PARTICLES AND COSMOLOGY

17th Baksan School on Astroparticle Physics

Modern Statistical Methods and Tools

Lecture 2

Grigory I. Rubtsov
Institute for Nuclear Research of the Russian Academy of Sciences

Terskol, Kabardino-Balkarian Republic, Russia
April 4-11, 2025

Update to lecture 1. Definition of discovery

- As you have noticed, a simplified definition of the discovery was shown yesterday
- It is possible to discover something that was not searched for and even something for which there is no model
- There are methods developed to search for unknown new physics.
  - General name for these methods is “semi-supervised anomaly detection”
  - These methods use M0 model, but no M model
- Larger statistics is required for discoveries of this type

M. Kuusela et al., J.Phys.Conf.Ser. 368 (2012) 012032; V. Belis et al., Rev.Phys. 12 (2024) 100091

Update to lecture 1. the opposite side: Blinding

- When searching for anomalies, one is exposed to fluctuations of different random processes
- These fluctuations make up a large background for a search
- To avoid that, the blinding technique is used
- Blinding practically means that the scientists do not have access to the data before certain point (e.g. Higgs@LHC)

1) The work is performed with simulations (M_0 and M). Then M is fixed based on simulations and published
2) Unblinding: the data are tested against M

- The data may be required for optimization on step 1. A part of data is used, which is then excluded on step 2.

Return to randomness: Gaussian random variables

- Multivariate Gaussian distribution

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{det\, C}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{b})^T C^{-1}(\boldsymbol{x}-\boldsymbol{b})\right)$$

- **b** – mean, C — covariance matrix
- For random Gaussian **x** with **b**=0 and any matrix A

$$Tr\, A = \langle \boldsymbol{x}^{\boldsymbol{T}} A C^{-1} \boldsymbol{x} \rangle$$

- For
- For random Gaussian **x** with **b**=0 and any matrix A

Return to randomness: Gaussian random variables

- Isserlis-Wick theorem for calculating the mean of the product of Gaussian variables
  - Isserlis – 1918 (mathematics)
  - Wick – 1950 (particle physics)
- Mean of the product of the Gaussian variables (assume **b**=0) is the sum of products of means over all possible pairings
- Example:

$$\langle x_1 x_2 x_3 x_4 \rangle = \langle x_1 x_2 \rangle \langle x_3 x_4 \rangle + \langle x_1 x_3 \rangle \langle x_2 x_4 \rangle + \langle x_1 x_4 \rangle \langle x_2 x_3 \rangle$$

- As a direct consequence:

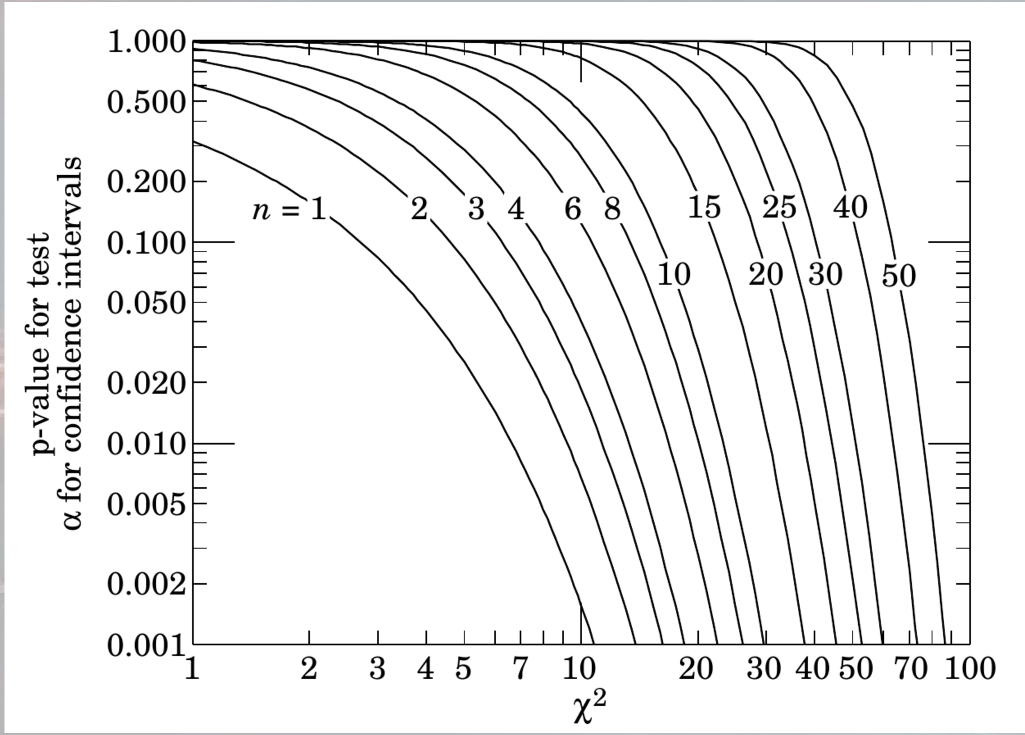
$$\langle x^4 \rangle = 3 \langle x^2 \rangle \langle x^2 \rangle = 3 \sigma^4$$

Return to randomness: Gaussian random variables

- For n Gaussian random variables x_i with zero mean one may define

$$\chi^2 = \sum_{i=0}^{n-1} x_i^2$$

- The χ^2 distribution depends on n (called d.o.f.) and is widely used



Particle Data Group

Frequentist vs Bayesian

The future is not set.

There is no fate but what we make for ourselves.

The past, present and future are not set.

The fate is a random hypothesis.

Bayesian approach

- Both model (M) and event (obs) are random

$$P(M|obs) = \frac{P(obs|M)\,P(M)}{P(obs)}$$

- P(M) – prior
- P(obs) – normalization constant we neglect at this step and recover later (by normalizing posterior)

$$P(M|obs) \sim P(obs|M)\,P(M)$$

Bayesian approach

$$P(M|obs) \sim P(obs|M)P(M)$$

- P(obs|M) is called likelihood $\mathcal{L}$(M,obs)
- P(M|obs) – posterior probability
- One often confuses the likelihood and the posterior probability.
- Q: What is the difference between them?

Bayesian approach

$$P(M|obs) \sim P(obs|M)\,P(M)$$

- P(obs|M) is called likelihood L(M,obs)
- P(M|obs) – posterior probability
- One often confuses the likelihood and the posterior probability.
- Q: What is the difference between them?
- A: These variables have a meaning of probability in different probability spaces

Bayesian approach

$$P(M|obs) \sim P(obs|M)\,P(M)$$

- The likelihood P(obs|M) is a probability in the space of random events (it is the probability in Frequentist’s approach)
- The posterior probability P(M|obs) is a probability in the space of random models

Bayesian approach

$$P(M|obs) \sim P(obs|M)\,P(M)$$

- Lost in spaces? Luckily, there is a clear way to identify the probability and it's space.
- Normalization condition

$$\int_{obs} P(obs|M) = 1$$

$$\int_{M} P(M|obs) = 1$$

Bayesian approach: work with posterior probability

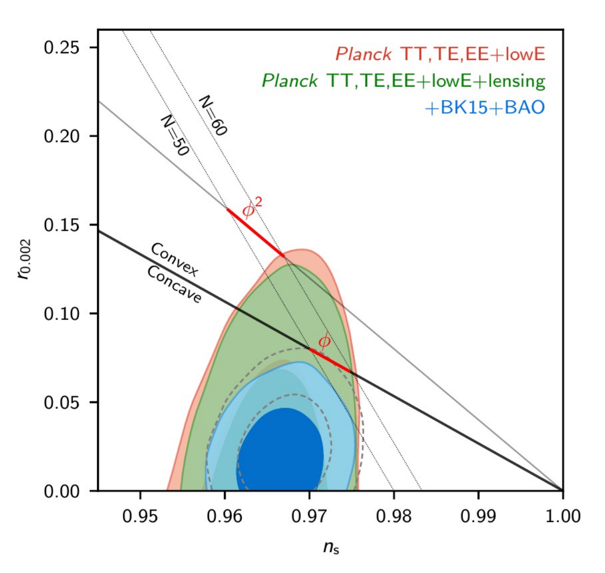
- Let us assume that M is parametrized by the K variables $\{m_k\}$
- Normalization condition may be written explicitly

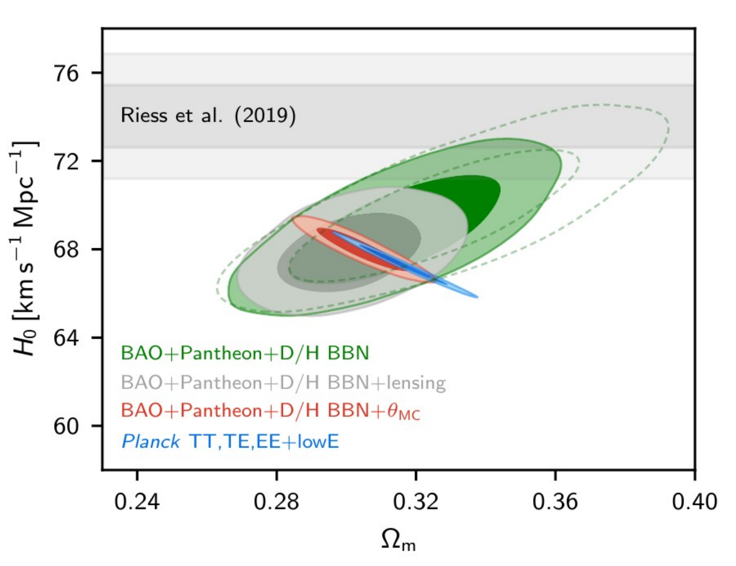
$$\iint_{m_1..m_K} P(M|obs)\,dM = 1$$

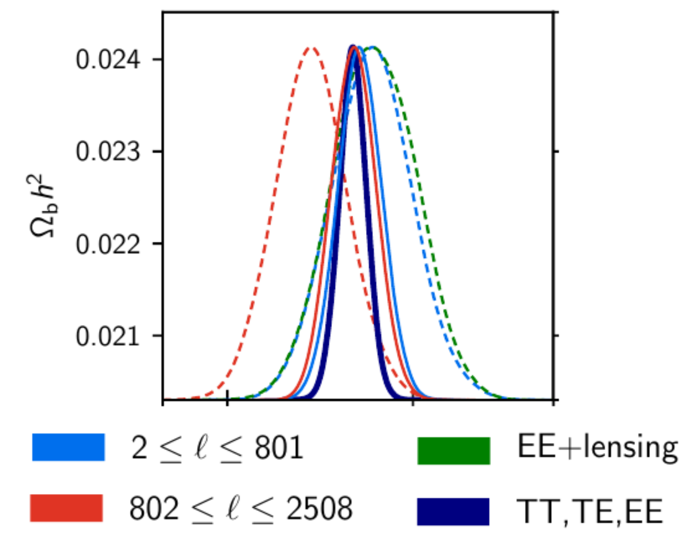
- Suppose we are exclusively interested in one or two parameters of the model. We calculate marginal distribution

$$p(m_l) = \frac{\iint_{m_1..m_K \backslash m_l} P(M|obs)\,dM}{\iint_{m_1..m_K} P(M|obs)\,dM} \qquad p(m_l, m_q) = \frac{\iint_{m_1..m_K \backslash m_l m_q} P(M|obs)\,dM}{\iint_{m_1..m_K} P(M|obs)\,dM}$$

Bayesian approach example: Planck 2018 results







Planck Collaboration, A&A 641, A6 (2020)

- These are 2D and 1D marginal distributions of posterior
- 1σ (2σ) contours – lines of equal probability, which include 68%, (95%) of the integral of posterior probability

Testing hypotheses: Bayesian approach

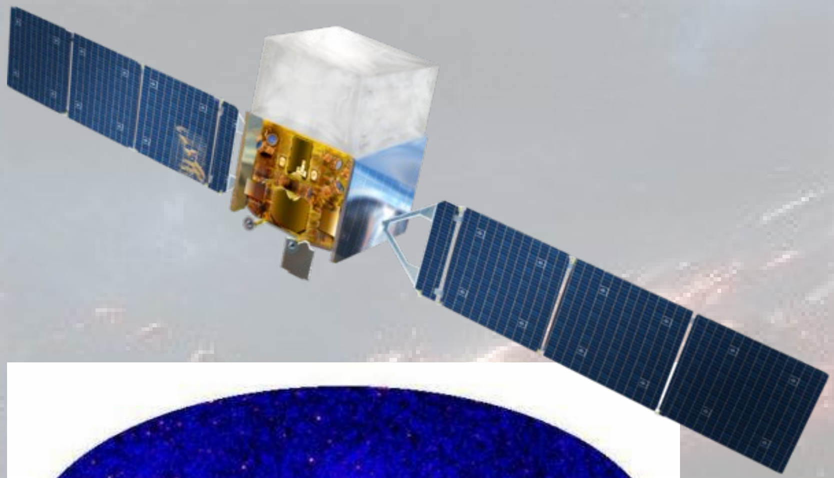
1) Define the space of models M
2) Define the likelihood function P(obs|M)
3) Define the prior P(M)
4) Calculate the posterior probability
5) Calculate marginal 1D or 2D distribution of the posterior
6) Plot the lines of equal probability, which include 68%, (95%) of the integral of posterior probability. These are the constraint we obtain

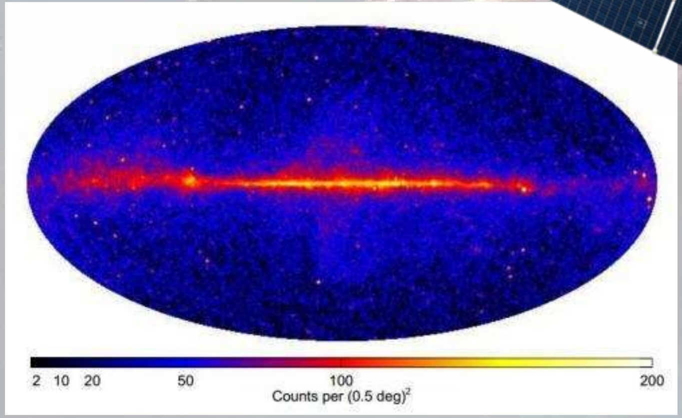
Takeout 2.1

- Gaussian random variables have unique properties and are widely used in the analysis
- Posterior probability and likelihood have a meaning of probability in different probability spaces
- The parameters of the models are studied in the Bayesian approach with the marginal distributions of the posterior probability
- The constraints on the parameters are obtained with the line of equal probability

Model example: gamma-ray sky observed by Fermi LAT

- Fermi LAT is a space gamma-ray telescope
- We will use the publicly available list of the photons and exposure to test the radiation models
- Fermi LAT observes photons starting from 100 MeV
- We'll constrain ourselves with the gamma-rays above 10 GeV for smaller data and computation volume





Fermi LAT Collaboration, E>10 GeV

Model example: Fermi LAT

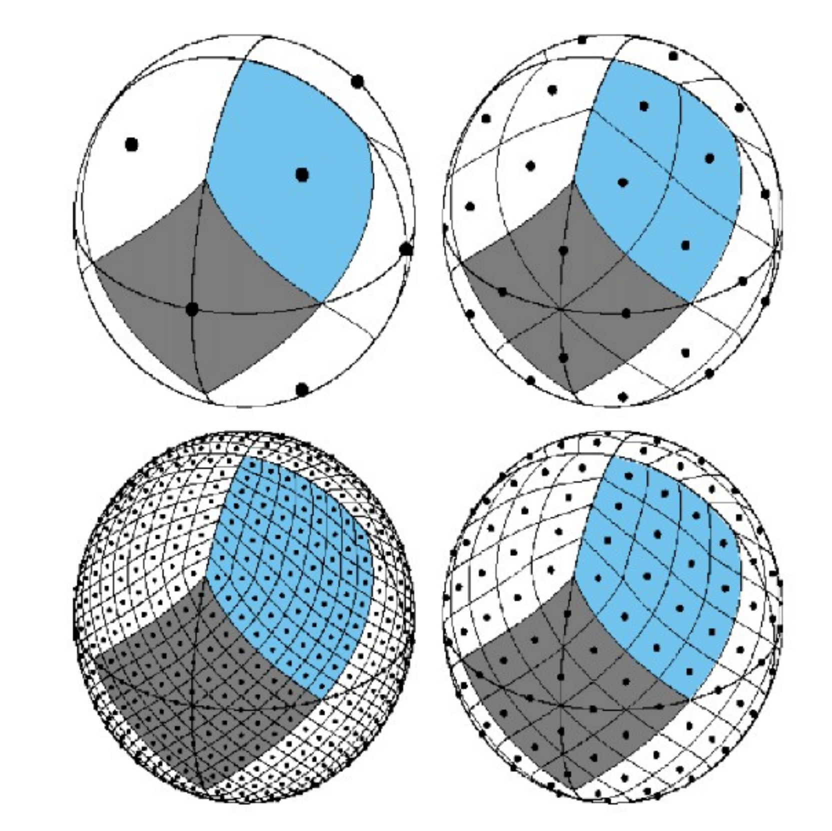
- The model of the gamma-ray emission is defined as a function on the position of the sphere f(Ω) in cm^{-2} s^{-1} sr^{-1}
- Will work in Galactic coordinates and use Ω for (l,b)
- We have an exposure X(Ω) of the experiment as a function of Ω for energy E=10 GeV in cm^{2} s^{2}
- The predicted probability density ρ(Ω) = f(Ω)X(Ω)
- The next step is to construct a likelihood

HEALPix: Pixelisation of the sphere

- HEALPix — Hierarchical Equal Area isoLatitude Pixelisation of a sphere
- Two types: ring or nested
- Npix = 12 $nside^2$

```
from healpy.pixelfunc:

  pix2ang(nside, ipix[, nest, lonlat])
  ang2pix(nside, theta, phi[, nest, lonlat])
```



Model example: Fermi LAT Likelihood

- We have pixels with area $\Delta\Omega$
- Expect $m_i = \rho(\Omega)\, \Delta\Omega$ events in a pixel
- Observe n_i events in a pixel
- Q: What is a likelihood?

Model example: Fermi LAT Likelihood

- We have pixels with area $\Delta\Omega$
- Expect $m_i = \rho(\Omega)\,\Delta\Omega$ events in a pixel
- Observe n_i events in a pixel
- Q: What is a likelihood?
- A: Binned likelihood is a product of Poisson distributions:

$$P(obs|M) = \prod_i W(m_i, n_i) = \prod_i \frac{m_i^{n_i}}{n_i!} \exp(-m_i) = \exp\left(-\sum m_i\right) \prod_i \frac{m_i^{n_i}}{n_i!}$$

Model example: Fermi LAT Likelihood

- Expect $m_i = \rho(\Omega_i)\,\Delta\Omega$, observe n_i events in a pixel
- Binned likelihood is a product of Poisson distributions:

$$P(obs|M) = \prod_i W(m_i, n_i) = \prod_i \frac{m_i^{n_i}}{n_i!} \exp(-m_i) = \exp\left(-\sum m_i\right) \prod_i \frac{m_i^{n_i}}{n_i!}$$

- Consider the limit $\Delta\Omega \to 0$, then n_i is either 0 or 1
- If $n_i = 0$, the term in a product equals to 1, keep only $n_i = 1$
- Let Ω_a be a coordinate of a-th event, a=1..N
- We arrive at unbinned likelihood

$$P(obs|M) = \exp\left(-\sum_i \rho(\Omega_i)\,\Delta\Omega\right) \prod_a \left(\rho(\Omega_a)\,\Delta\Omega\right)$$

Model example: Fermi LAT Likelihood

$$P(obs|M)=\exp\left(-\sum_i \rho(\Omega_i)\,\Delta\Omega\right)\prod_a \left(\rho(\Omega_a)\,\Delta\Omega\right)$$

$$P(obs|M)=\exp\left(-\int_\Omega \rho(\Omega)\,d\Omega\right)\Delta\Omega^N\prod_a \rho(\Omega_a)$$

- Removing constant normalization factor we arrive to final version of unbinned likelihood

$$P(obs|M)=\exp\left(-\int_\Omega \rho(\Omega)\,d\Omega\right)\prod_a \rho(\Omega_a)$$

Likelihood ratio test

- Suppose we have two models M_0 with N parameters and M_1 with N+q parameters
- We have best fit likelihoods for M_0 and M_1

$$\lambda = -2\left(\ln\left(L\left(M_0\right)\right) - \ln\left(L\left(M_1\right)\right)\right)$$

- If the L improvement is due to random fluctuation, λ is distributed according to χ^2 distribution with q degrees of freedom
- If λ value is improbable according to χ^2 distribution, the model extension is physics (e.g. new source exists)
- Confidence level is obtained from the above probability

Takeout 2.2

- One may use Bayesian approach to study gamma-ray sky
- The sky may be split into the pixels with the HEALPix library (healpy)
- Two types of likelihood may be constructed (binned and unbinned)
- The likelihood ratio test may be used to compare models with different number of parameters

Task for self-check

- Download the list of Fermi LAT photons and exposure from data directory at Yandex disk

fermi_photons_10GeV.dat - photons, registered by Fermi LAT with energy greater than 10 GeV

Time period:

2008-08-04T15:43:36.4941 - 2024-08-09T03:08:40.9339

File format (column description):

1. E, MeV
2. l, deg - Galactic longitude
3. b, deg - Galactic latitude
4. MET, s - photon arrival time

Task for self-check

- Download the exposure of Fermi LAT at 10 GeV

fermi_expo_10GeV.dat - exposure of Fermi LAT telescope for the total time period given below and energy equal to 10 GeV

Time period:

2008-08-04T15:43:36.4941 - 2024-08-09T03:08:40.9339

File format (column description):

1. l, deg - Galactic longitude
2. b, deg - Galactic latitude
3. exposure, cm^2 s

Task for self-check

- Construct a model of gamma-ray radiation with two sources:
  - Isotropic flux
  - Constant flux in a circle with a radius of 1° around Crab
- Calculate likelihood and posterior probability distribution
- Estimate the parameters of the model and significance of the Crab observation
- (*) extend the model making the source coordinates parameters of the model

Hands-on session

- Download the code
- https://disk.yandex.ru/d/bPrpOq2Z-oJIOw
- Run jupyter notebook
- Go through exercises in the notebook

Thank you!

Backup slides