-x_{i'j'})^2}{2\,\sigma_x^2} - \frac{(y_{ij}-y_{i'j'})^2}{2\,\sigma_y^2}\right], \quad (19)$$

The free parameters are the true image pixels μ_{ij}

If there is also the detector read-out noise (CCD cameras):

$$L(\boldsymbol{n}|\boldsymbol{\mu}) = \prod_{ij} \sum_m \frac{1}{\sqrt{2\pi}\sigma}\, \mathrm{e}^{(n_m - k_{ij})^2/\sigma^2}\, \frac{\nu_{ij}^{k_{ij}}}{k_{ij}!}\, \mathrm{e}^{-\nu_{ij}}$$

$$\ln L(\boldsymbol{n}|\boldsymbol{\mu}) = \sum_{ij} \left[-\ln\sqrt{2\pi}\sigma - \nu_{ij} + \ln \sum_m \left(\mathrm{e}^{(n_m - k_{ij})^2/\sigma^2}\, \frac{\nu_{ij}^{k_{ij}}}{k_{ij}!}\right)\right]$$

Regularization terms

The objective function to be minimized is

$$-F(\boldsymbol{\mu}) = -2\ln L(\boldsymbol{n}|\boldsymbol{\mu}) - \alpha\, C(\boldsymbol{\mu}) + \lambda(n_T - \sum_i \nu_i) \quad (20)$$

$$\mu_j = \mu_{\rm tot} p_j = \mu_{\rm tot} \int_{\rm bin\,j} f_t(y)\,{\rm d}y$$

where $\alpha > 0$. Some regularization terms:

- **minimum second derivative (Tichonov)**

$$C(\boldsymbol{\mu}) = -\int [f_t''(y)]^2\,{\rm d}y \simeq -\sum_{i=1}^{M-2} [-\mu_i + 2\mu_{i+1} - \mu_{i+2}]^2$$

- **minimum variance:**

$$C(\boldsymbol{\mu}) = -{\rm Var}[\boldsymbol{\mu}] = -\sum_i \mu_i^2$$

- **maximum entropy (MaxEnt)**

$$C(\boldsymbol{\mu}) = -\sum_i p_i \ln p_i = -\sum_i \frac{\mu_i}{\mu_T} \ln \frac{\mu_i}{\mu_T}$$

- **cross-entropy**

$$C(\boldsymbol{\mu}) = -\sum_i p_i \ln \frac{p_i}{q_i} = -\sum_i \frac{\mu_i}{\mu_T} \ln \frac{\mu_i}{\mu_T\, q_i}$$

  where $\boldsymbol{q} = (q_1, q_2, \ldots, q_n)$ is the most likely a priori shape for the true distribution μ_i.

The estimators $\hat{\mu}$

The estimators $\hat{\mu}$ of the true image are functions of the data. They are found by minimizing the function $-F(\boldsymbol{\mu})$ from (24). We have $N_{\text{bins}}+1$ equations:

$$\frac{\partial F}{\partial \mu_i}=0\ ,\quad i=1,2,\ldots,N_{\text{bins}}\ ;\quad \frac{\partial F}{\partial \lambda}=0\ ,\quad i=N_{\text{bins}}+1=M$$

Considering $\mu_M\equiv\lambda$ and by expanding around solutions μ, μ' where $F(\mu,\lambda,n)=0$, $F(\mu',\lambda',n')=0$:

$$\sum_j^M \frac{\partial^2 F_i}{\partial\mu_i\partial\mu_j}(\mu_j-\mu_i')+\sum_j^{N_{\text{bins}}}\frac{\partial^2 F_i}{\partial n_i\partial n_j}(n_j-n_i')=A_{ij}(\mu_j-\mu_i')+B_{ij}(n_j-n_i')=0$$

Solving for μ

$$\hat{\boldsymbol{\mu}}=\boldsymbol{\mu}-A^{-1}B(\boldsymbol{n}-\boldsymbol{n'})\equiv\boldsymbol{\mu}+G(\boldsymbol{n}-\boldsymbol{n'}) \tag{21}$$

One obtains variances

$$\operatorname{cov}[\hat{\mu}_i,\hat{\mu}_j]=\sum_{kl}^{N_{\text{bins}}}\frac{\partial\hat{\mu}_i}{\partial n_k}\frac{\partial\hat{\mu}_j}{\partial n_l}\operatorname{cov}[n_k,n_l] \tag{22}$$

and bias from (21):

$$b_i=\langle\hat{\mu}_i\rangle-\mu_i=G(\boldsymbol{n}-\boldsymbol{n'})\simeq\sum_j^{N_{\text{bins}}}G_{ij}(\nu_j-n_j) \tag{23}$$

Regularization paramter

The objective function to be minimized is

$$-F(\boldsymbol{\mu}) = -2\ln L(\boldsymbol{n}|\boldsymbol{\mu}) - \alpha\, C(\boldsymbol{\mu}) + \lambda(n_T - \sum_i \nu_i) \quad (24)$$

The choice of the regularization parameter $\alpha > 0$ is difficult. Some choices are:

- **Bayesian**

  $$\alpha = \frac{1}{\mu_T}$$

  usually too much smoothing

- **α is determined such as (zeroth order regularization)**

  $$2\ln L = \chi^2 \simeq N_{\text{bins}}$$

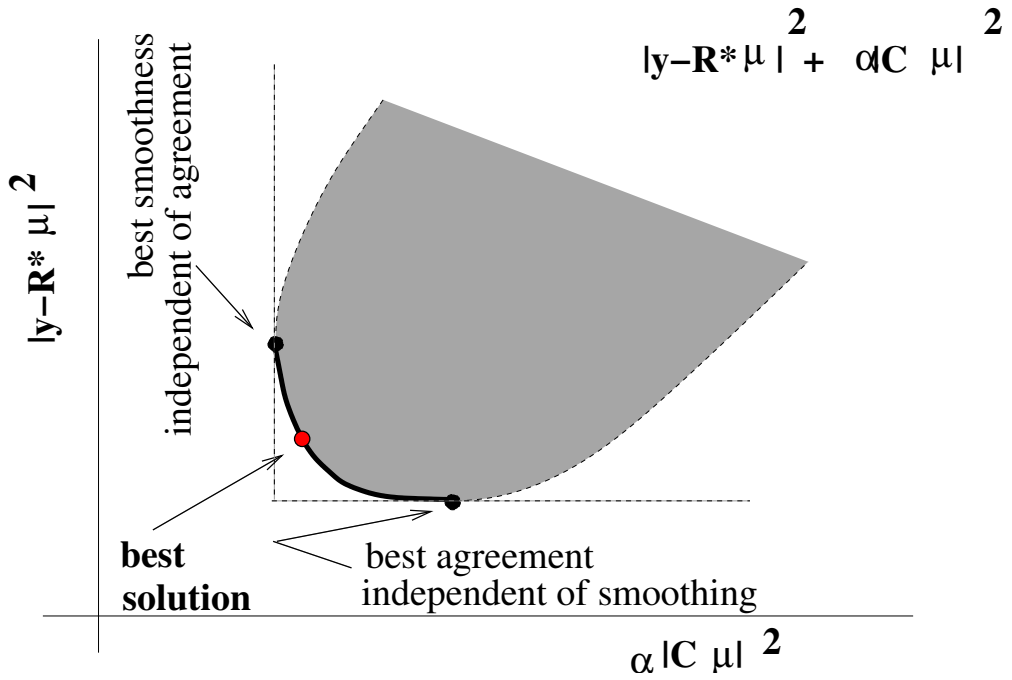
- **minimum Mean Square Error (MSE)**

  $$\text{MSE} = \frac{1}{N_{\text{bins}}} \sum_i^{N_{\text{bins}}} \text{Var}[\mu_i] + b_i^2$$

- **minimum relative Mean Square Error**

  $$\text{MSE} = \frac{1}{N_{\text{bins}}} \sum_i^{N_{\text{bins}}} \frac{\text{Var}[\mu_i] + b_i^2}{\mu_i}$$

Regularization paramter



With gaussian variables often we minimize

$$-F = ||y - R * \mu||^2 + \alpha||C\mu||^2$$

and the solution moves along the bold line.

In this case

$$A = \chi^2 + \alpha||C\mu||^2 > 0$$

we can regularize the solution by choosing $\chi^2 \simeq DoF \equiv$ number of pixel N, with the condition:

$$\alpha = \frac{A - N}{||C\mu||^2}$$

Regularization parameter

Two-peak deconvolution with the regularized ML methods (Glen Cowan, Statistical Data Analysis, Oxford (2000))

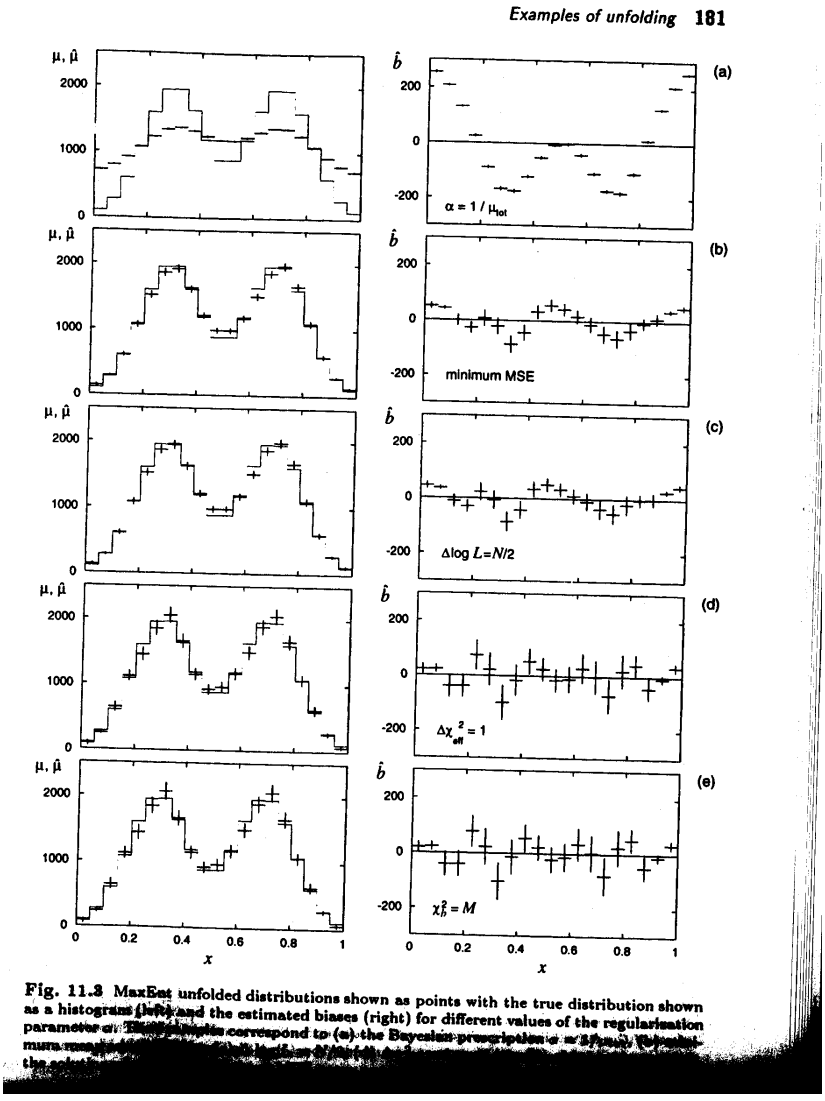
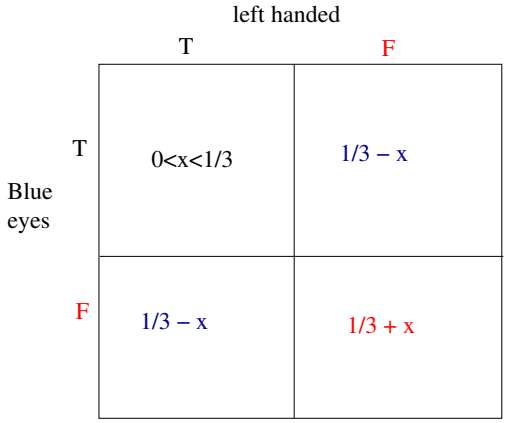


Fig. 11.3 MaxEnt unfolded distributions shown as points with the true distribution shown as a histogram (left) and the estimated biases (right) for different values of the regularization parameter α. The examples correspond to (a) the Bayesian prescription $\alpha = 1/\mu_{tot}$, (b) the minimum mean square, (c) $\Delta \log L = N/2$, (d) $\Delta\chi^2_{eff} = 1$, and (e) $\chi^2_b = M$. The data are the same as in Fig. 11.2.

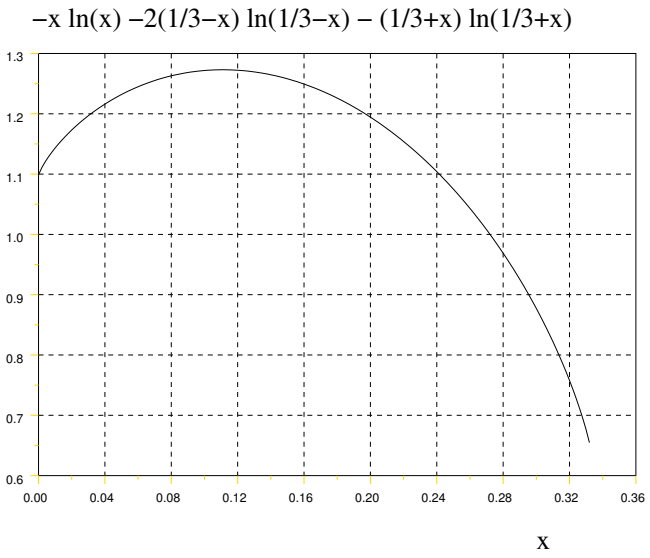
Entropy

The maximum entropy (MaxEnt) principle favours flat configurations

Example: 1/3 of Kangoos have blue eyes, 1/3 are left handed



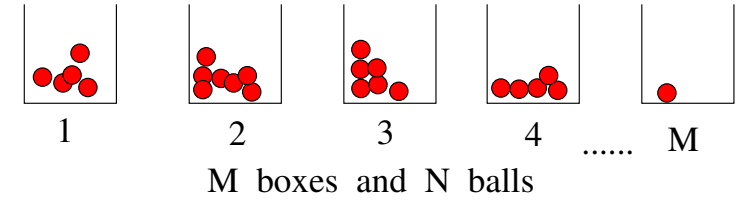
$S = -\sum p_i \ln p_i$. **MaxEnt gives** $x = 1/9$



Entropy: the monkey problem

We have M boxes and a monkey that throws N balls randomly into them.
What is the box-balls configuration of highest probability? Probability of a configuration:

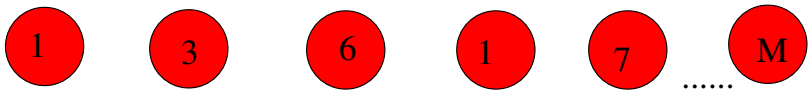
$$p = \frac{1}{M^N} \frac{N!}{n_1! n_2! \dots n_N!}$$



M boxes and N balls

equal to

N balls labelled randomly from 1 to M



$$\ln p = -N \ln M + \ln N! - \sum_i \ln(n_i!)$$

Stirling formula: $n! = \sqrt{2\pi n}\, n^n\, \mathrm{e}^{-n} \rightarrow \ln n! \simeq n \ln n - n$

$$\ln p = -N \ln M + N \ln N - N + \sum n_i - \sum_i n_i \ln n_i$$

$$\sum p_i = 1 \ , \qquad p_i = \frac{n_i}{N} \ , \ \ln p = -N \ln M - N \sum_i p_i \ln p_i$$

The most probable configuration means to maximize

$$S = -\sum_i p_i \ln p_i$$

The Metropolis algorithm

(N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller and E. Teller, 1953)
Useful for producing a sample of a random variable, according to

$$p(\boldsymbol{x}) = \frac{h(\boldsymbol{x})}{Z} \ ,$$

when other methods are impractical. When $Z = \int h(\boldsymbol{x})\, d\boldsymbol{x}$ unknown or very difficult to calculate. The algorithm is based on the following idea:
a random value x from $p(x)$ can migrate to another one y according to an arbitrary transition function $t(\boldsymbol{x} \to \boldsymbol{y})$. A sufficient condition for x to be distributed as $p(x)$ is the equilibrium condition assured by the detailed balance:

$$p(\boldsymbol{x})\, t(\boldsymbol{x} \to \boldsymbol{y}) = p(\boldsymbol{y})\, t(\boldsymbol{y} \to \boldsymbol{x}) \ , \tag{25}$$

where

$$t(\boldsymbol{x} \to \boldsymbol{y}) = q(\boldsymbol{x}, \boldsymbol{y})\, \alpha(\boldsymbol{x}, \boldsymbol{y}) \ . \tag{26}$$

where q is arbitrary and α is the transition probability
The algorithm:

- **generate $\boldsymbol{y}$ from $q(\boldsymbol{x}^{(i)}, \cdot)$**
- **calculate**

$$\alpha(\boldsymbol{x}^{(i)}, \boldsymbol{y}) = \min \left\{ 1, \frac{h(\boldsymbol{y})\, q(\boldsymbol{y}, \boldsymbol{x}^{(i)})}{h(\boldsymbol{x}^{(i)})\, q(\boldsymbol{x}^{(i)}, \boldsymbol{y})} \right\} \tag{27}$$

- **if $\texttt{random} \leq \alpha(\boldsymbol{x}^{(i)}, \boldsymbol{y})$, then $\boldsymbol{x}^{(i+1)} = \boldsymbol{y}$, otherwise set $\boldsymbol{x}^{(i+1)} = \boldsymbol{x}^{(i)}$.**

The Metropolis algorithm

The Metropolis sampled values follow asymptotically the density $p(x)$ because they satisfy to the detailed balance:

$$p(\boldsymbol{x})\, q(\boldsymbol{x},\boldsymbol{y})\, \alpha(\boldsymbol{x},\boldsymbol{y}) = p(\boldsymbol{y})\, q(\boldsymbol{y},\boldsymbol{x})\, \alpha(\boldsymbol{y},\boldsymbol{x}) \ ,$$

Indeed, from

$$\frac{p(\boldsymbol{y})\, q(\boldsymbol{y},\boldsymbol{x})}{[p(\boldsymbol{x})\, q(\boldsymbol{x},\boldsymbol{y})]} < 1$$

it follows asymptotically

$$\alpha(\boldsymbol{x},\boldsymbol{y}) = \frac{p(\boldsymbol{y}) q(\boldsymbol{y},\boldsymbol{x})}{[p(\boldsymbol{x})\, q(\boldsymbol{x},\boldsymbol{y})]}$$

when this condition is valid, the inverse Metropolis probability is $\alpha(\boldsymbol{y},\boldsymbol{x}) = 1$ and one obtains the identity

$$p(\boldsymbol{x})\, q(\boldsymbol{x},\boldsymbol{y})\, \frac{p(\boldsymbol{y})\, q(\boldsymbol{y},\boldsymbol{x})}{p(\boldsymbol{x})\, q(\boldsymbol{x},\boldsymbol{y})} = p(\boldsymbol{y})\, q(\boldsymbol{y},\boldsymbol{x}) \ .$$

Theorem: If $q(\boldsymbol{x},\boldsymbol{y}) > 0$ for every $\boldsymbol{x}$ and $\boldsymbol{y}$ belonging to the $\boldsymbol{X}$ spectrum, then

$$\lim_{N\to\infty} \frac{\sum_{i=1}^{N} g(\boldsymbol{x}^{(i)})}{N} = \langle g(\boldsymbol{X})\rangle = \begin{cases} \sum_{\boldsymbol{x}} g(\boldsymbol{x}) p(\boldsymbol{x}), & \text{discrete case;} \\ \int g(\boldsymbol{x}) p(\boldsymbol{x})\, d\boldsymbol{x}, & \text{continuum case.} \end{cases} \quad (28)$$

The Metropolis algorithm

Often one sets $q(x,y) \equiv U(a,b)$, so that
The algorithm:

- **generate $\boldsymbol{y}$ uniformly in (a,b)**
- **calculate**

$$\alpha(\boldsymbol{x}^{(i)}, \boldsymbol{y}) = \min\left\{1, \frac{p(\boldsymbol{y})}{p(\boldsymbol{x}^{(i)})}\right\} \qquad (29)$$

- **if** `random`$\leq \alpha(\boldsymbol{x}^{(i)}, \boldsymbol{y})$**, then $\boldsymbol{x}^{(i+1)} = \boldsymbol{y}$, otherwise set $\boldsymbol{x}^{(i+1)} = \boldsymbol{x}^{(i)}$.**

For gaussian variables:

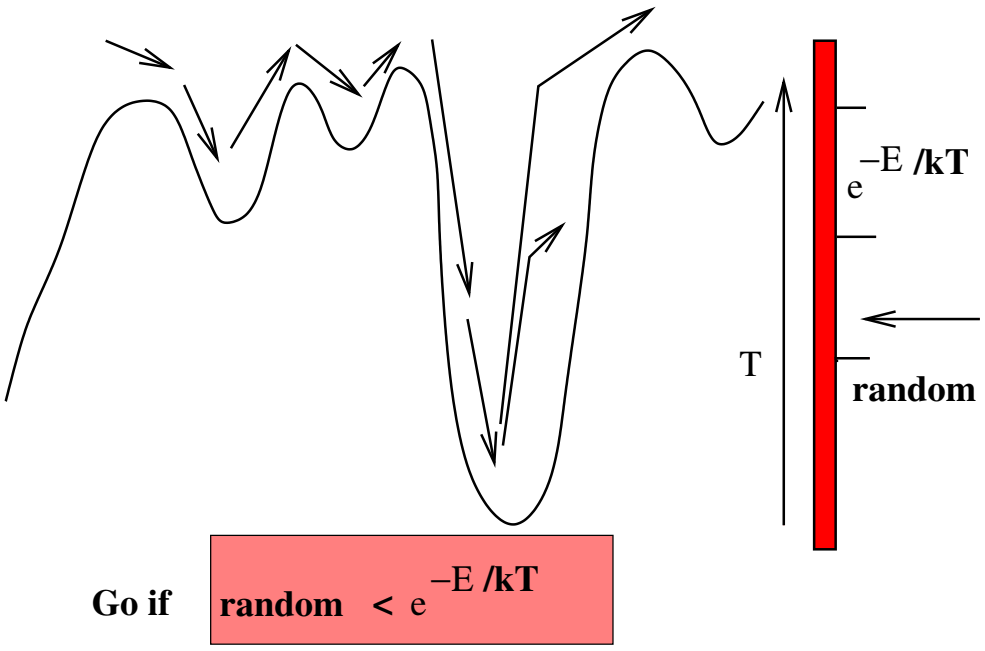
```
// ciclo di Metropolis
for k=2:N,
  i(k)=k;
  //campiona in +- ks sigma intorno alla media
  y= mu-ks*sigma + 2*ks*sigma*grand(1,1,'def');
  // rapporto di Metropolis tra gaussiane
  alpha= exp( -0.5*( (y-mu)^2 - (x(k-1)-mu)^2 )/sigma^2 );
  u= grand(1,1,'unf',0,1);
  x(k) = x(k-1);
  if(u<alpha) then x(k)=y; end;
  sumk = sumk+x(k);
  sumk2= sumk2+x(k)^2;
  plotk(k)  = sumk/k;                  // Metropolis per la media....
  plotk2(k) = sqrt(sumk2/k - plotk(k)^2);   // e deviazione standard
  xbasc();
  plot2d(i,plotk);   // display corrente sulla media
end;
```

Simulated annealing

This algorithm is inspired to the physical process of annealing, in which a metal acquires a stable energy configuration as the result of many Boltzmann statistical fluctuations.
This algorithm minimizes a function (seen as the energy of the system) having care to avoid local minima, thank to the "boiling" possibility offered by the Boltzmann function.

$$\mathrm{e}^{-E/kT}$$

In our case $E \to \chi^2$ or $E \to -2 \ln L(\mu)$



Simulated annealing

To minimize χ^2, at each iteration we have to consider the previous chi-square χ^2_1 and the actual one χ^2_2 and to follow the rule:

a) if $\chi^2_2 < \chi^2_1$ retain the last NP_{ij} function as the best one and update: $\chi^2_{\rm min} = \chi^2_2$;

b) if $\chi^2_2 > \chi^2_1$ calculate the Boltzmann probability

$$P(\Delta\chi^2) = \mathrm{e}^{-(\chi^2_2 - \chi^2_1)/kT} < 1 \qquad (30)$$

where kT is a parameter to be chosen (see below); retain the last NP_{ij} function as the best one and update the chi-square $\chi^2_{\rm min} = \chi^2_2$ only if a random number $0 < \mathcal{R} < 1$ is such as

$$\mathcal{R} < P(\Delta\chi^2)$$

In this step very bad χ^2 are sometime accepted (with low probability), thus allowing the possibility to "escape" from local minima (simulated annealing). Steps:

1) set configuration space (images)

2) decide the random choice

3) choose the objective function (E, χ^2, $2\ln L$)

4) manage control parameter T

Simulated annealing
A statistical interpretation

One can show that when

$$\alpha(\boldsymbol{x}^{(i)}, \boldsymbol{y}) = \min\left\{1, \frac{p(\boldsymbol{y})}{p(\boldsymbol{x}^{(i)})}\right\}$$

the sampled values from Metropolis are asymptotically distributed as $p(\boldsymbol{x})$
Since the simulated annealing uses the acceptation function

$$\frac{\mathrm{e}^{-\chi_2^2/kT}}{\mathrm{e}^{-\chi_1^2/kT}}$$

this is equivalent to consider

$$\mathrm{e}^{-0.5\sum_{ij}(\hat{\mu}_{ij}-\mu_{ij})^2/kT}$$

a gaussian sample and to find the best estimate of μ

Simulated annealing
Temperature managing

As a practical rule, the starting temperature should be chosen to allow about 80% of bad chi-square transitions. If K is the number of bad chi-square transitions, one sets

$$\frac{1}{K}\sum^{K}\exp\left[-\Delta\chi^2/kT\right] \simeq 0.8 \ .$$

With a Taylor expansion one obtains:

$$kT \simeq 5\left\langle\Delta\chi^2\right\rangle \ . \tag{31}$$

One sets $kT = 1$, calculates the average of the positive $\Delta\chi^2 = \chi^2_2 - \chi^2_1$ transitions and then applies (31).

Cooling schedule

- **the initial temperature is chosen by (31);**
- **the temperature is maintained until the improvement in the $\chi^2_{\rm min}$ is negligible. For example:**

$$\left|1 - \frac{\chi^2_{\rm min}}{\langle\chi^2_{\rm min}\rangle}\right| < 0.05 \tag{32}$$

  where the average is made on a fixed number of latest iterations (i.e 10,20)

- **then, the temperature is changed**

$$kT \to 0.95\ kT$$

  and a new cycle is initiated. The final stop is decided when, after a temperature changing, no improvements in χ^2 are detected. For example:

$$\left|1 - \frac{\chi^2_{\rm min}(kT_{\rm last})}{\chi^2_{\rm min}(kT_{\rm prev})}\right| < 0.05 \tag{33}$$

An example

Cat image with the regularized ML method with the ME principle (Cowan, 2000)

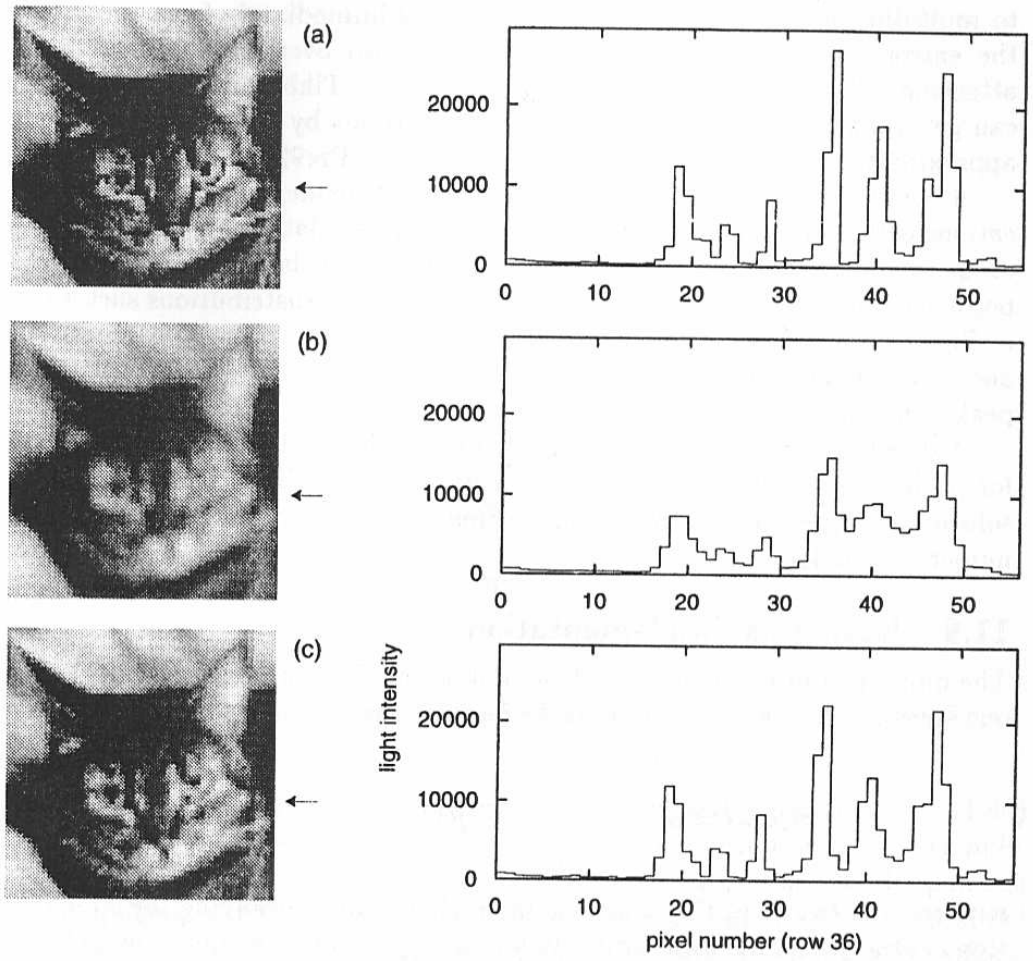
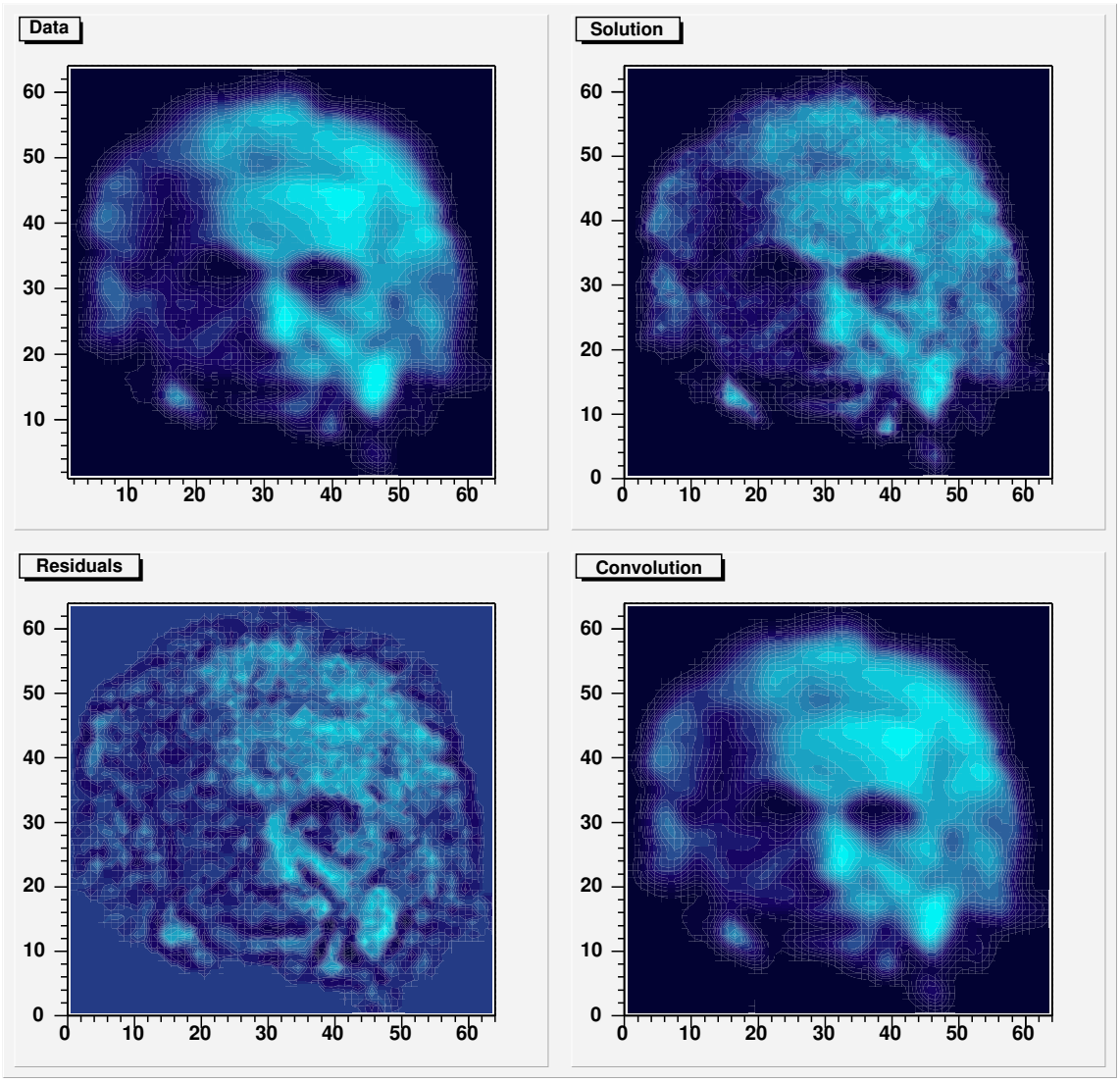


Fig. 11.5 (a) The original 'true' image μ. (b) The observed image **n**, blurred with a Gaussian point spread function with a standard deviation equal to 60% of the pixel size. (c) The maximum entropy unfolded image. The histograms to the right show the light intensity in pixel row 36 (indicated by arrows).

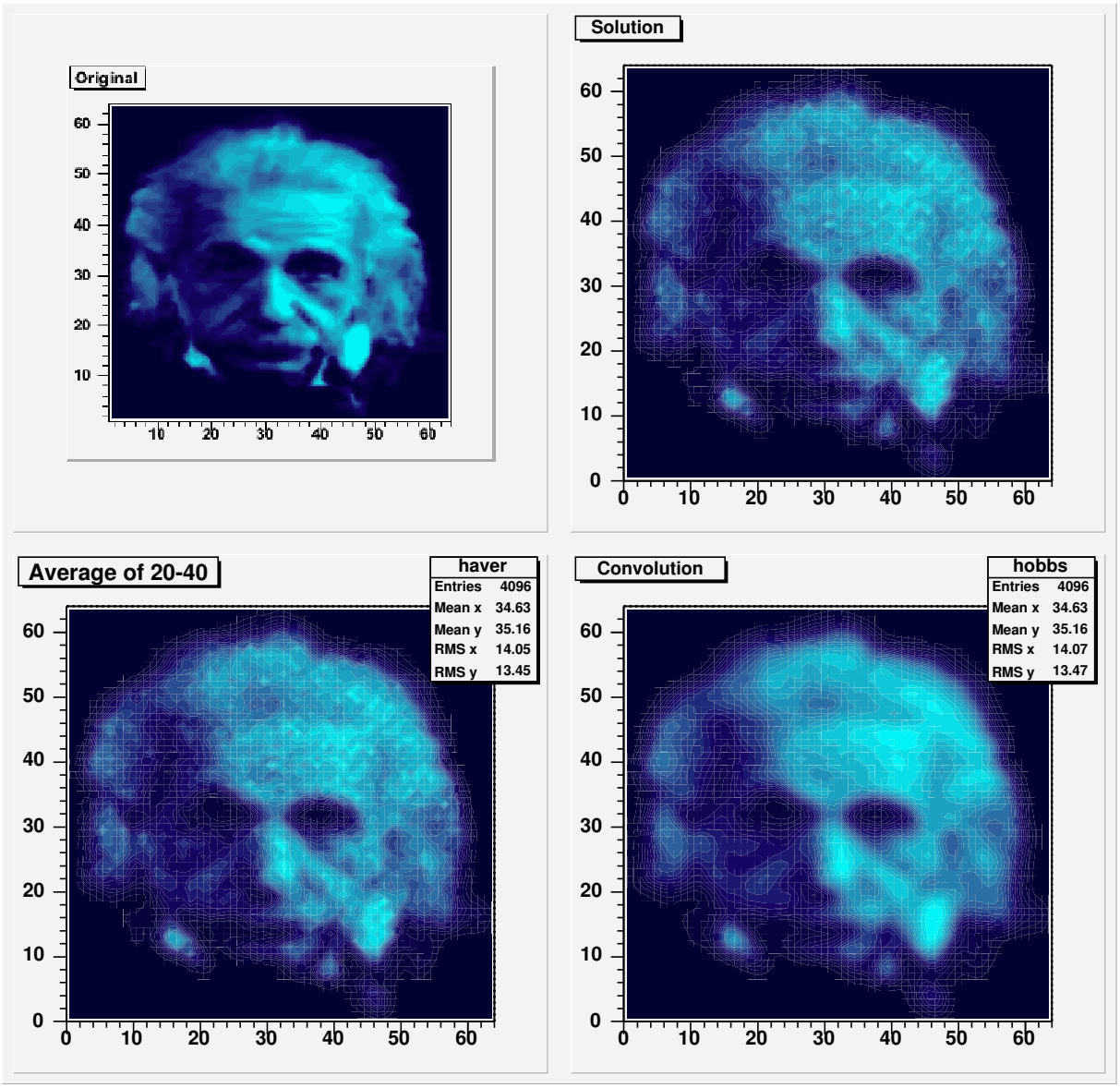
A personal example

ML method regularized with $\sum_i \mu_i^2$
Ratio $L(\boldsymbol{\mu})/\alpha \sum_i \mu_i^2 \simeq 1/30$. χ^2/number of pixels$\simeq 3308/3015$.
Minimization with Metropolis-Hastings algorithm with $q(y, x_i) = \mu_i/\mu_{\max}$, 10 pixels at a time, 12000 iterations.



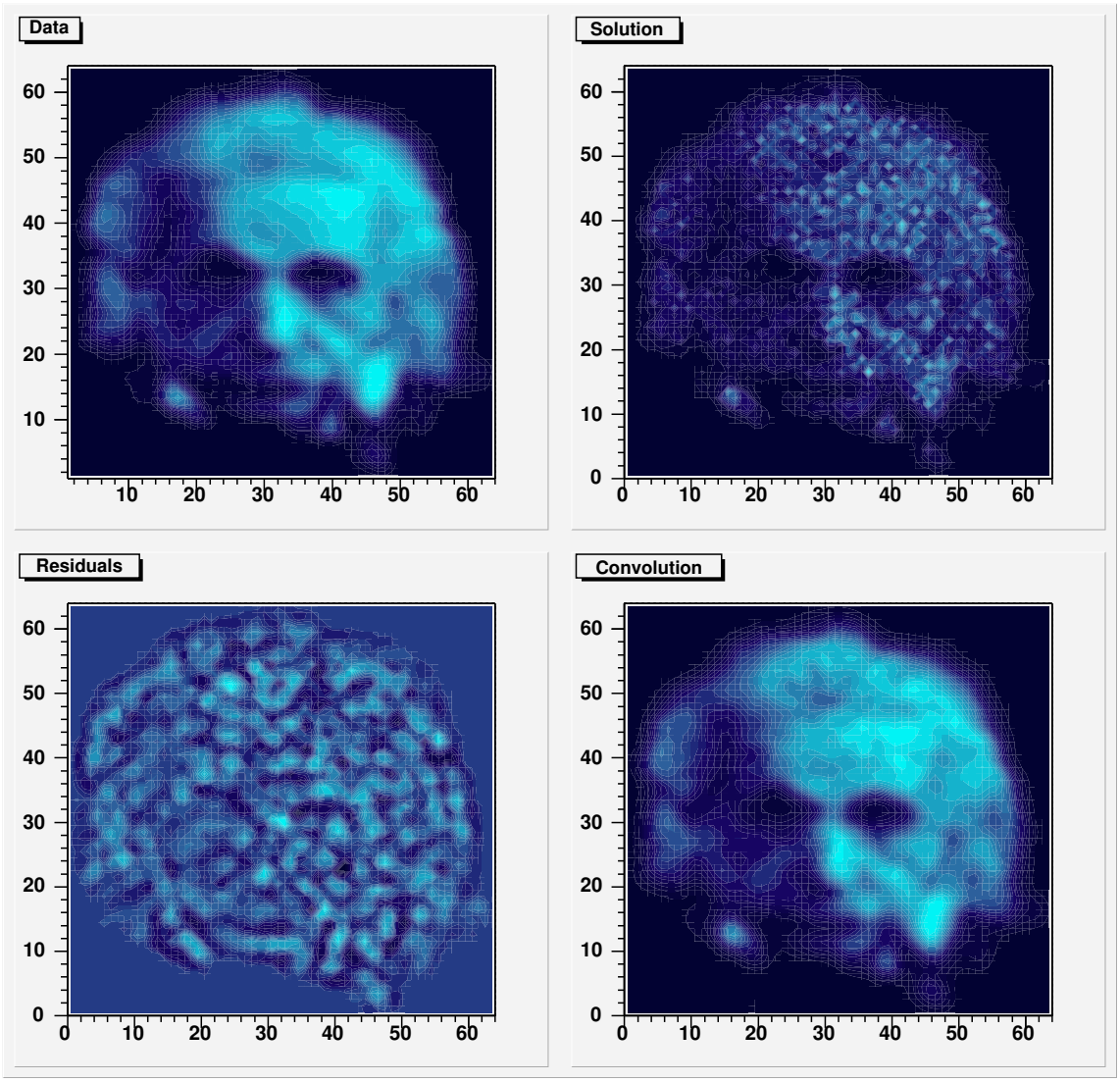
A personal example

The same enhanced



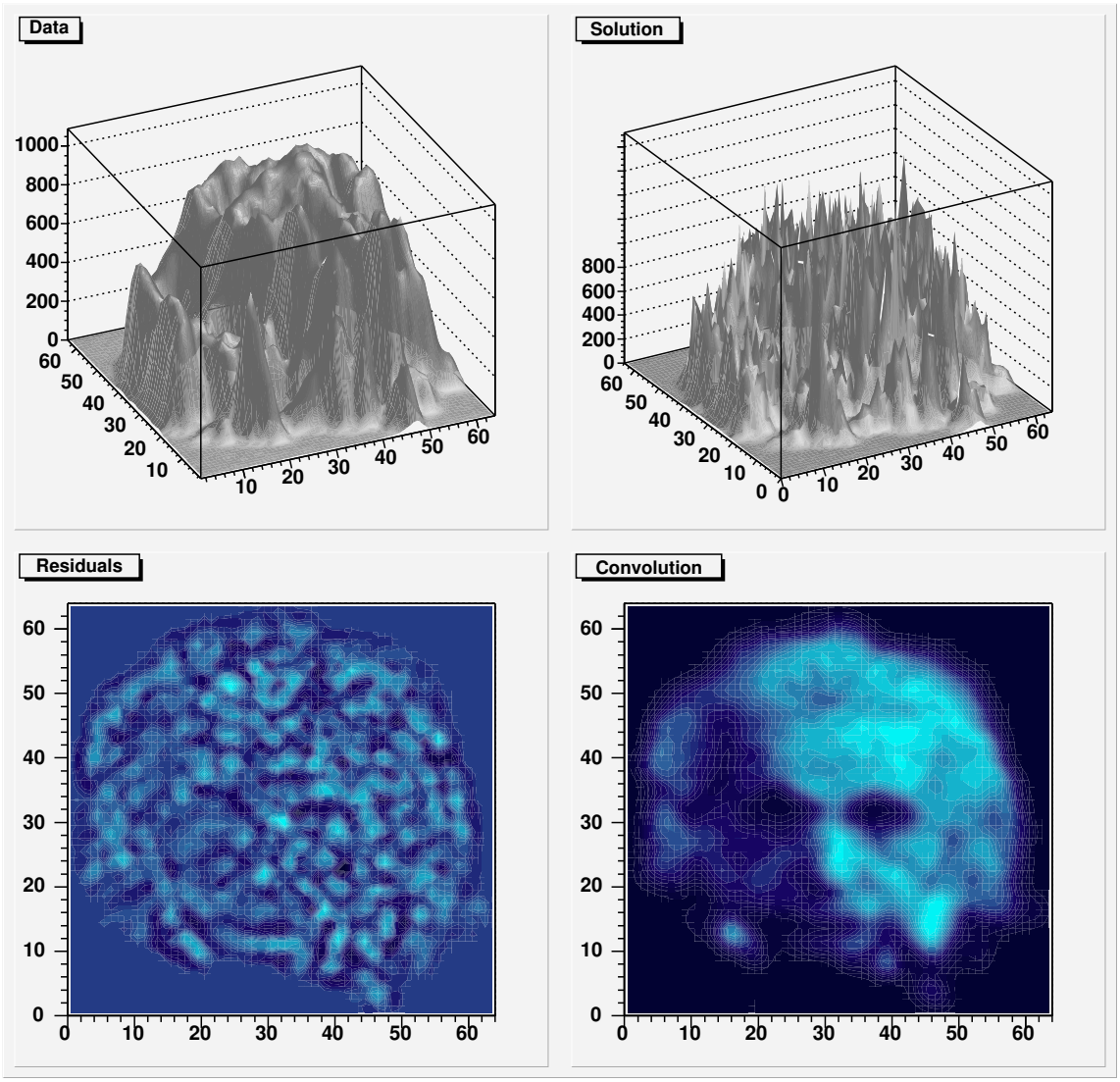
A personal example

The same without regularization



A personal example

The same without regularization



A new approach: iteration techniques

$$N_{ij}(\text{th}) = NP_{ij}(\text{obs}) = N \sum_{i'j'} P_{i'j'}(\text{true})\, P_v(\text{obs}_{ij}|\text{true}_{i'j'}) \quad (34)$$

We write this equation considering the operator R:

$$n = R * \mu$$

The iterative method (Van Cittert 1930) adds with a weight β the residual r_k to the current solution

$$\mu_{k+1} = \mu_k + \beta[n - R * \mu_k] \quad (35)$$

The method is based on the known equation

$$\sum_{i=0}^{k} q^n = \frac{1 - q^{k+1}}{1 - q} \quad (36)$$

for $k \to \infty$ the series converges if $|q < 1|$.
By applying iteratively (37):

$$\begin{aligned}
\mu_{k+1} &= \beta n + (1 - \beta R)\mu_k = \beta n + (1 - \beta R)(\beta n + (1 - \beta R)\mu_{k-1} \\
&= \beta n + \beta(1 - \beta R)n + (1 - \beta R)^2 \mu_{k-1} \\
&= \beta n + \beta(1 - \beta R)n + \beta(1 - \beta R)^2 n + (1 - \beta R)^3 \mu_{k-2} \ldots \\
&= \sum_{i=0}^{k} \beta(1 - \beta R)^i n \ .
\end{aligned}$$

From (36):

$$\mu_{k+1} = \frac{I - (I - \beta R)^{k+1}}{\beta R} \beta n \to R^{-1} n = \mu \ , \quad \text{for} \quad k \to \infty \ .$$

if $|I - \beta R| < 1$

Iteration techniques

In summary, we use

$$\mu_{k+1} = \mu_k + \beta[n - R * \mu_k] \tag{37}$$

with the initial condition

$$\mu_0 = n \ .$$

The convergence is assured if

$$|I - \beta R| < 1$$

Since $|1 - \beta x| < 1$ **implies** $0 < \beta < 2/x$**, in the case of the operator R, which can be transformed in a square matrix**

$$R' = (R * \mu)\, \mu^{-1}$$

we obtain the condition:

$$0 < \beta < \frac{2}{\text{max eigenvalue of R}'}$$

Note that we works always with square matrices $R * \mu$**,** μ**,** μ^{-1} **and** R'**.**

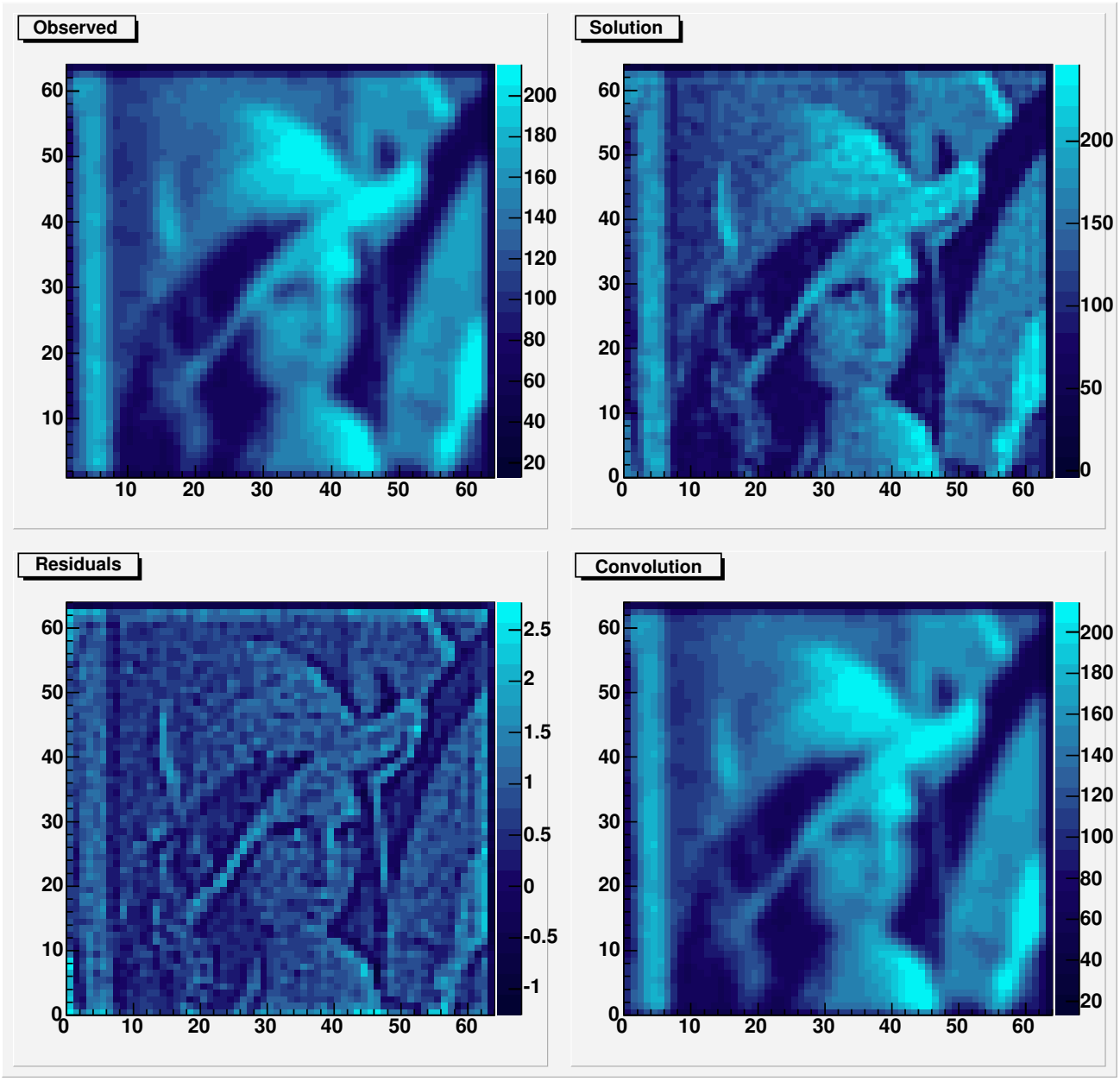
However, this step must be repeated at each iteration

This method sometimes gives spectacular result!

However, often it gives irregular solutions.

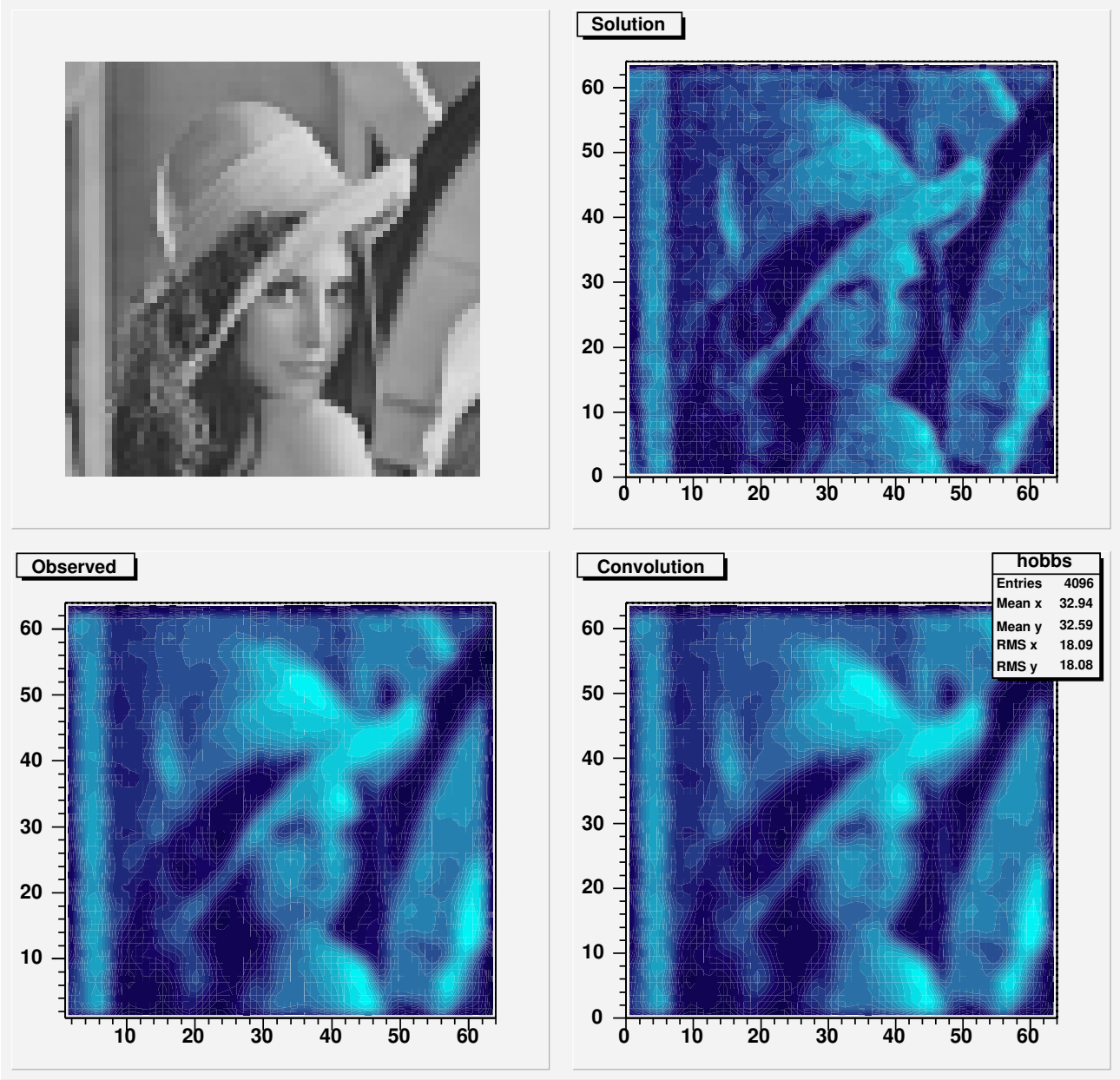
Iterative good solution

Without β $\quad \mu_{k+1} = \mu_k + [n - R * \mu_k]$



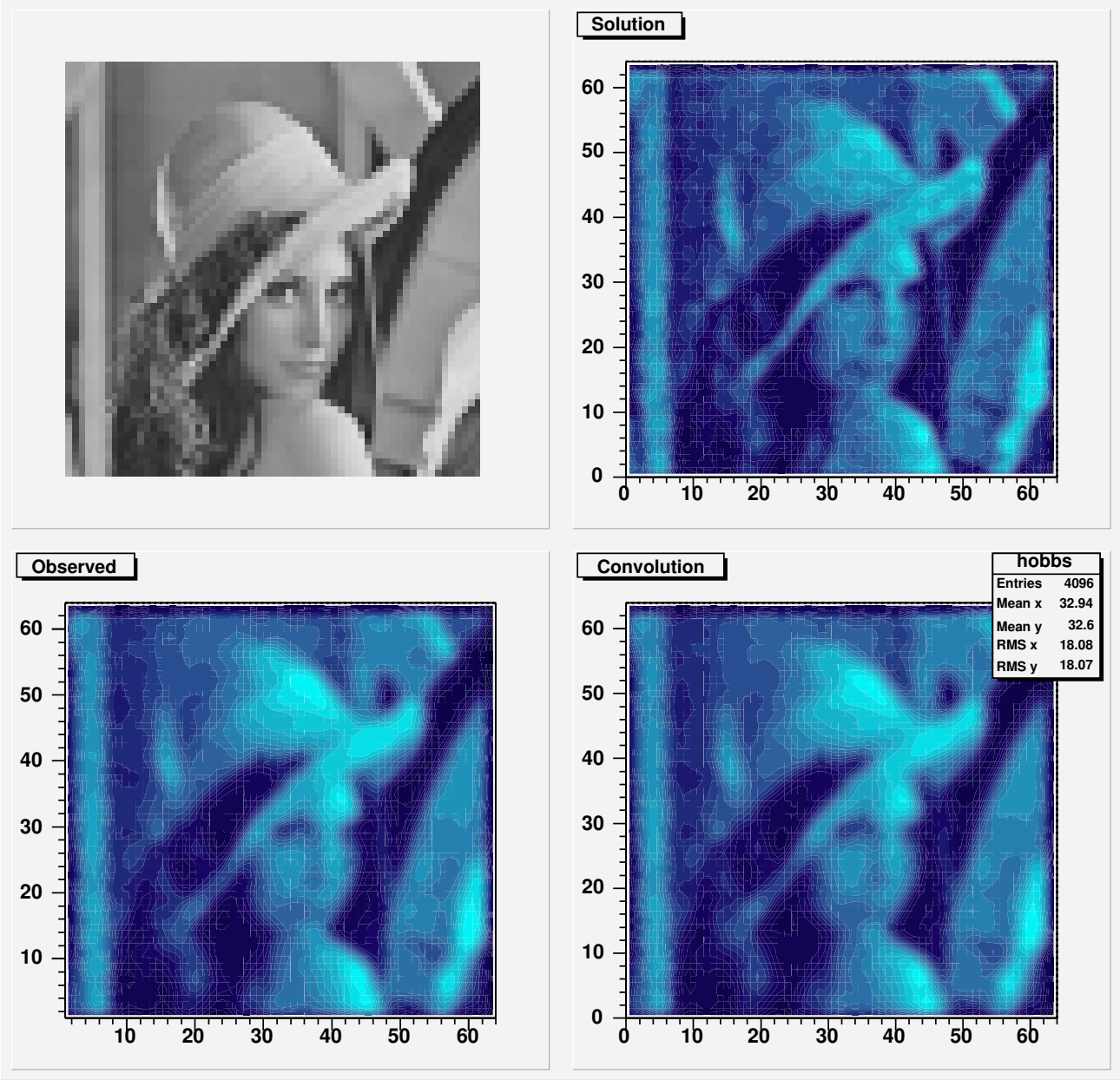
Iterative good solution

Without β but enhanced $\mu_{k+1} = \mu_k + [n - R * \mu_k]$



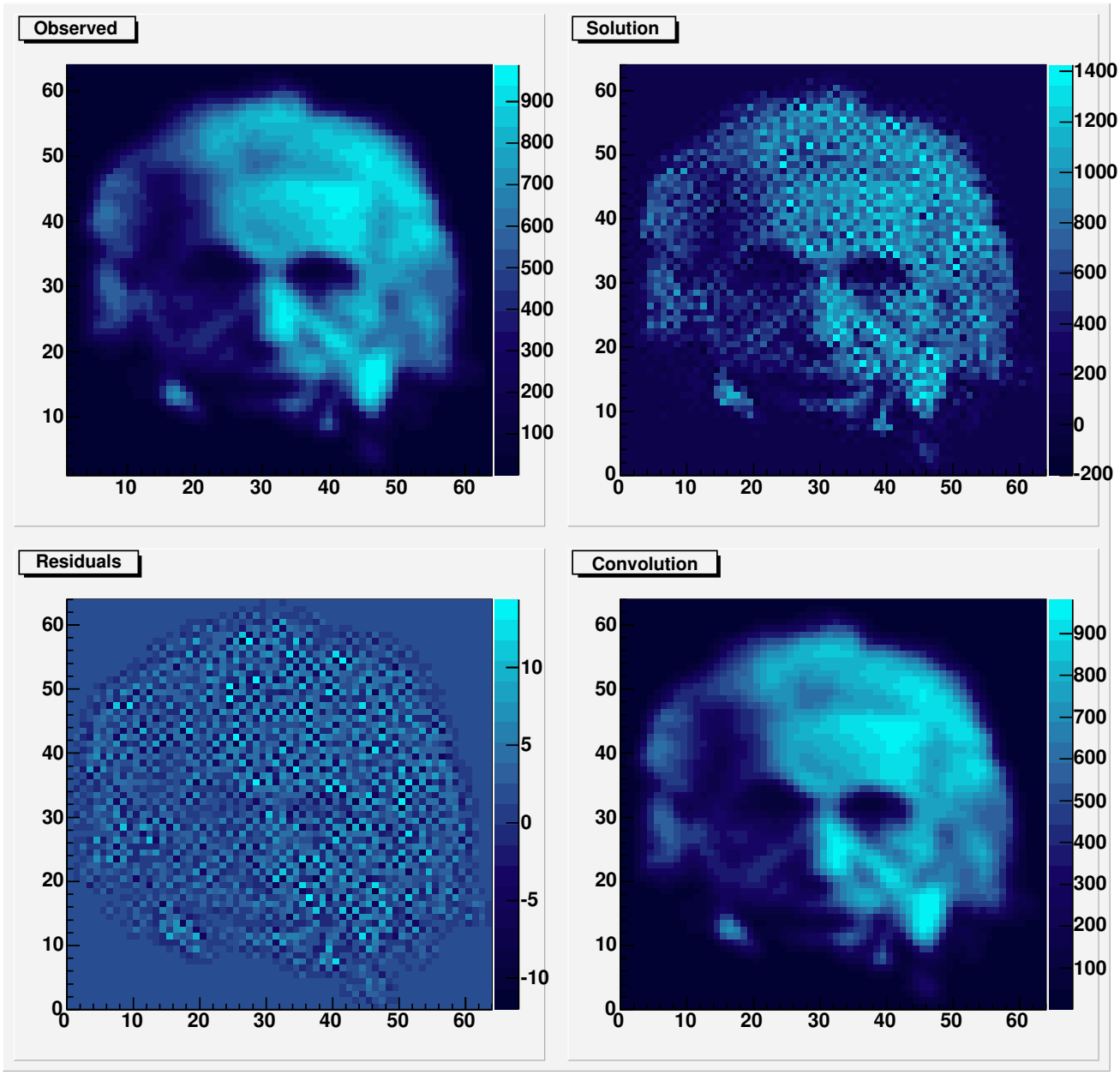
Iterative good solution

With β and enhanced: $\mu_{k+1} = \mu_k + \beta[n - R * \mu_k]$



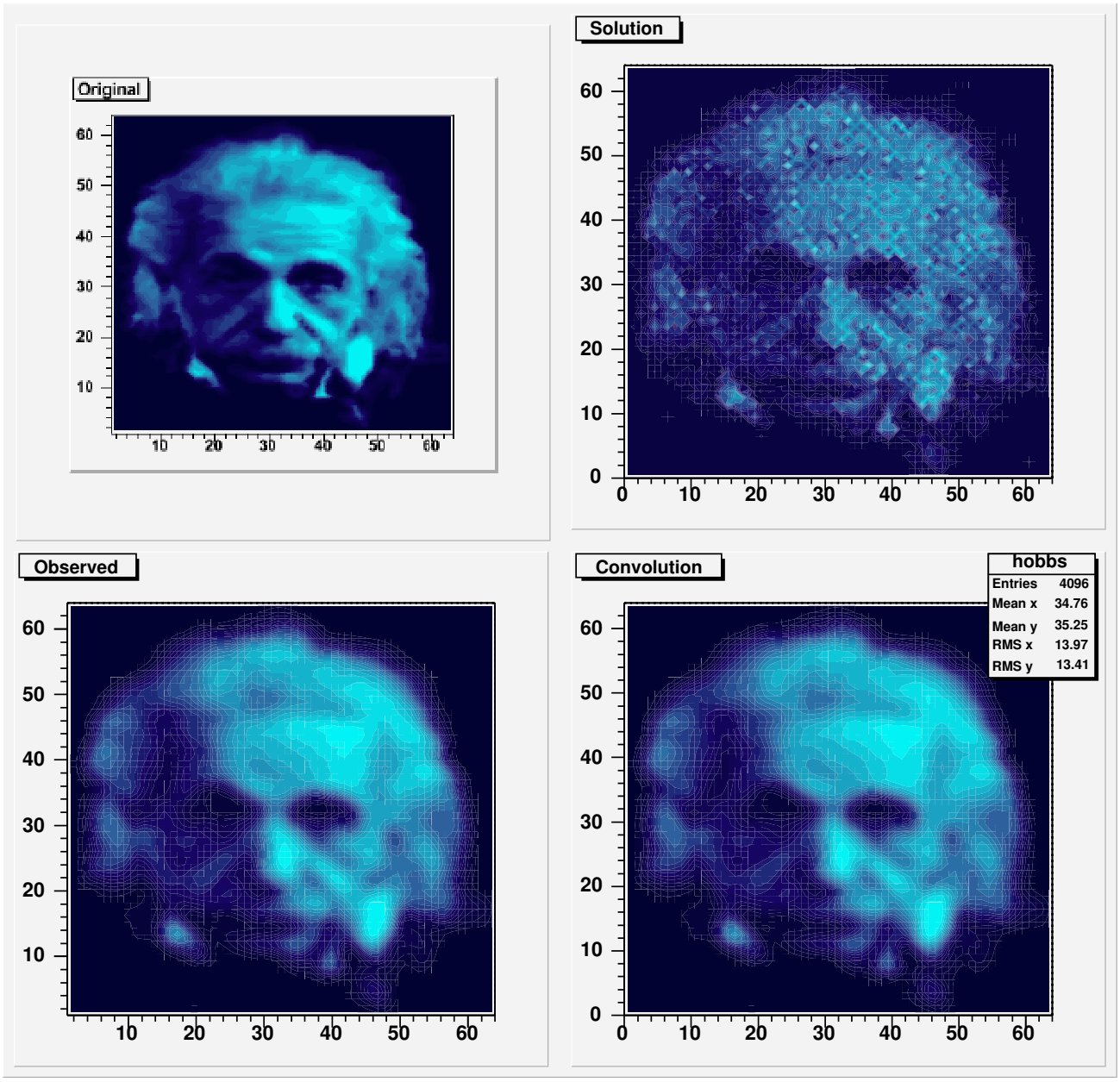
Iterative bad solution

Without β: $\mu_{k+1} = \mu_k + [n - R * \mu_k]$



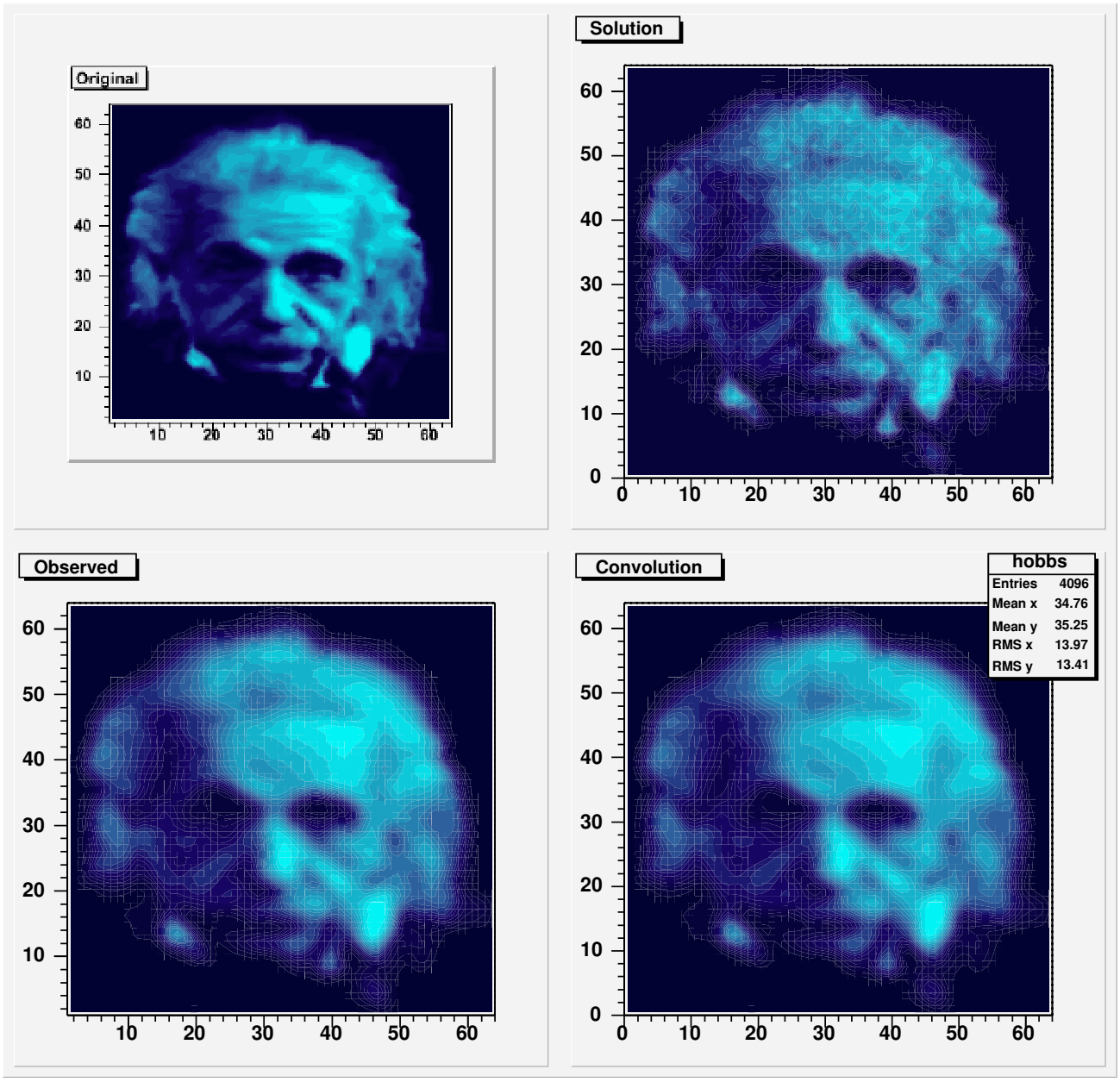
Iterative bad solution

Without β: $\mu_{k+1} = \mu_k + [n - R * \mu_k]$



Iterative acceptable solution

With β and enhanced: $\mu_{k+1} = \mu_k + \beta[n - R * \mu_k]$



Iteration techniques

Consider the case with statistical fluctuations

$$\mu_i \overset{PSF}{\rightarrow} \nu_i \overset{random}{\rightarrow} n_i$$

$$n = R{*}\mu + r$$

To have regular solutions and to make the method robust, we must search for an iterative solution which minimize the χ^2:

$$\chi^2 = ||R * \mu - n||^2 = \frac{1}{2}\sum_{ik}(\sum_{mn} R_{i-m,k-n}\,\mu_{mn} - n_{ik})^2 \ .$$

Note that is the case in which
the PSF depends on the pixel difference only (translational invariance)
In this case we consider R as an operator and we can work with symmetric $M \times N$ matrices.
Minimum χ^2 w.r.t μ_{ik} gives the equations:

$$\frac{\partial\chi^2}{\partial\mu_{mn}} = \sum_{ik}(\sum_{rs} R_{i-r,k-s}\,\mu_{rs} - n_{ik})\,R_{i-m,k-n} = 0$$

$$0 < m < M \ , \quad 0 < n < N \ .$$

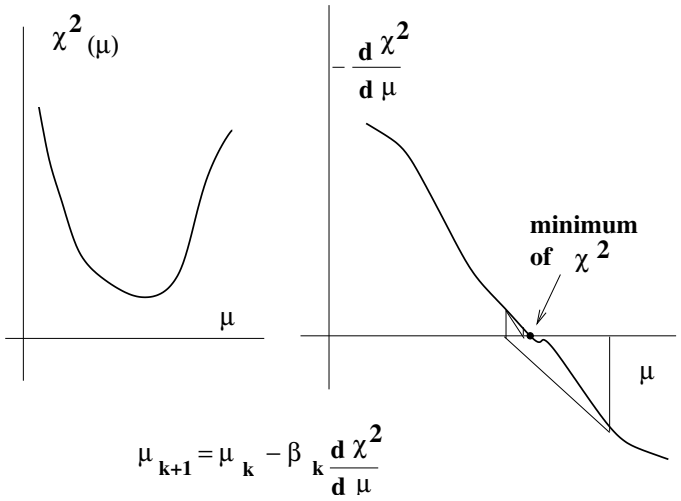
Hence :

$$\frac{\partial\chi^2}{\partial\mu} = R * [R * \mu - n] = 0 \qquad (38)$$

The iterative solution toward the μ's that minimize χ^2 is obtained with the Robbins-Monro algorithm (1951).

Robbins Monro method

It belongs to the family of the Newton-Raphson methods From (38):



$$\mu_{k+1} = \mu_k + \beta_k R * [n - R * \mu] \qquad (39)$$

This formula minimizes χ^2. For example when $\beta_k = \beta$:

$$\begin{aligned}
\mu_{k+1} &= \mu_k + \beta R * [n - R * \mu_k] \\
&= \beta R * n + (I - \beta R^2) * \mu_k \\
&= \beta R * n + \beta(I - \beta R^2) * R * n + (I - \beta R^2)^2 * \mu_{k-1} \\
&= \sum_{i=0}^{k} \beta(I - \beta R^2)^i * R * n = \frac{I - (I - \beta R^2)^{k+1}}{\beta R^2} \beta R * n \to R^{-1} n
\end{aligned}$$

Caution: since for $\mu = R^{-1}n$ the solution with 0 degrees of f. is irregular when n contains random noise, this method can give irregular solutions. However, the convergence is so slow that if one stops the process before, very often this is equivalent to smoothing.

Regularized Robbins Monro method

$$\mu_{k+1} = \mu_k + \beta_k R * [n - R * \mu_k] \tag{40}$$

Previous method converges if

$$||I - \beta R * R|| < 1$$

when β is independent of k. In this case

$$0 < \beta < \frac{2}{\text{max eigenvalue of } (R * R * \mu) * \mu^{-1}}$$

When β_k depends on k convergence is assured if (Robbins and Munro 1951)

$$\lim_{N\to\infty} \beta_N = 0 \ , \quad \sum_{N=1}^{\infty} \beta_N = \infty \ , \quad \sum_{N=1}^{\infty} \beta_N^2 < \infty \ ,$$

Next, the method must be regularized by adding to the χ^2 a term $\times$ a Lagrange multiplier α.

$$\chi^2 = ||R * \mu - n||^2 + \alpha ||C * \mu||^2 \tag{41}$$

For example,

$$||C * \mu||^2 = \sum_{ik} \mu_{ik}^2$$

$$||C * \mu||^2 = |\sum_{ik} \mu_{ik} \ln \mu_{ik}/\mu_{\text{tot}}| \ , \qquad \alpha < 0$$

The iterative solution becomes

$$\mu_{k+1} = \mu_k + \beta_k [R * n - (R * R + \alpha C * C)\mu_k] \tag{42}$$

Iterative methods vs Statistics

The iteration formula includes the χ^2 minimization

$$\chi^2 = ||R * \mu - n||^2 = \frac{1}{2}\sum_{ik}(\sum_{mn} R_{i-m,k-n}\,\mu_{mn} - n_{ik})^2 \ .$$

when one uses the the transformation

$$n_{ik} \rightarrow \frac{n_{ik}}{\sigma_{ik}}$$

$$R_{i-m,k-n} \rightarrow \frac{R_{i-m,k-n}}{\sigma_{ik}}$$

However, the convercence in this case results to be very slow
(This point should be better investigated)
The best technique seems to be to iterate on the difference and to set-up the regularization term to make in the objective function

$$A = \chi^2 + \alpha||C\mu||^2 > 0$$

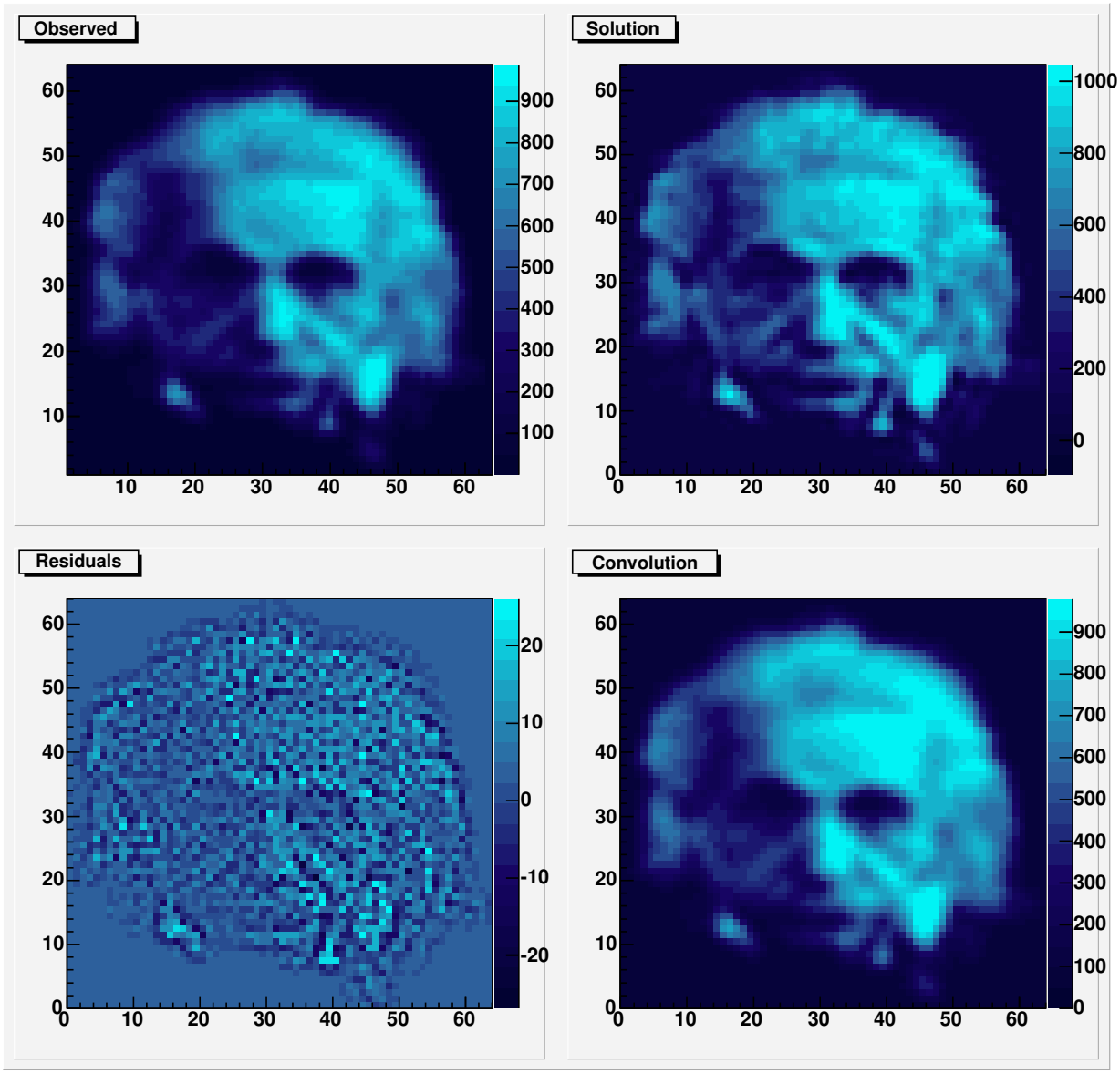
$\chi^2 \simeq DoF \equiv$ **number of pixel** N**:**

$$\alpha = \frac{A - N}{||C\mu||^2}$$

Iterative regularized good solution

$$\mu_{k+1} = \mu_k + \beta_k[R * n - (R * R + \alpha I) * \mu_k]$$

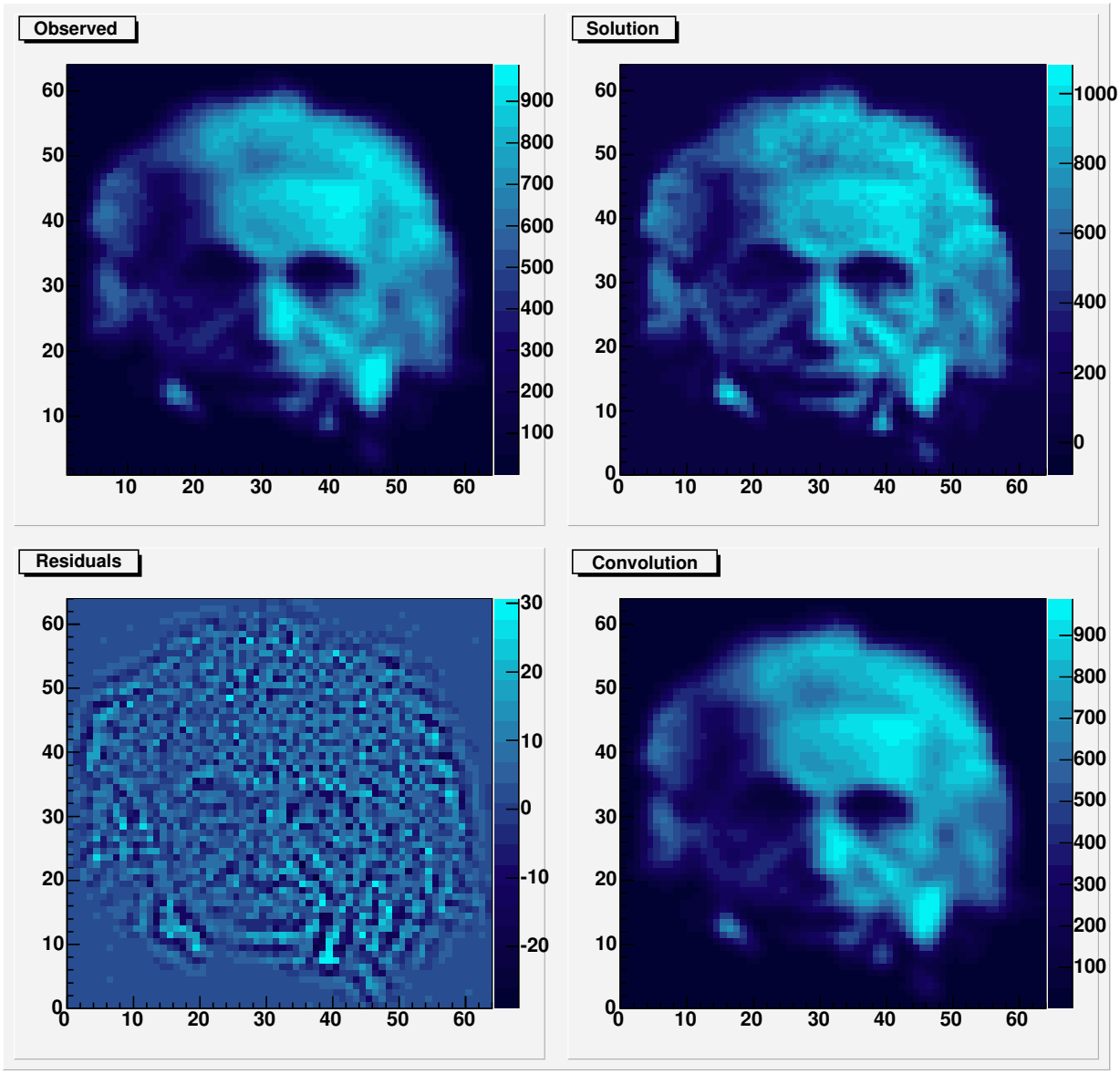
About 100 iterations, regularized with the sum of squares



Iterative regularized good solution

$$\mu_{k+1} = \mu_k + \beta_k[R * n - [R * R * \mu_k + \alpha(\ln \mu_k/\mu_T + I)]$$

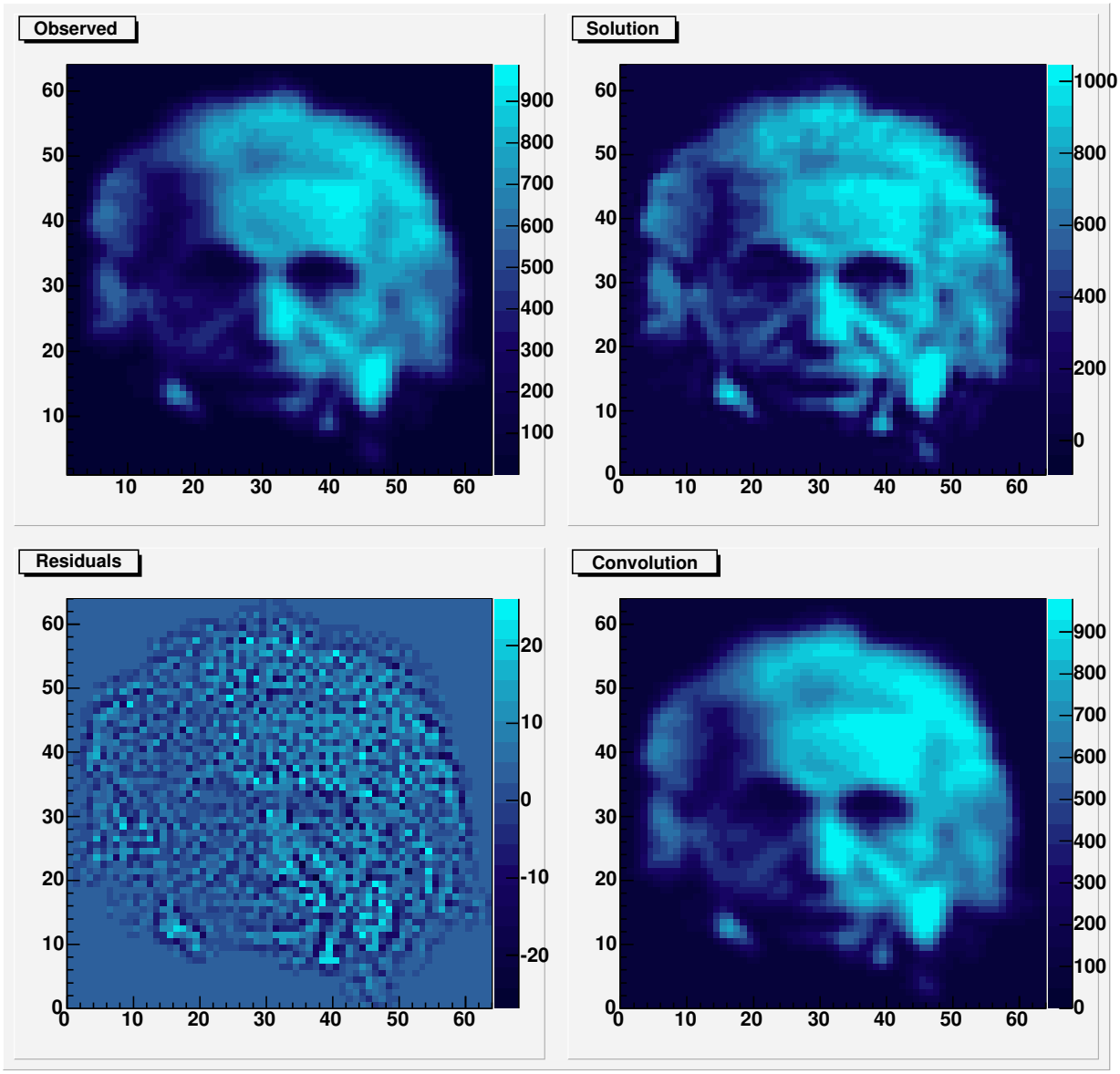
About 40 iterations, regularized with Maximum entropy



Iterative regularized good solution

$$\mu_{k+1} = \mu_k + \beta_k[R * n - (R * R + \alpha I) * \mu_k]$$

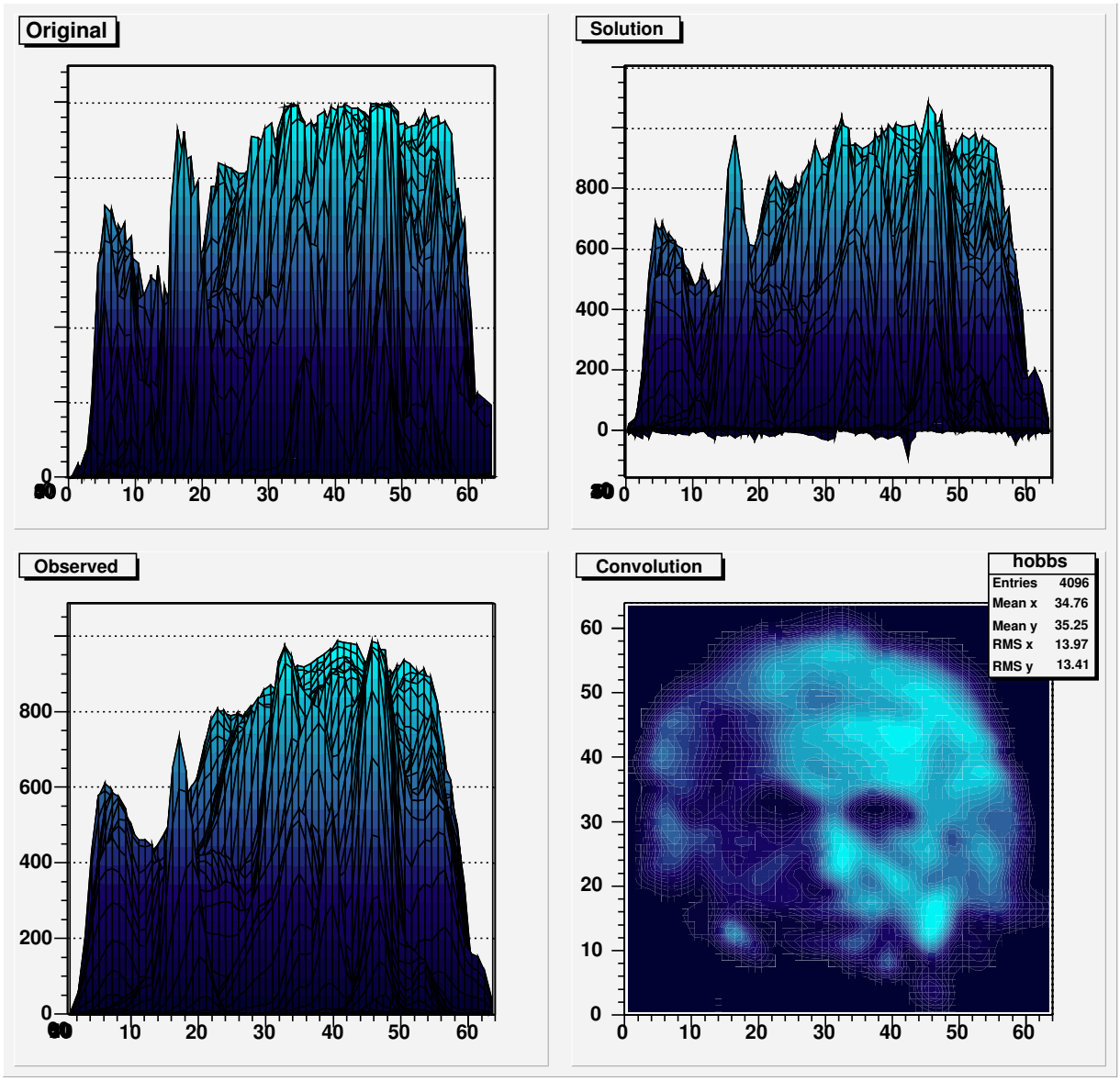
About 100 iterations, regularized with the sum of squares



Iterative regularized good solution

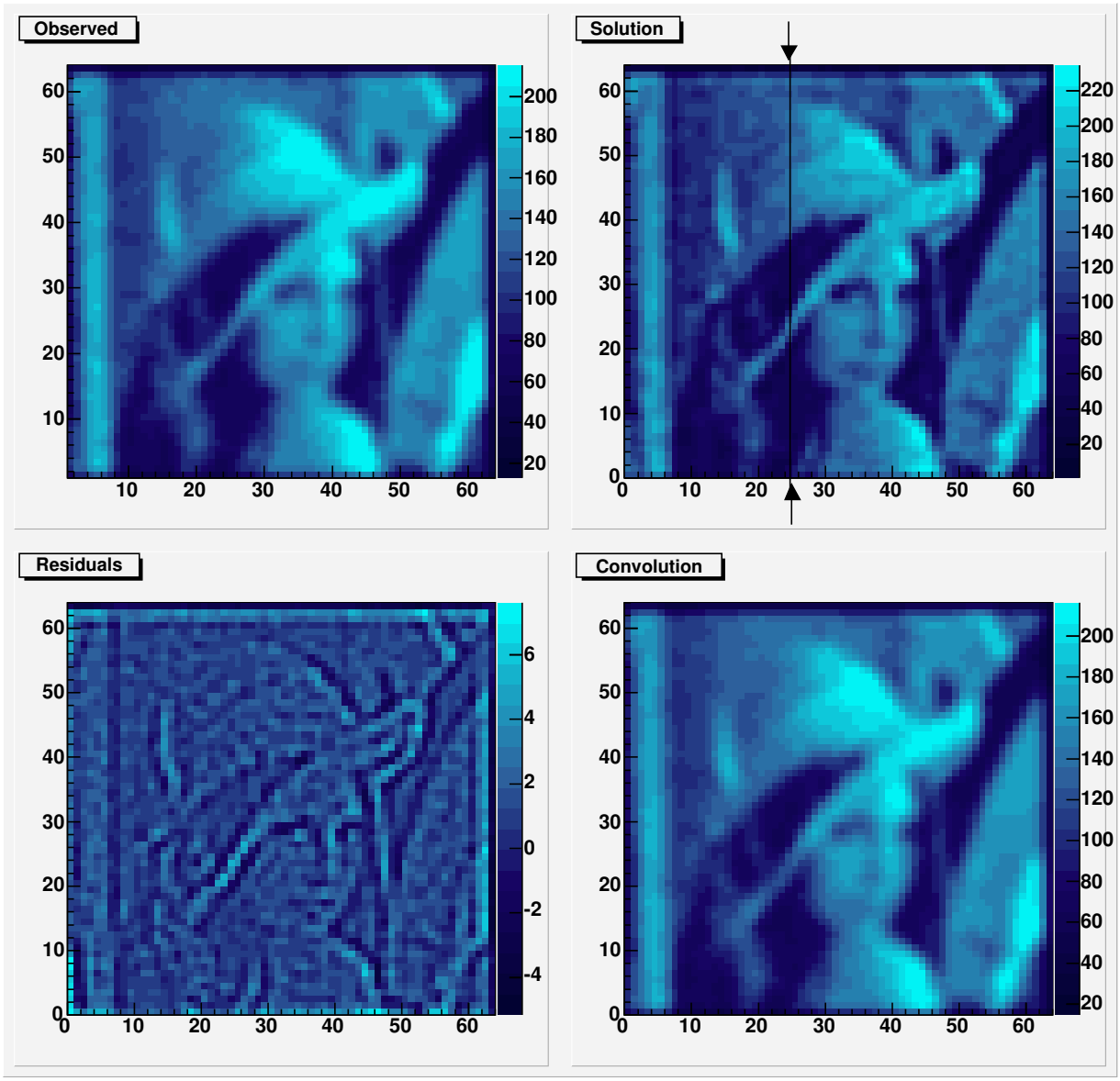
$$\mu_{k+1} = \mu_k + \beta_k[R * n - (R * R + \alpha I) * \mu_k]$$

Projection on a plane



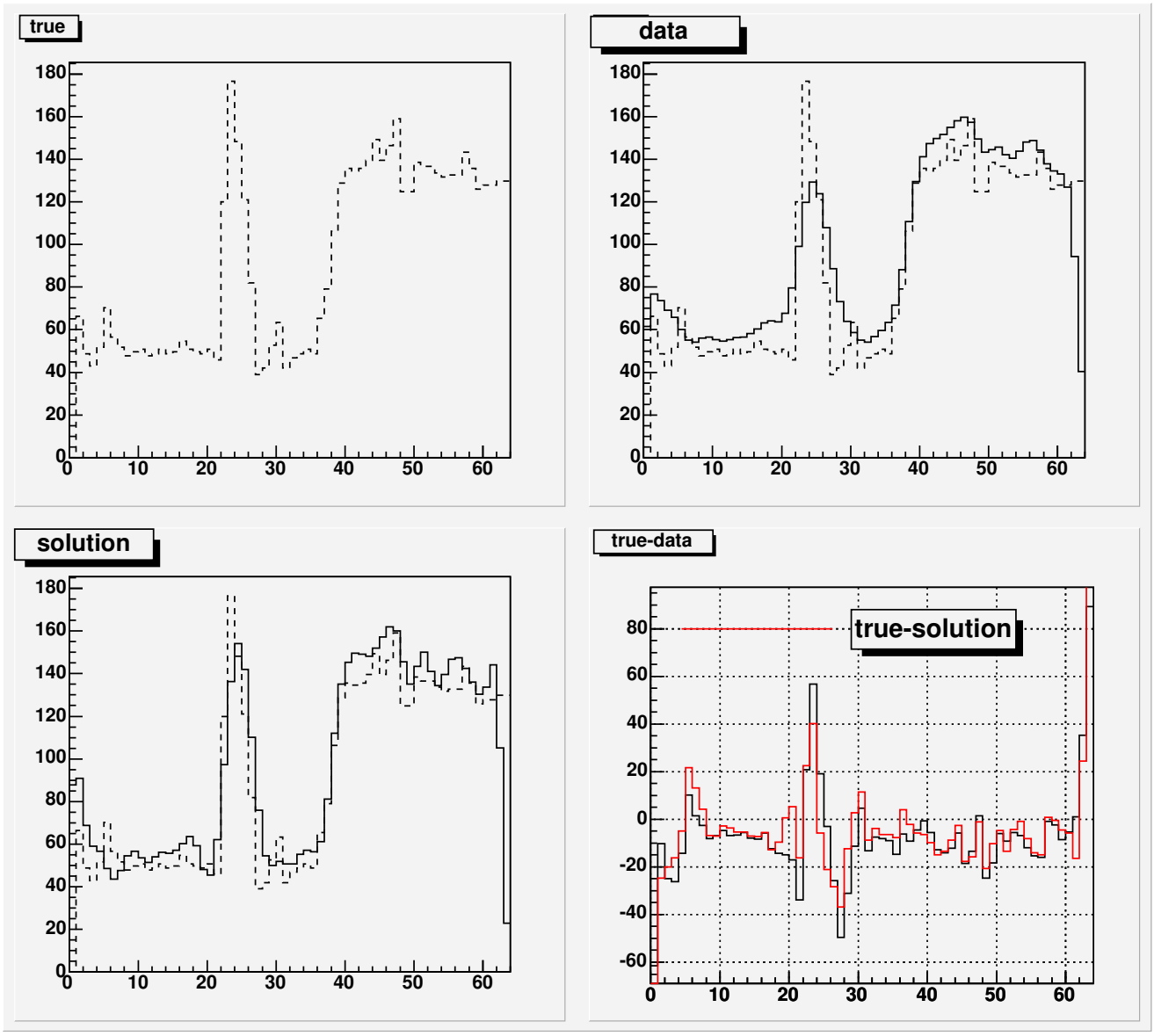
Iterative regularized good solution

$$\mu_{k+1} = \mu_k + \beta_k[R * n - (R * R + \alpha I) * \mu_k]$$



Iterative regularized good solution

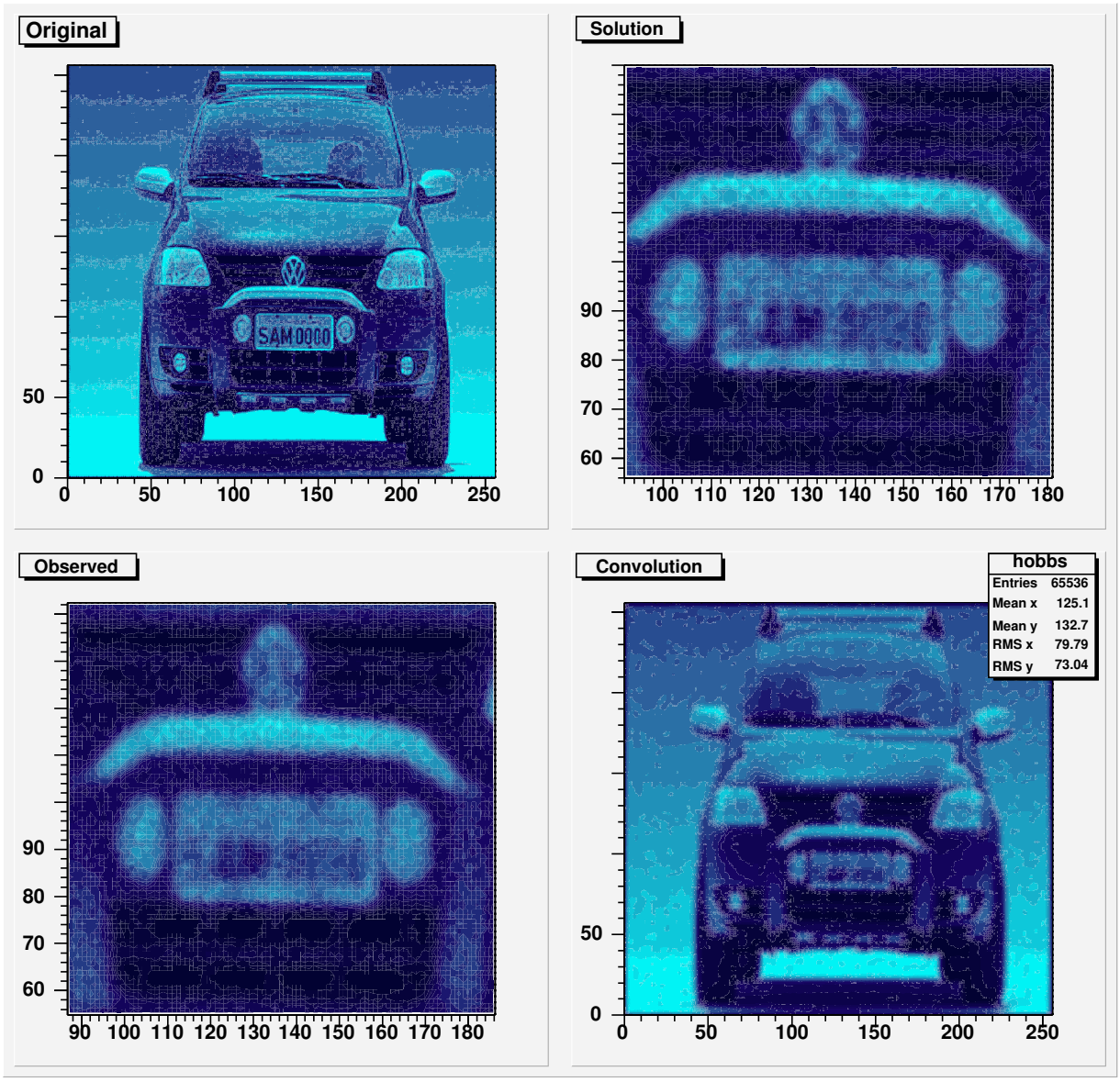
$$\mu_{k+1} = \mu_k + \beta_k[R * n - (R * R + \alpha I) * \mu_k]$$



Iterative good solution

$$\mu_{k+1} = \mu_k + \beta_k[R * n - (R * R + \alpha I) * \mu_k]$$

About 80 iterations ...



Stopping Rules

- χ^2 **variation**

$$\chi^2_k = ||y - R * \mu_k||^2 = \sum_i^{M\times N} \frac{(y_i - \sum_{ij} R_{ij}\mu_{j(k)})^2}{\sum_{ij} R_{ij}\mu_{j(k)}}$$

$$\frac{\chi^2_k - \chi^2_{k-1}}{\chi^2_{k-1}} < 10^{-6}$$

If the regularization is good, one has $\chi^2_k \simeq$ DoF.

- **Signal to noise ratio (usually measured in decibel)**

$$SNR = 10\log_{10}\left[\frac{\sum_i(\mu_i - y_i)^2}{\sum_i(\mu_i - \mu_{true\ i})^2}\right]$$

where μ_{true} is the true image. This quantity is used in the MC simulations during when the true image is known.

- **convergence of the solution**

$$\frac{||\mu_k - \mu_{k-1}||^2}{||\mu_{k-1}||^2} < 10^{-6}$$

More general iteration methods

When the PSF operator R is not traslationally invariant the PSF operation $R * \mu$

$$A_{ik} = \sum_{ik} (\sum_{rs} R_{i-r,k-s}\, \mu_{rs})$$
$$0 < r < M \ , \quad 0 < s < N \ .$$

can not be represented by a M $\times$ N square matrix multiplication.
The iterative methods can not be expressed any more in a M $\times$ N operatorial form
In this case the image is transformed in a vector of dimension MN (usually ordered by row, lexicographical order).

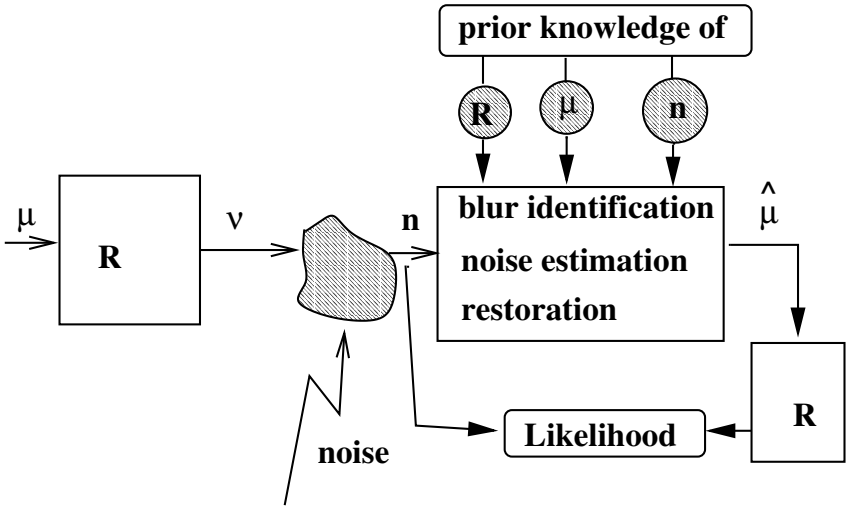
The PSF becomes a $MN \times MN$ matrix.
The advantage is that the PSF becomes a matrix R calculated only once
The disadvantage are the enormous dimensions of the system (megapixel matrices).
The general formula becomes:

$$\mu_{k+1} = \mu_k + \beta_k [R^T n - (R^T R + \alpha C^T C)\mu_k] \qquad (43)$$

$$||I - \beta R^T R|| < 1$$

$$0 < \beta < \frac{2}{\text{max eigenvalue of } (R^T R + \alpha C^T C)}$$

Conclusions



- maximum likelihood or minimum χ^2 methods requires regularization terms
- these terms are justified by Bayesian statistics
- alternatively, the DoF are introduced selecting the important cluster of pixels (pixons)
- stochastic minimization techniques (Metropolis) are sometimes used, in particular when the PSF is not known (blind deconvolution)
- iterative methods are more convenient with traslationally invariant PSF
- in general these methods require the manipulation of enormous matrices. In these cases stochastic minimization is preferred