Statistical Data Analysis
Discussion slides – week 10

- Problem sheet 7
- Bonus problem: Bayesian parameter estimation and MCMC

Problem sheet 7

Exercise 1: Suppose the random variable x follows a special case of the gamma pdf,

$$f(x;\theta) = \frac{x^2}{2\theta^3} e^{-x/\theta} \ ,$$

with $x \geq 0$ and $\theta > 0$. The expectation value and variance of x are $E[x] = 3\theta$, $V[x] = 3\theta^2$. Consider a sample of n independent values $x_1, \ldots, x_n$ from this pdf, with which we want to estimate θ. For parts (a)–(c), suppose that n is a fixed constant.

1(a) [3 marks] Write down the likelihood function $L(\theta)$ and show that the maximum-likelihood estimator for θ is

$$\hat{\theta} = \frac{1}{3n} \sum_{i=1}^{n} x_i \ .$$

Problem sheet 7

1a) $$L(\theta) = \prod_{i=1}^{n} \frac{x_i^2}{2\theta^3} e^{-x_i/\theta}$$

$$\ln L(\theta) = \sum_{i=1}^{n} \left[\ln x_i^2 - \ln 2 - \ln \theta^3 - \frac{x_i}{\theta} \right]$$

$$= -3n \ln \theta - \frac{1}{\theta} \sum_{i=1}^{n} x_i + C$$

$$\frac{\partial \ln L}{\partial \theta} = -\frac{3n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^{n} x_i \overset{\text{set}}{=} 0$$

$$\Rightarrow \quad \hat{\theta} = \frac{1}{3n} \sum_{i=1}^{n} x_i$$

Problem sheet 7

1(b) [**4 marks**] Show that $\hat{\theta}$ is unbiased, find its variance, and show that the variance is equal to the minimum variance bound.

b) $E[\hat{\theta}] = E\left[\frac{1}{3n}\sum_{i=1}^{n} x_i\right]$

$= \frac{1}{3n}\sum_{i=1}^{n} \underbrace{E[x_i]}_{= 3\theta} = \theta$

$\Rightarrow b = E[\hat{\theta}] - \theta = 0$

$V[\hat{\theta}] = V\left[\frac{1}{3n}\sum_{i=1}^{n} x_i\right] = \frac{1}{9n^2}\sum_{i=1}^{n} \underbrace{V[x_i]}_{= 3\theta^2} = \frac{\theta^2}{3n}$

Problem sheet 7

1(b) cont.

$$\text{MVB} = -\frac{\left(1+\frac{\partial b}{\partial \theta}\right)^2}{E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]}$$

$$\frac{\partial^2 \ln L}{\partial \theta^2} = \frac{3n}{\theta^2} - \frac{2}{\theta^3}\sum_{i=1}^{n} x_i$$

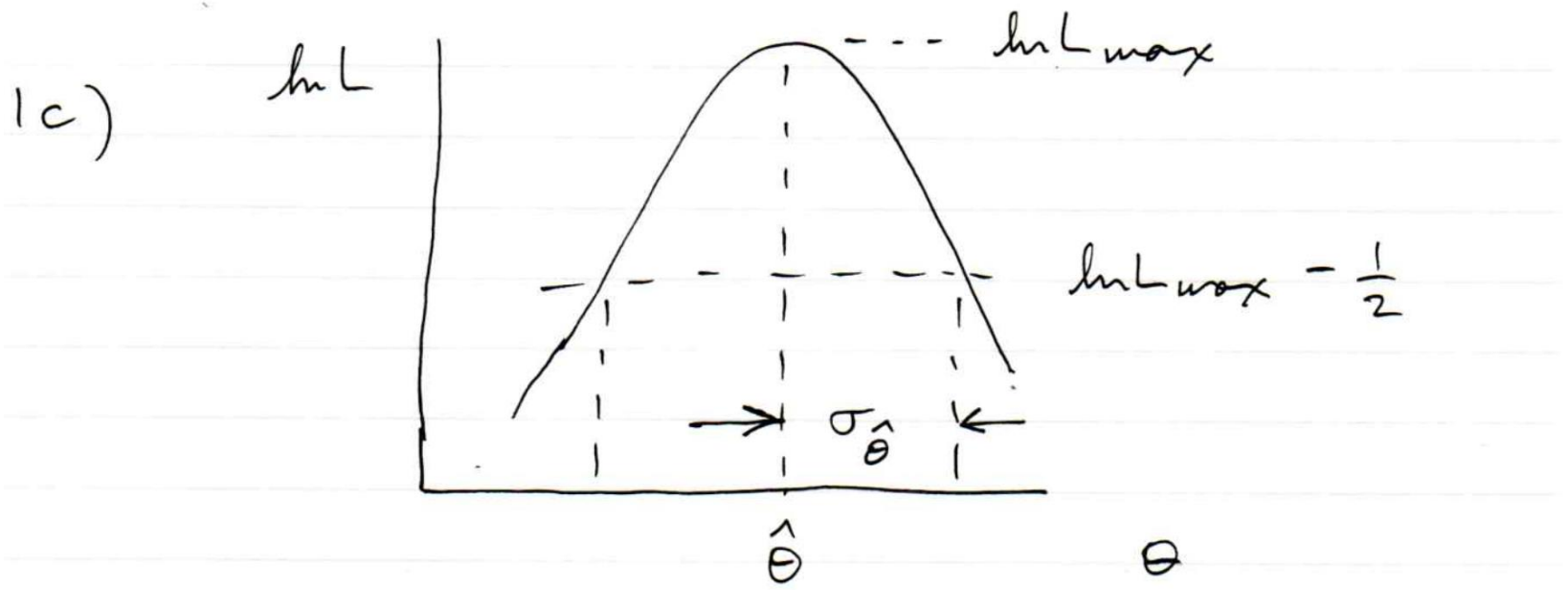
$$E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right] = \frac{3n}{\theta^2} - \frac{2}{\theta^3}\sum_{i=1}^{n} \underbrace{E[x_i]}_{\text{"}3\theta} = -\frac{3n}{\theta^2}$$

$$\Rightarrow \quad \text{MVB} = -\frac{(1+0)^2}{\left(-\frac{3n}{\theta^2}\right)} = \frac{\theta^2}{3n}$$

= same as $V[\hat{\theta}]$ $\Rightarrow$ $\hat{\theta}$ is efficient.

Problem sheet 7

1(c) [2 marks] Make a sketch of the log-likelihood function indicating the estimator $\hat{\theta}$ and indicate on the sketch how to find the standard deviation of $\hat{\theta}$.



For (d)-(f), treat $n \sim \text{Poisson}(\nu)$ i.e. $P(n|\nu) = \frac{\nu^n e^{-\nu}}{n!}$

$\nu = \alpha \theta^3$ (α known)

Problem sheet 7

For the rest of this question suppose that the sample size n is not fixed but rather follows a Poisson distribution with mean $\alpha\theta^3$, where α is a given constant. (Recall that the Poisson distribution for n with mean ν is $P(n;\nu) = \nu^n e^{-\nu}/n!$.)

1(d) [**4 marks**] Write down the full (i.e., extended) likelihood function for θ based on the Poisson distributed n and the n values $x_1, \ldots, x_n$. Show that the maximum-likelihood estimator for θ is

$$\hat{\theta} = \left(\frac{1}{3\alpha} \sum_{i=1}^{n} x_i \right)^{1/4} .$$

Problem sheet 7

1 (d)

$$L(\theta) = \frac{(\alpha\theta^3)^n}{n!} e^{-\alpha\theta^3} \prod_{i=1}^{n} \frac{x_i}{2\theta^3} e^{-x_i/\theta}$$

$$\ln L(\theta) = 3n \ln\theta - \alpha\theta^3 - 3n\ln\theta - \sum_{i=1}^{n} \frac{x_i}{\theta} + c$$

$$= -\alpha\theta^3 - \sum_{i=1}^{n} \frac{x_i}{\theta} + c$$

$$\frac{\partial \ln L}{\partial \theta} = -3\alpha\theta^2 + \frac{1}{\theta^2} \sum_{i=1}^{n} x_i \overset{\text{set}}{=} 0$$

$$\Rightarrow \hat{\theta} = \left(\frac{1}{3n} \sum_{i=1}^{n} x_i \right)^{1/4}$$

Problem sheet 7

1(e) [3 marks] Show that the expectation value of a function a of n and $\mathbf{x} = (x_1, \ldots, x_n)$ can be written

$$E[a(n, \mathbf{x})] = E_n \left[E_{\mathbf{x}}[a(n, \mathbf{x})|n] \right] \ ,$$

where E_n and $E_{\mathbf{x}}$ indicate the expectation values with respect to n and $\mathbf{x}$, respectively.

1e) $E[a(n,\vec{x})] = \sum_{n=0}^{\infty} \int a(n,\vec{x})\, P(n,\vec{x})\, d\vec{x}$

Joint probability for n & $\vec{x}$ is

$P(n,\vec{x}) = f(\vec{x}|n)\, P(n)$

$\Rightarrow E[a(n,\vec{x})] = \sum_{n=0}^{\infty} P(n) \int a(n,\vec{x})\, f(\vec{x}|n)\, d\vec{x}$

$= \sum_{n=0}^{\infty} P(n)\, E_{\vec{x}}[a(n,\vec{x})|n]$

$= E_n\left[E_{\vec{x}}[a(n,\vec{x})|n] \right]$

Problem sheet 7

1(f) [**4 marks**] Using the result from (e) and the second derivative of the log-likelihood function, show that the variance of $\hat{\theta}$ can be approximated as

$$V[\hat{\theta}] = \frac{1}{12\alpha\theta} \,,$$

stating any assumptions needed. Using the fact that the expectation value of n is $\alpha\theta^3$, compare the variance found here with that found in (b) for fixed n, and comment on why they are different.

Problem sheet 7

f) $\dfrac{\partial^2 \ln L}{\partial \theta^2} = -6\alpha\theta - \dfrac{2}{\theta^3}\sum_{i=1}^{n} x_i$ } function of both n and $\vec{x}$

$$E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right] = -6\alpha\theta - \frac{2}{\theta^3} E\left[\sum_{i=1}^{n} x_i\right]$$

use d) $E\left[\sum_{i=1}^{n} x_i\right] = E_n\left[E_x\left[\sum_{i=1}^{n} x_i \,|\, n\right]\right] = E_n\left[\sum_{i=1}^{n} \underbrace{E[x_i]}_{= 3\theta}\right]$

$$= E_n[3n\theta] = 3\nu\theta = 3\alpha\theta^4 \quad (\nu = \alpha\theta^3)$$

$$\Rightarrow E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right] = -6\alpha - \frac{2}{\theta^3} 3\alpha\theta^4 = -12\alpha\theta$$

$$\Rightarrow V[\hat{\theta}] = \frac{1}{12\alpha\theta} = \frac{\theta^2}{12\nu}$$

use $\alpha = \nu/\theta^3$

compare to $\theta^2/3n$ for fixed n case.

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

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

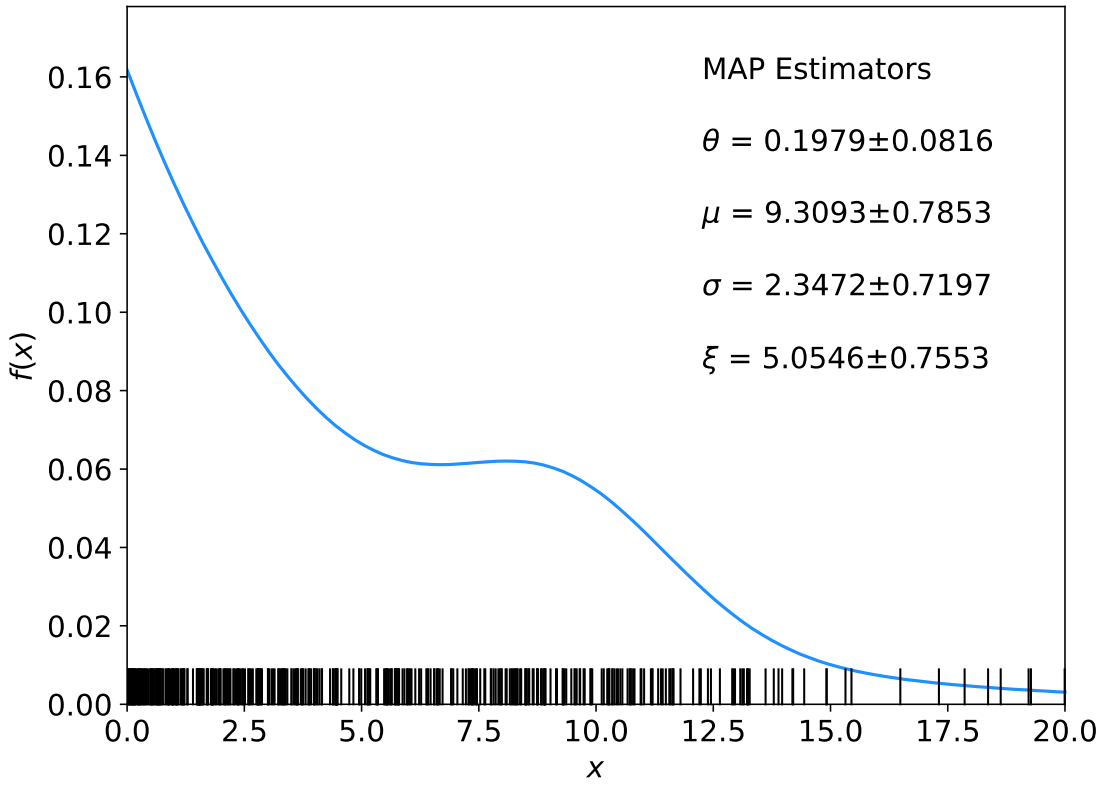
Posterior pdf for parameters $\boldsymbol{\lambda} = (\theta, \mu, \sigma, \xi)$ from Bayes theorem,

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

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



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.

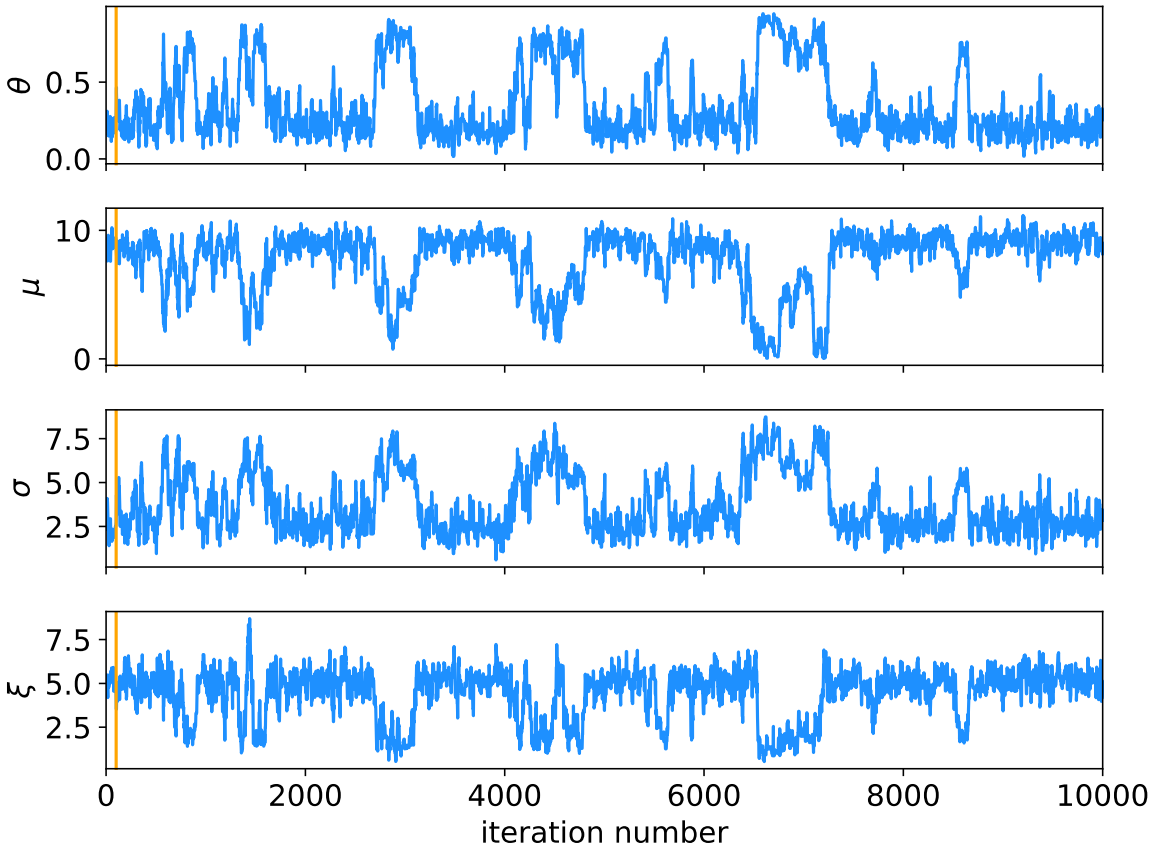
1e) Investigate point and interval estimates for θ

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.



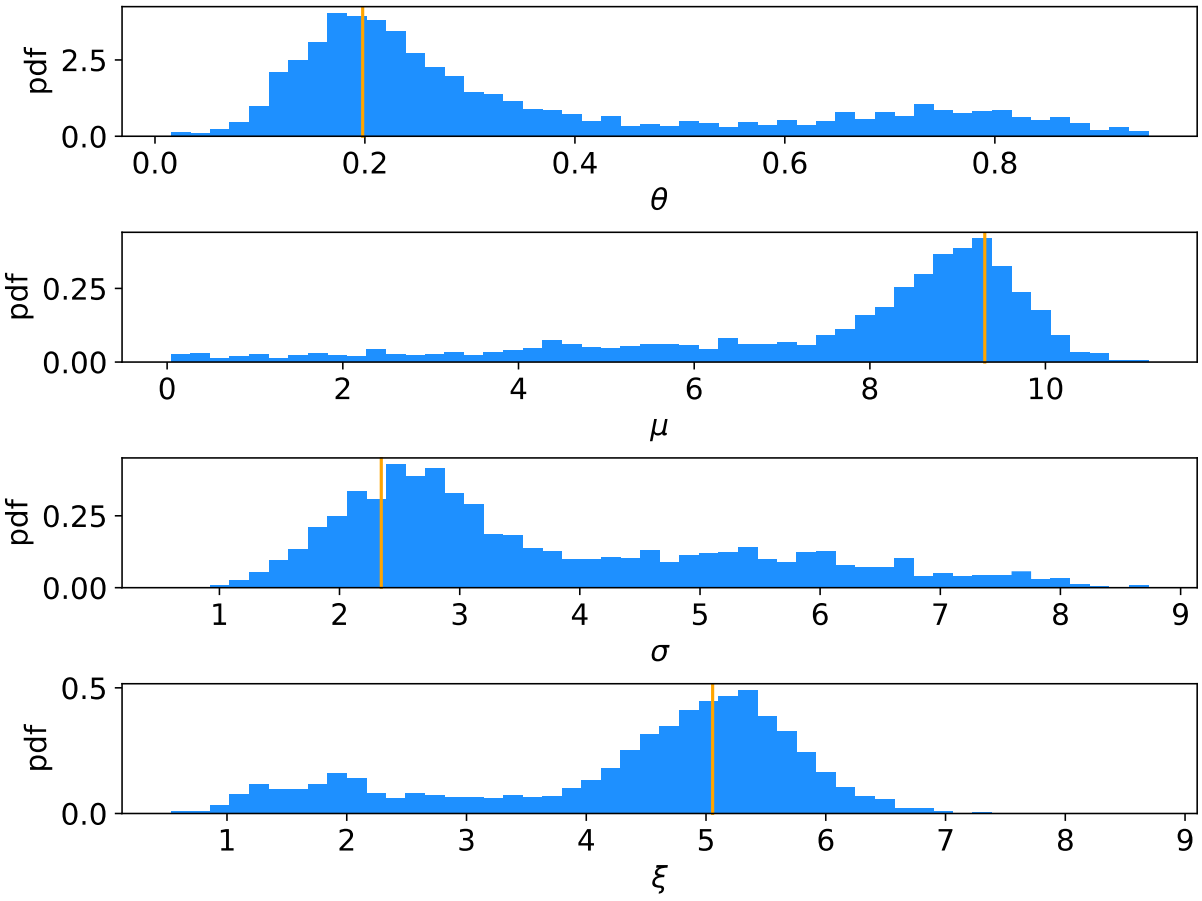
Gaussian proposal pdf, covariance $U = sV$, $s = (2.38)^2/N_{\text{par}} = 1.41$, gives acceptance probability ~ 0.24.

See, e.g., http://probability.ca/jeff/ftpdir/galinart.pdf

Here 10000 iterations (should use more).

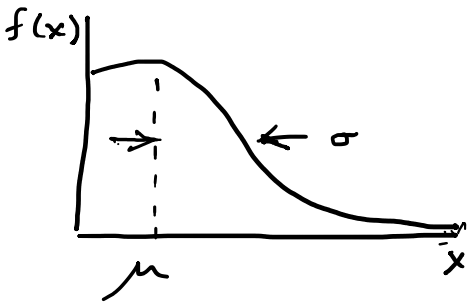
Marginal distributions

MAP estimates shown with vertical bars



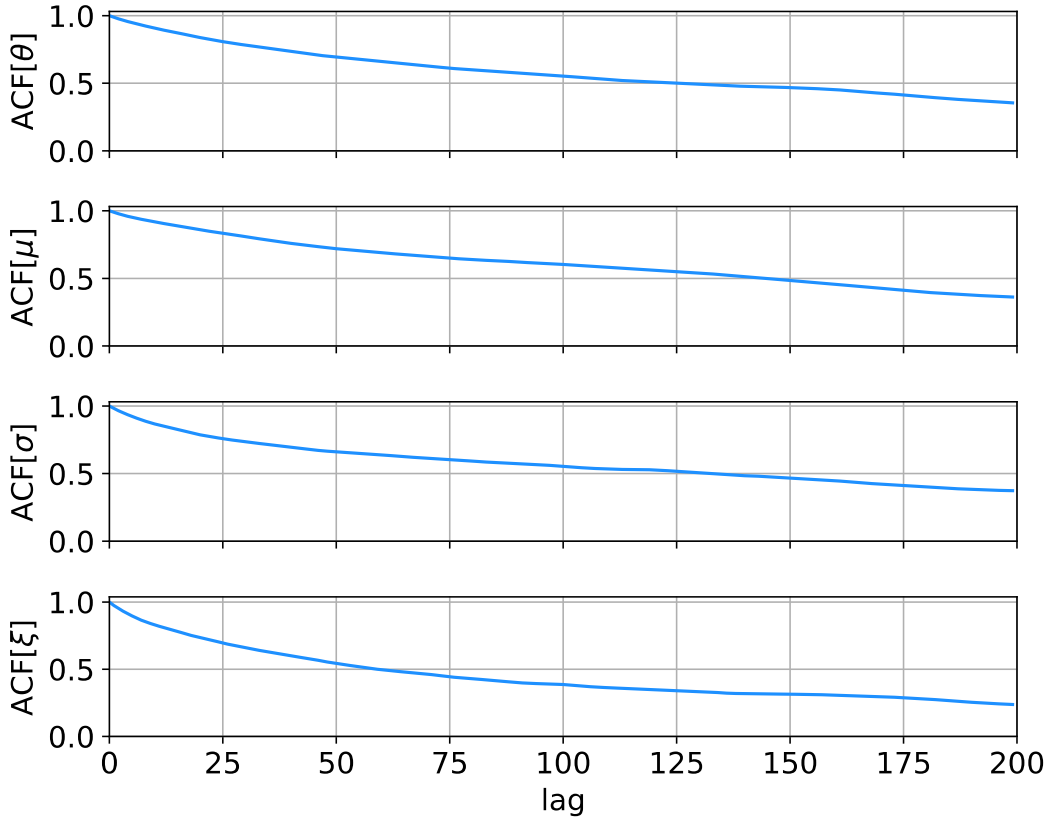
Note long tails.

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:



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

Effective sample size

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

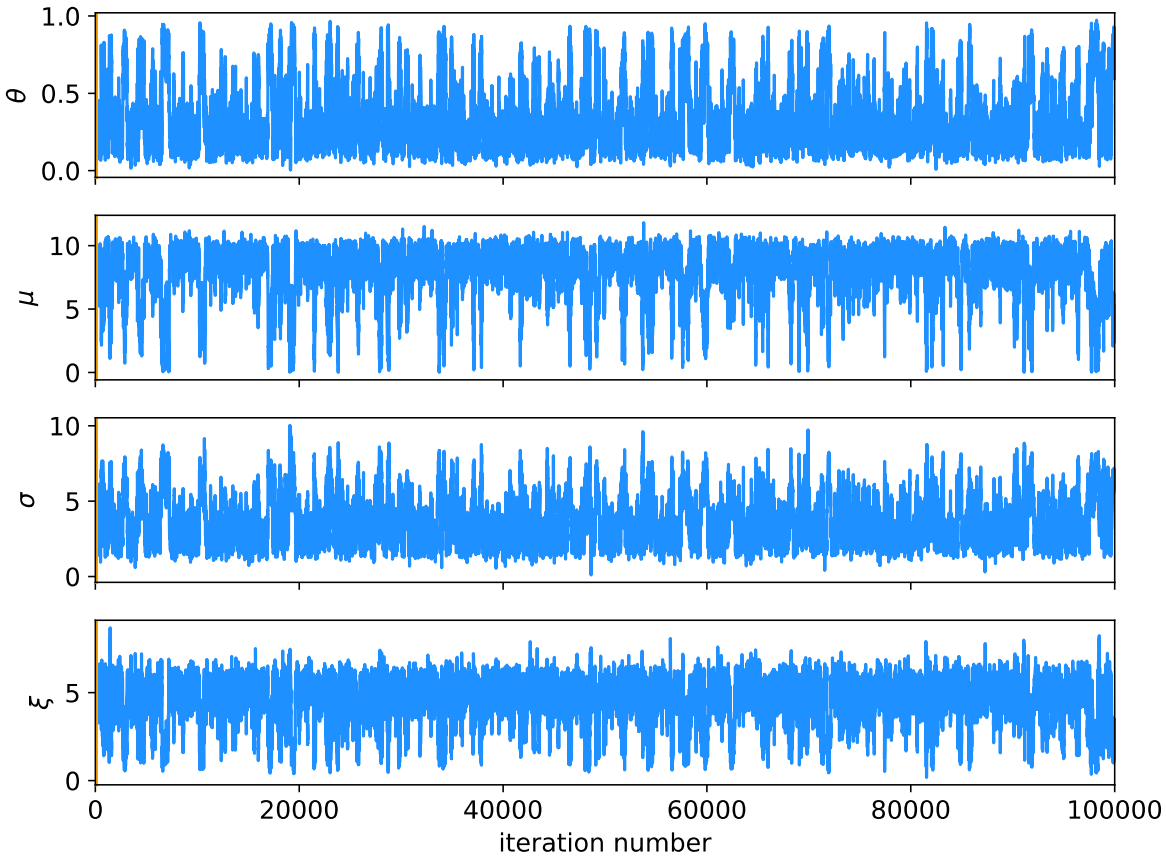
In stat. error estimates

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

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

MCMC trace plots

Increase number of iterations to 100000.



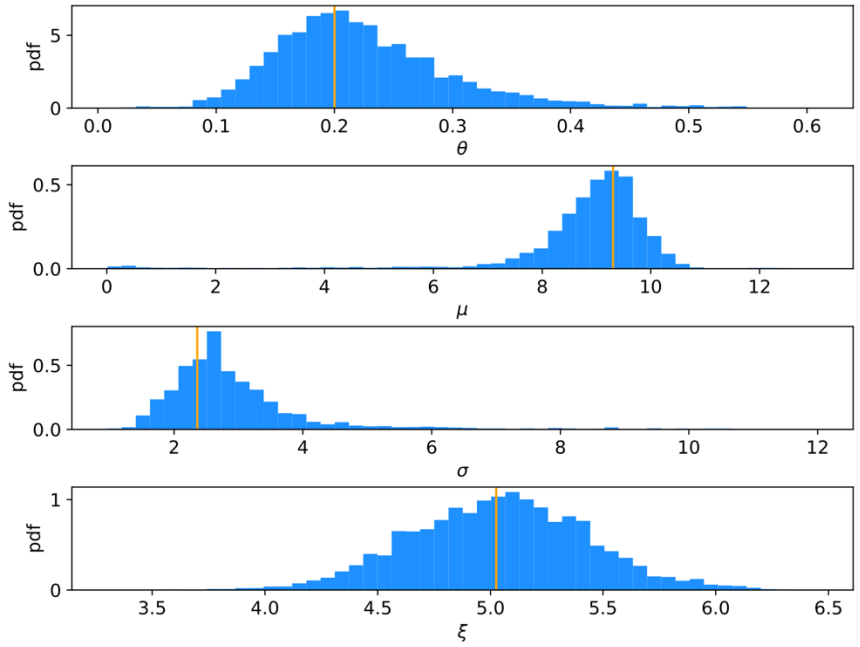
Regions of parameter space now sampled multiple times.

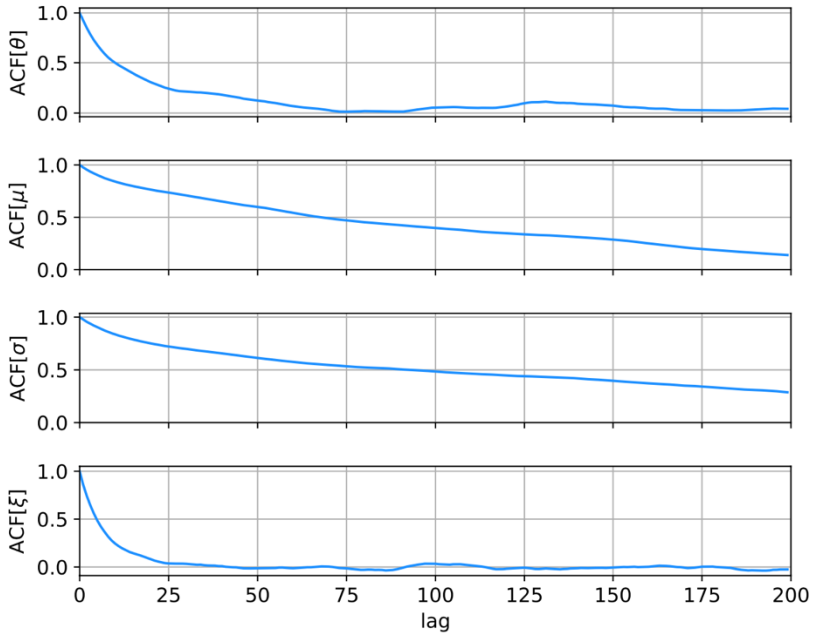
Auxiliary measurement for ξ

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 ~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 = $1 - \alpha$.

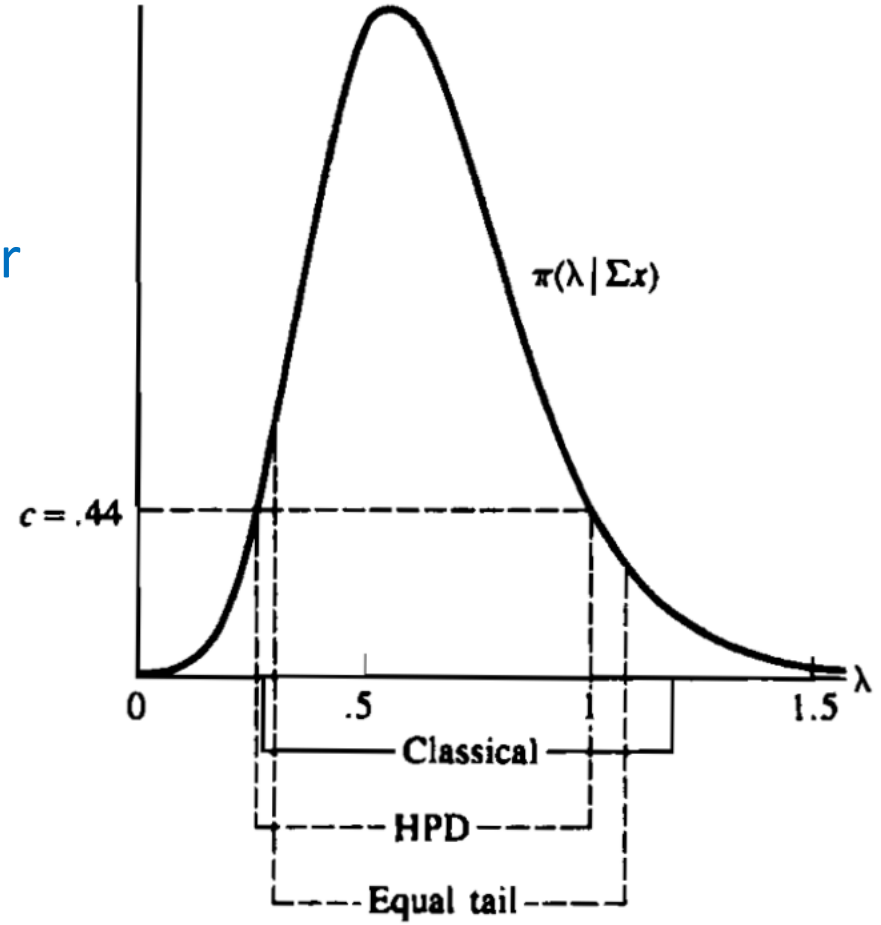
Point estimate +/- standard deviation, std. dev. from MCMC sample or by approximating core of posterior as Gaussian (from minuit); coincides with above in large-sample limit.

Types of intervals

HPD = Highest Posterior Density

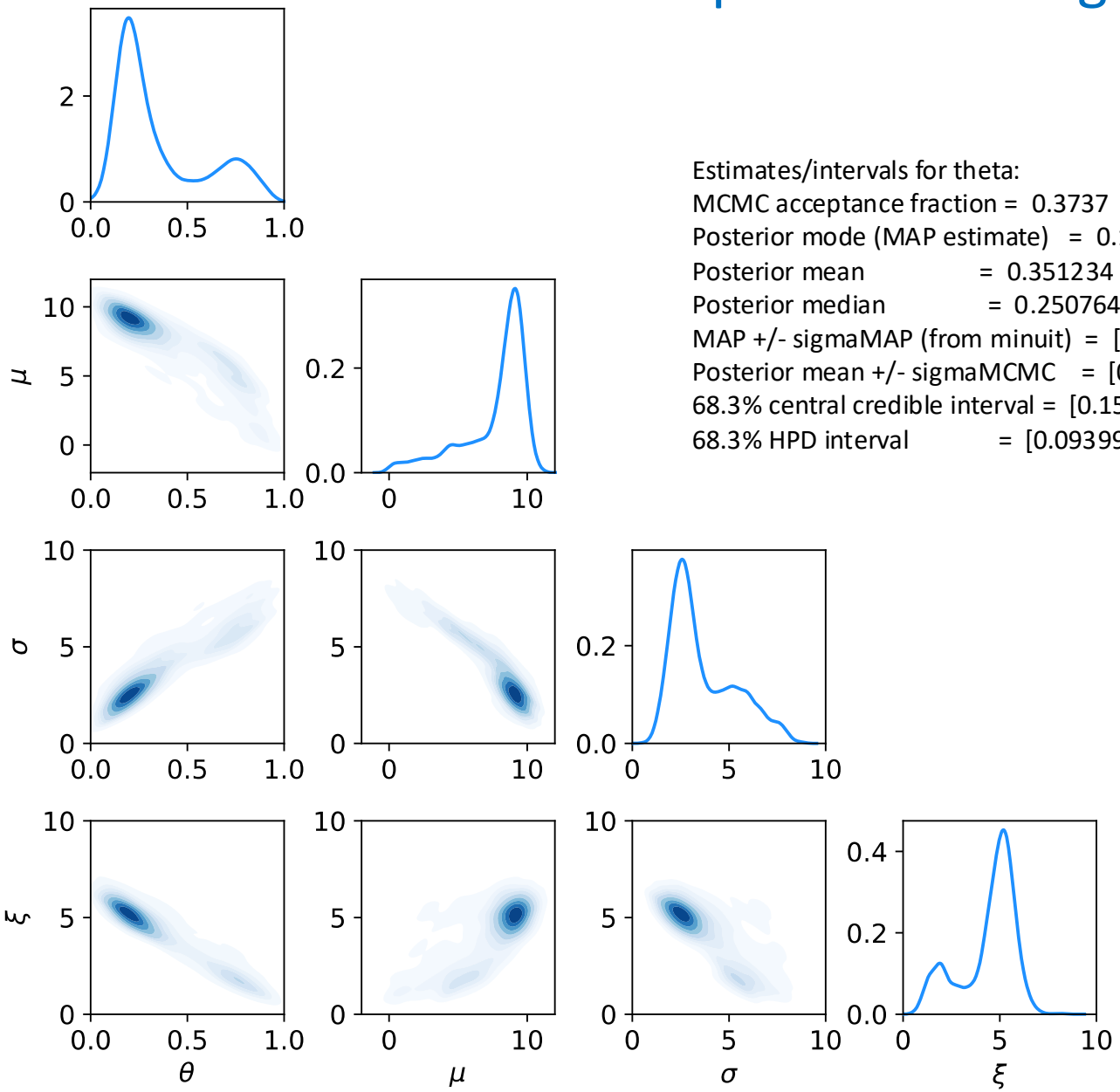
Equal tail (central) from posterior

Classical (frequentist)



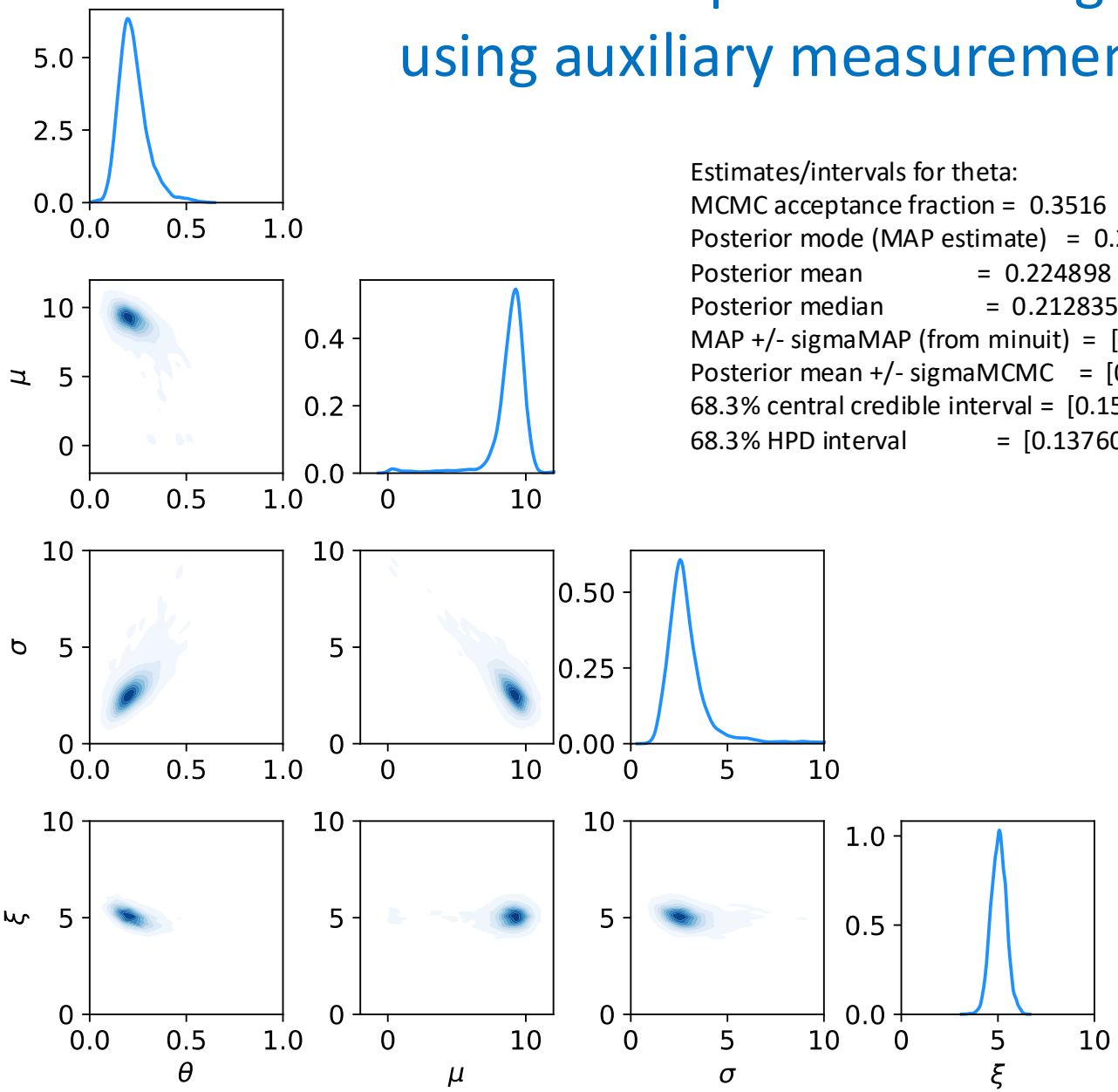
G. Casella and R. Berger, Statistical Inference, 2002

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]