Week 11: Bayesian Computations

MATH-517 Statistical Computation and Visualization

Linda Mhalla

2024-11-29

Section 1

Bayesian Inference

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 (θ fixed)
- likelihood & prior $=$ Bayesian model (θ random)

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 $\theta = (\psi, \lambda)$ can even be vectors:

$$f_{\psi|X=x_0}(\psi \mid x_0) \propto \int f_{X|\theta}(x_0 \mid \theta) f_\theta(\psi, \lambda) d\lambda$$

- our original (prior) information (belief) about θ was updated by observing $X = x_0$ into the (marginal) posterior
- this can be applied recursively (when a new Y independent of X arrives):

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

All available information about θ is summarized by the posterior (provides a complete inferential scope)

Prior Densities

- the prior distribution quantifies the researcher's uncertainty about parameters before observing data
- choice of the prior density is important: it is based on the best available information $\rightarrow$ subjective
- sometimes use an **improper** prior, which is not a true density (has infinite integral) but for which the posterior is a true density
- often use a **non-informative** proper prior, which inputs only **weak information** ($\neq$ ignorance) (e.g., normal density with very large (finite) variance, Jeffreys prior based on Fisher information matrix, or uniform prior though not transformation-invariant)
- **conjugate** priors make computations easy as they yield a posterior density of the same family (e.g. beta prior/binomial data $\rightarrow$ beta posterior or gamma prior/Poisson data $\rightarrow$ gamma posterior)

Note: Empirical Bayes can be used to select parameters of the prior (approximate prior distribution by frequentist methods). Other techniques: hierarchical Bayes or maximum entropy

Prior Densities

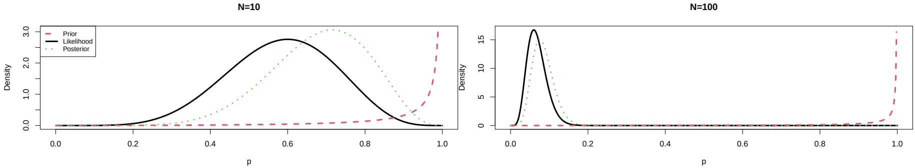
- as sample size increases, the effect of the prior is washed out

E.g., Bernoulli case

- likelihood: $\Pr(\mathbf{x}_{1:N}|p) = p^{\sum_{i=1}^{N} x_i}(1-p)^{N-\sum_{i=1}^{N} x_i}$
- beta prior $p \sim \text{Beta}(a, b)$:

$$\Pr(p) \propto p^{a-1}(1-p)^{b-1}$$

  on the interval $(0, 1)$
- posterior: $\Pr(p|\mathbf{x}_{1:N}) \propto \Pr(\mathbf{x}_{1:N}|p) \Pr(p)$



Prior Densities: Example

- **Improper Prior**

If $x \sim \mathcal{N}(\theta, 1)$ and $f_\theta(\cdot) \equiv \omega$ (constant), then posterior dist. of θ is

$$f_{\theta|x}(\theta \mid x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x-\theta)^2}{2}\right\},$$

i.e., corresponds to a $\mathcal{N}(x, 1)$ distribution $\Rightarrow$ independent of the prior

- **Non-informative (flat) prior**

Consider $x \sim \mathcal{N}(\theta, 1)$ and $\theta \sim \mathcal{N}(0, 10)$

$$\begin{aligned} f_{\theta|x}(\theta \mid x) \propto f(x \mid \theta) f_\theta(\theta) &\propto \exp\left\{-\frac{(x-\theta)^2}{2} - \frac{\theta^2}{20}\right\} \\ &\propto \exp\left(-\frac{11\theta^2}{20} + \theta x\right) \propto \exp\left[-\frac{11}{20}\{\theta - (10x/11)\}^2\right] \end{aligned}$$

and

$$\theta \mid x \sim \mathcal{N}\left(\frac{10}{11}x, \frac{10}{11}\right)$$

The Bayesian Approach

- let us denote the data set D, its realization d, and θ the parameter(s)
- the Bayesian model assumes
  - that nature picks θ from the prior distribution f_θ
  - that 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? a matter of choosing priors reflecting the hypotheses
- **Uncertainty:**
  - How much prior belief about θ changes in light of data (Bayesian)
  - How estimates vary in repeated sampling from the same population (frequentist)

Point estimation

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

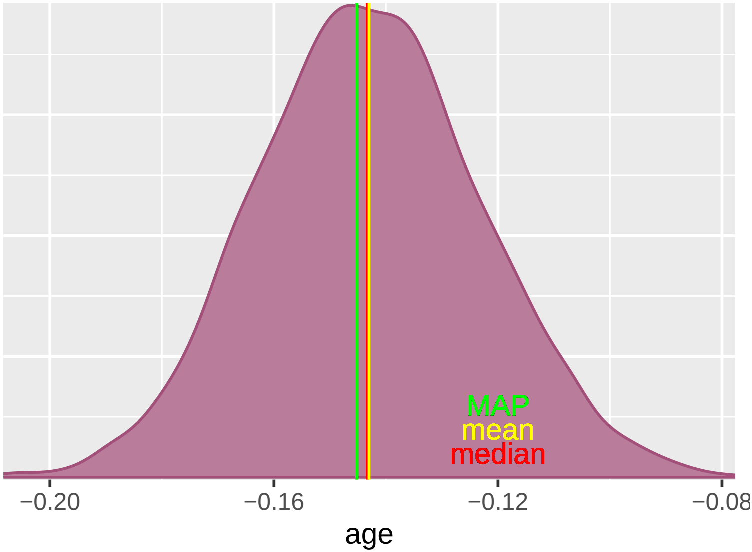
Frequentist approach:

- 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
  - the expectation is calculated under the posterior
  - e.g., for the squared error loss $L(\theta, a) = (\theta - a)^2$, the posterior expected loss $\int_\Theta (\theta - a)^2 dF_{\theta|D}(\theta)$ is minimized at the mean of the posterior distribution; $|\theta - a|$ yields the posterior median

Point Estimation



Interval Estimation

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

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

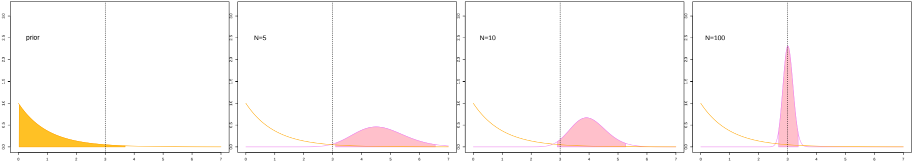
- dual to significance testing
- the probability that the interval contains the true parameter *under replication of the data* is $1-\alpha$

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

- a subset of Θ such that $P(\theta \in CR_{1-\alpha} \mid D) = 1-\alpha$
  - probability calculated under the posterior
- simple interpretation: given the model and data, the probability that the true parameter is in the credible set is $1-\alpha$
- Infinitely many such intervals/regions
- many options (just as in the frequentist context), most used: equal-tailed interval and the *highest posterior density set* (narrowest possible)

Interval Estimation: Equal-tailed Interval

- $CR_{1-\alpha} = [q_{\alpha/2}, q_{1-\alpha/2}]$ where q_α is the α-quantile of the posterior distribution $f_{\theta|D}$



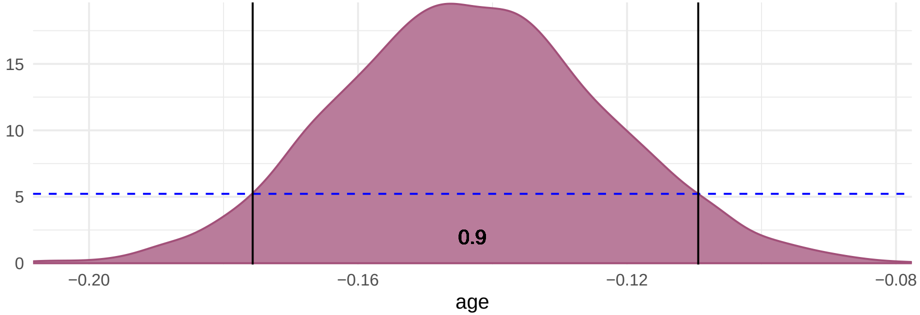
- credible interval influenced by the prior
- credible interval gets narrower with increasing N
- may include values with lower probability than those excluded, unless the posterior is unimodal and symmetric

Interval Estimation: Highest Posterior Density Set

- $\int_{\Theta \cap CR_{1-\alpha}} f_{\theta|D}(\theta \mid d) d\theta = 1 - \alpha$ such that

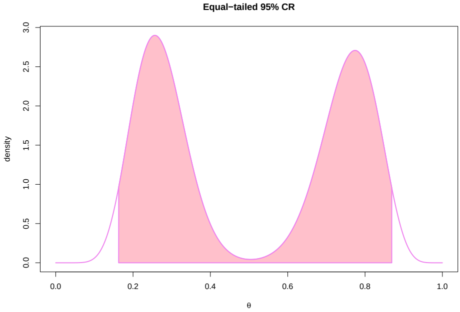
$$f_{\theta|D}(\theta \mid d) \geq f_{\theta|D}(\theta' \mid d)$$

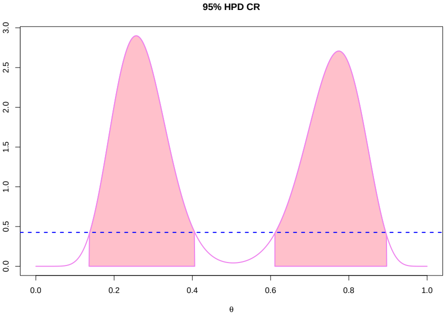
  for all $\theta \in CR_{1-\alpha}$ and $\theta' \notin CR_{1-\alpha}$



- not necessarily an interval: if the posterior is multimodal, the HPD set may be an union of distinct intervals (or distinct contiguous regions)

Interval Estimation: Bimodal Example





Interval Estimation: Example

- Data $X_i \mid \mu, \tau \sim N\left(\mu, \tau^{-1}\right), i = 1, \ldots, N$. Suppose τ is known, and that we use prior $\mu \sim N\left(\mu_0, \tau_0^{-1}\right)$ for some fixed values of μ_0 and $\tau_0 \geq 0$
- The corresponding posterior distribution of μ is

$$\mu \mid x \sim N\left(\mu_p, \tau_p^{-1}\right)$$

where $\tau_p = N\tau + \tau_0$ and $\mu_p = \frac{N\tau}{\tau_0 + N\tau}\bar{x} + \frac{\tau_0}{\tau_0 + N\tau}\mu_0$

Hence, 95% credible interval (in this case also 95% HPD region):

$$\left[\mu_p - z_{0.025}\sigma_p, \mu_p + z_{0.025}\sigma_p\right]$$

$\Rightarrow$ a priori information about a parameter decreases our (posterior) uncertainty about it

Note that the credible interval corresponding to the noninformative prior
$\left[\bar{x} - z_{0.025}\sigma/\sqrt{N}, \bar{x} + z_{0.025}\sigma/\sqrt{N}\right]$

coincides with the classical (frequentist) confidence interval

Prediction of Future Observations

Goal: posterior prediction, i.e., evaluating or sampling from the posterior predictive distribution $f_{\tilde{D}|D}$, where D is observed data and $\tilde{D}$ is yet to be observed data

Bayesian approach: prediction = estimation

- assume that likelihood satisfies $f_{D,\tilde{D}|\theta} = f_{D|\theta} \cdot f_{\tilde{D}|\theta}$, i.e., new and old data are independent given parameters
- then

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

- joint posterior: $f_{\theta,\tilde{D}|D} = f_{\tilde{D}|\theta} \cdot f_{\theta|D} \Rightarrow$ marginalize out θ

$$f_{\tilde{D}|D}(\tilde{d} \mid d) = \int_\Theta f_{\tilde{D}|\theta}(\tilde{d} \mid \theta) \cdot f_{\theta|D}(\theta \mid d) \, \mathrm{d}\theta$$

$\rightarrow$ estimated by MC if we can draw from posterior $f_{\theta|D}$

Model Selection

Consider a discrete set $\mathcal{M}$ of candidate models indexed by M (a parameter)

Goal: decide which candidate model fits best the data

E.g., $\mathcal{M}$ can be a mixture of K Gaussians (K is discrete random variable)

Frequentist approach: hypothesis testing, e.g., LRT

Bayesian approach: model selection = estimation (again)

- the data generation process is assumed to have additional level
  - the nature generates a model $M \in \mathcal{M}$ 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 out θ again
- select the MAP model

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 $\theta = (\beta, \sigma^2, \tau^2)^\top$ 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 2

Markov Chain Monte Carlo (MCMC)

MC versus MCMC

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

Monte Carlo (MC):

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

Markov Chain Monte Carlo (MCMC):

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

$$\frac{1}{T} \sum_t g(X_t) \underset{T \to \infty}{\overset{\text{a.s.}}{\to}} \mathbb{E}g(X),$$

for any bounded function $g : \mathcal{X} \to \mathbb{R}$

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 cond. density of $X^{(t+1)}$ is $k_x(y) := k(x, y)$
  - 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

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

  If $f_{t+1}(y) = \int k(x, y) f_t(x) dx = f_t(y)$, then we stay in the distribution f_t forever

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 **a** stationary distribution of the MC 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 $y \rightsquigarrow x$
- equilibrium distribution is preserved: if $x \sim f$ before a transition, then this is also true afterwards
- 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

Ergodicity

Definitions:

- A chain verifying the detailed balance condition (time-reversibility) is called *invariant*
- A chain is called *irreducible* if any point can be reached (using the kernel) starting from anywhere else

$$\forall u, v \in \mathcal{X}, \quad \exists\, t \text{ s.t. } P(X^{(t)} = u \mid X^{(0)} = v) > 0$$

Result: An irreducible and aperiodic Markov chain converges to a unique distribution, called stationary distribution

- An irreducible, aperiodic, and invariant chain is called *ergodic*

Theorem: If a chain is ergodic then its unique stationary distribution is the invariant distribution and $f_t \to f$ for $t \to \infty$ regardless of f_0

Running Monte Carlo via Markov Chains

Goal: For an arbitrary starting value $X^{(0)}$, construct a chain with a pre-specified stationary distribution f, typically the posterior $f_{\theta|D=d}$

- chain = function that generates $X^{(t+1)}$ depending on $X^{(t)}$
  - the transition kernel k is in the background
- produce a dependent sample $X^{(T_0)}, X^{(T_0+1)}, \ldots$, marginally generated from f, sufficient for most approximation purposes

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
  - need to ensure irreducibility
- after a *while* $f_{T_0} \approx f$ so we have our first draw $X^{(T_0)} \dot{\sim} f$
  - discard $X^{(0)}, \ldots, X^{(T_0-1)}$ and continue the chain (now stationary)
  - need to ensure invariance (detailed-balance)

Problem: How to build a Markov chain with a given stationary distribution? We will see some recipes

Metropolis–Hastings

Idea: construct a candidate new value y by drawing from arbitrary conditional density $q(y \mid x)$ (called *proposal* distribution)

- 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 density $q(y \mid x)$, the target f (up to a constant)
- **for** $t = 1, 2, \ldots$, update $X^{(t-1)}$ to $X^{(t)}$ by
  - generate $U^{(t)} \sim q(\cdot \mid X^{(t-1)})$
  - define
$$\alpha(X^{(t-1)}, U^{(t)}) = \min\left\{1, \frac{f(U^{(t)})q(X^{(t-1)} \mid U^{(t)})}{f(X^{(t-1)})q(U^{(t)} \mid X^{(t-1)})}\right\}$$
  - set $X^{(t)} := U^{(t)}$ with probability $\alpha(X^{(t-1)}, U^{(t)})$
  - otherwise set $X^{(t)} := X^{(t-1)}$

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

MH: Convergence Properties

- MH Markov Chain satisfies the detailed balance condition with

$$k(y,x) = \alpha(x,y)\, q(y \mid x) + \int \{1 - \alpha(x,\xi)\} q(\xi \mid x) d\xi \delta_x(y)$$

  where δ is the Dirac mass
- If $q(y \mid x) > 0$, $\forall x, y$, then the chain is irreducible
- If

$$P\left(\frac{f(U^{(t)}) q(X^{(t-1)} \mid U^{(t)})}{f(X^{(t-1)}) q(U^{(t)} \mid X^{(t-1)})} \geq 1 \right) < 1,$$

  that is the event $\{X^{(t+1)} = X^{(t)}\}$ is possible, then the chain is aperiodic

Thus, under the above two conditions, we have

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} g\left(X^{(t)}\right) = \int g(x) df(x),$$

for $\mathbb{E}_f |g(X)| < \infty$

MH with Random Walk

- use a local perturbation as proposal

$$U^{(t)} = X^{(t-1)} + \epsilon_t,$$

  where $\epsilon_t \sim g$, independent of $X^{(t-1)}$
- proposal q is a symmetric (around 0) density of the form $q(u - x)$
  - e.g., g is $\mathcal{N}(0, \sigma^2)$ hence $U^{(t)} \sim \mathcal{N}(X^{(t-1)}, \sigma^2)$
  - e.g., g is $\mathcal{U}[-\delta, \delta]$ hence $U^{(t)} \sim \mathcal{U}[X^{(t-1)} - \delta, X^{(t-1)} + \delta]$
- then,

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

MH with Random Walk

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 $\alpha(x,u) = \min\{1, f(u)/f(x)\}$
    - u is drawn from $q(u \mid x)$ a symmetric density around x
    - equal to $k(x,u) = \alpha(x,u)q(u \mid x) = \alpha(x,u)q(u-x) = \alpha(x,u)q(x-u)$
  - staying at x with probability $1 - \alpha(x,u)$
    - i.e., $x = u$ ... detailed balance trivially satisfied
- detailed balance: $\cancel{q(u \mid x)}\alpha(x,u)f(x) = \cancel{q(x \mid u)}\alpha(u,x)f(u)$
- 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

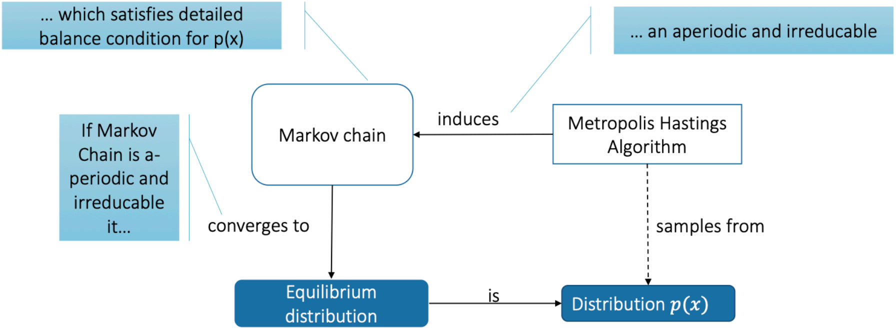
- f is usually a posterior, evaluations needed *up to normalization*
- MH similar in flavor to rejection sampling (RS) in MC
  - but RS needs a majorizing proposal g to decide accept vs. reject
  - MCMC instead moves vs. stays $\Rightarrow$ no majorization needed
- never moves to values with $f(y) = 0$
- the chain $\left(X^{(t)}\right)_t$ may take the same value several times in a row, even when f is a density wrt Lebesgue measure

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
- good rates: 25% (large dimension) - 50% (small dimension)

MH: Summary



credit: Marcel Lüthi,University of Basel

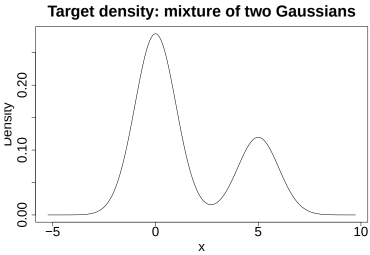
Example: MH with Random Walk

Consider the MH algorithm with

- a Gaussian mixture target model

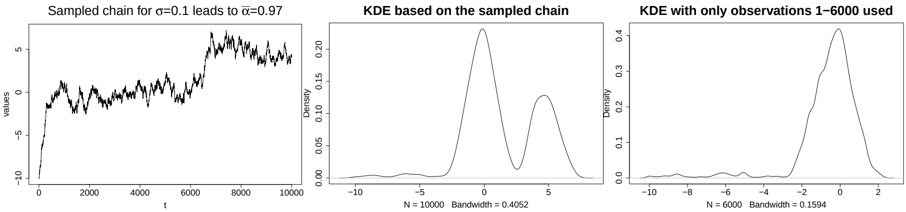
$$f_{\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\tau}(x) = \tau\varphi_{\mu_1,\sigma_1^2}(x) + (1-\tau)\varphi_{\mu_2,\sigma_2^2}(x)$$

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

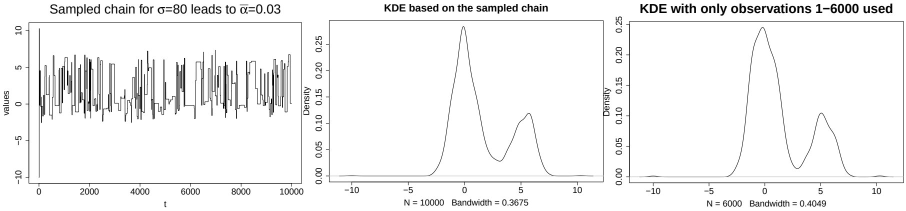


- a Gaussian random walk proposal $y \sim \mathcal{N}(x, \sigma^2)$, with $\sigma = 0.1, 3, 80$
- $x^{(0)} = -10$

Example: MH with Random Walk

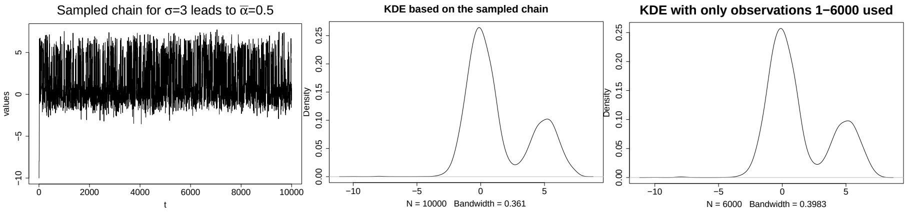


⇒ proposals often accepted but chain moves too slowly



⇒ chain gets stuck for too long

Example: MH with Random Walk



$\Rightarrow$ seems the best

References

C. P. Robert & G. Casella (1999) Monte Carlo Statistical Methods

C. P. Robert & G. Casella (2010) Introducing Monte Carlo Methods with *R*

Gelman & Carlin & Stern & Dunson & Vehtari & Rubin. 2013. Bayesian Data Analysis updated version