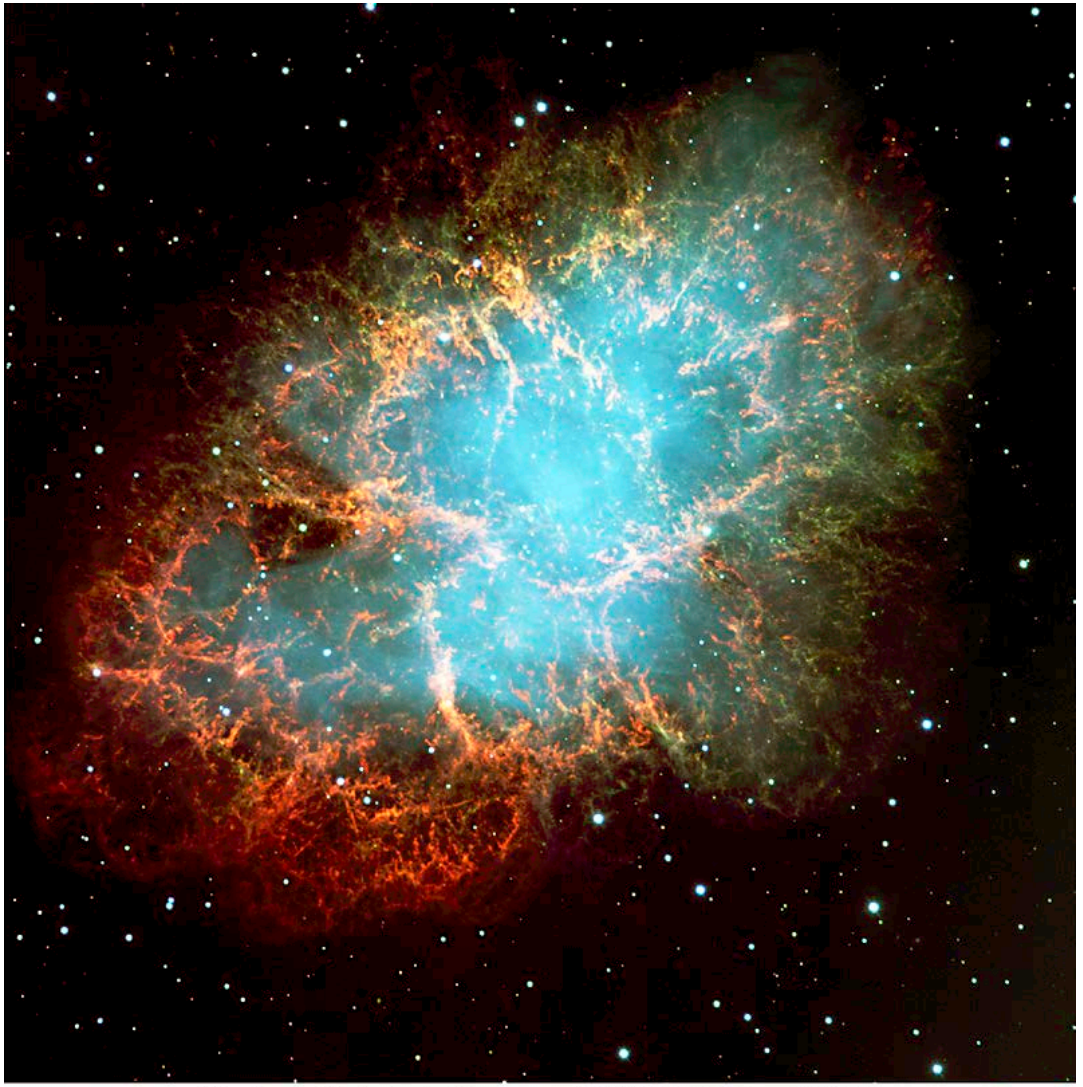
Introduction to Bayesian Statistics - 7

Edoardo Milotti

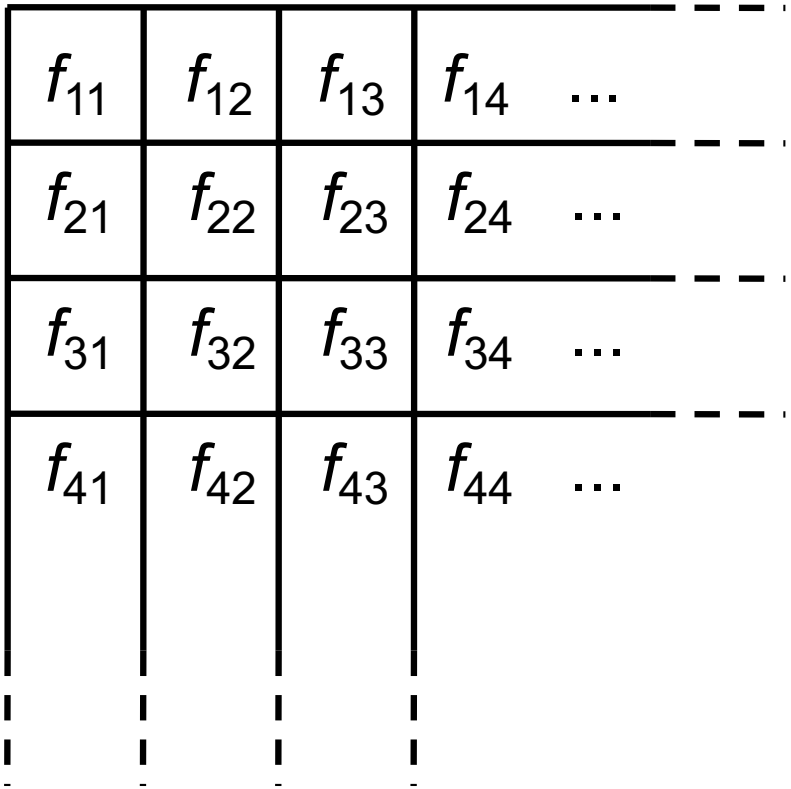
Università di Trieste and INFN-Sezione di Trieste

1. Image processing techniques (MLM, MEM)



The Crab Nebula in Taurus (VLT KUEYEN + FORS2)

ESO PR Photo 40f/99 (17 November 1999)



pixel map

true pixel vector **f**

posterior pixel distribution

likelihood

a priori pixel distribution

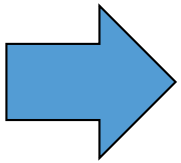
$$P(\mathbf{f}|\mathbf{g}) = \frac{P(\mathbf{g}|\mathbf{f})}{P(\mathbf{g})} P(\mathbf{f}) \propto P(\mathbf{g}|\mathbf{f}) P(\mathbf{f})$$

Bayesian estimate of true pixel vector from observed pixel vector

We estimate the true pixel distribution taking the pixel vector that maximizes the posterior distribution (MAP estimate: Maximum A Posteriori estimate).

This depends on the prior distribution

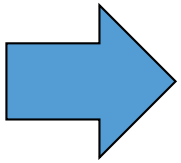
$P(\mathbf{f})$ flat prior



Maximum Likelihood Estimate (MLE)

$$P(\mathbf{f}|\mathbf{g}) \propto P(\mathbf{g}|\mathbf{f})P(\mathbf{f}) \propto P(\mathbf{g}|\mathbf{f})$$

$P(\mathbf{f})$ entropic prior



Maximum Entropy Method (MEM)

Notice that

$$\log P\left(\mathbf{f}|\mathbf{g}\right) \approx \log P\left(\mathbf{g}|\mathbf{f}\right) - \left[-\log P\left(\mathbf{f}\right)\right]$$

therefore we obtain the estimate $\hat{\mathbf{f}}$ by maximizing the likelihood with the *penalty function*

$$\left[-\log P\left(\mathbf{f}\right)\right]$$

Experiments have been tried with many different penalties, many of them barely justified on probabilistic grounds (or not at all!)

Let ***f*** be the vector of "true values" (uncorrupted intensities of an image, a spectrum, etc. ...), and translate these values into counts

$$n_i = \lfloor \alpha f_i \rfloor$$

(i = 1, ..., M). The least informative prior is that for a structureless image is uniform, and the probability of one count at the i-th position is just 1/M.

Likewise, the probability of a given vector of values where the total number of counts is N, is given by the multinomial probability

$$P(\mathbf{n}) = \frac{N!}{n_1!n_2!\ldots n_M!}\left(\frac{1}{M}\right)^N; \qquad \sum_k n_k = N$$

Using Stirling's approximation

$$n! \approx n^n e^{-n} \qquad \ln n! \approx n \ln n - n$$

we find, with the definition $p_i = f_i \Big/ \sum_{k=1}^{M} f_k$

$$\ln P(\mathbf{n}) \approx (N \ln N - N) - \sum_{i=1}^{M} (n_i \ln n_i - n_i)$$

$$= N \ln N - \sum_{i=1}^{M} n_i \ln n_i$$

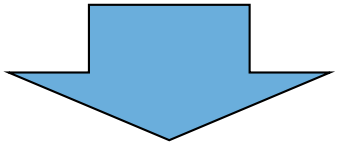
$$\approx -\alpha \sum_{i=1}^{M} f_i \ln f_i + \text{cost}.$$

entropic prior

$$P(\mathbf{n}) \propto \exp\left[-\alpha \sum_{i=1}^{M} f_i \ln f_i\right] \propto \exp\left[-\alpha \sum_{i=1}^{M} p_i \ln p_i\right] = \exp\left[\alpha S(\mathbf{f})\right]$$

Using the entropic prior and Bayes' theorem we find

$$P(\mathbf{f}) \propto \exp\left[\alpha S(\mathbf{f})\right]$$

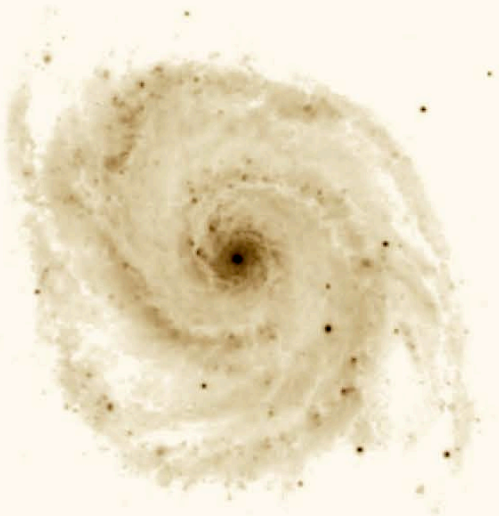


$$P(\mathbf{f}|\mathbf{g}) \propto P(\mathbf{g}|\mathbf{f})P(\mathbf{f}) \propto P(\mathbf{g}|\mathbf{f})\exp\left[\alpha S(\mathbf{f})\right]$$

$$\log P(\mathbf{f}|\mathbf{g}) \approx \log P(\mathbf{g}|\mathbf{f}) + \alpha S(\mathbf{f})$$

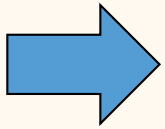
therefore we find the combination of pixels (i.e., the **f** vector) that maximizes the posterior distribution by maximizing a linear combination of likelihood and Shannon's entropy.

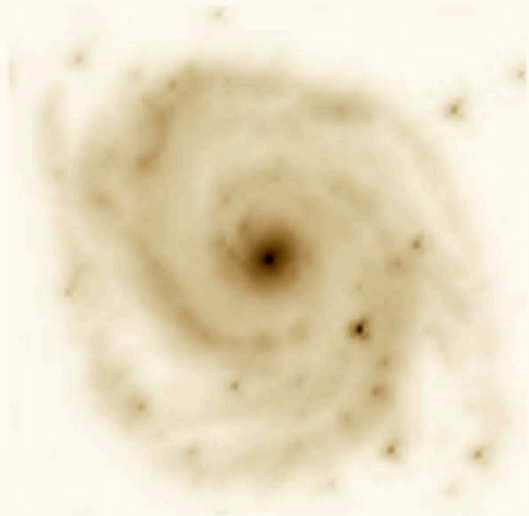
Image likelihood: 1. the observation model



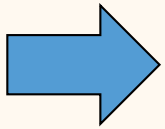
true image
of a galaxy

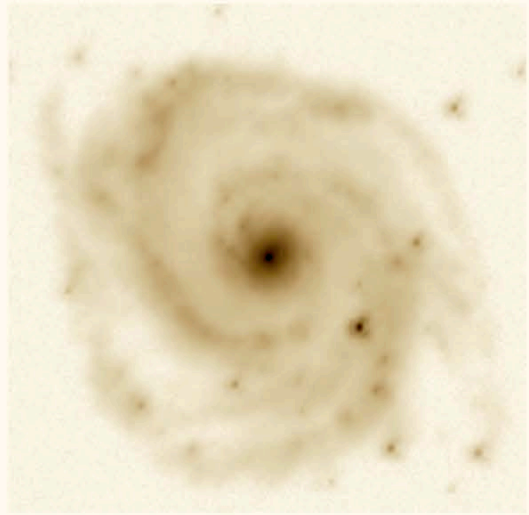
PSF
(Point Spread
Function





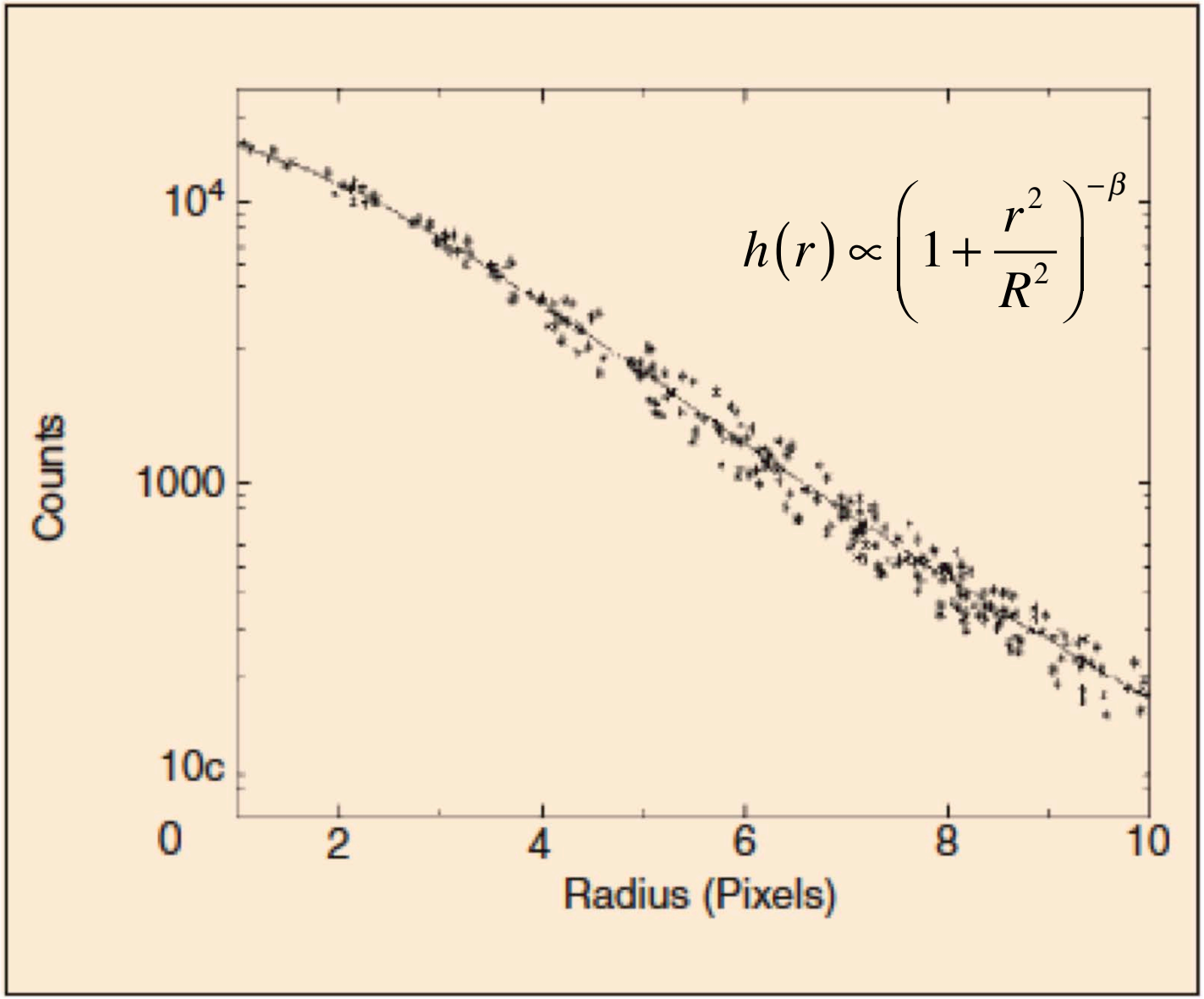
Noise





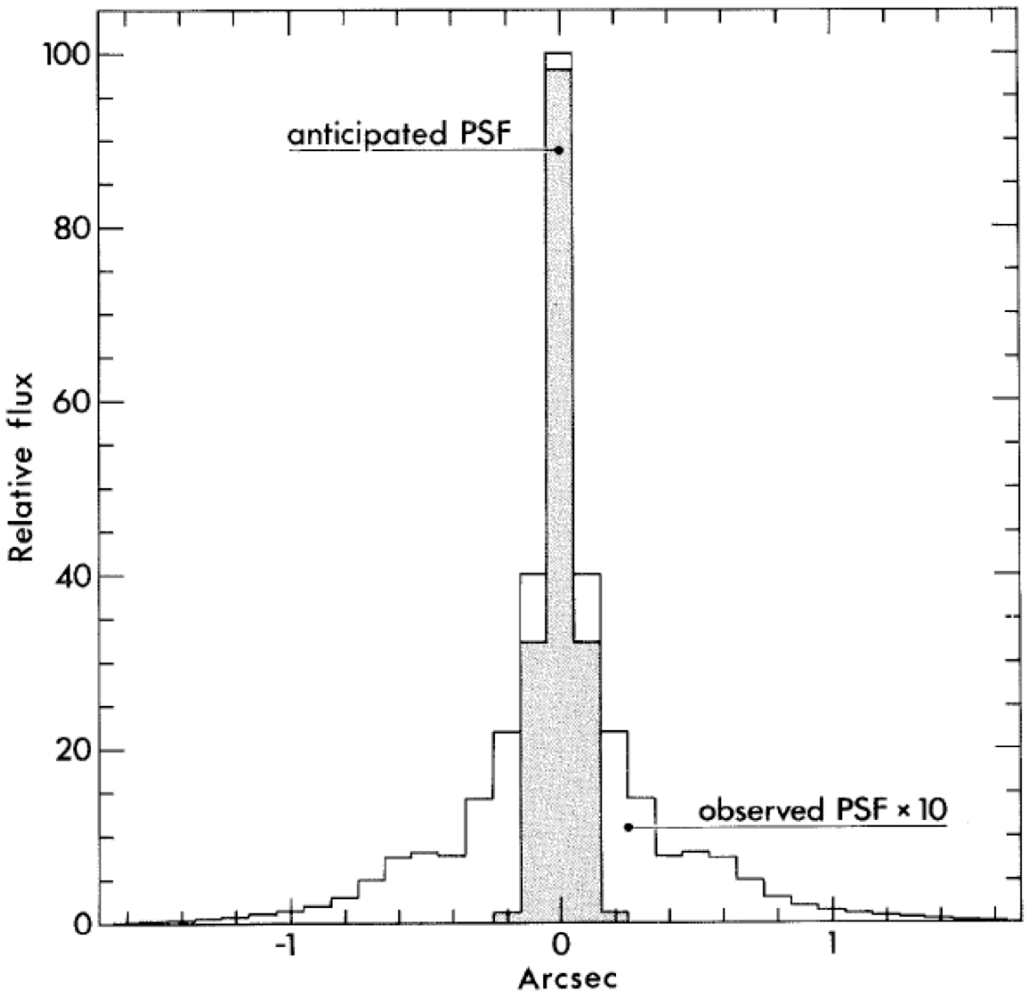
(example from Eric Thiebaut)

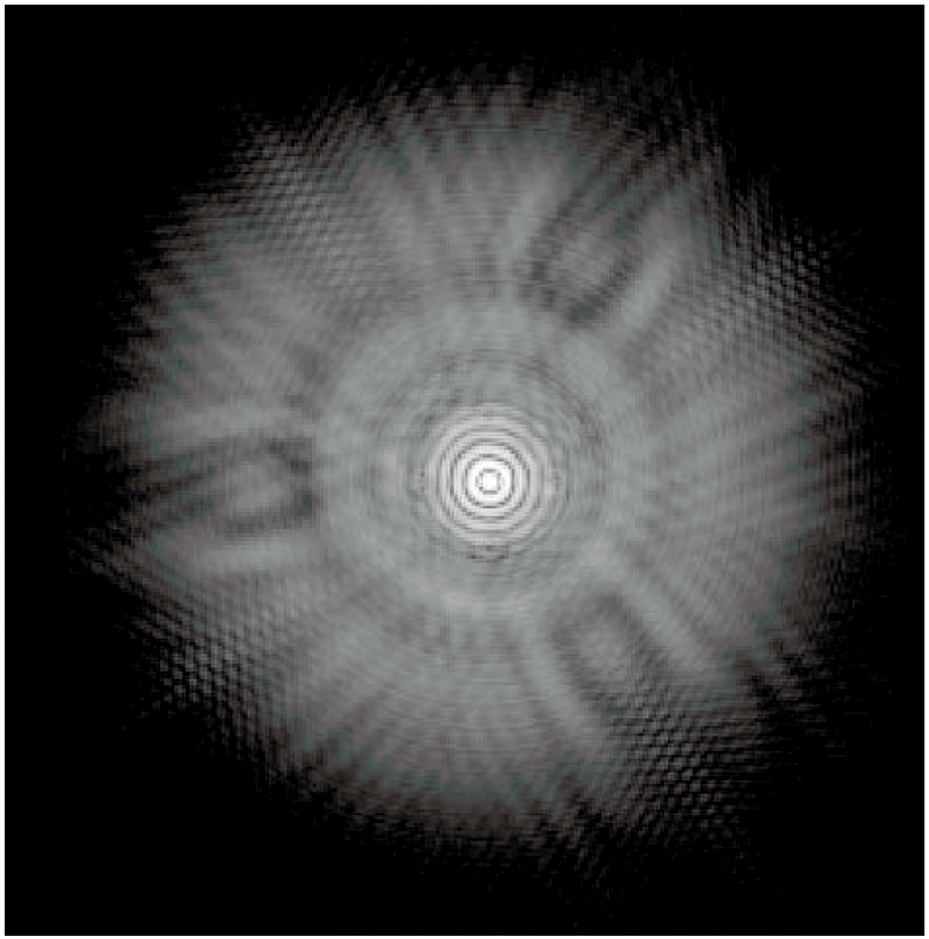
“dirty image”



PSF from atmospheric turbolence

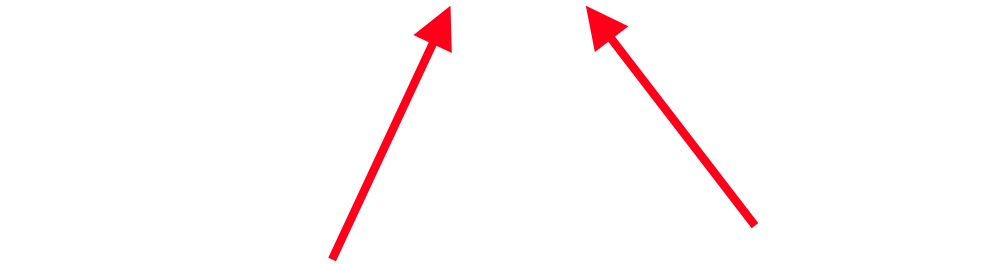
The Hubble PSF before the first servicing mission





In general the effect of the PSF is modeled by a linear operator

$$\mathbf{f} \rightarrow \mathbf{Hf}$$



action of optical system on true image is modeled by matrix **H**

"true" pixel vector

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

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

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

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

sometimes we can use the Gaussian approximation of Poisson noise

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

$$\approx \prod_n \exp\left[-\frac{\left(g_n - (\mathbf{Hf})_n\right)^2}{2(\mathbf{Hf})_n}\right]$$

$$= \exp\left[-\sum_n \frac{\left(g_n - (\mathbf{Hf})_n\right)^2}{2(\mathbf{Hf})_n}\right]$$

Gaussian noise only:

maximize linear combination of entropy and chi-square

$$\log P(\mathbf{f}|\mathbf{g}) \approx \alpha S(\mathbf{f}) - \frac{(\mathbf{g}-\mathbf{H}\mathbf{f})^2}{\sigma^2}$$

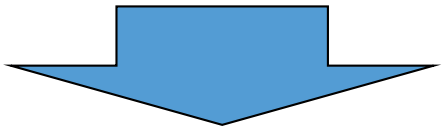
$$= \alpha S(\mathbf{f}) - \sum_n \frac{\left(g_n - (\mathbf{H}\mathbf{f})_n\right)^2}{\sigma^2}$$

$$= \alpha S(\mathbf{f}) - \chi^2(\mathbf{f})$$

Combined noise model

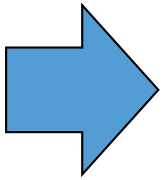
detector noise: Poisson noise
electronic noise: Gaussian noise

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



maximize

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



numerical maximization procedure

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

noise model: Poisson noise
prior: flat prior

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

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

maximize this posterior distribution

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

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

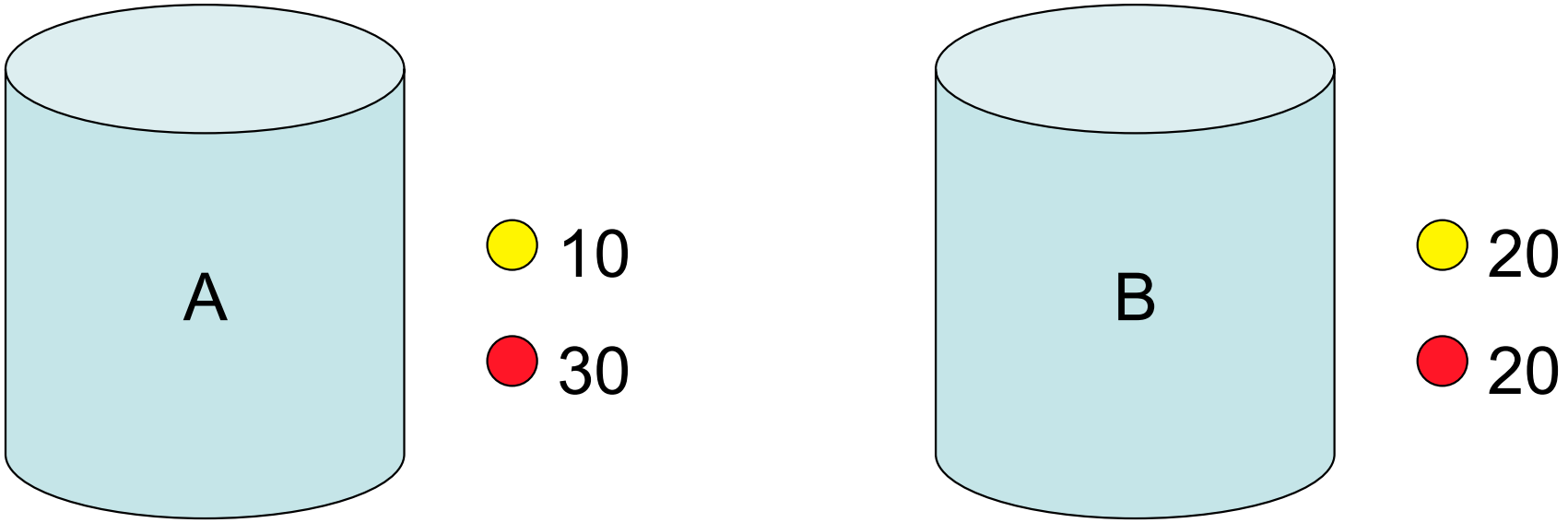
▲ *9. Reconstruction of the image of Saturn using the R-L algorithm.*

NGC 604 in Spiral Galaxy M33

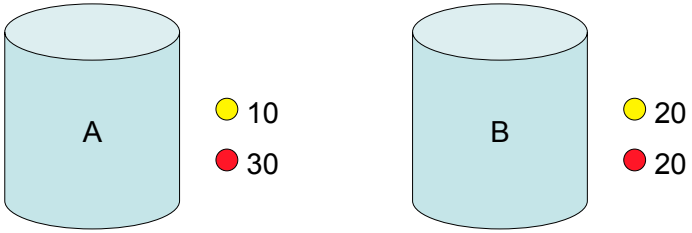
Edoardo Milotti - Bayesian Methods - May 2020

Hubble Heritage

2. A game of boxes: a simple example of Bayesian inference



- box A contains 10 yellow balls and 30 red balls;
- box B contains 20 yellow balls and 20 red balls;

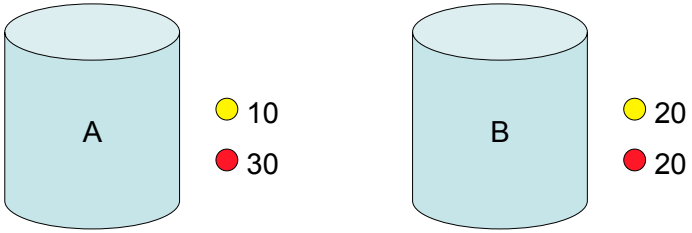


When we choose one of the two boxes at random, the probability of choosing a given box is 1/2

$$P(A|I) = P(B|I) = 1/2$$

and the probabilities of extracting a ball of a given color – yellow (Y) or red (R) – depend on the chosen box

$$P(Y|A, I) = 1/4; \quad P(R|A, I) = 3/4$$



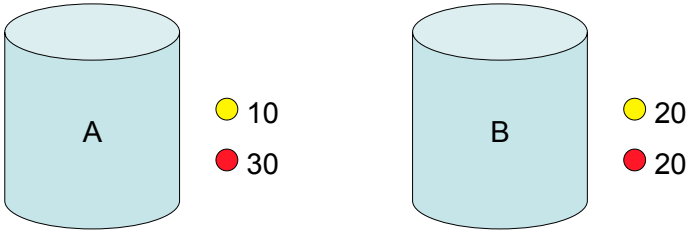
Guessing the box name

Now we play the reverse game with the help of a friend.

The friend hides the names of the boxes and places them in front of us. We extract one ball from the box on the left and we use all our prior knowledge and Bayes' theorem to infer its name.

Clearly, at the outset – i.e., *a priori*, given the information *I* that we initially have – the probability that the box on the left is actually box A is just $p_0(L = A|I) = 1/2$, because the two boxes A and B are equally probable.

We start the game by extracting the first ball.



Extraction 1

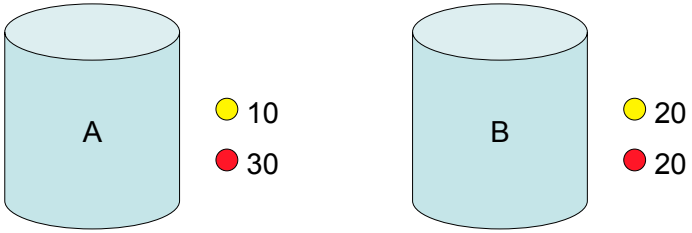
the ball is red. Using the prior information we find that the probabilities of extraction of a red or a yellow ball from any of the two boxes (the *evidences*) are

$$p_0(R) = P(R|L = A, I)p_0(L = A, I) + P(R|L = B, I)p_0(L = B, I) = 5/8 = 0.625$$

$$p_0(Y) = P(Y|L = A, I)p_0(L = A, I) + P(Y|L = B, I)p_0(L = B, I) = 3/8 = 0.375$$

Therefore, using Bayes' theorem, the posterior probability for A is

$$p_1(L = A|R, I) = \frac{P(R|L = A, I)}{p_0(R)} p_0(L = A|I) = \frac{3/4}{5/8}(1/2) = 3/5 = 0.6$$



After the first extraction the ball is reinserted in the box from which it was taken and we are ready for the next extraction.

Extraction 2

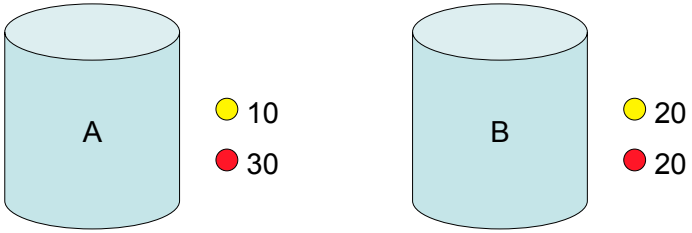
the ball is red, again. Then we use the old posterior probabilities as the new priors and we find

$$p_1(R) = P(R|L = A, I)p_1(L = A, I) + P(R|B, I)p_1(L = B, I) = 0.65$$

$$p_1(Y) = P(Y|L = A, I)p_1(L = A, I) + P(Y|B, I)p_1(L = B, I) = 0.35$$

and a repeated application of Bayes' theorem yields:

$$p_2(L = A|\{R, R\}, I) = \frac{P(R|L = A, I)}{p_1(R)} p_1(L = A|R, I) = 0.692308$$



Extraction *n*

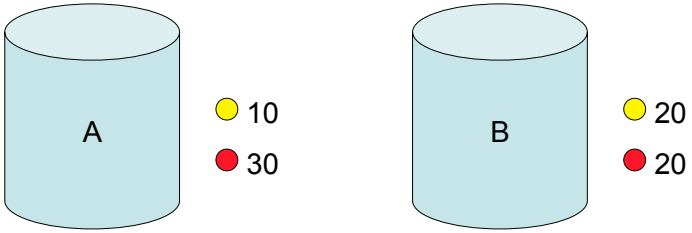
in the generic case we consider the *n*-th extraction and we do not specify explicitly the colors of the extractions. Using the posterior probabilities of the previous step we find

$$p_n(D_n|\{D_k\}_{k=1,n-1},I) = P(D_n|L=A,I)p_{n-1}(L=A|\{D_k\}_{k=1,n-1},I) + P(D_n|L=B,I)p_{n-1}(L=B|\{D_k\}_{k=1,n-1},I)$$

and from Bayes' theorem we find the new posterior probabilities:

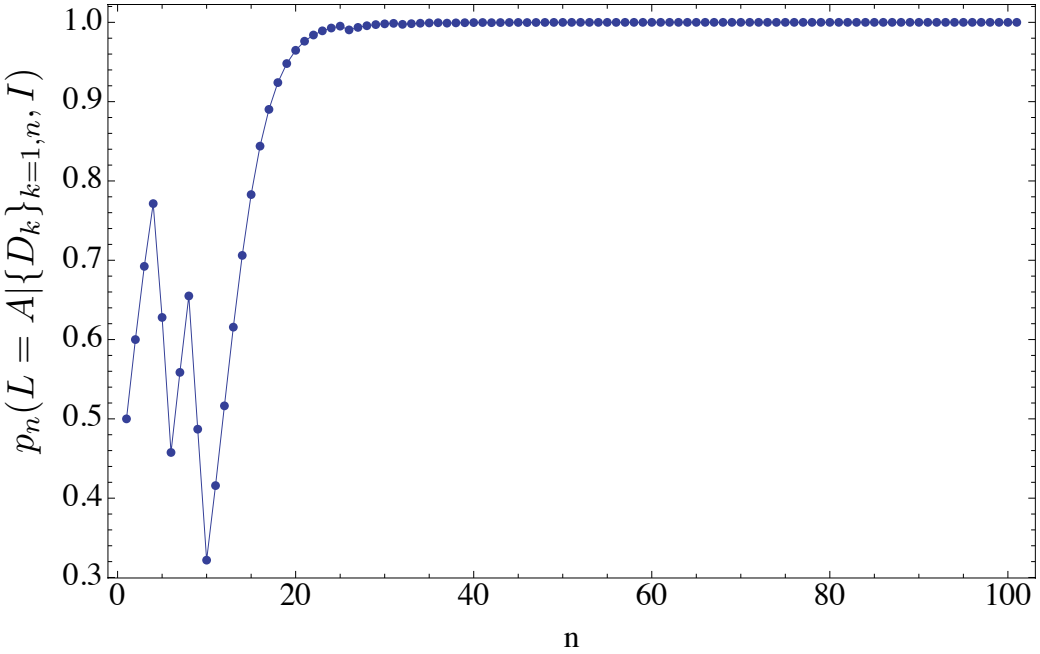
$$p_n(L=A|\{D_k\}_{k=1,n},I) = \frac{P(D_n|L=A,I)}{p_n(D_n|\{D_k\}_{k=1,n-1},I)}p_{n-1}(L=A|\{D_k\}_{k=1,n-1},I)$$

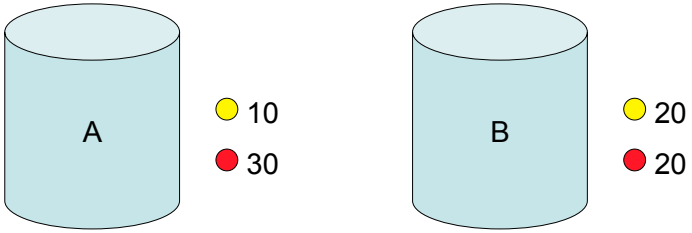
$$p_n(L=B|\{D_k\}_{k=1,n},I) = \frac{P(D_n|L=B,I)}{p_n(D_n|\{D_k\}_{k=1,n-1},I)}p_{n-1}(L=B|\{D_k\}_{k=1,n-1},I)$$



100 extractions

R, R, R, Y, Y, R, R, Y, Y, R, R, R, R, R, R, R, R, R, R, R, R, R, R, Y,
R, R, R, R, R, Y, R, R, R, R, Y, R, R, R, R, Y, R, R, R, Y, R, R, R, R, R,
R, Y, R, R, Y, R, R, R, R, R, R, Y, R, R, R, R, Y, R, R, Y, R, Y, R, R, Y,
Y, R, R, Y, R, R, R, Y, R, R, Y, R, R, R, R, R, R, R, R, R, R, Y, Y, R, R

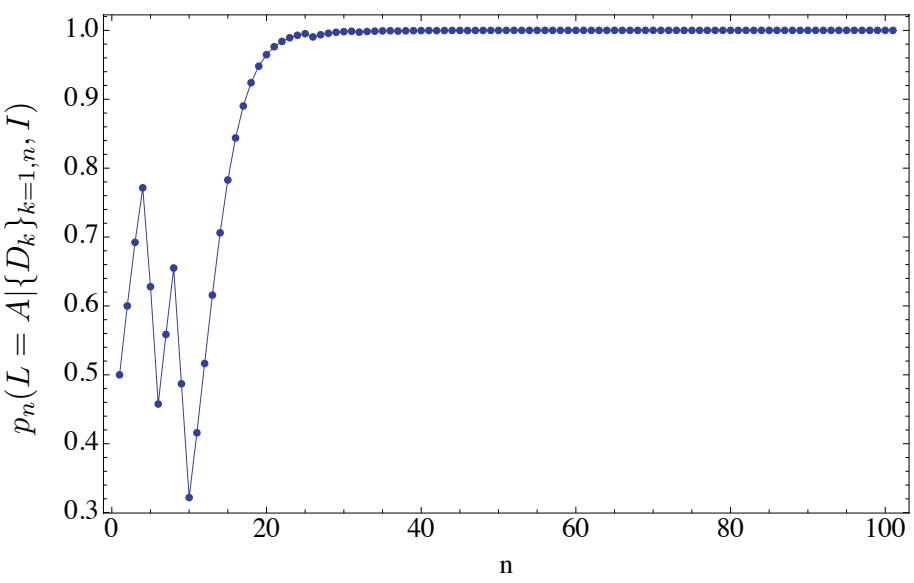


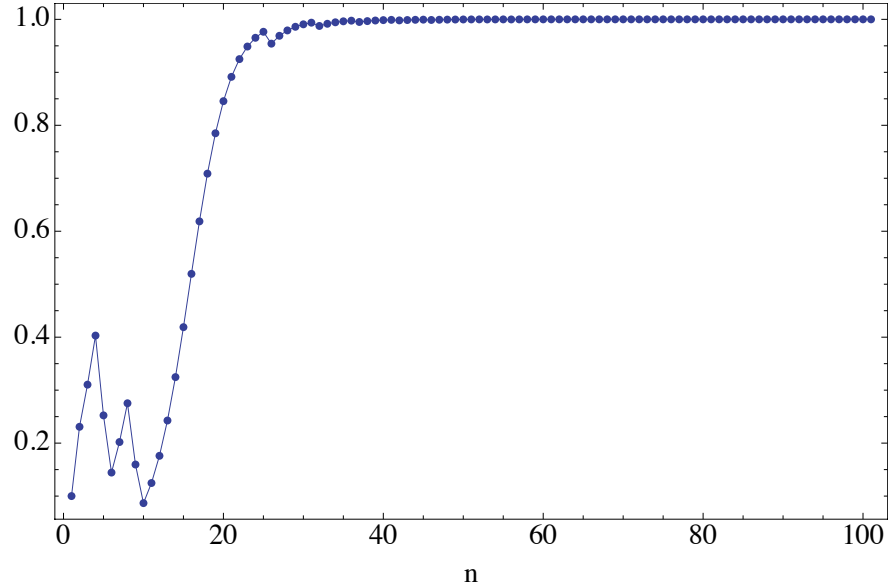


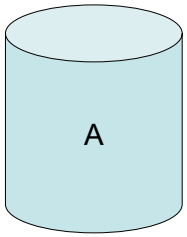
A different initial prior

$$p_0(L = A|I) = p_0(L = B|I) = 1/2,$$

$$p_0(L = A|I) = 0.1; \quad p_0(L = B|I) = 0.9$$







?

What if we do not know the number of balls?

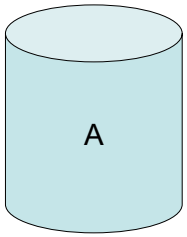
We only know the total number of balls (N-1), and the number of red balls ranges from 0 to N-1. Then

likelihood

$$P(Y|N_R, I) = (1 - N_R/N); \quad P(R|N_R, I) = N_R/N$$

and *we have N competing hypotheses*. Here we must use this version of Bayes theorem

$$P(A_{N_R}|D, I) = \frac{P(D|A_{N_R}, I)}{\sum_n P(D|A_n, I)P(A_n)} P(A_{N_R}|I)$$

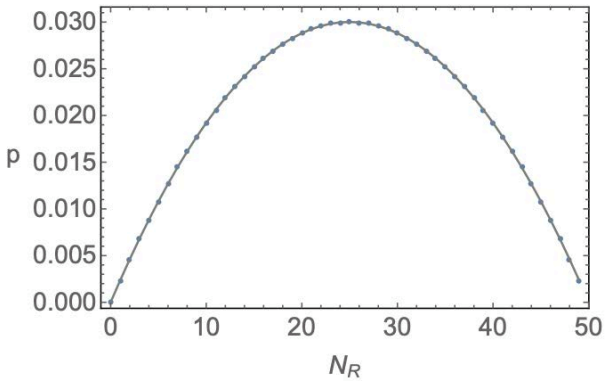


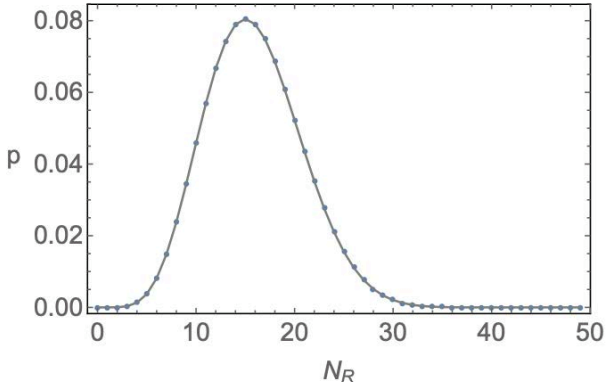
?

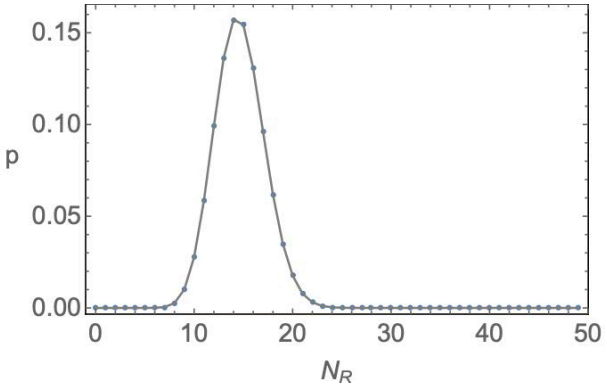
100 extractions

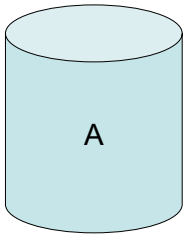
Y, R, Y, R, R, Y, Y, Y, Y, Y, Y, Y, R, Y, Y, R, Y, R, Y, Y, Y, R, Y, Y, R,
Y, Y, Y, Y, Y, Y, Y, Y, Y, R, Y, R, Y, Y, R, Y, Y, Y, Y, R, Y, R, R, R, Y,
R, Y, R, Y, Y, Y, R, Y, Y, R, Y, Y, Y, R, Y, Y, Y, Y, Y, Y, Y, Y, R, R, R,
Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, R, R, Y, Y, Y, Y, Y, Y, Y, R, Y, Y, Y, Y, R.

probability distributions for the different hypotheses after 2, 20, 80 draws



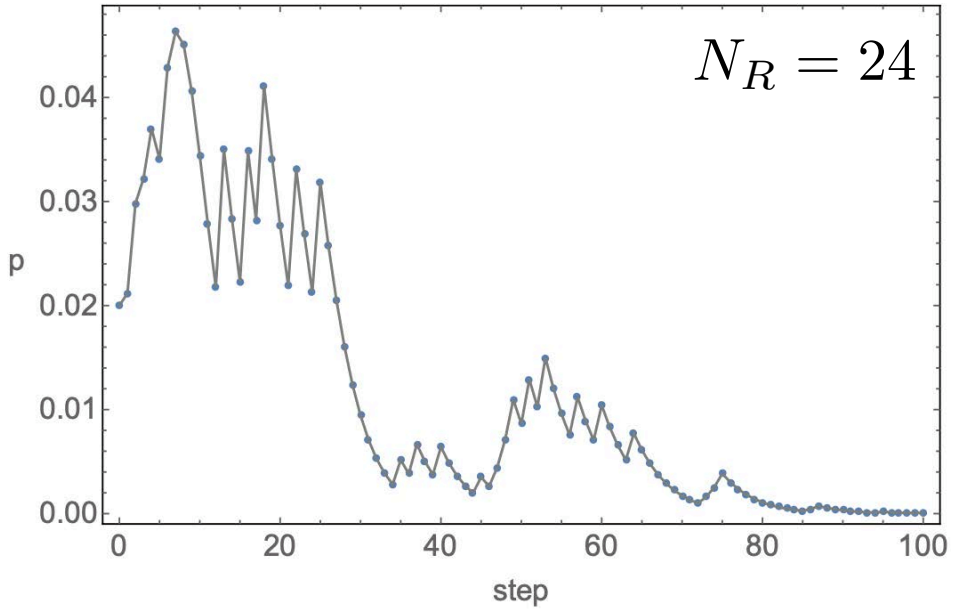


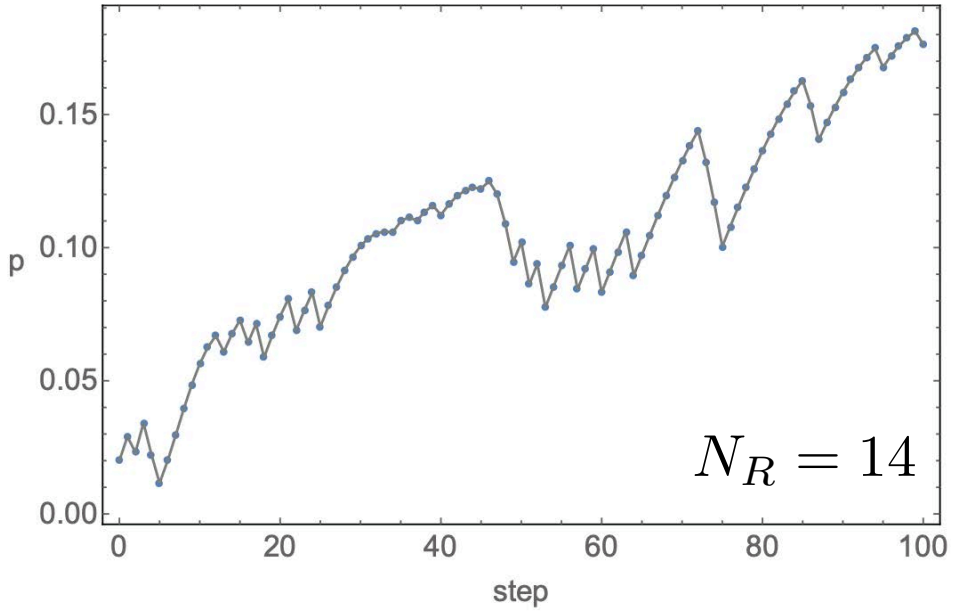


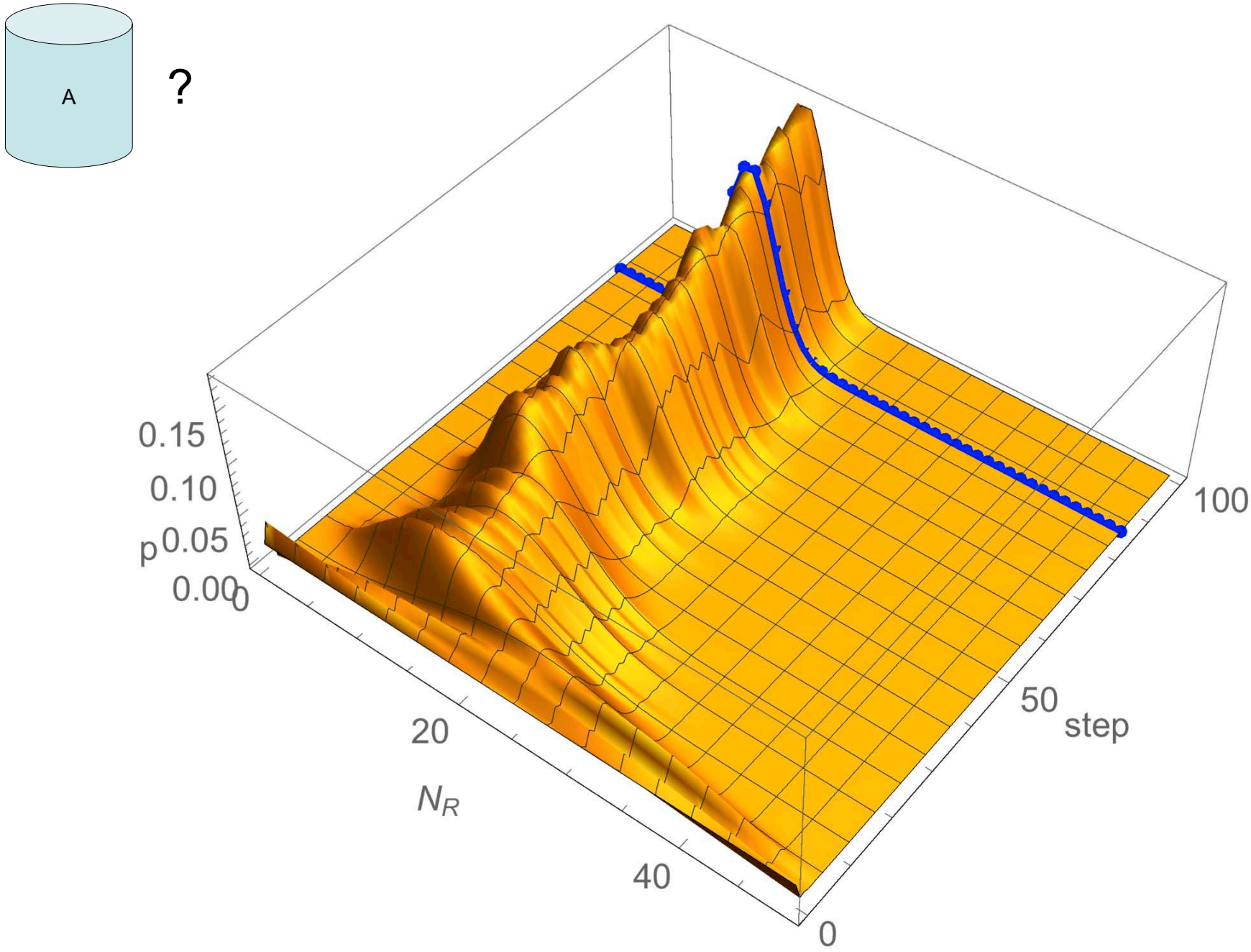


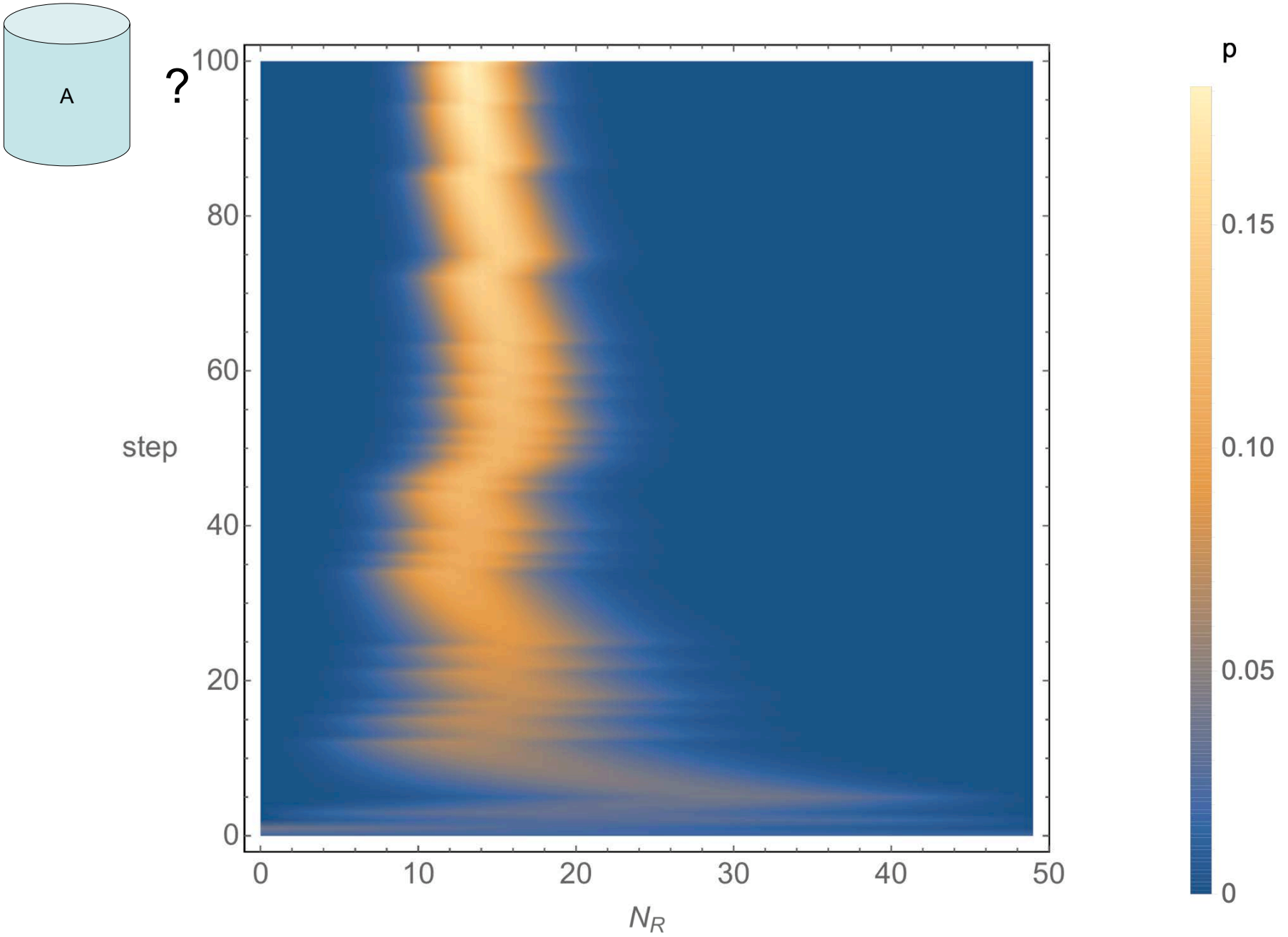
?

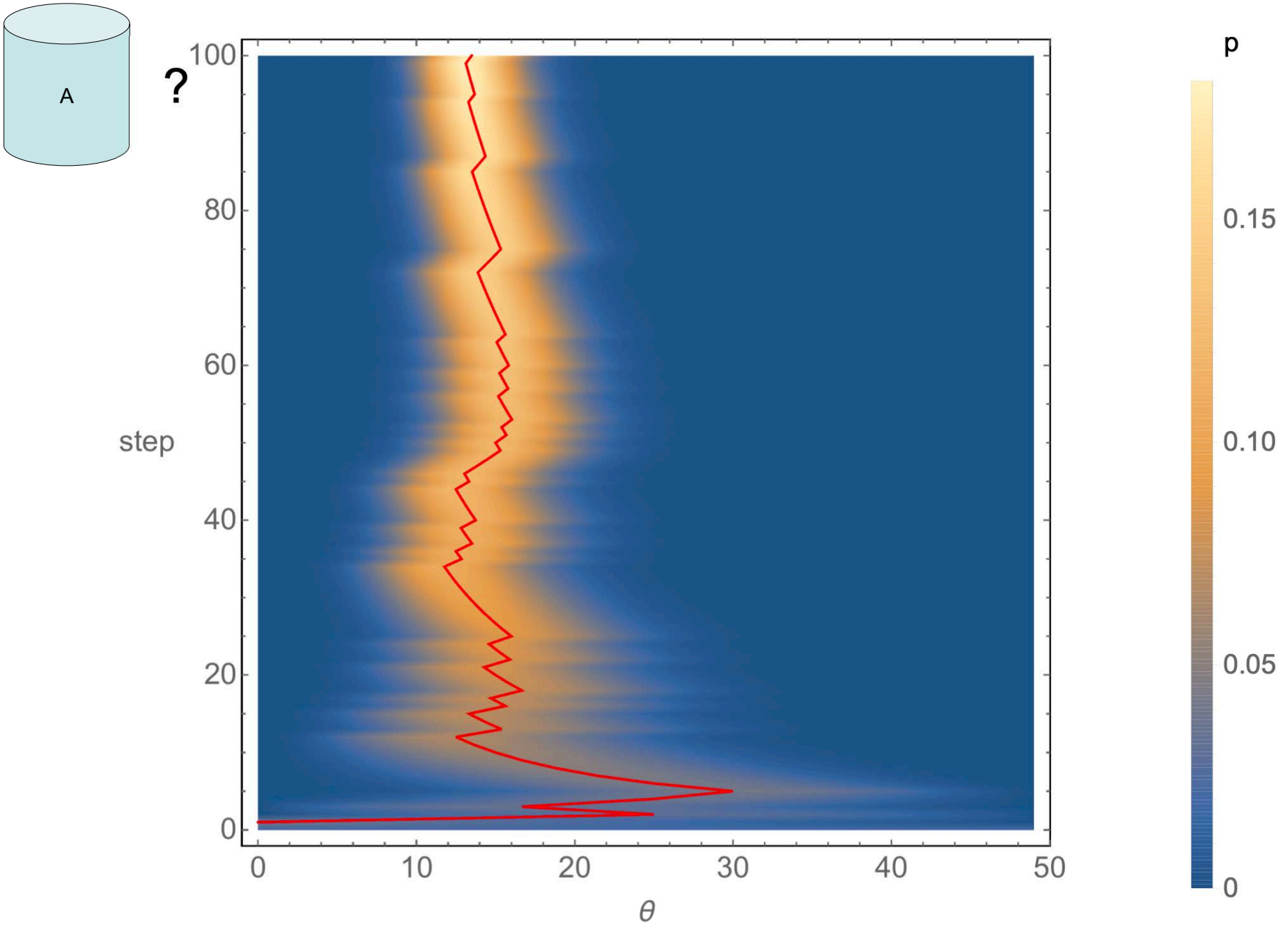
Evolution of the posterior probability as a function of the number of draws

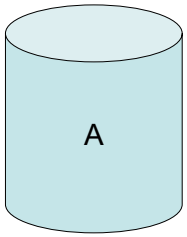





A
?
0.15
0.10
p 0.05
0.00
0
20
N_R
40
0
50
step
100


A
?
100
80
60
40
20
0
step
0
10
20
30
40
50
N_R
p
0.15
0.10
0.05
0


A
?
100
80
60
40
20
0
step
0
10
20
30
40
50
θ
p
0.15
0.10
0.05
0

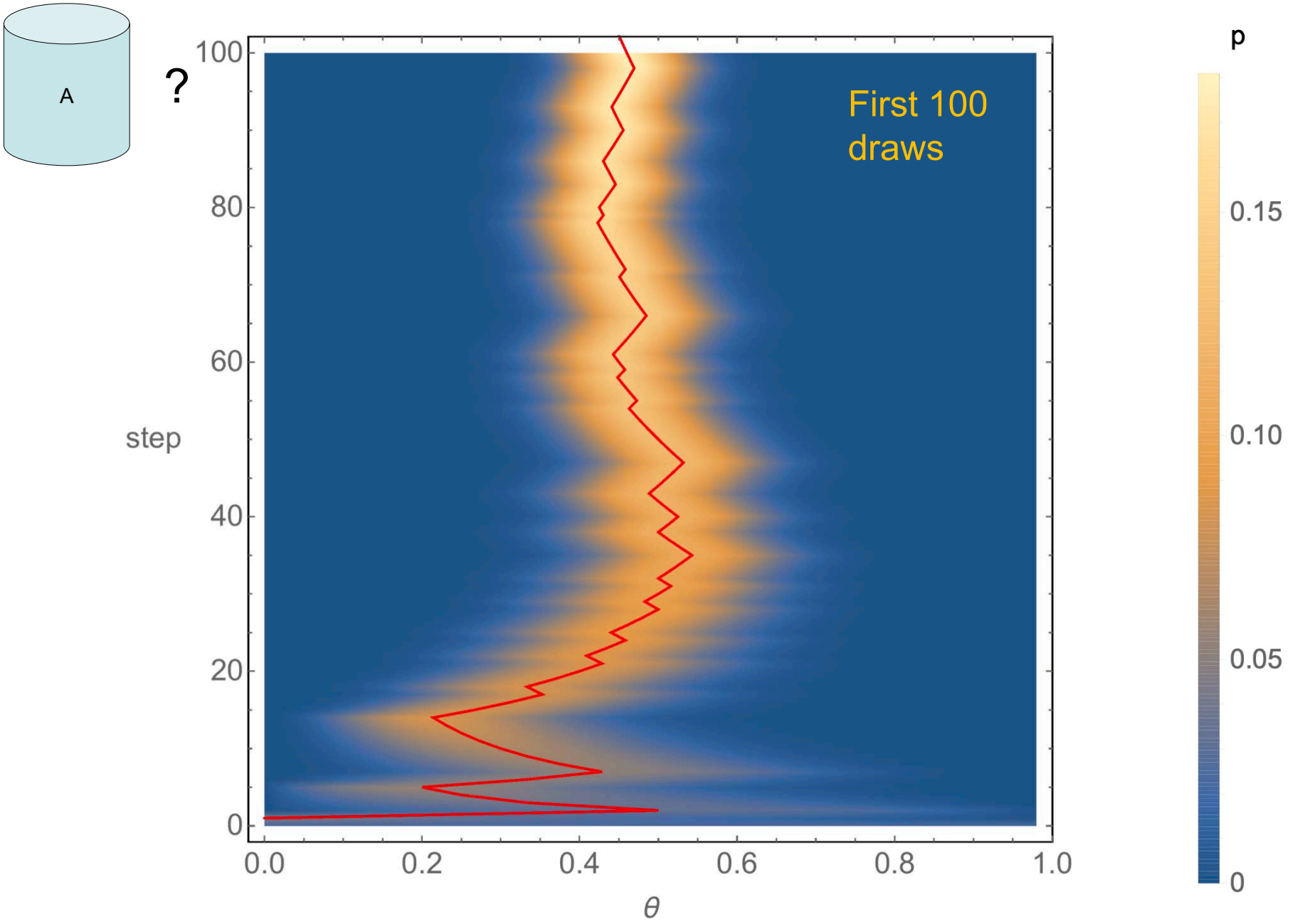


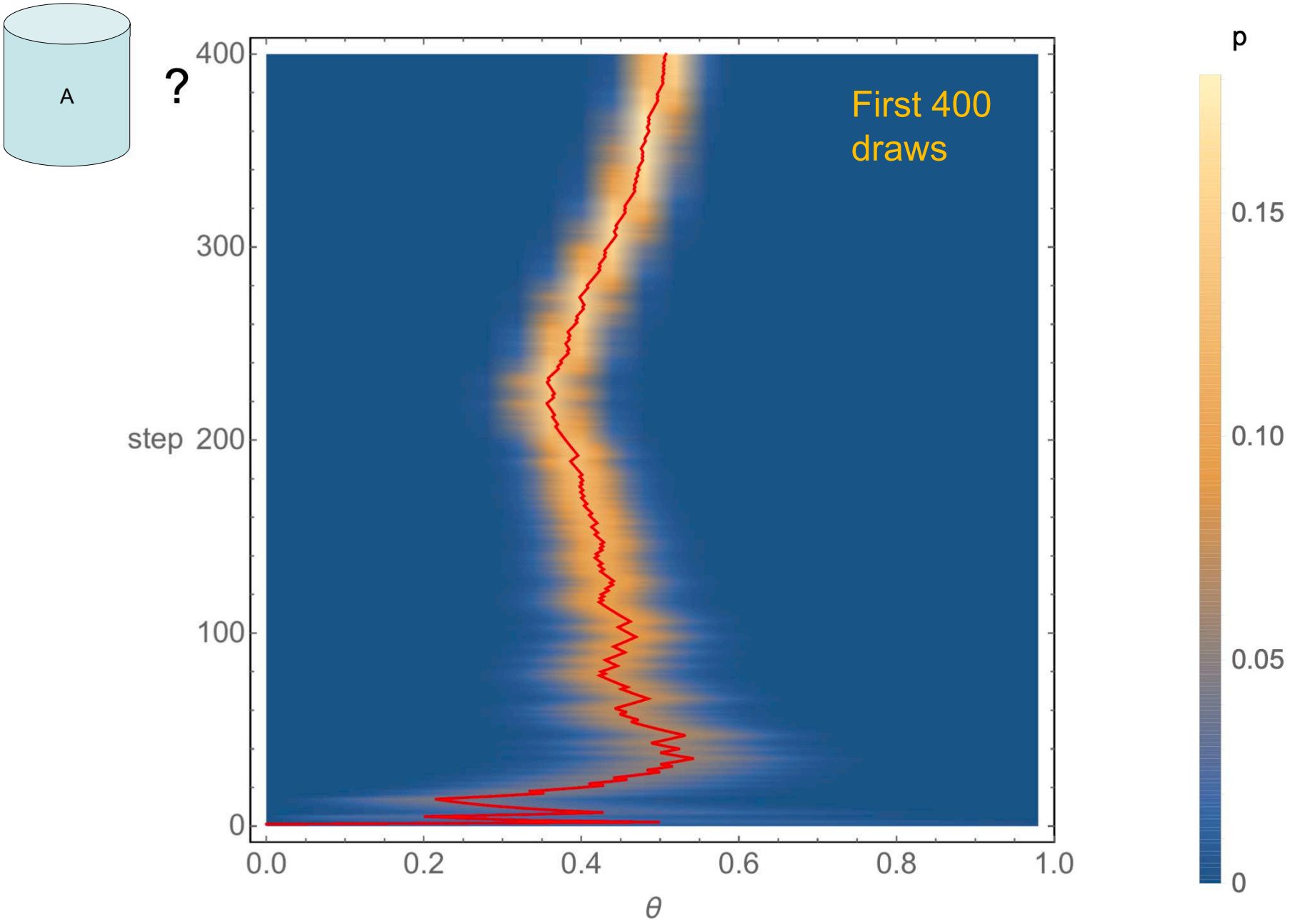
?

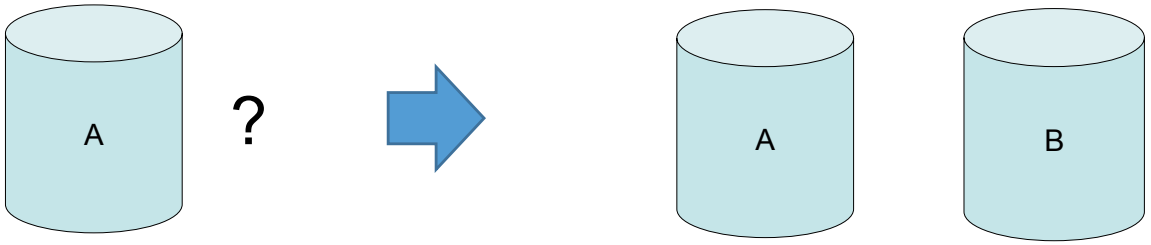
A mysterious result

```
Y, R, Y, Y, Y, R, R, Y, Y, Y, Y, Y, Y, Y, R, R, R, Y, R, R, R, Y, R, R, Y,
R, R, R, Y, R, R, Y, R, R, R, Y, Y, Y, R, R, Y, Y, Y, R, R, R, R, Y, Y, Y,
Y, Y, Y, Y, R, Y, Y, Y, R, Y, Y, R, R, R, R, R, Y, Y, Y, Y, Y, R, Y, Y, Y,
Y, Y, Y, R, Y, R, R, R, Y, Y, Y, R, R, R, R, Y, Y, Y, R, R, R, R, R, Y, Y.
```

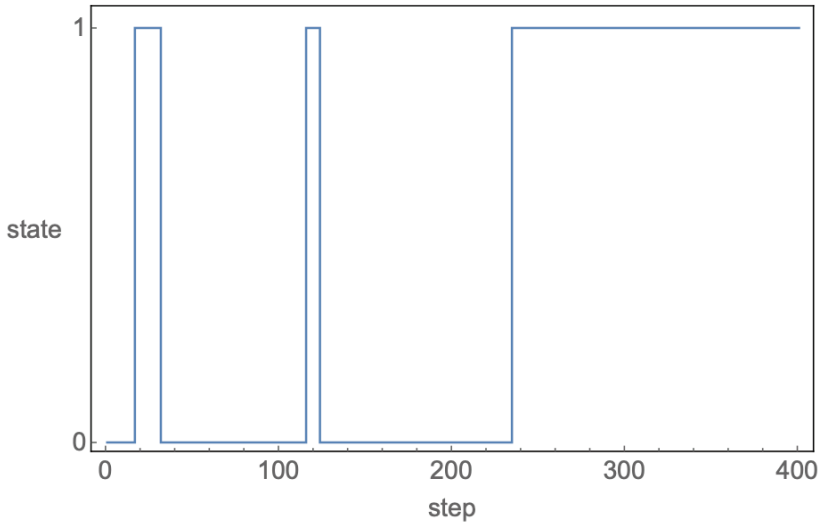
Is this the same game as before? We can repeat the latest analysis and we find the plots in the following slides.

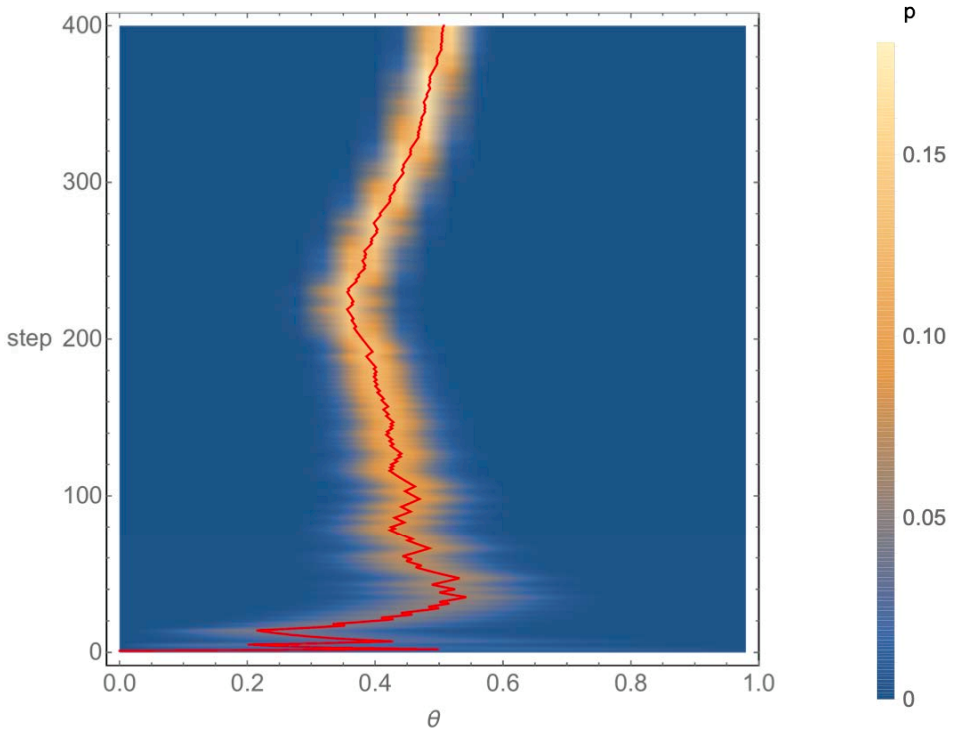

A
?
First 100
draws
step
100
80
60
40
20
0
0.0
0.2
0.4
0.6
0.8
1.0
θ
p
0.15
0.10
0.05
0


A
?
First 400
draws
step
400
300
200
100
0
0.0
0.2
0.4
0.6
0.8
1.0
θ
p
0.15
0.10
0.05
0



Answer: alternate draws from A and B (14 and 34 red balls out of 50)





This is an example of a Hidden Markov Model

3. Bayesian classification

data X, classes C

this likelihood is defined by training data

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

the prior is also defined by training data

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

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

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

The learning procedure must yield:

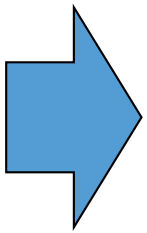
$P(y)$

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

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

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

How can we reduce the huge complexity of learning?



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

for instance, with just two attributes

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

conditional independence assumption

with more than 2 attributes

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

Therefore:

$$P(y_k|\mathbf{x}) = \frac{P(\mathbf{x}|y_k)}{P(\mathbf{x})} P(y_k) = \frac{P(\mathbf{x}|y_k)}{\sum_j P(\mathbf{x}|y_j) P(y_j)} P(y_k)$$

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

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

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

More general discrete inputs

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

Continuous inputs and discrete classes – the Gaussian case

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

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

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

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

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

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

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

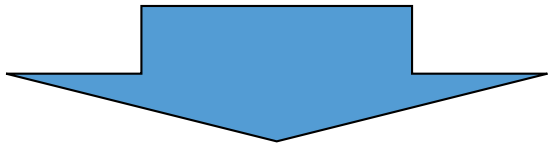
$$= \frac{1}{1+\frac{P(\mathbf{x}|y=1)P(y=1)}{P(\mathbf{x}|y=0)P(y=0)}}$$

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

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

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

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



logistic shape

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

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

Finally an input vector belongs to class y = 0 if

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

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

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

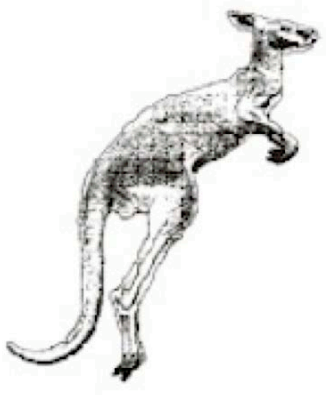
$$\exp\left(w_0 + \sum_{n=1}^{N} w_n x_n\right) < 1$$

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

4. The kangaroo problem with an extended contingency table (Cheesman and Stutz, 2004)

attributes (number of values):

- handedness (2)
- beer-drinking (2)
- state-of-origin (7)
- color (3)



4-dimensional contingency table
with 2x2x7x3 = 84 entries

The size of the contingency table increases exponentially as the number of attributes grows

If we are given the number of occurrencies $n_{i,j,k,l}$ for each position in the contingency table, we have the following constraints

$$0 \leq \theta_{i,j,k,l} \leq 1; \quad \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{7}\sum_{l=1}^{3}\theta_{i,j,k,l} = 1$$

$$0 \leq n_{i,j,k,l} \leq N; \quad \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{7}\sum_{l=1}^{3}n_{i,j,k,l} = N$$

with the likelihood

$$L\left(\mathbf{n}\middle|\boldsymbol{\theta}, N, I\right) = \frac{N!}{\prod_{i,j,k,l} n_{i,j,k,l}} \prod_{i,j,k,l} \theta_{i,j,k,l}^{n_{i,j,k,l}}$$

In a Bayesian context, data that depend on all the problem variables are sufficient statistics and we can estimate all the corresponding probabilities.

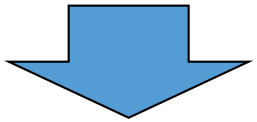
However, this is not always the case. If we are only given a set of marginals, i.e., of constraints, the marginals define a subspace of the whole parameter space, and ***within this subspace the distribution is eventually determined by the prior information only.***

With enough attributes, the contingency table becomes VERY large, and it becomes impossible to collect sufficient statistics, we are mostly limited to marginals.

The situation is very different if we assume independence: then the marginals are sufficient statistics. *E.g., if probabilities factorize, then kangaroos have only (2+2+7+3)-(1+1+1+1) = 10 independent values (using normalization) instead of 84.*

Maximum entropy approach to the kangaroo problem, given marginals

$$\sum_{j,k,l} n_{i,j,k,l} = n_i; \quad \sum_i n_i = N$$



$$\sum_{i,j,k,l} \theta_{i,j,k,l} = 1; \quad \sum_{j,k,l} \theta_{i,j,k,l} = \frac{n_i}{N}$$

Example with two marginals: we maximize the constrained entropy

$$S = -\sum_{i,j,k,l} \theta_{i,j,k,l} \log \theta_{i,j,k,l} + \lambda_0 \left(\sum_{i,j,k,l} \theta_{i,j,k,l} - 1 \right) + \lambda_1 \left(\sum_{j,k,l} \theta_{1,j,k,l} - \frac{n_1}{N} \right) + \lambda_2 \left(\sum_{i,k,l} \theta_{2,j,k,l} - \frac{n_2}{N} \right)$$

in the original kangaroo problem

$$S_V = \left(p_{bl} \log \frac{1}{p_{bl}} + p_{\bar{b}l} \log \frac{1}{p_{\bar{b}l}} + p_{b\bar{l}} \log \frac{1}{p_{b\bar{l}}} + p_{\bar{b}\bar{l}} \log \frac{1}{p_{\bar{b}\bar{l}}} \right)$$

$$+\lambda_1 \left(p_{bl} + p_{\bar{b}l} + p_{b\bar{l}} + p_{\bar{b}\bar{l}} - 1 \right) + \lambda_2 \left(p_{bl} + p_{b\bar{l}} - 1/3 \right) + \lambda_3 \left(p_{bl} + p_{\bar{b}l} - 1/3 \right)$$

$$\frac{\partial S_V}{\partial p_{bl}} = -\log p_{bl} - 1 + \lambda_1 + \lambda_2 + \lambda_3 = 0$$

$$\frac{\partial S_V}{\partial p_{\bar{b}l}} = -\log p_{\bar{b}l} - 1 + \lambda_1 + \lambda_3 = 0$$

$$\frac{\partial S_V}{\partial p_{b\bar{l}}} = -\log p_{b\bar{l}} - 1 + \lambda_1 + \lambda_2 = 0$$

$$\frac{\partial S_V}{\partial p_{\bar{b}\bar{l}}} = -\log p_{\bar{b}\bar{l}} - 1 + \lambda_1 = 0$$

$$\begin{cases} p_{\bar{b}l} = p_{\bar{b}\bar{l}} \exp(\lambda_3) \\ p_{b\bar{l}} = p_{\bar{b}\bar{l}} \exp(\lambda_2) \\ p_{bl} = p_{\bar{b}\bar{l}} \exp(\lambda_2 + \lambda_3) \end{cases} \Rightarrow \quad p_{\bar{b}l} p_{b\bar{l}} = p_{bl} p_{\bar{b}\bar{l}}$$

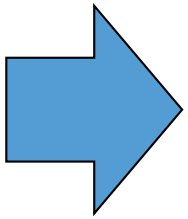
$$\begin{cases} p_{bl} + p_{\bar{b}l} + p_{b\bar{l}} + p_{\bar{b}\bar{l}} = 1 \\ p_{bl} + p_{b\bar{l}} = 1/3 \\ p_{bl} + p_{\bar{b}l} = 1/3 \\ p_{\bar{b}l} p_{b\bar{l}} = p_{bl} p_{\bar{b}\bar{l}} \end{cases} \Rightarrow \begin{cases} p_{b\bar{l}} = p_{\bar{b}l} = 1/3 - p_{bl} \\ p_{\bar{b}\bar{l}} = 1/3 + p_{bl} \\ (1/3 - p_{bl})^2 = p_{bl}/3 + p_{bl}^2 \\ 1/9 - 2p_{bl}/3 + p_{bl}^2 = p_{bl}/3 + p_{bl}^2 \end{cases}$$

$$\Rightarrow \quad p_{bl} = \frac{1}{9}; \quad p_{b\bar{l}} = p_{\bar{b}l} = \frac{2}{9}; \quad p_{\bar{b}\bar{l}} = \frac{4}{9}$$

this solution coincides with the independence hypothesis

In the extended kangaroo problem we find

$$\frac{\partial S}{\partial \theta_{m,j,k,l}} = -\left(\log \theta_{m,j,k,l} + 1\right) + \lambda_0 + \lambda_m = 0$$



$$\theta_{1,j,k,l} = \exp\left(\lambda_0 + \lambda_1 - 1\right)$$

$$\theta_{2,j,k,l} = \exp\left(\lambda_0 + \lambda_2 - 1\right)$$

thus we obtain again a multiplicative structure.

Whatever the choice of marginals, probabilities factorize, and the MaxEnt solution corresponds to a set of independent probabilities.

Thus independence is built-in the MaxEnt method, which is a sort of "generalized independence method".

References:

MaxEnt and image processing

- J. Skilling et al., Mon. Not. R. astr. Soc. **187** (1979) 145
- J. Skilling , Nature **309** (1984) 748
- R. Narayan and R. Nityananda, Ann. Rev. Astron. Astrophys. **24** (1986) 127
- R. Molina et al., IEEE Signal Proc. Magazine (marzo 2001) 13
- J. Skilling, A. W. Strong and K. Bennett, Mon. Not. R. astr. Soc. **187** (1979) 145
- J. Skilling and R. K. Bryan, Mon. Not. R. astr. Soc. **211** (1984) 11
- S. L. Bridle et al, Mon. Not. R. astr. Soc. **299** (1998) 895

Statistical inference principles

- E. Milotti, "The statistical eyeglasses: the math behind scientific knowledge", IOP Concise Physics (2018)
- P. Cheeseman and J. Stutz, “On the Relationship Between Bayesian and Maximum Entropy Inference”, in AIP Conf. Proc. ,Volume 735, pp. 445-461, BAYESIAN INFERENCE AND MAXIMUM ENTROPY METHODS IN SCIENCE AND ENGINEERING: 24th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (2004)