Week 12: Bayesian Computations

MATH-517 Statistical Computation and Visualization

Tomas Masak

December 9th 2022

Bayes' Rule

Let X be a random variable and θ a parameter, considered also a random variable:

$$f_{X,\theta}(x,\theta) = \underbrace{f_{X|\theta}(x \mid \theta)}_{\text{likelihood}} \underbrace{f_\theta(\theta)}_{\text{prior}} = \underbrace{f_{\theta|X}(\theta \mid x)}_{\text{posterior}} f_X(x).$$

- likelihood = frequentist model
- likelihood & prior = Bayesian model

Denoting by x_0 the observed value of X:

$$f_{\theta|X=x_0}(\theta \mid x_0) = \frac{f_{X|\theta}(x_0 \mid \theta) f_\theta(\theta)}{f_X(x_0)} = \frac{f_{X|\theta}(x_0 \mid \theta) f_\theta(\theta)}{\int f_{X|\theta}(x_0 \mid \theta) f_\theta(\theta) d\theta},$$

which is the Bayes' rule. Rewritten:

$$f_{\theta|X=x_0}(\theta \mid x_0) \propto f_{X|\theta}(x_0 \mid \theta) f_\theta(\theta),$$

in words: posterior $\propto$ likelihod $\times$ prior

$\propto$ ... proportional to

Information update

$X = x_0$ and/or θ can even be vectors:

$$f_{\theta|X=x_0}(\theta \mid x_0) \propto f_{X|\theta}(x_0 \mid \theta) f_\theta(\theta)$$

- our original (prior) information (belief) about θ was updated by observing $X = x_0$ into the posterior
- this can be applied recursively:

$$\begin{aligned} f_{\theta|X=x,Y=y}(\theta \mid x_0, y_0) &= f_{Y,X|\theta}(x_0, y_0 \mid \theta) f_\theta(\theta) \\ &= f_{Y|\theta}(y_0 \mid \theta) \underbrace{f_{X|\theta}(x_0 \mid \theta) f_\theta(\theta)}_{\text{old posterior}}, \end{aligned}$$

All available information about θ is summarized by the posterior.

The Bayesian Approach

- let us denote the data set D, its realization d, and θ the parameter(s).
- the Bayesian model assumes
  - the nature picks θ from the prior distribution f_θ
  - the nature generates data set $D = d$ from the likelihood $f_{D|\theta}$
- the posterior

$$f(\theta \mid D = d) \propto f(d \mid \theta) f(\theta).$$

  provides answers for all statistical tasks
  - point estimation
  - interval estimation
  - prediction
  - model selection
  - hypothesis testing?

Point estimation

Goal: a numerical value $\widehat{\theta}$ compatible with the data

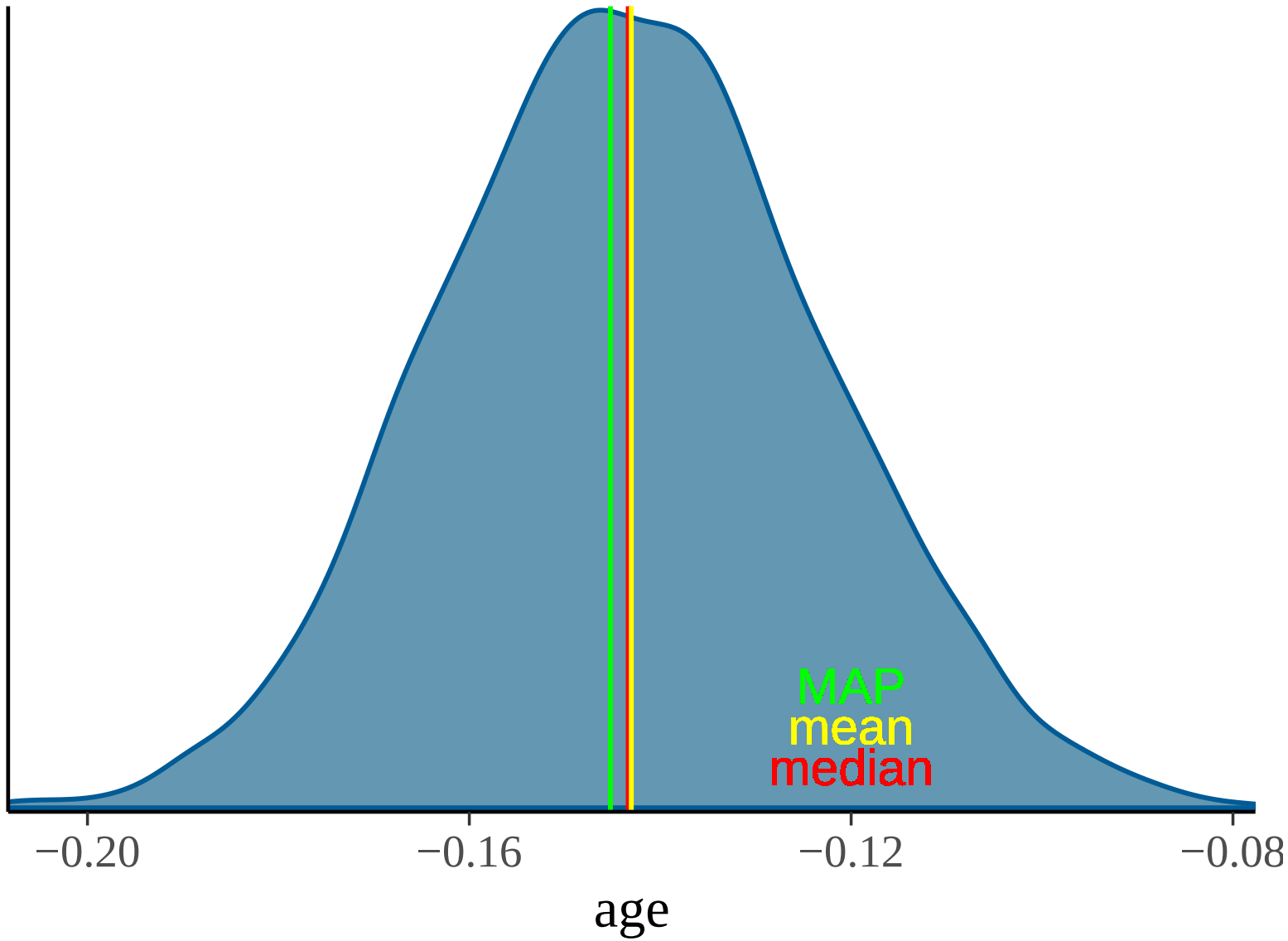
Frequentist approach:

- MLE
- method of moments
- optimization (e.g. penalized least squares), etc.

Bayesian approach:

- MAP - Maximum A Posterior estimate
  - the maximum of the posterior density
  - close to frequentist MLE
- posterior mean - the expected value of the posterior
- posterior median
- generally: minimizing the expected loss
  - any loss function we can come up with
  - the expectation is calculated under the posterior

Point Estimation



Interval Estimation

Goal: a range of values $\widehat{\theta}$ compatible with the data

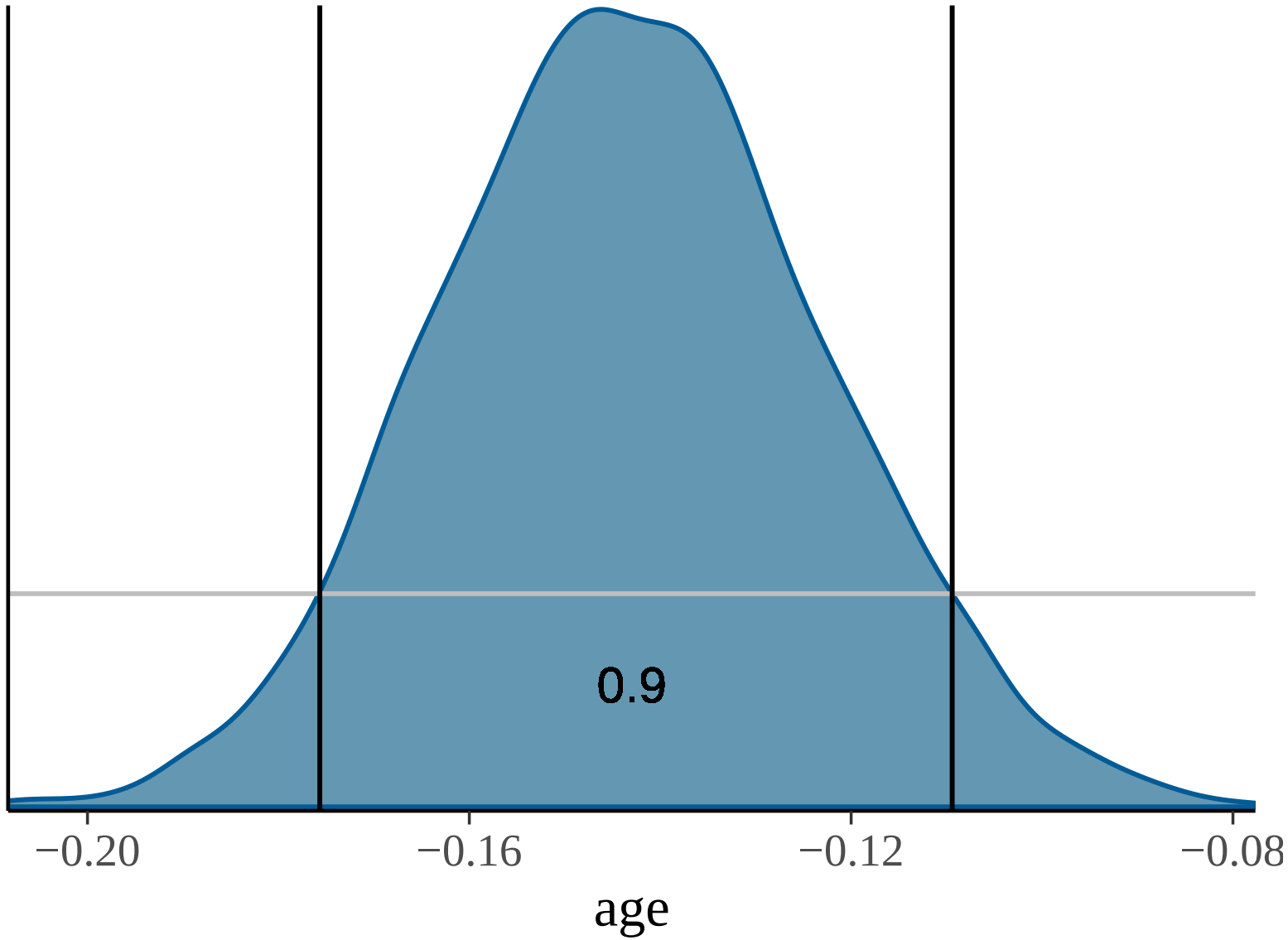
Frequentist approach: a confidence interval $CI_{1-\alpha}$

- connected to significance testing
- cannot be interpreted in simple probabilistic terms

Bayesian approach: a credible region $CR_{1-\alpha}$

- a subset of Θ such that $P(\theta \in CR_{1-\alpha}) = 1 - \alpha$
  - probability calculated under the posterior
- simple interpretation
- many options (just as in the frequentist context), the narrowest possible is called the *highest posterior density interval*

Interval Estimation



Prediction

Goal: predict new data points $D^\star$ based on the observed data $D = d$ and the model

Frequentist approach: varies

Bayesian approach: prediction = estimation

- treat $D^\star$ as parameters but the likelihood satisfies $f_{D,D^\star|\theta} = f_{D|\theta} \cdot f_{D^\star|\theta}$, i.e. new and old data are independent given parameters

$$\begin{aligned} f_{\theta,D,D^\star} &= f_{D|D^\star,\theta} \cdot f_{D^\star,\theta} = f_{D|\theta} \cdot f_{D^\star|\theta} \cdot f_\theta \\ &= f_{\theta,D^\star|D} \cdot f_D \end{aligned}$$

- posterior: $f_{\theta,D^\star|D} \propto f_{D|\theta} \cdot f_{D^\star|\theta} \cdot f_\theta$ marginalize for $D^\star$

Model Selection

Goal: decide which of a set M of candidate models fits the data

Frequentist approach: hypothesis testing

Bayesian approach: model selection = estimation (again)

- the data generation process is assumed to have additional level
  - the nature generates a model $M \in \Pi$ based on a prior f_M
  - then it generates θ conditionally on the model from $f_{\theta|M}$
  - finally the data are generated conditionally on the model and parameters from $f_{D|M,\theta}$
- calculate the posterior (now hierarchical):

$$\begin{aligned} f_{D,\theta,M} &= f_{D|\theta,M} \cdot f_{\theta,M} = f_{D|\theta,M} \cdot f_{\theta|M} \cdot f_M \\ &= f_{\theta,M|D} \cdot f_D \end{aligned}$$

posterior: $f_{\theta,M|D} \propto f_{D|\theta,M} \cdot f_{\theta|M} \cdot f_M$ ... marginalize for M again

Example: Bayesian Ridge

Consider a Gaussian linear model $Y = \mathbf{X}\beta + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2 I_{N\times N})$.
Consider the following priors:

- $\beta \sim \mathcal{N}(0, \tau^2 I_{p\times p})$
  - τ^2 is a hyperparameter - either fixed or with some hyperprior f_{τ^2}
- $f_{\sigma^2} \propto 1/\sigma^2$ (improper prior)

Then the posterior for $\beta = (\beta^\top, \sigma^2, \tau^2)$ is given by

$$f_{\theta|\mathbf{X},Y}(\beta, \sigma^2, \tau^2 \mid \mathbf{X}, Y) \propto \frac{1}{\sigma^n} e^{-\frac{1}{2\sigma^2}(Y-\mathbf{X}\beta)^\top(Y-\mathbf{X}\beta)} \frac{1}{\tau^p} e^{-\frac{1}{2\tau^2}\beta^\top\beta} \frac{1}{\sigma^2} f_{\tau^2}(\tau^2)$$

Interestingly, the log-posterior for β is

$$\log f_{\dots}(\beta \mid \mathbf{X}, Y, \sigma^2, \tau^2) \propto -\frac{1}{2\sigma^2}(Y - \mathbf{X}\beta)^\top(Y - \mathbf{X}\beta) - \frac{1}{2\tau^2}\beta^\top\beta$$

so MAP here gives the ridge estimator for $\lambda = \sigma^2/\tau^2$

Computational Difficulty

The Bayesian approach above is

- conceptually straightforward and holistic, but
- in practice requires computationally demanding integration
  - the normalization constant
  - marginalization
  - calculating expectations

Possible solutions:

- analytic approximations to the posterior (e.g. Laplace)
- Monte Carlo
  - but the MC techniques we saw already are useful mostly in low-dimensional problems
  - Markov Chain Monte Carlo (MCMC): explore the space in a dependent way, focusing on the important regions

Section 1

Markov Chain Monte Carlo (MCMC)

MCMC

Goal: calculate $\mathbb{E}g(X)$ for some function g

Monte Carlo (MC):

- draw independently $X_1, \ldots, X_N \overset{\perp\!\!\!\perp}{\sim} X$
- approximate $\mathbb{E}g(X)$ empirically by $N^{-1} \sum_n g(X_n)$
  - works due to LLN

Markov Chain Monte Carlo (MCMC):

- draw $X^{(1)}, X^{(2)}, \ldots, X^{(T)}$ as a Markov Chain with its *stationary distribution* equal to that of X
- approximate $\mathbb{E}g(X)$ empirically by $T^{-1} \sum_t g(X_n)$
  - works due to the *ergodic theorem* (LLN for Markov sequences)

Markov Chains

Definition (informal): A sequence of random variables $\{X^{(t)}\}_{t\geq 0}$ with values in $\mathcal{X} \subset \mathbb{R}^p$ such that

$$X^{(t+1)} \mid X^{(t)}, X^{(t-1)}, \ldots, X^{(0)} \sim X^{(t+1)} \mid X^{(t)}$$

is called a discrete-time *Markov chain*.

- the conditional distribution $X^{(t+1)} \mid X^{(t)}$ is given by the *transition kernel* $k(x, y)$
  - for $X^{(t)} = x$, the density for $X^{(t+1)}$ is $k_x(y) := k(x, y)$
  - formally, k has to meet some conditions on measurability and integrability
  - a Markov chain is fully determined by the transition kernel!
- a distribution f is called the stationary distribution of a Markov chain associated with a transition kernel k if

$$\int_{\mathcal{X}} k(x, y) f(x) dx = f(y).$$

Detailed Balance

Claim: If the following *detailed balance condition* holds

$$k(x, y)f(x) = k(y, x)f(y)$$

for a distribution f and a transition kernel k, then f is the stationary distribution of the Markov chain associated with k.

- k specifies the amount of flow between the points in the domain $\mathcal{X}$
- detailed balance: the forward flow $x \rightsquigarrow y$ is equal to the backward flow $y \rightsquigarrow x$
- let f_t denote the marginal distribution of $X^{(t)}$
  - f_0 is the initial distribution
  - the update $f_t \rightsquigarrow f_{t+1}$ is governed by k
  - no update $\Leftrightarrow$ f_t is the stationary distribution f
  - if $f_0 = f$, there will never be an update ... $f_t = f$ for all t
  - detailed balance: $f_t \to f$ for $t \to \infty$ regardless of f_0

MCMC

Goal: construct a chain with a pre-specified stationary distribution f, typically $f_{\theta|D=d}$

- **Q**: how to actually do this? (next slide)
- chain = function that generates $X^{(t-1)}$ depending on X
  - the transition kernel k is in the background

MCMC is more widely applicable than MC, but what about *mixing*?

- we initialize our chain from $f_0 \neq f$
  - because if we could draw from f, we would be doing MC instead
- after a *while* $f_t \approx f$ so we have our first draw $X^{(t)} \dot{\sim} f$
  - discard $X^{(0)}, \ldots, X^{(t-1)}$ and continue the chain (now stationary)
  - **Q**: but what is a *while*? (hard to tell...)

Metropolis-Hastings

Idea: start from a proposal chain with a wrong f (e.g. a random walk, which has no f) and tweak it to the target f.

- detailed balance requires the right amount of flow between all $x, y \in \mathcal{X}$
- if there is too much flow $x \rightsquigarrow y$, re-map some part of it as $x \rightsquigarrow x$

Metropolis-Hastings (MH) algorithm:

- **Input**: a proposal chain $\{U^{(t)}\}$ with kernel k, the target f
- **for** $t = 1, 2, \ldots$
  - set $X^{(t)} := U^{(t)}$ with probability

$$\alpha(X^{(t-1)}, U^{(t)}) = \min\left(1, \frac{f(U^{(t)})k(U^{(t)}, X^{(t-1)})}{f(X^{(t-1)})k(X^{(t-1)}, U^{(t)})}\right)$$

  - otherwise set $X^{(t)} := X^{(t-1)}$

(if the proposal is symmetric, k vanishes from the formula above)

MH with a Symmetric Proposal

Let $U^{(0)} = \epsilon_0$ and $U^{(t+1)} = X^{(t)} + \epsilon_t$, $t \geq 1$, where $\epsilon_0, \epsilon_1, \ldots$ drawn independently a density symmetric around zero.

Verifying detailed balance is relatively simple in this case:

- detailed balance: $k(x,u)f(x) = k(u,x)f(u)$ for $x \rightsquigarrow u$
- $k(x,u)$ is given implicitly as the mixture of
  - moving away $x \rightsquigarrow u$ with probability $\alpha(x,u) = \min\left(1, f(u)/f(x)\right)$
    - u is drawn from $\varphi_x(u)$ a symmetric density around x
    - equal to $\varphi_x(u)\alpha(x,u)$
  - staying at x with probability $1 - \alpha(x,u)$
    - i.e. $x = u$ ... detailed balance trivially satisfied
- detailed balance: $\cancel{\varphi_x(u)}\alpha(x,u)f(x) = \cancel{\varphi_u(x)}\alpha(u,x)f(u)$
- this is trivial since for $f(x) \neq f(u)$ it is
  - either $\alpha(u,x) = 1$ and $\alpha(x,u) = f(u)/f(x)$ leading to

$$\frac{f(u)}{f(x)} f(x) = f(u).$$

  - or the other way around

MH Remarks

- f is a posterior, evaluations needed *up to normalization*
- MH similar in flavor to rejection sampling (RS) in MC
  - but RS needs a majorizing proposal g to decide sample vs. reject
  - MCMC instead moves vs. stays $\Rightarrow$ no majorization needed
- when sampling from a continuous f, repeated values have probability 0
  - yet MH produces repeated values quite commonly

Def.: acceptance rate for MH is the average acceptance probability

$$\bar{\alpha} = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \alpha(X^{(t-1)}, U^{(t)})$$

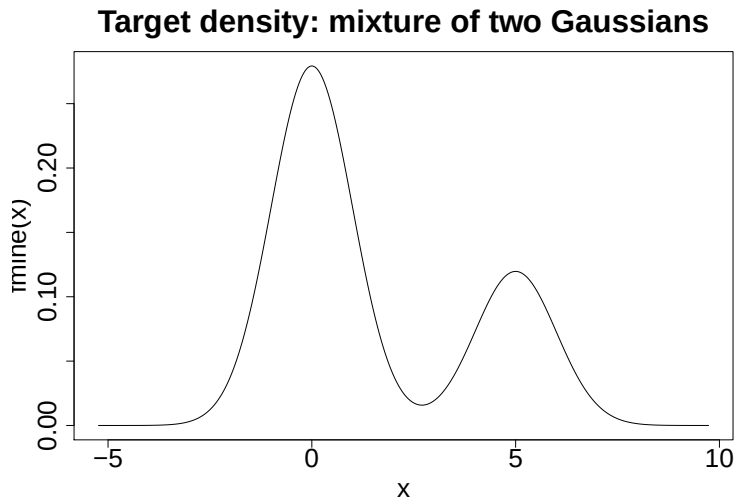
- if $\bar{\alpha}$ too large, we are probably not exploring the space, mostly staying close with our proposals to where we already were
- if $\bar{\alpha}$ too small, we have a lot of repeated values in our sample and hence the effective sample size is small even for large T
- 10-50% tends to be a good rate in practice

Example

Consider the MH algorithm with a Gaussian random walk proposal to sample from a Gaussian mixture model

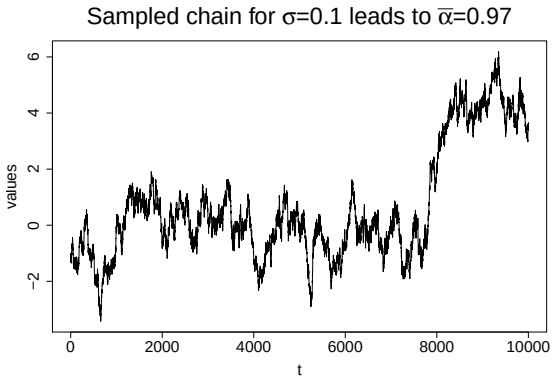
$$f_{\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\tau}(x) = \tau\varphi_{\mu_1,\sigma_1^2}(x) + (1-\tau)\varphi_{\mu_2,\sigma_2^2}(x)$$

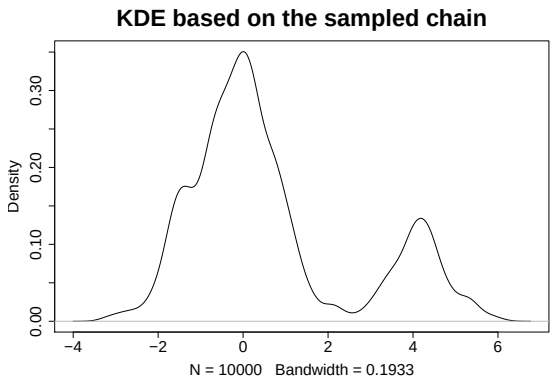
with $\mu_1 = 1, \mu_2 = 5, \sigma_1 = \sigma_2 = 1$ and $\tau = 0.7$.

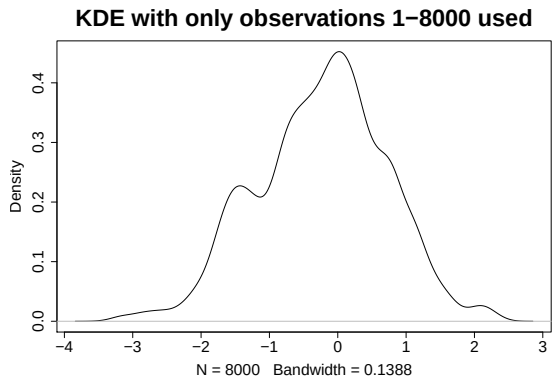


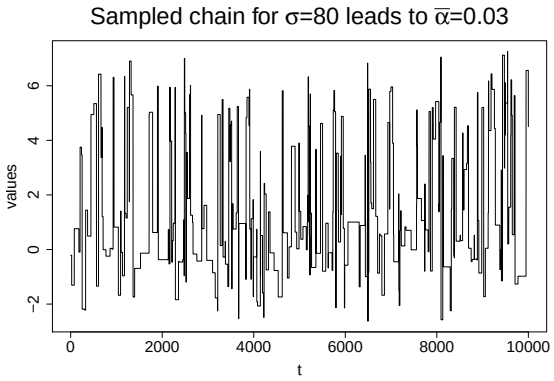
- Since we require the proposal to be symmetric
  - the mean of the proposal needs to be zero: $\mu = 0$
  - only σ^2 of the proposal has to be chosen

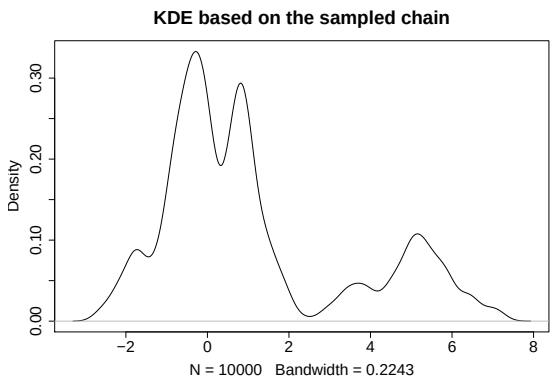
Example

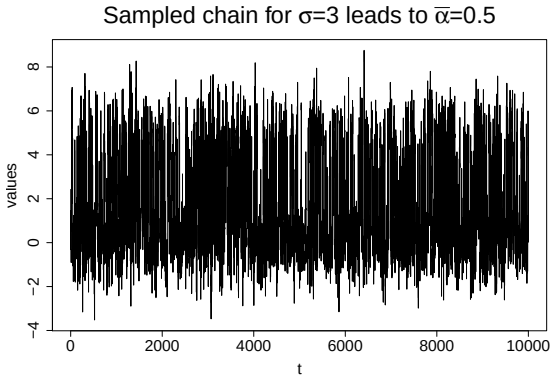

Sampled chain for σ=0.1 leads to ᾱ=0.97
values
t

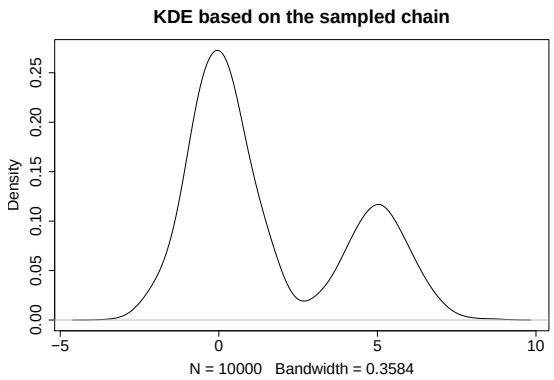

KDE based on the sampled chain
Density
N = 10000 Bandwidth = 0.1933


KDE with only observations 1−8000 used
Density
N = 8000 Bandwidth = 0.1388


Sampled chain for σ=80 leads to ᾱ=0.03
values
t


KDE based on the sampled chain
Density
N = 10000 Bandwidth = 0.2243


Sampled chain for σ=3 leads to ᾱ=0.5
values
t


KDE based on the sampled chain
Density
N = 10000 Bandwidth = 0.3584

Assignment 8: problem setting

Since black carbon (BC) is a pollutant known for its adverse health effects, it is of interest to monitor BC mass concentration in urban areas. While stationary measurement devices are able to precisely record BC concentrations (random variable X), simpler mobile devices provide more flexibility – at the cost of some measurement noise ε. For a mobile measurement, only $Y = X + \varepsilon$ is observed, where we may assume $\varepsilon \sim N(0, \sigma^2)$ with known standard deviation $\sigma = 0.6\,\mu g m^{-3}$ based on lab experiments.

To obtain a detailed overview over the BC concentrations in Lausanne, the aim is to add mobile measurements to those of available stationary devices. We want to predict X given a noisy observation $Y = y$ using a Bayesian approach, with stationary measurements motivating a priori a Weibull distribution for X with shape parameter 2 and scale parameter 1.2 (median $\approx 1\,\mu g m^3$).

Assignment 8: tasks [5%]

- Implement a Metropolis-Hastings-Algorithm for obtaining MCMC samples of $X \mid Y = y$ using $N(y, 0.6)$ as *fixed proposal distribution for all iterations* (note that this is an asymmetric proposal from the MH perspective!).
- Run your algorithm for $y = 0.5$, $y = 1$ and $y = 2$ for illustration. In each run, draw 10000 samples after a burn-in of 1000 (less if it takes too long).
- In this specific scenario, the posterior is in fact analytically available with the R code for the density function provided on the next slide. Graphically compare the empirical distributions of your MCMC samples with the true posterior densities and the proposal densities for the three considered values of y.

Hint: It might be wise to compute $\log(\alpha)$ first in dependence on log-densities to avoid numerical issues.

Assignment 8: true posterior density

```
dposterior <- function(x, y, scale = 1.2, sd = .6) {
  # x: evaluation points of the density
  # y: observation Y=y (length 1),
  # scale: scale parameter of Weibull prior (shape=2 fixed)
  # sd: standard deviation of Gaussian error (mean=0 fixed)
  a <- 1/2*1/sd^2; c <- 1/scale^2
  erf <- function(x) 2*pnorm(x*sqrt(2)) - 1
  k <- ifelse(x >= 0, x * exp( -a * (x-y)^2 - c*x^2 ), 0)
  n <- exp(-a*(y^2)) *
    (sqrt(pi) * a * y * exp(a^2*y^2 / (a+c)) *
       (erf(a*y/sqrt(a+c)) + 1) +
       sqrt(a + c) ) / (2* (a+c)^(3/2))
  k/n
}
```