Introduction to Probabilistic Reasoning

4. An overview of applications

Giulio D'Agostini

Dipartimento di Fisica

Università di Roma La Sapienza

Su lucidi tradizionali

- Inferenza parametrica: modello normale, binomiale, poissoniano
- Prior coniugate
- Modelli gerarchici (rappresentati graficamente come reti bayesiane)
- Incertezze dovute ad errori sistematici: caso generale, esempi su modelli semplici, approssimazioni.
- Raccomandazioni ISO (‘GUM’)

Per riferimenti, link, etc. vedi sul sito.

Bayes theorem on continuous variables

$$
\begin{aligned}
f(x,\mu\,|\,I) &= f(x\,|\,y,I)\cdot f(\mu\,|\,I) \\
f(x\,|\,I) &= \int f(x,\mu\,|\,I)\,dy
\end{aligned}
$$

$$
\begin{aligned}
f(\mu\,|\,x,I) &= \frac{f(x\,|\,\mu,I)\cdot f(\mu\,|\,I)}{f(x\,|\,I)} \\
&= \frac{f(x\,|\,\mu,I)\cdot f(\mu\,|\,I)}{\int f(x,\mu\,|\,I)\,dy}
\end{aligned}
$$

$$
\begin{aligned}
f(\mu\,|\,x,I) &\propto f(x\,|\,\mu,I)\cdot f(\mu\,|\,I) \\
&\propto \text{likelihood} \times \text{prior}
\end{aligned}
$$

Bayes theorem on continuous variables

$$
\begin{aligned}
f(x, \mu \,|\, I) &= f(x \,|\, y, I) \cdot f(\mu \,|\, I) \\
f(x \,|\, I) &= \int f(x, \mu \,|\, I)\, dy
\end{aligned}
$$

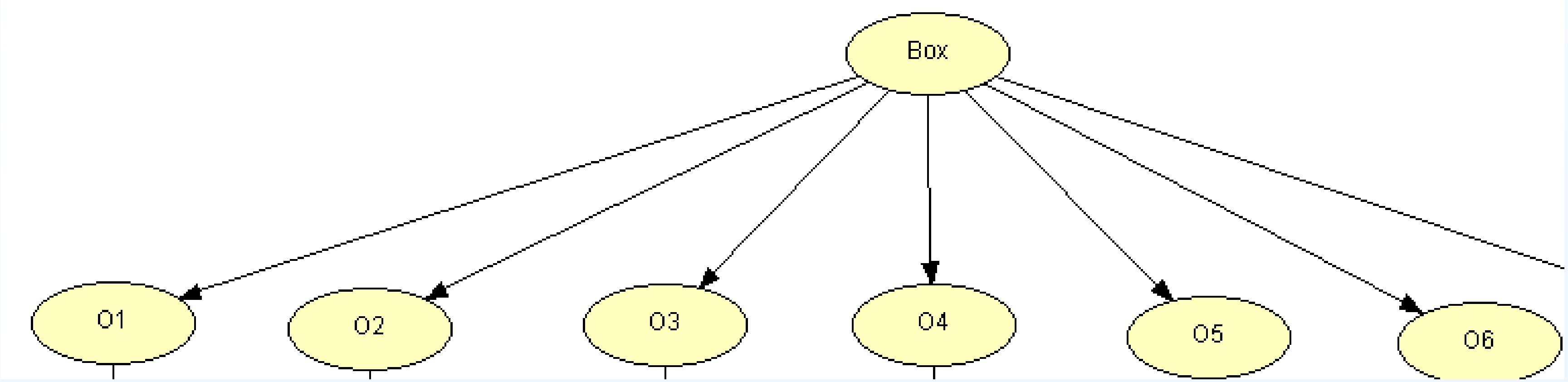
$$
\begin{aligned}
f(\mu \,|\, x, I) &= \frac{f(x \,|\, \mu, I) \cdot f(\mu \,|\, I)}{f(x \,|\, I)} \\
&= \frac{f(x \,|\, \mu, I) \cdot f(\mu \,|\, I)}{\int f(x, \mu \,|\, I)\, dy}
\end{aligned}
$$

$$
\begin{aligned}
f(\mu \,|\, x, I) &\propto f(x \,|\, \mu, I) \cdot f(\mu \,|\, I) \\
&\propto \text{likelihood} \times \text{prior}
\end{aligned}
$$

IF prior flat $\rightarrow$ inference dominated by Likelihood
Maximum of posterior $\Leftrightarrow$ Maximum of Likelihhod.
BUT $f(\mu \,|\, x, I)$ does not exist in ML methods!

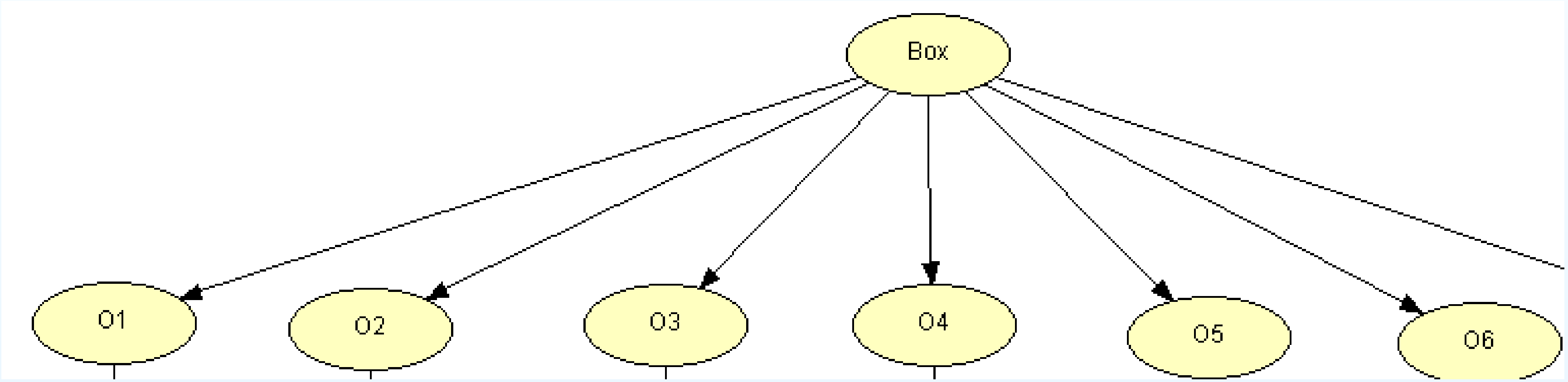
Cause-effect representation

box content $\longrightarrow$ observed color



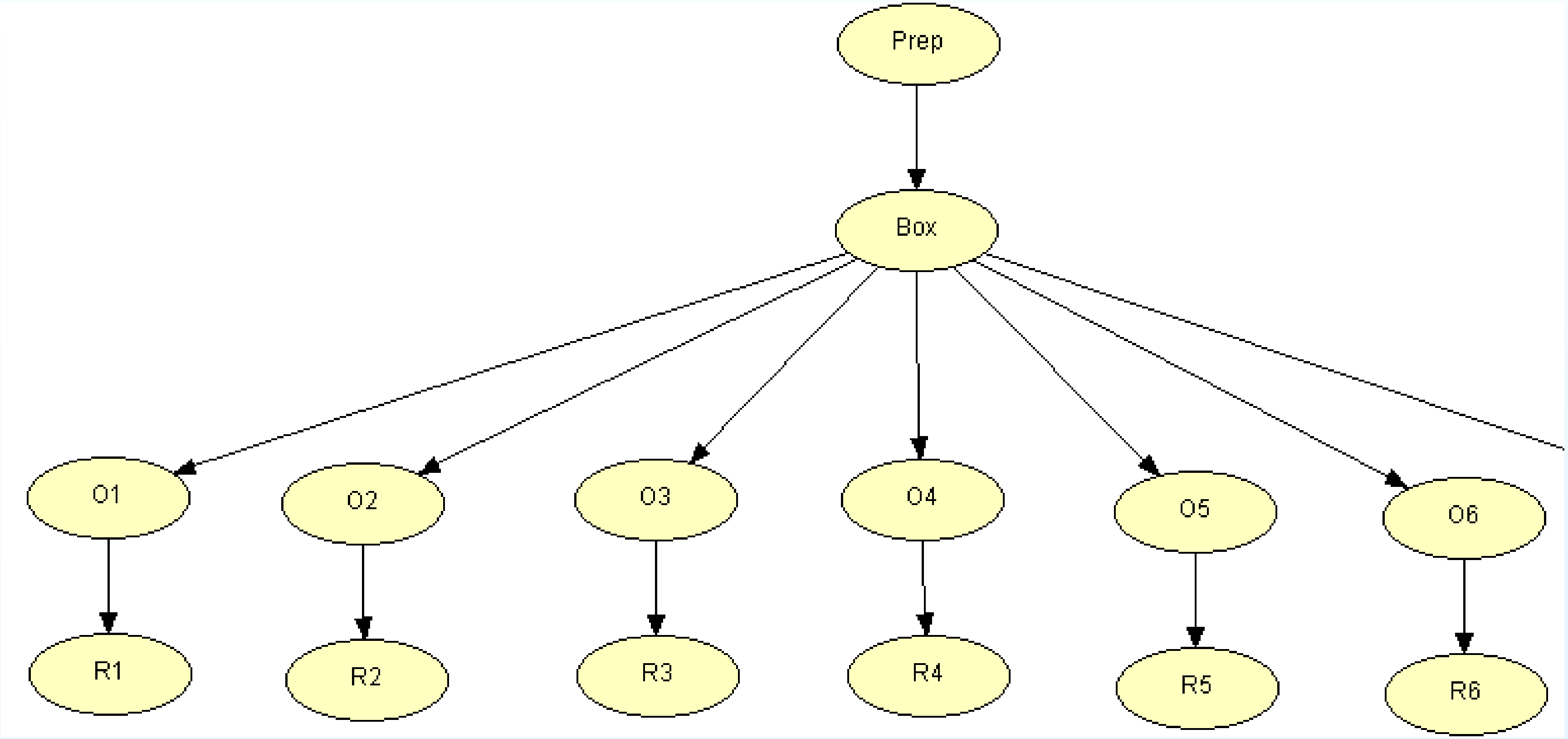
Cause-effect representation

box content $\longrightarrow$ observed color

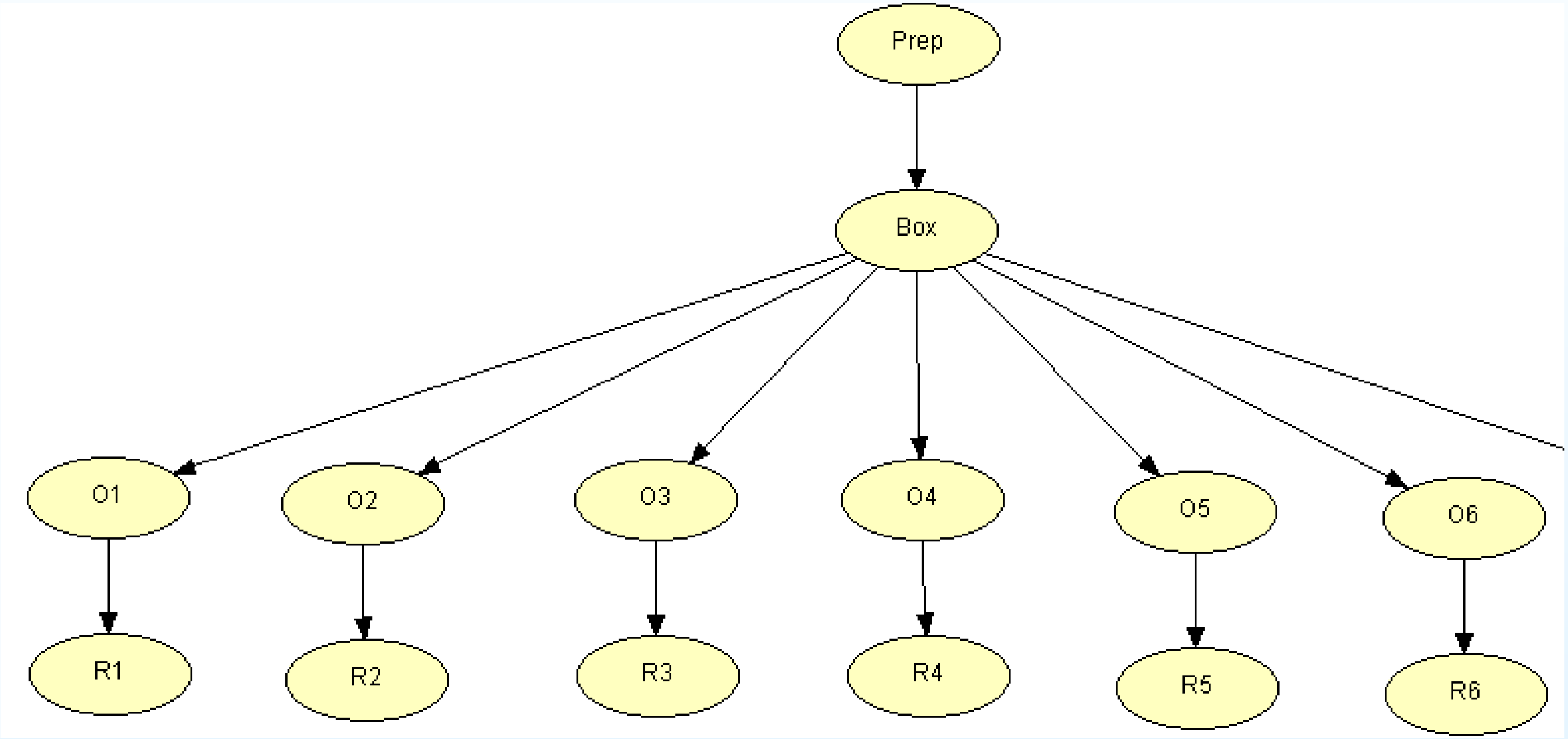


An effect might be the cause of another effect $\longrightarrow$

A network of causes and effects

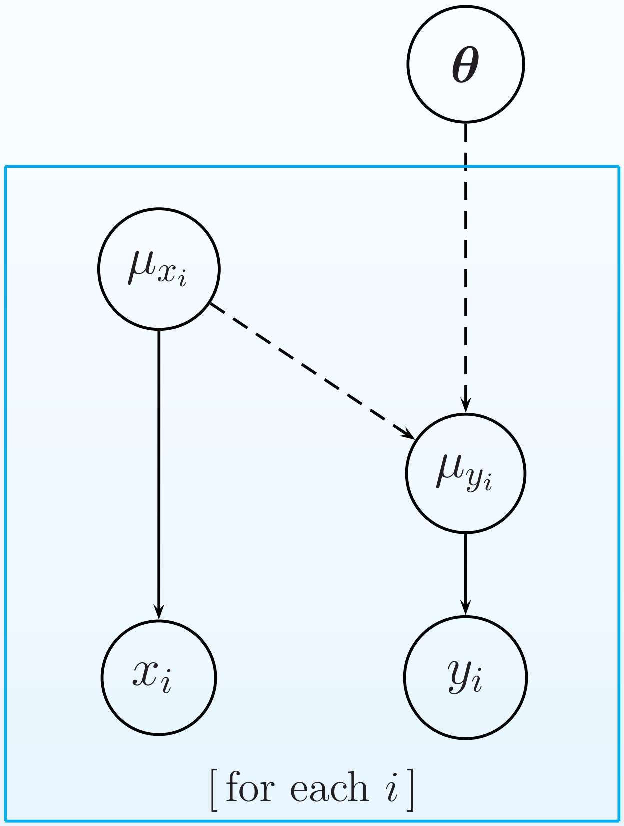


A network of causes and effects



and so on. . .　　　　$\Rightarrow$ Physics applications

A different way to view fit issues



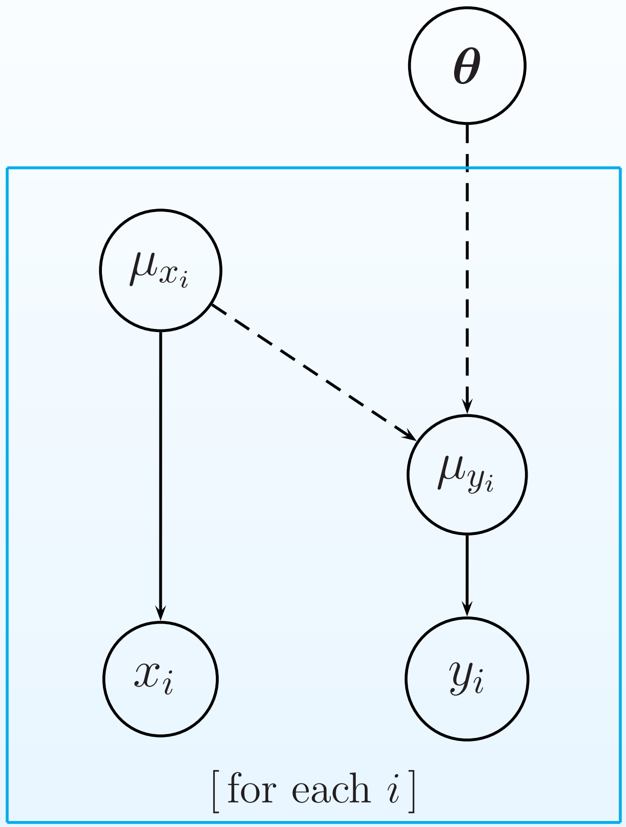
Determistic link μ_x's to μ_y's

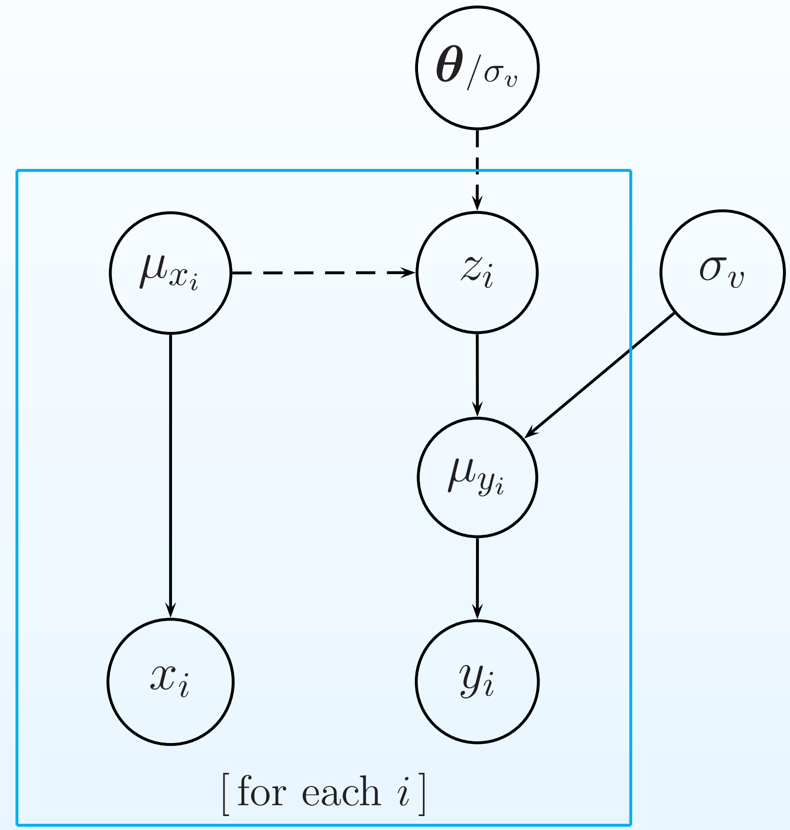
Probabilistic links $\mu_x \rightarrow x$, $\mu_y \rightarrow y$

(errors on both axes!)

$\Rightarrow$ aim of fit: $\{\boldsymbol{x}, \boldsymbol{y}\} \rightarrow \boldsymbol{\theta}$

A different way to view fit issues





Determistic link μ_x's to μ_y's

Probabilistic links $\mu_x \longrightarrow x$, $\mu_y \longrightarrow y$

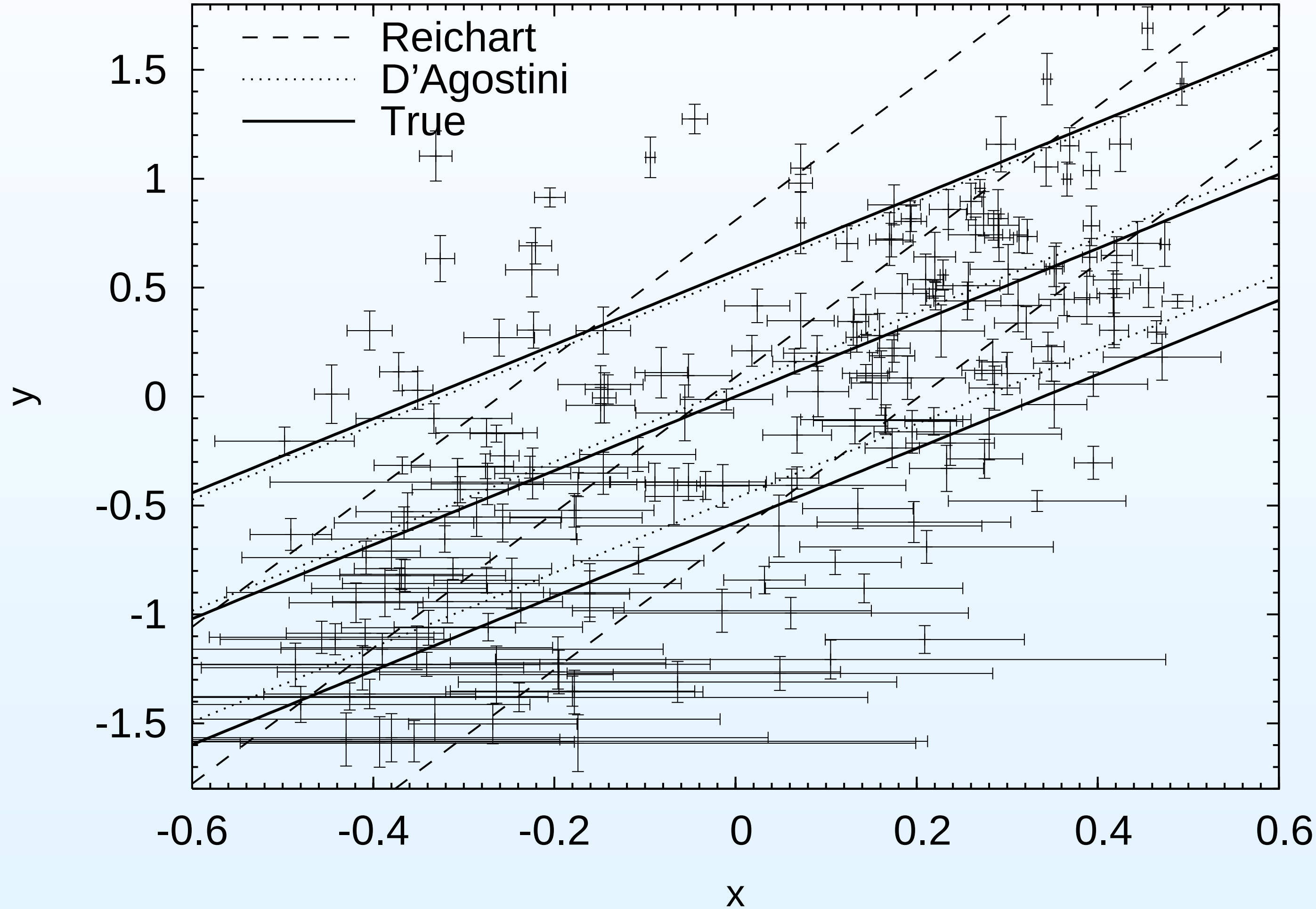
(errors on both axes!)

$\Rightarrow$ aim of fit: $\{\boldsymbol{x}, \boldsymbol{y}\} \longrightarrow \boldsymbol{\theta}$

Extra spread of the data points

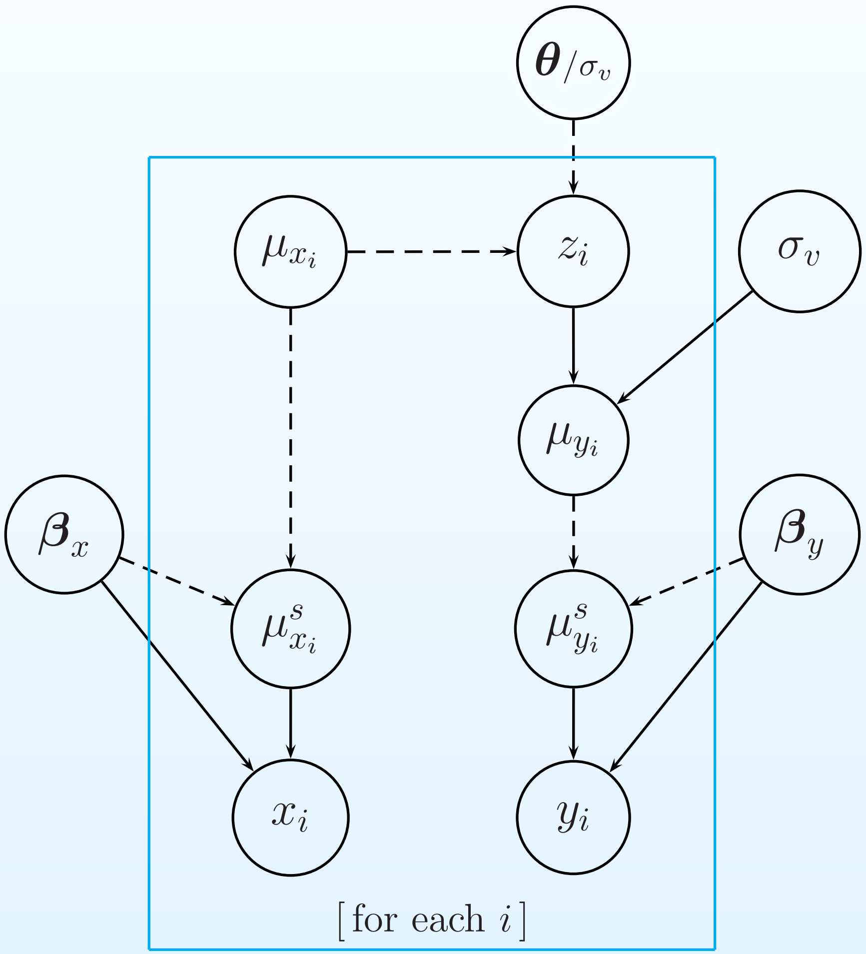
A different way to view fit issues

A physics case (from Gamma ray burts):



(Guidorzi et al., 2006)

A different way to view fit issues



Adding systematics

Conditional factorization of a Bayesian Network

$$
\begin{aligned}
f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta} \,|\, I) \;=\; & f(\boldsymbol{x} \,|\, \boldsymbol{y}, \boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta}, I) \\
& \cdot f(\boldsymbol{y} \,|\, \boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta}, I) \\
& \cdot f(\boldsymbol{\mu}_y \,|\, \boldsymbol{\mu}_x, \boldsymbol{\theta}, I) \\
& \cdot f(\boldsymbol{\mu}_x \,|\, \boldsymbol{\theta}, I) \\
& \cdot f(\boldsymbol{\theta} \,|\, I)
\end{aligned}
$$

with

$$
\begin{aligned}
f(\boldsymbol{x} \,|\, \boldsymbol{y}, \boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta}, I) &\;\longrightarrow\; f(\boldsymbol{x} \,|\, \boldsymbol{\mu}_x, I) \\
f(\boldsymbol{y} \,|\, \boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta}, I) &\;\longrightarrow\; f(\boldsymbol{y} \,|\, \boldsymbol{\mu}_y, I) \\
f(\boldsymbol{\mu}_y \,|\, \boldsymbol{\mu}_x, \boldsymbol{\theta}, I) &\;\longrightarrow\; \prod_i \delta[\,\mu_{y_i} - \mu_y(\mu_{x_i}, \boldsymbol{\theta})\,] \\
f(\boldsymbol{\mu}_x \,|\, \boldsymbol{\theta}, I) &\;\longrightarrow\; f(\boldsymbol{\mu}_x \,|\, I) \;\; \text{(prior on } \boldsymbol{\mu}_x) \\
f(\boldsymbol{\theta} \,|\, I) &\;\longrightarrow\; \text{prior on } \boldsymbol{\theta}
\end{aligned}
$$

Applying Bayes' theorem

Once we have built $f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta} \,|\, I)$, the rest is just math:

$$
\begin{aligned}
f(\boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta} \,|\, \boldsymbol{x}, \boldsymbol{y}, I) &= \frac{f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta} \,|\, I)}{f(\boldsymbol{x}, \boldsymbol{y} \,|\, I)} \\
&= \frac{f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta} \,|\, I)}{\int f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta} \,|\, I)\; d\boldsymbol{\mu}_x\, d\boldsymbol{\mu}_y\, d\boldsymbol{\theta}} \\
&\propto \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta} \,|\, I)
\end{aligned}
$$

$$
\begin{aligned}
f(\boldsymbol{\theta} \,|\, \boldsymbol{x}, \boldsymbol{y}, I) &= \int f(\boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta} \,|\, \boldsymbol{x}, \boldsymbol{y}, I)\; d\boldsymbol{\mu}_x\, d\boldsymbol{\mu}_y \\
&\propto \int f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta} \,|\, I)\; d\boldsymbol{\mu}_x\, d\boldsymbol{\mu}_y\,.
\end{aligned}
$$

Easy to built up the "kernel" $\Rightarrow$ the though task is normalization!

numerical methods $\longrightarrow$ best: MCMC

Getting a likelihood for approximative purposes

Using a flat prior for μ_{x_i} (quite assumption) and making use of independence amond the couples of data points, we get:

$$f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\mu}_x,\boldsymbol{\mu}_y,\boldsymbol{\theta}\,|\,I) \;\propto\; \prod_i f(x_i\,|\,\mu_{x_i},I)\cdot f(y_i\,|\,\mu_{y_i},I)\cdot \delta[\,\mu_{y_i}-\mu_y(\mu_{x_i},\boldsymbol{\theta})\,]\cdot f(\boldsymbol{\theta}\,|\,I)$$

and, hence,

$$\begin{aligned} f(\boldsymbol{\theta}\,|\,\boldsymbol{x},\boldsymbol{y},I) &\propto \left[\int \prod_i k_{x_i}\, f(x_i\,|\,\mu_{x_i},I)\cdot f(y_i\,|\,\mu_{y_i},I)\cdot \delta[\,\mu_{y_i}-\mu_y(\mu_{x_i},\boldsymbol{\theta})\,]\,d\boldsymbol{\mu}_x\,d\boldsymbol{\mu}_y\right]\cdot f(\boldsymbol{\theta}\,|\,I) \\ &\propto f(\boldsymbol{x},\boldsymbol{y}\,|\,\boldsymbol{\theta},I)\cdot f(\boldsymbol{\theta}\,|\,I) \\ &\propto \mathcal{L}(\boldsymbol{\theta}\,;\,\boldsymbol{x},\boldsymbol{y})\cdot f(\boldsymbol{\theta}\,|\,I) \end{aligned}$$

($\Rightarrow$ to those who like to think in terms of likelihoods)

Linear fit with Gaussian errors on both axes (and more)

A special case is the lineare fit (i.e. $\boldsymbol{\theta} = \{m, c\}$),
the previous formulae yield the following likelihood$\times$prior

$$f(m, c \,|\, \boldsymbol{x}, \boldsymbol{y}, I) \;\;\propto\;\; \prod_i \frac{1}{\sqrt{\sigma_{y_i}^2 + m^2\,\sigma_{x_i}^2}} \;\exp\left[-\frac{(y_i - m\,x_i - c)^2}{2\,(\sigma_{y_i}^2 + m^2\,\sigma_{x_i}^2)}\right] f(m, c$$

Linear fit with Gaussian errors on both axes (and more)

A special case is the lineare fit (i.e. $\boldsymbol{\theta} = \{m, c\}$), the previous formulae yield the following likelihood$\times$prior

$$f(m, c \,|\, \boldsymbol{x}, \boldsymbol{y}, I) \;\propto\; \prod_i \frac{1}{\sqrt{\sigma_{y_i}^2 + m^2\,\sigma_{x_i}^2}} \exp\left[-\frac{(y_i - m\,x_i - c)^2}{2\,(\sigma_{y_i}^2 + m^2\,\sigma_{x_i}^2)}\right] f(m, c$$

If also extra variability of the data is allowed, modelled with and intermediate ‘hidden variables’ z_i, around which the μ_{y_i} fluctuate normally with sigma σ_v, we get a three quantities inference:

$$f(m, c, \sigma_v \,|\, \boldsymbol{x}, \boldsymbol{y}, I) \;\propto\; \prod_i \frac{1}{\sigma_{eq}} \exp\left[-\frac{(y_i - m\,x_i - c)^2}{2\,\sigma_{eq}^2}\right] f(m, c, \sigma_v \,|\, I)$$

(with $\sigma_{eq} = \sqrt{\sigma_v^2 + \sigma_{y_i}^2 + m^2\,\sigma_{x_i}^2}$)

An easier example

Very basic problem:

- A sample of data comes from a true value μ according to a normal model with σ unknown:

$$X_i \sim \mathcal{N}(\mu, \sigma).$$

- A future measurement, Y, will be produced from the same μ with the same σ
- We are interested in
  - $f(\mu \,|\, \text{data})$,
  - $f(\sigma \,|\, \text{data})$,
  - $f(y \,|\, \text{data})$

  Note: we are interested in pdf's and not in 'estimators' and 'their errors'

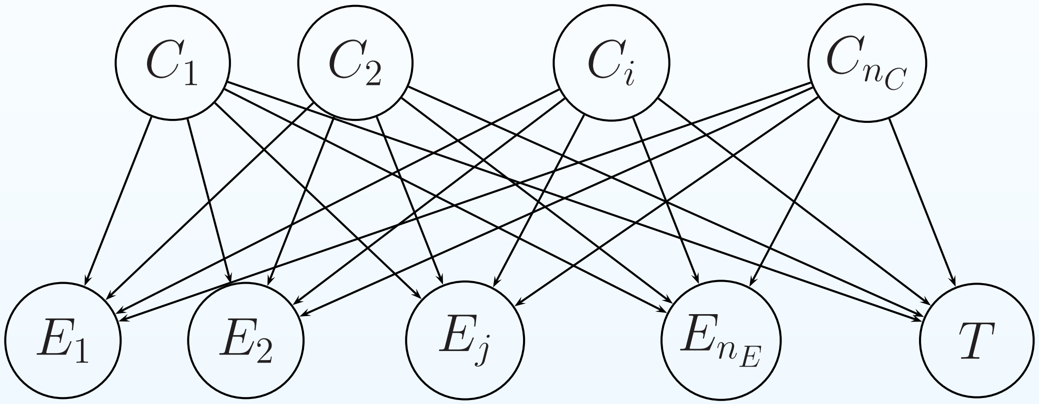
Problem modelled in OpenBUGS

BUGS: Bayesian analysis software Using Gibbs Sampling

```
model {
   for (i in 1:N) {
      X[i] ~ dnorm(mu,tau);
   }
   mu ~ dnorm(0.0, 1.0E-4);
   tau   ~ dgamma(1.0E-3, 1.0E-3);
   sigma <- 1.0/sqrt(tau);
   y ~ dnorm(mu,tau)
}
```

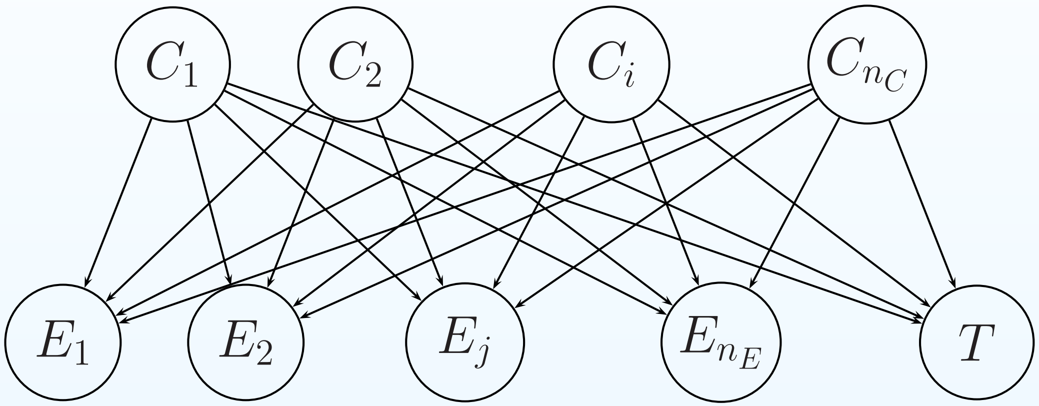
Unfolding a discretized spectrum

Probabilistic links: Cause-bins $\leftrightarrow$ effect-bins

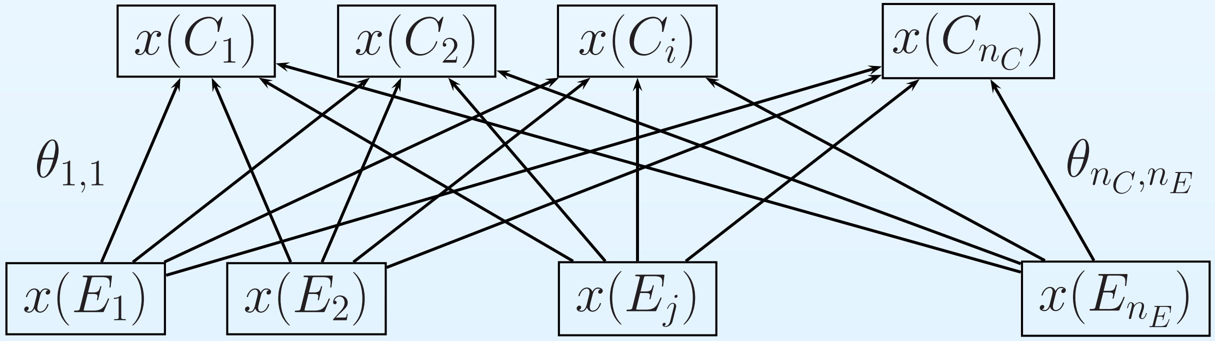


Unfolding a discretized spectrum

Probabilistic links: Cause-bins $\leftrightarrow$ effect-bins



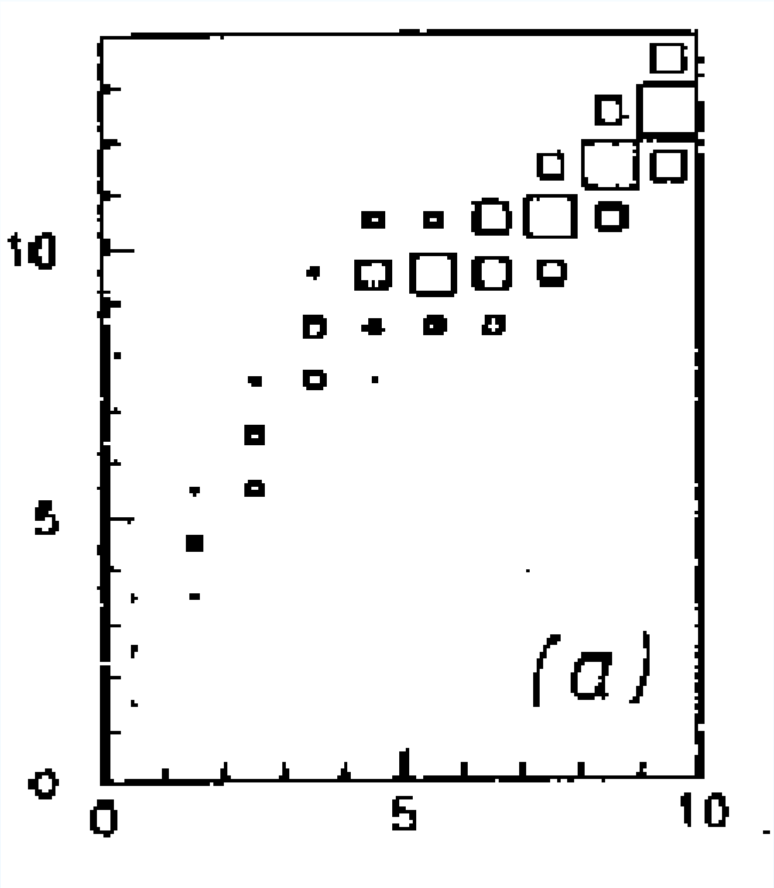
Sharing the observed events among the cause-bins

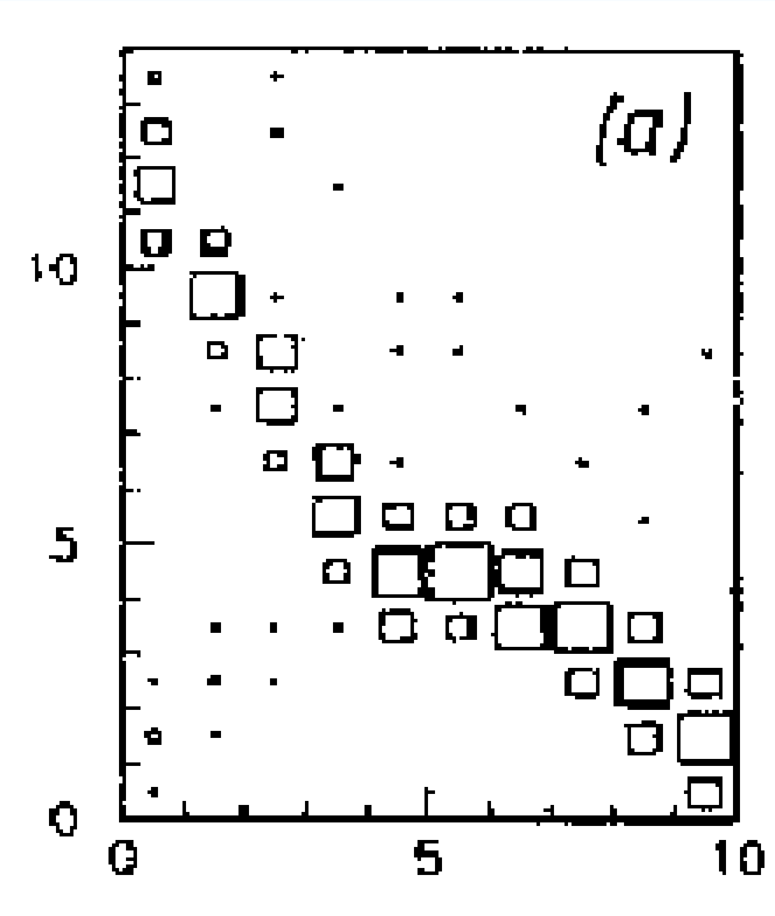


Unfolding a discretized spectrum

Need a smearing matrix (evaluated somehow)

Academic examples:





Why unfolding?

Te idea is to provide somethimg similar to an experimental spectrum, with a minimal interpretation by the experimentalist, a part from correcting from experimental distortions.

(The alternative would be to give a parametrized description of the true spectrum – a fit)

Smearing matrix $\longrightarrow$ unfolding matrix

Invert smearing matrix?

Smearing matrix $\longrightarrow$ unfolding matrix

Invert smearing matrix?

In general is a bad idea:

not a rotational problem

but an inferential problem!

Smearing matrix $\longrightarrow$ unfolding matrix

Imagine $S = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{pmatrix}$: $\longrightarrow U = S^{-1} = \begin{pmatrix} 1.33 & -0.33 \\ -0.33 & 1.33 \end{pmatrix}$

Let the true be $s_t = \begin{pmatrix} 10 \\ 0 \end{pmatrix}$: $\longrightarrow s_m = S \cdot s_t = \begin{pmatrix} 8 \\ 2 \end{pmatrix}$;

If we measure $s_m = \begin{pmatrix} 8 \\ 2 \end{pmatrix} \longrightarrow S^{-1} \cdot s_m = \begin{pmatrix} 10 \\ 0 \end{pmatrix}$ $\surd$

Smearing matrix $\longrightarrow$ unfolding matrix

Imagine $S = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{pmatrix}$: $\longrightarrow U = S^{-1} = \begin{pmatrix} 1.33 & -0.33 \\ -0.33 & 1.33 \end{pmatrix}$

Let the true be $s_t = \begin{pmatrix} 10 \\ 0 \end{pmatrix}$: $\longrightarrow s_m = S \cdot s_t = \begin{pmatrix} 8 \\ 2 \end{pmatrix}$;

If we measure $s_m = \begin{pmatrix} 8 \\ 2 \end{pmatrix} \longrightarrow S^{-1} \cdot s_m = \begin{pmatrix} 10 \\ 0 \end{pmatrix}$ $\surd$

BUT

if we had measured $\begin{pmatrix} 9 \\ 1 \end{pmatrix} \longrightarrow S^{-1} \cdot s_m = \begin{pmatrix} 11.7 \\ -1.7 \end{pmatrix}$

if we had measured $\begin{pmatrix} 10 \\ 0 \end{pmatrix} \longrightarrow S^{-1} \cdot s_m = \begin{pmatrix} 13.3 \\ -3.3 \end{pmatrix}$

Smearing matrix $\longrightarrow$ unfolding matrix

Imagine $S = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{pmatrix}$: $\longrightarrow U = S^{-1} = \begin{pmatrix} 1.33 & -0.33 \\ -0.33 & 1.33 \end{pmatrix}$

Let the true be $s_t = \begin{pmatrix} 10 \\ 0 \end{pmatrix}$: $\longrightarrow s_m = S \cdot s_t = \begin{pmatrix} 8 \\ 2 \end{pmatrix}$;

If we measure $s_m = \begin{pmatrix} 8 \\ 2 \end{pmatrix} \longrightarrow S^{-1} \cdot s_m = \begin{pmatrix} 10 \\ 0 \end{pmatrix}$ $\surd$

Indeed, matrix inversion is recognized to producing ‘crazy spectra’ and even negative values (unless such large numbers in bins such fluctuations around expectations are negligeable)

Probabilistic approach

(skipping the technical details)
Exact solution is difficult: solved by approximations:

- Apply Bayes's formula to get $P(C_i \,|\, E_j)$;
- Assign the events observed in bin E_j to all 'causes' according to $P(C_i \,|\, E_j)$;
- Take into account inefficiency.
- Evaluation of uncertainties:
  - in old program (1993) was done by linearization assuming normality of results (usual formulae of 'error propagation');
  - in new program it is done by (Monte Carlo) integrations over the various pdf's of interest

A difficult problem solved by iteration

Which initial spectrum? **Flat**?

A difficult problem solved by iteration

Which initial spectrum? **Flat**?

BUT *a flat spectrum does not model correctly our indifference on all possible spectra*!

(Just a starting point which will influence the solution!)

The solution depends on it!

A difficult problem solved by iteration

Which initial spectrum? **Flat**?

BUT *a flat spectrum does not model correctly our indifference on all possible spectra*!

(Just a starting point which will influence the solution!)

The solution depends on it!

- Problem solved by iteration:
- unfolded spectrum becomes next prior, etc.
- convergency very fast
- intermediate smoothing makes method very robust.

A difficult problem solved by iteration

Which initial spectrum? **Flat**?

BUT *a flat spectrum does not model correctly our indifference on all possible spectra*!

(Just a starting point which will influence the solution!)

The solution depends on it!

- Problem solved by iteration:
- unfolded spectrum becomes next prior, etc.
- convergency very fast
- intermediate smoothing makes method very robust.

$\Rightarrow$ see demo

Upper/lower limits

“Ogni limite ha una pazienza” (Totò)

Upper/lower limits

"Ogni limite ha una pazienza" (Totò)

A very simple problem:

- counting experiment described by a binomial of unkown p;
- our aim is to 'get' p, in the sense of evaluating $f(p\,|\,\text{data})$;
- we make n trials and get $x = 0$ successes.

Upper/lower limits

“Ogni limite ha una pazienza” (Totò)

A very simple problem:

- counting experiment described by a binomial of unkown p;
- our aim is to ‘get’ p, in the sense of evaluating $f(p\,|\,\text{data})$;
- we make n trials and get $x = 0$ successes.

Bayes’ theorem:

$$f(p\,|\,n, x=0, \mathcal{B}) \;=\; \frac{f(x=0\,|\,n,\mathcal{B})\, f_0(p)}{\int_0^1 f(x=0\,|\,n,\mathcal{B})\, f_0(p)\, dp}$$

with

$$f(x=0\,|\,n,\mathcal{B}) \;=\; (1-p)^n$$

Inference about p from 0 counts

Using flat prior, i.e. $f_0(p) = k$

$$
\begin{aligned}
f(p \,|\, n, x = 0, \mathcal{B}) &= (n+1)\,(1-p)^n \\
p_{max} &= 0 \\
E(p) &= \frac{1}{n+2} \longrightarrow \frac{1}{n} \\
\sigma(p) &= \sqrt{\frac{(n+1)}{(n+3)(n+2)^2}} \longrightarrow \frac{1}{n} \\
p_{95\%UL} &= 1 - \sqrt[n+1]{0.05}.
\end{aligned}
$$

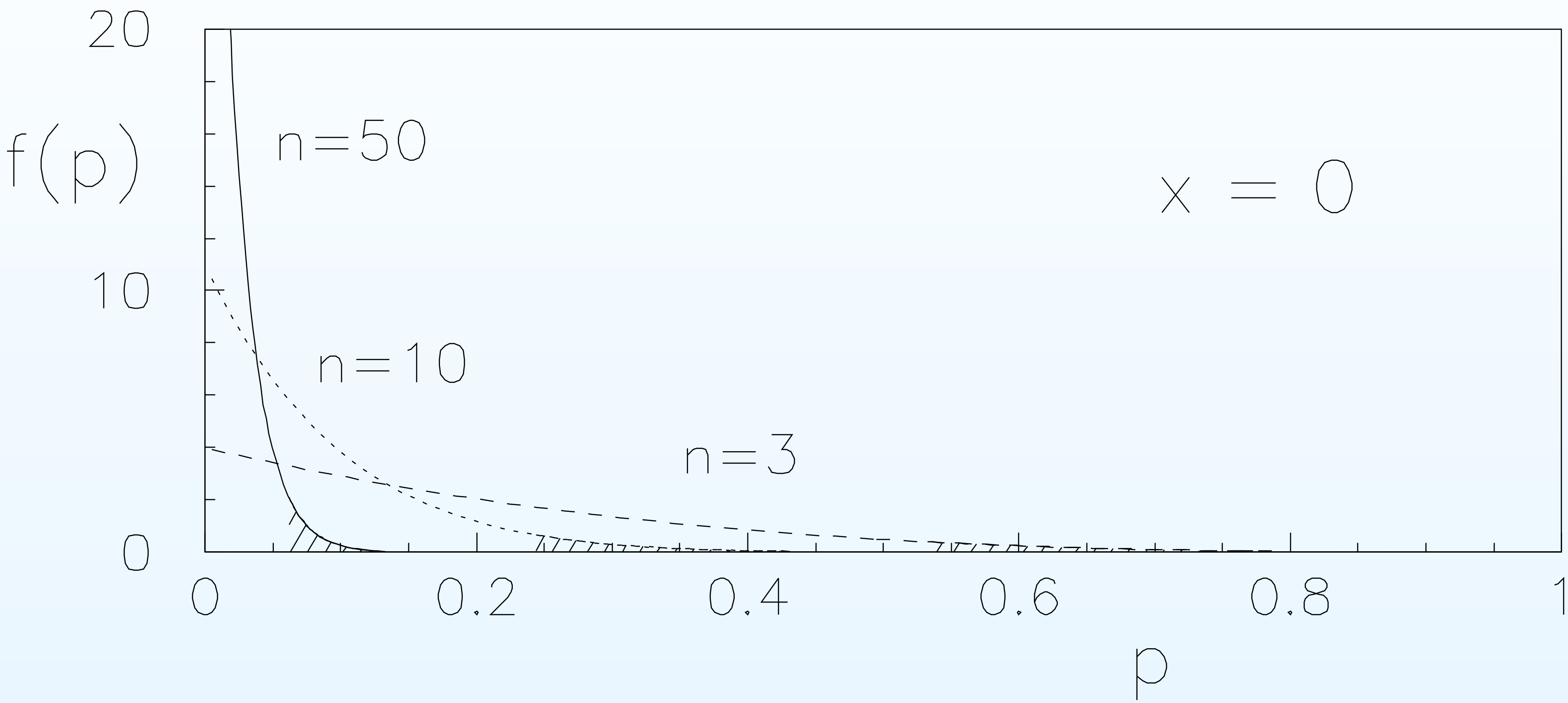
Inference about p from 0 counts

Using flat prior, i.e. $f_0(p) = k$

$$
\begin{aligned}
f(p \,|\, n, x=0, \mathcal{B}) &= (n+1)\,(1-p)^n \\
p_{max} &= 0 \\
E(p) &= \frac{1}{n+2} \rightarrow \frac{1}{n} \\
\sigma(p) &= \sqrt{\frac{(n+1)}{(n+3)(n+2)^2}} \rightarrow \frac{1}{n} \\
p_{95\%UL} &= 1 - \sqrt[n+1]{0.05}.
\end{aligned}
$$

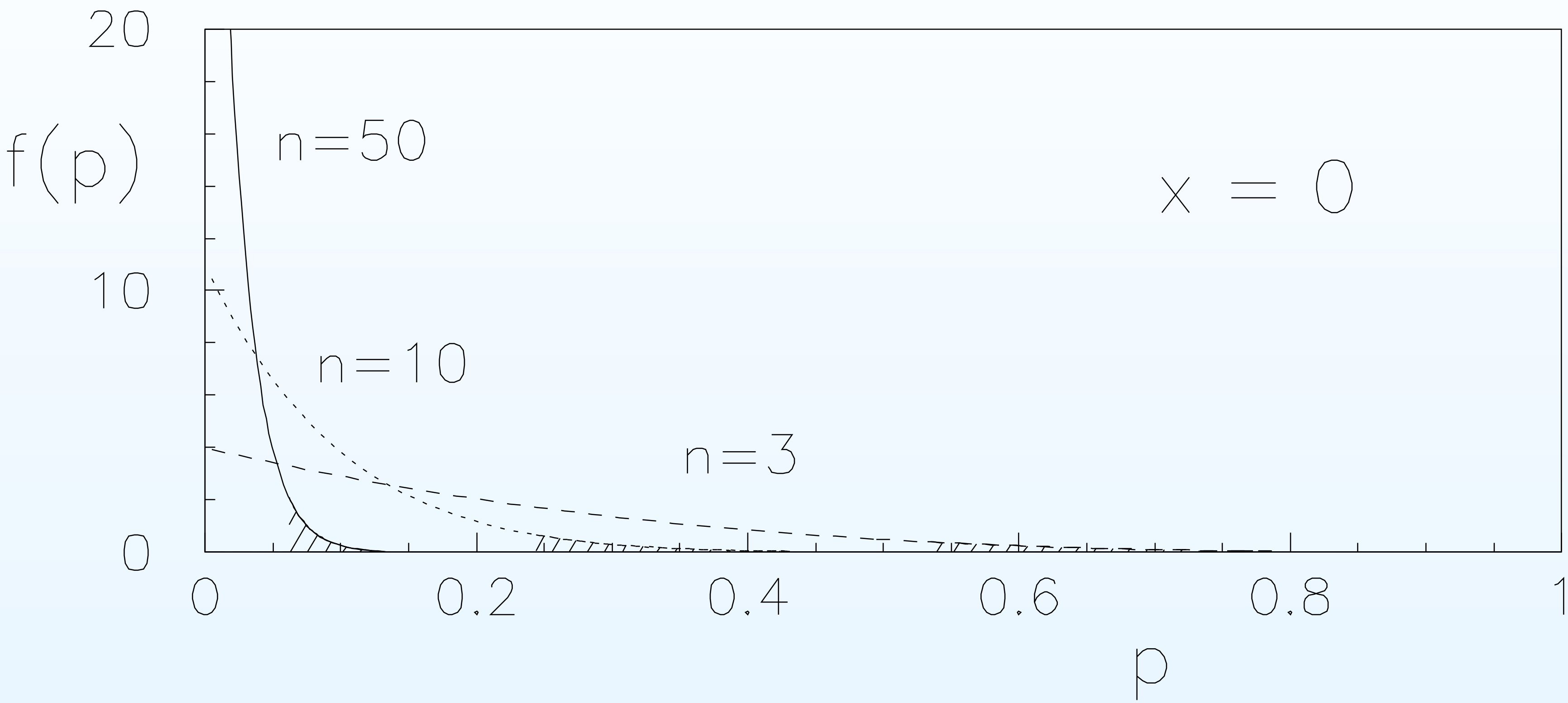
As n increases, we get more and more convinced that p has to be very small

Inference about p from 0 counts



$$f(p\,|\,n, x=0, \mathcal{B}) \;=\; (n+1)\,(1-p)^n$$
$$p_{95\%UL} \;=\; 1 - \sqrt[n+1]{0.05}.$$

Inference about p from 0 counts



Seems not problematic at all, but we have to remember that it relies on

$$f(x=0\,|\,n,\mathcal{B}) = (1-p)^n$$
$$f_0(p) = k$$

When likelihoods are non ‘closed’

Where is the problem? (Flat priors are regulary used, and are often assumed in other approaches, e.g. ML methods)

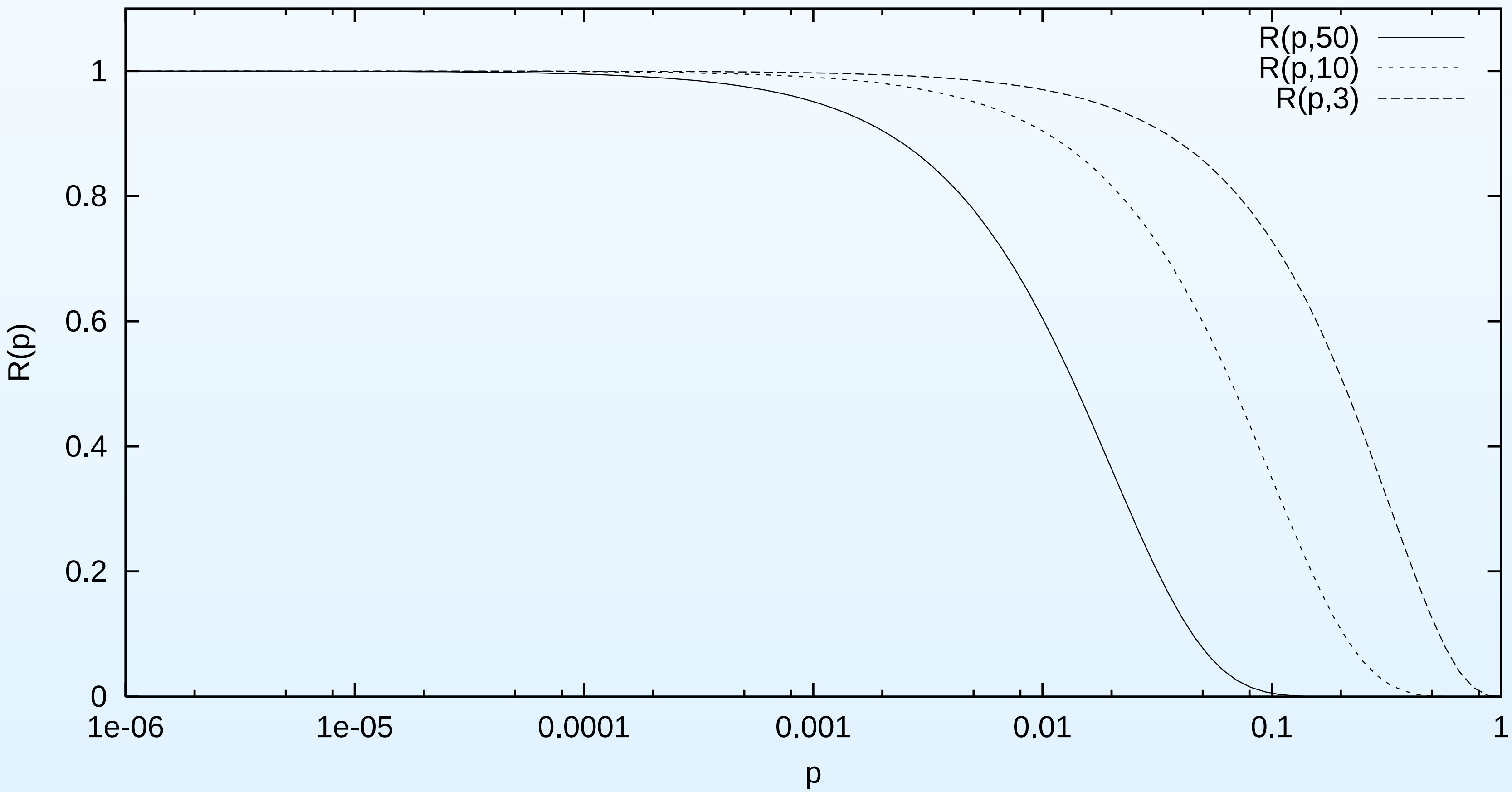
When likelihoods are non ‘closed’

The major problem is not in $f_0(p)$, but rather in the likelihood $f(x=0, \mid n, \mathcal{B})$ that does not go to zero on both sides!

When likelihoods are non ‘closed’

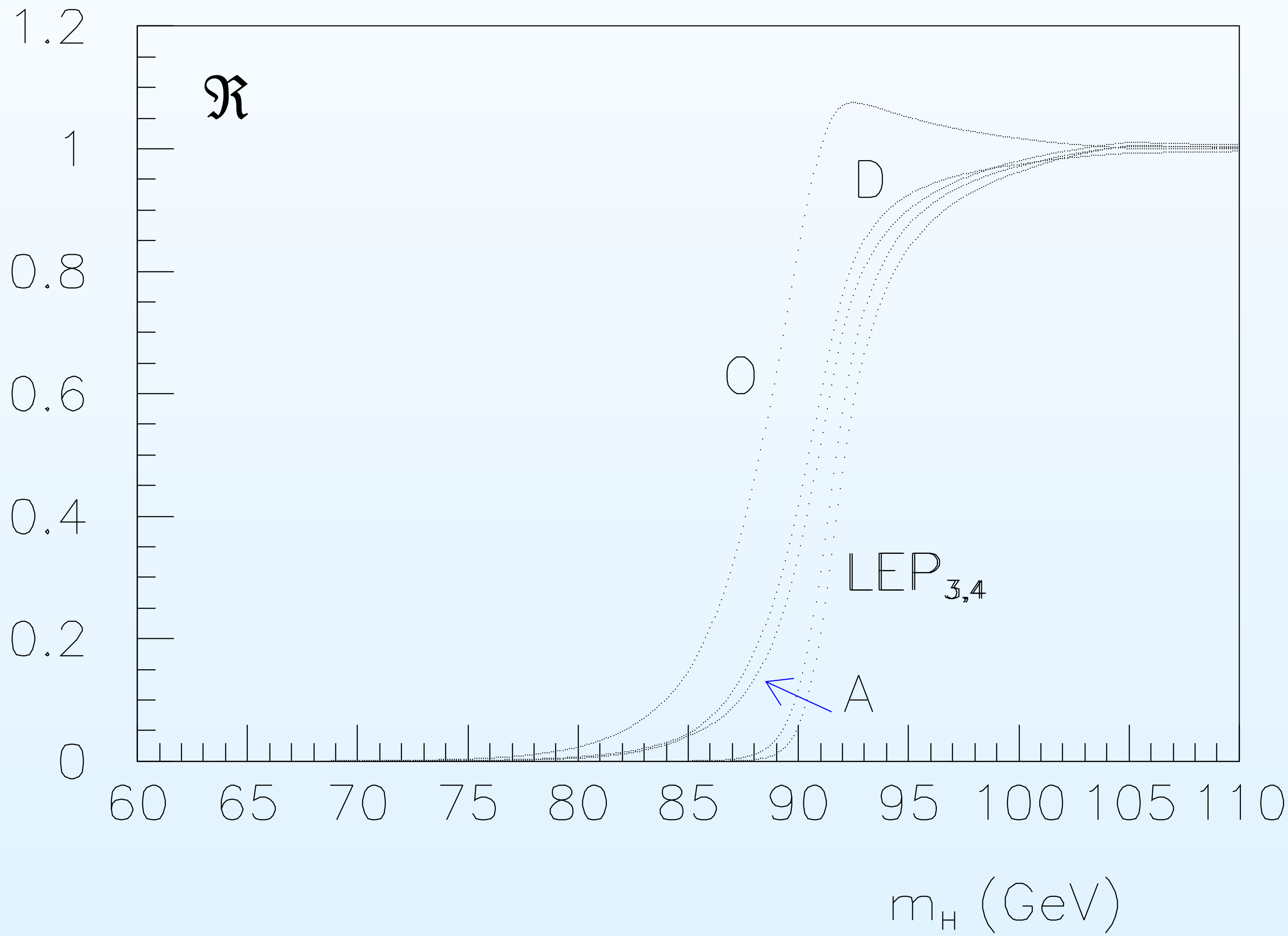
The major problem is not in $f_0(p)$, but rather in the likelihood $f(x=0,\,|\,n,\mathcal{B})$ that does not go to zero on both sides!

A different representation of the likelihood (properly rescaled) helps:



A probabilistic lower bound for the Higgs?

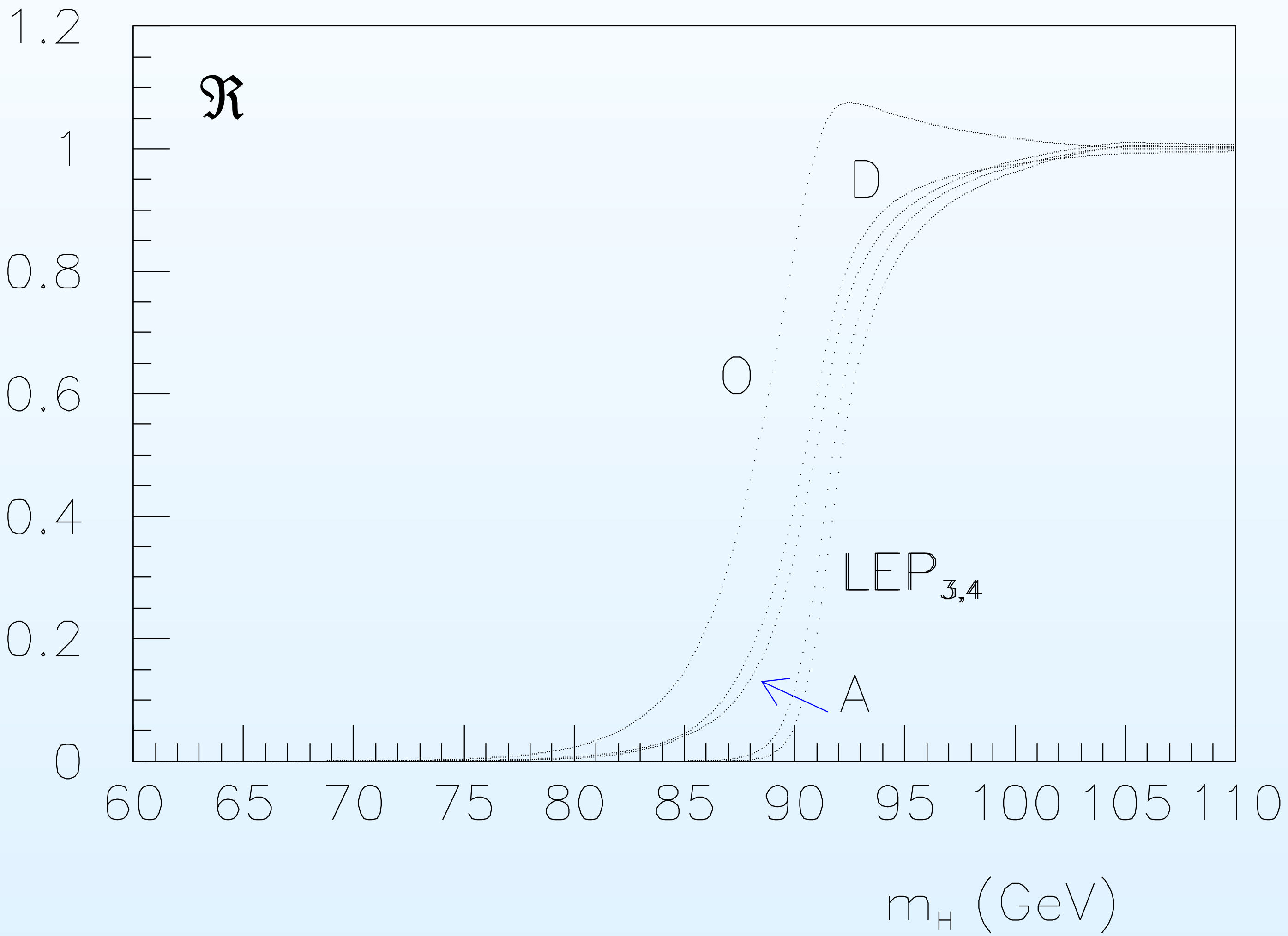
A similar think happens with the direct searches of the Higgs particle at LEP



(1999 figure, but substance unchanged)

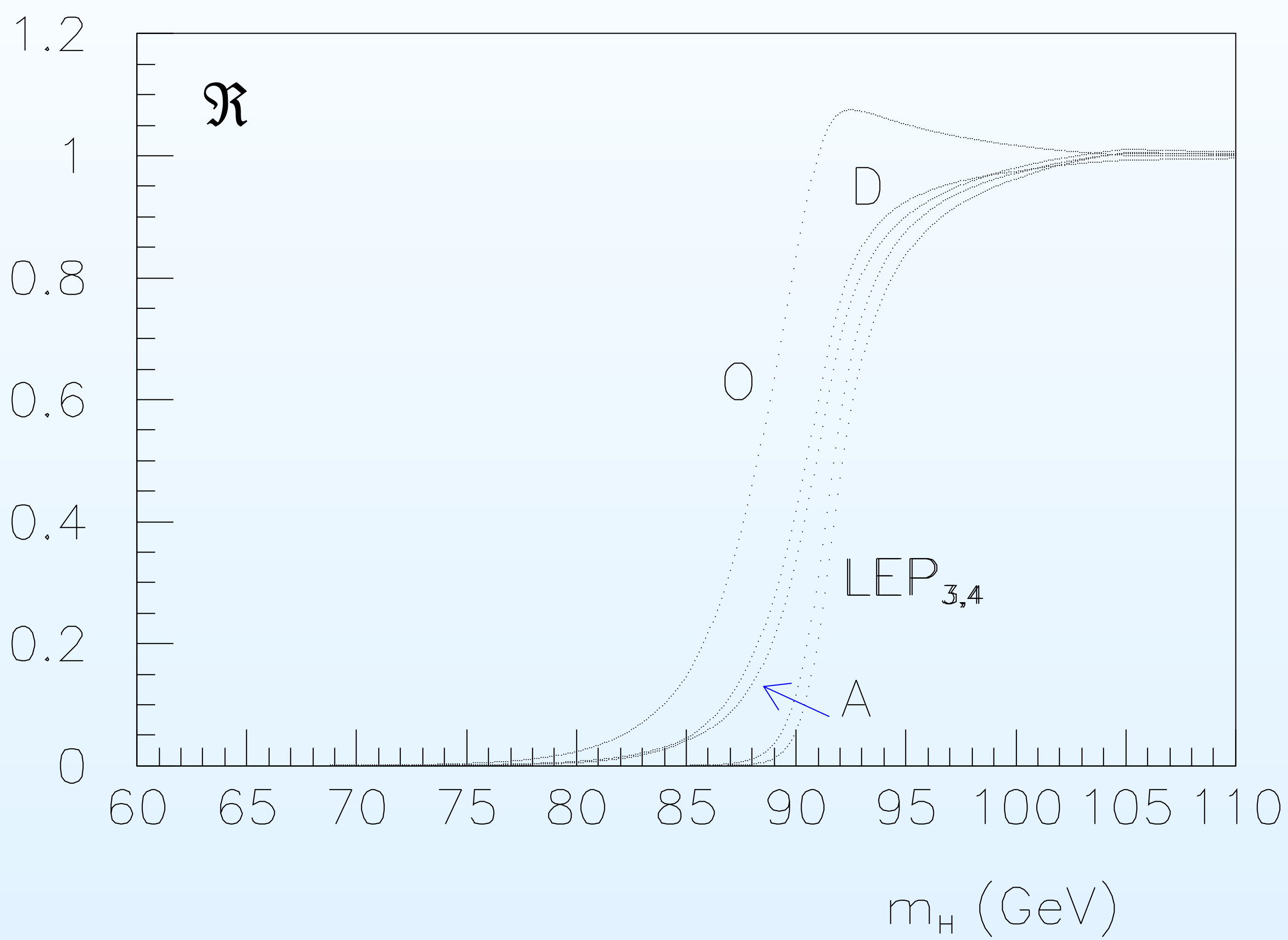
A probabilistic lower bound for the Higgs?

Impossible to express our confidence in probabilistic terms, unless we define an upper cut!



A probabilistic lower bound for the Higgs?

Confidence limit $\Rightarrow$ Sensitivity bound



Conclusions

- Probabilistic reasoning helps . . .
- . . . at least to avoid conceptual errors.

- Several ‘standard’ methods (like Least Square, etc.) can be easily recovered under well defined assumptions.

- But if this is not the case, nowdays there are no longer excuses to avoid the more general approach.

- Bayesian networks are a powerful conceptual and computational tool.

BAT - the Bayesian Analysis Toolkit

http://mppmu.mpg.de/bat/

End

FINE