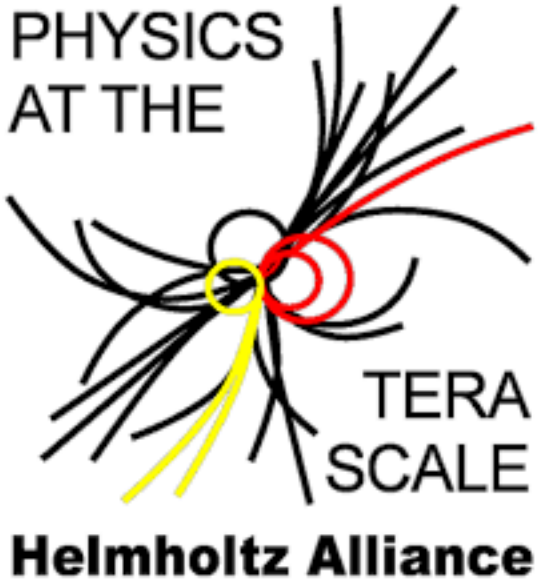
Statistics for Particle Physics
Lecture 3



Terascale Statistics School

https://indico.desy.de/event/46667/

DESY, Hamburg
24-28 Feb 2025

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

Outline

	Monday 15:30	Hypothesis testing
	Monday 16:30	Hypothesis testing tutorial
	Tuesday 10:30	Parameter estimation tutorial
	Tuesday 14:00	Setting limits
→	Wednesday 9:00	Bayesian parameter estimation
	Wednesday 11:00	Errors on errors

Reminder of 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 $p(x|\theta)$.

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

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

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

ML and Bayesian estimators

Purist Bayesian: $p(\theta|x)$ contains all knowledge about θ.

Pragmatist Bayesian: $p(\theta|x)$ could be a complicated function,

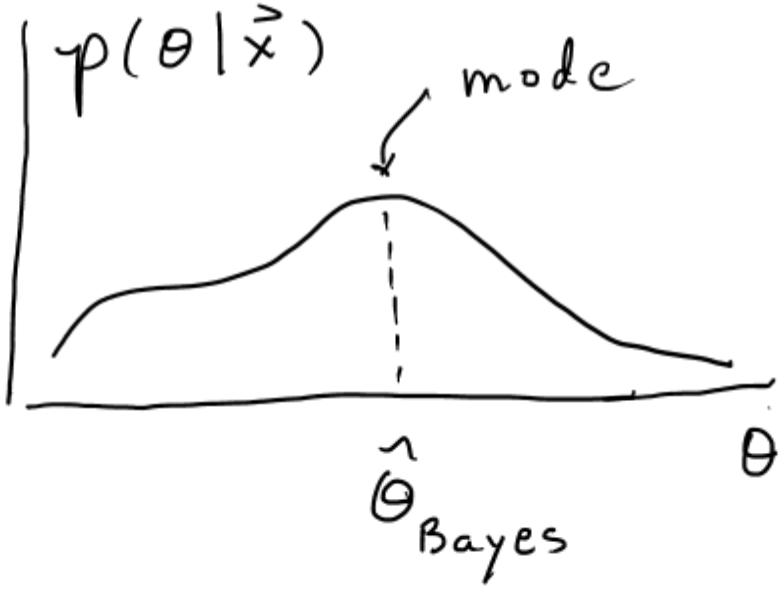
→ summarize using an estimator $\hat{\theta}_{\text{Bayes}}$

Take mode of $p(\theta|x)$, (could also use e.g. expectation value)

What do we use for $\pi(\theta)$?
No golden rule (subjective!),
often represent 'prior ignorance'
by $\pi(\theta)$ = constant, in which case

$$p(\theta|\mathbf{x}) \propto P(\mathbf{x}|\theta) = L(\theta)$$

$$\longrightarrow \quad \hat{\theta}_{\text{Bayes}} = \hat{\theta}_{\text{ML}}$$



ML and Bayesian estimators (2)

Note $\pi(\theta)$ = const. cannot be normalized – "improper prior".

Can be allowed for some problems; prior always appears multiplied by likelihood, so product $L(\theta)\pi_\theta(\theta)$ can result in normalizable posterior probability.

But... we could have used a different parameter, e.g., $\lambda = 1/\theta$, and if prior $\pi(\theta)$ is constant, then $\pi_\lambda(\lambda)$ is not:

$$\pi_\lambda(\lambda) = \pi_\theta(\theta)\left|\frac{d\theta}{d\lambda}\right| \propto \frac{1}{\lambda^2}$$

Maybe we know say we nothing about λ, so take $\pi_\lambda(\lambda)$ = const.

Then $\hat{\lambda}_{\rm Bayes} = \hat{\lambda}_{ML} \neq \dfrac{1}{\hat{\theta}_{\rm Bayes}}$

'Complete prior ignorance' is not well defined.

Example: fitting a straight line

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

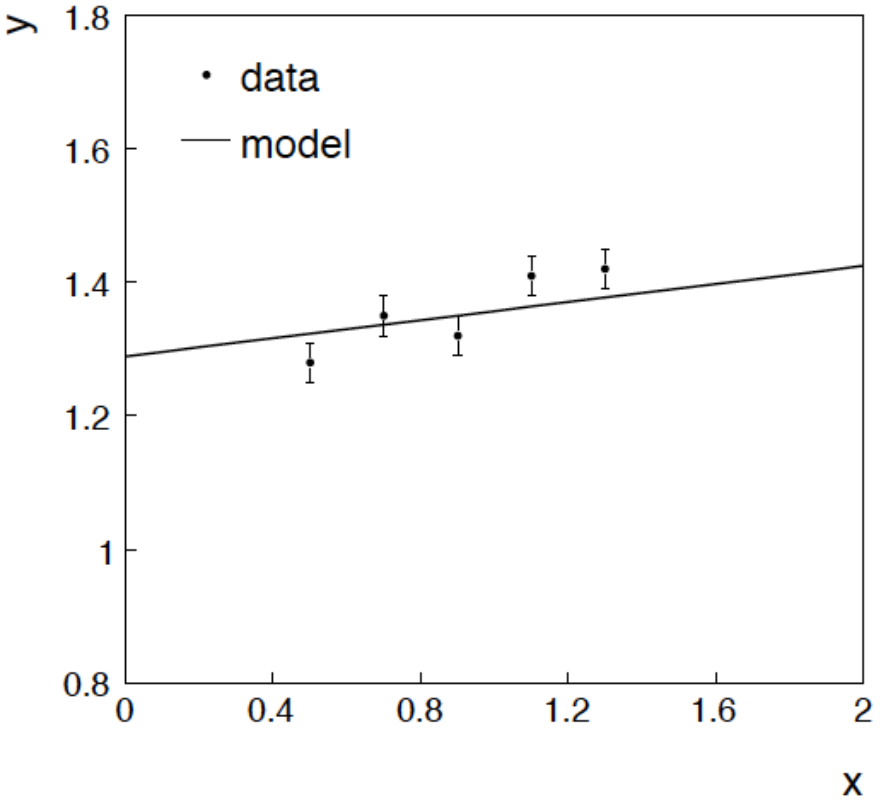
Model: y_i independent and all follow $y_i \sim \text{Gauss}(\mu(x_i), \sigma_i)$

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

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

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

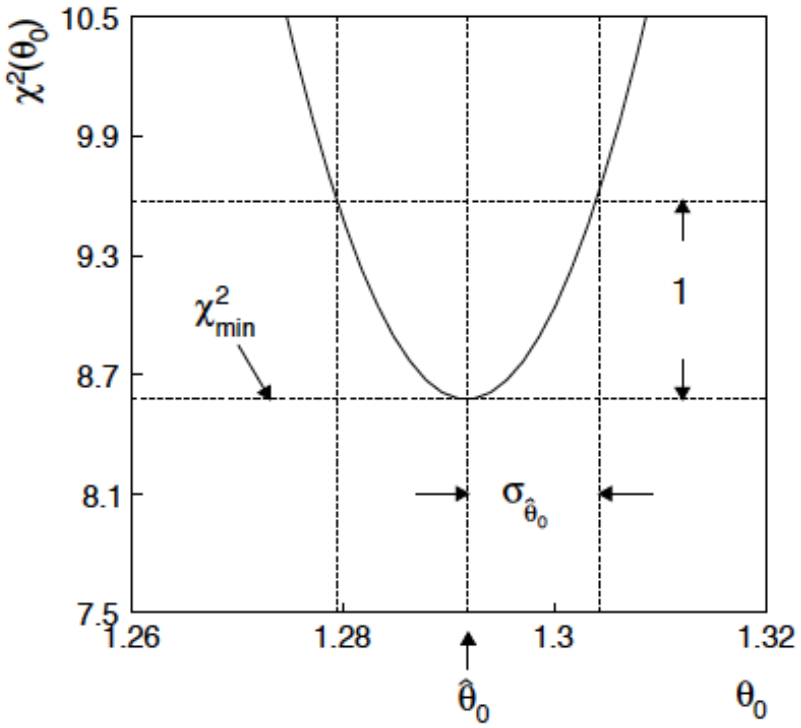
$$\chi^2(\theta_0) = -2\ln L(\theta_0) + \text{const} = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i;\theta_0,\theta_1))^2}{\sigma_i^2} .$$

For Gaussian y_i, ML same as LS

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

Come up one unit from χ^2_{min}

to find $\sigma_{\hat{\theta}_0}$.



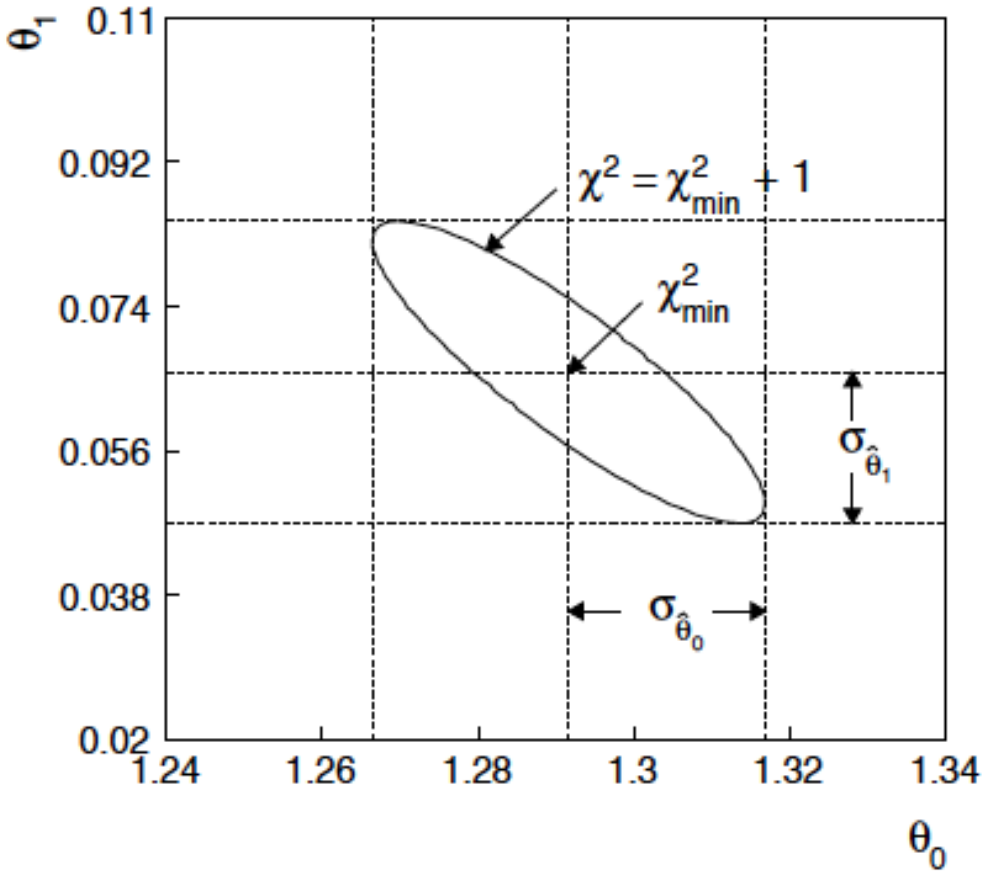
ML (or LS) fit of θ_0 and θ_1

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

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.

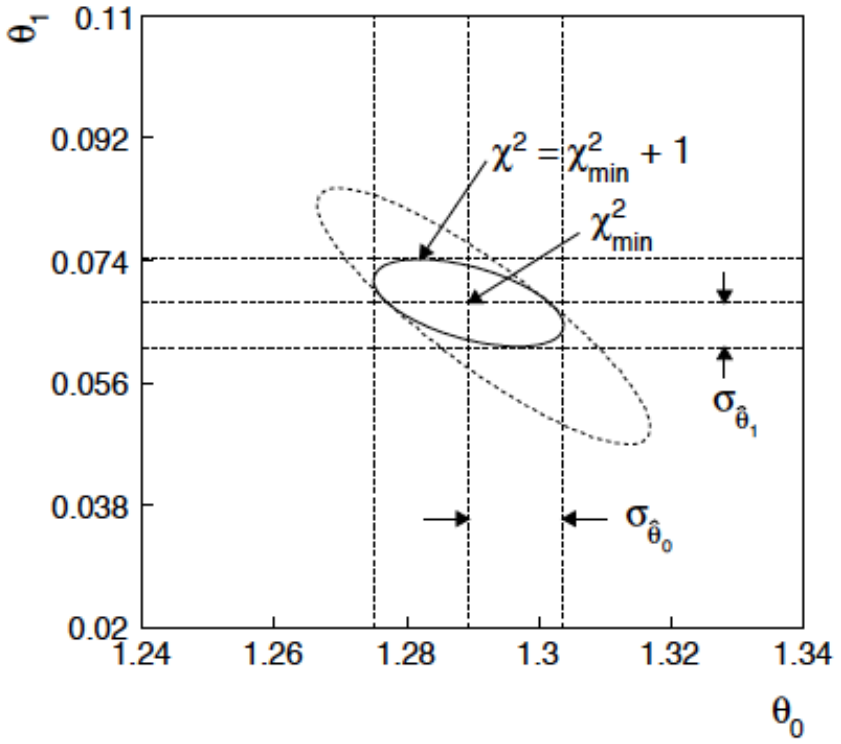


If we have a measurement $t_1 \sim \text{Gauss}\,(\theta_1, \sigma_{t_1})$

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

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

The information on θ_1 improves accuracy of $\hat{\theta}_0$.



Bayesian approach: $y_i \sim \text{Gauss}(\mu(x_i;\theta_0,\theta_1), \sigma_i)$

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

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

← suppose knowledge of θ_0 has no influence on knowledge of θ_1

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

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

$$\pi_1(\theta_1) = p(\theta_1|t_1) \propto p(t_1|\theta_1)\pi_{\text{Ur}}(\theta_1) = \frac{1}{\sqrt{2\pi}\sigma_t} e^{-(t_1-\theta_1)^2/2\sigma_t^2} \times \text{const.}$$

prior after t_1, before $\boldsymbol{y}$

Ur = "primordial" prior

Likelihood for control measurement t_1

Bayesian example: $y_i \sim \text{Gauss}(\mu(x_i;\theta_0,\theta_1), \sigma_i)$

Putting the ingredients 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

Note here the likelihood only reflects the measurements $\boldsymbol{y}$.

The information from the control measurement t_1 has been put into the prior for θ_1.

We would get the same result using the likelihood $P(\boldsymbol{y},t|\theta_0,\theta_1)$ and the constant "Ur-prior" for θ_1.

Here posterior only found as a proportionality.

Marginalizing the posterior pdf

We then integrate (marginalize) $p(\theta_0, \theta_1|\mathbf{y})$ to find $p(\theta_0|\mathbf{y})$:

$$p(\theta_0|\mathbf{y}) = \int p(\theta_0, \theta_1|\mathbf{y})\, d\theta_1$$

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

$$p(\theta_0|\mathbf{y}) = \frac{1}{\sqrt{2\pi}\sigma_{\theta_0}} e^{-(\theta_0 - \hat{\theta}_0)^2/2\sigma_{\theta_0^2}}$$

$$\hat{\theta}_0 = \text{same as MLE}$$

$$\sigma_{\theta_0} = \sigma_{\hat{\theta}_0} \qquad \text{(same as for MLE)}$$

For this example, numbers come out same as in frequentist approach, but interpretation different.

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 $\boldsymbol{\theta}$ but look only at distribution of parameters of interest.

MCMC basics: Metropolis-Hastings algorithm

Goal: given an n-dimensional pdf $p(\boldsymbol{\theta})$ up to a proportionality constant, generate a sequence of points $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \ldots$

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

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

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

3) Form 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).

Still works if $p(\boldsymbol{\theta})$ is known only as a proportionality, which is usually what we have from Bayes' theorem: $p(\boldsymbol{\theta}|\boldsymbol{x}) \propto p(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

The proposal density can be (almost) anything, but choose so as to minimize autocorrelation. Often take proposal density symmetric: $q(\boldsymbol{\theta}; \boldsymbol{\theta}_0) = q(\boldsymbol{\theta}_0; \boldsymbol{\theta})$

Test ratio is (*Metropolis*-Hastings): $\alpha = \min\left[1, \frac{p(\vec{\theta})}{p(\vec{\theta}_0)}\right]$

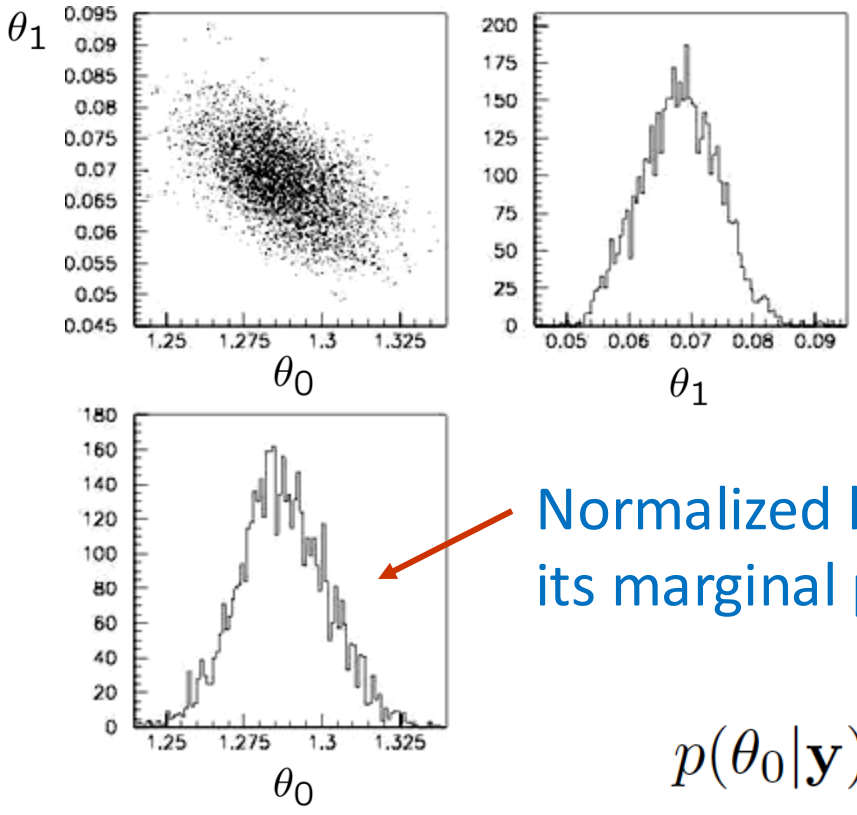
I.e. if the proposed step is to a point of higher $p(\boldsymbol{\theta})$, take it;

if not, only take the step with probability $p(\boldsymbol{\theta})/p(\boldsymbol{\theta}_0)$.

If proposed step rejected, repeat the current point.

Example: posterior pdf from MCMC

Sample the posterior pdf from previous example with MCMC:



Normalized histogram of θ_0 gives its marginal posterior pdf:

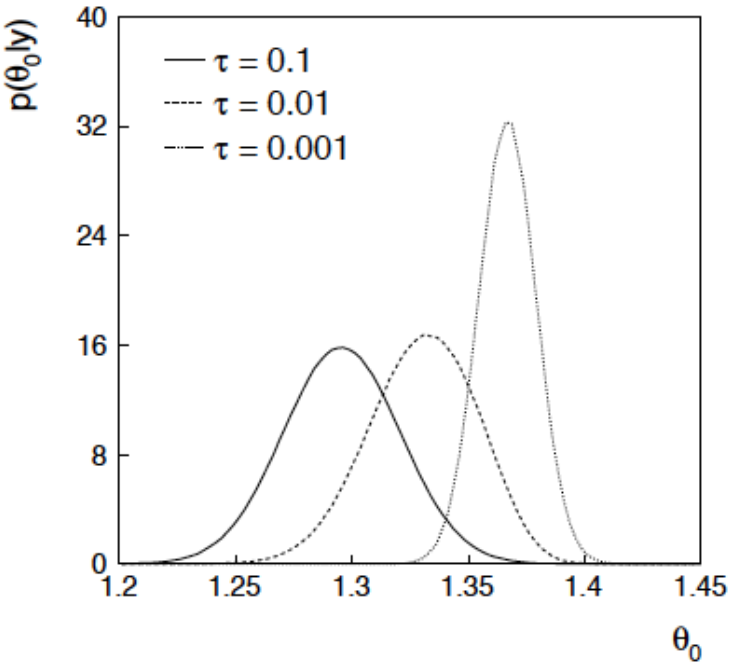
$$p(\theta_0|\mathbf{y}) = \int p(\theta_0, \theta_1|\mathbf{y})\, d\theta_1$$

Bayesian method with alternative priors

Suppose we don't have a previous measurement of θ_1 but rather, an "expert" 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.

Tutorial: Bayesian parameter estimation

The exercise is described

https://www.pp.rhul.ac.uk/~cowan/stat/exercises/bayesFit/

in the file bayes_fit_exercise.pdf.

The program is in bayesFit.py or bayesFit.ipynb.

This exercise treats the same fitting problem as seen with maximum likelihood, here using the Bayesian approach.

Bayes' theorem is used to find the posterior pdf for the parameters, and these are summarized using the posterior mode (MAP estimators).

The posterior pdf is marginalized over the nuisance parameters using Markov Chain Monte Carlo.

Gaussian signal on exponential background

Same pdf as from mlFit.py (see tutorial 1) with $n = 400$ independent values of x from

$$f(x|\boldsymbol{\lambda}) = \theta \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} + (1-\theta)\frac{1}{\xi} e^{-x/\xi}$$

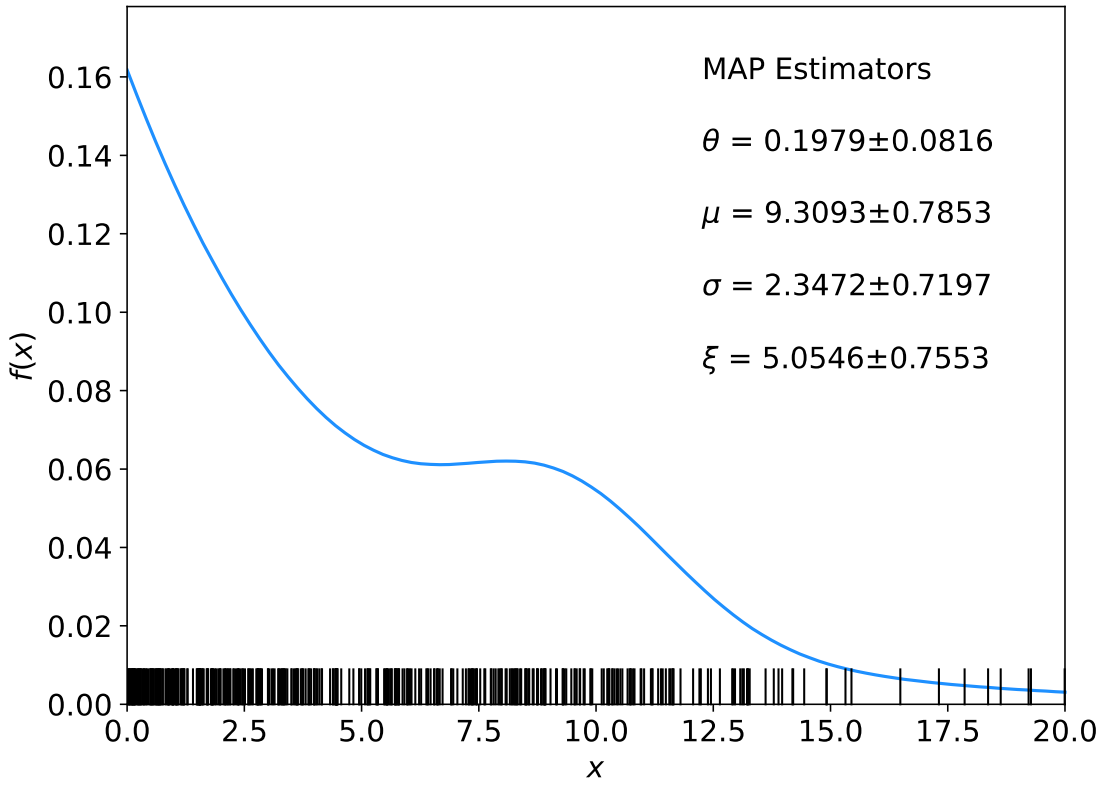
Posterior pdf for parameters $\boldsymbol{\lambda} = (\theta, \mu, \sigma, \xi)$ from Bayes theorem,

$$p(\boldsymbol{\lambda}|\mathbf{x}) \propto p(\mathbf{x}|\boldsymbol{\lambda})\pi(\boldsymbol{\lambda}) \,, \quad \text{where} \quad p(\mathbf{x}|\boldsymbol{\lambda}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\lambda})$$

At first take prior pdf constant for all parameters subject to $0 \leq \theta \leq 1$, $\sigma > 0$, $\xi > 0$ (later try different priors).

Data and MAP estimates

Maximize posterior with minuit (minimize $-\ln p(\boldsymbol{\lambda}|\boldsymbol{x})$).



Standard deviations from minuit correspond to approximating posterior as Gaussian near its peak.

Here priors constant so MAP estimates same as MLE, covariance matrix $V_{ij} = \mathrm{cov}[\theta_i, \theta_j]$ also same.

A look at bayesFit.py

Find maximum of posterior with iminuit (minimize $-\ln p(\boldsymbol{\lambda}|\boldsymbol{x})$), similar to maximum likelihood:

```
# Negative log-likelihood
def negLogL(par):
    fx = f(xData, par)
    return -np.sum(np.log(fx))

# Prior pdf
def prior(par):
    theta    = par[0]
    mu       = par[1]
    sigma    = par[2]
    xi       = par[3]
    pi_theta = 1. if theta >= 0. and theta <= 1. else 0.
    pi_mu    = 1. if mu >= 0. else 0.
    pi_sigma = 1. if sigma > 0. else 0.
    pi_xi    = 1. if xi > 0. else 0.
    piArr = np.array([pi_theta, pi_mu, pi_sigma, pi_xi])
    pi = np.product(piArr[np.array(parfix) == False])   # exclude fixed par
    return pi

# Negative log of posterior pdf
def negLogPost(par):
    return negLogL(par) - np.log(prior(par))
```

minimize with iminuit

Metropolis-Hastings algorithm in bayesFit.py

```
# Iterate with Metropolis-Hastings algorithm
chain = [np.array(MAP)]       # start point is MAP estimate
numIterate = 10000
numBurn = 100
numAccept = 0
print("Start MCMC iterations: ", end="")
while len(chain) < numIterate:
    par = chain[-1]
    log_post = -negLogL(par) + np.log(prior(par))
    par_prop = np.random.multivariate_normal(par, cov_prop)
    if prior(par_prop) <= 0:
        chain.append(chain[-1])    # never accept if prob<=0.
    else:
        log_post_prop = -negLogL(par_prop) + np.log(prior(par_prop))
        alpha = np.exp(log_post_prop - log_post)
        u = np.random.uniform(0, 1)
        if u <= alpha:
            chain.append(par_prop)
            numAccept += 1
        else:
            chain.append(chain[-1])
        if len(chain)%(numIterate/100) == 0:
            print(".", end="", flush=True)
chain = np.array(chain)
```

Try increasing number of iterations (10k runs in about 20 s).

Exercises on Bayesian parameter estimation (a)

1a) Run bayesFit.py, look at the plots

1(a) Run the program and examine the plots. These include:

1. The data values as ticks on the x axis together with the fitted curve evaluated with MAP estimators (Fig. 1 below). The uncertainties on the parameters correspond to the covariance $V_{ij} = \mathrm{cov}[\lambda_i, \lambda_j]$ that `iminuit` finds by approximating the posterior as a multivariate Gaussian near its maximum (similar to finding the covariance matrix of the MLEs).

2. Trace plots of each of the parameters (Fig. 2). In some problems it can be useful to discard a subset of the points (called "burn-in") if the starting point $\boldsymbol{\lambda}_0$ is too far from the main concentration of the target density's probability; this is indicated in the trace plots with a vertical yellow bar.

3. Marginal distributions of the individual parameters (Fig. 3). The histograms are normalized to unit area and the MAP estimates are indicated with the vertical bars.

4. The autocorrelation function for the parameters (Fig. 4).

Exercises on Bayesian parameter estimation (b,c)

1b) Investigate effect of data sample size, fixing parameters and length of MCMC chains.

1(b) Change the data sample size from $n = 400$ to 200 and 1000 and note the changes in the results.

Using again $n = 400$, fix the parameters μ and σ (by changing the corresponding elements in the array `parfix` from `False` to `True`) and note the changes in the results. When finished, go back to having all four parameters free.

Change the number of MCMC iterations from 10 000 to 100 000 and note the change in the results, particularly in the structures you see in the trace plots. (This probably takes some time to run; for the rest of the exercises it is probably best to change back to 10 000 iterations.

1c) Investigate changing the prior

1(c) Change the prior pdfs for ξ and σ to be $\pi(\xi) \propto 1/\xi$ and $\pi(\sigma) \propto 1/\sigma$ and note the change in the results. When finished, go back to constant priors.

Exercises on Bayesian parameter estimation (d)

1d) Include auxiliary measurement to constrain ξ

1(d) Suppose that one has an independent estimate u of the parameter ξ in addition to the $n = 400$ values of x. Treat u as Gaussian distributed with a mean ξ and standard deviation $\sigma_u = 0.5$ and take the observed value $u = 5$. Find the log-likelihood function that includes both the primary measurements $(x_1, \ldots, x_n)$ and the auxiliary measurement u and modify the fitting program accordingly. Investigate how the results are affected by including u.

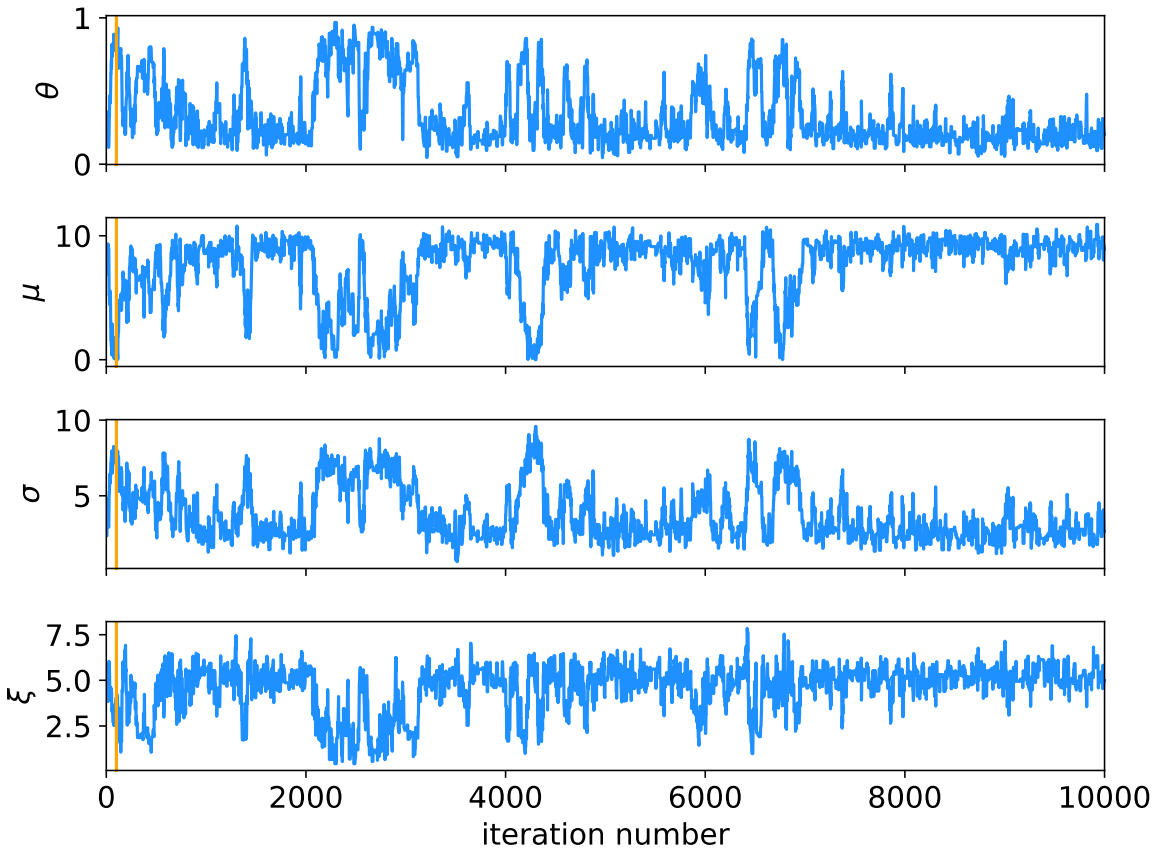
1e) Investigate point and interval estimates for θ

1(e) Using the functions `cc_interval` and `HPD_interval` provided in `bayesFit.py`, compute the central credible interval and HPD (highest probability density) interval for the parameter of interest θ using a credibility level of 68.3%. Compare these to the intervals one obtains from a point estimate (the MAP estimate, posterior median or posterior mean) plus or minus one standard deviation. For the standard deviation, try using both the sample standard deviation from the MCMC values and the standard deviation found by `iminuit`, which is based on a Gaussian approximation to the peak of the posterior. Find the estimates and intervales both with and without the auxiliary measurement of ξ as in (d) above and note how this effects the results.

MCMC trace plots

Take θ as parameter of interest, rest are nuisance parameters.

Marginalize by sampling posterior pdf with Metropolis-Hastings.

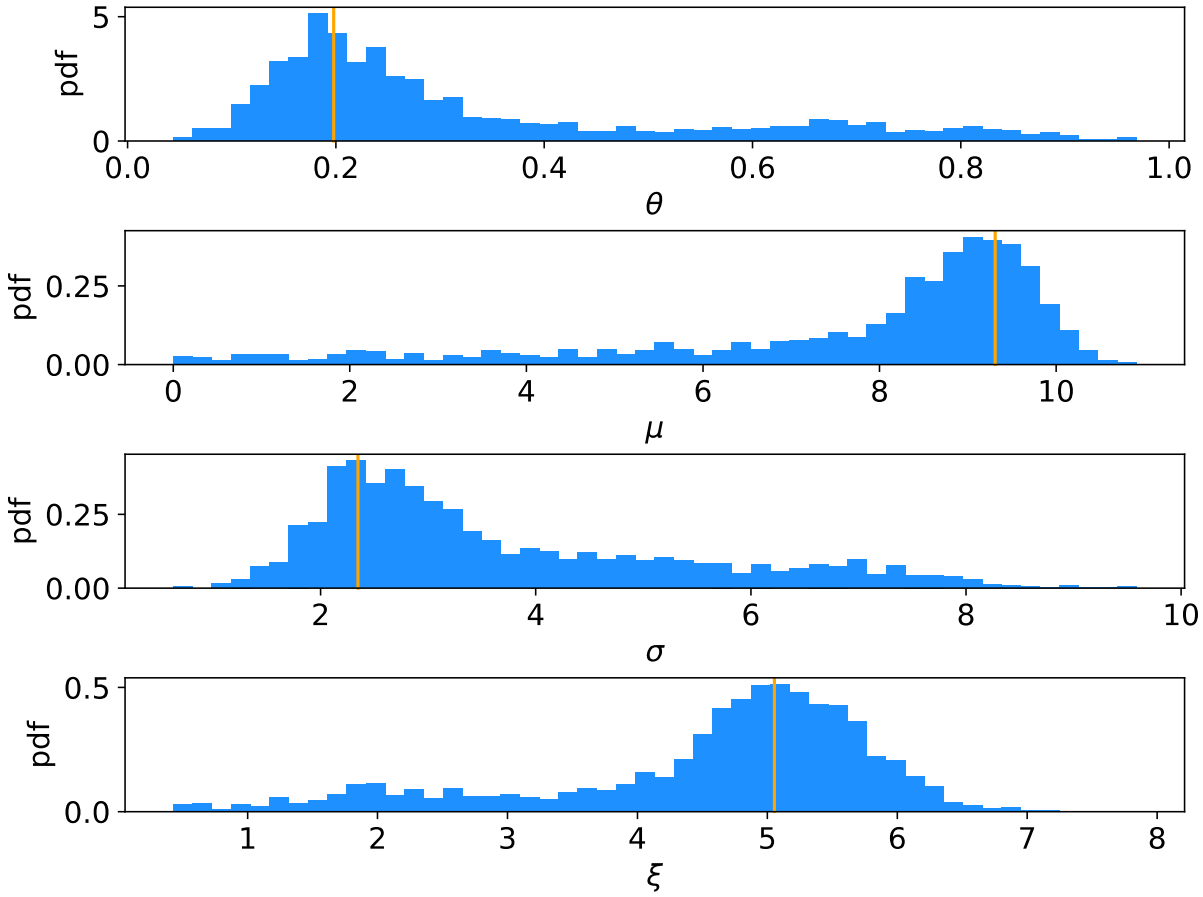


Gaussian proposal pdf, covariance $U = sV$, $s = (2.38)^2/N_{\text{par}} = 1.41$, gives acceptance probability ~ 0.24.

Here 10000 iterations (should use more).

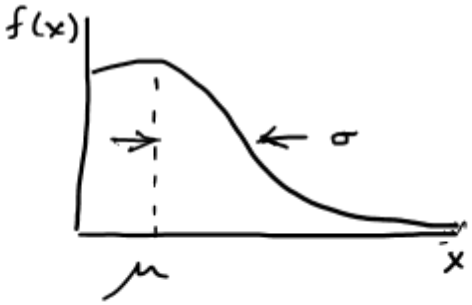
Marginal distributions

MAP estimates shown with vertical bars



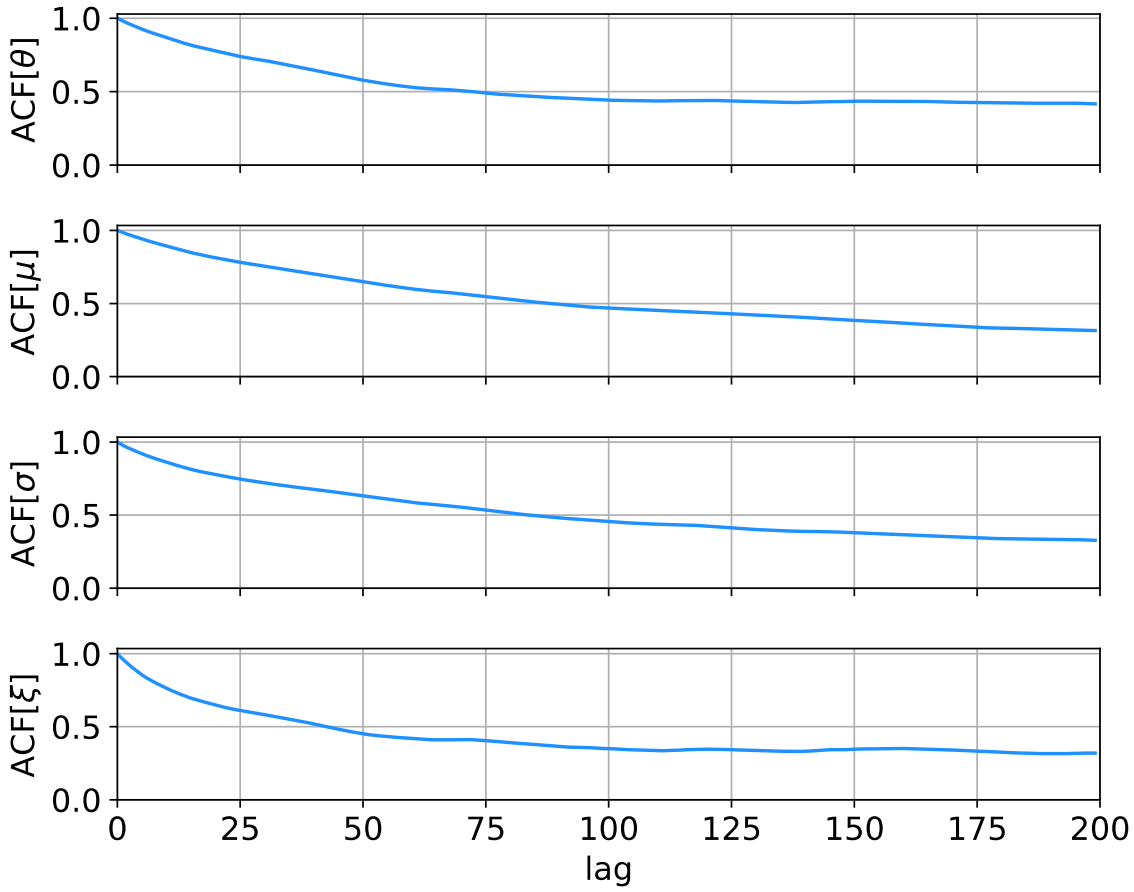
Note long tails.

Interpretation: data distribution can be approximated by Gaussian term only, (θ large, μ small) with large width ($\sigma \sim 4$-8) and a narrow exponential ($\xi \sim 1$-3).



Autocorrelation versus lag

MCMC samples are not independent, autocorrelation function = correlation coefficient of sample x_i with x_{i+l} as a function of the lag, l, where x = any of θ, μ, σ, ξ minus its mean:



$$\mathrm{ACF} = \frac{1}{N}\sum_{i=1}^{N}\frac{x_i x_{i+\ell}}{\sigma^2}$$

Effective sample size

$$N_{\mathrm{eff}} = \frac{N}{1 + 2\sum_{\ell=1}^{\infty}\mathrm{ACF}_\ell}$$

In stat. error estimates

$$\frac{1}{\sqrt{N}} \to \frac{1}{\sqrt{N_{\mathrm{eff}}}}$$

Ways to summarize the posterior

Point estimates:

Posterior mode (MAP, coincides with MLE for constant prior).

Posterior median (invariant under monotonic transformation of parameter).

Posterior mean; coincides with above in large-sample limit.

Intervals:

Highest Probability Density (HPD) interval, shortest for a given probability content, not invariant under param. trans.

Central credible intervals, equal upper and lower tail areas, e.g., $\alpha/2$ for CL = $1 - \alpha$.

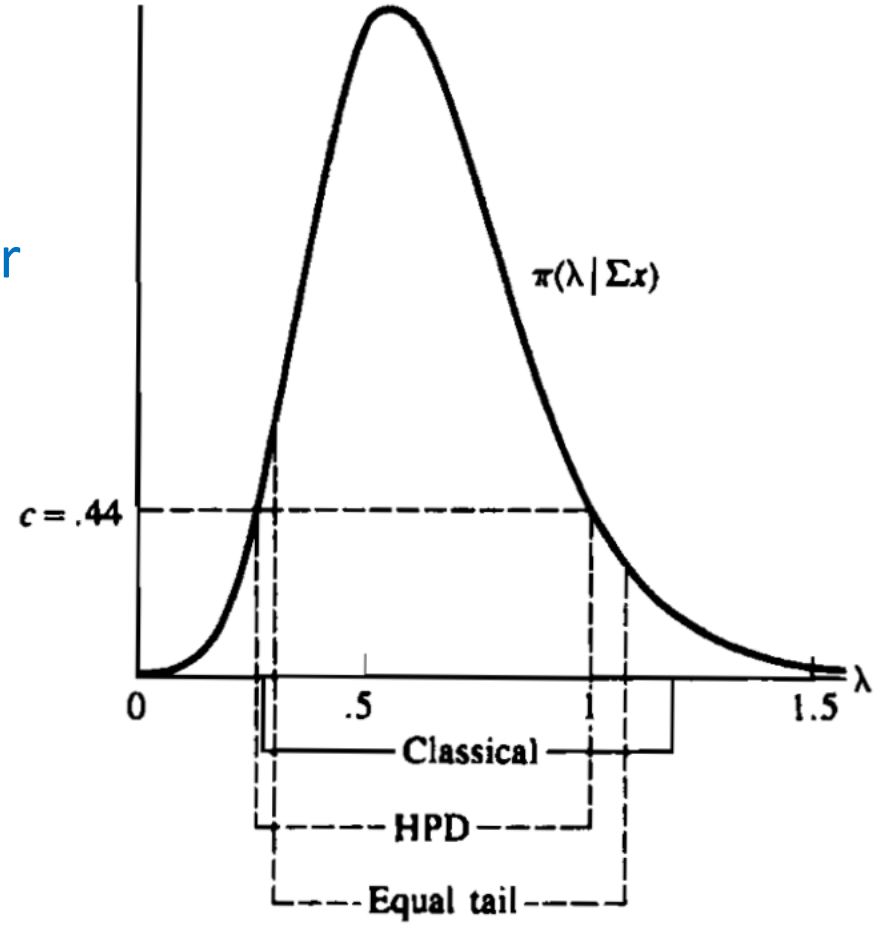
Point estimate +/- standard deviation, std. dev. from MCMC sample or by approximating core of posterior as Gaussian (from minuit); coincides with above in large-sample limit.

Types of intervals

HPD = Highest Posterior Density

Equal tail (central) from posterior

Classical (frequentist)



G. Casella and R. Berger, Statistical Inference, 2002