Practical Introduction to Bayesian (& Frequentist) analysis

Shantanu Desai
IIT Hyderabad

Data Analysis Problems in Astronomy or Physics

- Model Fitting or Regression
- Model Comparison (or Model Selection)
- Comparison of results from different experiments assessing the same model (tension metrics) M. Raveri's talk in busy week
- Forecasts for next generation experiments (not discussed today)

Model fitting and Model comparison techniques

- Model Fitting or Regression - two independent methods

  -Frequentist

  - Bayesian

- Model Comparison (or Model Selection) - three independent methods

  - Frequentist - not discussed today

  - Bayesian

  - Information theory

Goals in Parameter Estimation (Regression)

- Given some dataset with **error bars,** fit the data to a user-specified model (*)
- Obtain the best-fit parameters of the model and its error bars (or confidence intervals in case of multiple parameters)
- Determine if the resulting fit is good or bad and quantify it in terms of p-value

(*) Note that non-parametric regression is also becoming very popular and ubiquitous in Astronomy (Look for Gaussian Process Regression)

Frequentist Parameter Estimation References

References : Numerical Recipes Chapter 15 (in 1992 edition) or Bevington's book on Data Analysis for Physical sciences

arXiv: 1012.3754 Dos and don'ts of reduced chi-square

arXiv:1009.2755 Error estimation in astronomy : a guide

arXiv:1008.4686 Data analysis recipes : fitting a model to data - covers both frequentist & Bayesian viewpoints

Basics of Frequentist parameter estimation

- Create a likelihood function given some data (y_i), model $y(x_i)$ and error in y given by σ_i

  Assume Likelihood is Gaussian

$$P \propto \prod_{i=1}^{N} \exp\left[-\frac{1}{2}\left(\frac{y_i - y(x_i,\theta)}{\sigma_i}\right)^2\right]$$

Define $\chi 2 = -2 \ln L$

Least-squares minimization

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - y(x_i, \theta)}{\sigma_i} \right)^2$$

Not applicable for Poisson data

Use Cash statistics instead

Best-fit parameters are obtained by minimizing chi-square
For models that are linear in theta, probability distribution of $\chi 2$ is chi-square PDF with v= N-M degrees of freedom, where N is the number of data points and M is the number of free parameters.

A rule of thumb : $\chi 2 \sim v$ for a good fit (sometimes called "chi by eye")

$< \chi 2 > = v$ and its standard deviation = sqrt(v)

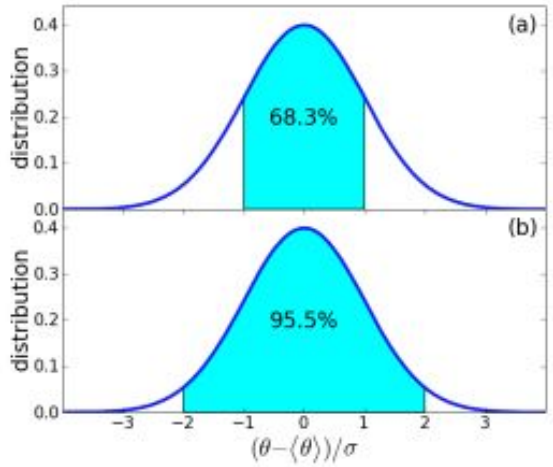
Reduced χ^2 = χ^2/v

Error Estimates/Confidence Intervals

$$\mathrm{prob}(\theta_- \leq \hat{\theta} \leq \theta_+) = \int_{\theta_-}^{\theta_+} d\theta\, \mathrm{prob}(\theta) = C\,,$$

where C is the confidence interval for which you want to estimate errors in parameters and *prob*(**θ**) is the likelihood used for parameter estimation

See 1009.2755 for more details



Confidence intervals in case of χ^2 minimization

- Assume μ is the number of free parameters for which you want to plot the joint confidence interval and p% is the confidence limit desired .

- $\Delta\chi 2$ is distributed as a chi-square distribution with μ degrees of freedom. $\Delta\chi 2 = \chi 2 - \chi 2_{min}$

Calculate $\Delta\chi 2$ such that chi-square probability for μ free parameters is less than p.

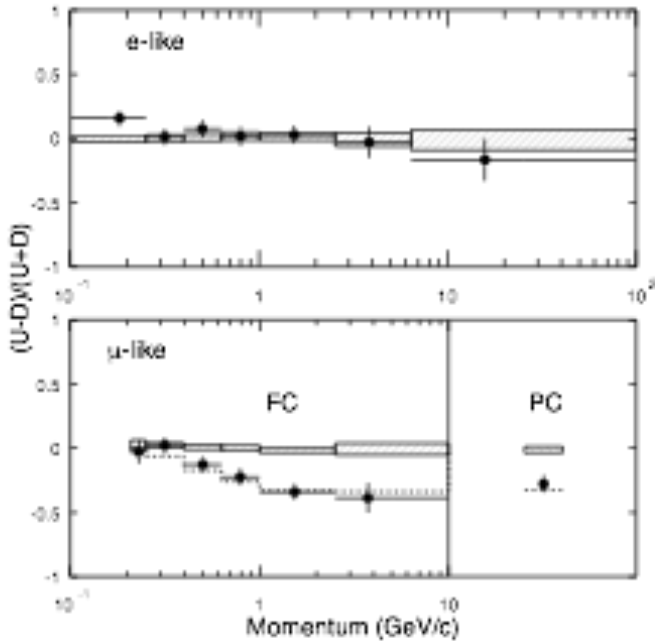
$P(\chi 2 < \Delta\chi 2 , \mu) = p\%$

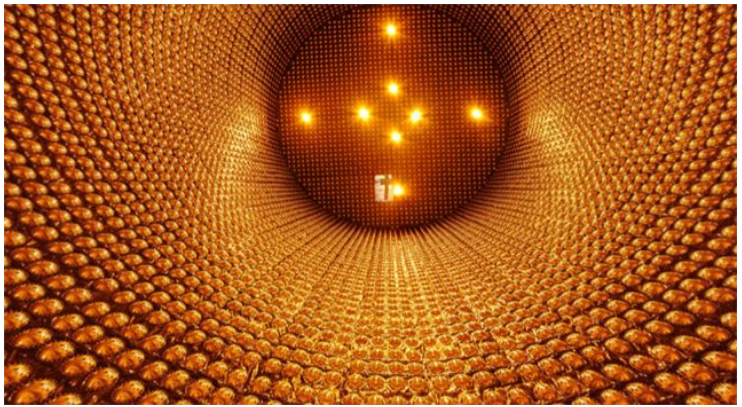
	ν					
p	1	2	3	4	5	6
68.3%	1.00	2.30	3.53	4.72	5.89	7.04
90%	2.71	4.61	6.25	7.78	9.24	10.6
95.4%	4.00	6.17	8.02	9.70	11.3	12.8
99%	6.63	9.21	11.3	13.3	15.1	16.8
99.73%	9.00	11.8	14.2	16.3	18.2	20.1
99.99%	15.1	18.4	21.1	23.5	25.7	27.8

$\Delta\chi 2$ tables as a function of number of free parameters (p) and confidence level (%)

Numerical recipes Sect 15.6

Examples from literature : Discovery of neutrino oscillations





Best-fit $\chi 2 = 65.2$ for 67 DOF

$$\chi^2 = \sum_{cos\Theta,p} (N_{DATA} - N_{MC})^2/\sigma^2 + \sum_j \epsilon_j^2/\sigma_j^2,$$

hep-ex/9807003

Example of confidence Intervals

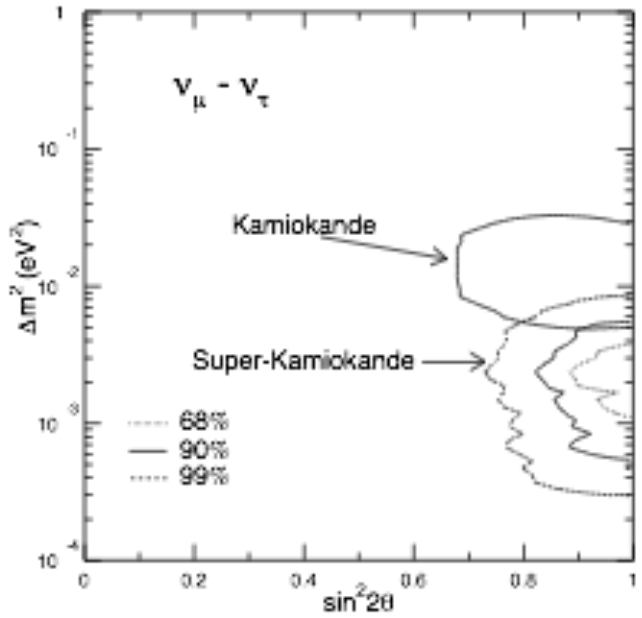
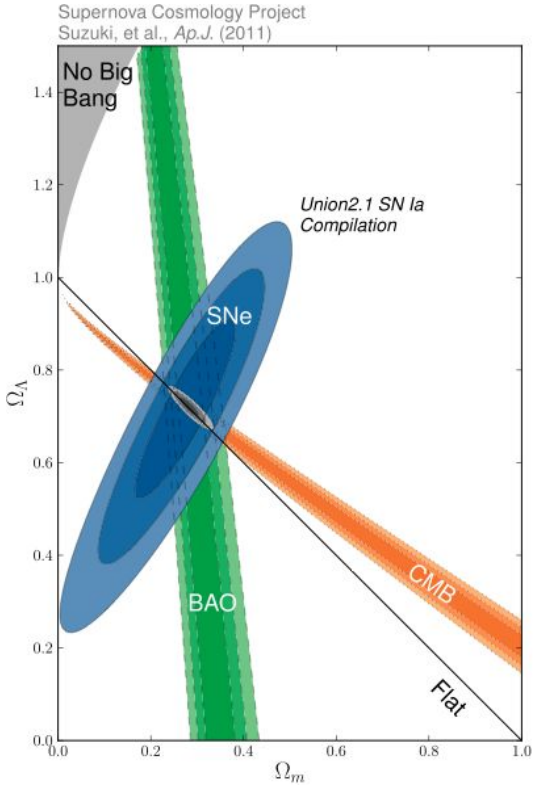


FIG. 2. The 68%, 90% and 99% confidence intervals are shown for $\sin^2 2\theta$ and Δm^2 for $\nu_\mu \leftrightarrow \nu_\tau$ two-neutrino oscillations based on 33.0 kiloton-years of Super–Kamiokande data. The 90% confidence interval obtained by the Kamiokande experiment is also shown.

$\Delta\chi 2$ intervals of 2.6,5.0, and 9.6
Correspond to 68%,90% and 99%
confidence intervals

2015 Nobel Prize to Takaaki Kajita

Examples from literature : Evidence for dark energy



$$\chi^2_{\rm stat} = \sum_{\rm SNe} \frac{[\mu_B(\alpha,\beta,\delta,M_{\rm B}) - \mu(z;\Omega_m,\Omega_w,w)]^2}{\sigma^2_{\rm lc} + \sigma^2_{\rm ext} + \sigma^2_{\rm sample}}.$$

2011 Nobel Prize to Saul Perlmutter, Adam Riess and Brian Schmidt

Model Comparison

- Frequentist Model Comparison See Ganguly & SD arXiv:1706.01202, if interested

- Information Theory Based Model Comparison : AIC and BIC Ref: Liddle astro-ph/070113

$$\mathrm{BIC} \equiv -2 \ln \mathcal{L}_{\max} + k \ln N \,,$$

$$\mathrm{AIC} \equiv -2 \ln \mathcal{L}_{\max} + 2k \,,$$

$$\mathrm{AIC_c} = \mathrm{AIC} + \frac{2k(k+1)}{N-k-1} \,.$$

N = no of data points
K = no of free parameters

Model with the smaller value of AIC or BIC is the preferred value.

Other measure also exist such as TIC, WAIC, DIC Krishak & SD arXiv:2003.10127

Strength of Evidence tests for AIC/BIC

ΔBIC	Evidence against Model i
0 – 2	Not Worth More Than A Bare Mention
2 – 6	Positive
6 – 10	Strong
> 10	Very Strong

ΔAIC	Level of Support For Model i
0 – 2	Substantial
4 – 7	Considerably Less
> 10	Essentially None

See arXiv:1901.07726 Kerscher & Weller for a holistic view of ALL model selection techniques

Difference between Frequentist and Bayesian viewpoints

Frequentist	Bayesian
Probabilities refer to relative frequencies of events	Refer to degree of subjective belief, not limiting frequency
Parameters are fixed unknown constants. Probability statements about parameters are meaningless	Probability statements can be made about things other than data including the model parameters and models themselves.
Frequentists consider model parameters to be fixed and data to be random	Bayesians consider data to be fixed and model parameters to be random
Confidence intervals should have well-defined long run frequency properties. 95% c.i. should bracket the true value of the parameter with a limiting frequency of at least 95%.	Inferences about a parameter are made by producing its probability. *Distribution quantifies amount of uncertainty of our knowledge about the parameter.* These are called credible intervals.

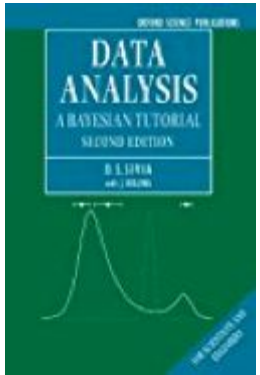
(Few) References on Bayesian Analysis

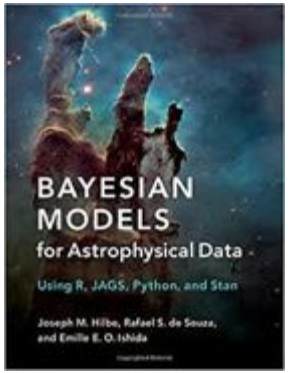
arXiv:0803.4089 and arXiv:1701.01467 Roberto Trotta ✔

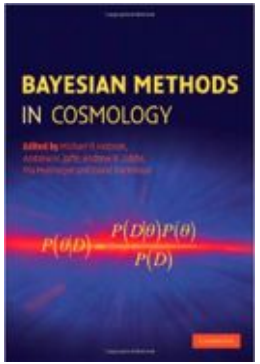
arXiv:0911.3105 Licia Verde

arXiv:1208.3036 Tom Loredo (check out his Bayesian statistics lectures in astrostatistics summer school at Penn State)

arXiv:1411.5018 - Jake Van Der Plas Practical Introduction (*Easy to read and digest along with code)* ✔







Basics of Bayesian Parameter Estimation (cf. Trotta)

$$P(\theta|d) = \frac{P(d|\theta)P(\theta)}{P(d)}.$$

P($\boldsymbol{\theta}$|d) is the posterior

P(d| $\boldsymbol{\theta}$) is the likelihood

P($\boldsymbol{\theta}$) is the prior

$$P(d) = \int d\theta P(d|\theta)P(\theta)$$

How to handle nuisance parameters in Bayesian analysis

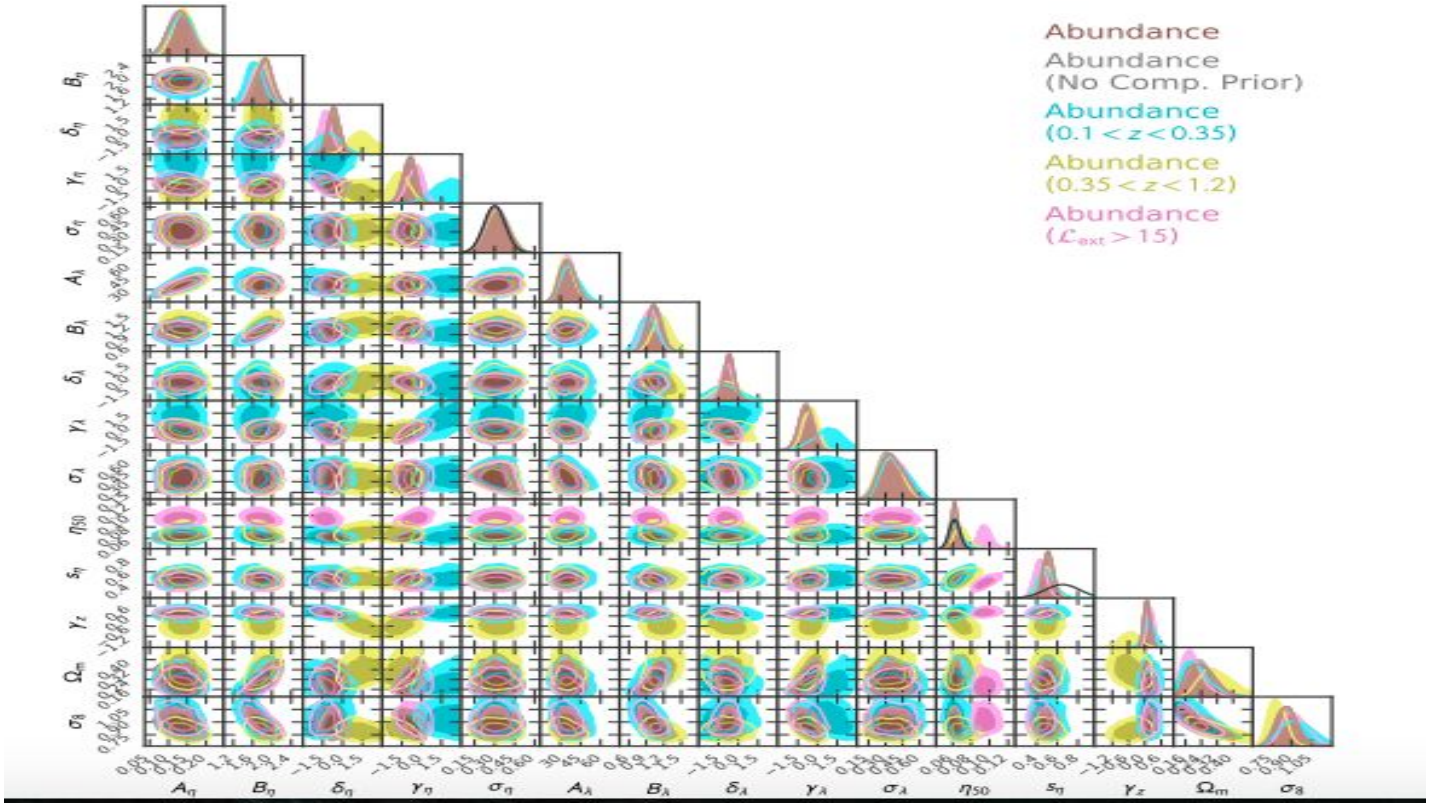
The posterior pdf for one parameter at the time is obtained by *marginalization*, i.e., by integrating over the uninteresting parameters. E.g., assume the the vector of parameters is given by $\theta = \{\phi, \psi\}$, then the 1D posterior pdf for ϕ alone is given by

$$p(\phi|d) \propto \int \mathscr{L}(\phi, \psi) p(\phi, \psi) \mathrm{d}\psi. \tag{79}$$

The final inference on ϕ from the posterior can then be communicated by plotting $p(\phi|d)$, with the other components marginalized over.

Frequentists handle nuisance parameters through Profile Likelihood (talk to experimental particle physics colleagues on how this is done)

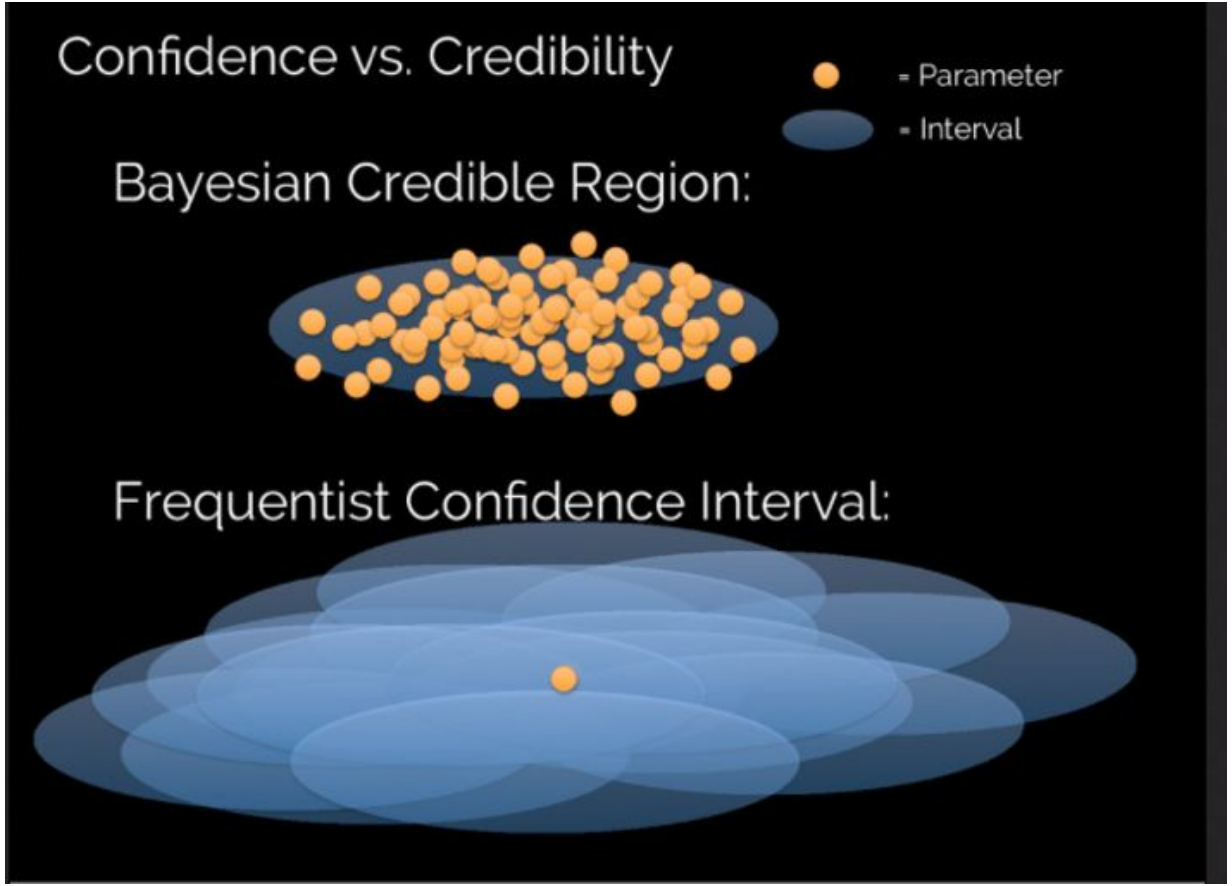
Example of Marginalization (arXiv:2207.12429)



Bayesian uncertainty estimation

- Evaluated in the same way as frequentist parameter estimation, except use posterior instead of likelihood
- For >= 2 parameters, contours obtained are called ``Credible'' Intervals instead of confidence intervals.

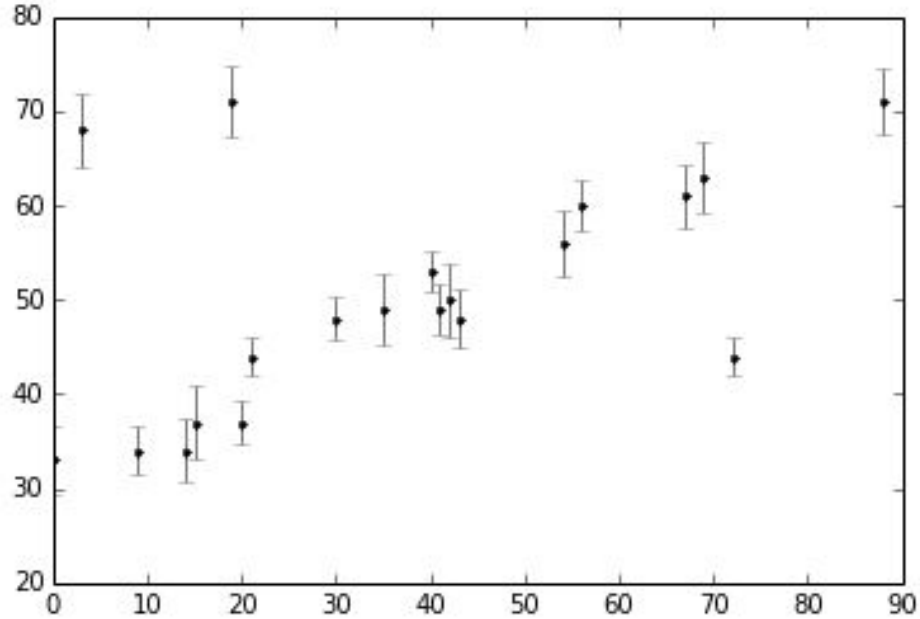
Difference between confidence and credible intervals



Credit: Jake Van Der Plas

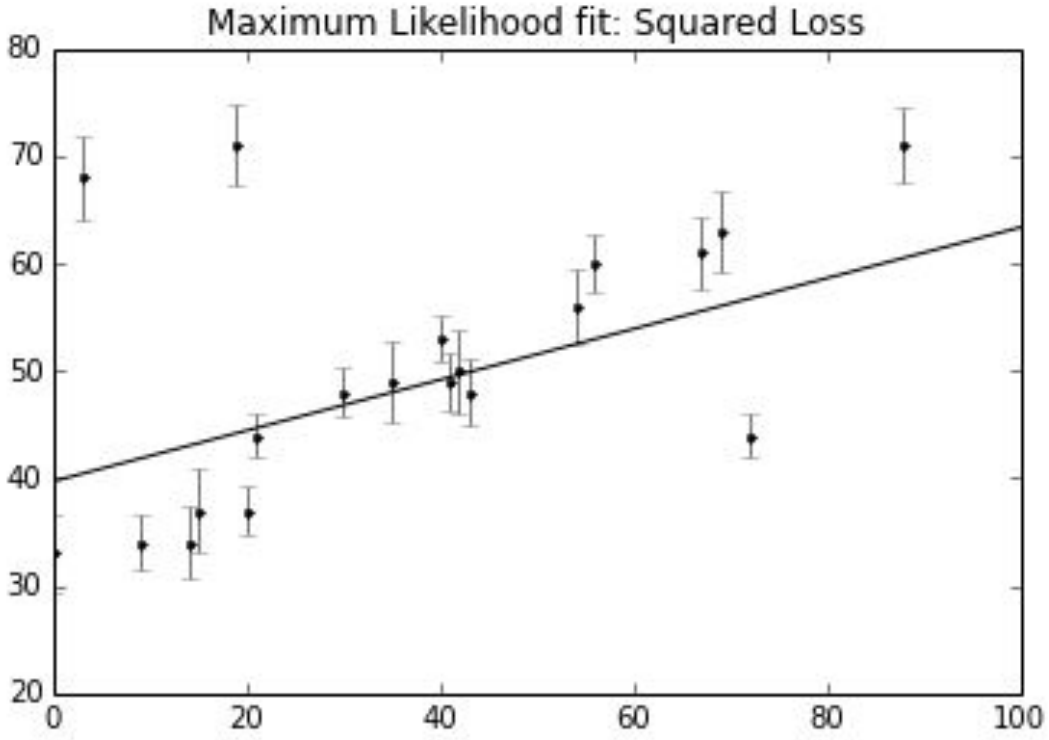
Example of Bayesian parameter estimation

Example : Linear fit with outliers *Credit: Jake Van der Plas (Pythonic Perambulations blog)*



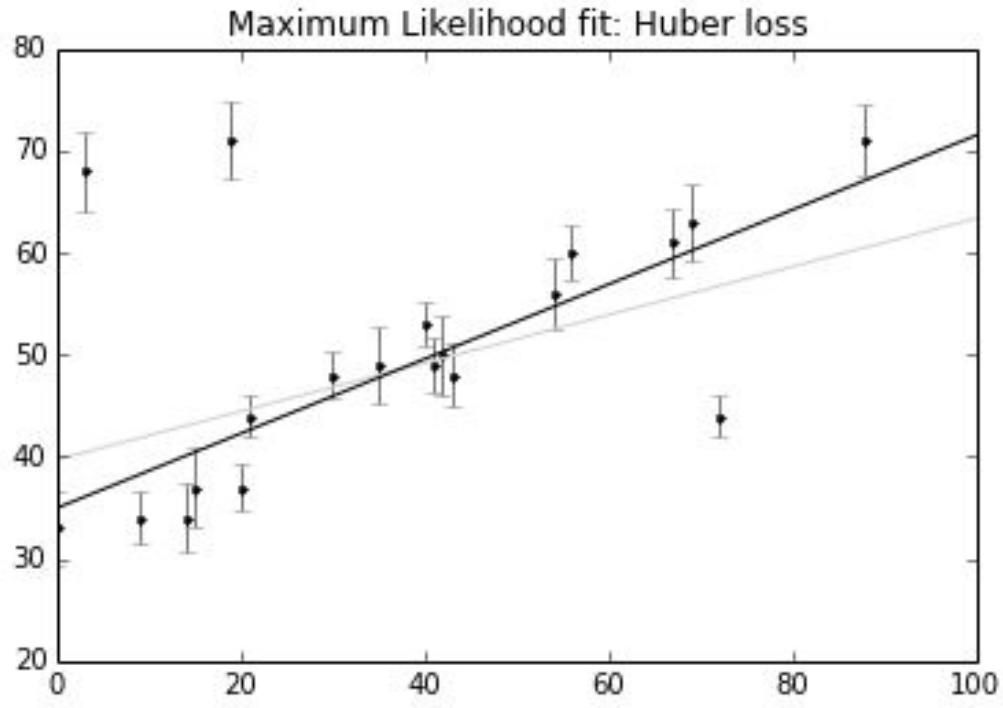
$$\hat{y}(x \mid \theta) = \theta_0 + \theta_1 x$$

Best-Fit “Textbook” Frequentist analysis



$$\text{loss} = \sum_i \frac{1}{2e_i^2}\left(y_i - \hat{y}(x_i \mid \theta)\right)^2$$

Results from Frequentist (Huber) Regression



Bayesian Analysis of this problem

Define an extended likelihood

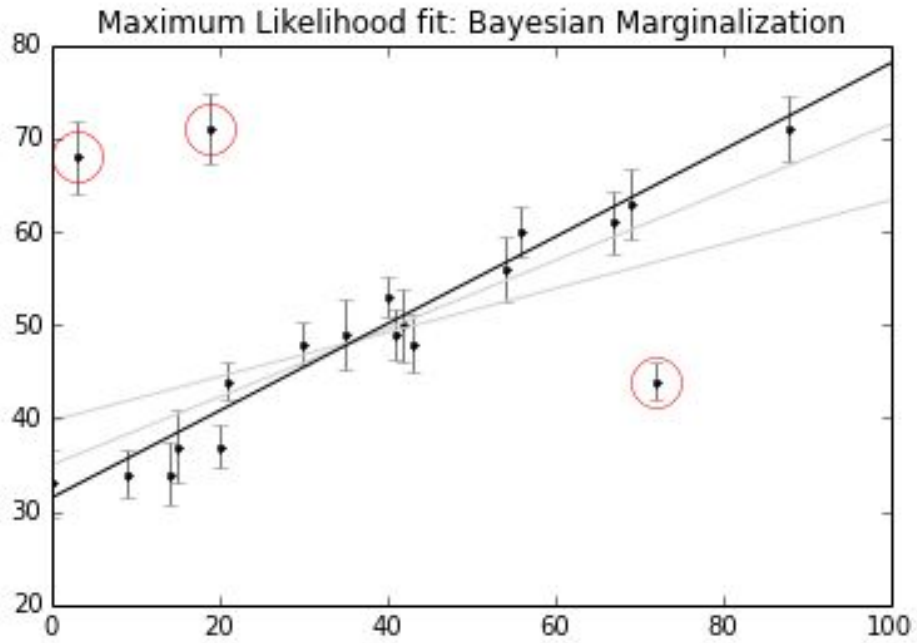
$$p(\{x_i\},\{y_i\},\{e_i\} \mid \theta,\{g_i\},\sigma,\sigma_b) = \frac{g_i}{\sqrt{2\pi e_i^2}}\exp\left[\frac{-(\hat{y}(x_i \mid \theta)-y_i)^2}{2e_i^2}\right] + \frac{1-g_i}{\sqrt{2\pi\sigma_B^2}}\exp\left[\frac{-(\hat{y}(x_i \mid \theta)-y_i)^2}{2\sigma_B^2}\right]$$

$g_{i=}1$ indicates data point is an outlier
g_i=0 indicates error bars are correct
$\boldsymbol{\sigma}_B$=50 (nuisance parameter)

Number of free parameters increases from 2 to 22

Assume uniform priors on all parameters between 0 and 1

Results from Bayesian analysis



Outliers identified by $g_i > 0.5$

Conclusions: Process of Bayesian marginalization gives a value closer to intuition and simultaneously allows us to identify outliers

Bayesian Goodness of Fit

????????

See however Lucy : arxiv:1511.02363

Bayesian Model Comparison

Ref: Kerscher & Weller arXiv:1901.07726 Krishak & SD arXiv:2003.10127

$$E(M) \equiv p(D|M) = \int p(D|M,\theta)p(\theta|M)d\theta$$

Is called Evidence or Marginal Likelihood or Bayesian Evidence

For comparing models M_1 and M_2

Odds Ratio becomes :

$$O_{21} = \frac{E(M_2)p(M_2)}{E(M_1)p(M_1)} = B_{21}\frac{p(M_2)}{p(M_1)}$$

where $$B_{21} = \frac{\int p(D|M_2,\theta_2)p(\theta_2|M_2)d\theta_2}{\int p(D|M_1,\theta_1)p(\theta_1|M_1)d\theta_1}$$ Bayes Factor

Interpretation of odds ratio/Bayes factor

- Jeffreys scale used as a qualitative guide to decide between model 2 and model 1

K	dHart	bits	Strength of evidence
$< 10^0$	< 0	< 0	Negative (supports M_2)
10^0 to $10^{1/2}$	0 to 5	0 to 1.6	Barely worth mentioning
$10^{1/2}$ to 10^1	5 to 10	1.6 to 3.3	Substantial
10^1 to $10^{3/2}$	10 to 15	3.3 to 5.0	Strong
$10^{3/2}$ to 10^2	15 to 20	5.0 to 6.6	Very strong
$> 10^2$	> 20	> 6.6	Decisive

Credit: wikipedia

Note that some works (eg Trotta 2017) use 150 instead of 100 for decisive evidence

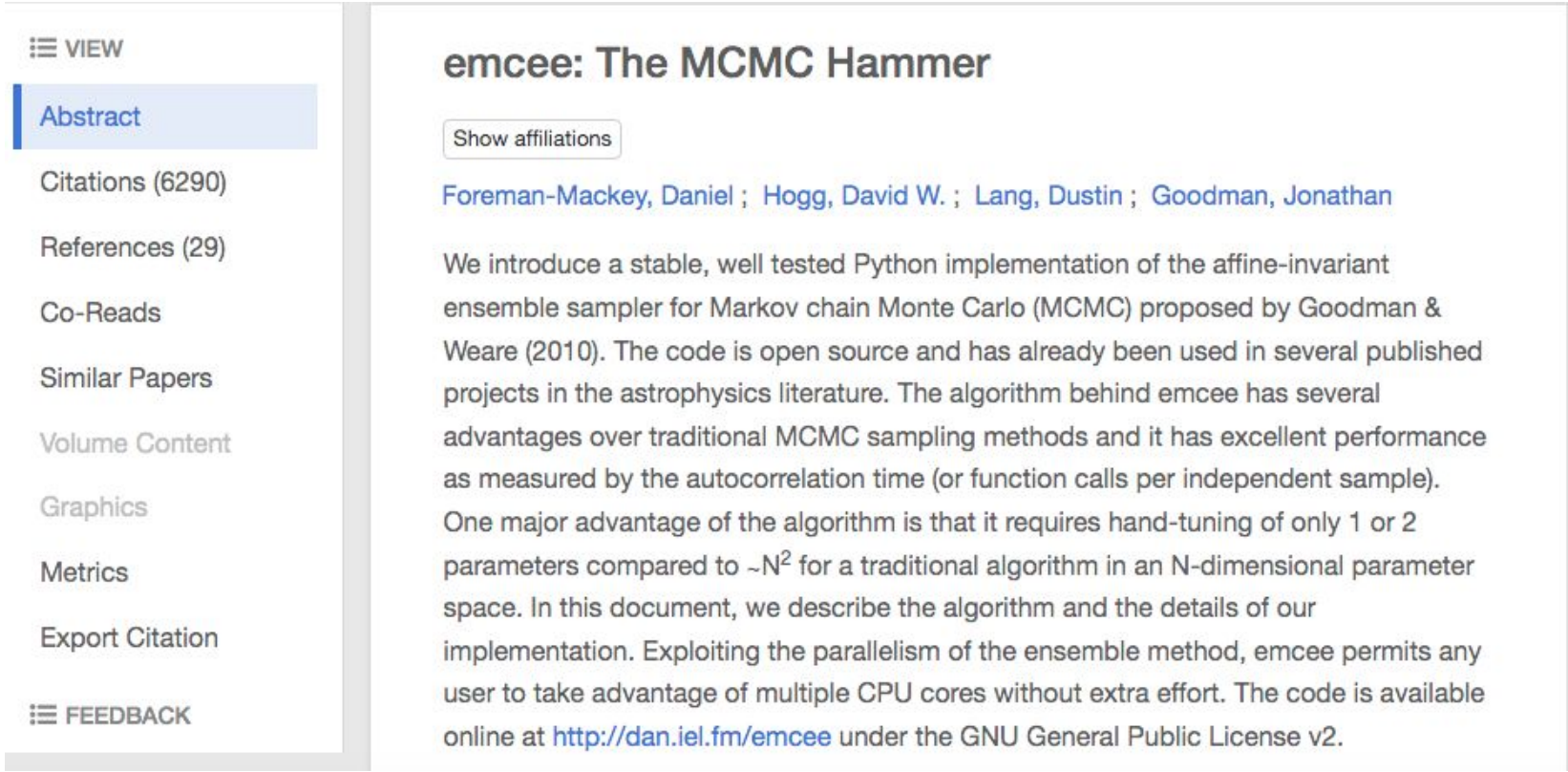
MCMC and Nested Sampling

MCMC : Sanjib Sharma arXiv:1706.01629 David Hogg & Foreman-Mackay 1710.06068
Nested Sampling : arXiv:2205.15570

- MCMC is the magic bullet needed to unleash the power of Bayesian analysis
- Originally designed to sample multi-modal likelihoods, Bayesian regression is the home turf for application of MCMC
- However it has also been (ab)used for frequentist analysis and for optimization
- Nested sampling is a numerical algorithm used for computation of multi-dimensional integration needed for Bayesian Model selection

http://mattpitkin.github.io/samplers-demo/pages/samplers-samplers-everywhere/ - All of you should bookmark this

Most widely used MCMC code - emcee



Pip install emcee

How to run MCMC in 2-3 steps

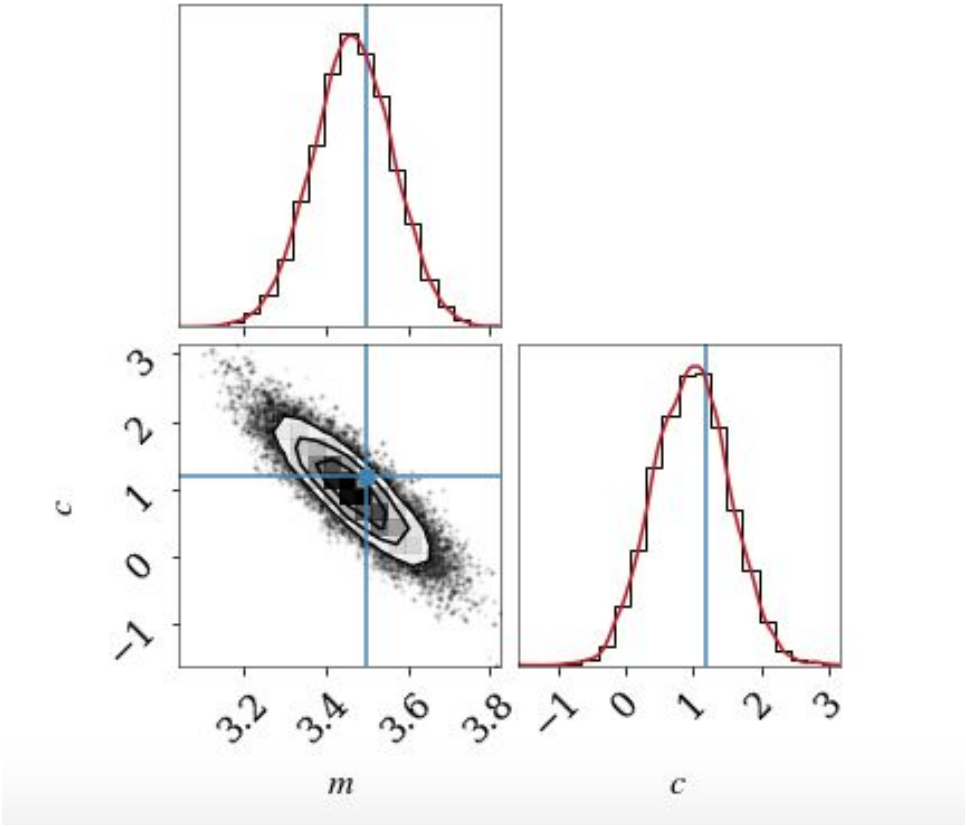
-> Write a function for log likelihood
-> Write a function for log prior
-> Define log posterior= log likelihood+ log prior
-> Choose initial guesses for free parameters
Steps for running emcee

```
sampler = emcee.EnsembleSampler(Nens, ndims, logposterior, args=argslist)
sampler.run_mcmc(inisamples, Nsamples + Nburnin);
samples_emcee = sampler.get_chain(flat=True, discard=Nburnin)
```

Post-processing of MCMC samples - corner, getdist, chainconsumer

```
fig = corner.corner(samples_emcee)
```

Example of MCMC corner plot



Most widely & (easy to use) Nested samplers

- Nestle
- Dynesty

```
sampler = NestedSampler(loglikelihood_dynesty, prior_transform, ndims,bound=bound,
sample=sample, nlive=nlive)
sampler.run_nested(dlogz=tol, print_progress=False)
res = sampler.results # get results dictionary from sampler
logZdynesty = res.logz[-1]
logZerrdynesty = res.logzerr[-1]
```

Advanced topics in Bayesian Analysis not covered

- Hierarchical Bayesian analysis (See Sanjib's paper)
- Likelihood-free inference (``forward modelling'') See arXiv:2109.05941 by Tam, Umetsu & Amara

Open questions/limitations/incorrect usage in Bayesian Analysis

- Criticism of Bayesian model selection in cosmology by G. Efsthathiou (0802.3185) and in particle physics by R. Cousins (0807.1330) , where results are sensitive to choice of priors. which are based unjustified assumptions.
- Criticism of Model Selection astro-ph/0702542 (See https://cosmocoffee.info/viewtopic.php?f=2&t=831 for discussions on this paper including rejoinders)
- How to deal with Jeffreys-Lindley paradox (1310.3791 & 1303.5973)
- Many posteriors used in Bayesian model selection are improper (1712.03549)
- One should also consider fluctuations in Bayesian evidence (2102.09547 & 2111.04231). See also Andrew Zic's latest paper
- Jeffreys scale is not always reliable (2102.09547)
- Is there such a thing as correct prior?