Likelihood Methods: A Companion to the NEPPSR analysis project

Colin Gay, Yale University

Outline

- We have arranged a “hands-on” mini-course on fitting techniques using the ROOT framework

- To fit data, we often use the MINUIT package from CERN, which is a fitting engine for numerically minimizing
- It is built-in to ROOT, or can be called stand-alone
- Instead of using it as a black box, we thought we’d show how conceptually simple its operation really is

- Quick review of some basics of probabilities
- Maximum Likelihood basics
- Properties of the ML method, using specific examples

and YES, there will be a test ☺

(Especially for repeat NEPPSR offenders)

Probability Basics

- Suppose X is a random variable. The probability of throwing X between x, x+dx is P(x)dx
  - P(x) is called the *probability density function (pdf)* for X

$$\int P(x)dx = 1$$

- The *Expectation value* of a function f(x) over P(x) is:

$$E(f) = \int f(x)P(x)dx$$

- The most common expectations we use are the first few moments of the pdf:

$$E(1) = \int 1P(x)dx = 1 \qquad \textbf{normalization}$$

$$E(x) = \int xP(x)dx = \overline{x} \qquad \textbf{mean}$$

$$E((x-\overline{x})^2) = \int (x-\overline{x})^2 P(x)dx = \sigma^2 \qquad \textbf{variance}$$

- See Craig Blocker's talk for a more detailed introduction

Probability Basics

- The *Conditional Probability* P(x|a) is the pdf for X, *given that a is true*

- For example, P(x|d) = probability that our detector measures a particle passing a wire at distance x, given that the particle is truly at distance d
- Or: $P(m|m_0)$ = probability of measuring mass m given the true mass is m_0

- We use pdfs all the time in our Monte Carlos: we know true value of masses, trajectories, etc, and turn into finite samples of quantities reconstructed by our detector via these pdfs

- Our job with real data is the inverse: Given a finite sample of measurements of a quantity, to infer our pdf and true value for the quantity

Samples

- Let X have pdf P(x). A Sample of size N is a set of $\{x_i\}$ of N throws of X.
- When we plot, e.g., the mass of all of our reconstructed top quarks, we are visualizing our sample pulled from the pdf $P(m|m_0)$, where m_0 is the true top mass
- Our job is, based on our finite sample, to estimate the true value m_0, and to quantify how certain our estimation is
- For this, we need an *estimator* for m_0
- Some estimators are better than others
  - e.g. My estimator is “150 GeV”, no matter what
  - This is an estimator, but not a very good one!

Estimators

- What properties would we like in our estimators?

- Consistent: As our sample size N increases, we'd like our estimator to converge to the true value. Such an estimator is called *consistent.*
- Efficient: There is a theoretical minimum for the variance of an estimator about the true value, given a sample of size N (called the Minimum Variance Bound)
  - An estimator with variance equal to the MVB is *efficient*
- Unbiased: An estimator whose expectation value (mean) is equal to the true value is called *unbiased*

Maximum Likelihood Estimators

- Suppose we have a sample of N measurements of a variable m, and we know the pdf for m is $P(m|m_0)$.
  - However, we don't know m_0 – in fact, it is the physics quantity we are interested in measuring
- Form the *likelihood* $$L(m_0) = \prod_{i=1}^{N} P(m_i \mid m_0)$$
- The Maximum Likelihood Estimator (MLE) m^* for m_0 is the value of m_0 that maximizes the joint likelihood
- MLEs are not always unbiased, but are consistent and efficient. They are also asymptotically normal.
- Extracting the MLE for a quantity is called "fitting" the data

Least-squares fit (review?)

- First fit most of us learn is a least-squares or χ^2 fit

- Put data into histogram bins: centers x_i, value y_i with uncertainty σ_i
- Choose function $f(x_i \mid \vec{\theta})$ which predicts the bin contents y_i as a function of the "fit parameters" $\vec{\theta}$

- Form $$\chi^2 = \sum \frac{(f(x_i \mid \vec{\theta}) - y_i)^2}{\sigma_i^2}$$

- The Least-Squares Principle states that the best estimate for the parameters $\vec{\theta}$ are the ones which minimize χ^2

Binned vs Unbinned fits

- The LSQ fit is an example of a binned fit
- A LSQ fit has the nice property that in addition to supplying the fit parameters, it also tells us how "good" the fit function approximates the data

- Binned fits with few (or zero) events per bin are problematic
- Gaussian approx of uncertainty σ_i on bin contents is 0
  - Contribution to χ^2 undefined
  - Root just ignores these bins => fit biased high

- Likelihood fits give us the ability to deal with data in an unbinned way
  - Ideal for small statistics, or sparse data

MLE Properties – NEPPSR project

- I'll develop the concepts of the ML fit using an analytically solvable case
- You will write a fitting program, fit to data supplied by Stephane, Kevin and John, and can compare to analytic result

- I'll use a common real-world case – fitting for the lifetime of a data sample of decay times

Lifetime Likelihood

- We can write the probability density function for an exponential decay in two ways:

$$P(t\,|\,\tau) = \frac{1}{\tau} e^{-t/\tau} \quad \text{or} \quad P(t\,|\,\Gamma) = \Gamma e^{-\Gamma t}$$

- These are properly normalized (more on this later)

$$\int P(t\,|\,\tau)\,dt = 1$$

- Construct the likelihood

$$L(\tau) = \prod_{i=1}^{N} P(t_i\,|\,\tau)$$

from the N time measurements $\{t_i\},\ i = 1, N$

Lifetime Likelihood

- Rather than Maximizing the likelihood, we usually take the log
- Numerically, the likelihood can get extremely small, resulting in precision issues. log is better.
- log is monotonic, so maximizing logL is equivalent to L
- Our main fitting package, MINUIT, finds Minima, so we multiply by -1 and Minimize:

$$
\begin{aligned}
-\log L(\tau) &= -\sum_{i=1}^{N} \log P(t_i \mid \tau) \\
&= -\sum_{i=1}^{N} \log\left(\frac{1}{\tau} e^{-t_i/\tau}\right) \\
&= \sum_{i=1}^{N} \left(\log \tau + \frac{t_i}{\tau}\right) \\
&= N \log \tau + \frac{1}{\tau} \sum_{i=1}^{N} t_i
\end{aligned}
$$

Maximizing the Likelihood

- To find the minimum, we set the first derivative to 0

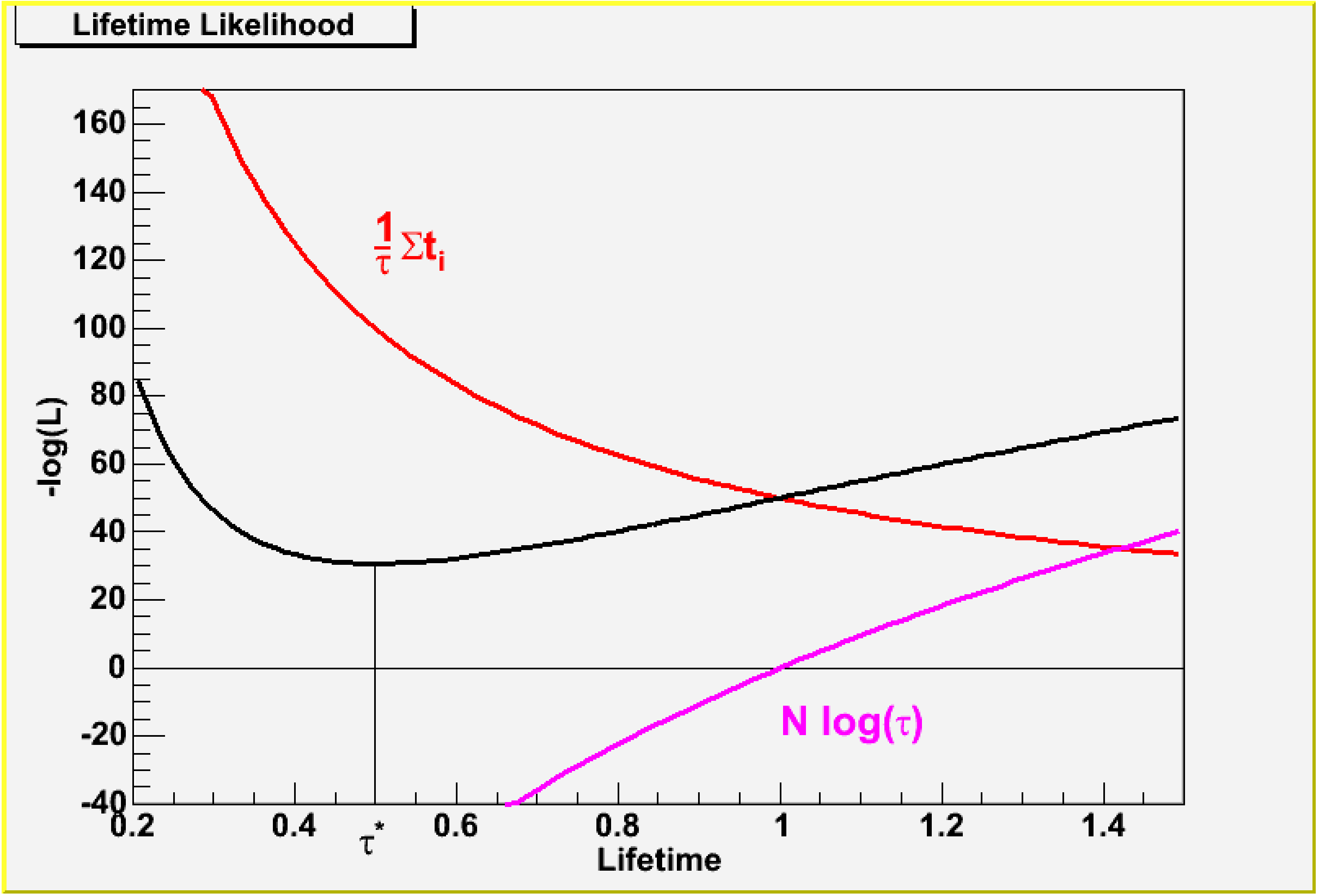
$$\frac{\partial(-\log L)}{\partial\tau} = 0 = \frac{N}{\tau} - \frac{1}{\tau^2}\sum_{i=1}^{N} t_i$$

$$\Rightarrow \quad N\tau = \sum_{i=1}^{N} t_i$$

$$\Rightarrow \quad \tau^* = \frac{1}{N}\sum_{i=1}^{N} t_i \quad (= E(t) = \overline{t} = \text{mean})$$

τ^* is the ML estimator for the true lifetime τ

Likelihood for a Lifetime



Numerical method

- While we can solve this system analytically, in general we must build up the likelihood numerically by looping over the events and adding up the −log P:

```
sum = 0;
for (i=0; i<N; i++) {
  sum += log(tau) + t[i]/tau;
}

Find minimum sum by scanning in tau
```

- This allows for very complicated probability functions to be handled in a straightforward way

What about the width?

- What if we considered the width to be our unknown, rather than the lifetime?

$$-\log L(\Gamma) = -\sum_{i=1}^{N} \log(\Gamma e^{-\Gamma t_i}) \;=\; \sum_{i=1}^{N} (-\log \Gamma + \Gamma t_i)$$

$$= -N \log \Gamma + \Gamma \sum_{i=1}^{N} t_i$$

$$\frac{\partial(-\log L)}{\partial \Gamma} = 0 = -\frac{N}{\Gamma} + \sum_{i=1}^{N} t_i$$

$$\Rightarrow \quad \Gamma^* = \frac{N}{\sum_{i=1}^{N} t_i} \qquad (\text{i.e. } \Gamma^* = 1/\tau^*)$$

Unbiasedness?

- Let's calculate the expectation (mean) value of our

  estimator $\tau^* = \frac{1}{N}\sum_{i=1}^{N} t_i$ given the true lifetime is τ

$$E(\tau^*) = E\left(\frac{1}{N}\sum_{i=1}^{N} t_i\right) = \frac{1}{N}\sum_{i=1}^{N} E(t_i)$$

$$= \frac{1}{N}\sum_{i=1}^{N}\int t_i \frac{1}{\tau} e^{-t_i/\tau} dt_i$$

$$= \int_0^\infty t \frac{1}{\tau} e^{-t/\tau} dt = te^{-t/\tau}\Big|_0^\infty - \int_0^\infty e^{-t/\tau} dt$$

$$= \tau$$

- The mean value of our fit result for τ^*, if we repeat the experiment many times, is the true value τ

  =>Unbiased

Bias (continued)

- What about Γ?

$$E(\Gamma^*) = E\left(\frac{N}{\sum_{i=1}^{N} t_i}\right) = N \int \ldots \int \frac{1}{\sum_{i=1}^{N} t_i} P(t_1 \mid \Gamma) \ldots P(t_N \mid \Gamma) dt_1 \ldots dt_N \neq \Gamma$$

- Hence Γ^* is a biased estimator. This sounds bad …

- However, fitting for lifetime or width gives the *same* answer: Remember

$$\Gamma^* = 1/\tau^*$$

- You can get a fine fit with either. It should be no surprise that $E\left(\frac{1}{x}\right) \neq \frac{1}{E(x)}$

  so that if estimator x is unbiased, 1/x must be biased

MLE transformation invariance

- In fact, the result of the likelihood fit is invariant under parameter transformation:
  - If f(x) is any function of the estimator x, then the MLE for f(x) satisfies

$$(f(x))^* = f(x^*)$$

  - That is, fitting for x, and then applying the transformation f gives the same result as fitting for f(x)
- Thus there's more than one way to skin any cat …

Uncertainty on ML Estimator

- Taylor expand likelihood about minimum:

$$\log L(\theta) = \log L|_{\theta^*} + \frac{1}{2}\frac{\partial^2 \log L}{\partial \theta^2}\bigg|_{\theta=\theta^*}(\theta-\theta^*)^2$$

$$L(\theta) = L|_{\theta^*}\, e^{\frac{1}{2}\frac{\partial^2 \log L}{\partial \theta^2}\big|_{\theta=\theta^*}(\theta-\theta^*)^2}$$

- **If we consider this a probability density for the true value of the parameter θ, we see it is a Gaussian, with mean θ^* and variance

$$\sigma^2 = V(\theta) = -\frac{1}{\frac{\partial^2 \log L}{\partial \theta^2}\Big|_{\theta=\theta^*}}$$

Meaning of Likelihood

- Depends on if you are Bayesian or Frequentist
  - While most particle physicists are professed frequentists, they also seem to me to be closet Bayesians at times
- Let's be a bit more careful in notation. The likelihood is

$$L(\{t_i\}|\tau) = \prod_{i=1}^{N} P(t_i|\tau)$$

- Bayes: We'd like to invert the probability, and consider this as the probability density for the *true value* τ given the observations t_i
- In general P(A|B) ≠ P(B|A). In fact, Bayes' thm is:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

(follows from $P(A|B)P(B) = P(B|A)P(A)$**)**

Meaning of Likelihood

- Thus $P(\tau \mid data) = \dfrac{P(data \mid \tau) P(\tau)}{P(data)}$
- $P(data)$ is some constant normalization
- If we take the so-called *prior probability* function $P(\tau)$ to be flat, then the likelihood *is* the probability distribution for the true value of τ

$$P(\tau \mid data) = P(data \mid \tau) = L(\tau)$$

ML Uncertainties

- Writing in a more suggestive way:

$$\log L(\theta) = \log L|_{\theta^*} - \frac{1}{2}\frac{(\theta - \theta^*)^2}{\sigma^2}$$

- Hence the values of the likelihood for the true value of θ being 1, 2, n σ from the central value are

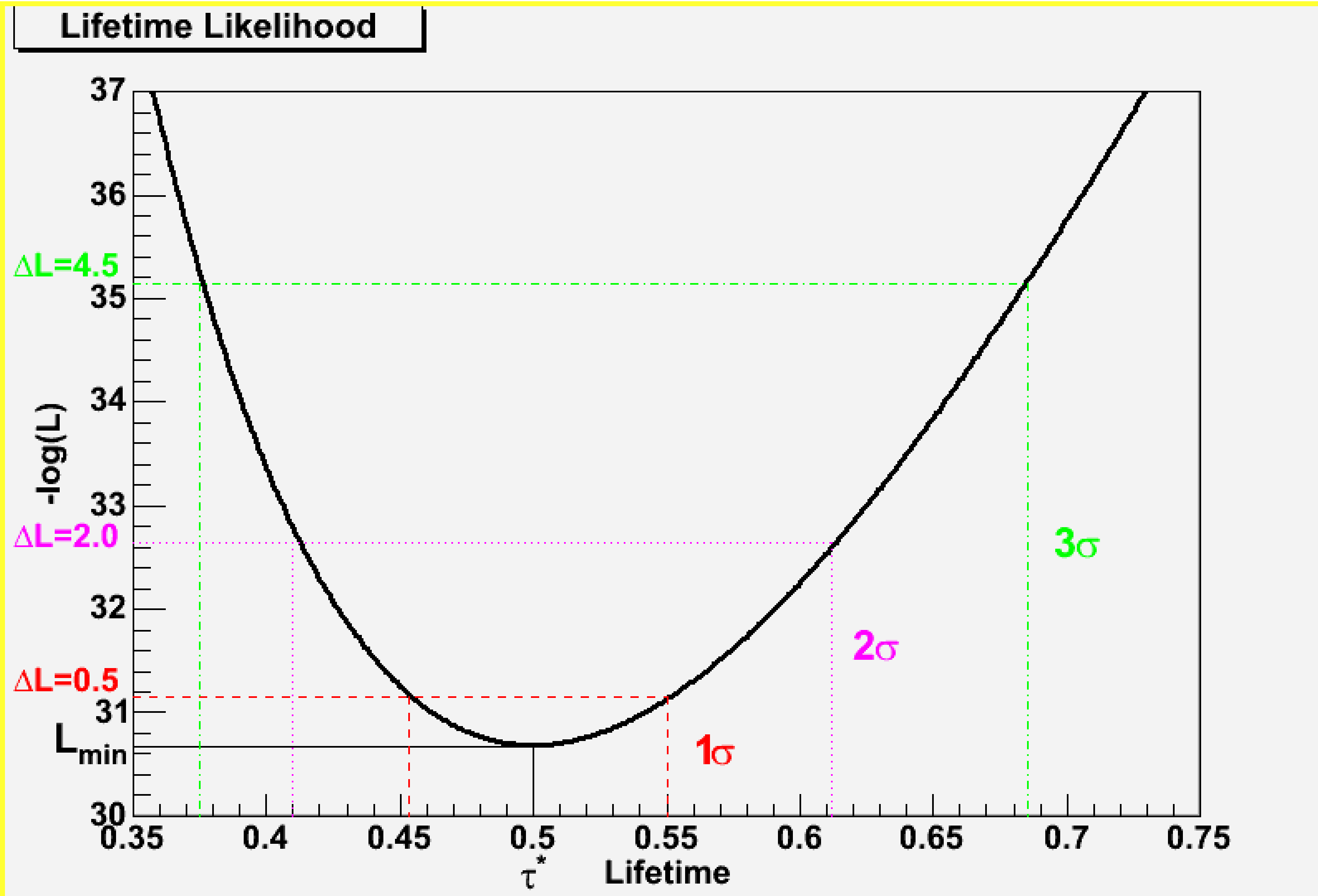
$$-\log L(\theta^* \pm 1\sigma) = -\log L|_{\theta^*} + \frac{1}{2}$$

$$-\log L(\theta^* \pm 2\sigma) = -\log L|_{\theta^*} + 2$$

$$-\log L(\theta^* \pm n\sigma) = -\log L|_{\theta^*} + \frac{1}{2}n^2$$

- For fits, $\chi^2 = \sum \frac{(f_i - x_i)^2}{\sigma_i^2}, \quad \Delta\chi^2 = 1,4,9$ correspond to 1, 2, 3 σ excursions
- Same for ML fits, except factor of 1/2

Likelihood Uncertainty



Estimated Variance of Lifetime fit

- In our case of the lifetime fit: $$-\log L = N \log \tau + \frac{1}{\tau}\sum_{i=1}^{N} t_i$$

$$\frac{\partial(-\log L)}{\partial \tau} = 0 = \frac{N}{\tau} - \frac{1}{\tau^2}\sum_{i=1}^{N} t_i$$

$$\Rightarrow \left.\frac{\partial^2(-\log L)}{\partial \tau^2}\right|_{\tau=\tau^*} = -\frac{N}{\tau^{*2}} + \frac{2}{\tau^{*3}}\sum_{i=1}^{N} t_i$$

$$= -\frac{N}{\tau^{*2}} + \frac{2N}{\tau^{*2}} = \frac{N}{\tau^{*2}} \Rightarrow$$

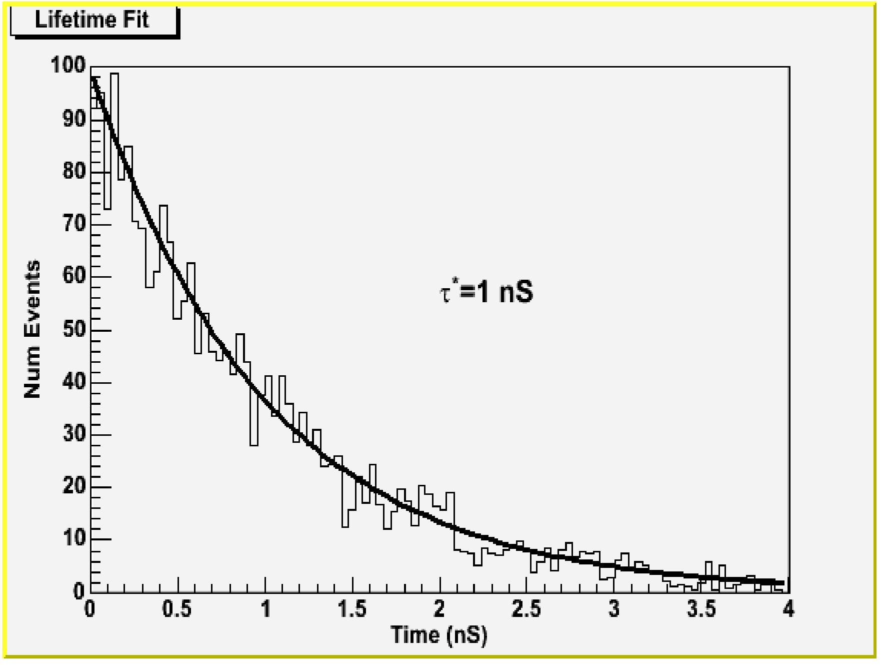
$$\sigma = \frac{1}{\sqrt{\left.\frac{\partial^2(-\log L)}{\partial \tau^2}\right|_{\tau=\tau^*}}} = \frac{\tau^*}{\sqrt{N}}$$

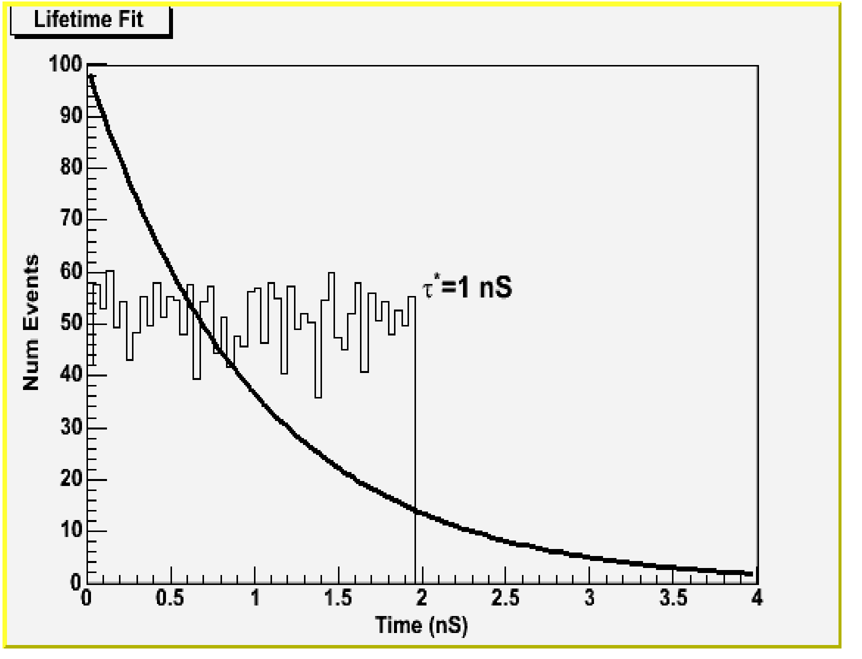
Lifetime fit Uncertainties

- Note that samples with fitted lifetimes that fluctuate low ALSO have the smallest uncertainty!
- This can easily (and did/does) cause world-averages of many measurements to be biased low, as low values get the largest weight
- More correct to combine the likelihood curves to average many experiments
- This is now done behind the scenes for many measurements
  - Experts supply sampling of the likelihood curve for their fit
  - L curves are ADDED
  - Minimum of summed L found, and Δ L=1/2 gives combined uncertainty

Goodness of Fit

- Consider the two data sets below, made into histograms for visualization. Both result in the same ML estimator for the lifetime.
- Surely, data set 1 is "more" likely than data set 2, right?
- Surely, since it is more exponential, the value of the likelihood function for the 1st should be larger than for the 2nd, right? Each events "probability" should be higher, resulting in a net larger likelihood.





Goodness of Fit

- Unfortunately, this is wrong. Recall:

$$-\log L = N\log\tau + \frac{1}{\tau}\sum_{i=1}^{N} t_i \quad \text{and} \quad \tau^* = \frac{1}{N}\sum_{i=1}^{N} t_i$$

$$-\log L_{\min} = N\log\tau^* + \frac{1}{\tau^*}\sum_{i=1}^{N} t_i = N\log(\frac{1}{N}\sum t_i) + \frac{N}{\sum t_i}\sum t_i$$

$$= N(-\log N + 1 + \log\sum t_i) = N(\log\tau^* + 1)$$

- ANY sample with the same $\sum t_i$ produces the SAME estimate $\tau^* = \sum t_i / N$ with the SAME value for the –logL of $N(\log\tau^* + 1)$

- The two previous fits are "equally likely"!
- This is a weakness of the ML method – it doesn't naturally supply a "goodness of fit" metric

Moral

- The ML method will not save you from yourself!
- Sanity check of results essential!
- How likely is likely needs care: e.g. from a self-help mathematics website:
  - **Definition of Unlikely Event**
  - The event that may not happen is an unlikely event.
  - In other words, unlikely event is an event that is not likely to happen.

Likelihood Normalization

- You must be vigilant that your likelihood is properly normalized (or at least its normalization doesn't change)
- Easiest way is to normalize each building-block probability distribution
- Ignoring this can cause real problems:
  - Consider a ML fit with 10,000 data events
  - Suppose the normalization of the underlying pdfs changes from 1 to 0.9999

$$-\log L = \sum_{i=1}^{10,000} \log P(t_i \mid \tau)$$

$$\Rightarrow \sum_{i=1}^{10,000} \log 0.9999 P(t_i \mid \tau)$$

$$= \sum_{i=1}^{10,000} (\log 0.9999 + \log P(t_i \mid \tau))$$

$$= \sum_{i=1}^{10,000} (-.0001 + \log P(t_i \mid \tau))$$

$$= \sum_{i=1}^{10,000} \log P(t_i \mid \tau) - 1$$

More than 1 sigma Change!

ML Fit with Constraint

- It is easy to add external constraints on fit parameters
- Let's assume Gaussian uncertainty on constraint
  - e.g. add in the world-average knowledge as given by PDG

$$L = P(\tau \mid \tau_{PDG}, \sigma_{PDG}) \prod_{i=1}^{N} P(t_i \mid \tau)$$

$$-\log L = -\log P(\tau \mid \tau_{PDG}, \sigma_{PDG}) + \sum_{i=1}^{N} \log P(t_i \mid \tau)$$

$$= -\log\left(\frac{1}{\sqrt{2\pi}\sigma_{PDG}} e^{-\frac{(\tau-\tau_{PDG})^2}{2\sigma_{PDG}^2}}\right) + \sum_{i=1}^{N} \log P(t_i \mid \tau)$$

$$= \sum_{i=1}^{N} \log P(t_i \mid \tau) + \frac{(\tau - \tau_{PDG})^2}{2\sigma_{PDG}^2} + \log \sqrt{2\pi}\sigma_{PDG}$$

Original

$\frac{1}{2}\chi^2$ **penalty**

Constant

Wandering by 1 sigma from constraint costs ½ unit of likelihood (=1 sigma)

Lifetime with imperfect Detector

- All detectors have non-zero time resolution

  -> Let's add this effect in to our fit
- Assume detector has a gaussian resolution function, with mean = 0 (i.e. unbiased) and width

$$P(t\,|\,\tau) = Exp(t',\tau) \otimes G(0,\sigma)$$

$$= \frac{1}{\sqrt{2\pi}\sigma\tau} \int_0^\infty e^{-t'/\tau} e^{-\frac{(t-t')^2}{2\sigma^2}} dt'$$

- Consider the exponentials. The exponent is:

$$-\frac{t'}{\tau} - \frac{(t-t')^2}{2\sigma^2} = -\frac{1}{2\sigma^2}\left(t'^2 - 2t't + t^2 + 2\frac{\sigma^2}{\tau}t'\right)$$

$$= -\frac{1}{2\sigma^2}\left(t'^2 - 2t'\left(t - \frac{\sigma^2}{\tau}\right) + t^2\right)$$

$$= -\frac{1}{2\sigma^2}\left[\left(t' - \left(t - \frac{\sigma^2}{\tau}\right)\right)^2 + 2\frac{\sigma^2 t}{\tau} + \frac{\sigma^4}{\tau^2}\right]$$

$$= -\frac{1}{2\sigma^2}\left(t' - \left(t - \frac{\sigma^2}{\tau}\right)\right)^2 - \frac{t}{\tau} - \frac{\sigma^2}{2\tau^2}$$

Lifetime with Imperfections

- PDF becomes

$$P(t\,|\,\tau) = \frac{1}{\sqrt{2\pi}\sigma\tau}\int_0^\infty e^{-t/\tau} e^{-\frac{\sigma^2}{2\tau^2}} e^{-\frac{(t'-(t-\frac{\sigma^2}{\tau}))^2}{2\sigma^2}} dt'$$

$$= \frac{1}{\tau} e^{-t/\tau} e^{-\frac{\sigma^2}{2\tau^2}} \frac{1}{\sqrt{2\pi}\sigma}\int_0^\infty e^{-\frac{(t'-(t-\frac{\sigma^2}{\tau}))^2}{2\sigma^2}} dt'$$

$$= \frac{1}{\tau} e^{-t/\tau} e^{-\frac{\sigma^2}{2\tau^2}} \frac{1}{\sqrt{2\pi}\sigma}\int_{-(t-\frac{\sigma^2}{\tau})}^\infty e^{-\frac{t'^2}{2\sigma^2}} dt'$$

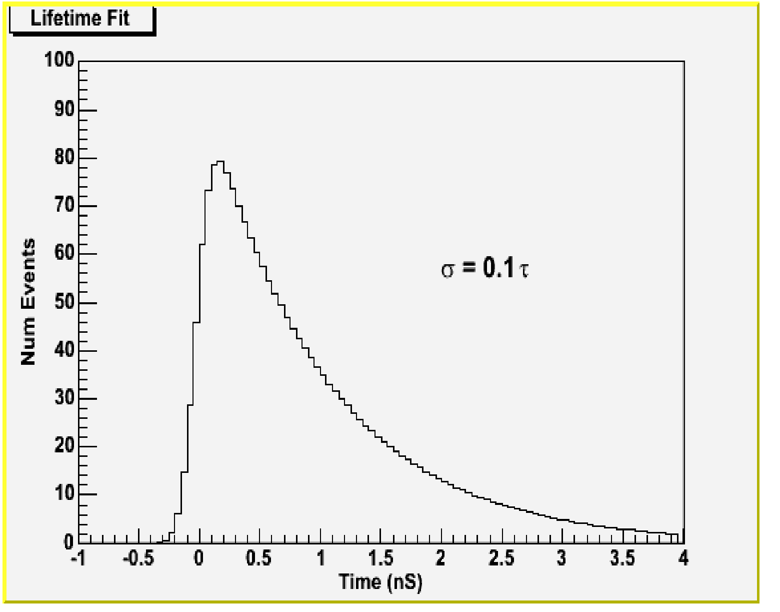
$$= \frac{1}{\tau} e^{-t/\tau} e^{-\frac{\sigma^2}{2\tau^2}} \frac{1}{\sqrt{\pi}}\int_{-\frac{1}{\sqrt{2}}(\frac{t}{\sigma}-\frac{\sigma}{\tau})}^\infty e^{-x^2} dx$$

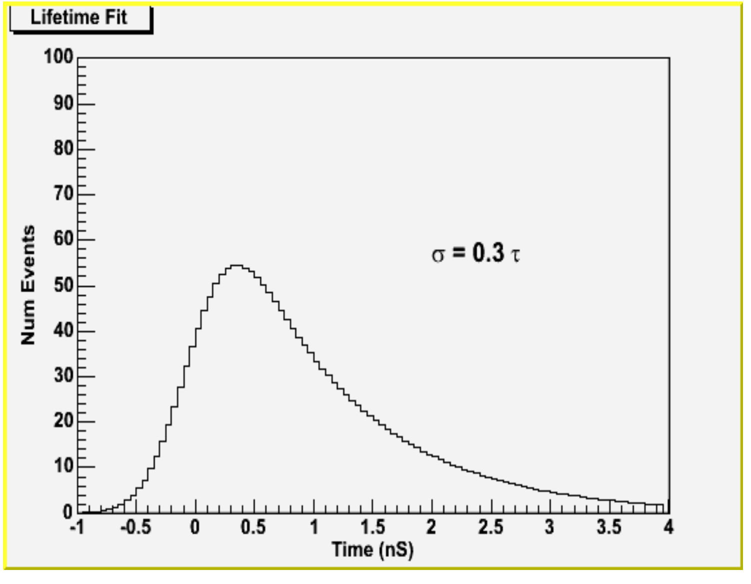
$$= \frac{1}{\tau} e^{-t/\tau} e^{-\frac{\sigma^2}{2\tau^2}} \frac{1}{2} erfc(-\frac{1}{\sqrt{2}}(\frac{t}{\sigma}-\frac{\sigma}{\tau}))$$

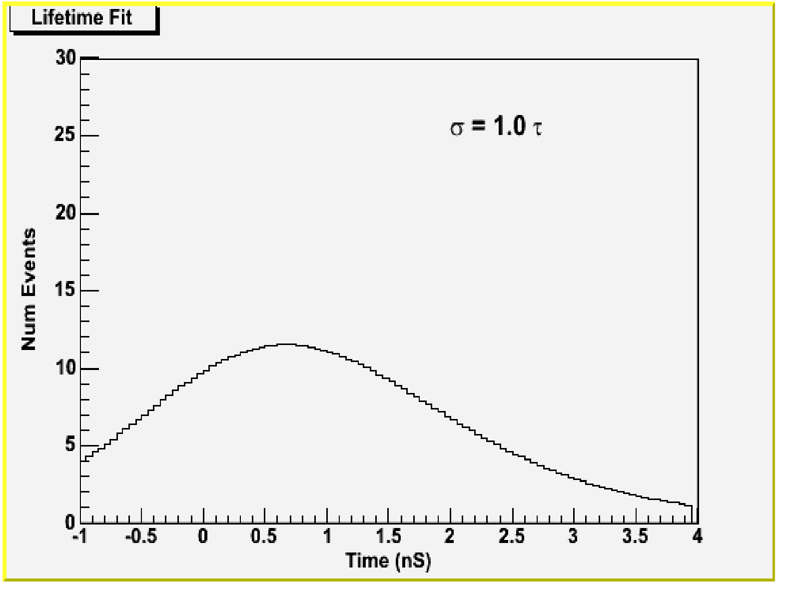
Effect of Worsening Resolution

$$P(t|\tau) = \frac{1}{\tau} e^{-t/\tau} e^{-\frac{\sigma^2}{2\tau^2}} \frac{1}{2} erfc(-\frac{1}{\sqrt{2}}(\frac{t}{\sigma} - \frac{\sigma}{\tau}))$$

- As σ grows, pdf looks less like exponential, more like the gaussian resolution function







Lifetime with Cutoff

- Suppose we are measuring the ½ life of some radioactive material, but we only have T days to run our experiment
- Our probability density for observing a decay must cut off at T days, affecting the normalization

$$P(t \mid \tau) = \frac{1}{1 - e^{-T/\tau}} \frac{1}{\tau} e^{-t/\tau}$$

$$-\log L(\tau) = -\sum_{i=1}^{N} \log\left(\frac{1}{1 - e^{-T/\tau}} \frac{1}{\tau} e^{-t_i/\tau}\right)$$

$$= -N\left(\log(1 - e^{-T/\tau}) + \log \tau\right) - \frac{1}{\tau}\sum_{i=1}^{N} t_i$$

- Now solve numerically … Note that the normalization changes with τ

Lifetime with Background

- Suppose we have a prompt (0 lifetime) background in our sample (measured with resolution σ)
- We can easily handle this by making our per-event probability:

$$P(t \mid \tau, f) = f \frac{1}{\tau} e^{-t/\tau} \otimes G(0,\sigma) + (1-f) G(t \mid 0,\sigma)$$

$$f = \text{fraction of signal}$$
$$G = \text{gaussian}$$

- Form –log L as usual, minimize wrt both τ, f

Multiple variables

- Maximum Likelihood fit easily handles multiple fit variables – simply multiply in the probability densities for each new variable
- Least-squares fit has problems with such fits
  - As we add more and more dimensions to our fit, binned methods encounter trouble
  - Data gets spread thin over large number of histogram bins, resulting in few entries (or zero) per bin => problematic for the fit
- e.g. Adding the reconstructed mass to the decay time as variables to fit and distinguish signal from background

$$P(t,m\,|\,\tau,f,m_B) = f\{\frac{1}{\tau}e^{-t/\tau} \otimes G(0,\sigma))G(m\,|\,m_B,\sigma_m)\}$$

$$+ (1-f)\{G(t\,|\,0,\sigma)(am+b)\}$$

Gaussian signal

Linear background

Summary

- Maximum Likelihood fit is a powerful, convenient technique for estimating parameters from finite samples
- Unbinned, so small statistics, sparse data ok
- Best choice for many-parameter fits
- For large N, gives unbiased value, converges to true value, and has minimum uncertainty possible

- Constant normalization critical
- Auxiliary goodness of fit required
- Visualization can be difficult

- Most serious fits you do in your career will be Maximum Likelihood fits

- Stephane will talk about the hands-on part of the project