

Lecture 7: Parameter Estimation and Uncertainty



D. Jason Koskinen
koskinen@nbi.ku.dk

Advanced Methods in Applied Statistics
Feb - Apr 2016

Outline

- Recap in 1D
- Extension to 2D
 - Likelihoods
 - Contours
 - Uncertainties

*Material derived from T. Petersen, D. R. Grant, and G. Cowan

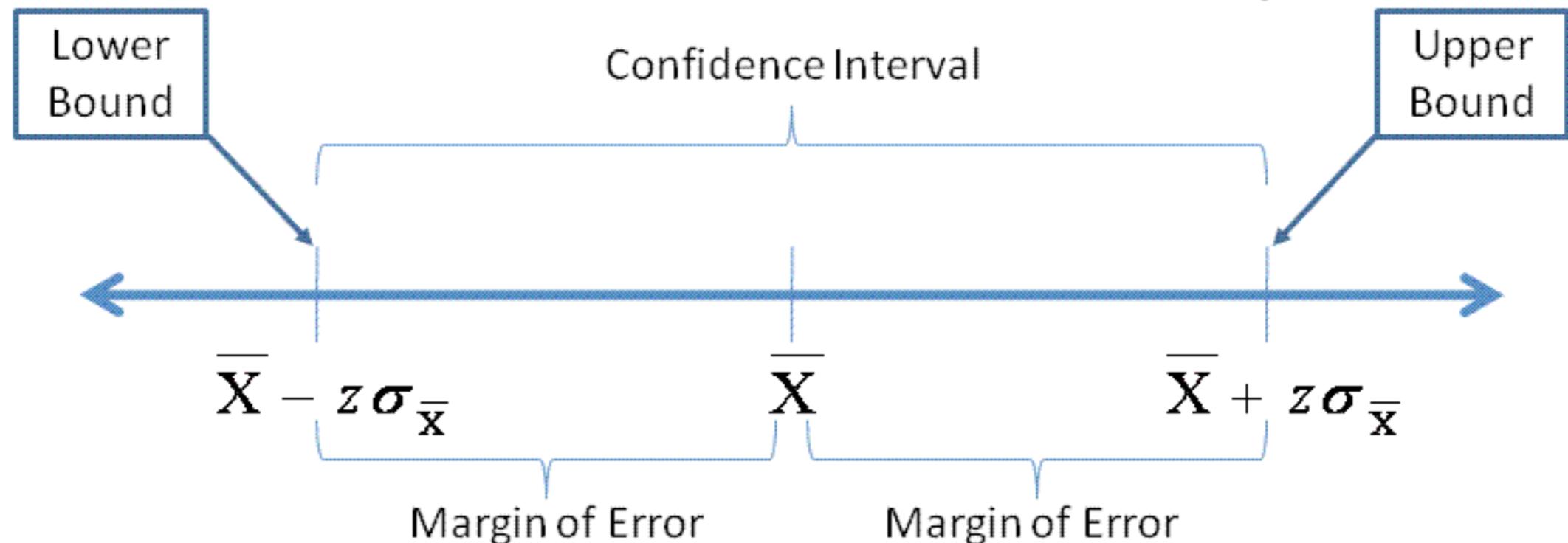
Confidence intervals

“Confidence intervals consist of a range of values (interval) that act as good estimates of the unknown population parameter.”

It is thus a way of giving a range where the true parameter value probably is.

A very simple confidence interval for a Gaussian distribution can be constructed as:
(z denotes the number of sigmas wanted)

$$\bar{x} \pm z \frac{s}{\sqrt{n}}$$



Confidence intervals

Confidence intervals are constructed with a certain **confidence level C**, which is roughly speaking the fraction of times (for many experiments) to have the true parameter fall inside the interval:

$$Prob(x_- \leq x \leq x_+) = \int_{x_-}^{x_+} P(x) dx = C$$

Typically, $C = 95\%$ (thus around 2σ), but 90% and 99% are also used occasionally.

There is a choice as follows:

1. Require symmetric interval (x_+ and x_- are equidistant from μ).
2. Require the shortest interval ($x_+ - x_-$ is a minimum).
3. Require a central interval (integral from x_- to μ is the same as from μ to x_+).

For the Gaussian, the three are equivalent!

Otherwise, 3) is usually used.

Variance of Estimators - Graphical Method

- Used for 1 or 2 parameters when the ML estimate and variance cannot be found analytically. Expand $\ln L$ about its maximum via a Taylor series:

$$\ln L(\theta) = \ln L(\hat{\theta}) + \left(\frac{\partial \ln L}{\partial \theta}\right)_{\theta=\hat{\theta}}(\theta - \hat{\theta}) + \frac{1}{2!} \left(\frac{\partial^2 \ln L}{\partial \theta^2}\right)_{\theta=\hat{\theta}}(\theta - \hat{\theta})^2 + \dots$$

- First term is $\ln L_{\max}$, 2nd term is zero, third term is used for information inequality.

- For 1 parameter:

- plot $\ln L$ as function of the θ and read off the value of $\hat{\theta}$ at the position where L is largest. Sometimes there is more than one peak — take the highest.
- Uncertainty deduced from positions where $\ln L$ is reduced by an amount $1/2$. For a Gaussian Likelihood function:

$$\ln L(\theta) = \ln L_{\max} - \frac{(\theta - \hat{\theta})^2}{2\hat{\sigma}_{\hat{\theta}}^2}$$

$$\ln L(\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}) = \ln L_{\max} - \frac{1}{2} \quad \text{or} \quad \ln L(\hat{\theta} \pm N\hat{\sigma}_{\hat{\theta}}) = \ln L_{\max} - \frac{N^2}{2} \quad \text{For } N \text{ standard deviations}$$

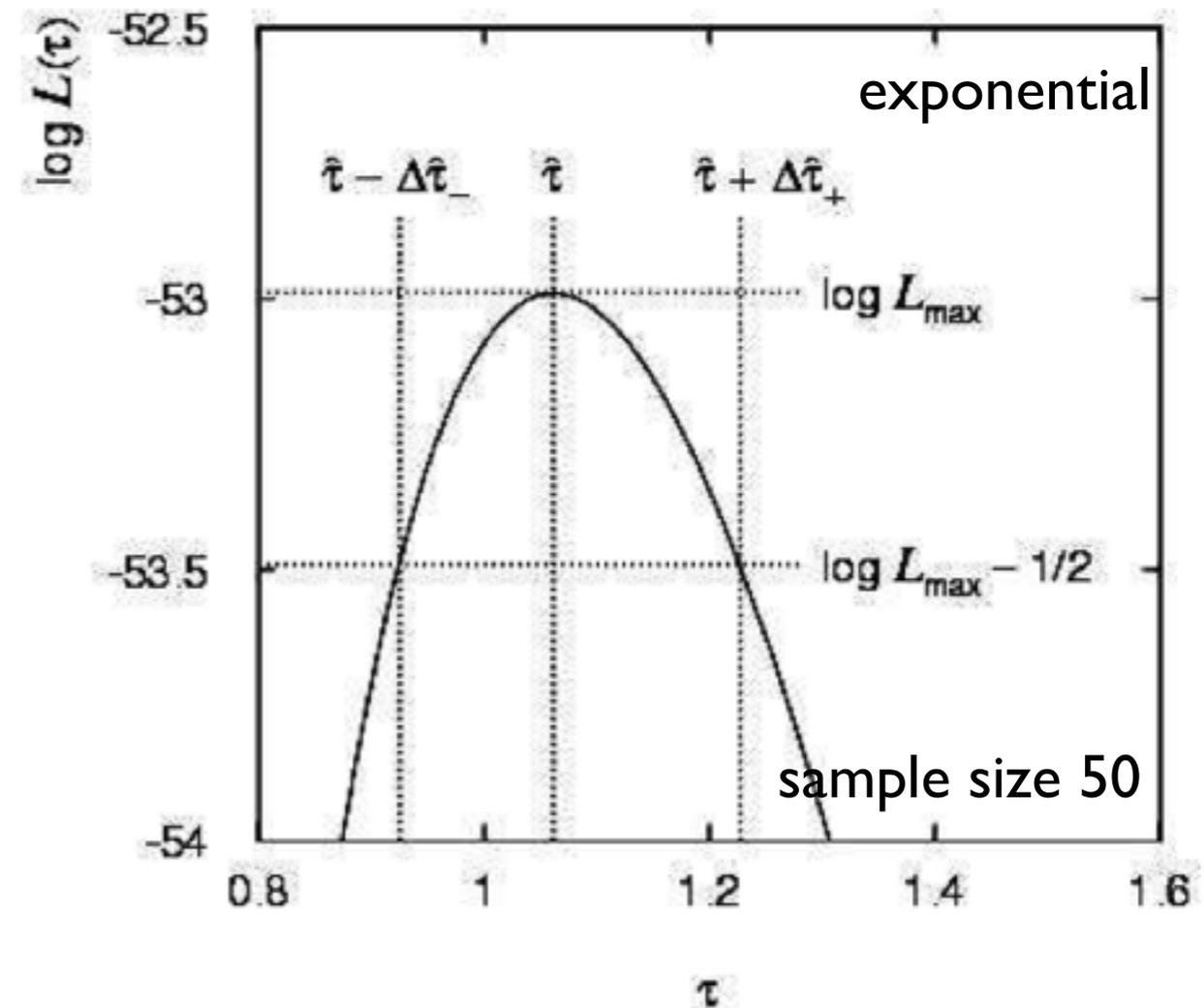
$\ln(\text{Likelihood})$ and $2 \cdot \text{LLH}$

- A change of 1 standard deviation (σ) in the maximum likelihood estimator (MLE) of the parameter θ leads to a decrease in the $\ln(\text{likelihood})$ of $1/2$ for a gaussian distributed estimator
 - Even for a non-gaussian MLE, the 1σ region defined as $\text{LLH}-1/2$ is a good approximation
 - Because the regions defined with $\Delta\text{LLH}=1/2$ are consistent with common χ^2 distributions multiplied by $1/2$, we often calculate the likelihoods as $2 \cdot \text{LLH}$
- Translates to >1 parameters too, with the appropriate change in $2 \cdot \text{LLH}$ confidence values
 - 1 parameter, $\Delta(2\text{LLH})=1$ for 68.3% C.L.
 - 2 parameter, $\Delta(2\text{LLH})=2.3$ for 68.3% C.L.

Variance of Estimators - Graphical

Method

- One Parameter cont.
 - The formula applies for non-Gaussian case, i.e. change variables to $g(\theta)$ which produces a Gaussian distribution. L is invariant under parameter transformation.
 - If the Likelihood function is asymmetric, as happens for small sample size, then an asymmetric interval about the most likely value may result.



$$\hat{\tau} = 1.062$$

$$\Delta\hat{\tau}_- = 0.137$$

$$\Delta\hat{\tau}_+ = 0.165$$

$$\hat{\sigma}_{\hat{\tau}} \approx \Delta\hat{\tau}_- \approx \Delta\hat{\tau}_+ \approx 0.15$$

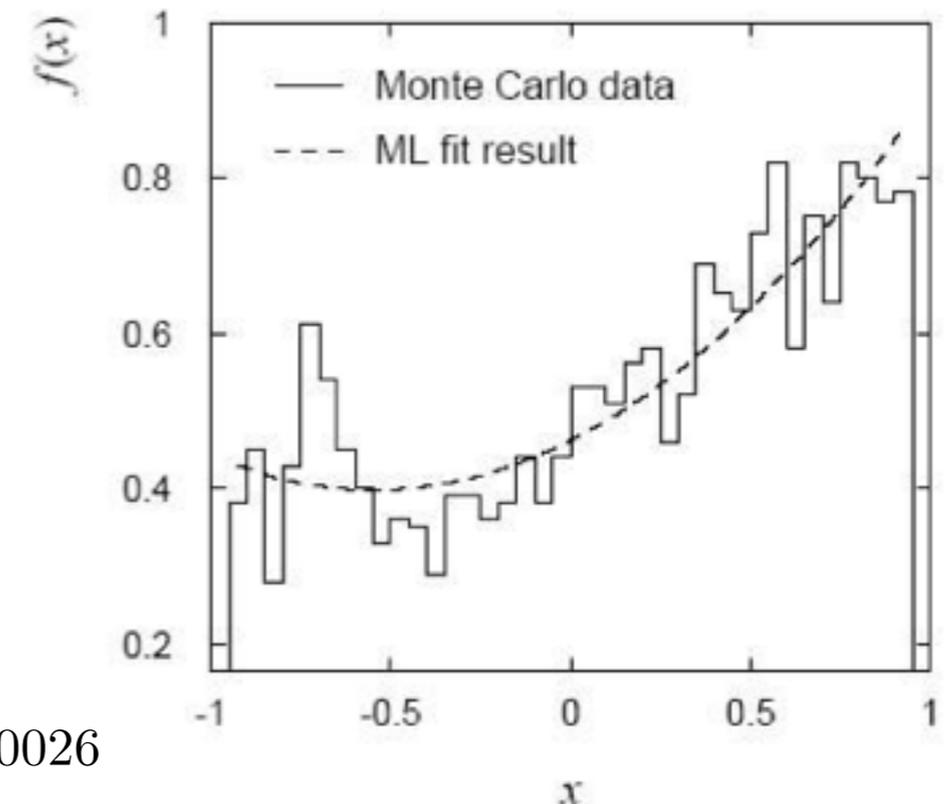
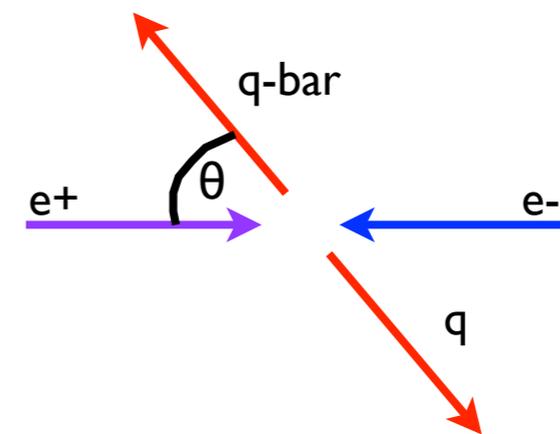
Variance of Estimators - Graphical Method

- Consider an example from scattering with an angular distribution given by $x = \cos\theta$
- if $x_{min} < x < x_{max}$ then the PDF needs to be normalized:

$$f(x; \alpha, \beta) = \frac{1 + \alpha x + \beta x^2}{2 + 2\beta/3} \quad \int_{x_{min}}^{x_{max}} f(x; \alpha, \beta) dx = 1$$

- Take the specific example where $\alpha=0.5$ and $\beta=0.5$ for 2000 points of $-0.95 \leq x \leq 0.95$
- The maximum may be found numerically, giving: $\alpha = 0.508, \beta = 0.47$
- The statistical errors can be estimated by numerically solving the 2nd derivative (shown here for completeness)

$$(V^{-1})_{ij} = -\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j} \Big|_{\vec{\theta}=\hat{\theta}} \quad \hat{\sigma}_{\hat{\alpha}} = 0.052, \quad \hat{\sigma}_{\hat{\beta}} = 0.11, \quad cov[\hat{\alpha}, \hat{\beta}] = 0.0026$$



Exercise #1

- Before we use the LLH values to determine the uncertainties for α and β , let's do it via Monte Carlo first
- Similar to the exercises 2-3 from Lecture 4, the theoretical prediction:

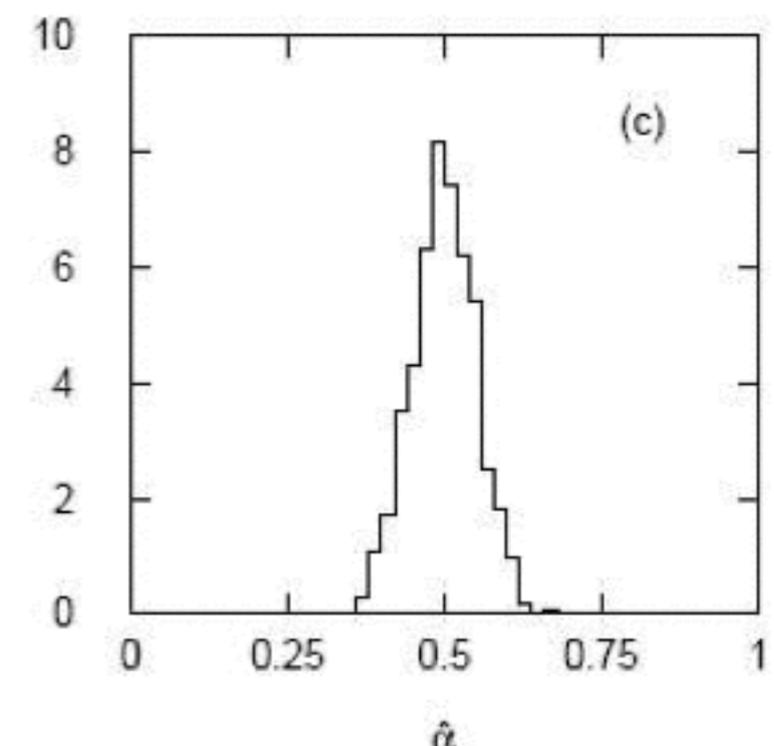
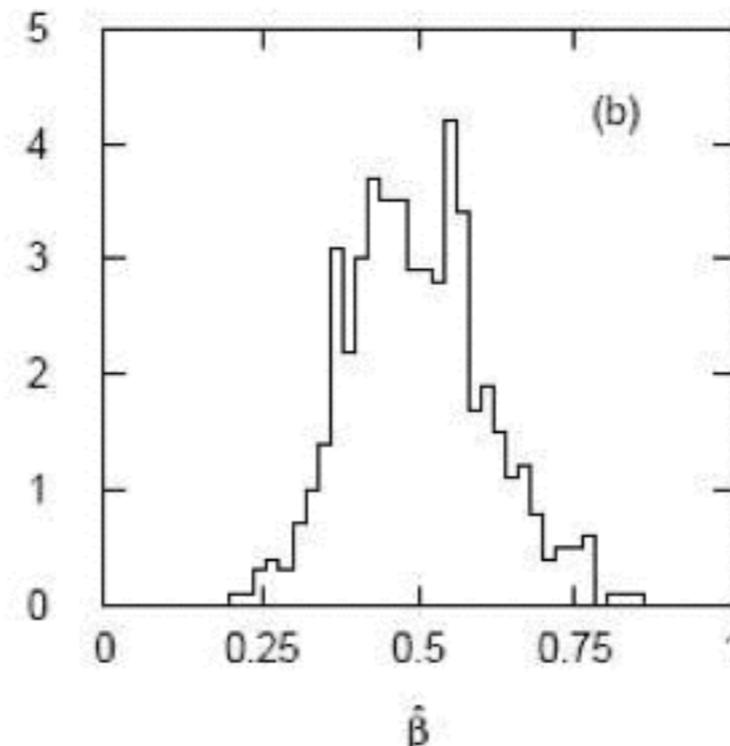
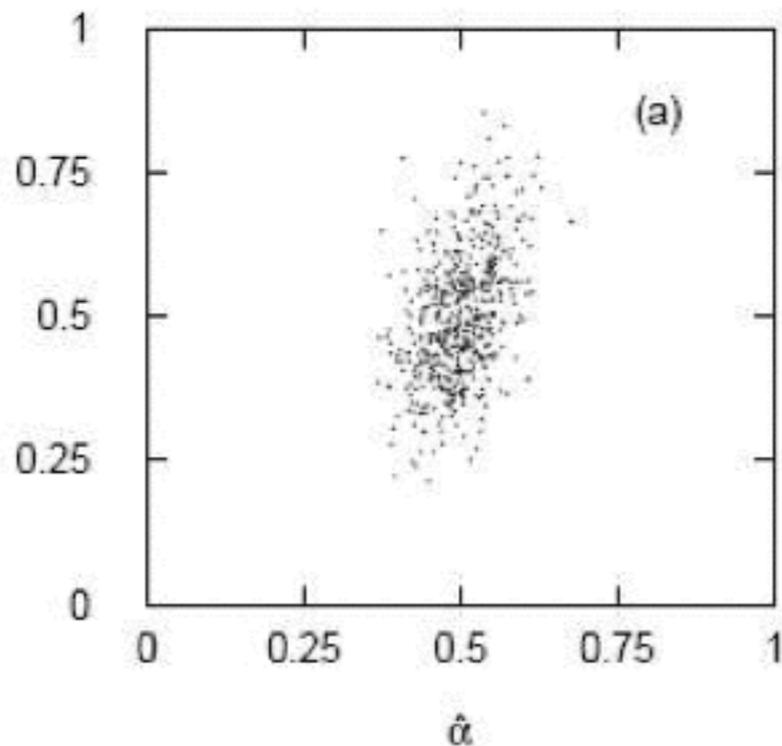
$$f(x; \alpha, \beta) = 1 + \alpha x + \beta x^2$$

- For $\alpha=0.5$ and $\beta=0.5$, generate 2000 Monte Carlo data points using the above function transformed into a PDF over the range $-0.95 \leq x \leq 0.95$
 - Remember to normalize the function properly to convert it to a proper PDF
 - Fit the MLE parameters $\hat{\alpha}$ and $\hat{\beta}$ using a minimizer/maximizer
 - Repeat 100 to 500 times plotting the distributions of $\hat{\alpha}$ and $\hat{\beta}$ as well as $\hat{\alpha}$ vs. $\hat{\beta}$

Exercise #1

- Shown are 500 Monte Carlo pseudo-experiments
- The estimates average to approximately the true values, the variances are close to initial estimates from slide 8 and the marginal pdfs are approximately Gaussian.

$$\begin{aligned}\bar{\hat{\alpha}} &= 0.499 \\ s_{\hat{\alpha}} &= 0.051 \\ \bar{\hat{\beta}} &= 0.498 \\ s_{\hat{\beta}} &= 0.111\end{aligned}$$



Comments

- After finding the best-fit values via $\ln(\text{likelihood})$ maximization/minimization from data, one of THE best and most robust calculations for the parameter uncertainties is to run numerous pseudo-experiments using the best-fit values for the Monte Carlo 'true' values and find out the spread in pseudo-experiment best fit values
 - MLEs don't have to be gaussian, i.e. uncertainty is accurate even if the Central Limit Theorem is invalid for your data/parameters
 - Monte Carlo plus fitting routine will take care of many parameter correlations
 - The problem is that it can be slow and gets exponentially slower with each dimension

Good?

- The LLH maximization/minimization will give the best parameters and often the uncertainty on those parameters. But, likelihood fits do not tell whether the data and the prediction agree.
 - Remember that the likelihood has a form (PDF) that is provided by you and may not be correct.
 - The physics PDF may be okay, but there may be some systematic that is unknown or at least unaccounted for which creates disagreement between the data and the best-fit prediction.
 - Likelihood *ratios* between two hypotheses are a good way to exclude models, and we'll cover hypothesis testing on Thursday.

Goodness-of-fit

- Pearson's Chi-square Test

- A goodness of fit test that could be applied to a histogram of observed values, x , with N bins. For the number of entries in bin i , n_i , and the number of expected entries for the same bin, λ_i , the test statistic becomes:

$$T = \chi^2 = \sum_{i=1}^N \frac{(n_i - \lambda_i)^2}{\lambda_i}$$

- If the data are Poisson distributed, and the number of entries is not too small in each bin (>5), then T follows a chi-square distribution of N degrees of freedom. This is true regardless of the distribution of x , implying the chi-square test is distribution free.
- Even though finding the maximum likelihood estimator (MLE) best-fits are often done using an unbinned likelihood, it is often useful to use histograms to get a (reduced) chi-squared value as a goodness-of-fit parameter

Variance of Estimators - Graphical Method

- Two Parameter Contours

- For $L(x; \theta_1, \theta_2)$ we can plot the contours of constant likelihood in the θ_1, θ_2 plane.
- For large n , $\ln L$ takes a quadratic form near the maximum:

$$\ln L(\alpha, \beta) \approx \ln L_{max} - \frac{1}{2(1 - \rho^2)} \left[\left(\frac{\alpha - \hat{\alpha}}{\sigma_{\hat{\alpha}}} \right)^2 + \left(\frac{\beta - \hat{\beta}}{\sigma_{\hat{\beta}}} \right)^2 - 2\rho \left(\frac{\alpha - \hat{\alpha}}{\sigma_{\hat{\alpha}}} \right) \left(\frac{\beta - \hat{\beta}}{\sigma_{\hat{\beta}}} \right) \right]$$

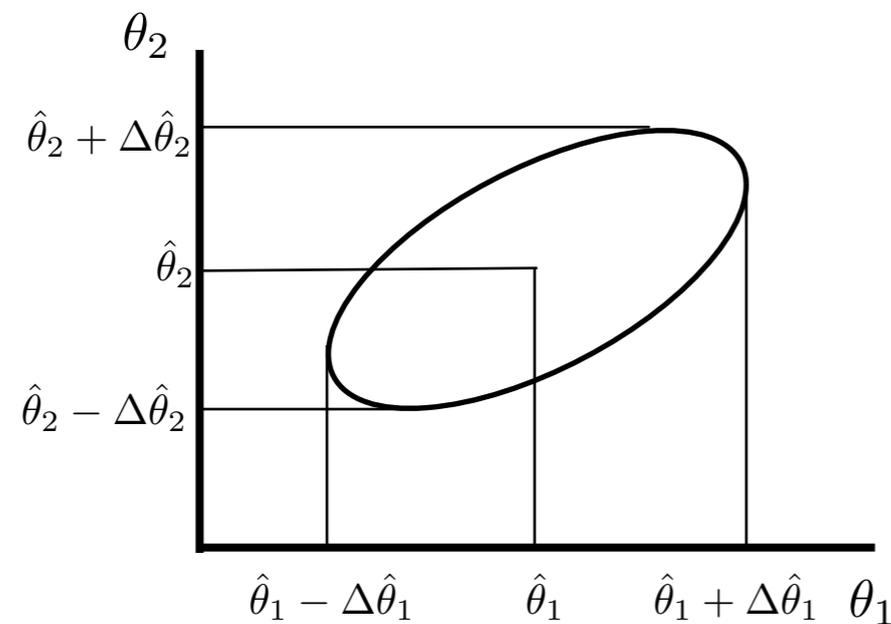
Then the contour given by $\ln L(\alpha, \beta) = \ln L_{max} - 1/2$

$$\frac{1}{(1 - \rho^2)} \left[\left(\frac{\alpha - \hat{\alpha}}{\sigma_{\hat{\alpha}}} \right)^2 + \left(\frac{\beta - \hat{\beta}}{\sigma_{\hat{\beta}}} \right)^2 - 2\rho \left(\frac{\alpha - \hat{\alpha}}{\sigma_{\hat{\alpha}}} \right) \left(\frac{\beta - \hat{\beta}}{\sigma_{\hat{\beta}}} \right) \right] = 1 \quad \text{is an ellipse, distributed as chi-square with 2 dof}$$

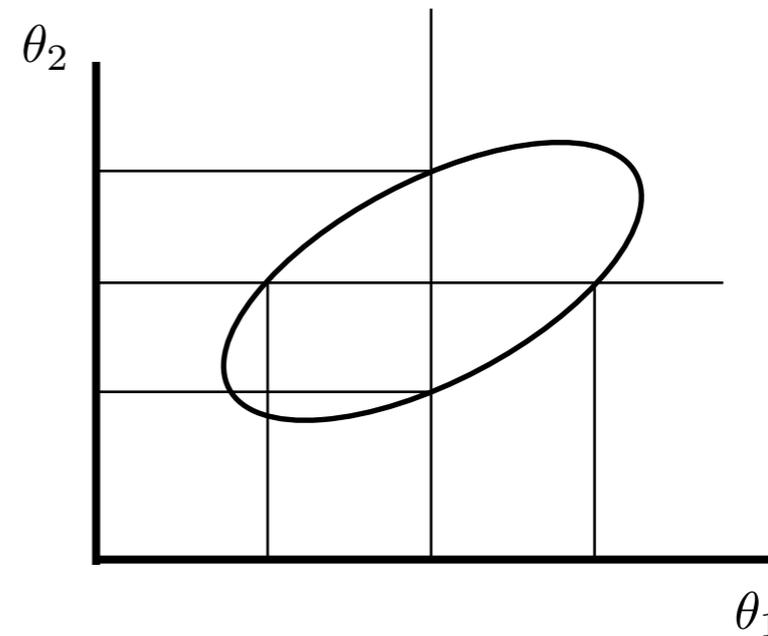
- There is often more than one maximum and if there is no clear peak over the others an additional experiment may be needed to identify which to consider.
- To find the uncertainty we plot the contour with $\ln L = \ln L_{max} - 1/2$ and examine the projection of the contour on the two axes.

Variance of Estimators - Graphical Method

- Two Parameter Contours



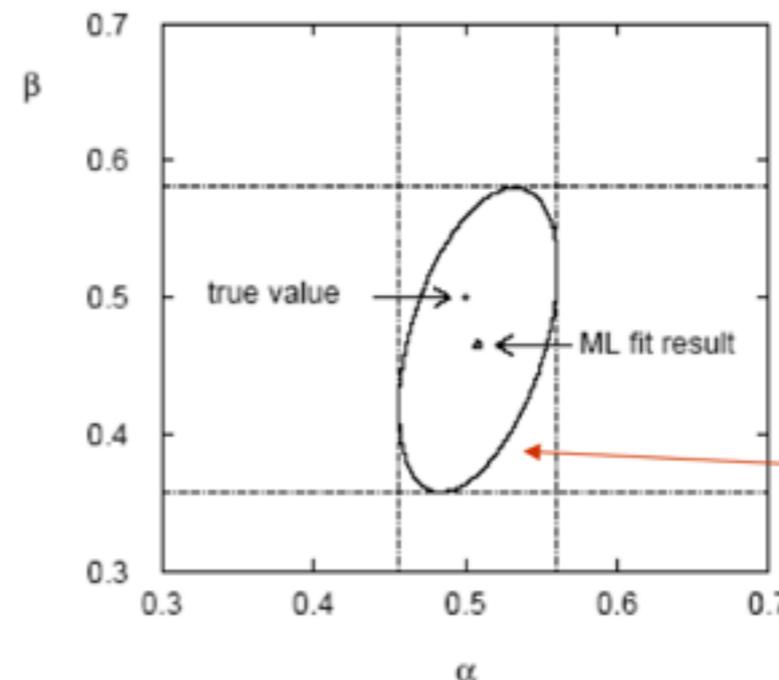
correct



incorrect

- Tangent lines to the contours give the standard deviations.
- Angle of ellipse, ϕ , is related to the correlation:

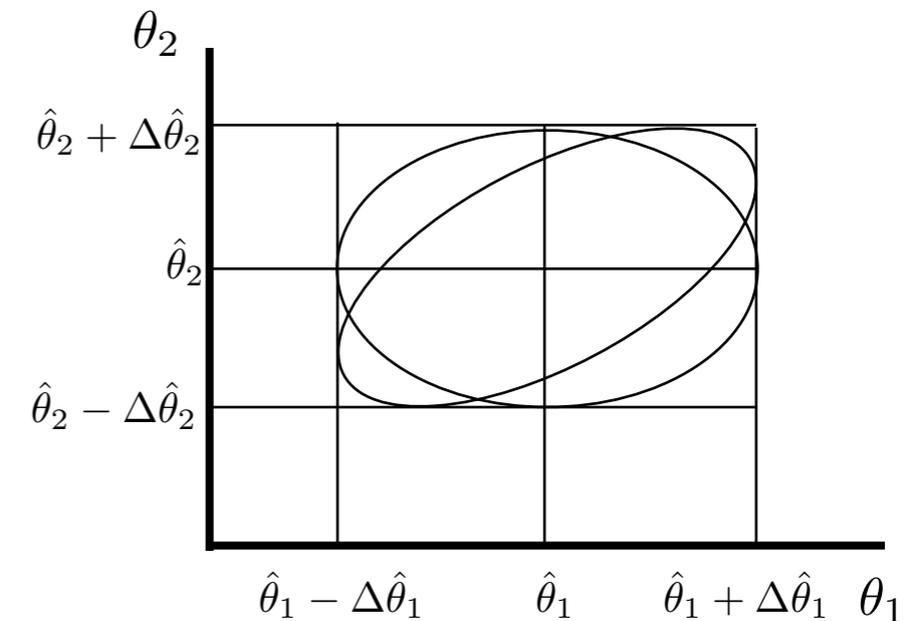
$$\tan 2\phi = \frac{2\rho\sigma_{\hat{\alpha}}\sigma_{\hat{\beta}}}{\sigma_{\hat{\alpha}}^2 - \sigma_{\hat{\beta}}^2}$$



$$\ln L(\alpha, \beta) = \ln L_{max} - 1/2$$

Variance of Estimators - Graphical Method

- Two Parameter Contour
 - When the correct, tangential, method is used then the uncertainties are not dependent on the correlation of the variables.
 - For a 2D Gaussian likelihood function, the probability to be in the error range is 0.683.
 - The probability the ellipses of constant $\ln L = \ln L_{max} - a$ contains the true point, θ_1 and θ_2 , is:



correct

a (1 dof)	a (2 dof)	σ
0.5	1.15	1
2.0	3.09	2
4.5	5.92	3

Variance of Estimators - Graphical Method

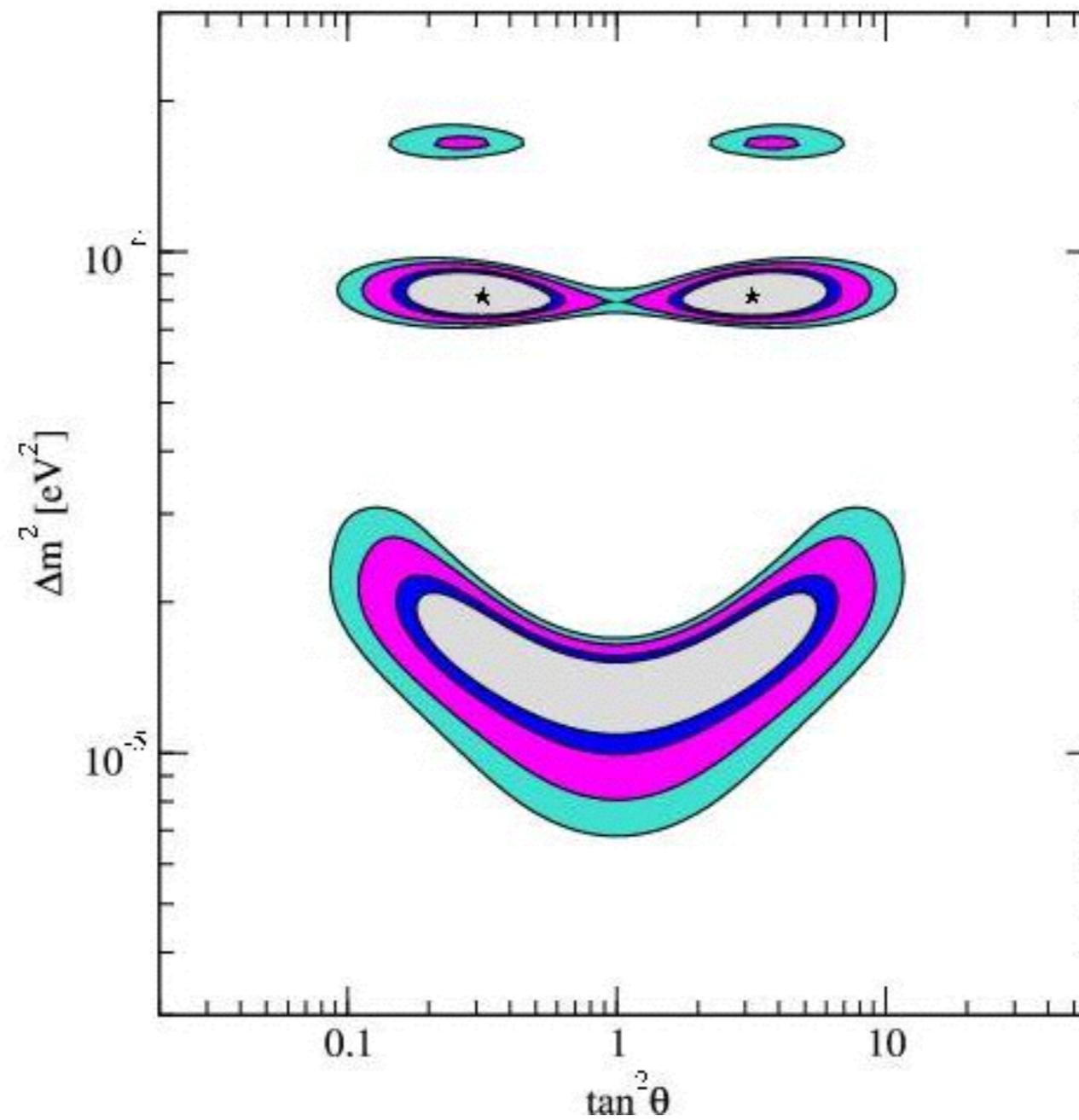
- Two Parameter Contour
 - If the likelihood function contours are very irregular so that a transformation to a 2D Gaussian is not possible, or if the contour consists of more than one closed curve, it is usually better to show the likelihood function contour directly instead of quoting intervals.
 - For three or more parameters, larger samples are necessary to have the likelihood function to be Gaussian
 - A general max/min program will probably be necessary to find the estimate and the uncertainties (ie. MINUIT from CERNLIB, BRENT and POWELL from Numerical Recipes).
 - Example: Region for mean and variance in a normal distribution:

• The joint maximum likelihood estimates of the mean and variance for the normal pdf are:

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_i x_i \qquad \hat{\sigma}^2 = \frac{n-1}{n} s^2 = \frac{1}{n} \sum_i (x_i - \bar{x})^2$$

Best Result Plot?

KamLAND: *"just smiling"*



Variance of Estimators - Graphical Method

- Two Parameter Contour - with provided co-variance matrix

- Example cont.

- with covariance elements given by:

$$c_{11} = \frac{\sigma^2}{n} \qquad c_{22} = \frac{2\sigma^4}{n}$$

- The ellipse which gives a 95% joint likelihood region is:

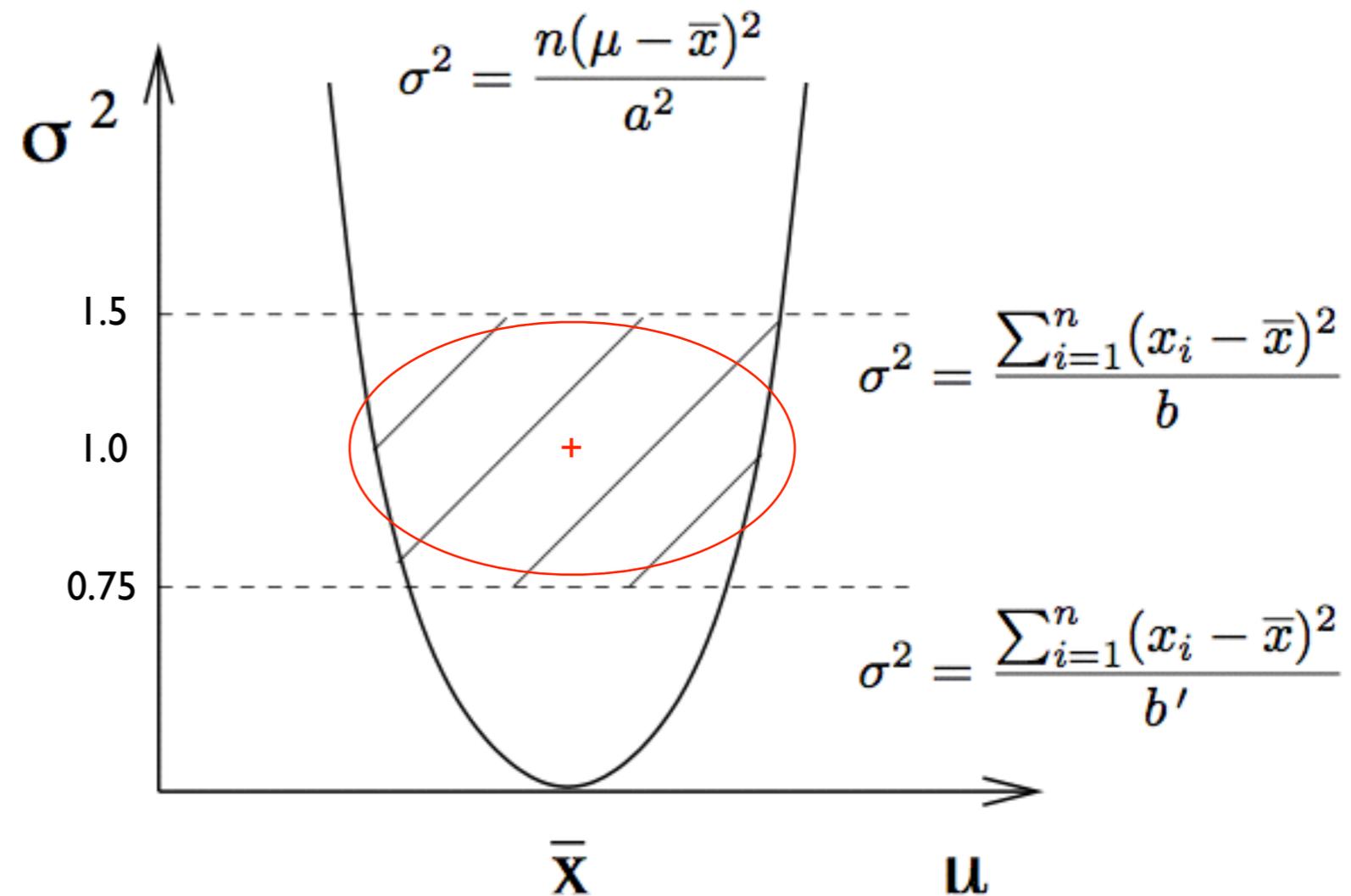
$$Q = (\mu - \hat{\mu})^2 \frac{n}{\hat{\sigma}^2} + (\sigma^2 - \hat{\sigma}^2)^2 \frac{n}{2\hat{\sigma}^4} \qquad Q = 2a \Rightarrow a = -\ln(1 - 0.95) = 2.996$$

- Consider a gaussian sample with $n=100$, sample mean=1 and sample variance=1. We want to find the ellipse or joint likelihood region. Recall the region bounded by a parabola for 2 parameters for the equal tail probability:

$$p = 1/2(1 - \sqrt{0.95})$$

Variance of Estimators - Graphical Method

- Two Parameter Contour
 - The likelihood region (ellipse) and confidence region (intersected parabola) for the 95% CL.
- Note the ellipse is smaller than the confidence region.
 - at $n=100$ one is not yet at the asymptotic limit (>500 usually). Thus, the likelihood region is an approximation of large n .



Variance/Uncertainty - Using LLH

Values

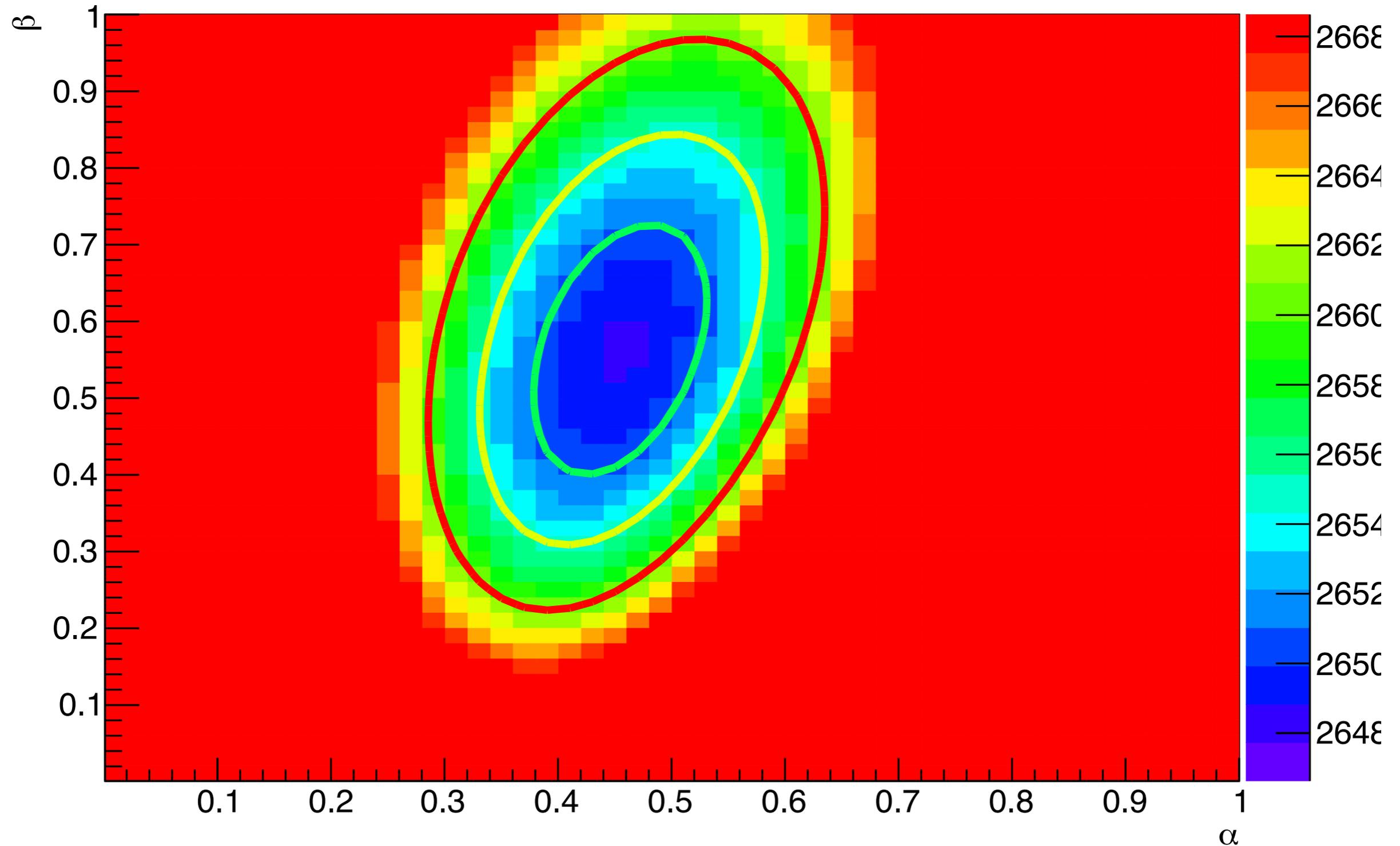
- The LLH (or $-2*LLH$) landscape provides the necessary information to construct 2+ dimensional confidence intervals, provided the respective MLEs are gaussian or well-approximated as gaussian
- Some minimization programs will return the uncertainty on the parameter(s) after finding the best-fit values
 - The `.migrad()` call in `iminuit`
 - It is possible to write your own code to do this as well

Exercise #2

- Using the same function and $\alpha=0.5$ and $\beta=0.5$ as Exercise #1, find the MLE values for a single Monte Carlo sample w/ 2000 points
- Plot the contours related to the 1σ , 2σ , and 3σ confidence regions
 - Remember that this function has 2 fit parameters
 - Because of different random number generators, your result is likely to vary from mine
- Calculate a goodness-of-fit using a reduced chi-squared

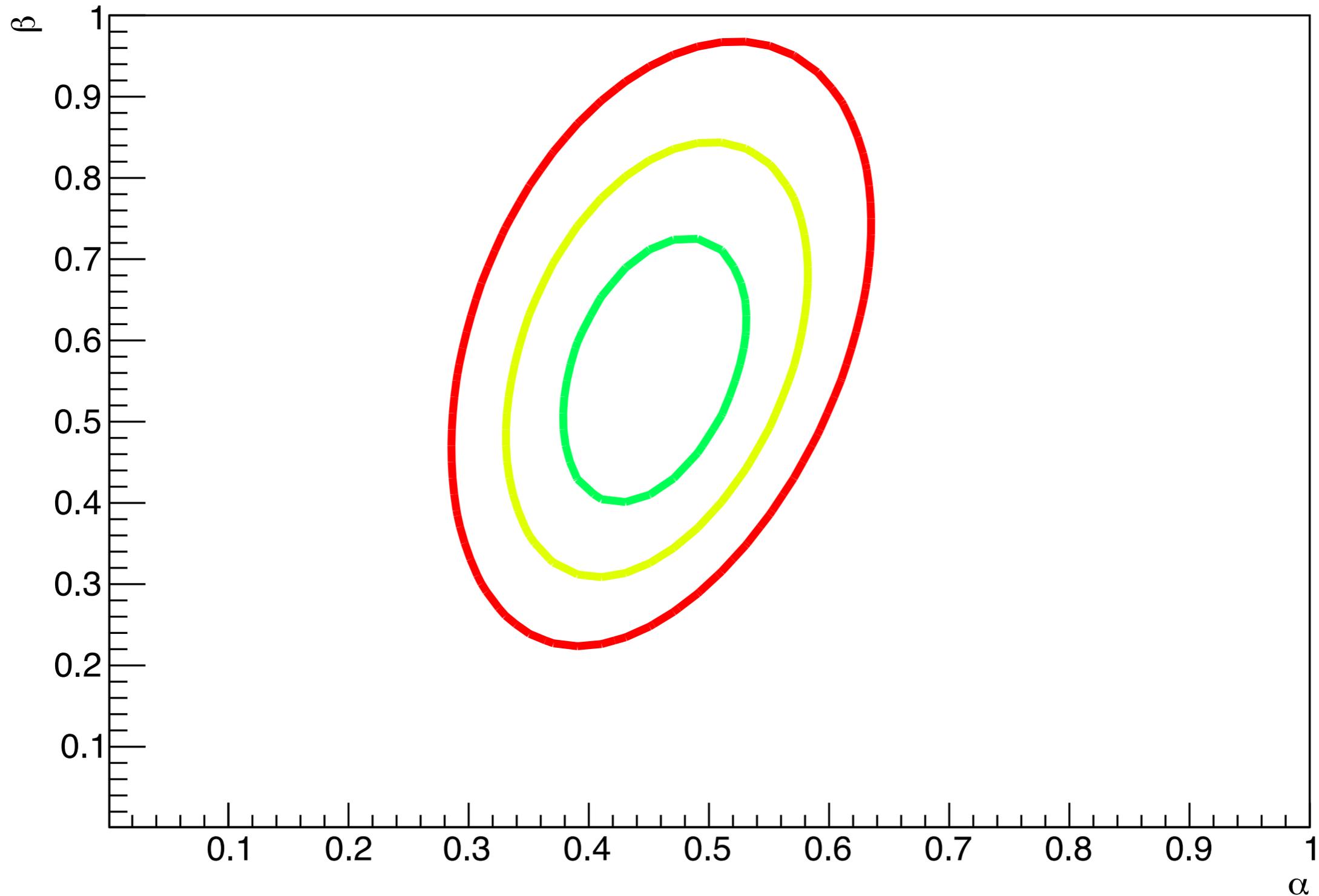
Contours on Top of the LLH Space

$-2*LLH$



Just the Contours

Contours from $-2*LLH$

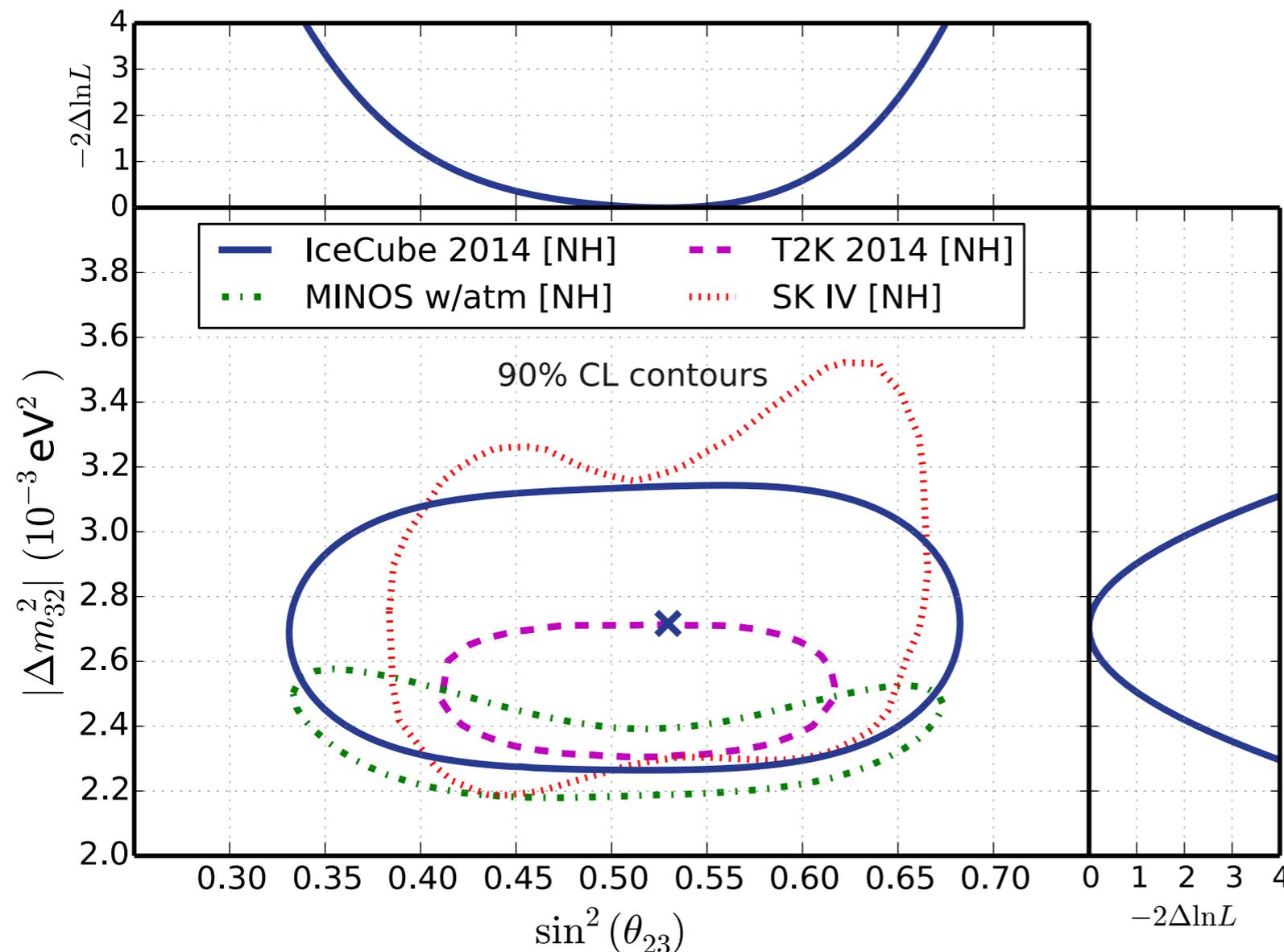


Real Data

- 1D projections of the 2D contour in order to give the best-fit values and their uncertainties

$$\sin^2 \theta_{23} = 0.53^{+0.09}_{-0.12}$$

$$\Delta m_{32}^2 = 2.72^{+0.19}_{-0.20} \times 10^{-3} \text{eV}^2$$



Remember, even though they are 1D projections the ΔLLH conversion to σ must use the degrees-of-freedom from the actual fitting routine

*arXiv:1410.7227

Exercise #3

- There is a file posted on the class webpage for “Class 7” which has two columns of x numbers (not x and y , only x for 2 pseudo-experiments) correspond to x over the range $-1 \leq x \leq 1$
- Using the function:

$$f(x; \alpha, \beta) = 1 + \alpha x + \beta x^2$$

- Find the best-fit for the unknown α and β
- Calculate the reduced chi-square goodness of fit by histogramming the data. The choice of bin width can be important.
 - Too narrow and there are not enough events in each bin for the statistical comparison.
 - Too wide and any difference between the ‘shape’ of the data and prediction histogram will be washed out, leaving the result uninformative and possibly misleading.

Extra

- Use a 3-dimensional function for $\alpha=0.5$, $\beta=0.5$, and $\gamma=0.9$ generate 2000 Monte Carlo data points using the function transformed into a PDF over the range $-1 \leq x \leq 1$

$$f(x; \alpha, \beta, \gamma) = 1 + \alpha x + \beta x^2 + \gamma x^5$$

- Find the best-fit values and uncertainties on α , β , and γ
- Similar to exercise #1, show that Monte Carlo re-sampling produces similar uncertainties as the Δ LLH prescription for the 3D hyper-ellipse
 - In 3D, are 500 Monte Carlo pseudo-experiments enough?
 - Are 2000 Monte Carlo data points per pseudo-experiment enough?
 - Write a profiler to project the 2D contour onto 1D, properly

Extra Extra

- Use Markov Chain to get the likelihood minimum and then use the LLH (or $-2*LLH$) values to get the uncertainties.
 - Is the MCMC quicker to converge to the 'best-fit' than using your LLH minimizer?
 - The Markov Chain estimator (maximum a posteriori - MAP) has a precision on the variance of $\mathcal{O}(1/n)$ for n simulation points, i.e. you can't get 99.9% interval without at least 1000 MCMC 'steps' after convergence. With a flat prior and using the 3-dimensional function the variance with an MCMC posterior distribution, do the best-fit values and uncertainties match what you get for the ΔLLH approach
 - Use the same 2000 data points for consistency from a single pseudo-experiment
 - Flat prior does not impact the $\mathcal{O}(1/n)$ variance, but just makes it easier to compare to the results already derived using the ΔLLH formulation for uncertainty

Extra

Variance of Estimators - Graphical Method

- More than one parameter
- For the case we estimate n parameters $\hat{\theta}$. The inverse minimum variance bound is given by the Fisher information matrix:

$$I_{ij} = E\left[-\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}\right] = -n \int f(x; \vec{\theta}) \frac{\partial^2 \ln f(x; \vec{\theta})}{\partial \theta_i \partial \theta_j} dx$$

- The information inequality state $V - I^{-1}$ is a positive semi-definite matrix with:

$$V_{ij} = \text{cov}[\hat{\theta}_i, \hat{\theta}_j] \rightarrow V[\hat{\theta}_i] \geq (I^{-1})_{ii}$$

- One will often find the inverse of the information matrix as an approximation for the covariance matrix, estimated using a matrix of 2nd derivatives at the maximum for the likelihood function L .

- Two Parameters

- For the 2D normal distribution, the consistent maximum likelihood estimators are:

$$\begin{aligned} \bar{x}_1 &= \frac{1}{N} \sum_{j=1}^N x_{1j} & \bar{x}_2 &= \frac{1}{N} \sum_{j=1}^N x_{2j} \\ s_1'^2 &= \frac{1}{N} \sum_{j=1}^N (x_{1j} - \bar{x}_1)^2 & s_2'^2 &= \frac{1}{N} \sum_{j=1}^N (x_{2j} - \bar{x}_2)^2 \\ r &= \frac{\sum_{j=1}^N (x_{1j} - \bar{x}_1)(x_{2j} - \bar{x}_2)}{N s_1' s_2'} \end{aligned}$$

sample correlation coeff.