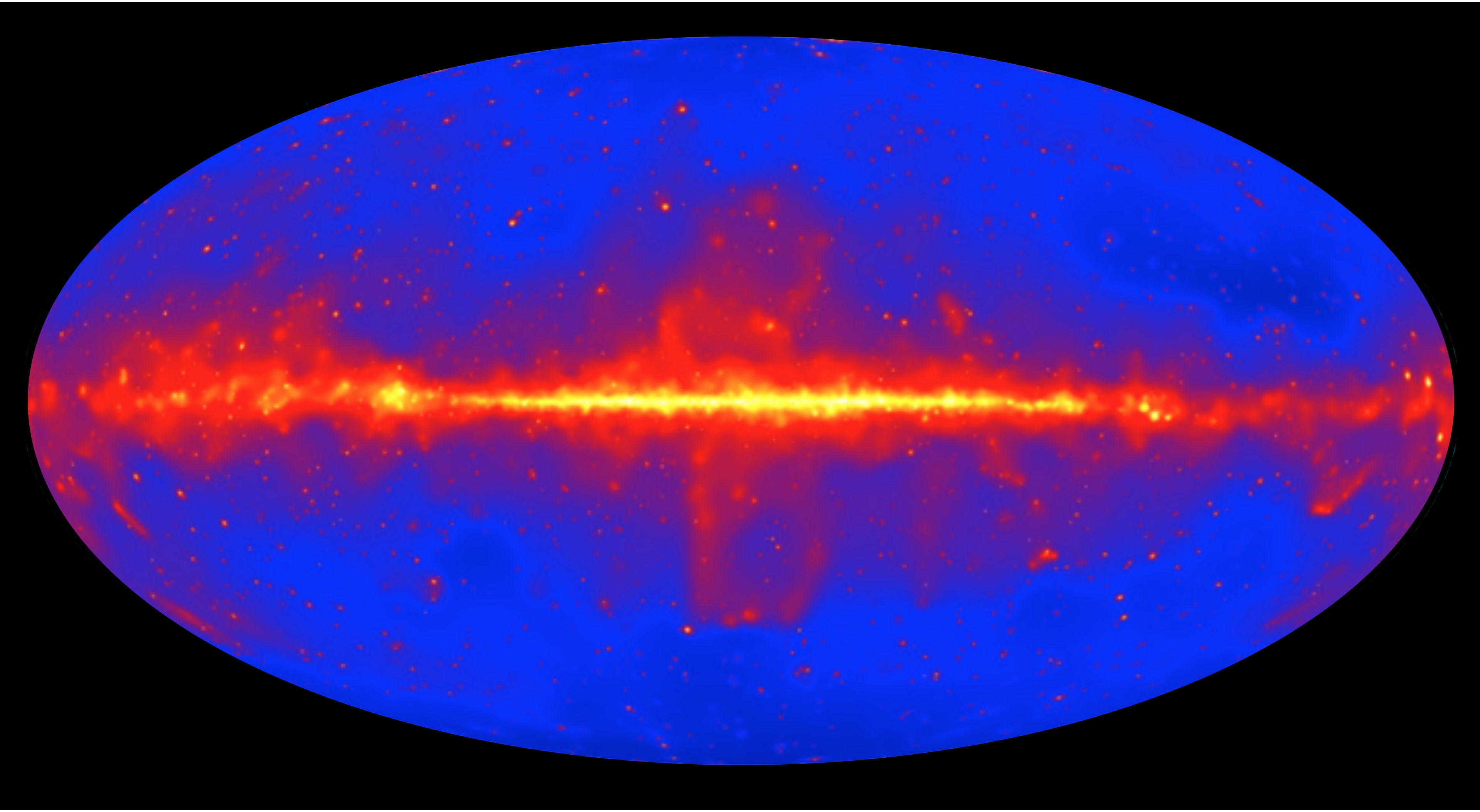
Intro to Maximum Likelihood

Mattia Di Mauro

(heavily inspired by Liz Hays and Steve Fegan’s 2013 notes)

Fermi-LAT Summer School 2019

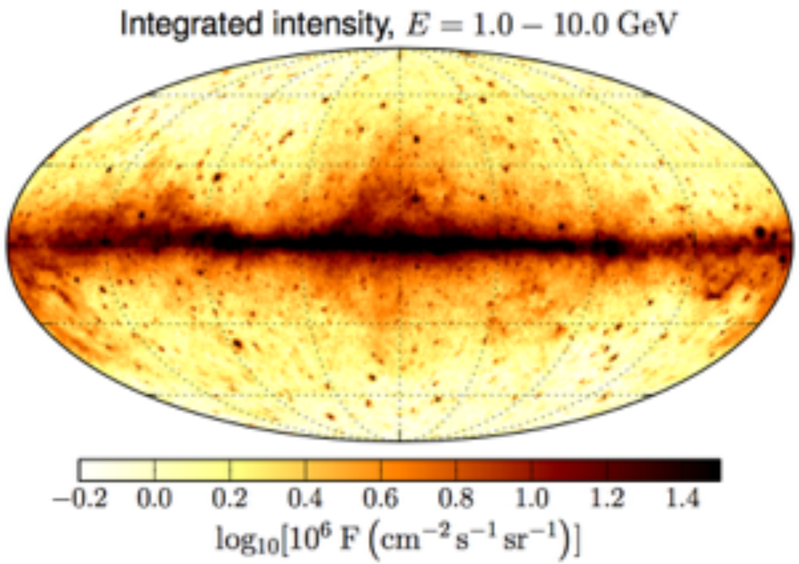
Fermi LAT E>10 GeV



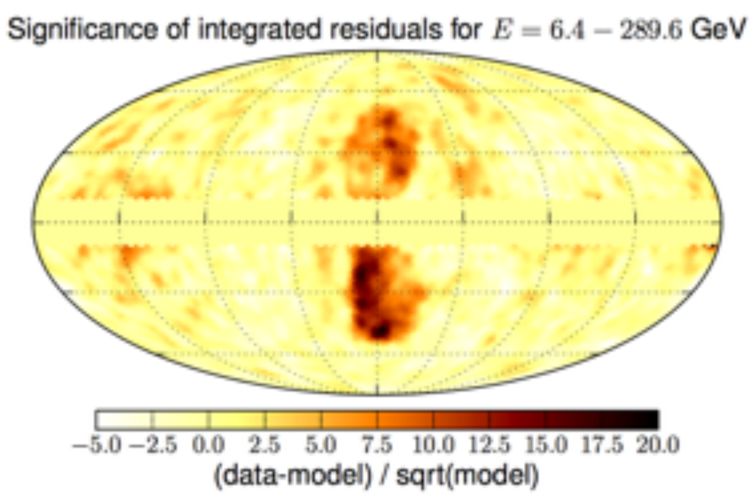
Fermi LAT E>10 GeV using 7 years of data (~700,000 photons)

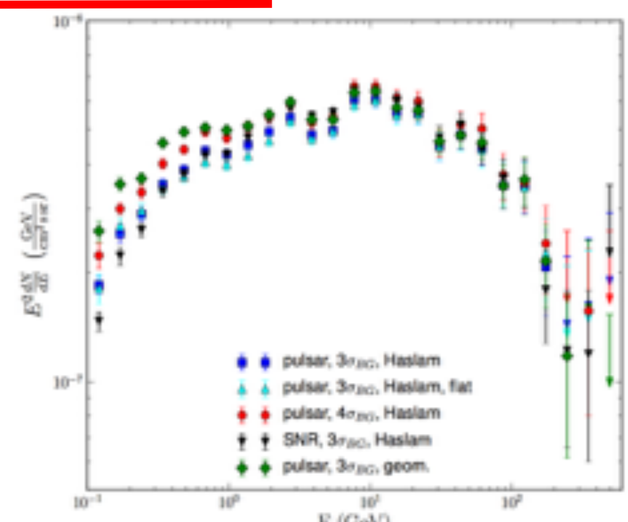
Motivation

Example: Fermi bubbles

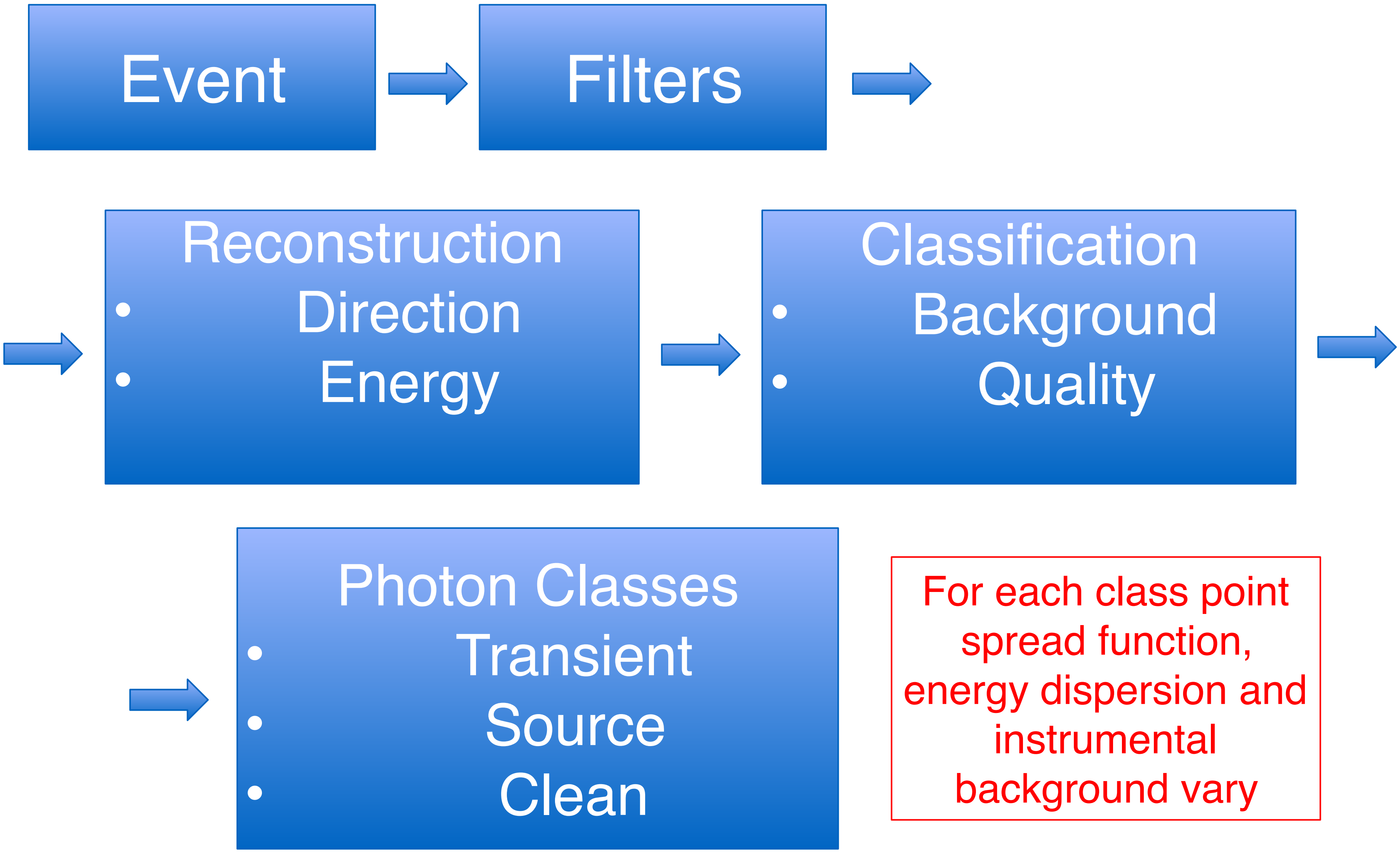


STATISTICAL METHODS





From LAT Event to Photon



From LAT Event to Photon 1

https://fermi.gsfc.nasa.gov/ssc/data/analysis/scitools/binned_likelihood_tutorial.html

Energy and time and zenith angle

```
prompt> gtselect evclass=128 evtype=3
Input FT1 file□ @binned_events.txt
Output FT1 file□ 3C279_binned_filtered.fits
RA for new search center (degrees) (0:360) □ INDEF
Dec for new search center (degrees) (-90:90) □ INDEF
radius of new search region (degrees) (0:180) □ INDEF
start time (MET in s) (0:) □ INDEF
end time (MET in s) (0:) □ INDEF
lower energy limit (MeV) (0:) □ 100
upper energy limit (MeV) (0:) □ 500000
maximum zenith angle value (degrees) (0:180) □ 90
Done.
prompt>
```

Photon quality

```
prompt>gtmktime
Spacecraft data file□ L181126210218F4F0ED2738_SC00.fits
Filter expression□ (DATA_QUAL>0)&&(LAT_CONFIG==1)
Apply ROI-based zenith angle cut□ no
Event data file□ 3C279_binned_filtered.fits
Output event file name□ 3C279_binned_gti.fits
prompt>
```

Bin the data in energy and space

```
prompt> gtbin
Type of output file (CCUBE|CMAP|LC|PHA1|PHA2|HEALPIX) [PHA2] CMAP
Event data file name□ 3C279_binned_gti.fits
Output file name□ 3C279_binned_cmap.fits
Spacecraft data file name□ NONE
Size of the X axis in pixels□ 150
Size of the Y axis in pixels□ 150
Image scale (in degrees/pixel)□ 0.2
Coordinate system (CEL - celestial, GAL -galactic)□ CEL
First coordinate of image center in degrees (RA or galactic l)□ 193.98
Second coordinate of image center in degrees (DEC or galactic b)□ -5.82
Rotation angle of image axis, in degrees□ 0.0
Projection method Projection method e.g. AIT|ARC|CAR|GLS|MER|NCP|SIN|STG|TAN:□ AIT
gtbin: WARNING: No spacecraft file: EXPOSURE keyword will be set equal to ontime.
prompt> ds9 3C279_binned_cmap.fits &
```

From LAT Event to Photon 2

https://fermi.gsfc.nasa.gov/ssc/data/analysis/scitools/binned_likelihood_tutorial.html

```
prompt> make4FGLxml.py gll_psc_v18.fit 3C279_binned_gti.fits -o 3C279_input_model.xml
-G $FERMI_DIR/refdata/fermi/galdiffuse/gll_iem_v07.fits -g gll_iem_v07
-I $FERMI_DIR/refdata/fermi/galdiffuse/iso_P8R3_SOURCE_V2_v1.txt
-i iso_P8R3_SOURCE_V2_v1 -s 120 -p TRUE
prompt>
```

Create the model

Livetime map

```
prompt> gtltcube zmax=90
Event data file[] 3C279_binned_gti.fits
Spacecraft data file[] L181126210218F4F0ED2738_SC00.fits
Output file[] 3C279_binned_ltcube.fits
Step size in cos(theta) (0.:1.)[] 0.025
Pixel size (degrees)[] 1
Working on file L181126210218F4F0ED2738_SC00.fits
.....................!
prompt>
```

Exposure map

```
prompt> gtexpcube2
Livetime cube file[] 3C279_binned_ltcube.fits
Counts map file[] none
Output file name[] 3C279_binned_expcube.fits
Response functions to use[] P8R3_SOURCE_V2
Size of the X axis in pixels[] 300
Size of the Y axis in pixels[] 300
Image scale (in degrees/pixel)[] .2
Coordinate system (CEL - celestial, GAL -galactic) (CEL|GAL) [] CEL
First coordinate of image center in degrees (RA or galactic l)[] 193.98
Second coordinate of image center in degrees (DEC or galactic b)[] -5.82
Rotation angle of image axis, in degrees[] 0
Projection method e.g. AIT|ARC|CAR|GLS|MER|NCP|SIN|STG|TAN[] AIT
Start energy (MeV) of first bin[] 100
Stop energy (MeV) of last bin[] 500000
Number of logarithmically-spaced energy bins[] 37
Computing binned exposure map....................!
```

Create the prediction for the model

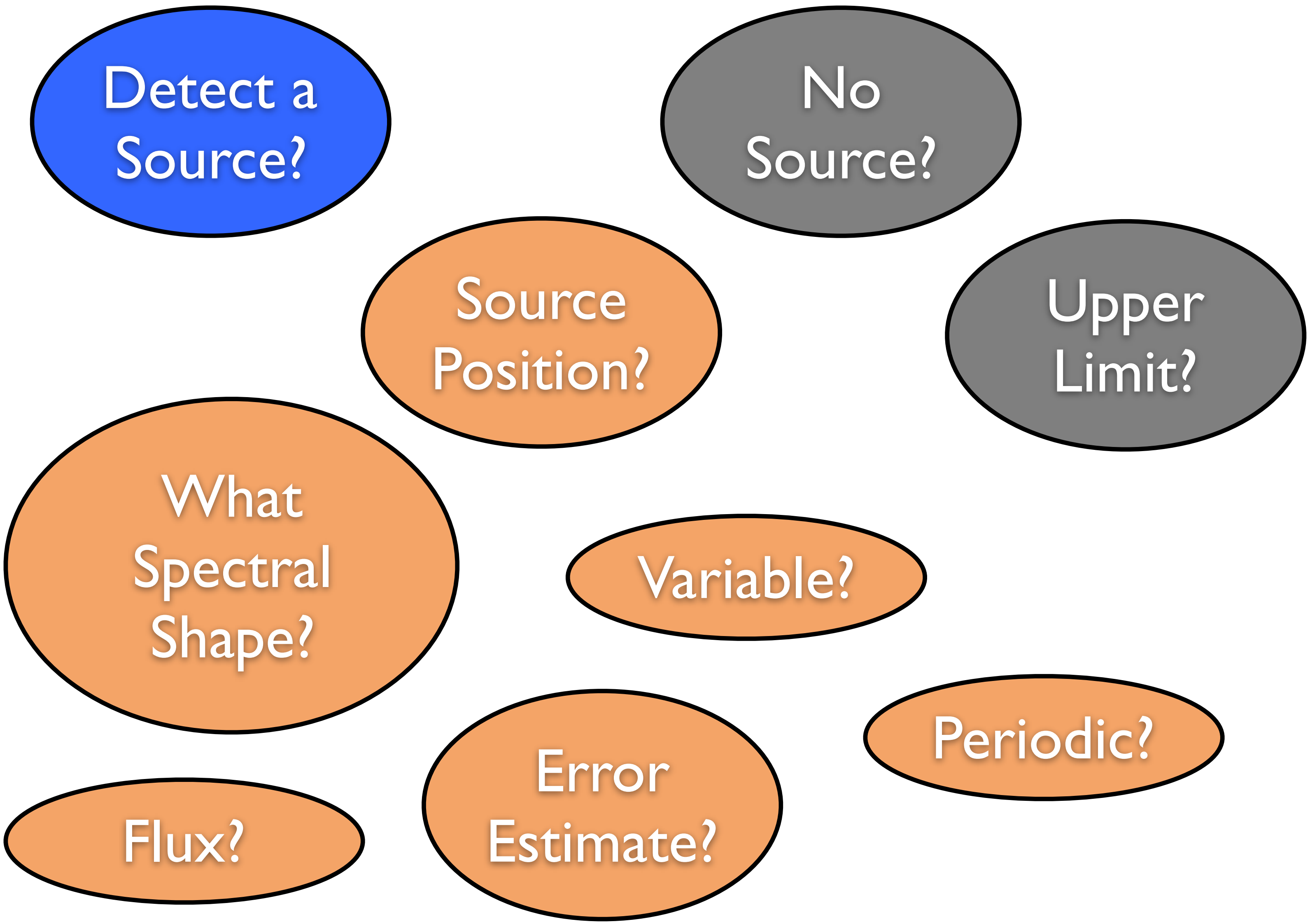
```
prompt> gtsrcmaps
Exposure hypercube file[] 3C279_binned_ltcube.fits
Counts map file[] 3C279_binned_ccube.fits
Source model file[] 3C279_input_model.xml
Binned exposure map[] 3C279_binned_allsky_expcube.fits
Source maps output file[] 3C279_binned_srcmaps.fits
Response functions[CALDB]
```

Perform the fit

```
prompt> gtlike refit=yes plot=yes sfile=3C279_binned_output.xml

Statistic to use (BINNED|UNBINNED)[] BINNED
Counts map file[] 3C279_binned_srcmaps.fits
Binned exposure map[] 3C279_binned_allsky_expcube.fits
Exposure hypercube file[] 3C279_binned_ltcube.fits
Source model file[] 3C279_input_model.xml
Response functions to use[] CALDB
Optimizer (DRMNFB|NEWMINUIT|MINUIT|DRMNGB|LBFGS)[] NEWMINUIT
```

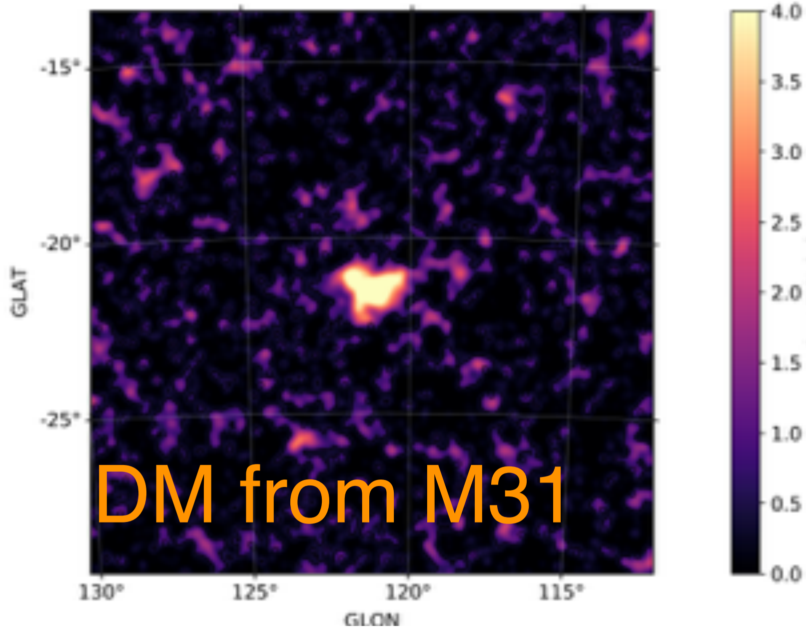
What can I infer from my observation?

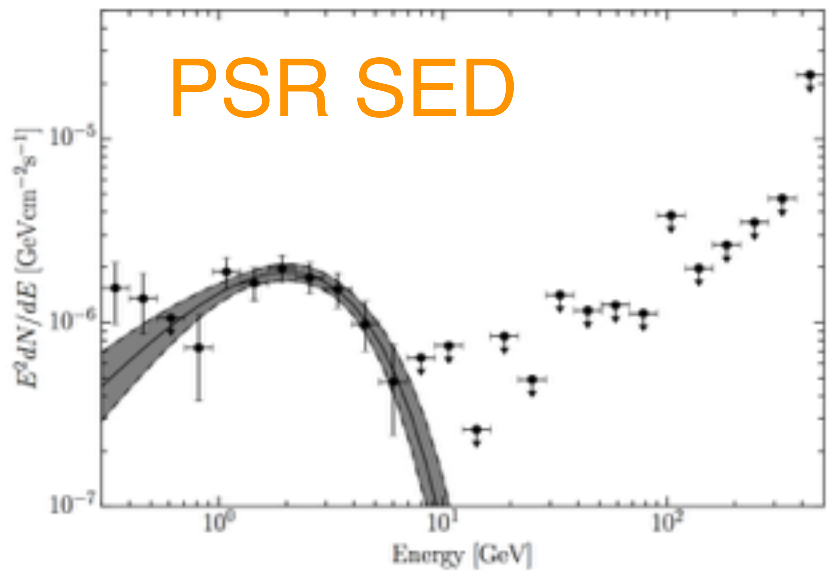


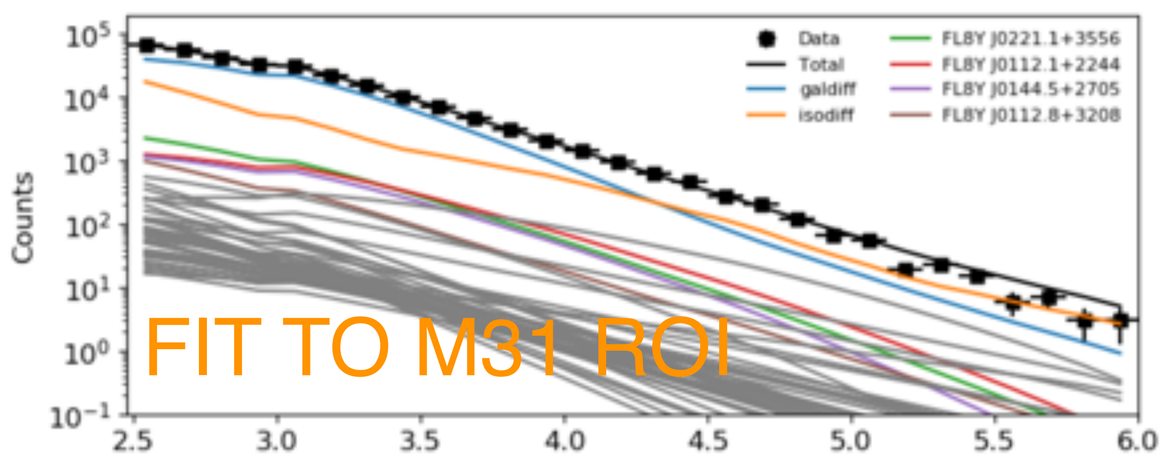
Measurements in γ-ray astronomy

- Is a source significantly detected?
  - If so, what is its flux?
  - If not, what is upper limit on the flux?
- What kind of spectrum does it have?
  - What is its spectral index?
- What is its location in the sky?
- What are the errors on these values?
- Is the source variable?

What can I infer from my observation?







The Method of Maximum Likelihood

"a simple recipe that purports to lead to the optimum solution for all parametric problems and beyond" ~ Stigler

Long history of evaluation: Gauss, Laplace, Fisher, Wilks…

Broad applicability to many measurement problems.

Good things about maximum likelihood

- General framework for statistical questions.
- Unbiased, minimum variance estimate as sample size increases.
- Asymptotically Gaussian: allows evaluation of confidence bounds & hypothesis testing.
- Well studied in the literature.
- Starting point for Bayesian analysis.

~~Good things~~ Cautions about maximum likelihood

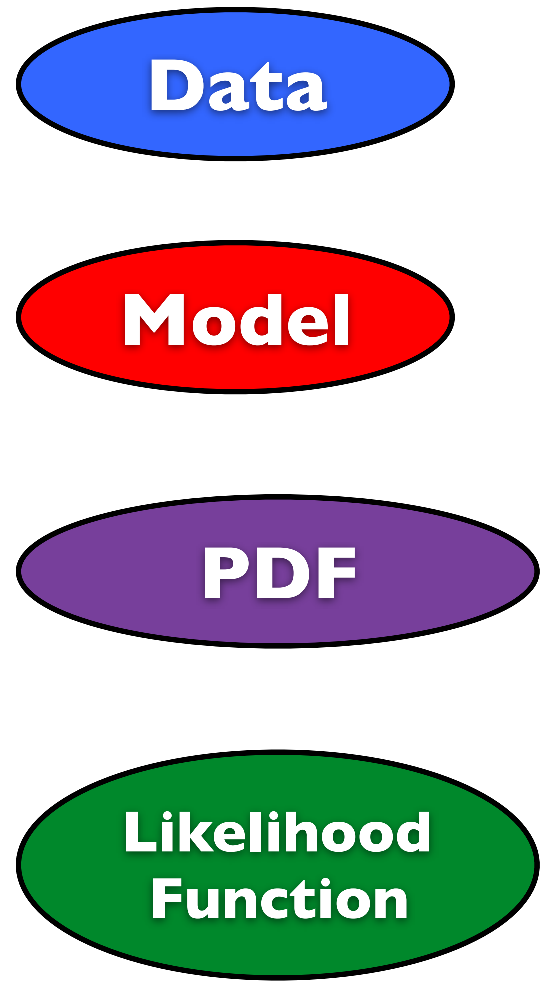
- General framework for statistical questions.
- Unbiased, minimum variance estimate as sample size increases.
- Asymptotically Gaussian: allows evaluation of confidence bounds & hypothesis testing.
- Well studied in the literature.
- Starting point for Bayesian analysis.

- Only answers the question asked.
- Be aware of small number regimes and departure from Gaussian assumption
- Starting point for Bayesian analysis.

Maximum likelihood technique

Given a set of observed data

- produce a model that *accurately* describes the data, including parameters that we wish to estimate,
- derive the probability (density) for the data given the model (probability density function, PDF),
- treat this as a function of the model parameters (likelihood function), and
- maximize the likelihood with respect to the parameters - ML estimation.



Maximum likelihood basics

Data $X = \{x_i\} = \{x_1, x_2, ..., x_N\}$

Model $\Theta = \{\theta_j\} = \{\theta_1, \theta_2, ..., \theta_M\}$

Likelihood Function $\mathcal{L}(\Theta|X) = P(X|\Theta)$

- Conditional probability rule for independent events: $P(A, B) \underbrace{=}_{\text{CPR}} P(A)P(B|A) \underbrace{=}_{\text{Independence}} P(A)P(B)$
- For independent data:

$$P(X|\Theta) = P(\{x_i\}|\Theta) = P(x_1|\Theta)P(x_2, .., x_N|\Theta) = \cdots$$

$$= P(x_1|\Theta)P(x_2|\Theta)\cdots P(x_N|\Theta) = \prod_i P(x_i|\Theta)$$

$$\mathcal{L}(\Theta|X) = \prod_i P(x_i|\Theta)$$

ML estimation (MLE)

- Parameters can be estimated by maximizing likelihood. Easier to work with log-likelihood:

$$\boxed{\ln \mathcal{L}(\Theta) = \ln \mathcal{L}(\Theta|X) = \sum_i \ln P(x_i|\Theta)}$$

- Estimates of parameters $\{\hat{\theta}_k\}$ from solving simultaneous equations: $$\left.\frac{\partial \ln \mathcal{L}}{\partial \theta_j}\right|_{\{\hat{\theta}_k\}} = 0$$

- For one parameter, if we have: $\mathcal{L}(\theta) \sim e^{-\frac{(\theta-\hat{\theta})^2}{2\sigma_\theta^2}}$

  then: $$\left.\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\right|_{\hat{\theta}} = -\frac{1}{\sigma_\theta^2}$$

  Gaussian approximation

  2nd derivative is related to "errors"

Example: Normal distribution

Data

- Suppose the data x1,x2,...,xn is drawn from a N(μ,σ^2) distribution, where μ and σ are unknown.

(x_i, σ_i)

Model

- all measurements are of a constant flux with Gaussian errors

F

Probabilities

$$P(x_i|F) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x_i - F)^2}{2\sigma_i^2}}$$

Likelihood Function

$$\ln \mathcal{L}(F) = -\sum \frac{(x_i - F)^2}{2\sigma_i^2} - \sum \ln \sigma_i - \frac{N}{2} \ln 2\pi$$

Example: χ^2 fit of constant

$$f_{X_i}(x_i) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}.$$

Since the X_i are independent their joint pdf is the product of the individual pdf's:

$$f(x_1,\ldots,x_n\,|\,\mu,\sigma) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n e^{-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}}.$$

For the fixed data $x_1,\ldots,x_n$, the likelihood and log likelihood are

$$f(x_1,\ldots,x_n|\mu,\sigma) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n e^{-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}}, \quad \ln(f(x_1,\ldots,x_n|\mu,\sigma)) = -n\ln(\sqrt{2\pi}) - n\ln(\sigma) - \sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}.$$

Since $\ln(f(x_1,\ldots,x_n|\mu,\sigma))$ is a function of the two variables μ, σ we use partial derivatives to find the MLE. The easy value to find is $\hat{\mu}$:

$$\frac{\partial f(x_1,\ldots,x_n|\mu,\sigma)}{\partial\mu} = \sum_{i=1}^n \frac{(x_i-\mu)}{\sigma^2} = 0 \;\Rightarrow\; \sum_{i=1}^n x_i = n\mu \;\Rightarrow\; \hat{\mu} = \frac{\sum_{i=1}^n x_i}{n} = \overline{x}.$$

To find $\hat{\sigma}$ we differentiate and solve for σ:

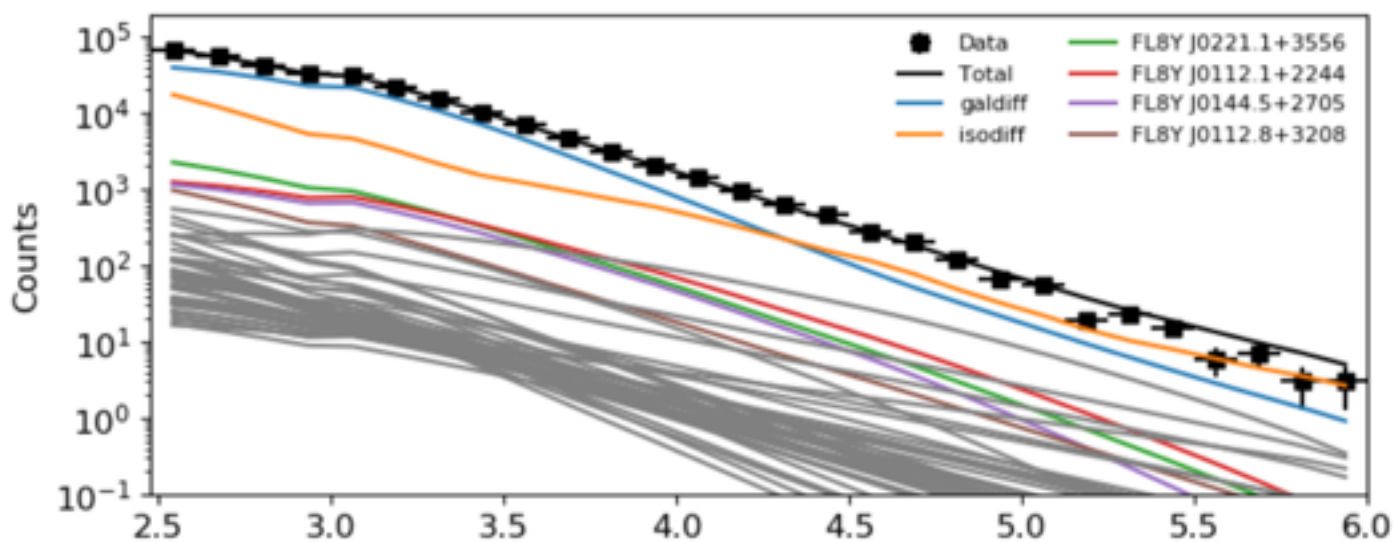
$$\frac{\partial f(x_1,\ldots,x_n|\mu,\sigma)}{\partial\sigma} = -\frac{n}{\sigma} + \sum_{i=1}^n \frac{(x_i-\mu)^2}{\sigma^3} = 0 \;\Rightarrow\; \hat{\sigma}^2 = \frac{\sum_{i=1}^n (x_i-\mu)^2}{n}.$$

We already know $\hat{\mu} = \overline{x}$, so we use that as the value for μ in the formula for $\hat{\sigma}$. We get the maximum likelihood estimates

$$\begin{aligned} \hat{\mu} &= \overline{x} && = \text{the mean of the data} \\ \hat{\sigma}^2 &= \sum_{i=1}^n \frac{1}{n}(x_i-\hat{\mu})^2 = \sum_{i=1}^n \frac{1}{n}(x_i-\overline{x})^2 && = \text{the variance of the data.} \end{aligned}$$

Fermi-LAT Analysis

- Fermi-LAT analysis is performed with photon counts.
- Photon counts are binned in energy and pixels.
- Photon counts of the data are compared to the ones from the model.
- This is done with the so called template fitting: a fit is performed varying the free parameters of the model in each energy bin independently and fitting the model to the data in each pixel.



Example: Event counting experiment

My Gamma-ray Counter ™

n events

- Model: Poisson process with mean of λ:

$$P(x|\theta) \rightarrow P(n|\lambda) = \frac{\lambda^n e^{-\lambda}}{n!}$$

- Log likelihood:

Constant WRT λ

$$\ln \mathcal{L}(\lambda) = \underbrace{n \ln \lambda}_{\text{Data cpt}} - \underbrace{\lambda}_{\text{Npred}} - \cancel{\ln n!}$$

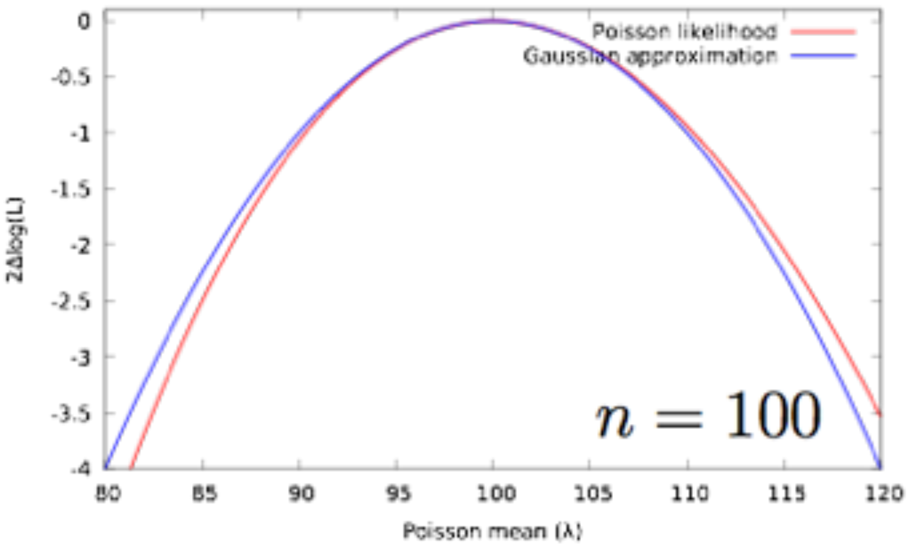
- ML estimate and error in Gaussian regime:

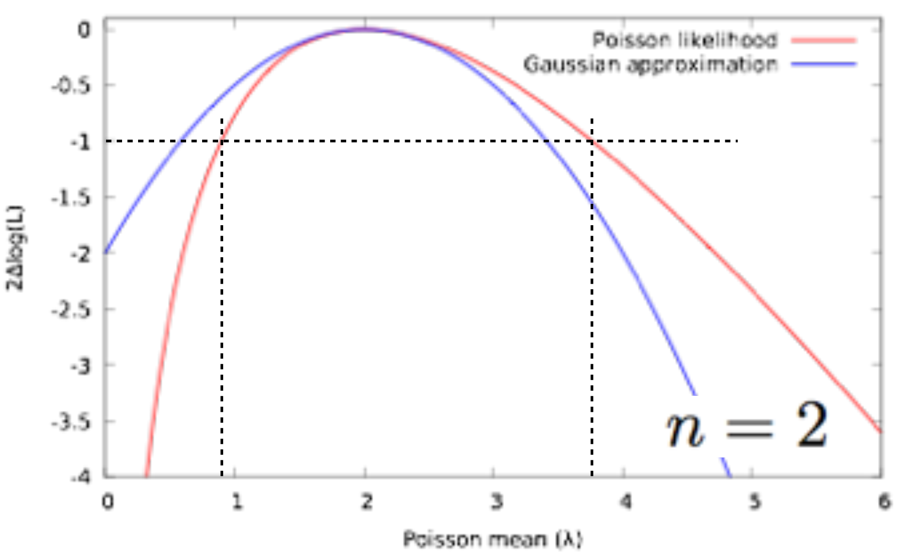
$$\frac{\partial \ln \mathcal{L}}{\partial \lambda} = \frac{n}{\lambda} - 1 \implies \boxed{\hat{\lambda} = n}$$

$$\frac{1}{\sigma_\lambda^2} = -\left.\frac{\partial^2 \ln \mathcal{L}}{\partial \lambda^2}\right|_{\hat{\lambda}} = \frac{n}{\hat{\lambda}^2} \implies \boxed{\sigma_\lambda^2 = n}$$

Gaussian approximation

Log-likelihood profile and errors





Large number of events – Gaussian approximation reasonably accurate

$$\sigma_\lambda^2 = n$$

Log-likelihood profile provides a more accurate estimate for small number of events

$$2\ln\mathcal{L}(\lambda) = 2\ln\mathcal{L}(\hat{\lambda}) - 1$$

n = 100; $\hat{\lambda} = 100.0^{+10.33}_{-9.67}$

Log-likelihood profile provides a better error estimate

n = 2; $\hat{\lambda} = 2.0^{+1.77}_{-1.10}$

About Wilks' Theorem

- **Likelihood ratio test** compares goodness of fit of a alternate model hypothesis to a null hypothesis

- Wilks' Theorem: in limit that sample size n approaches ∞**,** the test statistic TS for nested models* is distributed like χ^2 for the degrees of freedom different between the models

$$TS = 2 \ln \frac{\text{Likelihood for alternate hypothesis}}{\text{Likelihood for null hypothesis}}$$

We have a probability!

**Simulation checks highly encouraged for complicated applications*

Confidence regions

In problems with multiple parameters.

- Saw earlier that we can calculate “asymmetric errors” by finding points where $2\ln\mathcal{L}$ decreases by 1.0: 2-sided 1σ confidence interval (68%)
- Actually this comes from LRT (Wilks’ theorem). This is region where null hypothesis that parameter value has some value cannot be rejected at given confidence level.
- But what to do if likelihood depends on more than our parameter of interest?
- It depends...

Log-likelihood profile and errors

As in the single-variable case, because of the symmetry of the Gaussian function between $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}}$, one finds that contours of constant $\ln L$ or χ^2 cover the true values with a certain, fixed probability. That is, the confidence region is determined by

$$\ln L(\boldsymbol{\theta}) \geq \ln L_{\max} - \Delta \ln L \, , \tag{36.58}$$

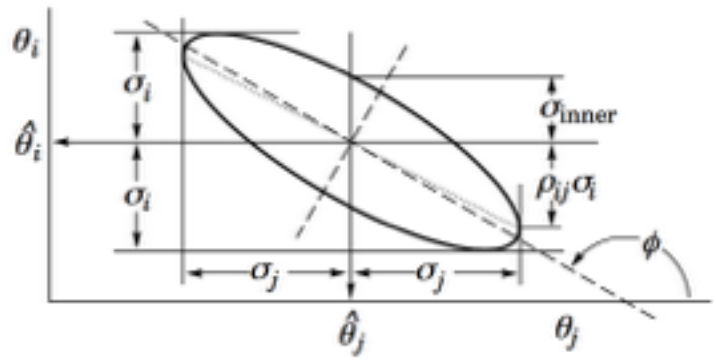


Figure 36.5: Standard error ellipse for the estimators $\widehat{\theta}_i$ and $\widehat{\theta}_j$. In this case the correlation is negative.

Table 36.2: $\Delta\chi^2$ or $2\Delta \ln L$ corresponding to a coverage probability $1-\alpha$ in the large data sample limit, for joint estimation of m parameters.

$(1-\alpha)$ (%)	$m=1$	$m=2$	$m=3$
68.27	1.00	2.30	3.53
90.	2.71	4.61	6.25
95.	3.84	5.99	7.82
95.45	4.00	6.18	8.03
99.	6.63	9.21	11.34
99.73	9.00	11.83	14.16

PDG
http://pdg.lbl.gov/2018/reviews/rpp2018-rev-statistics.pdf

Profile likelihood

Confidence regions with nuisance parameters

Rolke, et al., NIM A, 551, 493 (2005)

- Often we are either concerned only with the one parameter, or wish to treat the multiple parameters separately (ignore covariance).
- Produce "profile log-likelihood" curve, a function of only one parameter (at a time), maximized over all others.
- LRT says this should behave as $\chi^2(1)$.
- Define confidence region using this function exactly as before.

Hypothesis testing

- Compare likelihoods of two hypotheses to see which is better supported by the data.
- Likelihood-ratio test (LRT) & Wilks' theorem.
- Given a model with *N+M* parameters:

$$\Theta = \{\theta_1, \ldots, \theta_N, \theta_{N+1}, \ldots, \theta_{N+M}\}$$

where *N* have true values: $\theta_1^T, \ldots, \theta_N^T$

- Values of likelihood under two hypotheses:

$$\mathcal{L}_1 = \mathcal{L}(\hat{\theta}_1, \ldots, \hat{\theta}_N, \hat{\theta}_{N+1}, \ldots, \hat{\theta}_{N+M})$$

$$\mathcal{L}_0 = \mathcal{L}(\theta_1^T, \ldots, \theta_N^T, \hat{\theta}_{N+1}, \ldots, \hat{\theta}_{N+M})$$

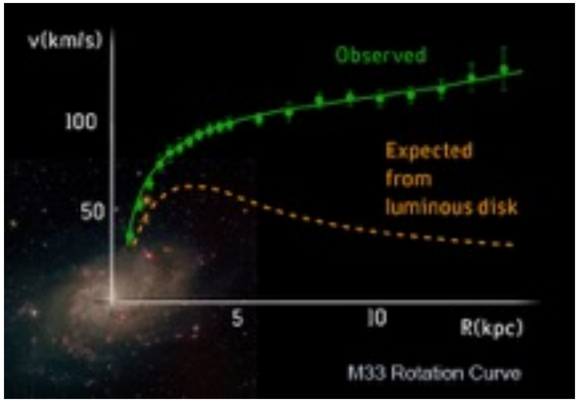
- "Ratio" distributed as: $\boxed{2(\ln \mathcal{L}_1 - \ln \mathcal{L}_0) \sim \chi^2(N)}$

Terms and conditions apply

Overwhelming astrophysical evidences of the existence of dark matter

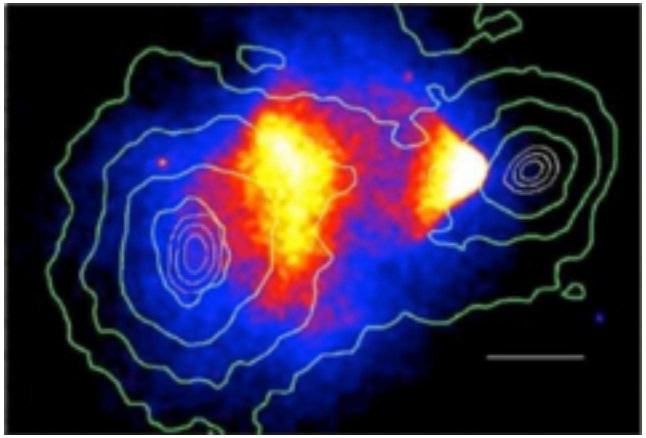
*Comprises **majority** of **mass** in Galaxies*
- Galaxy cluster dynamics

Zwicky (1937)



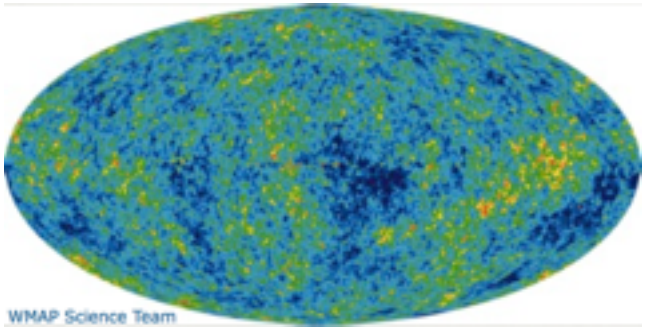
*Large **halos** around Galaxies*
- Galaxy rotation dynamics

Rubin+(1980)



*Almost **collisionless***
- “Bullet” cluster

Clowe+(2006)



“Cold” and not baryons (p, n)
- Deuterium abundance

Schramm and others (1980s)
- Cosmic background structure

WMAP(2010), *Planck*(2015)

Dark matter properties

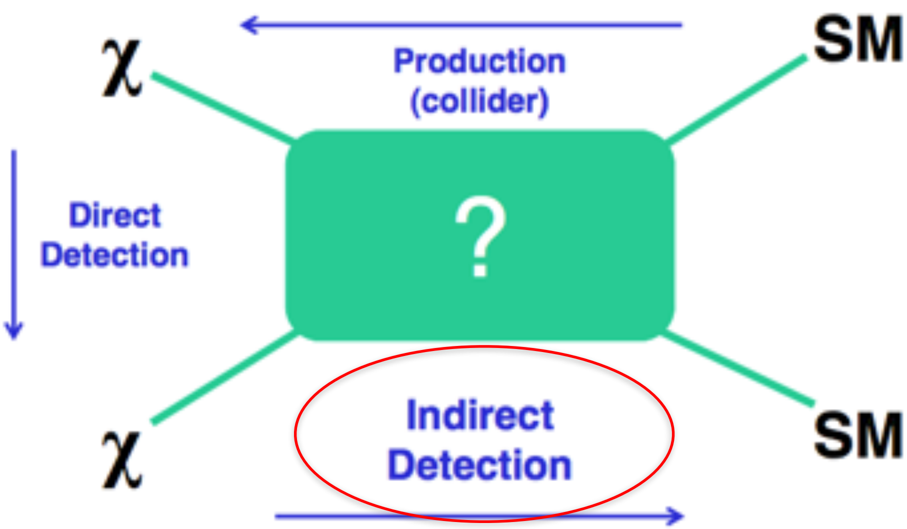
Requirements for a good dark matter candidate χ:

- Must have lifetime $\tau_\chi \gg \tau_U$.
- Must be electrically neutral.
- Must interact very weakly with ordinary matter.
- Must have correct relic density: $\Omega_\chi \simeq 0.22$.

$$\Downarrow$$

Weakly Interacting Massive Particles (WIMPs)

Indirect detection of dark matter

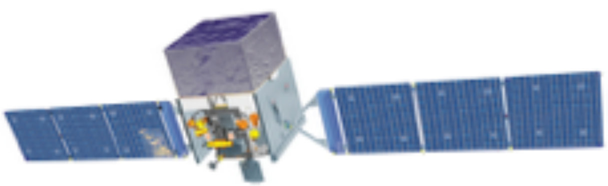


p-bar

e^+

gamma rays

anti-D



Fermi-LAT

AMS-02

CALET

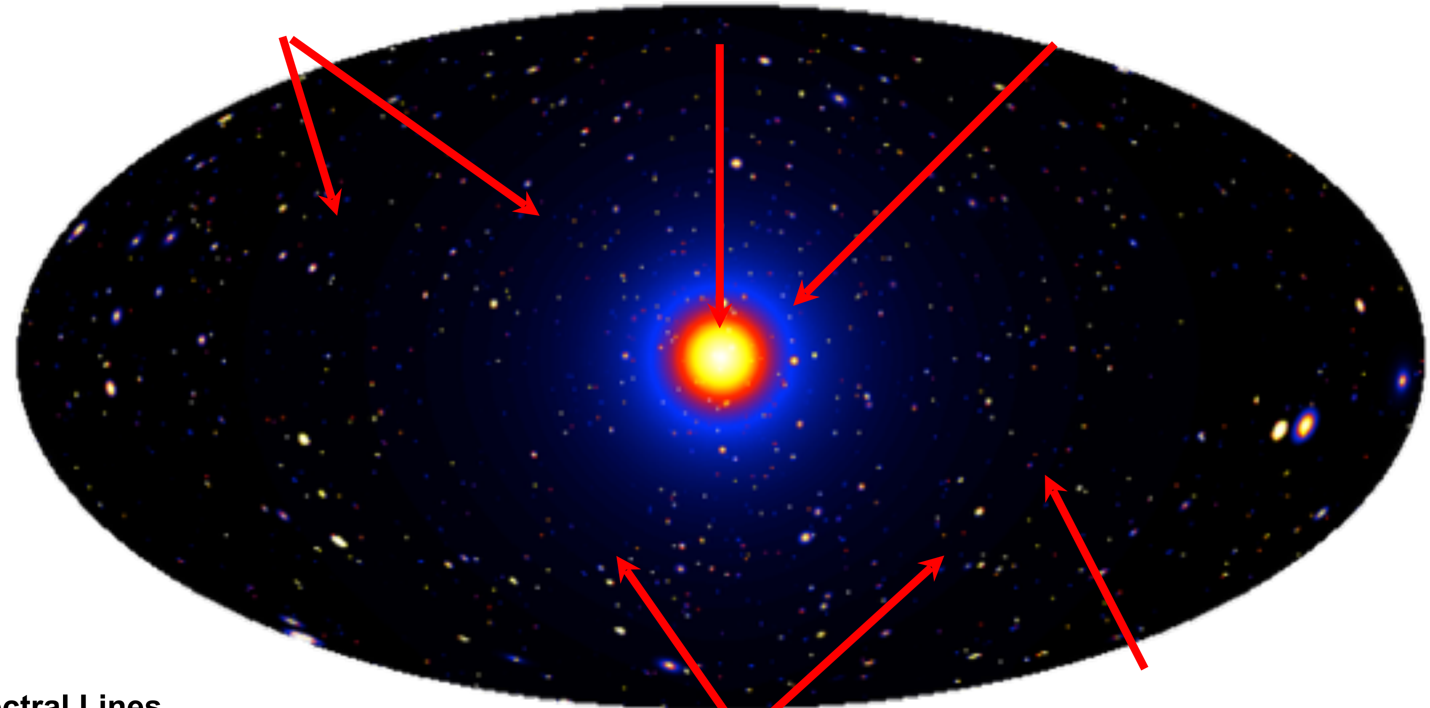
DAMPE

Gamma-ray sky from dark matter

Satellite galaxies
Low background and good source id, but low statistics

Galactic Center
Good statistics, but source confusion/diffuse background

Milky Way Halo
Large statistics, but diffuse background



Spectral Lines
Little or no astrophysical uncertainties, good source id, but low sensitivity because of expected small branching ratio

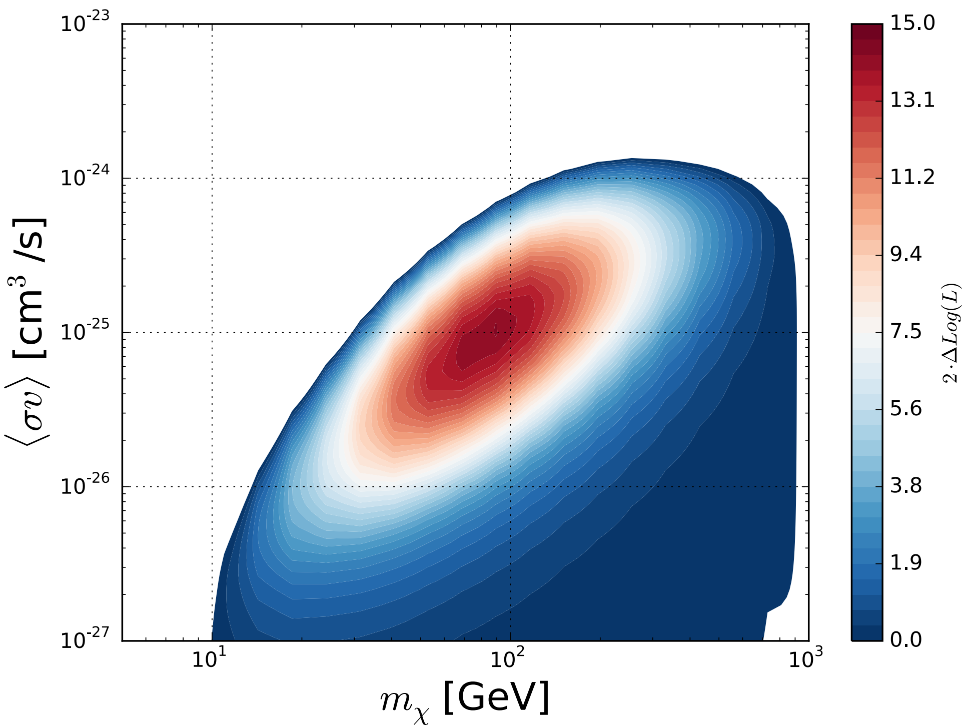
Galaxy Clusters
Low background, but low statistics

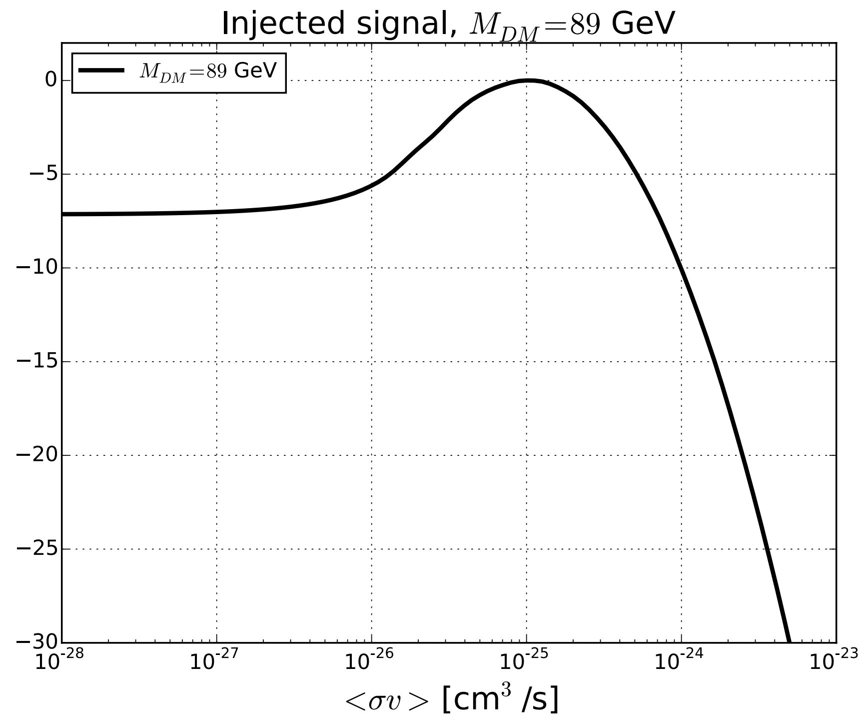
Isotropic contributions
Large statistics, but astrophysics, Galactic diffuse background

Simulation: Pieri+ 2011PhRvD..83b3518P

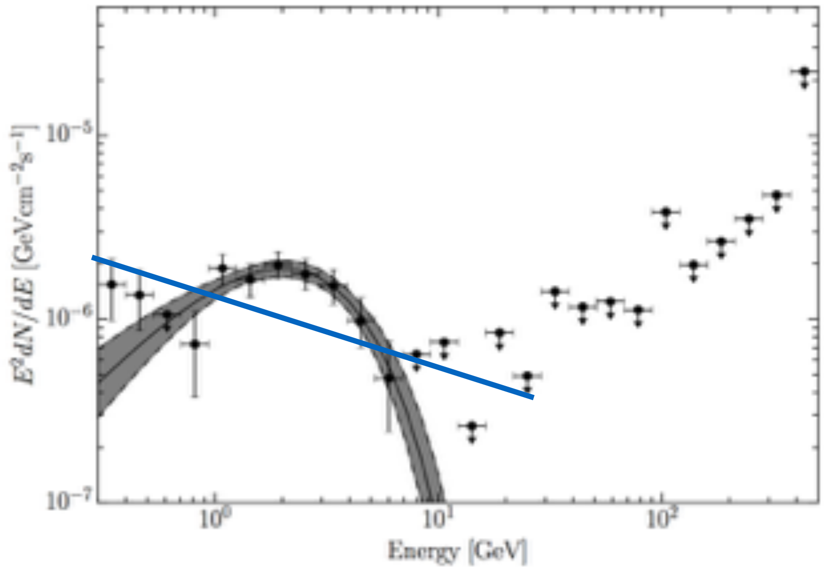
Example of profile likelihood

- The search for DM from M31 is performed by fixing the annihilation DM channel leaving free to vary the mass and cross section.
- Below I show the results for one simulation.
- With the profile likelihood you can find the best fit and error for the cross section.





Hypothesis testing: SED curvature



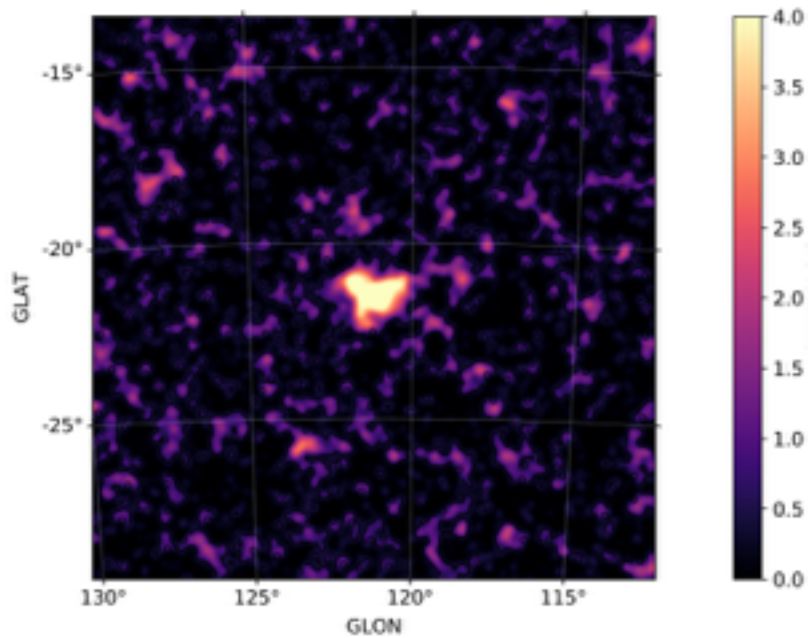
PL ($\mathcal{L}_{\mathrm{PL}}$)

PLE ($\mathcal{L}_{\mathrm{PLE}}$)

$$TS_{\mathrm{curv}}^{\mathrm{PLE}} = 2 \cdot (\log \mathcal{L}_{\mathrm{PLE}} - \log \mathcal{L}_{\mathrm{PL}})$$

$$TS_{\mathrm{curv}}^{\mathrm{PLE}} > 9$$

Hypothesis testing: TS of a source

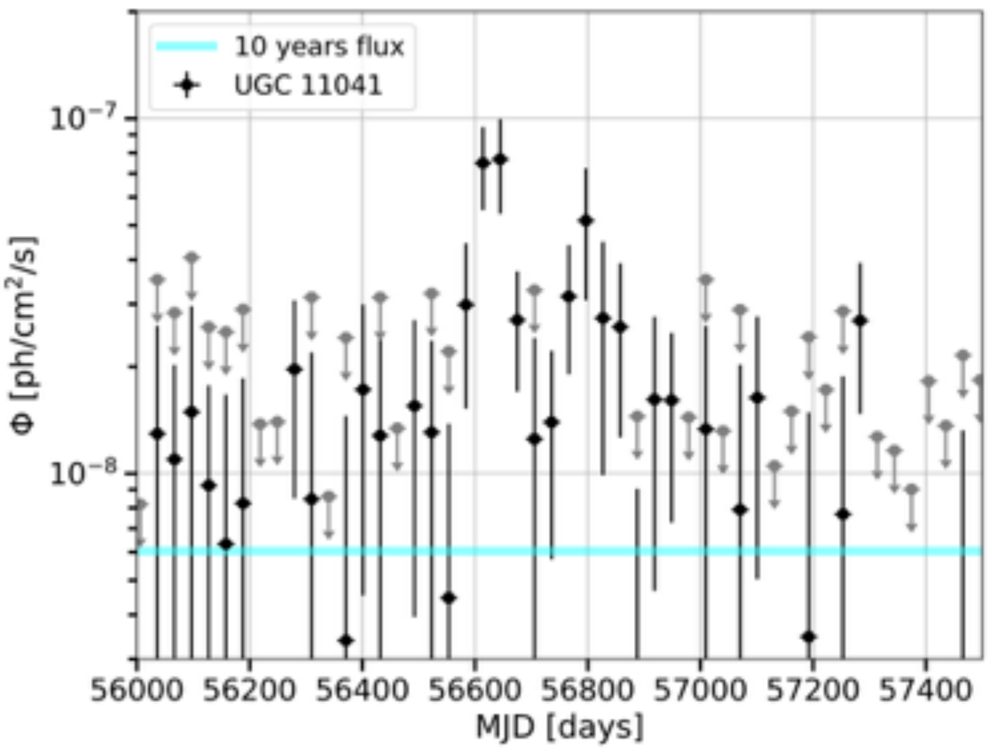


$\text{Log } L_0 = \text{Log } L_{IEM} + \text{Log } L_{ISO} + \text{Log } L_{SOURCES}$

$\text{Log } L_1 = \text{Log } L_{IEM} + \text{Log } L_{ISO} + \text{Log } L_{SOURCES} + \text{Log } L_{TESTSOURCES}$

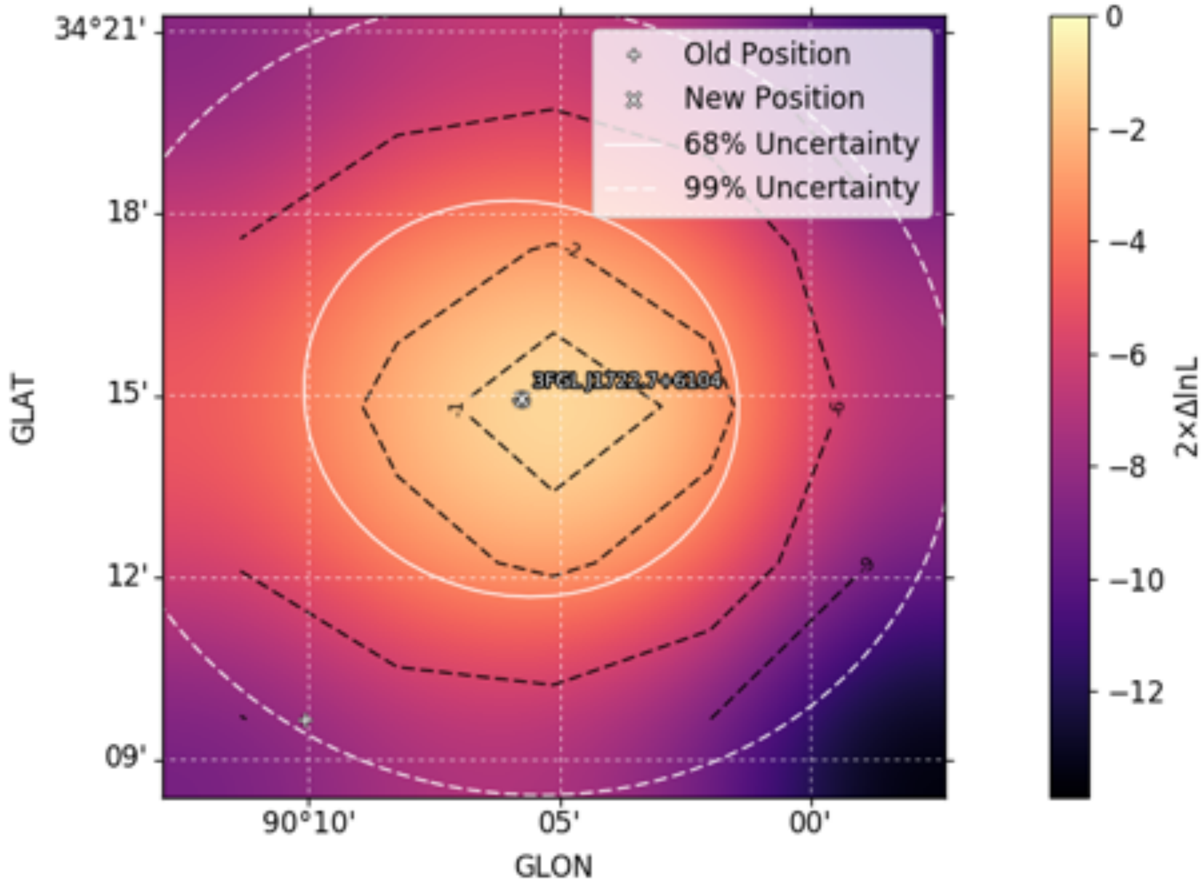
$$TS = 2 \cdot (\log \mathcal{L}_1 - \log \mathcal{L}_0)$$

Hypothesis testing: Variability of a source



$$TS_{var} = 2\left[\log \mathcal{L}(\{F_i\}) - \log \mathcal{L}(F_{Const})\right] = 2\sum_i \left[\log \mathcal{L}_i(F_i) - \log \mathcal{L}_i(F_{Const})\right]$$

Hypothesis testing: localization



Summary

MLE provides

- Framework for parameter estimation of a given model
- Covariant errors through inverse of Fisher matrix
- Asymmetric errors through profile likelihood
- Hypothesis testing of of models through Wilks' theorem

Example: χ^2 fit of constant

Data

- independent measurements of flux with errors (x_i, σ_i)

Model

- all measurements are of a constant flux with Gaussian errors F

Probabilities

$$P(x_i|F) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x_i - F)^2}{2\sigma_i^2}}$$

Likelihood Function

$$\ln \mathcal{L}(F) = -\sum \frac{(x_i - F)^2}{2\sigma_i^2} - \sum \ln \sigma_i - \frac{N}{2} \ln 2\pi$$

Example: χ^2 fit of constant

- Log likelihood:

Constant with respect to *F*

$$\ln \mathcal{L}(F) = -\sum \frac{(x_i - F)^2}{2\sigma_i^2} \boxed{\cancel{- \sum \ln \sigma_i - \frac{N}{2} \ln 2\pi}}$$

- Maximize for MLE of F:

$$\frac{\partial \ln \mathcal{L}}{\partial F} = \sum \frac{x_i - F}{\sigma_i^2} = 0 \implies \boxed{\hat{F} = \frac{\sum x_i/\sigma_i^2}{\sum 1/\sigma_i^2}}$$

- Curvature gives "error" on *F*:

$$\frac{1}{\sigma_F^2} = -\left.\frac{\partial^2 \ln \mathcal{L}}{\partial F^2}\right|_{\hat{F}} = \sum \frac{1}{\sigma_i^2} \implies \boxed{\sigma_F = \frac{1}{\sqrt{\sum 1/\sigma_i^2}}}$$