Introduction to Bayesian Statistics - 7

Edoardo Milotti

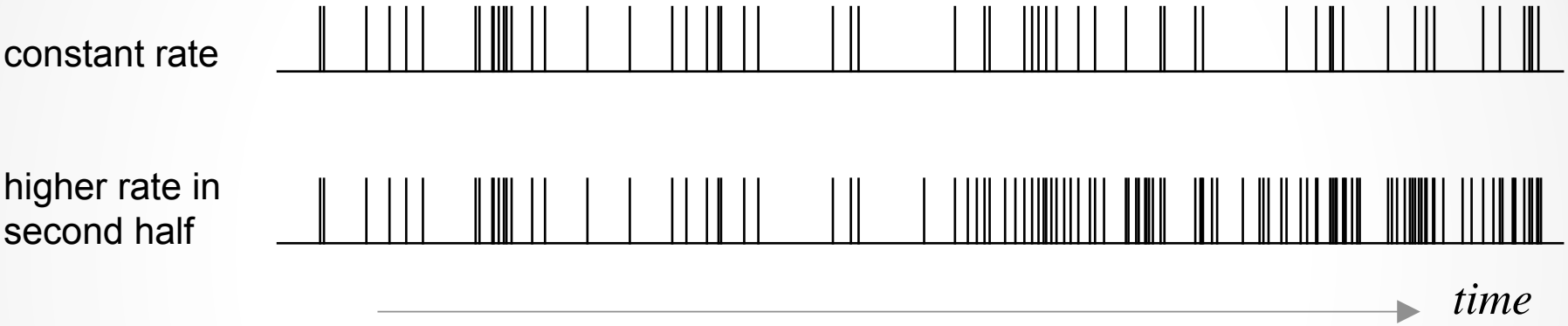
Università di Trieste and INFN-Sezione di Trieste

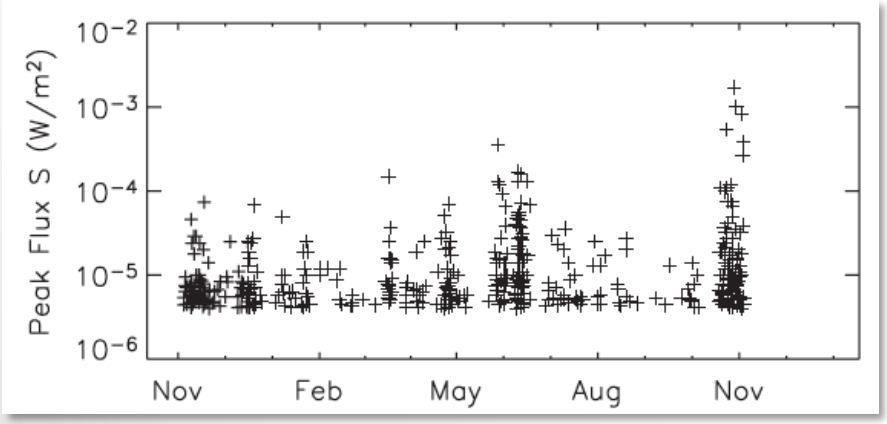
A few more applications of Bayesian methods, on the verge of epistemology.

- Bayesian blocks
- Solar flare statistics and prediction
- Bayes factors and Bell's inequalities
- Bayes classifiers
- The nature of learning in Bayesian and MaxEnt methods

Bayesian blocks (Scargle 1998)

Detection of bursts from piecewise change of (Poisson) event rate





Example of variable rate in solar flare events. Peak flux of 1–8 Å GOES events (crosses) above threshold versus time for one year prior to 4 November 2003. (from M. S. Wheatland, *Space Weather* **3** (2005) S07003)

When the (digital) system clock runs fast enough, for a given event rate, there is at most one event per clock tick.

When the event rate is λ (events/tick) $\left(\lambda \ll 1\right)$ and the time interval is *N* ticks, we find on average

$$n = \lambda N$$

events in the time interval. The average number of events in a clock tick is – obviously – λ again, and this is also the probability of finding an event in the time interval.

Therefore the probability model is binomial, with probability

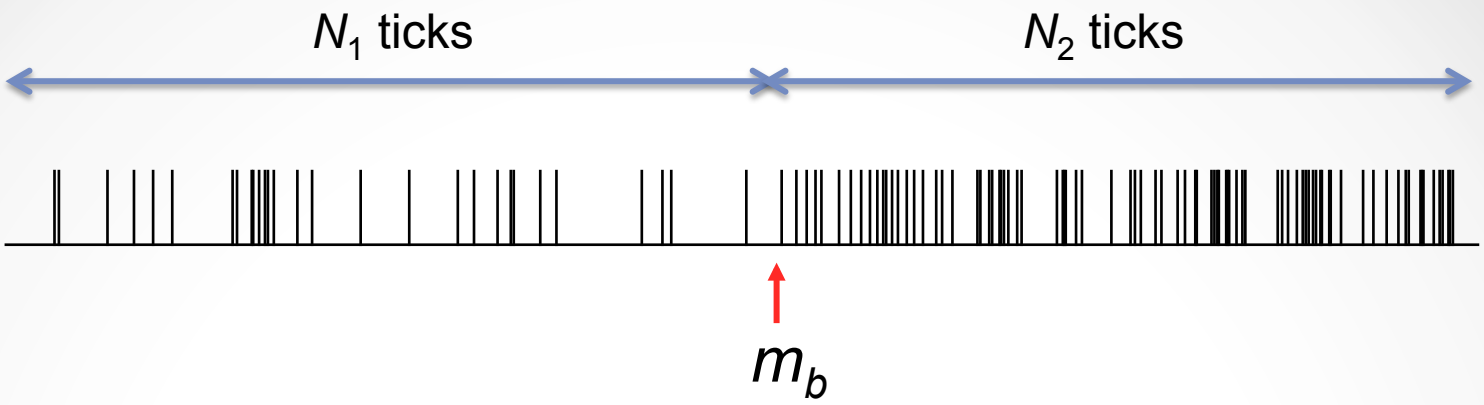
$$p = \lambda$$

This means that the likelihood of finding *n* events in the total time interval is the usual binomial expression

$$\binom{N}{n} p^n (1-p)^{N-n} = \binom{N}{n} \lambda^n (1-\lambda)^{N-n}$$

Notice that here the rate is given in clock ticks. If we use standard time units, we have (with clock tick duration δt)

$$\binom{N}{n} (\lambda \delta t)^n (1-\lambda \delta t)^{N-n}$$



When the rate is not constant, we can set a breakpoint at tick m_b, and the total likelihood becomes

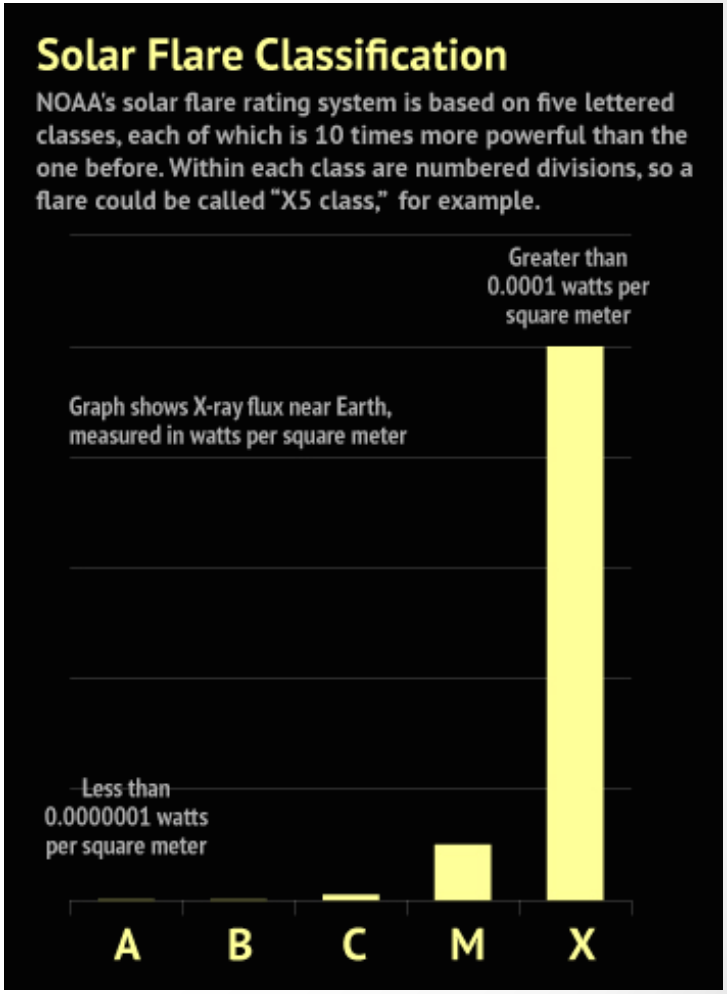
$$\left[\binom{N_1}{n_1}\lambda_1^{n_1}(1-\lambda_1)^{N_1-n_1}\right]\times\left[\binom{N_2}{n_2}\lambda_2^{n_2}(1-\lambda_2)^{N_2-n_2}\right]$$

The calculation then proceeds either with the full likelihood or with a marginalized likelihood

(see, e.g., J. Scargle, ApJ 504 (1998) 405 for many more details)

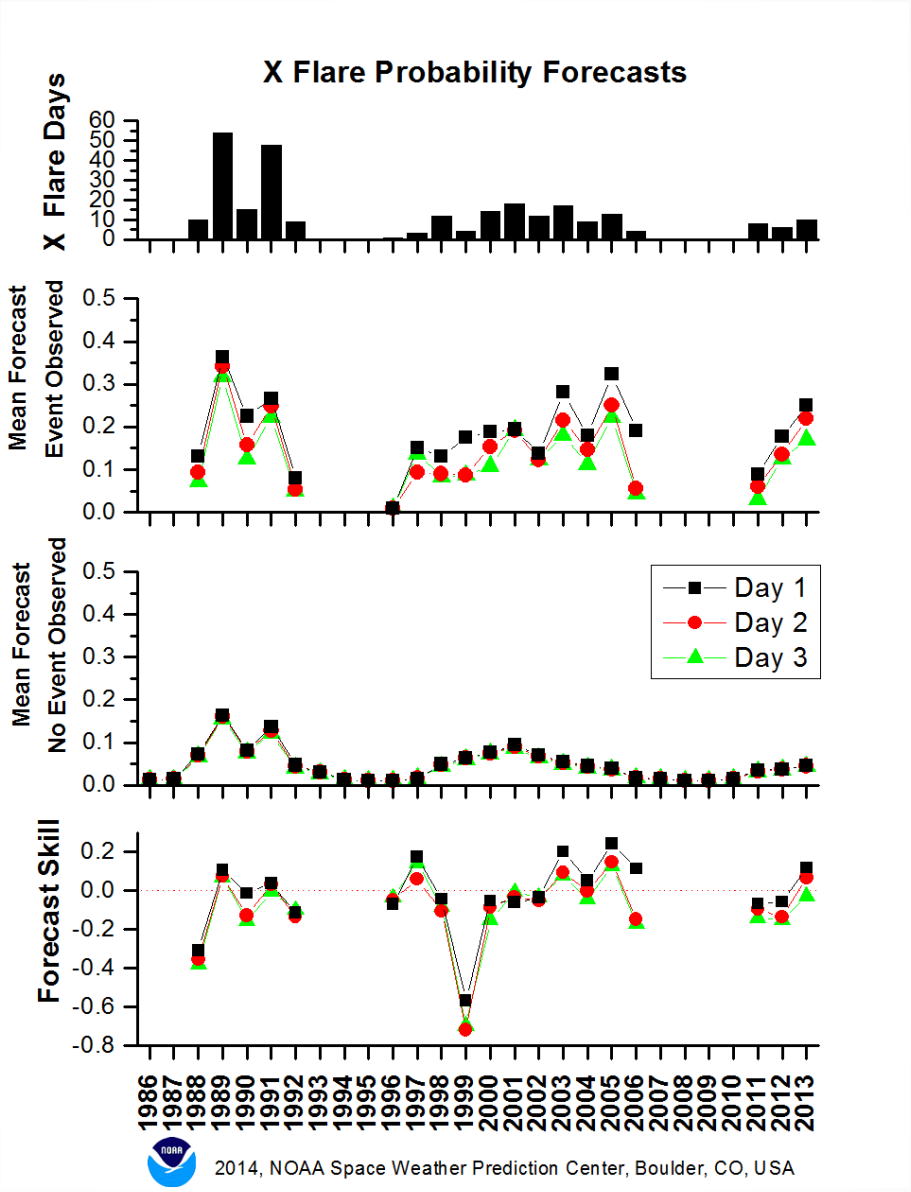
Solar flares

- Solar flares are magnetic explosions in the ionized outer atmosphere of the Sun, the solar corona.
- Flares occur in and around sunspots, where intense magnetic fields penetrate the visible surface of the Sun.
- During a flare some of the energy stored in the magnetic field is released and appears in accelerated particles, radiation, heating, and bulk motion.
- Large flares strongly influence our local “space weather.” They can lead to enhanced populations of energetic particles in the Earth’s magnetosphere and these particles can damage satellite electronics, and pose radiation risks to astronauts and to passengers on polar aircraft flights.
- **It is of great practical importance to construct predictive models of the occurrence of large solar flares.**



Graphics from space.com

The sun erupted with a massive solar at 0027 April 25 GMT, and ranked as an X1.3-class solar storm, one of the strongest types of flares the sun can experience, according to a report from the U.S. Space Weather Prediction Center. NASA's Solar Dynamics Observatory captured video of the intense solar flare in several different wavelengths.


X Flare Probability Forecasts
X Flare Days
60
50
40
30
20
10
0
Mean Forecast
Event Observed
0.5
0.4
0.3
0.2
0.1
0.0
Mean Forecast
No Event Observed
Day 1
Day 2
Day 3
Forecast Skill
0.2
0.0
-0.2
-0.4
-0.6
-0.8
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014, NOAA Space Weather Prediction Center, Boulder, CO, USA

Flare statistics

(from M. S. Wheatland: “A Bayesian approach to solar flare prediction”, ApJ **609** (2004) 1134)

Flare frequency-size distribution (*N* number of events per unit time)

$$N(S) = AS^{-\gamma}$$

where the power-law index is $\gamma \approx 1.5 - 2$

Moreover the statistics in time is Poissonian.

The total event rate for events larger than S_1 is

$$\lambda_1 = \int_{S_1}^{\infty} N(S)dS = A(\gamma - 1)^{-1} S_1^{-\gamma+1}$$

From

$$N(S) = AS^{-\gamma}$$

and

$$\lambda_1 = \int_{S_1}^{\infty} N(S)dS = A(\gamma - 1)^{-1} S_1^{-\gamma+1}$$

we find

$$N(S) = \lambda_1 (\gamma - 1) S_1^{\gamma-1} S^{-\gamma}$$

and likewise

$$\lambda_2 = \lambda_1 \left(\frac{S_1}{S_2} \right)^{\gamma-1}$$

if S_1 is the size of small events, this is an estimate of the rate of events larger than S_2

Using the Poisson model, the probability of finding at least one event larger than S_2 in the time interval ΔT is

$$\epsilon = 1 - \exp(-\lambda_2 \Delta T)$$

Thus we can estimate this useful probability from the rate of small events and from the spectral index (that we assume known).

In the work of Wheatland, the rate of small events is estimated using the Bayesian blocks method.

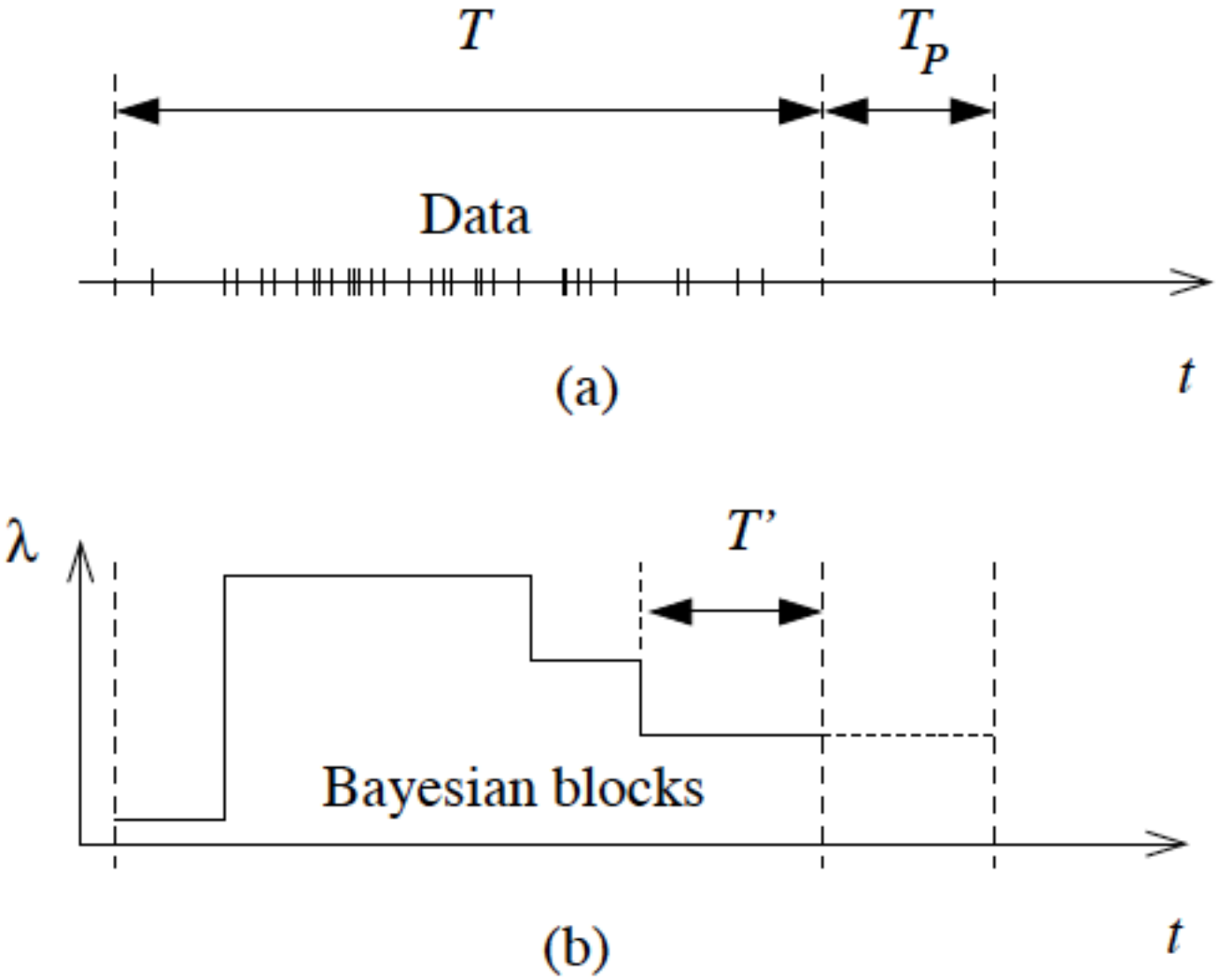


Fig. 1.4. Schematic illustration of Bayesian blocks determination of current rate. Panel (a): data, comnsisting of point events in time line during an observation interval T. The prediction interval T_P is also shown. Panel (b): Bayesian blocks decomposition of the rate λ, and identification of the most recent interval T' when the rate is approximately constant.

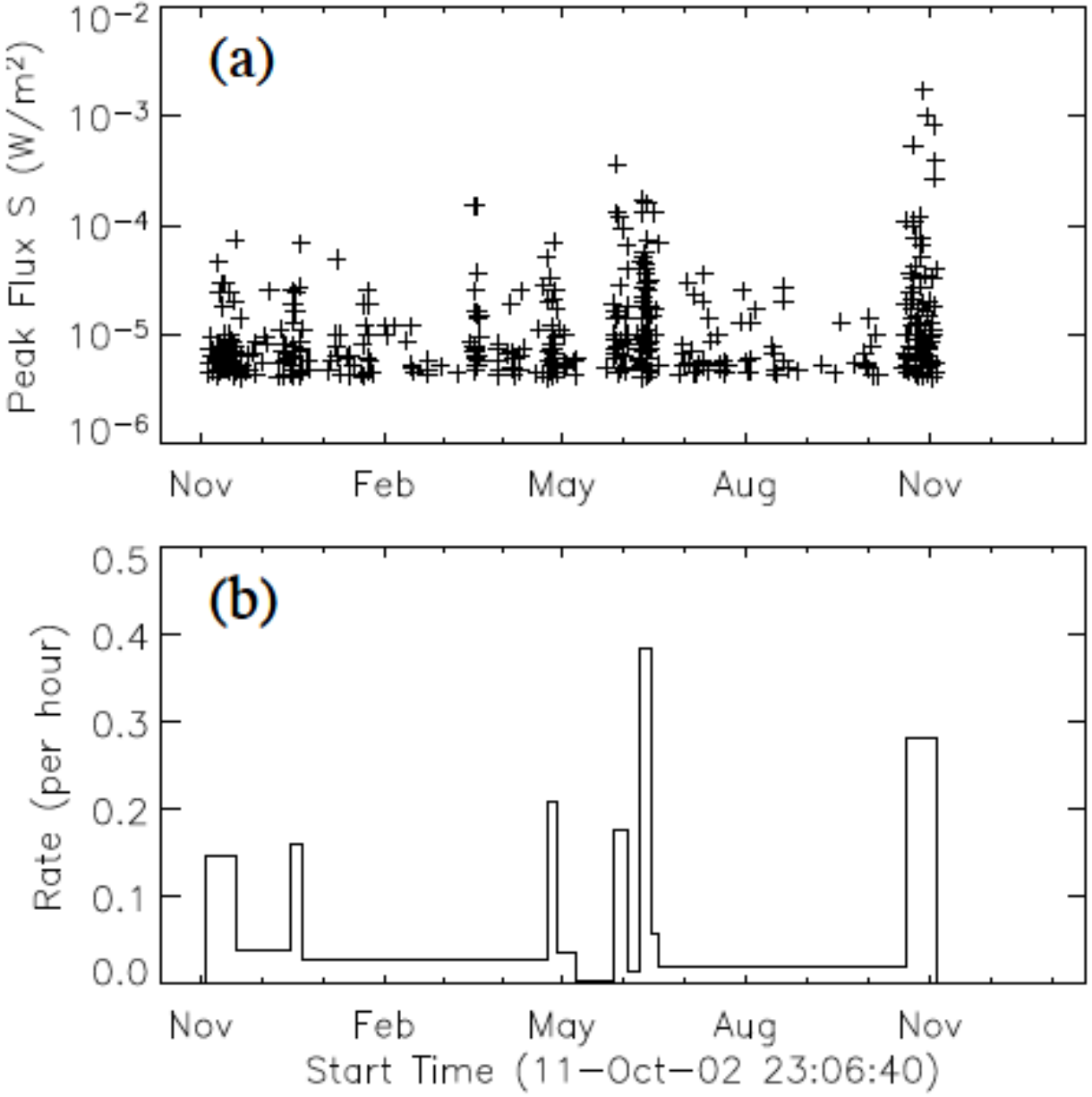


Fig. 1.5. Bayesian blocks applied to one year of GOES events prior to 4 November 2003.

Bayes factors and Bell's theorem

Obituary of J. S. Bell by Shimony, Telegdi and Veltman in Phys. Today

John S. Bell

John Stewart Bell died suddenly of cerebral hemorrhage on 1 October 1990, at the age of 62. The loss to physics, and to natural philosophy in general, is irreparable, for Bell not only made the most profound contribution of his generation to the foundations of quantum mechanics but had continued to explore new ideas on the subject.

...

...

In a sense John Bell had two careers. He contributed directly to the main mission of CERN by his research in nuclear physics, field theory, elementary-particle theory and accelerator design. But he also studied the foundations of quantum mechanics with great intensity, even though he jokingly referred to this work as his "hobby." His delightful exposition "Bertlman's Socks and the Nature of Reality" resulted from his attempt to explain his hobby to one of his collaborators in field theory. That article, together with other related papers by Bell, was reprinted in *Speakable and Unspeakable in Quantum Mechanics* (Cambridge University Press, 1987).

...

...

As an undergraduate Bell was already dissatisfied with textbook presentations of quantum mechanics, and was particularly disturbed by Niels Bohr's thesis that a measuring apparatus must be described classically and not treated quantum mechanically. Bell felt that there should be a unified description of the physical world applying to both microscopic and macroscopic systems. While at Birmingham, Bell was intrigued by two papers written by David Bohm in 1952, proposing a hidden-variables interpretation of quantum mechanics, which seemed a promising way to achieve the desired unification. According to Bohm's construction, something was amiss in John von Neumann's oft-cited demonstration of the impossibility of a hidden-variables interpretation. Bell seriously turned his attention to this matter after attending Josef Jauch's seminar in 1963 at the University of Geneva on the foundations of quantum mechanics. In his paper entitled "On the Problem of Hidden Variables in Quantum Mechanics," Bell proved the impossibility of simple hidden-variables theories, without relying on a dubious premise that von Neumann had used. In the same paper Bell also pointed to a more complex family of hidden-variables theories (later called "contextual") that are not excluded by his own theorem.

The fact that Bohm's construction required a kind of "action at a distance" between spatially separated particles led Bell to pose a penetrating and fruitful question: Is it possible for a hidden-variables theory to recover all the statistical predictions of quantum mechanics without postulating action at a distance? His negative answer to this question was published in 1964 in a paper called "On the Einstein–Podolsky–Rosen Paradox." The remarkable result contained therein is now commonly called Bell's theorem.

...

Einstein's dissatisfaction with quantum mechanics

MAY 15, 1935 PHYSICAL REVIEW VOLUME 47

Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?

A. EINSTEIN, B. PODOLSKY AND N. ROSEN, *Institute for Advanced Study, Princeton, New Jersey*

(Received March 25, 1935)

In a complete theory there is an element corresponding to each element of reality. A sufficient condition for the reality of a physical quantity is the possibility of predicting it with certainty, without disturbing the system. In quantum mechanics in the case of two physical quantities described by non-commuting operators, the knowledge of one precludes the knowledge of the other. Then either (1) the description of reality given by the wave function in quantum mechanics is not complete or (2) these two quantities cannot have simultaneous reality. Consideration of the problem of making predictions concerning a system on the basis of measurements made on another system that had previously interacted with it leads to the result that if (1) is false then (2) is also false. One is thus led to conclude that the description of reality as given by a wave function is not complete.

- **locality**: information cannot propagate faster than light
- **realism**: physical objects possess properties independently of measurements

Could quantum mechanics be just the phenomenology of a deeper classical theory with variables that we are unable to observe, i.e., with hidden variables?

If so, the hidden variables theory would satisfy both locality and realism.

John Bell displayed inequalities that are valid for any local, realistic theory, but are violated by quantum mechanics.

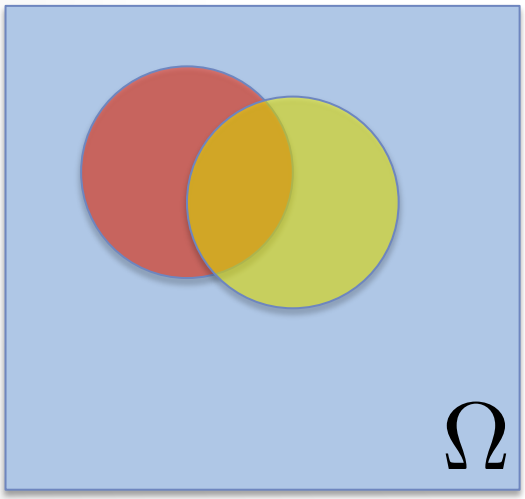
A simplified proof of Bell's theorem

(L. Maccone, arXiv:1212.5214 [quant-ph])

Take three objects with two-valued properties (values 0 and 1) A, B and C and let

$P_{same}(A,B)$ = prob. that property A of the first object has the same value as property B of the second object;

$P_{diff}(A,B)$ = prob. that property A of the first object differs from property B of the second object;



red area: $P_{same}(A,B)$

yellow area: $P_{same}(A,C)$

orange area: probability that A=B=C

blue area: $P_{same}(B,C)$

$$P_{\text{same}}(A,B) + P_{\text{same}}(A,C) + P_{\text{same}}(B,C) \geq 1$$

The inequality is violated by quantum mechanics.

Indeed, consider two two-level systems in the entangled state

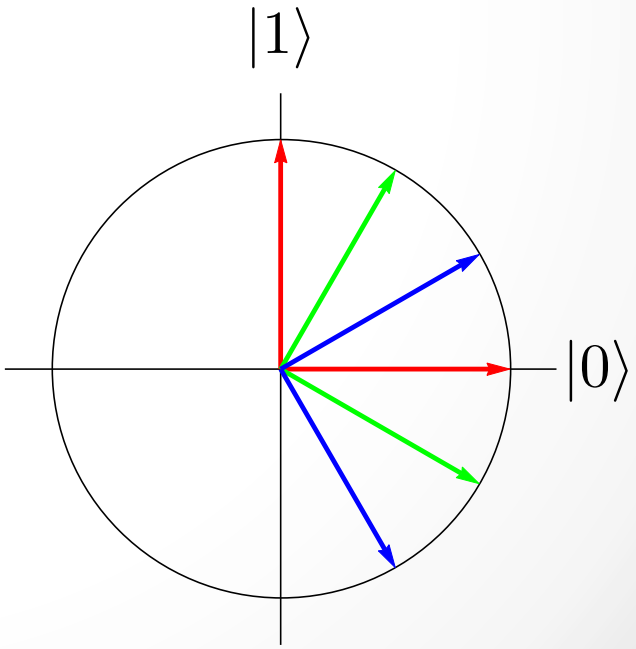
$$|\Phi^+\rangle = \frac{|00\rangle + |11\rangle}{\sqrt{2}}$$

and the properties A, B and C obtained by projecting the state on

$$A: \begin{cases} |a_0\rangle = |0\rangle \\ |a_1\rangle = |1\rangle \end{cases}$$

$$B: \begin{cases} |b_0\rangle = \frac{1}{2}|0\rangle + \frac{\sqrt{3}}{2}|1\rangle \\ |b_1\rangle = \frac{\sqrt{3}}{2}|0\rangle - \frac{1}{2}|1\rangle \end{cases}$$

$$C: \begin{cases} |c_0\rangle = \frac{1}{2}|0\rangle - \frac{\sqrt{3}}{2}|1\rangle \\ |c_1\rangle = \frac{\sqrt{3}}{2}|0\rangle + \frac{1}{2}|1\rangle \end{cases}$$



It is also easy to verify that

$$|\Phi^+\rangle = \frac{|a_0 a_0\rangle + |a_1 a_1\rangle}{\sqrt{2}} = \frac{|b_0 b_0\rangle + |b_1 b_1\rangle}{\sqrt{2}} = \frac{|c_0 c_0\rangle + |c_1 c_1\rangle}{\sqrt{2}}$$

This means that when we carry out a measurement of any property A we find that the subsystems always share the same property (whichever it is):

$$P_{\text{same}}(A, A) = P_{\text{same}}(B, B) = P_{\text{same}}(C, C) = 1$$

Now notice that

$$|a_0\rangle = \frac{|b_0\rangle + \sqrt{3}|b_1\rangle}{2} \qquad |a_1\rangle = \frac{\sqrt{3}|b_0\rangle - |b_1\rangle}{2}$$

and therefore

$$|\Phi^+\rangle = \frac{|a_0\rangle(|b_0\rangle + \sqrt{3}|b_1\rangle) + |a_1\rangle(\sqrt{3}|b_0\rangle - |b_1\rangle)}{2\sqrt{2}}$$

so that

$$\langle a_0 b_0|\Phi^+\rangle = \langle a_1 b_1|\Phi^+\rangle = \frac{1}{2\sqrt{2}} \quad \Rightarrow \quad P_{\text{same}}(A, B) = 1/4$$

The same can be done for the other properties, and one finds

$$P_{\text{same}}(A,B) = P_{\text{same}}(A,C) = P_{\text{same}}(B,C) = 1/4$$

and finally

$$P_{\text{same}}(A,B) + P_{\text{same}}(A,C) + P_{\text{same}}(B,C) = 3/4 < 1$$

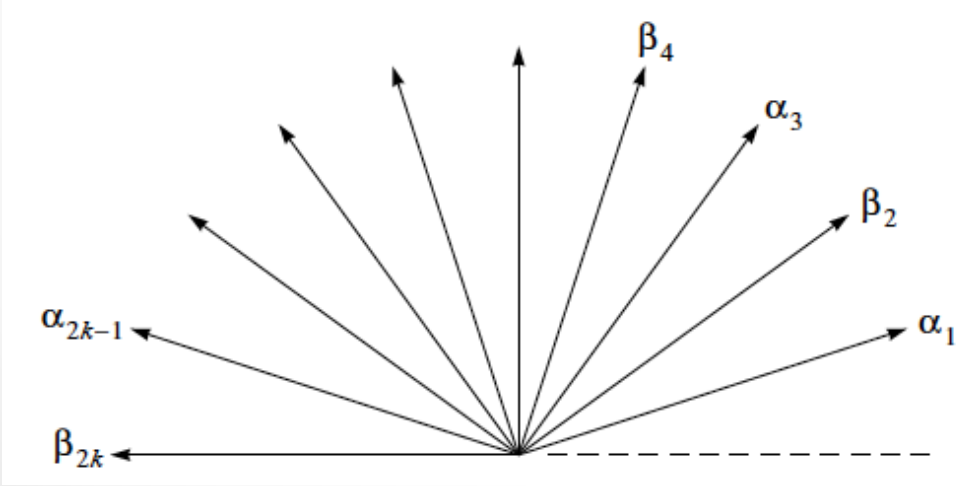
therefore QM violates the inequality and it cannot be both realistic and local.

Many other versions of this result have been devised after Bell ...

For instance, Clauser, Horne, Shimony, and Holt (CHSH), found that when two observers have a choice of several yes-no tests A_k (observer **a**) and B_k (observer **b**), then local realism implies

$$P(A_1 B_2) + P(B_2 A_3) + \cdots + p(A_{2k-1} B_{2k}) \geq p(A_1 B_{2k})$$

(Braunstein and Caves version of CHSH inequality, 2k-1 terms on the lhs)



Now take the measurements of the spins of two particles along the directions shown here as the measured “properties”.

Consecutive directions are separated by the angle

$$\theta = \pi / 2k$$

Quantum mechanics predicts that each probability of the lhs is

$$q = (1 - \cos\theta)/2$$

while the probability on the rhs of the inequality is just

$$(1 - q)$$

and this violates the inequality.

The closest a local realistic theory could get to quantum mechanics is by leading to an equality. If we further assume rotational symmetry, we can state that all probabilities have to be the same, and therefore

$$(2k - 1)r = 1 - r \quad \Rightarrow \quad r = 1/2k = \theta/\pi$$

Now we have two predictions

QM: probability *q* that two observers obtain the same result
LR: probability *r* that two observers obtain the same result

we compare the hypotheses using equal prior probabilities and the Bayes factor.

Since the underlying model is binomial (we find the same result or not in the two measurements), the likelihoods have the same functional form with different probabilities, i.e., the Bayes factor is

$$\text{Bayes factor} = \frac{q^n(1-q)^{N-n}}{r^n(1-r)^{N-n}} = \left(\frac{q}{r}\right)^n \left(\frac{1-q}{1-r}\right)^{N-n}$$

Assuming QM to be correct, we find that the number of positive tests is, on average

$$n = qN$$

and therefore

$$\text{Bayes factor} = \left[\left(\frac{q}{r} \right)^{q} \left(\frac{1-q}{1-r} \right)^{1-q} \right]^{N}$$

For example, if we wish a Bayes factor 10^4, and k=2, we must carry out at least $N \approx 287$ trials.

(further details in A. Peres, arXiv:quant-ph/9905084)

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:

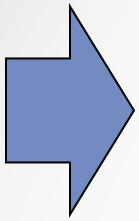
$P(y)$

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):

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

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

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

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

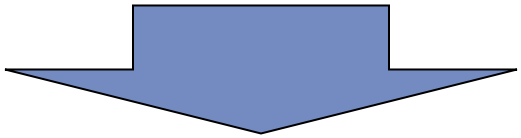
$$= \frac{1}{1+\frac{P(\mathbf{x}|y=1)P(y=1)}{P(\mathbf{x}|y=0)P(y=0)}}$$

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

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

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

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



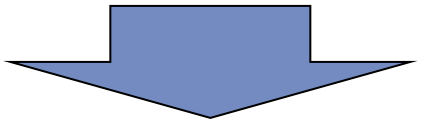
logistic shape

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

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

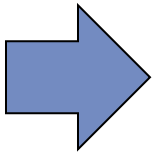
Finally an input vector belongs to class y = 0 if

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

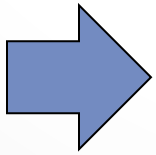


$$P(y=0|\mathbf{x}) = \frac{1}{1+\exp\left(w_0 + \sum_{n=1}^{N} w_n x_n\right)}$$

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



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



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

Naive Bayesian learning is an example of supervised learning, however there are also unsupervised Bayesian learning methods, such as the AUTOCLASS program and similar such projects.

On the nature of learning in Bayesian and MaxEnt Inference
(from Cheeseman & Stutz, 2004)

here we consider these three problems:

1. find the probabilities θ_i of getting face *i* in a throw of a possibly biased die, given the frequencies n_i of each face in a total of *N* throws;

2. find the probabilities when only the mean $M = \sum_{i=1}^{6} i n_i$,

   and the total number of throws *N*, are given;

3. analyze the kangaroo problem with a more complex contingency table

1. Find the probabilities θ_i of getting face *i* in a throw of a possibly biased die, given the frequencies n_i of each face in a total of *N* throws;

$$0 \le \theta_i \le 1; \qquad \sum_{i=1}^{6} \theta_i = 1; \qquad 0 \le n_i \le N; \qquad \sum_{i=1}^{6} n_i = N$$

likelihood is given by the multinomial probability

$$L\left(\{n_1,\ldots,n_6\}\middle|\boldsymbol{\theta},N,I\right) = \frac{N!}{\prod_{j=1}^{6} n_j!}\prod_{i=1}^{6}\theta_i^{n_i}$$

if, initially, we take a uniform prior, the posterior distribution from Bayes' theorem is

$$p\left(\boldsymbol{\theta}\middle|\left\{n_1,\ldots,n_6\right\},N,I\right)=\frac{\prod_{i=1}^{6}\theta_i^{n_i}\delta\left(\sum_{j=1}^{6}\theta_j-1\right)}{\int_0^1\prod_{i=1}^{6}\theta_i^{n_i}\delta\left(\sum_{j=1}^{6}\theta_j-1\right)d\theta_i}$$

$$=\frac{\Gamma\left(N+6\right)}{\prod_{j=1}^{6}\Gamma\left(n_j+1\right)}\prod_{i=1}^{6}\theta_i^{n_i}\delta\left(\sum_{j=1}^{6}\theta_j-1\right)$$

and we obtain a Dirichlet distribution (conjugate posterior of the multinomial distribution, just as the Beta distribution is the conjugate posterior of the binomial distribution).

Mathematical note on the normalization of the Dirichlet distribution:

$$B(m,n)=\int_0^1 t^{m-1}(1-t)^{n-1}\,dt=\frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)}$$

relationship between Beta and Gamma function

$$\int_{0\le\theta_i\le 1}\theta_1^{n_1}\theta_2^{n_2}\theta_3^{n_3}\delta(\theta_1+\theta_2+\theta_3-1)\,d\theta_1\,d\theta_2\,d\theta_3=\int_{0\le\theta_i\le 1}\theta_1^{n_1}\,d\theta_1\int_0^{1-\theta_1}p^{n_2}\left[(1-\theta_1)-p\right]^{n_3}dp$$

$$=\int_{0\le\theta_i\le 1}\theta_1^{n_1}\,d\theta_1(1-\theta_1)^{n_2+n_3+1}\int_0^1 x^{n_2}(1-x)^{n_3}\,dx$$

$$=B(n_2+1,n_3+1)\int_0^1\theta_1^{n_1}(1-\theta_1)^{n_2+n_3+1}\,d\theta_1=B(n_2+1,n_3+1)B(n_1+1,n_2+n_3+2)$$

$$=\frac{\Gamma(n_2+1)\Gamma(n_3+1)}{\Gamma(n_2+n_3+2)}\cdot\frac{\Gamma(n_1+1)\Gamma(n_2+n_3+2)}{\Gamma(n_1+n_2+n_3+3)}=\frac{\Gamma(n_2+1)\Gamma(n_3+1)\Gamma(n_1+1)}{\Gamma(n_1+n_2+n_3+3)}$$

$$\int_{0\le\theta_i\le 1}\prod_{i=1}^{M}\theta_i^{n_i}\,d\theta_i\,\delta\left(\sum_{j=1}^{M}\theta_j-1\right)=\frac{\prod_{i=1}^{M}\Gamma(n_i+1)}{\Gamma(N+M)}$$

normalization factor

thus, if we assume some prior information, we can start with a Dirichlet prior

$$p\left(\boldsymbol{\theta}\middle|\mathbf{w},I\right)=\frac{\Gamma\left(W\right)}{\prod_{j=1}^{6}\Gamma\left(w_j\right)}\prod_{i=1}^{6}\theta_i^{w_j-1}\delta\left(\sum_{j=1}^{6}\theta_j-1\right)\qquad\text{with}\quad W=\sum_{j=1}^{6}w_j$$

and obtain the posterior distribution

$$p\left(\boldsymbol{\theta}\middle|\mathbf{n},\mathbf{w},N,I\right)=\frac{\prod_{i=1}^{6}\theta_i^{n_i+w_i-1}\delta\left(\sum_{j=1}^{6}\theta_j-1\right)}{\int_0^1\prod_{i=1}^{6}\theta_i^{n_i+w_i-1}\delta\left(\sum_{j=1}^{6}\theta_j-1\right)d\theta_i}=\frac{\Gamma\left(N+W\right)}{\prod_{j=1}^{6}\Gamma\left(n_j+w_j\right)}\prod_{i=1}^{6}\theta_i^{n_i+w_i-1}\delta\left(\sum_{j=1}^{6}\theta_j-1\right)$$

$$=\frac{N!}{\prod_{j=1}^{6}n_j!}\cdot\frac{\Gamma\left(W\right)}{\prod_{j=1}^{6}\Gamma\left(w_j\right)}\prod_{i=1}^{6}\theta_i^{n_i+w_i-1}\delta\left(\sum_{j=1}^{6}\theta_j-1\right)$$

The inferred distribution can be used to compute averages, and also for prediction.

Indeed, the probability of observing r_i occurrences of the i-th face in the future is

$$P\left(\mathbf{r}\middle|\mathbf{n},N,R,\mathbf{w},I\right)=\int_{\boldsymbol{\theta}}P\left(\mathbf{r}\middle|\boldsymbol{\theta},N,R,I\right)p\left(\boldsymbol{\theta}\middle|\mathbf{n},N,\mathbf{w},I\right)d\boldsymbol{\theta}=$$

$$=\int_{\boldsymbol{\theta}}\frac{R!}{\prod_{j=1}^{6}r_j!}\prod_{i=1}^{6}\theta_i^{r_i}\,\frac{\Gamma\left(N+W\right)}{\prod_{j=1}^{6}\Gamma\left(n_j+w_j\right)}\prod_{i=1}^{6}\theta_i^{n_i+w_i-1}\delta\left(\sum_{j=1}^{6}\theta_j-1\right)d\boldsymbol{\theta}$$

$$=\frac{R!}{\prod_{j=1}^{6}r_j!}\cdot\frac{\Gamma\left(N+W\right)}{\prod_{j=1}^{6}\Gamma\left(n_j+w_j\right)}\cdot\frac{\prod_{j=1}^{6}\Gamma\left(n_j+r_j+w_j\right)}{\Gamma\left(N+R+W\right)}$$

so that we find, e.g.,

$$P\left(r_1 = 1 \middle| \mathbf{n}, N, R = 1, \mathbf{w}, I\right) = \frac{\Gamma\left(N + W\right)}{\prod_{j=1}^{6} \Gamma\left(n_j + w_j\right)} \cdot \frac{\prod_{j=1}^{6} \Gamma\left(n_j + w_j + \delta_{1j}\right)}{\Gamma\left(N + W + 1\right)}$$

$$= \frac{n_1 + w_1}{N + W}$$

2. Find the probabilities when only the total $M = \sum_{i=1}^{6} i n_i$, and the total of throws *N*, are given

Let $\langle \mathbf{n} \rangle_{NM}$ be the set of vectors that satisfy the conditions,

$$N = \sum_{i=1}^{6} n_i; \quad M = \sum_{i=1}^{6} i n_i$$

then the likelihood is

$$P\left(M \middle| \boldsymbol{\theta}, N, I\right) = \sum_{\langle \mathbf{n} \rangle_{NM}} P\left(\mathbf{n} \middle| \boldsymbol{\theta}, N, I\right) = \sum_{\langle \mathbf{n} \rangle_{NM}} \frac{N!}{\prod_{j=1}^{6} n_j!} \prod_{i=1}^{6} \theta_i^{n_i}$$

now notice that

$$P(\boldsymbol{\theta}|M,N,\mathbf{w},I) = \frac{P(M|\boldsymbol{\theta},N,I)P(\boldsymbol{\theta}|N,\mathbf{w},I)}{P(M|N,I)}$$

$$= \frac{\sum_{\langle\mathbf{n}\rangle_{NM}} P(\mathbf{n}|\boldsymbol{\theta},N,I)P(\boldsymbol{\theta}|N,\mathbf{w},I)}{\sum_{\langle\mathbf{n}\rangle_{NM}} \int_{\boldsymbol{\theta}} P(\mathbf{n}|\boldsymbol{\theta},N,I)P(\boldsymbol{\theta}|N,\mathbf{w},I)\,d\boldsymbol{\theta}}$$

$$\sum_{\langle\mathbf{n}\rangle_{NM}} P(\mathbf{n}|\boldsymbol{\theta},N,I)P(\boldsymbol{\theta}|N,\mathbf{w},I) = \sum_{\langle\mathbf{n}\rangle_{NM}} \frac{N!}{\prod_{j=1}^{6} n_j!} \frac{\Gamma(W)}{\prod_{j=1}^{6}\Gamma(w_j)} \prod_{i=1}^{6} \theta_i^{n_i+w_i-1}$$

$$\sum_{\langle\mathbf{n}\rangle_{NM}} \int_{\boldsymbol{\theta}} P(\mathbf{n}|\boldsymbol{\theta},N,I)P(\boldsymbol{\theta}|N,\mathbf{w},I)\,d\boldsymbol{\theta} = \sum_{\langle\mathbf{n}\rangle_{NM}} \frac{N!}{\prod_{j=1}^{6} n_j!} \frac{\Gamma(W)}{\prod_{j=1}^{6}\Gamma(w_j)} \frac{\prod_{i=1}^{6}\Gamma(n_i+w_i)}{\Gamma(N+W)}$$

from these formulas we can calculate all marginals and any expectation, although it is quite difficult to manipulate

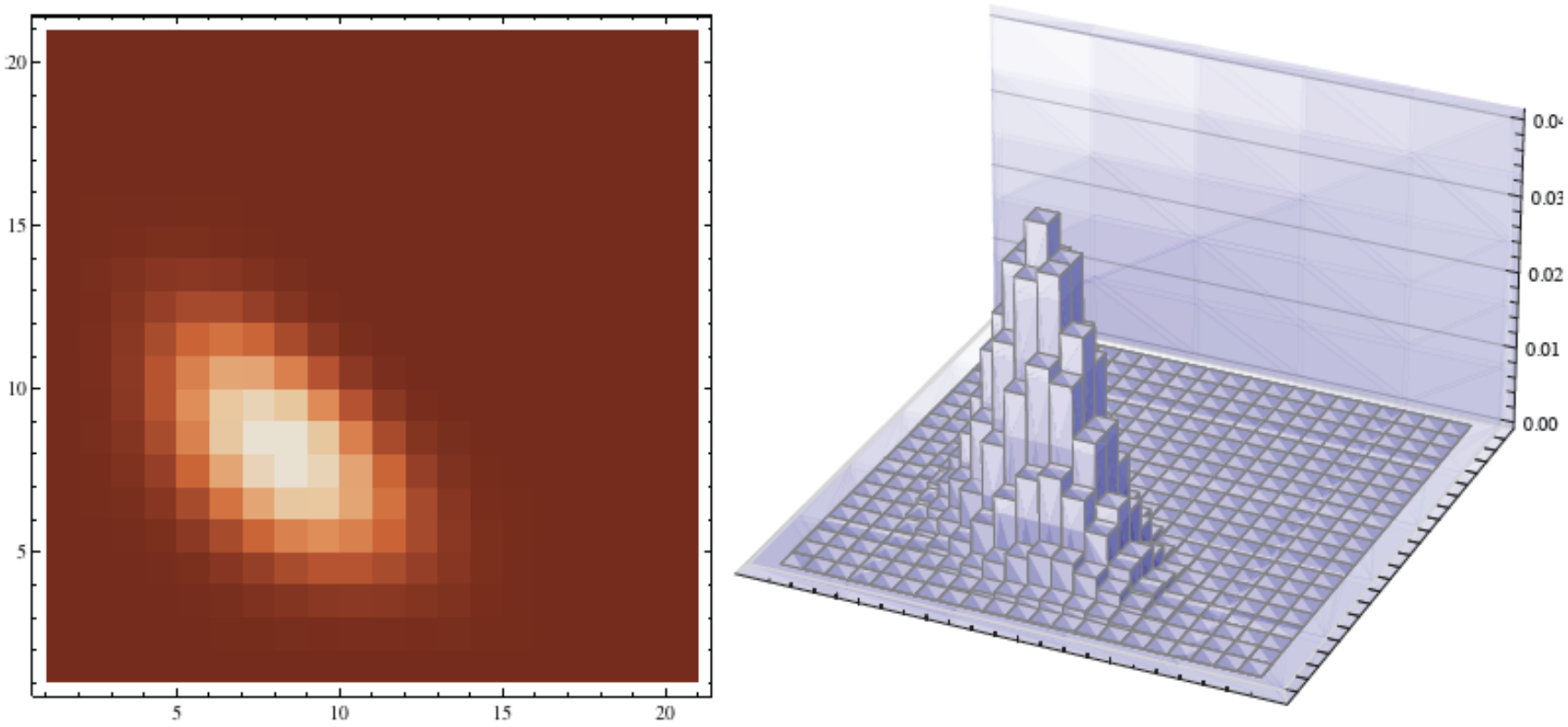


Figure 3: Multinomial distribution with $n = 20$, $k = 3$ and $p_1 = p_2 = p_3 = 1/3$, plotted as a function of the independent values n_1 and n_2. Density plot (left panel) and lego plot (right panel). As an exercise, explain why in this symmetrical case the distribution is not centered in the n_1, n_2 domain, and consider ways to represent multinomial distributions with $k > 3$.

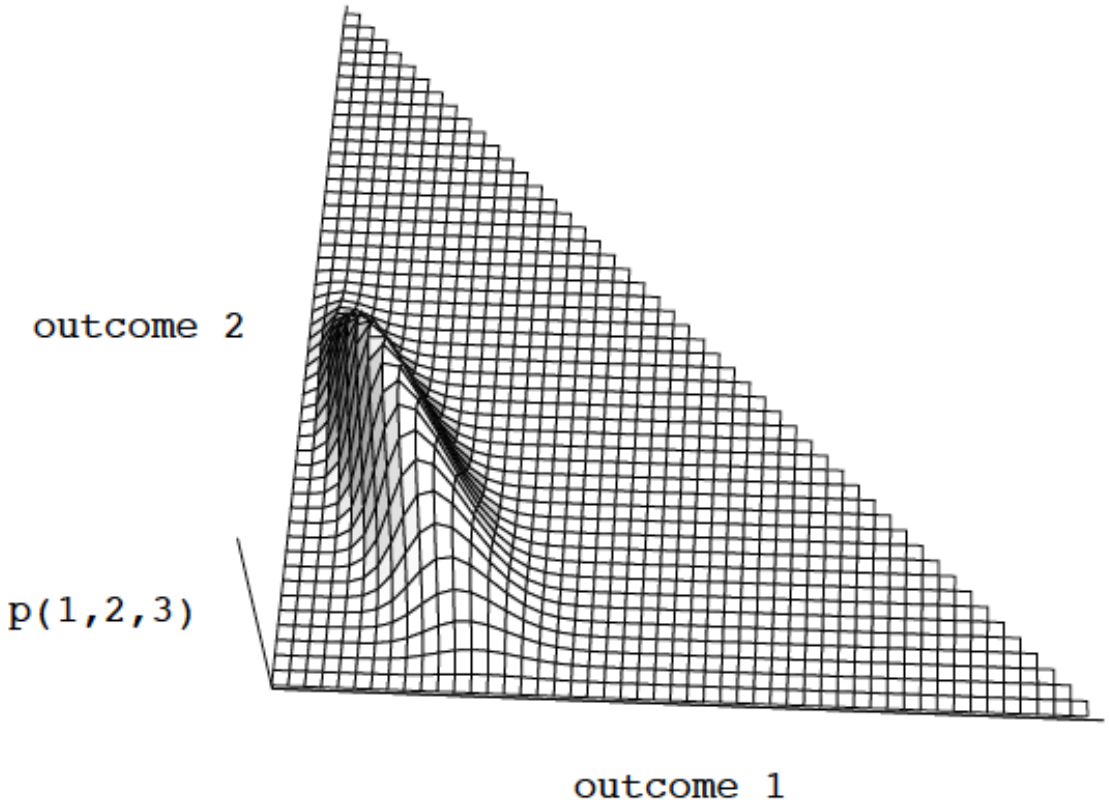


FIGURE 2. The posterior density for the 3-faces die example with a mean spot count of 2.5, N = 60, and prior weights of (1,1,1). Because of the normalization constraint, the third variable (not shown) is given by $\theta_3 = 1 - \theta_1 - \theta_2$.

The figure, from C&S, shows that the probability mass is concentrated close to the subspace defined by constraints, and becomes increasingly so as N increases. Bayesian inference tells us nothing on the distribution inside the subspace.The only information inside the subspace comes from priors.

3. The kangaroo problem with an extended contingency table

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 fall back to the first example of dice throw

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

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

with the likelihood

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

The $n_{i,j,k,l}$'s are sufficient statistics and we can estimate all the corresponding probabilities as in the first example.

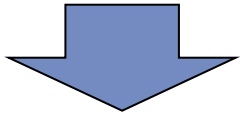
However if we are only given a set of marginals, i.e., of constraints, we are in the same situation as example 2, the marginals define a subspace of the whole parameter space, and in this subspace the distribution is eventually determined by the prior information only.

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

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

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

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

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

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

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

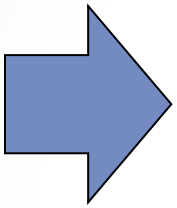
$$\begin{cases} p_{bl} + p_{\bar{b}l} + p_{b\bar{l}} + p_{\bar{b}\bar{l}} = 1 \\ p_{bl} + p_{b\bar{l}} = 1/3 \\ p_{bl} + p_{\bar{b}l} = 1/3 \\ p_{\bar{b}l} p_{b\bar{l}} = p_{bl} p_{\bar{b}\bar{l}} \end{cases} \Rightarrow \begin{cases} p_{b\bar{l}} = p_{\bar{b}l} = 1/3 - p_{bl} \\ p_{\bar{b}\bar{l}} =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

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

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".

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

Search: + GO

+ABOUT NASA | +LATEST NEWS | +MULTIMEDIA | +MISSIONS | +MY NASA | +WORK FOR NASA

+ NASA Home
+ Ames Home
+ Intelligent Systems Division

AutoClass

+ Home
+ AutoClass C
+ References

Introduction

In previous years, the Bayes group at Ames Research Center developed the basic theory and associated algorithms for various kinds of general data analysis techniques. Our earliest efforts were applied to the problem of automatic classification of data. We implemented this theory in the Autoclass series of programs. AutoClass takes a database of cases described by a combination of real and discrete valued attributes, and automatically finds the natural classes in that data. It does not need to be told how many classes are present or what they look like -- it extracts this information from the data itself. The classes are described probabilistically, so that an object can have partial membership in the different classes, and the class definitions can overlap. AutoClass generates reports on the classes it has found at the end of its search. AutoClass has been used and tested on many data sets, both within NASA and by industry, academia and other agencies. These applications typically find surprising classifications that show patterns in the data unknown to the user. Examples include: discovery of new classes of infra-red stars in the IRAS Low Resolution Spectral catalogue (see figure below; and see here and here for more information), new classes of airports in a database of all USA airports, discovery of classes of proteins, introns and other patterns in DNA/protein sequence data, and others.

The starting point of AUTOCLASS is a mixture model

$$dP(x) = \sum_k p_k dP_k(x|\theta); \qquad \sum_k p_k = 1$$

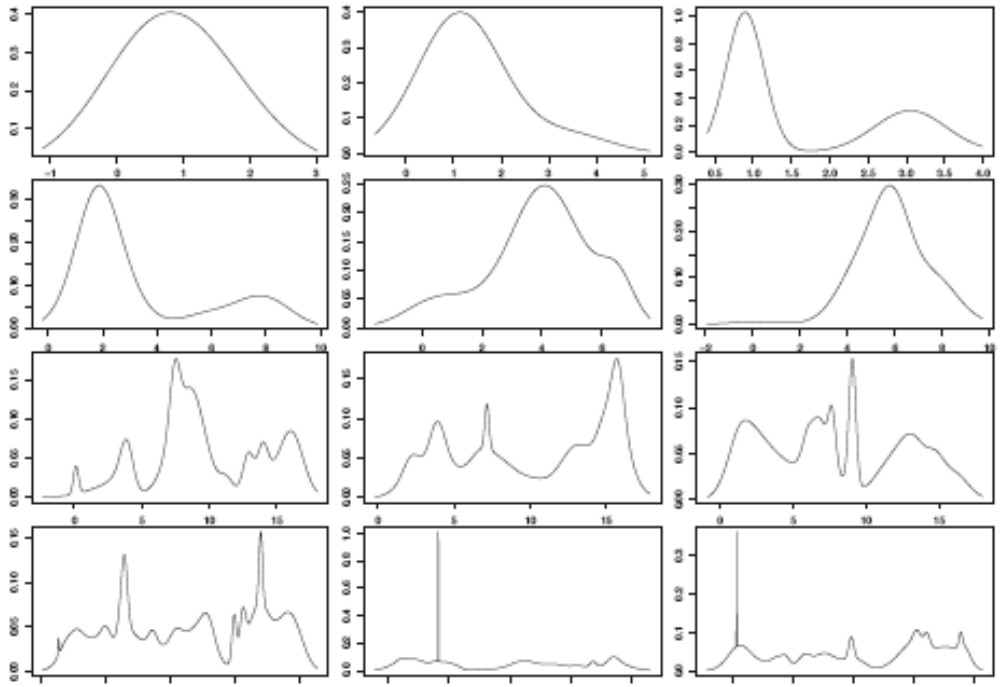


FIGURE 1. Some normal mixture densities for $K = 2$ *(first row)*, $K = 5$ *(second row)*, $K = 25$ *(third row)* and $K = 50$ *(last row)*.

$$dP(x) = \sum_k p_k dP_k(x|\theta)$$

there is a variable number of classes

the probabilities of belonging to a given class are drawn from a multinomial distribution

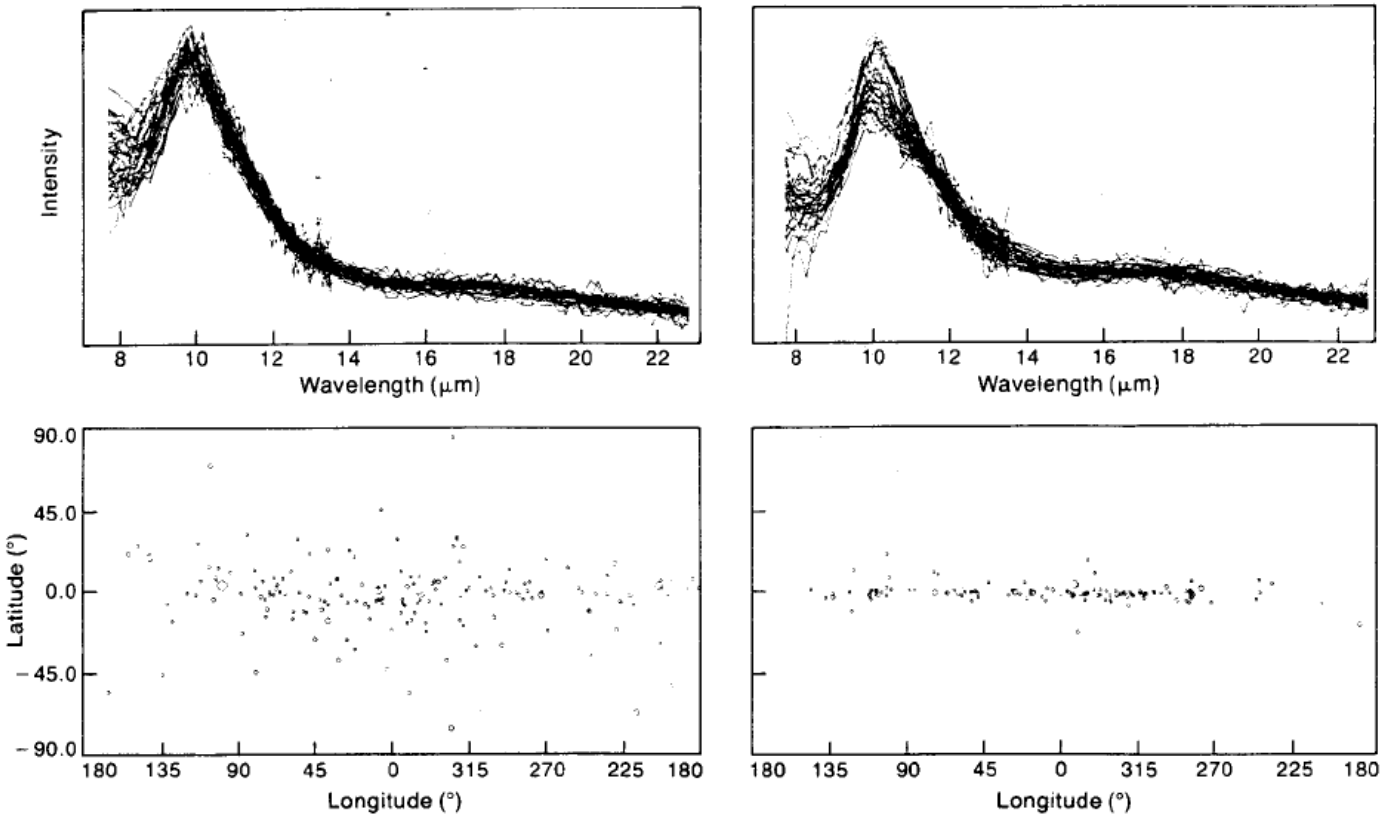
the component distributions are taken from a set of predefined distributions

the parameters define the shape of the component distribution

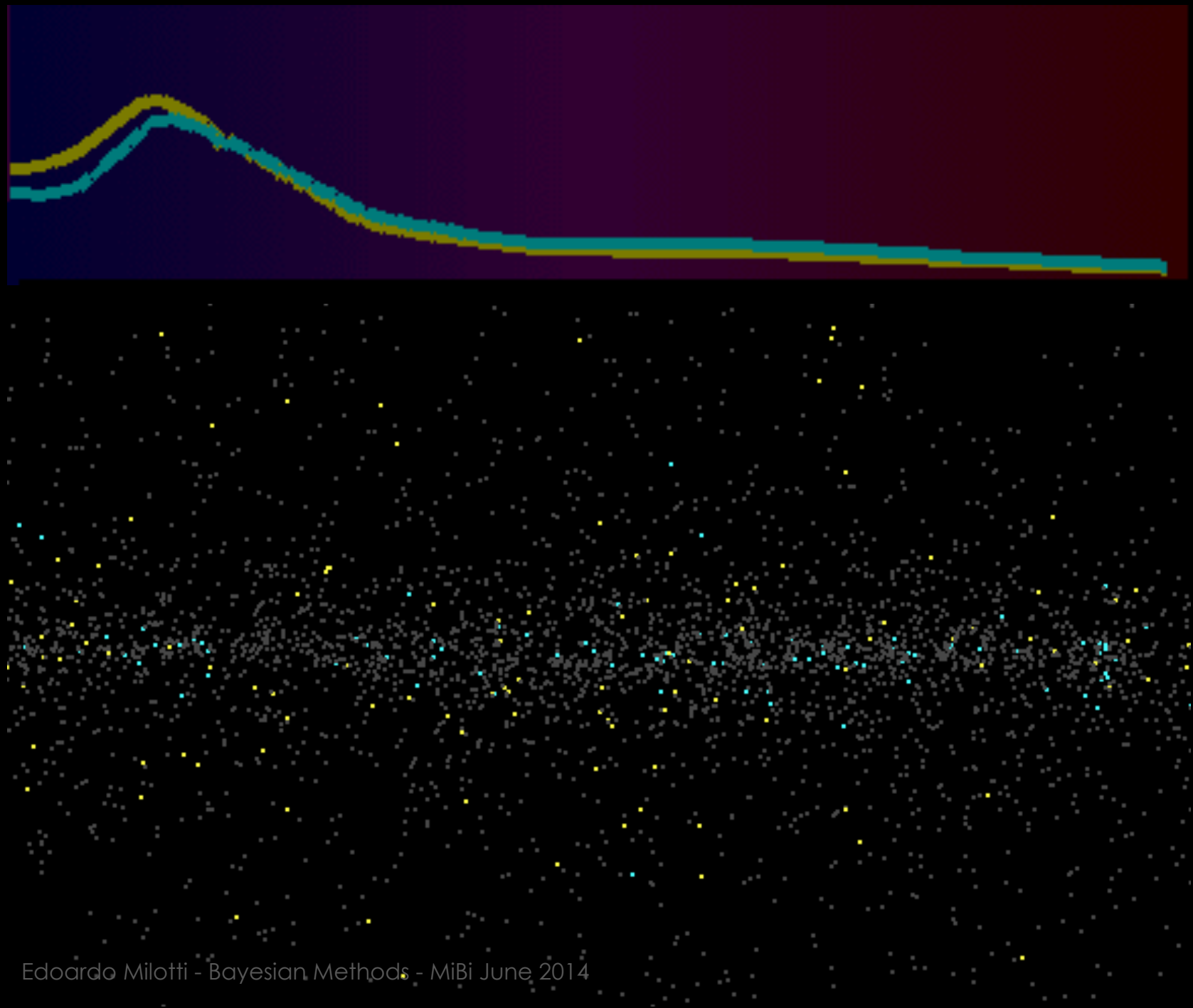
AUTOCLASS chooses a distribution and a parameter set for each class. Every data set determines a likelihood, and therefore a posterior distribution.

The class is selected by maximizing the posterior probability (MAP class estimate).

AUTOCLASS discoveries



In 1983 and 1984, the Infrared Astronomical Satellite (IRAS) detected 5,425 stellar objects and measured their infrared spectra. A program called AUTOCLASS used Bayesian inference methods to discover the classes present in the data and determine the most probable class of each object. It discovered some classes that were significantly different from those previously known to astronomers. One such discovery is illustrated above. Previous analysis had identified a set of 297 objects with strong silicate spectra. AUTOCLASS partitioned this set into two parts *(top)*. The class on the left (171 objects) has a peak at 9.7 microns and the class on the right (126 objects) a peak at 10.0 microns. When the objects are plotted on a star map by their celestial coordinates *(bottom)*, the right set shows a marked tendency to cluster around the galactic plane, confirming that the classification represents real differences between the classes of objects. AUTOCLASS did not use the celestial coordinates in its estimates of classes. Astronomers are studying the phenomenon further to determine the cause.



Welcome to AutoClass@IJM
the webserver for AutoClass Bayesian clustering system.

Developped by F. Achcar[1,2] and D. Mestivier[1] in collaboration with J.M. Camadro[2]

We kindly ask users to cite this paper when publishing results derived of the use of AutoClass@IJM.

References:

• 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)

• C. Elkan: “Naive Bayesian Learning”, CS97-557 tech. rep. UCSD

• Tom Mitchell: draft of new chapter for “Machine Learning”, http://www.cs.cmu.edu/%7Etom/NewChapters.html

• AUTOCLASS @ NASA: http://ti.arc.nasa.gov/tech/rse/synthesis-projects-applications/autoclass/

• AUTOCLASS @ IJM: http://ytat2.ijm.univ-paris-diderot.fr/