

Statistical Data Analysis 2023/24

Lecture Week 10



London Postgraduate Lectures on Particle Physics
University of London MSc/MSci course PH4515



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

Course web page via RHUL moodle (PH4515) and also
`www.pp.rhul.ac.uk/~cowan/stat_course.html`

Statistical Data Analysis

Lecture 10-1

- Upper limits on a Poisson rate parameter
 - Frequentist approach
 - Bayesian approach

Frequentist upper limit on Poisson parameter

Consider again observing $n \sim \text{Poisson}(s + b)$, with both $s, b \geq 0$.

Suppose $b = 4.5$, $n_{\text{obs}} = 5$. Find upper limit on s at 95% CL.

Relevant alternative is $s = 0$ (critical region at low n)

p -value of hypothesized s is $P(n \leq n_{\text{obs}}; s, b)$

Upper limit s_{up} at $\text{CL} = 1 - \alpha$ found from

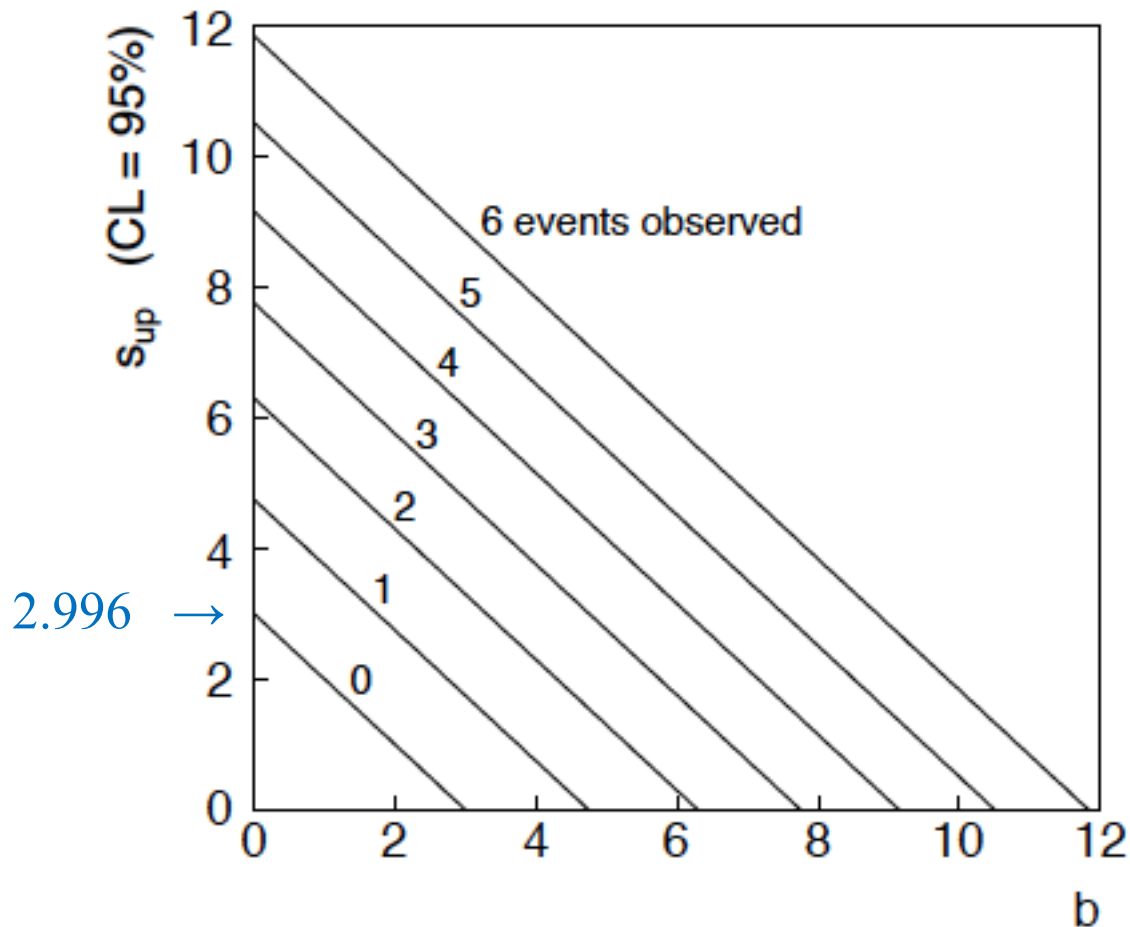
$$\alpha = P(n \leq n_{\text{obs}}; s_{\text{up}}, b) = \sum_{n=0}^{n_{\text{obs}}} \frac{(s_{\text{up}} + b)^n}{n!} e^{-(s_{\text{up}} + b)}$$

$$s_{\text{up}} = \frac{1}{2} F_{\chi^2}^{-1}(1 - \alpha; 2(n_{\text{obs}} + 1)) - b$$

$$= \frac{1}{2} F_{\chi^2}^{-1}(0.95; 2(5 + 1)) - 4.5 = 6.0$$

$n \sim \text{Poisson}(s+b)$: frequentist upper limit on s

For low fluctuation of n , formula can give negative result for s_{up} ; i.e. confidence interval is empty; all values of $s \geq 0$ have $p_s \leq \alpha$.



Limits near a boundary of the parameter space

Suppose e.g. $b = 2.5$ and we observe $n = 0$.

If we choose $CL = 0.9$, we find from the formula for s_{up}

$$s_{\text{up}} = -0.197 \quad (CL = 0.90)$$

Physicist:

We already knew $s \geq 0$ before we started; can't use negative upper limit to report result of expensive experiment!

Statistician:

The interval is designed to cover the true value only 90% of the time — this was clearly not one of those times.

Not uncommon dilemma – note probability to reject very small s (to which one has no sensitivity) is slightly greater than α .

Expected limit for $s = 0$

Physicist: I should have used $CL = 0.95$ — then $s_{up} = 0.496$

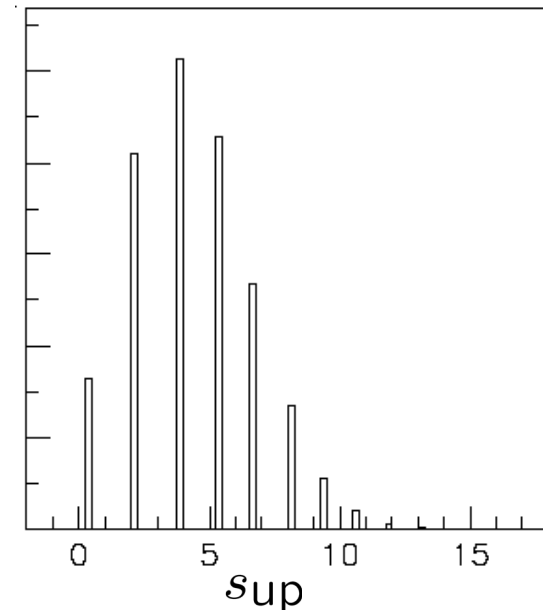
Even better: for $CL = 0.917923$ we get $s_{up} = 10^{-4}$!

Reality check: with $b = 2.5$, typical Poisson fluctuation in n is at least $\sqrt{2.5} = 1.6$. How can the limit be so low?

Look at the mean limit for the no-signal hypothesis ($s = 0$ sensitivity).

Distribution of 95% CL limits
with $b = 2.5, s = 0$.

Mean upper limit = 4.44



Report both observed and expected limits;
other types of limits (Bayesian, CLs, F-C) can mitigate the issue.

The Bayesian approach to limits

In Bayesian statistics need to start with ‘prior pdf’ $\pi(\theta)$, this reflects degree of belief about θ before doing the experiment.

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

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

Integrate posterior pdf $p(\theta|x)$ to give interval with any desired probability content.

For e.g. $n \sim \text{Poisson}(s+b)$, 95% CL upper limit on s from

$$0.95 = \int_{-\infty}^{s_{\text{sup}}} p(s|n) ds$$

Bayesian prior for Poisson parameter

Include knowledge that $s \geq 0$ by setting prior $\pi(s) = 0$ for $s < 0$.

Could try to reflect 'prior ignorance' with e.g.

$$\pi(s) = \begin{cases} 1 & s \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Not normalized; can be OK provided $p(n|s)$ dies off quickly for large s .

Not invariant under change of parameter — if we had used instead a flat prior for a nonlinear function of s , then this would imply a non-flat prior for s .

Doesn't really reflect a reasonable degree of belief, but often used as a point of reference; or viewed as a recipe for producing an interval whose frequentist properties can be studied (e.g., coverage probability, which will depend on true s).

Bayesian upper limit with flat prior for s

Put Poisson likelihood and flat prior into Bayes' theorem:

$$p(s|n) \propto p(n|s)\pi(s) = \frac{(s+b)^n}{n!} e^{-(s+b)} \times 1, \quad s \geq 0$$

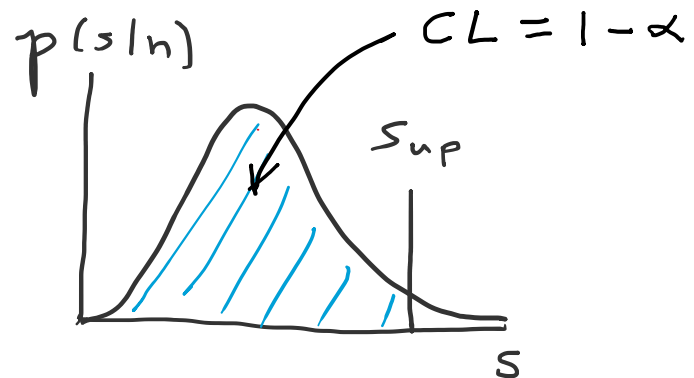
Normalize to unit area:

$$p(s|n) = \frac{(s+b)^n e^{-(s+b)}}{\Gamma(b, n+1)}$$

upper incomplete
gamma function

Upper limit s_{up} determined by

$$1 - \alpha = \int_0^{s_{\text{up}}} p(s|n) ds$$



Bayesian interval with flat prior for s

Solve to find limit s_{up} :

$$s_{\text{up}} = \frac{1}{2} F_{\chi^2}^{-1} [p, 2(n+1)] - b$$

where

$$p = 1 - \alpha \left(1 - F_{\chi^2} [2b, 2(n+1)] \right)$$

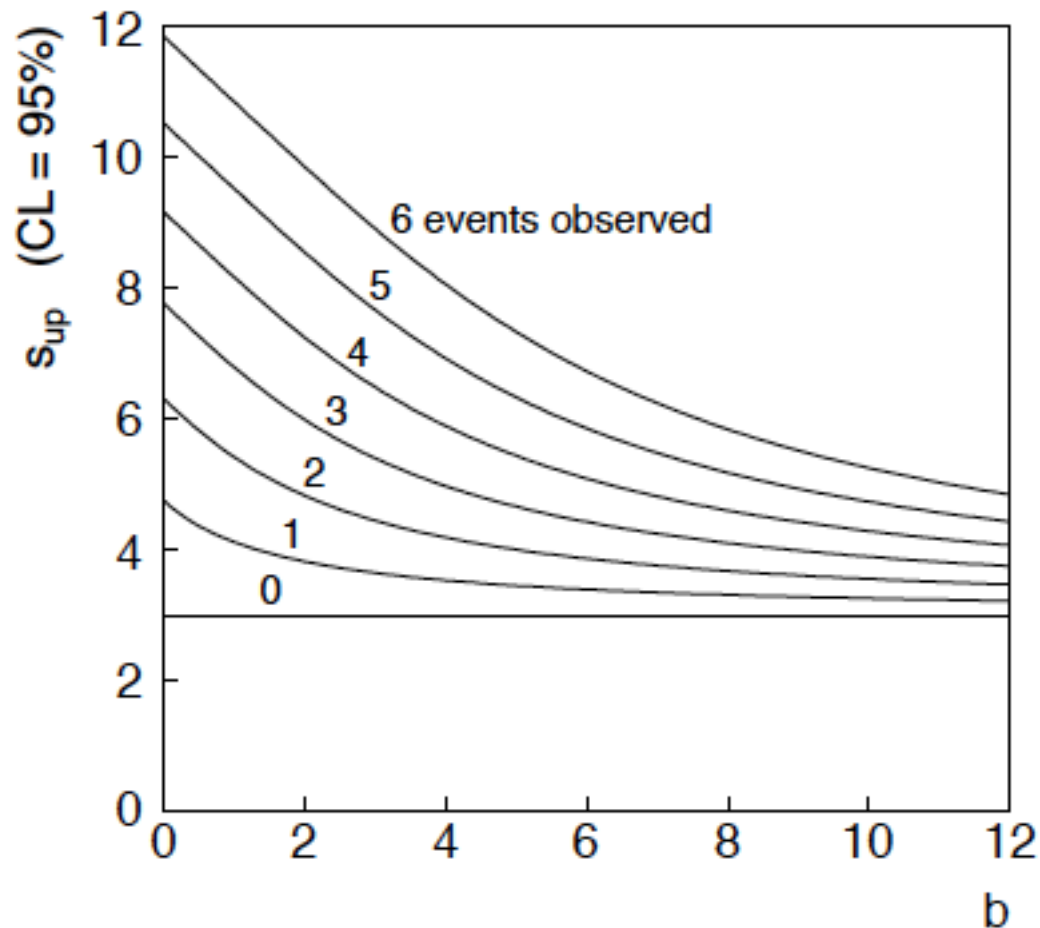
For special case $b = 0$, Bayesian upper limit with flat prior numerically same as one-sided frequentist case ('coincidence').

Bayesian interval with flat prior for s

For $b > 0$ Bayesian limit is everywhere greater than the (one sided) frequentist upper limit.

For $b = 0$, Bayesian and frequentist upper limits come out equal.

Never goes negative.
Doesn't depend on b if $n = 0$.



Statistical Data Analysis

Lecture 10-2

- Discussion on Bayesian prior probabilities
- Jeffreys' prior
- Example: Poisson mean

Priors from formal rules

Last time we took the prior for a Poisson mean to be constant to reflect a lack of prior knowledge; we noted this was not invariant under change of parameter.

Because of difficulties in encoding a vague degree of belief in a prior, one often attempts to derive the prior from formal rules, e.g., to satisfy certain invariance principles or to provide maximum information gain for a certain set of measurements.

Often called “objective priors”

Form basis of Objective Bayesian Statistics

The priors do not reflect a degree of belief (but might represent possible extreme cases).

In Objective Bayesian analysis, can use the intervals in a frequentist way, i.e., regard Bayes’ theorem as a recipe to produce an interval with a given coverage probability.

Priors from formal rules (cont.)

For a review of priors obtained by formal rules see, e.g.,

Robert E. Kass and Larry Wasserman, *The Selection of Prior Distributions by Formal Rules*, J. Am. Stat. Assoc., Vol. 91, No. 435, pp. 1343-1370 (1996).

Formal priors have not been widely used in Particle Physics, but there has been interest in this direction, especially the reference priors of Bernardo and Berger; see e.g.

L. Demortier, S. Jain and H. Prosper, *Reference priors for high energy physics*, Phys. Rev. D 82 (2010) 034002, arXiv:1002.1111.

D. Casadei, *Reference analysis of the signal + background model in counting experiments*, JINST 7 (2012) 01012; arXiv:1108.4270.

Jeffreys prior

According to *Jeffreys' rule*, take prior according to

$$\pi(\boldsymbol{\theta}) \propto \sqrt{\det(I(\boldsymbol{\theta}))}$$

where

$$I_{ij}(\boldsymbol{\theta}) = -E \left[\frac{\partial^2 \ln P(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right] = - \int \frac{\partial^2 \ln P(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} P(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x}$$

is the Fisher information matrix.

One can show that this leads to inference that is invariant under a transformation of parameters in the following sense:

Start with the Jeffreys prior for θ : $\pi_\theta(\theta) \sim \sqrt{(\det I(\theta))}$

Use it in Bayes' theorem to find:

$$P(\theta|\mathbf{x}) \propto P(\mathbf{x}|\theta)\pi_\theta(\theta)$$

Jeffreys prior (2)

Now consider a function $\eta(\theta)$. The posterior for η is

$$P(\eta|\mathbf{x}) = P(\theta|\mathbf{x}) \left| \frac{d\theta}{d\eta} \right|$$

Alternatively, start with η and use its Jeffreys' prior:

$$\pi_{\eta}(\eta) \propto \sqrt{\det I(\eta)}$$

Use this in Bayes' theorem: $P(\eta|\mathbf{x}) \propto P(\mathbf{x}|\eta)\pi_{\eta}(\eta)$

One can show that Jeffreys' prior results in the same $P(\eta|\mathbf{x})$ in both cases. For details (single-parameter case) see:

<https://www.pp.rhul.ac.uk/~cowan/stat/notes/JeffreysInvariance.pdf>

Jeffreys prior for Poisson mean

Suppose $n \sim \text{Poisson}(\mu)$. To find the Jeffreys' prior for μ ,

$$P(n|\mu) = \frac{\mu^n}{n!} e^{-\mu} \qquad \frac{\partial^2 \ln P(n|\mu)}{\partial \mu^2} = -\frac{n}{\mu^2}$$

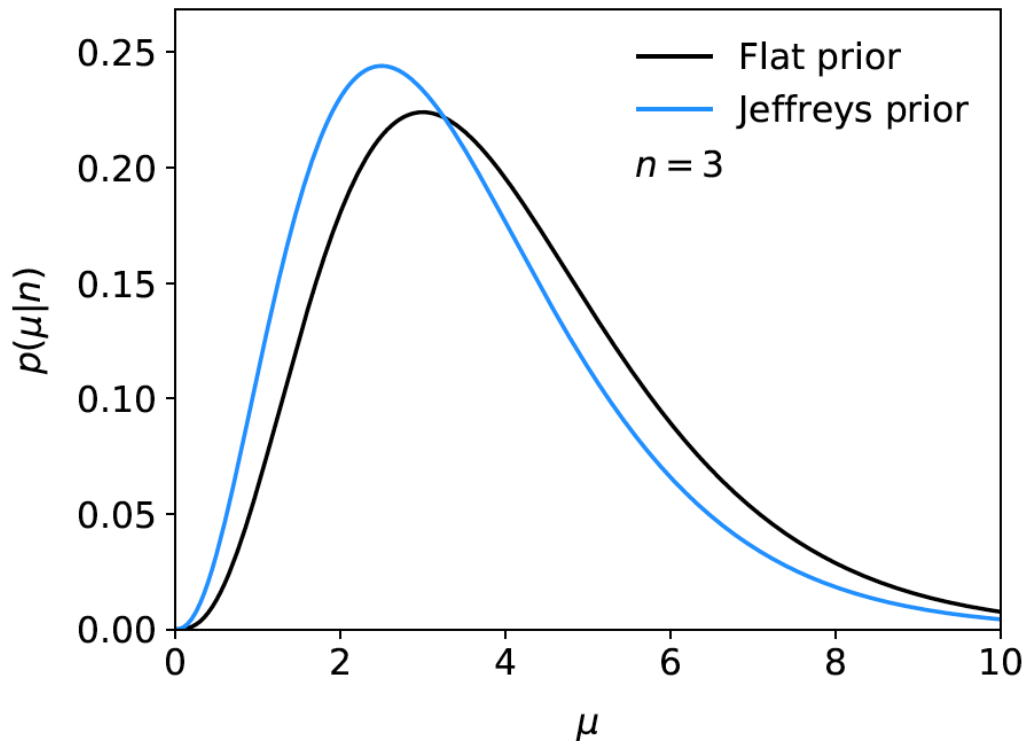
$$I = -E \left[\frac{\partial^2 \ln P(n|\mu)}{\partial \mu^2} \right] = \frac{E[n]}{\mu^2} = \frac{1}{\mu}$$

$$\pi(\mu) \propto \sqrt{I(\mu)} = \frac{1}{\sqrt{\mu}}$$

So e.g. for $\mu = s + b$, this means the prior $\pi(s) \sim 1/\sqrt{(s + b)}$, which depends on b . But this is not designed as a degree of belief about s .

Posterior pdf for Poisson mean

From Bayes' theorem, $p(\mu|n) \propto \mu^n e^{-\mu} \pi(\mu)$



Flat, $\pi(\mu) = \text{const.}$

$$p(\mu|n) = \frac{\mu^n e^{-\mu}}{\Gamma(n+1)}$$

Jeffreys, $\pi(\mu) \sim 1/\sqrt{\mu}$

$$p(\mu|n) = \frac{\mu^{n-\frac{1}{2}} e^{-\mu}}{\Gamma(n+\frac{1}{2})}$$

In both cases, posterior is special case of gamma distribution.

Upper limit for Poisson mean

To find upper limit at $CL = 1 - \alpha$, solve

$$1 - \alpha = \int_0^{\mu_{\text{up}}} p(\mu|n) d\mu$$

Jeffreys prior: $\mu_{\text{up}} = P^{-1}(n + \frac{1}{2}, 1 - \alpha) = 7.03$

Flat prior: $\mu_{\text{up}} = P^{-1}(n + 1, 1 - \alpha) = 7.75$

$n=3,$
 $CL=0.95$

where P^{-1} is the inverse of the normalized lower incomplete gamma function (see `scipy.special`)

$$P(a, \mu_{\text{up}}) = \frac{1}{\Gamma(a)} \int_0^{\mu_{\text{up}}} \mu^{a-1} e^{-\mu} d\mu$$

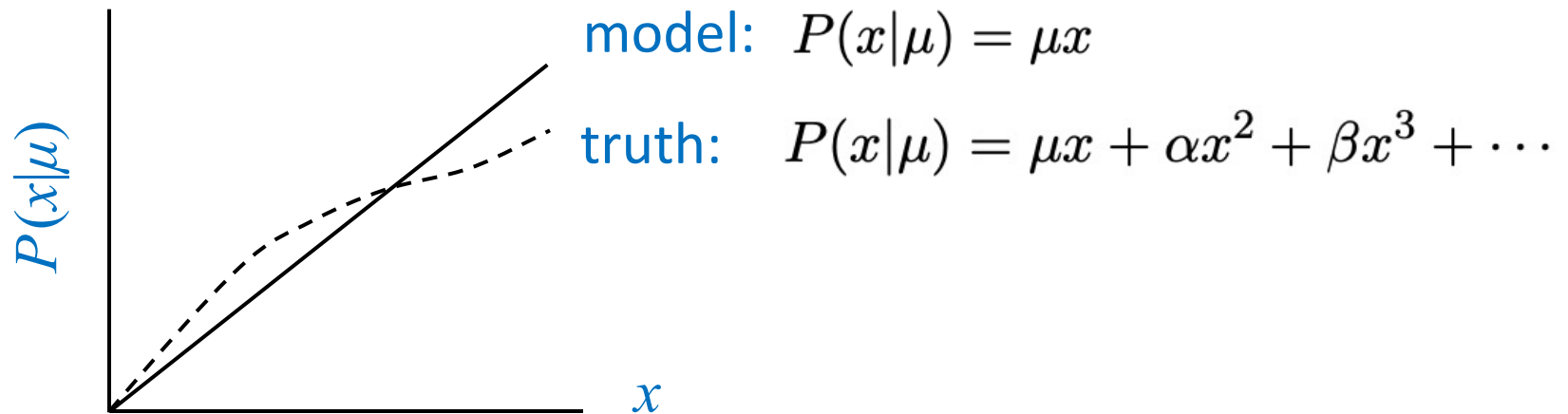
Statistical Data Analysis

Lecture 10-3

- Systematic uncertainties and nuisance parameters
- Profile likelihood

Systematic uncertainties and nuisance parameters

In general, our model of the data is not perfect:



Can improve model by including additional adjustable parameters.

$$P(x|\mu) \rightarrow P(x|\mu, \boldsymbol{\theta})$$

Nuisance parameter \leftrightarrow 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) , $i = 1, \dots, n$.

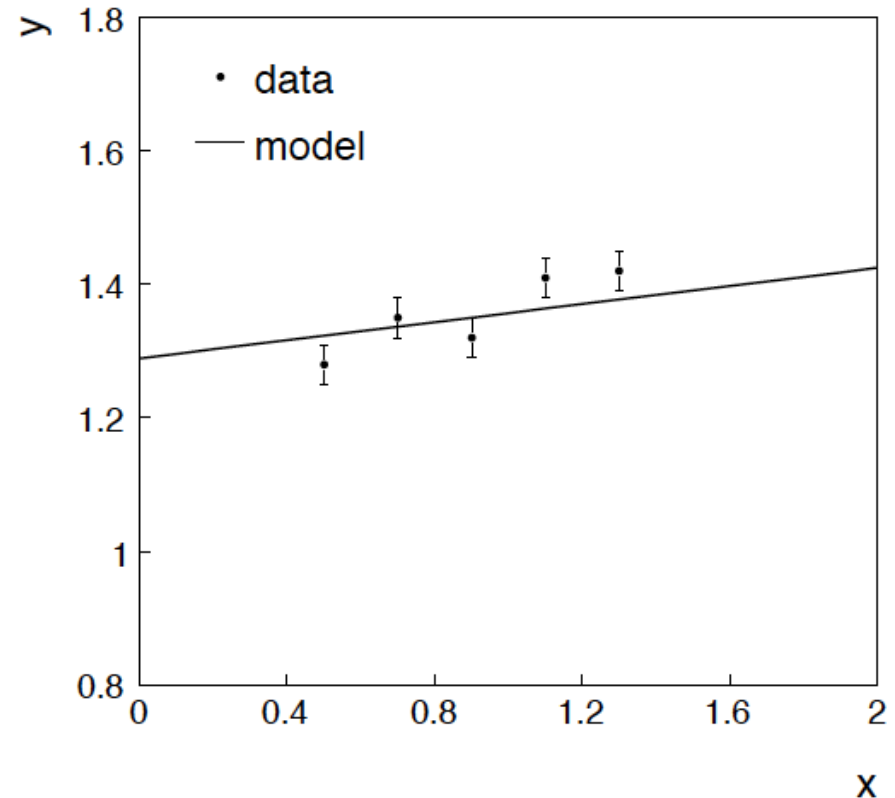
Model: y_i independent and all follow $y_i \sim \text{Gauss}(\mu(x_i), \sigma_i)$

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

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

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

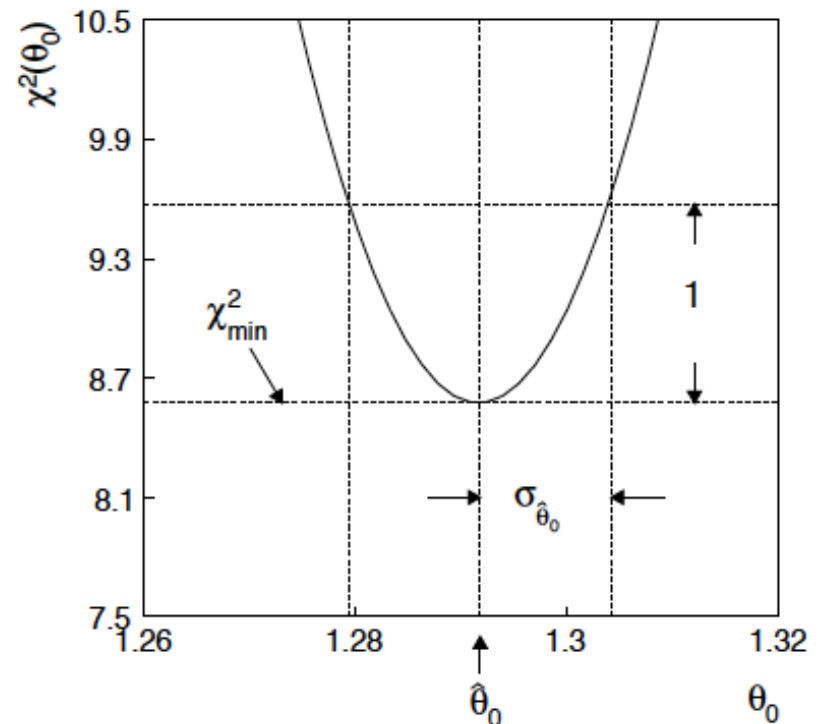
$$\chi^2(\theta_0) = -2 \ln L(\theta_0) + \text{const} = \sum_{i=1}^n \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2}.$$

For Gaussian y_i , ML same as LS

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

Come up one unit from χ_{\min}^2

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



ML (or LS) fit of θ_0 and θ_1

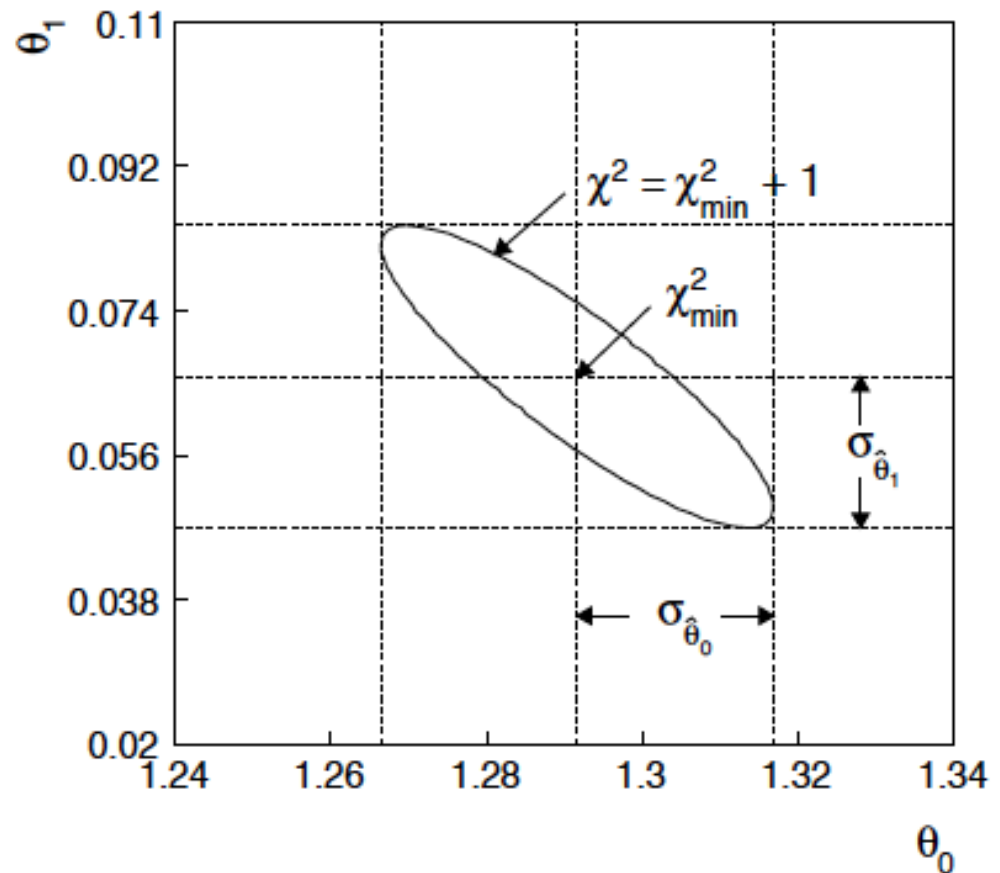
$$\chi^2(\theta_0, \theta_1) = -2 \ln L(\theta_0, \theta_1) + \text{const} = \sum_{i=1}^n \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} .$$

Standard deviations from
tangent lines to contour

$$\chi^2 = \chi_{\min}^2 + 1 .$$

Correlation between

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

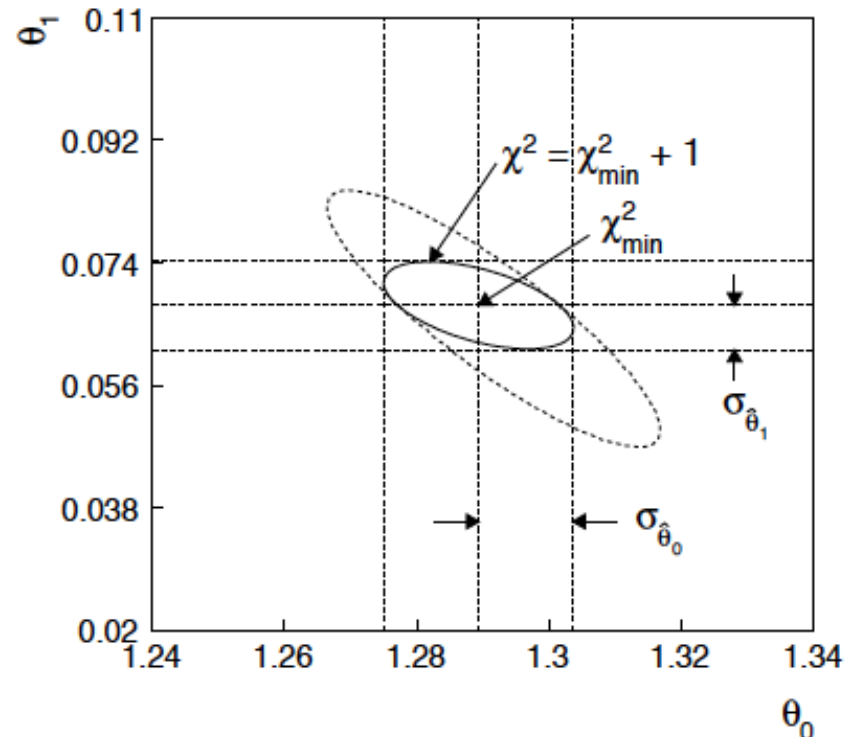


If we have a measurement $t_1 \sim \text{Gauss}(\theta_1, \sigma_{t_1})$

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

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

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



Profiling

The $\ln L = \ln L_{\max} - 1/2$ contour in the (θ_0, θ_1) plane is a confidence region at CL = 39.3%.

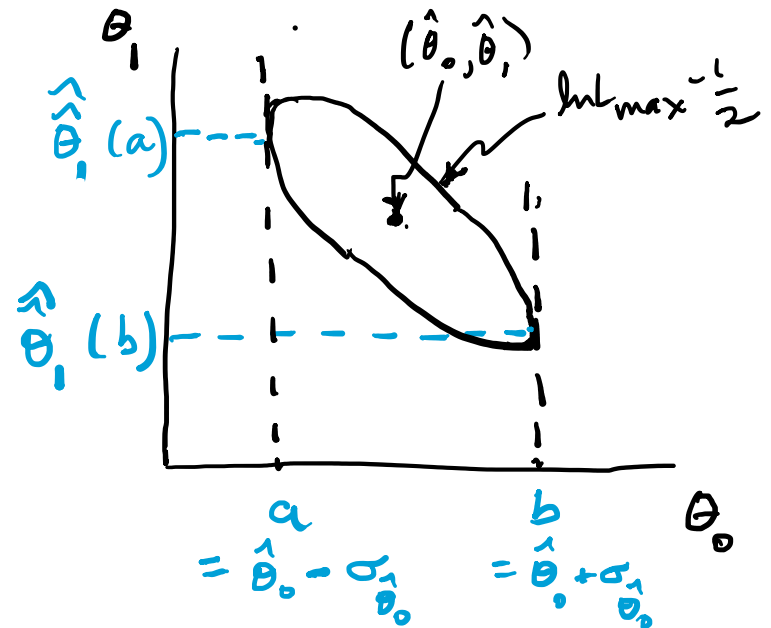
Furthermore if one wants to know only about, say, θ_0 , then the interval in θ_0 corresponding to $\ln L = \ln L_{\max} - 1/2$ is a confidence interval at CL = 68.3% (i.e., ± 1 std. dev.).

I.e., form the interval for θ_0 using

$$\ln L(\theta_0, \hat{\theta}_1(\theta_0)) = \ln L_{\max} - 1/2$$

where θ_1 is replaced by its “profiled” value

$$\hat{\theta}_1(\theta_0) = \operatorname{argmax}_{\theta_1} L(\theta_0, \theta_1)$$



Profile Likelihood

Suppose we have a likelihood $L(\boldsymbol{\mu}, \boldsymbol{\theta}) = P(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\theta})$ with N parameters of interest $\boldsymbol{\mu} = (\mu_1, \dots, \mu_N)$ and M nuisance parameters $\boldsymbol{\theta} = (\theta_1, \dots, \theta_M)$. The “profiled” (or “constrained”) values of $\boldsymbol{\theta}$ are:

$$\hat{\boldsymbol{\theta}}(\boldsymbol{\mu}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} L(\boldsymbol{\mu}, \boldsymbol{\theta})$$

and the profile likelihood is: $L_p(\boldsymbol{\mu}) = L(\boldsymbol{\mu}, \hat{\boldsymbol{\theta}})$

The profile likelihood depends only on the parameters of interest; the nuisance parameters are replaced by their profiled values.

The profile likelihood can be used to obtain confidence intervals/regions for the parameters of interest in the same way as one would for all of the parameters from the full likelihood.

Profile Likelihood Ratio – Wilks theorem

Goal is to test/reject regions of μ space (param. of interest).

Rejecting a point μ should mean $p_\mu \leq \alpha$ for all possible values of the nuisance parameters θ .

Test μ using the “profile likelihood ratio”:
$$\lambda(\mu) = \frac{L(\mu, \hat{\theta})}{L(\hat{\mu}, \hat{\theta})}$$

Let $t_\mu = -2 \ln \lambda(\mu)$. Wilks' theorem says in large-sample limit:

$$t_\mu \sim \text{chi-square}(N)$$

where the number of degrees of freedom is the number of parameters of interest (components of μ). So p -value for μ is

$$p_\mu = \int_{t_{\mu, \text{obs}}}^{\infty} f(t_\mu | \mu, \theta) dt_\mu = 1 - F_{\chi_N^2}(t_{\mu, \text{obs}})$$

Profile Likelihood Ratio – Wilks theorem (2)

If we have a large enough data sample to justify use of the asymptotic chi-square pdf, then if μ is rejected, it is rejected for any values of the nuisance parameters.

The recipe to get confidence regions/intervals for the parameters of interest at $CL = 1 - \alpha$ is thus the same as before, simply use the profile likelihood:

$$\ln L_p(\mu) = \ln L_{\max} - \frac{1}{2} F_{\chi_N^2}^{-1}(1 - \alpha)$$

where the number of degrees of freedom N for the chi-square quantile is equal to the number of parameters of interest.

If the large-sample limit is not justified, then use e.g. Monte Carlo to get distribution of t_μ .

Statistical Data Analysis

Lecture 10-4

- Bayesian parameter estimation
- Marginalization of posterior pdf
- Markov Chain Monte Carlo

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 .

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.,

$$\pi(\theta_0, \theta_1) = \pi_0(\theta_0)\pi_1(\theta_1) \quad \leftarrow \text{suppose knowledge of } \theta_0 \text{ has no influence on knowledge of } \theta_1$$

$$\pi_0(\theta_0) = \text{const.} \quad \leftarrow \text{'non-informative', in any case much broader than } L(\theta_0)$$

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

prior after t_1 ,
before \mathbf{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 \mathbf{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 $P(\mathbf{y}, t | \theta_0, \theta_1)$ 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

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

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

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

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 θ but look only at
distribution of parameters of interest.

MCMC basics: Metropolis-Hastings algorithm

Goal: given an n -dimensional pdf $p(\theta)$ up to a proportionality constant, generate a sequence of points $\theta_1, \theta_2, \theta_3, \dots$

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

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

Proposal density $q(\theta; \theta_0)$
e.g. Gaussian centred
about θ_0

3) Form test ratio

$$\alpha = \min \left[1, \frac{p(\vec{\theta})q(\vec{\theta}_0; \vec{\theta})}{p(\vec{\theta}_0)q(\vec{\theta}; \vec{\theta}_0)} \right]$$

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

The proposal density can be (almost) anything, but choose so as to minimize autocorrelation. Often take proposal density symmetric: $q(\theta; \theta_0) = q(\theta_0; \theta)$

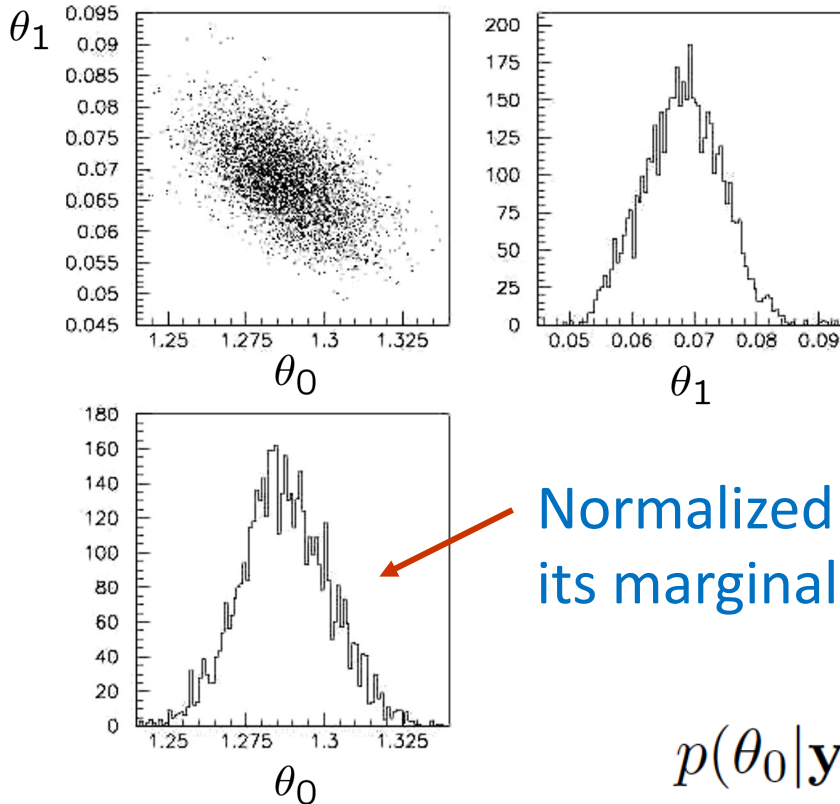
Test ratio is (*Metropolis-Hastings*): $\alpha = \min \left[1, \frac{p(\vec{\theta})}{p(\vec{\theta}_0)} \right]$

I.e. if the proposed step is to a point of higher $p(\theta)$, take it; if not, only take the step with probability $p(\theta)/p(\theta_0)$.

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:

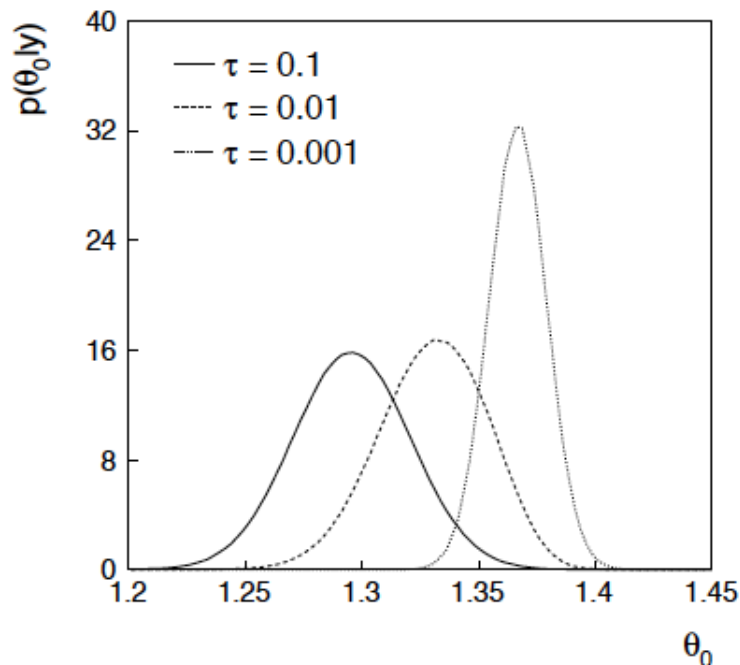
$$p(\theta_0|\mathbf{y}) = \int p(\theta_0, \theta_1|\mathbf{y}) d\theta_1$$

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.

Extra slides

Bonus exercise: Bayesian parameter estimation

The exercise is described in

<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

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

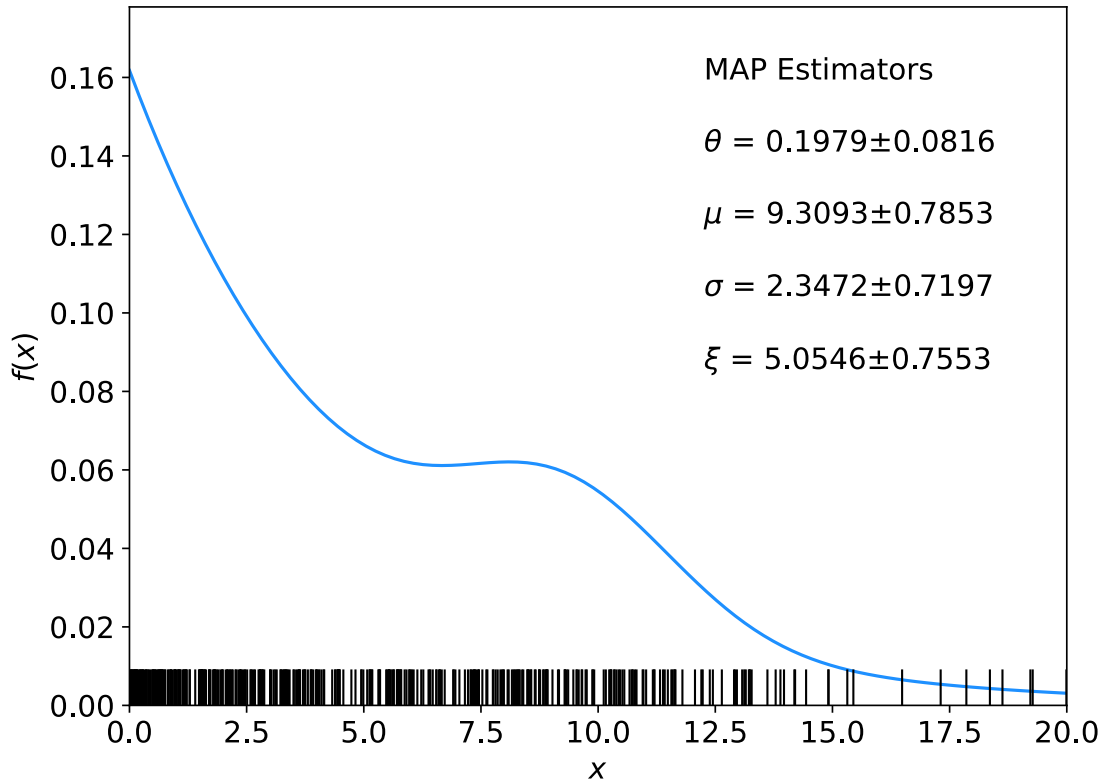
Posterior pdf for parameters $\boldsymbol{\lambda} = (\theta, \mu, \sigma, \xi)$ from Bayes theorem,

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

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 $-\ln p(\lambda|\mathbf{x})$).



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

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


A look at bayesFit.py

Find maximum of posterior with iminuit (minimize $-\ln p(\lambda|\mathbf{x})$), 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 $V_{ij} = \text{cov}[\lambda_i, \lambda_j]$ 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 λ_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, \dots, x_n) and the auxiliary measurement u and modify the fitting program accordingly. Investigate how the results are affected by including u .

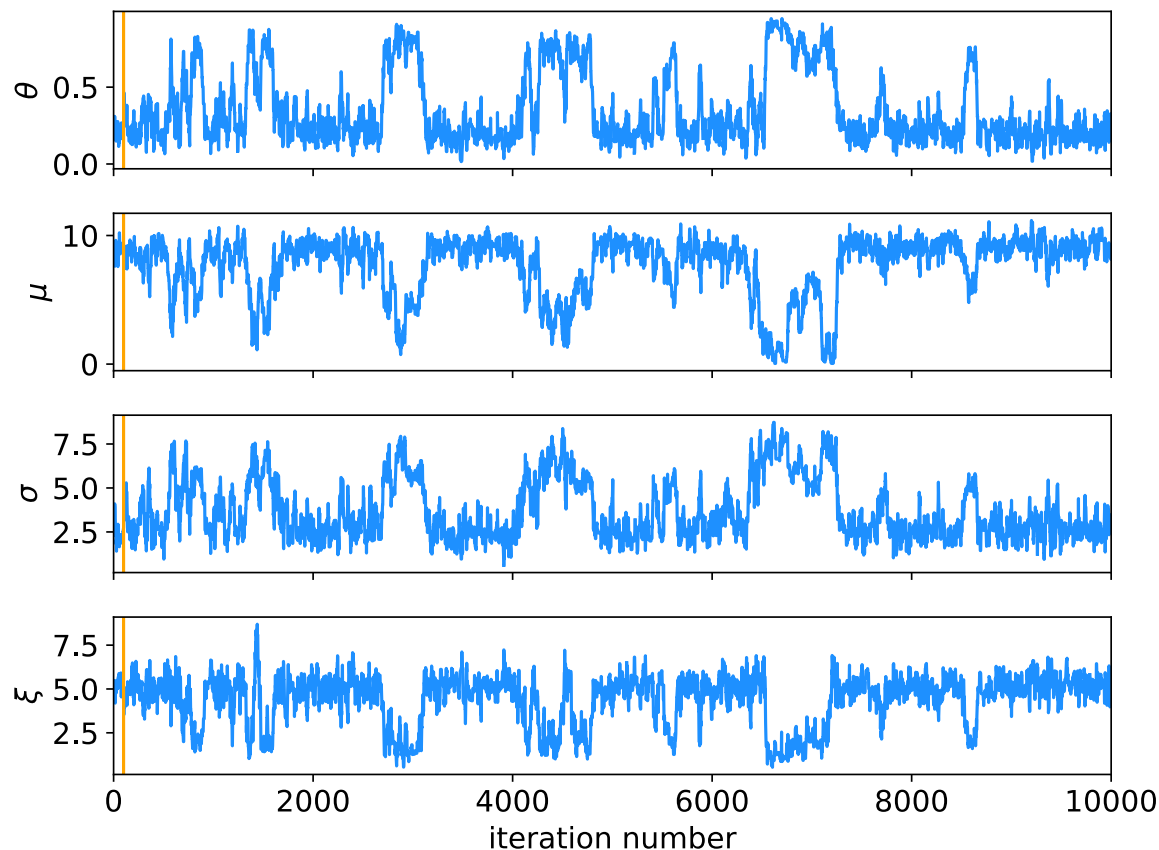
1e) Investigate point and interval estimates for θ

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 intervals 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.



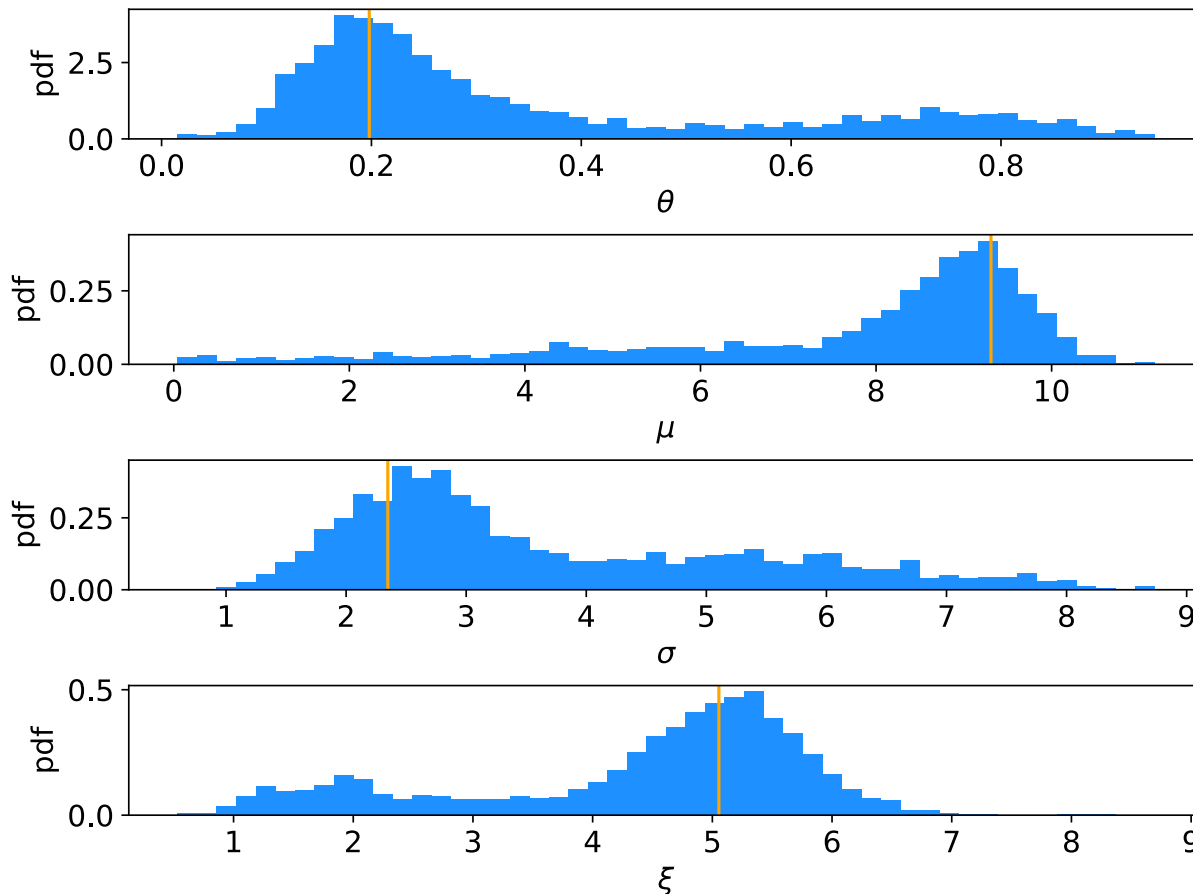
Gaussian proposal pdf,
covariance $U = sV$,
 $s = (2.38)^2/N_{\text{par}} = 1.41$,
gives acceptance
probability ~ 0.24 .

See, e.g.,
<http://probability.ca/jeff/ftplib/galinart.pdf>

Here 10000 iterations
(should use more).

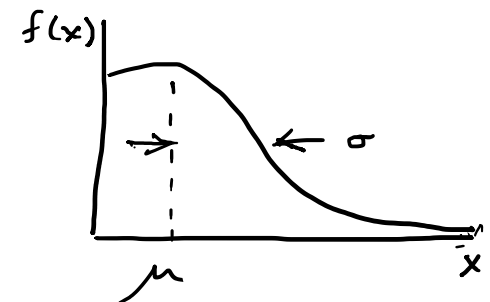
Marginal distributions

MAP estimates shown with vertical bars



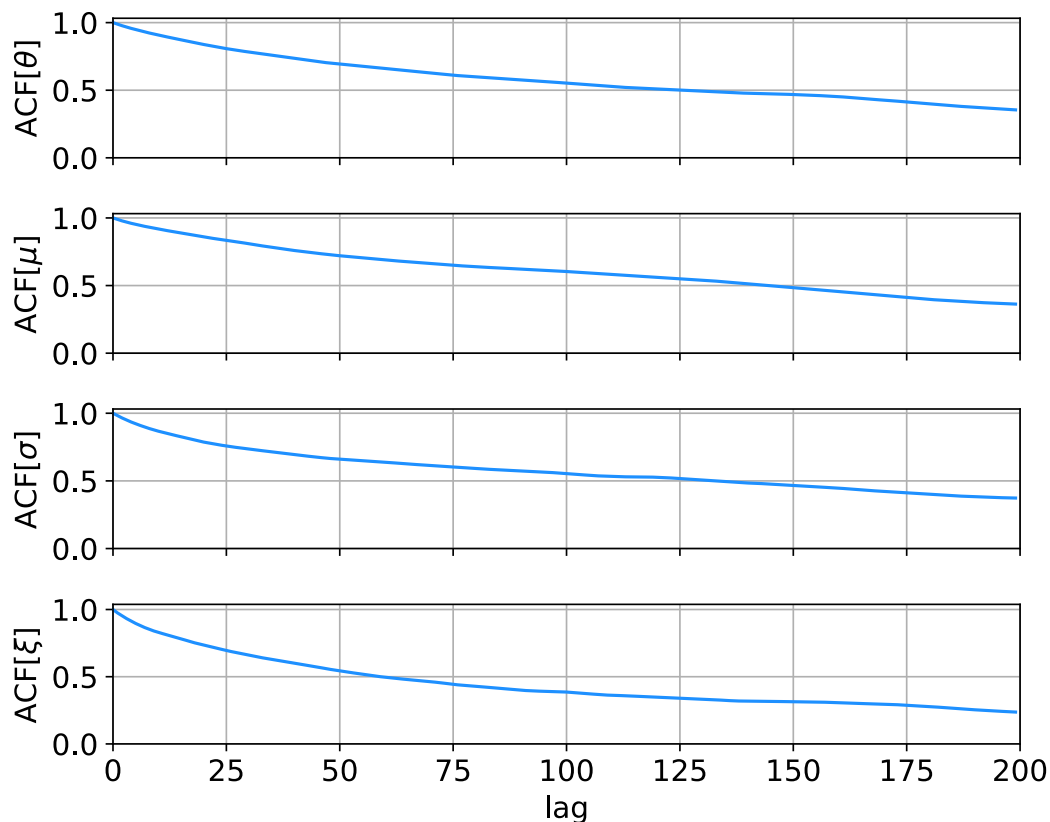
Note long tails.

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:



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

Effective sample size

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

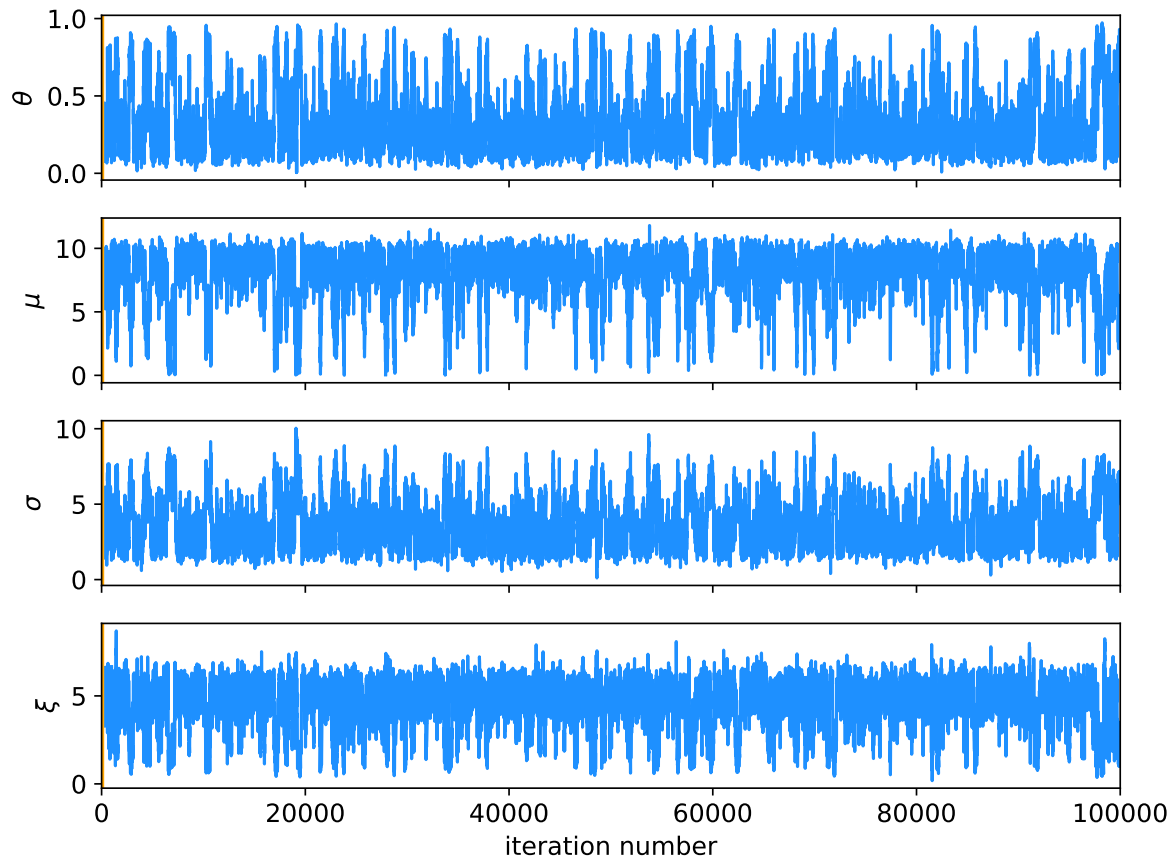
In stat. error estimates

$$\frac{1}{\sqrt{N}} \rightarrow \frac{1}{\sqrt{N_{\text{eff}}}}$$

See, e.g., <https://mc-stan.org/docs/reference-manual/effective-sample-size.html>

MCMC trace plots

Increase number of iterations to 100000.



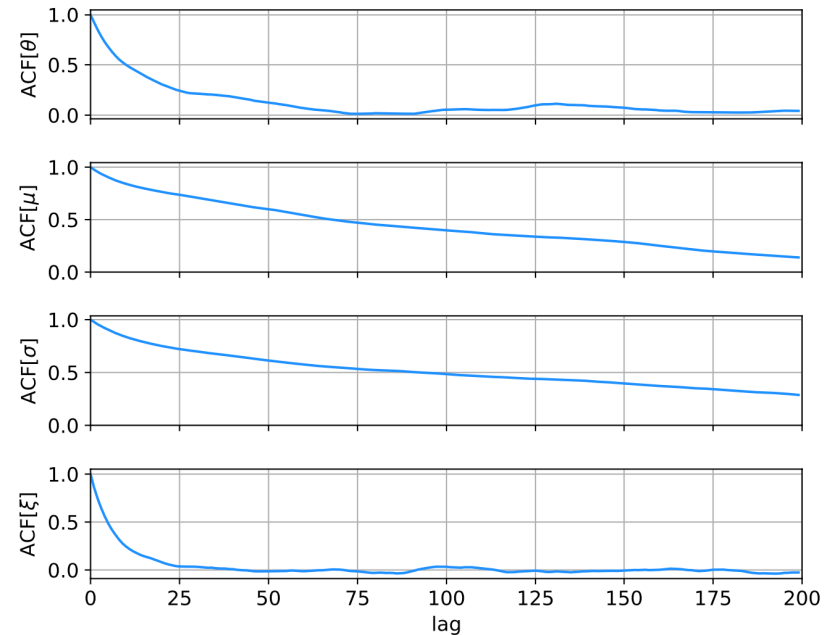
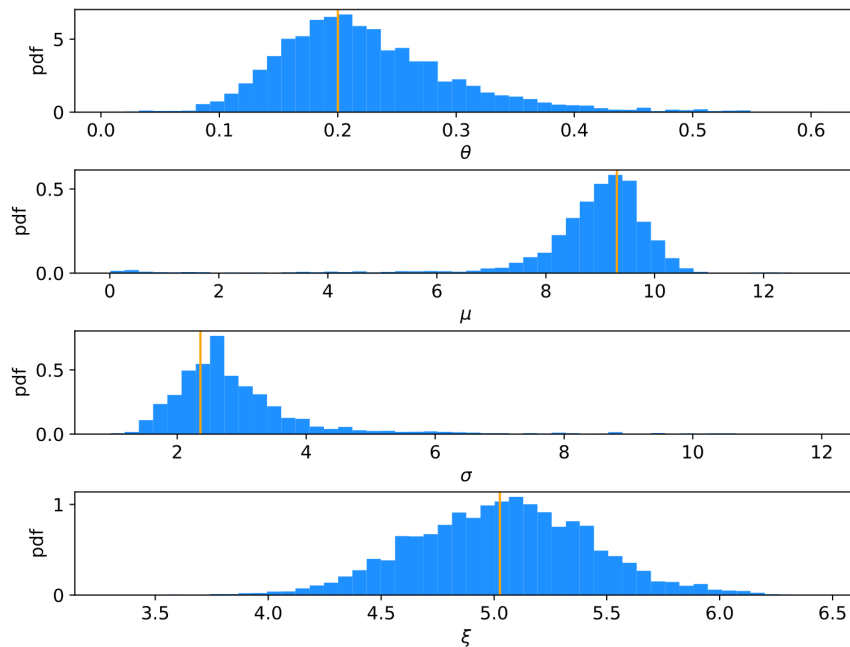
Regions of parameter space now sampled multiple times.

Auxiliary measurement for ξ

Suppose we have an auxiliary measurement $u \sim \text{Gauss}(\xi, \sigma_u)$ with $\sigma_u = 0.5$ and we observe $u = 5$.

Build into likelihood:
$$p(\mathbf{x}, u | \boldsymbol{\lambda}) = \frac{1}{\sqrt{2\pi}\sigma_u} e^{-(u-\xi)^2/2\sigma_u^2} \prod_{i=1}^n f(x_i | \boldsymbol{\lambda})$$

Marginals closer to Gaussian, ACF falls more quickly to \sim zero.



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 $CL = 1 - \alpha$.

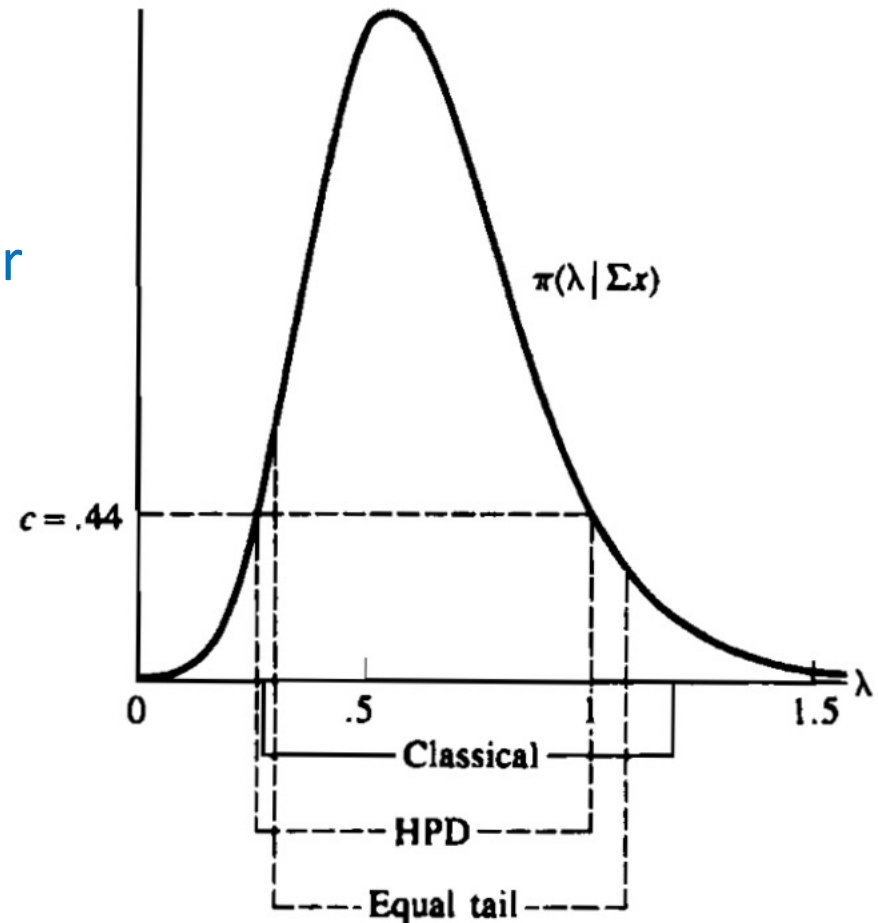
Point estimate +/- standard deviation, std. dev. from MCMC sample or by approximating core of posterior as Gaussian (from minuit); coincides with above in large-sample limit.

Types of intervals

HPD = Highest Posterior Density

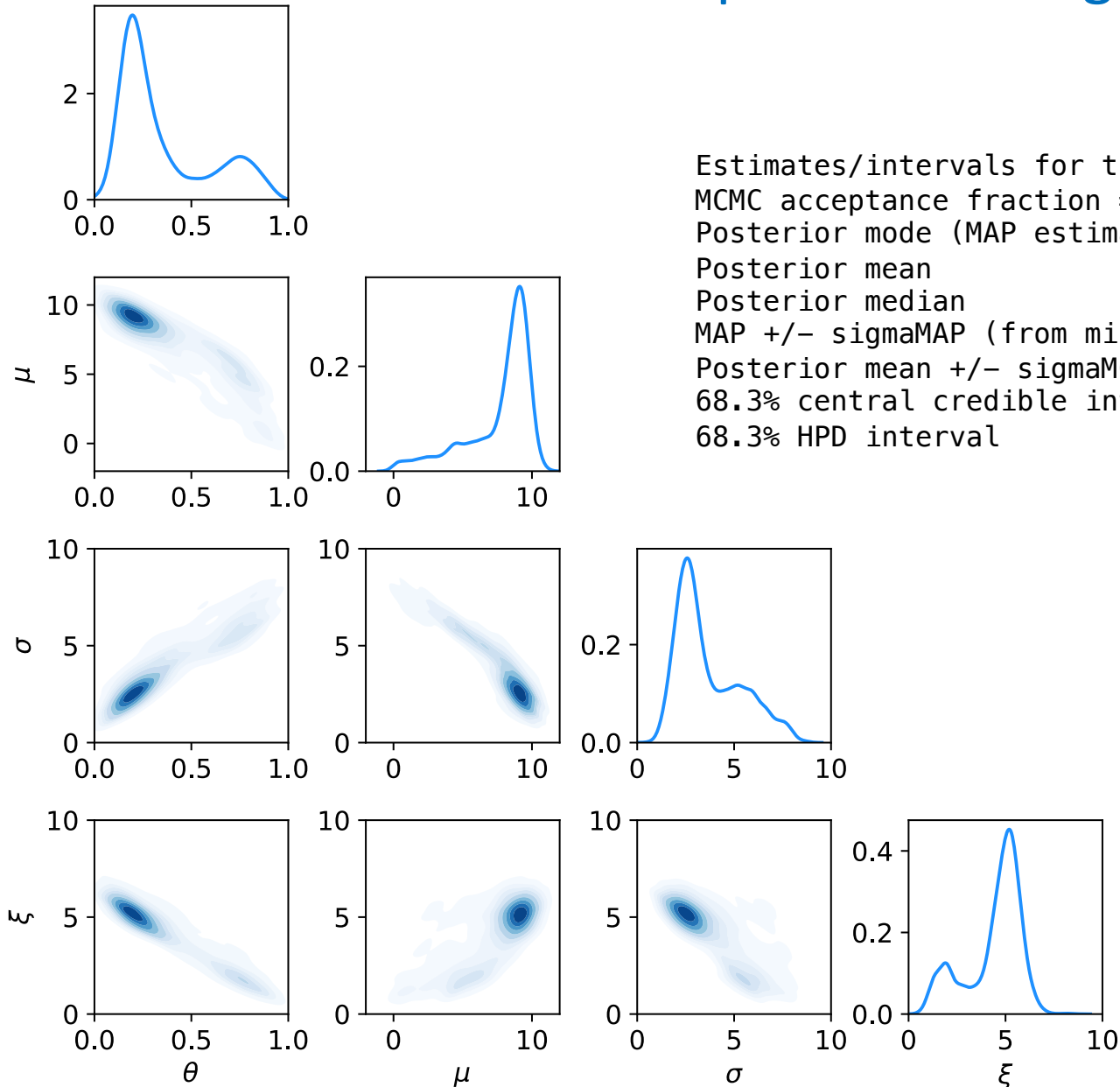
Equal tail (central) from posterior

Classical (frequentist)



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

Correlation plots and marginal distributions



Estimates/intervals for theta:

MCMC acceptance fraction = 0.3737

Posterior mode (MAP estimate) = 0.197936

Posterior mean = 0.351234

Posterior median = 0.250764

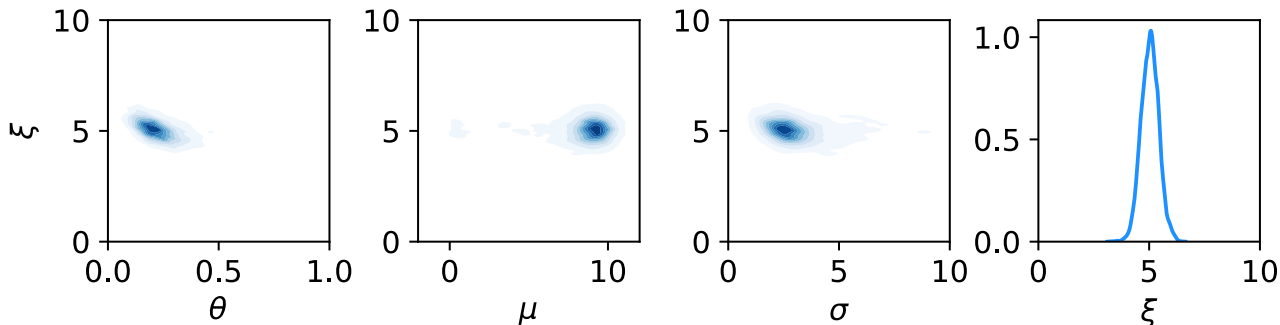
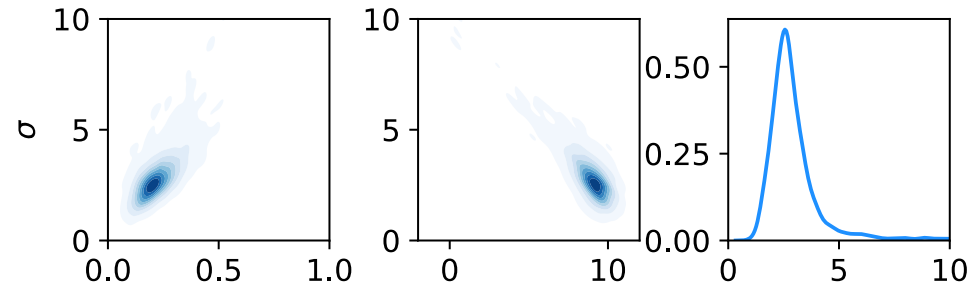
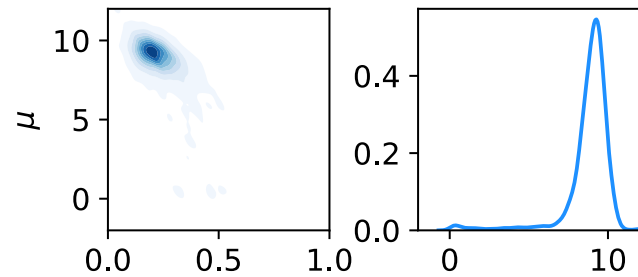
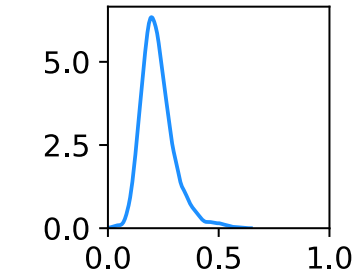
MAP +/- sigmaMAP (from minuit) = [0.140056, 0.255816]

Posterior mean +/- sigmaMCMC = [0.118503, 0.583964]

68.3% central credible interval = [0.158557, 0.685847]

68.3% HPD interval = [0.093997, 0.392642]

Correlation plots and marginal distributions using auxiliary measurement for ξ



Estimates/intervals for theta:

MCMC acceptance fraction = 0.3516

Posterior mode (MAP estimate) = 0.200077

Posterior mean = 0.224898

Posterior median = 0.212835

MAP +/- sigmaMAP (from minuit) = [0.149282, 0.250873]

Posterior mean +/- sigmaMCMC = [0.148481, 0.301314]

68.3% central credible interval = [0.156736, 0.291246]

68.3% HPD interval = [0.137607, 0.267077]