Mathematical techniques in data science

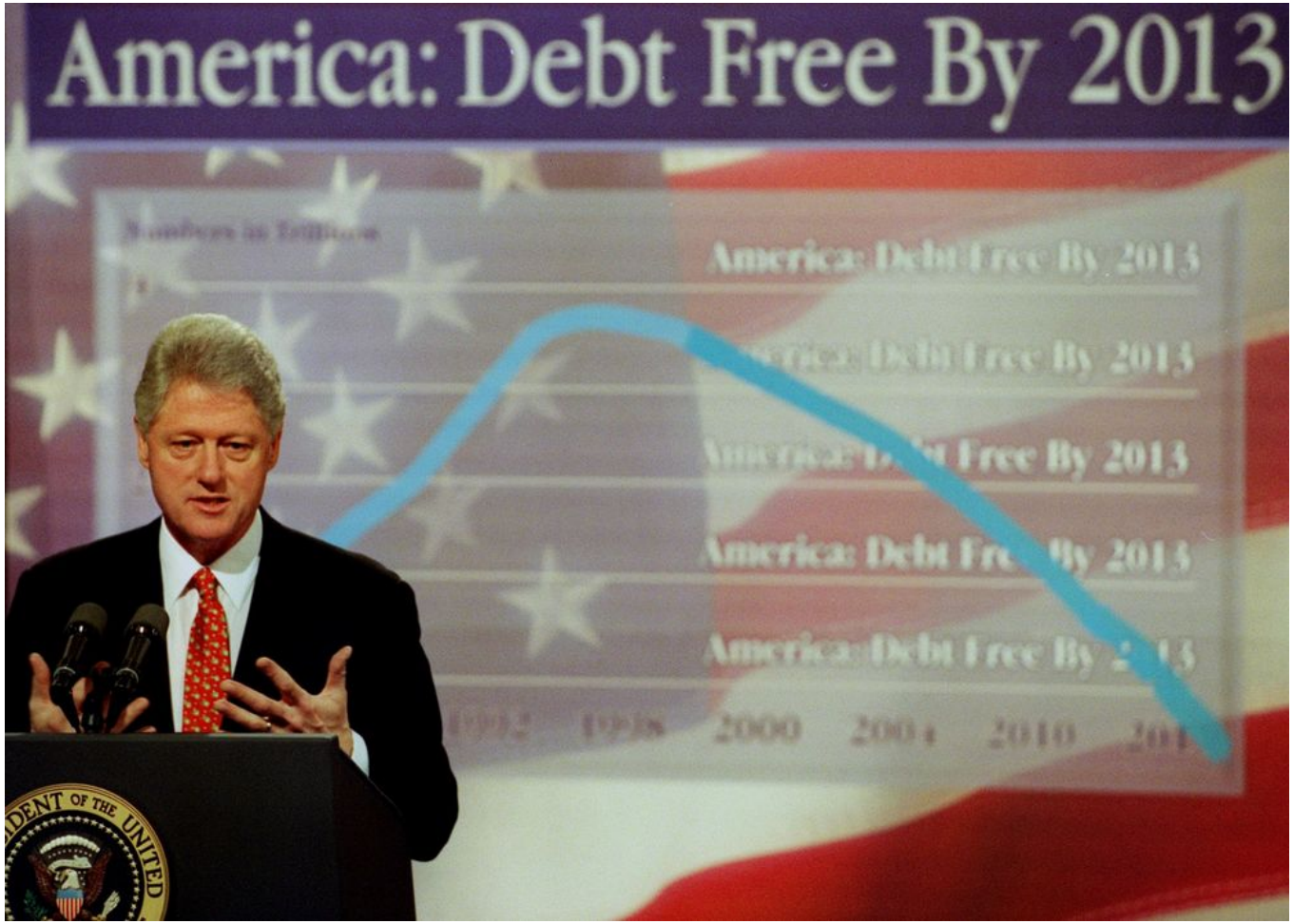
Lecture 28: Bayesian inference and MCMCs

April 29th, 2019

Schedule

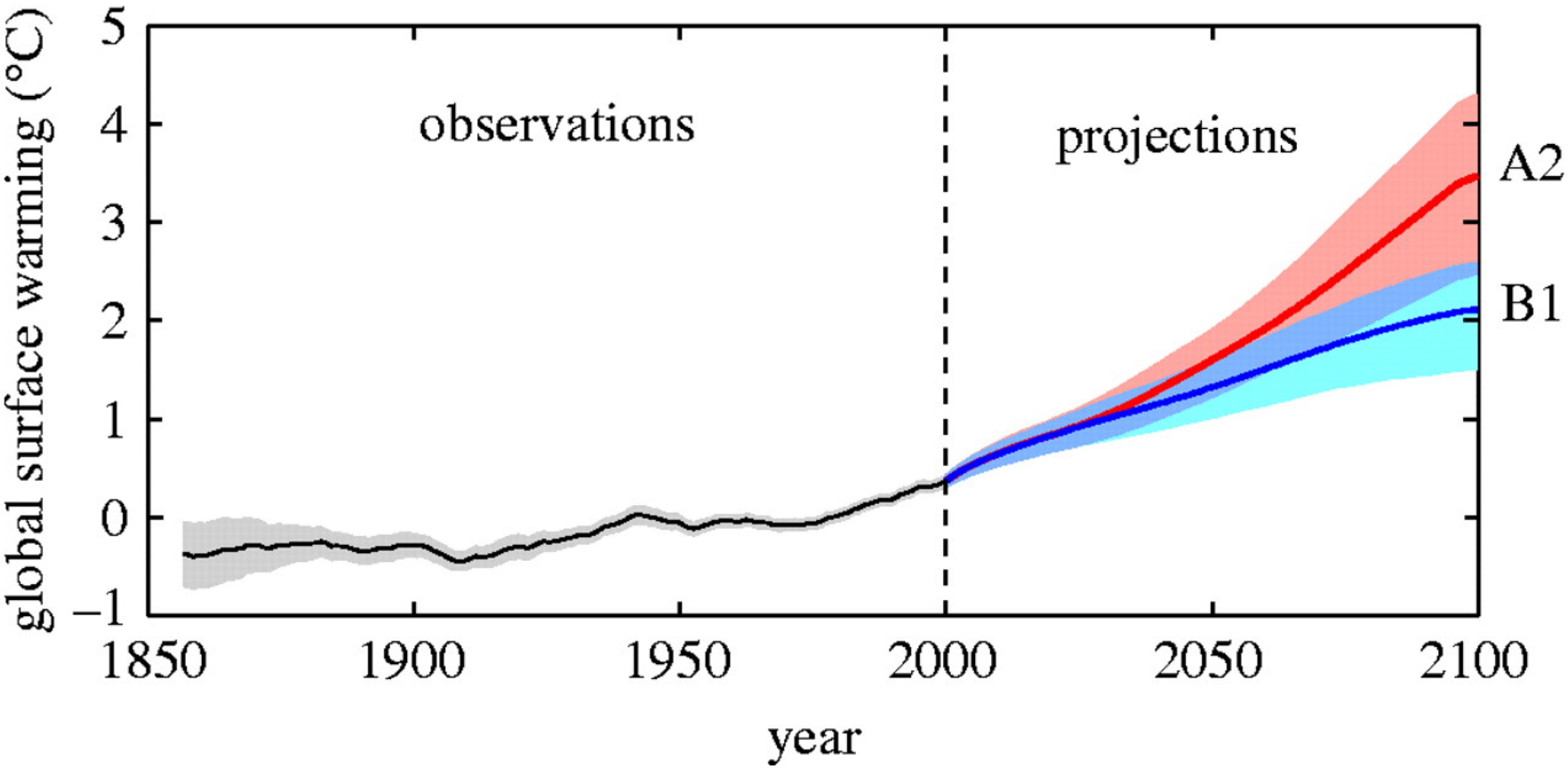
Week	Chapter
1	Chapter 2: Intro to statistical learning
3	Chapter 4: Classification
4	Chapter 9: Support vector machine and kernels
5, 6	Chapter 3: Linear regression
7	Chapter 8: Tree-based methods + Random forest
8	
9	Neural networks
12	PCA $\rightarrow$ Manifold learning
11	Clustering: K-means $\rightarrow$ Spectral Clustering
10	**Bayesian methods + UQ**
13	Reinforcement learning/Online learning/Active learning
14	Project presentation

— It is difficult to make predictions, especially about the future.

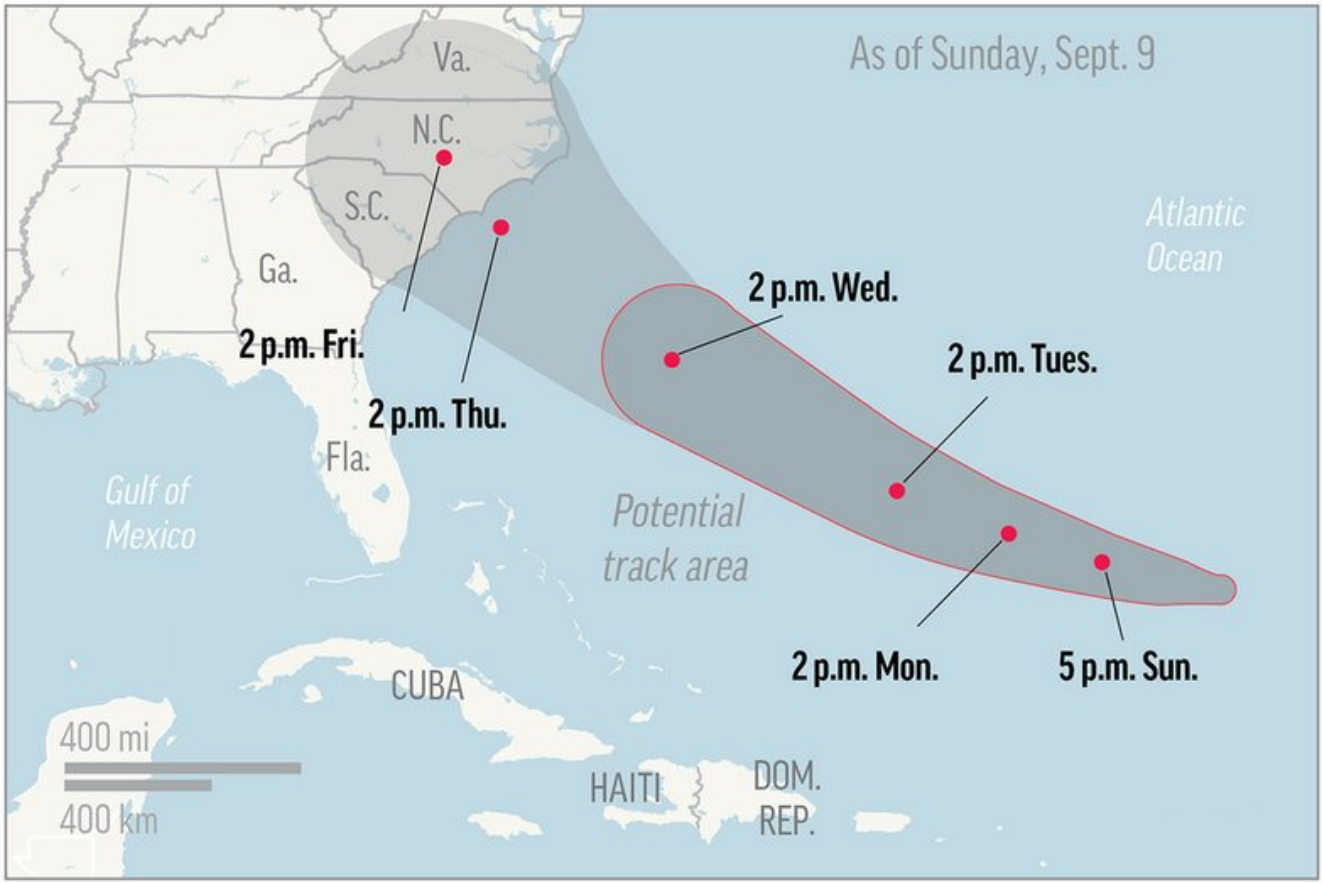


Modelling uncertainties

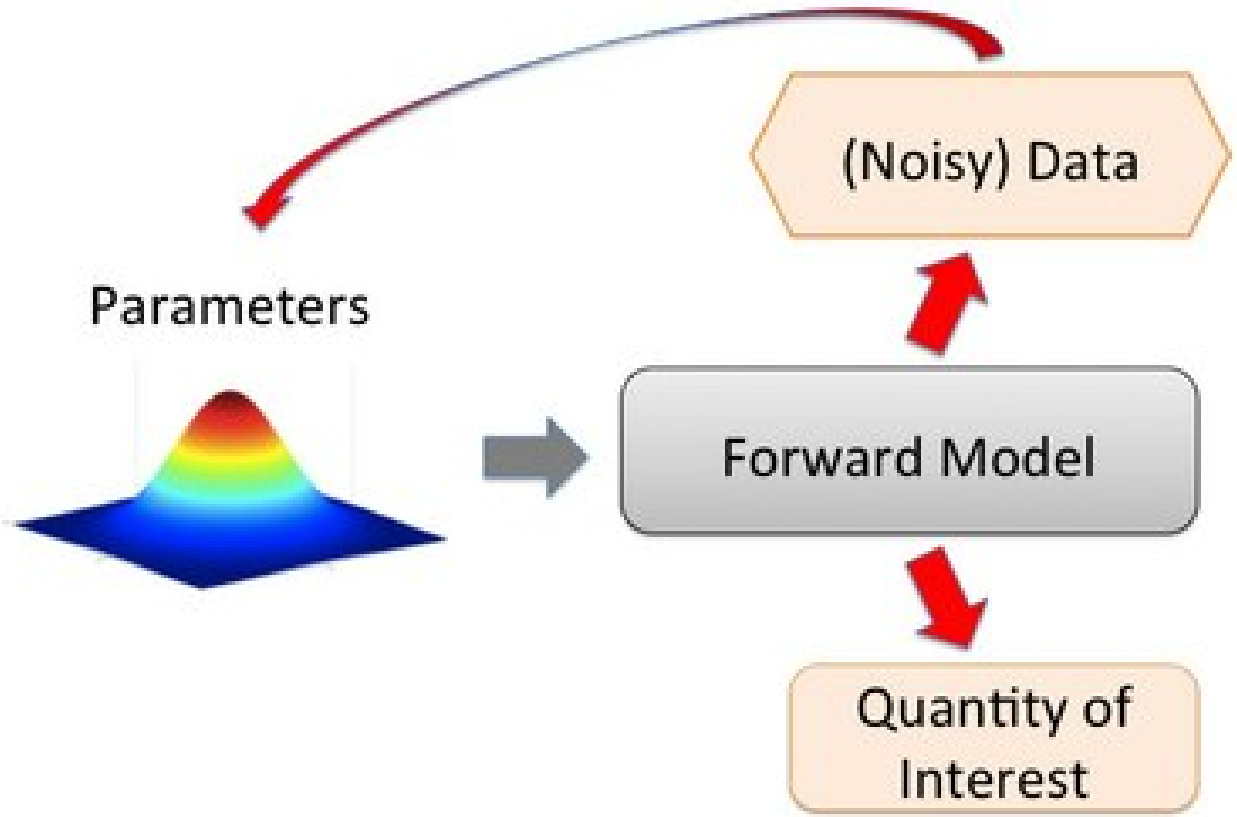
— Data science is about making predictions in the presence of uncertainties



Modelling uncertainties



Uncertainty quantification



Frequentist vs. Bayesian

Frequentist statistics:

- Compute *point* estimates (e.g. maximum likelihood).
- Define probabilities as the long-run frequency of events .

Bayesian statistics:

- Probabilities are a "state of knowledge" or a "state of belief".
- Parameters have a probability distribution.
- Prior knowledge is updated in the light of new data.

Example

You flip a coin 14 times. You get `head` 10 times. What is $p := P(head)$?

- Frequentist approach: estimate p using, say maximum likelihood:

$$p \approx \frac{10}{14} \approx 0.714.$$

Example

You flip a coin 14 times. You get head 10 times. What is $p := P(head)$?

- Frequentist approach: estimate p using, say maximum likelihood:

$$p \approx \frac{10}{14} \approx 0.714.$$

- Bayesian approach: we treat p as a *random variable*.
  1. Choose a *prior* distribution for p, say $P(p)$.
  2. Update the prior distribution using the data via *Bayes' theorem*:

$$P(p|data) = \frac{P(data|p)P(p)}{P(data)} \propto P(data|p)P(p).$$

Example

- Bayesian approach: we treat p as a *random variable*.
  1. Choose a *prior* distribution for p, say $P(p)$.
  2. Update the prior distribution using the data via *Bayes' theorem*:

$$P(p|data) = \frac{P(data|p)P(p)}{P(data)} \propto P(data|p)P(p).$$

Note: "$data|p$" $\sim$ Binomial$(14, p)$. Therefore:

$$P(data|p) = \binom{14}{10} p^{10}(1-p)^4.$$

What should we choose for $P(p)$?

Example

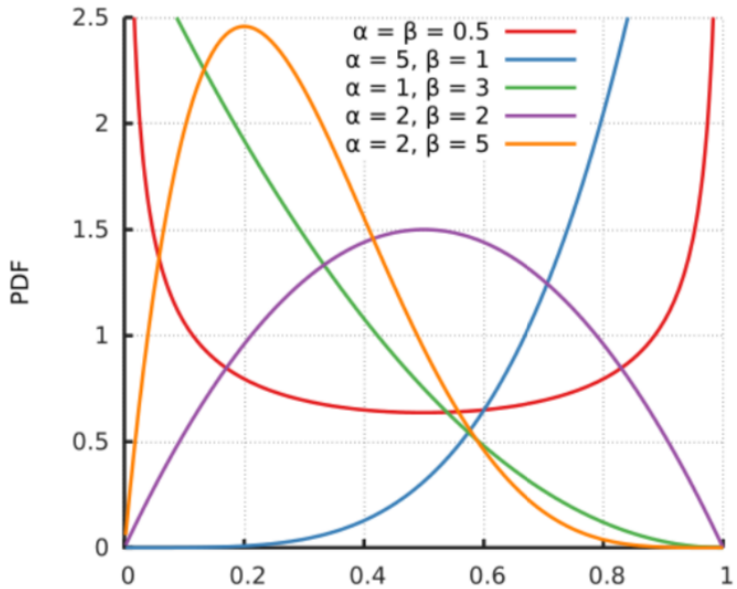
Note: "$data|p$" $\sim \text{Binomial}(14, p)$. Therefore:

$$P(data|p) = \binom{14}{10} p^{10}(1-p)^4.$$

What should we choose for $P(p)$?

The beta distribution $\text{Beta}(\alpha, \beta)$:

$$P(p; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \qquad (p \in (0,1)).$$



Example

- Suppose we decide to pick $p \sim \mathrm{Beta}(\alpha, \beta)$. Then:

$$\begin{aligned} P(p|data) &\propto P(data|p)P(p) \\ &= \binom{14}{10} p^{10}(1-p)^4 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \\ &\propto p^{10}(1-p)^4 p^{\alpha-1}(1-p)^{\beta-1} \\ &= p^{10+\alpha-1}(1-p)^{4+\beta-1}. \end{aligned}$$

Remark: We don't need to worry about the *normalization constant* since it is uniquely determined by the fact that $P(p|data)$ is a probability distribution.

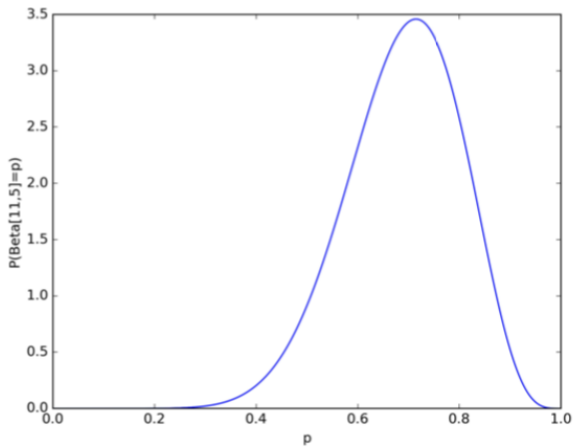
- Conclusion: $P(p|data) \sim \mathrm{Beta}(10+\alpha, 4+\beta)$.

Example

- How should we choose α, β?

  According to our *prior knowledge* of p.

- Suppose we have no prior knowledge: use a *flat* prior: $\alpha = \beta = 1$ (Uniform distribution).
- The resulting *posterior distribution* is $p|data \sim \text{Beta}(11, 5)$:



Our "knowledge" of p has now been updated using the observed data (or evidence).

Bayesian analysis

More generally: suppose we have a model for X that depends on some parameters θ. Then:

1. Choose a prior $P(\theta)$ for θ.
2. Compute the posterior distribution of θ using

$$p(\theta|X) \propto P(X|\theta) \cdot P(\theta).$$

Note: $\text{Posterior} = \text{Prior} \times \text{Likelihood}$.

Bayesian analysis

Advantages:

- Mimics the scientific method: formulate hypothesis, run experiment, update knowledge.
- Can incorporate prior information (e.g. the range of variables).
- Automatically provides uncertainty estimates.

Drawbacks:

- Not always obvious how to choose priors.
- Can be difficult to compute the posterior distribution.
- Can be computationally intensive to sample from the posterior distribution (when not available in closed form).

Central questions

1. How do we sample from the posterior distribution

$$P(\theta \,|data) \propto P(data|\theta).P(p)$$

assuming that we can evaluate $P(\theta \,|data)$ point-wise (up to a normalizing constant)

2. Consider a quantity of interest $y = F(p)$, how can we compute

$$E_{P(\theta|data)}[F(p)],$$

and

$$Var_{P(\theta|data)}[F(p)],$$

and quantify related probabilistic properties of the quantity of interest?

Monte Carlo methods

Monte Carlo methods

If we can sample independent samples $X_1, X_2 \ldots, X_n$ from

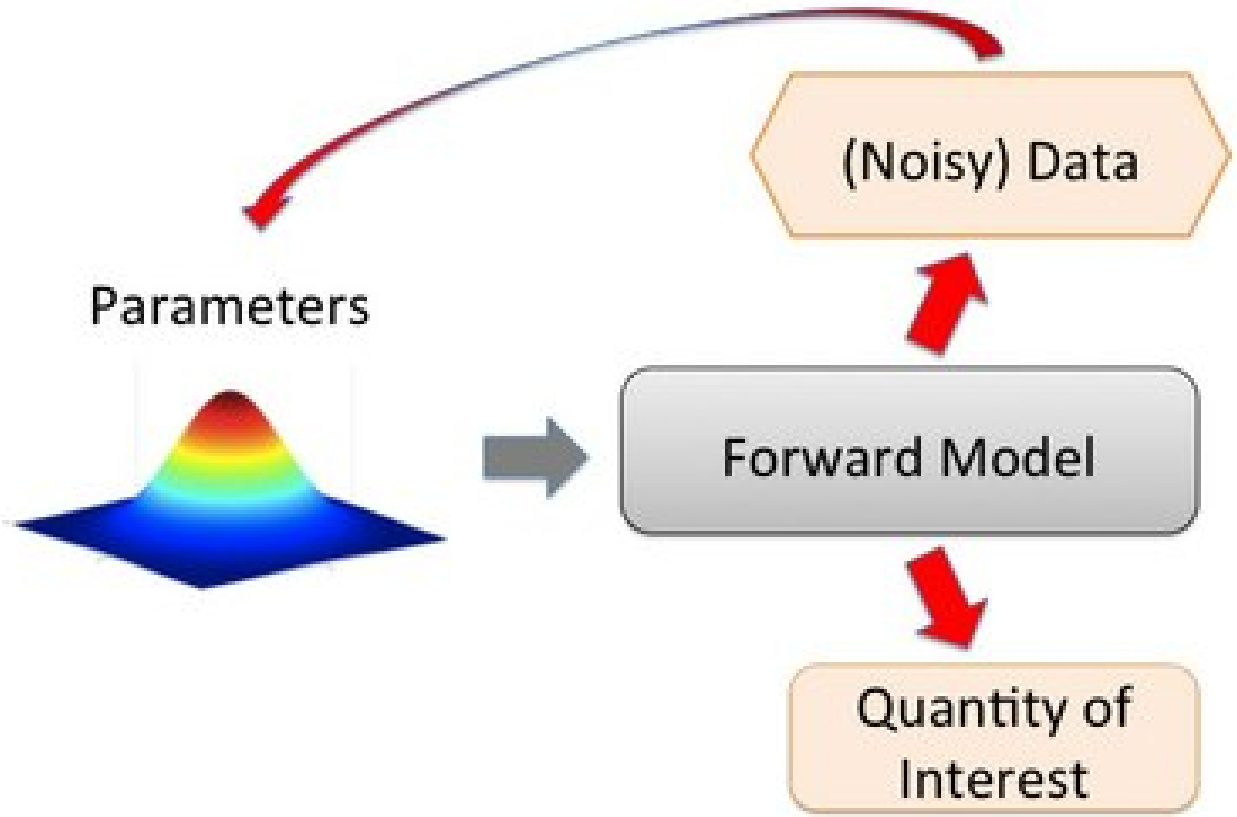
$$P(\theta \mid data) \propto P(data|\theta).P(p)$$

then for any function $F(\theta)$

$$\frac{F(X_1) + F(X_2) + \ldots F(X_n)}{n} \quad \rightarrow_{a.s.} \quad E_{P(\theta|data)}[F(p)],$$

as n goes to infinity.

Uncertainty quantification



Inverse transform sampling

- Let $F(x)$ by the cdf of some 1D distribution
- Claim: If U is a uniform random variable on $[0, 1]$, then $F^{-1}(U)$ has F as its cdf

Proof:

$$\begin{aligned}
&\Pr(F^{-1}(U) \leq x) \\
&= \Pr(U \leq F(x)) && (\text{applying } F, \text{ to both sides}) \\
&= F(x) && (\text{because } \Pr(U \leq y) = y, \text{ when U is uniform on} (0, 1))
\end{aligned}$$

Rejection sampling

Ideas:

- Suppose we want to sample from a distribution $p(x)$, which is known up to a proportional constant
- If we know another easy-to-sample proposal distribution $q(x)$ that satisfies
$$p(x) \leq Mq(x)$$
- then we can sample from $p(x)$ as follows:
  - sample $x \sim q(x)$, and $u \sim U([0,1])$ (the uniform distribution in $[0,1]$)
  - If
$$u < \frac{p(x)}{Mq(x)}$$
then accept the sample
  - otherwise, reject it

Rejection sampling

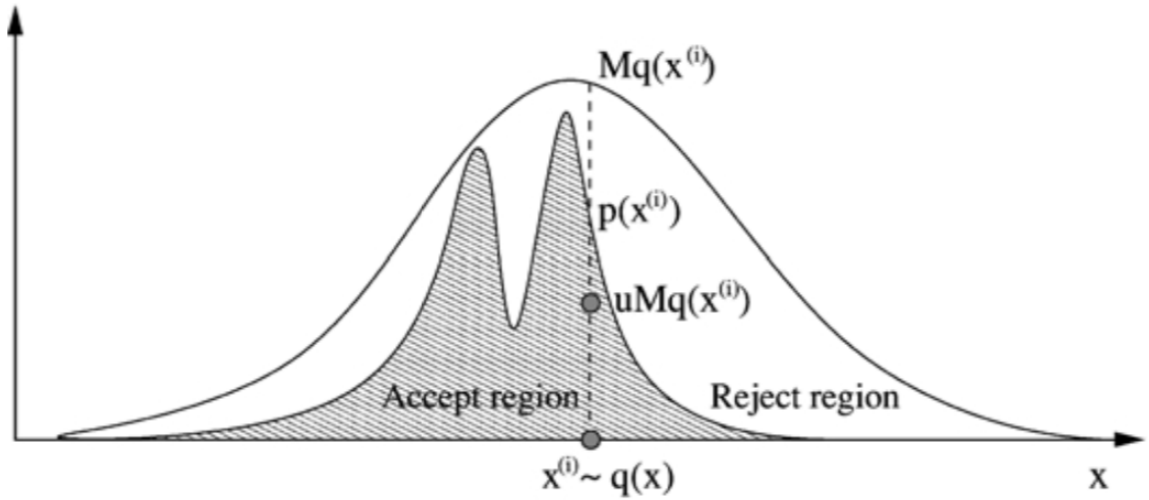


Figure 2. Rejection sampling: Sample a candidate $x^{(i)}$ and a uniform variable u. Accept the candidate sample if $uMq(x^{(i)}) < p(x^{(i)})$, otherwise reject it.

Markov chain Monte Carlo

MCMC

- Markov chain Monte Carlo (MCMC) methods are popular ways of sampling from complicated distributions (e.g. the posterior distribution of a complicated model).
- Idea:
  1. Construct a **Markov chain** with the desired distribution as its **stationary distribution** π.
  2. Burn (e.g. forget) a given number of samples from the Markov chain (while the chain converges to its stationary distribution).
  3. Generate a sample from the desired distribution (approximately).
- One generally then compute some *statistics* of the sample (e.g. mean, variance, mode, etc.).

Markov chain

- Let $S := \{s_1, s_2, \dots\}$ be a countable set.
- A (discrete time) **Markov chain** is a discrete stochastic process $\{X_n : n = 0, 1, \dots\}$ such that
  1. X_n is an S-valued random variable $\forall n \geq 0$.
  2. (Markov Property) For all $i, j, i_0, \dots, i_{n-1} \in S$, and all $n \geq 0$:

$$P(X_{n+1} = j | X_0 = i_0, \dots, X_{n-1} = i_{n-1}, X_n = i) = P(X_{n+1} = j | X_n = i).$$

Interpretation: Given the present X_n, the future X_{n+1} is independent of the past $(X_0, \dots, X_{n-1})$.

- The elements of S are called the **states** of the Markov chain.
- When $X_n = j$, we say that the process is in state j at time n.

Stationarity and transition probabilities

- A Markov chain is **homogeneous** (or **stationary**) if for all $n \geq 0$ and all $i, j \in S$,

$$P(X_{n+1} = j | X_n = i) = P(X_1 = j | X_0 = i) =: p(i, j).$$

In other words, the **transition probabilities** do not depend on time.

- We will only consider homogeneous chains in what follows.
- We denote by $P := (p(i, j))_{i,j \in S}$ the **transition matrix** of the chain.
- Note: P is a **stochastic matrix**, i.e.,

$$\forall i, j \in S,\ p(i, j) \geq 0, \quad \text{and} \quad \forall i \in S,\ \sum_{j \in S} p(i, j) = 1.$$

- Conversely, every stochastic matrix is the transition matrix of some homogeneous discrete time Markov chain.

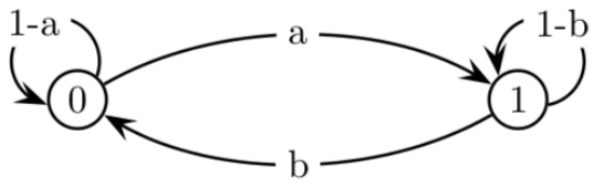
Example

Example 1: (Two-state Markov chain)

$$S = \{0, 1\}, \quad p(0,1) = a, \quad p(1,0) = b, \quad a, b \in [0,1]$$

$$P = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}.$$

We naturally represent P using a transition (or state) diagram:



Example

Example 2: (Simple random walk) Let $\xi_1, \xi_2, \xi_3, \ldots$ be iid random variables such that $\forall i \geq 1$,

$$\xi_i = \begin{cases} +1 & P(\xi_i = +1) = p \\ 0 & P(\xi_i = 0) = r \\ -1 & P(\xi_i = -1) = q \end{cases},$$

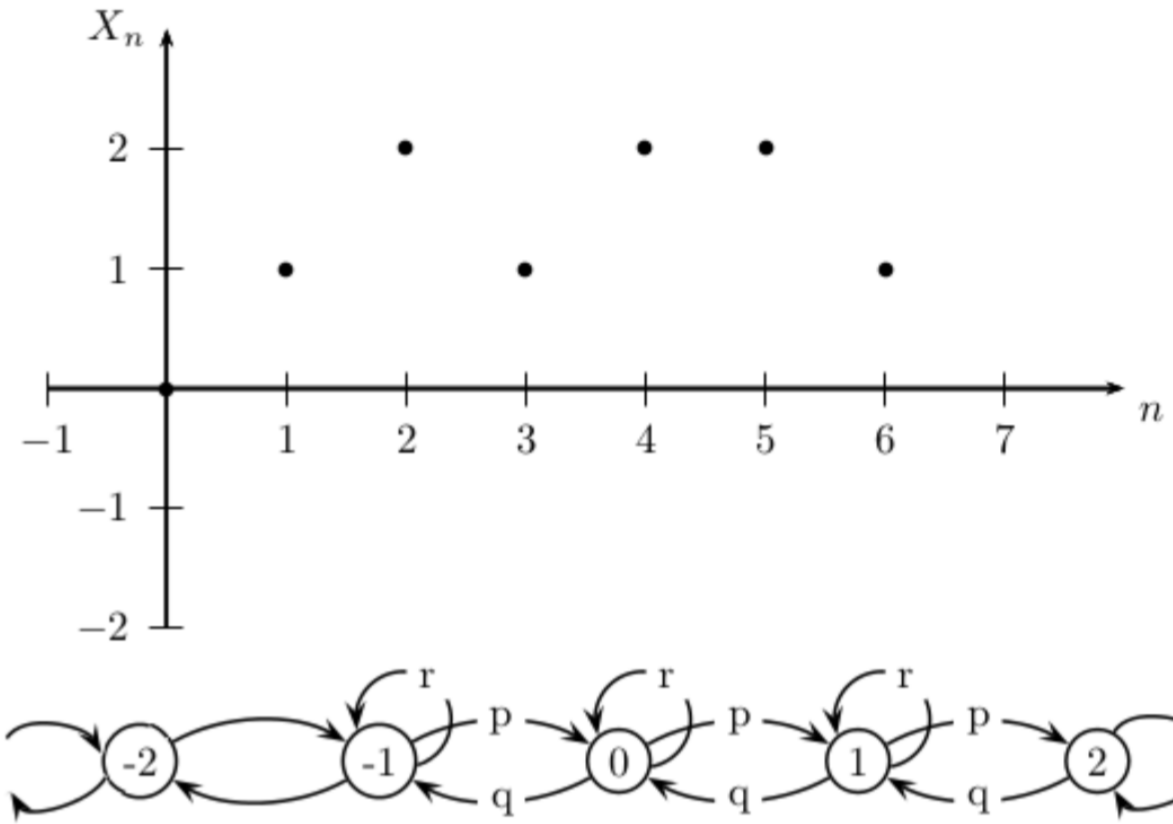
where $p + r + q = 1$, $p, r, q \geq 0$.

- Let X_0 be an integer valued random variable independent of the ξ_i's.
- Define $\forall n \geq 1$,

$$X_n = X_0 + \sum_{i=1}^{n} \xi_i.$$

Example

$$S = \{0, \pm 1, \pm 2, \dots\}.$$



n-step transitions

Let $\{X_n : n \geq 0\}$ be a Markov chain.

- We define the **initial distribution** of the chain by

$$\mu_0(i) := P(X_0 = i) \qquad (i \in S).$$

- All distributional properties of a (homogeneous) Markov Chain are determined by its initial distribution and transition probability matrix.
- For $n \geq 1$, we define the n-**step transition probability** $p^n(i,j)$ by

$$p^n(i,j) := P(X_n = j | X_0 = i) = P(X_{n+m} = j | X_m = i).$$

Also, define

$$p^0(i,j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}.$$

- We define the n-**step transition matrix** by

$$P^{(n)} := (p^n(i,j) : i,j \in S).$$

Chapman-Kolmogorov

Theorem: (The Chapman-Kolmogorov Equations) We have for all $m, n \geq 1$:

$$P^{(n+m)} = P^{(n)} \cdot P^{(m)}.$$

In particular, for all $n \geq 1$,

$$P^{(n)} = P \cdot P^{(n-1)} = \ldots = P^n.$$

Moral: n-step transition probabilities are computed using matrix multiplications.

Chapman-Kolmogorov

- Let $\mu_n := (\mu_n(i) : i \in S)$ denote the **distribution of** X_n:

$$\mu_n(i) := P(X_n = i).$$

Proposition: We have

$$\mu_{m+n} = \mu_m P^n, \quad \text{and} \quad \mu_n = \mu_0 P^n.$$

Moral: Distributional computations for Markov Chains are just matrix multiplications.