General overview of data analysis

Gravitational wave mini-school
at
Beijing Normal University
15-18 September, 2016

Heng Ik Siong (王毅雄)

University of Glasgow

Outline of data analysis lectures

- **Introduction**
  - Normal, Student-t and χ^2 distributions
  - Power spectral density and time-frequency plots
- **Bayesian inference**
  - parameter estimation
- **Bayesian model selection**
- **Methods for identifying gravitational wave transients**
  - generic transients
  - matched filtering

Reading

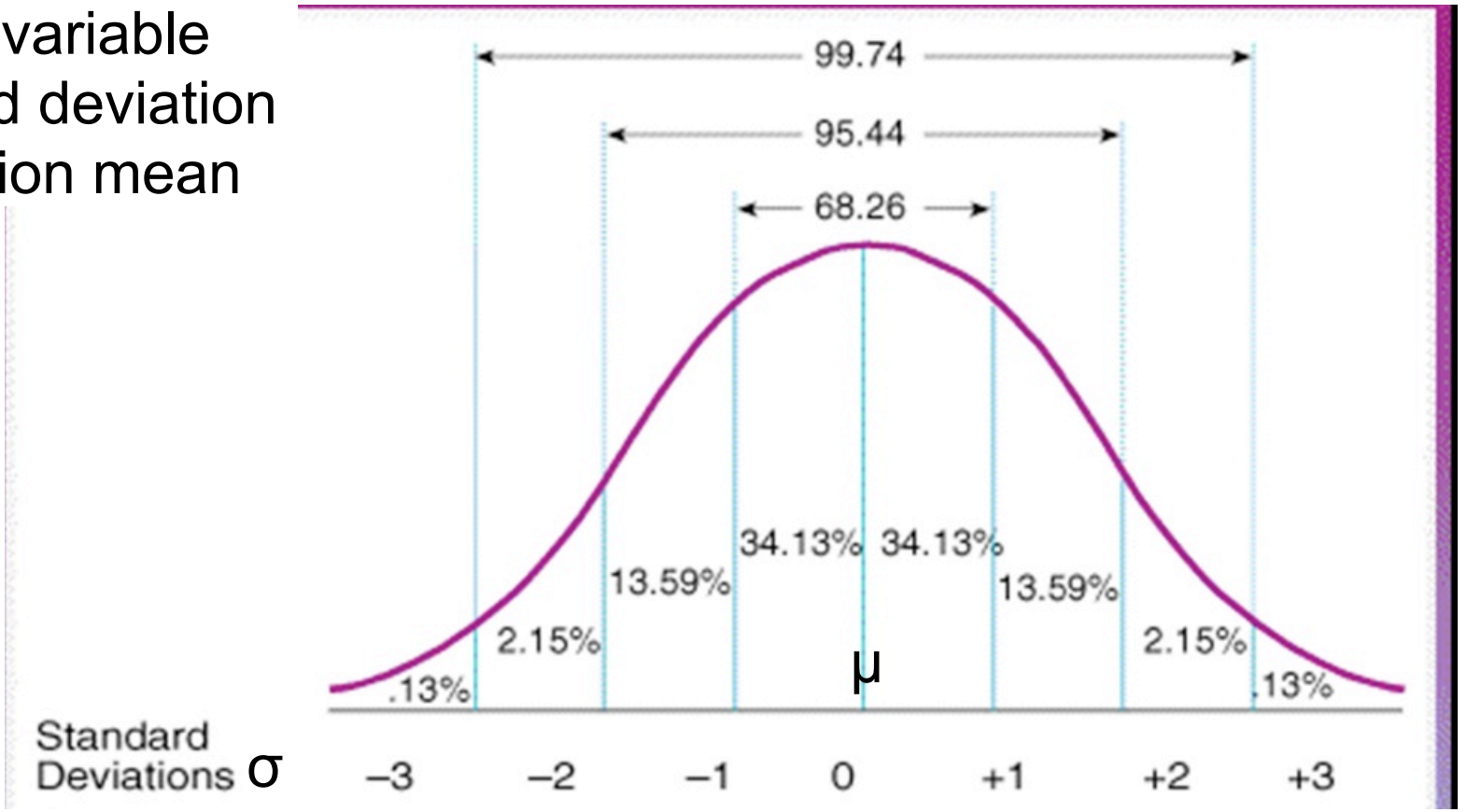
- **Further reading**
  - Living reviews: http://relativity.livingreviews.org/
    - data analysis: http://relativity.livingreviews.org/Articles/lrr-2012-4/
    - astrophysics: http://relativity.livingreviews.org/Articles/lrr-2009-2/
  - Data Analysis: A Bayesian Tutorial by D.S. Sivia & J. Skilling
  - Bayesian Logical Data Analysis for the Physical Sciences by P. Gregory
  - Probability Theory by E.T. Jaynes
  - Bayesian Spectrum and Parameter Estimation by G.L. Bretthorst (Free!)
    - http://bayes.wustl.edu/glb/book.pdf
    - http://bayes.wustl.edu/

Normal distribution

Probability of drawing the value x from a Normal distribution is given by

$$p(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

x - random variable
σ - standard deviation
μ - distribution mean



Student's t distribution

- What if we do not assume that the variance is known?
- We can estimate this from the data (provided n>1)
- We form the statistic

$$t_{\mathrm{obs}} = \left(\frac{\bar{x}_{\mathrm{obs}} - \bar{x}_{\mathrm{model}}}{\hat{\sigma}_\mu / \sqrt{n}} \right)$$

where

$$\hat{\sigma}_\mu = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x}_{\mathrm{obs}})^2$$

Number of independent variables reduced since we deduce the mean from observed data

- However, t_{obs} no longer has a Gaussian distribution

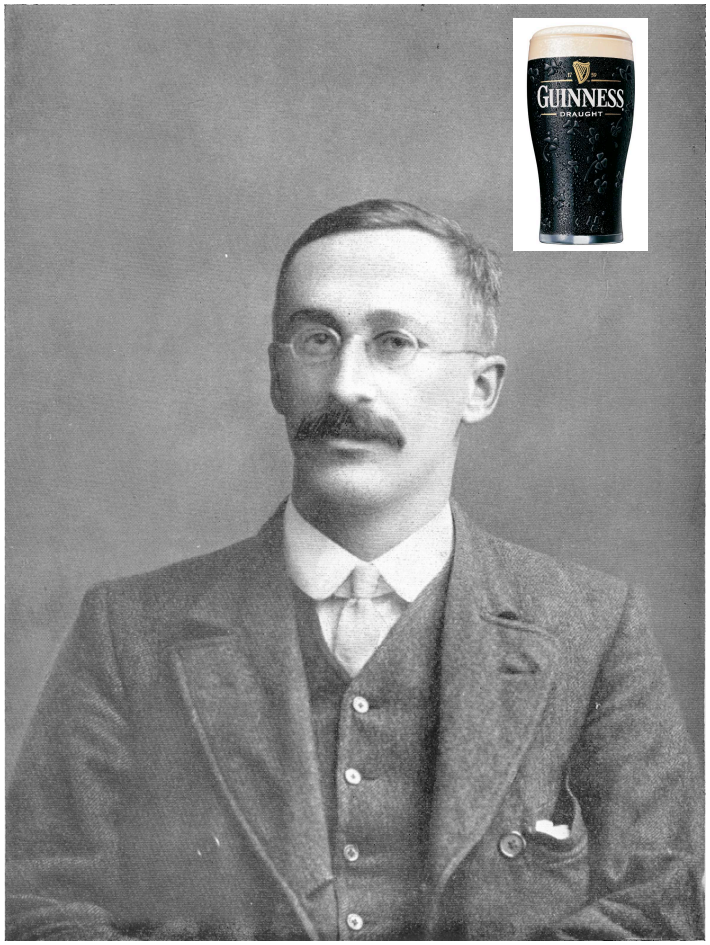
Student's t distribution

- **In fact, t_{obs} has a PDF known as Student's t distribution**

$$p(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

- **where ν = *n*-1 is its number of degrees of freedom and Γ is the Gamma function**

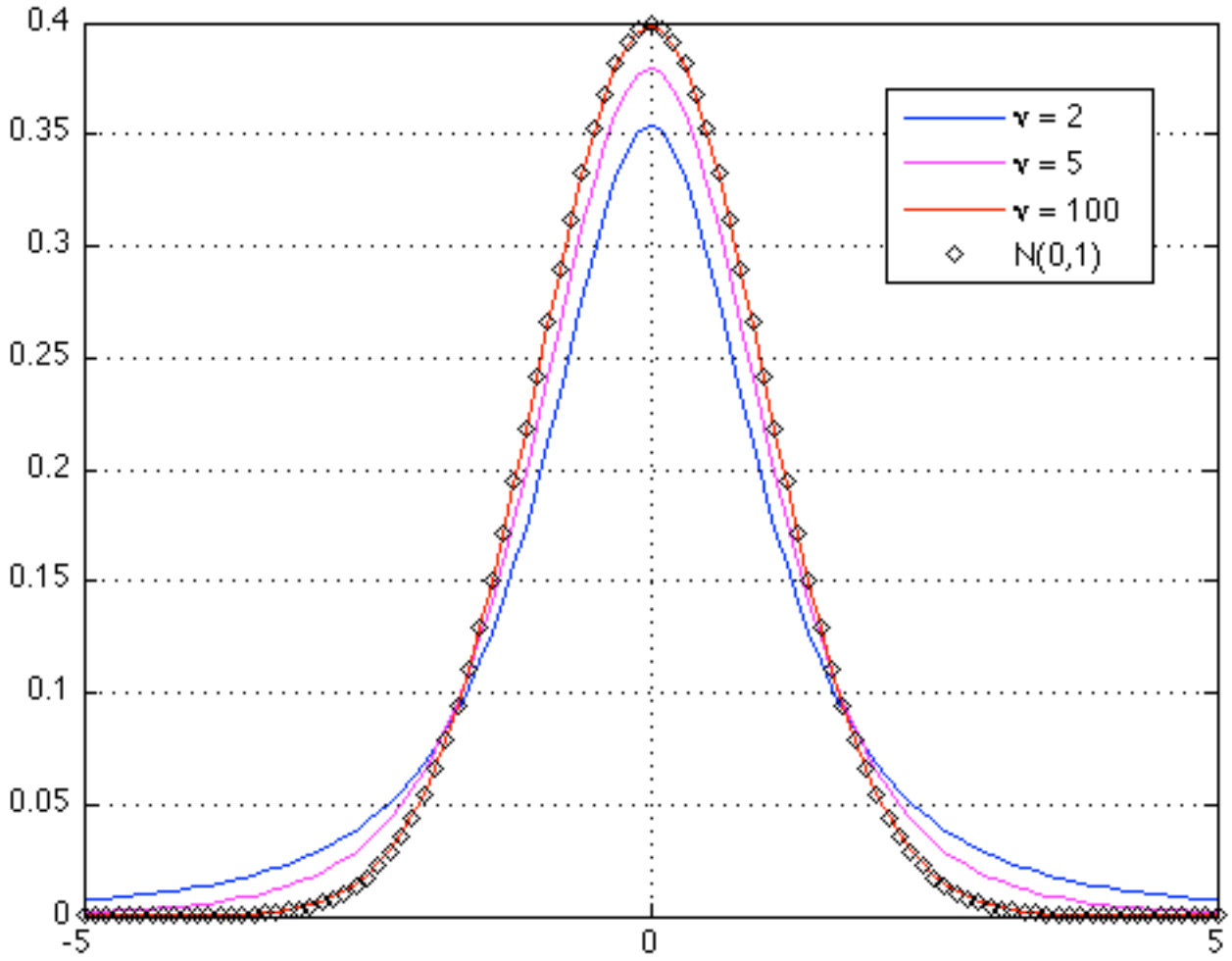
$$\Gamma(\nu) = \int_0^\infty x^{\nu-1}e^{-x}dx$$



William Sealy Gosset

Student's t distribution

- When *n* is large, the Student's t distribution approximates a Gaussian distribution



χ^2 distribution

- **Consider a single observation from a Gaussian likelihood**

$$p(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

- **If we have *k* observations, then the likelihood function becomes**

$$p(\{x_i\}|\{\mu_i\},\{\sigma_i\}) = \prod_{i=1}^{k} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left[-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}\right]$$

χ^2 distribution

- The log likelihood is therefore

$$\log_e p(\{x_i\}|\{\mu_i\},\{\sigma_i\}) \propto -\sum_{i=1}^{k} \frac{(x_i-\mu_i)^2}{2\sigma_i^2}$$

- The RHS belongs to a χ^2 distribution with *k* degrees of freedom with a probability density function such that

$$p(y|k) = \frac{1}{2^{k/2}\Gamma(k/2)} y^{k/2-1} e^{-y/2}$$

- Here, *y* corresponds to the sum of *k* Gaussian variables (RHS in top equation)
- Note that if the mean, μ_i, is not subtracted from x_i, then we obtain a non-central χ^2 distribution where the non-centrality parameter is the sum of all μ_i

Interesting properties

- In general, if we divide a Gaussian random variable with a scaled random variable from a χ^2 distribution, the resulting statistic follows Student's t distribution

- That is, if *Y* and *Z* belong to Gaussian and χ^2 distributions respectively, and

$$X = \frac{Y}{\sqrt{Z/n}}$$

- then *X* will have Student's t distribution with *n*-1 degrees of freedom

Interesting properties

- **Comparing**

$$t_{\mathrm{obs}} = \left(\frac{\bar{x}_{\mathrm{obs}} - \bar{x}_{\mathrm{model}}}{\hat{\sigma}_\mu / \sqrt{n}} \right) \quad \text{with} \quad X = \frac{Y}{\sqrt{Z/n}}$$

- **we note that σ_μ^2 must have a χ^2 distribution**
- **For the biased and unbiased sample variance,**

$$\sigma = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right)^2 \qquad \hat{\sigma}_\mu = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x}_{\mathrm{obs}})^2$$

- **both sum the square of Gaussian variables which gives a χ^2 distribution with *n* degrees of freedom**

Summary of distributions

- **Normal (Gaussian) distribution**

$$p(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

- **χ^2 distribution: Obtained by summing the square of *k* Gaussian variables; $\mu = k$, $\sigma^2 = 2k$**

$$p(x|k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$$

- **Student's *t* distribution: $t = x/\sqrt{(yn)}$ where *x* and *y* are Gaussian and χ^2 variables, *n* is the total number of variables and $\nu = n-1$; $\mu = 0$, $\sigma^2 = \nu/(\nu-2)$ for $\nu > 2$**

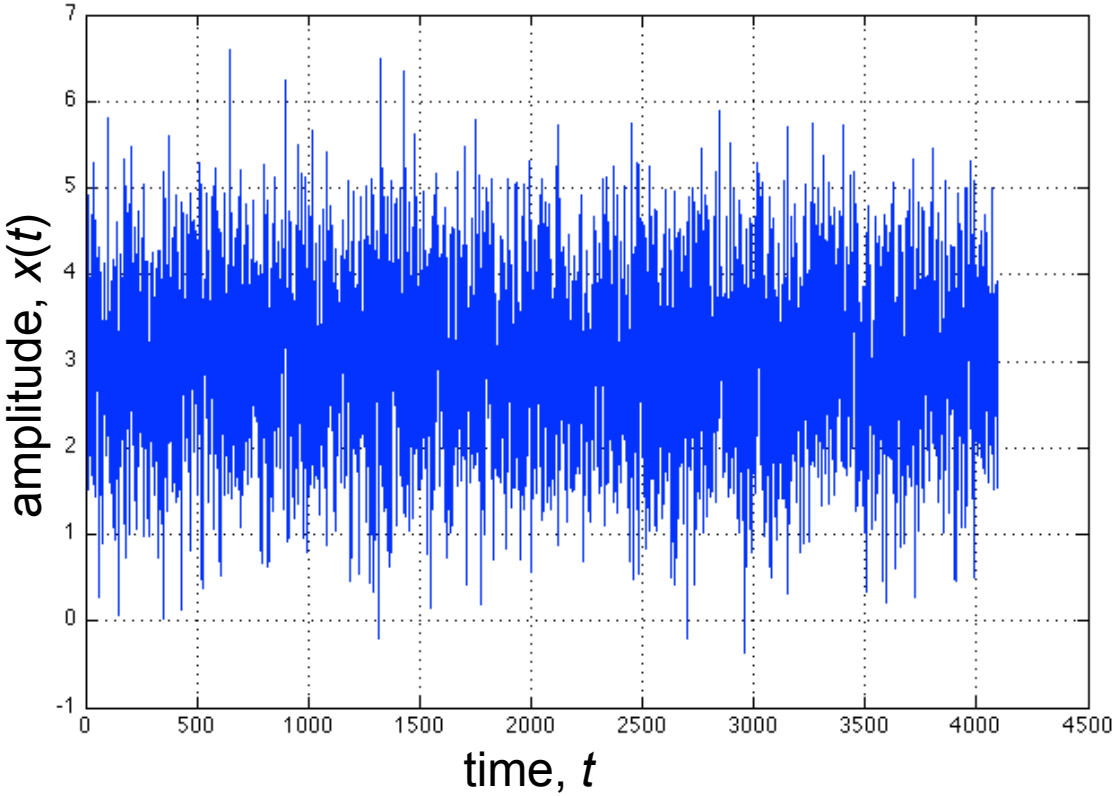
$$p(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(k/2)} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Power spectral density

-Consider a stream of data with amplitude x(t)

-The average power in this data is

$$P = \frac{1}{2T}\int_0^T x(t)^2 dt$$



-Also, the Fourier transform of the data is

$$\hat{x}_T(2\pi f) = \frac{1}{\sqrt{T}}\int_0^T x(t)e^{-i2\pi f t} dt$$

Power spectral density

- The power spectral density is defined as

$$S_{xx}(2\pi f) = \lim_{T\to\infty} \mathbf{E}\left[|\hat{x}_T(2\pi f)|^2\right]$$

- The expectation value is expanded to become

$$\mathbf{E}\left[|\hat{x}_T(2\pi f)|^2\right] = \mathbf{E}\left[\frac{1}{T}\int_0^T x^*(t)e^{i2\pi ft}dt \int_0^T x(t')e^{-i2\pi ft}dt'\right]$$

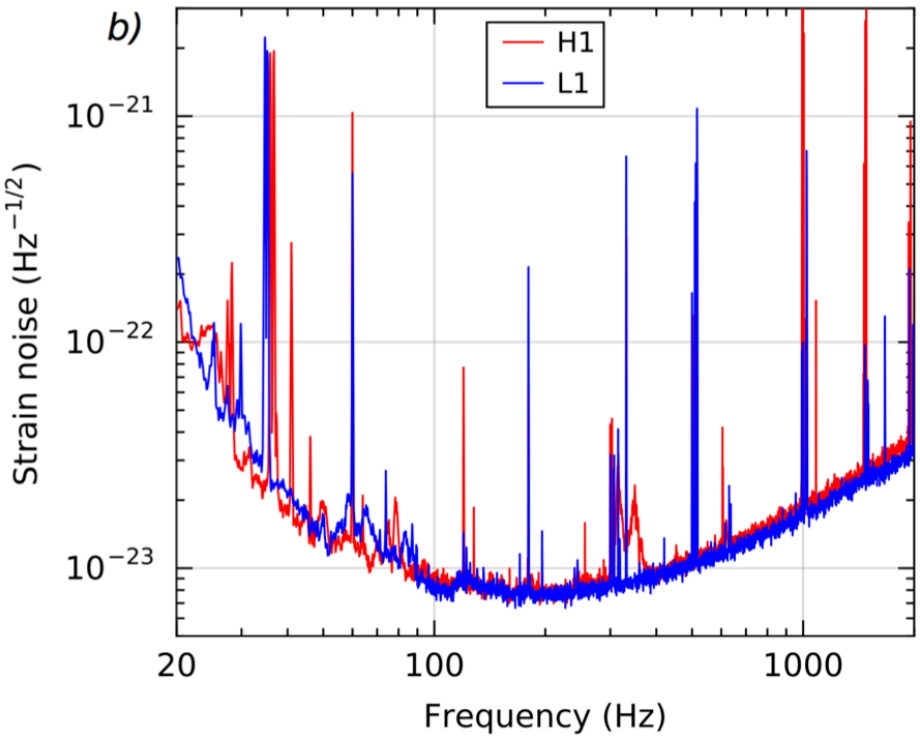
$$= \frac{1}{T}\int_0^T\int_0^T \mathbf{E}\left[x^*(t)x(t')\right] e^{i2\pi ft}e^{-i2\pi ft}dtdt'$$

Autocorrelation function of $x(t)$

- So, power spectral density is the Fourier transform of the autocorrelation function of the data

Power spectral density

- The power spectral density is the amount of power in the data in each frequency bin
- The power spectral density can be expressed as the Fourier transform of the autocorrelation function, γ(τ),

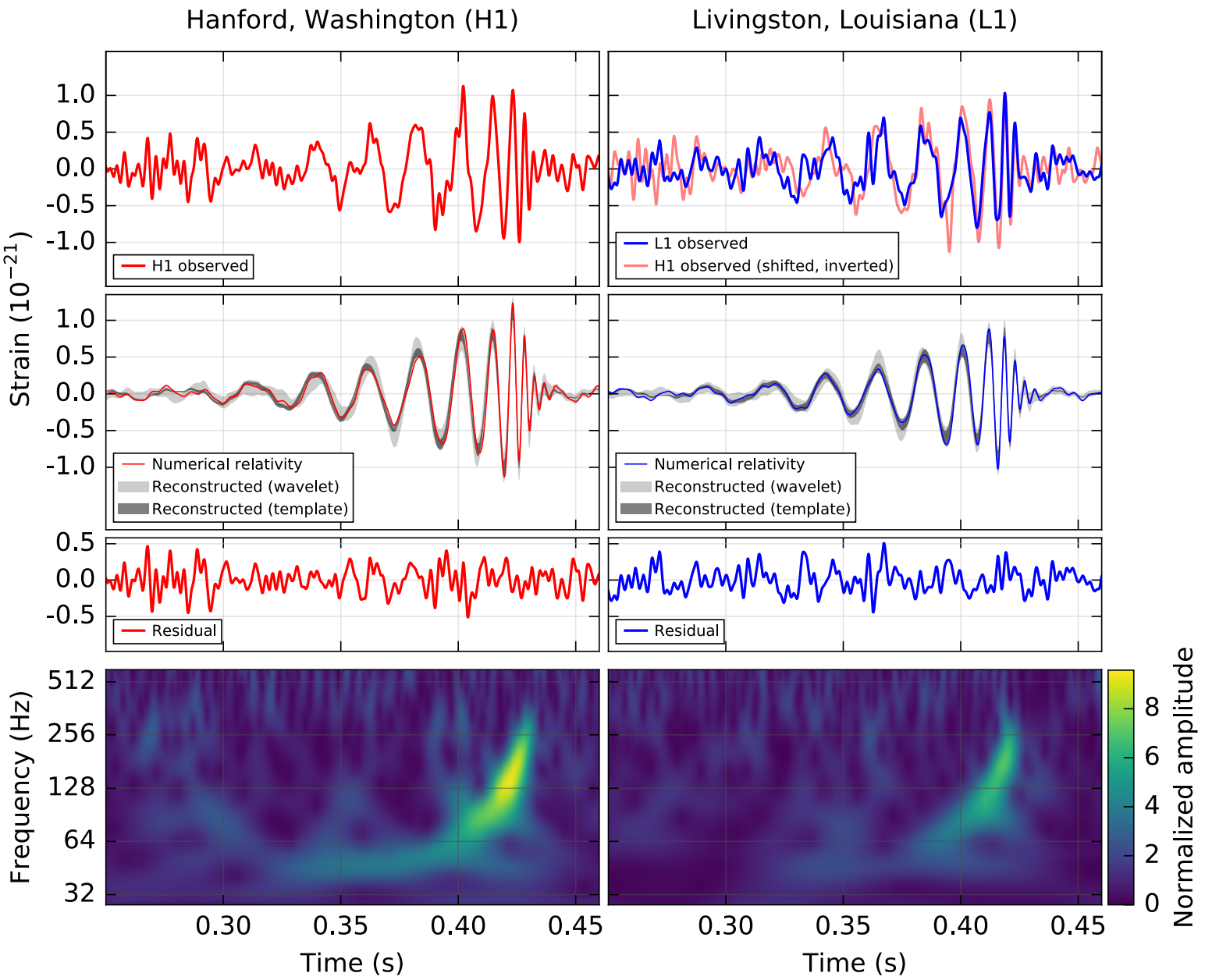


$$S_{xx}(2\pi ft) = \int_{-\infty}^{\infty} \gamma(\tau) e^{-i2\pi f\tau} d\tau$$

- The amplitude spectral density (ASD) is the square root of the power spectral density (PSD)

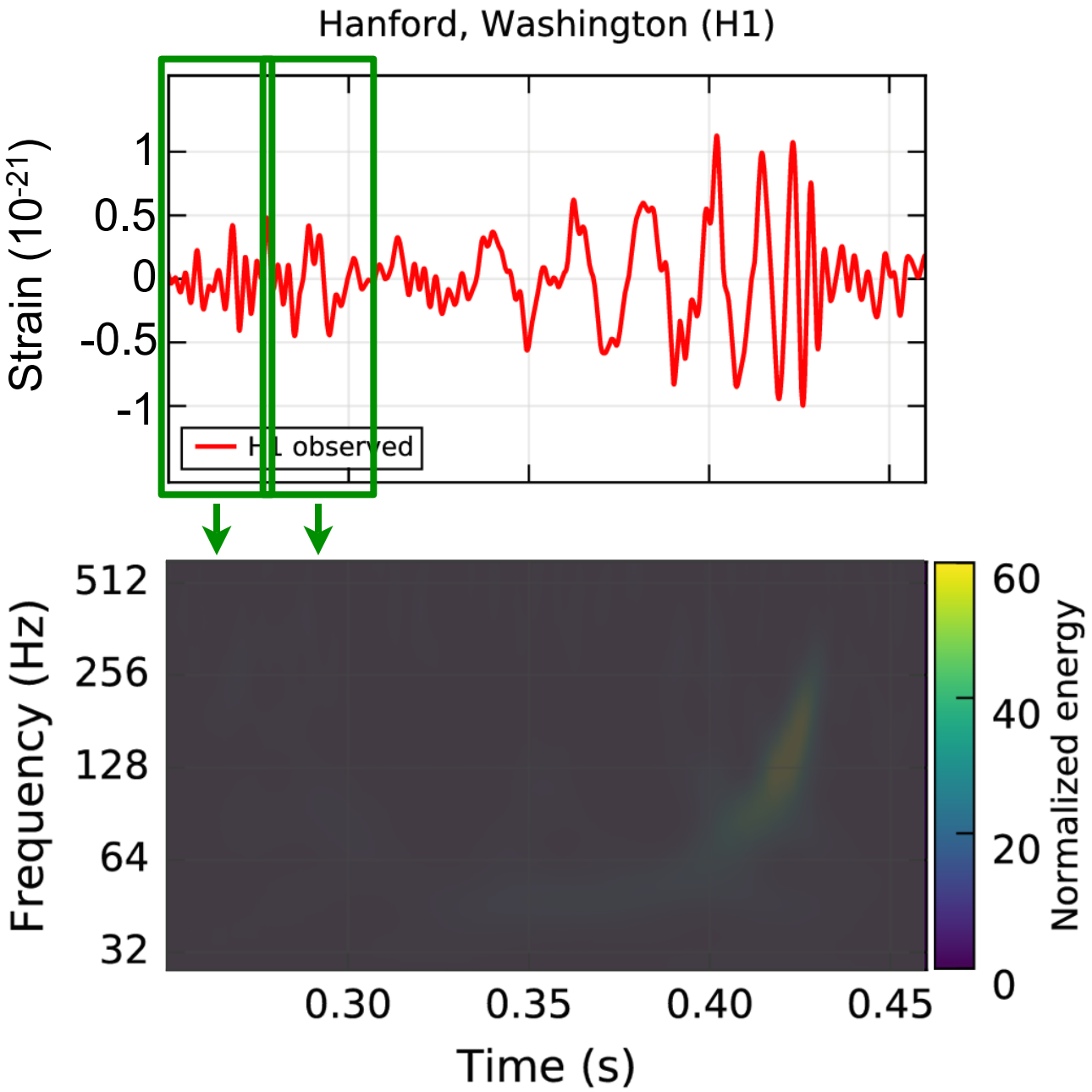
Time-frequency representations

Power spectral densities can also be used to build time-frequency representations of the data



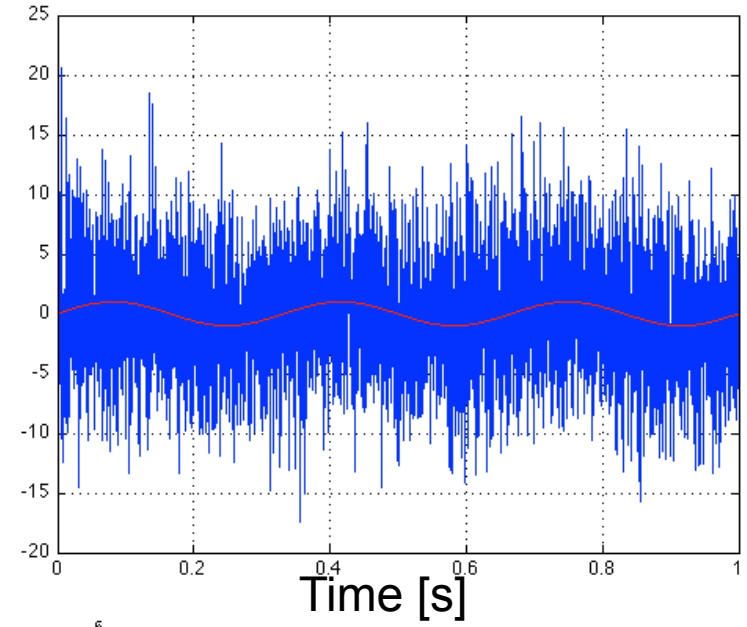
Time-frequency map

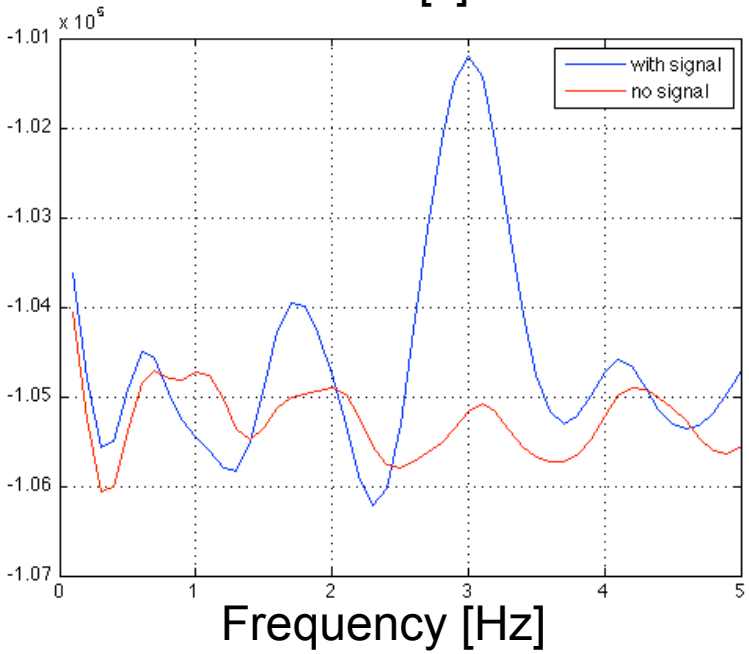
- To construct a time-frequency map, one does the following:
  - split time series into multiple segments
  - calculate PSD of each segment
  - plot each segment in order, with the colours representing the power in each time-frequency bin



What is data analysis

- Consider noisy data where you would like to determine the presence of a signal
- Compare the expected signal with the data
- Obtain a quantitative way of determining the presence of the signal (or lack thereof)





Bayesian inference

- The probability that both *X* and *Y* are true can be written as

$$p(X, Y|I) = p(X|Y, I) \times p(Y|I)$$

- where *I* is relevant background information
- We expect that

$$p(X, Y|I) = p(Y, X|I)$$

- So, by expanding both sides and rearranging, we obtain

$$p(X|Y, I) = \frac{p(Y|X, I) \times p(X|I)}{p(Y|I)}$$

Bayesian inference

- **This is known as Bayes' theorem**

$$\underbrace{p(X|Y,I)}_{\text{posterior}} = \frac{\overbrace{p(Y|X,I)}^{\text{likelihood}} \times \overbrace{p(X|I)}^{\text{prior}}}{\underbrace{p(Y|I)}_{\text{evidence}}}$$

Thomas Bayes
1701-1761

- **For the purposes of statistical data analysis, we can interpret the above as**

$$p(model|data,I) \propto p(data|model,I) \times p(model|I)$$

- **Note that evidence is not used here and we will discuss this later**

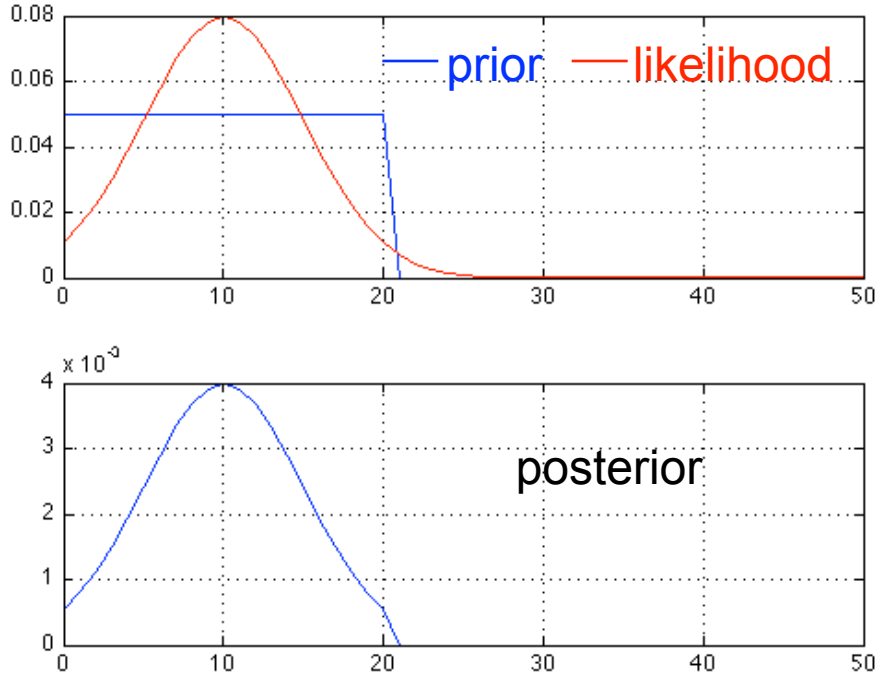
Bayesian inference

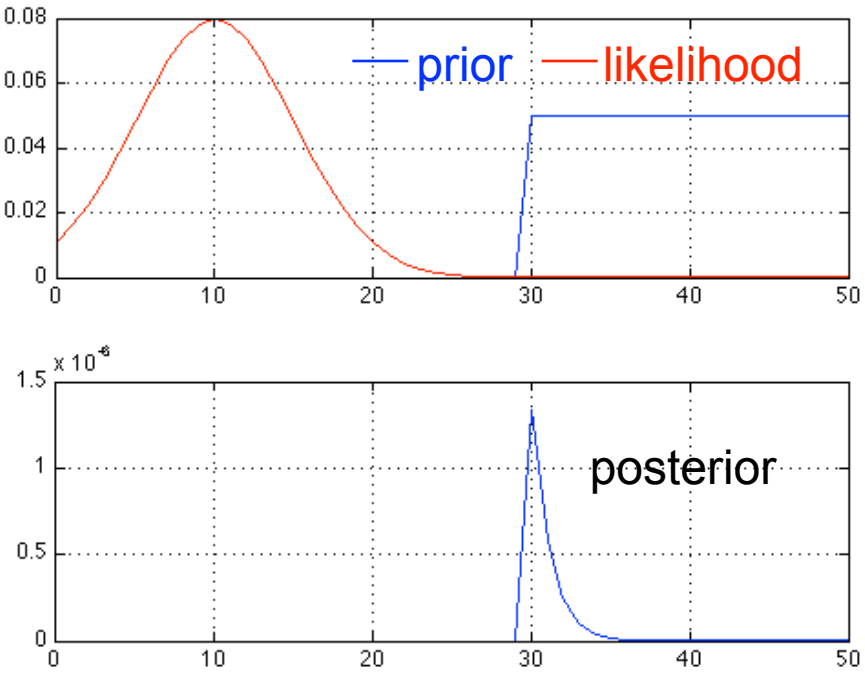
$$p(model|data, I) \propto p(data|model, I) \times p(model|I)$$

- Here, the posterior probability represents our knowledge about the model given the data we have acquired
- The prior probability represents our state of knowledge about the model before any analysis of the data and this is modified by the acquisition of data through the likelihood function

Bayesian inference

$$p(model|data, I) \propto p(data|model, I) \times p(model|I)$$





- Inference about the model should be made using the posterior
- If we consider a broad, uniform prior, then the posterior is proportional to the likelihood

Parameter estimation

- We can find the most probable parameters by finding the maximum of the posterior probability density function (PDF)
- Note that this leads to maximising the likelihood since it is proportional to the posterior PDF
- So, a posterior, ***P*(*X*)**, for a model, ***X***, will take the form

$$P(X) = p(X|\{data\}, I)$$

- If X_0 is the best estimate of the model parameters, then

$$\left.\frac{dP(X)}{dX}\right|_{X=X_0} = 0 \quad \text{and} \quad \left.\frac{d^2P(X)}{dX^2}\right|_{X=X_0} < 0$$

Parameter estimation

- Instead of looking at $P(X)$ directly, we consider its logarithm, $L(X) = \log_e P(X)$, which is smoother
- Since we are considering $L(X)$ about the maximum point X_0, we can perform a Taylor expansion so that

$$L(X) = L(X_0) + \underbrace{\frac{dL(X)}{dX}\Big|_{X=X_0}}_{0}(X - X_0) + \frac{1}{2}\frac{d^2L(X)}{dX^2}\Big|_{X=X_0}(X - X_0)^2 + \ldots$$

- The quadratic term dominates the expansion, so

$$P(X) \approx A \exp\left[\frac{1}{2}\frac{d^2L(X)}{dX^2}\Big|_{X=X_0}(X - X_0)^2\right]$$

- where A is a normalising constant

Parameter estimation

$$P(X) \approx A \exp\left[\frac{1}{2}\frac{d^2 L(X)}{dX^2}\Big|_{X=X_0}(X - X_0)^2\right]$$

- **Comparing this with a Gaussian**

$$p(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

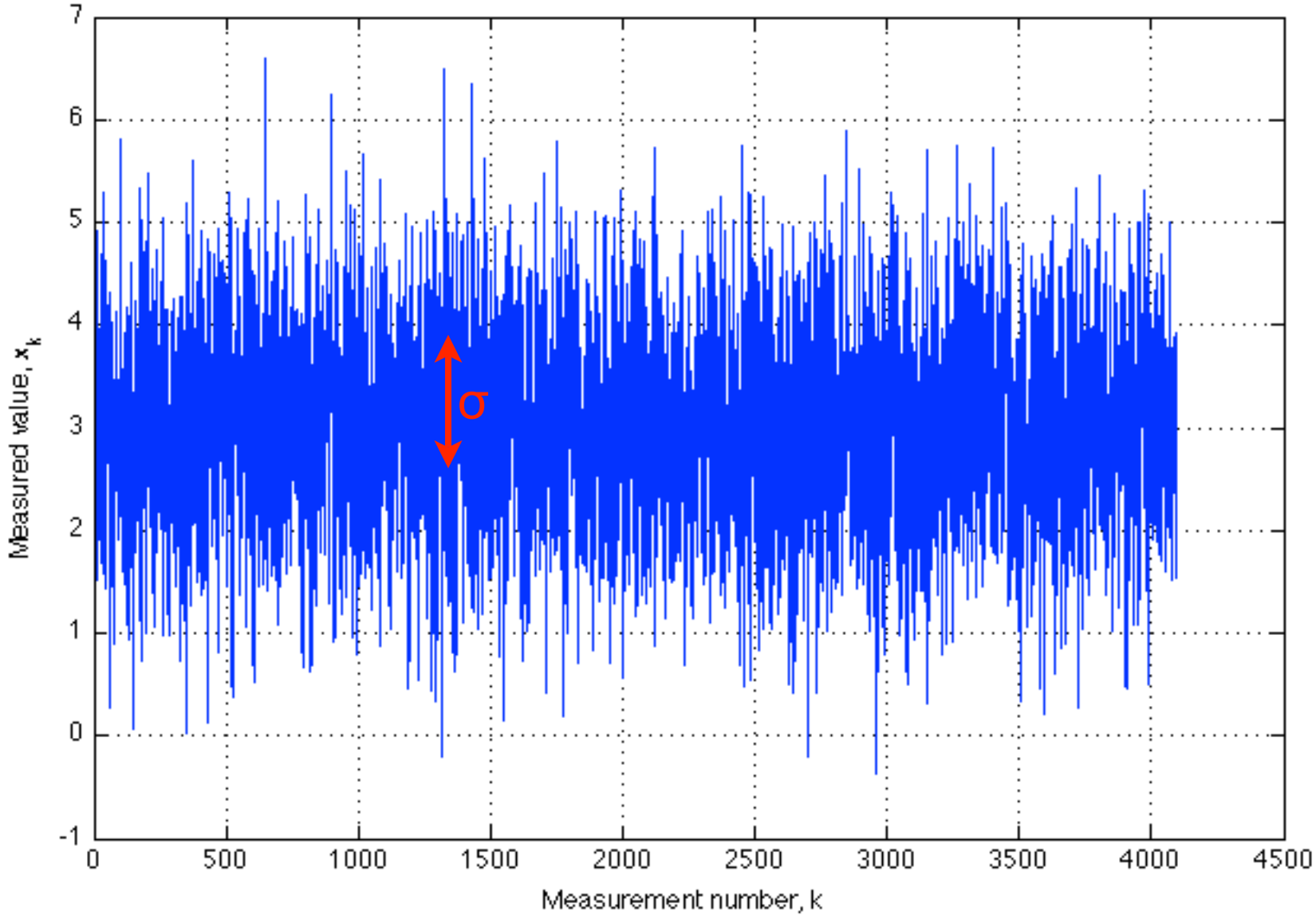
- **We see that our approximation of the posterior PDF is equivalent to a Gaussian distribution with**

$$\sigma_X^{-2} = -\frac{d^2 L(X)}{dX^2}\Big|_{X=X_0}$$

and our best estimate parameters are inferred such that

$$X = X_0 \pm \sigma_X$$

Parameter estimation example



μ?

Parameter estimation example

- **For our data set of k independent measurements**

$$p(\{x_k\}|\mu,\sigma,I) = \prod_{i=1}^{N} p(x_k|\mu,\sigma,I)$$

$$= \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x_k-\mu)^2}{2\sigma^2}\right]$$

- **Also, assign a uniform prior such that**

$$p(\mu|\sigma,I) = p(\mu|I) = \begin{cases} a & \mu_{\min} \le \mu \le \mu_{\max} \\ 0 & \text{otherwise} \end{cases}$$

where *a* is a normalisation constant

Parameter estimation example

- We want the log of the posterior pdf which, in this case, is directly proportional to the likelihood function

$$L = \ln p(\mu|\{x_k\}, \sigma, I) = \ln a + \sum_{i=1}^{N} \ln[p(x_k|\mu, \sigma, I)]$$

$$= \text{constant} - \sum_{i=1}^{N} \frac{(x_k - \mu)^2}{2\sigma^2}$$

- The constant contains all terms that do not involve μ

Parameter estimation example

- We denote the best estimate of μ with μ_0 and we differentiate *L* and set it to 0

$$\left.\frac{dL}{d\mu}\right|_{\mu_0} = \sum_{i=1}^{N} \frac{(x_k - \mu_0)}{\sigma^2} = 0$$

- From this we find that

$$\sum_{i=1}^{N} x_k = \sum_{i=1}^{N} \mu_0 = N\mu_0$$

or

$$\mu_0 = \frac{1}{N} \sum_{i=1}^{N} x_k$$

Credible regions

- Lets think about the log likelihood

$$L(\theta_1, \theta_2) = \text{constant} - \frac{1}{2}\chi^2(\theta_1, \theta_2)$$

- Here, the exponent has been written as a χ^2
- We define $\chi^2 = \chi^2_{\text{min}}$ when $(\theta_1, \theta_2) = (\theta_{01}, \theta_{02})$
- So, we can write

$$\Delta\chi^2 = \chi^2 - \chi^2_{\text{min}}$$

- which tells us the value of the χ^2 when the parameters values are chosen do not correspond to their *true* values
- what to maximise the likelihood, so minimise χ^2

Credible regions

- We can rewrite our log likelihood in terms of the minimised χ^2

$$L(\theta_1, \theta_2) - L(\theta_{01}, \theta_{02}) = -\frac{1}{2}(\chi^2 - \chi^2_{\rm min}) = -\frac{1}{2}\Delta\chi^2$$

- Rearranging, we get

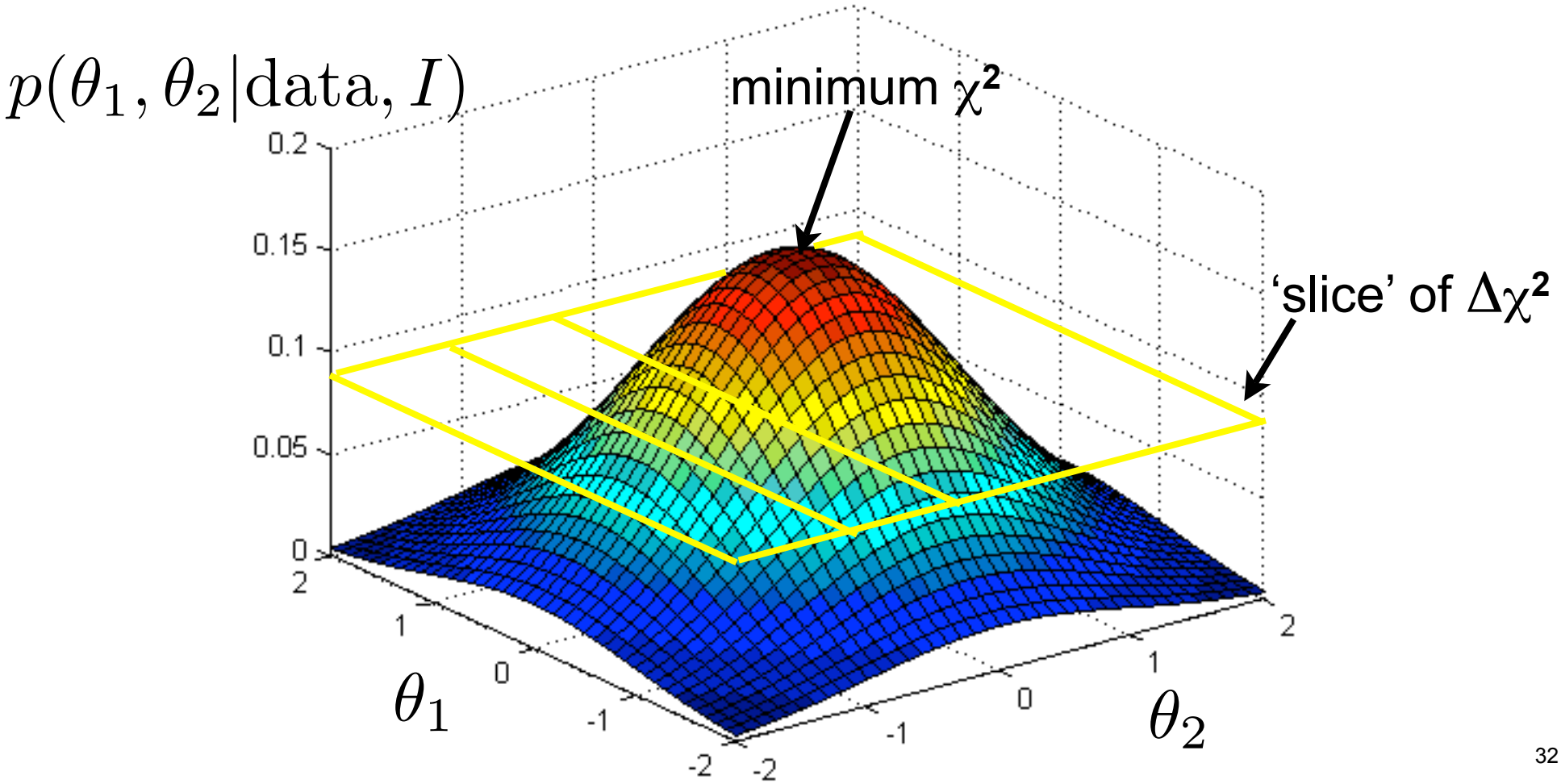
$$L(\theta_1, \theta_2) = L(\theta_{01}, \theta_{02}) - \frac{1}{2}\Delta\chi^2$$

- Therefore, the posterior can be written as

$$p(\theta_1, \theta_2|{\rm data}, I) = p(\theta_{01}, \theta_{02}|{\rm data}, I)\exp\left[-\frac{1}{2}\Delta\chi^2\right]$$

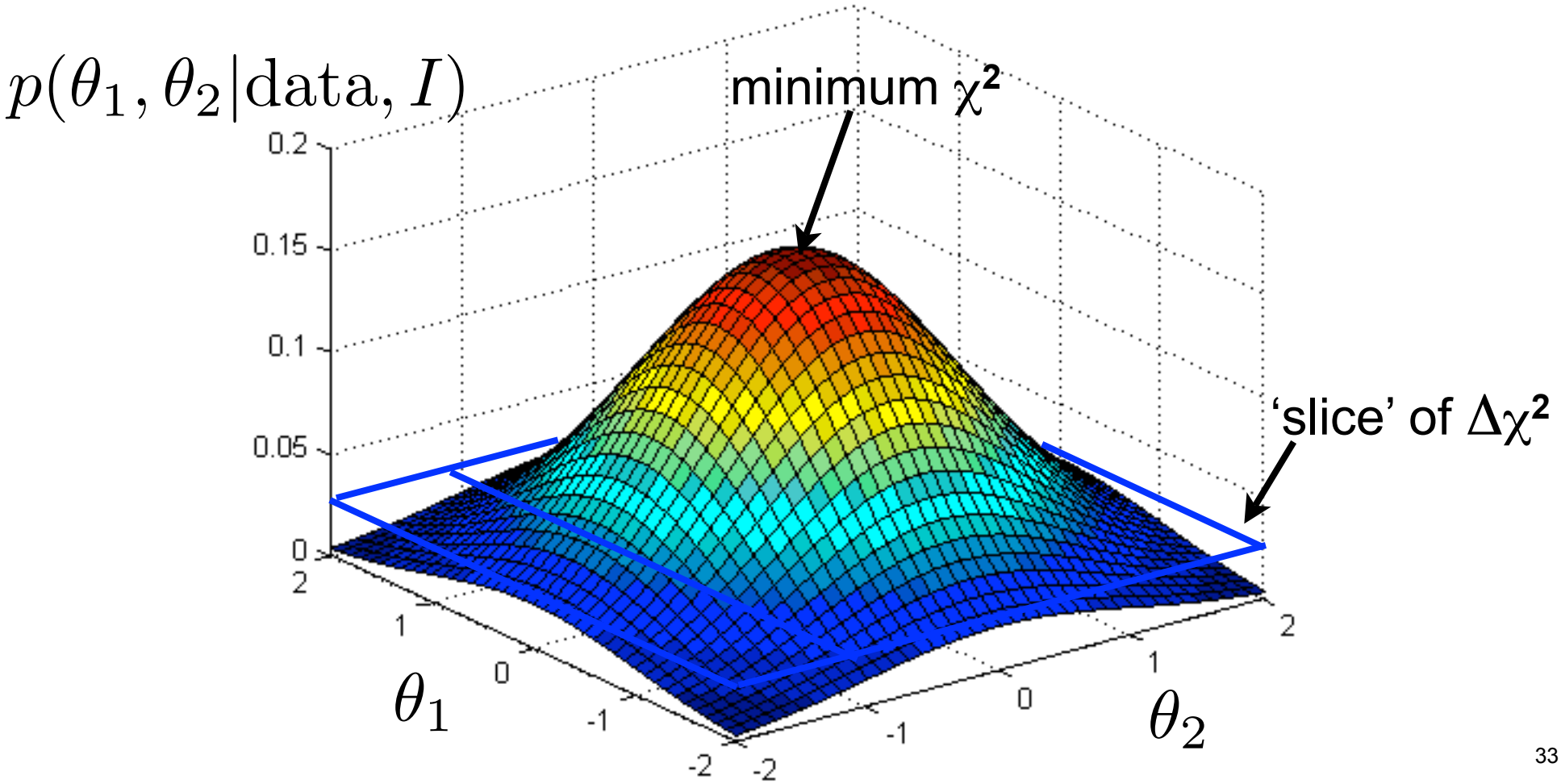
Credible regions

- **Each 'level' of posterior probability corresponds to a unique value of $\Delta\chi^2$**



Credible regions

- **The further away the chosen parameters are from $(\theta_{01}, \theta_{02})$ the broader the credible regions**



Credible regions

- **We can compute $\Delta\chi^2$ that enclose eg. 68%, 90%, 99% of the posterior PDF**
- **These $\Delta\chi^2$ slices are called credible regions**

  -also referred to as confidence intervals

- **Table of $\Delta\chi^2$ values for different number of parameters**

	Degrees of freedom (number of parameters)		
p	**1**	**2**	**3**
0.68	**1.00**	**2.30**	**3.53**
0.90	**2.71**	**4.61**	**6.25**
0.99	**6.63**	**9.21**	**11.3**

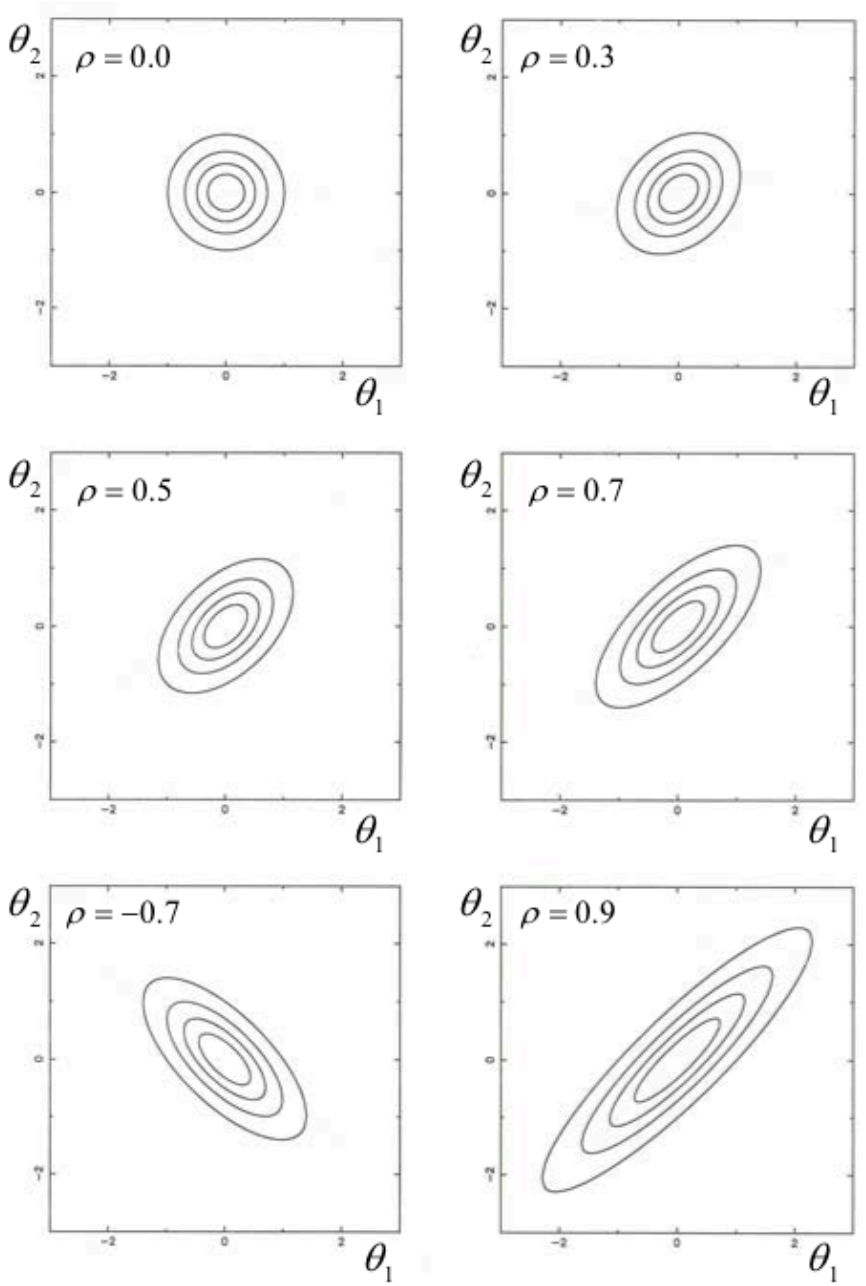
- **For a bivariate normal posterior, the covariance matrix is, in general, not diagonal**

Correlation coefficient

- **We define a correlation coefficient, ρ, for 2 variables *x* and *y* so that**

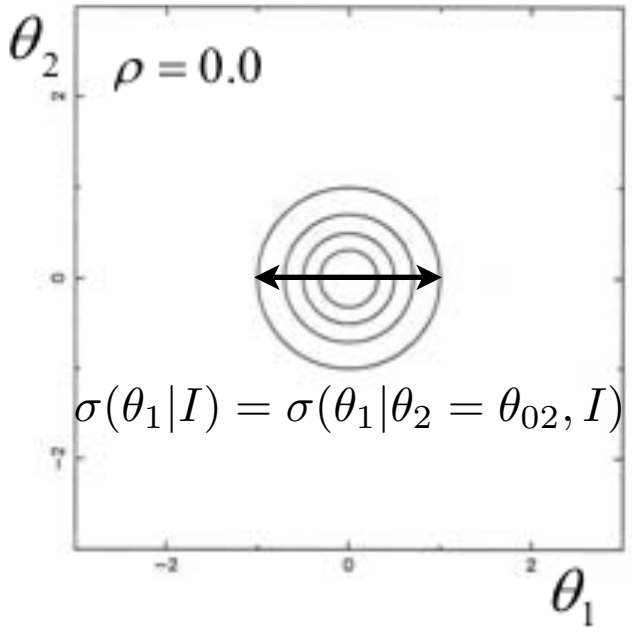
$$\rho = \frac{\sigma_{xy}}{\sqrt{\sigma_x \sigma_y}}$$

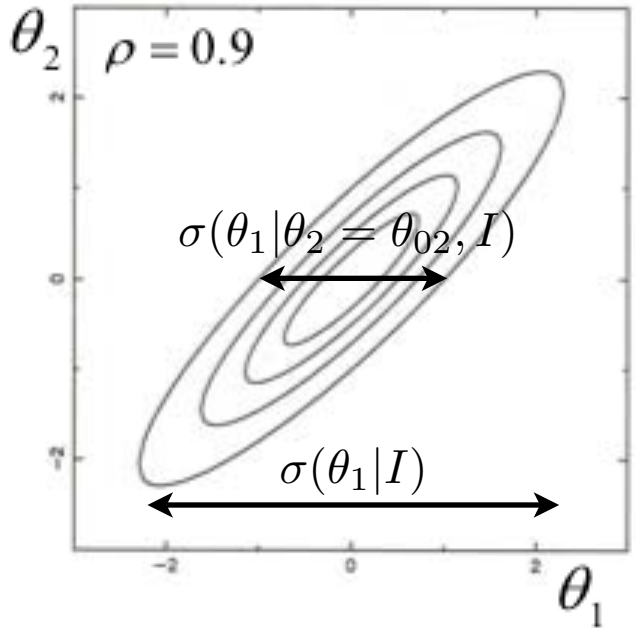
- **As the covariance matrix becomes 'less diagonal'**
  - -|ρ| increases
  - -isoprobability contours elongate
- **This is important if we are interested in only 1 of many parameters**



Correlation coefficient

- If we ignore the correlation coefficient, we can severely underestimate the uncertainty on a single parameter
- When ρ = 0, the marginal and conditional error bars are equal
- Otherwise, using the conditional value will give an error bar that is too small!

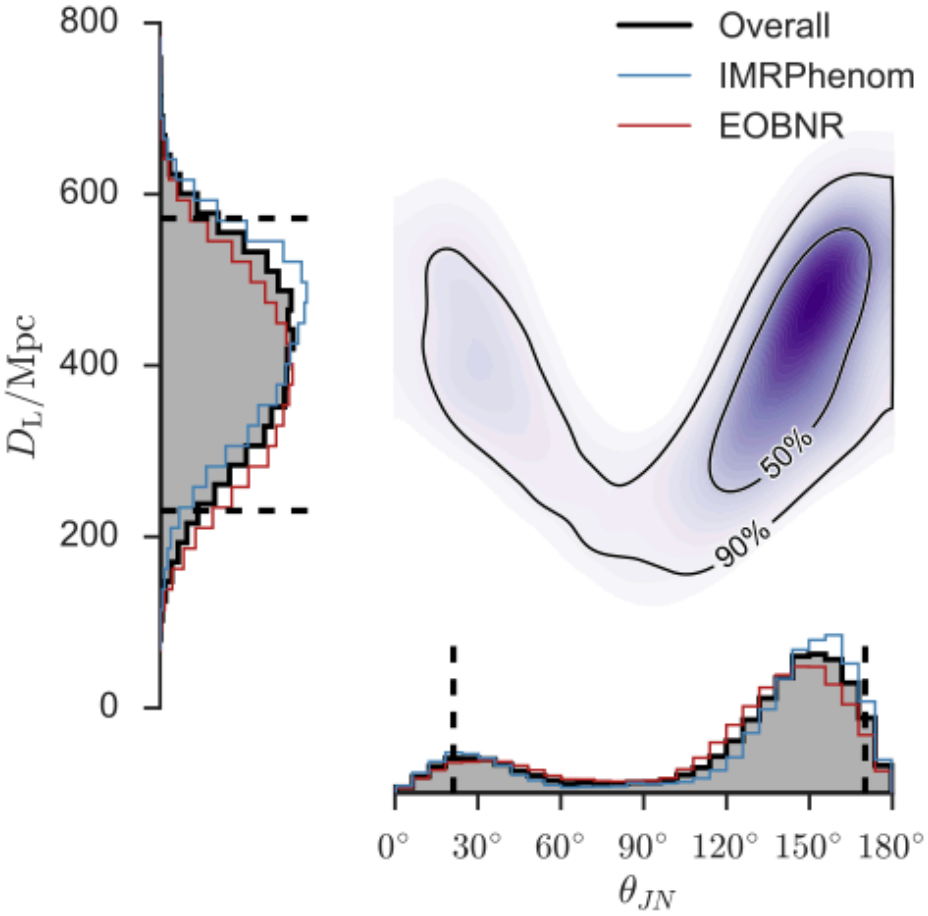




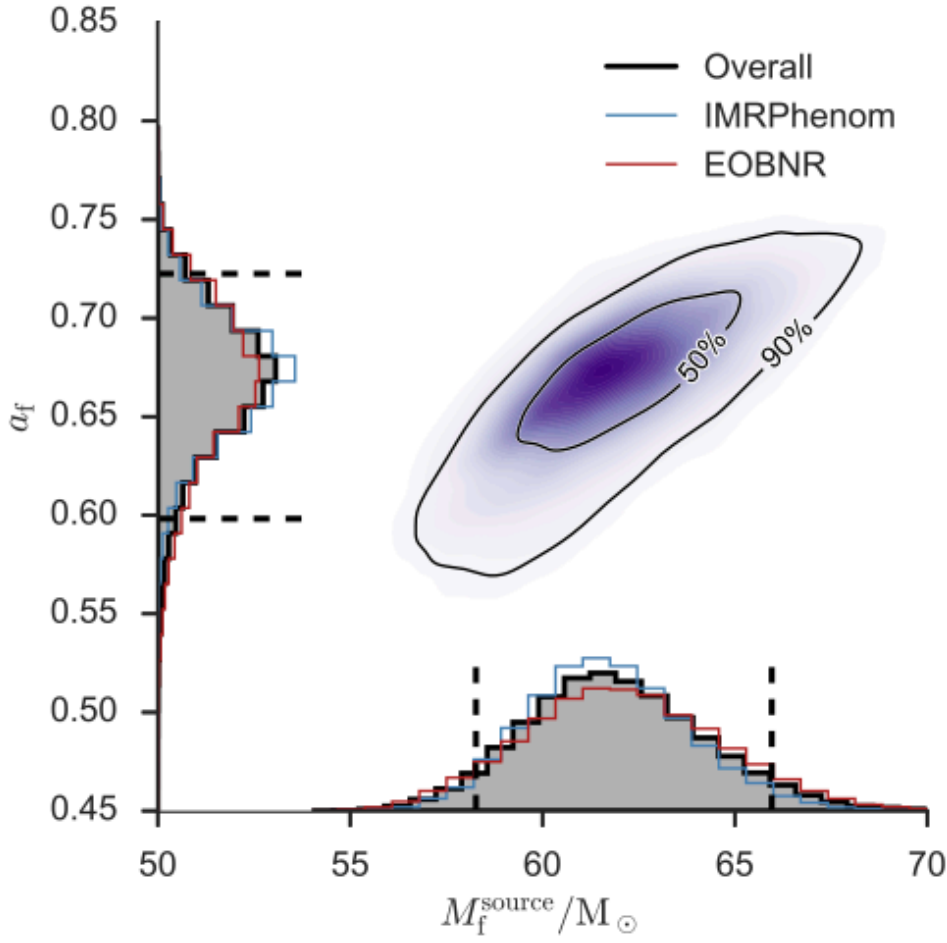
Example posterior distributions

- **Parameter estimation posteriors for GW150914**

Distance-inclination angle



final BH spin-final BH mass



Example posterior distributions

- Posterior on sky location estimate of GW150914