Statistical Methods for Particle Physics
Lecture 3: Further Topics

http://indico.ihep.ac.cn/event/5966/

iSTEP 2016
Tsinghua University, Beijing
July 10-20, 2016

Glen Cowan (谷林·科恩）
Physics Department
Royal Holloway, University of London
g.cowan@rhul.ac.uk
www.pp.rhul.ac.uk/~cowan

Outline

Lecture 1: Introduction and review of fundamentals
   Probability, random variables, pdfs
   Parameter estimation, maximum likelihood
   Statistical tests

Lecture 2: Discovery and Limits
   Comments on multivariate methods (brief)
   p-values
   Testing the background-only hypothesis: discovery
   Testing signal hypotheses: setting limits

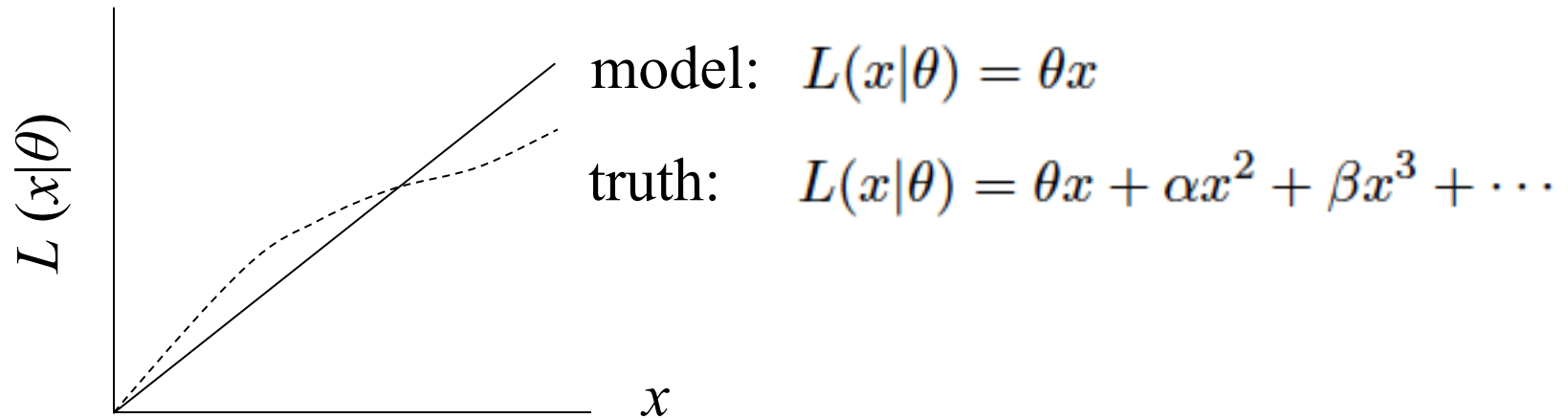➡ **Lecture 3: Systematic uncertainties and further topics**
   Nuisance parameters (Bayesian and frequentist)
   Experimental sensitivity
   The look-elsewhere effect

Systematic uncertainties and nuisance parameters

In general our model of the data is not perfect:

model: $L(x|\theta) = \theta x$

truth: $L(x|\theta) = \theta x + \alpha x^2 + \beta x^3 + \cdots$

(plot with axes labeled $L(x|\theta)$ and x, showing model as solid line and truth as dashed line)

Can improve model by including additional adjustable parameters.

$$L(x|\theta) \rightarrow L(x|\theta, \nu)$$

Nuisance parameter ↔ systematic uncertainty. Some point in the parameter space of the enlarged model should be "true".

Presence of nuisance parameter decreases sensitivity of analysis to the parameter of interest (e.g., increases variance of estimate).

Example: fitting a straight line

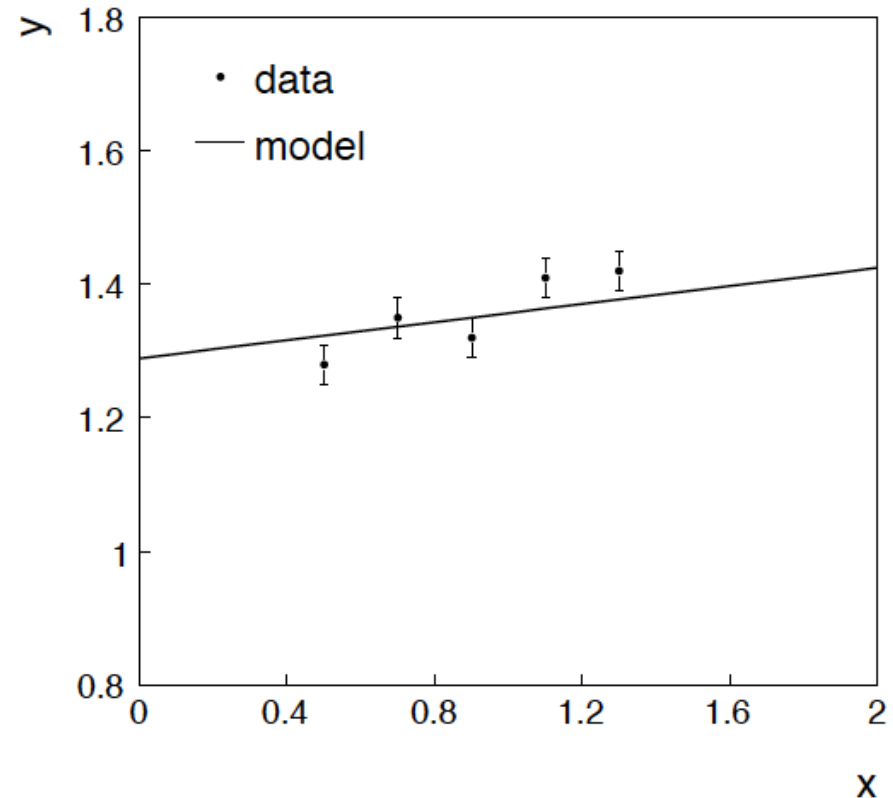Data: (x_i, y_i, σ_i), $i = 1, \ldots, n$.

Model: y_i independent and all follow $y_i \sim \text{Gauss}(\mu(x_i), \sigma_i)$

$$\mu(x; \theta_0, \theta_1) = \theta_0 + \theta_1 x,$$

assume x_i and σ_i known.

Goal: estimate θ_0

Here suppose we don't care about θ_1 (example of a "nuisance parameter")

Maximum likelihood fit with Gaussian data

In this example, the y_i are assumed independent, so the likelihood function is a product of Gaussians:

$$L(\theta_0, \theta_1) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2}\right],$$

Maximizing the likelihood is here equivalent to minimizing

$$\chi^2(\theta_0, \theta_1) = -2\ln L(\theta_0, \theta_1) + \text{const} = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2}.$$

i.e., for Gaussian data, ML same as Method of Least Squares (LS)

θ_1 known a priori

$$L(\theta_0) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2}\right] \, .$$
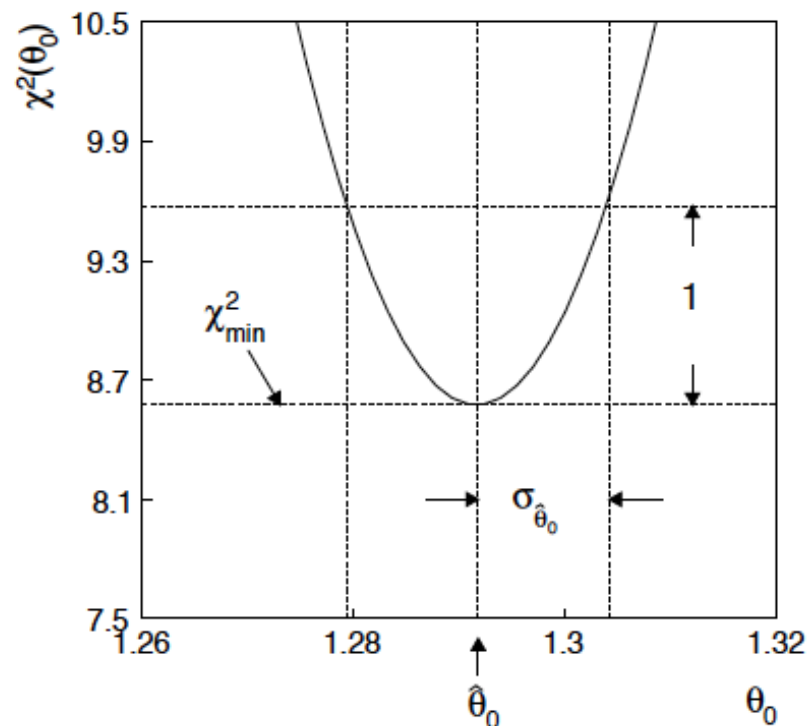
$$\chi^2(\theta_0) = -2\ln L(\theta_0) + \text{const} = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} \, .$$

For Gaussian y_i, ML same as LS

Minimize $\chi^2 \rightarrow$ estimator $\hat{\theta}_0$.

Come up one unit from $\chi^2_{\min}$
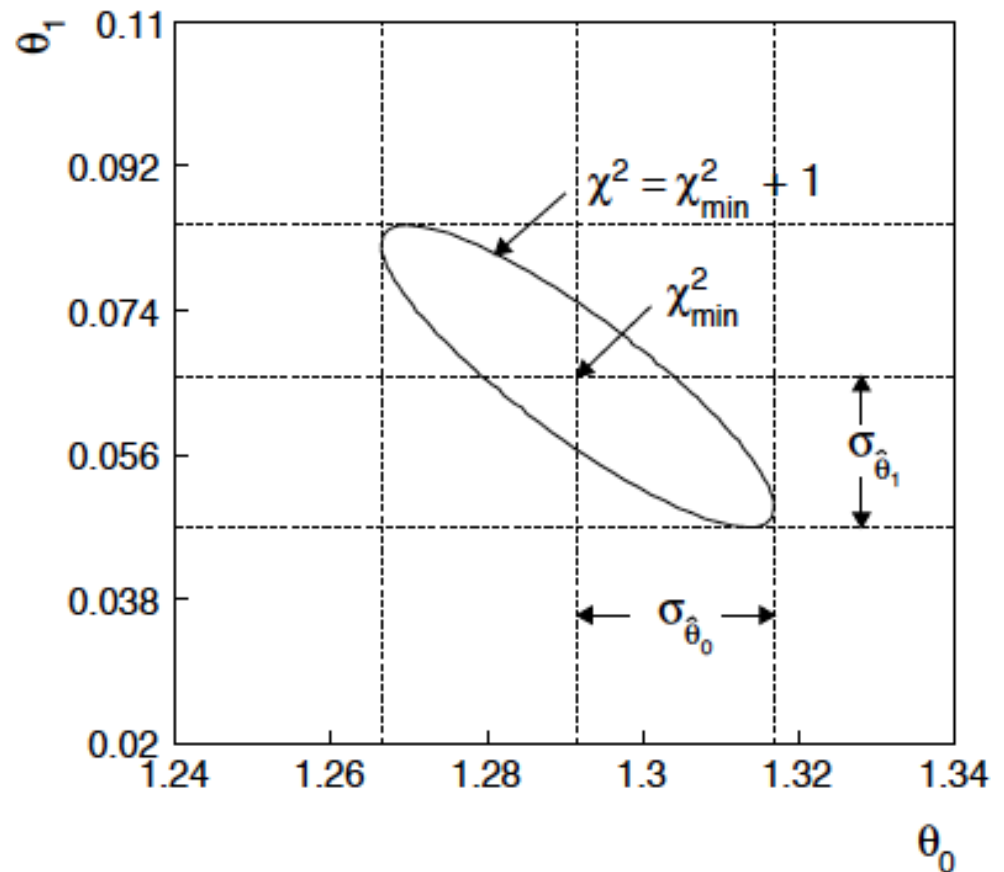
to find $\sigma_{\hat{\theta}_0}$.

ML (or LS) fit of θ_0 and θ_1

$$\chi^2(\theta_0, \theta_1) = -2 \ln L(\theta_0, \theta_1) + \text{const} = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} \; .$$

Standard deviations from tangent lines to contour

$$\chi^2 = \chi^2_{\text{min}} + 1 \; .$$

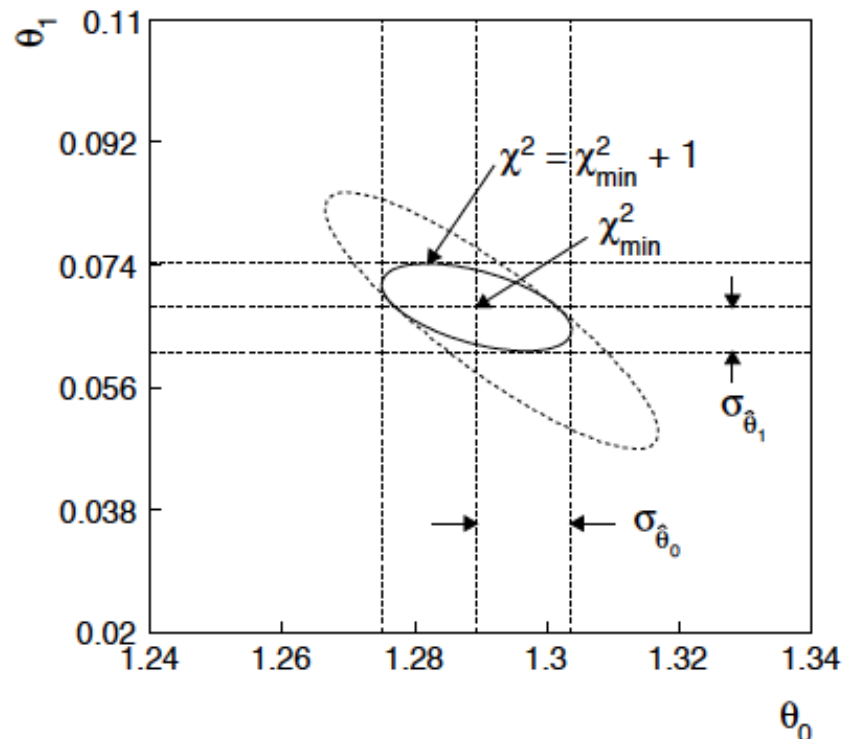Correlation between $\hat{\theta}_0, \; \hat{\theta}_1$ causes errors to increase.

If we have a measurement $t_1 \sim \text{Gauss}\,(\theta_1, \sigma_{t1})$

$$L(\theta_0, \theta_1) = \frac{1}{\sqrt{2\pi}\sigma_t} e^{-(t_1-\theta_1)^2/2\sigma_{t_1}^2} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\frac{(y_i - \mu(x_i;\theta_0,\theta_1))^2}{\sigma_i^2}\right]$$

$$\chi^2(\theta_0, \theta_1) = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i;\theta_0,\theta_1))^2}{\sigma_i^2} + \frac{(t_1 - \theta_1)^2}{\sigma_{t_1}^2}$$

The information on θ_1

improves accuracy of $\hat{\theta}_0$.

The Bayesian approach

In Bayesian statistics we can associate a probability with a hypothesis, e.g., a parameter value θ.

Interpret probability of θ as 'degree of belief' (subjective).

Need to start with 'prior pdf' $\pi(\theta)$, this reflects degree of belief about θ before doing the experiment.

Our experiment has data x, $\rightarrow$ likelihood function $L(x|\theta)$.

Bayes' theorem tells how our beliefs should be updated in light of the data x:

$$p(\theta|x) = \frac{L(x|\theta)\pi(\theta)}{\int L(x|\theta')\pi(\theta')\,d\theta'} \propto L(x|\theta)\pi(\theta)$$

Posterior pdf $p(\theta|x)$ contains all our knowledge about θ.

Bayesian method

We need to associate prior probabilities with θ_0 and θ_1, e.g.,

$$\pi(\theta_0, \theta_1) = \pi_0(\theta_0)\,\pi_1(\theta_1)$$

$$\pi_0(\theta_0) = \text{const.}$$

'non-informative', in any case much broader than $L(\theta_0)$

$$\pi_1(\theta_1) = \frac{1}{\sqrt{2\pi}\sigma_{t_1}} e^{-(\theta_1 - t_1)^2/2\sigma_{t_1}^2}$$

$\leftarrow$ based on previous measurement

Putting this into Bayes' theorem gives:

$$p(\theta_0, \theta_1 | \vec{y}) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(y_i - \mu(x_i; \theta_0, \theta_1))^2/2\sigma_i^2}\,\pi_0\,\frac{1}{\sqrt{2\pi}\sigma_{t_1}} e^{-(\theta_1 - t_1)^2/2\sigma_{t_1}^2}$$

posterior $\propto$ likelihood $\times$ prior

Bayesian method (continued)

We then integrate (marginalize) $p(\theta_0, \theta_1 \mid x)$ to find $p(\theta_0 \mid x)$:

$$p(\theta_0|x) = \int p(\theta_0, \theta_1|x)\, d\theta_1 \ .$$

In this example we can do the integral (rare). We find

$$p(\theta_0|x) \ = \ \frac{1}{\sqrt{2\pi}\sigma_{\theta_0}} e^{-(\theta_0 - \hat{\theta}_0)^2/2\sigma_{\theta_0}^2} \quad \text{with}$$

$$\hat{\theta}_0 \ = \ \text{same as ML estimator}$$

$$\sigma_{\theta_0} \ = \ \sigma_{\hat{\theta}_0} \ (\text{same as before})$$

Usually need numerical methods (e.g. Markov Chain Monte Carlo) to do integral.

Digression: marginalization with MCMC

Bayesian computations involve integrals like

$$p(\theta_0 | x) = \int p(\theta_0, \theta_1 | x) \, d\theta_1 .$$

often high dimensionality and impossible in closed form, also impossible with 'normal' acceptance-rejection Monte Carlo.

Markov Chain Monte Carlo (MCMC) has revolutionized Bayesian computation.

MCMC (e.g., Metropolis-Hastings algorithm) generates correlated sequence of random numbers:
    cannot use for many applications, e.g., detector MC;
    effective stat. error greater than if all values independent .

Basic idea: sample multidimensional $\vec{\theta}$,
look, e.g., only at distribution of parameters of interest.

MCMC basics: Metropolis-Hastings algorithm

Goal: given an n-dimensional pdf $p(\vec{\theta})$,

generate a sequence of points $\vec{\theta}_1, \vec{\theta}_2, \vec{\theta}_3, \ldots$

1) Start at some point $\vec{\theta}_0$

Proposal density $q(\vec{\theta}; \vec{\theta}_0)$
e.g. Gaussian centred about $\vec{\theta}_0$

2) Generate $\vec{\theta} \sim q(\vec{\theta}; \vec{\theta}_0)$

3) Form Hastings test ratio $\alpha = \min\left[1, \dfrac{p(\vec{\theta})q(\vec{\theta}_0; \vec{\theta})}{p(\vec{\theta}_0)q(\vec{\theta}; \vec{\theta}_0)}\right]$

4) Generate $u \sim \text{Uniform}[0, 1]$

5) If $u \leq \alpha$, $\vec{\theta}_1 = \vec{\theta}$, ← move to proposed point

    else $\vec{\theta}_1 = \vec{\theta}_0$ ← old point repeated

6) Iterate

Metropolis-Hastings (continued)

This rule produces a *correlated* sequence of points (note how each new point depends on the previous one).

For our purposes this correlation is not fatal, but statistical errors larger than if points were independent.

The proposal density can be (almost) anything, but choose so as to minimize autocorrelation. Often take proposal density symmetric: $q(\vec{\theta}; \vec{\theta}_0) = q(\vec{\theta}_0; \vec{\theta})$
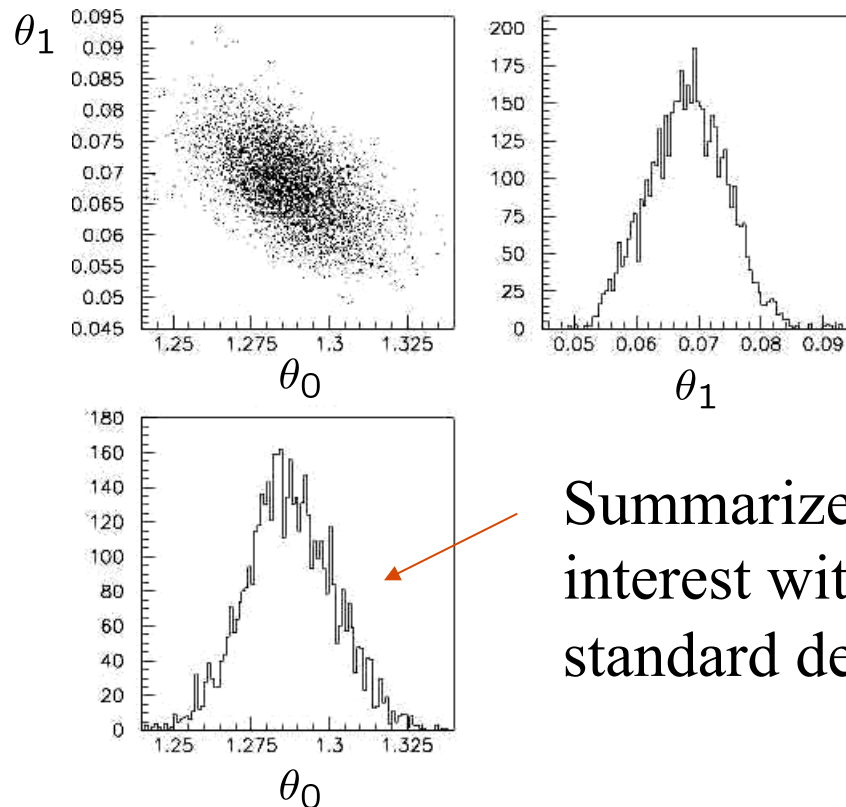
Test ratio is (*Metropolis*-Hastings): $\quad \alpha = \min \left[1, \dfrac{p(\vec{\theta})}{p(\vec{\theta}_0)} \right]$

I.e. if the proposed step is to a point of higher $p(\vec{\theta})$, take it; if not, only take the step with probability $p(\vec{\theta})/p(\vec{\theta}_0)$.

If proposed step rejected, hop in place.

Example: posterior pdf from MCMC

Sample the posterior pdf from previous example with MCMC:

Summarize pdf of parameter of interest with, e.g., mean, median, standard deviation, etc.

Although numerical values of answer here same as in frequentist case, interpretation is different (sometimes unimportant?)

Bayesian method with alternative priors

Suppose we don't have a previous measurement of θ_1 but rather, e.g., a theorist says it should be positive and not too much greater than 0.1 "or so", i.e., something like

$$\pi_1(\theta_1) = \frac{1}{\tau}e^{-\theta_1/\tau}\ , \quad \theta_1 \geq 0\ , \quad \tau = 0.1\ .$$

From this we obtain (numerically) the posterior pdf for θ_0:
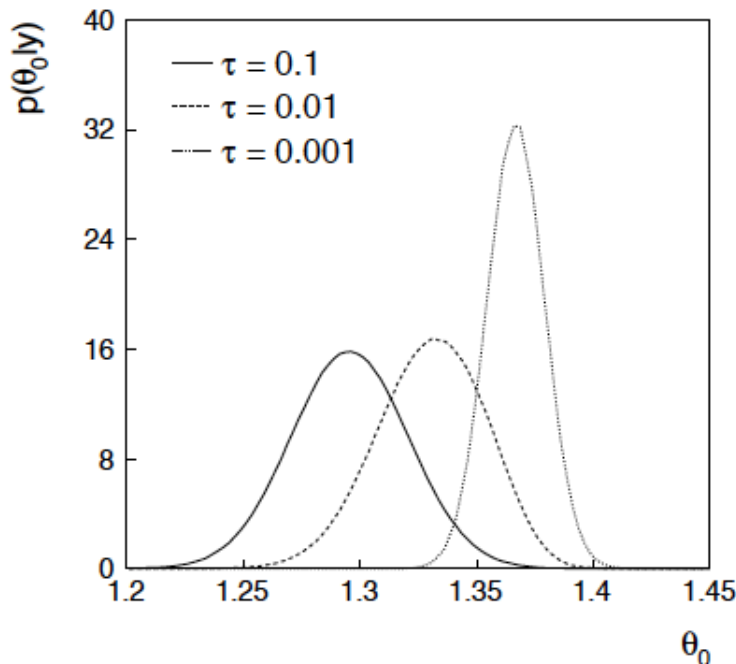
This summarizes all knowledge about θ_0.

Look also at result from variety of priors.

Expected discovery significance for counting experiment with background uncertainty

I. Discovery sensitivity for counting experiment with b known:

(a) $\dfrac{s}{\sqrt{b}}$

(b) Profile likelihood ratio test & Asimov:

$$\sqrt{2\left((s+b)\ln\left(1+\frac{s}{b}\right)-s\right)}$$

II. Discovery sensitivity with uncertainty in b, σ_b:

(a) $\dfrac{s}{\sqrt{b+\sigma_b^2}}$

(b) Profile likelihood ratio test & Asimov:

$$\left[2\left((s+b)\ln\left[\frac{(s+b)(b+\sigma_b^2)}{b^2+(s+b)\sigma_b^2}\right]-\frac{b^2}{\sigma_b^2}\ln\left[1+\frac{\sigma_b^2 s}{b(b+\sigma_b^2)}\right]\right)\right]^{1/2}$$

Counting experiment with known background

Count a number of events $n \sim$ Poisson($s+b$), where

s = expected number of events from signal,

b = expected number of background events.

To test for discovery of signal compute p-value of $s = 0$ hypothesis,

$$p = P(n \geq n_{\text{obs}}|b) = \sum_{n=n_{\text{obs}}}^{\infty} \frac{b^n}{n!} e^{-b} = 1 - F_{\chi^2}(2b; 2n_{\text{obs}})$$

Usually convert to equivalent significance: $Z = \Phi^{-1}(1 - p)$ where Φ is the standard Gaussian cumulative distribution, e.g., $Z > 5$ (a 5 sigma effect) means $p < 2.9 \times 10^{-7}$.

To characterize sensitivity to discovery, give expected (mean or median) Z under assumption of a given s.

$s/\sqrt{b}$ for expected discovery significance

For large $s + b$, $n \rightarrow x \sim$ Gaussian(μ, σ), $\mu = s + b$, $\sigma = \sqrt{(s + b)}$.

For observed value x_{obs}, p-value of $s = 0$ is Prob$(x > x_{\text{obs}} \mid s = 0)$,:

$$p_0 = 1 - \Phi \left(\frac{x_{\text{obs}} - b}{\sqrt{b}} \right)$$

Significance for rejecting $s = 0$ is therefore

$$Z_0 = \Phi^{-1}(1 - p_0) = \frac{x_{\text{obs}} - b}{\sqrt{b}}$$

Expected (median) significance assuming signal rate s is

$$\text{median}[Z_0 | s + b] = \frac{s}{\sqrt{b}}$$

Better approximation for significance

Poisson likelihood for parameter *s* is

$$L(s) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

For now no nuisance params.

To test for discovery use profile likelihood ratio:

$$q_0 = \begin{cases} -2\ln\lambda(0) & \hat{s} \geq 0 \,, \\ 0 & \hat{s} < 0 \,. \end{cases} \qquad \lambda(s) = \frac{L(s, \hat{\hat{\boldsymbol{\theta}}}(s))}{L(\hat{s}, \hat{\boldsymbol{\theta}})}$$

So the likelihood ratio statistic for testing *s* = 0 is

$$q_0 = -2\ln\frac{L(0)}{L(\hat{s})} = 2\left(n\ln\frac{n}{b} + b - n\right) \quad \text{for } n > b, \ 0 \text{ otherwise}$$

Approximate Poisson significance (continued)

For sufficiently large $s + b$, (use Wilks' theorem),

$$Z = \sqrt{2\left(n\ln\frac{n}{b} + b - n\right)} \quad \text{for } n > b \text{ and } Z = 0 \text{ otherwise.}$$
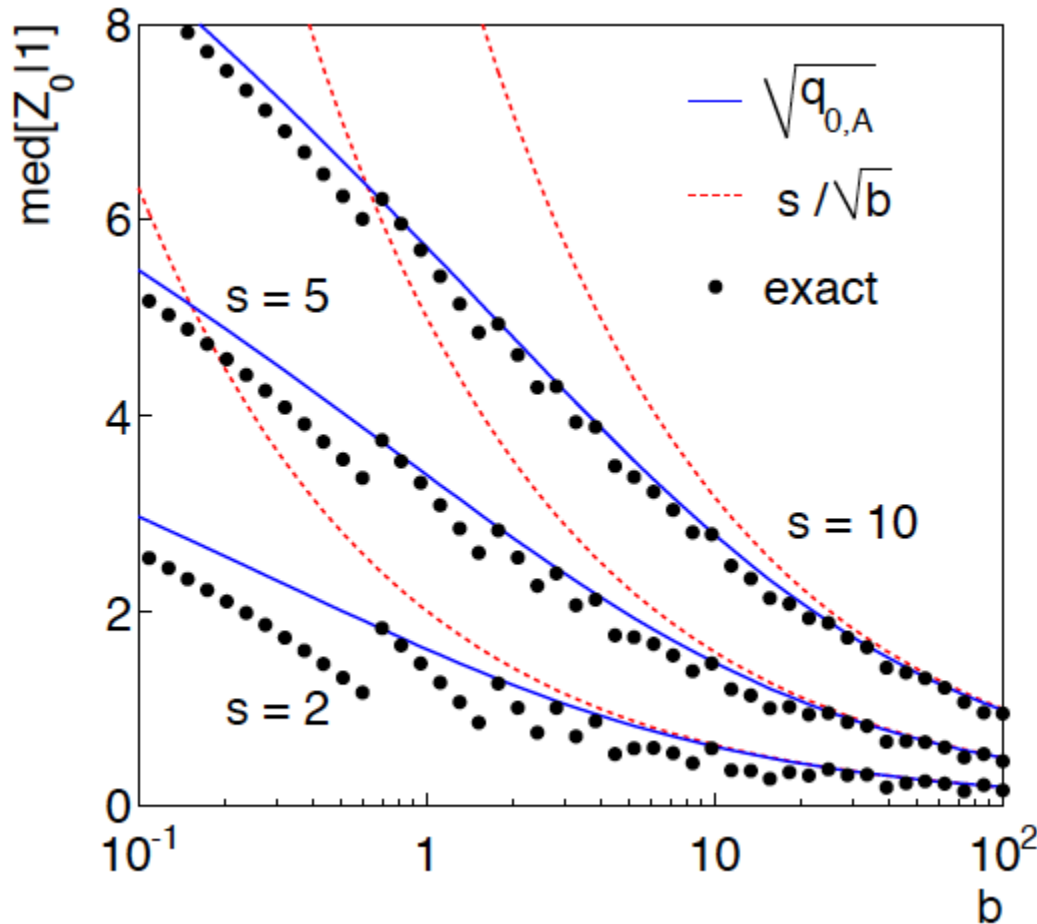
To find median$[Z|s]$, let $n \to s + b$ (i.e., the Asimov data set):

$$Z_{\mathrm{A}} = \sqrt{2\left((s+b)\ln\left(1 + \frac{s}{b}\right) - s\right)}$$

This reduces to $s/\sqrt{b}$ for s << b.

$n \sim \text{Poisson}(s+b)$, median significance, assuming s, of the hypothesis $s = 0$

CCGV, EPJC 71 (2011) 1554, arXiv:1007.1727

"Exact" values from MC, jumps due to discrete data.

Asimov $\sqrt{q_{0,A}}$ good approx. for broad range of s, b.

$s/\sqrt{b}$ only good for $s \ll b$.

Extending $s/\sqrt{b}$ to case where b uncertain

The intuitive explanation of $s/\sqrt{b}$ is that it compares the signal, s, to the standard deviation of n assuming no signal, $\sqrt{b}$.

Now suppose the value of b is uncertain, characterized by a standard deviation σ_b.

A reasonable guess is to replace $\sqrt{b}$ by the quadratic sum of $\sqrt{b}$ and σ_b, i.e.,

$$\mathrm{med}[Z|s] = \frac{s}{\sqrt{b + \sigma_b^2}}$$

This has been used to optimize some analyses e.g. where σ_b cannot be neglected.

Profile likelihood with b uncertain

This is the well studied "on/off" problem: Cranmer 2005; Cousins, Linnemann, and Tucker 2008; Li and Ma 1983,...

Measure two Poisson distributed values:

$n \sim$ Poisson($s+b$)   (primary or "search" measurement)

$m \sim$ Poisson(τb)   (control measurement, τ known)

The likelihood function is

$$L(s,b) = \frac{(s+b)^n}{n!} e^{-(s+b)} \frac{(\tau b)^m}{m!} e^{-\tau b}$$

Use this to construct profile likelihood ratio (b is nuisance parmeter):

$$\lambda(0) = \frac{L(0, \hat{\hat{b}}(0))}{L(\hat{s}, \hat{b})}$$

Ingredients for profile likelihood ratio

To construct profile likelihood ratio from this need estimators:

$$
\hat{s} = n - m/\tau ,
$$

$$
\hat{b} = m/\tau ,
$$

$$
\hat{\hat{b}}(s) = \frac{n + m - (1+\tau)s + \sqrt{(n+m-(1+\tau)s)^2 + 4(1+\tau)sm}}{2(1+\tau)} .
$$

and in particular to test for discovery ($s = 0$),

$$
\hat{\hat{b}}(0) = \frac{n+m}{1+\tau}
$$

Asymptotic significance

Use profile likelihood ratio for q_0, and then from this get discovery significance using asymptotic approximation (Wilks' theorem):

$$Z = \sqrt{q_0}$$

$$= \left[-2 \left(n \ln \left[\frac{n+m}{(1+\tau)n} \right] + m \ln \left[\frac{\tau(n+m)}{(1+\tau)m} \right] \right) \right]^{1/2}$$

for $n > \hat{b}$ and $Z = 0$ otherwise.

Essentially same as in:

Robert D. Cousins, James T. Linnemann and Jordan Tucker, NIM A 595 (2008) 480–501; arXiv:physics/0702156.

Tipei Li and Yuqian Ma, Astrophysical Journal 272 (1983) 317–324.

Asimov approximation for median significance

To get median discovery significance, replace n, m by their expectation values assuming background-plus-signal model:

$$n \to s + b$$

$$m \to \tau b$$

$$Z_A = \left[-2 \left((s+b) \ln \left[\frac{s + (1+\tau)b}{(1+\tau)(s+b)} \right] + \tau b \ln \left[1 + \frac{s}{(1+\tau)b} \right] \right) \right]^{1/2}$$

Or use the variance of $\hat{b} = m/\tau$, $V[\hat{b}] \equiv \sigma_b^2 = \dfrac{b}{\tau}$, to eliminate τ:

$$Z_A = \left[2 \left((s+b) \ln \left[\frac{(s+b)(b+\sigma_b^2)}{b^2 + (s+b)\sigma_b^2} \right] - \frac{b^2}{\sigma_b^2} \ln \left[1 + \frac{\sigma_b^2 s}{b(b+\sigma_b^2)} \right] \right) \right]^{1/2}$$
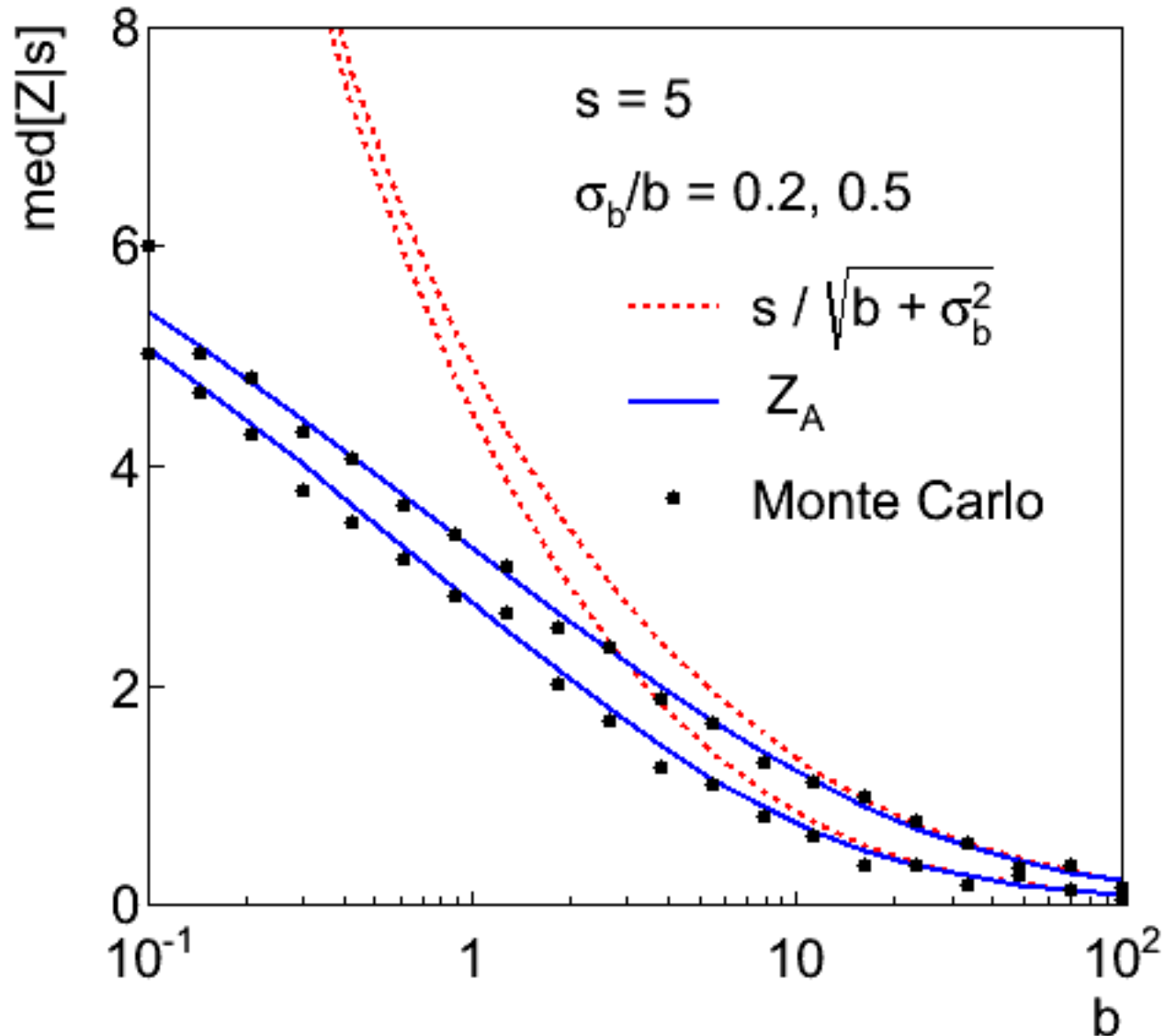
Limiting cases

Expanding the Asimov formula in powers of *s*/*b* and σ_b^2/b (= $1/\tau$) gives

$$Z_A = \frac{s}{\sqrt{b + \sigma_b^2}} \left(1 + \mathcal{O}(s/b) + \mathcal{O}(\sigma_b^2/b) \right)$$

So the "intuitive" formula can be justified as a limiting case of the significance from the profile likelihood ratio test evaluated with the Asimov data set.

Testing the formulae: $s = 5$

$s = 5$

$\sigma_b/b = 0.2, 0.5$

$\cdots\cdots \quad s / \sqrt{b + \sigma_b^2}$

—— Z_A

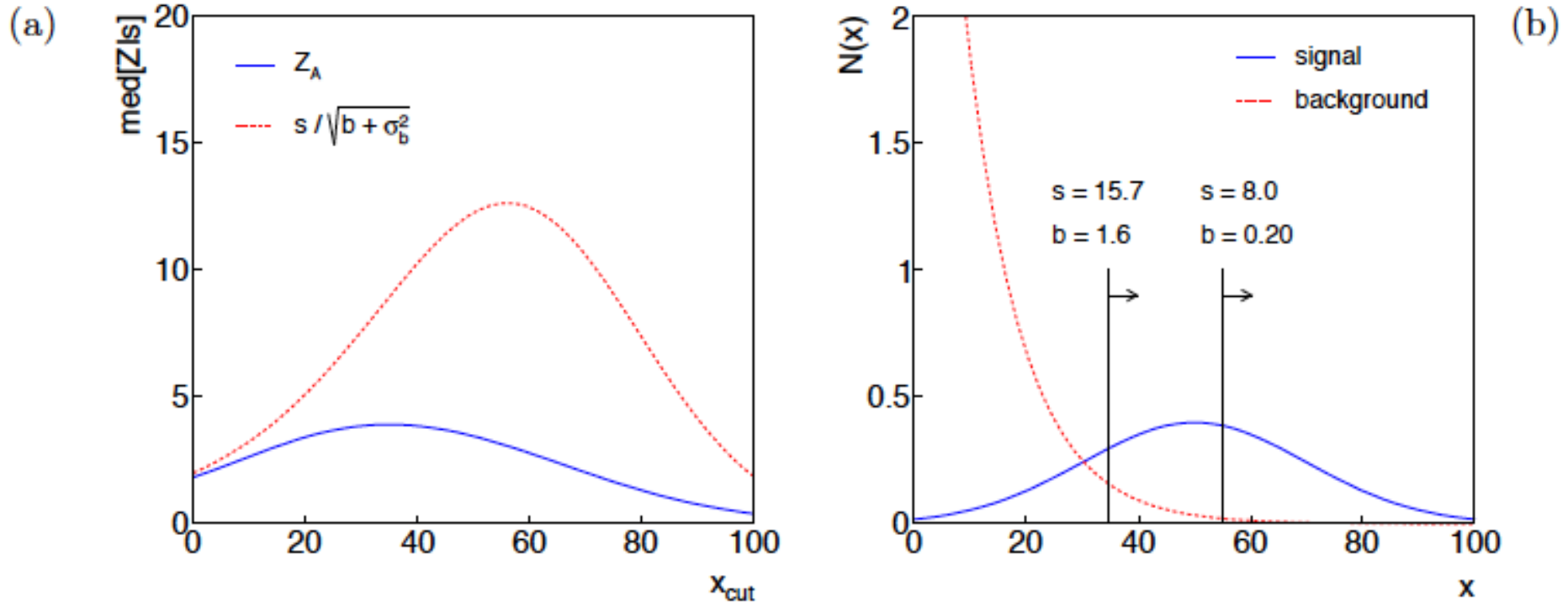■ Monte Carlo

Using sensitivity to optimize a cut

Figure 1: (a) The expected significance as a function of the cut value x_{cut}; (b) the distributions of signal and background with the optimal cut value indicated.

Summary on discovery sensitivity

Simple formula for expected discovery significance based on profile likelihood ratio test and Asimov approximation:

$$Z_A = \left[2 \left((s+b) \ln \left[\frac{(s+b)(b+\sigma_b^2)}{b^2 + (s+b)\sigma_b^2} \right] - \frac{b^2}{\sigma_b^2} \ln \left[1 + \frac{\sigma_b^2 s}{b(b+\sigma_b^2)} \right] \right) \right]^{1/2}$$

For large b, all formulae OK.

For small b, $s/\sqrt{b}$ and $s/\sqrt{(b+\sigma_b^2)}$ overestimate the significance.

Could be important in optimization of searches with low background.

Formula maybe also OK if model is not simple on/off experiment, e.g., several background control measurements (checking this).

Finally

Three lectures only enough for a brief introduction to:

Statistical tests for discovery and limits

Multivariate methods

Bayesian parameter estimation, MCMC

Experimental sensitivity

No time for many important topics

Properties of estimators (bias, variance)

Bayesian approach to discovery (Bayes factors)

The look-elsewhere effect, etc., etc.

Final thought: once the basic formalism is understood, most of the work focuses on writing down the likelihood, e.g., $P(x|q)$, and including in it enough parameters to adequately describe the data (true for both Bayesian and frequentist approaches).

Extra slides

Choice of test for limits (2)

In some cases $\mu = 0$ is no longer a relevant alternative and we want to try to exclude μ on the grounds that some other measure of incompatibility between it and the data exceeds some threshold.

If the measure of incompatibility is taken to be the likelihood ratio with respect to a two-sided alternative, then the critical region can contain both high and low data values.

$\rightarrow$ unified intervals, G. Feldman, R. Cousins,
Phys. Rev. D 57, 3873–3889 (1998)

The Big Debate is whether to use one-sided or unified intervals in cases where small (or zero) values of the parameter are relevant alternatives. Professional statisticians have voiced support on both sides of the debate.

Unified (Feldman-Cousins) intervals

We can use directly

$$t_\mu = -2 \ln \lambda(\mu) \qquad \text{where} \qquad \lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}$$

as a test statistic for a hypothesized μ.

Large discrepancy between data and hypothesis can correspond either to the estimate for μ being observed high or low relative to μ.

This is essentially the statistic used for Feldman-Cousins intervals (here also treats nuisance parameters).

   G. Feldman and R.D. Cousins, Phys. Rev. D 57 (1998) 3873.

Lower edge of interval can be at $\mu = 0$, depending on data.

Distribution of t_μ

Using Wald approximation, $f(t_\mu | \mu')$ is noncentral chi-square for one degree of freedom:

$$f(t_\mu | \mu') = \frac{1}{2\sqrt{t_\mu}} \frac{1}{\sqrt{2\pi}} \left[\exp\left(-\frac{1}{2}\left(\sqrt{t_\mu} + \frac{\mu - \mu'}{\sigma}\right)^2\right) + \exp\left(-\frac{1}{2}\left(\sqrt{t_\mu} - \frac{\mu - \mu'}{\sigma}\right)^2\right) \right]$$

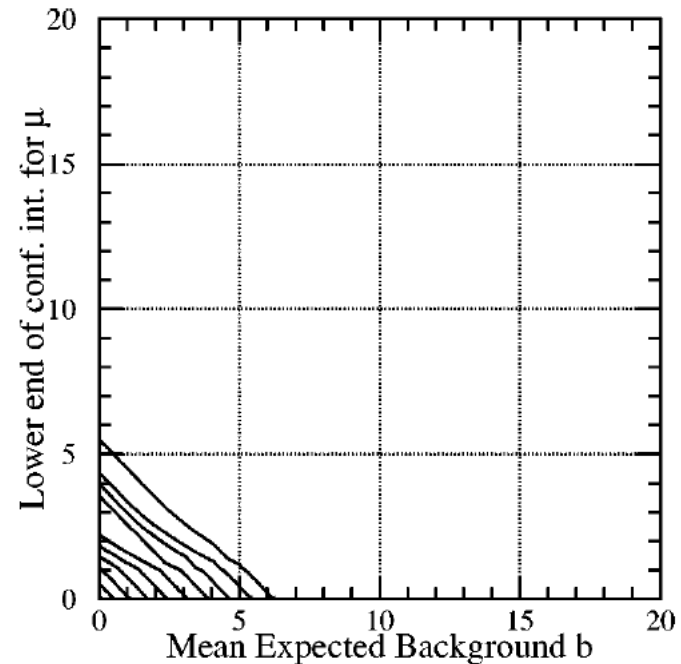Special case of $\mu = \mu'$ is chi-square for one d.o.f. (Wilks).

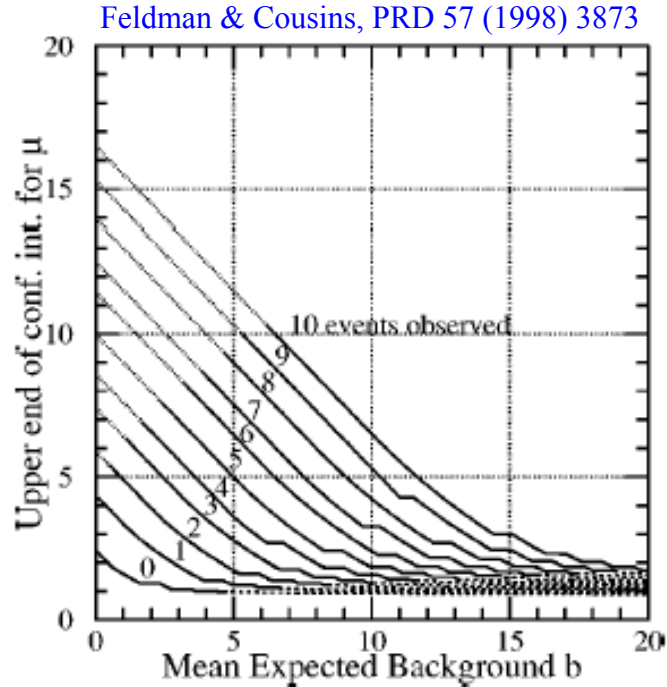The p-value for an observed value of t_μ is

$$p_\mu = 1 - F(t_\mu | \mu) = 2\left(1 - \Phi\left(\sqrt{t_\mu}\right)\right)$$

and the corresponding significance is

$$Z_\mu = \Phi^{-1}(1 - p_\mu) = \Phi^{-1}\left(2\Phi\left(\sqrt{t_\mu}\right) - 1\right)$$

Upper/lower edges of F-C interval for μ versus b for $n \sim$ Poisson($\mu+b$)

Feldman & Cousins, PRD 57 (1998) 3873

Lower edge may be at zero, depending on data.

For $n = 0$, upper edge has (weak) dependence on b.

Feldman-Cousins discussion

The initial motivation for Feldman-Cousins (unified) confidence intervals was to eliminate null intervals.

The F-C limits are based on a likelihood ratio for a test of μ with respect to the alternative consisting of all other allowed values of μ (not just, say, lower values).

The interval's upper edge is higher than the limit from the one-sided test, and lower values of μ may be excluded as well. A substantial downward fluctuation in the data gives a low (but nonzero) limit.

This means that when a value of μ is excluded, it is because there is a probability α for the data to fluctuate either high or low in a manner corresponding to less compatibility as measured by the likelihood ratio.

The Look-Elsewhere Effect

Suppose a model for a mass distribution allows for a peak at a mass m with amplitude μ.

The data show a bump at a mass m_0.
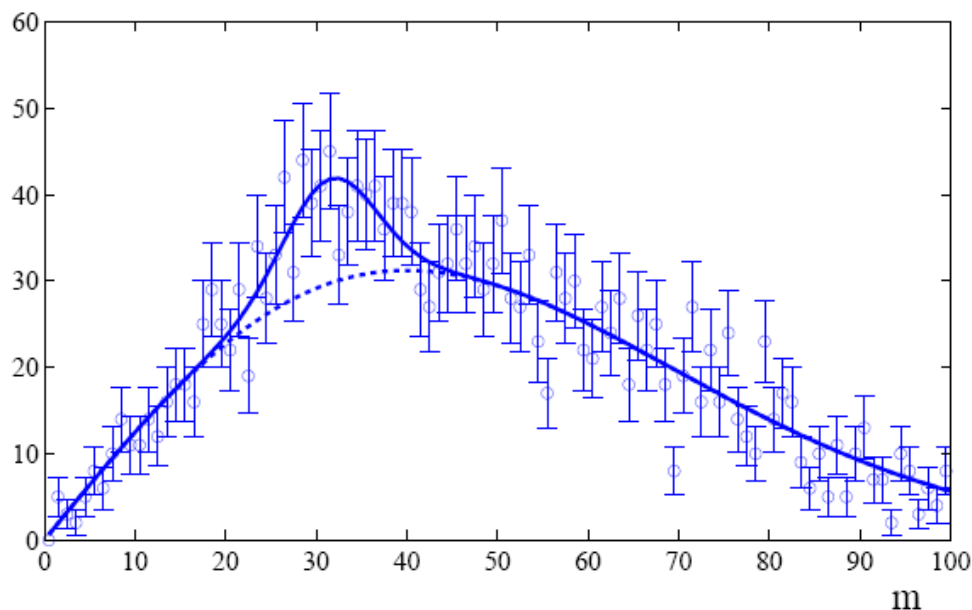
How consistent is this with the no-bump ($\mu = 0$) hypothesis?

Local *p*-value

First, suppose the mass m_0 of the peak was specified a priori.

Test consistency of bump with the no-signal ($\mu = 0$) hypothesis with e.g. likelihood ratio

$$t_{\text{fix}} = -2 \ln \frac{L(0, m_0)}{L(\hat{\mu}, m_0)}$$

where "fix" indicates that the mass of the peak is fixed to m_0.

The resulting *p*-value

$$p_{\text{local}} = \int_{t_{\text{fix,obs}}}^{\infty} f(t_{\text{fix}}|0) \, dt_{\text{fix}}$$

gives the probability to find a value of t_{fix} at least as great as observed at the specific mass m_0 and is called the local *p*-value.

Global *p*-value

But suppose we did not know where in the distribution to expect a peak.

What we want is the probability to find a peak at least as significant as the one observed anywhere in the distribution.

Include the mass as an adjustable parameter in the fit, test significance of peak using

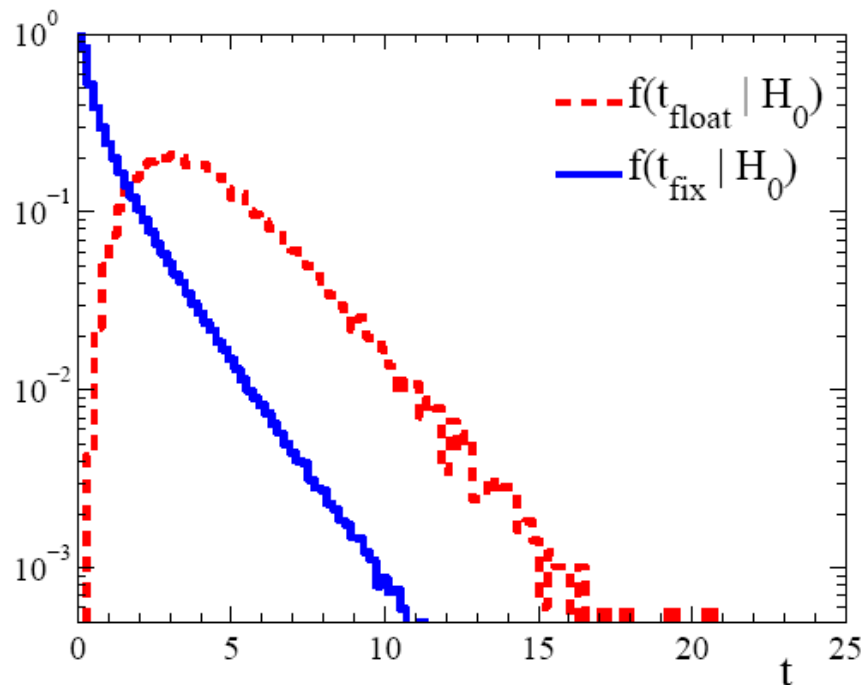$$t_{\text{float}} = -2 \ln \frac{L(0)}{L(\hat{\mu}, \hat{m})}$$

(Note *m* does not appear in the $\mu = 0$ model.)

$$p_{\text{global}} = \int_{t_{\text{float,obs}}}^{\infty} f(t_{\text{float}}|0)\, dt_{\text{float}}$$

Distributions of t_{fix}, t_{float}

For a sufficiently large data sample, t_{fix} ~chi-square for 1 degree of freedom (Wilks' theorem).

For t_{float} there are two adjustable parameters, μ and m, and naively Wilks theorem says t_{float} ~ chi-square for 2 d.o.f.

In fact Wilks' theorem does not hold in the floating mass case because on of the parameters (m) is not-defined in the $\mu = 0$ model.

So getting t_{float} distribution is more difficult.

Approximate correction for LEE

We would like to be able to relate the p-values for the fixed and floating mass analyses (at least approximately).

Gross and Vitells show the p-values are approximately related by

$$p_{\text{global}} \approx p_{\text{local}} + \langle N(c) \rangle$$

where $\langle N(c) \rangle$ is the mean number "upcrossings" of $t_{\text{fix}} = -2\ln \lambda$ in the fit range based on a threshold

$$c = t_{\text{fix,obs}} = Z^2_{\text{local}}$$

and where $Z_{\text{local}} = \Phi^{-1}(1 - p_{\text{local}})$ is the local significance.

So we can either carry out the full floating-mass analysis (e.g. use MC to get p-value), or do fixed mass analysis and apply a correction factor (much faster than MC).
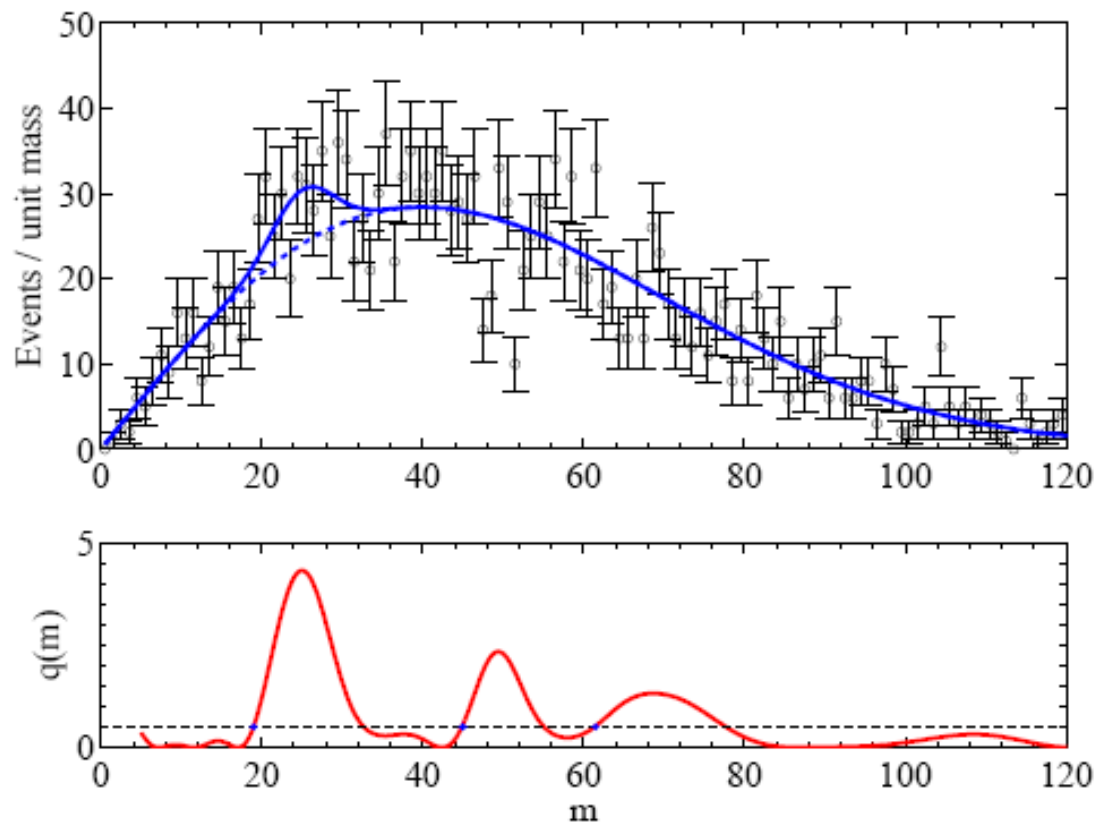
Upcrossings of $-2\ln L$

The Gross-Vitells formula for the trials factor requires $\langle N(c) \rangle$, the mean number "upcrossings" of $t_{\text{fix}} = -2\ln \lambda$ in the fit range based on a threshold $c = t_{\text{fix}} = Z_{\text{fix}}^2$.

$\langle N(c) \rangle$ can be estimated from MC (or the real data) using a much lower threshold c_0:

$$\langle N(c) \rangle \approx \langle N(c_0) \rangle e^{-(c-c_0)/2}$$

In this way $\langle N(c) \rangle$ can be estimated without need of large MC samples, even if the the threshold c is quite high.

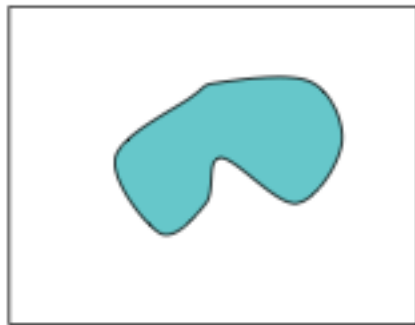# Multidimensional look-elsewhere effect

Generalization to multiple dimensions: number of upcrossings replaced by expectation of Euler characteristic:

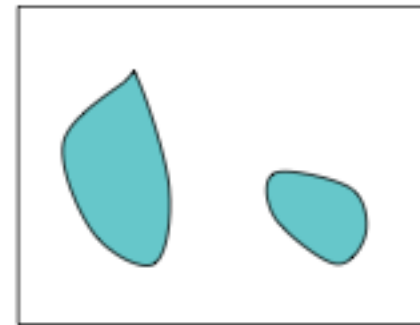$$E[\varphi(A_u)] = \sum_{d=0}^{n} \mathcal{N}_d \rho_d(u)$$

○ Number of disconnected components minus number of 'holes'

φ=1   φ=0   φ=2

Applications: astrophysics (coordinates on sky), search for resonance of unknown mass and width, ...

Summary on Look-Elsewhere Effect

Remember the Look-Elsewhere Effect is when we test a single model (e.g., SM) with multiple observations, i..e, in mulitple places.

Note there is no look-elsewhere effect when considering exclusion limits. There we test specific signal models (typically once) and say whether each is excluded.

With exclusion there is, however, the also problematic issue of testing many signal models (or parameter values) and thus excluding some for which one has little or no sensitivity.

Approximate correction for LEE should be sufficient, and one should also report the uncorrected significance.

"There's no sense in being precise when you don't even know what you're talking about." — John von Neumann

Why 5 sigma?

Common practice in HEP has been to claim a discovery if the p-value of the no-signal hypothesis is below 2.9×10^{-7}, corresponding to a significance $Z = \Phi^{-1}(1 - p) = 5$ (a 5σ effect).

There a number of reasons why one may want to require such a high threshold for discovery:

The "cost" of announcing a false discovery is high.

Unsure about systematics.

Unsure about look-elsewhere effect.

The implied signal may be a priori highly improbable (e.g., violation of Lorentz invariance).

Why 5 sigma (cont.)?

But the primary role of the *p*-value is to quantify the probability that the background-only model gives a statistical fluctuation as big as the one seen or bigger.

It is not intended as a means to protect against hidden systematics or the high standard required for a claim of an important discovery.

In the processes of establishing a discovery there comes a point where it is clear that the observation is not simply a fluctuation, but an "effect", and the focus shifts to whether this is new physics or a systematic.

Providing LEE is dealt with, that threshold is probably closer to 3σ than 5σ.

Bayesian model selection ('discovery')

In Bayesian statistics, all information about a hypothesis H given data x is encapsulated in the posterior probability, $P(H|x)$.

If the model contains internal parameters $\boldsymbol{\theta}$, then to find the probability of H for any $\boldsymbol{\theta}$, marginalize over $\boldsymbol{\theta}$:

$$P(H|\mathbf{x}) = \int P(H, \boldsymbol{\theta}|\mathbf{x}) \, d\boldsymbol{\theta}$$

From Bayes' theorem, the probability $P(H, \boldsymbol{\theta}|x)$ is related to the likelihood $P(x|H, \boldsymbol{\theta})$ by

$$P(H, \boldsymbol{\theta}|\mathbf{x}) = \frac{P(\mathbf{x}|H, \boldsymbol{\theta})\pi(H, \boldsymbol{\theta})}{P(\mathbf{x})} = \frac{P(\mathbf{x}|H, \boldsymbol{\theta})\pi(H, \boldsymbol{\theta})}{\sum_i \int P(\mathbf{x}|H_i, \boldsymbol{\theta}_i)\pi(H_i, \boldsymbol{\theta}_i) \, d\boldsymbol{\theta}_i}$$

In the denominator on the right, the sum is over all possible models H_i and each is integrated over its internal parameters $\boldsymbol{\theta}_i$.

Bayes factors

In practice it may not be possible to carry out the sum over hypotheses for $P(x)$ because we can't enumerate all possible models and assign to them meaningful prior probabilities.

The situation is simpler if we compare any pair of models and ask to what degree the data favour one over the other. This can be done with a quantity called the Bayes factor.

First write the joint prior probability as $\pi(H, \boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|H)\,\pi(H)$, where $\pi(\boldsymbol{\theta}|H)$ is the conditional prior probability of $\boldsymbol{\theta}$ given H, and $\pi(H)$ is the overall prior of H (i.e., for any $\boldsymbol{\theta}$).

Therefore the posterior probability of H is

$$P(H|\mathbf{x}) = \frac{\int P(\mathbf{x}|H, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|H)\pi(H)\,d\boldsymbol{\theta}}{P(\mathbf{x})}$$

Bayes factors (2)

Now suppose there are two models H_i and H_j, each with a set of internal parameters $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$.

The posterior odds ratio is

$$\frac{P(H_i|\mathbf{x})}{P(H_j|\mathbf{x})} = \frac{\int P(\mathbf{x}|H_i, \boldsymbol{\theta}_i)\pi(\boldsymbol{\theta}_i|H_i)\,d\boldsymbol{\theta}_i}{\int P(\mathbf{x}|H_j, \boldsymbol{\theta}_j)\pi(\boldsymbol{\theta}_j|H_j)\,d\boldsymbol{\theta}_j} \times \frac{\pi(H_i)}{\pi(H_j)}$$

posterior odds   Bayes factor B_{ij} prior odds

That is, the Bayes factor is what the posterior odds would be if we would take the overall prior odds to be one.

Note the Bayes factor is independent of the overall priors, but still needs the conditional priors for the models' internal parameters.

Assessing Bayes factors

One can use the Bayes factor much like a p-value (or Z value).

The Jeffreys scale, analogous to HEP's 5σ rule:

B_{10}	Evidence against H_0
1 to 3	Not worth more than a bare mention
3 to 20	Positive
20 to 150	Strong
> 150	Very strong

Kass and Raftery, *Bayes Factors*, J. Am Stat. Assoc 90 (1995) 773.

Numerical determination of Bayes factors

Both numerator and denominator of B_{ij} are of the form

$$m = \int L(\vec{x}|\vec{\theta})\pi(\vec{\theta})\,d\vec{\theta} \quad \longleftarrow \quad \text{'marginal likelihood'}$$

Various ways to compute these, e.g., using sampling of the posterior pdf (which we can do with MCMC).

Harmonic Mean (and improvements)
Importance sampling
Parallel tempering (~thermodynamic integration)
Nested Samplying (MultiNest), ...

Kass and Raftery, *Bayes Factors*, J. Am. Stat. Assoc. 90 (1995) 773-795.

Cong Han and Bradley Carlin, *Markov Chain Monte Carlo Methods for Computing Bayes Factors: A Comparative Review*, J. Am. Stat. Assoc. 96 (2001) 1122-1132.

Phil Gregory, *Bayesian Logical Data Analysis for the Physical Sciences*, Cambridge University Press, 2005.

Priors for Bayes factors

Note that for Bayes factors (unlike Bayesian limits), the prior $\pi(\boldsymbol{\theta}|H)$ cannot be improper. If it is, the posterior is only defined up to an arbitrary constant, and so the Bayes factor is ill defined

> Possible exception allowed if both models contain *same* improper prior; but having same parameter name (or Greek letter) in both models does not fully justify this step.

If improper prior is made proper e.g. by a cut-off, the Bayes factor will retain a dependence on this cut-off.

In general or Bayes factors, all priors must reflect "meaningful" degrees of uncertainty about the parameters.

Harmonic mean estimator

E.g., consider only one model and write Bayes theorem as:

$$\frac{\pi(\boldsymbol{\theta})}{m} = \frac{p(\boldsymbol{\theta}|\mathbf{x})}{L(\mathbf{x}|\boldsymbol{\theta})}$$

$\pi(\boldsymbol{\theta})$ is normalized to unity so integrate both sides,

posterior
expectation

$$m^{-1} = \int \frac{1}{L(\mathbf{x}|\boldsymbol{\theta})} p(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta} = E_p[1/L]$$

Therefore sample $\boldsymbol{\theta}$ from the posterior via MCMC and estimate m with one over the average of $1/L$ (the harmonic mean of L).

M.A. Newton and A.E. Raftery, *Approximate Bayesian Inference by the Weighted Likelihood Bootstrap*, Journal of the Royal Statistical Society B 56 (1994) 3-48.

Improvements to harmonic mean estimator

The harmonic mean estimator is numerically very unstable; formally infinite variance (!). Gelfand & Dey propose variant:

Rearrange Bayes thm; multiply both sides by arbitrary pdf $f(\boldsymbol{\theta})$:

$$\frac{f(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{x})}{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})} = \frac{f(\boldsymbol{\theta})}{m}$$

Integrate over $\boldsymbol{\theta}$:

$$m^{-1} = \int \frac{f(\boldsymbol{\theta})}{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}p(\boldsymbol{\theta}|\mathbf{x}) = E_p\left[\frac{f(\boldsymbol{\theta})}{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}\right]$$

Improved convergence if tails of $f(\boldsymbol{\theta})$ fall off faster than $L(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$

Note harmonic mean estimator is special case $f(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$.

A.E. Gelfand and D.K. Dey, *Bayesian model choice: asymptotics and exact calculations,* Journal of the Royal Statistical Society B 56 (1994) 501-514.

Importance sampling

Need pdf $f(\boldsymbol{\theta})$ which we can evaluate at arbitrary $\boldsymbol{\theta}$ and also sample with MC.
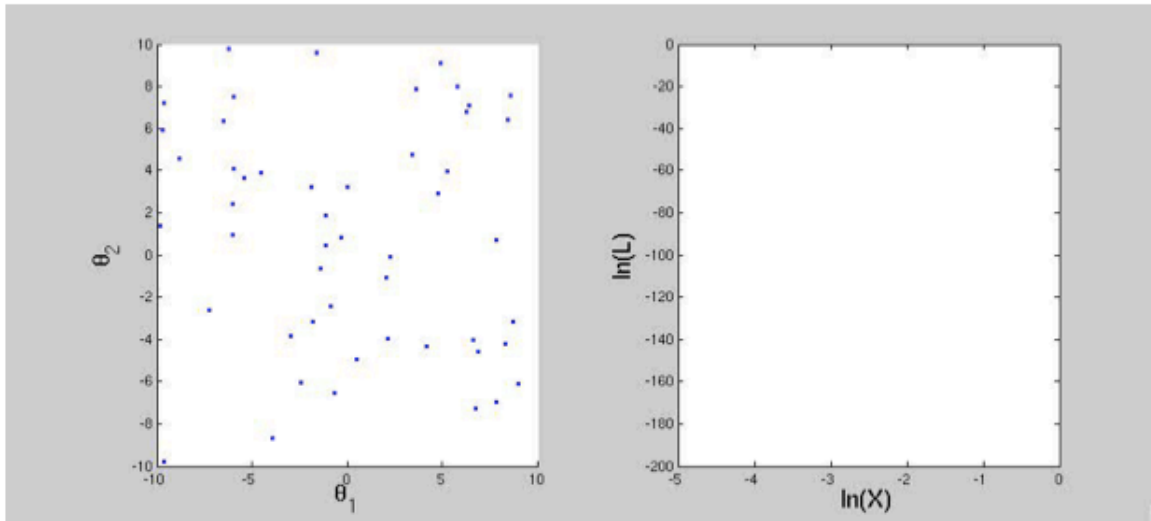
The marginal likelihood can be written

$$m = \int \frac{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{\theta})} f(\boldsymbol{\theta}) \, d\boldsymbol{\theta} = E_f \left[\frac{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{\theta})} \right]$$

Best convergence when $f(\boldsymbol{\theta})$ approximates shape of $L(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

Use for $f(\boldsymbol{\theta})$ e.g. multivariate Gaussian with mean and covariance estimated from posterior (e.g. with MINUIT).

The nested sampling algorithm

(animation courtesy of David Parkinson)

An algorithm originally aimed primarily at the Bayesian evidence computation (Skilling, 2006):

$$X(\lambda) = \int_{\mathcal{L}(\theta) > \lambda} P(\theta) d\theta$$

$$P(d) = \int d\theta \mathcal{L}(\theta) P(\theta) = \int_0^1 X(\lambda) d\lambda$$

Feroz et al (2008), arxiv: 0807.4512, Trotta et al (2008), arxiv: 0809.3792