Maximum Likelihood:

Statistics in photon counting experiments

Stephen Fegan
LLR/Ecole Polytechnique, France

sfegan@llr.in2p3.fr

Questions in γ-ray astronomy

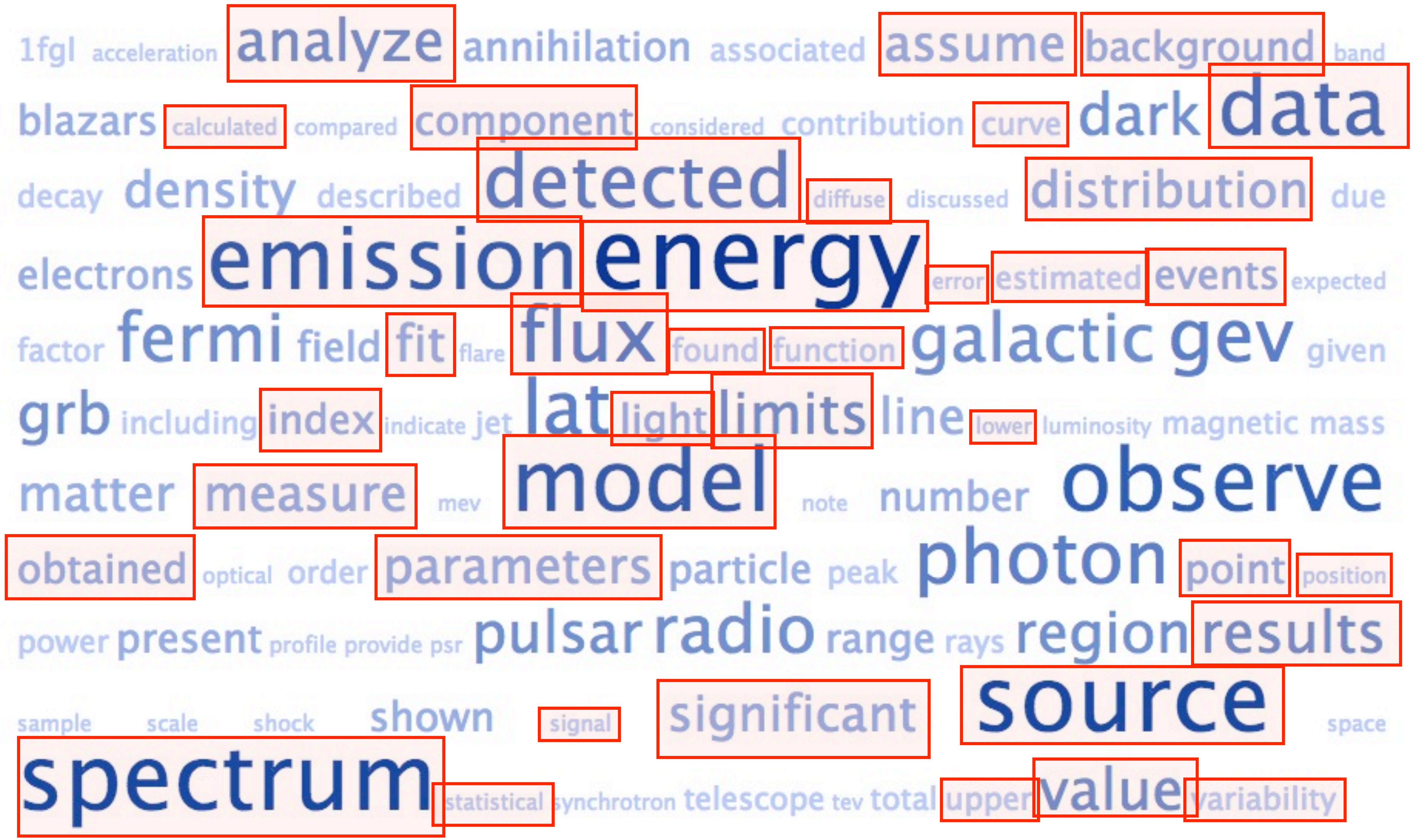
- Is a source significantly detected?
- If so, what is its flux?
- If not, what is upper limit on the flux?
- What kind of spectrum does it have?
- What is its spectral index?
- What is its location in the sky?
- What are the errors on these values?
- Is the source variable?

Questions in DM astrophysics

- Does Fermi detect γ-ray line emission from DM particle annihilation?
- With what significance?
- What is the energy of the line?
- What is the measurement error?
- What is the spatial distribution?
- What kind of systematic errors may be present?

LAT paper tag cloud

100 most frequently used words from 875 papers mentioning LAT and γ-ray on arXiv



Questions in γ-ray astronomy

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

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 (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 parameters: $\Theta = \{\theta_j\} = \{\theta_1, \theta_2, ..., \theta_M\}$
- Likelihood: $\mathcal{L}(\Theta|X) = P(X|\Theta)$
- Conditional probability rule for independent events: $P(A,B) \underset{\text{CPR}}{=} P(A)P(B|A) \underset{\text{Independence}}{=} P(A)P(B)$
- 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 $\{\hat{\theta}_k\}$ from solving simultaneous equations:

$$\boxed{\left.\frac{\partial \ln \mathcal{L}}{\partial \theta_j}\right|_{\{\hat{\theta}_k\}} = 0}$$

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

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

Gaussian approximation

so 2^{nd} derivative is related to "errors"

Why maximum likelihood...

...rather than some ad-hoc estimation method?

- ML framework provides a “cookbook” through which problems can be solved.
  In other methods ad-hoc choices may have to be made.
- ML provides unbiased, minimum variance estimate as sample size increases.
  Same may not be case for ad-hoc methods.
- Asymptotically Gaussian: evaluation of confidence bounds & hypothesis testing.
- Well studied in the literature.
- Starting point for Bayesian analysis.

MLE example 1:

χ^2 fit of constant - I

- Data: independent measurements of flux of some source with errors - (x_i, σ_i)
- Model: all measurements are of a constant flux F with Gaussian errors.
- Probabilities: $P(x_i|F) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x_i-F)^2}{2\sigma_i^2}}$
- Log likelihood:

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

MLE example 1:

χ^2 fit of constant - II

- Log likelihood:

Constant with respect to *F*

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

- 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*:

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

MLE example 2:

Event counting experiment

- Experiment detects *n* events (e.g. γ rays)
- Model: Poisson process with mean of of *λ*:

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

- Log likelihood: $\ln \mathcal{L}(\lambda) = \underbrace{n \ln \lambda}_{\text{Data cpt}} - \underbrace{\lambda}_{\text{Npred}} \overbrace{\cancel{- \ln n!}}^{\text{Constant WRT } \lambda}$
- ML estimate and error in Gaussian regime:

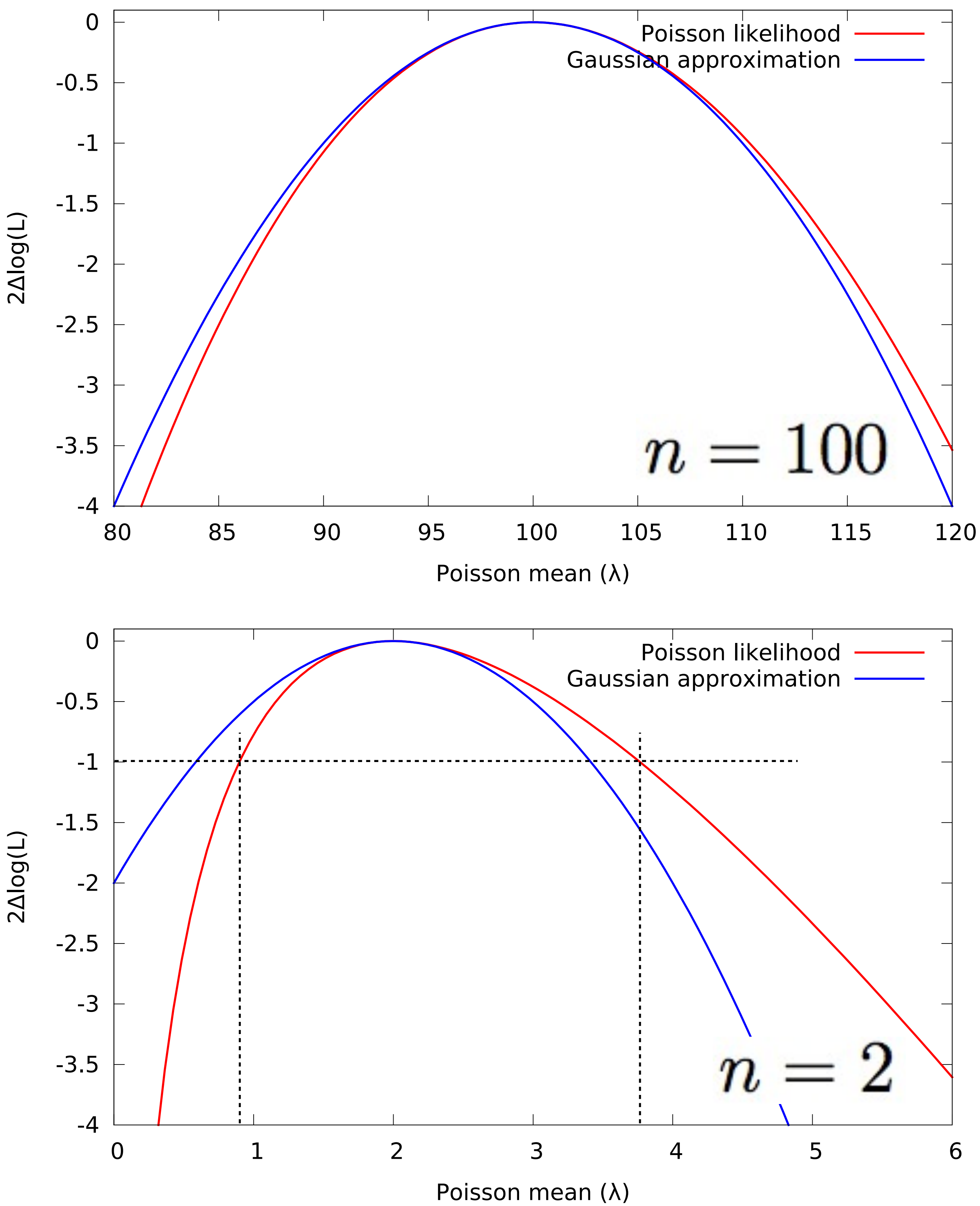
$$\frac{\partial \ln \mathcal{L}}{\partial \lambda} = \frac{n}{\lambda} - 1 \implies \boxed{\hat{\lambda} = n}$$

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

Gaussian approximation

MLE example 2:

Log-likelihood profiles



- Gaussian approximation is reasonable when *n* is "large enough". In this case $\sigma_\lambda^2 = n$ is a good estimate of the "error".

- If not, estimate errors by finding points where $2\ln\mathcal{L}(\lambda)$ decreases by 1.0 from maximum, i.e.,

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

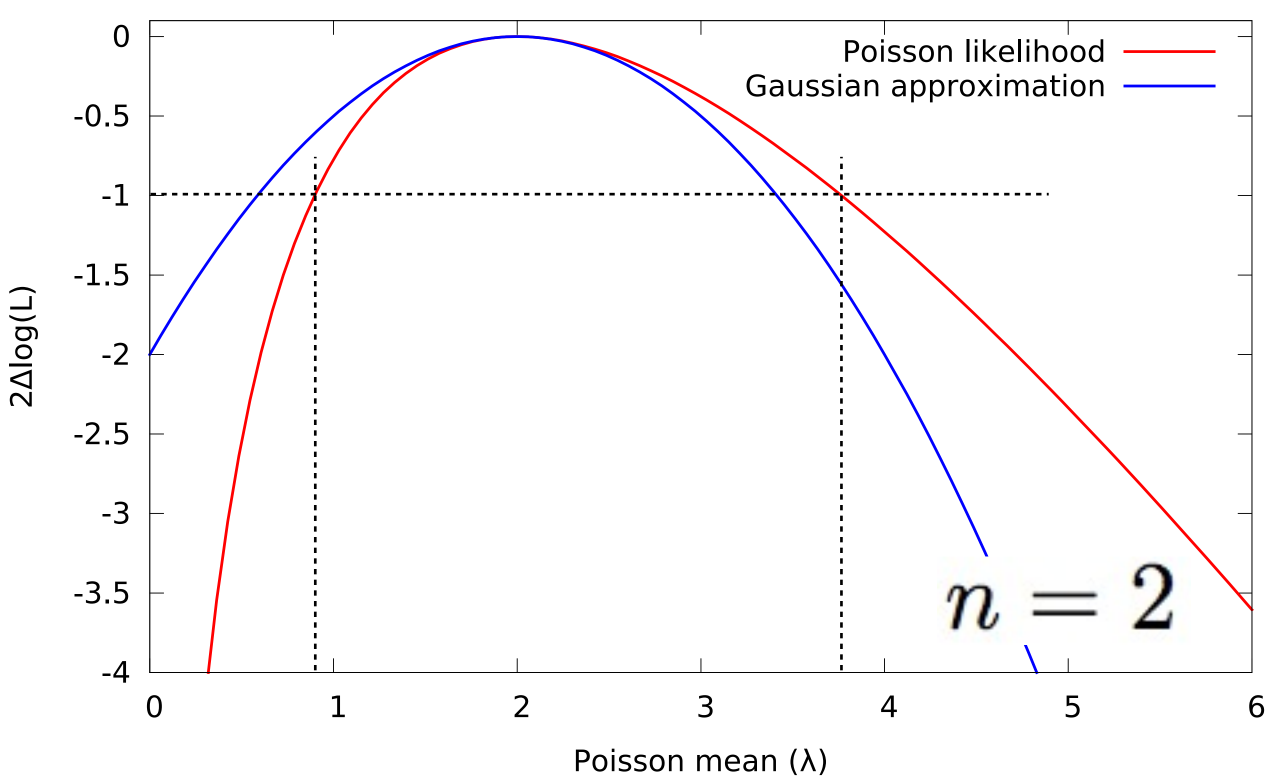
- *n=100* : $\hat{\lambda} = 100.0^{+10.33}_{-9.67}$

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

MLE example 2:

Log-likelihood profiles

```
# errors_poisson.py - 2013-05-07 SJF
# Evaluate the errors on the Poisson mean
import math, scipy.optimize
n_meas   = 2
logL     = lambda lam: n_meas*math.log(lam)-lam
opt_fn   = lambda lam: -logL(lam)
opt_res  = scipy.optimize.minimize(opt_fn, 1e-8)
lam_est  = opt_res.x[0]
logL_max = logL(lam_est)
root_fn  = lambda lam: 2.0*(logL(lam)-logL_max)+1.0
lam_lo   = scipy.optimize.brentq(root_fn, 1e-8, lam_est)
lam_hi   = scipy.optimize.brentq(root_fn, lam_est, 1e8)
print lam_est, lam_lo-lam_est, lam_hi-lam_est
```



$2 \ln \mathcal{L}(\lambda)$ decreases by 1.0 from maximum, i.e.,

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

- *n=100* : $\hat{\lambda} = 100.0^{+10.33}_{-9.67}$
- *n=2* : $\hat{\lambda} = 2.0^{+1.77}_{-1.10}$

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

Why is that useful?

(We don't know the true values of any parameters!)

- We make an assumption about the model (*the null hypothesis*), in which the parameters have some presumed "true" values.
- Compute $\mathcal{L}_0$ from these values and $\mathcal{L}_1$ using MLE for all params.
- Hope to show that $2(\ln \mathcal{L}_1 - \ln \mathcal{L}_0)$ is so large that it is improbable from $\chi^2(N)$,
- and, hence, reject the null hypothesis. Usually cannot say hypothesis is true!

http://xkcd.com/892/

MLE example 3:

Source & Background

- Data: events detected in two independent "channels": $X = \{n_1, n_2\}$
- Model: Poisson process with...
  - Unknown "source" and "background":

$$\Theta = \{\theta_1, \theta_2\} = \{S, B\} \qquad \vec{\Theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} S \\ B \end{pmatrix}$$

  - Response matrix (presumed known) $\mathbf{R} = \begin{pmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{pmatrix}$
  - Poisson means: $\vec{\lambda} = \mathbf{R}\vec{\Theta}$

$$\begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = \begin{pmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{pmatrix} \begin{pmatrix} S \\ B \end{pmatrix}$$

MLE example 3:

MLE

- Log likelihood:

$$\ln \mathcal{L}(S,B) = \overbrace{n_1 \ln(r_{11}S + r_{12}B) + n_2 \ln(r_{21}S + r_{22}B)}^{\text{Data component}} \\ \underbrace{- (r_{11}+r_{21})S - (r_{12}+r_{22})B}_{\text{Npred}} + \cancel{const}$$

- MLE: $$\frac{\partial \ln \mathcal{L}}{\partial S} = \frac{\partial \ln \mathcal{L}}{\partial B} = 0 \implies \hat{\vec{\Theta}} = \mathbf{R}^{-1}\vec{n}$$

$$\begin{pmatrix} \hat{S} \\ \hat{B} \end{pmatrix} = \frac{1}{r_{11}r_{22} - r_{12}r_{21}} \begin{pmatrix} r_{22} & -r_{12} \\ -r_{21} & r_{11} \end{pmatrix} \begin{pmatrix} n_1 \\ n_2 \end{pmatrix}$$

$$\ln \mathcal{L}_1 = \ln \mathcal{L}(\hat{S}, \hat{B}) = n_1 \ln n_1 + n_2 \ln n_2 - (n_1 + n_2)$$

- If likelihood: $\mathcal{L}(\vec{\Theta}) \sim e^{-\frac{1}{2}(\vec{\Theta}-\hat{\vec{\Theta}})^T \mathbf{\Sigma}^{-1} (\vec{\Theta}-\hat{\vec{\Theta}})}$

Gaussian approximation

"errors" are: $$\left.\frac{\partial^2 \ln \mathcal{L}}{\partial \theta_i \partial \theta_j}\right|_{\hat{\vec{\Theta}}} = -(\mathbf{\Sigma}^{-1})_{ij} = -\mathcal{I}_{ij}$$

Covariance matrix

Fisher information matrix

MLE example 3:

Covariances and errors

- Calculate Fisher information matrix and invert:

$$\boldsymbol{\mathcal{I}}_{ij} = - \left. \frac{\partial^2 \ln \mathcal{L}}{\partial\theta_i \partial\theta_j} \right|_{\hat{\vec{\Theta}}} \rightarrow \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_S^2 & \mathrm{cov}(S,B) \\ \mathrm{cov}(S,B) & \sigma_B^2 \end{pmatrix} = \boldsymbol{\mathcal{I}}^{-1}$$

- For our example we get:

$$\boldsymbol{\mathcal{I}} = \frac{1}{n_1 n_2} \begin{pmatrix} r_{21}^2 n_1 + r_{11}^2 n_2 & r_{21} r_{22} n_1 + r_{11} r_{12} n_2 \\ r_{21} r_{22} n_1 + r_{11} r_{12} n_2 & r_{22}^2 n_1 + r_{12}^2 n_2 \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \frac{1}{\det(\mathbf{R})^2} \begin{pmatrix} r_{22}^2 n_1 + r_{12}^2 n_2 & -r_{21} r_{22} n_1 - r_{11} r_{12} n_2 \\ -r_{21} r_{22} n_1 - r_{11} r_{12} n_2 & r_{21}^2 n_1 + r_{11}^2 n_2 \end{pmatrix}$$

- In general parameters are correlated, but can choose set that is uncorrelated. Here they are $\{\lambda_1, \lambda_2\}$ giving $\hat{\lambda}_1 = n_1, \hat{\lambda}_2 = n_2, \boldsymbol{\Sigma}_\lambda = \mathrm{diag}(n_1, n_2)$

MLE example 3:

Source significance

- Null hypothesis: suppose $S = 0$, then:

$$\begin{aligned}\ln \mathcal{L}_0(B) &= \ln \mathcal{L}(S=0, B) \\ &= n_1 \ln r_{12} B + n_2 \ln r_{22} B - (r_{12} + r_{22}) B\end{aligned}$$

- MLE for *B* gives: $\dfrac{\partial \ln \mathcal{L}_0}{\partial B} = 0 \implies \hat{B}_0 = \dfrac{n_1 + n_2}{r_{12} + r_{22}}$

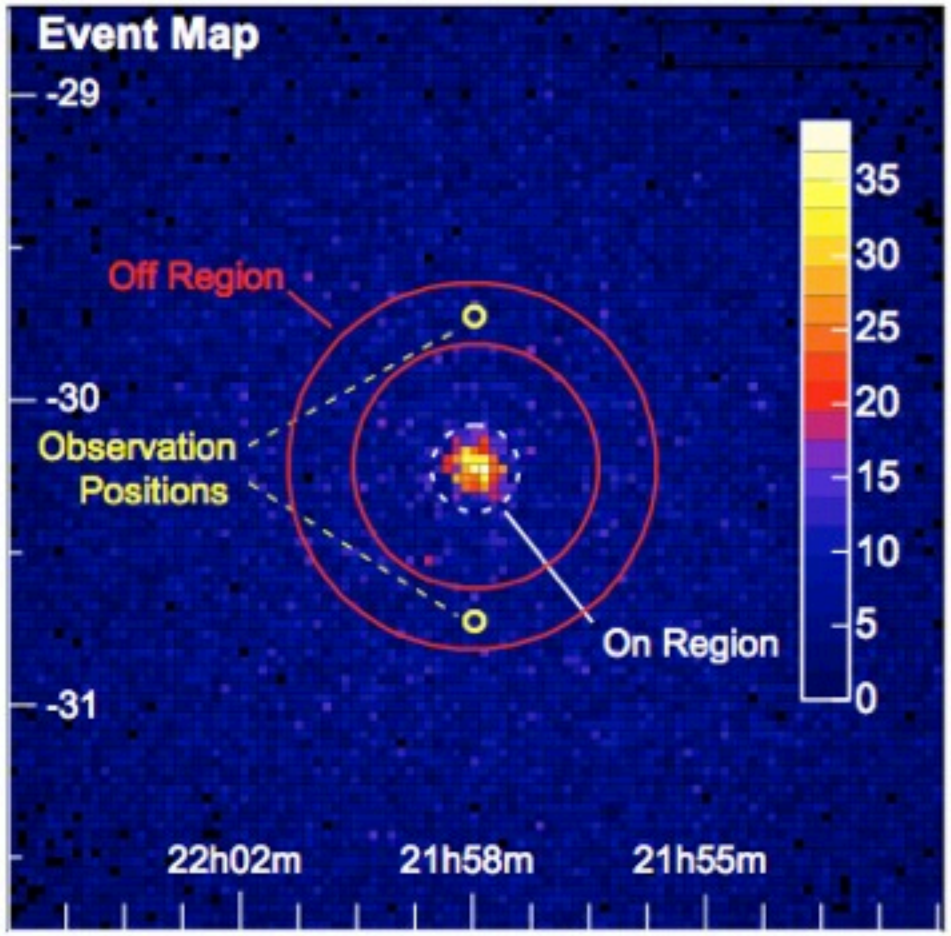
$$\begin{aligned}\ln \mathcal{L}_0 &= \ln \mathcal{L}_0(\hat{B}_0) \\ &= n_1 \ln \frac{r_{12}(n_1 + n_2)}{r_{12} + r_{22}} + n_2 \ln \frac{r_{22}(n_1 + n_2)}{r_{12} + r_{22}} - (n_1 + n_2)\end{aligned}$$

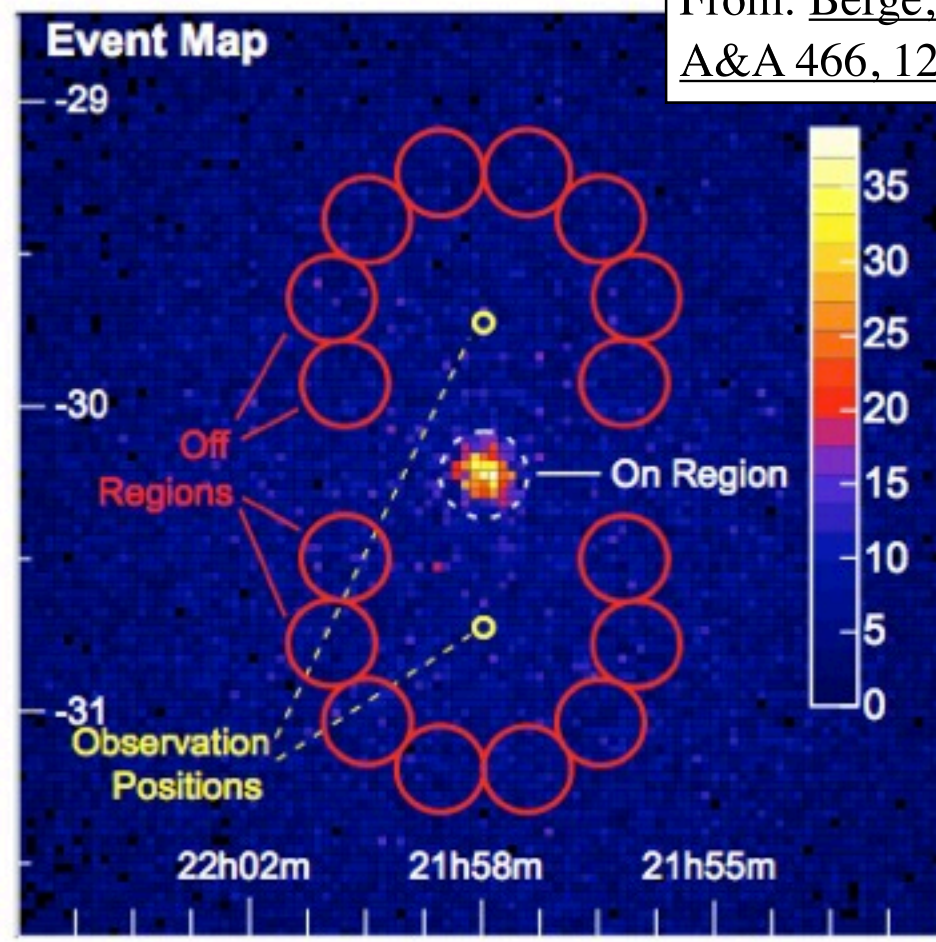
- Test statistic: $TS = 2(\ln \mathcal{L}_1 - \ln \mathcal{L}_0) \sim \chi^2(1)$

$$TS = 2\left[n_1 \ln \frac{(r_{12} + r_{22}) n_1}{r_{12}(n_1 + n_2)} + n_2 \ln \frac{(r_{12} + r_{22}) n_2}{r_{22}(n_1 + n_2)}\right]$$

On/Off problems

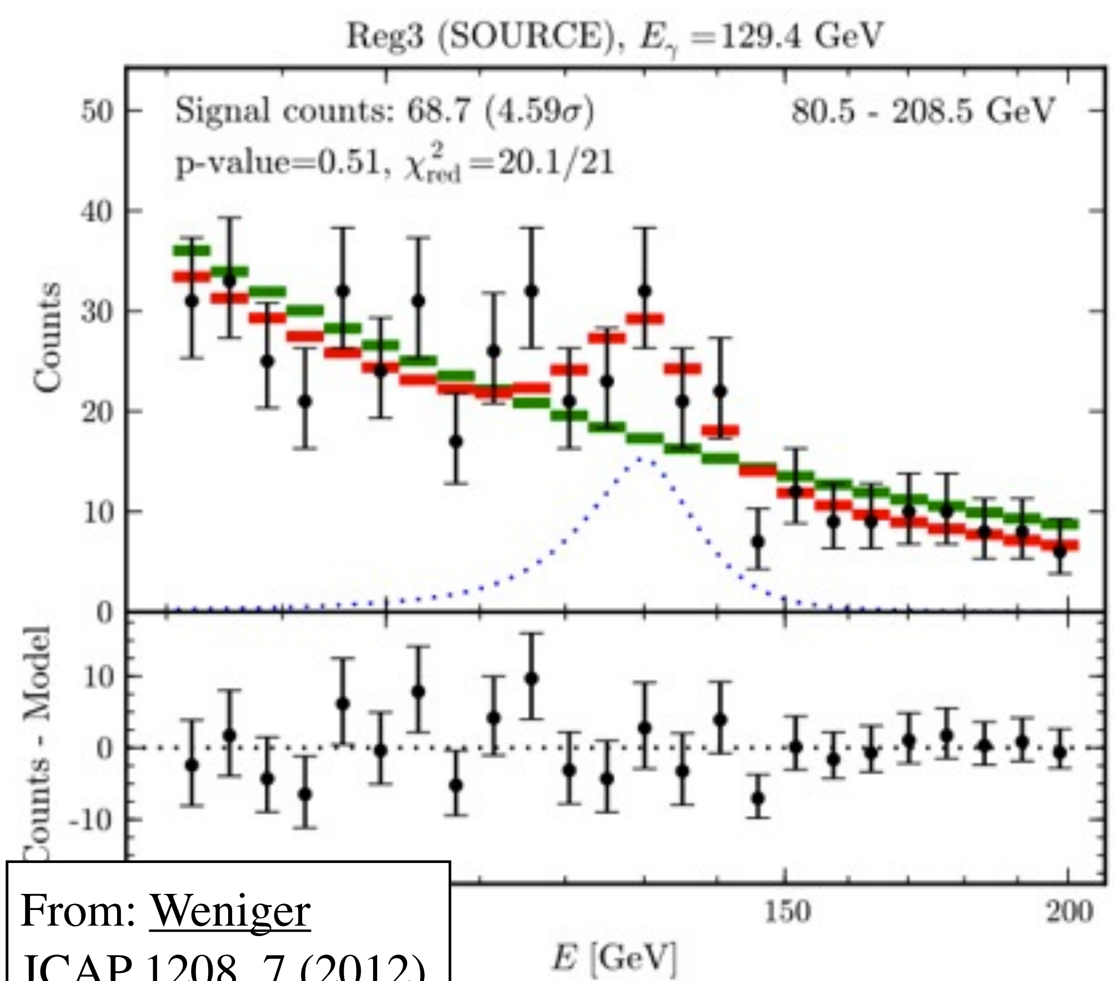
From: Berge, Funk, Hinton
A&A 466, 1219–1229 (2007)



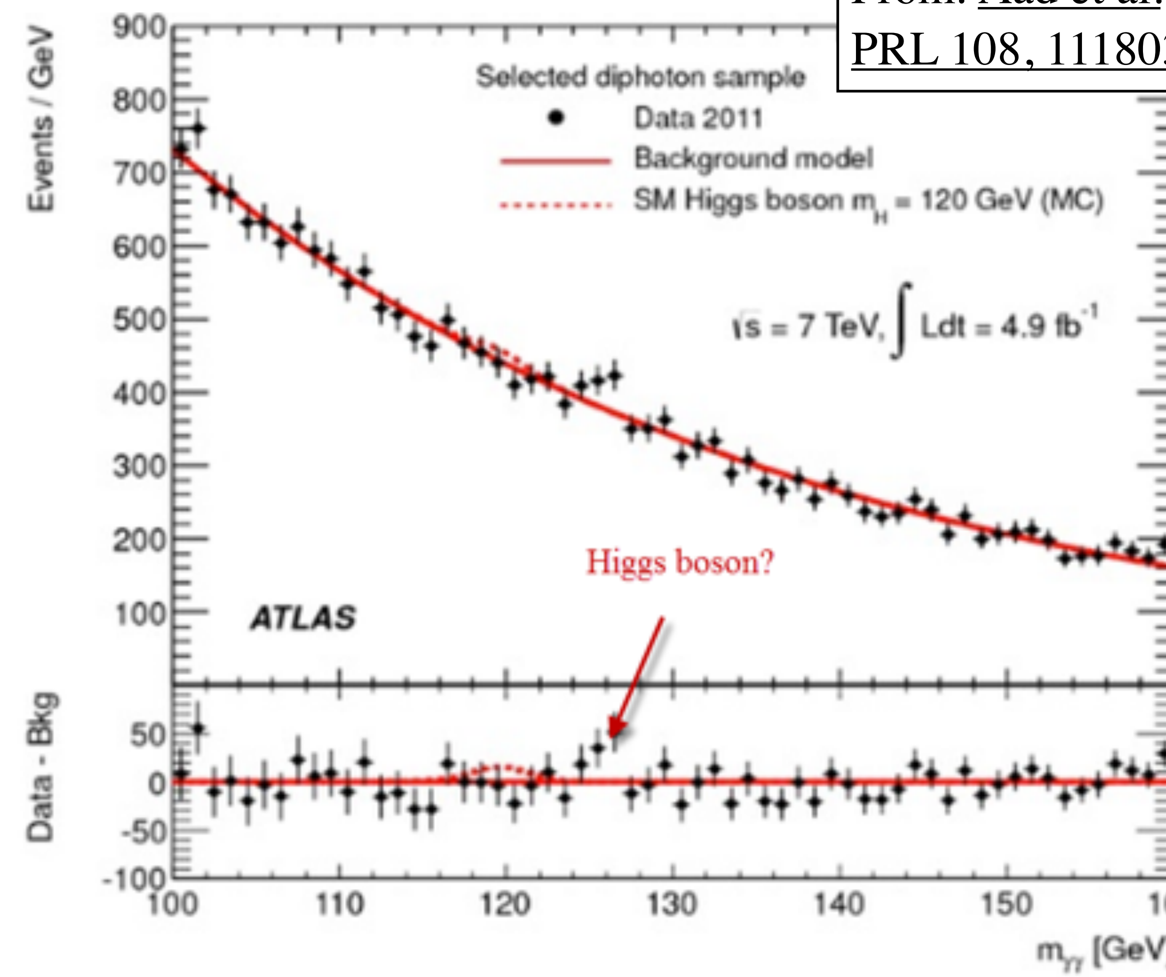


- VHE astronomy - gamma-ray sources and a background of cosmic rays.
- Problem - to evaluate flux of source and its statistical significance. Define on-source (source+background) and off-source (background) channels.

On/Off problems



From: Weniger
JCAP 1208, 7 (2012)



From: Aad et al.
PRL 108, 111803 (2012)

- Line searches - DM with Fermi, or Higgs with ATLAS.
- Problem - detect line signal on top of spectrum of background events. Define “on-source” and “off-source” regions. Must assume that spectrum of background is known or calculable.

On/Off problems

- General set of problems where:

$$n_2 \rightarrow n_{off}$$
$$n_1 \rightarrow n_{on}$$
$$\lambda_2 \rightarrow \lambda_{off} = BT$$
$$\lambda_1 \rightarrow \lambda_{on} = (S + \alpha B)T$$

- and where these are assumed to be known:
  - α - on to off-source background ratio
  - T - observation time (or other detector factors)

MLE for On/Off problems

- Then: $\mathbf{R} = T \begin{pmatrix} 1 & \alpha \\ 0 & 1 \end{pmatrix} \quad \mathbf{R^{-1}} = \frac{1}{T} \begin{pmatrix} 1 & -\alpha \\ 0 & 1 \end{pmatrix}$

$$\begin{aligned} \ln \mathcal{L}(S,B) =& n_{on} \ln[(S+\alpha B)T] + n_{off} \ln BT \\ & - (S + (1+\alpha)B)T \end{aligned}$$

- MLE & (co)variances of S and B are:

$$\hat{B} = \frac{1}{T} n_{off} \qquad \sigma_B^2 = \frac{1}{T^2} n_{off}$$

$$\hat{S} = \frac{1}{T}(n_{on} - \alpha n_{off}) \qquad \sigma_S^2 = \frac{1}{T^2}(n_{on} + \alpha^2 n_{off})$$

This is what you would expect!

$$\mathrm{cov}(\hat{S}, \hat{B}) = -\frac{1}{T^2} \alpha n_{off}$$

TS for On/Off problems

- Test statistic for source detection in On/Off problems is:

$$TS = 2\left[n_{on}\ln\frac{(1+\alpha)n_{on}}{\alpha(n_{on}+n_{off})} + n_{off}\ln\frac{(1+\alpha)n_{off}}{(n_{on}+n_{off})}\right]$$

- Significance is: $\sigma = \sqrt{TS}$
- This is the famous “Li & Ma” formula from: ApJ 272, 317 (1983) - 493 citations on ADS
- Probably, you wouldn’t arrive at this formula using ad hoc estimation methods
- P-values: `scipy.stats.chi2.sf(`*TS*`,1)`

Eg: 1ES1218+304 w/VERITAS

Discovery of Variability in the Very High Energy γ-Ray Emission of 1ES 1218+304 with VERITAS

Acciari, et al., ApJ, 709, 163 (2010)

Table 1 summarizes the results of the VERITAS observations of 1ES 1218+304. For the spectral analysis, we report an excess of 1155 events with a statistical significance of 21.8 standard deviations, σ, from the direction of 1ES 1218+304 during the 2008-2009 campaign (2808 signal events, 4959 background events with a normalization of 0.33). Figure 2 shows the corresponding time-averaged differential energy spectrum. The spectrum extends from 200 GeV to 1.8 TeV and is well described ($\chi^2/\mathrm{dof} = 8.2/7$) by a power law,

$$n_{off} = 4959$$
$$n_{on} = 2808$$
$$\alpha = 1/3$$
$$T = 27.2\,\mathrm{hr}$$

Table 1. Summary of observations and analysis of 1ES 1218+304[a] .

	Live Time [hours]	Zenith [°]	Significance [σ]	$\Phi(> 200$ GeV) $[10^{-12}$ cm^{-2} s$^{-1}]$	Units of Crab Nebula flux (E > 200 GeV)
2006-2007[b]	17.4	2–35	10.4	$12.2 \pm 2.6_{stat}$	0.05 ± 0.011
2008-2009	27.2	2–30	21.8	$18.4 \pm 0.9_{stat}$	0.07 ± 0.004

$$\hat{S} = 42.5\,\mathrm{hr}^{-1}$$
$$\sigma_S = 2.1\,\mathrm{hr}^{-1}$$
$$TS = 474.9$$
$$\sigma = 21.8$$
$$\mathrm{P-value} = 2.8 \times 10^{-105}$$

$$\sigma_{POE} = \frac{\hat{S}}{\sigma_S} = 19.9 \approx \frac{18.4}{0.9}$$

Ratio of value to error - used as “significance” before Li&Ma

Eg: 1ES1218+304 w/VERITAS

```
# lima.py - 2013-05-15 SJF
# Example of Li & Ma significance calculation
import math, scipy.stats

def ts_lima(non,noff,alpha):
    opa  = 1.0+alpha
    ntot = non+noff
    return 2.0*(non*math.log(opa*non/alpha/ntot) \
                + noff*math.log(opa*noff/ntot))

non    = 2808
noff   = 4959
alpha  = 1.0/3
T      = 27.2

S_hat  = (non - noff*alpha)/T
sig2_S = (non + noff*alpha**2)/T**2
ts     = ts_lima(non,noff,alpha)
signif = math.sqrt(ts)
Pval   = scipy.stats.chi2.sf(ts,1)

print S, math.sqrt(sig2_S), ts, signif, Pval
```

Ratio of value to error - used as "significance" before Li&Ma

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...

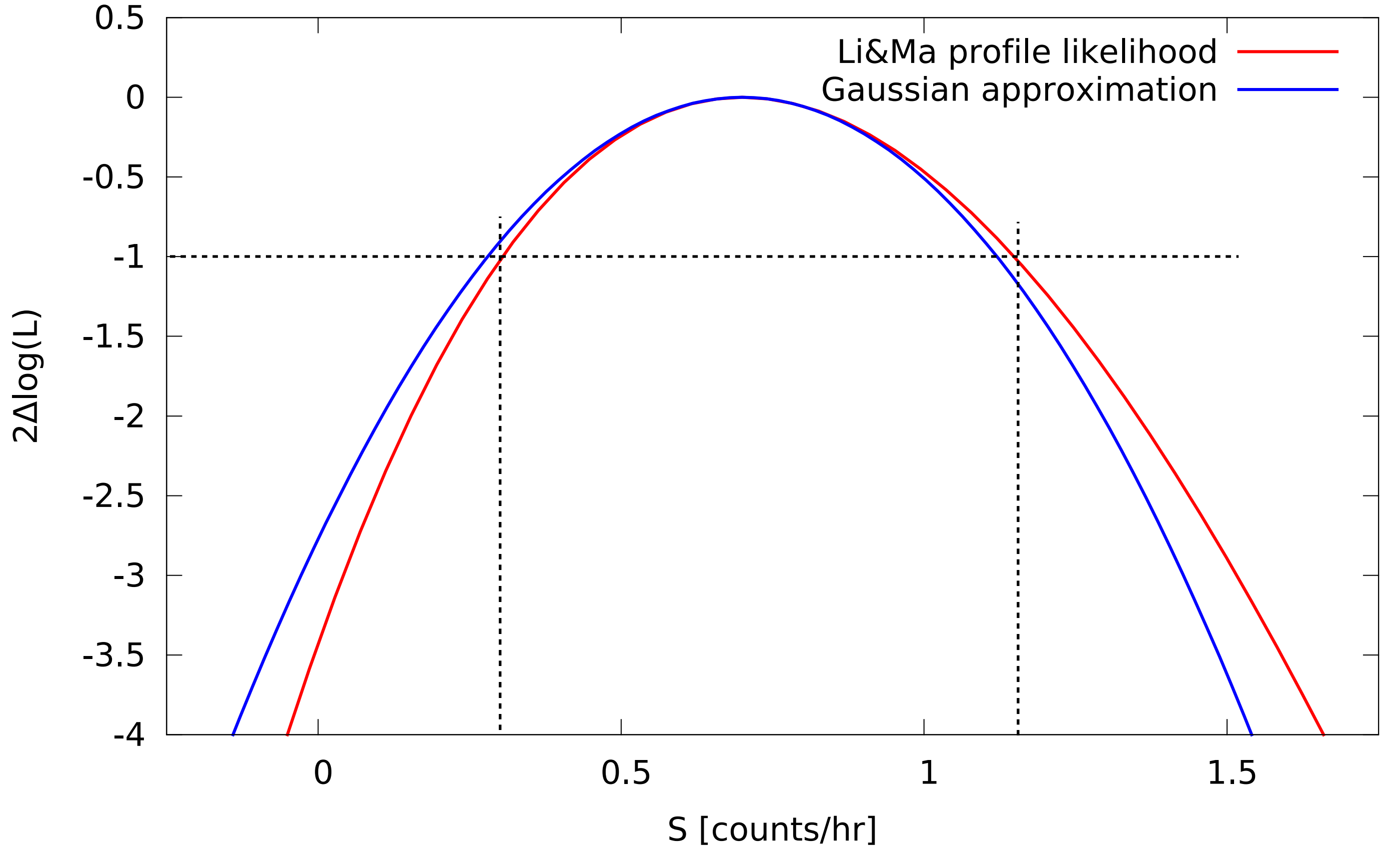
Profile likelihood

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.

Example of profile likelihood



$$\hat{S} = 0.7^{+0.45}_{-0.39}\,\mathrm{hr}^{-1}$$

- Our 1ES1218 example isn't very enlightening here, so take:

$$n_{off} = 24$$
$$n_{on} = 15$$
$$\alpha = 1/3$$
$$T = 10.0\,\mathrm{hr}$$

- Giving:

$$\hat{S} = 0.7\,\mathrm{hr}^{-1}$$
$$\sigma_S = 0.42\,\mathrm{hr}^{-1}$$
$$TS = 3.43$$
$$\sigma = 1.85$$

This is not a significant result, so we would usually not claim a detection. Provide an upper limit instead.

Example of profile likelihood

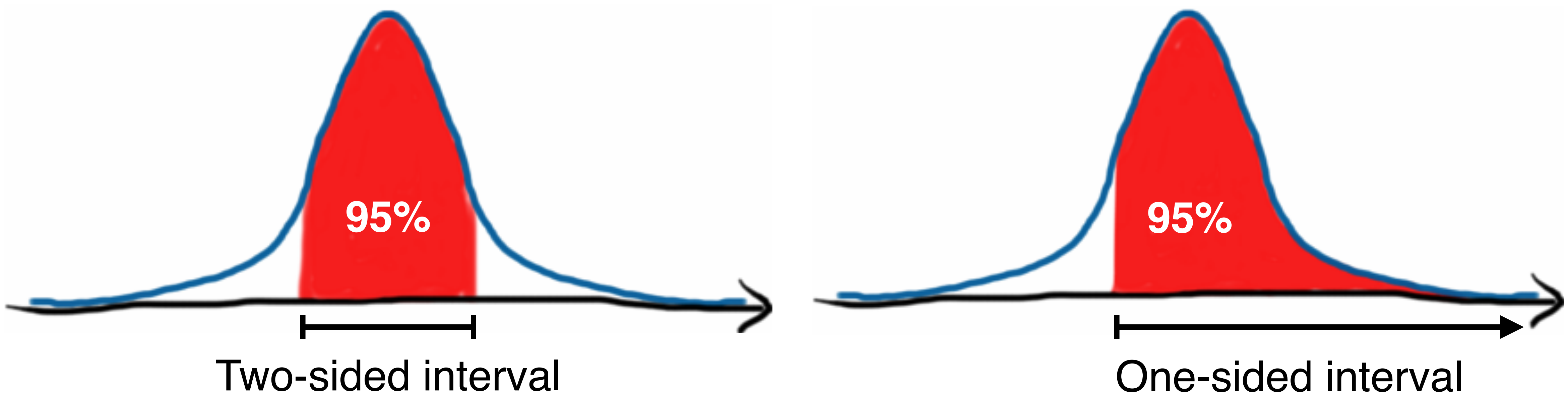
```
# conf_lima_1d.py - 2013-05-25 SJF
# 1-D 2-sided confidence interval in Li & Ma problem
from math import *
import scipy.stats, scipy.optimize, sys
# non, noff, alpha, T = (2808, 4959, 1.0/3, 27.2)
non, noff, alpha, T = (15, 24, 1.0/3, 10.0)
C = 0.68; # Use 1-sigma confidence region
d2logL = scipy.stats.chi2.ppf(C,1)
def logL(S,B):
    return non*log(max((S+alpha*B)*T,sys.float_info.min)) + \
    noff*log(max(B*T,sys.float_info.min))-(S+(1+alpha)*B)*T
def profileLogL(S):
    opt_fn = lambda B: -logL(S,B)
    opt_res = scipy.optimize.minimize(opt_fn, 1)
    return -opt_res.fun
S_hat    = (non-noff*alpha)/T
B_hat    = noff/T
logL_max = logL(S_hat,B_hat)
sig_S    = sqrt(non+noff*alpha**2)/T
TS       = -2.0*(profileLogL(0)-logL_max)
root_fn  = lambda S: 2.0*(profileLogL(S)-logL_max)+d2logL
S_lo     = scipy.optimize.brentq(root_fn, 1e-8, S_hat)
S_hi     = scipy.optimize.brentq(root_fn, S_hat, 1e8)
print S_hat, S_lo-S_hat, S_hi-S_hat, sig_S, TS, sqrt(TS)
```

2Δlog(L)

Frequentist upper limits

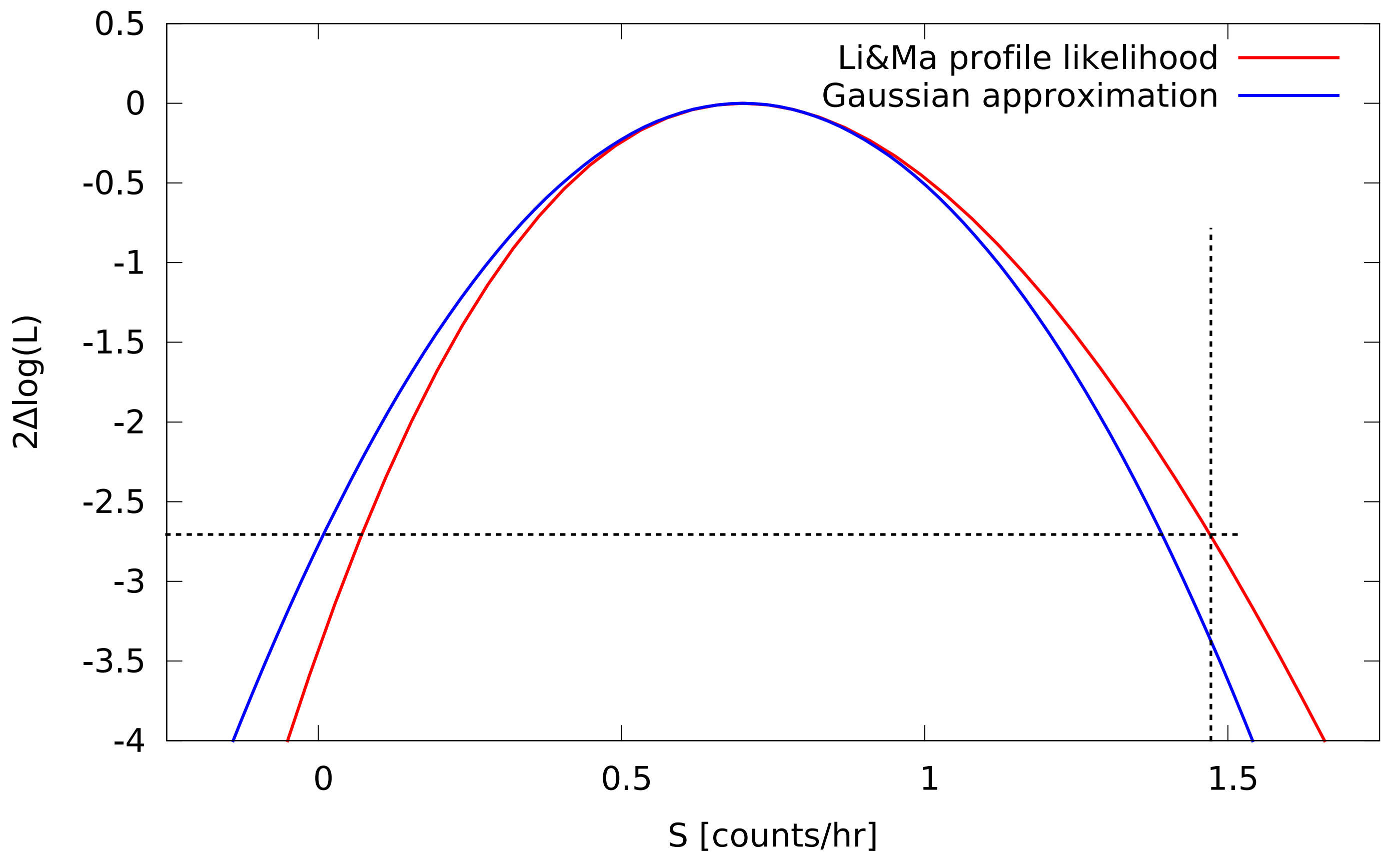
One-sided confidence region using profile likelihood

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



- In two-sided interval search for two points $S_{1,2}$ where $-2\Delta \ln \mathcal{L}(S_{1,2}) = x$ with $\chi^2(x,1) = C$
- For one-sided interval (with $C>0.5$) we need to find single such point with $S_{UL} > \hat{S}$ and for which $0.5 + \chi^2(x,1)/2 = C$ (or $\chi^2(x,1) = 2C - 1$)
- E.g. for $C=0.95$ we search $-2\Delta \ln \mathcal{L}(S_{UL}) = 2.71$

Example of profile likelihood



$$\hat{S} = 0.7^{+0.45}_{-0.39}\,\mathrm{hr}^{-1}$$

- Frequentist upper limit at 95% confidence level:

$$S_{<95\%} = 1.47\,\mathrm{hr}^{-1}$$

Exercise: adapt 2-sided interval code to calculate this

- Our 1ES1218 example isn't very enlightening here, so take:

$$n_{off} = 24$$
$$n_{on} = 15$$
$$\alpha = 1/3$$
$$T = 10.0\,\mathrm{hr}$$

- Giving:

$$\hat{S} = 0.7\,\mathrm{hr}^{-1}$$
$$\sigma_S = 0.42\,\mathrm{hr}^{-1}$$
$$TS = 3.43$$
$$\sigma = 1.85$$

Bayesian statistics

- Likelihood function has no meaning itself, e.g., it is not a probability. Its usefulness comes from theorems such as the LRT.
- MLE belongs to the class of “frequentist” statistical methods: talk about the results of repeated hypothetical experiments.
- Saw how to produce confidence intervals: true parameter value would lie inside the interval in a certain % of hypothetical expts.
- Somewhat awkward language ???

Bayesian statistics

- In Bayesian statistics we talk about the “probability” that the parameters have certain values.

Prior probability density

- Bayes’ theorem:

Likelihood

Posterior probability density

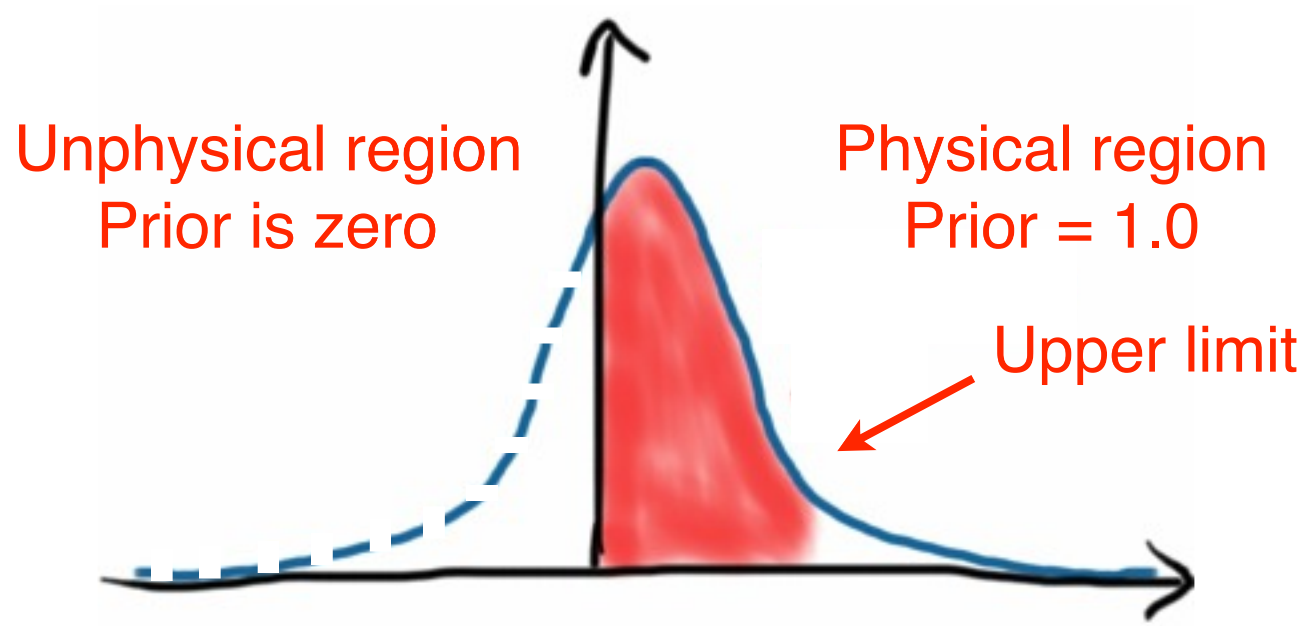
$$P(\Theta|X) = \frac{P(\Theta)P(X|\Theta)}{P(X)} \propto P(\Theta)\mathcal{L}(\Theta|X)$$

relates probability after experiment has been done to probability before.

- Can think of this as refining our belief about the model through experimental results.

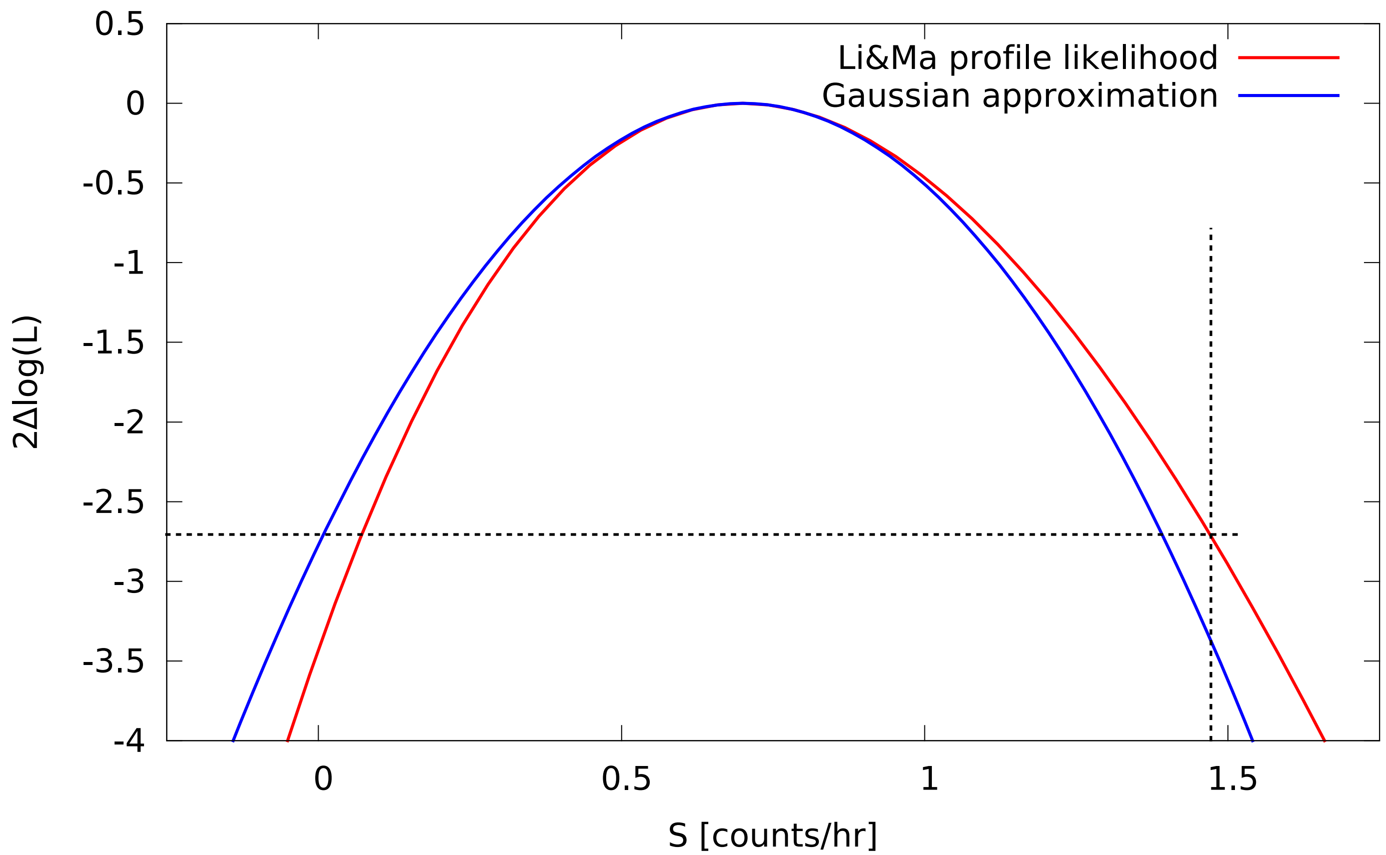
Bayesian upper limits

Or more correctly “Quasi-Bayesian” or “Bayesian-like”



- Bayesian confidence regions correspond to what you would expect...
- ... they are regions that contain a certain fraction of the posterior probability.
- Integrate over parameter from lower bound to find point where integral reaches C% of total.
- In case of multiple parameters, use the profile likelihood. Not strictly a Bayesian approach.

Example of profile likelihood



$$\hat{S} = 0.7^{+0.45}_{-0.39}\,\mathrm{hr}^{-1}$$

- Frequentist upper limit at 95% confidence level:

$$S_{<95\%} = 1.47\,\mathrm{hr}^{-1}$$

- Bayesian 95% upper limit:

$$S_{<95\%} = 1.54\,\mathrm{hr}^{-1}$$

- Our 1ES1218 example isn't very enlightening here, so take:

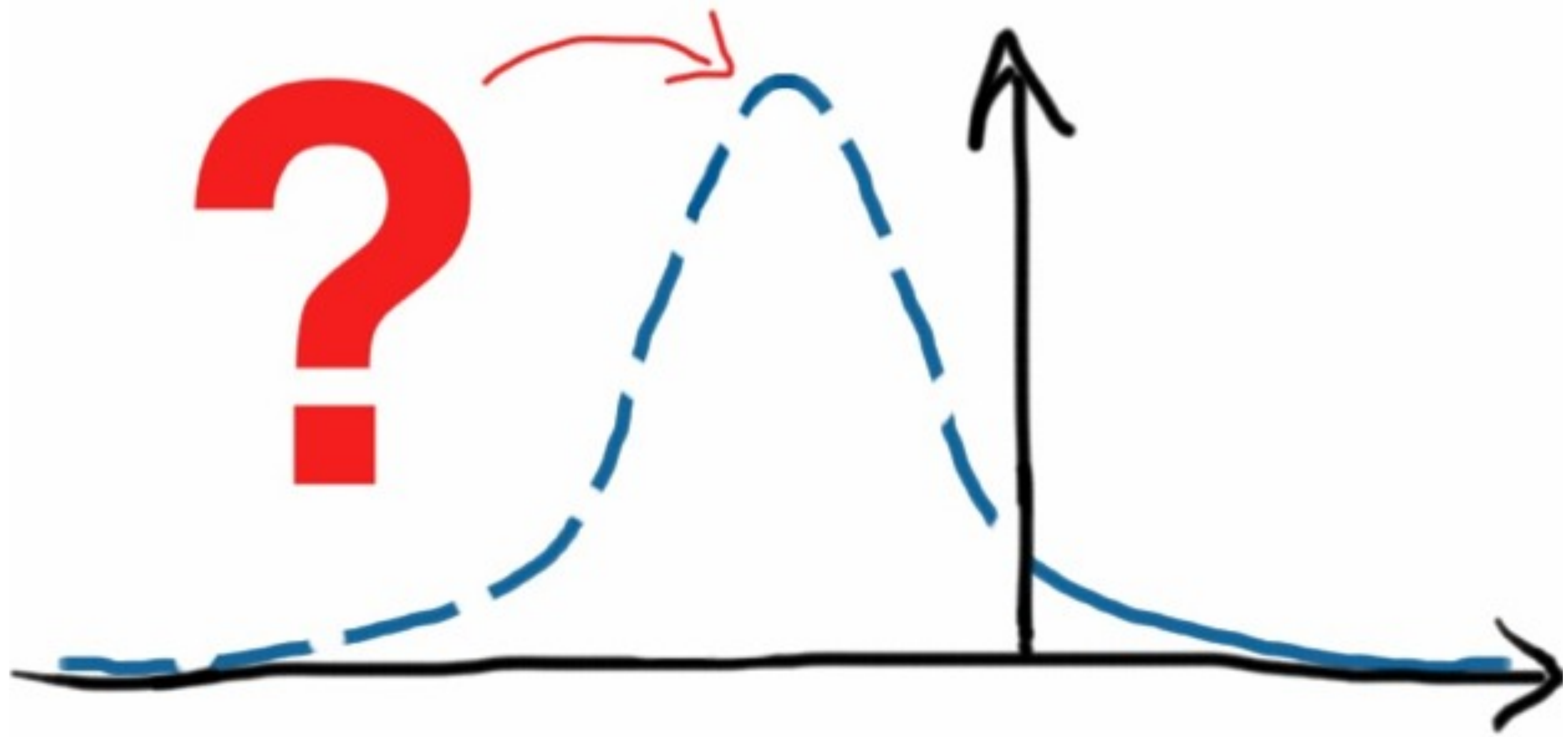
$$n_{off} = 24$$
$$n_{on} = 15$$
$$\alpha = 1/3$$
$$T = 10.0\,\mathrm{hr}$$

- Giving:

$$\hat{S} = 0.7\,\mathrm{hr}^{-1}$$
$$\sigma_S = 0.42\,\mathrm{hr}^{-1}$$
$$TS = 3.43$$
$$\sigma = 1.85$$

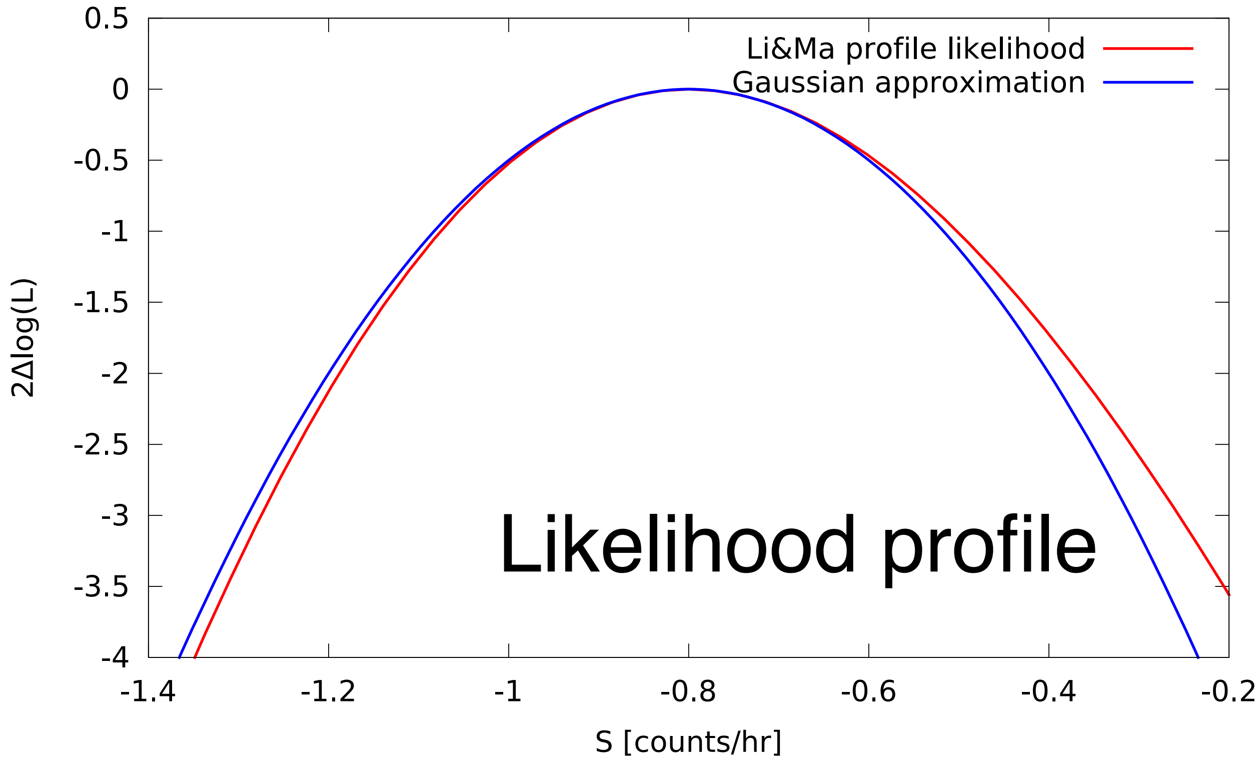
Why have two methods?

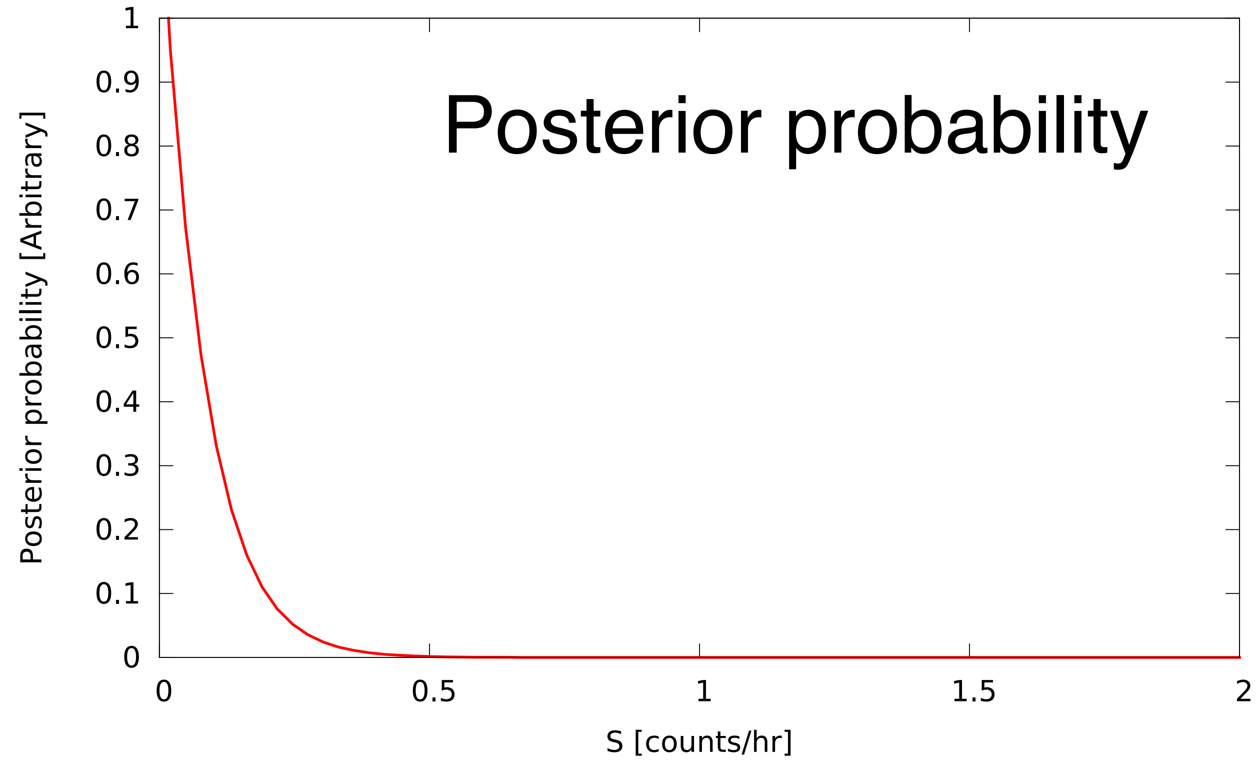
The problem of unphysical upper limits



- Unphysical frequentist upper limits occur can occur if the peak of the likelihood is in an unphysical region of the parameter space.
- More complex (or ad hoc) approaches fix this.
- But Bayesian upper limits are not affected.

Example of unphysical MLE





$$n_{off} = 36$$
$$n_{on} = 4$$
$$\alpha = 1/3$$
$$T = 10.0\,\mathrm{hr}$$

MLE is negative - not physical for source flux.

"Background fluctuation", fewer "On" counts than expected given "Off"

$$\hat{S} = -0.8\,\mathrm{hr}^{-1}$$
$$\sigma_S = 0.28\,\mathrm{hr}^{-1}$$
$$TS = 5.80$$
$$\sigma = -2.41$$

Frequentist UL: $S_{<95\%} = -0.29\,\mathrm{hr}^{-1}$ - unphysical

Bayesian UL: $S_{<95\%} = 0.43\,\mathrm{hr}^{-1}$ - OK!

```
# ul_lima_bayes_1d.py - 2013-05-25 SJF
# Bayesian upper limit in Li & Ma problem
from math import *
import scipy.stats, scipy.optimize, scipy.integrate, sys
# non, noff, alpha, T = (2808, 4959, 1.0/3, 27.2)
# non, noff, alpha, T = (15, 24, 1.0/3, 10.0)
non, noff, alpha, T = (4, 36, 1.0/3, 10.0)
C = 0.95; # Use 95% confidence region (C must be >0.5)
def logL(S,B):
    return non*log(max((S+alpha*B)*T,sys.float_info.min)) + \
    noff*log(max(B*T,sys.float_info.min))-(S+(1+alpha)*B)*T
def profileLogL(S):
    opt_fn = lambda B: -logL(S,B)
    opt_res = scipy.optimize.minimize(opt_fn, 1)
    return -opt_res.fun
S_hat    = (non-noff*alpha)/T
sig_S    = sqrt(non+noff*alpha**2)/T
logL_max = profileLogL(S_hat)
def logPrior(S):
    return log(1);
def logPosterior(S):
    return logPrior(S)+profileLogL(S)-logL_max
def integralPosterior(Smax):
    integrand = lambda S: exp(logPosterior(S))
    y, err = scipy.integrate.quad(integrand,0,Smax)
    return y
total_integral = integralPosterior(S_hat+100*sig_S);
root_fn  = lambda S: integralPosterior(S) - total_integral*C
S_ul = scipy.optimize.brentq(root_fn, 0, S_hat+100*sig_S)
print S_ul, integralPosterior(S_ul)/total_integral, total_integral
```

Good practices

- It is always best to define all the parameters of an analysis before looking at the data.
  - Data selection “cuts”
  - Thresholds for claiming detection.
- It is tempting to adjust the analysis procedure to enhance some small signal, **BUT THIS IS FRAUGHT WITH DANGER!**
- Best practice is to do a blind analysis. Use MC or side-band data to refine analysis in advance.
- But this is not always possible...

Trials factors

Or the “look-elsewhere effect”

- Often you simply don’t know enough in advance to fully determine the analysis, e.g.
  - the mass of the DM particle (or Higgs)
  - the locations of sources in the sky etc...
- So, you must look through the data and search for a significant excess signal ...
- ... and unfortunately you must pay a statistical penalty for doing so.

Trials factors

Or the “look-elsewhere effect”

- If after making N independent tests of for a significant event (e.g N energy channels)
- the most significant test had a P-value of: P_{pre}
- then to account for the number of “trials” you must scale the P-value as: $P_{post} = 1 - (1 - P_{pre})^N$
- For example, a 4σ event has a P-value of $P_{pre} = 6.3 \times 10^{-5}$. With 1000 trials, the post-trial P-value of $P_{post} = 1 - (1 - 6.3 \times 10^{-5})^{1000} = 0.06$ which is equivalent to a 1.9σ event.

Detectability / Sensitivity

- Interested in detectability of sources, i.e. sensitivity of instrument for given threshold.
- Consider "no fluctuations" case where:

$$n_{on}^{NF} = (S_t + \alpha B_t)T, \quad n_{off}^{NF} = B_t T$$

- Then test statistic is:

$$TS^{NF} = 2\left[(S_t + \alpha B_t)T \ln \frac{(1+\alpha)(S_t + \alpha B_t)T}{\alpha(S_t + (1+\alpha)B_t)T} + B_t T \ln \frac{(1+\alpha)B_t T}{(S_t + (1+\alpha)B_t)T}\right]$$

Detectability / Sensitivity

- Weak source case: $S_t \ll \alpha B_t$

$$\sigma^{NF} = \sqrt{TS^{NF}} \approx \frac{\sqrt{T}}{\sqrt{1+\alpha}} \frac{S_t}{\sqrt{\alpha B_t}}$$

Grows as sqrt(T)

- Weak background case: $S_t \gg \alpha B_t$

$$\sigma^{NF} = \sqrt{TS^{NF}} \approx \sqrt{2S_t T \ln(1+1/\alpha)}$$

Note what happens here when $\alpha \to 0$ (which corresponds to perfectly well determined "zero" on-source background) the significance becomes infinite. If you have no background then even one event is a significant.

Detectability / Sensitivity

Minimum source strength to achieve detection at some threshold σ_{det}

- Weak source case: $S_t \ll \alpha B_t$

$$S_t > \frac{\sigma_{det}\sqrt{\alpha B_t}}{\sqrt{T}}\sqrt{1+\alpha}$$

Minimum detectable flux decreases as 1/sqrt(T) and depends on B_t: “Background-dominated regime”

- Weak background case: $S_t \gg \alpha B_t$

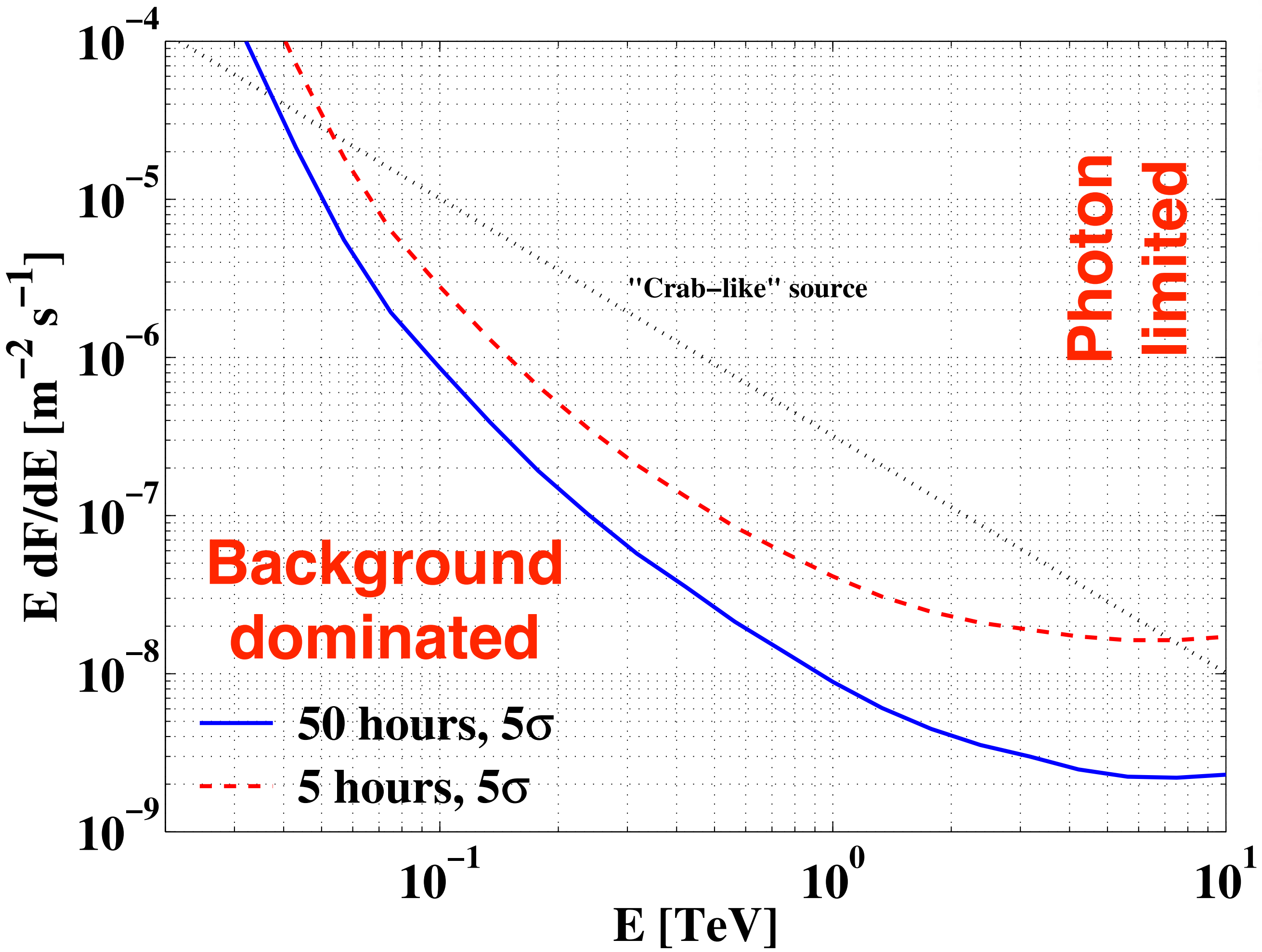
$$S_t > \frac{\sigma_{det}^2}{T}\frac{1}{2\sqrt{1+1/\alpha}}$$

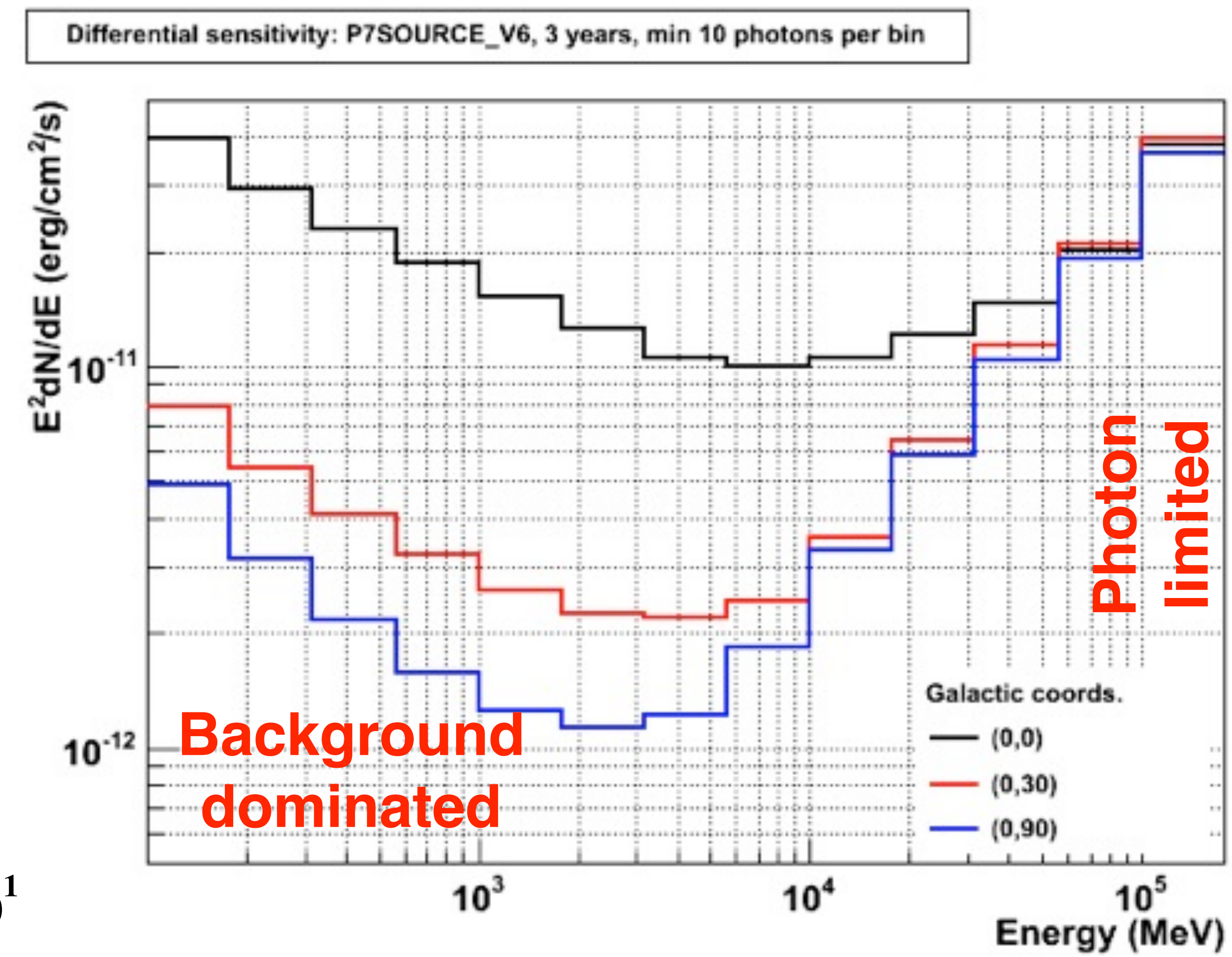
Roughly this says that the number of detected photons must be larger than σ^2 (times some constant): $S_t T = n_{det} > C\sigma_{det}^2$
eg. must detect 25 photons for 5σ.

Minimum detectable flux decreases as 1/T and is independent of B_t : “Photon-limited regime”

Detectability / Sensitivity

"Differential sensitivity" plots, i.e. sensitivity in logarithmic energy bands





Sensitivity for ACT array of 4 telescopes for 5 and 50 hours of observation.

Low energies: sqrt(10) × improvement.
High energies: 10 × improvement.

LAT sensitivity from FSSC site for different background levels (Galactic or extra-galactic).

Low energies: big dependency
High energies: almost no dependency.

Systematic errors

What if assumed value of alpha is incorrect?

- Assume there is no real source:

$$n_{on}^{NF} = \alpha_t B_t T = \alpha(1+\delta) B_t T, \; n_{off}^{NF} = B_t T$$

where the error in alpha is small: $\delta \ll 1$

- Then: $\hat{S}^{NF} = B_t \alpha \delta$

$$\sigma^{NF} = \sqrt{T S^{NF}} \approx \frac{\sqrt{T}}{\sqrt{1+\alpha}} \delta \sqrt{\alpha B_t}$$

- This looks like a real signal. Accurate knowledge of experimental response is critical. MLE is only as good as the model!

Review

- ML provides “cookbook” for estimation and hypothesis testing:
  - estimates: maximum of likelihood
  - errors: curvature of log-likelihood surface
  - TS and significance: is improvement in log-$\mathcal{L}$ over null hypothesis consistent with χ^2?
- Significance expected to grow as sqrt(T), but sensitivity can improve as 1/T if photon limited.
- Systematic errors important to consider

Onwards to LAT analysis...

- LAT ML analysis is fundamentally the same a what we have seen here (but more complex).
- Channels organized by sky position and energy (i.e. 3-dimensions). Million channels typical.
- Model is Poisson for each channel with mean determined by:
  - spatial-spectral model provided by user
  - observational response (calculated by software from IRFs provided by LAT team)
- MLE by software: errors, covariances, TS, etc